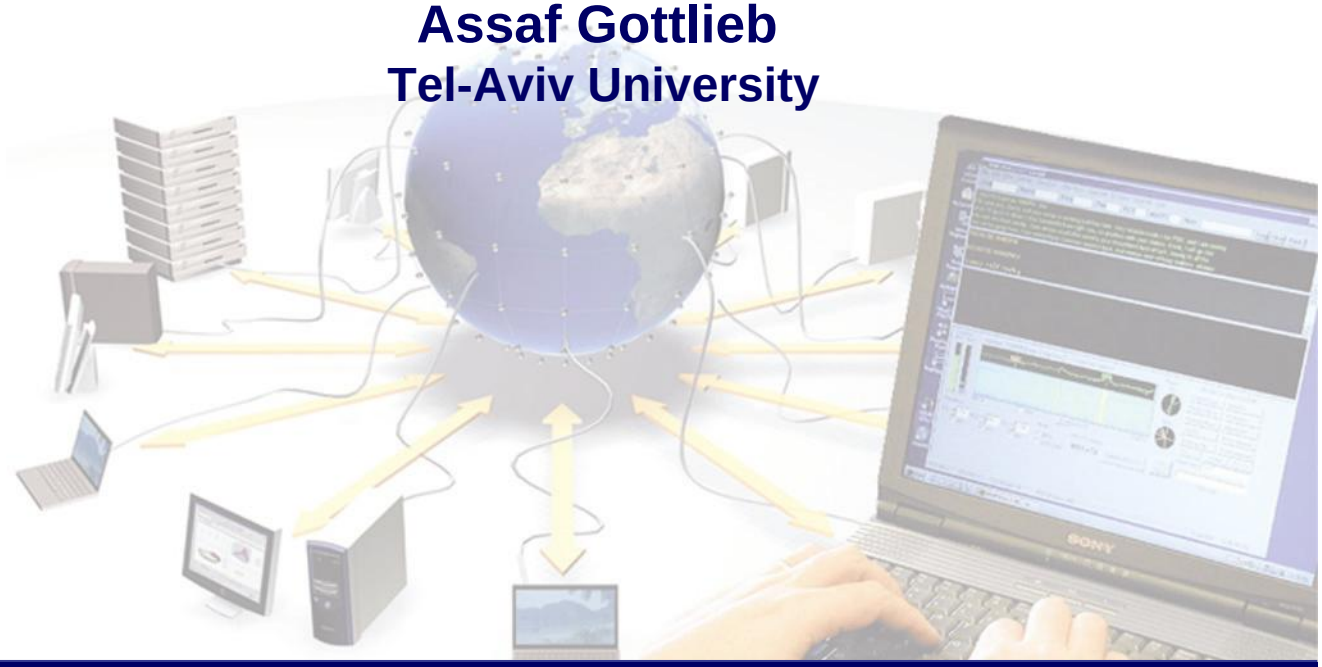
www.eu-egee.org

Job Description Language - more control over your Job

Assaf Gottlieb
Tel-Aviv University



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

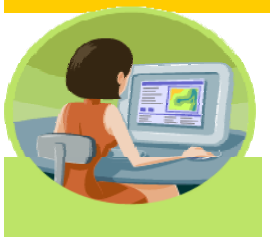
Outline

- Introduction
- Job submission services – what is really going on inside...
- JDL syntax

The use of jobs for running applications

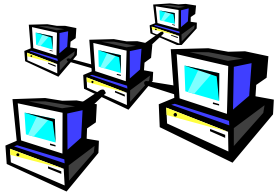
- Jobs are the way users execute applications on the grid.
- Information to be specified when a job has to be submitted:
  - Job characteristics
  - Job requirements and preferences on the computing resources
    - Also including software dependencies
  - Job data requirements
- Information specified using a Job Description Language (JDL)
  - Based upon Condor's *CLASSified ADvertisement language (ClassAd)*
    - Fully extensible language
    - A ClassAd is a sequence of attributes separated by semi-colon (;).

How does it work? Main components

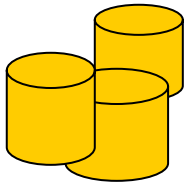


User Interface (UI): The place where users logon to the Grid

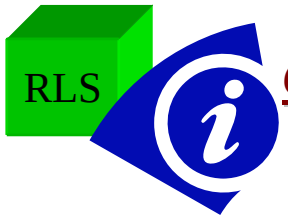
Resource Broker (RB): Matches the user requirements with the available resources on the Grid



Computing Element (CE): A batch queue on a farm of computers where the user Job gets executed



Storage Element (SE): A storage server where Grid files are stored (read/write/copy) or replicated.

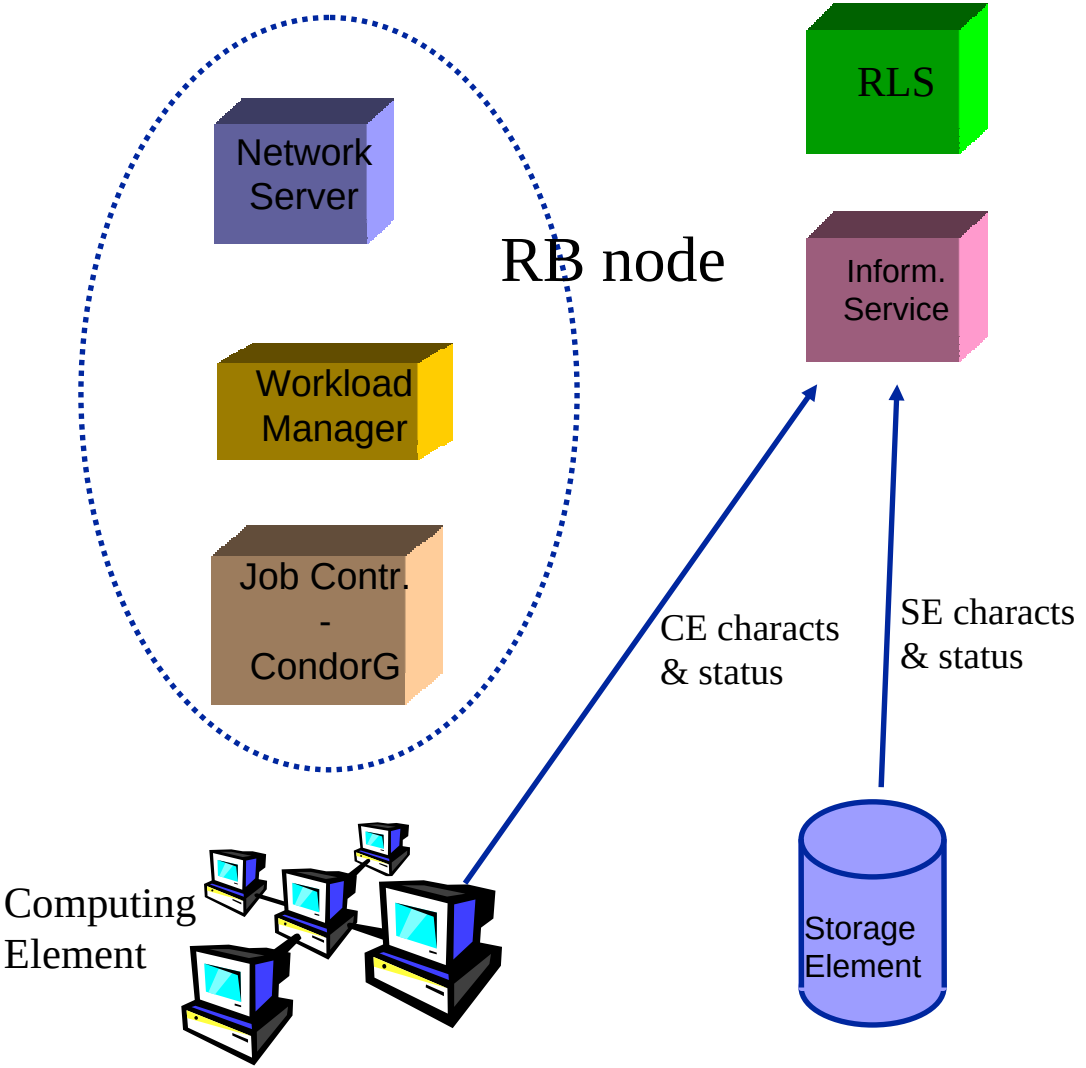


Catalogues(MDS/RLS): A storage server where Grid files are stored (read/write/copy) or replicated.

EGEE/LCG Workload Management System

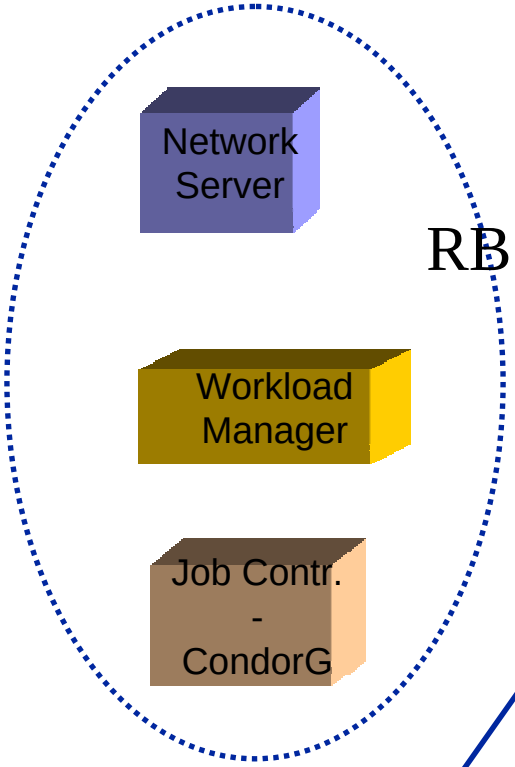
- The user interacts with Grid via a **Workload Management System (WMS)**
- The Goal of WMS is the **distributed scheduling and resource management in a Grid environment**.
- What does it allow Grid users to do?
  - To submit their jobs
  - To execute them on the “best resources”
    - The WMS tries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

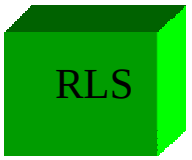
Job Submission



Job Submission

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



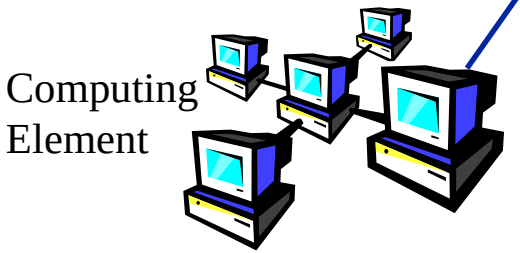


Job Status

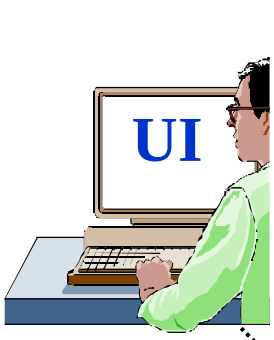
submitted

CE characts & status

SE characts & status



Job Submission



Network Server

RB node

Workload

RLS

Job Status

submitted

Inform. Service

glite-job-submit myjob.jdl

Myjob.jdl

```
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;
```

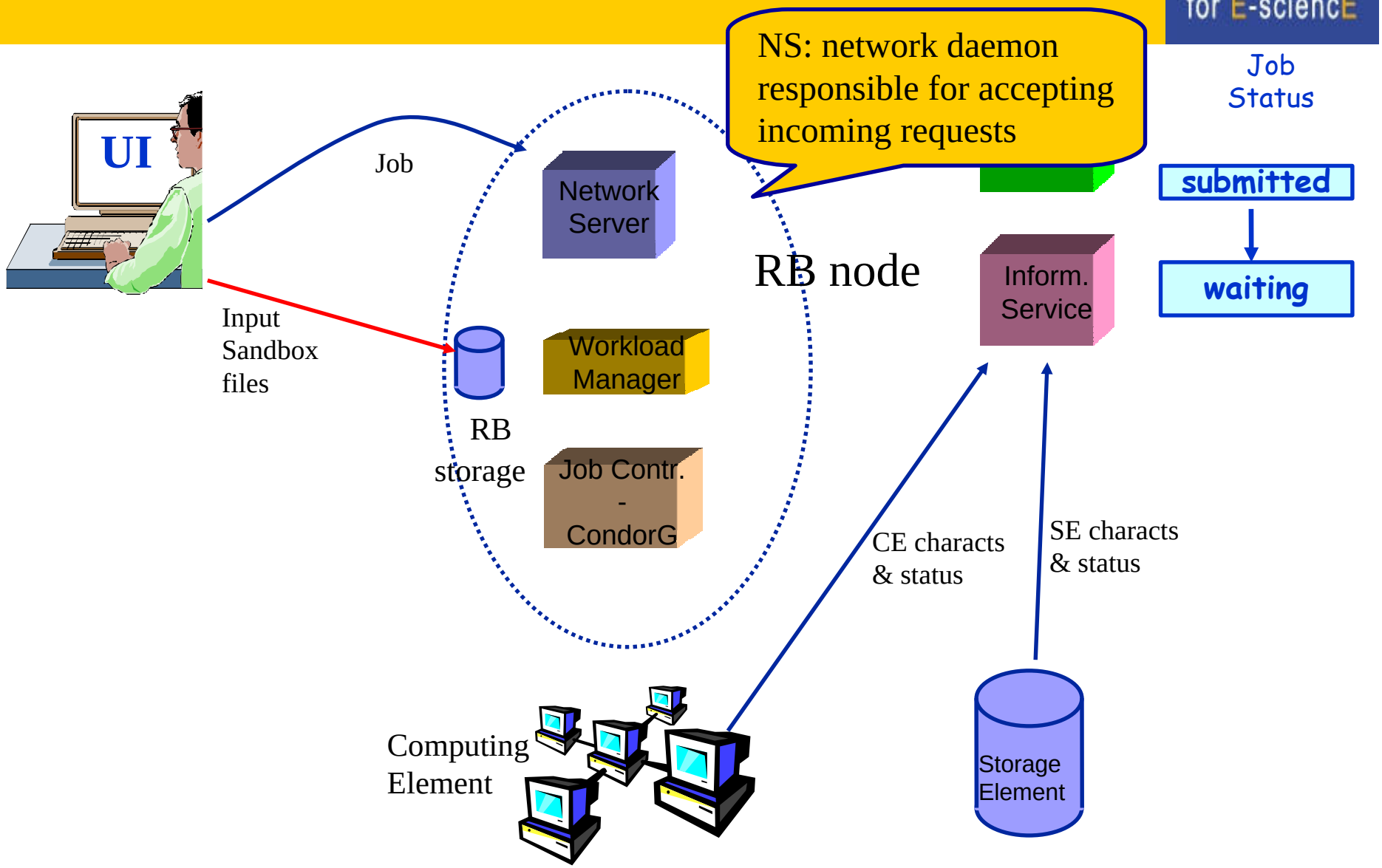
Job Description Language (JDL) to specify job characteristics and requirements

& status

Computing Element

Storage Element

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 Contr. - CondorG

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

Computing Element

RLS

Inform. Service

CE characts & status

SE characts & status

Storage Element

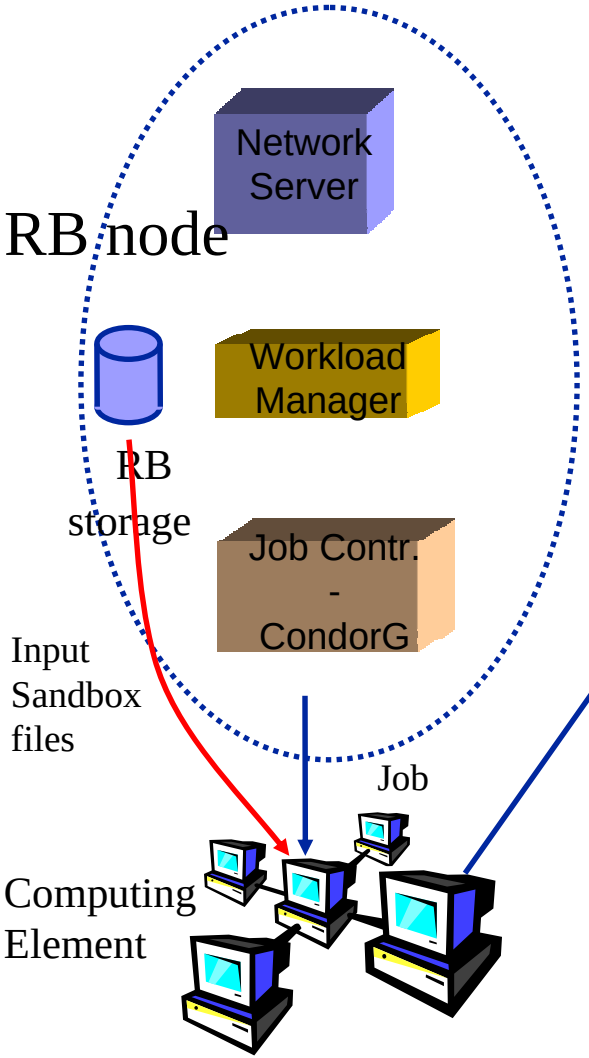
Job Status

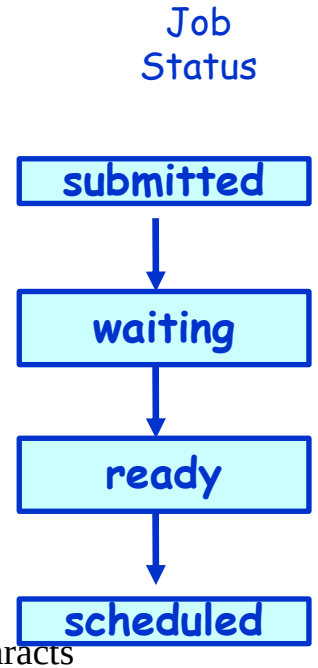
submitted

waiting

ready

Job Submission





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

RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Output Sandbox files

Computing Element

RLS

Inform. Service

Storage Element

Job Status

submitted

waiting

ready

scheduled

running

done

Job Submission

RB node

Network Server

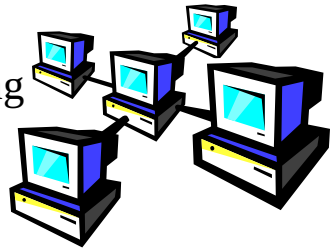
Workload Manager

RB storage

Job Contr. - CondorG

glite-job-output <job-id>

Computing Element



RLS

Inform. Service

Storage Element

Job Status

submitted

waiting

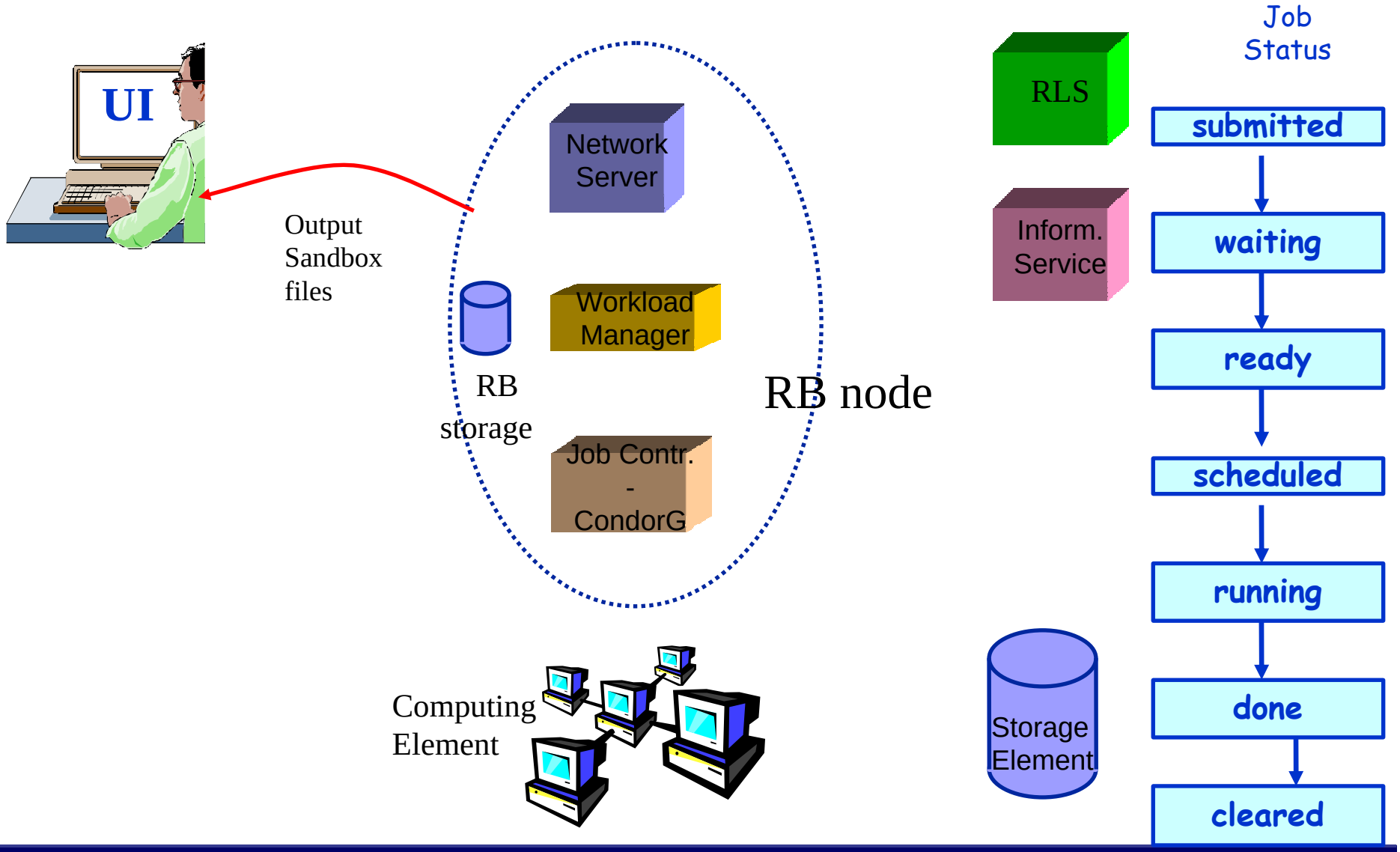
ready

scheduled

running

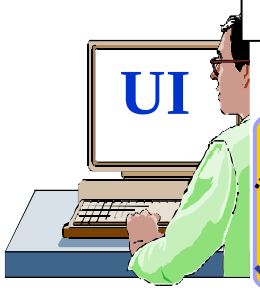
done

Job Submission



Job monitoring

glite-job-status <job-id>
glite-job-logging-info <job-id>



LB: receives and stores job events; processes corresponding job status

Job status

Network Server

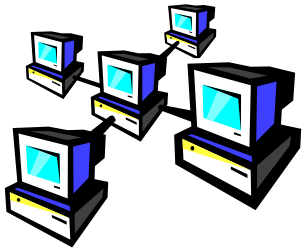
Workload Manager

RB node

Logging & Bookkeeping

Job Contr. - CondorG

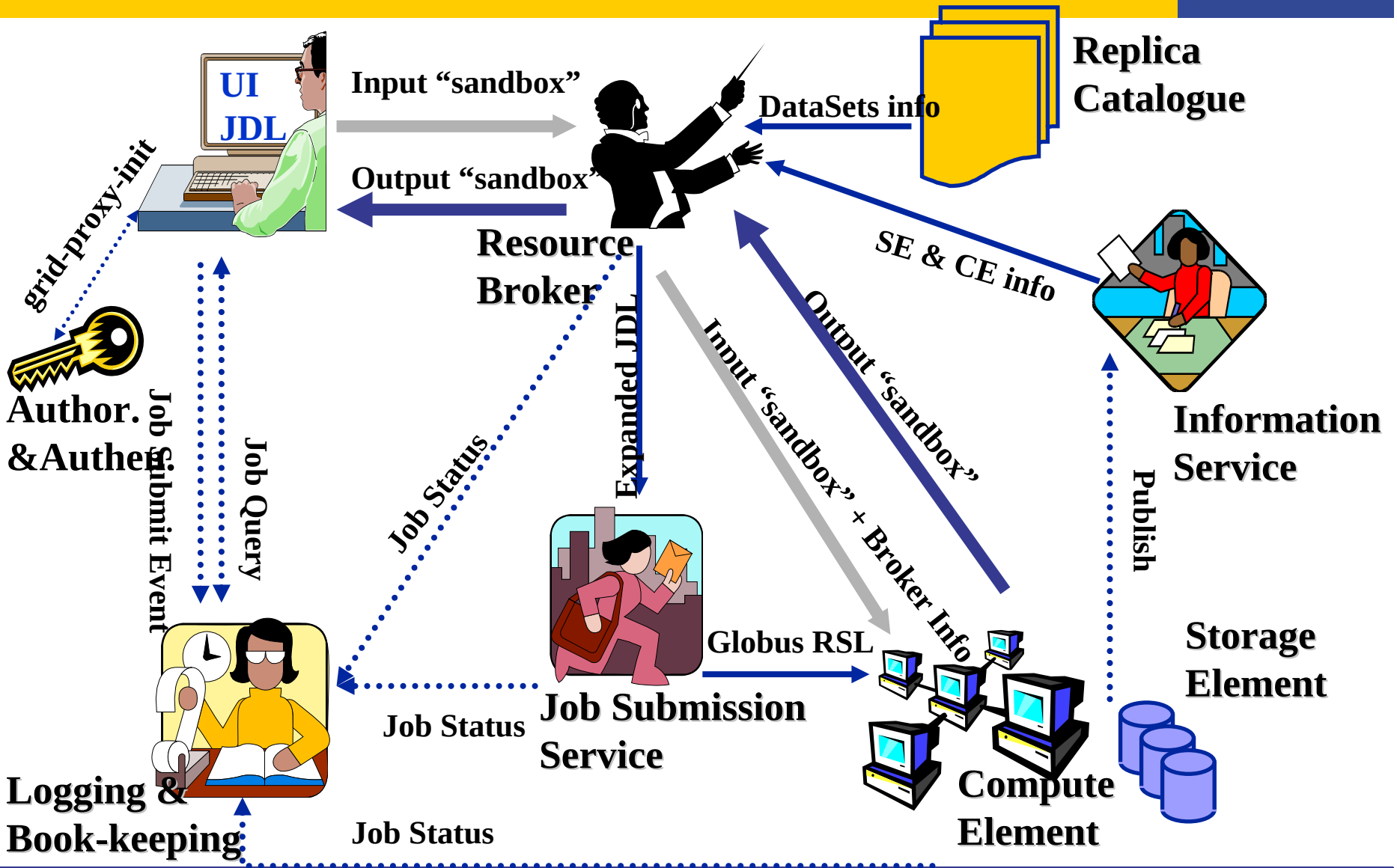
Computing Element



Log Monitor

LM: parses CondorG log file (where CondorG logs info about jobs) and notifies LB

A typical job workflow



Essential JDL - syntax

- 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

Essential JDL

- 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

Essential JDL (contd.)

- **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 (recommended, not mandatory)
  - 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"};
]
```

Job Description Language: 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

Job Description Language: 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)

Job Description Language: 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)

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

`glite-job-submit [-r <res_id>] [-vo <VO>] [-o <output file>] <job.jdl>`

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

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

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

Useful for other commands, e.g.:

`glite-job-status -i` `<input file> (or jobId)`

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

Possible job states

Flag	Meaning
SUBMITTED	submission logged in the LB
WAIT	job match making for resources
READY	job being sent to executing CE
SCHEDULED	job scheduled in the CE queue manager
RUNNING	job executing on a WN of the selected CE queue
DONE	job terminated without grid errors
CLEARED	job output retrieved
ABORT	job aborted by middleware, check ***reason***

Other (most relevant) UI commands

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

Let's try it out!
(But questions first…)

Additional material: Relevant Glue Attributes

- **State (objectclass GlueCEState)**
  - GlueCEStateRunningJobs:
    - number of running jobs
  - GlueCEStateWaitingJobs:
    - number of jobs not running
  - GlueCEStateTotalJobs:
    - total number of jobs (running + waiting)
  - GlueCEStateStatus:
    - queue status: queueing (jobs are accepted but not run), production (jobs are accepted and run), closed (jobs are neither accepted nor run), draining (jobs are not accepted but those in the queue are run)
  - GlueCEStateWorstResponseTime:
    - worst possible time between the submission of a job and the start of its execution
  - GlueCEStateEstimatedResponseTime:
    - estimated time between the submission of a job and the start of its execution
  - GlueCEStateFreeCPUs:
    - number of CPUs available to the scheduler

Relevant Glue Attributes (contd.)

- **Policy (objectclass GlueCEPolicy)**
  - **GlueCEPolicyMaxWallClockTime:**
    - maximum wall clock time available to jobs submitted to the CE, in seconds
  - **GlueCEPolicyMaxCPUTime:**
    - maximum CPU time available to jobs submitted to the CE, in seconds
  - **GlueCEPolicyMaxTotalJobs:**
    - maximum allowed total number of jobs in the queue
  - **GlueCEPolicyMaxRunningJobs:**
    - maximum allowed number of running jobs in the queue
  - **GlueCEPolicyPriority:**
    - information about the service priority

Relevant Glue Attributes (contd.)

- **Architecture (objectclass GlueHostArchitecture)**
  - **GlueHostArchitecturePlatformType:**
    - platform description
  - **GlueHostArchitectureSMPSize:**
    - number of CPUs
- **Processor (objectclass GlueHostProcessor)**
  - **GlueHostProcessorVendor:**
    - name of the CPU vendor
  - **GlueHostProcessorModel:**
    - name of the CPU model
  - GlueHostProcessorVersion:
    - version of the CPU
  - **GlueHostProcessorOtherProcessorDescription:**
    - other description for the CPU
  - […]

Relevant Glue Attributes (contd.)

- **Application software (objectclass GlueHostApplicationSoftware)**
  - GlueHostApplicationSoftwareRunTimeEnvironment:
    - list of software installed on this host
- **Main memory (objectclass GlueHostMainMemory)**
  - GlueHostMainMemoryRAMSize:
    - physical RAM
  - GlueHostMainMemoryVirtualSize:
    - size of the configured virtual memory
- **Benchmark (objectclass GlueHostBenchmark)**
  - GlueHostBenchmarkSI00:
    - SpecInt2000 benchmark
  - GlueHostBenchmarkSF00:
    - SpecFloat2000 benchmark
- **Network adapter (objectclass GlueHostNetworkAdapter)**
  - […]
  - GlueHostNetworkAdapterOutboundIP:
    - permission for outbound connectivity
  - GlueHostNetworkAdapterInboundIP:
    - permission for inbound connectivity

Rank

- **Possible rank values are below reported:**
  - **-other.GlueCEStateEstimatedResponseTime**
    (the lowest estimated traversal time)
  - **other.GlueCEStateFreeCPUs**
    (the highest number of free CPUs)
    - Bad idea: number of free CPU published per QUEUE, not per VO
  - **(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)