Workload Management System

Jason Shih
WLCG T2 Asia Workshop
Dec 2, 2006: TIFR

Academia Sinica Grid Computing

Outline

- WMS introduction
- Job Submission Sequence and WMS Components
- User Job submit

Need Workload Management System

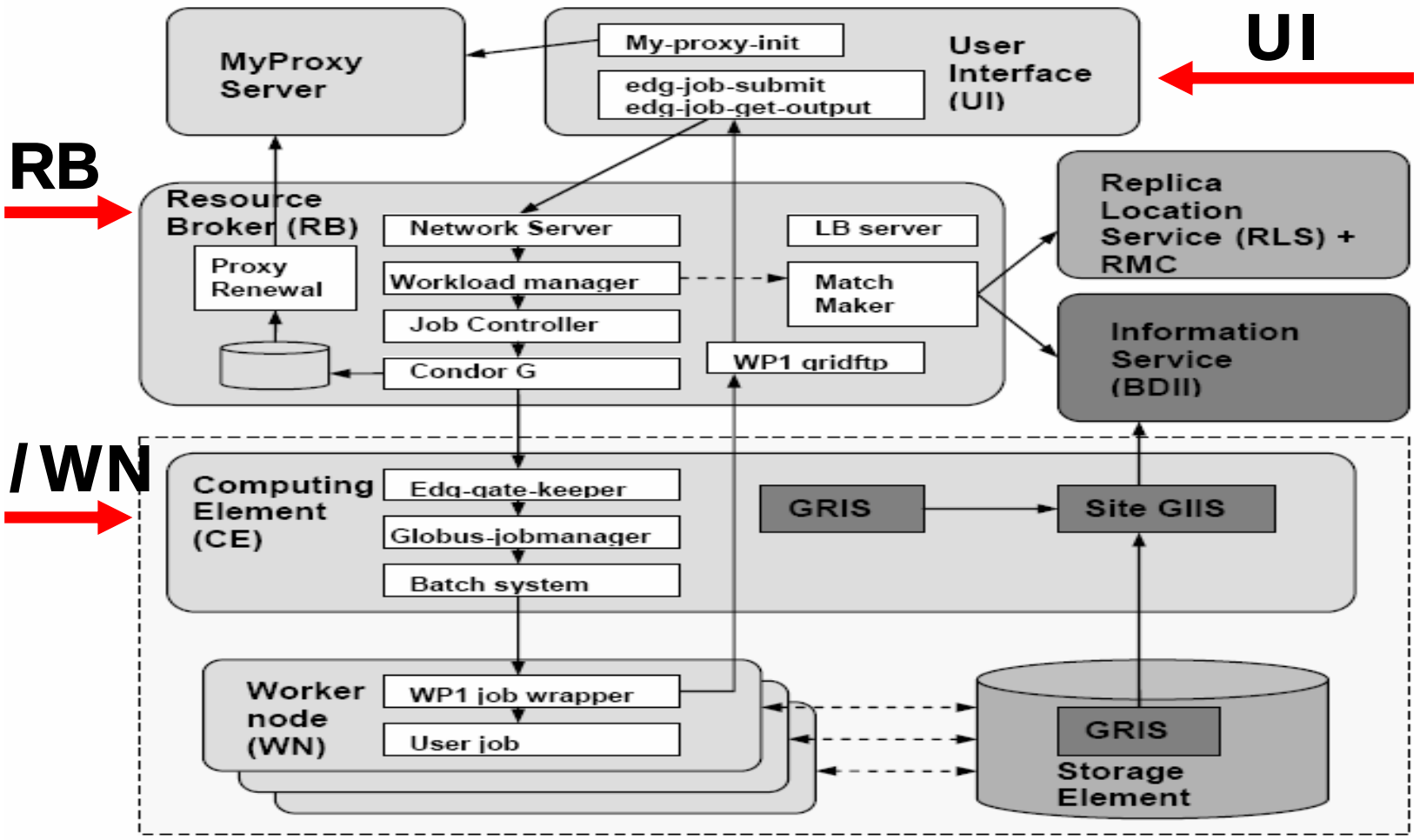
Why we need workload management system?

- **For Grid environment:**

  need distributed scheduling and resource management.

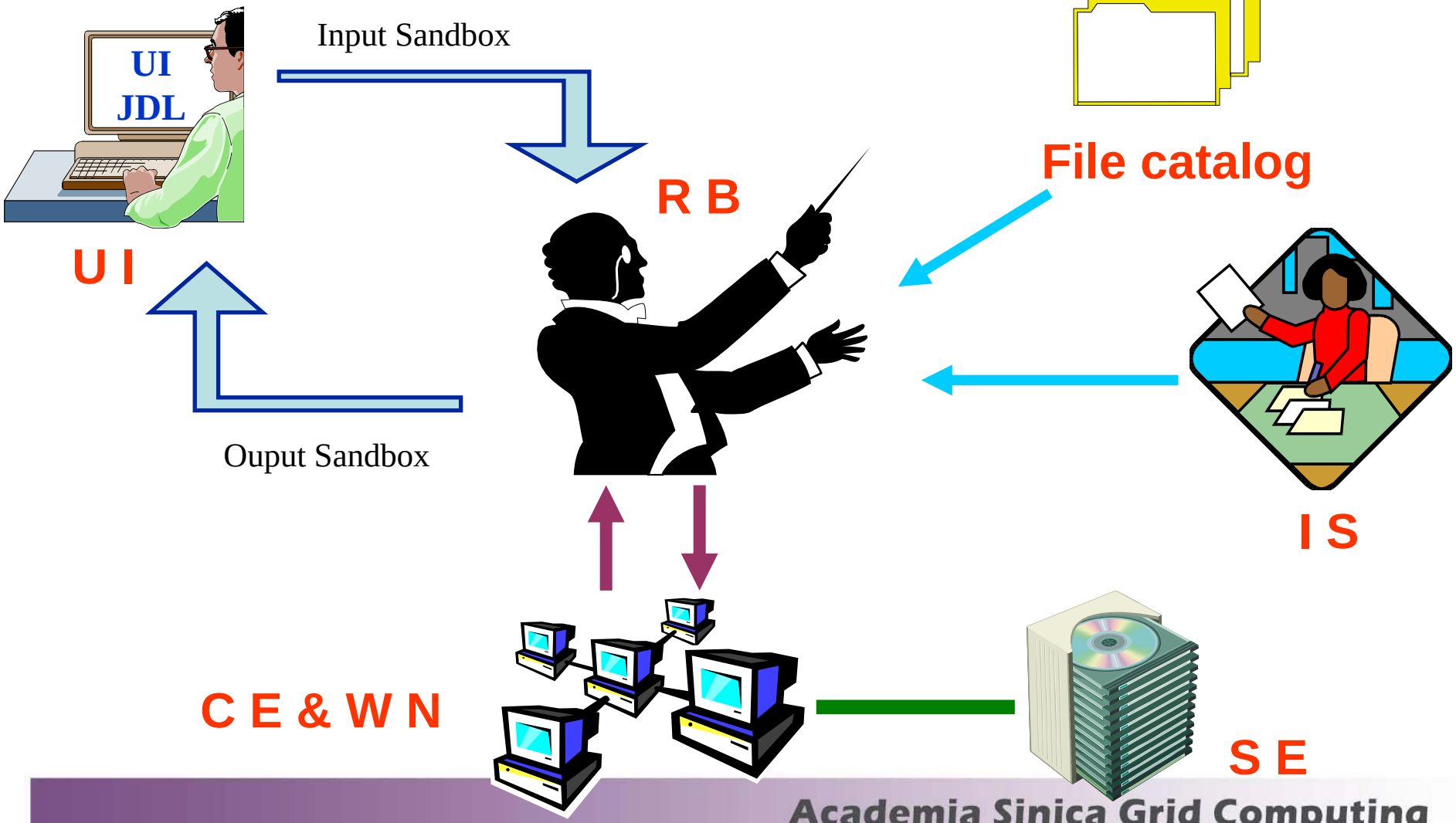
- **For a user:**
  - To submit their jobs.
  - To execute them on the “best resources”.
  - To get information about their status.
  - To retrieve their output.

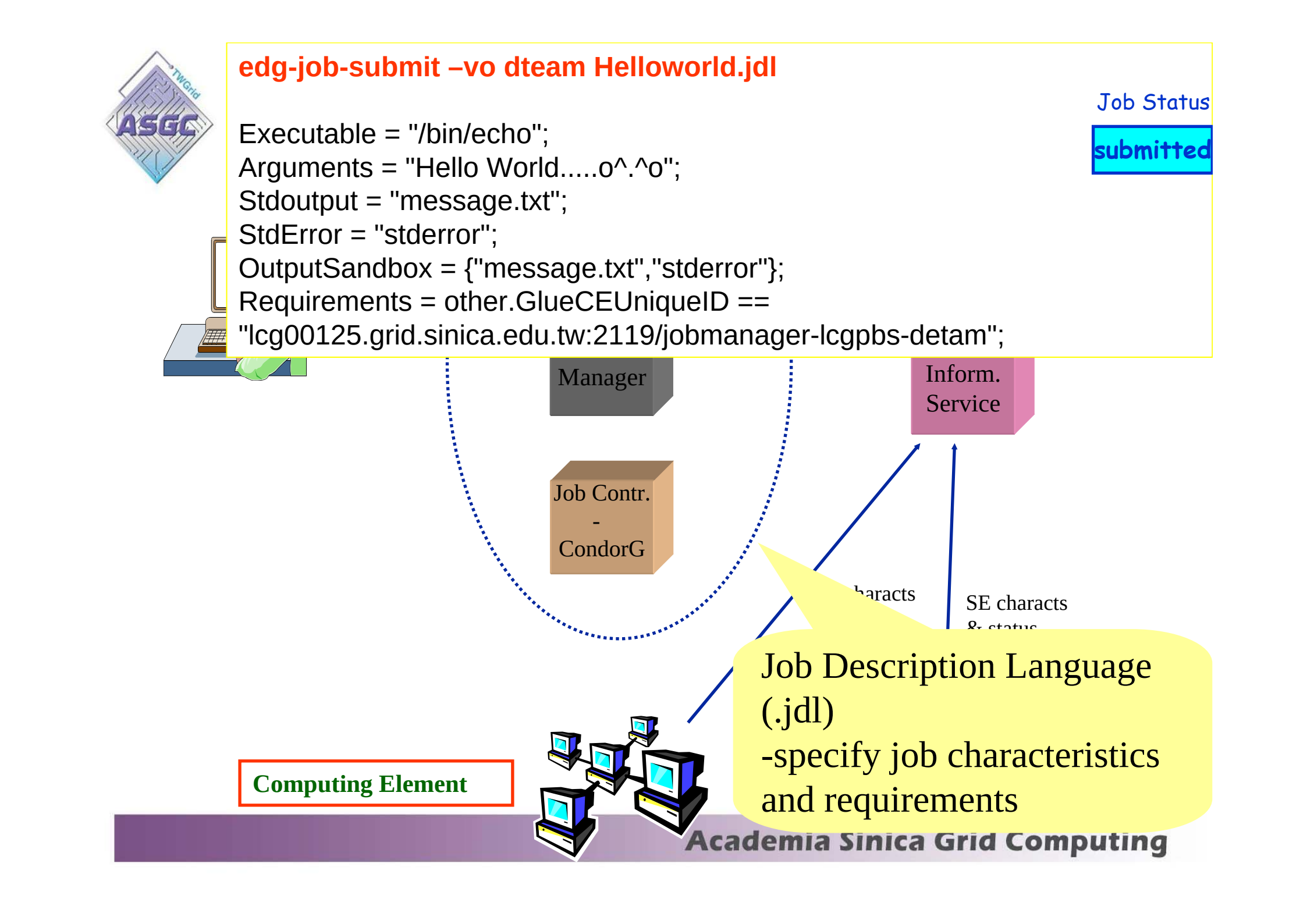
WMS Architecture



- WMS introduction
- **Job Submission Sequence and WMS Components**
- User Job submit

Job Submission Flow

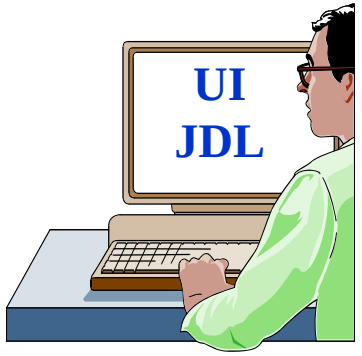


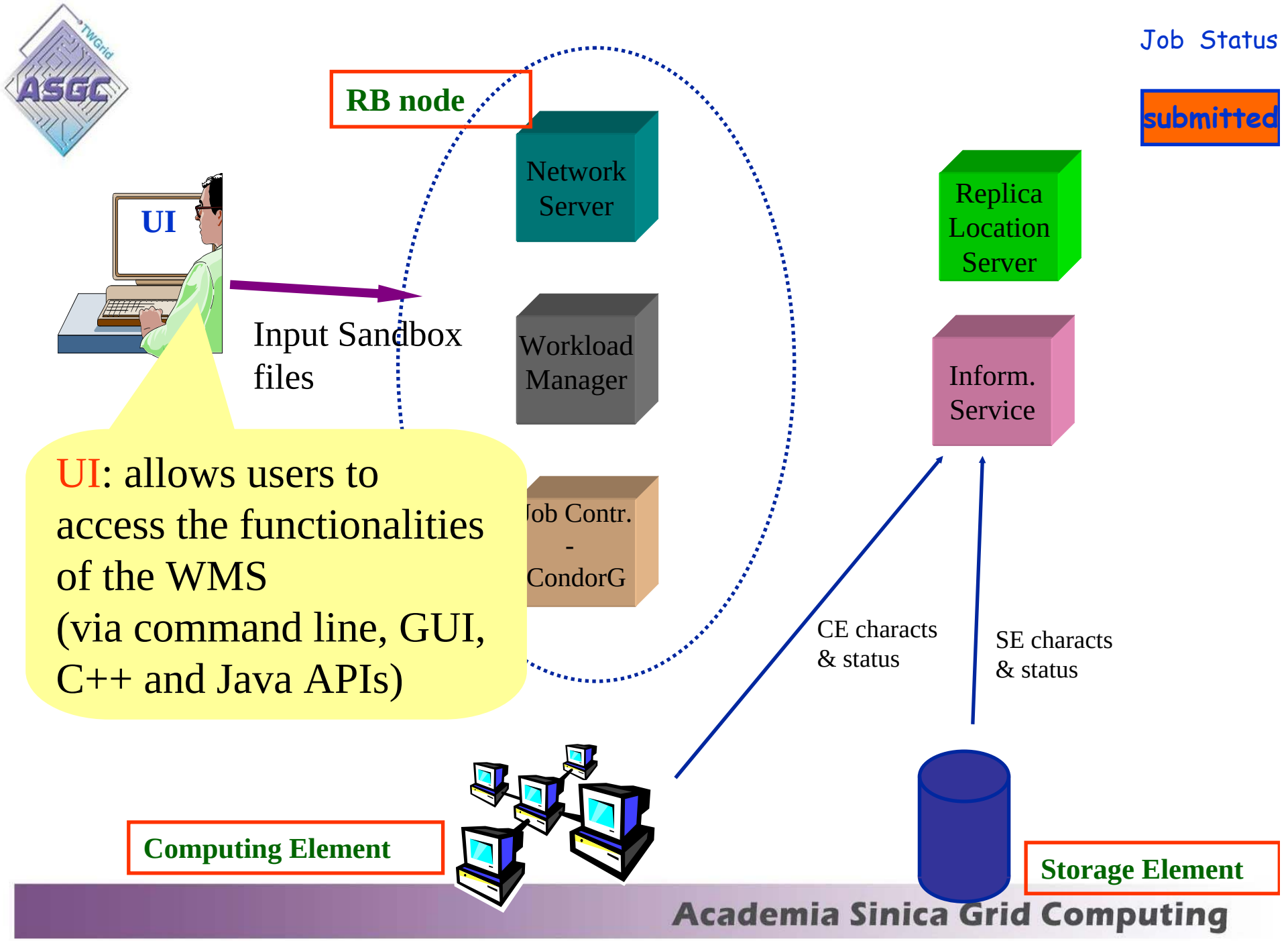

edg-job-submit –vo dteam Helloworld.jdl
Executable = "/bin/echo";
Arguments = "Hello World.....o^.^o";
Stdoutput = "message.txt";
StdError = "stderror";
OutputSandbox = {"message.txt","stderror"};
Requirements = other.GlueCEUniqueID ==
"lcg00125.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-detam";
Job Status
submitted
Manager
Inform.
Service
Job Contr.
-
CondorG
SE characts
Job Description Language
(.jdl)
-specify job characteristics
and requirements
Computing Element
Academia Sinica Grid Computing

User Interface

- **The user's interface to the Grid.**
- The basic functionalities are:

- list the computing resources
- submit a job,
- get the job status,
- cancel a job,
- retrieve the output of a job.




Job Status
submitted
RB node
Network Server
Replica Location Server
UI
Input Sandbox files
Workload Manager
Inform. Service
UI: allows users to access the functionalities of the WMS (via command line, GUI, C++ and Java APIs)
Job Contr. - CondorG
CE characts & status
SE characts & status
Computing Element
Storage Element
Academia Sinica Grid Computing

Resource Broker

- Run the Workload Management System
  - To accept job submissions
- It provides a matchmaking service:
  - Dispatch jobs to appropriate Compute Element (CE)
  - Allow users
    - To get information about their status
    - To retrieve their output
- A configuration file on each UI node determines which RB node(s) will be used.

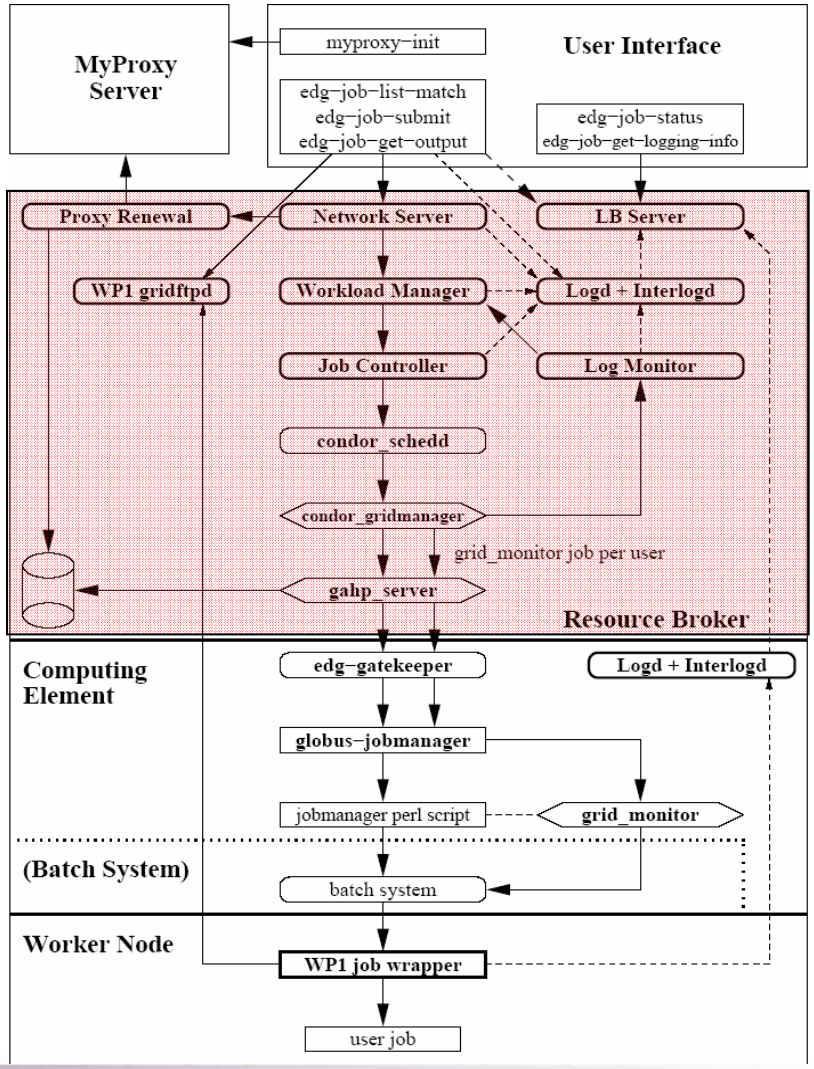
Resource Broker (NS & WM)

Network Server (NS)

- Accepting incoming requests from the UI.
- **Authenticates the user.**
- Obtains a delegated full proxy from the user proxy.
- Enqueues the job to the Workload Manager..

Workload Manager (WM)

- Calls **Matchmaker** to find the resource which best matches the job requirements.
- Interacting with Information System and File catalog.
- Calculates the ranking of all the matchmaked resource.



Resource Broker (JC & CondorG)

Job Controller (JC)

- Converts the condor submit file into ClassAd
- **hands over the job to CondorG.**

Condor-G

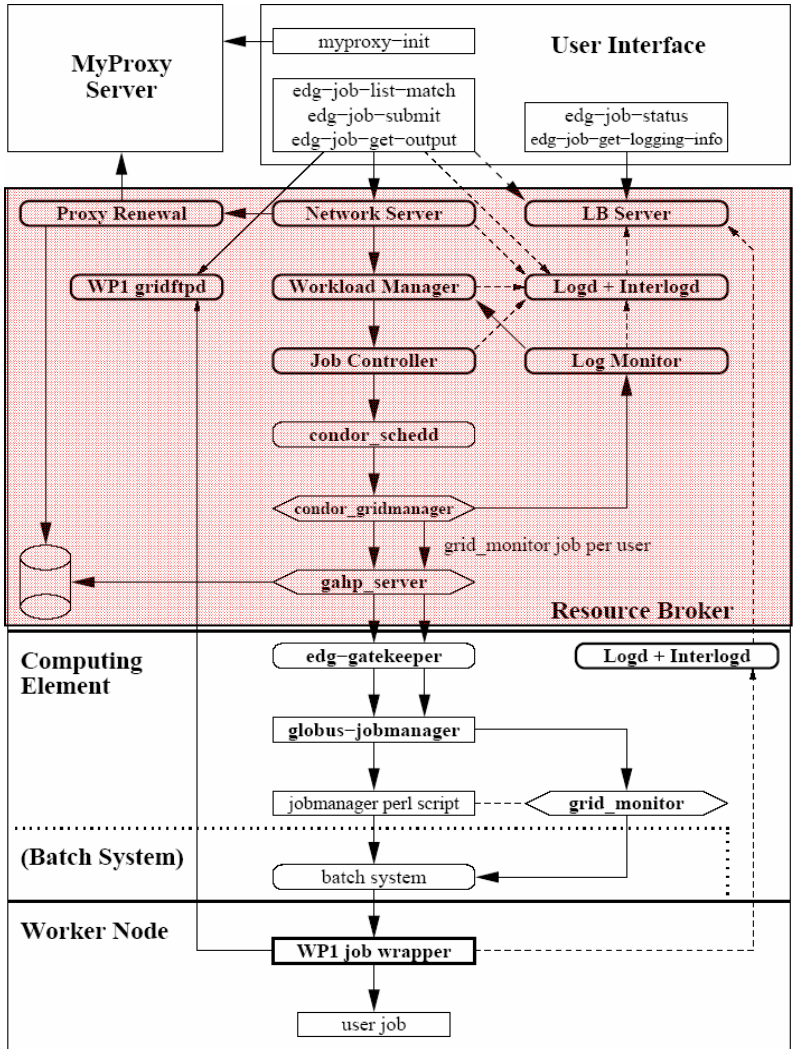
- Condor-G is a Globus-enabled version of the Condor scheduler.
- CondorG consists two elements:

➢ **condor_gridmanager process:**

- Interprets the ClassAD description and traslates it into RSL .
- submits the job to the CE; and it submits an extra job (the ***grid monitor***) per CE and per user to monitor the user jobs.

➢ **The GAHP server**

- It is a GRAM client to contact the edg-gatekeeper .
- It is a GASS server for the results from the grid monitor job.



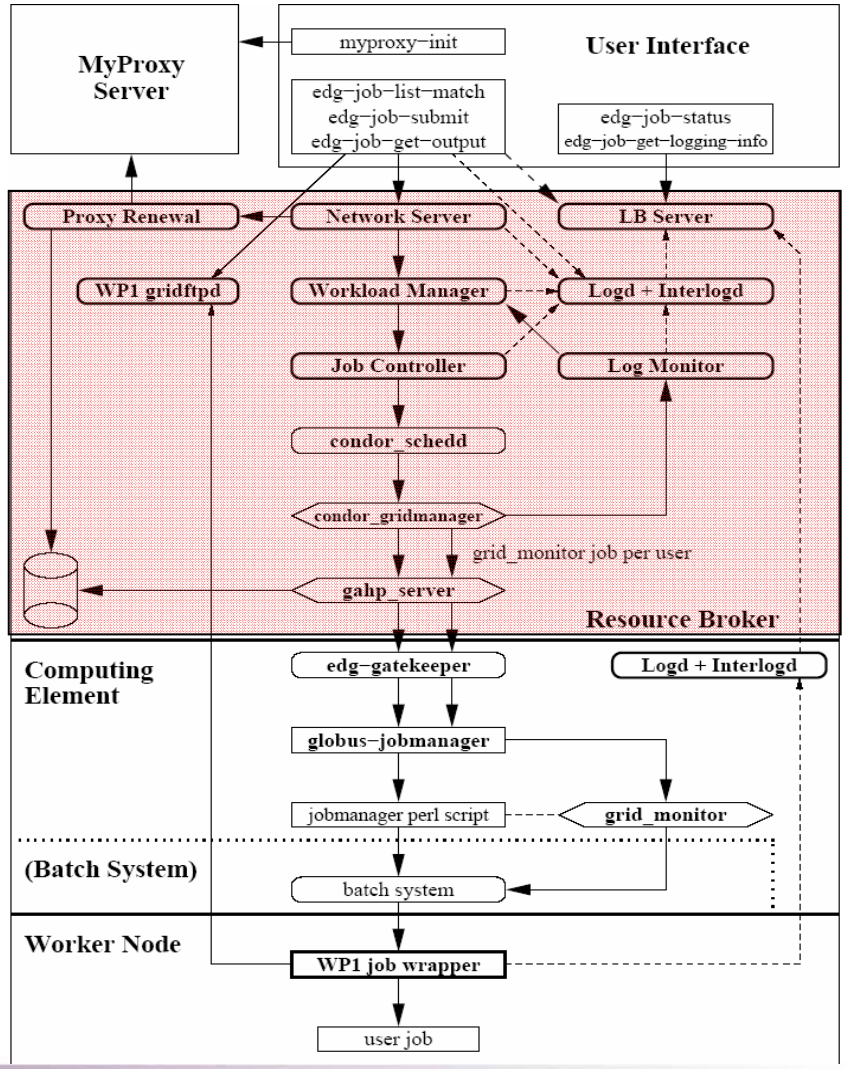
Resource Broker (LM & LB)

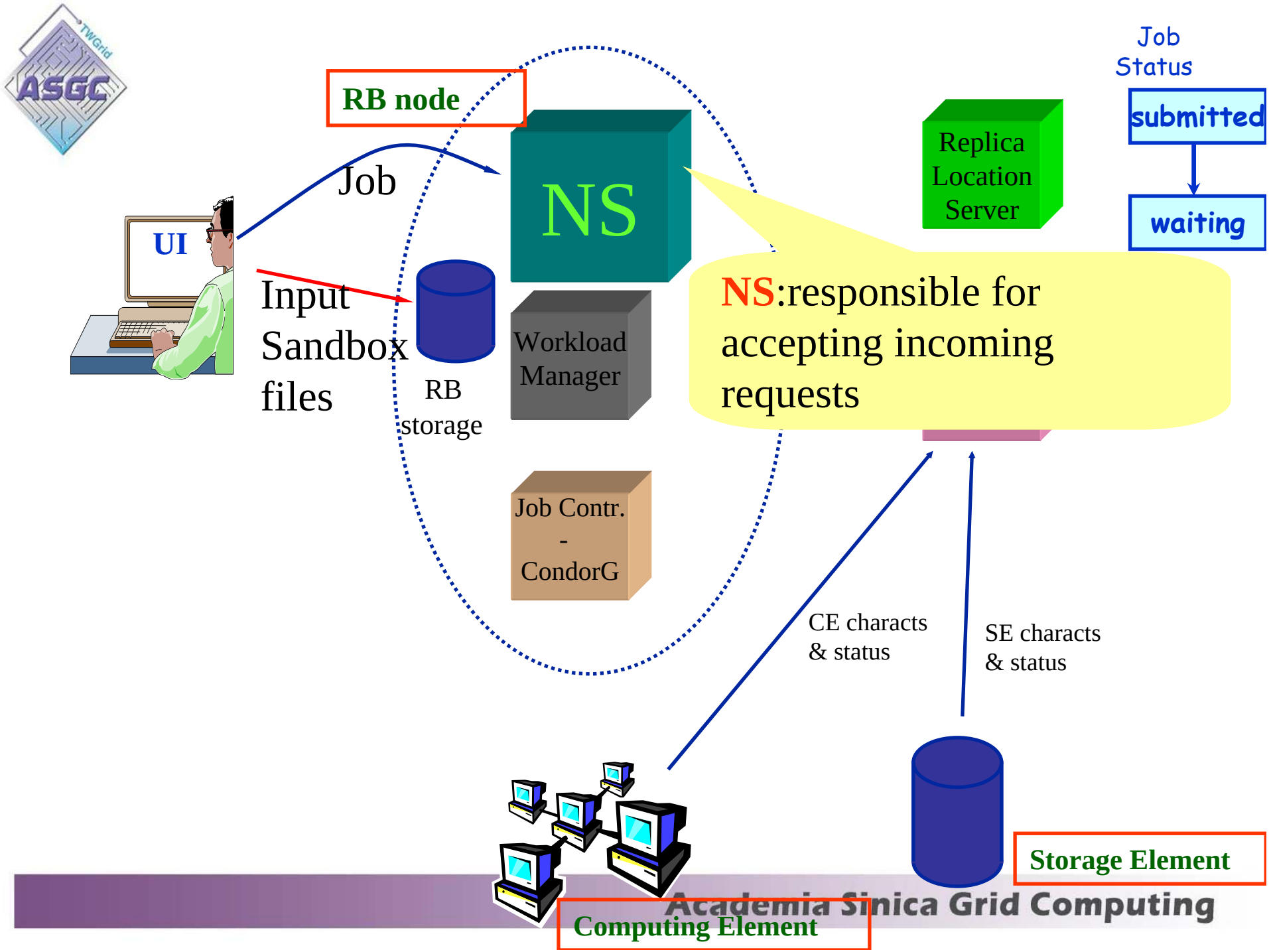
Log Monitor (LM)

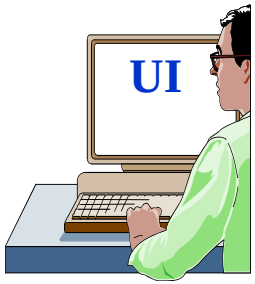
- Continuously parses Condor-G logs.
- Looks for events concerning active jobs

Logging and Bookkeeping (LB)

- All those information are stored by the **logging and bookkeeping** service.
- Collection is done by LB local-loggers




RB node
Job
NS
Replica Location Server
Job Status
submitted
waiting
UI
Input Sandbox files
RB storage
Workload Manager
NS:responsible for accepting incoming requests
Job Contr. - CondorG
CE characts & status
SE characts & status
Storage Element
Computing Element
Academia Sinica Grid Computing



RB node

Network Server

Job

RB storage

WM

Job Contr. - CondorG

WM: acts to satisfy the request

Replica Location Server

Job Status

submitted

waiting

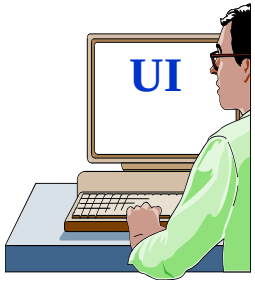
Inform. Service

CE characts & status

SE characts & status

Computing Element

Storage Element



RB node

Network Server

Match-Maker

Replica Location Server

Workload Manager

RB storage

Inform.

Where must this job be executed ?

Job Contr. - CondorG

CE characts & status

SE characts & status

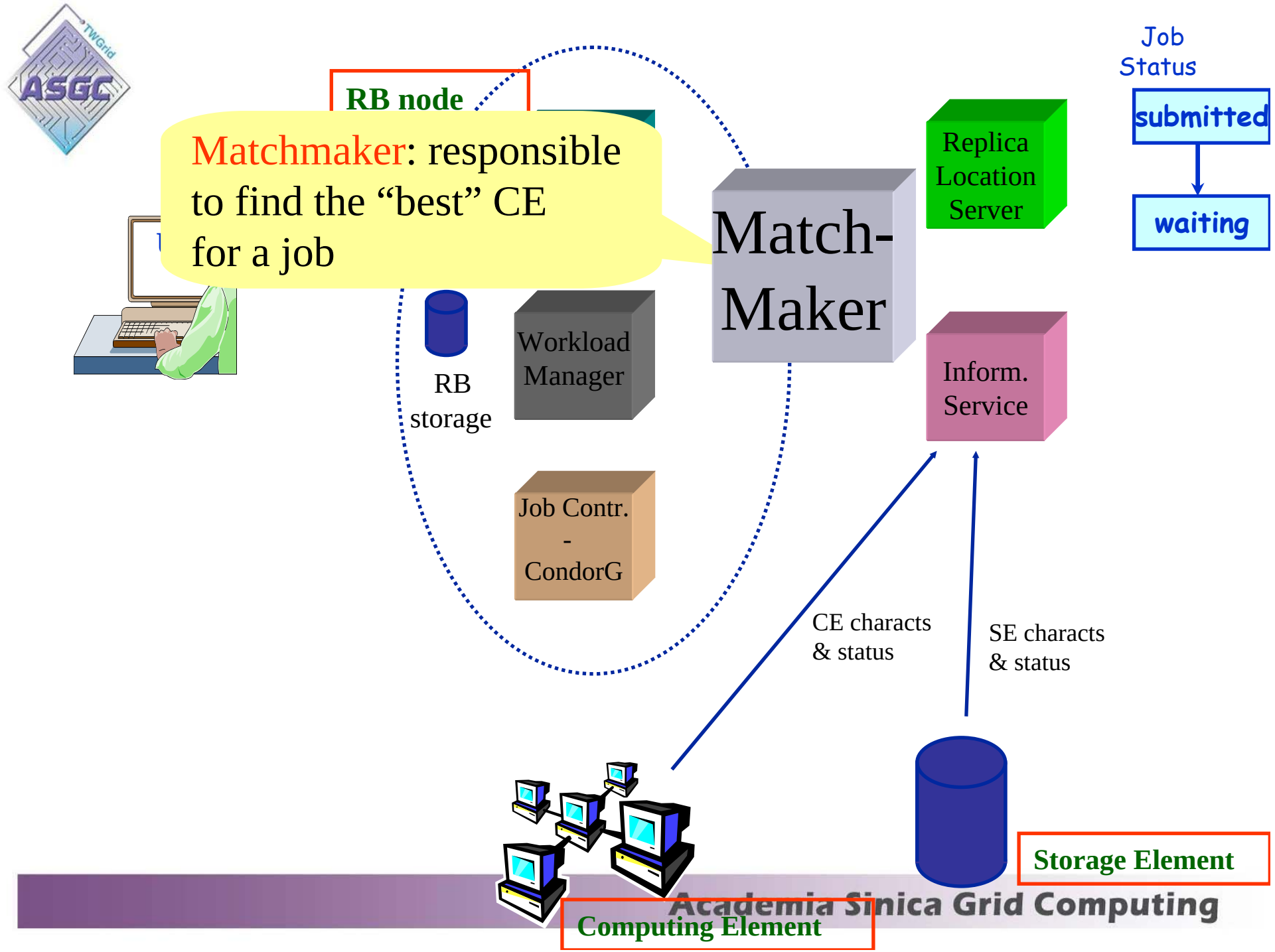
Job Status

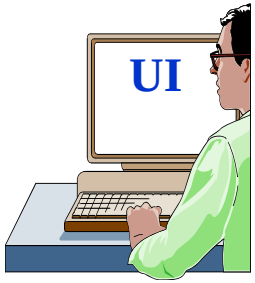
submitted

waiting

Storage Element

Computing Element


RB node
Matchmaker: responsible to find the "best" CE for a job
Match-Maker
Replica Location Server
Job Status
submitted
waiting
Workload Manager
RB storage
Inform. Service
Job Contr. - CondorG
CE characts & status
SE characts & status
Storage Element
Computing Element
Academia Sinica Grid Computing



Where are (which SEs) the needed data ?

Network Server

Replica Location Server

Match-Maker

Workload Manager

RB storage

Inform. Service

Job

Cor

What is the status of the Grid ?

Job Status

submitted

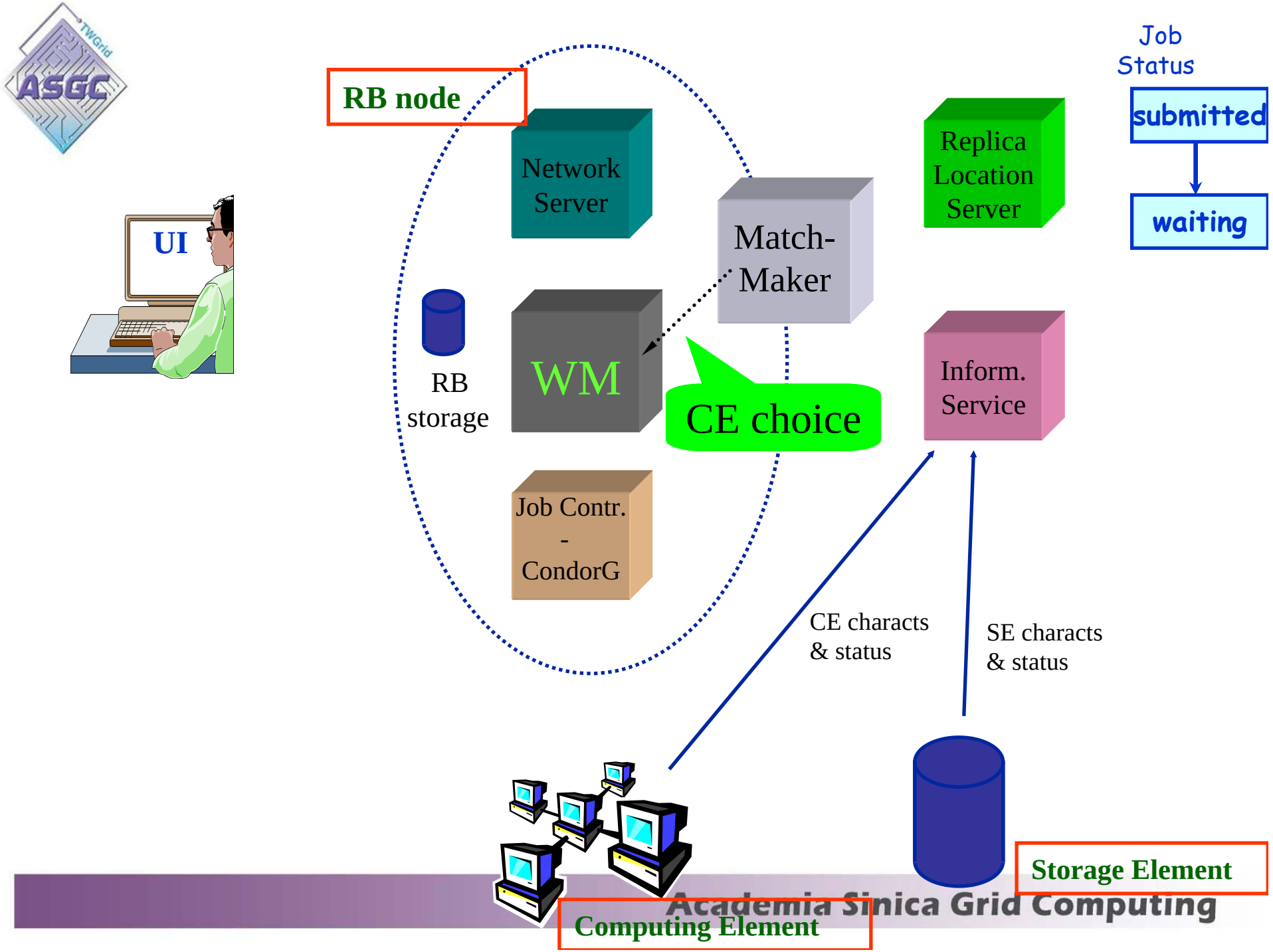
waiting

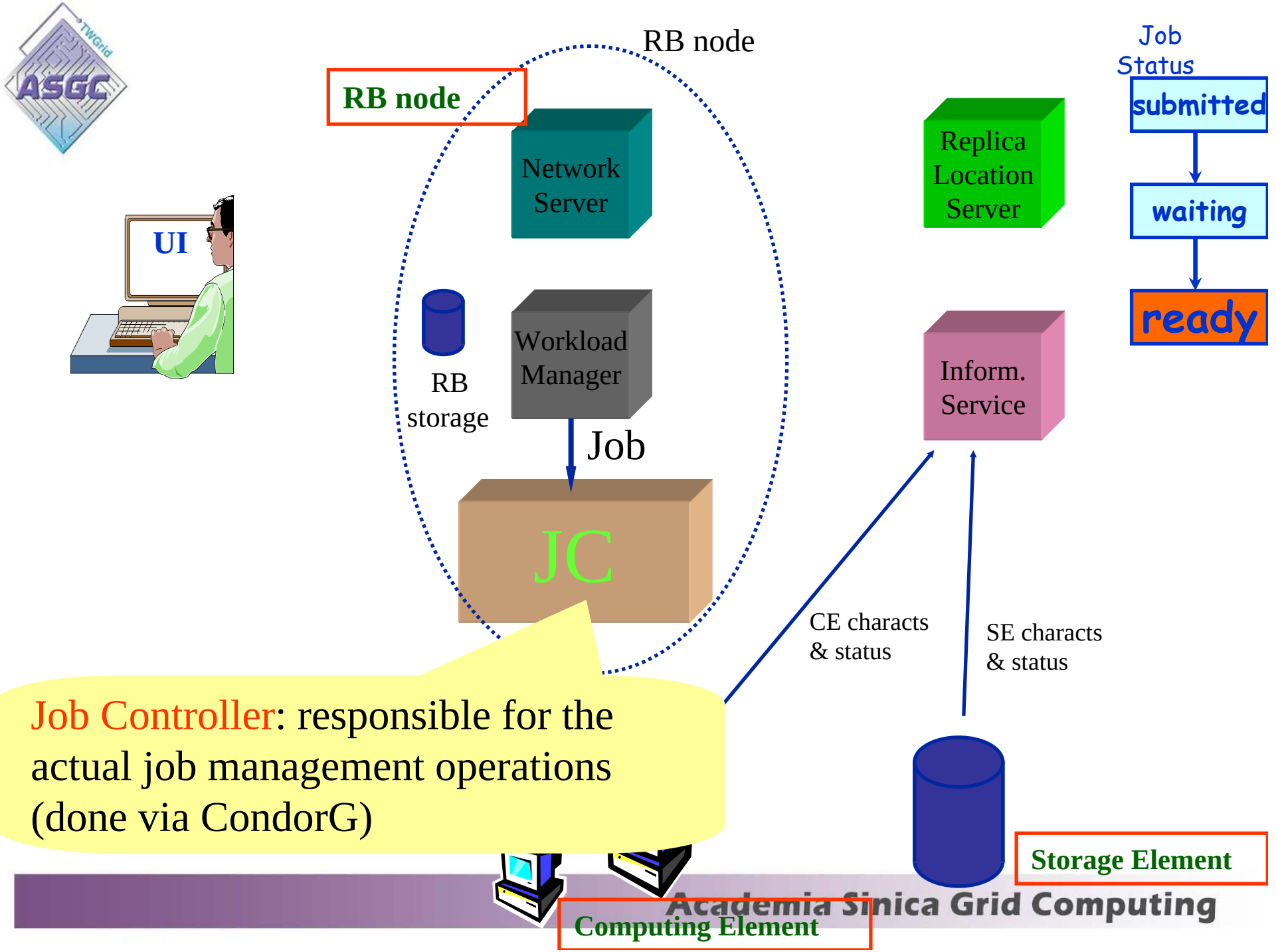
CE characts & status

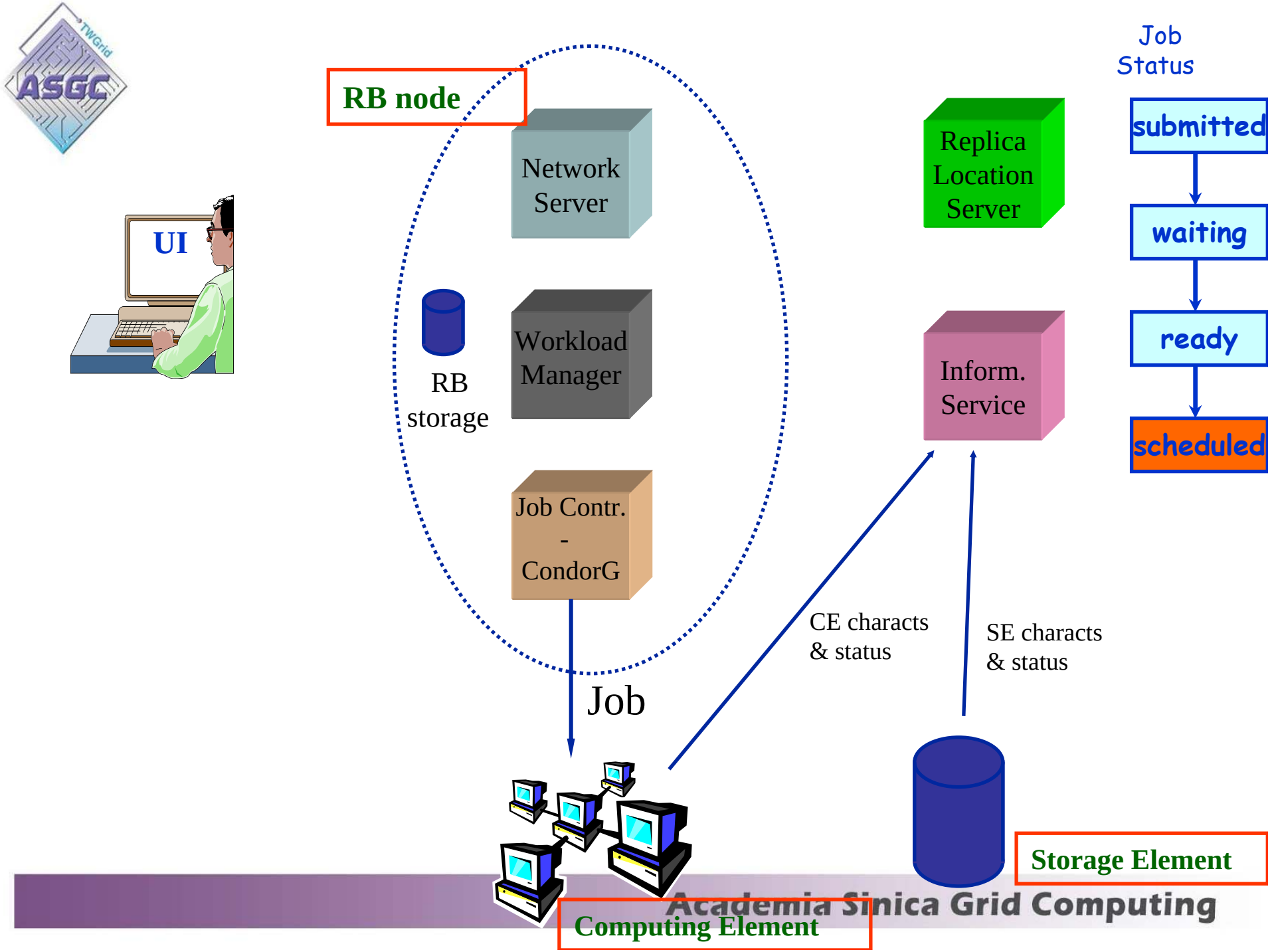
SE characts & status

Storage Element

Computing Element


RB node
Network Server
Match-Maker
Replica Location Server
Job Status
submitted
waiting
UI
RB storage
WM
CE choice
Inform. Service
Job Contr. - CondorG
CE characts & status
SE characts & status
Storage Element
Computing Element
Academia Sinica Grid Computing


RB node
RB node
Network Server
Replica Location Server
Job Status
submitted
waiting
ready
UI
RB storage
Workload Manager
Job
Inform. Service
JC
CE characts & status
SE characts & status
Job Controller: responsible for the actual job management operations (done via CondorG)
Storage Element
Computing Element
Academia Sinica Grid Computing

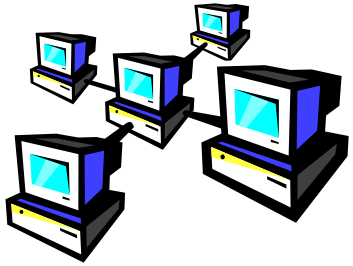

RB node
Network Server
Workload Manager
RB storage
Job Contr. - CondorG
Job
UI
Replica Location Server
Inform. Service
CE characts & status
SE characts & status
Job Status
submitted
waiting
ready
scheduled
Storage Element
Computing Element
Academia Sinica Grid Computing

Computing Element (CE)

- is the interface to a Grid computing nodes.
- The admitted format for CEId is:

  <hostname>:<port>/jobmanager-<service><queue name>

  - i.e :lcg00125.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-dteam
- A Computing Element is built on a homogeneous farm of computing nodes (called **Worker Nodes**)
  - Each LCG-2 site runs at least one CE and a farm of WNs behind it.



Computing Element
(Gatekeeper & Clobus-jobmanager)

Gatekeeper

- **Grants access to the CE**

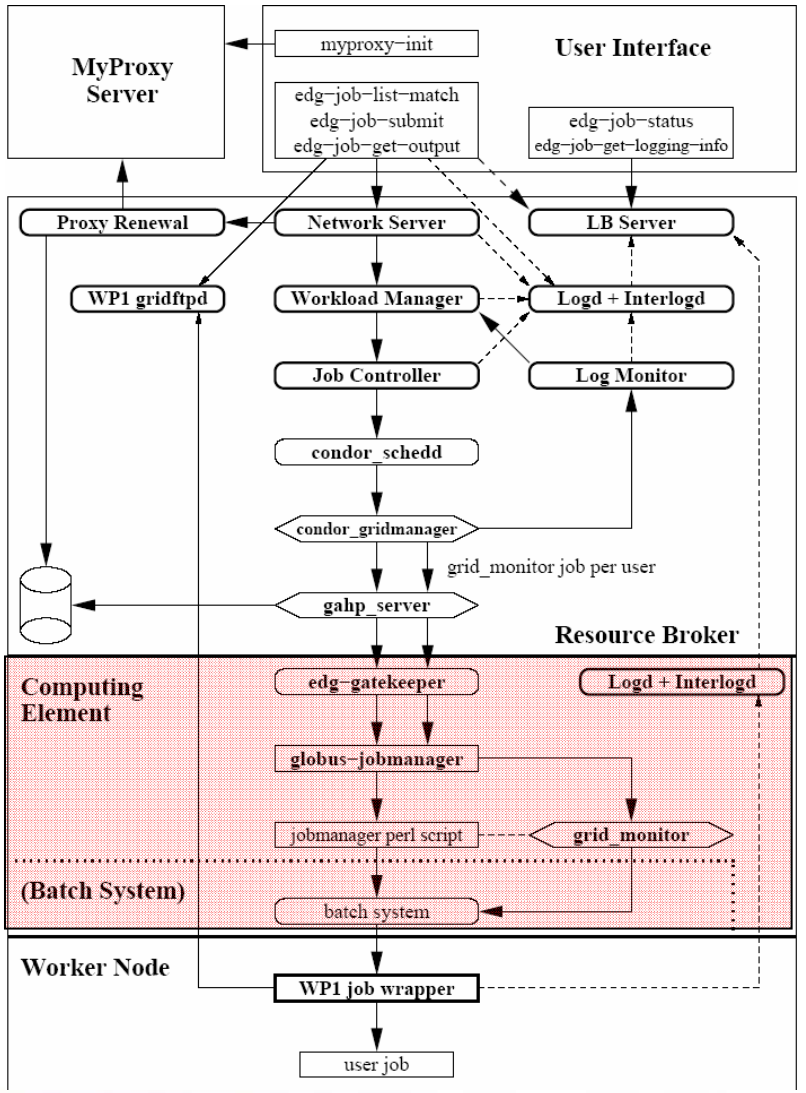
  Authentication and authorization more complicate (compare to RB)

- **the gatekeeper**

  accepts requests from Condor-G, **forks the globus-jobmanager.**

Globus-jobmanager

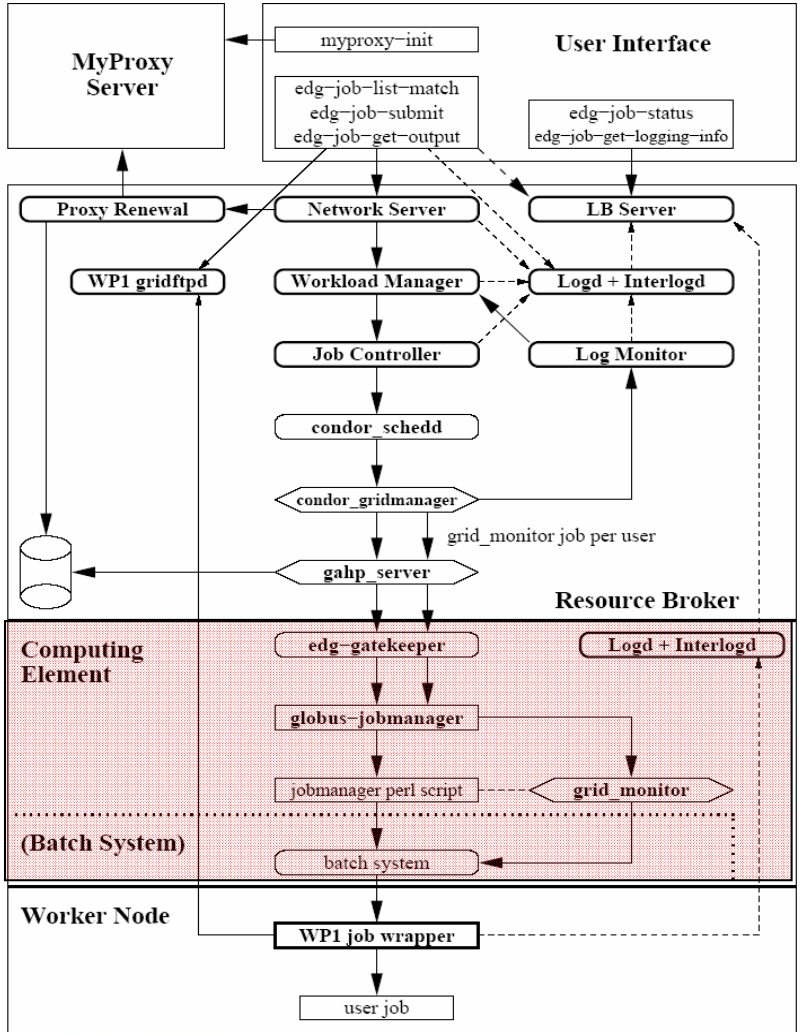
- Offers an interface to the **local batch system.**
- submits or cancel a job.



Computing Element (Batch System)

Batch System

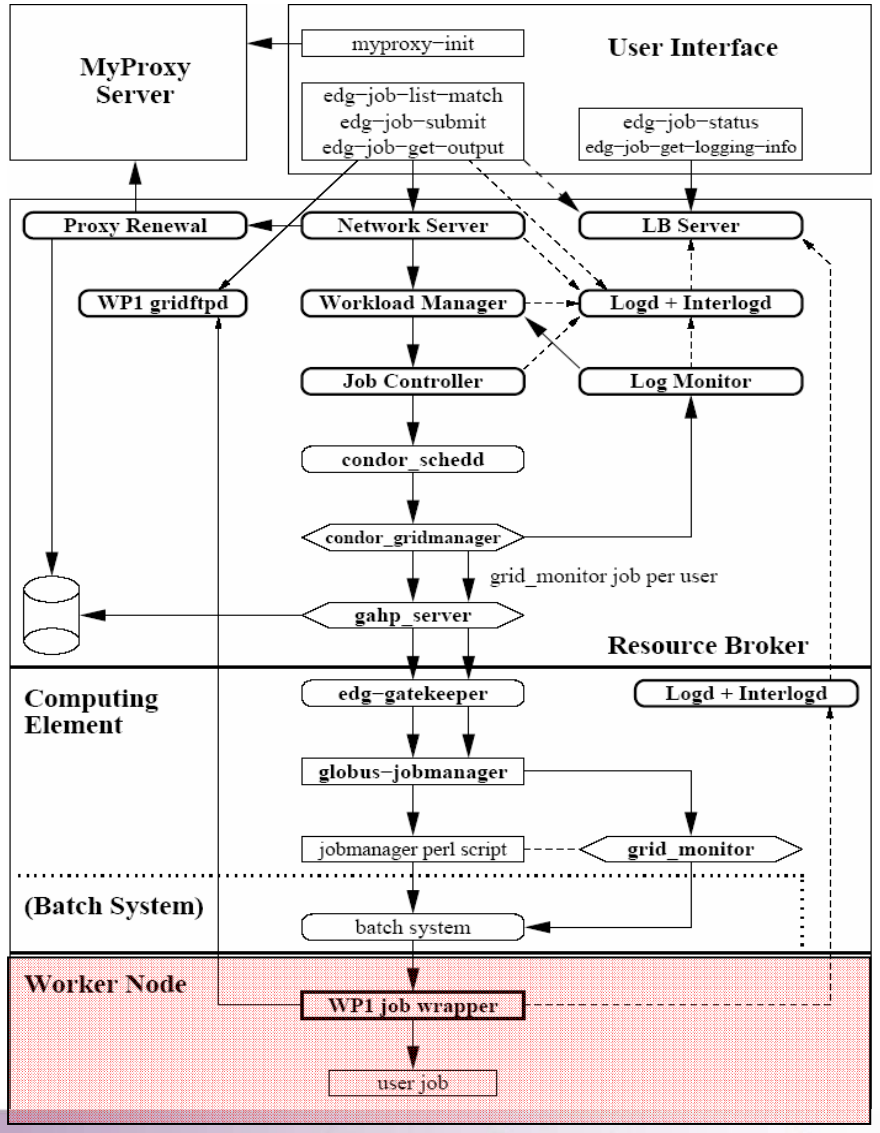
- handles the job execution on the available local farm worker nodes.
- Batch System consists of:
  - torque (formerly known as OpenPBS) resource manager .
  - maui job scheduler .

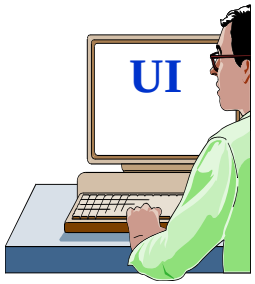


Worker Node

Worker nodes

- It is the host executing the job .
- A set of WNs managed by a CE constitues a computing cluster.
- A cluster MUST be homogeneous.
- is probably the simplest part of the Grid .
- The WN runs the job wrapper





RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Replica Location Server

Inform. Service

Job Status

submitted

waiting

ready

scheduled

running

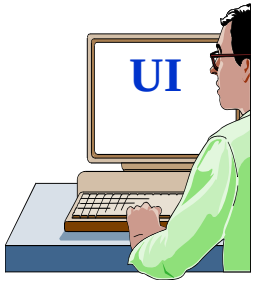
Input Sandbox files

"Grid enabled" data transfers/ accesses

Computing Element

Job

Storage Element



RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Replica Location Server

Inform. Service

Job Status

submitted

waiting

ready

scheduled

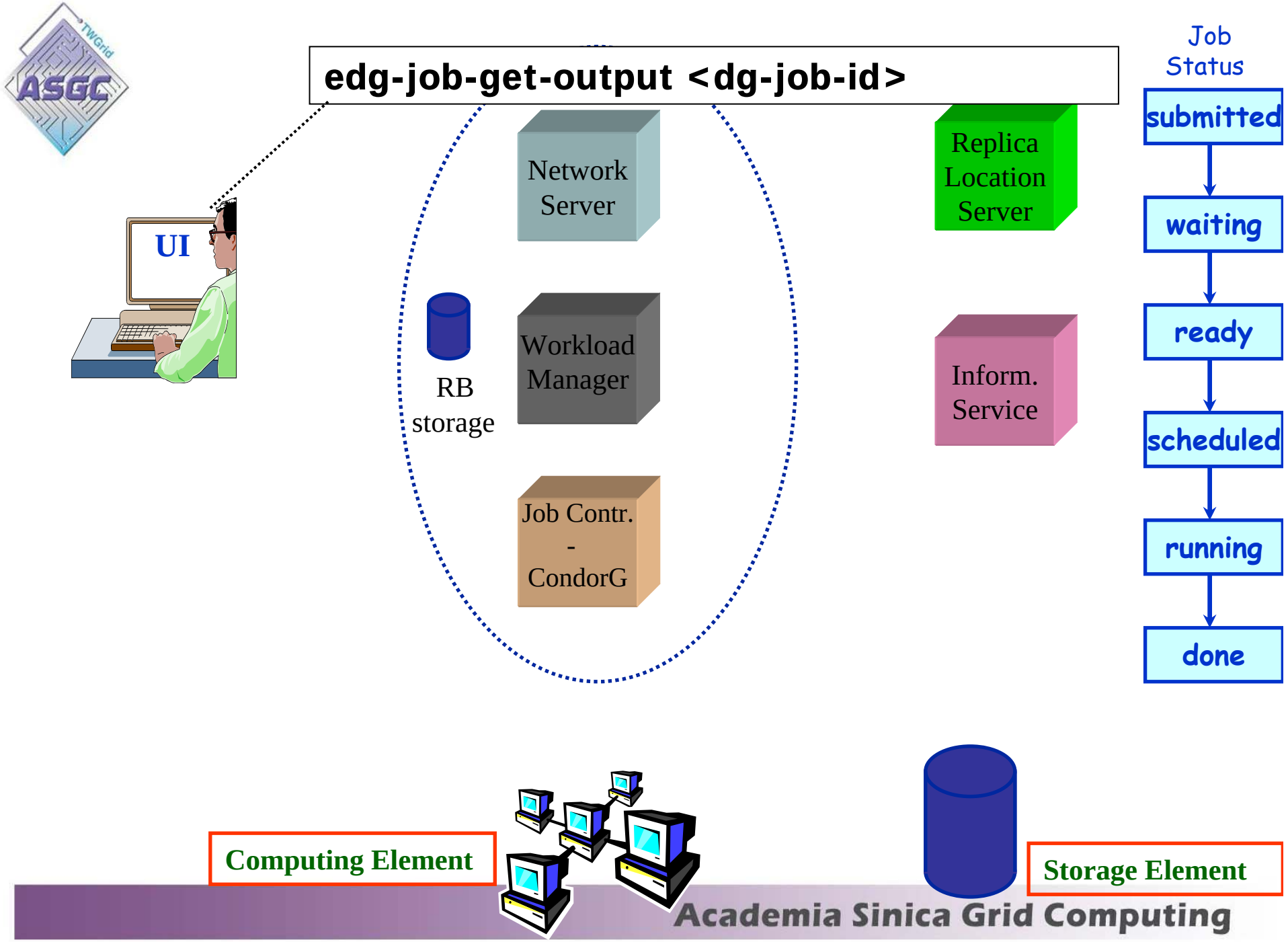
running

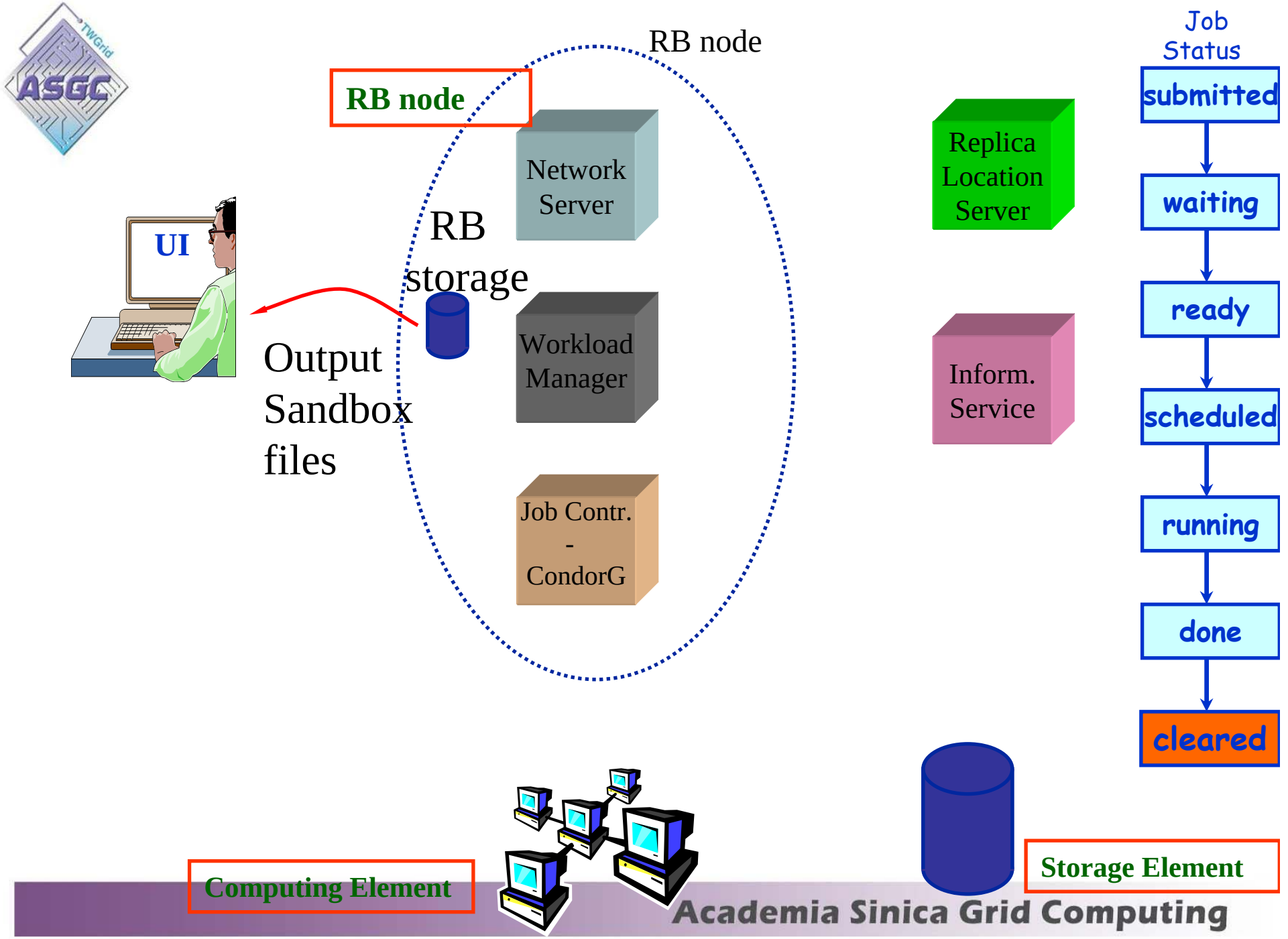
done

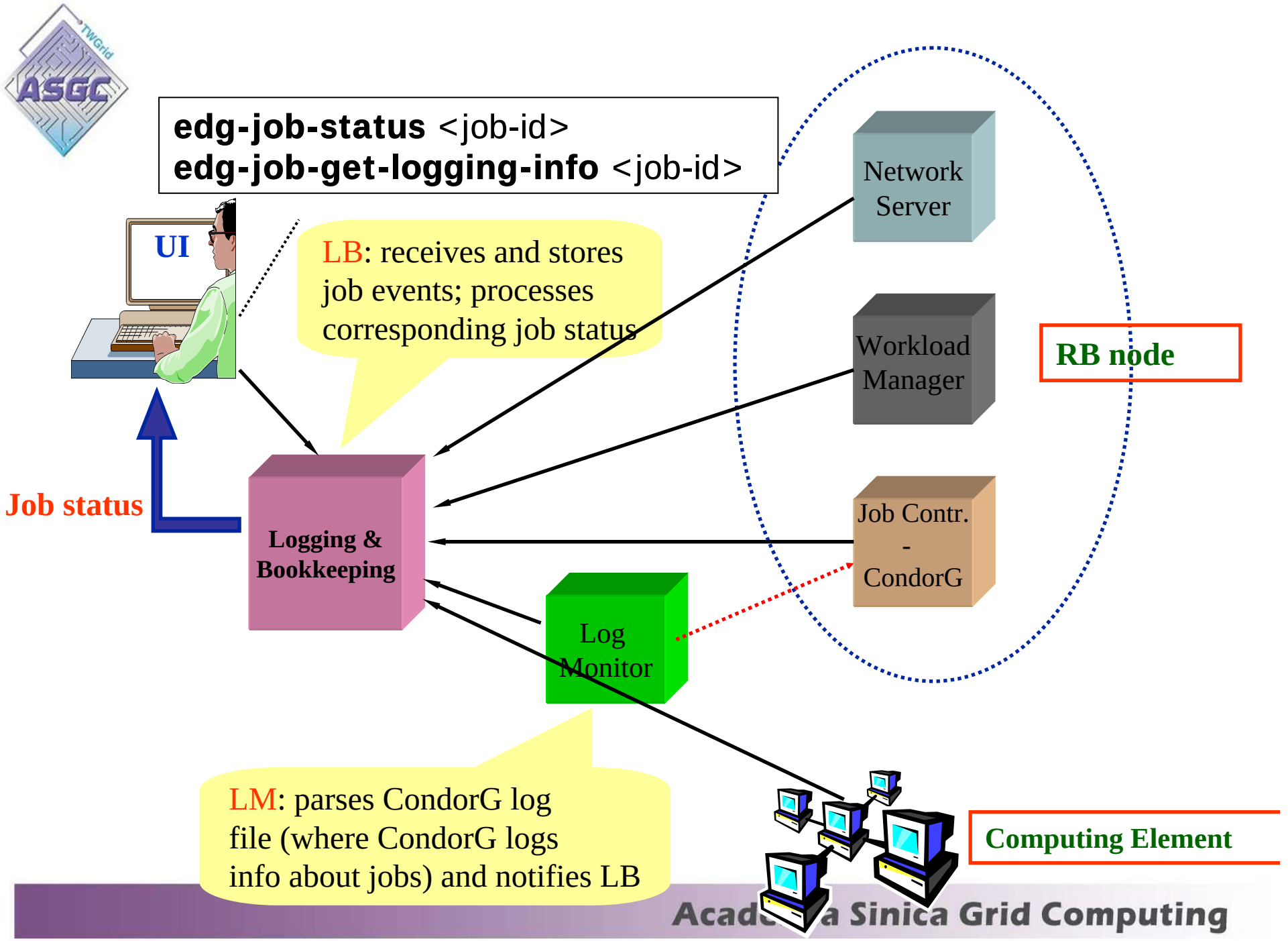
Output Sandbox files

Computing Element

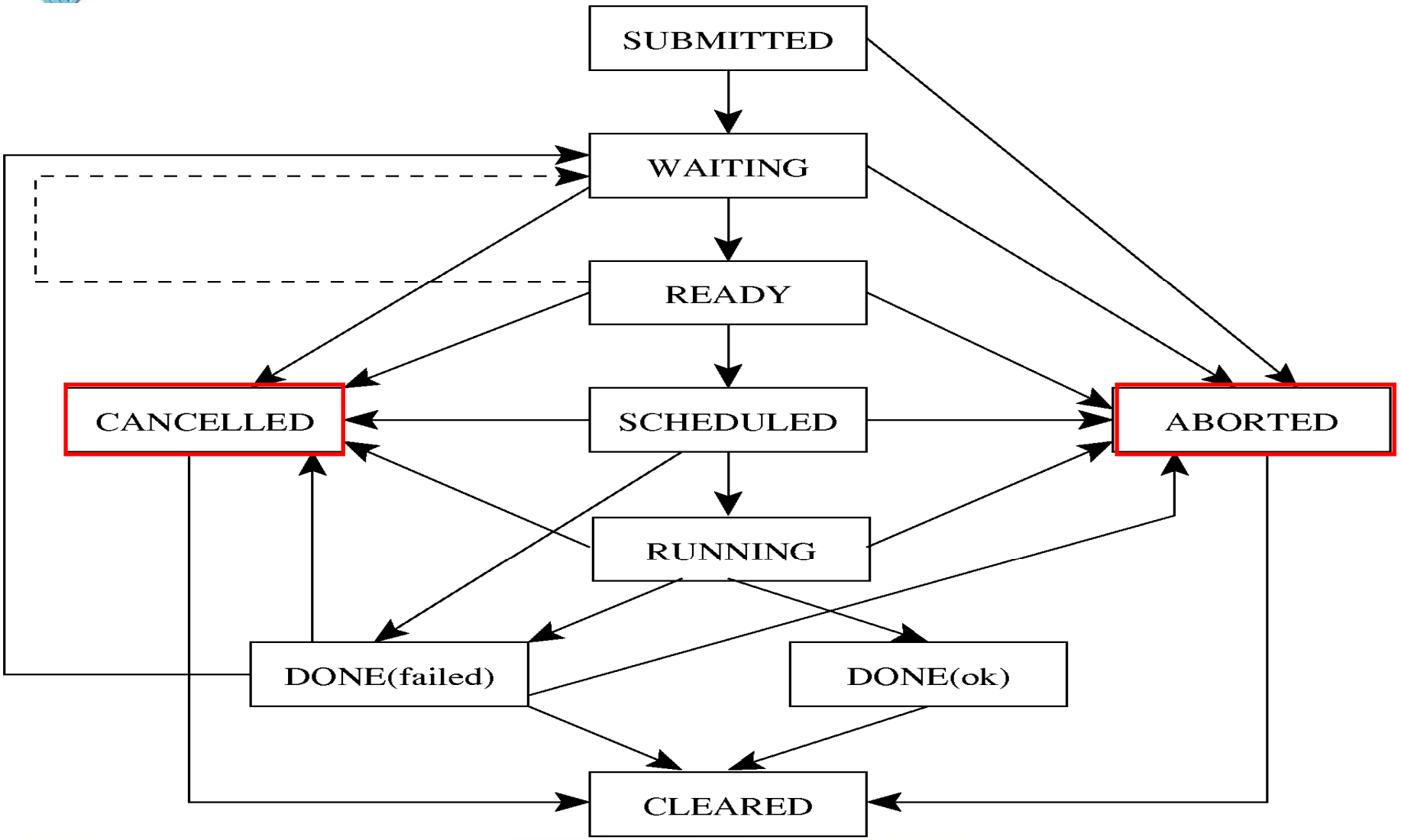
Storage Element


edg-job-get-output <dg-job-id>
UI
Network Server
Replica Location Server
RB storage
Workload Manager
Inform. Service
Job Contr. - CondorG
Job Status
submitted
waiting
ready
scheduled
running
done
Computing Element
Storage Element
Academia Sinica Grid Computing


RB node
RB node
Network Server
RB storage
Workload Manager
Job Contr. - CondorG
UI
Output Sandbox files
Replica Location Server
Inform. Service
Job Status
submitted
waiting
ready
scheduled
running
done
cleared
Computing Element
Storage Element
Academia Sinica Grid Computing


edg-job-status <job-id>
edg-job-get-logging-info <job-id>
UI
LB: receives and stores job events; processes corresponding job status
Network Server
Workload Manager
RB node
Job status
Logging & Bookkeeping
Job Contr. - CondorG
Log Monitor
LM: parses CondorG log file (where CondorG logs info about jobs) and notifies LB
Computing Element
Academia Sinica Grid Computing

Possible job states



Job resubmission

- If something goes wrong, the WMS tries to reschedule and resubmit the job.
- Maximum number of resubmissions:
  - **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

- WMS introduction
- Job Submission Sequence and WMS components
- **User Job submit**

Job Preparation

- Some issues :

➢ What are the characteristics of the job ?

➢ What are the computational requirements?

➢ What are the data requirements of the job?

➢ Are there any software dependencies?

Job Description Language (JDL)

- **Using a Job Description Language (JDL) to describe a job.**
- Based upon **Condor's *CLASSified ADvertisement*** *language (ClassAd)*
- A ClassAd syntax :

  <attribute> = <value>;

How to write a Job Description

- Here is a minimal job description

```
Executable= "/bin/echo";
Arguments = "Hello World!";
StdError  = "stderr";
StdOutput = "stdout";
OutputSandbox = {"stderr", "stdout"};
```

- We specified
  - The program to run and its arguments
    - Executable is already on (any) computing node
  - Directed the standard error and output streams to files
  - Told it what to do with the output

JDL: relevant attributes

- **Executable** (mandatory)
  - The command name
- **Arguments** (optional)
  - Job command line arguments
- **StdInput**, **StdOutput**, **StdError** (optional)
  - Standard input/output/error of the job
- **Environment**
  - List of environment settings needed by the job to run properly
- **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

JDL: relevant attributes

- **Requirements**
  - Job requirements on computing resources
  - Specified using attributes of all the GLUE attributes of the IS can be used.
  - If not specified, default value defined in UI configuration file is considered
  - Its value is a Boolean expression.
- **Rank**
  - Expresses preference (how to rank resources that have already met the Requirements expression)
  - Specified using attributes of resources published in the Information Service
  - If not specified, default value defined in the UI configuration file is considered

JDL: relevant attributes

- **InputData**
  - Refers to data used as input by the job: these data are published in the Replica Location Service (RLS) and stored in the SEs)
  - LFNs and/or GUIDs
- **DataAccessProtocol**
  - The protocol or the list of protocols which the application is able to speak with for accessing *InputData* on a given SE
- **OutputSE**
  - RB uses it to choose a CE that is compatible with the job and is close to SE

JDL: important notes

- Input and output sandboxes are intended for relatively small files (few megabytes).
- Large input files or generating large output files should insteadly read from or write to SE.

Other 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
  - Very useful for debug purposes

Job submission

$ grid-proxy-init

```
Your identity:/C=TW/O=AS/OU=CC/CN=Horng-Liang
Shih/Email=hlshih@gate.sinica.edu.tw
Enter GRID pass phrase for this identity:
Creating proxy ..................................................... Done
Your proxy is valid until: Sun Mar 12 16:03:30 2006
```

$ edg-job-submit -o id.txt -vo dteam HelloWorld.jdl

```
The job has been successfully submitted to the Network Server.
Use edg-job-status command to check job current status. Your job
      identifier (edg_jobId) is:

- https://lcg00124.grid.sinica.edu.tw:9000/QUMY4Dxg4TVVLvCaDDd2KA

The edg_jobId has been saved in the following file:
/home/hlshih/JSexercise1/id.txt
```

==========

Checking the status

$ edg-job-status -i id.txt

OR

$ edg-job-status https://lcg00124.grid.sinica.edu.tw:9000/QUMY4Dxg4TVVLvCaDDd2KA

```
**********************************************************************
    BOOKKEEPING INFORMATION:

    Status info for the Job :
    https://lcg00124.grid.sinica.edu.tw:9000/QUMY4Dxg4TVVLvCaDDd2KA
    Current Status:     Done (Success)
    Exit code:          0
    Status Reason:      Job terminated successfully
    Destination:        lcg00125.grid.sinica.edu.tw:2119/jobmanager-lcgpbs-
    dteam
    reached on:         Sun Mar 12 04:30:41 2006
**********************************************************************
```

Getting the Output

```
$ edg-job-get-output -i id.txt –dir $PWD
Retrieving files from host: lcg00124.grid.sinica.edu.tw ( for
    https://lcg00124.grid.sinica.edu.tw:9000/QUMY4Dxg4TVVLvCaDDd2KA )

**********************************************************************************
    JOB GET OUTPUT OUTCOME
    Output sandbox files for the job:
    - https://lcg00124.grid.sinica.edu.tw:9000/QUMY4Dxg4TVVLvCaDDd2KA
    have been successfully retrieved and stored in the directory:
    /home/hlshih/hlshih_QUMY4Dxg4TVVLvCaDDd2KA
**********************************************************************************

  $ ls -l  /home/hlshih/hlshih_QUMY4Dxg4TVVLvCaDDd2KA
  total 4
  -rw-r--r--    1 hlshih hlshih         0 Mar 12 04:54 stderr
  -rw-r--r--    1 hlshih hlshih        22 Mar 12 04:54 stdout
```

Reference

Job submit

- explains step-by-step how to submit your job
  https://edms.cern.ch/document/498081/1.0
- Job Description language How To.
  http://server11.infn.it/workload-grid/docs/DataGrid-01-TEN-0102-0_2-Document.pdf

Resource Broker

- Resource Broker Achitecture and APIs
  http://server11.infn.it/workload-grid/docs/20010613-RBArch-2.pdf

WMS

- WP1 Workload Management Software - Administrator and User Guide.
  http://server11.infn.it/workload-grid/docs/DataGrid-01-TEN-0118-1_2.pdf
- WP1 internal documents - more complete list of documents
  http://server11.infn.it/workload-grid/internal-documents.html