Grid & Cloud Computing

3. Grilles de calcul

Grid Computing : analogie électrique (1)

- Réseau électrique (*Power Grid*) :
  - Des centrales qui produisent l’électricité
  - Un réseau haute-tension pour l'acheminement

Grid Computing : analogie électrique (2)

- Grille informatique :
  - Des ressources informatiques : calculs, stockage
  - Un réseau pour relier ces ressources : internet
  - Des services

Grid Computing : définition

«*A system that coordinates resources which are not subject to centralized control, using standard, open, general-purpose protocols and interfaces to deliver nontrivial qualities of service*»

Ian Foster, 2002.

Grid Computing : caractéristiques

- Coordination de ressources non centralisée
  - Répartition des ressources hétérogènes à **grande échelle**
  - Utilisateurs hétérogènes (structures, domaines, logiciels...)
- Protocoles et interfaces ouverts et standardisés
  - Interopérabilité des grilles
  - Organisme de standardisation : **Open Grid Forum** (http://www.ogf.org)
- Différents niveaux de Qualité de Service
  - QoS : débit, disponibilité, sécurité, ram, cpu...
  - Variabilité de la QoS importante dûe à l'hétérogénéité

Différents type de grilles

- *Desktop grids* (ou *peer to peer*)

- **Grilles de production (eScience Grids)**

- Grilles de recherche

- Grilles d'Entreprise

Grille de production

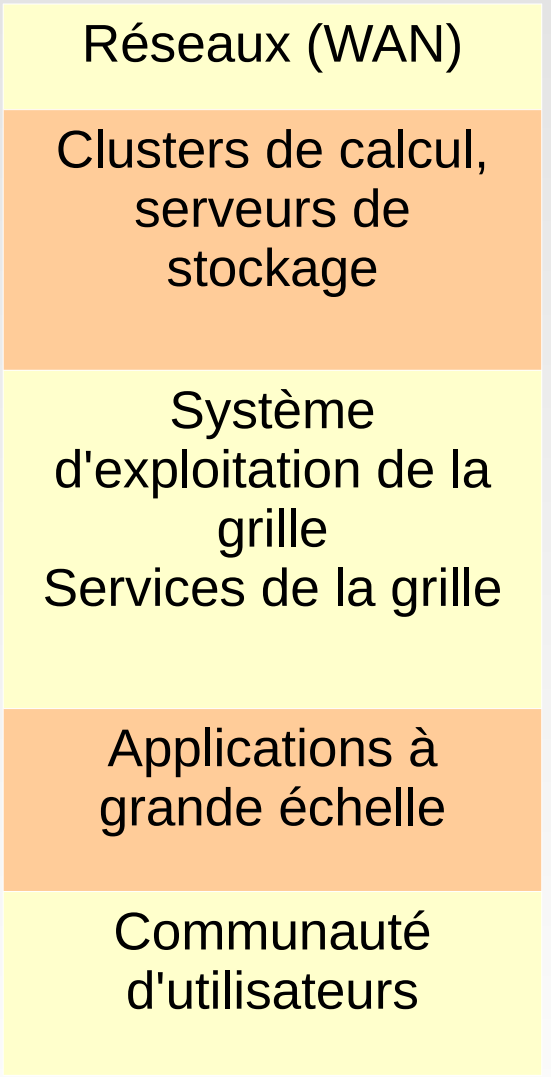
➡Grille distribuée sur des sites distants interconnectés par des réseaux à haut débit

➡Des ressources importantes de calcul et de stockage

➡Gérées par un système d’exploitation commun (middleware)

➡Offrant un ensemble de services permettant de déployer des applications à grande échelle

➡Pour servir des communautés d’utilisateurs



Middleware

- Logiciel d'exploitation de la grille qui s'interface entre les ressources informatique et les applications
  - Gestion des jobs : soumission, planification, ordonnancement
  - Gestion des données : stockage, transferts, réplication
  - Gestion de la sécurité
  - Gestion des utilisateurs

➡ Globus, gLite, ARC, Unicore, DIET, OAR, iRODS, EMI

Communauté d'utilisateurs

- Ensemble d'utilisateurs partageant un intérêt commun sur la grille
  - Une expérience
  - Des données
  - Des logiciels
  - Des ressources
- Collaboration à grande échelle
- Les communautés d'utilisateurs se retrouvent au sein d'Organisations Virtuelles (VO)

Open Grid Forum

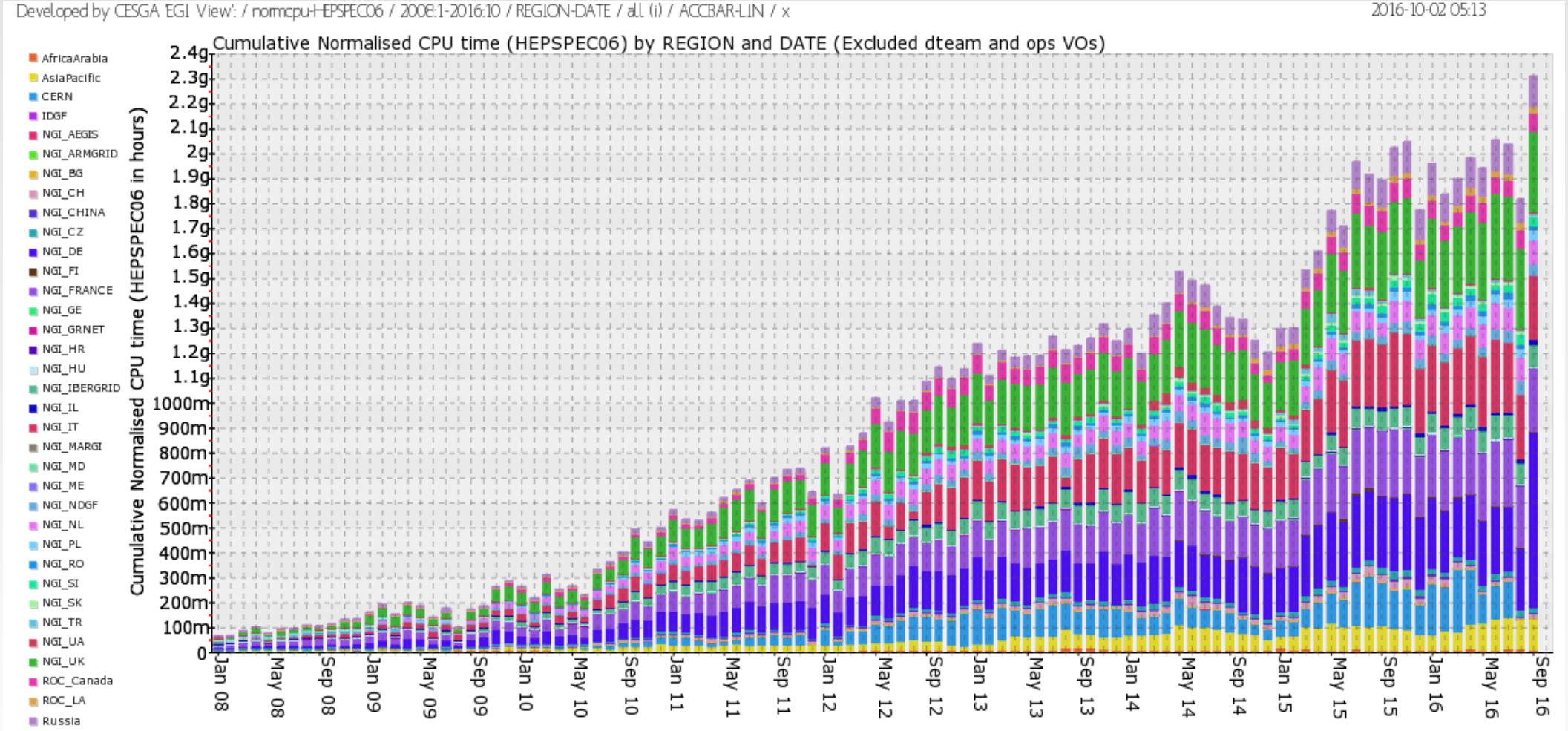
- Consortium regroupant des utilisateurs, des développeurs, des entreprises
- Recommandation de standards pour les grilles :
  - OGSA : *Open Grid Services Architecture*
  - OGSI : *Open Grid Services Infrastructure*
  - JSDL : *Job Submission Description Language*
  - GLUE : *Grid Laboratory Uniform Environment*
  - DRMAA : *Distributed Resource Management Application API*
  - SAGA : *Simple API for Grid Applications*

Infrastructure de grille européenne EGI

EGI : Historique

- Milieu des années 1990 : naissance du concept de grille
- Globus, développé par Ian Foster et Carl Kasselman (US)
- 1999 : CERN adopte le principe de grille pour traiter les données du futur LHC (LCG : LHC Computing Grid)
- 2001-2003 : grille de test DATAGRID
- 2004-2009 : EGEE 3 projets européens de 2 ans pour la mise en place d'une infrastructure de grille de production
- ~2007 : grille devient véritablement opérationnelle
- 2010 : lancement du LHC
- 2010 : EGI (European Grid Initiative)

Evolution consommation CPU

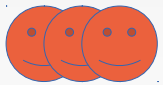


Source: http://accounting.egi.eu/

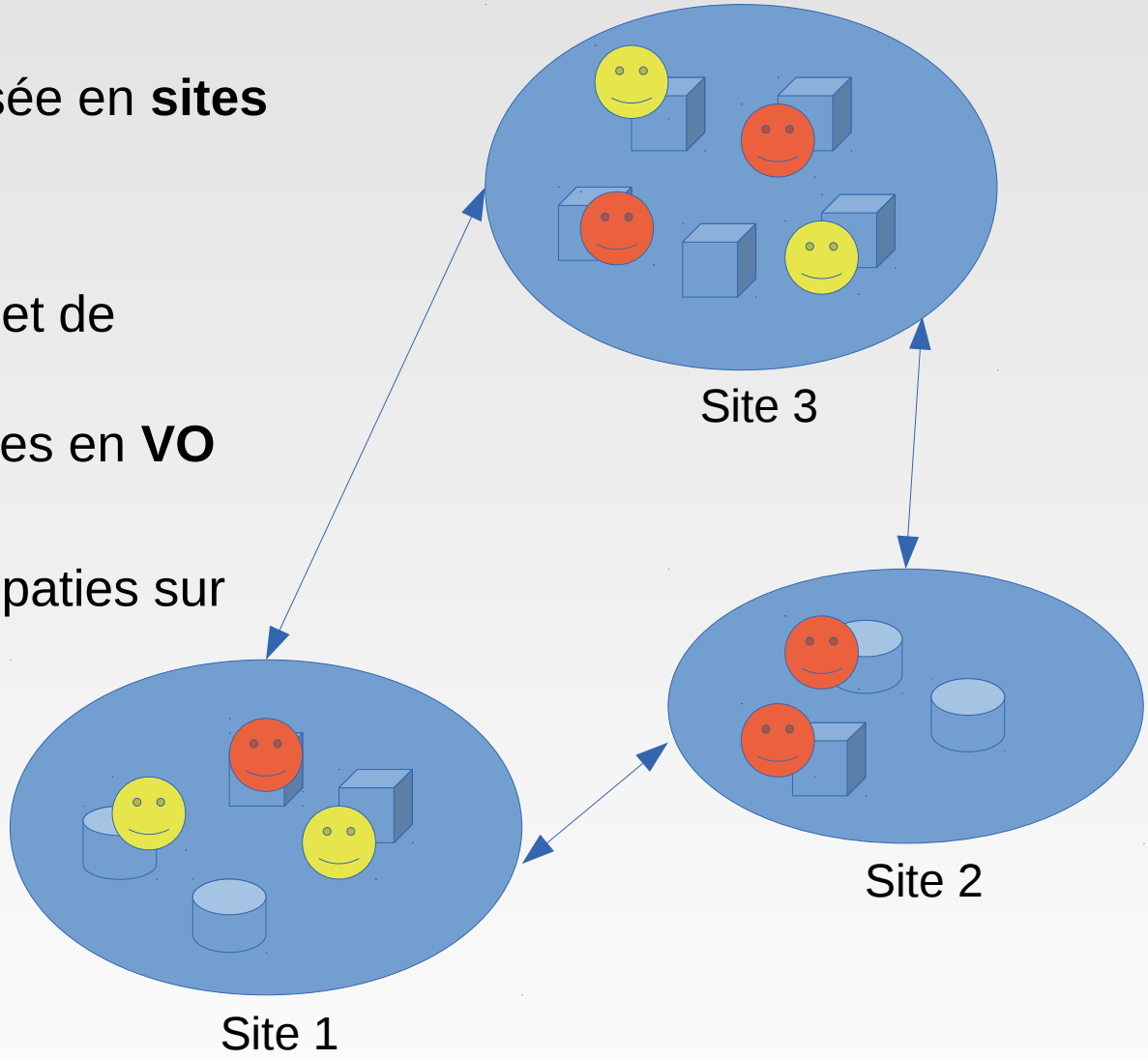
Organisations des ressources

- Ressources de calcul organisée en **sites**
  - Unité géographique
  - Unité administrative
  - Ensemble de ressources et de services
- Les utilisateurs sont organisées en **VO**
  - Unité thématique
  - Ressources identiques répaties sur plusieurs sites

VO1



VO2



Organisation virtuelle (VO)

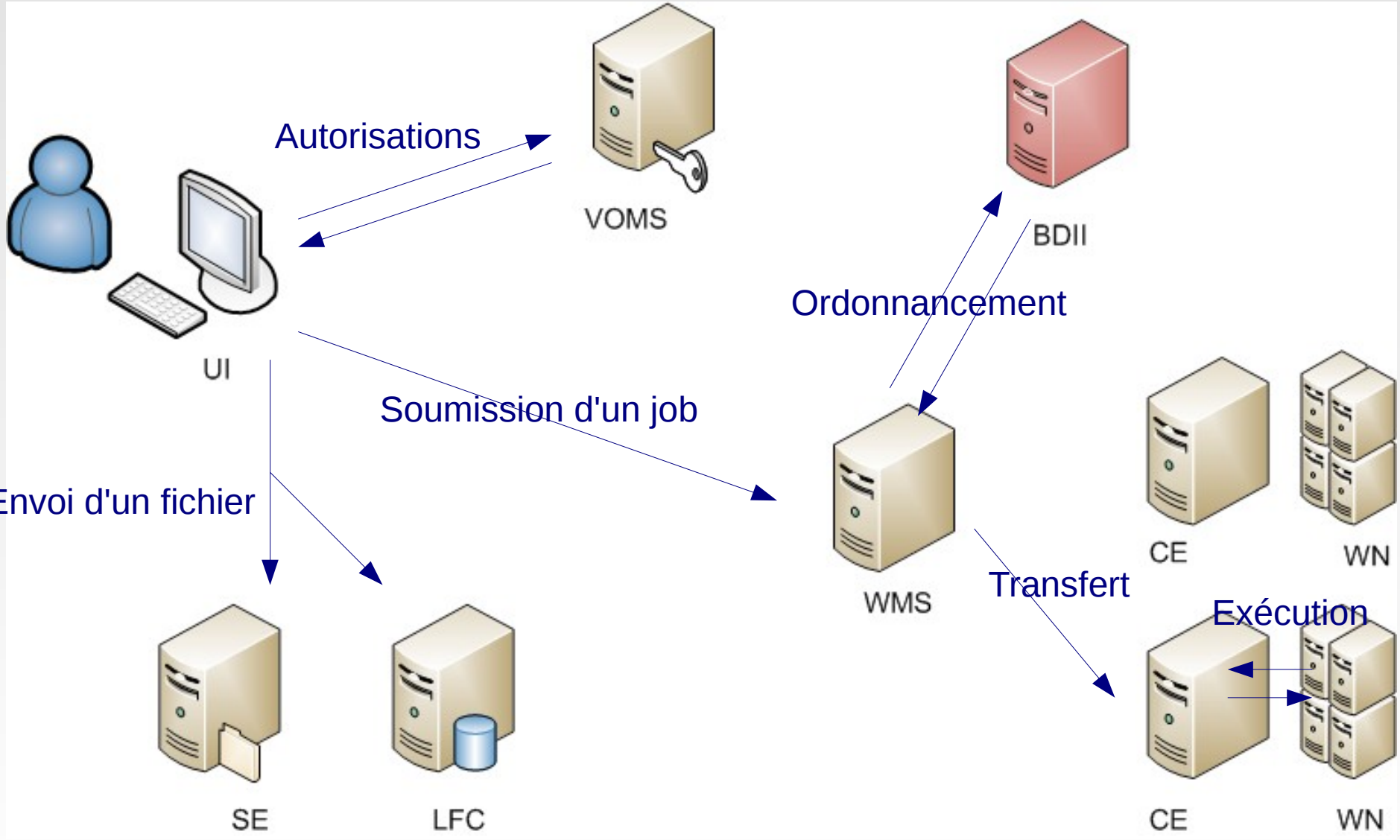
- Les utilisateurs sont regroupés par communautés
  - Sciences du vivant : Biomed
  - HEP : Alice, Atlas, Babar, CMS, D0, LHCb, …
  - Observation de la Terre : ESR
- Les autorisations sont fonction de l’Organisation Virtuelle
- Un administrateur par Organisation Virtuelle

  C’est le gestionnaire des utilisateurs de sa VO
- Les ressources se déclarent utilisables pour certaines VO
  - Affectation des ressources aux VO par l'administrateur d'un site

Éléments constitutifs principaux

CE	Computing Element	Ressources de calcul (cluster)
SE	Storage Element	Serveur de stockage
UI	User Interface	Interface d'accès à la grille
WMS	Workload Management Server	Système de gestion des jobs
LFC	LCG File Catalog	Catalogue des fichiers (pour une VO)
VOMS	Virtual Organization Membership Service	Gestion des utilisateurs d'une VO
BDII	Berkeley Database Information Index	Système d'information de la grille

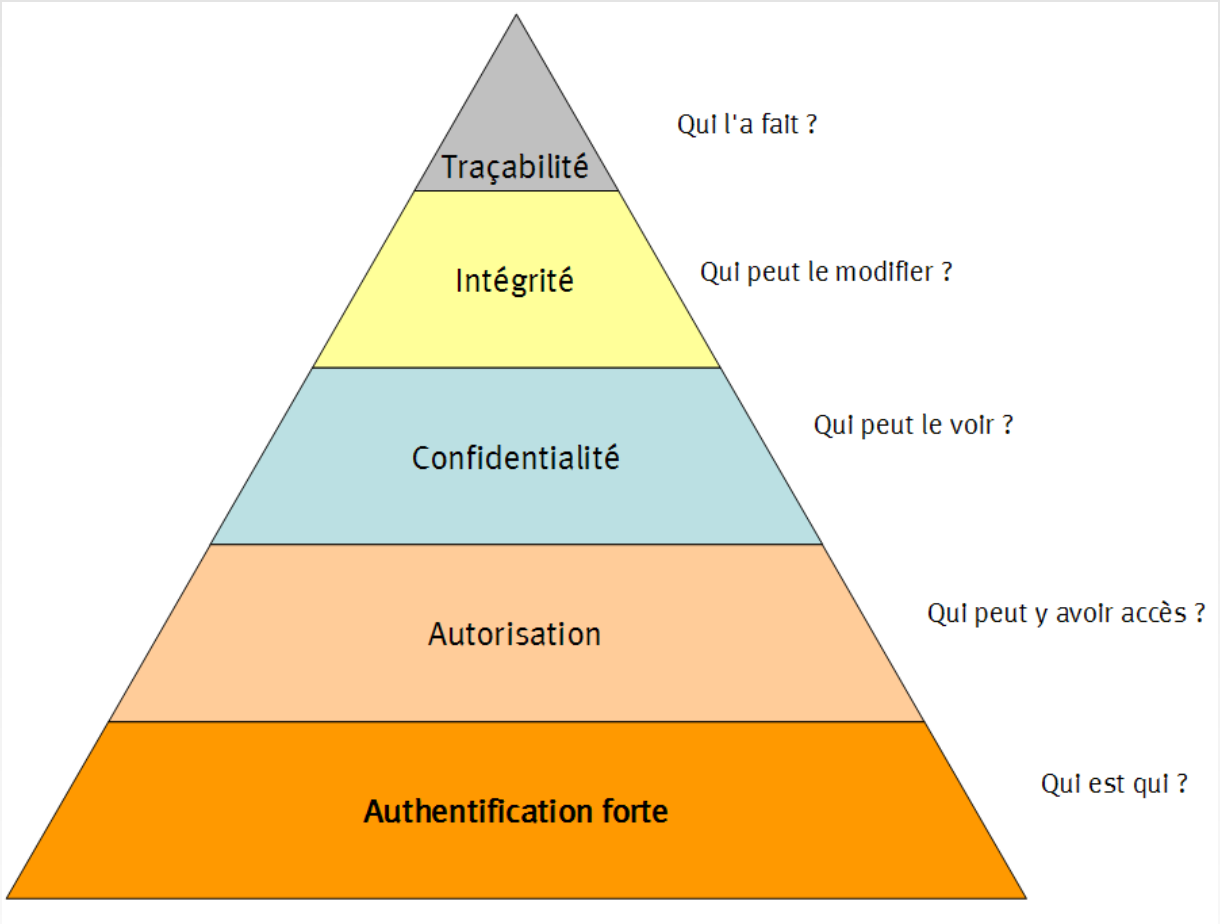
Fonctionnement général



Sécurité sur la grille

- Comme les ressources et les utilisateurs distribués et appartenant à des établissement différents, il faut garantir que seuls les utilisateurs autorisés peuvent y accéder
- Une infrastructure d'authentification est donc nécessaire.
- Les utilisateurs et les propriétaires des ressources doivent être protégés les uns des autres
- Les utilisateurs doivent assurer la sécurité de :
  - Leurs données
  - Leurs codes
  - Leurs messages

Sécurité



LOG

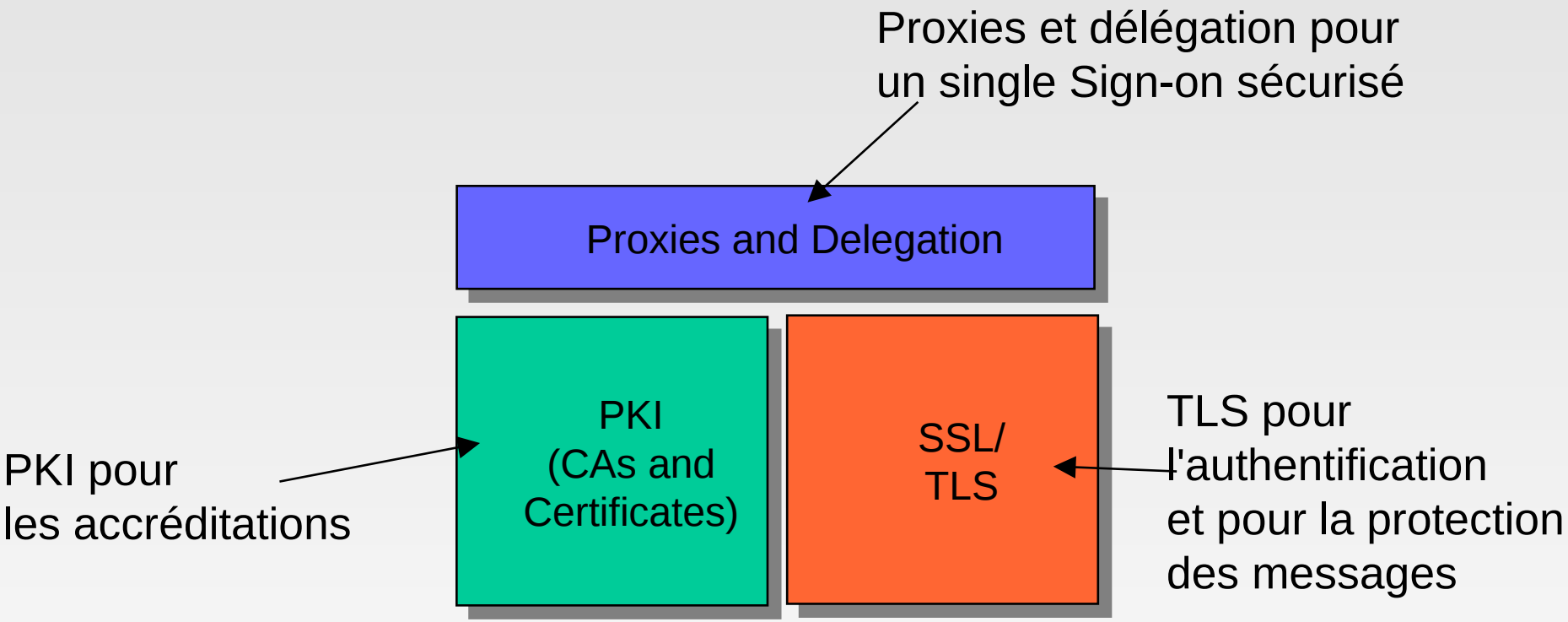
ACL

ACL

VOMS

PKI / GSI

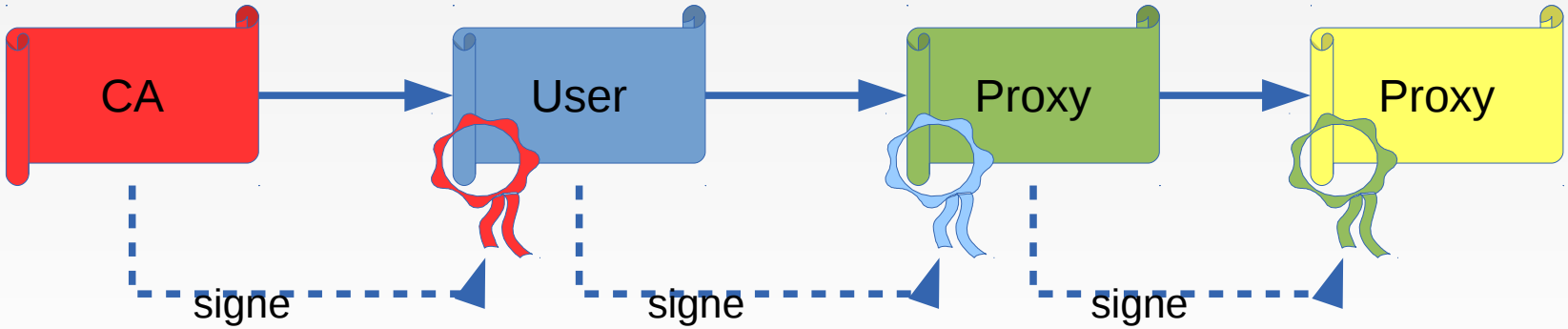
Grid Security Infrastructure (GSI)



PKI: Public Key Infrastructure,
SSL: Secure Socket Layer
TLS: Transport Level Security

Grid Proxy

- Authentification par certificats non adaptée sur la grille
  - **Délégation dynamique** : délégation de certains privilèges à une entité tierce, pour un temps limité (ex WMS)
  - **Entités dynamique** : les entités tierces ne sont pas connus à l'avance (ex CE)
  - **Authentifications répétées** : chaque opération d'authentification va demander le mot de passe à l'utilisateur...
- Proxy = certificat/clé privée dont l'émetteur et le signataire est un certificat user X.509 ou un certificat proxy
- Durée validité courte (12 heures à 36 heures max)

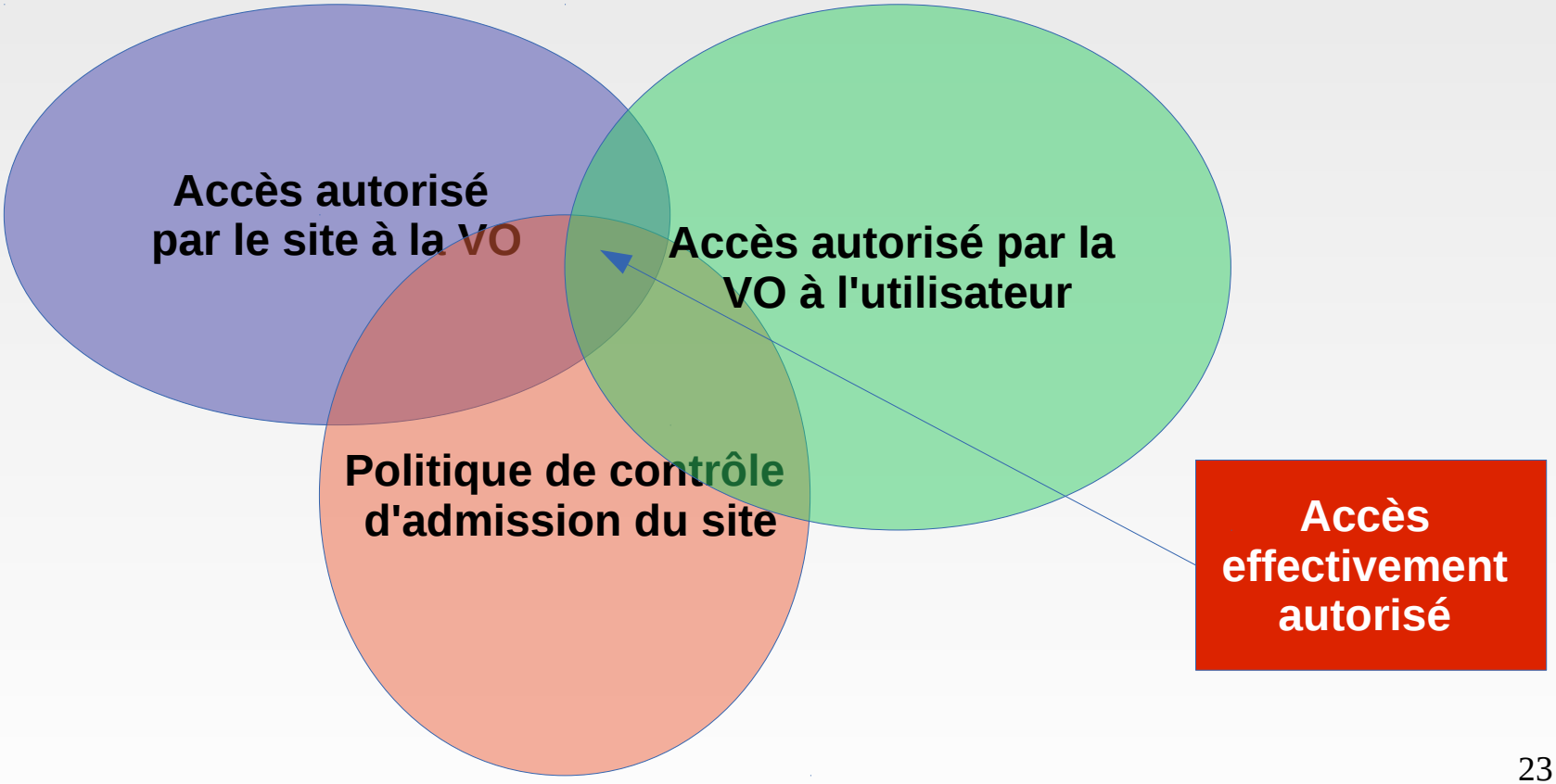


Certificat Proxy

```
$ openssl x509 -in /tmp/x509up_u6xx -text
Certificate:
    Data:
        Version: 3 (0x2)
        Serial Number: 163570064 (0x9bfe190)
        Signature Algorithm: sha1WithRSAEncryption
        Issuer: O=GRID-FR, C=FR, O=UBP, OU=CIRI, CN=Antoine Mahul
        Validity
            Not Before: Oct  3 15:55:02 2010 GMT
            Not After : Oct  4 16:00:02 2010 GMT
        Subject: O=GRID-FR, C=FR, O=UBP, OU=CIRI, CN=Antoine Mahul,
CN=proxy
        Subject Public Key Info:
            Public Key Algorithm: rsaEncryption
            RSA Public Key: (1024 bit)
                Modulus (1024 bit):
                    00:d5:42:59:8b:70:20:f2:1a:4f:b5:40:e4:bc:5b:
                    ...
                Exponent: 65537 (0x10001)
        X509v3 extensions:
            1.3.6.1.4.1.8005.100.100.5:
```

Gestion des autorisations : VOMS

- Un certificat grille ne donne pas l'accès aux ressources
- Les autorisations sont définies par les VO



Virtual Organization Membership Service (VOMS)

- Extension de la GSI pour gérer les autorisations
- Attributs du proxy avec les informations d'autorisation :
  - Appartenance à une (ou plusieurs) VO
  - Groupe, rôles et privilèges
- Chaque VO a une base de données des utilisateurs définissant les droits de chaque utilisateur
- L'utilisation contacte le serveur VOMS pour obtenir ses autorisation
- Le serveur VOMS Server retourne les informations d'autorisation et un certificat proxy

Création d'un proxy VOMS

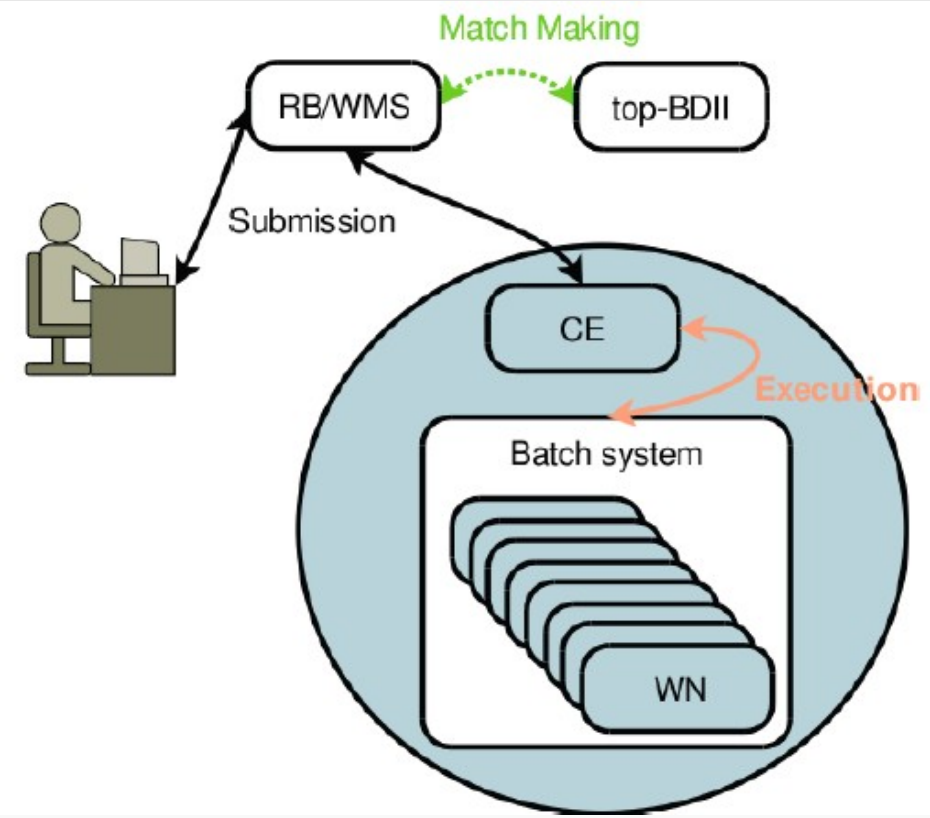
```
$ voms-proxy-init --voms auvergrid -valid 24:00
Enter GRID pass phrase:
Your identity: /O=GRID-FR/C=FR/O=UBP/OU=CIRI/CN=Antoine Mahul
Creating temporary proxy ..................................... Done
Contacting  cclcgvomsli01.in2p3.fr:15002
[/O=GRID-FR/C=FR/O=CNRS/OU=CC-LYON/CN=cclcgvomsli01.in2p3.fr] "auvergrid"
Done
Creatingproxy.................................................. Done
Your proxy is valid until Mon Oct  4 19:46:19 2010

$ voms-proxy-info -all
subject   : /O=GRID-FR/C=FR/O=UBP/OU=CIRI/CN=Antoine Mahul/CN=proxy
issuer    : /O=GRID-FR/C=FR/O=UBP/OU=CIRI/CN=Antoine Mahul
identity  : /O=GRID-FR/C=FR/O=UBP/OU=CIRI/CN=Antoine Mahul
type      : proxy
strength  : 1024 bits
path      : /tmp/x509up_u600
timeleft  : 23:57:08
=== VO auvergrid extension information ===
VO        : auvergrid
subject   : /O=GRID-FR/C=FR/O=UBP/OU=CIRI/CN=Antoine Mahul
issuer    : /O=GRID-FR/C=FR/O=CNRS/OU=CC-LYON/CN=cclcgvomsli01.in2p3.fr
attribute : /auvergrid/Role=NULL/Capability=NULL
timeleft  : 23:57:08
uri       : cclcgvomsli01.in2p3.fr:15002
```

Récapitulatif

- Installation des certificats sur l'UI :
  - Clé publique au format PEM : ~/.globus/usercert.pem
  - Clé privée au format PEM : ~/.globus/userkey.pem
    - Protégée par un mot de passe
    - Avec les droits de lecture uniquement pour le propriétaire

- Création d'un proxy : `voms-proxy-init --voms...`
- Infos sur un proxy : `voms-proxy-info -all`
- Destruction d'un proxy : `voms-proxy-destroy`

Jobs sur la grille



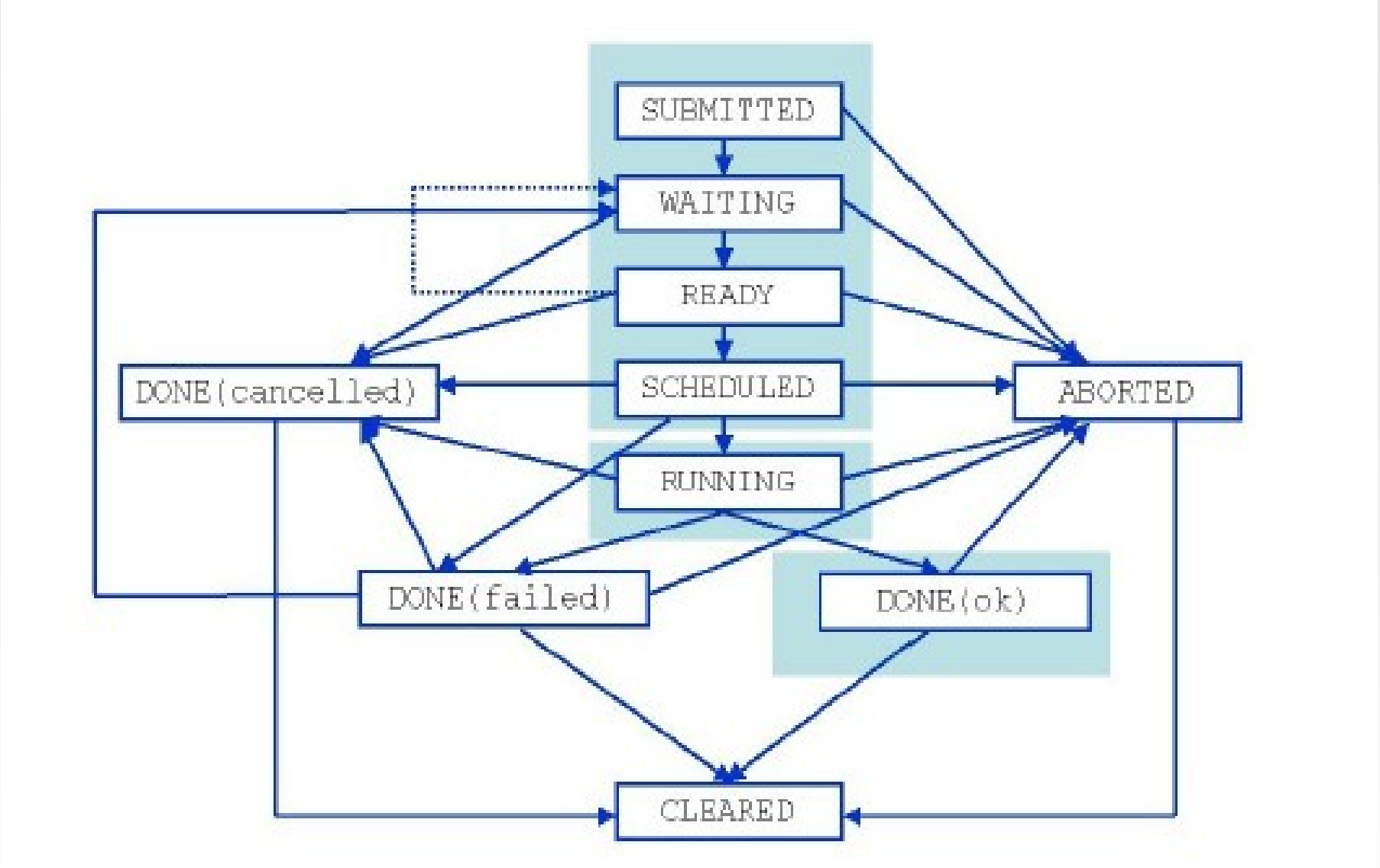
- Description du job
- Soumission du Job
- Allocation (Match Making)
- Exécution du job
- Suivi du Job

Workload Management System (WMS)

- *Job Scheduler* de la grille
  - Masque la complexité de la grille
  - Se charge de l'ordonnancement et de l'allocation des ressources (*Match Making*)
  - Permet de suivre le jobs jusqu'à la fin de son exécution
  - Est capable d'interagir avec les autres services de la grille et éventuellement avec d'autres grilles

- Redondance : plusieurs WMS par VO

Diagramme d'état d'un job



Description des états

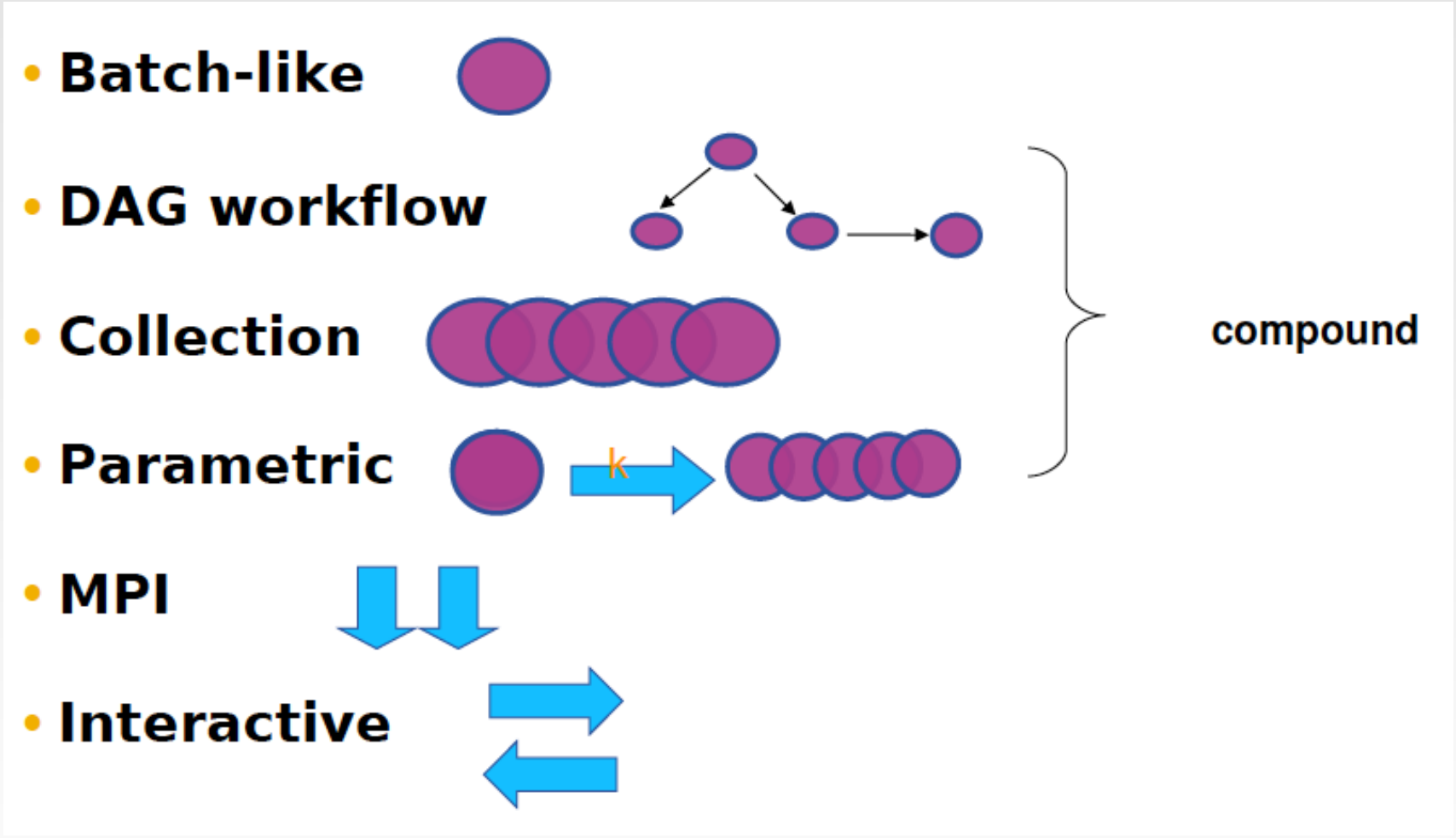
Etat	Description
Submitted	Job enregistré dans le LB
Waiting	Recherche des CE éligibles (« match making »)
Ready	Job envoyé au CE
Scheduled	Job en attente sur le CE
Running	Job en cours d'exécution sur un WN
Done	Job terminé (avec ou sans erreur)
Aborted	Job avorté par le middleware
Cleared	Job nettoyé (sorties récupérées par l'utilisateur)

Description d'un Job

- Langage de description : JDL (Job Description Language)
  - Type du job (simple, DAG, collection)
  - Exécutable et arguments
  - *Input Sandbox* : fichiers à envoyer avec le job (< 10 Mo)
  - *Output Sandbox* : fichiers à récupérer à la fin du job (< 10 Mo)
  - *Requirements* : exigences, contraintes

Remarque : l'exécutable doit être portable, sans lien vers des bibliothèques dynamiques, sans chemin absolu

Types de Jobs



Exemple 1 : Hello World

```
Executable = “/bin/echo”;
Arguments = “Hello World”;
StdError = “stderr.log”;
StdOutput = “stdout.log”;
OutputSandbox = {“stderr.log”, “stdout.log”};
```

- Pas de fichier d'entrée ni d'exécutable à transférer sur le WN
- Fichiers de sortie (transférés du WN vers l'UI après l’exécution)
  - sortie standard
  - erreur standard

Exemple 2 : script sh

```
Executable = “script.sh”;
Arguments = “arg1 arg2”;
StdError = “stderr.log”;
StdOutput = “stdout.log” ;
InputSandbox = {“script.sh”, “datafile1”}
OutputSandbox = {“stderr.log”, “stdout.log”} ;
RetryCount = 3 ;
```

- Fichiers d'entrée (transférés de l'UI vers le WN avant l'exécution)
  - script à exécuter (`script.sh`)
  - un fichier de données (`datafile1`)
- Fichiers de sortie (transférés du WN vers l'UI après l’exécution)
  - sortie standard
  - erreur standard

Attributs JDL (1)

- **Executable** (obligatoire) : nom de la commande à exécuter.
- **Arguments** (option) : arguments de la commande à exécuter.
- **StdInput**, **StdOutput**, **StdErr** (option) : définition des entrées/sorties/erreurs standards.
- **Environment** (option) : Ensemble de valeurs lié à l'environnement d'exécution.
- **InputSandbox** (option) : Liste des fichiers se trouvant sur l'UI qui seront transférés avec le job. Ces fichiers seront copiés sur le CE cible.
- **OutputSandbox** (option) : Liste des fichiers générés par le job qui seront accessibles via l'output sandbox

Attributs JDL (2)

- **Requirements** (option) : besoins du jobs
  - architecture, logiciel, mémoire...
- **Rank** (Option) : critère de classement des ressources éligibles
- **InputData** (option) : Référence des données utilisées en entrée par le job. SURL et/ou LFNs.
- **DataAccessProtocol** (seulement si InputData est spécifié) : protocole ou la liste de protocoles de communication utilisables par l'application pour accéder aux INPUTDATA.
- **OutputData** (option) : Référence des données de sorties qui seront récupérables

JDL : choix des ressources

```
Requirements =  RegExp("univ-bpclermont.fr",other.GlueCEUniqueID);

Requirements = Member("VO-auvergrid-blast",
other.GlueHostApplicationSoftwareRunTimeEnvironment);

Requirements = (other.GlueHostArchitecturePlatformType == "x86_64");



Rank = other.GlueCEStateFreeCPUs;

Rank = ( other.GlueCEStateWaitingJobs == 0 ) ?
other.GlueCEStateFreeCPUs : -other.GlueCEStateWaitingJobs;
```

WMS: glite-wms-job-submit

- Avec délégation automatique du proxy :

```
$ glite-wms-job-submit -a helloworld.jdl
================== glite-wms-job-submit Success ==================
The job has been successfully submitted to the WMProxy
Your job identifier is:
https://marlb.in2p3.fr:9000/wAkOlUJWWlUXI_wGQDOvBQ
==========================================================
```

```
$ glite-wms-job-delegate-proxy -d zz3
$ glite-wms-job-submit -d zz3 helloworld.jdl
```

WMS : glite-wms-job-status

- Récupère l'état d'un ou plusieurs jobs
- Syntaxe : **`glite-wms-job-status [options] jobid1 jobid2...`**

```
$ glite-wms-job-status https://marlb.in2p3.fr:9000/-_6d3JQRx8RR2SF6EZuK5g

======================== glite-wms-job-status Success =====================

BOOKKEEPING INFORMATION:


Status info for the Job : https://marlb.in2p3.fr:9000/-
_6d3JQRx8RR2SF6EZuK5g

Current Status:     Ready

Status Reason:      unavailable

Destination:        clrlcgce01.in2p3.fr:2119/jobmanager-lcgpbs-auvergrid

Submitted:          Sun Oct 17 21:45:55 2010 CEST

==========================================================================
```

WMS : glite-wms-job-output

- Récupère l'*Output Sandbox* d'un job terminé
- Syntaxe :
  glite-wms-job-output [--dir <localdir>] <jobid>

```
$ glite-wms-job-output --dir ~/tmp/ https://marlb.in2p3.fr:9000/n8Oj5613pwZ-aAeYzcGu_w

Connecting to the service https://marwms.in2p3.fr:7443/glite_wms_wmproxy_server

================================================================================

JOB GET OUTPUT OUTCOME

Output sandbox files for the job:

https://marlb.in2p3.fr:9000/n8Oj5613pwZ-aAeYzcGu_w

have been successfully retrieved and stored in the directory:

/home/mahul/tmp/mahul_n8Oj5613pwZ-aAeYzcGu_w

================================================================================

$ ls ~/tmp/mahul_n8Oj5613pwZ-aAeYzcGu_w

hello.err  hello.out
```

Gestion des données sur EGI (1)

- Pas d’espace disque partagé !
- Entité de données élémentaire : le fichier
- Le système de gestion de données (DMS) de la grille c'est :
  - Des éléments de stockage (SE) pour stocker physiquement les fichiers sur des disques ou des bandes
  - Des protocoles de transferts chiffrés qui s'appuie sur la GSI
  - Un catalogue des fichiers (LFC) pour gérer les noms de fichiers et les réplicas

Gestion des données sur EGI (2)

- Le DMS fournit toutes les opérations utiles pour:
  - Uploader / télécharger des fichiers
  - Créer des fichiers / répertoires
  - Renommer des fichiers / répertoires
  - Supprimer des fichiers / répertoires
  - Déplacer des fichiers / répertoires
  - Lister des répertoires
  - Créer des liens symboliques
- Note: Les fichiers ne sont inscriptibles qu’une fois, mais visibles plusieurs fois
  - Les fichiers ne peuvent être changés que s’ils sont supprimés ou déplacés

Schéma de nommage (1)

- **SURL** (Storage URL) : identifiant d'une instance physique d'un fichier sur un SE

```
srm://cirigridse01.univ-bpclermont.fr/dpm/univ-bpclermont.fr/home/
auvergrid/generated/2009-03-07/filefa99337c-37b6-4071-b36c-068748b79572
```

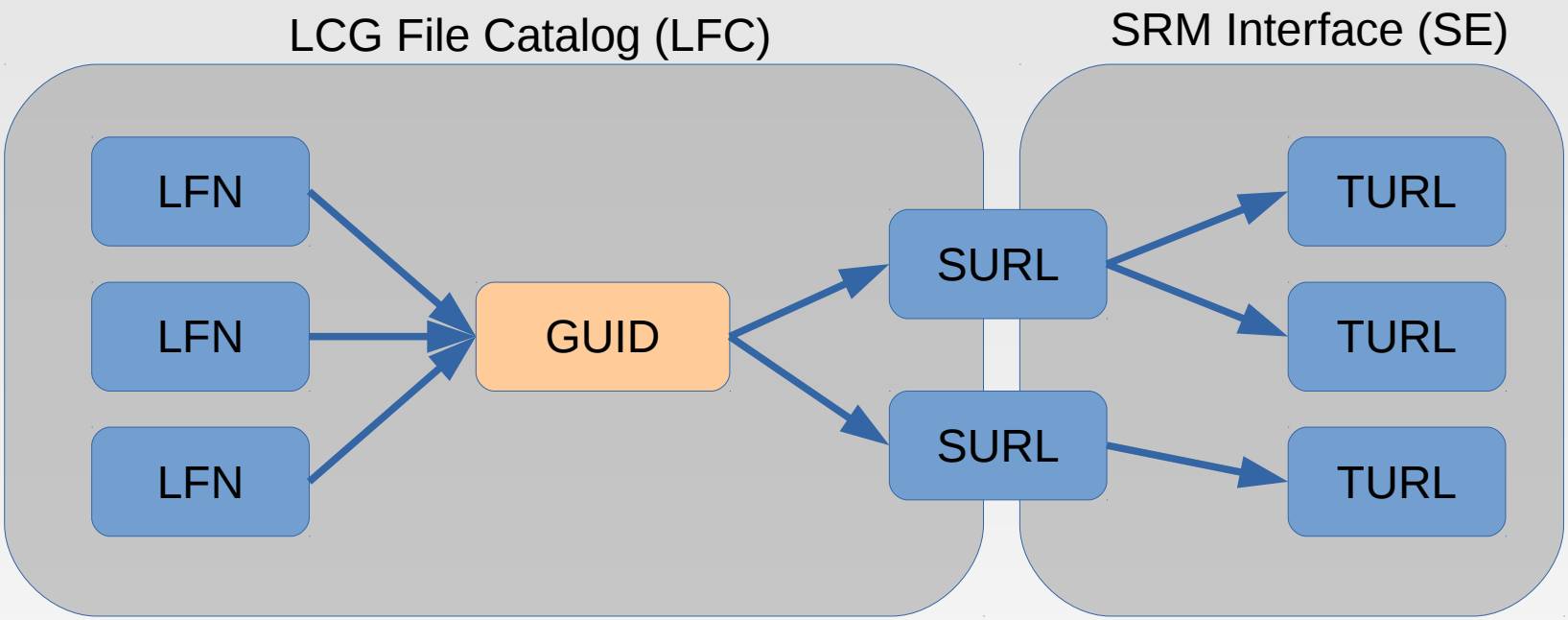
- **TURL** (Transport URL) : identifiant d'une instance physique d'un fichier sur un SE donné avec un protocole donné

```
gsiftp://cirigridse01.univ-bpclermont.fr/cirigridse01.univ-bpclermont.fr:/
storage/auvergrid/auvergrid/2009-03-07/filefa99337c-37b6-4071-b36c-
068748b79572.2651013.0
```

```
lfn:/grid/auvergrid/zz.txt
```

```
guid:0595e47a-a912-440c-9ec3-ddc9985ce7ee
```

Schéma de nommage (2)



LCG File Catalog (LFC)

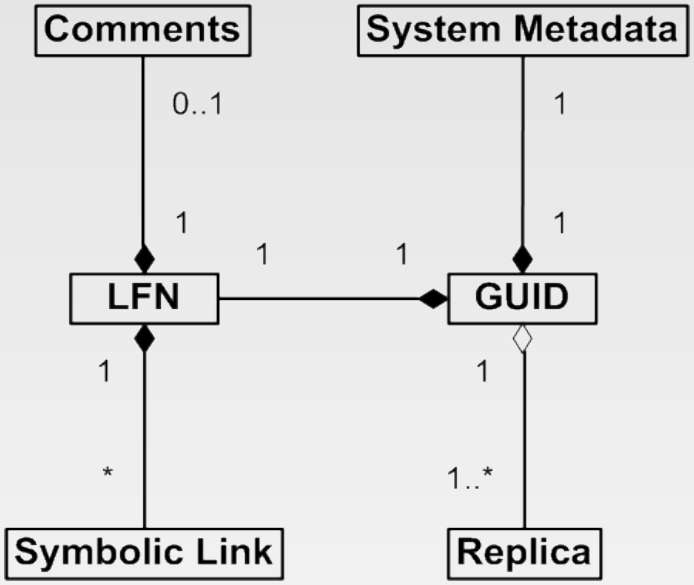
- Service qui conserve le lien entre LFN(s), GUID et SURL(s)
- Le catalogue publie son point d'accès dans le BDII et peut donc être découvert par les autres services de la grille (WMS notamment)

```
$ lcg-infosites --vo atlas lfc
prod-lfc-atlas-central.cern.ch
```

- Un catalogue unique par VO, accessible à tous les utilisateurs de la VO
- Organisation hiérarchique des fichiers avec des répertoires et des liens symboliques

Architecture du catalogue

- Lien symboliques
  (LFN supplémentaires)
- Métadonnées système
  (taille, propriétaire, checksum)
- Informations sur les réplicas
- Un champ de métadonnées utilisateur
  (commentaire)
- ACL
- Intégration à VOMS (UID, GID)
- API C



- Organisation arborescente des LFN de la forme :

/grid/<voname>/<subpaths>

Familles de commandes

- Les commandes LFC
  - Interaction uniquement avec le catalogue
  - Permet de gérer les noms logiques, les alias, les répertoires, les droits
  - Définir la variable d'environnement **LFC_HOST**

```
$ export LFC_HOST=lfc-biomed.in2p3.fr
```

- Les commandes SRM
  - Interaction uniquement avec un SE
  - Permet de gérer les réplicas, les répertoires, les droits
  - A éviter pour conserver la cohérence du LFC
- Les commandes LCG
  - Couple les opérations sur le catalogue et les opérations de gestions
  - Gère la réplication

Commandes LFC

lfc-ls	Liste les entrées (fichiers et répertoire) d'un répertoire
lfc-mkdir	Crée un nouveau répertoire
lfc-rm	Supprime un fichier ou un répertoire
lfc-rename	Renomme un fichier ou un répertoire
lfc-ln	Crée un lien symbolique vers un fichier ou un répertoire
lfc-chown	Change le propriétaire ou le groupe du fichier ou d'un rép.
lfc-chmod	Change les permissions d'un fichiers ou d'un répertoire
lfc-getacl	Récupère des ACL
lfc-setacl	Définit les ACL
lfc-setcomment	Définit ou remplace le commentaire associé à un fichier
lfc-delcomment	Supprime le commentaire

Commandes LCG

lcg-cr	Enregistre un fichier sur la grille (Local => SE)
lcg-cp	Copie un fichier depuis la grille (SE => Local)
lcg-rep	Crée un nouveau réplica (SE => SE)
lcg-del	Supprime un réplica ou tous les réplicas d'un fichier
lcg-la	Liste les alias d'un LFN, GUID ou SURL
lcg-lg	Récupère le GUID d'un LFN ou d'un SURL
lcg-lr	Liste les réplicas (SURL) d'un LFN, d'un GUID ou d'un SURL
lcg-aa	Crée un nouvel alias (LFN) d'un GUID
lcg-ra	Supprime un alias pour un GUID donné
lcg-rf	Enregistre un SURL dans le catalogue
lcg-uf	Supprime l'enregistrement d'un fichier dans le catalogue

Autres services

- AMGA
  - Serveur de métadonnées
  - Support de la GSI
  - Clé / Valeur avec une organisation hiérarchique
  - Langage de requêtage propre
- Hydra
  - Outils de chiffrement de fichiers
  - Distribution de la clé en 3 parties (2 parties suffisent)
- FTS (File Transfert Service)
  - Gestion asynchrone des transferts similaire à la gestion de jobs
  - Création de canaux de transferts entre sites