Enabling Grids for E-sciencE

www.eu-egee.org

EGEE Tutorial
Rome, 03 February 2005

Grid Workload Management

Valeria Ardizzone
INFN Catania

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

Contents

- The Workload Management System
- Job Preparation
  - Job Description Language
- Job submission and job status monitoring
- WMS Matchmaking
- Different job types
  - Interactive jobs
  - Checkpointable jobs
  - MPI jobs
  - DAG jobs

EGEE/LCG Workload Management System

- The user interacts with Grid via a **Workload Management System (WMS)**
- The Goal of WMS is the **distributed scheduling and resource management in a Grid environment**.
- What does it allow Grid users to do?
  - To submit their jobs
  - To execute them on the “best resources”
    - The WMS tries to optimize the usage of resources
  - To get information about their status
  - To retrieve their output

Job Preparation

- Information to be specified when a job has to be submitted:
  - Job characteristics
  - Job requirements and preferences on the computing resources
    - Also including software dependencies
  - Job data requirements
- Information specified using a Job Description Language (JDL)
  - Based upon Condor's *CLASSified ADvertisement language (ClassAd)*
    - Fully extensible language
    - A ClassAd
      - Constructed with the classad construction operator []
      - It is a sequence of attributes separated by semi-colon (;).
- So, the JDL allows definition of a set of attribute, the WMS takes into account when making its scheduling decision

Job Preparation

- An attribute is a pair (key, value), where value can be a Boolean, an Integer, a list of strings, ...
  - <attribute> = <value>;
- In case of literal string for values:
  - if a string itself contains double quotes, they must be escaped with a backslash
    - `Arguments = " \"Hello\" 10";`
  - the character "'" cannot be specified in the JDL
  - special characters such as &, |, >, < are only allowed
    - if specified inside a quoted string
    - if preceded by triple \
      - `Arguments = "-f file1\\\&file2";`
- Comments must be preceded by a sharp character (#) or have to follow the C++ syntax
- The JDL is sensitive to blank characters and tabs
  - they should not follow the semicolon (;) at the end of a line

Job Description Language

- The supported attributes are grouped in two categories:
  - **Job Attributes**
    - Define the job itself
  - **Resources**
    - Taken into account by the RB for carrying out the matchmaking algorithm (to choose the “best” resource where to submit the job)
    - *Computing Resource*
      - Used to build expressions of Requirements and/or Rank attributes by the user
      - Have to be prefixed with “other.”
    - *Data and Storage resources (see talk Job Services With Data Requirements)*
      - Input data to process, SE where to store output data, protocols spoken by application when accessing SEs

JDL: Relevant attributes

- **JobType**
  - *Normal* (simple, sequential job), *Interactive*, *MPICH*, *Checkpointable*
  - Or combination of them
- **Executable** (mandatory)
  - The command name
- **Arguments** (optional)
  - Job command line arguments
- **StdInput**, **StdOutput**, **StdError** (optional)
  - Standard input/output/error of the job
- **Environment (optional)**
  - List of environment settings
- **InputSandbox** (optional)
  - List of files on the UI local disk needed by the job for running
  - The listed files will automatically staged to the remote resource
- **OutputSandbox** (optional)
  - List of files, generated by the job, which have to be retrieved
- **VirtualOrganisation** (optional)
  - A different way to specify the VO of the user

JDL: Relevant attributes

- **Requirements**
  - Job requirements on the resources
  - Specified using GLUE attributes of resources published in the Information Service
  - Its value is a boolean expression
  - Only one requirements can be specified
    - if there are more than one, only the last one is taken into account
  - If not specified, default value defined in UI configuration file is considered
    - Default: *other.GlueCEStateStatus == "Production"* (the resource has to be able to accept jobs and dispatch them on WNs)

JDL: Relevant attributes

- **Requirements**
  - Other possible requirements values are below reported:
    - *other.GlueCEInfoLRMSType == "PBS" && other.GlueCEInfoTotalCPUs > 1* (the resource has to use PBS as the LRMS and whose WNs have at least two CPUs)
    - *Member("CMSIM-133", other.GlueHostApplicationSoftwareRunTimeEnvironment)* (a particular experiment software has to run on the resource and this information is published on the resource environment)
      - The *Member* operator tests if its first argument is a member of its second argument
    - *RegExp("cern.ch", other.GlueCEUniqueId)* (the job has to run on the CEs in the domain cern.ch)
    - *(other.GlueHostNetworkAdapterOutboundIP == true) && Member("VO-alice-Alien", other.GlueHostApplicationSoftwareRunTimeEnvironment) && Member("VO-alice-Alien-v4-01-Rev-01", other.GlueHostApplicationSoftwareRunTimeEnvironment) && (other.GlueCEPolicyMaxWallClockTime > 86000)* (the resource must have some packages installed VO-alice-Alien and VO-alice-Alien-v4-01-Rev-01 and the job has to run for more than 86000 seconds)

JDL: Relevant attributes

- **Rank**
  - Expresses preference (how to rank resources that have already met the Requirements expression)
  - It is expressed as a floating-point number
  - The CE with the highest rank is the one selected
  - Specified using GLUE attributes of resources published in the Information Service
  - If not specified, default value defined in the UI configuration file is considered
    - Default: - *other.GlueCEStateEstimatedResponseTime* (the lowest estimated traversal time)
    - Default: *other.GlueCEStateFreeCPUs* (the highest number of free CPUs)
  - Other possible rank value is below reported:
    - *(other.GlueCEStateWaitingJobs == 0 ? other.GlueCEStateFreeCPUs : -other.GlueCEStateWaitingJobs)* (the number of waiting jobs is used if this number is not null and the rank decreases as the number of waiting jobs gets higher; if there are not waiting jobs, the number of free CPUs is used)

Essential JDL

- **At least one has to specify the following attributes:**
  - the name of the executable
  - the files where to write the standard output and standard error of the job
  - the arguments to the executable, if needed
  - the files that must be transferred from UI to WN and viceversa

```
[
Executable = "ls -al";
StdError = "stderr.log";
StdOutput = "stdout.log";
OutputSandbox = {"stderr.log", "stdout.log"};
]
```

Example of JDL file

```
[
JobType = “Normal”;
Executable = "$(CMS)/exe/sum.exe";
InputSandbox = {"/home/user/WP1testC","/home/file*”,
"/home/user/DATA/*"};
OutputSandbox = {“sim.err”, “test.out”, “sim.log"};
Requirements  = (other. GlueHostOperatingSystemName
== “linux") && (other.GlueCEPolicyMaxWallClockTime >
10000);
Rank = other.GlueCEStateFreeCPUs;
]
```

Job Submission

edg-job-submit [-r *<res_id>*] [-c *<config file>*] [-vo *<VO>*] [-o *<output file>*] *<job.jdl>*

-r the job is submitted directly to the computing element identified by *<res_id>*

-c the configuration file *<config file>* is pointed by the UI instead of the standard configuration file

-vo the Virtual Organisation (if user is not happy with the one specified in the UI configuration file)

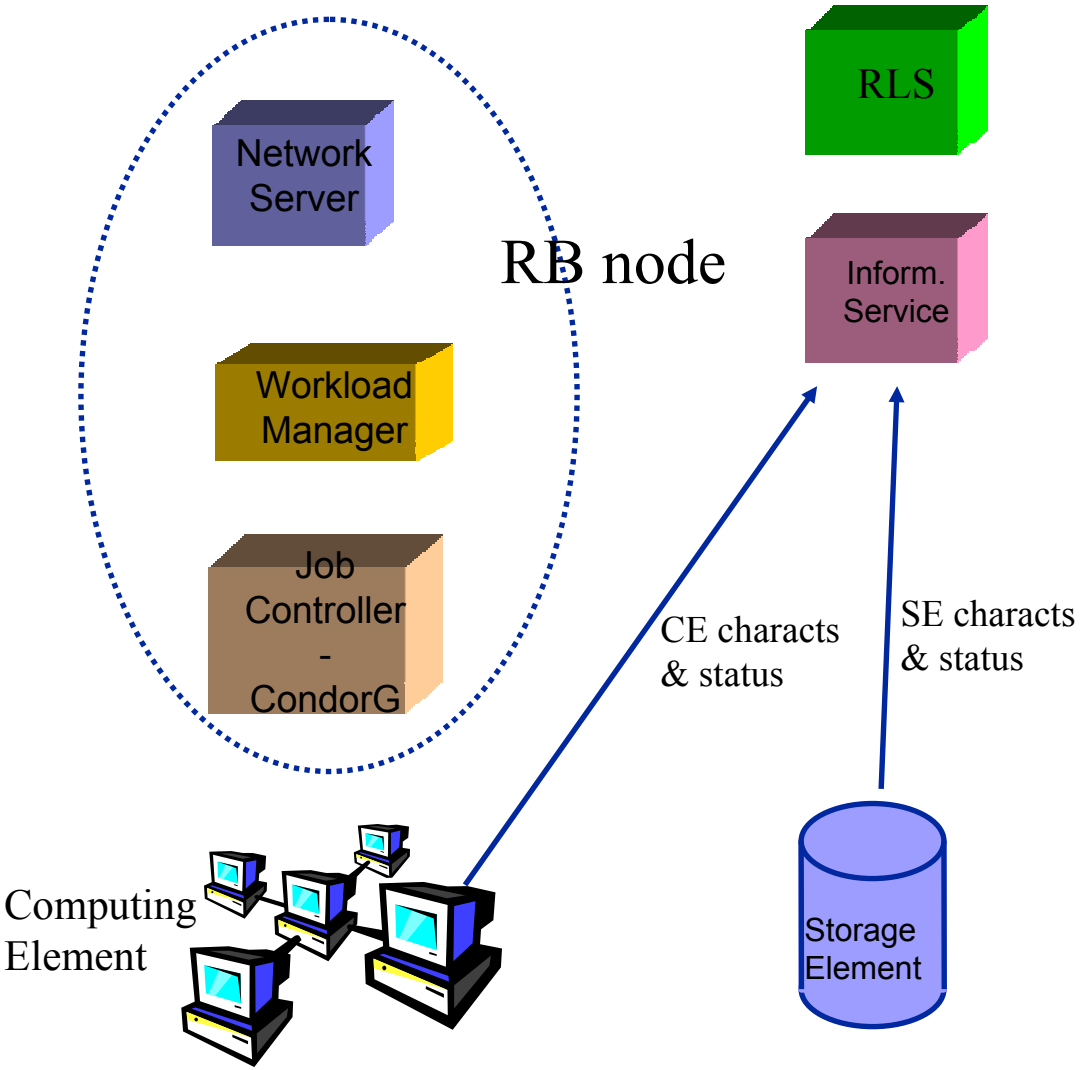
-o the generated edg_jobId is written in the *<output file>*

Useful for other commands, e.g.:

edg-job-status -i *<input file>* (or edg_jobId)

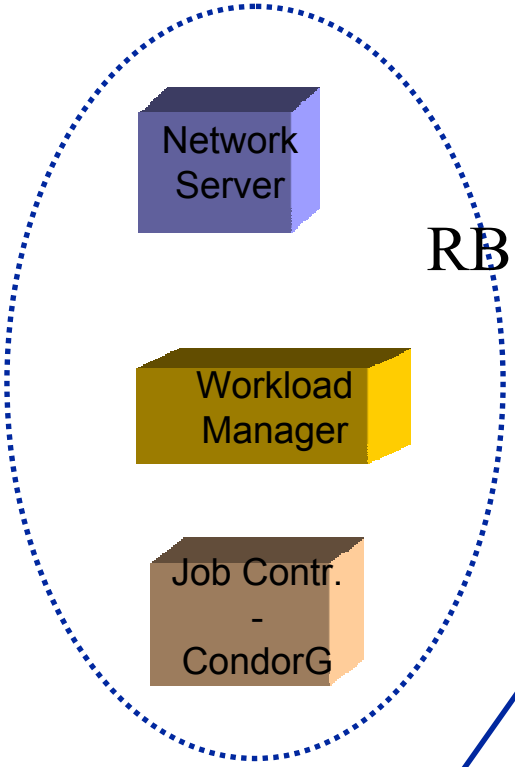
-i the status information about edg_jobId contained in the *<input file>* are displayed

Job Submission



Job Submission

UI: allows users to access the functionalities of the WMS
(via command line, GUI, C++ and Java APIs)



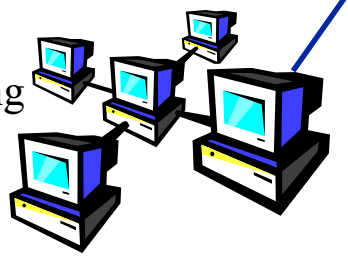
RB node

Job Status

submitted

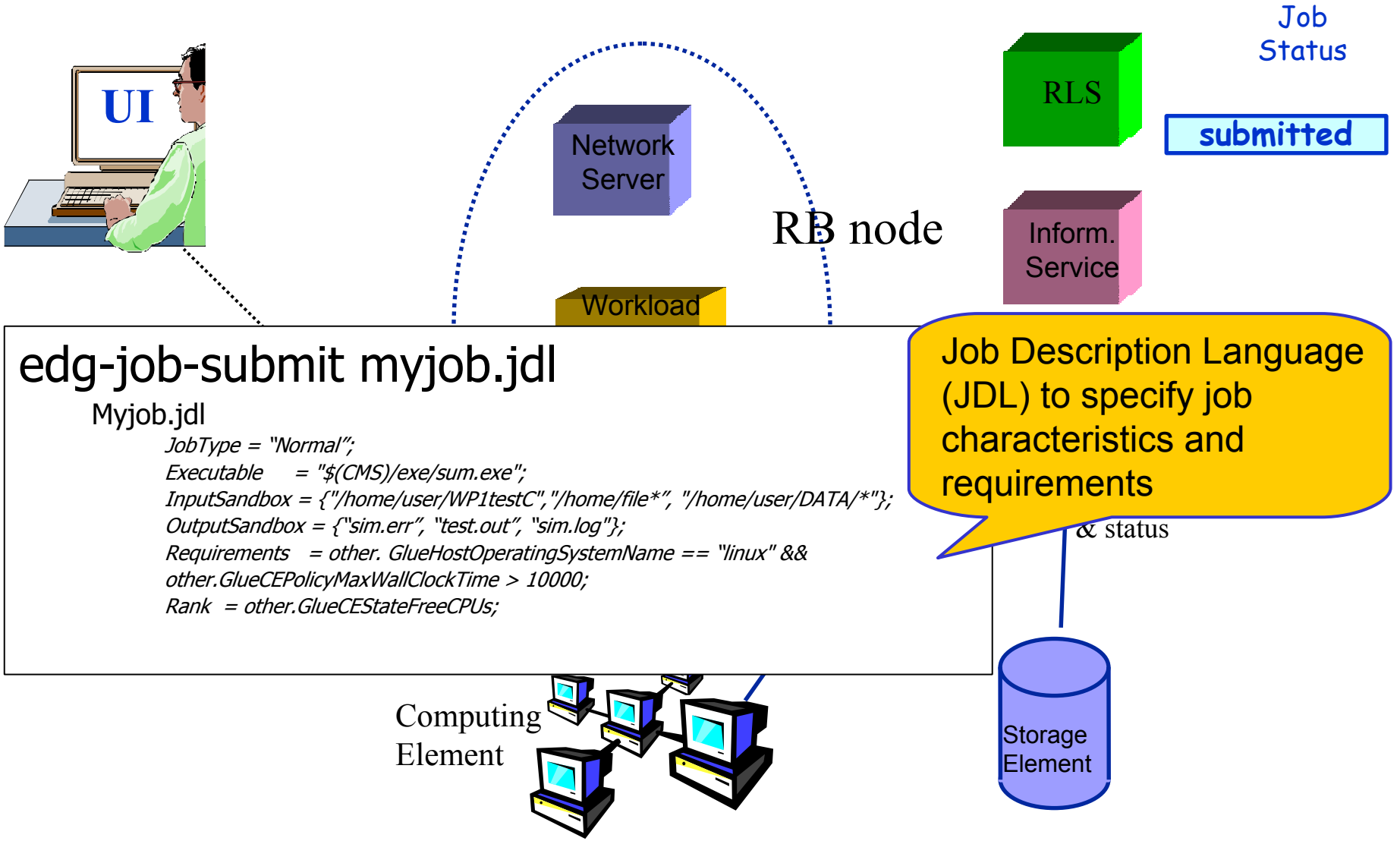
CE characts & status

SE characts & status

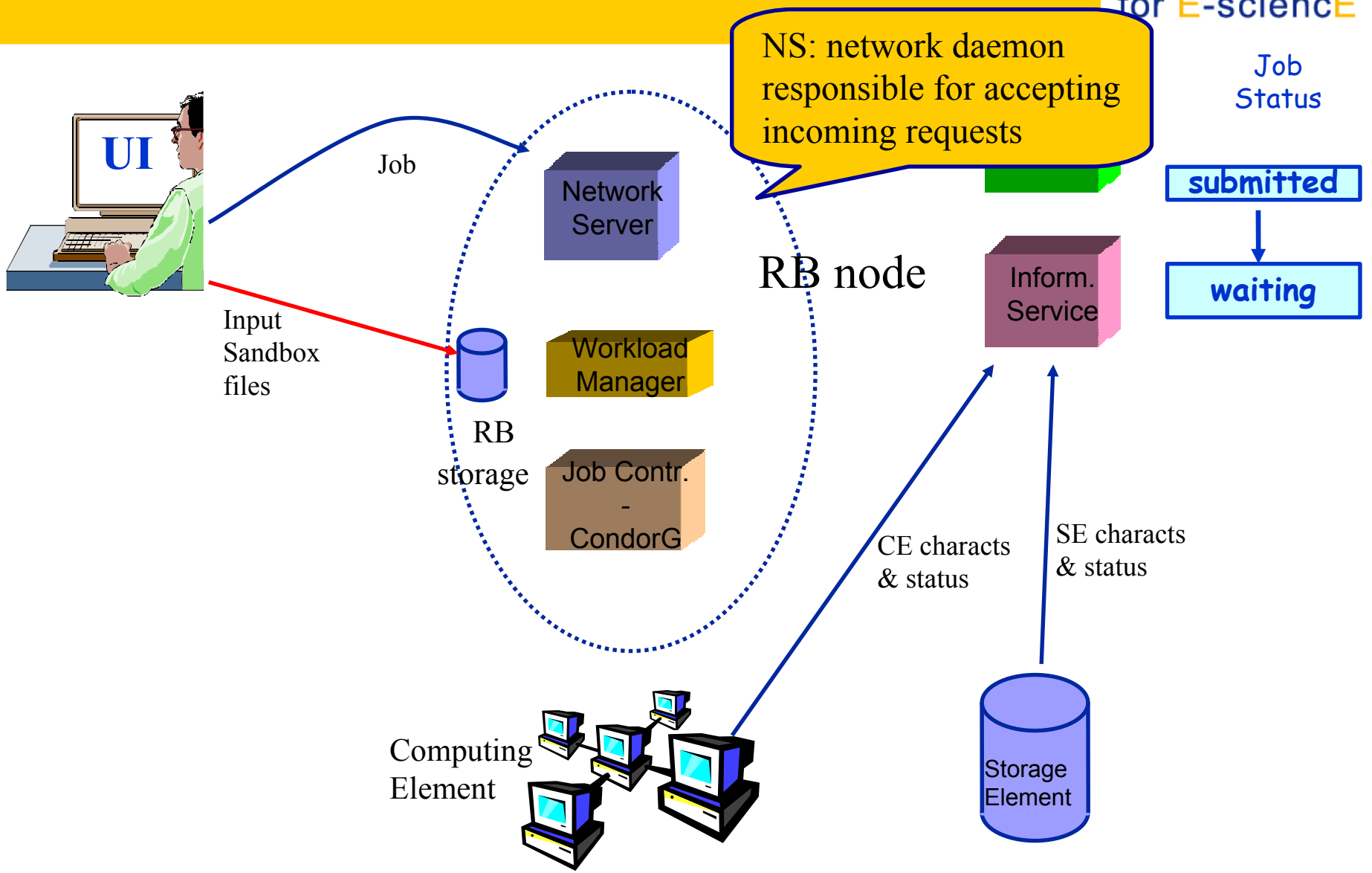


Computing Element

Job Submission



Job Submission



Job Submission

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

RB node

RLS

Inform. Service

Job Status

submitted

waiting

WM: responsible to take the appropriate actions to satisfy the request

CE characts & status

SE characts & status

Computing Element

Storage Element

Job Submission

RB node

Network Server

Match-Maker/ Broker

Workload Manager

RB storage

Job Contr. - CondorG

Where must this job be executed ?

RLS

Inform. Service

Job Status

submitted

waiting

CE characts & status

SE characts & status

Computing Element

Storage Element

Job Submission

RB node

Network Server

Match-Maker/ Broker

Workload Man

RB storage

Job Con - CondorG

RLS

Inform. Service

Job Status

submitted

waiting

Matchmaker: responsible to find the "best" CE where to submit a job

CE characts & status

SE characts & status

Computing Element

Storage Element

Job Submission

RB node

Network Server

Match-Maker/ Broker

Workload Manager

RB storage

Job Contr. - CondorG

RLS

Inform. Service

Job Status

submitted

waiting

What is the status of the Grid ?

CE characts & status

SE characts & status

Where are (which SEs) the needed data ?

Storage Element

Computing Element

Job Submission

RB node

Network Server

Match-Maker/ Broker

Workload Manager

CE choice

RB storage

Job Contr. - CondorG

RLS

Inform. Service

Job Status

submitted

waiting

CE characts & status

SE characts & status

Computing Element

Storage Element

Job Submission

RB node

Network Server

Workload Manager

Job Adapter

RB storage

Job Contr. -

RLS

Inform. Service

Job Status

submitted

waiting

CE characts

SE characts & status

JA: responsible for the final "touches" to the job before performing submission (e.g. creation of wrapper script, etc.)

Computing Element

Storage Element

Job Submission

RB node

Network Server

Workload Manager

RB storage

Job Controller - CondorG

JC: responsible for the actual job management operations (done via CondorG)

Computing Element

CE characts & status

SE characts & status

RLS

Inform. Service

Storage Element

Job Status

submitted

waiting

ready

Job Submission

RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Input Sandbox files

Job

Computing Element

RLS

Inform. Service

CE characts & status

SE characts & status

Storage Element

Job Status

submitted

waiting

ready

scheduled

Job Submission

RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

RLS

Inform. Service

Job Status

submitted

waiting

ready

scheduled

running

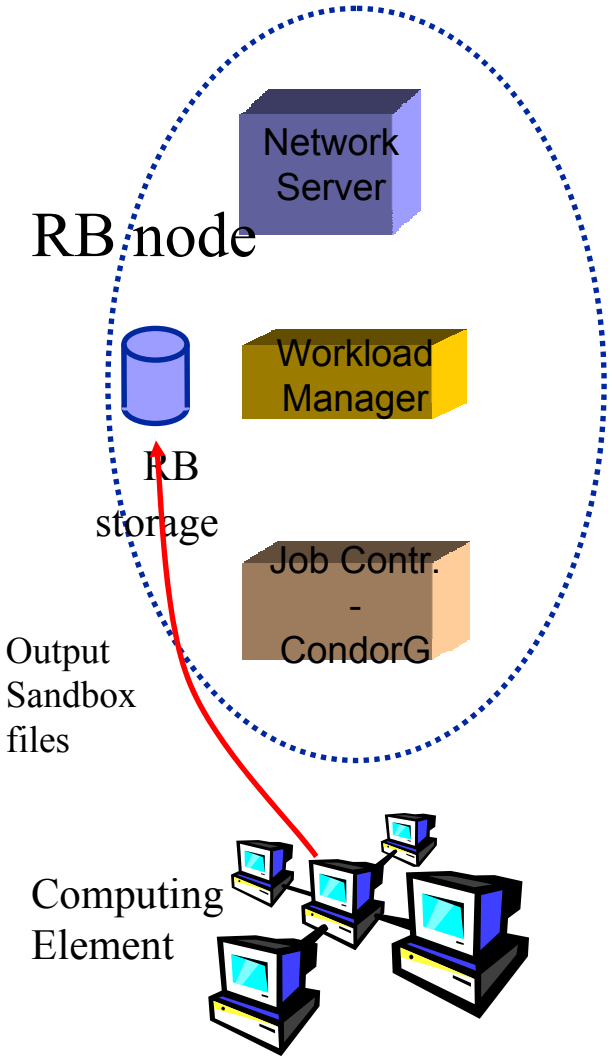
Computing Element

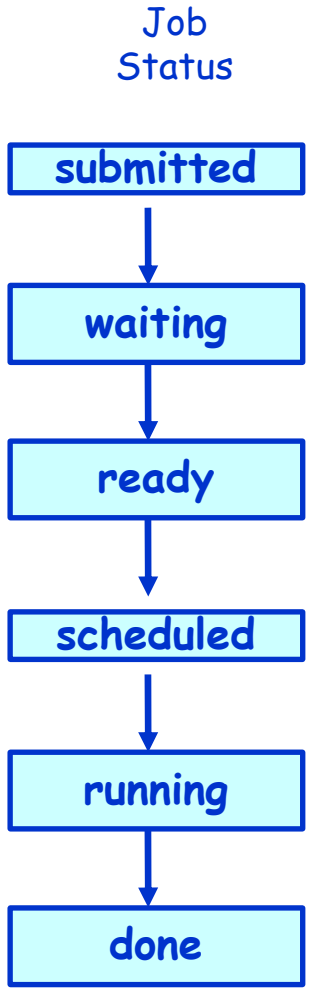
Job

"Grid enabled" data transfers/ accesses

Storage Element

Job Submission





Job Submission

edg-job-get-output <dg-job-id>

RB node

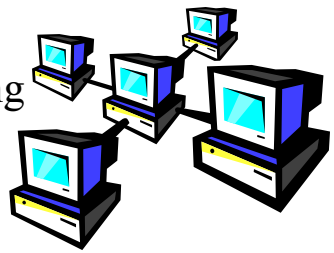
Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Computing Element



RLS

Inform. Service

Storage Element

Job Status

submitted

waiting

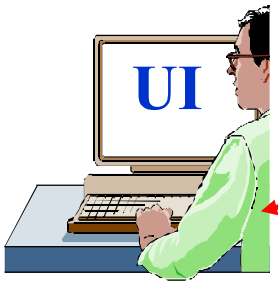
ready

scheduled

running

done

Job Submission



Output Sandbox files

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

RB node

RLS

Inform. Service

Computing Element

Storage Element

Job Status

submitted

waiting

ready

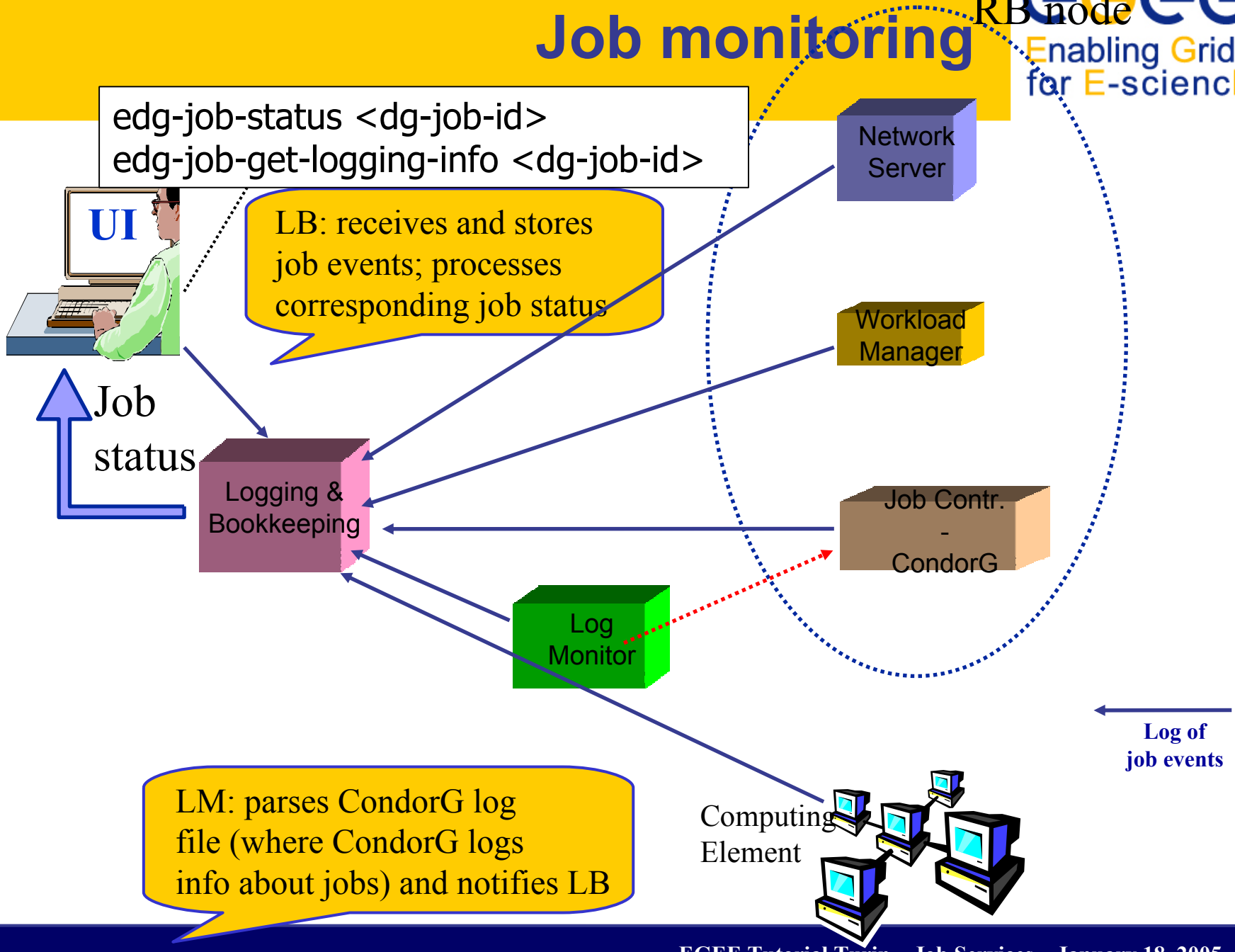
scheduled

running

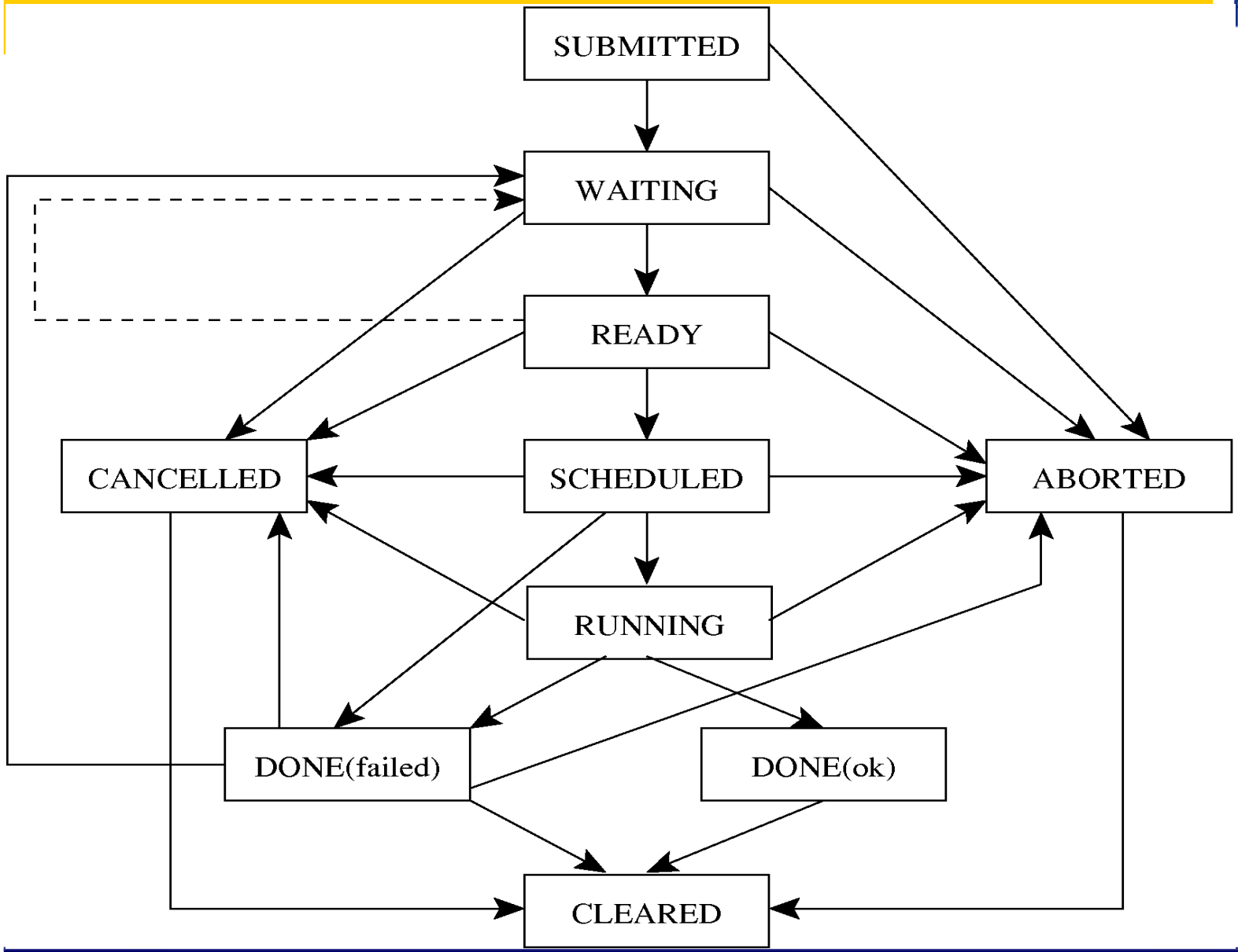
done

cleared

Job monitoring



Possible job states



Job resubmission

- If something goes wrong, the WMS tries to reschedule and resubmit the job (possibly on a different resource satisfying all the requirements)
- Maximum number of resubmissions: min(RetryCount, MaxRetryCount)
  - RetryCount: JDL attribute
  - MaxRetryCount: attribute in the “RB” configuration file
- e.g., to disable job resubmission for a particular job: *RetryCount=0;* in the JDL file

Other (most relevant) UI commands

- **edg-job-list-match**
  - Lists resources matching a job description
  - Performs the matchmaking without submitting the job
- **edg-job-cancel**
  - Cancels a given job
- **edg-job-status**
  - Displays the status of the job
- **edg-job-get-output**
  - Returns the job-output (the OutputSandbox files) to the user
- **edg-job-get-logging-info**
  - Displays logging information about submitted jobs (all the events “pushed” by the various components of the WMS)
  - Very useful for debug purposes

The Matchmaking algorithm

- The matchmaker has the goal to find the best suitable CE where to execute the job
- To accomplish this task, the WMS interacts with the other EGEE/LCG components (Replica location Service, and Information Service)
- There are three different scenarios to be dealt with separately:
  - Direct job submission
  - Job submission without data-access requirements
  - Job submission with data-access requirements (*see talk Job Services With Data Requirements*)

The Matchmaking algorithm: direct job submission

- The user JDL contains a link to the resource to submit the job
- The WMS does not perform any matchmaking algorithm at all
- The job is simply submitted to the specified CE

IMPORTANT:

- If the CEId is specified then the WMS
  - neither checks whether the user who submitted the job is authorised to access the given CE, nor interacts with the RLS for the resolution of files requirements, if any
  - Only checks the JDL syntax, while converting the JDL into a ClassAd
- The user run the edg-job-submit –resource <ce_id> <nome.jdl> command

  ce_id = hostname:port/jobmanager-lsf-grid01

The Matchmaking algorithm: job submission without data access requirements

- The user JDL contains some requirements
- Once the JDL has been received by the WMS and converted in ClassAd, the WMS invokes the matchmaker
- The matchmaker has to find if the characteristics and status of Grid resources match the job requirements
- There are two phases:
  - Requirements check:
    - The Matchmaker contacts the GOUT/II in order to create a set of suitable CEs compliant with user requirements and where the user is authorized to submit jobs
    - The Matchmaker creates the set of suitable CEs
  - Ranking phase:
    - The Matchmaker contacts directly the LDAP (GRIS) server of the involved CEs to obtain the values of those attributes that are in the rank JDL expression

The Matchmaking algorithm: job submission without data access requirements

- The matchmaker can select a CE randomly, if there are two or more CEs that meet all the requirements and have the same rank
- In general, the CE with maximum rank value is selected

- IMPORTANT:
  - The CE attributes involved in the JDL requirements refers to static information
  - All the information cached in the IS represent a good source for matches among job requirements and CE features
  - In the first phase it is more efficient to contact the GOUT/II, than querying each CE
  - The rank attributes refers to variable varying in time very frequently
  - In the second phase it is more efficient to contact each suitable CE, rather than using the GOUT/II as source of information

The Matchmaking algorithm: job submission without data access requirements

- The matchmaker can adopt a stochastic selection while searching for the best matching CE, enabling fuzzyness in the matchmaking algorithm
- The user has to set the JDL FuzzyRank attribute to true
- The rank value represents the probability that each CE has to be selected as the best matching one
- The higher the probability is, the higher the rank value is

Interactive Job

- The Interactive job is a job whose standard streams are forwarded to the submitting client
- The user has to set the JDL JobType attribute to interactive
- When an interactive job is submitted, the edg-job-submit command
  - starts a Grid console shadow process in the background that listens on a port assigned by the Operating System
    - The port can be forced through the ListenerPort attribute in the JDL
  - opens a new window where the incoming job streams are forwarded
- The DISPLAY environment variable has to be set correctly, because an X window is open
- The user can specify --nogui option, which makes the command provide a simple standard non-graphical interaction with the running job
- It is not necessary to specify the OutputSandbox attribute in the JDL because the output will be sent to the interactive window

Logical Checkpointing Job

- The Checkpointing job is a job that can be decomposed in several steps
- In every step the job state can be saved in the LB and retrieved later in case of failures
- The job state is a set of pairs <key, value> defined by the user
- The job can start running from a previously saved state and not from the beginning again
- The user has to set the JDL JobType attribute to checkpointable

Logical Checkpointing Job

- When a checkpointable job is submitted and starts from the beginning, the user run simply the edg-job-submit command
  - the number of steps, that represents the job phases, can be specified by the JobSteps attribute
    - e.g. JobSteps = 2;
  - the list of labels, that represents the job phases, can be specified by the JobSteps attribute
    - e.g. JobSteps = {“genuary”, “february”};
- The latest job state can be obtained by using the edg-job-get-chkpt <jobid> command
- A specific job state can be obtained by using the edg-job-get-chkpt –cs <state_num> <jobid> command
- When a checkpointable job has to start from an intermediate job state, the user run the edg-job-submit command using the –chkpt <state_jdl> option where <state_jdl> is a valid job state file, where the state of a previously submitted job was saved

Other (most relevant) UI commands

- `edg-job-attach`
  - Starts an interactive session for previously submitted interactive jobs
  - Srarts a listener process on the UI machine
- `edg-job-get-chkpt`
  - Allows the user to retrieve one or more checkpoint states by a previously submitted job

MPI Job

- There are a lot of libraries supporting parallel jobs, but we decided to support MPICH.
- The MPI job is run in parallel on several processors
- The user has to set the JDL **JobType** attribute to MPICH and specify the **NodeNumber** attribute that's the required number of CPUs
- When a MPI job is submitted, the UI adds
  - in the **Requirements** attribute

    `Member("MpiCH", other.GlueHostApplicationSoftwareRunTimeEnvironment)` (the MPICH runtime environment must be installed on the CE)

    `other.GlueCEInfoTotalCPUs >= NodeNumber` (a number of CPUs must be at least be equal to the required number of nodes)
  - In the Rank attribute

    `other.GlueCEStateFreeCPUs` (it is chosen the CE with the largest number of free CPUs)

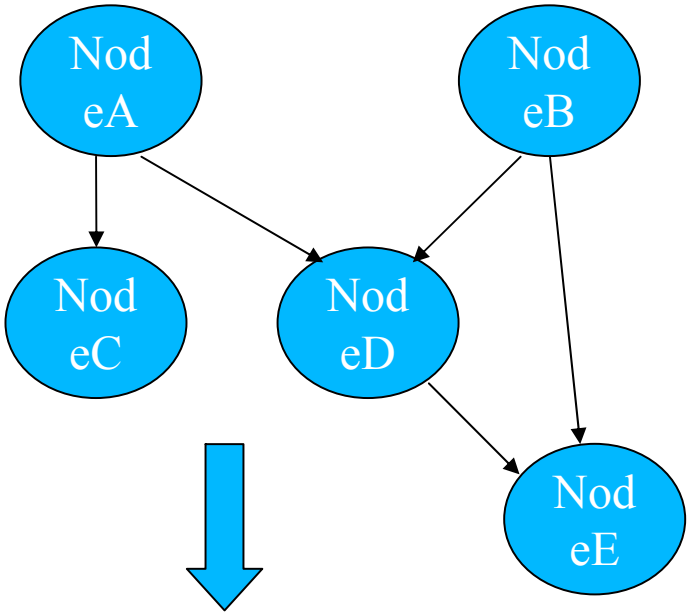
MPI Job

```
[
JobType = "MPICH";
NodeNumber = 4;
Executable = "MPItest.sh";
Argument = "cpi 4";
InputSandbox = {"MPItest.sh", "cpi"};
OutputSandbox = "executable.out";
Requirements = other.GlueCEInfoLRMSType == “PBS” ||
other.GlueCEInfoLRMSType == “LSF”;
]
```

- The NodeNumber entry is the number of threads of MPI job
- The MPItest.sh script only works if PBS or LSF is the local job manager
- If you want to submit your MPI programs you have to compile them

What is a DAG

- DAG means Directed Acyclic Graph
- Each node represents a job
- Each edge represents a temporal dependency between two nodes
  - e.g. NodeC starts only after NodeA has finished
- A dependency represents a constraint on the time a node can be executed
  - Limited scope, it may be extended in the future
- Dedendencies are represented as “expression lists” in the ClassAd language




```
dependencies = {
{NodeA, {NodeC, NodeD}},
{NodeB, NodeD},
{NodeB, NodeD}, NodeE}
}
```

DAG Job

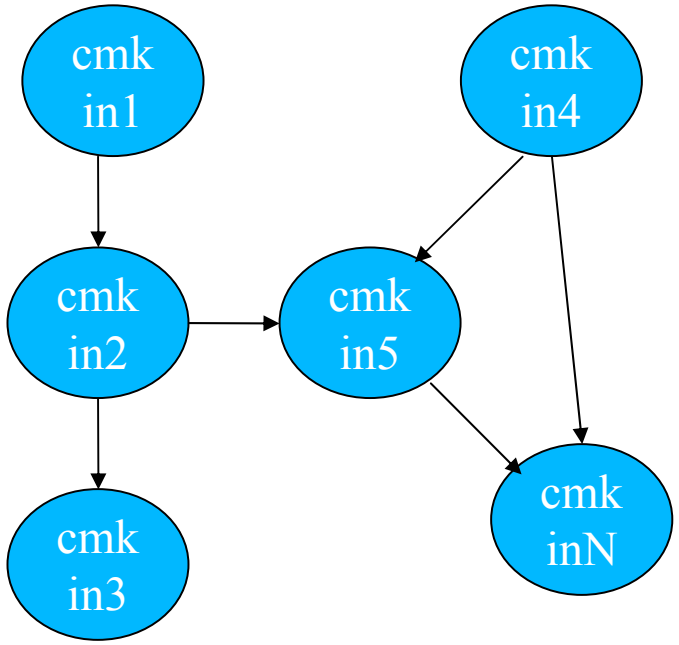
- The DAG job is a Directed Acyclic Graph Job
- The sub-jobs are scheduled only when the corresponding DAG node is ready
- The user has to set the JDL JobType attribute to dag, nodes attributes that contains the description of the nodes, and dependencies attributes

NOTE:

- A plug-in has been implemented to map an EGEE DAG submission to a Condor DAG submission
- Some improvements have been applied to the ClassAd API to better address WMS need

DAG Job

```
nodes = {
   cmkin1 = [
      file = "bckg_01.jdl" ;
   ],
   cmkin2 = [
      file = "bckg_02.jdl" ;
   ],
   ……
   cmkinN = [
      file = "bckg_0N.jdl" ;
   ]
};
dependencies = {
   {cmkin1, cmkin2},
   {cmkin2, cmkin3},
   {cmkin2, cmkin5},
   {{cmkin4, cmkin5}, cmkinN}
}
```



Summary

- We explained the main functionality of the Workload Management System
- The JDL file describes a user job
- A set of commands allow the user to get status information and retrieve relevant data
- Questions ?