NORDUGRID-MANUAL-20
25/4/2013

ARC 12.05 Computing Element

System Administrator Guide

F. Paganelli, Zs. Nagy, O. Smirnova,
and various contributions from all ARC developers

Contents

Chapter 1

Overview

The *ARC middleware* [**?**] by *NorduGrid* [**?**] is a software solution that uses grid technologies to enable *sharing* and *federation* of *computing* and *storage* resources distributed across different administrative and application domains. ARC is used to create grid infrastructures of various scope and complexity, from campus to national grids.

This document gives a detailed overview of the ARC *Computing Element* (CE), along with step-by-step installation and configuration instructions and a full reference of the configuration commands.

1.1 The grid

An ARC-based grid aggregates computing and storage resources, making them accessible through standard interfaces, and using a common information system to optimize access.

Client tools can query this information system to see what kind of resources are available, match user's tasks to best available resources, submit computing jobs, which are smaller or bigger tasks (scripts and/or binaries, often processing defined input data) to run on computing nodes in the grid, they can access files on and upload results to storage resources.

For users, all this complexity is hidden: they simply formulate their tasks in a special language and send them to the grid, not even knowing which computing or storage resources are out there. ARC takes care of the rest.

While submitting jobs, users must specify requirements for each job, namely, what software should it execute, what data to process, what kind of software environment it needs on the computing node, how much memory, how strong CPU, etc. — these are specified in the formal job description. They can use various client tools, like the native command-line interface supplied along with the ARC middleware [**?**], GUI tools, web portals or specialized clients as part of a bigger software tool. All users must be authenticated by grid services using *X.509 certificate*s signed by trusted Certificate Authorities. ARC also uses short-lived *proxy certificates* to delegate users' rights to various activities performed by Grid services on their behalf, such as job execution or data transfer. Authentication alone is not sufficient: users must also be authorized to perform such activities. Typically, users form groups (called *Virtual Organization*s, VOs) to ease to process of getting authorized on the several computing resources.

In order to handle all the computing resources in a uniform way, there is a need for a layer ("middleware") between the client tools and the resources: the *Computing Element* (CE). This document describes how to use the CE functionality of the ARC middleware to make a computing resource accessible for grid users.

1.2 The ARC services

Grid computing has three big areas: *computation*, *storage* and *information*. The server side of the ARC middleware provides services for all three main areas:

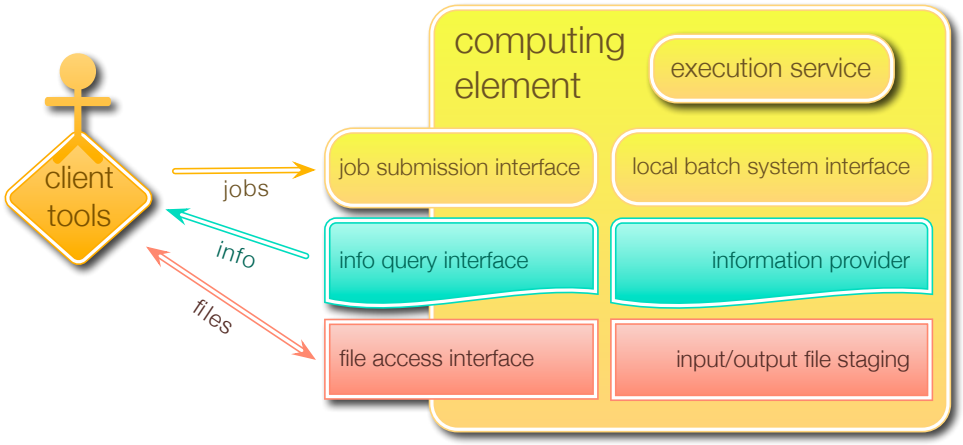


Figure 1.1: The interfaces and internal components of a generic grid computing element

- The **Computing Element (CE)**. By installing the ARC *Computing Element* (CE), a computing resource (usually, computing clusters managed by a batch system—*LRMS*—or a standalone workstation) will gain standard grid interfaces, through which users (authenticated using their *X.509 certificates*) can get information about the resource, submit, query and manage computing jobs with the help of client tools. The computing resource will also gain a capability to register itself to several different grid information system such that client tools would discover it.
- The **Storage Element (SE)**. The ARC *GridFTP Server* [**?**] besides being an important part of the ARC Computing Element, can also be installed as a standalone storage solution.
- The **Indexing Service (EGIIS)**. The ARC *Enhanced Grid Information Indexing Service* (EGIIS) is capable of collecting registrations from computing elements and storage elements equipped with the ARC Resource Information Service (ARIS) and providing these resource pointers to the client tools. There are several EGIIS instances deployed all around the world. New resources usually register themselves to one or more of the existing indexes.

These three functionalities are implemented by one or more ARC services, which can be installed separately in a standalone manner, or all of them can reside on the same machine. *This document only describes the ARC Computing Element (CE).* For the description of the standalone GridFTP Storage Element, please refer to the The NorduGrid GridFTP Server document [**?**].

There is a very important fourth area: the client side. The **ARC command line clients** [**?**] are able to fully interact with the A-REX or other computing elements, they support several data transfer protocols to be able to upload and download files from all kinds of storage resources. They are querying the available computing resources from the information system, doing brokering based on the requirements specified in the job description (languages supported: XRSL [**?**], JSDL [**?**] and JDL [**?**]), they are able to query the status of jobs and manage their lifecycle, and to handle all aspects of the secure communication including delegation of the user's credentials.

1.3 The functionality of the ARC Computing Element

Figure 1.1 shows the interfaces and the internal components of a generic grid computing element. An ARC **Computing Element (CE)** has these interfaces and components, and with them it is capable of the following:

- to advertise (register) itself in an information system to make the clients tools know about its location and capabilities
- to accept job execution requests coming through the *job submission interface* and to process the jobs (written in standard job description languages) handled by the *execution service*

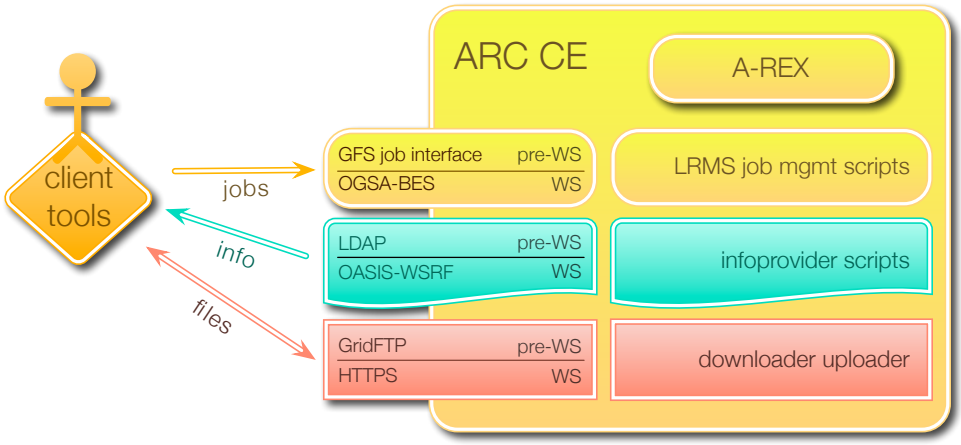


Figure 1.2: The interfaces and components of the ARC Computing Element

- to accept the files requested by the jobs from the user through the *file access interface* or to download them from remote storages (*input file staging*) and to avoid downloading the same files over and over again by caching them
- to forward the jobs to the *local resource management system (LRMS)* (such as Condor [**?**], Torque [**?**], OpenPBS [**?**], Sun Grid Engine [**?**], etc.), which will schedule and execute them on the computing nodes in the local cluster
- to monitor the status of the jobs by running the *information provider* scripts and make this information available through the *information query interface*.
- to make the results (output files) of the jobs accessible through the *file access interface* or upload them to a remote storage *output file staging*

1.4 The A-REX, the execution service

The most important component of the ARC Computing Element is the **A-REX** (ARC Resource-coupled EXecution service). The A-REX accepts requests containing a description of generic computational jobs and executing it in the underlying local batch system. It takes care of the pre- and post-processing of the jobs: *staging in* (downloading) files containing input data or program modules from a wide range of sources and storing or *staging out* (uploading) the output results.

The ARC Computing Element with the help of A-REX and some other services provides two distinct set of interfaces: the pre-web service interfaces, which are based on LDAP and GridFTP, and are currently widely deployed and in production; and the web service interfaces, which are based on grid standards, are also well-tested and production-quality but not yet widely used. Figure 1.2 shows the interfaces and also the other components.

1.4.1 The pre-web service interfaces

The pre-web service *job submission interface* uses the GridFTP protocol in a special way. It is provided by a separate component, the ARC *GridFTP Server* (GFS) has a *job plugin* which accepts job descriptions in the *XRSL* job description language. The A-REX works together with the GridFTP Server to get notified about new jobs.

The pre-web service *information query interface* of the ARC CE is an LDAP/BDII based interface, which is provided by a separate component, called the *ARIS* (the ARC Resource Information System).

The pre-web service *file access interface* uses the GridFTP protocol, and is served by the same ARC *GridFTP Server* (GFS) which provides the job submission interface too.

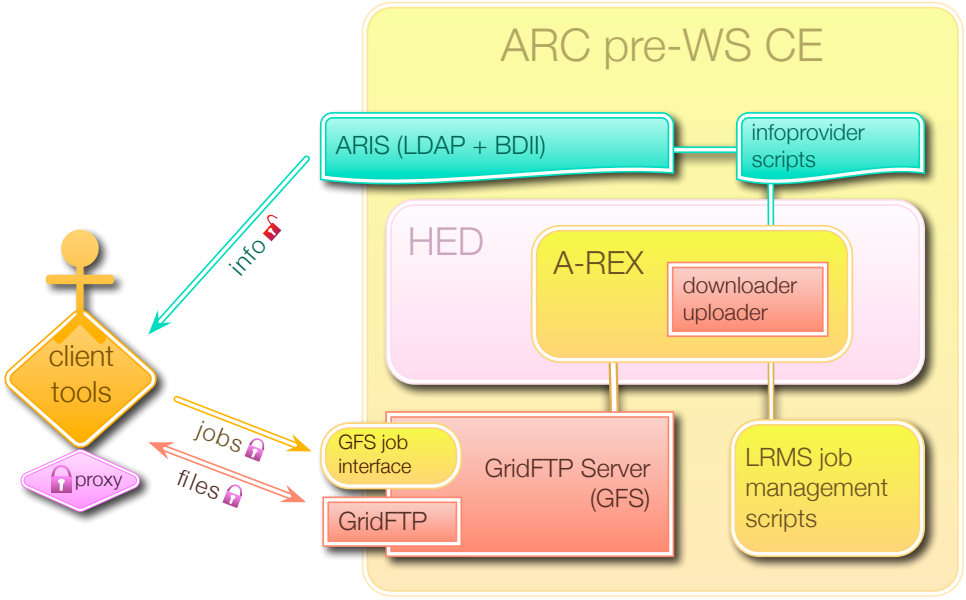


Figure 1.3: The services and components of the pre-web service ARC CE

The A-REX service itself has no direct interface to the clients in the pre-web service case, it communicates through the GridFTP Server (GFS). Figure 1.3 shows the services and the components of the pre-web service ARC CE.

1.4.2 The web service interfaces

The web service *job submission interface* of the ARC CE is provided by the A-REX itself, and it is a standard-based interface: an enhancement of the OGSA Basic Execution Service recommendation [**?**].

The web service *information query interface* of the ARC CE is also provided by the A-REX itself, and it is also a standard-based interface, called *LIDI* (Local Information Description Interface), which is an implementation of the OASIS Web Services Resource Properties specification [**?**].

The *file access interface* is technically not a web service, but it is the well-known HTTPS interface provided by the A-REX itself.

In the web service case, all the interfaces are provided by the A-REX itself, there is no need of separate services. Figure 1.4 shows the components of the web service ARC CE.

The web service and the pre-web service interfaces are capable to work together: an ARC CE can provide both interfaces at the same time.

1.5 Security on the Grid

Security on the grid is achieved using *X.509 certificates*. Any grid service needs to have a certificate issued by a trusted *Certificate Authority (CA)*. A single machine, like a front-end running a CE, is identified by a **host certificate**. A single user accessing the grid is identified by a **user certificate** also issued by a trusted CA.

Grid CAs are often established in each country, though there are also CAs issuing certificates for specific organizations (like CERN), or for several countries (like TERENA). Each CA has its own certification policies and procedures: to access/setup a grid service, one has to contact the relevant Certificate Authority in order to obtain the needed certificates.

When a user wants to access the grid, the client tools generate a short-lived *proxy certificate* to delegate user's rights to jobs or other activities performed by grid services on the user's behalf.

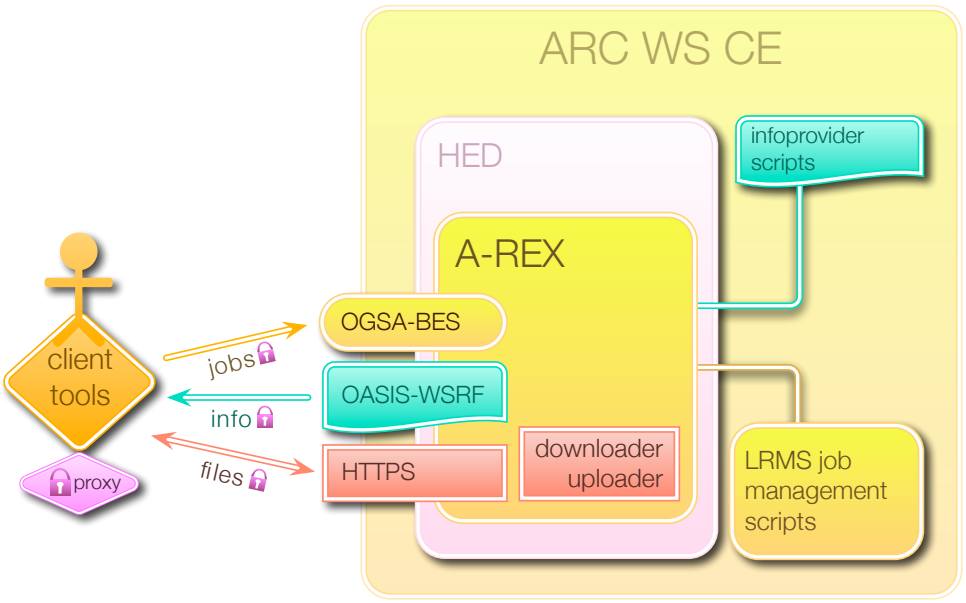


Figure 1.4: The components of the web service ARC CE

In order for the server to authenticate the client, the certificate of the CA issuing the user's certificate has to be installed on the server machine. In the same manner in order for the client to authenticate the server, the certificate of the CA issuing the host's certificate should be installed on the client machine.

On the server side it is the responsibility of the system administrator to decide which authorities to trust, by installing each authority's certificate. On the client side, the user decides which CA certificates she installs. The user cannot access a grid resource, if the issuer CA certificate of the host is not installed.

Figure 1.5 shows an overview of the required keys and certificates, and also the process of creating a client proxy certificate using the user's credentials, and optionally collecting more information about the Virtual Organization (VO) the user belongs by connecting to a Virtual Organization Membership Service (VOMS).

1.6 Handling jobs

A job is described as a set of input files (which may include executables), a main executable and a set of output files. The job's life cycle (its session) starts with the arrival of the job description to the Computing Element (CE), next comes the gathering of the input files, then follows the execution of the job, then the handling of the output files and finally job ends with the removal of the session contents by either the user or after a specified amount of days by the CE.

Each job gets a directory on the CE called the *session directory* (SD). Input files are gathered in the SD. The job may also produce new data files in the SD. The A-REX does not guarantee the availability of any other places accessible by the job other than SD (unless such a place is part of a requested *Runtime Environment*, see section **1.7**, *Application software in ARC: The RunTime Environments*).

Each job gets a globally unique identifier (*jobid*). This *jobid* is effectively a URL, and can be used to access the session directory (to list, download and even upload files into the SD) from outside, either through the HTTP(S) interface or through the GridFTP Server.

1.6.1 A sample job processing flow

The jobs in the ARC Computing Element usually go through these steps:

1. The client (such as the ARC command line tools [**?**]) connects to the *job submission interface* (either to the web service interface of A-REX or to the GridFTP Server).

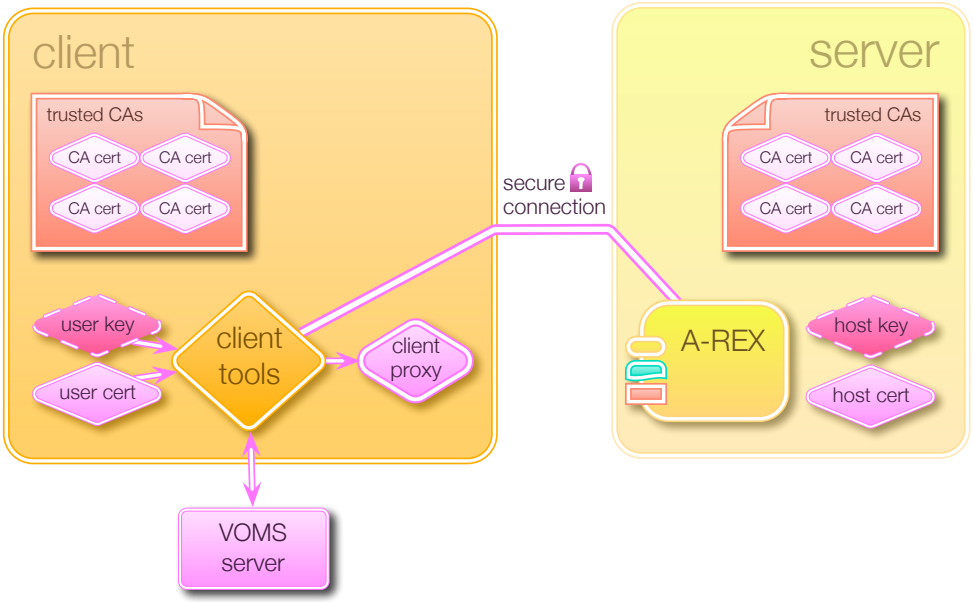


Figure 1.5: Certificates on the client side and on the server side. The client tools create a proxy certificate using the user's credentials, and optionally collect more information about the Virtual Organization (VO) the user belongs by connecting to a Virtual Organization Membership Service (VOMS).

2. Using the well-established processes of the X.509 Public-Key Infrastructure [**?**], the client and the server both authenticate each other, based on the trusted CA credentials which were previously installed on both ends.

3. The A-REX authorizes the user based on configurable rules, and maps the grid identity to a local username which should be available also on all the worker nodes.

4. The client tool delegates user's credentials to the A-REX to enable it to act on behalf of the user when transferring files. (See Figure 1.6.)

5. A job description written in one of the supported languages (XRSL [**?**] or JSDL [**?**]) is sent from the client to the server. (The client itself understands the JDL [**?**] language also, and it translates it to either XRSL or JSDL for the A-REX to understand.)

6. The job is accepted and a directory (the *session directory, SD*) is created which will be the home of the session. Metadata about the job is written into the *control directory* of the A-REX.

7. The client tool receives the location of the session directory (SD), and if there are local input files, those will be uploaded into the SD through the *file access interface* (either through the HTTP(S) interface of the A-REX, or through the GridFTP Server).

8. If the job description specifies input files on remote locations, the A-REX fetches the needed files and puts them into the SD. If the caching is enabled, the A-REX checks first if the file was already downloaded recently, and uses the cached version if possible.

9. When all the files prescribed in the job description are present (either uploaded by the client tool or downloaded by the A-REX), a suitable job script is created for and submitted to the configured batch system (LRMS).

10. During this time, the SD of the job is continuously accessible by the client tool, thus any intermediate result can be checked.

11. The *information provider* scripts periodically monitor the job status, updating the information in the control directory.

12. When the job in the LRMS is finished, the A-REX uploads, keeps or removes the resulted output files according to the job description.

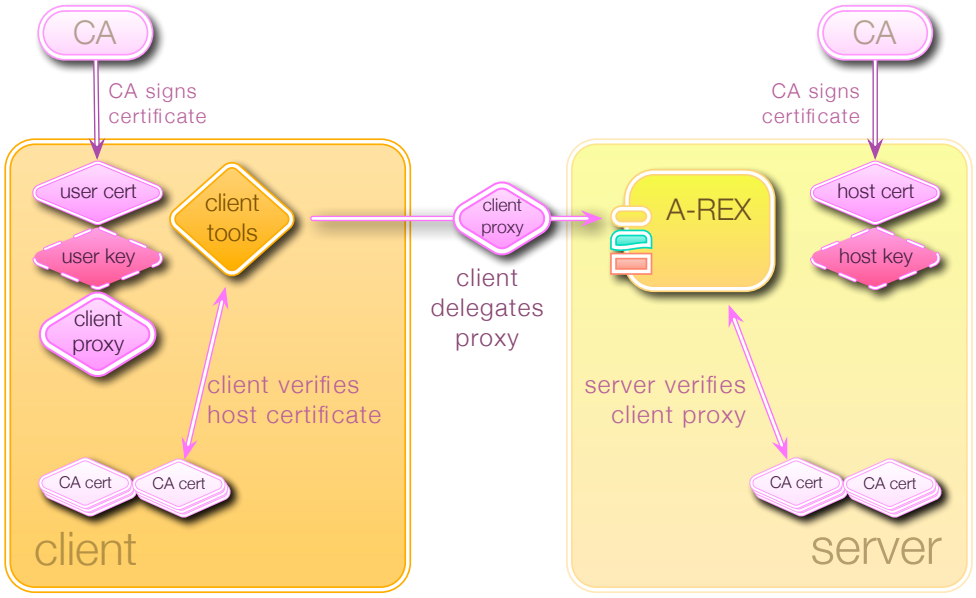


Figure 1.6: The client delegates the client proxy to the Computing Element, while both parties verifies that the credentials are signed by a trusted Certificate Authority (CA)

13. The client tool may also download the output files through the *file access interface*, and remove the job from the Computing Element (CE).

During the whole lifetime of the job, its status can be queried through the *information query interface* (either through the LDAP interface or through the LIDI web service interface).

Figure 1.7 and Figure 1.8 shows the staging process.

1.7 Application software in ARC: The RunTime Environments

Code development in science but also in specific knowledge areas always demands specific software, libraries and tools to be used. A common task when offering computational power is to recreate such environments for each specific knowledge domain.

To provide such software environments and tools in the grid world, ARC enforces the concept of the RunTime Environment (RTE).

ARC RunTime Environments (RTEs) provide user interfaces to application software and other resources in a way that is independent of the details of the local installation of the application and computing platform (OS, hardware, etc.).

It addresses setups typically required by large research groups or user bases, dealing with a common set of software.

The actual implementation of particular RTE may differ from site to site as necessary. However, it should be designed so that resource providers with different accounting, licence or other site-specific implementation details can advertise the same application interface (RE) for all users. It is always up to the local system administrators to take a decision whether to install and enable a particular runtime environment or not.

A RTE, as conceptualized in `http://pulse.fgi.csc.fi/gridrer/htdocs/intro.phtml`, is defined by two items:

1. RTE Homepage
   - describes the users' application interface
   - provides application installation instructions for the site administrators

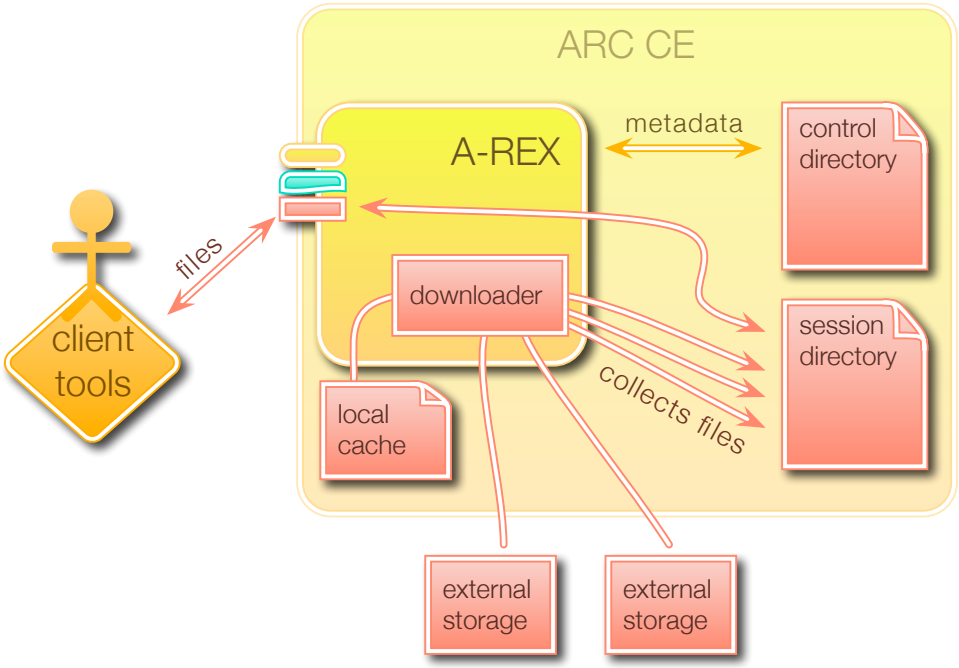


Figure 1.7: The process of staging in the input files of a job

- links to the application support information

2. RTE itself
   - is a shell environment initialization script
   - is installed on computing resources
   - initializes variables that point to the application software

Let's have an example from the user perspective:

A user has a script written in python 2.6 that she wishes to execute in some remote computing node in Grid. She requests `PYTHON-2.6` Runtime Environment in the job-description file and passes that file to the command arcsub.

Upon submission, arcsub parses the job description, notices the RTE request and submits the job only to sites advertising that RTE. After job submission A-REX on the chosen site initializes the environment in the computing node before local execution of the job. It initializes the environment so that python interpreter and standard libraries are in the PATH and executable/readable by the user as described in the RTE Homepage.

What does this give to the users:

- easier access to a large software resource base
- identical interface to applications independent of the computing platform

What does this do for resource providers and application developers:

- opens the application to a large user base
- reduces overlapping work with application support

More information on how to setup RTEs can be found in Section **4.5.2**, *Runtime Environments*.

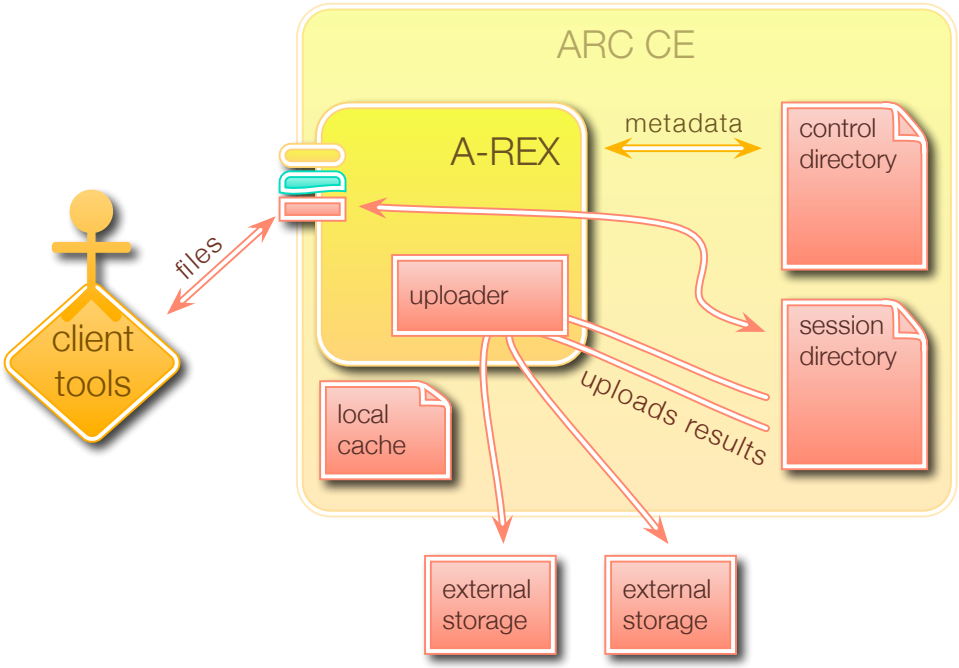


Figure 1.8: The process of staging out the output files of a job

1.8 The local information

In order to create a Grid infrastructure using ARC-enabled computing resources, information description and aggregation services need to be deployed. ARIS is coupled to a computing resource and collects information about it. EGIIS keeps a list of ARIS instances, and eventually, of other EGIIS instances lower down in hierarchy. Top-level EGIIS instances thus serve as an entry point to the Grid, allowing to discover all the resources.

While ARIS is *coupled* to a resource, EGIIS is an *independent* service. A typical Grid resource owner always has to deploy ARIS*. EGIIS servers, on the other hand, are normally deployed by the overall Grid infrastructure operators.

A system effectively created by ARIS and EGIIS services is called the *ARC Information System*. Being based on OpenLDAP [**?**], it can be accessed in a standard manner by a variety of LDAP clients, giving a full overview of the infrastructure resources.

ARIS instances are responsible for resource (e.g. computing or storage) description and characterization. The local information is generated on the resource, and it can be cached. Upon client requests it is presented via LDAP interface.

1.8.1 Overview of ARC LDAP Infosys schemas

ARC information system currently can present information in three different formats, or schemas. These can be enabled simultaneously. The schemas are:

1. NorduGrid-ARC schema – this is the NorduGrid default schema, described in detail in this document. It was inspired by Globus MDS, but has been improved a lot over the years and due to incompatible changes was moved into the NorduGrid LDAP namespace. In order for standard NorduGrid clients to submit jobs to a resource, this schema must be published.

2. Glue 1.2 – This is the schema that is used by gLite [**?**]. Currently, gLite supports Glue 1.3 schema, but Glue 1.2 is sufficient to be compatible. If ARC is configured to publish information in the Glue 1.2

*Without ARIS, a resource is still functional, but is not a Grid resource

format, it will first produce data in the NorduGrid-ARC schema which will then be translated to Glue 1.2. To allow gLite clients to submit to a resource, this schema must be published. Please note, that the gLite information system must also be hooked into the resource in order for this interoperability to work.

3. Glue 2.0 – This is the common schema for the EMI [**?**]. This schema can be published both through LDAP and XML interfaces of the ARC Compute Element.

ARIS is the information service that is installed on the ARC Compute Element. It publishes via LDAP interface information about the local computing cluster, like: operating system, amount of main memory, computer architecture, information about running and finished jobs, users allowed to run and trusted certificate authorities. The information can be published in either NorduGrid-ARC schema, Glue 1.2 schema or Glue 2.0 schema.

The dynamic resource state information is generated on the resource. Small and efficient programs, called information providers, are used to collect local state information from the batch system, from the local Grid layer (e.g. A-REX or GridFTP server) or from the local operating system (e.g. information available in the `/proc` area). Currently, ARC is capable interfacing to the following batch systems (or local resource management system LRMS in the ARC terminology): UNIX fork, the PBS-family (OpenPBS, PBS-Pro, Torque), Condor, Sun Grid Engine, IBM LoadLeveler and SLURM.

The output of the information providers (generated in LDIF format) is used to populate the local LDAP tree. This OpenLDAP back-end implements two things: it is capable caching the providers output and upon client query request it triggers the information providers unless the data is already available in its cache. The caching feature of the OpenLDAP back-end provides protection against overloading the local resource by continuously triggering the information providers.

1.9 LRMS, Queues and execution targets

Usually the A-REX is installed on top of an existing local resource management system (LRMS). The A-REX has to interfaced to the LRMS in order to be able to submit jobs and query their information.

The A-REX assumes that the LRMS has one or more **queue**s, which is a couple of (usually homogeneous) worker nodes grouped together. These queues should not overlap. The different LRMSes have different concepts of queues (or have no queues at all). Nevertheless, in the A-REX configuration, the machines of the LRMS should be mapped to A-REX queues. The details can be found in Section **4.4.2**, *Connecting to the LRMS*.

The client side job submission tools query the information system for possible places to submit the jobs, where each queue on a CE is represented as an **execution target**, and treated separately.

Chapter 2

Requirements

To properly configure an ARC CE the following prerequisites are needed:

- **Administrators installing ARC CE must have access to network firewall configuration**: Several ports will need to be open for the ARC services to work (see **4**, *Configuration* and **4.1.4**, *Firewalls*
- **Time Synchronization** of the system that will run an ARC CE must be setup, by using the NTP protocol [**?**] or similar. The grid relies on synchronization for the jobs to be correctly submitted and for the security infrastructure to work properly.

The following is optional but suggested to be on the machines running an ARC CE:

- **A networked filesystem** such as NFS or similar, to connect storage and share job data between the ARC middleware and the LRMS system behind it.

2.1 Software Requirements

ARC services can be built mainly for GNU/Linux and Unix systems.

Table 2.1 shows the current officially supported ones.

Operating System	Version/Distribution	Supported Architectures
GNU/Linux	Scientific Linux 5.5+	i386, x86_64
	RedHat 5+	i386, x86_64
	Debian 6+	i386, x86_64
	Ubuntu 10.04+	i386, x86_64

Table 2.1: Supported operating systems

For a detailed list of the software libraries needed to compile and install ARC services, please refer to the README included in the source tarball. See Chapter **3**, *Installation* for details.

2.2 Hardware Requirements

The NorduGrid middleware does not impose heavy requirements on hardware. The choice is only bound to the computational needs of your organization.

Table 2.2 shows the minumum requirements.

Architecture	32 or 64 bits
CPU families	$\geq$ i386 , PowerPC
CPU Speed	$\geq$ 300 MHz
Memory Size	$\geq$ 128MB
Disk space for binaries	$\leq$ 30MB
Disk space including development files	160MB
Disk space including external software (such as Globus Toolkit 5)	+10MB
Network connectivity	a public IP on the front-end cluster is strongly encouraged. Worker nodes can be on a private or local network.

Table 2.2: Hardware Requirements

2.3 Certificates

To run an ARC CE and have it servicing the grid, a **host certificate** provided by a Certificate Authority (CA) is needed.

A request for such a certificate must be sent to the National Grid Infrastructure organization or to any local organization entitled to provide grid services.

The CA certificate is needed as well, this is public and can be usually obtained from either the CA itself, of fetched from the EMI repository, IGTF repository, NorduGrid yum/apt repositories, or from the NorduGrid Downloads area. These are needed to verify that the service and the users connecting to it have valid credentials, to perform mutual authentication.

If this is the **first time** the reader sets up an ARC CE, we suggest to obtain temporary test certificates for hosts, users and a temporary CA via the InstantCA service:

`https://arc-emi.grid.upjs.sk/instantCA/instantCA`

Such certificates cannot be used in production environments and can only be used for testing purposes.
Once the system administrator feels comfortable with an ARC CE setup, InstantCA certificates can be substituted with actual ones from trusted production CAs.

Installation of certificates is discussed in Section **3.3**, *Installation of certificates*.

Chapter 3

Installation

3.1 Installation for commom GNU/Linux Distributions

The preferred installation method for ARC middleware is by installing packages from **repositories**. The currently supported distributions are those based on **YUM-RPM** (Red Hat, CentOS, Fedora, Scientific Linux) and those based on **APT** (Debian, Ubuntu).

The packaging systems will automatically download additional libraries and dependencies for all the ARC middleware components to work properly. You can choose to install single packages one by one and add functionalities in a step-by-step fashion. Please refer to table 3.1 if you plan to do so.

ARC provides also meta-packages that are shortcuts to install a group of packages that provide a single functionality. It is strongly recommended to use this functionality for a quick start.

3.1.1 Setting up the repositories

The current repository is the official NorduGrid one. To configure NorduGrid repositories please follow the up-to-date instructions at:

`http://download.nordugrid.org/repos.html`

If ARC CE is to be used together with other European grid products, for example to join European scientific experiments such as ATLAS or ALICE, then the suggested repository is the EMI repository.

The EMI consortia provides also official production level customer support for distributions such as Scientific Linux 5.5 and Debian 6 and above, so it is strongly recommended to install from EMI if you are planning to use an ARC CE on these systems.

To install such repositories, please follow the instructions at EMI official website at this link:

`http://emisoft.web.cern.ch/emisoft/index.html`

3.1.2 Performing the installation

To perform the installation, follow these steps:

1. Configure a repository (see above for details)
2. Install the ARC CE using meta-packages: issue the following command as root:
   For RPM-Based distros:

```
yum install nordugrid-arc-compute-element
```

For APT-Based distros:

```
apt-get install nordugrid-arc-compute-element
```

This will install the packages marked with * in table 3.1.

3. (optional) if you want to customize your setup with individual packages, issue:
For RPM-Based distros:

```
yum install <packagename>
```

For APT-Based distros:

```
apt-get install <packagename>
```

Package	Content
All	
nordugrid-arc*!	All components
ARC CE	
nordugrid-arc-arex*!	ARC Remote EXecution service
nordugrid-arc-hed*!	ARC Hosting Environment Daemon
nordugrid-arc-plugins-needed*!	ARC base plugins
nordugrid-arc-gridftpd*!	ARC GridFTP server
nordugrid-arc-plugins-globus*	ARC Globus plugins
nordugrid-arc-python	ARC Python wrapper
nordugrid-arc-java	ARC Java wrapper
nordugrid-arc-cache-service	ARC cache service
nordugrid-arc-aris*+	ARC LDAP information service
ARC SE	
nordugrid-arc-gridftpd	ARC GridFTP server
ARC IS	
nordugrid-arc-egiis+!	ARC EGIIS service
Security	
nordugrid-arc-gridmap-utils*!	NorduGrid authorization tools
nordugrid-arc-ca-utils*!	NorduGrid authentication tools
Monitoring	
nordugrid-arc-ldap-monitor	ARC LDAP monitor service
nordugrid-arc-ws-monitor	ARC WS monitor service
Documentation	
nordugrid-arc-doc	ARC documentation

Figure 3.1: **ARC packages**: the table shows a brief description of each package and the components they belong to. Packages marked with "!" are mandatory to have a working functionality. Packages marked with "*" are automatically installed by ARC-CE nordugrid-arc-compute-element metapackage, packages marked with "+" are automatically installed by ARC Infosys nordugrid-arc-information-index metapackage

3.2 Installation for other systems and distributions

Packages are not provided for platforms other than GNU/Linux, so for the moment being the only way of installing ARC services is by compiling from source. Please refer to the README file* in the source code repository for more details.

3.3 Installation of certificates

A description of what certificates are and why they are needed can be found in Section **1.5**, *Security on the Grid*.

Information about reading the contents of the certificates, changing their formats and more can be found in the ARC certificate mini how-to document†.

In case ARC was installed using meta-packages (see Chapter **3**, *Installation*) all the required CAs are already installed and a script will automatically update them together with system updates.

If you want to install or remove specific CAs, NorduGrid repositories contain packaged CAs for ease of installation. By installing these packages, all the CA credentials will get updated by system updates. These packages are named in this format:

```
ca_<CA name>
```

Example:

```
ca_nordugrid
```

You can install them as you would install any package by APT or YUM.

In case your resource is in a Nordic country (Denmark, Finland, Norway, Iceland or Sweden), install the `certrequest-config` package from the NorduGrid Downloads area. It is also in the NorduGrid repositories with name `ca-nordugrid-certrequest-config`. This contains the default configuration for generating certificate requests for Nordic-based services and users. If you are located elsewhere, contact your local CA for details.

For example, in Nordic countries, generate a host certificate request with

```
grid-cert-request -host <my.host.fqdn>
```

and a LDAP certificate request with

```
grid-cert-request -service ldap -host <my.host.fqdn>
```

and send the request(s) to the NorduGrid CA for signing.

3.3.1 Installing host certificates

Once an host certificate is obtained from a CA, it has to be installed for the CE to use it.

When generating a certificate, two files will be created: a certificate file (public), typically `hostcert.pem`; and a key file (private), typically `hostkey.pem`.

Installation is as follows:

1. Copy the two files `hostcert.pem` and `hostkey.pem` into the standard ARC location: `/etc/grid-security`.
2. Both files must be owned by root.

*http://svn.nordugrid.org/trac/nordugrid/browser/arc1/trunk/README
†http://www.nordugrid.org/documents/certificate_howto.html

3. The private key (`hostkey.pem`) must be readable only by root.
4. The two files MUST NOT have executable permissions.
5. The key file MUST NOT be password protected. This is especially important if a tool other than `grid-cert-request` was used.

If the ARC services will be run as a different user than root, then these files should be owned and accessible by this other user.

3.3.2 Installing custom CA certificates

If you're planning to install custom certificates such as the one provided by InstantCA (See **2.3**, *Certificates*) then the files must usually be copied into the `/etc/grid-security/certificates/` directory.

3.3.3 Authentication Policy

The credential-level authentication policy is just a decision on which certificates the CE will accept. Only those users whose CAs are installed will be able to connect to the CE. (This does not mean they will be authorized to submit jobs, but at least they can establish the connection.) It is strongly advised to obtain a certificate from each CA by contacting it. To simplify this task, the NorduGrid Downloads area has a non-authoritative collection of CA credentials approved by EUGridPMA. As soon as you decide on the list of trusted certificate authorities, you simply download and install the packages containing their public keys and certificates. Before installing any CA package, you are advised to check the credibility of the CA and verify its policy!

> **Example** If your host certificate is issued by the NorduGrid CA, and your user has a certificate issued by the Estonian CA, and she is going to transfer files between your site and Slovakia, you need the NorduGrid, Estonian and Slovak CA credentials.

3.3.4 Revocation lists

The Certificate Authorities are responsible for maintaining lists of revoked personal and service certificates, known as CRLs (Certificate Revocation Lists). It is the CE administrator responsibility to check the CRLs regularly and deny access to Grid users presenting a revoked certificate. Outdated CRLs will render your site unusable. A tool called `fetch-crl` exists to get the latest CRLs, which can be installed from the `fetch-crl` package which is included with the `nordugrid-arc-compute-element` meta-package and also available from major repositories (this package is not provided by NorduGrid). The tool is intended to run as a cron job. There are 2 init scripts available:

```
/etc/init.d/fetch-crl-boot
/etc/init.d/fetch-crl-cron
```

The `fetch-crl-boot` script enables CRL downloads during boot while `fetch-crl-cron` enables scheduled download of CRLs. Detailed configuration can be tuned via `/etc/fetch-crl.conf`.
More information can be found here: `http://vdt.cs.wisc.edu/components/fetch-crl.html`.
Automatic startup of these services are distribution dependent and the administrator should take care of running these scripts by the means offered by their OS distribution.

3.3.5 Authorization policy

The authorization policy is a decision on which grid users or groups of grid users (Virtual Organizations) are allowed to use a resource. Configuration of this will be discussed in the following sections: Section **4.4.1**, *Access control: users, groups, VOs* and Section **6.10**, *Structure of the grid-mapfile*.

Chapter 4

Configuration

This section leads through the following steps:

1. Prepare the system to run ARC services (Section **4.1**, *Preparing the system*)
2. Configure a basic CE (Section **4.2**, *Configuration file formats* and Section **4.3**, *Setting up a basic CE*)
3. Make it production-ready (Section **4.4**, *Production CE setup*)
4. Add optional features (Section **4.5**, *Enhancing CE capabilities*)

4.1 Preparing the system

4.1.1 Users and groups

ARC services are run by the `root` user by default, and this is the most convenient way for normal operation. But it is also possible to run them as a non-privileged user (see Section **4.1.3**, *Permissions*).

Users accessing the grid have a *grid identity* (see Section **1.5**, *Security on the Grid*) and will submit and run jobs on different physical machines. In ARC, each grid identity is mapped to a local UNIX user on the front-end machine (the one that runs A-REX) and eventually on the machine actually performing the job (worker nodes, managed by the LRMS). Hence, one or more local UNIX users need to be created in the system, to run the jobs submitted by grid clients.

It is possible to map all grid users to the same local user. For a basic CE setup, this will be sufficient. Later however for security reasons it is better to have a pool of local users to choose from, or the have actual local users for each grid user. To anticipate more users in the future, it is a good practice to create a dedicated local group for these mapped users, so that is possible to use local UNIX authorization methods to restrict the grid accounts.

For the basic CE setup, let's create a new group called **grid** and new user called **griduser1** that belongs to this group. Later more users can be created.

More advanced user configuration setups are discussed in Section **4.4.1**, *Access control: users, groups, VOs.*

4.1.2 Disk, partitioning, directories

The ARC CE uses separate directories to store the data files of the jobs, the metadata about the jobs, and the cached input files. It also requires a directory with the installed CA certificates and optionally can use a directory of runtime environments.

Figure 4.1 shows these directories, Table 4.1 summarizes how these directories should be configured.

Some of these directories are suggested to be local to the front-end, other can be on shared or networked filesystems on external storage. The following is a description of the important directories for ARC CE. Note: some of them are **Required** for the ARC CE to work.

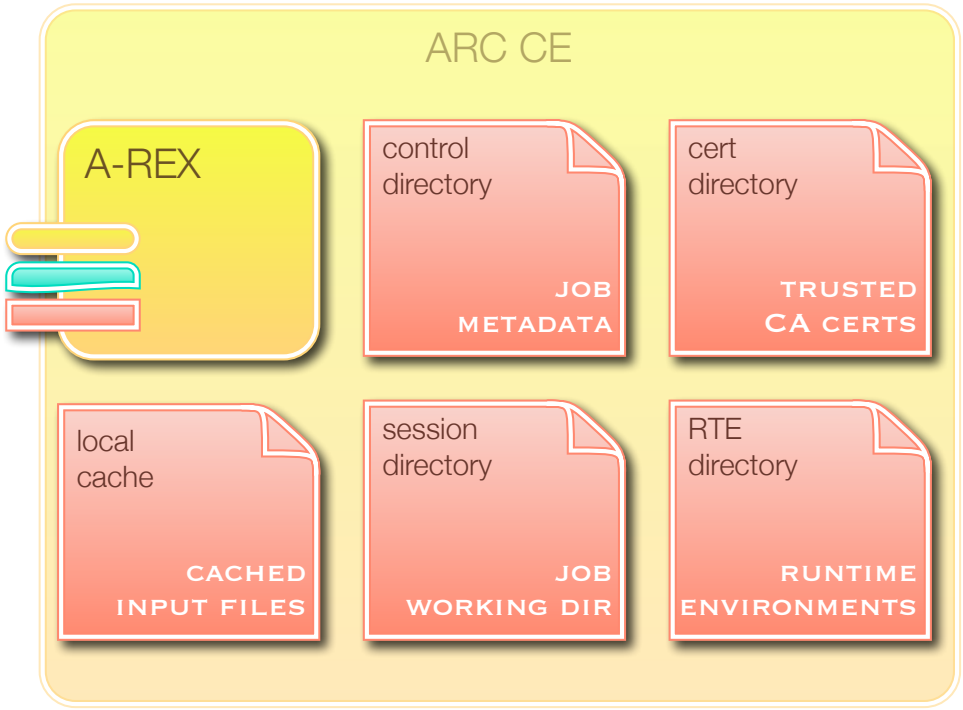


Figure 4.1: The directories on an ARC CE

Control Directory (CD) [Required] contains all the information about jobs handled by the A-REX, such as job specification files and LRMS submission scripts. The information provider scripts also use this directory to get information about jobs. This directory is heavily accessed by the A-REX, hence it should not be on a slow remote storage.

Session Directory (SD) [Required] contains the executable and data files of the jobs. This is where the jobs run, and this is the only area where they can produce results. Each job is assigned a unique directory within the session directory. This is usually shared among the worker nodes and the frontend, and can be remote for the frontend also. (See also Section **6.13**, *Using a scratch area.*)

Grid Certificates Directory [Required] contains the certificates of and other information about the trusted CAs. It is usually located at `/etc/grid-security/certificates`. (For setup instructions, see Section **3.3**, *Installation of certificates.*)

Cache Directory [Optional] can be used to cache downloaded input files, so if a new job requires the same file, it doesn't have to be downloaded again. Can reside on a shared filesystem. Caching is discussed in sections Section **4.4.3**, *Enabling the cache* and Section **6.4**, *Cache.*

Runtime Environments Scripts directory [Optional] contains special scripts that setup a particular runtime enviroment for a job to access. These include environment variables and software selections. Can reside on a shared filesystem. Runtime Environments are explained in Section **4.5.2**, *Runtime Environments.*

When partitioning disks and connecting shared storage, keep in mind the following things:

- The control directory (CD) is frequently accessed by the CE, so it is strongly advised to have it on a local hard disk. It can, however, grow pretty much with the number of jobs, so it is better to allocate a separate partition for it. The amount of data per job is generally around 50-100kb, but depending on the configured log level and the amount of data transfer, the data transfer log for each job can be much larger than this.
- The session directory (SD) stores all the executables, input and output files, and intermediate results of the jobs. It should be on a separate partation or even on a remote storage.

For more details please refer to sections Section **6.13**, *Using a scratch area*, Section **4.4.3**, *Enabling the cache*.

The ARC suggested setup for these directories is summarized in table 4.1.

Directory	**Suggested Location**	**Example**	Required?
session directory	NFS or shared FS, can be also on a separate disk partition	`/var/spool/arc/session`	Required
control directory	local to the front-end, also in a separate disk partition	`/var/spool/arc/control`	Required
CA certificates	local to the front-end	`/etc/grid-security/certificates`	Required
RTE scripts	NFS or shared FS	`/SOFTWARE/runtime`	Optional
cache directory	local, NFS, local and published via NFS	`/var/spool/arc/cache`	Optional

Table 4.1: Summary of ARC CE directories setup

4.1.3 Permissions

By default, the ARC services are run by root. In this case the control directory (CD) and the session directory (SD) should be writable, readable and executable by the root user, and then the A-REX will set all the other permissions as needed.

In case the ARC services should be run as a non-privileged (non-root) user, they cannot modify permissions of directories as easily. After the grid users are mapped to local users, they have to be able to access the job's session directory, hence the suggested setup is:

- put all the local users into the same group (e.g. `grid`)
- to set group ownership of the SD to this group
- the SD has to be writable, readable and executable by members of this group
- the SD and the CD have to be writable, readable and executable by the user running the ARC services

The host credentials need to have special permissions (see Section **3.3**, *Installation of certificates*).

4.1.4 Networking

DNS Requirements For the ARC middleware, the frontend has to have a public IP and a Fully Qualified Domain Name (FQDN) in order to join an indexing service and thus the grid (more on this on chapter Section **4.4.5**, *How to join the grid: registering to an index service*). This means that a reverse DNS lookup for the frontend's IP has to return the FQDN.

Basic networking recommendations are the following:

- Make sure your frontend has a FQDN. Issuing `hostname -f` should print it.
- In the `/etc/hosts` file, make sure that the FQDN of your machine comes **first**, before other network names. Example: if `130.235.185.195` is the IP address and `gridtest.hep.lu.se` is the FQDN assigned to it, `/etc/hosts` should look like:

  ```
  130.235.185.195 gridtest.hep.lu.se gridtest
  ```

  while the following could lead to problems:

  ```
  # wrong!
  130.235.185.195 gridtest gridtest.hep.lu.se
  ```

Firewalls ARC-CE needs the following incoming and outgoing ports to be opened:

- For the web service interface: HTTP(s), default 80 and 443
- For LDAP Information System, default 2135 (see also Section **4.3.5**, *The [infosys] section: the local information system*)
- For the gridftp service interface: GridFTP,
  - default 2811
  - a range of ports for GridFTP data channels, typically 9000-9300
- For HTTPg, default 8443 (outgoing only)
- For SMTP, default 25 (outgoing only)
- For NTP, default 123 (outgoing only, in case NTP is used for time synchronisation, see **2**, *Requirements*)
- For webservices, the port defined for A-REX. See Section **4.5.3**, *Enabling the Web Services interface.*

Most ports, including 2135 and 2811, are registered with IANA and should normally not be changed. The ports for GridFTP data channels can be chosen arbitrary, based on following considerations: gridftpd by default handles 100 connections simultaneously; each connection should not use more than 1 additional TCP port. Taking into account that Linux tends to keep ports allocated even after the handle is closed for some time, it is a good idea to triple that amount. Hence about 300 data transfer ports should be enough for the default configuration. Typically, the range of ports from 9000 to 9300 is being opened. Remember to specify this range in the ARC configuration file (see Section **4.2**, *Configuration file formats*, `globus_tcp_port_range` attribute) later on.

For using legacy Globus components it is also worth to read information at this URL: `http://dev.globus.org/wiki/FirewallHowTo`

Other network related Internal cluster nodes (i.e. LRMS nodes) are NOT required to be fully available on the public internet (however, user applications may require it). For information about publishing nodes' network connectivity please refer to Section **4.3.5.1**, *The [cluster] section: information about the host machine.*

4.1.5 Security considerations

SELinux If the system uses SELinux, the startup script should be usually able to create profiles for the services.

To fine tune LDAP information system permissions, see **5.7.2**, *How to configure SELinux to use a port other than 2135 for the LDAP information system.*

If any problem in connecting to or starting up services arises, submit a bug report to the ARC bugzilla*.

If problems arise and it is suspected they are due to SELinux, the best is to set SELinux in permissive mode and check if the problem persists.

AppArmor On Ubuntu and Debian machines AppArmor profiles have been reported to prevent the infosystem starting. AppArmor profiles are currently not shipped for ARC components. Therefore for the time being:

- Remove `/etc/apparmor.d/usr.sbin.slapd` and restart AppArmor.
- If the above doesn't exist or doesn't help, disable AppArmor completely or put all the profiles in complain mode†.

*`http://bugzilla.nordugrid.org/`
†`https://help.ubuntu.com/community/AppArmor`

4.2 Configuration file formats

Configuration of ARC can be done with a single configuration file usually located at `/etc/arc.conf`.

This configuration file format is fully compatible with the one for ARC middleware version 0.8.x.

> ⋆ If you have a legacy file from an ARC 0.8.x version,
> you can directly use that file for the new A-REX-based ARC CE.

Using the the `arc.conf` is sufficient for the majority of use cases, however there is a possibility to use a lower-level XML-based configuration format (and a corresponding higher-level INI format) in special cases. For more details, see Section **6.7**, *The XML and the INI configuration formats*.

4.2.1 Structure of the `arc.conf` configuration file

An ARC configuration file is a text file containing **sections** and related **commands**.

Each **section** identifies one or more components/features of ARC, and **commands** are used to modify the behaviour of these component/features.

A **section name** is sourrounded by square brackets and can contain slashes. Names after the slashes identify subsections. Examples:

```
[cluster]
[infosys]
[infosys/glue12]
[queue/fork]
[infosys/cluster/registration/toPGS1]
```

As a general rule, a section name containing a subsection has to appear **after** its section. Examples in Figure 4.2.

```
...
[infosys]
...
[infosys/glue12]
...
[queue/fork]
...
[infosys/cluster/registration/toPGS1]
...
```

Correct

```
...
[infosys/cluster/registration/toPGS1]
...
[infosys/glue12]
...
[infosys]
...
[queue/fork]
...
```

Wrong

Figure 4.2: Ordering of section names.

A **configuration command** is a one-line `command="value"` expression. Examples:

```
hostname="gridtest.hep.lu.se"
nodecpu="2"
resource_location="Lund, Sweden"
mail="gridmaster@hep.lu.se"
```

Comments can be added one per line by putting a # at the beginning of the line.

A section starts with a section name and ends at another section name or if the end of the configuration file is reached. Configuration commands always belong to one section.

Here is an overall example:

```
# this is a comment, at the beginning of the [common] section
[common]
hostname="piff.hep.lu.se"
x509_user_key="/etc/grid-security/hostkey.pem"
x509_user_cert="/etc/grid-security/hostcert.pem"
x509_cert_dir="/etc/grid-security/certificates"
gridmap="/etc/grid-security/grid-mapfile"
lrms="fork"

# since there is a new section name below, the [common] section ends
# and the grid-manager section starts
[grid-manager]
user="root"
controldir="/tmp/control"
sessiondir="/tmp/session"
# cachedir="/tmp/cache"
debug="3"

# other commands...

[queue/fork]

# other commands till the end of file.
# This ends the [queue/fork] section.
```

4.2.2 Description of configuration items

In the descriptions of commands, the following notation will be used:

command=*value [value]* – where the values in square brackets *[...]* are *optional*. They should be inserted without the square brackets!

A pipe "|" indicates an exclusive option. Example:

securetransfer=*yes|no* – means that the value is either `yes` or `no`.

For a complete list and description of each configuration item, please refer to Section **6.1**, *Reference of the `arc.conf` configuration commands*.

The configuration commands are organized in sections. The following is a description of the main **mandatory** sections and of the components and functionalities they apply to, in the order they should appear in the configuration file. These are needed for minimal and basic functionalities (see Section **4.3**, *Setting up a basic CE*).

[common] Common configuration affecting networking, security, LRMS. These commands define defaults for all the ARC components (A-REX, GridFTPd, ARIS), which can be overridden by the specific sections of the components later. Always appears at the beginning of the config file.

Discussed in Section **4.3.2**, *The [common] section*.

[group] This section and its subsections define access control mappings between grid users and local users. Applies to all ARC components. Usually follows the `[common]` section. If there are `[vo]` sections, they should come before the `[group]` section.

Discussed in Section **4.4.1**, *Access control: users, groups, VOs*.

If no access control is planned (for example for tests) this section can be omitted but the administrator must manually edit the grid-mapfile (see Section **6.10**, *Structure of the grid-mapfile*)

[grid-manager] This section configures the A-REX, including job management behavior, directories, file staging and logs.

Discussed in Section **4.3.3**, *The [grid-manager] section: setting up the A-REX and the arched.*

[gridftpd] This section configures the GridFTPd, which is the server process running the GridFTP protocol. Its subsections configure the different plugins of the GridFTPd, in particular the job submission interface: **[gridftpd/jobs]**.

Discussed in Section **4.3.4**, *The [gridftpd] section: the job submission interface.*

[infosys] This section configures the local information system (ARIS) and the information provider scripts. (This section also can be used to configure an information index server, see [**?**].) The commands affect the data published by the information system, the behaviour of the publishing server and its networking options. The subsections configure registration to information index servers, and extra information for different information schemas.

Discussed in Section **4.3.5**, *The [infosys] section: the local information system.*

[cluster] Configures the A-REX information provider scripts. The commands here affect the data published by the local information system, mostly regarding the front-end machine. Must appear after the `[infosys]` section.

Discussed in Section **4.3.5.1**, *The [cluster] section: information about the host machine*

[queue/queuename] Configures the queues provided by A-REX. At least one [queue/...] section must exist. The commands here affect the data published by the information system, mostly regarding the LRMS queues A-REX is serving. Must appear after the `[infosys]` section.

Discussed in Section **4.3.5.2**, *The [queue/fork] section: configuring the fork queue.*

Generic commands These commands specify common defaults in the `[common]` section, and also can be used to set different values per component in the following sections: `[grid-manager]`, `[gridftpd]` and its subsections and `[infosys]`.

logfile=*path* – where the logs will be written.

pidfile=*path* – where the PID of the process will be written.

debug=*number* – specifies the level of logging from 5 (DEBUG) to 0 (FATAL).

4.3 Setting up a basic CE

A basic CE is the starting point of every ARC setup. A basic CE is a stand-alone machine ready to accept job submission. A basic CE will not be connected to an information index, so clients will have to explicitly specify its job submission interface URL to connect to. This chapter will show a basic configuration of the main sections seen in chapter Section **4.2.2**, *Description of configuration items.*

Please make sure all the steps in chapter Section **4.1**, *Preparing the system* are done before proceeding.

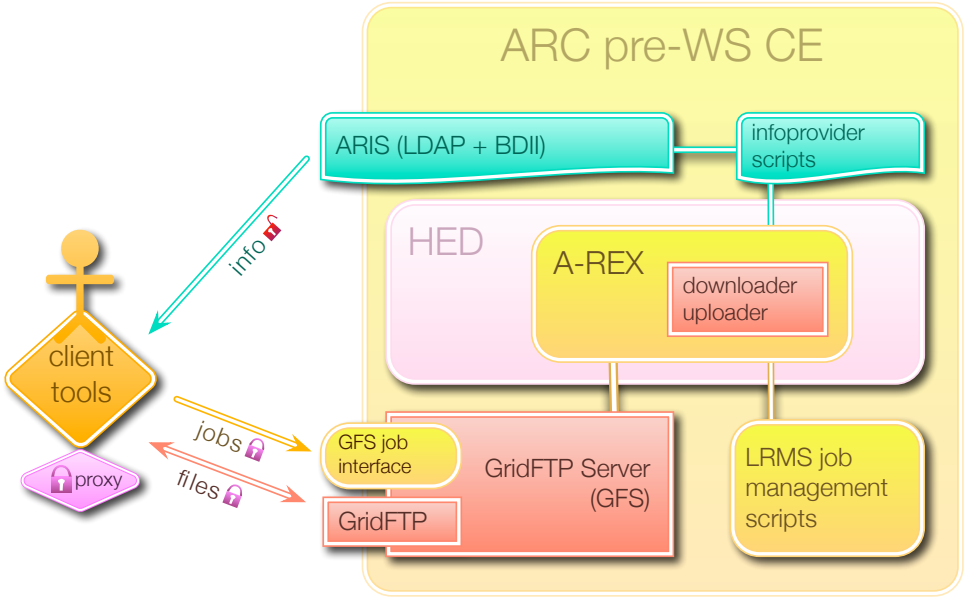
The basic CE will have **fork** as an LRMS, which will allow the machine to process jobs in the environment provided by the operating system of the front-end machine. Connecting to real LRMSes is discussed in Section **4.4.2**, *Connecting to the LRMS.*

4.3.1 Creating the arc.conf file

ARC will by default search for its configuration file in the following location:

```
/etc/arc.conf
```

The minimal configuration file described in the following is usually installed here:




```
/usr/share/doc/nordugrid-arc-doc<version>/examples/arc_computing_element.conf
```

where <version> varies with every update of the documentation.

The latest one can be downloaded from the ARC Configuration Examples web page[‡].

Copy this file into `/etc` with the name `arc.conf`, then customize its contents following the instructions below, although it should work without any customization.

4.3.2 The [common] section

The `[common]` section maintains informations that will be used by any subsystem of the CE. It has to appear as the first item in the configuration file.

A minimal configuration for this section is shown here:

```
[common]
x509_user_key="/etc/grid-security/hostkey.pem"
x509_user_cert="/etc/grid-security/hostcert.pem"
x509_cert_dir="/etc/grid-security/certificates"
gridmap="/etc/grid-security/grid-mapfile"
lrms="fork"
```

Here we specify the path of the host's private key and certificate, the directory where the certificates of the trusted Certificate Authorities (CAs) are located, the path of the grid map file, which defines mapping of grid users to local users, and the name of the default LRMS, which is "fork" in the basic case, when we only want to use the frontend as a worker node, not a real cluster.

For details about these configuration commands, please see Section **6.1.1**, *Generic commands in the [common] section*

For the basic CE, let's create a "grid map file" which looks like this:

```
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo1" griduser1
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo2" griduser1
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo3" griduser1
```

[‡] `http://www.nordugrid.org/arc/configuration-examples.html`

4.3.3 The [grid-manager] section: setting up the A-REX and the arched

The `[grid-manager]` section configures `A-REX` and `arched`. Its commands will affect the behaviour of the startup scripts and the `A-REX` and `arched` processes.

A sample section would look like this:

```
[grid-manager]
user="root"
controldir="/tmp/jobstatus"
sessiondir="/tmp/grid"
debug="3"
logfile="/tmp/grid-manager.log"
pidfile="/tmp/grid-manager.pid"
mail="grid.support@somewhere.org"
joblog="/tmp/gm-jobs.log"
```

Here we specify which user the A-REX should be run as, where should be the directory for the job's metadata (the control dir) and data (the session dir), what level of log message we want, where should be the log file and where should the process ID of the arched daemon be written. We also specify an e-mail contact address and the path of the "joblog" file, which will contain information about each job's lifecycle.

For details about these configuration commands, please see Section **6.1.12**, *Commands in the [grid-manager] section*

4.3.4 The [gridftpd] section: the job submission interface

Currently, the production level job submission interface uses the **gridftp** protocol which is served by the GridFTP Server (GFS) running on the frontend.

The `[gridftpd]` section configures the behaviour of the gridftpd daemon and its startup scripts.

A sample section for a basic CE is the following:

```
[gridftpd]
user="root"
debug="3"
logfile="/tmp/gridftpd.log"
pidfile="/tmp/gridftpd.pid"
port="2811"
allowunknown="no"
```

Here we specify which user the GridFTP server should run as, the verbosity of the log messages, the path of the logfile and the pidfile, the port of the GridFTP server, and that only "known" users (specified in the grid map file) should be allowed to connect.

For a minimal ARC CE to work, we need the configure the job interface with setting up the "job plugin" of the GridFTP server in a configuration subsection:

`[gridftpd/jobs]` controls how the virtual path /jobs for job submission will behave. These paths can be thought of as those of a UNIX mount command. The name *jobs* itself is not relevant, but the contents of the section and especially the `plugin` command determine the path behaviour.

For a minimal CE to work, it is sufficient to configure the following:

```
[gridftpd/jobs]
path="/jobs"
plugin="jobplugin.so"
allownew="yes"
```

Here we specify the virtual path where the job plugin will sit, the name of the library of the plugin, and that new jobs can be submitted (turning `allownew` to "no" would stop accepting new jobs, but the existing jobs would still run.)

For a more complex configuration example with fine-grained authentication based on groups see **6.15.4**, *Configuration Examples* and for full details on all configuration commands, please see Section **6.1.4**, *Commands in the [gridftpd] section*

As GridFTPd interface is planned to be phased out and replaced by the web service interface, no big changes will be done in the future.

4.3.5 The [infosys] section: the local information system

The `[infosys]` section and its subsections control the behaviour of the information system. This includes:

- configuration of ARIS and its infoproviders
- customization of the published information
- configuration of the `slapd` server to publish information via LDAP
- configuration of BDII to generate ldif trees for LDAP
- selection of the LDAP schema(s) to publish
- registration to an EGIIS index service (see Section **4.4.5**, *How to join the grid: registering to an index service*)
- running a EGIIS IS (not covered in this manual, please refer to [**?**])

After this section, several subsections will appear as well as some other sections which are related to the information system, such as `[cluster]` and `[queue/...]` sections. More on these will be explained later.

A sample configuration for a basic CE would be the following:

```
[infosys]
user="root"
overwrite_config="yes"
port="2135"
debug="1"
slapd_loglevel="0"
registrationlog="/tmp/inforegistration.log"
providerlog="/tmp/infoprovider.log"
provider_loglevel="2"
```

Here we specify which user the slapd server, the infoproviders, the BDII and the registration scripts should run, then we specify that we want the low-level slapd configs to be regenerated each time, then the port number, the debug verbosity of the startup script, the slapd server and the infoproviders, and the logfiles for the registration messages and the infoprovider messages.

For details about these configuration commands, please see Section **6.1.5**, *Commands in the [infosys] section.*

4.3.5.1 The [cluster] section: information about the host machine

This section has to follow the `[infosys]` section and it is used to configure the information published about the host machine running ARC CE.

A sample configuration can be seen below:

```
[cluster]
cluster_alias="MINIMAL Computing Element"
comment="This is a minimal out-of-box CE setup"
homogeneity="True"
architecture="adotf"
nodeaccess="inbound"
nodeaccess="outbound"
```

Here we specify the alias of the cluster, a comment about it, that the worker nodes are homogeneous, that we want infoprovider scripts to determine the architecture automatically on the frontend ("adotf"), and that the worker nodes have inbound and outbound network connectivity.

For details about these configuration commands, please see Section **6.1.9**, *Commands in the [cluster] section.*

4.3.5.2 The [queue/fork] section: configuring the fork queue

Each `[queue/queuename]` section configures the information published about computing queues. At least one queue must be specified for a CE to work. In this chapter a configuration for the **fork** LRMS will be shown.

The fork LRMS is just a simple execution environment provided by the means of the underlying operating system, that is, usually a shell with the standard linux environment variables provided to the mapped UNIX user.

A special section name [queue/fork] is used to configure such information, some of its commands can be used for any queue section, some are specific for the fork queue. More about this will be explained in Section **4.4.2**, *Connecting to the LRMS.*

A minimal CE configuration for this section would look like this:

```
[queue/fork]
name="fork"
fork_job_limit="cpunumber"
homogeneity="True"
scheduling_policy="FIFO"
comment="This queue is nothing more than a fork host"
nodecpu="adotf"
architecture="adotf"
```

Here we specify that this is a "fork" queue, that the number of allowed concurent jobs should equal the number of CPUs, that the queue is homogeneous, the scheduling policy, an informative comment, and that the type of the cpu and the architecture should be determined automatically on the frontend. The only fork-specific command is the `fork_job_limit` command, the others can be used for other LRMSes also. See sections Section **4.4.2**, *Connecting to the LRMS* and Section **6.1.10**, *Commands in the [queue] subsections.*

4.3.6 A basic CE is configured. What's next?

A basic CE is now set. To test its functionality, it must be started first. Please refer to Section **5.1.2**, *Starting the CE* to start the CE. If none of the startup scripts give any error, the testing can be started. Please follow the testing suggestions in Section **5.2**, *Testing a configuration.*

If everything works as expected, the next step is the turn the basic CE into a production level CE: connecting it to the LRMS, turning on input file caching, and registering it to an information index service. Please follow the instructions in Section **4.4**, *Production CE setup.*

For some additional (optional) features, please proceed to Section **4.5**, *Enhancing CE capabilities.*

4.4 Production CE setup

Once a basic CE is in place and its basic functionalities have been tested, these things are usually needed to make it production-ready:

Configure access control to streamline the maintenance of the authentication and authorization of users, VOs and authorization groups should be defined and the `nordugridmap` tool should be utilized to generate the grid map file automatically. See Section **4.4.1**, *Access control: users, groups, VOs*.

Connect to the LRMS to be able to use the underlying batch system, ARC support several famous clustering and load balancing systems such as Torque/PBS, Sun Grid Engine, LSF, and others. See Section **4.4.2**, *Connecting to the LRMS*.

Enabling the cache to keep a copy of the downloaded input files in case the next job needs the same, which greatly decreases wait time for jobs to start. See Section **4.4.3**, *Enabling the cache*

Configure data staging Staging data in and out for jobs is a critical part of the CE, and it is important that it is correctly configured to optimise performance. See Section **4.4.4**, *Configuring Data Staging*.

Register to an index service NorduGrid provides an index service that will publish the CE to all the grid clients that have access to the NorduGrid network. In this way the CE will be part of the GRID. See Section **4.4.5**, *How to join the grid: registering to an index service*.

Accounting the A-REX is capable of sending usage records to the SGAS accounting service. See Section **4.4.7**, *Sending usage records to SGAS with urlogger*.

Monitoring Nagios plugins exist for monitoring the ARC Computing Element. See Section **4.4.8**, *Monitoring the ARC CE: Nagios probes*.

4.4.1 Access control: users, groups, VOs

The grid mappings between grid users and local unix accounts are listed in the so-called `grid map file`, usually located in the directory `/etc/grid-security/`. By default this file also serves as list of authorized users. While this text file can be edited by hand this is not advisible in production environments. To ease the security administrator's job, NorduGrid provides a collection of scripts and cron jobs that automatically keeps the local grid map files synchronized to a central user database. If the CE has to join the Grid, it is suggested to install the `nordugrid-arc-gridmap-utils` package from the NorduGrid Downloads area or EMI repository, see Chapter **3**, *Installation* for details. Once installed, the `[groups]` and `[vo]` sections in the configuration file can be edited as well as optionally the location of the file representing the local list of mappings (can have any name, but usually called `/etc/grid-security/local-grid-mapfile`). For the description of the grid map file, please refer to Section **6.10**, *Structure of the grid-mapfile*.

The two sections `[group]` and `[vo]` configure basic access control policies. The `[vo]` section may be also used to control automatic mapping of GRID identities to local UNIX users:

`[vo]` defines Virtual Organizations (VOs). A VO is a simple way of grouping sets of users belonging to different (real) organizations and, for example, willing to use the same set of software. A common use of this section is to include users published by VOMS servers [**?**]. `[vo]` sections can be referred by `[group]` sections. If this happens, it is important that the corresponding `[vo]` definition appears **before** the `[group]` section that refers to it.

`[group]` defines *authorization rules* to access the CE for users or set of users defined by [vo] sections.

The configuration presented here is sufficient for a simple production setup where the identities are known or are already contained in a file or a collection of files, eventually located and updated remotely.

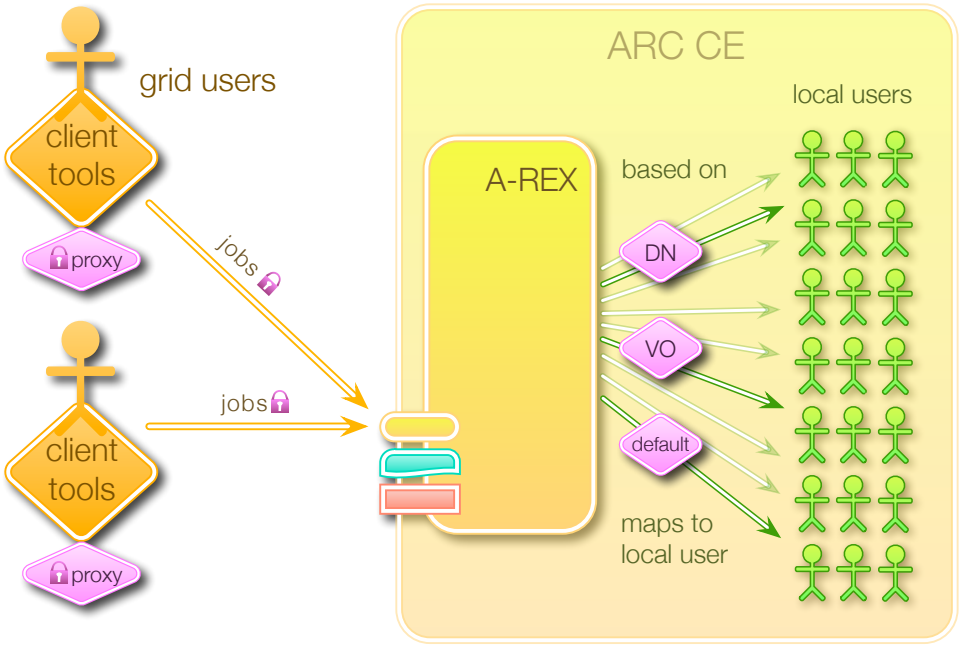


Figure 4.3: The A-REX maps the grid users to local users based on information about their identity and Virtual Organization membership. It's also possible to do default mapping.

4.4.1.1 [vo] configuration commands

The following is a sample [vo] section for a minimal CE:

```
[vo]
id="vo_1"
vo="TestVO"
source="file:///etc/grid-security/local-grid-mapfile"
mapped_unixid="griduser1"
require_issuerdn="no"
```

We define a VO here with the name of `TestVO` and the id of `vo_1`, the list of members comes from a URL (which here points to a local file, see example below), and all members of this VO will be mapped to the local user `griduser1`.

Here's an example of the file with the list of members:

```
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo1"
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo2"
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo3"
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo4"
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo5"
```

For more configuration options, please see Section **6.1.2**, *Commands in the [vo] section.*

To generate the actual grid map file from these [vo] settings, we need the `nordugridmap` utility, described below.

4.4.1.2 Automatic update of the mappings

The package `nordugrid-arc-gridmap-utils` contains a script to automatically update user mappings (usually located in `/usr/sbin/nordugridmap`). It does that by fetching all the sources in the `source` commands and writing their contents adding the mapped user `mapped_unixid` in the grid-mapfile and each file specified by the `file` command. The script is executed from time to time as a cron job.

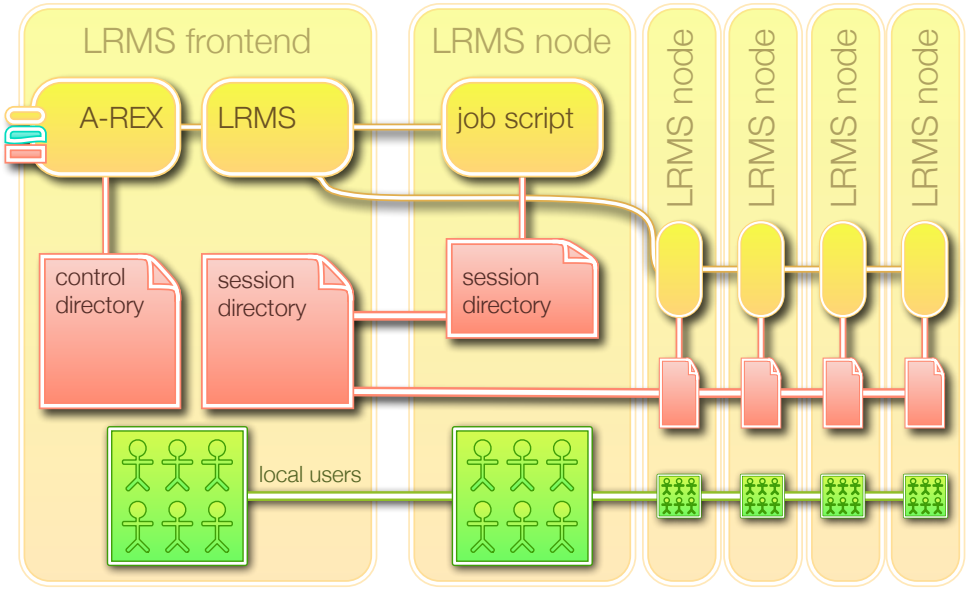


Figure 4.4: The LRMS frontend and the nodes sharing the session directory and the local users

4.4.1.3 [group] configuration commands

[group] defines *authorizations* for users accessing the grid.

There can be more than one group in the configuration file, and there can be subsections identified by the group name such as `[group/users]`.

For a minimal CE with no authorization rules, it is sufficient to have something like the following, preceeded with the `[vo]` section previously defined in this chapter:

```
[group/users]
name="users"
vo="TestVO"
```

where the `name` could be omitted and then would be automatically taken from the subsection name.

For more about authorization, please read Section **6.1.3**, *Commands in the [group] section.*

4.4.2 Connecting to the LRMS

A-REX supports several Local Resource Management Systems, with which it interacts by several backend scripts.

Connecting A-REX to one of these LRMS involves the following steps:

1. creation of shared directories between A-REX, the LRMS frontend and its working nodes. It might involve setup of shared filesystems such as NFS or similar.
2. configuration of the behaviour of a-rex with respect to the shared directories in the `[grid-manager]` section.
3. configuration of the following `arc.conf` sections: `[common]`, `[grid-manager]`, `[queue/*]`.

In the `[common]` section the name of the LRMS has to be specified:

lrms=*default_lrms_name [default_queue_name]* – specifies the name of the LRMS and optionally the queue.

The following [grid-manager] configuration commands affect how A-REX interacts with the LRMS:

gnu_time=*path* – path to *time* utility.

tmpdir=*path* – path to directory for temporary files. Default is /tmp.

runtimedir=*path* – path to directory which contains *runtimenvironment* scripts.

shared_filesystem=*yes|no* – if computing nodes have an access to session directory through a shared file system like NFS. If set to "*no*", this means that the computing node does not share a filesystem with the frontend. In this case the content of the SD is moved to a computing node using means provided by the LRMS. Results are moved back after the job's execution in a similar way. Sets the environment variable RUNTIME_NODE_SEES_FRONTEND

scratchdir=*path* – path on computing node where to move session directory before execution. If defined should contain the path to the directory on computing node which can be used to store a job's files during execution. Sets the environment variable RUNTIME_LOCAL_SCRATCH_DIR .

shared_scratch=*path* – path on frontend where ***scratchdir*** can be found. If defined should contain the path corresponding to that set in scratchdir as seen on the **frontend** machine. Sets the environment variable RUNTIME_FRONTEND_SEES_NODE .

nodename=*command* – command to obtain hostname of computing node.

For additional details, see Section **6.1.12.10**, *Substitutions in the command arguments* and Section **6.13**, *Using a scratch area.*

Each LRMS has his own peculiar configuration options.

4.4.2.1 PBS

The Portable Batch System (PBS) is one of the most popular batch systems. PBS comes in many flavours such as OpenPBS (unsupported), Terascale Open-Source Resource and QUEue Manager (TORQUE) and PBSPro (currently owned by Altair Engineering). ARC supports all the flavours and versions of PBS.

Recommended batch system configuration PBS is a very powerful LRMS with dozens of configurable options. Server, queue and node attributes can be used to configure the cluster's behaviour. In order to correctly interface PBS to ARC (mainly the information provider scripts) there are a couple of configuration REQUIREMENTS asked to be implemented by the local system administrator:

1. The computing nodes MUST be declared as cluster nodes (job-exclusive), at the moment time-shared nodes are not supported by the ARC setup. If you intend to run more than one job on a single processor then you can use the virtual processor feature of PBS.

2. For each queue, one of the max_user_run or max_running attributes MUST be set and its value SHOULD BE IN AGREEMENT with the number of available resources (i.e. don't set the max_running = 10 if there are only six (virtual) processors in the system). If both max_running and max_user_run are set then obviously max_user_run has to be less or equal to max_running.

3. For the time being, do NOT set server limits like max_running, please use queue-based limits instead.

4. Avoid using the max_load and the ideal_load directives. The Node Manager (MOM) configuration file (<PBS home on the node>/mom_priv/config) should not contain any max_load or ideal_load directives. PBS closes down a node (no jobs are allocated to it) when the load on the node reaches the max_load value. The max_load value is meant for controlling time-shared nodes. In case of job-exclusive nodes there is no need for setting these directives, moreover incorrectly set values can close down a node.

5. Routing queues are now supported in a simple setup were a routing queue has a single queue behind it. This leverages MAUI work in most cases.
   Other setups (i.e. two or more execution queues behind a routing queue) cannot be used within ARC.

Additional useful configuration hints:

- If possible, please use queue-based attributes instead of server level ones (for the time being, do not use server level attributes at all).
- The "acl_user_enable = True" attribute may be used with the "acl_users = user1,user2" attribute to enable user access control for the queue.
- It is advisory to set the max_queuable attribute in order to avoid a painfully long dead queue.
- Node properties from the `<PBS home on the server>/server_priv/nodes` file together with the `resources_default.neednodes` can be used to assign a queue to a certain type of node.

Checking the PBS configuration:

- The node definition can be checked by `<PBS installation path>/bin/pbsnodes -a`. All the nodes MUST have ntype=cluster.
- The required queue attributes can be checked as `<PBS installation path>/bin/qstat -f -Q queuename`. There MUST be a max_user_run or a max_running attribute listed with a REASONABLE value.

Configuration commands in `arc.conf` Below the PBS specific configuration variables are collected.

`lrms`=*"pbs"* – in the `[common]` section enables the PBS batch system back-end. No need to specify the flavour or the version number of the PBS, simply use the `"pbs"` keyword as LRMS configuration value.

For each grid-enabled (or grid visible) PBS queue a corresponding `[queue/queuename]` subsection must be defined. `queuename` should be the PBS queue name.

`pbs_bin_path`=*path* – in the `[common]` section should be set to the path to the qstat,pbsnodes,qmgr etc. PBS binaries.

`pbs_log_path`=*path* – in the `[common]` sections should be set to the path of the PBS server logfiles which are used by A-REX to determine whether a PBS job is completed. If not specified, A-REX will use the `qstat` command to find completed jobs.

For additional configuration commands, please see Section **6.1.14**, *PBS specific commands*.

Known limitations Some of the limitations are already mentioned under the PBS deployment requirements. No support for routing queues, difficulty of treating overlapping queues, the complexity of node string specifications for parallel jobs are the main shortcomings.

4.4.2.2 Condor

The Condor [**?**] system, developed at the University of Wisconsin-Madison, was initially used to harness free cpu cycles of workstations. Over time it has evolved into a complex system with many grid-oriented features. Condor is available on a large variety of platforms.

Recommended batch system configuration Install Condor on the A-REX node and configure it as a submit machine. Next, add the following to the node's Condor configuration (or define `CONDOR_IDS` as an environment variable):

```
CONDOR_IDS = 0.0
```

`CONDOR_IDS` has to be 0.0, so that Condor will be run as root and can then access the Grid job's session directories (needed to extract various information from the job log).

Make sure that no normal users are allowed to submit Condor jobs from this node. If normal user logins are not allowed on the A-REX machine, then nothing needs to be done. If for some reason users are allowed to log into the A-REX machine, simply don't allow them to execute the condor_submit program. This can be done by putting all local Unix users allocated to the Grid in a single group, e.g. 'griduser', and then setting the file ownership and permissions on condor_submit like this:

```
chgrp griduser $condor_bin_path/condor_submit
chmod 750 $condor_bin_path/condor_submit
```

Configuration commands in `arc.conf` The Condor-specific configuration commands:

`lrms`=*"condor"* – in the `[common]` section enables the Condor batch system back-end.

`condor_bint_path`=*path* – in the `[common]` section should be set to the directory containing Condor binaries (f.ex., /opt/condor/bin). If this parameter is missing, ARC will try to guess it out of the system path, but it is highly recommended to have it explicitly set.

For additional configuration commands, please see Section **6.1.15**, *Condor specific commands*.

Known limitations Only Vanilla universe is supported. MPI universe (for multi-CPU jobs) is not supported. Neither is Java universe (for running Java executables). ARC can only send jobs to Linux machines in the Condor pool, therefore excluding other unixes and Windows destinations.

4.4.2.3 LoadLeveler

LoadLeveler(LL), or Tivoli Workload Scheduler LoadLeveler in full, is a parallel job scheduling system developed by IBM.

Recommended batch system configuration The back-end should work fine with a standard installation of LoadLeveler. For the back-end to report the correct memory usage and cputime spent, while running. LoadLeveler has to be set-up to show this data in the llq command. Normally this is turned off for performance reasons. It is up to the cluster administrator to decide whether or not to publish this information. The back-end will work whether or not this is turned on.

Configuration commands in `arc.conf` Only the two basic LRMS config options are relevant for LoadLeveler:

`lrms`=*"ll"* – in the `[common]` section enables the LoadLeveler batch system.

`ll_bin_path`=*path* – in the `[common]` section must be set to the path of the LoadLeveler binaries.

Known limitations There is at the moment no support for parallel jobs on the LoadLeveler back-end.

4.4.2.4 Fork

The Fork back-end is a simple back-end that interfaces to the local machine, i.e.: there is no batch system underneath. It simply forks the job, hence the name. The back-end then uses standard posix commands (e.g. ps or kill) to manage the job.

Recommended batch system configuration Since fork is a simple back-end and does not use any batch system, there is no specific configuration needed for the underlying system.

Configuration commands in `arc.conf` Only these commands are applied:

`lrms`=*"fork"* – in the `[common]` section enables the Fork back-end. The queue must be named `"fork"` in the `[queue/fork]` subsection.

`fork_job_limit`=*cpunumber* – sets the number of running grid jobs on the fork machine, allowing a multi-core machine to use some or all of its cores for Grid jobs. The default value is 1.

Known limitations Since Fork is not a batch system, many of the queue specific attributes or detailed job information is not available. The support for the "Fork batch system" was introduced so that quick deployments and testing of the middleware can be possible without dealing with deployment of a real batch system since fork is available on every UNIX box. The "Fork back-end" is not recommended to be used in production. The back-end by its nature, has lots of limitations, for example it does not support parallel jobs.

4.4.2.5 LSF

Load Sharing Facility (or simply LSF) is a commercial computer software job scheduler sold by Platform Computing. It can be used to execute batch jobs on networked Unix and Windows systems on many different architectures.

Recommended batch system configuration Set up one or more LSF queues dedicated for access by grid users. All nodes in these queues should have a resource type which corresponds to the one of the the frontend and which is reported to the outside. The resource type needs to be set properly in the `lsb.queues` configuration file. Be aware that LSF distinguishes between 32 and 64 bit for Linux. For a homogeneous cluster, the `type==any` option is a convenient alternative.

Example: In `lsb.queues` set one of the following:

```
RES_REQ = type==X86_64
RES_REQ = type==any
```

See the `-R` option of the `bsub` command man page for more explanation.

Configuration commands in `arc.conf` The LSF back-end requires that the following options are specified:

`lrms`=*"lsf"* – in the `[common]` section enables the LSF back-end

`lsf_bin_path`=*path* – in the `[common]` section must be set to the path of the LSF binaries

`lsf_profile_path`=*path* – must be set to the filename of the LSF profile that the back-end should use.

Furthermore it is very important to specify the correct architecture for a given queue in arc.conf. Because the architecture flag is rarely set in the xRSL file the LSF back-end will automatically set the architecture to match the chosen queue. LSF's standard behaviour is to assume the same architecture as the frontend. This will fail for instance if the frontend is a 32 bit machine and all the cluster resources are 64 bit. If this is not done the result will be jobs being rejected by LSF because LSF believes there are no useful resources available.

Known limitations Parallel jobs have not been tested on the LSF back-end.

The back-end does not at present support reporting different number of free CPUs per user.

4.4.2.6 SGE

Sun Grid Engine (SGE, Oracle Grid Engine, Codine) is an open source batch system maintained by Sun (Oracle). It is supported on Linux, and Solaris in addition to numerous other systems.

Recommended batch system configuration Set up one or more SGE queues for access by grid users. Queues can be shared by normal and grid users. In case it is desired to set up more than one ARC queue, make sure that the corresponding SGE queues have no shared nodes among them. Otherwise the counts of free and occupied CPUs might be wrong. Only SGE versions 6 and above are supported.

Configuration commands in `arc.conf` The SGE back-end requires that the following options are specified:

`lrms`=*"sge"* – in the `[common]` section enables the SGE batch system back-end.

`sge_root`=*path* – in the `[common]` section must be set to SGE's install root.

`sge_bin_path`=*path* – in the `[common]` section must be set to the path of the SGE binaries.

`sge_jobopts`=*options* – in the `[queue/queuename]` section can be used to add custom SGE options to job scripts submitted to SGE. Consult SGE documentation for possible options. Example:

```
lrms="sge"
sge_root="/opt/n1ge6"
sge_bin_path="/opt/n1ge6/bin/lx24-x86"

...

[queue/long]
sge_jobopts="-P atlas -r yes"
```

For additional configuration commands, please see Section **6.1.19**, *SGE specific commands*.

Known limitations Multi-CPU support is not well tested. All users are shown with the same quotas in the information system, even if they are mapped to different local users. The requirement that one ARC queue maps to one SGE queue is too restrictive, as the SGE's notion of a queue differs widely from ARC's definition. The flexibility available in SGE for defining policies is difficult to accurately translate into NorduGrid's information schema. The closest equivalent of nordugrid-queue-maxqueuable is a per-cluster limit in SGE, and the value of nordugrid-queue-localqueued is not well defined if pending jobs can have multiple destination queues.

4.4.2.7 SLURM

SLURM is an open-source (GPL) resource manager designed for Linux clusters of all sizes. It is designed to operate in a heterogeneous cluster with up to 65,536 nodes. SLURM is actively being developed, distributed and supported by Lawrence Livermore National Laboratory, Hewlett-Packard, Bull, Cluster Resources and SiCortex.

Recommended batch system configuration The backend should work with a normal installation using only SLURM or SLURM+moab/maui. Do not keep nodes with different amount of memory in the same queue.

Configuration commands in `arc.conf` The SLURM back-end requires that the following options are specified:

`lrms`=*"SLURM"* – in the `[common]` section enables the SLURM batch system back-end.

`slurm_bin_path`=*path* – in the `[common]` section must be set to the path of the SLURM binaries.

Known limitations If you have nodes with different amount of memory in the same queue, this will lead to miscalculations. If SLURM is stopped, jobs on the resource will get canceled, not stalled. The SLURM backend is only tested with SLURM 1.3, it should however work with 1.2 as well.

4.4.3 Enabling the cache

The A-REX can cache input files, so that subsequent jobs requiring the same file don't have to wait for downloading it again: the cached file will be symlinked (or copied) into the session directory of the job (but only after the permissions of this user and the modification date of the file are checked).

Enabling caching is as simple as providing a directory path with the `cachedir` configuration command in the `[grid-manager]` section and turning on the cache cleaning mechanism with the `cachesize` command:

```
cachedir=path
cachesize=high_mark low_mark
```

Here `path` points to a directory which will be used by the A-REX to store the cached files. A-REX will create this directory when the first job is submitted, it should be owned by the same user as which the A-REX is running. The size of the cache directory is maintained by removing the least recently accessed files. If the cache size exceeds a given percentage ("high mark") of the available space, the oldest files will be removed until the size goes below another given percentage ("low mark").

A sample section is shown here:

```
[grid-manager]
user="root"
controldir="/tmp/control"
sessiondir="/tmp/session"
mail="grid.support@somewhere.org"
joblog="/tmp/gm-jobs.log"
securetransfer="no"
cachedir="/tmp/cache"
cachesize="80 70"
```

It is possible to use more than one cache directory by simply specifing more than one `cachedir` command in the configuration file. When multiple caches are used, a new cache file will go to a randomly selected cache where each cache is weighted according to the size of the file system on which it is located (e.g. if there are two caches of 1TB and 9TB then on average 10% of input files will go to the first cache and 90% will go to the second cache).

By default the files will be soft-linked into the session directory of the job. If it is preferred to copy them (because e.g. the cache directory is not accessible from the worker nodes), a dot (.) should be added after the path:

```
cachedir="path ."
```

If the cache directory is accessible from the worker nodes but on a different path, then this path can be specified also:

```
cachedir="path link_path"
```

With large caches mounted over NFS and an A-REX heavily loaded with data transfer processes, cache cleaning can become slow, leading to caches filling up beyond their configured limits. For performance reasons it may be advantageous to disable cache cleaning by the A-REX (by removing the `cachesize` command from the config), and run the *cache-clean* tool independently on the machine hosting the file system. (Please refer to Section **5.3**, *Administration tools*.)

Caches can be added to and removed from the configuration as required without affecting any cached data, but after changing the configuration file, the A-REX should be restarted. If a cache is to be removed and all data erased, it is recommended that the cache be put in a *draining* state until all currently running jobs possibly accessing files in this cache have finished. This can be done by putting the word "drain" as the *link_path*:

```
cachedir="path drain"
```

For more details about the mechanisms of the cache, please refer to Section **6.4**, *Cache*.

4.4.3.1 The Cache Service

The ARC caching system automatically saves to local disk job input files for use with future jobs. The cache is completely internal to the computing element and cannot be accessed or manipulated from the outside. The ARC Cache Service exposes various operations of the cache and can be especially useful in a pilot job model where input data for jobs is not known until the job is running on the worker node.

It is packaged as `nordugrid-arc-cache-service`, and it can either be started with its own init script, or it can be configured to run in the same container as A-REX. For more information about the cache service, please visit the NorduGrid wiki:

`http://wiki.nordugrid.org/index.php/Cache_Service`

4.4.3.2 The ARC Cache Index (ACIX)

There is another option for locating already cached files: the ARC Cache Index (ACIX). It consists of two components, one on the computing resource: the *Cache Server*, and the *Index Server* which indexes the cache locations retrieved from the Cache Servers. Installation instructions can be found in the `README.txt` files in the NorduGrid subversion:

`http://svn.nordugrid.org/trac/nordugrid/browser/contrib/acix/trunk`

The Cache Server periodically scans the A-REX cache and constructs a Bloom filter of cache content. This filter is a way of representing the cache content in an extremely compressed format, which allows fast query of any element of the filter and efficient upload of the content to an index server. This type of compression however has the possibility of giving false-positives, i.e. a certain file may appear to be present in the cache according to the filter when it is not. The Cache Server runs in an HTTPS server and the filter is accessible at the endpoint `https://hostname:5443/data/cache`. It does not require any configuration (but make sure port 5443 is open in the firewall) and it uses the caches specified in the A-REX arc.conf.

The Index Server runs independently of the Cache Servers and A-REX, but can be deployed on the same host as both of them. It is configured with a list of Cache Servers and periodically pulls the cache filter from each one. It runs within an HTTPS server through which users can query the cached locations of files. Configuration uses the regular arc.conf file in the section `[acix/indexserver]`. Here Cache Servers are specified by the `cacheserver` option. For example:

```
[acix/indexserver]
cacheserver="https://my.host:5443/data/cache"
cacheserver="https://another.host:5443/data/cache"
```

The Index Server can be queried at the endpoint `https://hostname:6443/data/index` and the list of URLs to check are given as comma-separated values to the option “url” of this URL, e.g.

```
https://hostname:6443/data/index?url=http://www.nordugrid.org:80/data/echo.sh,\
 http://my.host/data1,lfc://lfc.org/grid/data2
```

A JSON-formatted response is returned, consisting of a dictionary mapping each URL to a list of locations. Any HTTP client can be used to query for cache locations and this makes it easy to use data-based brokering for ARC jobs. (It is planned to write an ACIX-based broker plugin for ARC).

Figure 4.5 shows an example ACIX set up. Each CE runs a Cache Server and there is a central Index Server which pulls content from all CEs. In addition there is one site with two CEs, CE 1a and CE 1b. In order to do data-based brokering on just those two sites (and ease the load on the global Index Server), a local Index Server is running which pulls content from only these two sites.

4.4.4 Configuring Data Staging

The CE is responsible for collecting input data for a job before submission to the LRMS, and for staging out data after the job has finished. The component which performs data staging is called DTR (Data Transfer Reloaded). Its architecture is shown in Figure 4.6.

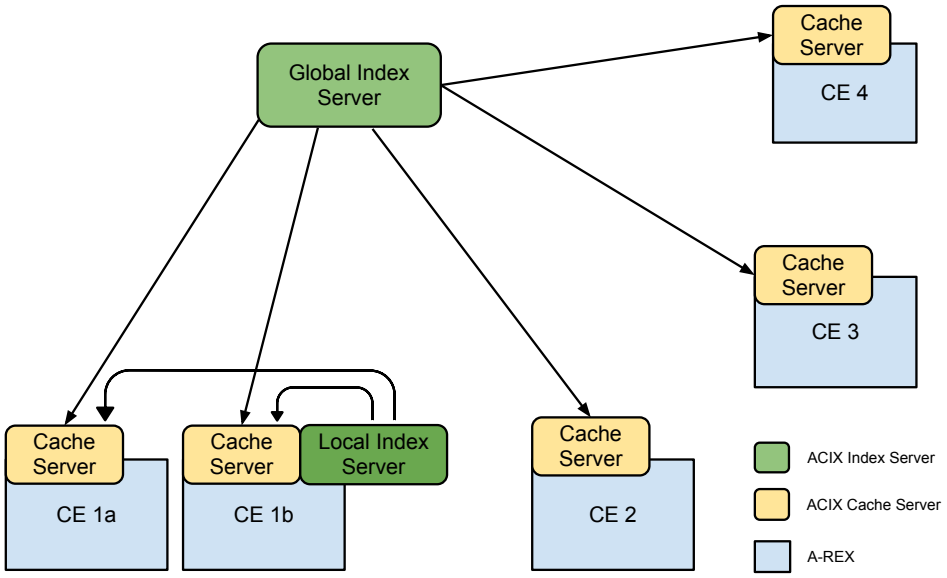


Figure 4.5: ACIX deployment scenario, with one global Index Server and a local Index Server for CE 1a and CE 1b.

A-REX sends each job that requires data staging before or after execution to the Generator, which constructs a Data Transfer Request (DTR) per file that needs to be transferred. These DTRs are sent to the Scheduler for processing. The Scheduler sends DTRs to the Pre-processor for anything that needs to be done up until the physical transfer takes place (e.g. cache check, resolve replicas) and then to Delivery for the transfer itself. Once the transfer has finished the Post-processor handles any post-transfer operations (e.g. register replicas, release requests). The number of slots available for each component is limited, so the Scheduler controls queues and decides when to allocate slots to specific DTRs, based on the prioritisation algorithm implemented.

DTR configuration is specified in the [data-staging] section, and each parameter is explained in detail in Section **6.1.13**, *Commands in the [data-staging] section.* Reasonable (conservative) default values exist which allow safe operation without any configuration being set, but it is better to tune values according to each set up. Example:

```
[data-staging]
maxdelivery="40"
maxprocessor="20"
maxemergency="2"
maxprepared="200"
sharetype="voms:role"
definedshare="myvo:production 80"
definedshare="myvo:student 20"
```

DTR also features a priorities and shares system, as well as the ability to distribute data transfers over multiple nodes. For more information on this and all other aspects of DTR, please consult the data staging page of the NorduGrid wiki[§].

4.4.5 How to join the grid: registering to an index service

Once a cluster is setup, it needs to communicate to some index service to join the grid. Joining an index will let clients query the index to find the CE without specifying the CE hostname on the command line.

In the grid world, this is crucial as the user is agnostic about the server his/her jobs will run.

[§]http://wiki.nordugrid.org/index.php/Data_Staging

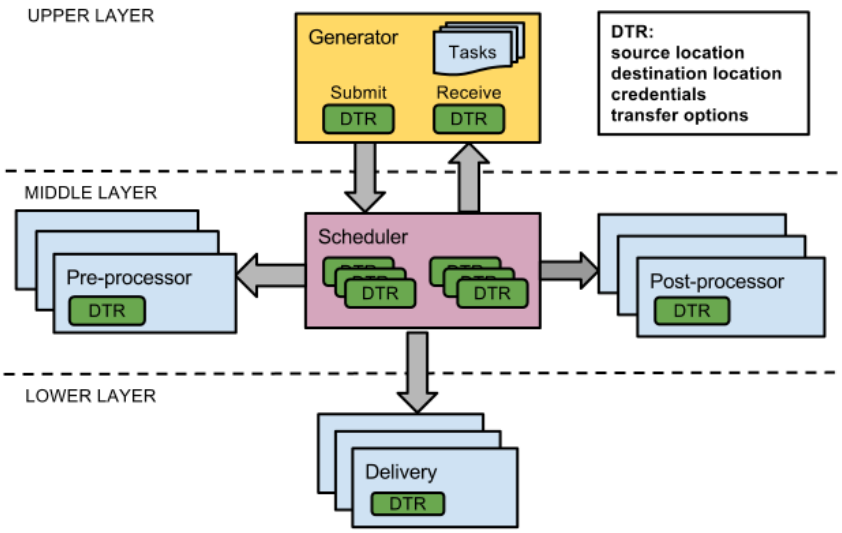


Figure 4.6: The architecture of DTR.

Connection to an index will enable resource sharing in a federated way, among users accepted by the rules in the `[group]` and `[vo]` sections.

National Grid Infrastructures usually run their own index, and NorduGrid runs several:

ldap://index1.nordugrid.org:2135

ldap://index2.nordugrid.org:2135

ldap://index3.nordugrid.org:2135

To connect to an index, add the following to a basic CE configuration file, **after** all the other existing `[infosys]` related sections:

```
...
[infosys/cluster/registration/toPGS1]
targethostname="quark.hep.lu.se"
targetport="2135"
targetsuffix="mds-vo-name=PGS,o=grid"
regperiod="300"
...
```

The special section name `[infosys/cluster/registration/toIndex]` is used to configure registration of a **cluster** (a CE) to an **index service** (an IS).

Registration commands explained:

`targethostname`=*`FQDN`* – The FQDN of the host running the target index service.

`targetport`=*`portnumber`* – Port where the target Index Service is listening. Defaults to 2135.

`targetsuffix`=*`ldapsuffix`* – ldap suffix of the target index service. This has to be provided by a manager of the index service, as it is a custom configuration value of the Index Service. Usually is a string of the form `"mds-vo-name=<custom value>,o=grid"`

`regperiod`=*`seconds`* – the registration script will be run each number of *seconds*. Defaults to 120.

These commands will affect the way the registration script is run. Logs about registration information can be found by looking at the file configured by the `registrationlog` command in the `[infosys]` section (see Section **4.3.5**, *The [infosys] section: the local information system*). For information on how to read the logs see Section **5.4**, *Log files*

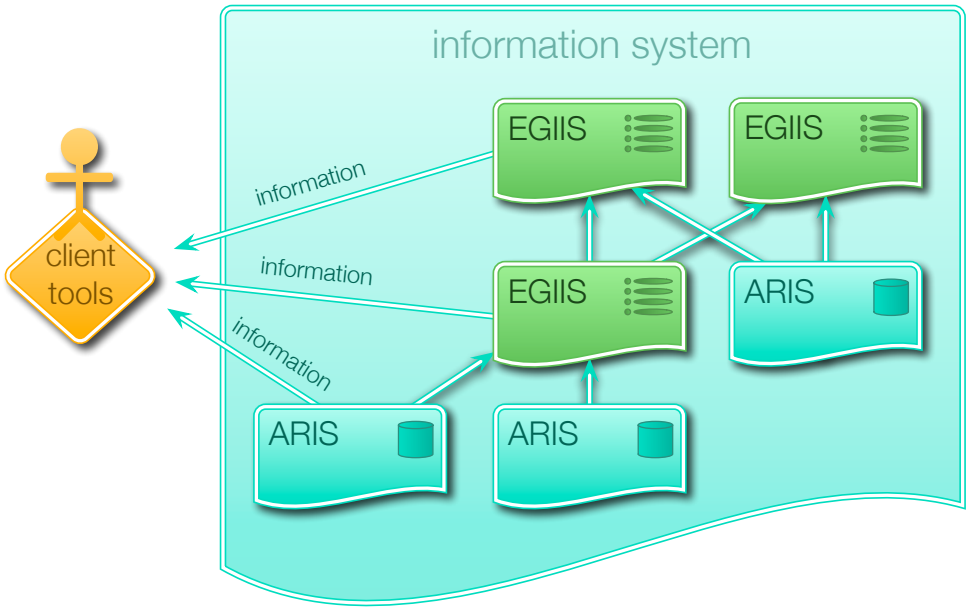


Figure 4.7: The components of the ARC information system: the ARIS which sits next to a computing element (or a storage resource) and advertises information about it; and the EGIIS which indexes the location of ARISes and other EGIIS, creating a hierarchical information mash, where querying the top nodes would provide information about all the resources.

The registration script is called `grid-info-soft-register`. Once registration to an index is configured, parameters of this script can be checked on the system by issuing at the shell:

```
[root@piff tmp]# ps aux | grep reg
root     29718  0.0  0.0  65964  1316 pts/0    S    14:36   0:00
  /bin/sh /usr/share/arc/grid-info-soft-register
  -log /var/log/arc/inforegistration.log
  -f /var/run/arc/infosys/grid-info-resource-register.conf -p 29710

root     29725  0.0  0.0  66088  1320 pts/0    S    14:36   0:00
  /bin/sh /usr/share/arc/grid-info-soft-register
  -log /var/log/arc/inforegistration.log
  -register -t mdsreg2 -h quark.hep.lu.se -p 2135 -period 300
  -dn Mds-Vo-Op-name=register, mds-vo-name=PGS,o=grid -daemon
  -t ldap -h piff.hep.lu.se -p 2135 -ttl 600
  -r nordugrid-cluster-name=piff.hep.lu.se,Mds-Vo-name=local,o=Grid
  -T 45 -b ANONYM-ONLY -z 0 -m cachedump -period 0
```

Other less relevant options are available for registration, please refer to Section **6.1.11**, *Commands in the [infosys/cluster/registration/registrationname] subsections.*

If the registration is successful, the cluster will be shown on the index. To find out that, please refer to the Index Service documentation [**?**].

4.4.6 Accounting with JURA

The A-REX can be configured to periodically execute an external usage reporting utility which should create standard-complient usage records from job usage information provided by the A-REX (called the "job log files") and send the records to remote accounting services. The JURA is such an external utility which capable of doing this. It is distributed with the A-REX.

JURA is capable of creating two types of usage records from the job log files:

- Usage Record 1.0 (UR) XML format [**?**]
- Compute Accounting Record (CAR) XML format [**?**]

After creating these usage records, JURA can archive them to a given directory and it can send them to remote services:

- The UR can be sent to an SGAS LUTS (Logging and Usage Tracking Service) [**?**]
- Experimental feature: The CAR usage record can be sent to the new version of APEL or any other service supporting the format.

To enable reporting, the `jobreport` and the `jobreport_publisher` configuration commands in the `[grid-manager]` section has to be set to the URL of an accounting destination and the name of the executable publisher. Multiple URLs can be specified in one `jobreport` command, and multiple `jobreport` commands can be used. The usage record of each job will be reported to all of the destinations. Currently if the URL starts with "`CAR:`", then a Compute Accounting Record (CAR) will be created, and logged, but it will not be sent anywhere. When the URL starts with "`APEL:`", then a Compute Accounting Record (CAR) will be reported directly to APEL. If a HTTPS URL is given, then then a Usage Record 1.0 will be created and sent to an SGAS LUTS destination. (The experimental APEL support can be utilized by running JURA separately and specifying a "topic" with the `-t` command line options.) A number can be specified after the URLs: how many days the job log files will be kept if the reporting fails.

The credentials for the HTTPS connection should be set using the `jobreport_credentials` command, specifying first the path to the key then the path to the certificate and the path to the CA certificates directory, separated by space.

Additional options can be given to JURA in the form of comma-separated `key:value` pairs by setting the `jobreport_options` configuration command. Currently these options are recognized:

- **urbatch**:size – JURA sends usage records not one-by-one, but in batches. This options sets the maximum size of a batch. Zero value means unlimited batch size. Default is 50.
- **archiving**:dir – JURA can archive the generated usage records to a given directory. This options specifies the directory and enables archiving. If the directory does not exist, an attempt is made to create it. If this option is absent, no archiving is performed.
- **topic**:name of the topic – here can be set a name of the APEL topic where would like to publish an accounting records. When not set this option then a JURA will be use a default APEL topic.
- **gocdb_name**:GOCDB name of CE – here can be set a GOCDB name of the resource that would be see as Site attribute in the generated APEL record. When not set this option then a JURA will be use a hostname (uppercase, "." replaced with "-") as default Site name.

An example configuration which will report all jobs to both destinations using the given credentials, sending them in batches of 50, and archiving them into `var/urs`:

```
[grid-manager]
jobreport="https://luts1.grid.org:8443/wsrf/services/sgas/LUTS"
jobreport="https://luts2.grid.org:8443/wsrf/services/sgas/LUTS 7"
jobreport="APEL:https://apel.cern.ch:2170"
jobreport_publisher="jura"
jobreport_credentials="/etc/grid-security/hostkey.pem
       /etc/grid-security/hostcert.pem /etc/grid-security/certificates"
jobreport_options="urbatch:50,archiving:/var/urs,topic:/queue/cpu,
                   gocdb_name:SE-NGI-CE-GOCDB-NAME"
```

For the configuration commands, see also **6.1.12.7**, *Commands related to usage reporting.*

It is also possible to run JURA separately from the A-REX (e.g. a cron job can be set to execute it periodically). The command line options of JURA are the following:

```
jura -E <days> -u <url> -t <topic> -o <path> <control dir>
```

- `-E <days>` – for how many days should failed-to-send records be kept
- `-u <url>` – runs JURA in "interactive mode", which sends usage reports only to the URLs given as command line arguments (and not those which were put into the job log files by the A-REX), and does not delete job log files after a successful report. Multiple `-u` can be given.
- `-t <topic>` – after each `-u <url>` a topic can be specified. This topic is needed for publishing to APEL. If the URL does not start with "CAR" and a topic is specified, the report will be sent to APEL, if a topic is not specified, the report will be sent to SGAS.
- `-o <path>` – specifies the path of the archiving directory, which will be used only for this run of JURA, and the usage records will be put into this directory.
- `<control dir> [<control dir> ...]` – one or more control directories has to be specified. JURA looks for the job log files in the "logs" subdirectory of the control directories given here.

For more details about JURA, see **6.6**, *JURA: The Job Usage Reporter for ARC*.

4.4.7 Sending usage records to SGAS with urlogger

The `urlogger` component of the A-REX is capable of generating and sending job usage records to the SGAS [**?**] accounting service.

The following libraries need to be installed:

- Python 2.4 or later
- Twisted Core and Web (http://twistedmatrix.com/)
- PyOpenSSL (https://launchpad.net/pyopenssl)
- ElementTree (http://effbot.org/zone/element-index.htm - only needed with Python 2.4)

Debian/Ubuntu package names:

`python-twisted-core, python-twisted-web, python-openssl`
`python-elementtree` (only needed with Python 2.4)

RPM based distributions (e.g., RHEL, CentOS, SL, Fedora, etc.):

`python-twisted-core python-twisted-web pyOpenSSL`
`python-elementtree` (only needed with Python 2.4)

The `urlogger` reads its configuration from the `arc.conf`. It should be specified for the A-REX that the `urlogger` generator script should be run for each job, so the following should be put into the `[grid-manager]` section:

```
[grid-manager]
authplugin="FINISHED timeout=10,onfailure=pass /usr/libexec/arc/arc-ur-logger %C %I %S %U"
```

The plugin will log to the file: `/var/log/arc/ur-logger.log` (configurable). This file will not appear until a job has finished.

Then the SGAS service should be specified in a `[logger]` section:

```
[logger]
log_all="https://sgas.ndgf.org:6143/sgas"
```

This will send all records to the given address. It is possible to specify separate SGASes for separate VOs:

```
log_vo="bio.ndgf.org https://biosgas.ndgf.org:6143/sgas"
```

For more options, see Section **6.1.21**, *Commands for the `urlogger` accounting component.*

The A-REX needs to be restarted after the configuration is finished.

A cron job should be set up to send the usage records to SGAS periodically, e.g.:

```
0 * * * * /usr/libexec/arc/arc-ur-registrant
```

To ensure that the registrant is working, you can run the script from the command line first. Note that the script will still write its log to /var/log/arc/ur-registration.log (configurable). By running the script with -s its ouput will be directed to stdout.

4.4.8 Monitoring the ARC CE: Nagios probes

Nagios scripts (probes) exist that allow monitoring of ARC-CEs. The scripts are available in the EGI repository[¶].

NorduGrid provides a set of Nagios tests that can be used to monitor the functionality of an ARC computing element. These tests were originally developed by the NDGF in order to provide availability monitoring to WLCG. The maintenance of the tests has since been taken over by the EMI project.

The tests are available in the workarea of the nordugrid subversion server:

`http://svn.nordugrid.org/trac/workarea/browser/nagios`

They are also available packaged as an RPM: `grid-monitoring-probes-org.ndgf`.

The configuration of the tests is collected in one configuration file called org.ndgf.conf. Make sure that the user configured to run the tests is authorized at the CEs under test and has the necessary access rights to the storage locations and catalogues configured.

Some of the tests send test jobs to the CE and will report the result when the test job has finished. If the job does not complete within 12 hours it will be killed and a warning is reported in Nagios.

More information about the tests can be found here:

`http://wiki.nordugrid.org/index.php/Nagios_Tests`

4.5 Enhancing CE capabilities

Once a basic CE is in place and its basic functionalities have been tested, is possible to add more features to it.

These include:

Enable glue1.2/1,3, GLUE2 LDAP schemas To be compliant with other grid systems and middlewares, ARC CE can publish its information in these other schemas. In this way its information can show up also in information systems compliant with gLite [**?**]. ARC CE can act as a resource-BDII, to be part of a site-BDII and join the European grid.
See Section **4.5.1**, *Enabling or disabling LDAP schemas*

Provide customized execution environments on-demand As every experiment can have its own libraries, dependencies and tools, ARC provides a means of creating such environments on demand for each user. This feature is called Runtime Environment (RTE). See Section **4.5.2**, *Runtime Environments.*

Use web services instead/together with of GridFTPd/LDAP Next generation ARC Client and servers are Web Service ready. Job submission and the Information System can now be run as a single standardized service using the https protocol. See Section **4.5.3**, *Enabling the Web Services interface.*

[¶] `https://wiki.egi.eu/wiki/EMI_Nagios_probes`

4.5.1 Enabling or disabling LDAP schemas

ARIS, the cluster information system, can publish information in three schemas and two protocols. Information published via the LDAP protocol can follow the following three schemas:

NorduGrid Schema The default NorduGrid schema, mostly used in Nordic countries and within all the NorduGrid Members. Definition and technical information can be found in [**?**].

Glue 1.2 / 1.3 schema Default currently used by gLite middleware[**?**] and the European grids. Specifications can be found here: [**? ?**].

GLUE 2 schema Next generation glue schema with better granularity. Will be the next technology used in production environments. Specification can be found here: [**?**].

The benefits of enabling these schemas are the possibility to join grids other than NorduGrid, for example to join machines allotted to do special e-Science experiments jobs, such as the ATLAS experiment[**?**].

To enable or disable schema publishing, the first step is to insert the enable commands in the `[infosys]` section as explained in **6.1.5**, *Commands in the [infosys] section.*

The Glue 1.2/1.3 schemas carry geographical information and have to be configured in a separate section, **`[infosys/glue12]`**.

If the nordugrid-arc-doc package is installed, two arc.conf examples are available in

`/usr/share/doc/nordugrid-arc-doc/examples/`

Glue 1.2/1.3 `arc_computing_element_glue12.conf`

Glue 2 `arc_computing_element_glue2.conf`

More examples can be found on svn:

`http://svn.nordugrid.org/repos/nordugrid/doc/trunk/examples/`

An example configuration of the [infosys/glue12] section is given in Figure 4.8.

```
[infosys/glue12]
resource_location="Somewhere, Earth"
resource_latitude="54"
resource_longitude="25"
cpu_scaling_reference_si00="2400"
processor_other_description="Cores=1,Benchmark=9.8-HEP-SPEC06"
glue_site_web="http://www.eu-emi.eu"
glue_site_unique_id="MINIMAL Infosys configuration"
provide_glue_site_info="true"
```

Figure 4.8: An example [infosys/glue12] configuration section

Explanation of the commands can be found in the technical reference, section **6.1.7**, *Commands in the [infosys/glue12] section.*

For the GLUE 2.0 it is enough set the command to enable. However, there are other options to let the system administrator configure more features, like the AdminDomain information used for a cluster to join a domain that might be distributed across different geographical sites. A minimal example is detailed in Figure 4.9 and it just contains the domain name. For detailed information please see **6.1.6**, *Commands in the [infosys/admindomain] section.*

```
[infosys/admindomain]
name="ARC-TESTDOMAIN"
```

Figure 4.9: An example [infosys/admindomain] configuration section

4.5.1.1 Applying changes

Once arc.conf is modified, restart the infosystem as explained in Section **5.1**, *Starting and stopping CE services.*

To test information is being published, follow the instructions in Section **5.2.1**, *Testing the information system.*

4.5.2 Runtime Environments

A general description of Runtime Environments (RTEs) can be found in Section **1.7**, *Application software in ARC: The RunTime Environments.*

The A-REX can run specially prepared *BASH* scripts prior to creation of the job's script, before and after executing job's main executable. These scripts are usually grouped in a directory and called RTE scripts.

To configure a RTE, it is enough to add to the [grid-manager] block the following:

```
runtimedir="/SOFTWARE/runtime"
```

where `/SOFTWARE/runtime` is a directory that contains different RTEs, usually organized in different directories.

Each RTE SHOULD have its own directory containing its scripts. A proposal on how to organize such directories can be seen here: `http://pulse.fgi.csc.fi/gridrer/htdocs/concept.phtml` .

It is important that each directory is replicated or accessible by all the computing nodes in the LRMS that are intended to use those Runtime Environments. A-REX will scan each directory and identify the different RTEs.

A specific set of scripts for an RTE is requested by client software in the job description, through the *runtimeenvironment* attribute in XRSL, JSDL or ADL, with a value that identifies the name of the RTE.

The scripts are run with first argument set to '0','1' or '2', and executed in specific moments of the job's lifetime, in this way:

'0' is passed during creation of the job's LRMS submission script. In this case the scripts are run by A-REX on the frontend, before the job is sent to the LRMS. Some enviroment variables are defined in this case, and can be changed to influence the job's execution later. A list is presented in table 4.2.

'1' is passed before execution of the main executable. The scripts are executed on the computing node of the LRMS. Such a script can prepare the environment for some third-party software package. The current directory in this case is the one which would be used for execution of the job. Variable `$HOME` also points to this directory.

'2' is passed after the main executable has finished. The scripts are executed on the computing node of the LRMS. The main purpose is to clean possible changes done by scripts run with '1' (like removing temporary files). Execution of scripts on computing nodes is in general not reliable: if the job is killed by LRMS they most probably won't be executed.

If the job description specifies additional arguments for corresponding RTE those are appended starting at second position.

The scripts all are run through BASH's 'source' command, and hence can manipulate shell variables.

For a description on how to organize and create a RTE, please follow the instructions here: `http://pulse.fgi.csc.fi/gridrer/htdocs/maintainers.phtml`

For publicly available runtime environments please see the RTE Registry at `http://pulse.fgi.csc.fi/gridrer/htdocs/index.phtml`.

4.5.3 Enabling the Web Services interface

A-REX provides a standard-compliant Web Service (WS) interface to handle job submission/management. The WS interface of A-REX is however disabled by default in ARC and EMI distributions as of 2011. To experiment with this advanced A-REX feature, setting the option `arex_mount_point` in the `[grid-manager]` section of `arc.conf` enables the web service interface, e.g.

```
arex_mount_point="https://your.host:60000/arex"
```

Remember to enable incoming and outgoing traffic in the firewall for the chosen port; in the example above, port 60000.

Then jobs can be submitted through this new WS interface with the `arcsub` command (available in the ARC client package) and jobs can be managed with other `arc*` commands.

A-REX also has an EMI Execution Service interface. To enable it, in addition to the above option the following option must be specified

```
enable_emies_interface="yes"
```

IMPORTANT: this web service interface does not accept legacy proxies created by `voms-proxy-init` by default. RFC proxies must be used, which can be created by specifying `voms-proxy-init -rfc` or using `arcproxy`.

The WS interface can run alongside the GridFTP interface. Enabling the WS interface as shown above does not disable the GridFTP interface - if desired "gridftpd" service must be explicitly stopped.

4.5.4 Virtual Organization Membership Service (VOMS)

Classic authentication of users in grid environment is based on his/her certificate subject name (SN). Authorization of users is performed by checking the lists of permitted user SNs, also known as grid-mapfiles. The classic scheme is the simplest to deal with, but it may have scalability and flexibility restrictions when operating with dynamic groups of researchers – Virtual Organizations (VO).

From the computing element perspective, all members of a particular VO are involved in the same research field having common predictable requirements for resources that allows flexibly configured LRMS scheduler policies. In general, VOs have an internal structure that regulate relationships between members that is implemented via groups, roles and attributes. VO membership parameters are controlled by means of the VOMS specialized software[‖].

VOMS consists of two parts:

- VO Management interface (VOMS-Admin) – web-based solution to control membership parameters. Along with the management interface, the service provides a SOAP interface to generate lists of VO members' SNs. EDG VOMS-Admin is a classic VO Management solution distributed by EMI [**?**]. There is also alternative lightweight solution available – PHP VOMS-Admin [**?**].
- Credentials signing service (vomsd) – standalone daemon that fortifies user VO membership and its parameters. A credentials signing daemon issues an Attribute Certificate (AC) extension attached to the user's proxy-certificate and is used in a delegation process. VOMS processing API of the middleware or some external authorization processing executables may parse and verify the VOMS AC extension and make a decision taking into account group affiliation instead of just using the personal certificate SN.

[‖]There are other existing technologies for group management, but VOMS is the most popular and widely supported

Variable	Description
joboption_directory	session directory of job.
joboption_controldir	control directory of job. Various internal information related to this job is stored in file in this directory under names job.*job_gridid*.*. For more information see section 6.11.
joboption_arg_#	command with arguments to be executed as specified in the JD (**not** bash array).
joboption_arg_code	exit code expected from executable if execution succeeded.
joboption_pre_#_#	command with arguments to be executed before main executable (**not** bash array). There may be multiple such pre-executables numbered from 0.
joboption_pre_#_code	exit code expected from corresponding pre-executable if execution succeeded.
joboption_post_#_#	command with arguments to be executed after main executable (**not** bash array). There may be multiple such post-executables numbered from 0.
joboption_post_#_code	exit code expected from corresponding post-executable if execution succeeded.
joboption_stdin	name of file to be attached to stdin handle.
joboption_stdout	same for stdout.
joboption_stderr	same for stderr.
joboption_env_#	array of 'NAME=VALUE' environment variables (**not** bash array).
joboption_cputime	amount of CPU time requested (minutes).
joboption_walltime	amount of execution time requested (minutes).
joboption_memory	amount of memory requested (megabytes).
joboption_count	number of processors requested.
joboption_runtime_#	array of requested *runtimeenvironment* names (**not** bash array).
joboption_num	*runtimeenvironment* currently beeing processed (number starting from 0).
joboption_jobname	name of the job as given by user.
joboption_lrms	LRMS to be used to run job.
joboption_queue	name of a queue of LRMS to put job into.
joboption_starttime	execution start time as requested in the JD in MDS format.
joboption_gridid	identifier of the job assigned by A-REX. It is an opaque string representing the job inside the A-REX service. It may be not same as the job identifier presented to an external client.
joboption_inputfile_#	local name of pre-staged file (**not** bash array).
joboption_outputfile_#	local name of file to be post-staged or kept locally after execution (**not** bash array).
joboption_localtransfer	if set to 'yes' data staging is done on computing node.
joboption_nodeproperty_#	array of properties of computing nodes (LRMS specific, **not** bash array).

For example *joboption_arg_#* could be changed to wrap the main executable. Or *joboption_runtime* could be expanded if the current one depends on others.

Table 4.2: RTEs predefined environment variables when the scripts are run with option '0'

To maintain the grid-mapfiles based on information in the VOMS database (using the SOAP interface of the VO Management service), use `voms://` or `vomss://` sources in the [vo] configuration block for the `nordugridmap` utility (see section **6.1.2**, *Commands in the [vo] section* for details).

A VOMS credentials signing daemon is used directly by client tools (see `arcproxy` manual) to create a VOMS AC-enabled proxy. The computing element does not interact with the credentials signing daemon directly, but verifies the digital signature of the VOMS server against a configured list of trusted VOMS AC issuers instead.

All general VOMS-related configurations described below are supported by the ARC API as well as other EMI products based on classic VOMS libraries.

4.5.4.1 Configuring trusted VOMS AC issuers

The VOMS AC signature included in a client's proxy certificate can be verified in two ways:

1. Get the issuing VOMS server certificate to trust beforehand and use it for signature verification.
2. Configure the lists of certificates (LSC) to verify the certificate chain in the VOMS AC.

In case of errors detected in VOMS AC processing, A-REX behavior depends on the `voms_processing` configuration variable (see Section **6.1.12.2**, *Commands affecting the A-REX Web Service communication interface* and **6.1.4.1**, *General commands*).

Getting the VOMS server certificate. This was historically the first method of VOMS server signature verification based on retrieval of the server public key.

THIS CONFIGURATION METHOD IS NOW OBSOLETE AND UNSUPPORTED SINCE ARC 1.0.0!

Among all EGI-supported grid services there are only few that do not support LSC files configuration – glite-FTS and glite-WMS for gLite 3.1. If legacy VOMS credentials setup is required for those services, please refer to appropriate documentation.

Configure lists of certificates. The trust chain from the Certificate Authority (CA) to the VOMS AC issuing server certificate needs to be described in order to verify ACs in clients' proxies issued by that server. Generally, the VOMS server certificate is signed by the CA directly, so there is only two certificates in the chain of trust, but it can be much longer in other cases.

Chains of trust are configured in *.LSC files. Each line of the LSC file lists a single certificate SN starting from the VOMS server and continues up the trust chain ending with the root CA certificate SN. The following is an example of an LSC file for the voms.ndgf.org server:

```
/O=Grid/O=NorduGrid/CN=host/voms.ndgf.org
/O=Grid/O=NorduGrid/CN=NorduGrid Certification Authority
```

In some rare cases (e.g. host certificate change and/or moving to different CA) it is possible to specify several lists per hostname, separating them with `------ NEXT CHAIN ------` line (the ARC parser uses `NEXT CHAIN` match only, but classic VOMS libraries require exactly six dashes and space around it, so it is better to put it there for compatibility).

The following is an example of several chains in a single LSC-file:

```
/DC=org/DC=ugrid/O=hosts/O=KNU/CN=host/grid.org.ua
/DC=org/DC=ugrid/CN=UGRID CA
------ NEXT CHAIN ------
/DC=org/DC=ugrid/O=hosts/O=KNU/CN=grid.org.ua
/DC=org/DC=ugrid/CN=UGRID CA
```

To get the trust chain of SNs for the VOMS server either contact the VO manager or use `openssl` for known VOMS servers:

```
echo | openssl s_client -connect <server:port> 2>/dev/null \
| openssl x509 -noout -subject -issuer
```

Here `<port>` is typically the standard VOMS-Admin https interface port – 8443. Port of the `vomsd` daemon listed in the `vomses` file can also be used.

The location of LSC files for ARC is fixed and compatible with other EMI software's default setup:

```
/etc/grid-security/vomsdir/<VO>/<hostname>.lsc
```

Creation of an additional LSC file or modifying an old can be performed without A-REX restart.

Another ARC-specific way exists to configure trust chains without creation of *.LSC files for each VOMS server – define `voms_trust_chain` configuration options that contain information about all trusted issuers in one place.

This approach is more useful with A-REX standalone installations that provide resources for a few VOs. In contrast the LSC files based solution is more scalable and compatible with other EMI software.

These variables can be specified in the `[common]` configuration block and extended in `[grid-manager]` and/or `[gridftpd]` blocks:

```
voms_trust_chain="/O=Grid/O=NorduGrid/CN=host/arthur.hep.lu.se" "/O=Grid/O=NorduGrid/CN=NorduGrid Certification Authority"
voms_trust_chain="/O=Grid/O=NorduGrid/CN=host/emi-arc.eu" "/O=Grid/O=NorduGrid/CN=NorduGrid Certification Authority"
voms_trust_chain="^/O=Grid/O=NorduGrid"
```

NOTE! A defined `voms_trust_chain` option will override the information in *.LSC files.

Unlike LSC files the `voms_trust_chain` option supports regular expressions syntax. After `voms_trust_chain` modification services should be restarted to apply changes.

4.5.4.2 Configuring VOMS AC signing servers to contact

Clients rely on VOMSes configuration. VOMSes refers to a list of VOMS servers that are used to manage the supported VOs, more precisely speaking – VOMS AC signing daemons' contact parameters.

The old way of specifying VOMSes is to put all VOs configuration into a single file `/etc/vomses`. Each line should be written in the following format:

```
"alias" "host address" "TCP port" "host certificate SN" "official VO name"
```

It is advised to have *alias* the same as *official VO name*: several VOMS client versions mix them. If several VOMS servers are used by the VO for redundancy, specify them on separate lines. These parameters can be found in the "Configuration" section of VOMS-Admin labeled "VOMSES string for this VO".

With recent versions of grid software it is possible to maintain a separate VOMSes files for each VO. This files should be placed in the VOMSes directory – `/etc/grid-security/vomses/` is used by default but can be redefined with `X509_VOMSES` environmental variable. Please refer to client documentation for more information. For example, to configure support of the `nordugrid.org` VO, create a file `/etc/grid-security/vomses/nordugrid.org` with the following content:

```
"nordugrid.org" "voms.ndgf.org" "15015" "/O=Grid/O=NorduGrid/CN=host/voms.ndgf.org" "nordugrid.org"
```

4.5.4.3 Configuring ARC to use VOMS extensions

From the client side, `arcproxy` already has built-in support for VOMS AC extensions, so no additional configuration is required unless it is desired to redefine VOMSes path.

To utilize VOMS AC extensions in A-REX there are several possibilities:

- using an access control filter based on VOMS AC (see section **4.4.1**, *Access control: users, groups, VOs* for details)
- using LCAS/LCMAPS authorization and mapping (see section **4.5.7**, *Using LCAS/LCMAPS* for details)
- using external plugins that operate with VOMS AC (e.g. `arc-vomsac-check`)

4.5.5 Dynamic vs static mapping

There are many debates on using static or dynamic local account mapping policy. Historically, ARC initially supported only static mapping. Currently, ARC and all middlewares involved in the EMI project support and can be configured to use any combination of the two.

4.5.5.1 Static mapping

The main reason of using a static account mapping policy is to *simplify administration of grid services*. Static mapping works by assigning a fixed operating system account to a grid user identified by his/her SN. General practice is to map all grid users to one or a few operating system accounts dedicated to grid jobs.

The most significant drawback of sharing local accounts is that different grid users are *indistinguishable* for the underlying system infrastructure. There is no easy way to securely isolate different jobs running with the same credentials or implement complex scheduling policy in the LRMS with reservations and priorities as well as employ flexible disk space allocation policy.

On the other hand, if every grid user is mapped to a dedicated local account, there is significant increase of administration burden. Individual mappings and their permissions may need to be manually synchronized with grid user directories (like VOMS or Globus CAS).

4.5.5.2 Dynamic mapping

A dynamic mapping policy allows to provide every grid user with a separate dynamically leased local account and to deploy more secure and flexible configurations. Generally, dynamic mapping involves using multiple pools of local accounts for different classes of grid users.

Common examples of such classes include VOs and groups/roles in the VOs. This allows for building authorization and mapping policies in terms of VOMS FQANs additionally to user SNs, which is very common as a site usually provides resources for more than one VO.

Each grid user accessing some site service gets mapped to a local account which is leased from an appropriate pool. Policy rules define how to select that pool depending on the VOMS AC presented by the user as a part of his/her proxy-certificate. A user accessing the same site with different credentials generally will be mapped differently, depending on FQANs included.

Each grid user gets separated from other local and grid users by means of the underlying operating system because with dynamic mapping every grid user is mapped to a dedicated local account. If the local account lease is not used for some period of time, it is released and can be assigned to another grid user.

Pool accounts can belong to specific local groups which can be a subject of LRMS scheduling policy or disk quotas. Authorization and mapping policies should be updated only in the case when a new role or group is introduced in a VO, update in case of user membership changes is not necessary.

There are different approaches to the implementation of a dynamic mapping policy, including:

- deploying the ARC built-in `simplepool` mapping plugin
- using LCMAPS from Site Access Control framework (see section **4.5.7**, *Using LCAS/LCMAPS*)
- using the Argus dedicated authorization service (see section **4.5.6**, *Using Argus authorization service*)
- using any third-party solution which can be implemented through a call to an external executable

Please note that to completely disable static mapping, an empty grid-mapfile needs to be specified in the configuration. This is needed because users are always mapped to accounts in the grid-mapfile by default. And because the grid-mapfile is used as the primary authorization list by default the option `allowunknown="yes"` must be specified in the `[gridftpd]` section to turn that check off.

Also for security purposes it is advisable to always provide a fallback mapping rule to map the user to a safe or nonexiting local account in case all the dynamic mapping rules failed for some reason.

4.5.6 Using Argus authorization service

A-REX with the Web Service (WS) interface enabled (see section **4.5.3**, *Enabling the Web Services interface*) may directly use the Argus service [**?**] for requesting authorization decisions and performing client mapping to a local user account. To make A-REX communicate to Argus PEP or PDP service for every operation requested through WS interface add the following option to the `[grid-manager]` section of `arc.conf`:

```
arguspep_endpoint="https://arguspep.host:8154/authz"
```

or

```
arguspdp_endpoint="https://arguspdp.host:8154/authz"
```

A-REX can use different XACML profiles for communicating to Argus. Available are

- direct - pass all authorization attributes (only for debugging). No deployed Argus service implements this profile.
- subject - pass only subject name of client. This is a simplified version of the 'cream' profile.
- cream - makes A-REX pretend it is a gLite CREAM service. This is currently the recommended profile for interoperability with gLite based sites.
- emi - a new profile developed in the EMI project. This is the default choice.

Example:

```
arguspep_profile="cream"
```

or

```
arguspdp_profile="cream"
```

To choose whether the username of the local account provided by Argus PEP should be accepted, the `arguspep_usermap` option is used. By default the local account name provided by Argus is ignored. This can be changed by setting

```
arguspep_usermap="yes"
```

Although a corresponding option for Argus PDP server exists, the Argus PDP server itself does not provide a local user identity in its response yet.

IMPORTANT: note that first mapping rules defined in the `[grid-manager]` section are processed and then Argus is contacted. Hence the account name provided by Argus will overwrite the one defined by local rules.

IMPORTANT: although direct communication with the Argus PEP server is only possible for a WS enabled A-REX server it is possible to use Argus command line utilities as authorization and account mapping plugins in the `[grid-manager]` section of the configuration file. For example:

```
  [grid-manager]
  authplugin="ACCEPTED timeout=20 pepcli_wrapper.sh %C/job.%I.proxy"

Content of pepcli_warpper.sh:
  #!/bin/sh
  pepcli --pepd https://arguspep.host:8154/authz --certchain "$1" -v --cert \
    /etc/grid-security/hostcert.pem --key /etc/grid-security/hostkey.pem \
    --capath /etc/grid-security/certificate | grep -F "Permit"
```

The example above uses the authplugin feature of A-REX to perform authorization for the job submission operation. More sophisticated scenarios may be covered by a more complex pepci-wrapper.sh. For more information see Argus documentation [**?**] and description of various plugins sections: **6.1.3**, *Commands in the [group] section*, **6.1.4**, *Commands in the [gridftpd] section* and **6.1.12.8**, *Other general commands in the [grid-manager] section.*

4.5.7 Using LCAS/LCMAPS

LCAS stands for Local Centre Authorization Service. Based on configured policies, LCAS makes binary authorization decisions. Most of LCAS functionality is covered by ARC's internal authorization mechanism (see section **6.1.3**, *Commands in the [group] section*), but it can be used for interoperability to maintain a common authorization policy across different Grid Middlewares.

LCMAPS stands for Local Credential Mapping Service, it takes care of translating Grid credentials to Unix credentials local to the site. LCMAPS (as well as LCAS) is modular and supports flexible configuration of complex mapping policies. This includes not only classical mapping using a grid-mapfile generated by `nordugridmap` (see section **4.4.1**, *Access control: users, groups, VOs*) but primarily using dynamic pools and VOMS AC-based mapping using FQAN match which differs in some aspects from the functionality provided by ARC natively. LCMAPS can be used to implement VO integration techniques and also for interoperability to maintain a common account mapping policy.

LCAS/LCMAPS libraries are provided by the Site Access Control (SAC) framework [**?**] that was originally designed to be called from the gLite middleware stack and the pre-WS part of Globus Toolkit version 4. ARC can also be configured to employ these libraries.

The main goal of using SAC is to maintain common authorization and Unix account mapping policies for a site and employ them on multiple services of a site. The framework allows to configure site-wide autorization policies independently of the contacted service and consistent identity mapping among different services that use LCAS/LCMAPS libraries, e.g. A-REX, LCG CE (GT4), CREAM CE or GSISSH.

Additionally, the SAC framework provides the SCAS mapping service, an ARGUS client and the gLExec enforcement executable. More information about its functionality and configuration can be found in the SAC documentation [**? ? ?**].

4.5.7.1 Enabling LCAS/LCMAPS

LCAS and LCMAPS can be used by configuring them in the corresponding sections of the configuration file and will be used by the `gridftpd` jobplugin or fileplugin and the A-REX WS interface. To avoid undesired behavior of the SAC framework - changing user identity of running process, use and manipulation of environment variables, etc. - which is harmful for a multithreaded execution environment, mediator executables are used called `arc-lcas` and `arc-lcmaps` correspondingly. They are located at `<ARCinstallation path>/libexec/arc` and are invoked by ARC services with grace 60 seconds timeout to avoid hanging connections. Both executables invoke appropriate functions from shared libraries (usually `liblcas.so` and `liblcmaps.so` respectively), so LCAS/LCMAPS must be installed to use it. Installing the SAC framework is not covered by this manual, please refer to the corresponding EMI documentation [**? ?**].

Although the advised way to use LCAS and LCMAPS is through corresponding dedicated authorization and mapping rules it is also possible to use the generic plugin capability of ARC and call those executables directly. Their arguments syntax is the same as one of the corresponding configuration rules with two additional arguments prepended - subject name of user and path to file containing user credentials. Credentials must

include the full chain of user credentials with optional CA certificate. If file containing X.509 proxy is used its private key is ignored.

Using LCAS LCAS is configured in the `[group]` section using an `lcas` authorization rule. This command requires several parameters:

```
lcas=<LCAS library name> <LCAS library path> <LCAS policy description file>
```

The corresponding system command to call the mediator executable is

```
arc-lcas <user subject> <user credentials> <LCAS library name> <LCAS library path> \
             <LCAS policy description file>
```

This command can be invoked manually to check the desired operation of LCAS.

The user subject and credentials path can be substituted by A-REX using `%D` and `%P` syntax. It is also necessary to pass the LCAS library name and path to the SAC installation location. The syntax of the LCAS policy description file is provided later in this section.

Enabling LCAS in `arc.conf` example:

```
[group/users]
lcas="liblcas.so /opt/glite/lib /etc/lcas.db"

[gridftpd/jobs]
groupcfg="users"
path="/jobs"
plugin="jobplugin.so"
```

And if using authorization plugin functionality section `[group/users]` can be written

```
[group/users]
plugin="5 /opt/arc/libexec/arc/arc-lcas %D %P liblcas.so /opt/glite/lib /etc/lcas.db"
```

As one can see this syntax may be used to achieve an even higher degree of flexibility by tweaking more parameters.

Using LCMAPS LCMAPS is configured with an `lcmaps` rule for one of the identity mapping commands - `unixmap`, `unixgroup` or `unixvo` - in the `[gridftpd]` section. This rule requires several parameters:

```
lcmaps <LCMAPS library name> <LCMAPS library path> <LCMAPS policy description file> \
    <LCMAPS policy name> [<LCMAPS policy name>...]
```

The corresponding system command to call the mediator executable is

```
arc-lcmaps <user subject> <user credentials> <LCMAPS library name> \
    <LCMAPS library path> <LCMAPS policy description file> \
    <LCMAPS policy name> [<LCMAPS policy name>...]
```

An LCMAPS policy description file can define multiple policies, so additional `LCMAPS policy name` parameter(s) are provided to distinguish between them. The syntax of LCMAPS policy description is provided later in this section.

Enabling LCMAPS in `arc.conf` example:

```
[gridftpd]
gridmap="/dev/null"
allowunknown="yes"
unixmap="* lcmaps liblcmaps.so /opt/glite/lib /etc/lcmaps.db voms"
```

And if using generic plugin functionality section `unixmap` command can be written

```
unixmap="* mapplugin 30 /opt/arc/libexec/arc/arc-lcmaps %D %P liblcmaps.so \
    /opt/glite/lib /etc/lcmaps.db voms"
```

4.5.7.2 LCAS/LCMAPS policy configuration

LCAS and LCMAPS provide a set of plugins to be used for making the policy decisions. All configuration is based on the plugins used and their parameters.

LCAS configuration To create an access control policy using LCAS, the following set of basic plugins is needed:

lcas_userallow.mod allows access if SN of the user being checked is listed in the config file provided.

lcas_userban.mod denies access if SN of the user being checked is listed in the config file provided.

lcas_voms.mod checks if FQANs in user's proxy certificate VOMS AC match against config file provided.

lcas_timeslots.mod makes authorization decisions based on available time slots (as mentioned in LCAS documentation "the most useless plugin ever" :-))

The LCAS configuration file (`lcas.db`) contains several lines with the following format:

```
pluginname="<module name/path to plugin file>", pluginargs="<arguments>"
```

Each line represents an authorization policy rule. A positive decision is only reached if all the modules listed permit the user (logical AND).

LCMAPS configuration LCMAPS plugins can belong to one of two classes, namely acquisition and enforcement. Acquisition modules gather the information about user credentials or find mapping decisions that determine the user's UID, primary GID and secondary GIDs that can then be assigned by enforcement modules.

LCMAPS basic acquisition modules:

lcmaps_localaccount.mod uses account name corresponding to user's SN in static mapfile (mostly like classic grid-mapfile).

lcmaps_poolaccount.mod allocates account from a pool corresponding to user's SN in static mapfile (like grid-mapfile with "dot-accounts" for Globus with `GRIDMAPDIR` patch).

lcmaps_voms.mod parses and checks proxy-certificate VOMS AC extension and then fills internal LCMAPS data structures with that parsed information, which can be used by other plugins invoked later.

lcmaps_voms_localaccount.mod uses static UID value corresponding to user's VOMS FQAN.

lcmaps_voms_localgroup.mod uses static GID value corresponding to user's VOMS FQAN.

lcmaps_voms_poolaccount.mod allocates account from a pool corresponding to user's VOMS FQAN.

lcmaps_voms_poolgroup.mod allocate GID from a pool corresponding to user's VOMS FQAN.

lcmaps_scas_client.mod passes request to a SCAS server for making the decision.

LCMAPS basic enforcement modules:

lcmaps_posix_enf.mod sets UID/GID by POSIX `setreuid()/setregid()` calls so that the LCMAPS caller process after successful enforcement continues running with credentials of an account mapped.

lcmaps_ldap_enf.mod change an information about an account in the LDAP database (uidnumber, gidnumber, memberuid, etc.).

lcmaps_dummy_good.mod does not perform enforcing and returns success in case the mapping was found.

The LCMAPS configuration file (`lcmaps.db`) is more complex than LCAS one due to flexibility of policies.

```
# define path to plugable modules
path = /path/to/lcmaps/modules
# define actions
<action1 name> = "<module1 name> [<module1 options>]"
<action2 name> = "<module2 name> [<module2 options>]"
...
<actionM name> = "<moduleN name> [<moduleN options>]"
# define policies
<policy1 name>:
<actionX1> -> <action on success> [| <action on fault>]
<actionX2> -> <action on success> [| <action on fault>]
...
<actionXN> -> <action on success> [| <action on fault>]
...
<policyN name>:
<actionY1> -> <action on success> [| <action on fault>]
<actionY2> -> <action on success> [| <action on fault>]
...
<actionYN> -> <action on success> [| <action on fault>]
```

After specifying the path to LCMAPS modules, several actions need to be defined. Each action can be either an acquisition or enforcement action, depending on the specific module used. If a module requires parameters, they are specified just after the module filename.

Then defined actions are combined into sequences defining the mapping policy. The first line after policy name starts the sequence. A module defined by action from the left side of the arrow "->" is executed and depending on the execution result (positive or negative) another action gets called. The action sequence ends on enforcement module execution.

To find more information about available pluggable modules and their configuration options, please follow the LCMAPS documentation [**?**].

Environment variables To fine-tune or debug LCAS/LCMAPS framework operation, special environmental variables should be used. There is no another way to change e.g. debug level.

LCAS environmental variables:

LCAS_LOG_FILE sets location of the logfile

LCAS_LOG_TYPE determines method of logging (`logfile`, `syslog`, `both` or `none`)

LCAS_LOG_STRING specifies text to be prepended to each line to be logged

LCAS_DB_FILE specifies location of lcas policy file (either absolute or relative to `LCAS_DIR`)

LCAS_DEBUG_LEVEL sets debug level (0-5)

LCAS_MOD_DIR sets location of the LCAS plugins (`/modules` will be added to the end of value specified)

LCAS_DIR sets location of LCAS configuration files

LCAS_ETC_DIR can be used alternatively to `LCAS_DIR` for the same purpose

LCMAPS enviromental variables:

LCMAPS_LOG_FILE sets location of the logfile

LCMAPS_LOG_TYPE determines method of logging (`logfile`, `syslog`, `both` or `none`)

LCMAPS_LOG_STRING specifies text to be prepended to each line to be logged

LCMAPS_DB_FILE specifies location of lcas policy file (either absolute or relative to `LCMAPS_DIR`)

LCMAPS_DEBUG_LEVEL sets debug level (0-5)

LCMAPS_MOD_DIR sets location of the LCMAPS plugins (/modules will be added to the end of value specified)

LCMAPS_DIR sets location of LCMAPS configuration files

LCMAPS_ETC_DIR can be used alternatively to LCMAPS_DIR for the same purpose

LCMAPS_POLICY_STRING determines the list of policies to apply from a configuration file

4.5.7.3 Example LCAS configuration

Here is an example of LCAS configuration file:

```
pluginname=lcas_userban.mod,pluginargs=/etc/grid-security/lcas/ban_users.db
pluginname=lcas_voms.mod,pluginargs="-vomsdir /etc/grid-security/vomsdir/"
" -certdir /etc/grid-security/certificates/"
" -authfile /etc/grid-security/voms-user-mapfile"
" -authformat simple"
```

There are two modules used: lcas_userban.mod and lcas_voms.mod. The list of particular users to ban (their certificate SNs) is stored in the file /etc/grid-security/lcas/ban_users.db that is passed to lcas_userban.mod.

If the user's certificate SN is not directly banned, then VO membership check is performed by lcas_voms.mod. The plugin accepts several parameters: vomsdir and certdir paths used to check proxy-certificate and VOMS AC extension; authfile contains allowed FQANs specified in a format set by authformat.

Example content of /etc/grid-security/voms-user-mapfile:

```
"/dteam" .dteam
"/dteam/Role=lcgadmin" .sgmdtm"
"/dteam/Role=NULL/Capability=NULL" .dteam
"/dteam/Role=lcgadmin/Capability=NULL" .sgmdtm
"/VO=dteam/GROUP=/dteam" .dteam
"/VO=dteam/GROUP=/dteam/ROLE=lcgadmin" .sgmdtm
"/VO=dteam/GROUP=/dteam/ROLE=NULL/Capability=NULL" .dteam
"/VO=dteam/GROUP=/dteam/ROLE=lcgadmin/Capability=NULL" .sgmdtm
```

Only the first parameter (FQAN) is used. The second parameter is valuable only for LCMAPS, when it is configured to use the same file. The several FQAN specification formats are used to support different versions of the VOMS library. If the latest VOMS library (later than version 2.0.2 from EMI-1) is installed on a site then just the first two lines are enough, but to keep things safe and support older VOMS, all of them should be given.

A GACL format of authfile can also be used as well as more options and plugins. Please refer LCAS documentation for more information.

4.5.7.4 Example LCMAPS configuration

LCMAPS configuration for ARC is not an enforcing configuration (it means that LCMAPS does not actually apply UID/GID assignment on execution), so the lcmaps_dummy_good.mod plugin must be used to accomplish this. The arc-lcmaps executable returns the user name and optionally group name to stdout which is then used by ARC to perform enforcing by itself.

Simple gridmap behavior For gridmap behaviour the lcmaps_localaccount.mod plugin can be used with a grid-mapfile, where the users are mapped to some Unix account(s).

Example lcmaps.db configuration file:

```
path = /opt/glite/lib/modules
# ACTIONS
```

```
# do not perform enforcement
good = "lcmaps_dummy_good.mod"
# statically mapped accounts
localaccount = "lcmaps_localaccount.mod"
" -gridmapfile /etc/grid-security/grid-mapfile"

# POLICIES
staticmap:
localaccount -> good
```

There is only one policy `staticmap` defined: after `localaccount` action is called, LCMAPS execution gets finished.

VOMS AC-based mapping to pools Parsing the VOMS AC is accomplished via the `lcmaps_voms` family of plugins. Account pools and `gridmapdir` should be created beforehand.

```
path = /opt/glite/lib/modules
# ACTIONS
# do not perform enforcement
good = "lcmaps_dummy_good.mod"
# parse VOMS AC to LCMAPS data structures
vomsextract = "lcmaps_voms.mod"
" -vomsdir /etc/grid-security/vomsdir"
" -certdir /etc/grid-security/certificates"
# FQAN-based pool account mapping
vomspoolaccount = "lcmaps_voms_poolaccount.mod"
" -override_inconsistency"
" -max_mappings_per_credential 1"
" -do_not_use_secondary_gids"
" -gridmapfile /etc/grid-security/voms-user-mapfile"
" -gridmapdir /etc/grid-security/gridmapdir"
# FQAN-based group mapping
vomslocalgroup = "lcmaps_voms_localgroup.mod"
" -groupmapfile /etc/grid-security/voms-group-mapfile"
" -mapmin 1"

#POLICIES
voms:
vomsextract -> vomspoolaccount | good
vomspoolaccount -> vomslocalgroup | good
vomslocalgroup -> good
```

Configuration requires `voms-group-mapfile` which maps FQANs to groups and `voms-user-mapfile` which maps FQANs to accounts from pools. Directories `vomsdir` and `certdir` in `vomsextract` configuration are used to check VOMS AC validity.

Example content of `/etc/grid-security/voms-user-mapfile` is provided in section **4.5.7.3**, *Example LCAS configuration*. The second parameter indicates the Unix account used to accomplish mapping for specified FQAN. Notice the dot prepending an account name – that means that a free pool account will be used instead of a single account.

For example, there are 50 pool accounts named `dteam01`, `dteam02` ...`dteam50`. Specifying `.dteam` in `voms-user-mapfile` means that LCMAPS needs to get any unused account from the pool and assign it to the user's SN:FQAN pair. Already leased accounts are tracked by hard-linking the account file to a url-encoded SN:FQAN pair file in the `gridmapdir`.

File `voms-group-mapfile` is simillar to `voms-user-mapfile`, but the second parameter indicates a group name. If the parameter has a dot prepended like in `voms-user-mapfile`, it defines the name of the pool of groups that can be used with the `lcmaps_voms_poolgroup.mod` plugin.

There is only one policy defined – `voms`. Action `vomsextract` gets executed first and on success `vomspoolaccount` module is used. If validation and parsing the VOMS AC has failed then LCMAPS execution finishes. When

called `vomspoolaccount` allocates an account from the pool, LCMAPS moves on to finding an appropriate group by `vomslocalgroup` action. LCMAPS execution is finished if `vomspoolaccount` had no success and after `vomslocalgroup` has finished operation successfully.

Chapter 5

Operations

5.1 Starting and stopping CE services

5.1.1 Overview

There are three components needed for a production level CE to work:

- `gridftpd` : Starts the gridftpd interface. Brings up the server (configured in the `[gridftpd]` block) and all the services related to it (configured in all the `[gridftpd/subsection]` blocks).
  Usually located in `/etc/init.d/`
  See Section **4.3.4**, *The [gridftpd] section: the job submission interface* for configuration details.
- `a-rex` : Starts A-REX, the grid manager (configured in the `[grid-manager]` block). It prepares the configuration files and starts the `arched` hosting environment process.
  Starts the Web Services interface **only if** it has been enabled. See Section **4.5.3**, *Enabling the Web Services interface*
  Starts LRMS scripts. See Section **4.4.2**, *Connecting to the LRMS* for configuration details.
  Usually located in `/etc/init.d/`
  See Section **4.3.3**, *The [grid-manager] section: setting up the A-REX and the arched* for configuration details.
- `grid-infosys` : Starts the LDAP server and the infosystem scripts (configured in the `[infosys]` configuration block and its subsections).
  Usually located in `/etc/init.d/`
  See Section **4.3.5**, *The [infosys] section: the local information system* for configuration details.

5.1.2 Starting the CE

To start a CE, issue the following commands with root rights in the following order:

1. `# service gridftpd start`
2. `# service a-rex start`
3. `# service grid-infosys start`

Alternatively the exact same procedure can be used calling the scripts directly:

1. `# /etc/init.d/gridftpd start`
2. `# /etc/init.d/a-rex start`
3. `# /etc/init.d/grid-infosys start`

Note: If ARC-related environment variables are set, for example $ARC_LOCATION or $ARC_CONFIG, then the second form must be used in order to pass those variables through to the script.

5.1.3 Stopping the CE

To stop a CE, issue the following commands with root rights in the following order:

1. `# service grid-infosys stop`
2. `# service a-rex stop`
3. `# service gridftpd stop`

Alternatively the exact same procedure can be used calling the scripts directly:

1. `# /etc/init.d/grid-infosys stop`
2. `# /etc/init.d/a-rex stop`
3. `# /etc/init.d/gridftpd stop`

5.1.4 Verifying the status of a service

To check the status of a service, issue the command:

```
# service <servicename> status
```

Alternatively the exact same procedure can be used calling the scripts directly:

```
# /etc/init.d/<servicename> status
```

where `<servicename>` is one of `gridftpd`, `a-rex`, `grid-infosys`
Depending on the security configuration, root permissions might be needed to execute these commands.
A CE is fully funtional when all the three scripts return an OK status.

5.2 Testing a configuration

This chapter gives instructions on how to test and troubleshoot that a given configuration is correct, and that everything is running properly.

Things to check are, in order of importance:

1. **The information system is running and publishing the correct information.** Without a properly configured information system, the clients will not be able to query the cluster for its resources and do an efficient brokering.
   See Section **5.2.1**, *Testing the information system*
2. **A-REX is running with valid certificates installed**.
   See Section **5.2.2**, *Testing whether the certificates are valid*
3. **The job submission interface is listening and accepting jobs**.
   See Section **5.2.3**, *Testing the job submission interface*
4. **LRMS configuration is correct and a job can be executed on the queues**.
   See Section **5.2.4**, *Testing the LRMS*

5.2.1 Testing the information system

The ARC-CE information system publishes in LDAP and WebServices/XML format.

To test if the LDAP information system is running, ldap tools must be installed. In particular the tool called `ldapsearch` [**?**].

To test if the WS information system is running, ARC suggests its own tool called `arcwsrf` [**?**].

5.2.1.1 Check NorduGrid Schema publishing

To check if the information system is creating the needed ldap trees and publishing them, issue the following:

```
ldapsearch -x -H ldap://localhost:2135 -b 'mds-vo-name=local,o=grid'
```

and the result should be something like the one in Figure 5.1.

To check that the information system is publishing **outside** the cluster, i.e. on its public IP, execute the same query on its hostname, preferrably from a remote machine:

```
ldapsearch -x -H ldap://<hostname>:2135 -b 'mds-vo-name=local,o=grid'
```

The result must be the similar to the one in Figure 5.1.

All the values must be consistent with the setup. For example, nordugrid-cluster-name must be the machine's hostname.

5.2.1.2 Check Glue 1.x Schema publishing

To check if the information system is creating the needed ldap trees and publishing them, issue the following:

```
ldapsearch -x -H ldap://localhost:2135 -b 'mds-vo-name=resource,o=grid'
```

and the result should be something like the one in Figure 5.2.

To check that the information system is publishing **outside** the cluster, i.e. on its public IP, execute the same query on its hostname, preferrably from a remote machine:

```
ldapsearch -x -H ldap://<hostname>:2135 -b 'mds-vo-name=resource,o=grid'
```

The result must be the similar to the one in Figure 5.2.

```
# extended LDIF
#
# LDAPv3
# base <mds-vo-name=local,o=grid> with scope subtree
# filter: (objectclass=*)
# requesting: ALL
#

# local, Grid
dn: Mds-Vo-name=local,o=Grid
objectClass: Mds
objectClass: MdsVo
Mds-Vo-name: local
Mds-validfrom: 20110811172014Z
Mds-validto: 201108111820l4Z

# piff.hep.lu.se, local, grid
dn: nordugrid-cluster-name=piff.hep.lu.se,Mds-Vo-name=local,o=grid
nordugrid-cluster-totalcpus: 2
nordugrid-cluster-homogeneity: TRUE
nordugrid-cluster-name: piff.hep.lu.se
nordugrid-cluster-lrms-version: 0.9
nordugrid-cluster-middleware: nordugrid-arc-1.0.1
nordugrid-cluster-middleware: globus-5.0.3
nordugrid-cluster-trustedca: /O=Grid/O=NorduGrid/CN=NorduGrid Certification Au
 thority
nordugrid-cluster-cpudistribution: 2cpu:1
nordugrid-cluster-sessiondir-lifetime: 10080
nordugrid-cluster-issuerca: /DC=eu/DC=KnowARC/CN=LUEMI-1310134495.12
nordugrid-cluster-credentialexpirationtime: 20110807141455Z
nordugrid-cluster-lrms-type: fork
nordugrid-cluster-sessiondir-free: 129566
nordugrid-cluster-sessiondir-total: 143858
nordugrid-cluster-architecture: x86_64
nordugrid-cluster-prelrmsqueued: 0
nordugrid-cluster-comment: This is a minimal out-of-box CE setup
nordugrid-cluster-contactstring: gsiftp://piff.hep.lu.se:2811/jobs
nordugrid-cluster-issuerca-hash: 8050ebf5
nordugrid-cluster-totaljobs: 0
nordugrid-cluster-aliasname: MINIMAL Computing Element
nordugrid-cluster-usedcpus: 0
objectClass: Mds
objectClass: nordugrid-cluster
Mds-validfrom: 20110811172104Z
Mds-validto: 201108111722O4Z

# fork, piff.hep.lu.se, local, grid
dn: nordugrid-queue-name=fork,nordugrid-cluster-name=piff.hep.lu.se,Mds-Vo-nam
 e=local,o=grid
nordugrid-queue-running: 0
```

Figure 5.1: Output of an ldapsearch on a CE

```
 ldapsearch -x -h piff.hep.lu.se -p 2135 -b 'mds-vo-name=resource,o=grid'
# extended LDIF
#
# LDAPv3
# base <mds-vo-name=resource,o=grid> with scope subtree
# filter: (objectclass=*)
# requesting: ALL
#

# resource, Grid
dn: Mds-Vo-name=resource,o=Grid
objectClass: Mds
objectClass: MdsVo
Mds-Vo-name: resource
Mds-validfrom: 20110822130627Z
Mds-validto: 20110822140627Z

# piff.hep.lu.se, resource, grid
dn: GlueClusterUniqueID=piff.hep.lu.se,Mds-Vo-name=resource,o=grid
objectClass: GlueClusterTop
objectClass: GlueCluster
objectClass: GlueSchemaVersion
objectClass: GlueInformationService
objectClass: GlueKey
GlueClusterUniqueID: piff.hep.lu.se
GlueClusterService: piff.hep.lu.se
GlueSchemaVersionMinor: 2
GlueForeignKey: GlueCEUniqueID=piff.hep.lu.se:2811/nordugrid-fork-arc
GlueForeignKey: GlueSiteUniqueID=MINIMAL Infosys configuration
GlueSchemaVersionMajor: 1
GlueClusterName: MINIMAL Infosys configuration

# MINIMAL Infosys configuration, resource, grid
dn: GlueSiteUniqueID=MINIMAL Infosys configuration,Mds-Vo-name=resource,o=grid
GlueSiteDescription: ARC-This is a minimal out-of-box CE setup
GlueSiteSecurityContact: mailto: -1
objectClass: GlueTop
objectClass: GlueSite
objectClass: GlueKey
objectClass: GlueSchemaVersion
GlueSiteSysAdminContact: mailto: -1
GlueSiteName: MINIMAL Infosys configuration
GlueSiteUniqueID: MINIMAL Infosys configuration
GlueSchemaVersionMinor: 2
GlueSiteLongitude: 25
GlueSiteLatitude: 54
GlueSchemaVersionMajor: 1
GlueForeignKey: None
GlueSiteOtherInfo: Middleware=ARC
GlueSiteUserSupportContact: mailto: -1
GlueSiteWeb: http://www.eu-emi.eu
GlueSiteLocation: Somewhere, Earth

# piff.hep.lu.se:2811/nordugrid-fork-arc, resource, grid
dn: GlueCEUniqueID=piff.hep.lu.se:2811/nordugrid-fork-arc,Mds-Vo-name=resource
 ,o=grid
GlueCEStateStatus: Production
GlueCEStateTotalJobs: 0
GlueCEInfoJobManager: arc
GlueCEInfoHostName: piff.hep.lu.se
GlueCEUniqueID: piff.hep.lu.se:2811/nordugrid-fork-arc
GlueCEStateFreeJobSlots: 2
GlueForeignKey: GlueClusterUniqueID=piff.hep.lu.se

   ...

# search result
search: 2
result: 0 Success

# numResponses: 9
# numEntries: 8
```

Figure 5.2: Sample glue 1.x infosystem output on a ldap query. The output has been shortened for ease of reading.

5.2.1.3 Check LDAP GLUE2 Schema publishing

To check if the information system is creating the needed ldap trees and publishing them, issue the following:

```
ldapsearch -x -H ldap://localhost:2135 -b 'o=glue'
```

and the result should be something like the one in Figure 5.3.

To check that the information system is publishing **outside** the cluster, i.e. on its public IP, execute the same query on its hostname, preferrably from a remote machine:

```
ldapsearch -x -H ldap://<hostname>:2135 -b 'o=glue'
```

The result must be the similar to the one in Figure 5.3.

5.2.1.4 Check WS/XML GLUE2 Schema publishing

First a proxy certificate is needed and these credentials must be authorised on the CE to test, see [**?**].

Call the `arcwsrf` test tool:

```
$ arcwsrf https://<hostname>:<a-rex port>/<a-rex path>
```

where `<a-rex port>` `<a-rex path>` are those specified in Section **4.5.3**, *Enabling the Web Services interface*.

The output should look like in Figure 5.4

5.2.1.5 Further testing hints

If nothing is published or the query hangs, then there can be something wrong with ldap or A-REX.

Check `slapd` logs to find out the problem in the former case, A-REX logs in the latter. Please see also Section **5.4**, *Log files*.

5.2.2 Testing whether the certificates are valid

While A-REX is running, check the logfile specified with the `logfile` option in the `[grid-infosys]` block in `/etc/arc.conf`:

```
[grid-infosys]
 ...
 logfile="/tmp/grid-manager.log"
 ...
```

It will contain information on expired certificates or certificates about to expire, see Figure 5.5.

While ARIS is running, is possible to get that information as well from its logfiles specified with the `providerlog` option in the `[infosys]` block in `/etc/arc.conf` :

```
[infosys]
...
providerlog="/tmp/infoprovider.log"
...
```

It will contain information about expired certificates, see Figure 5.6.

The certificates' dates can be inspected by using openssl commands. Please refer to the certificate mini How-to

To understand how to read the logs please refer to Section **5.4**, *Log files*

```
$ ldapsearch -x -h piff.hep.lu.se -p 2135 -b 'o=glue'
[...]
# glue
dn: o=glue
objectClass: top
objectClass: organization
o: glue

# urn:ogf:AdminDomain:hep.lu.se, glue
dn: GLUE2DomainID=urn:ogf:AdminDomain:hep.lu.se,o=glue
objectClass: GLUE2Domain
objectClass: GLUE2AdminDomain
GLUE2EntityName: hep.lu.se
GLUE2DomainID: urn:ogf:AdminDomain:hep.lu.se

# urn:ogf:ComputingService:hep.lu.se:piff, urn:ogf:AdminDomain:hep.lu.se, glue
dn: GLUE2ServiceID=urn:ogf:ComputingService:hep.lu.se:piff,
    GLUE2DomainID=urn:ogf:AdminDomain:hep.lu.se,o=glue
GLUE2ComputingServiceSuspendedJobs: 0
GLUE2EntityValidity: 60
GLUE2ServiceType: org.nordugrid.execution.arex
GLUE2ServiceID: urn:ogf:ComputingService:hep.lu.se:piff
objectClass: GLUE2Service
objectClass: GLUE2ComputingService
GLUE2ComputingServicePreLRMSWaitingJobs: 0
GLUE2ServiceQualityLevel: development
GLUE2ComputingServiceWaitingJobs: 0
GLUE2ServiceComplexity: endpoint=1,share=1,resource=1
GLUE2ComputingServiceTotalJobs: 0
GLUE2ServiceCapability: executionmanagement.jobexecution
GLUE2ComputingServiceRunningJobs: 0
GLUE2ComputingServiceStagingJobs: 0
GLUE2EntityName: piff
GLUE2ServiceAdminDomainForeignKey: urn:ogf:AdminDomain:hep.lu.se
GLUE2EntityCreationTime: 2011-08-22T13:23:24Z

# urn:ogf:ComputingEndpoint:piff.hep.lu.se:443,
  urn:ogf:ComputingService:hep.lu.se:piff, urn:ogf:AdminDomain:hep.lu.se, glue
dn: GLUE2EndpointID=urn:ogf:ComputingEndpoint:piff.hep.lu.se:443,
  GLUE2ServiceID=urn:ogf:ComputingService:hep.lu.se:piff,
  GLUE2DomainID=urn:ogf:AdminDomain:hep.lu.se,o=glue
GLUE2ComputingEndpointRunningJobs: 0
GLUE2ComputingEndpointStaging: staginginout
GLUE2EntityValidity: 60
GLUE2EndpointQualityLevel: development
GLUE2EndpointImplementor: NorduGrid
GLUE2EntityOtherInfo: MiddlewareName=EMI
GLUE2EntityOtherInfo: MiddlewareVersion=1.1.2-1
GLUE2EndpointCapability: executionmanagement.jobexecution
GLUE2EndpointHealthState: ok
GLUE2EndpointServiceForeignKey: urn:ogf:ComputingService:hep.lu.se:piff
GLUE2EndpointTechnology: webservice
GLUE2EndpointWSDL: https://piff.hep.lu.se/arex/?wsdl
GLUE2EndpointInterfaceName: ogf.bes
GLUE2ComputingEndpointWaitingJobs: 0
GLUE2ComputingEndpointComputingServiceForeignKey: urn:ogf:ComputingService:hep.lu.se:piff
GLUE2EndpointURL: https://piff.hep.lu.se/arex
GLUE2ComputingEndpointSuspendedJobs: 0
GLUE2EndpointImplementationVersion: 1.0.1
GLUE2EndpointSemantics: http://www.nordugrid.org/documents/arex.pdf
GLUE2ComputingEndpointPreLRMSWaitingJobs: 0
GLUE2EndpointIssuerCA: /DC=eu/DC=KnowARC/CN=LUEMI-1313588355.29
GLUE2EndpointServingState: production
GLUE2ComputingEndpointStagingJobs: 0
objectClass: GLUE2Endpoint
objectClass: GLUE2ComputingEndpoint
GLUE2EndpointInterfaceVersion: 1.0
GLUE2EndpointSupportedProfile: http://www.ws-i.org/Profiles/BasicProfile-1.0.html
GLUE2EndpointSupportedProfile: http://schemas.ogf.org/hpcp/2007/01/bp
GLUE2EndpointImplementationName: ARC
GLUE2EndpointTrustedCA: /DC=eu/DC=KnowARC/CN=LUEMI-1313588355.29
GLUE2EndpointTrustedCA: /O=Grid/O=NorduGrid/CN=NorduGrid Certification Authority
GLUE2ComputingEndpointJobDescription: ogf:jsdl:1.0
GLUE2ComputingEndpointJobDescription: nordugrid:xrsl
GLUE2EndpointID: urn:ogf:ComputingEndpoint:piff.hep.lu.se:443
GLUE2EntityCreationTime: 2011-08-22T13:23:24Z
[...]
# search result
search: 2
result: 0 Success
# numResponses: 6
# numEntries: 5
```

Figure 5.3: Sample LDAP search output on GLUE2 enabled infosystem. The output has been shortened with [...] for ease of reading.

```
<wsrf-rp:GetResourcePropertyDocumentResponse><InfoRoot>
  <Domains xmlns="http://schemas.ogf.org/glue/2008/05/spec_2.0_d41_r01" [...]>
    <AdminDomain BaseType="Domain">
      <ID>urn:ogf:AdminDomain:hep.lu.se</ID>
      <Name>hep.lu.se</Name>
      <Services>
        <ComputingService BaseType="Service" CreationTime="2011-08-22T13:34:56Z" Validity="60">
          <ID>urn:ogf:ComputingService:hep.lu.se:piff</ID>
          <Name>piff</Name>
          <Capability>executionmanagement.jobexecution</Capability>
          <Type>org.nordugrid.execution.arex</Type>
          <QualityLevel>development</QualityLevel>
          <Complexity>endpoint=1,share=1,resource=1</Complexity>
          <TotalJobs>0</TotalJobs>
          <RunningJobs>0</RunningJobs>
          <WaitingJobs>0</WaitingJobs>
          <StagingJobs>0</StagingJobs>
          <SuspendedJobs>0</SuspendedJobs>
          <PreLRMSWaitingJobs>0</PreLRMSWaitingJobs>
          <ComputingEndpoint BaseType="Endpoint" CreationTime="2011-08-22T13:34:56Z" Validity="60">
            <ID>urn:ogf:ComputingEndpoint:piff.hep.lu.se:60000</ID>
            <OtherInfo>MiddlewareName=EMI</OtherInfo>
            <OtherInfo>MiddlewareVersion=1.1.2-1</OtherInfo>
            <URL>https://piff.hep.lu.se:60000/arex</URL>
            <Capability>executionmanagement.jobexecution</Capability>
            <Technology>webservice</Technology>
            <InterfaceName>ogf.bes</InterfaceName>
            <InterfaceVersion>1.0</InterfaceVersion>
            <WSDL>https://piff.hep.lu.se:60000/arex/?wsdl</WSDL>
            <SupportedProfile>http://www.ws-i.org/Profiles/BasicProfile-1.0.html</SupportedProfile>
            <SupportedProfile>http://schemas.ogf.org/hpcp/2007/01/bp</SupportedProfile>
            <Semantics>http://www.nordugrid.org/documents/arex.pdf</Semantics>
            <Implementor>NorduGrid</Implementor>
            <ImplementationName>ARC</ImplementationName>
            <ImplementationVersion>1.0.1</ImplementationVersion>
            <QualityLevel>development</QualityLevel>
            <HealthState>ok</HealthState>
            <ServingState>production</ServingState>
            <IssuerCA>/DC=eu/DC=KnowARC/CN=LUEMI-1313588355.29</IssuerCA>
            <TrustedCA>/DC=eu/DC=KnowARC/CN=LUEMI-1313588355.29</TrustedCA>
            <TrustedCA>/O=Grid/O=NorduGrid/CN=NorduGrid Certification Authority</TrustedCA>
            <Staging>stagininout</Staging>
            <JobDescription>ogf:jsdl:1.0</JobDescription>
            <JobDescription>nordugrid:xrsl</JobDescription>
            <TotalJobs>0</TotalJobs>
            <RunningJobs>0</RunningJobs>
            <WaitingJobs>0</WaitingJobs>
            <StagingJobs>0</StagingJobs>
            <SuspendedJobs>0</SuspendedJobs>
            <PreLRMSWaitingJobs>0</PreLRMSWaitingJobs>
            <Associations>
              <ComputingShareID>urn:ogf:ComputingShare:hep.lu.se:piff:fork</ComputingShareID>
            </Associations>
            <ComputingActivities>
            </ComputingActivities>
          </ComputingEndpoint>
          <ComputingShare BaseType="Share" CreationTime="2011-08-22T13:34:56Z" Validity="60">
            <ID>urn:ogf:ComputingShare:hep.lu.se:piff:fork</ID>
            <Name>fork</Name>
            <Description>This queue is nothing more than a fork host</Description>
            <MappingQueue>fork</MappingQueue>
[...]
            <PreLRMSWaitingJobs>0</PreLRMSWaitingJobs>
            <FreeSlots>2</FreeSlots>
            <FreeSlotsWithDuration>2</FreeSlotsWithDuration>
            <UsedSlots>0</UsedSlots>
            <RequestedSlots>0</RequestedSlots>
            <Associations>
              <ComputingEndpointID>urn:ogf:ComputingEndpoint:piff.hep.lu.se:60000</ComputingEndpointID>
              <ExecutionEnvironmentID>urn:ogf:ExecutionEnvironment:hep.lu.se:piff:fork</ExecutionEnvironmentID>
            </Associations>
          </ComputingShare>
          <ComputingManager BaseType="Manager" CreationTime="2011-08-22T13:34:56Z" Validity="60">
            <ID>urn:ogf:ComputingManager:hep.lu.se:piff</ID>
[...]
          </ComputingManager>
        </ComputingService>
      </Services>
    </AdminDomain>
  </Domains>
</InfoRoot>
</wsrf-rp:GetResourcePropertyDocumentResponse>
```

Figure 5.4: Sample ARC WS information system XML output. The output has been shortened with [...] for ease of reading.

```
...
[2011-08-05 11:12:53] [Arc] [WARNING] [3743/406154336] Certificate /DC=eu/DC=KnowARC/CN=LUEMI-1310134495.12
  will expire in 2 days 5 hours 2 minutes 1 second
[2011-08-05 11:12:53] [Arc] [WARNING] [3743/406154336] Certificate /DC=eu/DC=KnowARC/O=Lund University/CN=demo1
  will expire in 2 days 5 hours 2 minutes 1 second
...
```

Figure 5.5: A sample certificate information taken from A-REX logs.

```
...
[2011-08-12 10:39:46] HostInfo: WARNING: Host certificate is expired in file: /etc/grid-security/hostcert.pem
[2011-08-12 10:39:46] HostInfo: WARNING: Certificate is expired for CA: /DC=eu/DC=KnowARC/CN=LUEMI-1305883423.79
[2011-08-12 10:39:46] HostInfo: WARNING: Certificate is expired for CA: /DC=eu/DC=KnowARC/CN=LUEMI-1301496779.44
[2011-08-12 10:39:46] HostInfo: WARNING: Issuer CA certificate is expired in file:
                                         /etc/grid-security/certificates/8050ebf5.0
[2011-08-12 10:39:46] HostInfo: WARNING: Certificate is expired for CA: /DC=eu/DC=KnowARC/CN=LUEMI-1310134495.12
[2011-08-12 10:39:46] HostInfo: WARNING: Issuer CA certificate is expired in file:
                                         /etc/grid-security/certificates/917bb2c0.0
[2011-08-12 10:39:46] HostInfo: WARNING: Certificate is expired for CA: /DC=eu/DC=KnowARC/CN=LUEMI-1310134495.12
[2011-08-12 10:39:46] HostInfo: WARNING: Certificate is expired for CA: /DC=eu/DC=KnowARC/CN=LUEMI-1305883423.79
[2011-08-12 10:39:46] HostInfo: WARNING: Certificate is expired for CA: /DC=eu/DC=KnowARC/CN=LUEMI-1301496779.44
...
```

Figure 5.6: A sample certificate information taken from ARIS logs.

5.2.3 Testing the job submission interface

To test the job submission interface an ARC Client is needed, such as the arc* tools.

To install an ARC Client refer to `http://www.nordugrid.org/documents/arc-client-install.html`.

Once the clients are installed, the **`arctest`** utility can be used to submit test jobs.

Usage of this tool is out of the scope of this manual. Refer to [**?**] for further information.

To test basic job submission try the following command:

```
arctest -c <hostname fqdn> -J 1
```

The job should at least be submitted succesfully.

5.2.4 Testing the LRMS

Each LRMS has its own special setup. Nevertheless it is good practice to follow this approach:

1. submit a job that includes at least these two lines:

   ```
   ("stderr" = "stderr" )
   ("gmlog" = "gmlog" )
   ```

   The first one will pipe all standard errors to a file called stderr, while the second will generate all the needed debugging information in a folder called `gmlog`.

2. retrieve the job with `arcget -a`.

3. In the job session folder just downloaded, check the `gmlog/errors` file to see what the job submission script was and if there are some LRMS related errors.

The rest of LRMS troubleshooting is LRMS dependent, so please refer to each LRMS specific guide and logs.

5.3 Administration tools

A-REX comes with some administration utilities to help the system administrator. These tools are located at $ARC_LOCATION/libexec/arc ($ARC_LOCATION is normally /usr for standard installation from packages on Linux). Most of the utilities in this directory are for A-REX's own internal use, but the following may also be used by humans:

- *gm-jobs* – displays information related to jobs handled by A-REX. Different types of information may be selected by using various options. This utility also can perform simple management operations - currently cancelling processing of specific jobs and removing them. Default behavior is to print minimal information about all jobs currently handled by A-REX and some statistics. See *gm-jobs -h* for a list of possible options.
- *cache-clean* – This tool is used periodically by A-REX to keep the size of each cache within the configured limits. *cache-clean -h* gives a list of options. The most useful option for administrators is *-s*, which does not delete anything, but gives summary information on the files in the cache, including information on the ages of the files in the cache.
  It is not recommended to run *cache-clean* manually to clean up the cache, unless it is desired to temporarily clean up the cache with different size limits to those specified in the configuration, or to improve performance by running it on the file system's local node as mentioned in **4.4.3**, *Enabling the cache*.
- *cache-list* – This tool is used to list all files present in each cache or, given a list of URLs as arguments, shows the location of each URL in the cache if present. In the first case it simply reads through all the cache *.meta* files and prints to stdout a list of all URLs stored in each cache and their corresponding cache filename, one per line. In the second case the cache filename of each URL is calculated and then each cache is checked for the existence of the file.

5.4 Log files

ARC CE log files paths are configured in `arc.conf` for each component according to the following table:

Component	Configuration section	Default Location	More information
A-REX	`[grid-manager]`	`/var/log/arc/grid-manager.log`	in subsection 4.3.3
gridftpd interface	`[gridftpd]`	`/var/log/arc/gridftpd.log`	in subsection 4.3.4
infoproviders	`[infosys]`	`/var/log/arc/infoprovider.log`	in subsection 4.3.5
information system	`[infosys]`	`/var/log/arc/bdii/bdii-update.log`	in subsection 4.3.5
information registration	`[infosys]`	`/var/log/arc/inforegistration.log`	in subsection 4.3.5
cache cleaning	`[grid-manager]`	`/var/log/arc/cache-clean.log`	in section 5.3

5.4.1 The format of the log files

The format of arc log files is the following:

A-REX	[Date]	[Component name]	[error level]	[pid/thread]	Message
gridftpd	[Date]	[Component name]	[error level]	[pid/thread]	Message
infoproviders	[Date]	infprovider script name:	error level:	Message	
infoprovider registration	Date	pid file of script process	Message		

5.5 Modules of the A-REX

The A-REX consists of several separate modules. These are:

- *libarex.so* – The main module providing main functionality and web interface. It is implemented as HTTP and SOAP service inside HED. It is responsible for processing jobs, moving them through states and running other modules.
- *downloader* – This is a legacy module responsible for gathering input files in the SD. It processes the *job.ID.input* file and updates it. The downloader has been replaced by Data Transfer Reloaded (see Section **4.4.4**, *Configuring Data Staging*).
- *uploader* – This legacy module is responsible for delivering output files to the specified SEs and registration at an Indexing Service (like LFC) as needed. It processes and updates the *job.ID.output* file. The uploader has also been replaced by Data Transfer Reloaded.
- *gm-kick* – Sends a signal to the A-REX though a FIFO file to wake it up. It's used to increase responsiveness of A-REX.
- *CEinfo.pl* – Collects and generates information about computing resource as XML document in NorduGrid and Glue 2 format.

The following modules are always run under the Unix account to which a Grid user is mapped.

- *smtp-send.sh* and *smtp-send* – These are the modules responsible for sending e-mail notifications to the user. The format of the mail messages can be easily changed by editing the simple shell script *smtp-send.sh*.
- *submit-*-job* – Here * stands for the name of the LRMS. Currently supported LRMS are PBS/Torque, Condor, LoadLeveler, LSF, SLURM, and SGE. Also *fork* pseudo-LRMS is supported for testing purposes. This module is responsible for job submission to the LRMS.
- *cancel-*-job* – This script is for canceling jobs which have been already submitted to the LRMS.

- *scan-*-job* -This shell script is responsible for notifying the A-REX about completion of jobs. Its implementation for PBS uses server logs to extract information about jobs. If logs are not available it uses the less reliable *qstat* command for that. Other backends use different techniques.

5.6 Migration of an A-REX service to another host

It is possible to move A-REX and all its managed jobs to another host, if for example the machine or the disk hosting the service has issues and need to be replaced.

When planning migration, the system administrator has to take into account the following facts:

- A-REX does not stop jobs running in the underlying LRMS when it stops. Jobs in the LRMS will continue being processed until they finish or fail. A-REX will check their status when restarted. The outcome of this check depends on LRMS scripts implementation.
- The SD or Cache Directory (see **4.1.2**, *Disk, partitioning, directories*) cannot be moved while there are running jobs. The reason for this is in most setups cached files are accessed via symlinks in the SD, so moving both or just one of them to a new filesystem, eventually with different paths, will cause the soft linking between files will be lost. This restriction does not apply if A-REX is set up so that cached files are copied to the SD.
- In general, A-REX does NOT support hotswapping. SD, CD, Cache Directory, Runtime Environments Scripts directory should not be moved while A-REX is running.

As a consequence of the above facts, A-REX is best migrated when all these conditions hold:

1. All managed grid jobs are finished
2. SD and Cache Directory are on a storage that does **not** belong to the machine from which A-REX will be removed, for example a NFS share and they will be remounted on the target migration machine with the same filesystem paths
3. A-REX is stopped

If SD is hosted on the A-REX machine all grid jobs must be completed **and** results retrieved by the users before migration. If Cache Directory is hosted on the A-REX machine it may be copied to the new machine, otherwise it will be lost during migration. In the procedure described below, it is assumed that SD and Cache directory are on a separate NFS share.

5.6.1 Planned Service Migration

In this scenario, A-REX is planned to be moved from an old cluster (OLD) to a new one (NEW). The service running on OLD can be shut down on a maintenance schedule. SD and Cache directory are on a separate NFS share, that does not reside on OLD.

1. Open arc.conf and add the keyword `drain` to ALL the *sessiondir* configuration commands in the `[gridmanager]` section:

   ```
   [gridmanager]
   ...
   sessiondir="/mnt/grid drain"
   ...
   ```

   See section **6.1.12.3**, *Commands setting control and session directories* for description of the `drain` behaviour.

2. If the `[gridftpd/jobs]` section is present, set `allownew=no` to prevent A-REX accepting new jobs via the *org.nordugrid.gridftpjob* interface.

3. Restart the *a-rex* service. *gridftpd* reads configuration dynamically and does not need to be restarted. At this point A-REX will not accept any new jobs.

4. Wait for A-REX submitted jobs to finish. Checking that A-REX managed jobs are done can be done in three ways:

   - Using the *gm-jobs* command line utility directly on the cluster, and verify there is 0 Running jobs:

     ```
     # /usr/libexec/arc/gm-jobs 2>/dev/null | grep Running
     Running: 0/-1
     ```

   - Using the LDAP information system and ldapsearch command:

     ```
     # ldapsearch -x -LLL -h hostname -p 2135 -b o=glue \
     '(objectclass=GLUE2ComputingService)' GLUE2ComputingServiceRunningJobs

     dn: GLUE2ServiceID=urn:ogf:ComputingService:hostname:arex,GLUE2GroupID=services,o=glue

     GLUE2ComputingServiceRunningJobs: 0
     ```

     Or, using the NorduGrid schema:

     ```
     ldapsearch -x -LLL -h hostname -p 2135 -b mds-vo-name=local,o=grid \
     '(objectclass=nordugrid-cluster)' nordugrid-cluster-totaljobs \
     nordugrid-cluster-prelrmsqueued

     dn: nordugrid-cluster-name=hostname,Mds-Vo-name=local,o=grid
     nordugrid-cluster-totaljobs: 0
     nordugrid-cluster-prelrmsqueued: 0
     ```

5. Shut down the services (in this order): *grid-infosys*, *a-rex*, *gridftpd*.

6. Backup host certificate files and custom grid-mapfile. If you have customized information system scripts, remember to backup those as well.

7. Copy arc.conf and CD from OLD to NEW, and reconfigure arc.conf to point at the correct CD path in NEW if it's different from the previous one. Remove the drain option from SDs.

8. Mount sessiondir(s) and cachedir(s) in NEW. Be sure that NEW has the same permissions, UIDs and GIDs on files as in OLD. This might require some work with the /etc/passwd, /etc/shadow and /etc/group in the NEW to replicate IDs from OLD, or reassigning permissions to the session directories.

9. Copy the backed up certificates and grid-mapfiles.

10. Restart the services (in this order): *gridftpd*, *a-rex*, *grid-infosys*.

If the migration resulted in a change of hostname as seen by the outside world, users who wish to retrieve results from jobs that completed during the migration may have to use the *arcsync* command to synchronise their local job store.

5.7 Common tasks

In this section the sysadmin will find some acknowledged ways of performing common tasks on an ARC CE. The information gathered here has been collected over time by ARC experts to fulfill the needs of the communities using ARC.

Tags on each task will show what is the related area of expertise.

5.7.1 How to ban a single user based on his/her subject name

Tags: Security, Authorization

The first step would be to prevent user from accessing resource by disallowing any remote requests identified by given subject name. This task can be different ways described below.

Solution 1 Quick and simple

If You use grid-mapfile for authorization and [**vo**] section for generation of the grid-mapfile, then **filter=** command can be used to prevent some subject names from being accepted. After modifying **vo** section it is advisable to run **nordugridmap** utility to initiate immediate re-generation of the grid-mapfile. This solution to be used for quick result when there is no time for establishing more sophisticated authorization setup.

Solution 2 Local

1. Create a file containing the subject names of the users to ban, say, `/etc/grid-security/banned`, one per line. Use quotes to handle subject names with spaces in them.
2. In the [**group**] section used for authorization, add a line:

   ```
   -file=/etc/grid-security/banned
   ```

   Remember that the rules are processed in order of comparison, so this rule must appear *before* a rule that will allow users to access the cluster. See **6.1.3**, *Commands in the [group] section* for a detailed explanation of the rule parsing process. Make sure You have this line in all relevant [**group**] sections. Maybe consider using agregating [**group**] section which gathers results of processing of other groups by using **group=** keyword.
3. If You modified configuration file restart the A-REX service. The gridftpd service does not need to be restarted because it re-reads configuration on every connection made. If You only modified file containing subject names of banned users then You do not need to restart anything.

   This solution to be used when pure local authorization solution is needed. It allows every site to have own set of banned users.

Solution 3 Distributed

1. Setup Argus PEP or PDP service locally.
2. Adjust configuration to use your local Argus service (see section **4.5.6**, *Using Argus authorization service*).
3. Restart A-REX service after You changed configuration file.

This solution has an advantage in case You need to handle more than one service. You may also integrate your Argus service into hierarchy of other Argus services and use advantage of quick automatic propagation of information about banned users from participating authorities. For more information see Argus documentation at [**?**].

Solution 4 Relaying to LCAS

In a way similar to Argus one may use setup based on LCAS. If You are familiar with LCAS or need to integrate with gLite infrastructure this may be solution for You.

The next step is to identify and cancel all activity already initiated by banned user. For that *gm-jobs* utility may be used. See *gm-jobs -h* and *man gm-jobs* for the avaialble options. You may check all jobs belonging to user by calling *gm-jobs -f subject_name* and cancel active ones with *gm-jobs -K subject_name*. Cancel request is passed to A-REX. So it may take some time till jobs are canceled. If You want to *immediately* cancel all running jobs of the user *gm-jobs -l -f subject_name* can be used to obtain identifier of job in LRMS aka batch system (look for values labeled *LRMS id*). You may then use LRMS tools to cancel those jobs. It is still advisable to use *gm-jobs -K* first to avoid new jobs being started and canceled ones being re-started.

When You are done investigating harm caused by the banned user You may wipe his/her jobs with *gm-jobs -R subject_name*.

5.7.2 How to configure SELinux to use a port other than 2135 for the LDAP information system

The defined SELinux rules for the default port 2135 are as follows:

```
semanage port -a -t ldap_port_t -p tcp 2135 2>/dev/null || :
semanage fcontext -a -t slapd_db_t "/var/run/arc/bdii(/.*)?" 2>/dev/null || :
```

To use a port other than 2135, change the port number in the above in SELinux configuration.

NOTE: ARC packages postinstall scripts will always default to 2135, so make sure the specific SELinux configuration is loaded independently from ARC.

5.7.3 How to debug the ldap subsystem

In case there are problems with ldap publishing, it's strongly advised not to turn on slapd logs, as they will slow down performance. Most of the problems can arise in the process of updating the LDAP trees, for example due to odd values in some of the attributes. This process is performed by the BDII component, in particular by the bdii-update script.

To increase verbosity of such script, modify or add the value of the *bdii_debug_level* option in the **[infosys]** block.

1. Stop the ldap subsystem by running

   ```
   # service grid-infosys stop
   ```

2. edit *arc.conf* so to have:

   ```
   [infosys]
   ...
   bdii_debug_level="ERROR"
   ...
   ```

3. Restart the ldap subsystem by running

   ```
   # service grid-infosys start
   ```

`/var/log/arc/bdii/bdii-update.log` will contain relevant LDAP errors.

5.7.4 Missing information in LDAP or WSRF

A known issue in ARC new infoproviders is some slowdown in the information system when a huge amount of jobs are sitting in the control directory. Symptoms of this issue are unresponsive ldap server and job status not retrieved by arc tools.

Also by looking at A-REX logs, it's possible to see the error message:

```
... Resource information provider timeout: ...
```

To overcome this limitation there is a current workaround, which allows a system administrator to let infoproviders run for more time. This is done by increasing the timeout.

The default 600 seconds should be suitable up to 5000 jobs. The rule of thumb is to increase this value of 600 seconds each 5000 jobs.

1. Stop a-rex and the ldap subsystem by running

```
# service a-rex stop; service grid-infosys stop
```

2. edit *arc.conf* so to have:

```
[infosys]
...
infoproviders_timeout="1200"
...
```

   the value is in seconds. One may need to fine tune it.

3. Restart A-REX and the ldap subsystem by running

```
# service a-rex start; service grid-infosys start
```

`/var/log/arc/grid-manager.log` should not show the above error anymore. If this happens, increase the timeout.

The ARC team is working on a smarter solution that will let A-REX infoproviders process this information faster and eventually change the timeout automatically.

Chapter 6

Technical Reference

6.1 Reference of the `arc.conf` configuration commands

6.1.1 Generic commands in the [common] section

x509_user_key=*path* – sets the path to the host private key, usually `/etc/grid-security/hostkey.pem`

x509_user_cert=*path* – sets the path to the host public certificate, usually `/etc/grid-security/hostcert.pem`

x509_cert_dir=*path* – sets the path to the CA certificates, usually `/etc/grid-security/certificates`

gridmap=*path* – the path of the "grid map file", which maps Grid users to local unix accounts. This has to be set even if the mappings are dynamically created by the `nordugrid-arc-gridmap-utils` package is installed (see sections Section **4.4.1**, *Access control: users, groups, VOs* and Section **6.10**, *Structure of the grid-mapfile* for a brief explanation).

hostname=*hostname* – sets the hostname of the front-end. *hostname* is just a FQDN string. If not specified, *hostname* will be the one returned by the shell command `hostname -f`. Make sure this hostname is the same listed in `/etc/hosts` (see also Section **4.1.4**, *Networking*). This hostname has to be the same FQDN in the host certificates (see also Section **3.3**, *Installation of certificates*).

lrms=*lrms_name [default_queue_name]* – specifies names for the LRMS and queue. Queue name can also be specified in the JD.

x509_voms_dir=*path* – sets the path to the directory containing *.lsc files needed for checking validity of VOMS extensions. If not specified default value `/etc/grid-security/vomsdir` is used.

6.1.2 Commands in the [vo] section

These sections are also used by the `nordugridmap` utility which reads sources and generates list of Grid users belonging to particular VO or some other group.

vo=*vo_name* – specifies name of VO. It is required.

id=*unique_id* – defines an unique id for the VO.

file=*path* – path to file which contains list of users' DNs belonging to VO and their mappings. This file follows the format stated in Section **6.10**, *Structure of the grid-mapfile*. If `nordugridmap` is used it fills that file. If this VO is used as a `vo` option in a `[group]` block then `file` must be specified.

source=*URL* – specifies the URL from which a list of users may be obtained. There can be more than one `source` entries in the same `[vo]` section. URL is in the form $<$ *protocolname* : // $<$ *path* $>$ where $<$ *protocolname* $>$ is one of: *vomss, http, https, ldap, file* and $<$ *path* $>$ is a path in the form accepted by the protocol standard. In production environments, this URL to source files can

be requested to the Grid organization who hosts the CA or the Grid computing organizations the CE is meant to be part of.

Some examples:

```
source="http://www.nordugrid.org/community.dn"
source="vomss://sample.hep.lu.se:8443/voms/knowarc.eu?/knowarc.eu"
source="file:///etc/grid-security/local-grid-mapfile"
```

mapped_unixid=*uid* – This is the local UNIX user account to which the DNs contained in the `source` command will be mapped. Only one `mapped_unixid` can be defined per `[vo]` section!

require_issuerdn=*[yes|no]* – *yes* would map only those DNs obtained from the urls which have the corresponding public CA packages installed. Default is *no*.

6.1.3 Commands in the [group] section

The [group] sections and subsections define authorization unities called groups.

name=*group_name* – specifies the name of an authorization group. If used within a `[group/subsection]` it has to be the same as the subsection name. If this command is omitted, `name` will implicitly taken from the subsection name.

Authorization is performed by applying a set of rules to users credentials. Credentials are certificates or certificates content (DN subject name, VO the user belongs to, CA that released the certificate...). Rules have the same $<$ *command* $>=<$ *value* $>$ format as rest of configuration file, with the difference that each rule command is prepended with optional modifiers: [+|-][!].

The rules are process sequentially in same order as presented in configuration. Processing stops at first matched rule.

A rule is said to *match* a credential if the credential **"satisfies"** the value specified by the command. By prepending rule with **!** matching is reversed. Matching rules turns into non-matching and non-matching into matching.

There are two kinds of matching. Rule prepended by + sign is called to produce *positive match* and matched credentials are considered to be belonging to this group. If rule is prepended with **-** sign it produces *negative match* and credentials are considered *not* belonging to this group. In both cases processing of rules for this groups is stopped. By default rule produces positive match - so + is optional.

Examples:

vo=*TESTVO* – This rule matches all the users belonging to the *TESTVO* Virtual Organization.

!vo=*TESTVO* – This rule matches all the users NOT belonging to the *TESTVO* Virtual Organization.

A credential (and therefore the user presenting it) can be *accepted* or *rejected*.

Accepted means that the credential becomes member of the group being processed - *positive match*.

Rejected means that the credential does **not** become member of group being processed - *negative match*.

Examples:

+vo=*TESTVO* – all the users belonging to the TESTVO Virtual Organization are Accepted into group. It can also be written as `vo = TESTVO`

-vo=*TESTVO* – all the users belonging to the TESTVO Virtual Organization are Rejected from this group.

+!vo=*TESTVO* – all the users NOT belonging to the TESTVO Virtual Organization are Accepted into group. It can also be written as `!vo = TESTVO`

-!vo=*TESTVO* – all the users NOT belonging to the TESTVO Virtual Organization are Rejected from group.

Note that `-vo = TESTVO` and `+!vo = TESTVO` do *not* do same thing. In first case user is rejected from group immediately. In second case following rules will be processed - if any - and user may finally be accepted.

A summary of the modifiers is on **6.1**, *Basic Access Control modifiers and their meaning.*

Group membership does not automatically mean user is allowed to access resources served by A-REX. Whenever a GRID user submits a job to or requests information from the CE, A-REX will try to find a rule that matches that credential, for every [group...] section. Groups and rules will be processed in the order they appear in the `arc.conf` file.

Processing of rules in every group stops after the first positive or negative match, or when failure is reached. All groups are always processed. Failures are rule-dependent and may be caused by conditions like missing files, unsupported or mistyped rule, etc.

The following *rule words* and arguments are supported:

subject=*subject [subject [...]]* – match user with one of specified subjects

file=*[filename [...]]* – read rules from specified files. Format of file similar to format of commands in *group* section with = replaces with space. Also in this file subject becomes default command and can be omitted. So it becomes possible to use files consisting of only subject names of user credentials and Globus grid-mapfiles can be used directly.

remote=*[ldap://host:port/dn [...]]* – match user listed in one of specified LDAP directories (uses network connection hence can take time to process)

voms=*vo group role capabilities* – accept user with VOMS proxy with specified vo, group, role and capabilities. * can be used to accept any value.

vo=*[vo [...]]* – match user belonging to one of specified Virtual Organizations as defined in *vo* section configuration section (see `[vo]` above). Here VO membership is determined from corresponding *vo* section by comparing subject name of credentials to one stored in VO list file.

group=*[groupname [groupname [...]]]* – match user already belonging to one of specified groups.

plugin=*timeout plugin [arg1 [arg2 [...]]]* – run external plugin (executable or function in shared library) with specified arguments. Execution of plugin may **not** last longer than *timeout* seconds.
If plugin looks like `function@path` then function `int function(char*,char*,char*,...)` from shared library `path` is called (timeout has no effect in that case). Rule matches if plugin or executable exit code is 0. Following substitions are applied to arguments before plugin is started:

- %D - subject of userss certificate,
- %P - name of credentials proxy file.

lcas=*library directory database* – - call LCAS functions to check rule. Here *library* is path to shared library of LCAS, either absolute or relative to *directory*; *directory* is path to LCAS installation directory, equivalent of LCAS_DIR variable; *database* is path to LCAS database, equivalent to LCAS_DB_FILE variable. Each arguments except library is optional and may be either skipped or replaced with *.

all accept any user

Here is an example of authorization group:

! invert matching. Match is treated as non-match. Non-match is treated as match, either positive (+ or nothing) or negative (-).

+ accept credential if matches following rule (positive match, default action);

- reject credential if matches following rule (negative match);

Figure 6.1: Basic Access Control modifiers and their meaning

```
(1)    [group/admins]
(2)    -subject=/O=Grid/OU=Wrong Place/CN=Bad Person
(3)    file=/etc/grid-security/internal-staff
(4)    voms=nordugrid admin * *
```

The processing will work in the following way:

Let credential has subject `/O=Grid/OU=Wrong Place/CN=Bad Person`. Then, subject matches (1) and the credential is Rejected from this group, processing of this group will stop.

Let credential has subject `/O=Grid/OU=Internal-Staff/CN=Good Person`, and let this subject be inside the file `/etc/grid-security/internal-staff`. Then, (1) doesn't match, processing continues to (2). Since subject is present inside the file specified by the `file` command, then the credential is Accepted in this group and the processing of this group stops.

Let credential has subject `/O=Grid/OU=SomeoneNotStaffButInnordugridVO/CN=Loyal Person`, and supplied credentials contain VOMS extension issued by `nordugrid` VO with group `admin` assigned. Let this credential be NOT present in the `internal-staff` file. Then, neither (1) nor (2) match and processing passes to (3). Since the credential belongs to that VO and group matches, the credential is Accepted in this group and processing of this group stops.

Let credential be `/O=Grid/OU=SomeOrg/CN=UN Known`, not present in the file neither belonging to VO. Processing passes through (1), (2), (3) without matching. Credential is Rejected from this group.

6.1.4 Commands in the [gridftpd] section

This section describes configuration options used in the [gridftpd] section, including how to set up fine-grained authorisation based on groups and VOs explained in the previous sections.

6.1.4.1 General commands

daemon=*yes|no* – defines if GFS must run in daemon mode. Default is `yes`.

logfile=*path* – specifies log file for GFS. Default is `/var/log/arc/gridftpd.log`.

logsize=*size [number]* – restricts log file size to *size* and keeps *number* archived log files. If installed from packages, the log properties are managed by logrotate. If logrotate or another external log management tool is used then *logsize* should not be used.

logreopen=*yes|no* – defines if GFS closes log file after every write into log file and reopens it for every write. Default is `no`.

user=*username[:groupname]* – tell GFS to switch to specified user `username` and optionally to group `groupname` after start. Default is not to change user account.

pidfile=*path* – file containing the PID of the gridftpd process. Default is `/var/run/gridftpd.pid`.

debug=*number* – defines numerical - from 0 to 5 - verbosity of messages written into log file. Default is 3.

pluginpath=*path* – non-standard location of plugins location directory. Default is ARC_LOCATION/lib/arc.

port=*number* – specifies TCP/IP port number. Default is 2811.

maxconnections=*number* – limits number of simultaneously served clients. Default is 100. Specifying 0 removes this limit. Connections over limit are rejected.

defaultbuffer=*number* – defines size of every buffer for data reading/writing. Default is 64kB. Actual value may decrease if cumulative size of all buffers exceeds value specified by `maxbuffer`.

maxbuffer=*number* – defines maximal amount of memory in bytes to be allocated for all data reading/writing buffers. Default is 640kB. Number of buffers is parallelism level requested by connecting client multiplied by 2 and increased by 1. Final number is limited by 41 from top and 3 from bottom. Hence even without parallel streams enabled number of buffers will be 3.

firewall=*hostname* – defines hostname or IP address of firewall interface in case of GFS situated behind firewall with Network Address Translation functionality. If this command is specified GFS

will use corresponding IP address instead of IP address of interface used for accepting client request in response to PASV and similar commands.

encryption=*yes|no* – specifies if encryption of data channel is allowed. Default is `yes`. Data encryption is a heavy operation which can create significant load on the GFS host and increase data transfer time.

include=*path* – include contents of another config file.

allowunknown=*yes|no* – if set to `no` all clients with subject not in grid-mapfile are immediately rejected. If set to `yes`, this check is not performed. Default is no.

voms_processing=*relaxed|standard|strict|noerrors* – specifies how to behave if failure happens during VOMS processing. See description of this option in Section **6.1.12.2**, *Commands affecting the A-REX Web Service communication interface.*

voms_trust_chain=*subject [subject [...]]* – specifies chain of VOMS credentials subject names tobe trusted during VOMS processing. See description of this option in Section **6.1.12.2**, *Commands affecting the A-REX Web Service communication interface.*

globus_tcp_port_range=*min, max* – Globus TCP port range. Equivalent to the *$GLOBUS_TCP_PORT_RANGE* environment variable. By default, leave it up to Globus.

globus_udp_port_range=*min, max* – Globus UDP port range. Equivalent to the *$GLOBUS_UDP_PORT_RANGE* environment variable. By default, leave it up to Globus.

x509_user_key=*path* – path to the X509 certificate key file. Equivalent to the *$X509_USER_KEY* environment variable. Default is `/etc/grid-security/hostkey.pem`.

x509_user_cert=*path* – path to the X509 certificate file. Equivalent to the *$X509_USER_CERT* environment variable. Default is `/etc/grid-security/hostcert.pem`.

x509_cert_dir=*path* – path to a directory where to search for X509 CA certificates. Equivalent to the *$X509_CERT_DIR* environment variable. Default is `/etc/grid-security/certificates`.

http_proxy=*url* – HTTP proxy setting. Equivalent to the *$ARC_HTTP_PROXY* environment variable. Default is not to use HTTP proxy.

maxconnections=*number* – defines the maximum number of simultaneous connections. Default is 100.

defaultbuffer=*number* – default buffer size. Default is 65536.

maxbuffer=*number* – maximum buffer size. Default is 655360.

6.1.4.2 Commands for fine-grained authorisation

These commands use the VOs and groups set up in Sections **6.1.2**, *Commands in the [vo] section* and **6.1.3**, *Commands in the [group] section* to perform user mapping.

unixgroup=*group rule* – define local UNIX user and optionally UNIX group to which user belonging to specified authorization *group* is mapped. Local names are obtained from the specified *rule*. If the specified rule could not produce any mapping, the next command is used. Mapping stops at first matched rule. The following rules are supported:

mapfile=*file* – the user's subject is matched against a list of subjects stored in the specified file, one per line followed by a local UNIX name.

simplepool=*directory* – the user is assigned one of the local UNIX names stored in a file *directory/pool*, one per line. Used names are stored in other files placed in the same *directory*. If a UNIX name was not used for 10 days, it may be reassigned to another user.

lcmaps=*library directory database policy_name [policy_name [...]]* – call LCMAPS functions to do mapping. Here *library* is the path to the shared library of LCMAPS, either absolute or relative to *directory*; *directory* is the path to the LCMAPS installation directory, equivalent to the LCMAPS_DIR variable; *database* is the path to the LCMAPS database, equivalent to the LCMAPS_DB_FILE variable. The *policy_name* refers to name of policy as defined in *database*. The must be at least one *policy_name* specified. Each argument except

library is optional and may be either skipped or replaced with '*'. See Section **4.5.7**, *Using LCAS/LCMAPS* for more information on LCMAPS.

mapplugin=*timeout plugin [arg1 [arg2 [...]]]* – run external *plugin* executable with specified arguments. Execution of *plugin* may not last longer than *timeout* seconds. A rule matches if the exit code is 0 and there is a UNIX name printed on *stdout*. A name may be optionally followed by a UNIX group separated by ':'. In arguments the following substitutions are applied before the plugin is started:

- %D – subject of user's certificate,
- %P – name of credentials' proxy file.

unixvo=*vo rule* – same as ***unixgroup*** for users belonging to Virtual Organization (VO) *vo*.

unixmap=*[unixname][:unixgroup] rule* – define a local UNIX user and optionally group used to represent connected client. *rule* is one of those allowed for authorization groups and for ***unixgroup*/*unixvo***. In case of a mapping rule, username is the one provided by the rule. Otherwise the specified *unixname:unixgroup* is taken. Both *unixname* and *unixgroup* may be either omitted or set to '*' to specify missing value.

6.1.4.3 Commands to configure the jobplugin

The GFS comes with 3 plugins: *fileplugin.so*, *gaclplugin.so* and *jobplugin.so*. For the computing element, only the jobplugin (*jobplugin.so*) is needed - the others are described in [**?**]. The jobplugin and it is configured in a subsection **[gridftpd/jobs]** (the name `jobs` is used by convention but is not enforced).

Configuration options for all plugins:

path=*path* – virtual path to which the service will be associated (by convention "/jobs" for jobplugin)

plugin=*library_name* – use plugin library_name to serve virtual path ("jobplugin.so" for jobplugin)

groupcfg=*[group [group [...]]]* – defines authorization groups which are allowed to use functionality of this plugin. Default is to allow any client.

Configuration options for jobplugin:

configfile=*path* – defines non-standard location of the arc.conf file. You only need this option if GFS and A-REX are configured in separate configuration files.

allownew=*yes|no* – specifies if new jobs can be submitted. Default is yes

unixgroup/unixvo/unixmap=*rule* – same options as in the top-level GFS configuration, but can be specified here to allow different rules per GFS plugin. If the mapping succeeds, the obtained local user will be used to run the submitted job.

remotegmdirs=*control_dir session_dir [drain]* – specifies control and session directories under the control of another A-REX to which jobs can be assigned. This option is useful if several A-REX services should be accessed via one GFS interface. Remote directories can be added and removed without restarting the GFS. However, it may be desirable to drain them prior to removal by adding the drain option. In this case no new jobs will be assigned to these directories but their contents will still be accessible.

maxjobdesc=*size* – specifies maximal allowed size of job description in bytes. Default value is 5MB. If value is missing or set to 0 no limit is applied.

For information about how to communicate with the **jobplugin.so** see section **6.15**, *GridFTP Interface (jobplugin)*.

Simple and detailed configuration examples for the jobplugin are given in Section **6.15.4**, *Configuration Examples*.

6.1.5 Commands in the [infosys] section

The `user` command here defines the UNIX user ID with which the slapd server, the infoproviders, BDII and registration scripts will run.

oldconfsuffix=*.suffix* – sets the suffix of the backup files of the low-level slapd config files in case they are regenerated. Default is ".oldconfig".

overwrite_config=*yes|no* – determines if the grid-infosys startup script should generate new low-level slapd configuration files. By default the low-level configuration files are regenerated with every server startup making use of the values specified in the arc.conf.

hostname=*FQDN* – the hostname of the machine running the slapd service.

port=*port_number* – the port number where the slapd service runs. Default infosys port is 2135.

debug=*0|1* – sets the debug level/verbosity of the startup script. Default is 0.

slapd_loglevel=*verbosity_level* – sets the native slapd syslog loglevel (see man slapd for verbosity_level values). The default is set to no-logging (0) and it is RECOMMENDED not to be changed in a production environment. Non-zero slap_loglevel value causes serious performance decrease.

slapd_hostnamebind=**|* – may be used to set the hostname part of the network interface to which the slapd process will bind. Most of the cases no need to set since the hostname config parameter is already sufficient. The default is empty, but this can lead to problems in systems where slapd is set by security policies to be run only on the localhost interface. The wildcard will bind the slapd process to all the network interfaces available on the server. However, this is not recommended in clusters with many network interfaces. The recommended setting is the hostname assigned to the interface that will be publicly accessible.

threads=*num_threads* – the native slapd threads parameter, default is 32. If you run an Index Service too (see [**?**]) you should modify this value.

timelimit=*seconds* – the native slapd timelimit parameter. Maximum number of seconds the slapd server will spend answering a search request. Default is 3600.

slapd_cron_checkpoint=*enabled|disabled* – introduced to solve bug 2032 to prevent BDII to create huge logs with lots of files. Enable this if you're experiencing this problem. Default is disabled as latest versions of BDII doesn't seem to be affected.

providerlog=*path* – Specifies log file location for the information provider scripts. Default is `/var/log/arc/infoprovider.log`.

provider_loglevel=*[0-5]* – loglevel for the infoprovider scripts (0, 1, 2, 3, 4, 5). The default is 1 (critical errors are logged). This corresponds to different verbosity levels, from less to maximum, namely: FATAL, ERROR, WARNING, INFO, VERBOSE, DEBUG

infoproviders_timeout=*seconds* – this only applies to new infoproviders. It changes A-REX behaviour with respect to a single infoprovider run. Increase this value if you have many jobs in the controldir and infoproviders need more time to process. The value is in seconds. Default is 600 seconds. See also **5.7.4**, *Missing information in LDAP or WSRF*.

registrationlog=*path* – specifies the logfile for the registration processes initiated by your machine. Default is `/var/log/arc/inforegistration.log`. For registration configuration, see Section **4.4.5**, *How to join the grid: registering to an index service*.

infosys_nordugrid=*enable|disable* – Activates or deactivates NorduGrid infosys schema [**?**] data generation and publishing. Default is *enabled*. This schema doesn't need further configuration.

infosys_glue12=*enable|disable* – Activates or deactivates Glue 1.x [**?**] infosys schema data generation and publishing. Default is *disabled*. For configuration of this schema, see Section **4.5.1**, *Enabling or disabling LDAP schemas*.

infosys_glue2_ldap=*enable|disable* – Activates or deactivates Glue 2 [**?**] infosys schema data generation and publishing. Default is *disabled*. For configuration of this schema, see Section **4.5.1**, *Enabling or disabling LDAP schemas*.

infosys_compat=*enable|disable* – If *enabled*, A-REX will use old infoproviders (ARC $\leq$ 0.8 infoproviders). These infoproviders cannot produce GLUE2 data, therefore default is *disabled*. If there is a need for features belonging to this set of infoproviders, or if there is some performance issue with the new infoproviders, it is possible to enable this compatibility mode, but it will only allow NorduGrid and Glue1 schema.

BDII related commands

bdii_debug_level=*CRITICAL | ERROR | WARNING | INFO | DEBUG* – Defines verbosity of BDII debug messages, from less verbose (*CRITICAL*) to maximum verbosity (*CRITICAL*). Default is (*ERROR*). Set this to (*DEBUG*) if experiencing LDAP publication problems. See also **5.7.3**, *How to debug the ldap subsystem*.

provider_timeout=*seconds* – This variable allows a system administrator to modify the behaviour of the `bdii-update` script. This is the time BDII waits for the scripts generated by A-REX inforproviders to produce their output. Default is 300 seconds. The total waiting time is given by this value plus A-REX wakeupperiod, usually 450 seconds in total.
Use this command to tweak bdii in case inforproviders are producing a lot of data. This has reported to happen when `infosys_glue2_ldap_showactivities` is enabled.

Legacy commands

The following command only affect old infoproviders, that is, when `infosys_compat=enable`

cachetime=*seconds* – The validity time in seconds that will be used to fill information system records about the cluster.

6.1.6 Commands in the [infosys/admindomain] section

Commands in this subsection are used to fill the AdminDomain GLUE2 data. This extends and generalizes the glue12 "site" concept.

name=*domain_name* – The string that identifies uniquely a domain name. Case is considered. This is mandatory if the [infosys/admindomain] block is enabled.

description=*text* – A human-readable description of the domain. Optional.

www=*domain_url* – A url pointing to relevant Domain information. Optional.

distributed=*yes|no* – A flag indicating the nature of the domain. Yes if services managed by the AdminDomain are considered geographically distributed by the administrator themselves. Most likely this has to be set to no only if the CE is standalone and not part of other organizations. Optional

owner=*string* – A string representing some person or legal entity which pays for the services or resources. Optional.

otherinfo=*text* – This field is only for further development purposes. It can fit all the information that doesn't fit above.

6.1.7 Commands in the [infosys/glue12] section

All the commands are mandatory if infosys_glue12 is enabled in the `[infosys]` section.

resource_location=*City, Country* – The field is free text but is a common agreement to have the City and the Country where the CE is located, separated by comma.

resource_latitude=*latitudevalue* – latitude of the geolocation where the CE is, expressed as degrees, e.g. 55.34380

resource_longitude=*longitudevalue* – latitude of the geolocation where the CE is, expressed as degrees, e.g. 12.41670

cpu_scaling_reference_si00=*number* – number represent the scaling reference number wrt si00. Please refer to the GLUE schema specification [] to know which value to put.

processor_other_description=*string* – String representing information on the processor, i.e. number of cores, benchmarks.... Please refer to the GLUE schema specification [] to know which value to put. Example: `Cores=3,Benchmark=9.8-HEP-SPEC06`

glue_site_web=*url* – full url of the website of the site running the CE. Example: `http://www.ndgf.org`

glue_site_unique_id =*siteID* – Unique ID of the site where the CE runs. Example: `NDGF-T1`

provide_glue_site_info=*true|false* – This variable decides if the GlueSite should be published, in case a more complicated setup with several publishers of data to a GlueSite is needed.

6.1.8 Commands in the [infosys/site/*sitename*] section

This command is used for an ARC CE to mimic the Glue1.2/1.3 Site BDII behavior. This option was an early attempt to overcome the needs to install a Site BDII, so that ARC CE information could directly be fetched by a Top BDII. This worlaround has performance issues on the CE, as it increases the amount of information contained in the local LDAP server. Thus is strongly discouraged. If one wants a similar functionality, should enable the GLUE2 rendering instead.

Site information will be rendered according to the specified *sitename* and the previously described [infosys/glue12] section. For a correct behavior, Glue1.2/1.3 rendering MUST be enabled (see above blocks)

unique_id=*sitename* – a string, the unique id of the Site, MUST be the same as *sitename* in the block identifier. Example: "LundTestSite"

Administrator should take care that this value is the same as with glue_site_unique_id in the [infosys/glue12] block.

url=*ldap url* – a url to an existing Glue1.2/1.3 resource bdii. In the case of ARC, the Glue1.2/1.3 renderings presents information as a resource bdii. Therefore the url MUST be:

"ldap://localhost:2135/mds-vo-name=resource,o=grid".

The URL value can be changed to any valid LDAP URL that returns a Glue1.2/1.3 compliant rendering. This is, however, strongly discouraged. ARC CE is not a Site BDII.

6.1.9 Commands in the [cluster] section

These commands will affect the GLUE2 ComputingService, ComputingManager, ExecutionEnvironment, Policies objects, in how the information providers will fetch information about the frontend and computing nodes managed by the LRMS.

For a decent brokering, at least *architecture*, *nodecpu*, *nodememory* and *opsys* should be published.

cluster_alias=*name* – an arbitrary alias name of the cluster, optional

comment=*text* – a free text field for additional comments on the cluster in a single line, no newline character is allowed!

lrmsconfig=*description* – an optional free text field to describe the configuration of your Local Resource Management System (batch system).

homogeneity=*True|False* – determines whether the cluster consists of identical NODES with respect to cputype, memory, installed software (opsys). The frontend is NOT needed to be homogeneous with the nodes. In case of inhomogeneous nodes, try to arrange the nodes into homogeneous groups assigned to a queue and use queue-level attributes. Default is True. If set to False, the infosystem will try to fill GLUE2 ExecutionEnvironment information for each inhomogeneous node.

architecture=*string|adotf* – sets the hardware architecture of the NODES. The "architecture" is defined as the output of the "uname -m" (e.g. i686). Use this cluster attribute if only the NODES are homogeneous with respect to the architecture. Otherwise the queue-level attribute may be used for inhomogeneous nodes. If the frontend's architecture agrees to the nodes, the "adotf" (Automatically Determine On The Frontend) can be used to request automatic determination.

opsys=*string|adotf* – this multivalued attribute is meant to describe the operating system of the computing NODES. Set it to the opsys distribution of the NODES and not the frontend! opsys can also be used to describe the kernel or libc version in case those differ from the originally shipped ones. The distribution name should be given as distroname-version.number, where spaces are not allowed. Kernel version should come in the form kernelname-version.number. If the NODES are inhomogeneous with respect to this attribute do NOT set it on cluster level, arrange your nodes into homogeneous groups assigned to a queue and use queue-level attributes.

nodecpu=*string|adotf* – this is the cputype of the homogeneous nodes. The string is constructed from the /proc/cpuinfo as the value of "model name" and "@" and value of "cpu MHz". Do NOT set this attribute on cluster level if the NODES are inhomogeneous with respect to cputype, instead arrange the nodes into homogeneous groups assigned to a queue and use queue-level attributes. Setting the nodecpu="adotf" will result in Automatic Determination On The Frontend, which should only be used if the frontend has the same cputype as the homogeneous nodes. String can be like: "AMD Duron(tm) Processor @ 700 MHz"

nodememory=*number_MB* – this is the amount of memory (specified in MB) on the node which can be guaranteed to be available for the application. Please note in most cases it is less than the physical memory installed in the nodes. Do NOT set this attribute on cluster level if the NODES are inhomogeneous with respect to their memories, instead arrange the nodes into homogeneous groups assigned to a queue and use queue-level attributes.

defaultmemory=*number_MB* – If a user submits a job without specifying how much memory should be used, this value will be taken first. The order is: job specification → defaultmemory → nodememory → 1GB. This is the amount of memory (specified in MB) that a job will request(per count for parallel jobs).

nodeaccess= *|inbound|outbound* – determines how the nodes can connect to the internet. Not setting anything means the nodes are sitting on a private isolated network. "outbound" access means the nodes can connect to the outside world while "inbound" access means the nodes can be connected from outside. inbound & outbound access together means the nodes are sitting on a fully open network.

cluster_location=*XX-postalcode* – The geographycal location of the cluster, preferably specified as a postal code with a two letter country prefix, like "DK-2100"

cluster_owner=*name* – it can be used to indicate the owner of a resource, multiple entries can be used

clustersupport=*email* – this is the support email address of the resource, multiple entries can be used

authorizedvo=*string* – a free-form string used to advertise which VOs are authorized on all the services on the CE. Multiple entries are allowed, each authorizedvo= entry will add a VO name to the infosystem. On the GLUE2 schema, this information will be published in the AccessPolicies and MappingPolicies. Example:

```
authorizedvo=VO:ATLAS
authorizedvo=support.nordugrid.org
```

cpudistribution=*[ncpu:m, ...]ncpu:m* – This is the CPU distribution over nodes given in the form: `ncpu:m` where

- n is the number of CPUs per machines
- m is the number of such computers

Example: `1cpu:3,2cpu:4,4cpu:1`
represents a cluster with 3 single CPU machines, 4 dual CPU machines, one machine with 4 CPUs.
This command is needed to tweak and overwrite the values returned by the underlying LRMS. In general there is no need to configure it.

6.1.10 Commands in the [queue] subsections

These commands will affect the ComputingShare GLUE2 object. Special GLUE2 MappingPolicies publishing configuration per queue is not yet supported.

fork_job_limit=*number*|cpunumber – sets the allowed number of concurrent jobs in a fork system, default is 1. The special value cpunumber can be used which will set the limit of running jobs to the number of cpus available in the machine. This parameter is used in the calculation of freecpus in a fork system.

name=*queuename* – The name of the grid-enabled queue, it must also be in the queue section name[queue/queuename]. Use "fork" for the fork LRMS.

homogeneity=*True*|*False* – - determines whether the queue consists of identical NODES with respect to cputype, memory, installed software (opsys). In case of inhomogeneous nodes, try to arrange the nodes into homogeneous groups and assigned them to a queue. Default is True.

scheduling_policy=*FIFO*|*MAUI* – this optional parameter tells the scheduling policy of the queue, PBS by default offers the FIFO scheduler, many sites run the MAUI. At the moment FIFO & MAUI is supported. If you have a MAUI scheduller you should specify the "MAUI" value since it modifies the way the queue resources are calculated. BY default the "FIFO" scheduler is assumed. More about this in chapter Section **4.4.2**, *Connecting to the LRMS*.

comment=*text* – a free text field for additional comments on the queue in a single line, no newline character is allowed!

The following commands only apply to old infoproviders, that is, when infosys_compat=enable:

cachetime=*seconds* – The validity time in seconds that will be used to fill information system records about the queue.

6.1.11 Commands in the [infosys/cluster/registration/registrationname] subsections

Computing resource (cluster) registration block, configures and enables the registration process of a Computing Element to an Index Service. The string infosys/cluster/registration/ identifies the block, while registrationname is a free form string used to identify a registration to a specific index. A cluster can register to several Index Services. In this case, each registration process should have its own block, each with its own registrationname.

Registration commands explained:

targethostname=*FQDN* – The FQDN of the host running the target index service.

targetport=*portnumber* – Port where the target Index Service is listening. Defaults to 2135.

targetsuffix=*ldapsuffix* – ldap suffix of the target index service. This has to be provided by a manager of the index service, as it is a custom configuration value of the Index Service. Usually is a string of the form "mds-vo-name=<custom value>,o=grid"

regperiod=*seconds* – the registration script will be run each number of *seconds*. Defaults to 120.

registranthostname=*FQDN* – the registrant FQDN. This is optional as ARC will try to guess it from the system of from then [common] block. Example: registranthostname="myhost.org"

registrantport=*port* – the port where the local infosystem of the registrant is running. Optional, as this port is already specified in the [infosys] block. Example: registrantport="2135"

registrantsuffix=*ldap_base_string* – the LDAP suffix of the registrant cluster resource. Optional, as it is automatically determined from the [infosys] block and the registration blockname. In this case the default registrantsuffix will be: nordugrid-cluster-name=FQDN,Mds-Vo-name=local,o=Grid. Please mind uppercase/lowercase characters above if defining allowreg in an index! Don't set it unless you want to overwrite the default. Example: registrantsuffix="nordugrid-cluster-name=myhost.org,Mds-Vo-name=local,o=grid"

6.1.12 Commands in the [grid-manager] section

6.1.12.1 Commands affecting the A-REX process and logging

pidfile=*path* – specifies file where process id of A-REX process will be stored. Defaults to `/var/run/arched-arex.pid` if running as root and `$HOME/arched.pid` otherwise.

logfile=*path* – specifies name of file for logging debug/informational output. Defaults to `/var/log/arc/grid-manager.log`. Note: if installed from binary packages, ARC comes with configuration for *logrotate* log management utility and A-REX log is managed by *logrotate* by default.

logsize=*size number* – restricts log file size to *size* and keeps *number* archived log files. This command enables log rotation by ARC and should only be used if *logrotate* or other external log rotation utility is not used. Using ARC log rotation and external log management simultaneously may result in strange behaviour.

logreopen=*yes|no* – specifies if log file must be opened before writing each record and closed after that. By default log file is kept open all the time (default is no).

debug=*number* – specifies level of debug information. More information is printed for higher levels. Currently the highest effective number is 5 (DEBUG) and lowest 0 (FATAL). Defaults to 2 (WARNING).

user=*username[:groupname]* – specifies username and optionally groupname to which the A-REX must switch after reading configuration. Defaults to *not switch*.

6.1.12.2 Commands affecting the A-REX Web Service communication interface

voms_processing=*relaxed|standard|strict|noerrors* – specifies how to behave if failure happens during VOMS processing.

- relaxed – use everything that passed validation.
- standard – same as relaxed but fail if parsing errors took place and VOMS extension is marked as critical. This is a default.
- strict – fail if any parsing error was discovered.
- noerrors – fail if any parsing or validation error happened.

Default is *standard*. This option is effective only if A-REX is started using startup script.

voms_trust_chain=*subject [subject [...]]* – specifies chain of VOMS credentials to trust during VOMS processing. There can be multiple `voms_trust_chain` commands one per trusted chain/VOMS server. Content of this command is similar to the information *.lsc file, but with two differences with one `voms_trust_chain` corresponding to one *.lsc file. Differently from *.lsc this command also accepts regular expressions - one per command. If this command is specified information in *.lsc files is not used even if *.lsc exist. This option is effective only if A-REX is started using startup script.

fixdirectories=*yes|missing|no* – specifies during startup A-REX should create all directories needed for it operation and set suitable default permissions. If *no* is specified then A-REX does nothing to prepare its operational environement. In case of *missing* A-REX only creates and sets permissions for directories which are not present yet. For *yes* all directories are created and permisisons for all used directories are set to default safe values. Default behavior is as if *yes* is specified.

arex_mount_point=*URL* – specifies URL for accessing A-REX through WS interface. This option is effective only if A-REX is started using startup script. The presence of this option enables WS interface. Default is to not provide WS interface for communication with A-REX.

enable_arc_interface=*yes|no* – turns on or off the ARC own WS interface based on OGSA BES and WSRF. If enabled the interface can be accessed at the URL specified by *arex_mount_point* (this option must also be specified). Default is *yes*.

enable_emies_interface=*yes|no* – turns on or off the EMI Execution Service interface. If enabled the interface can be accessed at the URL specified by *arex_mount_point* (this option must also be specified). Default is *no*.

max_job_control_requests=*number* – specifies maximal number of simultaneously processed job control requests. Requests above that threshold are put on hold and client is made to wait for response. Default value is 100. Setting value to -1 turns this limit off. This option is effective only if A-REX is started using startup script.

max_infosys_requests=*number* – specifies maximal number of simultaneously processed information requests. Requests above that threshold are put on hold and client is made to wait for response. Default value is 1. Setting value to -1 turns this limit off. This option is effective only if A-REX is started using startup script.

max_data_transfer_requests=*number* – specifies maximal number of simultaneously processed data read/write requests. Requests above that threshold are put on hold and client is made to wait for response. Default value is 100. Setting value to -1 turns this limit off. This option is effective only if A-REX is started using startup script.

arguspep_endpoint=*URL* – specifies URL of Argus PEP service to use for authorization and user mapping. If this option is set Argus is contacted for every operation requested through WS interface.

arguspep_profile=*profile name* – - defines which communication profile to use for passing information to Argus PEP service. Possible values are:
direct – pass all authorization attributes (only for debugging),
subject – pass only subject name of client,
cream – makes A-REX pretend it is gLite CREAM service,
emi – new profile developed in EMI project. This is default choice.

arguspep_usermap=*yes|no* – - specifies either response from Argus servie may define mapping of client to local account. Possible values are 'yes' and 'no'. Default is 'no' and that means any user mapping information is ignored. Argus is contacted after all other user mapping is performed. Hence it overwrites all other decisions.

arguspdp_endpoint=*URL* – specifies URL of Argus PDP service to use for authorization and user mapping. If this option is set Argus is contacted for every operation requested through WS interface.

arguspdp_profile=*profile name* – - defines which communication profile to use for passing information to Argus PEP service. Possible values are:
subject – pass only subject name of client,
cream – makes A-REX pretend it is gLite CREAM service,
emi – new profile developed in EMI project. This is default choice.

6.1.12.3 Commands setting control and session directories

controldir=*path* – sets the directory for A-REX to store control files containing information on jobs. Since this directory is heavily accessed by A-REX it should not be on a remote file system. This is a required command.

sessiondir=*path [drain]* – specifies the path to the directory in which the session directory (SD) is created. A-REX creates subdirectories in the SD for each job. Multiple SDs may be specified by specifying multiple *sessiondir* commands. In this case jobs are spread evenly over the directories. If the path is * the default sessiondir is used for each locally-mapped user - *$HOME/.jobs*. When adding a new SD, ensure to restart the A-REX so that jobs assigned there are processed. A SD can be drained prior to removal by adding the *"drain"* option (no restart is required in this case if *gridftp* interface is used). No new jobs will be assigned to this SD but running jobs will still be accessible. When all jobs are processed and the SD is empty, it can be removed from configuration and the A-REX should be restarted. This is a required command.

defaultttl=*ttl [ttr]* – specifies the time in seconds for a job's SD to be available after the job

finishes (*ttl*). A second optional number (*ttr*) defines the time from removal of the SD until all information about the job is discarded - the job stays in DELETED state during that period. Defaults are 7 days for *ttl* and 30 days for *ttr*. The minumum value for both parameters is 2 hours.

6.1.12.4 Commands to configure the cache

cachedir=*path [link_path]* – specifies a directory to store cached data (see Section Section **6.4**, *Cache*). Multiple cache directories may be specified by specifying multiple *cachedir* commands. Cached data will be distributed over multiple caches according to free space in each. Specifying no *cachedir* command or commands with an empty path disables caching. The optional *link_path* specifies the path at which *path* is accessible on computing nodes, if it is different from the path on the A-REX host. If *link_path* is set to '.' files are not soft-linked, nor are per-job links created, but files are copied to the session directory. If a cache directory needs to be drained, then *cachedir* should specify *"drain"* as the *link_path*.

remotecachedir=*path [link_path]* – specifies caches which are under the control of other A-REXs, but which this A-REX can have read-only access to (see Section 6.4.3). Multiple remote cache directories may be specified by specifying multiple *remotecachedir* commands. If a file is not available in paths specified by cachedir, the A-REX looks in remote caches. *link_path* has the same meaning as in *cachedir*, but the special path "replicate" means files will be replicated from remote caches to local caches when they are requested.

cachesize=*high_mark [low_mark]* – specifies high and low watermarks for space used on the file system on which the cache directory is located, as a percentage of total file system capacity. When the max is exceeded, files will be deleted to bring the used space down to the min level. It is a good idea to have each cache on its own separate file system. If no *cachesize* is specified, or it is specified without parameters, no cleaning is done. These cache settings apply to all caches specified by *cachedir* commands.

cachelifetime=*lifetime* – if cache cleaning is enabled, files accessed less recently than the *lifetime* time period will be deleted. Example values of this option are 1800, 90s, 24h, 30d. When no suffix is given the unit is seconds.

cachelogfile=*path* – specifies the filename where output of the *cache-clean* tool should be logged. Defaults to */var/log/arc/cache-clean.log*.

cacheloglevel=*number* – specifies the level of logging by the *cache-clean* tool, between 0 (FATAL) and 5 (DEBUG). Defaults to 3 (INFO).

cachecleantimeout=*timeout* – the timeout in seconds for running the cache-clean tool. If using a large cache or slow file system this value can be increased to allow the cleaning to complete. Defaults to 3600 (1 hour).

6.1.12.5 Commands setting limits

maxjobs=*max_processed_jobs [max_running_jobs [max_jobs_per_dn [max_jobs_total]]]* – specifies maximum number of jobs being processed by the A-REX at different stages:
max_processed_jobs – maximum number of concurrent jobs processed by A-REX. This does not limit the amount of jobs which can be submitted to the cluster.
max_running_jobs – maximum number of jobs passed to Local Resource Management System
max_jobs_per_dn – maximum number of concurrent jobs processed by A-REX per user DN. If this option is used the total maximum number of jobs processed is still *max_processed_jobs*.
max_jobs_total – total maximum number of jobs associated with service. It is advised to use this limit only in exceptional cases because it also accounts for finished jobs.

Missing value or -1 means no limit.

maxload=*max_frontend_jobs [emergency_frontend_jobs [max_transferred_files]]* – specifies maximum load caused by data staging being processed on frontend:
max_frontend_jobs – maximum number of jobs in PREPARING and FINISHING states (downloading and uploading files). Jobs in these states can cause a heavy load on the A-REX host. This

limit is applied before moving jobs to PREPARING and FINISHING states.
emergency_frontend_jobs – if limit of *max_frontend_jobs* is used only by PREPARING or by FINISHING jobs, aforementioned number of jobs can be moved to another state. This is used to avoid the case where jobs cannot finish due to a large number of recently submitted jobs.
max_transferred_files – maximum number of files being transferred in parallel by every job. Used to decrease load on not so powerful frontends.
Missing value or -1 means no limit.
If new data staging (DTR) is enabled this command is replaced by commands in the [data-staging] section (**6.1.13**, *Commands in the [data-staging] section*).

maxloadshare=*max_share share_type* – specifies a sharing mechanism for data transfer. *max_share* is the maximum number of processes that can run per transfer share and *share_type* is the scheme used to assign jobs to transfer shares.
for possible values and more details.
If new data staging (DTR) is enabled this command is replaced by commands in the [data-staging] section (**6.1.13**, *Commands in the [data-staging] section*).

share_limit=*name limit* – specifies a transfer share that has a number of processes different from the default value in maxloadshare. *name* is the name of the share and *limit* is the number of processes for this share. In the configuration it should appear after maxloadshare. Can be repeated several times for different shares.
If new data staging (DTR) is enabled this command is replaced by commands in the [data-staging] section (**6.1.13**, *Commands in the [data-staging] section*).

maxrerun=*number* – specifies maximal number of times job will be allowed to rerun after it failed at any stage. Default value is *5*. This only specifies a upper limit. The actual number is provided in job description and defaults to 0.

maxtransfertries=*number* – specifies the maximum number of times download and upload will be attempted per file (retries are only performed if an error is judged to be temporary, for example a communication error with a remote service). This number must be greater than 0 and defaults to 10.

6.1.12.6 Commands related to file staging

securetransfer=*yes|no* – specifies whether to use encryption while transferring data. Currently works for GridFTP only. Default is *no*. It may overridden for every source/destination by values specified in URL options.

passivetransfer=*yes|no* – specifies whether GridFTP transfers are passive. Setting this option to yes can solve transfer problems caused by firewalls. Default is no.

localtransfer=*yes|no* – specifies whether to pass file downloading/uploading task to computing node. If set to yes the A-REX will not download/upload files but compose script submitted to the LRMS in order that the LRMS can execute file transfer. This requires installation of A-REX and all related software to be accessible from computing nodes and environment variable ARC_LOCATION to be set accordingly. Default is *no*.

speedcontrol=*min_speed min_time min_average_speed max_inactivity* –
specifies how long or slow data transfer is allowed to take place. Transfer is canceled if transfer rate (bytes per second) is lower than *min_speed* for at least *min_time* seconds, or if average rate is lower than *min_average_speed*, or no data is received for longer than *max_inactivity* seconds. To allow statistics to build up, no transfers will be stopped within the first 3 minutes.

preferredpattern=*pattern* – specifies how to order multiple replicas of an input file according to preference. It consists of one or more patterns (strings) separated by a pipe character (|) listed in order of preference. Input file replicas will be matched against each pattern and then ordered by the earliest match. If the dollar character ($) is used at the end of a pattern, the pattern will be matched to the end of the hostname of the replica.

newdatastaging=*yes|no* – turns on or off the new data staging framework (DTR), which replaces the downloader and uploader utilities. See **4.4.4**, *Configuring Data Staging*. Default is *no*.

copyurl=*template replacement* – specifies that URLs starting from *template* should be accessed at *replacement* instead. The *template* part of the URL will be replaced with *replacement.* This option is useful when for example a Grid storage system is accessible as a local file system on the A-REX host. *replacement* can be either a URL or a local path starting from '/'. It is advisable to end template with '/'.

linkurl=*template replacement [node_path]* – mostly identical to *copyurl* but file will not be copied. Instead a soft-link will be created. *replacement* specifies the way to access the file from the frontend, and is used to check permissions. The *node_path* specifies how the file can be accessed from computing nodes, and will be used for soft-link creation. If *node_path* is missing, *local_path* will be used instead. Neither *node_path* nor *replacement* should be URLs.

NOTE: URLs which fit into *copyurl* or *linkurl* are treated as more easily accessible than other URLs. That means if A-REX has to choose between several URLs from which should it download input file, these will be tried first.

6.1.12.7 Commands related to usage reporting

jobreport=*URL ... number* – specifies that A-REX has to report information about jobs being processed (started, finished) to a remote service running at the given *URL*. Multiple entries and multiple URLs are allowed. *number* specifies how long (in days) old records have to be kept if failed to be reported. The last specified value becomes effective.

jobreport_publisher=*filename* – specifies the name of the executable which will be run by the A-REX periodically to publish job reports if a jobreport URL is specified. The executable will be searched in the nordugrid `libexec` directory. The default name is `jura`.

jobreport_credentials=*key_file [cert_file [ca_dir]]* – specifies the credentials for accessing the accounting service.

jobreport_options=*options* – specifies additional options for the usage reporter (e.g. JURA). See **4.4.6**, *Accounting with JURA*.

6.1.12.8 Other general commands in the [grid-manager] section

wakeupperiod=*time* – specifies how often the A-REX checks for job state changes (like new arrived job, job finished in LRMS, etc.). *time* is a minimal time period specified in seconds. Default is *3 minutes*. The A-REX may also be woken up by external processes such as LRMS scripts before this time period expires.

authplugin=*state options plugin* – specifies *plugin* (external executable) to be run every time job is about to switch to *state*. The following states are allowed: ACCEPTED, PREPARING, SUBMIT, FINISHING, FINISHED and DELETED. If exit code is not 0 job is canceled by default. *Options* consists of *name=value* pairs separated by commas. The following *name*s are supported:
timeout – specifies how long in seconds execution of the plugin allowed to last (mandatory, "*timeout=*" can be skipped for backward compatibility).
onsuccess, *onfailure* and *ontimeout* – defines action taken in each case (*onsuccess* happens if exit code is 0). Possible actions are:
pass – continue execution,
log – write information about result into log file and continue execution,
fail – write information about result into log file and cancel job. Default actions are *fail* for *onfailure* and *ontimeout*, *pass* for *onsuccess*.

localcred=*timeout plugin* – specifies *plugin* (external executable or function in shared library) to be run every time job has to do something on behalf of local user. Execution of *plugin* may not last longer than *timeout* seconds. If *plugin* looks like *function@path* then function *int function(char*,char*,char*,...)* from shared library *path* is called (*timeout* is not functional in that case). If exit code is not 0 current operation will fail. This functionality was introduced for acquiring Kerberos tickets for local filesystem access and is currently deprecated.

norootpower=*yes|no* – if set to *yes* all processes involved in job management will use local identity of a user to which Grid identity is mapped in order to access file system at path specified in

session command (see below). Sometimes this may involve running temporary external process under specified account.

joblog=*path* – specifies where to store log file containing information about started and finished jobs. This file contains one line per every started and evety finished job. Default is not to write such file.

mail=*e-mail_address* – specifies an email address **from** which notification mails are sent.

helper=*username command [argument [argument [...]]]* – associates an external program with A-REX. This program will be kept running under account of the user specified by *username*. Currently only '.' is supported as username, corresponding to the user running A-REX. *command* is an executable and *argument*s are passed as arguments to it. This executable is started under the specified local account and re-started every time it exits.

6.1.12.9 Global commands specific to communication with the underlying LRMS

gnu_time=*path* – path to *time* utility.

tmpdir=*path* – path to directory for temporary files.

runtimedir=*path* – path to directory which contains *runtimenvironment* scripts.

shared_filesystem=*yes|no* – if computing nodes have an access to session directory through a shared file system like NFS.
Corresponds to an environment variable RUNTIME_NODE_SEES_FRONTEND.
(See Section **6.12**, *Environment variables set for the job submission scripts*).

nodename=*command* – command to obtain hostname of computing node.

scratchdir=*path* – path on computing node where to move session directory before execution. Default is not to move session directory before execution.

shared_scratch=*path* – path on frontend where ***scratchdir*** can be found. Only needed if *scratchdir* is used.

6.1.12.10 Substitutions in the command arguments

In command arguments (paths, executables, ...) the following substitutions can be used:

%R – session root – see command *sessiondir*

%C – control dir – see command *controldir*

%U – username of mapped local user

%u – userid – numerical user id of mapped local user

%g – groupid – numerical group id of mapped local user

%H – home dir – home of username as specified in /etc/passwd

%Q – default queue – see command *lrms*

%L – lrms name – see command *lrms*

%W – installation path – ${ARC_LOCATION}

%F – path to configuration file of this instance

%I – job ID (for plugins only, substituted in runtime)

%S – job state (for *authplugin* plugins only, substituted in runtime)

%O – reason (for *localcred* plugins only, substituted in runtime). Possible reasons are:

new – new job, new credentials

renew – old job, new credentials

write – write/delete file, create/delete directory

read – read file, directory, etc.

extern – call external program

6.1.13 Commands in the [data-staging] section

The design of the new data staging framework (DTR) is described in Section **4.4.4**, *Configuring Data Staging*.

maxdelivery=*number* – maximum number of files in delivery, i.e. the maximum number of physical transfer slots.

maxprocessor=*number* – maximum number of files in each processing state. These states are for example checking the cache or resolving replicas in index services.

maxemergency=*number* – maximum emergency slots for delivery and processor. If the maximum slots as defined above are already used, a new transfer share can still proceed by using an emergency slot. These slots therefore stop shares from being blocked by others in busy situations.

maxprepared=*number* – maximum number of files prepared for transfer. For protocols such as SRM files are pinned for transfer. This parameter determines the maximum number of files that are pinned. It should normally be set a few times larger than maxdelivery so that new pinned transfers are ready to start when others finish, but not so high as to pass the limits of the storage systems. This number is applied per-transfer share and only for applicable protocols.

sharetype=*type* – sharing mechanism for data transfer. This is the scheme used to assign transfers to different shares. Possible values are:

- *dn* - each job is assigned to a share based on the DN of the user sumbitting the job.
- *voms:vo* - if the user's proxy is a VOMS [**?**] proxy the job is assigned to a share based on the VO specified in the proxy. If the proxy is not a VOMS proxy a default share is used.
- *voms:role* - if the user's proxy is a VOMS proxy the job is assigned to a share based on the role specified in the first attribute found in the proxy. If the proxy is not a VOMS proxy a default share is used.
- *voms:group* - if the user's proxy is a VOMS proxy the job is assigned to a share based on the group specified in the first attribute found in the proxy. If the proxy is not a VOMS proxy a default share is used.

definedshare=*share priority* – a transfer share with a specific priority, different from the default. This is used to give more or less transfer slots to certain shares. `priority` should be a number between 1 and 100 (higher number is higher priority). The default priority for a share is 50.

deliveryservice=*url* – remote delivery service for executing transfers. More details are available in the NorduGrid wiki.*

localdelivery=*yes|no* – in case remote delivery services are configured using the previous option, this option specifies whether or not delivery should also be done locally on the A-REX host. Default is *no*.

remotesizelimit=*number* – file size in bytes under which data transfer will always use local delivery rather than a remote delivery service. This can optimise performance when many small files are being transferred, where communication overhead with remote services can become a bottleneck.

usehostcert=*yes|no* – specifies whether to use the A-REX host certificate when contacting remote delivery services. This can be done for added security but only with host certificates which can be used as client certificates. Default is *no*.

dtrlog=*path* – file in which to periodically dump data staging information. This data can be used to monitor the system. Defaults to *controldir/dtrstate.log*.

*http://wiki.nordugrid.org/index.php/Data_Staging/Multi-host

6.1.14 PBS specific commands

For each grid-enabled (or Grid visible) PBS queue a corresponding `[queue/queuename]` subsection must be defined. `queuename` should be the PBS queue name.

lrms=*"pbs"* – enables the PBS batch system back-end

pbs_bin_path=*path* – in the `[common]` section should be set to the path to the qstat,pbsnodes,qmgr etc. PBS binaries.

pbs_log_path=*path* – in the `[common]` sections should be set to the path of the PBS server logfiles which are used by A-REX to determine whether a PBS job is completed. If not specified, A-REX will use the `qstat` command to find completed jobs.

lrmsconfig=*text* – in the `[cluster]` block can be used as an optional free text field to describe further details about the PBS configuration (e.g. `lrmsconfig="single job per processor"`). This information is then exposed through information interfaces.

dedicated_node_string=*text* – in the `[cluster]` block specifies the string which is used in the PBS node config to distinguish the Grid nodes from the rest. Suppose only a subset of nodes are available for Grid jobs, and these nodes have a common `node property` string, this case the `dedicated_node_string` should be set to this value and only the nodes with the corresponding PBS `node property` are counted as Grid enabled nodes. Setting the `dedicated_node_string` to the value of the PBS `node property` of the grid-enabled nodes will influence how the totalcpus, user freecpus is calculated. No need to set this attribute if the cluster is fully available for the Grid and the PBS configuration does not use the `node property` method to assign certain nodes to Grid queues.

scheduling_policy=*FIFO|MAUI* – in the `[queue/queuename]` subsection describes the scheduling policy of the queue. PBS by default offers the FIFO scheduler, many sites run the MAUI. At the moment `FIFO` & `MAUI` are supported values. If you have a MAUI scheduler you should specify the `"MAUI"` value since it modifies the way the queue resources are calculated. By default the `"FIFO"` scheduler type is assumed.

maui_bin_path=*path* – in the `[queue/queuename]` subsection sets the path of the MAUI commands like `showbf` when `"MAUI"` is specified as `scheduling_policy` value. This parameter can be set in the `[common]` block as well.

queue_node_string=*text* – in the `[queue/queuename]` block can be used similar to the configuration command `dedicated_node_string`. In PBS you can assign nodes to a queue (or a queue to nodes) by using the `node property` PBS node configuration method and assigning the marked nodes to the queue (setting the `resources_default.neednodes = queue_node_string` for that queue). This parameter should contain the `node property` string of the queue-assigned nodes. Setting the `queue_node_string` changes how the queue-totalcpus, user freecpus are determined for this queue.

6.1.15 Condor specific commands

lrms=*"condor"* – in the `[common]` section enables the Condor batch system back-end.

condor_bin_path=*path* – in the `[common]` section specifies location of Condor executables. If not set then ARC will try to guess it out of the system path.

condor_rank=*ClassAd float expression* – in the `[common]` section, if defined, will cause the `Rank` attribute to be set in each job description submitted to Condor. Use this option if you are not happy with the way Condor picks out nodes when running jobs and want to define your own ranking algorithm. `condor_rank` should be set to a ClassAd float expression that you could use in the `Rank` attribute in a Condor job description. For example:

```
condor_rank="(1-LoadAvg/2)*(1-LoadAvg/2)*Memory/1000*KFlops/1000000"
```

condor_requirements=*constraint string* – in the `[queue/queuename]` section defines a subpool of condor nodes. Condor does not support queues in the classical sense. It is possible,

however, to divide the Condor pool in several sub-pools. An ARC "queue" is then nothing more than a subset of nodes from the Condor pool.

Which nodes go into which queue is defined using the `condor_requirements` configuration option in the corresponding `[queue/queuename]` section. Its value must be a well-formed constraint string that is accepted by a `condor_status -constraint '...'` command. Internally, this constraint string is used to determine the list of nodes belonging to a queue. This string can get quite long, so, for readability reasons it is allowed to split it up into pieces by using multiple `condor_requirements` options. The full constrains string will be reconstructed by concatenating all pieces.

Queues should be defined in such a way that their nodes all match the information available in ARC about the queue. A good start is for the `condor_requirements` attribute to contain restrictions on the following: `Opsys`, `Arch`, `Memory` and `Disk`. If you wish to configure more than one queue, it's good to have queues defined in such a way that they do not overlap. In the following example disjoint memory ranges are used to ensure this:

```
[queue/large]
condor_requirements="(Opsys == "linux" && (Arch == "intel" || Arch == "x86_64")"
condor_requirements=" && (Disk > 30000000 && Memory > 2000)"
[queue/small]
condor_requirements="(Opsys == "linux" && (Arch == "intel" || Arch == "x86_64")"
condor_requirements=" && (Disk > 30000000 && Memory <= 2000 && Memory > 1000)"
```

Note that `nodememory` attribute in arc.conf means the maximum memory available for jobs, while the Memory attribute in Condor is the physical memory of the machine. To avoid swapping (and these are probably not dedicated machines!), make sure that `nodememory` is smaller than the minimum physical memory of the machines in that queue. If for example the smallest node in a queue has 1Gb memory, then it would be sensible to use `nodememory="850"` for the maximum job size.

In case you want more precise control over which nodes are available for Grid jobs, using pre-defined ClassAds attributes (like in the example above) might not be sufficient. Fortunately, it's possible to mark nodes by using some custom attribute, say `NORDUGRID_RESOURCE`. This is accomplished by adding a parameter to the node's local Condor configuration file, and then adding that parameter to `STARTD_EXPRS`:

```
NORDUGRID_RESOURCE = True
STARTD_EXPRS = NORDUGRID_RESOURCE, $(STARTD_EXPRS)
```

Now queues can be restricted to contain only "good" nodes. Just add to each `[queue/queuename]` section in arc.conf:

```
condor_requirements=" && NORDUGRID_RESOURCE"
```

6.1.16 LoadLeveler specific commands

lrms=*"ll"* – in the `[common]` section enables the LoadLeveler batch system.

ll_bin_path=*path* – in the `[common]` section must be set to the path of the LoadLeveler binaries.

6.1.17 Fork specific commands

lrms=*"fork"* – in the `[common]` section enables the Fork back-end. The queue must be named `"fork"` in the `[queue/fork]` subsection.

fork_job_limit=*cpunumber* – sets the number of running Grid jobs on the fork machine, allowing a multi-core machine to use some or all of its cores for Grid jobs. The default value is 1.

6.1.18 LSF specific commands

lrms="*lsf*" – in the [common] section enables the LSF back-end

lsf_bin_path=*path* – in the [common] section must be set to the path of the LSF binaries

lsf_profile_path=*path* – must be set to the filename of the LSF profile that the back-end should use.

Furthermore it is very important to specify the correct architecture for a given queue in arc.conf. Because the architecture flag is rarely set in the xRSL file the LSF back-end will automatically set the architecture to match the chosen queue. LSF's standard behaviour is to assume the same architecture as the frontend. This will fail for instance if the frontend is a 32 bit machine and all the cluster resources are 64 bit. If this is not done the result will be jobs being rejected by LSF because LSF believes there are no useful resources available.

6.1.19 SGE specific commands

lrms="*sge*" – in the [common] section enables the SGE batch system back-end.

sge_root=*path* – in the [common] section must be set to SGE's install root.

sge_bin_path=*path* – in the [common] section must be set to the path of the SGE binaries.

sge_cell=*cellname* – in the [common] section can be set to the name of the SGE cell if it's not the default

sge_qmaster_port=*port* – in the [common] section can be set to the qmaster port if the sge command line clients require the SGE_QMASTER_PORT environment variable to be set

sge_execd_port=*port* – in the [common] section can be set to the execd port if the sge command line clients require the SGE_EXECD_PORT environment variable to be set

sge_jobopts=*options* – in the [queue/queuename] section can be used to add custom SGE options to job scripts submitted to SGE. Consult SGE documentation for possible options.

6.1.20 SLURM specific commands

lrms="*SLURM*" – in the [common] section enables the SLURM batch system back-end.

slurm_bin_path=*path* – in the [common] section must be set to the path of the SLURM binaries.

slurm_wakeupperiod=*seconds* – How long should infosys wait before querying SLURM for new data.

6.1.21 Commands for the **urlogger** accounting component

Note that prefered way for handling accounting in A-REX is to use Jura components described in section 6.6.

These commands are in the [logger] section:

log_dir=*path* – sets the top directory for the generated usage records. The option will default to /var/spool/arc/usagerecords/ and can be left out. We suggest you leave out this option, unless you have a reason not to.

log_all=*URL [URL...]* – configures where the usage records are sent to. All usage records will be registered to the URLs specified with the log_all option. It is possible to specifiy multiple URLs separated by a space. There can be only one log_all command.

log_vo=*VO URL [, VO URL...]* – makes it possible to only register usage records run with certain VO users to a given URL. It is possible to have multiple entries, by seperating entries with comma. There can be only one log_vo command.

ur_lifetime=*days* – specifies how many days usage records are kept after being archived. The default is 30.

urlogger_logfile=*path* – specifies the logfile the `urlogger` should write its logs. Default is `/var/log/arc/ur-logger.log`.

urlogger_loglevel=*debug|info|warning* – specifies the verbosity of logging.

registrant_logfile=*path* – specifies the logfile for the `urlogger` registration module, default is `/var/log/arc/ur-registration.log`.

6.2 Handling of the input and output files

One of the most important tasks of the A-REX is to take care of processing of the input and output data (files) of the job. Input files are gathered in the session directory (SD) or in the associated cache area. If caching is enabled, then the A-REX checks the cache whether a requested input file is already present (with proper authorization checks and timely invalidation of cached data), and links (or copies) it to the SD of the job without re-downloading it. If file is present in the cache but not marked as authorized for specific grid identity then connection to data server is performed for authorization check.

There are two ways to put a file into the SD:

- If a file is specified in the job description as an input file with a remote source location: the A-REX contacts the remote location (using the user's delegated credentials) and downloads the file into the session directory or cache location using one of the supported protocols (GridFTP, FTP, HTTP, HTTPS, and also communicating with Grid file catalogs is supported).
- If a file is specified in the job description as an input file without a remote source location: the A-REX expects the client tool to upload the file to the session directory (SD) using the URL provided by A-REX. The client tools should do this step automatically.

The A-REX does not provide any other reliable way to obtain input data.

After the job finishes, the files in the session directory are treated in three possible ways:

- If a file is specified in the job description as an output file with a remote target location: the A-REX uploads the results to the remote storage (optionally register the file to a catalog), then it will remove the file from the session directory. If the job execution fails, these files will not be uploaded (but they will be kept for the user to download). Depending of used submission interface and corresponding job description format it is possible to request more sophisticated conditions for files processing in case of failure.
- If a file is specified in the job description as an output file without a target location: the A-REX will keep the files, and the user can download them by accessing the session directory. The client tools usually support downloading these files.
- If a file is not specified in the job description as an output file: the A-REX will remove the file from the session directory after the job finished.

It is possible to specify in job description an option to keep a whole directory, but if a file is not specified in the job description as an output file and it is not in a directory which is requested to be kept, it will be removed when the job is finished.

6.3 Job states

Figure 6.2 shows the *internal* states a job goes through, also listed here:

- **Accepted**: the job has been submitted to the CE but hasn't been processed yet.
- **Preparing**: the input data is being gathered.
- **Submitting**: the job is being submitted to the local resource management system (LRMS).

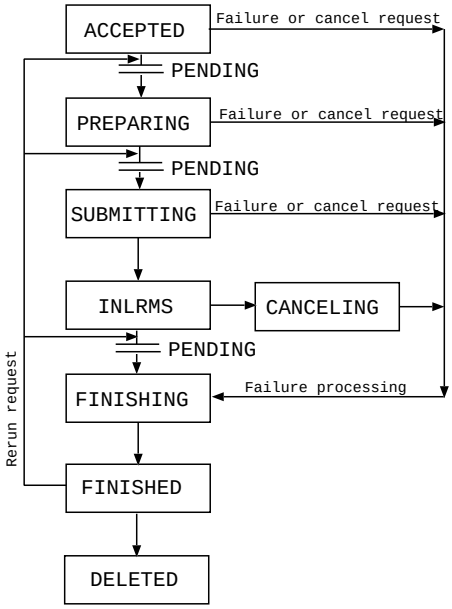


Figure 6.2: Job states

- **Executing (InLRMS)**: the job is queued or being executed in the LRMS.
- **Killing (Canceling)**: the job is being canceled.
- **Finishing**: the output data is being processed (even if there was a failure).
- **Finished**: the job is in this state either it finished successfully or there was an error during one of the earlier steps.
- **Deleted**: after specified amount of days the job gets deleted and only minimal information is kept about it.

Limits can be configured on the CE for the number of jobs in some states. If the limit reached, new jobs would stay in the preceding state (indicated by the **Pending** prefix). It is possible to re-run a job which is in the **Finished** state because of a failure. In this case the job would go back to the state where the failure happened.

These are internal states which are translated into more user-friendly external states when presented to the users. These external states take into account additional information, not just the internal state, so one internal state can correspond to multiple external states. In this list every row starts with the internal state followed by a colon and then the possible external states:

- ACCEPTED: ACCEPTING, ACCEPTED
- PREPARING: ACCEPTED, PREPARING, PREPARED
- SUBMITTING: PREPARED, SUBMITTING
- INLRMS: INLRMS, EXECUTED
- FINISHING: EXECUTED, FINISHING
- FINISHED: FINISHED, FAILED, KILLED
- CANCELING: KILLING
- DELETED: DELETED

6.4 Cache

6.4.1 Structure of the cache directory

Cached files are stored in sub-directories under the *data* directory in each main cache directory. Filenames are constructed from an SHA-1 hash of the URL of the file and split into subdirectories based on the two initial characters of the hash. In the extremely unlikely event of a collision between two URLs having the same SHA-1 hash, caching will not be used for the second file.

When multiple caches are used, a new cache file goes to a randomly selected cache, where each cache is weighted according to the size of the file system on which it is located. For example: if there are two caches of 1TB and 9TB then on average 10% of input files will go to the first cache and 90% will go to the second cache.

Some associated metadata including the corresponding URL and an expiry time, if available, are stored in a file with the same name as the cache file, with a *.meta* suffix.

For example, with a cache directory */cache*, the file

lfc://atlaslfc.nordugrid.org/grid/atlas/file1
is mapped to
/cache/data/78/f607405ab1df6b647fac7aa97dfb6089c19fb3

and the file */cache/data/78/f607405ab1df6b647fac7aa97dfb6089c19fb3.meta* contains the original URL and an expiry time if one is available.

At the start of a file download, the cache file is locked, so that it cannot be deleted and so that another download process cannot write the same file simultaneously. This is done by creating a file with the same name as the cache filename but with a *.lock* suffix. This file contains the process ID of the process and the hostname of the host holding the lock. If this file is present, another process cannot do anything with the cache file and must wait until the cache file is unlocked (i.e. the *.lock* file no longer exists). The lock is continually updated during the transfer, and is considered stale if 15 minutes have passed since the last update. These stale locks, caused for example by a download process exiting abnormally, will therefore automatically be cleaned up. Also, if the process corresponding to the process ID stored inside the lock is no longer running on the host specified in the lock, it is safe to assume that the lock file can be deleted. If a file is requested which already exists in the cache (and is not locked), the cache file is not locked, but checks are done at the end of cache processing to ensure the file was not modified during the processing.

6.4.2 How the cache works

If a job requests an input file which can be cached or is allowed to be cached, it is stored in the selected cache directory, then a hard link is created in a per-job directory, under the *joblinks* subdirectory of the main cache directory. Then depending on the configuration, either the hard-link is copied or soft-linked to the SD. The former option is advised if the cache is on a file system which will suffer poor performance from a large number of jobs reading files on it, or the file system containing the cache is not accessible from worker nodes. The latter option is the default option. Files marked as executable in the job will be stored in the cache without executable permissions, but they will be copied to the SD and the appropriate permissions applied to the copy.

The per-job directory is only readable by the local user running the job, and the cache directory is readable only by the A-REX user. This means that the local user cannot access any other users' cache files. It also means that cache files can be removed without needing to know whether they are in use by a currently running job. However, as deleting a file which has hard links does not free space on the disk, cache files are not deleted until all per-job hard links are deleted. **IMPORTANT:** If a cache is mounted from an NFS server and the A-REX is run by the root user, the server must have the *no_root_squash* option set for the A-REX host in the */etc/exports* file, otherwise the A-REX will not be able to create the required directories. Note that when running A-REX under a non-privileged user account, all cache files will be owned and accessible by the same user, and therefore modifiable by running jobs. This is potentially dangerous and so caching should be used with caution in this case.

If the file system containing the cache is full and it is impossible to free any space, the download fails and is retried without using caching.

Before giving access to a file already in the cache, the A-REX contacts the initial file source to check if the user has read permission on the file. In order to prevent repeated checks on source files, this authentication information is cached for a limited time. On passing the check for a cached file, the user's DN is stored in the *.meta* file, with an expiry time equivalent to the lifetime remaining for the user's proxy certificate. This means that the permission check is not performed for this user for this file until this time is up (usually several hours). File creation and validity times from the original source are also checked to make sure the cached file is fresh enough. If the modification time of the source is later than that of the cached file, the file will be downloaded again. The file will also be downloaded again if the modification date of the source is not available, as it is assumed the cache file is out of date. These checks are not performed if the DN is cached and is still valid.

The A-REX checks the cache periodically if it is configured to do automatic cleaning. If the used space on the file system containing the cache exceeds the high water-mark given in the configuration file it tries to remove the least-recently accessed files to reduce size to the low water-mark.

6.4.3 Remote caches

If a site has multiple A-REXs running, an A-REX can be configured to have its own caches and have read-only access to caches under the control of other A-REXs (remote caches). An efficient way to reduce network traffic within a site is to configure A-REXs to be under control of caches on their local disks and have caches on other hosts as remote caches. If an A-REX wishes to cache a file and it is not available on the local cache, it searches for the file in remote caches. If the file is found in a remote cache, the actions the A-REX takes depends on the policy for the remote cache. The file may be replicated to the local cache to decrease the load on the remote file system caused by many jobs accessing the file. However, this will decrease the total number of cache files that can be stored. The other policy is to use the file in the remote cache, creating a per-job directory for the hard link in the remote cache and bypassing the local cache completely. The usual permission and validity checks are performed for the remote file. Note that no creation or deletion of remote cache data is done - cache cleaning is only performed on local caches.

6.4.4 Cache cleaning

The cache is cleaned automatically periodically (every 5 minutes) by the A-REX to keep the size of each cache within the configured limits. Files are removed from the cache if the total size of the cache is greater than the configured limit. Files which are not locked are removed in order of access time, starting with the earliest, until the size is lower than the configured lower limit. If the lower limit cannot be reached (because too many files are locked, or other files outside the cache are taking up space on the file system), the cleaning will stop before the lower limit is reached.

Since the limits on cache size are given as a percentage of space used on the filesystem on which the cache is located, it is recommended that each cache has its own dedicated file system. If the cache shares space with other data on a file system, changes in the amount of non-cache data will result in changes in the available cache space.

With large caches mounted over NFS and an A-REX heavily loaded with data transfer processes, cache cleaning can become slow, leading to caches filling up beyond their configured limits. For performance reasons it may be advantageous to disable cache cleaning by the A-REX, and run the *cache-clean* tool independently on the machine hosting the file system. Please refer to Section **5.3**, *Administration tools*.

Caches can be added to and removed from the configuration as required without affecting any cached data, but after changing the configuration file, the A-REX should be restarted. If a cache is to be removed and all data erased, it is recommended that the cache be put in a *draining* state until all currently running jobs possibly accessing files in this cache have finished. In this state the cache will not be used by any new jobs, but the hard links in the *joblinks* directory will be cleaned up as each job finishes. Once this directory is empty it is safe to delete the entire cache

6.5 Batch system back-ends implementation details

The batch system back-ends are what tie the ARC Grid middleware (through the ARC Resource-coupled EXecution service, A-REX to the underlying cluster management system or LRMS. The back-ends consist of set of a shell and Perl scripts whose role are twofold:

1. to allow A-REX, to control jobs in the LRMS including job submit, status querying and cancel operations.

2. to collect information about jobs, users, the batch system and the cluster itself for the Information System.

The job control part of the LRMS interface is handled by the A-REX. It takes care of preparing a native batch system submission script, managing the actual submission of the batch system job, cancellation of job on request and scanning for completed batch jobs. Besides the LRMS job control interface it is also A-REX which provides e.g. the data staging and communication with the Grid client, provides RTE environments, arranges file staging (to the node via LRMS capability), dealing with stdout/stderr, etc. The job control batch system interface of A-REX requires three programs. These programs can be implemented any way the designer sees it fits, but all the existing back-end interfaces use shell scripting for portability and ease of tailoring to a specific site. A-REX will call the following programs: cancel-LRMS-job, submit-LRMS-job, and scan-LRMS-job where LRMS is replaced with the short hand name for the LRMS; e.g. cancel-pbs-job. The scripts are described one by one in the following subsections.

6.5.1 Submit-LRMS-job

The submit program is the most involved. It is called by A-REX once a new job arrives and needs to be submitted to the LRMS. It is given the GRAMi file as argument on execution. The GRAMi file is a file in the job control directory containing the job description in a flat list of key-value pairs. This file is created by A-REX and is based on the JSDL job description. Submit-LRMS-job then has to set up the session directories, run-time environment and anything else needed. Then it submits the job to the local LRMS. This is normally done by generating a native job script for the LRMS and then running the local submit command, but it can also be done through an API if the LRMS supports it.

6.5.2 Cancel-LRMS-job

If a Grid user cancels his job, the message will reach the grid-manager. The manager will then call the cancel-LRMS-job for the suitable back-end. The cancel script is called with the GRAMi file containing information about the job such as the job id in the LRMS. Cancel-LRMS-job must then use that information to find the job and remove it from the queue or actually cancel it if it is running in the LRMS.

6.5.3 Scan-LRMS-job

The scan-LRMS-job is run periodically. Its job is to scan the LRMS for jobs that have finished. Once it has found a finished job it will write the exit-code of that job to the file job.{gridid}.lrms_done in the ARC job status directory[†]. Then it will call the gm-kick program to notify A-REX about the finished job. Subsequently, A-REX starts finalizing the job.

Generally, two approaches are taken to find jobs which are finished in LRMS. One is to directly ask the LRMS. Since all started Grid jobs have its own status file[‡] found in the job status directory, this can be done by checking if the status is "INLRMS" in this file. If so, a call to the LRMS is made asking for the status of the job (or jobs if several jobs have status "INLRMS"). If it is finished, it is marked as such in the job status directory, and the gm-kick program is activated. For most LRMSs the information about finished jobs are only available for a short period of time after the job finished. Therefore appropriate steps have

[†]normally /var/spool/nordugrid/jobstatus/, but can be set via the controldir variable of arc.conf
[‡]job.{gridid}.status

to be taken if the job has the status "INLRMS" in the job status directory, but is no longer present in the LRMS. The normal approach is to analyze the job's status output in the session directory.

The second approach is to parse the LRMSs log files. This method has some drawbacks like e.g.: A-REX has to be allowed read access to the logs. The back-end will then have to remember where in the log it was last time it ran. This information will have to be stored in a file somewhere on the front-end.

6.5.4 PBS

The job control batch interface makes use of the `qsub` command to submit native PBS job scripts to the batch system. The following options are used:

`-l nodes, cput, walltime, pvmem, pmem,`

`-W stagein, stageout`

`-e, -j eo`

`-q`

`-A`

`-N`

For job cancellation the `qdel` command is used. To find completed jobs, i.e. to scan for finished jobs the `qstat` command or the `PBS server log file` is used.

The information system interface utilizes the `qstat -f -Q queuename` and `qstat -f queuename` commands to obtain detailed job and queue information. `qmgr -c "list server"` is used to determine PBS flavour and version. The `pbsnodes` command is used to calculate total/used/free cpus within the cluster. In case of a MAUI scheduler the `showbf` command is used to determine user freecpu values. All these external PBS commands are interfaced via parsing the commands' output.

6.5.5 Condor

The job control part of the interface uses the `condor_submit` command to submit jobs. Some of the options used in the job's ClassAd are:

`Requirements` – is used to select the nodes that may run the job. This is how ARC queues are implemented for Condor.

`Periodic_remove` – is used to enforce cputime and walltime limits.

`Log` – the job's condor log file is parsed by the information scripts to find out whether the job was suspended.

The information system component uses the following Condor commands:

`condor_status -long` – for collecting information about nodes

`condor_status -format "%s\n" Machine -constraint '...'` – for listing nodes that make up an ARC queue.

`condor_q -long -global` – for monitoring running jobs.

`condor_history -l jobid` – for collecting information about finished jobs. Further cues are taken from the job's condor log file and the body of the email sent by Condor when a job completes.

6.5.6 LoadLeveler

The LoadLeveler back-end uses LoadLeveler's command line interface(CLI) commands to submit and cancel jobs. All information in the information system is similarly parsed from the output of CLI commands. It does not parse any log files, nor does it use the binary APIs. The reason that the back-end is completely

based on the CLI is that the log files are normally kept on another machine than the front end and that the binary API for LL changes quite often. Often with each new version of LL.

6.5.7 Fork

The Fork back-end implements an interface to the "fork" UNIX command which is not a batch system. Therefore the back-end should rather be seen as an interface to the operating system itself. Most of the "batch system values" are determined from the operating system (e.g. cpu load) or manually set in the configuration file.

6.5.8 LSF

The LSF implementation of the back-end are based solely on parsing and running LSF's command line interface commands. No log files or other methods are used. To get the correct output o any output at all the back-end needs to have an appropriate LSF profile. The path to this profile must be set in arc.conf. It will then be executed by the back-end before running any of LSF's CLI commands.

6.5.9 SGE

The SGE back-end's commands are similar to the PBS commands. These commands are used in the code:

Submit job:

- qsub -S /bin/sh (specifies the interpreting shell for the job)

Get jobs status:

If the job state is not suspended, running or pending then its state is failed.

- qstat -u '*' -s rs (show the running and suspended jobs status)
- qstat -u '*' -s p (show the pending jobs status)
- qstat -j job_id (long job information)
- qacct -j job_id (finished job report)

Job terminating:

- qdel job_id (delete Sun Grid Engine job from the queue)

Queue commands:

- qconf -spl (show a list of all currently defined parallel environments)
- qconf -sql (show a list of all queues)
- qconf -sep (show a list of all licensed processors/slots)
- qstat -g c (display cluster queue summary)
- qconf -sconf global (show global configuration)
- qconf -sq queue_name (show the given queue configuration)

Other:

- qstat -help (show Sun Grid Engine's version and type)

A-REX log entry	OGF-UR	CAR	ARC-UR
nodename, ngjobid (hyphen-separated, with custom prefix)	RecordIdentity:*	RecordIdentity*	
nodename		Host	
globalid	GlobalJobId:	GlobalJobId	globaljobid
localid	LocalJobId:	LocalJobId*	localid
[MISSING]	ProcessId:		processid
usersn	GlobalUserName:	GlobalUserName	globaluserid
localuser	LocalUserId:	LocalUserId*	localuserid
jobname	JobName:	JobName	jobname
[MISSING]	Charge:	Charge	charge
status (no conversion yet!)	Status:*	Status*	status
usedwalltime (sec)	WallDuration: (ISO)	WallDuration* (ISO)	usedwalltime
usedcputime (sec)	CpuDuration: (ISO)	CpuDuration* (ISO) all	usedcputime
usedusercputime (sec)	user	user	
usedkernelcputime (sec)	kernel	kernel	
submissiontime	StartTime: (UTC)	StartTime* (UTC)	submissiontime
endtime	EndTime: (UTC)	EndTime* (UTC)	endtime
nodename	Host:	Host	nodename
clienthost (port number removed)	SubmitHost:		submithost
headnode	MachineName:	MachineName*, SubmitHost, Site*	
lrms		Infrastructure* (desc.)	
queue	Queue:	Queue*	queue (lrms)
projectname	ProjectName:	GroupAttribute	projectname, usercert
usercert	VO	Group	usercert
exitcode		ExitStatus	
[MISSING]	ServiceLevel:	ServiceLevel*	
[MISSING]	Network:		network
[MISSING]	Disk:		useddisk
usedmemory	Memory: virtual,average	Memory: Share, average	usedmemory
usedmaxresident	Memory: physical,max	Memory: Physical, max	
usedaverageresident	Memory: physical,average	Memory: Physical, average	
[MISSING]	Swap:	Swap	usedswap
nodecount	NodeCount:	NodeCount	nodecount
[MISSING]	Processors:	Processors	processors
[MISSING]	TimeDuration:		
[MISSING]	TimeInstant:	TimeInstant	
[MISSING]	Extension:		
[MISSING]		Middleware	

Figure 6.3: Mapping of usage record properties, mandatory properties are flagged with an asterisk (*)

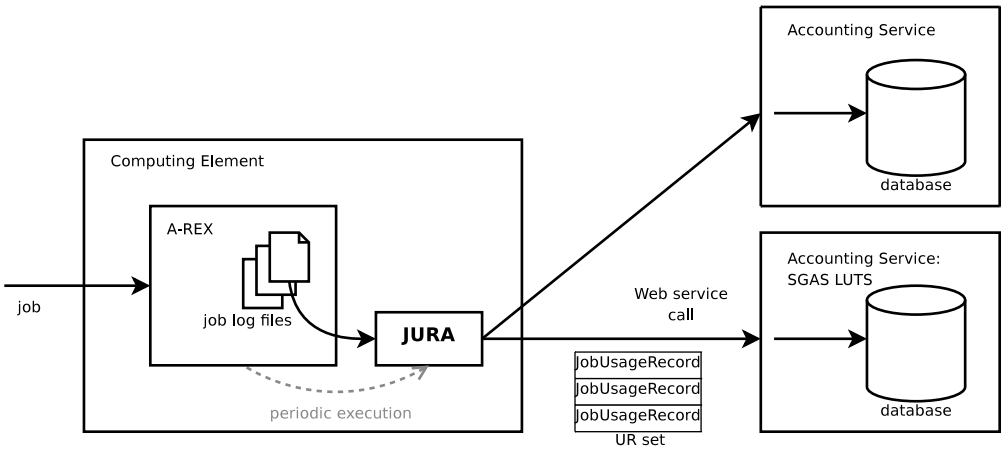


Figure 6.4: The usage reporting mechanism

6.6 JURA: The Job Usage Reporter for ARC

6.6.1 Overview

JURA is a stand-alone binary application which is periodically run by the A-REX (see Figure 6.4). There is no designated configuration file for JURA, nor is the configuration file of A-REX read directly by the application. Instead, options related to reporting are included within the job log files generated by the A-REX or supplied via command line argument. The primary purpose of these job log files is to hold metadata about jobs starting, running and stopping. This is the main input of JURA.

The application is run periodically. First, it processes the job log files, and based on the target accounting service specified in them, JURA creates usage records in a format appropriate for the target accounting service. Then these records are sent to one or more accounting services, referred to as *reporting destinations* in this document. Several reporting destinations are supported, these can be configured by the system administrator in the A-REX configuration file, and in addition, the user submitting the job can specify destinations in the job description.

About configuration of JURA, see **4.4.6**, *Accounting with JURA*.

6.6.2 Job log files

The A-REX puts the following extra information into the job log files:

- ***key_path*** – Path to the private key file used when submitting records.
- ***certificate_path*** – Path to the certificate file used when submitting records.
- ***ca_certificates_dir*** – Directory holding the certificates of trusted CAs.
- ***accounting_options*** – Additional configuration options for JURA.

The A-REX generates at least two job log files for each job and for each reporting destination: one at the time of job submission, another one after the job finishes, and possibly others at each start and stop event.

The job log files generated by A-REX reside under the directory `<control_dir>/logs`. The name of the job log files consist of the ID of the job and a random string to avoid collision of multiple job log files for the same job: `<jobid>.<random>`.

The job log file consists of "`name=value`" lines, where "`value`" is either a job-related resource usage data or a configuration parameter.

If interactive mode is not activated by the "`-u`" option, after successful submission to a reporting destination, the job log file is deleted, thus preventing multiple insertion of usage records. If submission fails, the log files

are kept, so another attempt is made upon a subsequent run of JURA. This mechanism will be repeated until the expiration time passes at which point the next execution of JURA removes the file without processing.

6.6.3 Archiving

The archiving functionality allows to store generated usage records in a specified directory on the disk. If enabled, the generated usage records are written to files named "`usagerecord[CAR].<jobid>.<random>`". If a job log file is processed repeatedly – for example because of temporary connection failures to the remote accounting service – and a respective usage record archive file already exists, then the usage record is not generated again. Instead, the contents of the archive file are used without change (the creation time stamp is also retained).

6.6.4 Reporting to LUTS

In case of non-interactive invocation of JURA by A-REX, the generated URs are submitted to the accounting services specified by the reporting destination configuration parameters and if present, to the destinations specified in the job description as well. Under interactive mode of operation, they are submitted to the services given via the "*-u*" command line option. Reporting URs to several destinations is possible.

LUTS has a simple custom web service interface loosely based on WS-ResourceProperties[**?**]. JURA uses the insertion method of this interface to report URs. The corresponding job log files are deleted after receiving a non-fault response from the service.

To increase communication efficiency JURA can send URs in batches provided that the server side supports this feature. LUTS accepts a batch of URs in a single request. The batch is an XML element called *UsageRecords*, containing elements representing URs.

The process of handling batches is the following: JURA does not send all usage records immediately after generation, but instead collects them in a batch until reaching the maximal number of records or until running out of job log files. The maximal number of URs in a batch can be set as a configuration parameter of JURA ("*jobreport_options*", see Section 6.1.12.7).

6.6.5 Reporting to APEL

Reporting mechanism is almost same as the LUTS but here are the different things:

- generated messages are XML based CAR records
- APEL publisher[**?**] transfer the generated records
- APEL publisher[**?**] use different message path location for every destination where to put the generated messages that will be transfer.
  - path format: */var/spool/arc/ssm/<destination host name>/outgoing/00000000/*
  - generated file name format in this location: *<YYYYMMDDhhmmss>*

6.6.6 Security

The JURA executable runs with the same user privileges as the A-REX. The owner of a job log file is the local user mapped for the submitter entity of the corresponding job. Since these files contain confidential data, A-REX restricts access to them allowing only read access for the job owner, thus when JURA is executed by A-REX it is allowed to read and delete job log files.

All usage records are submitted using the X.509 credentials specified by the value of the `jobreport_credentials` value of the A-REX configuration file. No proxies are used.

6.6.7 Mapping of job log entries to usage record properties

See **6.3**, *Mapping of usage record properties, mandatory properties are flagged with an asterisk (*).*

6.7 The XML and the INI configuration formats

This section clarifies the roles of the different configuration file formats used in ARC.

The main service of the ARC Computing Element is the A-REX, which runs in a *service container* called the *HED*. The HED is part of the ARC middlware, it contains a daemon and several loadable modules. The configuration of the HED can be done with an XML configuration file, which describes how to connect the several internal components to eachother.

Administrators of the ARC CE usually only uses the `arc.conf` configuration file. The HED service container does not understand the format of the `arc.conf` file, that's why the init script of the A-REX has to parse the `arc.conf` and generate a configuration file which is suitable for the HED service container.

The low-level configuration format of the HED service container is an XML-based format, which allows for very fine-grained configuration of all the internal components and other services than the A-REX.

There is a higher-level configuration possibility, which has the INI format, and which is transformed into the XML configuration when the HED service container is started. Putting configuration options of the INI config enables and sets sections of an XML configuration profile.

Although the original `arc.conf` also has an INI-like format, it should not be confused with the high-level INI configuration of the HED service container.

To summarize the three configuration formats:

arc.conf parsed by the A-REX init script and the A-REX itself. The init script generates an XML configuration from it to configure the HED service container

XML is the low-level configuration format of the HED service container

INI is the high-level configuration format of the HED service container (and it has nothing to do with the `arc.conf`)

For details on the XML and INI configuration, please see [**?**].

6.8 The internals of the service container of ARC (the HED)

This section describes the internal components of the ARC service container.

The main service of the ARC Computing Element is the A-REX, which runs in a *service container* called the *HED*. The HED is part of the ARC middlware, it contains a daemon and several loadable modules. The configuration of the HED can be done with an XML configuration file, which describes how to connect the several internal components to eachother. Here follows a short description of these internal components.

6.8.1 The MCCs

The HED service container has zero or more *message chains*. A message chain can contain *Message Chain Components (MCCs)*, *Plexers* and *Services*. These components are passing messages to eachother, which messages can have various extra attributes which also can be read and set by these components.

An MCC gets a message from a previous MCC, does something with it, passes it to the next MCC (or Plexer or Service), then "waits" for the response of this next MCC, then it does something with the response and passes it back to the previous MCC. (In case of the server-side TCP MCC, it is listening on a network port, and gets the message from there, not from another MCC, and the response will be sent back to the network, not to another MCC—the client-side TCP MCC is not listening but opening a connection to a host and port and sending the message into it.)

The Plexer checks the path of the destination of the message, and based on the matching of regular expressions it sends the message to one of the configured Services (or other MCCs or Plexers). The path checked by the Plexer comes from a message attribute called "ENDPOINT". The Plexer treats this as a URL, and uses the "path" part of it. This message attribute is set by the MCCs before the Plexer, usually by the HTTP MCC, which puts the URL of the HTTP request there. (But it is possible that a SOAP request

contains a WS-Addressing information with an endpoint in it, then this will be set as "ENDPOINT" by the SOAP MCC.)

A Service processes the message and produces a result which it will pass back to the previous element (Plexer, MCC). If we only have one web service in a chain which lives at the root HTTP path (e.g. https://hostname:port), then we don't need a Plexer. If we want to run a Service without any interface, then we don't need MCCs, we can have a chain with a single Service inside.

The XML configuration of the HED has "Chain" XML elements which contains the MCCs (in XML elements called "Component") and Plexers ("Plexer") and Services ("Service").

The MCCs:

TCP handles the external network communication (IPv4 and IPv6)

TLS handles the TLS/SSL connection, it is usually directly after the TCP MCC, it extracts information from the message like the Subject Name (DN) of the client

HTTP handles HTTP communication, should be directly after the TLS MCC (or can be directly after the TCP MCC, if we want HTTP without TLS security). The HTTP MCC can route the message to different components depending on the HTTP method (POST, GET, PUT, HEAD). Usually the POST method goes through the SOAP MCC, while the others skip the SOAP handling.

SOAP handles the SOAP communication. To use it, the HTTP MCC should directly route the POST message to the SOAP MCC.

If there is a Plexer, a useful configuration could be the following: the HTTP MCC sends the POST messages through the SOAP MCC to the Plexer, and sends the GET, PUT and HEAD messages directly to the Plexer. Then there are several Services sitting after the Plexer, and the Plexer would route the message to a selected service based on the configured regular expressions. The Plexer can handle messages from the HTTP MCC and from the SOAP MCC as well (actually, from the TCP MCC also, but that's currently not useful at all, because there is no endpoint path in that case). This enables services like the A-REX to handle both SOAP messages and non-SOAP requests (e.g. uploading input files through HTTP).

6.8.2 The SecHandlers

All MCCs and Services can have zero or more *Security Handlers (SecHandlers)* in one or more queues. Usually an MCC (or a Service) has two queues, one for the incoming message, and the other for the outgoing response. When the message comes into an MCC, the first SecHandler in the incoming queue gets it, and checks it. If it says "denied" then the message will be discarded and an error message will be sent back as a respone. If it says "ok" then the next SecHandler in the queue gets it, etc.

The SecHandlers adds security attributes to the message, which can be used later by other SecHandlers or Services (e.g. the A-REX can use them to figure out which user to map to).

In the XML configuration of the HED each MCC or Service element can have zero or more SecHandler elements, which specifies the name of the SecHandler (based on this name, the binary plugin will be loaded) and the name of the queue (called "event": incoming or outgoing).

ARC has the following SecHandlers:

- ArgusPEPClient (`arguspepclient.map`)
- ArgusPDPClient (`arguspdpclient.map`)
- IdentityMap (`identity.map`)
- DelegationCollector (`delegation.collector`)
- ArcAuthZ (`arc.authz`)
- DelegationSH (`delegation.handler`)
- LegacyMap (`arclegacy.map`)

- LegacySecHandler (`arclegacy.handler`)
- SAML2SSO_AssertionConsumerSH (`saml2ssoassertionconsumer.handler`)
- SAMLTokenSH (`samltoken.handler`)
- UsernameTokenSH (`usernametoken.handler`)

Each of them has an associated PluginDescriptor object, which are used the create the `.apd` files (Arc Plugin Description) which can be found next the the loadable modules (libraries on linux) in the installed location (e.g. `/usr/local/lib/arc`). The `.apd` files are generated by the `arcplugin` utility, and they are used by the HED when it tries to find the plugin based on its name. (If there are no `.apd` files, then all the modules have to be loaded in order to find the plugins, which takes more time.)

The names of the SecHandlers sometimes contains `.map` or `.handler` or `.authz` as a suffix. This is just a naming convention, they are all Security Handlers.

The most important SecHandlers:

IdentityMap

This security handler tries to map the Grid user (the DN which comes from the TLS MCC) to a local user. If it finds a local user, it puts the username into a security attribute which can be used later by other components. (This SecHandler never denies the message going forward, the worst thing can happen is that it doesn't find a local user, so the security attribute will be empty.)

It has three ways to do the mapping:

LocalName always maps to the specified local user regardless of the DN

LocalList uses the familiar grid-map file format to find the local name

LocalSimplePool maps to a pool of local user in a way that only one DN should be mapped to one local user (and a directory will contain the current mappings)

This SecHandler uses plugins called PDPs, see later.

ArcAuthZ

This is the main ARC security handler, which uses a couple of plugins (PDPs) to decide if a connection should go through or should be stopped and denied. (This does not do any Grid user – local user mapping.) So when a message arrives to the ArcAuthZ SecHandler, it will run the configured PDPs, and if any of them says "denied" then the message will be denied. If all the PDPs say "ok", then the message can go forward (this behaviour can be changed by configuration). About the PDPs, see later.

LegacyMap

This SecHandler does user mapping, it uses the `arc.conf` and does the mapping according to the `unixgroup`, `unixvo` and `unixmap` configuration parameters.

LegacySecHandler

This SecHandler does not map to local user or deny messages, only collects information which will be used later by the LegacyMap SecHandler (and by the Legacy PDP, see next section). It uses the `arc.conf`, and figures out which VOs and Groups the user belongs to.

ArgusPEPClient and ArgusPDPClient

These modules communicate to Argus PEP and PDP servers to obtain authorization and optional user mapping decision.

6.8.3 The PDPs

Some of the SecHandlers use another layer of plugins, which are called *Policy Decision Points (PDPs)*. (Currently only the IdentityMap and the ArcAuthZ SecHandlers use PDPs.) These SecHandlers have a queue of PDPs and they run them one by one, doing different things based on the results. The IdentityMap plugin has a given mapping policy (user, list, pool) for each PDP, and at the first PDP which returns "ok", it will stop running further PDPs and use the mapping policy configured for the given PDP. The ArcAuthZ runs all the PDPs, and only accepts the message if all of them returns "ok". (Although it can be configured differently, e.g. to accept the message if at least one PDP says "ok", or only accept a message if a PDPs says "deny", etc.)

The current PDPs:

- LegacyPDP (`arclegacy.pdp`)
- SimpleListPDP (`simplelist.pdp`)
- ArcPDP (`arc.pdp`)
- XACMLPDP (`xacml.pdp`)
- PDPServiceInvoker (`pdpservice.invoker`)
- DelegationPDP (`delegation.pdp`)
- AllowPDP (`allow.pdp`)
- DenyPDP (`deny.pdp`)

The most important ones:

LegacyPDP

This one check the previously set (by the LegacySecHandler) Group and VO attributes, and it also checks the `arc.conf`, and figures out if the given user is allowed or not.

SimpleListPDP

This one checks a given file with a list of DNs (can be a grid-map file), and only accepts messages from DNs listed in the file.

ArcPDP

This one parses policy file written in a general purpose policy language (developed by KnowARC) and makes a decision based on it.

AllowPDP

This one always allows.

DenyPDP

This one always denies.

6.9 How the a-rex init script configures the HED

The `a-rex` init script extracts information from the arc.conf, and creates an XML configuration for the HED. The A-REX service (living inside the HED) itself uses the `arc.conf` to configure itself, but there is a higher layer of configuration options which has to be set in the HED directly (e.g. authentication of the TLS communication), this configuration parameters has to be extracted from the `arc.conf` before the A-REX can even be started, and a proper XML configuration has to be assembled to configure the HED itself.

The `a-rex` init script first decides if the A-REX would have a web service interface or not. If the web service is disabled, then the XML configuration of the HED would look like this:

```
<?xml version="1.0"?>
<ArcConfig
xmlns="http://www.nordugrid.org/schemas/ArcConfig/2007"
xmlns:arex="http://www.nordugrid.org/schemas/a-rex/Config">
<Server>
  <PidFile>$PID_FILE</PidFile>
  <Logger>
    <File>$LOGFILE</File>
    <Level>$LOGLEVEL</Level>
    <Backups>$LOGNUM</Backups>
    <Maxsize>$LOGSIZE</Maxsize>
    <Reopen>$LOGREOPEN</Reopen>
  </Logger>
</Server>
<ModuleManager>
  <Path>$ARC_LOCATION/@pkglibsubdir@/</Path>
</ModuleManager>
<Plugins><Name>arex</Name></Plugins>
<Chain>
  <Service name="a-rex" id="a-rex">
    <arex:gmconfig>$ARC_CONFIG</arex:gmconfig>
  </Service>
</Chain>
</ArcConfig>
```

The variables (names starting with a dollar sign) are substituted with values from the `arc.conf`. Here the message chain contains only a single A-REX service, which has one single config parameter: “gmconfig”, which points to the location of the `arc.conf`. In this case the A-REX does not have any HTTP or SOAP interfaces, no SecHandlers, no PDPs, because everything is done by the GridFTP Server, which has a separate init script, it is a separate process, and it has all the authentication and authorization mechanisms built-in.

When the web service interface is enabled, then the job submission through the web service interface would go through through the following components:

- a TCP MCC listening on the given port:

  ```
  <Component name="tcp.service" id="tcp">
    <next id="tls"/>
    <tcp:Listen><tcp:Port>$arex_port</tcp:Port></tcp:Listen>
  </Component>
  ```

- a TLS MCC using the key and certificate and CA paths from the arc.conf, trusting all the VOMS servers, having a specific VOMSProcessing (relaxed, standard, strict, noerrors), having an IdentityMap SecHandler which uses the given gridmapfile to map the Grid users and maps to “nobody” in case of error, then having a LegacySecHandler which uses the arc.conf to match the client to groups and VOs configured there:

```
<Component name="tls.service" id="tls">
  <next id="http"/>
  <KeyPath>$X509_USER_KEY</KeyPath>
  <CertificatePath>$X509_USER_CERT</CertificatePath>
  <CACertificatesDir>$X509_CERT_DIR</CACertificatesDir>
  <VOMSCertTrustDNChain>
    <VOMSCertTrustRegex>.*</VOMSCertTrustRegex>
  </VOMSCertTrustDNChain>
  <VOMSProcessing>$VOMS_PROCESSING</VOMSProcessing>
  <!-- Do initial identity mappping by gridmap file -->
  <SecHandler name="identity.map" id="map" event="incoming">
    <PDP name="allow.pdp"><LocalList>$GRIDMAP</LocalList></PDP>
    <PDP name="allow.pdp"><LocalName>nobody</LocalName></PDP>
  </SecHandler>
  <!-- Match client to legacy authorization groups -->
  <SecHandler name="arclegacy.handler" event="incoming">
    <ConfigFile>$ARC_CONFIG</ConfigFile>
  </SecHandler>
</Component>
```

- one HTTP MCC, one SOAP MCCs, and the Plexer, with POST messages going through SOAP to the Plexer, GET/PUT/HEAD messages going directly to the Plexer, which checks if the path is the configured `arex_path`, if yes, it sends the message to the A-REX, otherwise fails:

```
<Component name="http.service" id="http">
  <next id="soap">POST</next>
  <next id="plexer">GET</next>
  <next id="plexer">PUT</next>
  <next id="plexer">HEAD</next>
</Component>
<Component name="soap.service" id="soap">
  <next id="plexer"/>
</Component>
<Plexer name="plexer.service" id="plexer">
  <next id="a-rex">^/$arex_path</next>
</Plexer>
```

- then the A-REX itself, with ArcAuthZ SecHandler containing a single LegacyPDP which will decide based on the `[gridftpd/jobs]` section of `arc.conf` if this message can go through or should be denied, then a LegacyMap SecHandler which uses the `[gridftpd]` section of arc.conf to figure out which local user should the Grid user be mapped to, then the full URL of the A-REX is given to the service (which in theory could be figured out from the incoming messages, but it is safer to be set explicitly), then the location of the `arc.conf` is given to the service (otherwise it wouldn't know), then some extra limits are set:

```
<Service name="a-rex" id="a-rex">
  <!-- Do authorization in same way as jobs plugin of gridftpd does -->
  <!-- Beware of hardcoded block name -->
  <SecHandler name="arc.authz" event="incoming">
    <PDP name="arclegacy.pdp">
      <ConfigBlock>
        <ConfigFile>$ARC_CONFIG</ConfigFile>
        <BlockName>gridftpd/jobs</BlockName>
      </ConfigBlock>
    </PDP>
  </SecHandler>
  <!-- Perform client mapping according to rules of gridftpd -->
  <SecHandler name="arclegacy.map" event="incoming">
```

```
        <ConfigBlock>
          <ConfigFile>$ARC_CONFIG</ConfigFile>
          <BlockName>gridftpd</BlockName>
        </ConfigBlock>
      </SecHandler>
      <arex:endpoint>$arex_mount_point</arex:endpoint>
      <arex:gmconfig>$ARC_CONFIG</arex:gmconfig>
      <arex:InfosysInterfaceMaxClients>
        $MAX_INFOSYS_REQUESTS
      </arex:InfosysInterfaceMaxClients>
      <arex:JobControlInterfaceMaxClients>
        $MAX_JOB_CONTROL_REQUESTS
      </arex:JobControlInterfaceMaxClients>
      <arex:DataTransferInterfaceMaxClients>
        $MAX_DATA_TRANSFER_REQUESTS
      </arex:DataTransferInterfaceMaxClients>
    </Service>
```

In summary, A-REX is usually started with the `a-rex` init script, which parses the `arc.conf` and creates an XML configuration, then starts the HED. This configuration uses the IdentityMap SecHandler to do an initial user mapping based on the configured grid-map file, if it fails, it maps to "nobody", then it uses the LegacySecHandler to match the user to Groups and VOs configured in arc.conf, then it uses the ArcAuthZ SecHandler with a LegacyPDP inside to allow or deny connections based on the authorization configured in the `[gridftpd/jobs]` section of the `arc.conf` (and the previously collected Group and VO information), then the LegacyMap SecHandler tries to map the Grid user to a local user based on the `[gridftpd]` section of `arc.conf` (and the previously collected Group and VO information).

6.10 Structure of the grid-mapfile

The following is not needed to setup a production environment but is described here as a reference.

A grid-mapfile is a simple text file. Each line is a record of the form

```
<grid identity certificate DN> <unix local account>
```

For each user that will connect to the CE, a Distinguished Name or DN contained in each user's certificate will be needed. Many Grid users can map to the same unix account.

A sample grid-mapfile is shown below:

```
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo1" griduser1
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo2" griduser1
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo3" griduser2
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo4" griduser2
"/DC=eu/DC=KnowARC/O=Lund University/CN=demo5" griduser2
```

Please refer to the certificate mini How-to to strip out the subject from Grid identity certificates.

6.11 Internal files of the A-REX

A-REX stores information about jobs in files in the *control directory*. Information is stored in files to make it easier to recover in case of failure, but for faster processing job state is also held in memory while A-REX is running. All files belonging to the same job have names starting with **job.ID.,** where ID is the job identifier.

The files in the control directory and their formats are described below:

- *job.ID.status* – current state of the job. This is a plain text file containing a single word representing the internal name of current state of the job. Possible values and corresponding external job states are:
  - ACCEPTED
  - PREPARING
  - SUBMIT
  - INLRMS
  - FINISHING
  - FINISHED
  - CANCELING
  - DELETED

  See Section 6.3 for a description of the various states. Additionally each value can be prepended the prefix "PENDING:" (like PENDING:ACCEPTED, see Section 6.3). This is used to show that a job is *ready* to be moved to the next state but it has to stay in it's current state *only* because otherwise some limits set in the configuration would be exceeded.

  This file is not stored directly in the *control directory* but in the following sub-directories:
  - accepting - for jobs in ACCEPTED state
  - finished - for jobs in FINISHED and DELETED states
  - processing - for other states
  - restarting - temporary location for jobs being restarted on user request or after restart of A-REX
- *job.ID.description* – contains the description of the job (JD).
- *job.ID.local* – information about the job used by the A-REX. It consists of lines of format *"name = value"*. Not all of them are always available. The following names are defined:
  - *globalid* – job identifier as seen by user tools. Depending on used interface it is either BES ActivityIdentifier XML tree, GUID of EMI ES or GridFTP URL.
  - *headnode* – URL of service interface used to submit this job.
  - *interface* – name of interface used for jobs submission - *org.nordugrid.xbes*, *org.ogf.glue.emies.activitycreation* or *org.nordugrid.gridftpjob*.
  - *lrms* – name of the LRMS backend to be used for local submission
  - *queue* – name of the queue to run the job at
  - *localid* – job id in LRMS (appears only after the job reached state **InLRMS**)
  - *args* – main executable name followed by a list of command-line arguments
  - *argscode* – code which main executable returns in case of success
  - *pre* – executable name followed by a list of command-line arguments for executable to run before main executable. There maybe few of them
  - *precode* – code which pre-executable returns in case of success
  - *post* – executable name followed by a list of command-line arguments for executable to run after main executable. There maybe few of them
  - *postcode* – code which post-executable returns in case of success
  - *subject* – user certificate's subject, also known as the distinguished name (DN)
  - *starttime* – GMT time when the job was accepted represented in the Generalized Time format of LDAP
  - *lifetime* – time period to preserve the SD after the job has finished in seconds
  - *notify* – email addresses and flags to send mail to about the job specified status changes
  - *processtime* – GMT time when to start processing the job in Generalized Time format

- *exectime* – GMT time when to start job execution in Generalized Time format
- *clientname* – name (as provided by the user interface) and IP address:port of the submitting client machine
- *clientsoftware* – version of software used to submit the job
- *rerun* – number of retries left to rerun the job
- *priority* – data staging priority (1 - 100)
- *downloads* – number of files to download into the SD before execution
- *uploads* – number of files to upload from the SD after execution
- *jobname* – name of the job as supplied by the user
- *projectname* – name of the project as supplied by the user. There may be few of them
- *jobreport* – URL of a user requested *logger service*. The A-REX will also send job records to this service in addition to the default logger service configured in the configuration. There may be few of them
- *cleanuptime* – GMT time when the job should be removed from the cluster and it's SD deleted in Generalized Time format
- *expiretime* – GMT time when the credentials delegated to the job expire in Generalized Time format
- *gmlog* – directory name which holds files containing information about the job when accessed through GridFTP interface
- *sessiondir* – the job's SD
- *failedstate* – state in which job failed (available only if it is possible to restart the job)
- *failedcause* – contains *internal* for jobs failed because of processing error and *client* if client requested job cancellation.
- *credentialserver* – URL of MyProxy server to use for renewing credentials.
- *freestagein* – *yes* if client is allowed to stage-in any file
- *activityid* – Job-id of previous job in case the job has been resubmitted or migrated. This value can appear multiple times if a job has been resubmitted or migrate more than once.
- *migrateactivityid* –
- *forcemigration* – This boolean is only used for migration of jobs. It determines whether the job should persist if the termination of the previous job fails.
- *transfershare* – name of share used in **Preparing** and **Finishing** states.

This file is filled partially during job submission and fully when the job moves from the **Accepted** to the **Preparing** state.

- *job.ID.input* – list of input files. Each line contains 3 values separated by a space. First value contains name of the file relative to the SD. Second value is a URL or a file description. Example:

  input.dat gsiftp://grid.domain.org/dir/input_12378.dat

  A URL represents a location from which a file can be downloaded. Each URL can contain additional options.

  A file description refers to a file uploaded from the UI and consists of [size][.checksum] where

  size - size of the file in bytes.

  checksum - checksum of the file identical to the one produced by ***cksum*** (1).

  These values are used to verify the transfer of the uploaded file. Both size and checksum can be left out. A special kind of file description *.* is used to specify files which are **not** required to exist.

  The third optional value is path to delegated credentials to be used for communication with remote server.

  This file is used by the data staging subsystem of the A-REX. Files with *URL* will be downloaded to the SD or cache and files with 'file description' will simply be checked to exist. Each time a new **valid** file appears in the SD it is removed from the list and *job.ID.input* is updated.

- *job.ID.input_status* – contains list of files uploaded by client to the SD.
- *job.ID.output* – list of output files. Each line contains 1, 2 or 3 values separated by a space. First value is the name of the file relative to the SD. The second value, if present, is a URL. Supported URLs are the same as those supported by job.ID.input. Optional 3rd value is path to delegated credentials to be used while accessing remote server.

  This file is used by the data staging subsystem of the A-REX. Files with *URL* will be uploaded to SE and remaining files will be left in the SD. Each time a file is uploaded it is removed from the list and *job.ID.output* is updated. Files not mentioned as output files are removed from the SD at the beginning of the **Finishing** state.
- *job.ID.output_status* – list of output files successfully pushed to remote locations.
- *job.ID.failed* – the existence of this file marks the failure of the job. It can also contain one or more lines of text describing the reason of failure. Failure includes the return code different from zero of the job itself.
- *job.ID.errors* – this file contains the output produced by external utilities like ***downloader***, ***uploader***, script for job submission to LRMS, etc on their stderr handle. Those are not necessarily errors, but can be just useful information about actions taken during the job processing. In case of problem include content of that file while asking for help.
- *job.ID.diag* – information about resources used during execution of job and other information suitable for diagnostics and statistics. It's format is similar to that of *job.ID.local*. The following names are at least defined:
  - *nodename* – name of computing node which was used to execute job,
  - *runtimeenvironments* – used runtime environments separated by ';',
  - *exitcode* – numerical exit code of job,
  - *frontend_distribution* – name and version of operating system distribution on frontend computer,
  - *frontend_system* – name of operating on frontend computer,
  - *frontend_subject* – subject (DN) of certificate representing frontend computer,
  - *frontend_ca* – subject (DN) of issuer of certificate representing frontend computer,

  and other information provided by GNU *time* utility. Note that some implementations of *time* insert unrequested information in their output. Hence some lines can have broken format.
- *job.ID.proxy* – delegated X509 credentials or chain of public certificates.
- *job.ID.proxy.tmp* – temporary X509 credentials with different UNIX ownership used by processes run with effective *user id* different from job owner's *id*.
- *job.ID.statistics* – statistics on input and output data transfer
- *delegations* – sub-directory containing collection of delegated credentials.
- *logs* – sub-directory with information prepared for reporting plugins.

There are other files with names like job.ID.* which are created and used by different parts of the A-REX. Their presence in the *control directory* can not be guaranteed and can change depending on changes in the A-REX code.

6.12 Environment variables set for the job submission scripts

The A-REX comes with support for several LRMS. Features explained below are for **PBS/Torque** backend, but for the other backends the behaviour is similar. This support is provided through *submit-pbs-job*, *cancel-pbs-job*, *scan-pbs-job* scripts. *submit-pbs-job* creates job's script and submits it to PBS. Created job's script is responsible for moving data between frontend machine and cluster node (if required) and execution of

actual job. Alternatively it can download input files and upload output if *"localtransfer=no"* is specified in the configuration file.

Behavior of submission script is mostly controlled using environment variables. Most of them can be specified on frontend in A-REX's environment and overwritten on cluster's node through PBS configuration. Some of them may be set in configuration file too.

PBS_BIN_PATH – path to PBS executables. Like */usr/local/bin* for example. Corresponds to *pbs_bin_path* configuration command.

PBS_LOG_PATH – path to PBS server logs. Corresponds to *pbs_log_path* configuration command.

TMP_DIR – path to directory to store temporary files. Default value is */tmp*. Corresponds to *tmpdir* configuration command.

RUNTIME_CONFIG_DIR – path where runtime setup scripts can be found. Corresponds to *runtimedir* configuration command.

GNU_TIME – path to GNU time utility. It is important to provide path to utility compatible with GNU time. If such utility is not available, modify *submit-pbs-job* to either reset this variable or change usage of available utility. Corresponds to *gnu_time* configuration command.

NODENAME – command to obtain name of cluster's node. Default is */bin/hostname -f*. Corresponds to *nodename* configuration command.

RUNTIME_LOCAL_SCRATCH_DIR – if defined should contain path to the directory on computing node, which can be used to store job's files during execution. *scratchdir* configuration command.

RUNTIME_FRONTEND_SEES_NODE – if defined should contain path corresponding to *RUNTIME_LOCAL_SCRATCH_DIR* as seen on **frontend** machine. Corresponds to *shared_scratch* configuration command.

RUNTIME_NODE_SEES_FRONTEND – if set to "*no*" means computing node does not share file system with frontend. In that case content of the SD is moved to computing node by using means provided by the LRMS. Results are moved back after job's execution in a same way. Corresponds to *shared_filesystem* configuration command.

For the last options, see Section **6.13**, *Using a scratch area*

6.13 Using a scratch area

Figures 6.5, 6.6 and 6.7 present some possible combinations for RUNTIME_LOCAL_SCRATCH_DIR and RUNTIME_FRONTEND_SEES_NODE and explain how data movement is performed. Figures a) correspond to the situation right after all input files are gathered in the session directory and actions taken right after the job script starts. Figures b) show how it looks while the job is running and actions which are taken right after it has finished. Figures c) show the final situation, when job files are ready to be uploaded to external storage elements or be downloaded by the user.

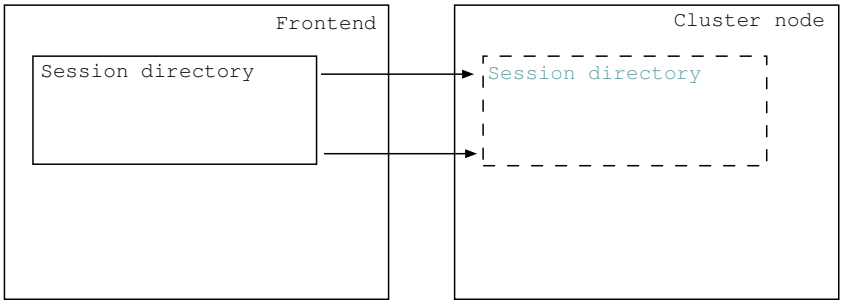


Figure 6.5: Both RUNTIME_LOCAL_SCRATCH_DIR and RUNTIME_FRONTEND_SEES_NODE undefined. Job is executed in a session directory placed on the frontend.

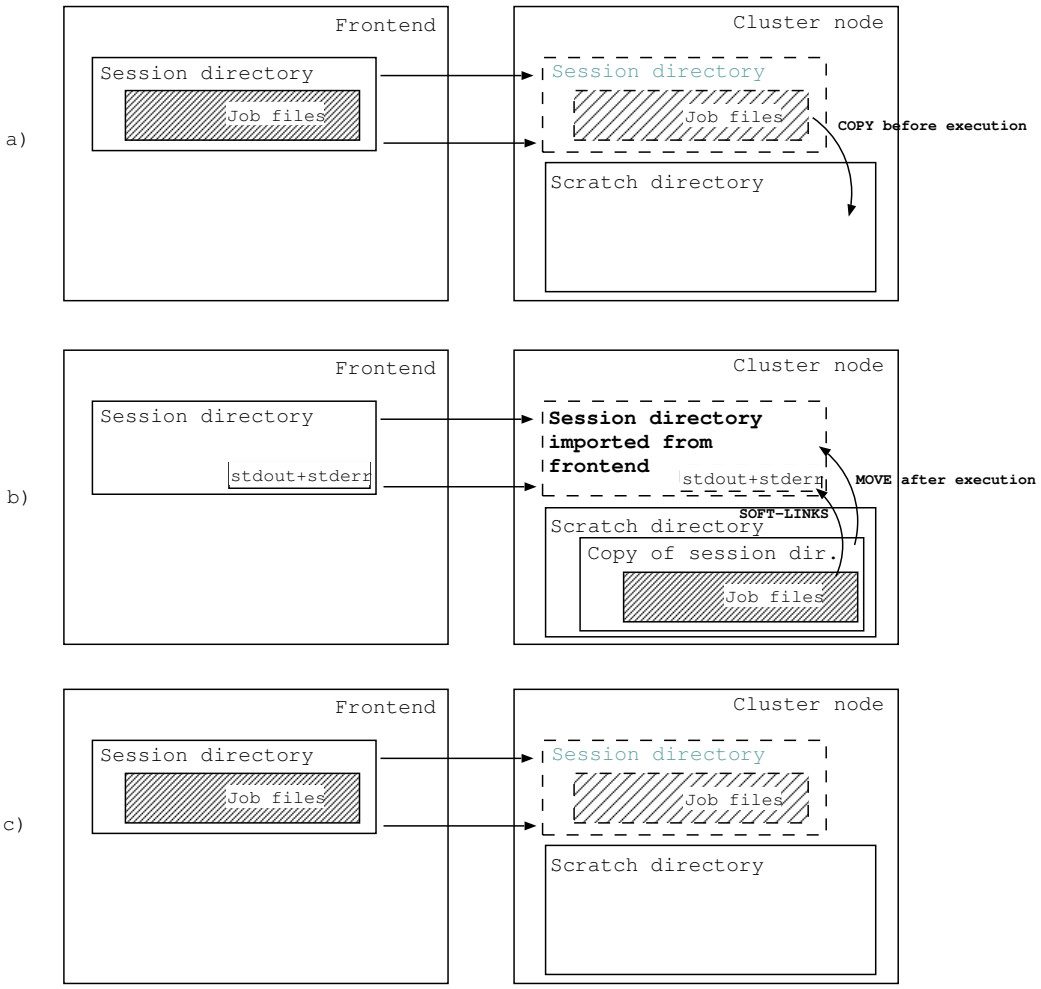


Figure 6.6: RUNTIME_LOCAL_SCRATCH_DIR is set to a value representing the scratch directory on the computing node, RUNTIME_FRONTEND_SEES_NODE is undefined.

a) After the job script starts all input files are moved to the "scratch directory" on the computing node.

b) The job runs in a separate directory in "scratch directory". Only files representing the job's *stdout* and *stderr* are placed in the original "session directory" and soft-linked in "scratch". After execution all files from "scratch" are moved back to the original "session directory".

c) All output files are in "session directory" and are ready to be uploaded/downloaded.

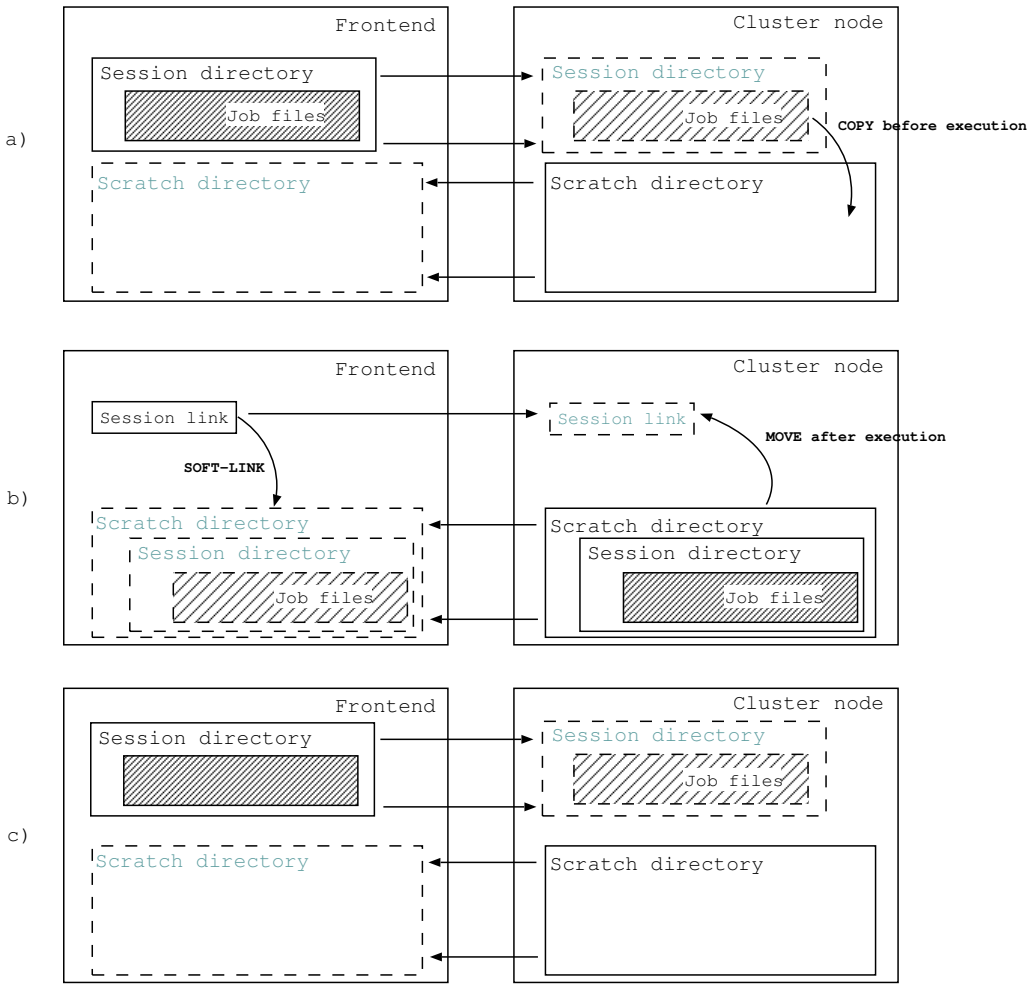


Figure 6.7: RUNTIME_LOCAL_SCRATCH_DIR and RUNTIME_FRONTEND_SEES_NODE are set to values representing the scratch directory on the computing node and a way to access that scratch directory from the frontend respectively.

a) After the job script starts, all input files are moved to "scratch directory" on the computing node. The original "session directory" is removed and replaced with a soft-link to a copy of the session directory in "scratch" as seen on the frontend.

b) The job runs in a separate directory in "scratch directory". All files are also available on the frontend through a soft-link. After execution, the soft-link is replaced with the directory and all files from "scratch" are moved back to the original "session directory"

c) All output files are in "session directory" and are ready to be uploaded/downloaded.

6.14 Web Service Interface

The A-REX Web Service Interface provides means to submit a description of a computational job to a computing resource, to stage-in additional data, to monitor and control processing of jobs, and obtain data produced during the execution of a job. The WS Interface is built and deployed inside the Hosting Environment Daemon (HED) infrastructure [**?**].

6.14.1 Basic Execution Service Interface

The job submission and control interface is based on a document produced by the OGF OGSA Basic Execution Services (BES) Working Group [**?**].

The exchange of SOAP messages is performed via HTTP(S). The BES interface is represented by two port-types – BES-Management and BES-Factory. The former is made to control the A-REX service itself and thus defines operations to start and stop the functionality of the BES service. The A-REX does not implement remote control of service functionality. Hence the BES-Management port-type is not functional. The BES-Factory port-type provides operations to submit new jobs (to create an activity in terms of BES) and to monitor its state. It also has an ability to provide information about the service. A-REX fully implements the functionality of this port-type.

For job descriptions A-REX accepts the Job Submission Description Language (JSDL) [**?**] documents as defined by the OGF Job Submission Description Language Working Group. Supported elements and extensions are described below.

6.14.2 Extensions to OGSA BES interface

A-REX introduces two new operations in addition to those provided by BES. It does that by defining its own port-type with new operations *ChangeActivityStatus* and *MigrateActivity*.

The *ChangeActivityStatus* operation provides a way to request simple transfers between states of jobs and corresponding actions.

- *ChangeActivityStatus*
  - Input
    - *ActivityStatusType OldStatus*: Description of the state the job is supposed to be in during execution of this request. If the current state of the job is different from the one having been given, the operation is aborted and a fault is returned. This parameter is optional.
    - *ActivityStatusType NewStatus*: Description of the state the job is to be put into.
  - Output
    - *ActivityStatusType NewStatus*: Description of the current state of the job.
  - Fault(s)
    - *NotAuthorizedFault*: Indicates that the client is not allowed to do this operation.
    - *InvalidActivityIdentifierFault*: There is no such job/activity.
    - *CantApplyOperationToCurrentStateFault*: The requested transition is not possible.

  On result of this command, the job should be put into the requested state. If such a procedure cannot be performed immediately then the corresponding sequence is initiated and fault OperationWillBeAppliedEventuallyFault will be returned.

  Since BES allows implementations to extend their initial activity states with additional sub-states, A-REX defines a set of sub-states of activity processing in addition to those defined by the BES, as listed in Table 6.1. Their meaning is described in Section 6.3.

  The *MigrateActivity* operation generates a request to migrate a grid job from another A-REX, i.e. the operation will get input files and possibly job description from the cluster currently holding the job and create the job as a new activity at the present cluster. Currently only migration of queuing jobs is supported.

- *MigrateActivity*
  - Input
    - *wsa:EndpointReferenceType ActivityIdentifier*: This element should contain the *wsa:EndpointReference* of the job to be migrated.
    - *ActivityDocument*: JSDL document of the job to be migrated. This element is optional.
    - *Boolean ForceMigration*: Boolean that determines whether the job will persist on the new cluster if the termination of the previous job fails.
  - Output
    - *wsa:EndpointReferenceType ActivityIdentifier*: This element should contain the *wsa:EndpointReference* of the new activity.
    - *ActivityDocument*: Contains the JSDL document of the new activity.
  - Fault(s)
    - *NotAuthorizedFault*: Indicates that the client is not allowed to do this operation.
    - *NotAcceptingNewActivitiesFault*: A fault that indicates that A-REX currently is not accepting new activities.
    - *UnsupportedFeatureFault*: This fault indicates that an sub-element in the JDSL document is not supported or the ActivityDocument has not been recognised as JSDL.
    - *InvalidRequestMessageFault*: This fault indicates that an element in the request is either missing or has an invalid format. Typically this would mean that the job-id cannot be located in the ActivityIdentifier of the old job.

The *ActiviterIdentifier* specifies the URL of the job which will be migrated. In case the *ActivityDocument* is filled this document will be used to create a new activity otherwise an attempt will be made to retrieve the job description through the BES operation *GetActivityDocument*.

Once the input files have been downloaded from the other cluster, a request will be send to terminate the old job. If this request fails the new activity at the present cluster will be terminate unless the *ForceMigration* is true. This is to prevent the job from being executed at two different places at the same time.

6.14.3 Delegation Interface

The A-REX also supports the Delegation Interface.This is a common purpose interface to be used by ARC services which accepts delegated credentials from clients. The Delegation Interface implements two operations: initialization of credentials delegation (DelegateCredentialsInit) and update/renewal of credentials (UpdateCredentials).

- *DelegateCredentialsInit* operation – this operation performs the first half of the credentials delegation sequence.
  - Input
    - None. On this request the service generates a pair of *public* and private keys. The public key is then sent to the client in response.
  - Output(s)
    - *TokenRequestType TokenRequest*: Contains the public key generated by the service as a Value element. It also provides an identifier in the Id element which should be used to refer to the corresponding private key.
  - Fault(s)
    - *UnsupportedFault*: Indicates that the service does not support this operation despite supporting the port-type.
    - *ProcessingFault*: Internal problems during generation of the token.
- *UpdateCredentials* operation – this operation makes it possible to update the content of delegated credentials (like in the case of credentials being renewed) unrelated to other operations of the service.

Table 6.1: Job states definitions and mappings

Applicable BES state	ARC BES sub-state	EMI ES state	ARIS state	A-REX internal state	Description
Pending	Accepting	ACCEPTED		ACCEPTED	Job is in the process of being submitted. This state is not recognised by the A-REX yet. *Accepted* is first reported state
	Accepted	ACCEPTED		ACCEPTED	Job was submitted
Running	Preparing	PREPROCESSING		PREPARING	Stage-in process is going on
	Prepared	PREPROCESSING		PREPARING + PENDING	Stage-in process has finished
	Submitting	PROCESSING-ACCEPTING		SUBMIT	Communication with local batch system is in process
	Queued	PROCESSING-RUNNING		INLRMS	Job entered local batch system but is not runnning now. This state is not recognised by the A-REX yet. *Executing* is reported instead
	Executing	PROCESSING-RUNNING		INLRMS	Job is being executed in local batch system
	Executed	PROCESSING-RUNNING		INLRMS, INLRMS + PENDING	Job execution in local batch system has finished. The A-REX dos not detect job states inside local batch system yet. As result this state is reported only if job is *Pending*.
	Killing	PROCESSING		CANCELING	Communication with local batch system to terminate execution is in process
	Finishing	POSTPROCESSING		FINISHING	Stage-out process is going on
Cancelled	Killed	TERMINAL		FINISHED	Job was stopped by explicit user request. The A-REX currently does not remember this request. *Failed* is reported instead.
Failed	Failed	TERMINAL		FINISHED	There was a failure during execution
Finished	Finished	TERMINAL		FINISHED	Job finished successfully
Finished	Deleted	TERMINAL		DELETED	Job finished and was left in A-REX too long
All	Pending			PENDING	Job is prevented from going to the next state due to some internal limits; this sub-state appears in parallel with other sub-states
All	Held				Job processing is suspended on client request; this sub-state appears in parallel with other sub-states. This state is reserved for future and is not implemented yet.

- Input
  - *DelegatedTokenType DelegatedToken*: Contains an X509 proxy certificate based on the public key from the DelegateCredentialsInit signed by the user's proxy certificate. Also includes the Id element which identifies the private key stored at the service side associated with these credentials. The reference element refers to the object to which these credentials should be applied in a way specific to the service. The same element must also be used for delegating credentials as part of other operations on service.
- Output(s)
  - None.
- Fault(s)
  - *UnsupportedFault*: Indicates that service does not support this operation despite supporting the port-type.
  - *ProcessingFault*: Internal problems during generation of the token.

Additionally, A-REX Web Service Interface allows delegation to be performed as part of the *CreateActivity* operation of the BES-Factory port-type. For this it accepts the element *DelegatedCredentials* inside the *CreateActivity* element. The *Id* element of *DelegatedCredentials* must contain an identifier obtained in response to the previous *DelegateCredentialsInit* operation. For more information about delegations and delegation interface refer to [**?**].

6.14.4 Local Information Description Interface

The A-REX implements the Local Information Description Interface (LIDI) interface common for all ARC services. This interface is based on OASIS Web Services Resource Properties specification [**?**]. Information about resources and maintained activities/jobs are represented in a *WS-Resource Properties* informational XML document. The document type is defined in the A-REX WSDL as a *ResourceInformationDocument-Type*. It contains the following elements/resources:

nordugrid – description of computing resource that uses NorudGrid LDAP schema [**?**] converted to XML document.

Domains – description of a computation resource that uses Glue2 schema.

All information can be accessed either through requests on particular resources or through XPath queries using WS-Resource Properties operations.

6.14.5 Supported JSDL elements

A-REX supports the following elements from the JSDL version 1.0 specification [**?**] including POSIX Applications extension and JSDL HPC Profile Application Extension [**?**]:

JobName – name of the job as assigned by the user.

Executable (POSIX,HPC) – name of the executable file.

Argument (POSIX,HPC) – arguments the executable will be launched with.

DataStaging

Filename – name of the data file on the executing node.

Source – source where the file will be taken from before execution.

Target – destination the file will be delivered to after execution.

Input (POSIX,HPC) – file to be used as standard input for the executable.

Output (POSIX,HPC) – file to be used as standard output for the executable.

Error (POSIX,HPC) – file to be used as standard error for the executable.

MemoryLimit (POSIX) – amount of physical memory needed for execution.

TotalPhysicalMemory – same as MemoryLimit.

IndividualPhysicalMemory – same as MemoryLimit.

CPUTimeLimit (POSIX) – maximal amount of CPU time needed for execution.

TotalCPUTime – same as CPUTimeLimit.

IndividualCPUTime – same as CPUTimeLimit.

WallTimeLimit (POSIX) – amount of clock time needed for execution.

TotalCPUCount – number of CPUs needed for execution.

IndividualCPUCount – same as *TotalCPUCount*.

6.14.6 ARC-specific JSDL Extensions

A-REX accepts JSDL documents having the following additional elements

IsExecutable – marks file to become executable after being delivered to the computing resource.

RunTimeEnvironment – specifies the name of the Runtime Environment needed for job execution.

Middleware – request for specific middleware on the computing resource frontend.

RemoteLogging – destination for the usage record report of the executed job.

LocalLogging – name for the virtual directory available through job interface and containing various debug information about job execution.

AccessControl – ACL expression which describes the identities of those clients who are allowed to perform operations on this job.

Notify – Email destination for notification of job state changes.

SessionLifeTime – duration for the directory containing job-related files to exist after the job finished executing.

JoinOutputs – specifies if standard output and standard error channels must be merged.

Reruns – defines how many times a job is allowed to rerun in case of failure.

CredentialServer – URL of MyProxy service which may be used for renewing the expired delegated job credentials.

CandidateTarget – specifies host name and queue of a computing resource.

OldJobID – specifies the previous job-ids in case the job has been resubmitted or migrated.

6.15 GridFTP Interface (jobplugin)

6.15.1 Virtual tree

The GFS with *jobplugin* plugin configured presents virtual file tree under its mount point through GridFTP protocol. A user connecting with a gridftp client will see virtual directories representing the jobs belonging to him/her. Directory names are job identifiers, each representing one job. These directories are directly connected to session directories of jobs and contain the files and subdirectories that are visible on the frontend. Client can access content of the session directory through these directories in virtual file tree.

If a job's xRSL description has *gmlog* attribute specified, then the job's directory also contains a virtual subdirectory of that name holding files with information about the job as created by the A-REX. Those are same files described in the section 6.11 with *job.ID.* part removed from their names. The 'proxy' file is notaccessible due to security reasons. The 'status' file is not accessible too. The 'errors' file is especially usefull for troubleshooting because it contains the stderr output of the various modules run by the A-REX during various job processing stages (downloader, uploader, job's submission to LRMS).

Also directly under the jobplugin's mount point there is another virtual directory named 'new' – used to accept new job descriptions – and another directory named 'info'. The latter has subdirectories named after job ids, each of which contains files with information about a job. These are the same files that can be accessed in the subdirectory specified through *gmlog* as described above.

6.15.2 Submission

Each file containing xRSL job description – name of file is not relevant – put into the 'new' directory by a gridftp client is treated as a new job's description. GFS's jobplugin parses the job description and returns to the client a positive response if there were no errors in the request.

The new job gets an identifier and a directory with the corresponding name appears under mount point. If the job's description contains input files which should be transferred from the client's machine, the client must upload them to that directory under specified names.

Please note that there is no predefined format or embedded information for assigned job identifiers. Those are opaques strings of characters suitable for being used for directory names in FTP protocol. One should not make any assumptions on how job identifier looks like also because it may change between different versions of jobplugin.

The job identifier reserved by GFS for a new job must be somehow communicated back to the client. Within the bounds of the FTP protocol, this is achieved in the following way. Prior to uploading the xRSL, the client issues a CWD command to change the current directory to 'new'. The jobplugin reserves new job identifier and responds with a redirect to the new session directory named after the reserved identifier. The client now knows the job's id, and proceeds with uploading the xRLS to the current directory or to the 'new'. If job description is accepted

6.15.3 Actions

Various actions to affect processing of an existing job are requested by uploading special xRSL files into directory 'new'. Such anxRSL must contain only 2 parameters - *action* for action to be performed, and *jobid* to identify the job to be affected. All other parameters are ignored.

The currently supported actions are:

- *cancel* to cancel a job
- *clean* to remove a job from computing resource
- *renew* to renew credentials delegated to a job
- *restart* to restart a job after failure at some phases

Alternatively, it is also possible to perform some of these actions by using the shortcut FTP operations described below.

6.15.3.1 Cancel

A job is canceled by performing a DELE (delete file) operation on the virtual directory representing the sessin directory of the job. It can take some time (a few minutes) before the job is actually canceled. Nevertheless, the client gets a response immediately.

6.15.3.2 Clean

A job's content is cleaned by performing a RMD (remove directory) operation on the virtaul directory representing the job. If the job is in FINISHED state it will be cleaned immediately. Otherwise it will be cleaned after it reaches state FINISHED.

6.15.3.3 Renew

If a client requests CWD to a job's session directory, credentials passed during authentication are compared to the currently assigned credentials of the job. If the validity time of the new credentials is longer, the job's current credentials are replaced with the new ones.

6.15.4 Configuration Examples

The examples presented below contain full configuration examples for the GridFTP server configured with the jobplugin. For clarity other sections such as those configuring A-REX and the information system are not shown.

6.15.4.1 Simple Example

In the following minimal example we use a single static mapfile which contains all possible user mappings for this site.

```
[common]
hostname="myhost.org"
lrms="fork"
gridmap="/etc/grid-security/grid-mapfile"

[gridftpd]
debug="3"
encryption="no"
allowunknown="no"
maxconnections="200"

[gridftpd/jobs]
path="/jobs"
plugin="jobplugin.so"
```

6.15.4.2 Detailed Example

Here we configure a simple PBS based cluster according to the following use case. John is member of the VO "smscg" where he belongs to the group "atlas" and has been assigned the roles "production" and "test". Since groups and roles are fully decoupled, John can request proxies that can include one (or several) of the following different group-role combinations (termed "Fully Qualified Names" (FQAN)):

- /smscg (notice it's the same as /smscg/Role=NULL)
- /smscg/Role=production
- /smscg/Role=test
- /smscg/atlas
- /smscg/atlas/Role=production
- /smscg/atlas/Role=test

A-REX serves as front-end to a batch-system that provides a "low_prio_queue" and a "high_prio_queue". Assignment to the different queues is done via local user identites. More precisely, the local users "smscg001, smscg002, smscg003" will be assigned to the low_prio_queue, whereas users "smscgP001, smscgP002, smscgP003" to the high_prio_queue (the configuration of the batch-system to support this is out of scope of this example).

Users sending jobs to A-REX should be assigned to one of the queues depending on the credentials they present in their proxy certificate. The assignment shall look as follows:

- /smscg , /smscg/Role=test , /smscg/Role=production =¿ shall map to one of the smscg00[1-3] local identities (thus low_prio_queue)
- /smscg/atlas , /smscg/atlas/Role=test , /smscg/atlas/Role=production =¿ shall map to one of the smscgP00[1-3] local identities (thus high_prio_queue)

The following usage pattern is considered. User John first wants to run a monitoring job on the high_prio_queue. He performs a voms-proxy-init and specifies his "/smscg/atlas/Role=test" FQAN to be used. When he submits his monitoring-job, John will be mapped to one of the smscgP001, smscgP002, smscgP003 accounts. John's job will thus run on the high_prio_queue.

After submitting the monitoring job, John submits regular jobs with his FQAN "/smscg". These jobs will run on the low_prio_queue. Later John switches back to the FQAN "/smscg/atlas/Role=test" to fetch the result of his monitoring job.

The discrimination to what queue John is to be mapped is done with VO information only and not on the basis of the DN of John's certificate. Hence the choice to what queue to be mapped is under control of John (we silently presumed John knows the mappings at the source).

Notes:

- a DN based grid-mapfile is generated on the front-end with a default mapping entry for John. The grid-mapfile is only used by the information system (GIIS) to make the grid resource look eligible for jobs submitted by John.
- the DN based grid-mapfile per se does not permit John to access the grid resource under different local identities (e.g. once as smscg001 and later as smscgP001), since the first matching DN defines the local identity John is to be mapped to. This is not a flaw since NorduGrid has support for lcmaps, which allows a 're-mapping' of a user.
- the mapping of the FQAN to the local user identity (e.g. "/smscg" to local user "smscg001") shall be done with lcmaps (in detail the lcmaps framework + lcmaps voms plugins). Direct VOMS based mapping is also possible.

If user John creates a proxy certificate with the "grid-proxy-init" command instead of "voms-proxy-init", hence the proxy certificate will not contain any VO information and submits a job to A-REX (the matchmaking will still work, since it's done with John's DN) he shall not be authorized.

Example configuration:

```
[common]
pbs_bin_path="/usr/bin"
pbs_log_path="/var/spool/pbs/server_logs"
hostname="myhost.org"
lrms="pbs"


[vo]
# We will use this configuration block for a few purposes.
# 1. To generate grid-mapfile needed for information system.
#    For that purpose nordugridmap utility will have to be
#    run periodically.
# 2. To provide coarse-grained information to authorization
```

```
#    rules used to define authorization groups. If needed of
#    course.
id="smscg_vo"
vo="smscg_vo"

# Here we define path to file to which nordugridmap will write DNs of
# users matching rules below. Because we are going to use it as
# grid-mapfile for other purposes it is going to reside at default
# location.
file="/etc/grid-security/grid-mapfile"

# Now we tell nordugridmap to pull information from
# VOMRS/VOMSS/or_whatever_it_is_called_now service and to ask for
# users belonging to smscg VO.
source="vomss://voms.smscg.org:8443/voms/smscg"

# Now we specify default mapping to local *NIX id. It is possible to
# completely redefine mapping in [gridftpd] block. But this one will
# be used by information system to compute and present resources
# available to user.  Let's use one of lowest priority account defined
# in use-case.
mapped_unixid="smscg001"


[group]
# In this authorization group we are going to check if user presents
# any proof that he belongs to 'smscg' VO. We can use that information
# later to explicitly limit access to resources. If such access
# control is not needed this group can be removed.
name="smscg_auth"

# Here we can use internal support of ARC for VOMS attributes
# voms="smscg * * *"
# If we want to limit access to resources also by other VOMS
# attributes then other voms rules similar to those defined
# below in [gridftpd] section may be used.

# Or we can ask some external executable to analyze delegated
# credentials of user. In this example executable vomatch
# is called with first argument containing path to delegated
# proxy certificate and second - required VO name.
# plugin="10 /opt/external/bin/vomatch %P smscg"

# Or - probably preferred way in this use case - we can use
# LCAS to analyze delegated proxy.
# First element after '=' sign is path to LCAS library whatever
# it is called in current implementation. Second is LCAS installation
# path - it will be used to set environment variable LCAS_DIR.
# And third element is path to LCAS database file - it will be passed
# to environment variable LCAS_DB_FILE.
# Function 'lcas_get_fabric_authorization' of specified LCAS library
# will be called with following 3 arguments
#  1. char* pointer to string containing DN of user
#  2. gss_cred_id_t variable pointing at delegated credentials of user
#  3. char* pointer to empty string
# Returned 0 int value is treated as positive response
lcas="/opt/glite/lib/liblcas.so /opt/glite /opt/glite/share/lcas.db"
```

```
# As coarse grained solution it is also possible to check if user
# belongs to one of defined VOs as specified in _previously_ defined
# [vo] group. Here we refer to VO group smscg_vo defined above.
#vo="smscg_vo"


[gridftpd]
debug="2"
logfile="/var/log/arc/gridftpd.log"
logsize="10000000 2"
pidfile="/var/run/gridftpd.pid"
port="2811"
pluginpath="/usr/local/lib/arc"
encryption="no"

# By specifying 'no' here we limit users allowed to establish
# connection to this server to those specified in grid-mapfile. This
# may be not necessary if additional authorization is applied as done
# below. But this provides additional layer of protection so let it
# be.
allowunknown="no"

maxconnections="200"

# Here we start fine-grained user mapping. Let's first define few VOMS
# mappings using embedded functionality of ARC.  These lines should
# map Grid users to high-priority and low-priority *NIX users smscg001
# and smscgP001. Mind order - those with more attributes defined come
# first.  I do not know if missing attribute is passed by VOMS as
# empty string or as string containing NULL keyword. Here I assume
# empty string. If it is NULL then "" has to be replaced with NULL.
#unixmap="smscgP001 voms smscg atlas test *
#unixmap="smscgP001 voms smscg atlas production *
#unixmap="smscgP001 voms smscg atlas "" *
# These 3 lines are not needed if grid-mapfile defines default mapping
# to smscg001 user. But we can have them for consistence and if mapping
# to nobody is defined below for safety reasons.
#unixmap="smscg001 voms smscg "" test *
#unixmap="smscg001 voms smscg "" production *
#unixmap="smscg001 voms smscg "" "" *

# Instead of using multiple unixmap commands above we may define
# 2 authorization groups using [group] blocks. Let's say their
# names are smscg_low and smscg_high. Then 'group' matching rule
# may be used.
#unixmap="smscgP001 group smscg_high"
#unixmap="smscg001 group smscg_low"

# Or if we want to use all 6 local accounts and let mapping choose
# randomly within 2 group accounts 'simplepool' may be used. In
# example below 'unixgroup' ensures proper choice of group and
# 'simplepool' makes a choise from accounts in pool. Last argument
# specifies directory containing file named 'pool'. That file contains
# list of local user accounts. Also this directory will be used for
# writing information about current mappings.
#unixgroup="smscg_high simplepool /var/nordugrid/smscg_high"
#unixgroup="smscg_low simplepool /var/nordugrid/smscg_low"
```

```
# And mapping preferred in this use case - through LCMAPS.  First
# element after '=' sign is path to LCMAPS library whatever it is
# called in current implementation. Second is LCMAPS installation path
# - it will be used to set environment variable LCMAPS_DIR.  And third
# element is path to LCMAPS database file - it will be passed to
# environment variable LCMAPS_DB_FILE. Those 3 arguments are followed
# list of policy names.
# Function 'lcmaps_run_and_return_username' of specified LCMAPS library
# will be called with following arguments
#  1. char* pointer to string containing DN of user
#  2. gss_cred_id_t variable pointing at delegated credentials of user
#  3. char* pointer to empty string
#  4. char** pointer for chosen username.
#  5. int variable containing number of policies
#  6. char** list of policy names
# Expected 0 int value returned and argument 4 set. Value returned in
# 4th argument is used as username of local account.
unixmap="* lcmaps /opt/glite/lib/liblcmaps.so /opt/glite \
  /opt/glite/share/lcmaps.db policy1 policy2"

# Here we can specify mapping to some harmless local user account for
# safety reasons. If that account is not allowed to submit jobs to
# LRMS then this will also work as authorization effectively cutting
# off users without proper VOMS attributes.
unixmap="nobody:nobody all"


[gridftpd/jobs]
# This block defines job submission service
path="/jobs"
plugin="jobplugin.so"

# Line below specifies that this plugin/service is only available to
# users belonging to authorization group. If such behavior is not
# required then this line must be commented.
groupcfg="smscg_auth"


[queue/low_prio_queue]
name="low_prio_queue"
homogeneity="True"
scheduling_policy="FIFO"
comment="This queue is low priority"
nodecpu="adotf"
nodememory="512"
architecture="adotf"
opsys="Mandrake 8.0"
opsys="Linux-2.4.19"
benchmark="SPECINT2000 222"
benchmark="SPECFP2000 333"
cachetime="30"
timelimit="30"
sizelimit="5000"


[queue/high_prio_queue]
name="high_prio_queue"
homogeneity="True"
```

```
scheduling_policy="FIFO"
comment="This queue is high priority"
nodecpu="adotf"
nodememory="512"
architecture="adotf"
opsys="Mandrake 8.0"
opsys="Linux-2.4.19"
benchmark="SPECINT2000 222"
benchmark="SPECFP2000 333"
```

Acknowledgements

This work was supported in parts by: the Nordunet 2 program, the Nordic DataGrid Facility, the EU KnowARC project (Contract nr. 032691), the EU EMI project (Grant agreement nr. 261611) and the Swedish Research council via the eSSENCE strategic research program.

Bibliography