Enabling Grids for E-science in Europe

www.eu-egee.org

Soumission de jobs

E.Fede (CNRS/LAPP)

Tutorial Utilisateurs EGEE
26 Septembre 2007

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

Plan

- Workload Management System d'EGEE
  - Généralités
- Soumission de Job sur EGEE (LCG, Nordu grid…)
  - Le cycle de vie du job
  - Mécanisme de sandbox
  - Préparation d'un job
  - Exemple de base
  - Langage de description des jobs (JDL)
  - Quelques commandes
- Ordonnancement
- Nouveau WMS gLite
- Exemples de jobs
  - MPI

Le WMS

Enabling Grids for E-science in Europe

- Les utilisateurs soumettent leurs jobs via le **Workload Management System**
- Le but du WMS est d'ordonnancer, distribuer et manager les ressources d'une grille de calcul.
- Il permet aux utilisateurs de :
  - Soumettre leurs jobs
  - Connaître leurs statuts
  - Récupérer les sorties
- Le WMS essaye d'optimiser l'utilisation des ressources de la grille

Sous-ensembles du WMS

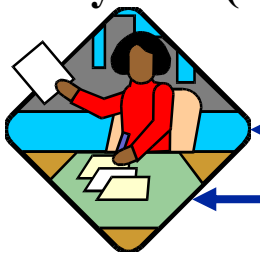
- WMS est composé de différentes parties:

1. **User Interface** (UI) : Point d'entrée pour les utilisateurs
2. **Resource Broker** (RB) : L'ordonnanceur, responsable de trouver LA meilleure ressource où le job pourra être exécuté
3. **Job Submission Service** (JSS) : Le système de soumission
4. **Information Index** (BDII) : Un serveur qui collecte les informations sur l'état de la grille pour approvisionner le RB en information.
5. **Logging and Bookkeeping** services (LB) : Base de données sur l'état des jobs
6. **Computing Element** : Element local (site) de job management

Job Submission Scenario

Replica Location Service (RLS)

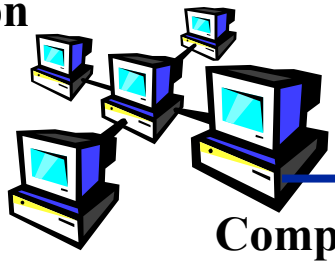
Information System (IS)



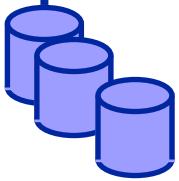
Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

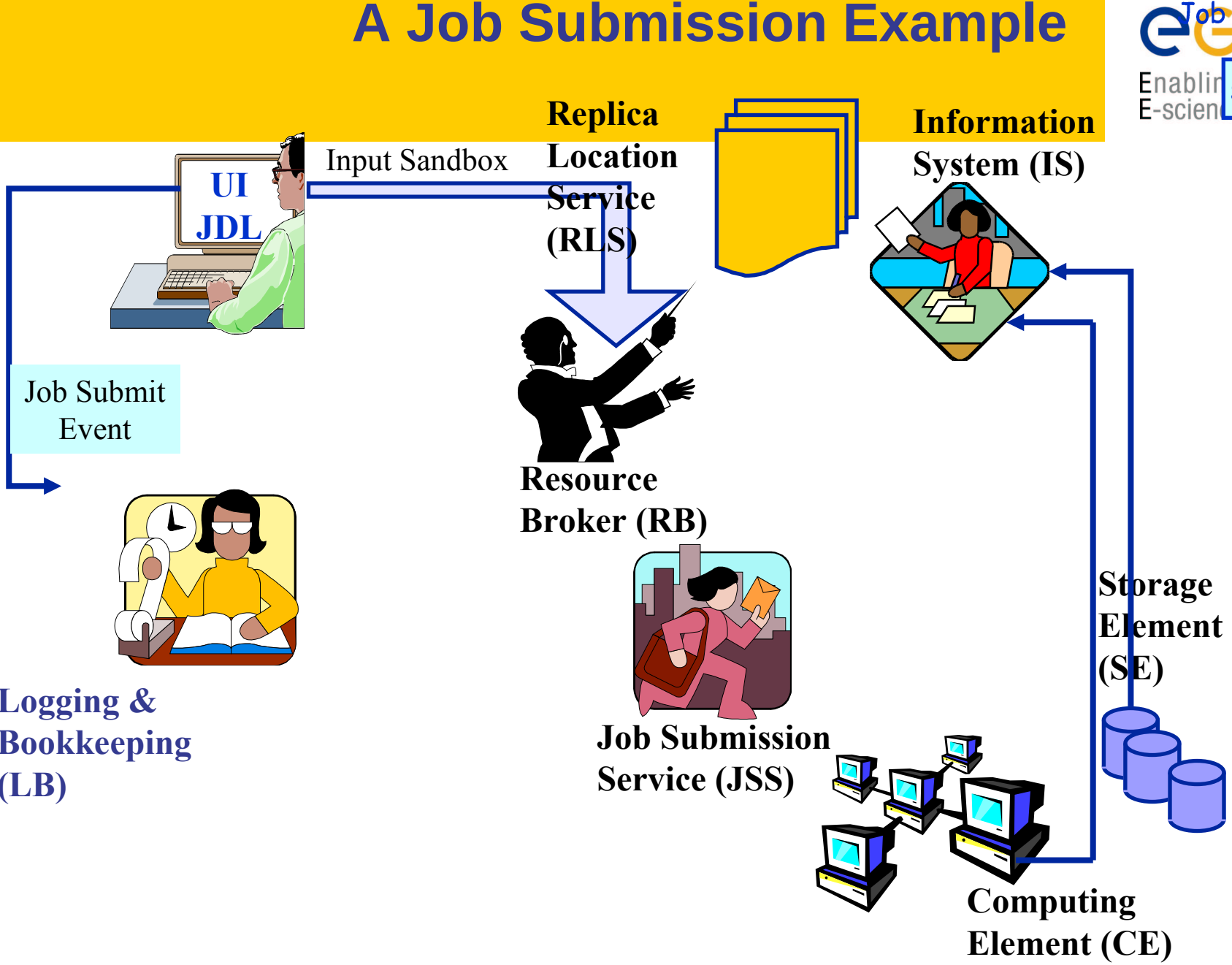


Computing Element (CE)



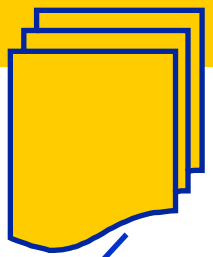
Storage Element (SE)

A Job Submission Example



A Job Submission Example

Replica Location Service (RLS)

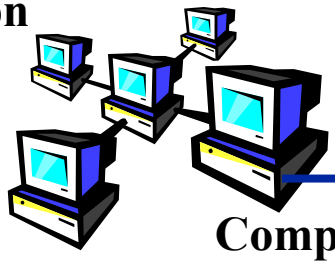


Information System (IS)

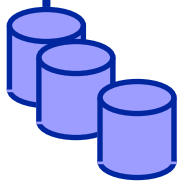
Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)



Computing Element (CE)



Storage Element (SE)

Job Status

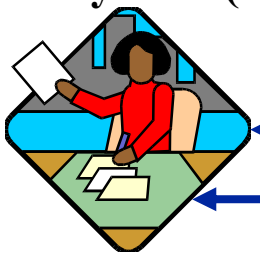
submitted

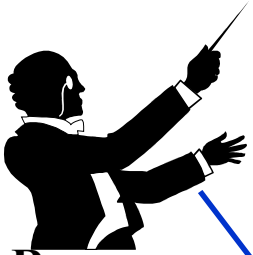
waiting

A Job Submission Example

Replica Location Service (RLS)

Information System (IS)



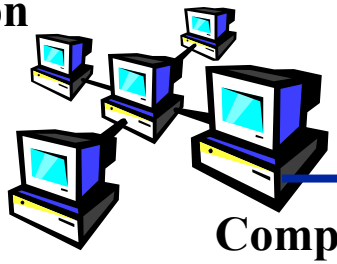


Resource Broker (RB)

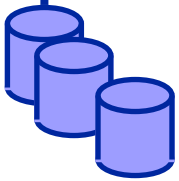
Logging & Bookkeeping (LB)

Job Status

Job Submission Service (JSS)



Computing Element (CE)



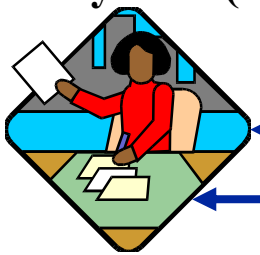
Storage Element (SE)

Job Status

A Job Submission Example

Replica Location Service (RLS)

Information System (IS)



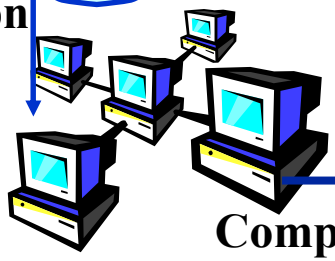
Resource Broker (RB)

Logging & Bookkeeping (LB)

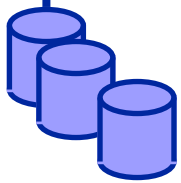
Job Status

Job Submission Service (JSS)

BrokerInfo

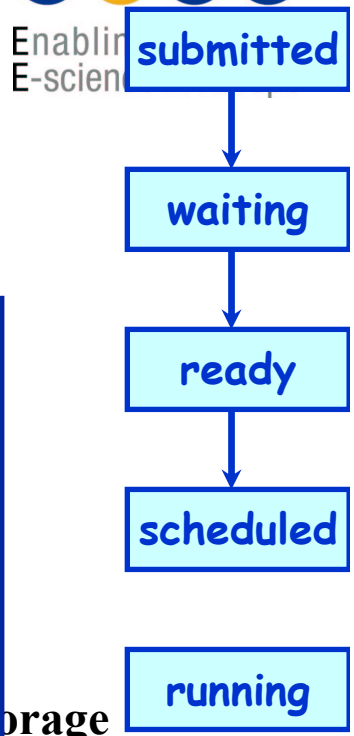


Computing Element (CE)



Storage Element (SE)

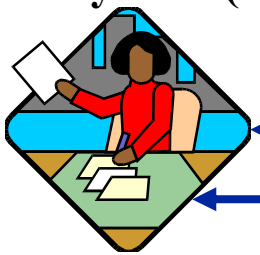
Job Status



A Job Submission Example

Replica Location Service (RLS)

Information System (IS)

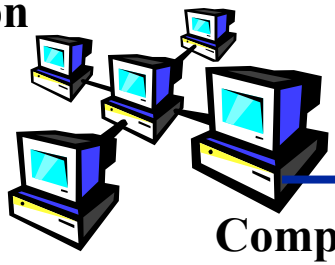


Resource Broker (RB)

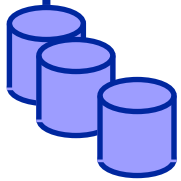
Input Sandbox

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

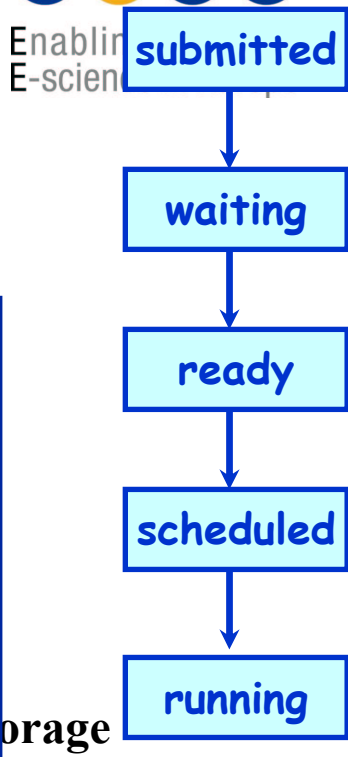


Computing Element (CE)



Storage Element (SE)

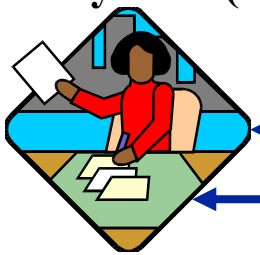
Job Status



A Job Submission Example

Replica Location Service (RLS)

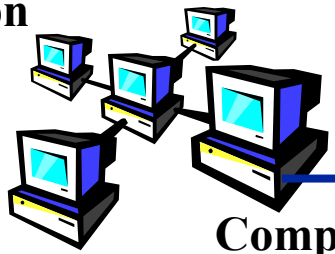
Information System (IS)



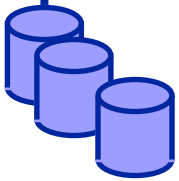
Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

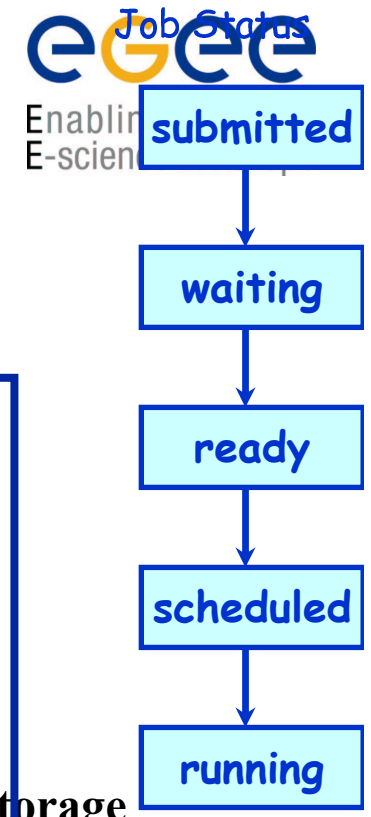


Computing Element (CE)



Storage Element (SE)

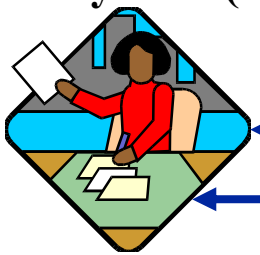
Job Status



A Job Submission Example

Replica Location Service (RLS)

Information System (IS)

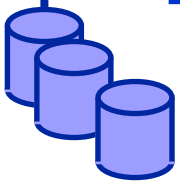


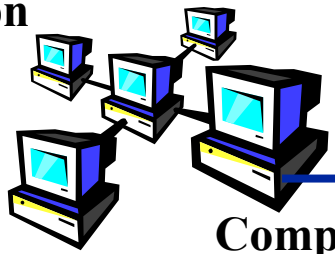
Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

Storage Element (SE)

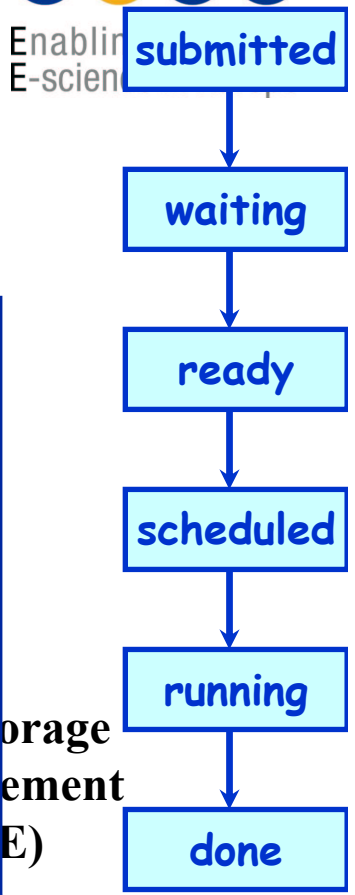




Computing Element (CE)

Job Status

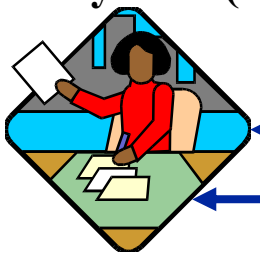
Job Status



A Job Submission Example

Replica Location Service (RLS)

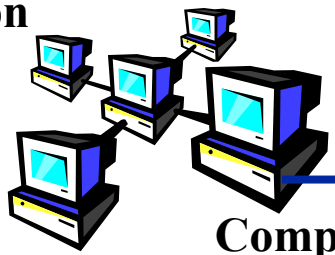
Information System (IS)



Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)



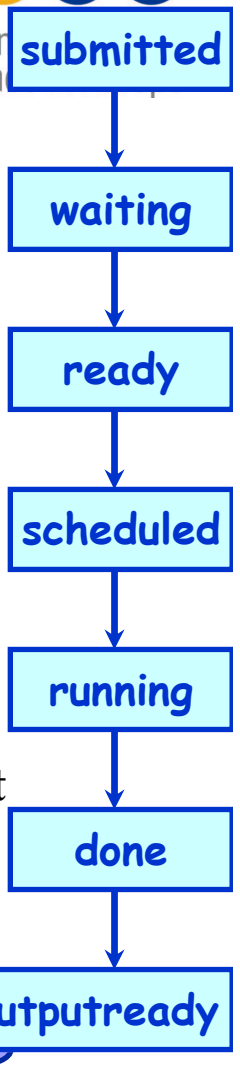
Computing Element (CE)

Storage Element (SE)

Job Status

Output Sandbox

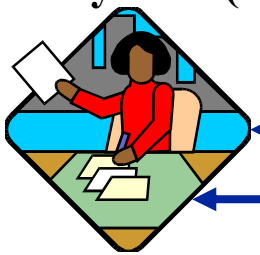
Job Status



A Job Submission Example

Replica Location Service (RLS)

Information System (IS)



Output Sandbox

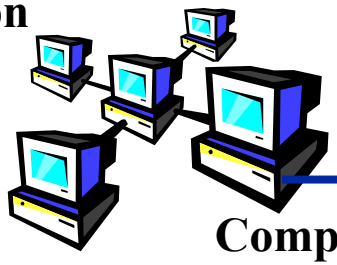
Resource Broker (RB)

Job Status

Logging & Bookkeeping (LB)

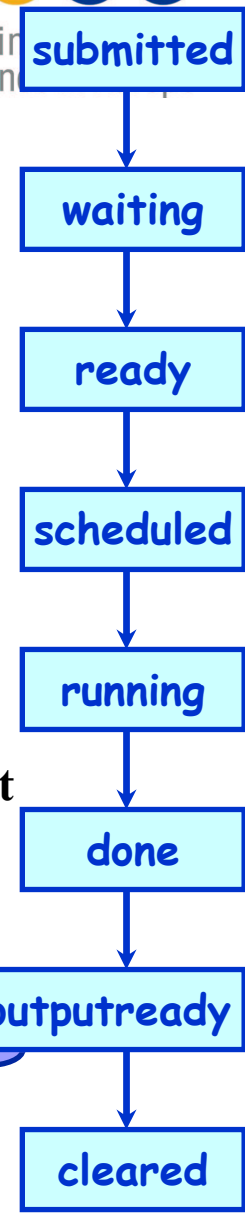
Job Submission Service (JSS)

Storage Element (SE)



Computing Element (CE)

Job Status



Mécanisme de sandbox

Enabling Grids for E-science in Europe

- Permet le transport de fichiers en même temps que le job
  - Inputsandbox
    - Mécanisme permettant de transporter des fichiers (exécutables, données,..) avec le job
    - UI ⇨ RB ⇨ CE ⇨ WN
    - Limitée en taille
  - Outputsandbox
    - Mécanisme permettant de récupérer des fichiers (exécutables, données, output,..) du job
    - WN ⇨ CE⇨ RB⇨UI
    - Limitée en taille
    - Utilisée notamment pour récupérer les stdout et stderr

Préparation d'un job

Enabling Grids for E-science in Europe

- Pour pouvoir soumettre un job on doit :
  - Donner une description de celui-ci
    - Quel programme?
    - Quelles données?
    - Quels softs, quels OS, quels besoins spécifiques?
  - Avoir un programme autant que possible:
    - Exécutable sur une plate-forme inconnue
    - Portable autant que possible
    - Sans liens absolus, path spécifiques et autres liens « rigides »
  - Avoir des données entrantes (pas nécessaire)

Exemple de Job Soumission

- Créer un proxy : *voms-proxy-init –voms VO*
  - Entrer son mot de passe de certificat
  - On récupère un proxy
- Soumet le job: *edg-job-submit mytest.jdl*
  On récupère du système un identifiant de job unique (JobId)
- Requête de status : *edg-job-status JobId*
  Pour émettre une requête au système de L&B afin de connaître l'état du job
- Quand le job entre dans l'état "OutputReady" on peut récupérer les output : *edg-job-get-output JobId*

  Et le système retourne le nom d'un répertoire temporaire (sur l'UI) où se trouvent nos fichiers de sortie

edg-job-submit <jdl_file>
edg-job-status <job_Id>
edg-job-get-output <job_Id>
edg-job-cancel <job_Id>
edg-job-xxxxxx

glite-job-submit <jdl_file>
glite-job-status <job_Id>
glite-job-output <job_Id>
glite-job-cancel <job_Id>
glite-job-xxxxxx

La soumission

Enabling Grids for E-science in Europe

```
fede@lapp-ui01:~/GRID/JDL >voms-proxy-init --voms dteam
Enter GRID pass phrase:
Your identity: /O=GRID-FR/C=FR/O=CNRS/OU=LAPP/CN=Eric Fede
Cannot find file or dir: /home1/fede/.glite/vomses
Creating temporary proxy ..................................... Done
Contacting  lcg-voms.cern.ch:15004 [/DC=ch/DC=cern/OU=computers/CN=lcg-voms.cern.ch] "dteam" Done
Creating proxy ..................................... Done
Your proxy is valid until Fri Sep 21 00:14:30 2007


fede@lapp-ui01:~/GRID/JDL >edg-job-submit HelloWorld.jdl

Selected Virtual Organisation name (from proxy certificate extension): dteam
Connecting to host lapp-rb01.in2p3.fr, port 7772
Logging to host lapp-rb01.in2p3.fr, port 9002


**********************************************************************************************
                        JOB SUBMIT OUTCOME
 The job has been successfully submitted to the Network Server.
 Use edg-job-status command to check job current status. Your job identifier is:

 - https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg

**********************************************************************************************
**********************************************************************************************
```

JobId

Connaître le statut

Enabling Grids for E-science in Europe

fede@lapp-ui01:~/GRID/JDL >edg-job-status https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg

****
BOOKKEEPING INFORMATION:

Status info for the Job : https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg
Current Status: Scheduled
Status Reason: Job successfully submitted to Globus
Destination: mars-ce2.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-24hr
reached on: Thu Sep 20 10:16:10 2007
****

fede@lapp-ui01:~/GRID/JDL >edg-job-status https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg

****
BOOKKEEPING INFORMATION:

Status info for the Job : https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg
Current Status: Done (Success)
Exit code: 0
Status Reason: Job terminated successfully
Destination: mars-ce2.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-24hr
reached on: Thu Sep 20 10:17:54 2007
****

Récupérer les sorties

Enabling Grids for E-science in Europe

fede@lapp-ui01:~/GRID/JDL >edg-job-get-output https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg

Retrieving files from host: lapp-rb01.in2p3.fr (for https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg)

****

JOB GET OUTPUT OUTCOME

Output sandbox files for the job:
- https://lapp-rb01.in2p3.fr:9000/XNSXcZdnNOKcWaDPZQxreg
have been successfully retrieved and stored in the directory:
/scratch/lcg/fede_XNSXcZdnNOKcWaDPZQxreg

****

fede@lapp-ui01:~/GRID/JDL >more /scratch/lcg/fede_XNSXcZdnNOKcWaDPZQxreg/stdout.log
Hello World

Job Description Language (JDL)

Enabling Grids for E-science in Europe

- Basé sur Condor's *CLASSified ADvertisement language (ClassAd)*
- Mapping entre attributs et expressions
- ClassAd est un langage extensible.
- Construit sur des séquences d'attributs.
- Application de ClassAds : match-making

```
Executable = "/bin/echo";
Arguments = "Bonjour";
StdError = "stderr.log";
StdOutput = "stdout.log";
OutputSandbox = {"stderr.log", "stdout.log"};
```

- On spécifie ici :
  - Le programme et ses arguments
  - On définie les STDOUT et STDERR
  - On dit que faire des outputs

Différents types d'attributs

- Deux grandes catégories d'attributs:
  - Ceux concernant le job

    Définissent le job lui-même.
  - Ceux concernant les ressources
    - Utilisés par l'ordonnanceur (RB) pour définir la ressource utilisée pour l'exécution du job.
    - Permettent de définir les caractéristiques de calcul requises

      Utilisées par l'utilisateur pour construire des demandes précises (librairies disponibles,etc)

      Ces attributs sont définis à l'aide du préfixe "other."
    - Définissent les caractéristiques liées aux données

      Ce sont : les données entrantes , l'élément de stockage où les données sont prises ou bien mises, les protocoles,…

Attributs de jobs

Enabling Grids for E-science in Europe

- **JobType** (facultatif)
  - Défini le type du job (normal, mpich,…)
- **Executable** (obligatoire)
  - Le nom de la commande à exécuter.
- **Arguments** (facultatif)
  - Les arguments de la commande à exécuter.
- **StdInput, StdOutput, StdErr** (facultatif)
  - La définition des entrées/sorties/erreurs standards.
- **Environment** (facultatif)
  - Ensemble de valeurs liées à l'environnement d'exécution.
- **InputSandbox** (facultatif)
  - Liste des fichiers se trouvant sur l'UI qui seront transférés avec le job.
  - Ces fichiers seront copiés sur le CE cible.
- **OutputSandbox** (facultatif)
  - Liste des fichiers générés par le job qui seront accessibles via l'output sandbox
- **RetryCount** (facultatif)
  - Nombre de fois ou le RB resoumet le job dans le cas ou quelque chose se passe mal (du point de vue de la grille)

Attributs de ressources

- **Requirements (pré-requis)**
  - Besoins du job vis à vis des ressources de calcul.
  - Sont spécifiés à partir des attributs qui sont définis dans le système d'information de la grille.
  - S'ils ne sont pas définis dans le JDL, ce sont les valeurs définies par défaut dans l'UI qui sont utilisées
    - Default: other.GlueCEStateStatus == "Production" (Les ressources utilisables devront absolument présenter l'attribut demandé)
- **Rank**
  - C'est une préférence, concernant l'ordre de rangement des ressources qui remplissent les « requirements »
  - Si non spécifié, utilise la valeur définie sur l'UI
    - Default: - other.GlueCEStateFreeCPUs (le plus grand nombre de CPU libres)

Attributs de ressources (données)

- **InputData** (facultatif)
  - Réfère les données utilisées en entrée par le job .
  - PFNs et/ou LFNs.

- **DataAccessProtocol** (seulement si InputData est spécifié)
  - Le protocole ou la liste de protocoles de communication utilisables par l'application pour accéder aux *INPUTDATA.*

Exemple de fichier JDL

Enabling Grids for E-science in Europe

```
Executable = “Test.sh”;
StdError = “stderr.log”;
StdOutput = “stdout.log”;
InputSandbox = {“/home/fede/script/Test.sh”};
OutputSandbox = {“stderr.log”, “stdout.log”};
InputData = “lfn:testbed0-00019”;
DataAccessProtocol = “gridftp”;
Requirements = other.Architecture==“INTEL” && \
               other.OpSys==“LINUX” && other.FreeCpus >=4;
Rank = “other.GlueCEStateFreeCPUs”;
```

Commande de Job Soumission

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

  -r le job est directement envoyé par le RB sur le CE identifié par <*res_id*>

  -c utilise le fichier de configuration <*config file*> afin de surcharger les valeurs par défaut de l'UI

  -o renvoi l'identifiant de job dans <*output file*>

  --vo la virtual organization sous laquelle le job doit être exécuté

Autres commandes du WMS UI

- **`edg-job-list-match`**

  Liste les ressources correspondantes à la description du job.

  Permet de connaître le résultat de l'ordonnancement sans soumettre le job.

- **`edg-job-cancel`**

  Annule le job.

- **`edg-job-status`**

  Affiche l'état du job.

- **`edg-job-get-output`**

  Récupère le contenu de OutputSandbox

- **`edg-job-get-logging-info`**

  Affiche des informations sur les divers états pris par le job tout le long de son existence.

  Utilisée essentiellement pour le debugging.

Le système d'ordonnancement

Enabling Grids for E-science in Europe

- L'ordonnanceur est le cœur du WMS.
- Il doit trouver la meilleure ressource où le job sera exécuté.
- Il interagit avec le système de management des données et le système d'information.
- Le CE choisi par le RB doit remplir les conditions requises par le job (environnement d'exécution, accès aux données,…)
- Si plus de un CE satisfait les contraintes, c'est celui proposant la meilleure condition pour le requirement rank qui est choisi.

Soumission de job

Enabling Grids for E-science in Europe

- Trois types de scénario sont possibles.
  - En utilisant le Ressources Broker, c'est à dire en excluant le cas ou l'on soumet le job directement au job manager d'un site.

Scénario 1: Soumission directe

- Le job est directement soumis au CE (spécifié par le paramètre –r de la commande edg-job-submit).
- Le RB n'effectue aucune recherche de ressources.
- Peut (et généralement c'est le cas) générer des erreurs si on utilise l'attribut InputData.
- L'utilisateur est responsable de la cohérence de son job.

Job Soumission, sans InputData

Enabling Grids for E-science in Europe

Scénario 2: Soumission de job sans Requirements liés aux données

- Aucun CE, ni données entrantes (input data) sont précisés
- Le RB utilise l'algorithme de recherche de ressources qui comporte deux phases:
  - Le RB contact le système d'information afin de déterminer quels CE peuvent satisfaire les demandes.
  - Si plus de deux sont proposés alors on utilise l'attribut rank pour faire le choix.

Job soumission avec des données sur la grille

Enabling Grids for E-science in Europe

Scénario 3: Le CE n'est pas spécifié et l'on a des données en entrée

- Le RB contacte le service de management des données afin de déterminer quels SE peuvent satisfaire les besoins (quels SE possèdent les données requises)
- Le RB cherche le meilleur compromis (best effort) entre :
  - Les CE où l'utilisateur à le droit de soumettre ces jobs.
  - Les SE qui ont été déterminés préalablement.
- La stratégie du RB est de soumettre les jobs au plus près des données.
- Les deux phases suivantes sont les mêmes que dans le scénario précédent (uniquement pour les CE qui satisfont les requirements de données)
  - Requirements check
  - Rank computation

WMS gLite

Enabling Grids for E-science in Europe

- Nouvelle version du middelware
  - Un nouveau CE (en discussion)
  - Le nouveau RB (gLite WMS)
    - Tout ce que l'on a vu précédemment reste valide
    - Nouvelles fonctionnalitées
      - Bulk submission (soumission par lot)
      - Job paramétrique
      - Pull submission (Dépend du nouveau composent CE)
      - Performances accrues
      - …

WMS gLite

Enabling Grids for E-science in Europe

- Délégation de proxy **WMProxy**
  - C'est une autre « façon » d'utiliser l'ordonnanceur
  - C'est notamment grâce à ce service que de nouvelles fonctionnalités sont disponibles au niveau de l'ordonnanceur.
  - Les performances accrues.

glite-wms-job-submit <jdl_file>

glite-wms-job-status <jdl_file>

glite-wms-job-list-match <jdl_file>

glite-wms-job-cancel <job_Ids>

glite-wms-job-output <job_Ids>

WMS gLite

Enabling Grids for E-science in Europe

- Nouveaux JobType
  - Bulk submission : soumission par lot
    - JobType = « collection »
    - Permet l'envoi et la gestion de plusieurs jobs comme un seul
  - Type DAG
    - JobType=« DAG » (direct acyclic graph)
    - Permet de définir des dépendances entre jobs
  - Type paramétrique
    - JobType = « Parametric »
    - Permet l'envoi et la gestion d'un job avec plusieurs jeux de paramètres

Exemple de job de type MPI

Enabling Grids for E-science in Europe

```
Type = "Job";
JobType = "MPICH";
NodeNumber = 16;
Executable    = "MPItest.sh";
Arguments     = "mpiok 16";
InputSandbox = {"/home/fede/SCRIPT/ok.c","/home/fede/SCRIPT/MPItest.sh"};
StdOutput     = "std.out";
StdError      = "std.err";
OutputSandbox = {"std.out","std.err","info"};
Requirements =  other.GlueCEInfoTotalCPUs > 16;
```

Informations utiles

Enabling Grids for E-science in Europe

- ClassAd

  https://www.cs.wisc.edu/condor/classad
- gLite 3.0 User Guide
  - https://edms.cern.ch/file/722398/1.1/gLite-3-UserGuide.pdf
- JDL attributes specification for WM proxy
  - https://edms.cern.ch/document/590869/1
- WMProxy quickstart
  - http://egee-jra1-wm.mi.infn.it/egee-jra1-wm/wmproxy_client_quickstart.shtml
- WMS user guides
  - https://edms.cern.ch/document/572489/1