Instituts thématiques

Inserm

Institut national
de la santé et de la recherche médicale

Les grilles et le *Cloud Computing*

PARTIE 2 – La grille : Détails de fonctionnement, opérations et outils

Gilles Mathieu – gilles.mathieu@inserm.fr
Coordination de l'Informatique Scientifique de l'Inserm

Cette présentation…

… fait partie d'une série de 3 :

- Partie 1 : La grille : Concepts, architectures et fonctionnement général
- **Partie 2 : La grille : Détails de fonctionnement, opérations et outils**
- Partie 3 : Le *Cloud Computing* : principes et utilisation

La grille…

- **Détails de fonctionnement**
  - Services d'authentification et d'autorisation
  - Système d'information
  - Sites grille, calcul et stockage
- **Outils d'opérations**
  - Monitoring et Accounting
  - Autres outils
- **Outils d'utilisation**
  - Job schedulers et gateways
  - Gestionnaires de données

DETAILS DE FONCTIONNEMENT

La pile grille

Applications

Standalone apps

Portails & gateways

Middleware

File catalogs

Information System

Storage Element

Workload management

Computing Element

User Interface

Auth. service

Ressources

Disque/bande

Cluster

Serveur

Réseau

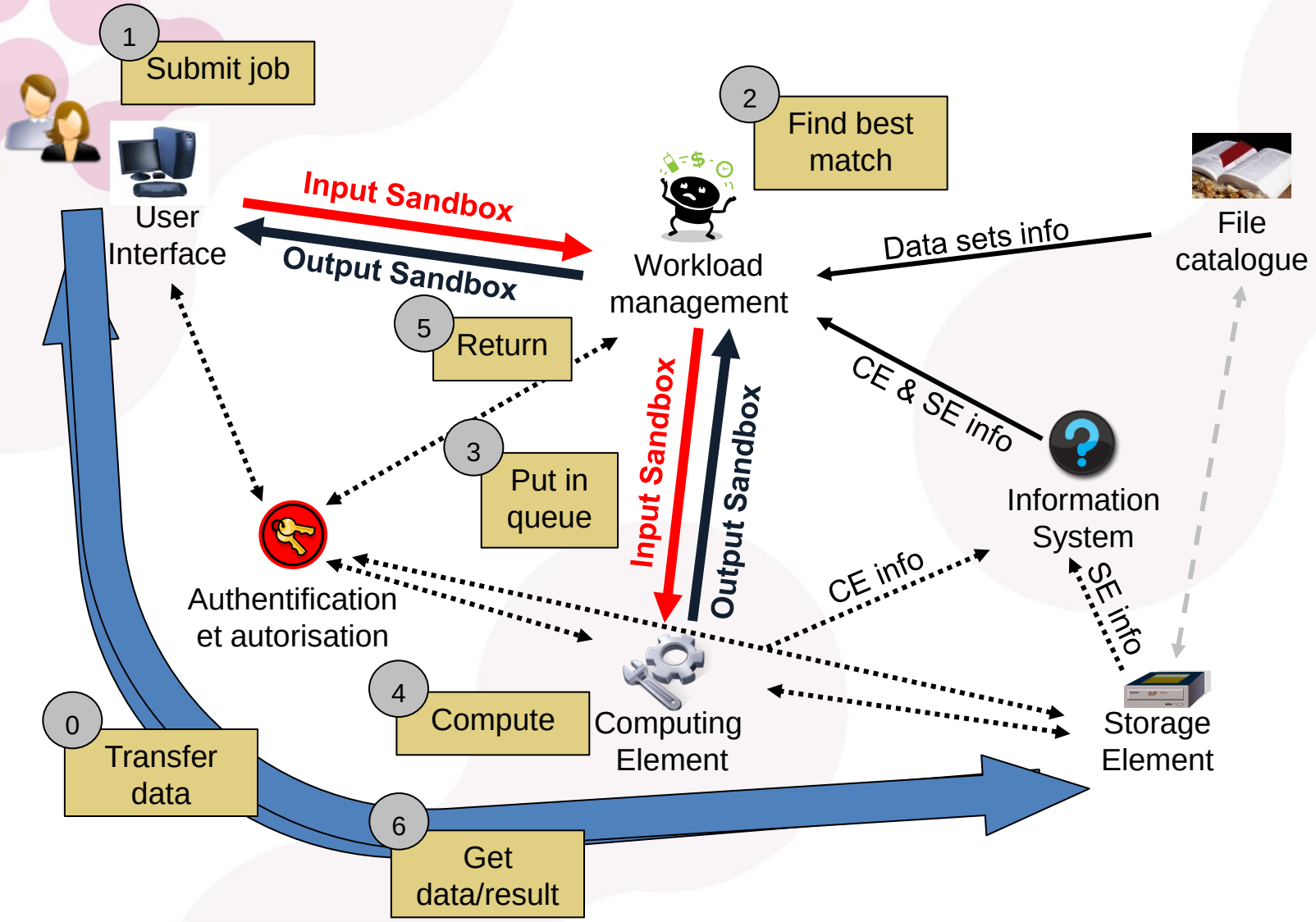
Internet public

Réseaux académiques

Lignes dédiées

5

En un coup d'oeil…



6

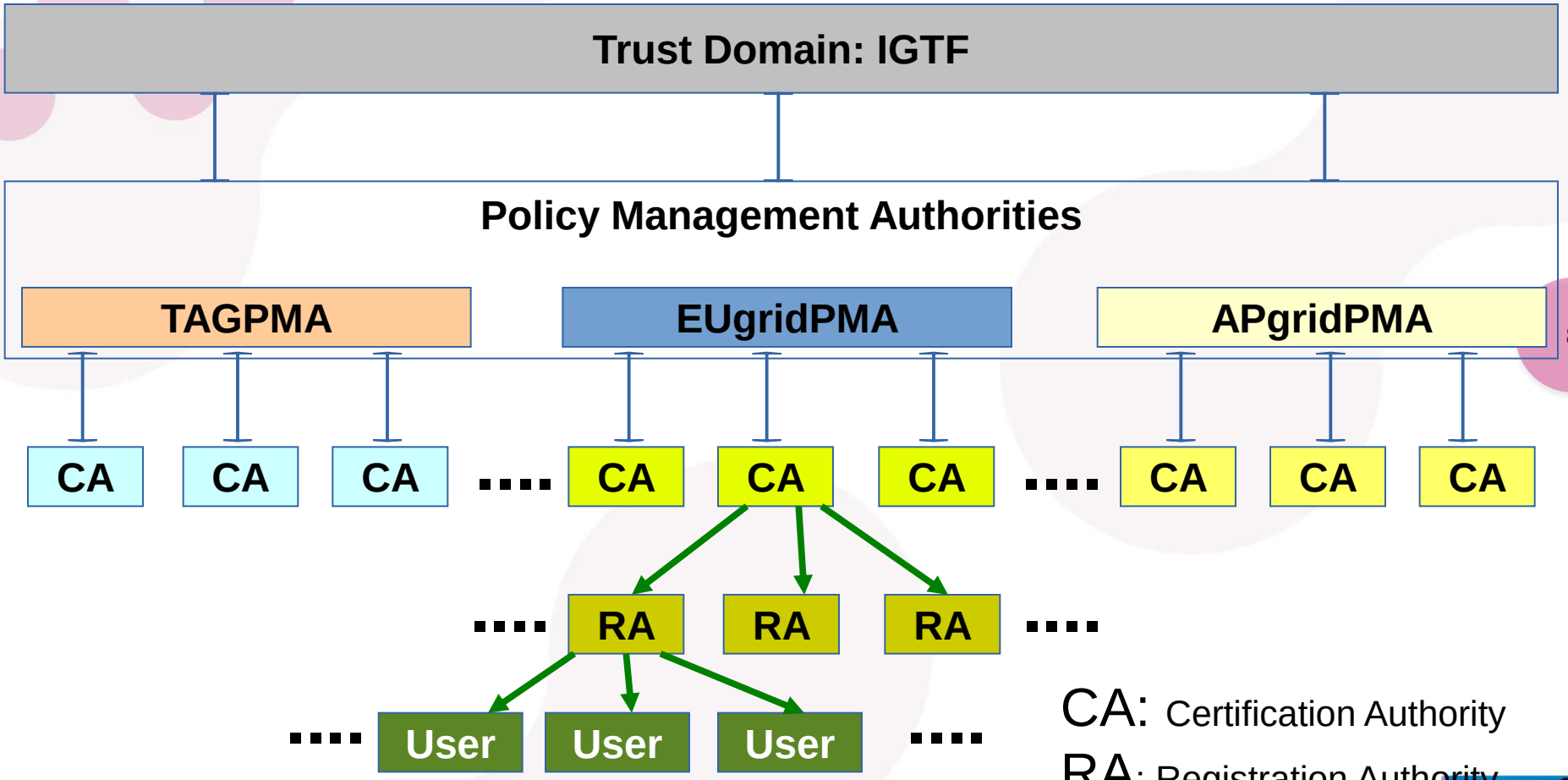
Authentification et autorisation

- **Authentification basée sur certificats X509**
  - Délivrés par une autorité de certification reconnue
  - Authentification mutuelle (utilisateur/service)

- **Autorisation basée sur l'appartenance à une VO**
  - Enregistrement dans un service dédié (VOMS)
  - Validation régulière par les VO managers
  - Etre enregistré signifie accepter les conditions d'utilisation

7

Authentification et autorisation : la chaine de certification



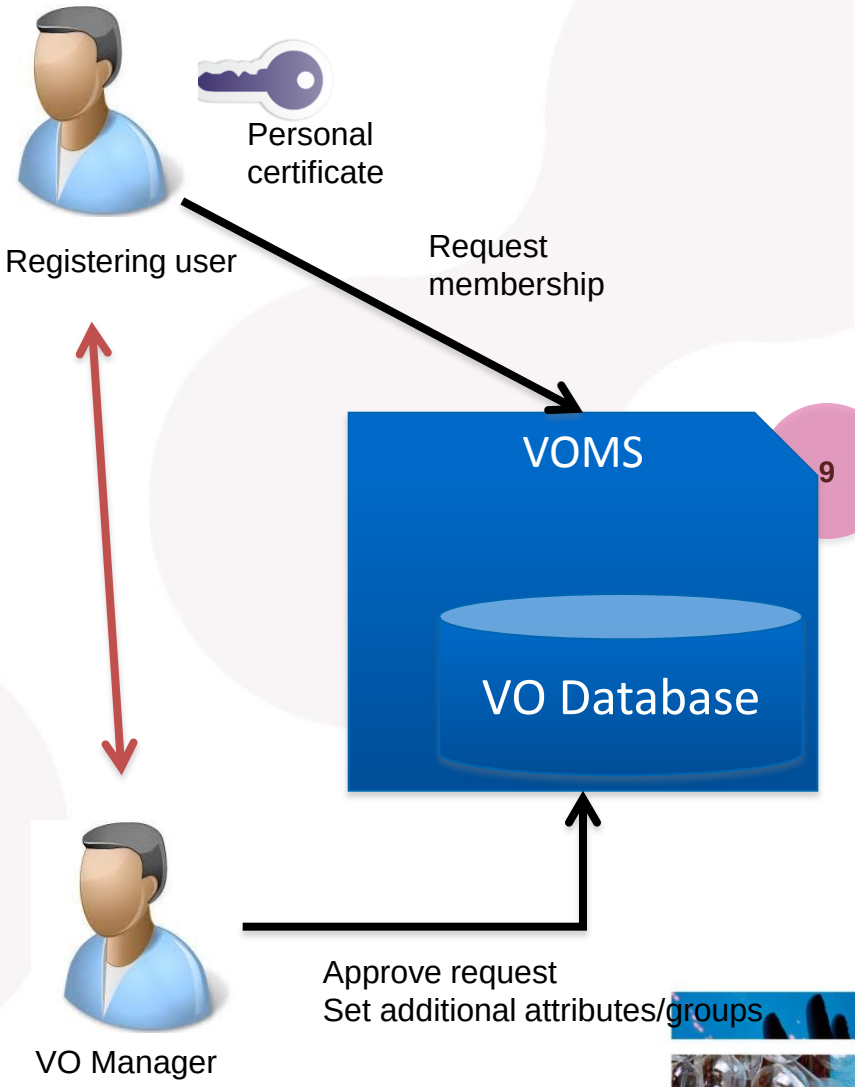
CA: Certification Authority

RA: Registration Authority

Crédits : Peter Solonga, EGI.eu

L'enregistrement dans une VO

- User registers at the VO via VOMS
- VO manager authorizes the user via VOMS
- VO manager can give specific attributes to users, or insert them in specific groups
- Specific VOMS service is configured in all the services supporting the VO



Crédits : Peter Solagna, EGI.eu

Le proxy de certificat

- **Le proxy de certificat est une version courte durée du certificat de l'utilisateur (signée par son certificat)**
- **Les proxies sont utilisés pour tout les services non interactifs et pour la délégation**
  - Un calcul est envoyé avec le proxy de l'utilisateur, et peut ainsi stocker des données de sortie en son nom
- **Un proxy est "self contained" : il contient toutes les infos nécessaires pour authentifier et autoriser l'utilisateur au niveau service**
  - Identité de l'utilisateur
  - Appartenance à une VO signée par le service VOMS de cette VO

X509 Proxy DN: User Certificate info | VO Information

Authentification : une variante

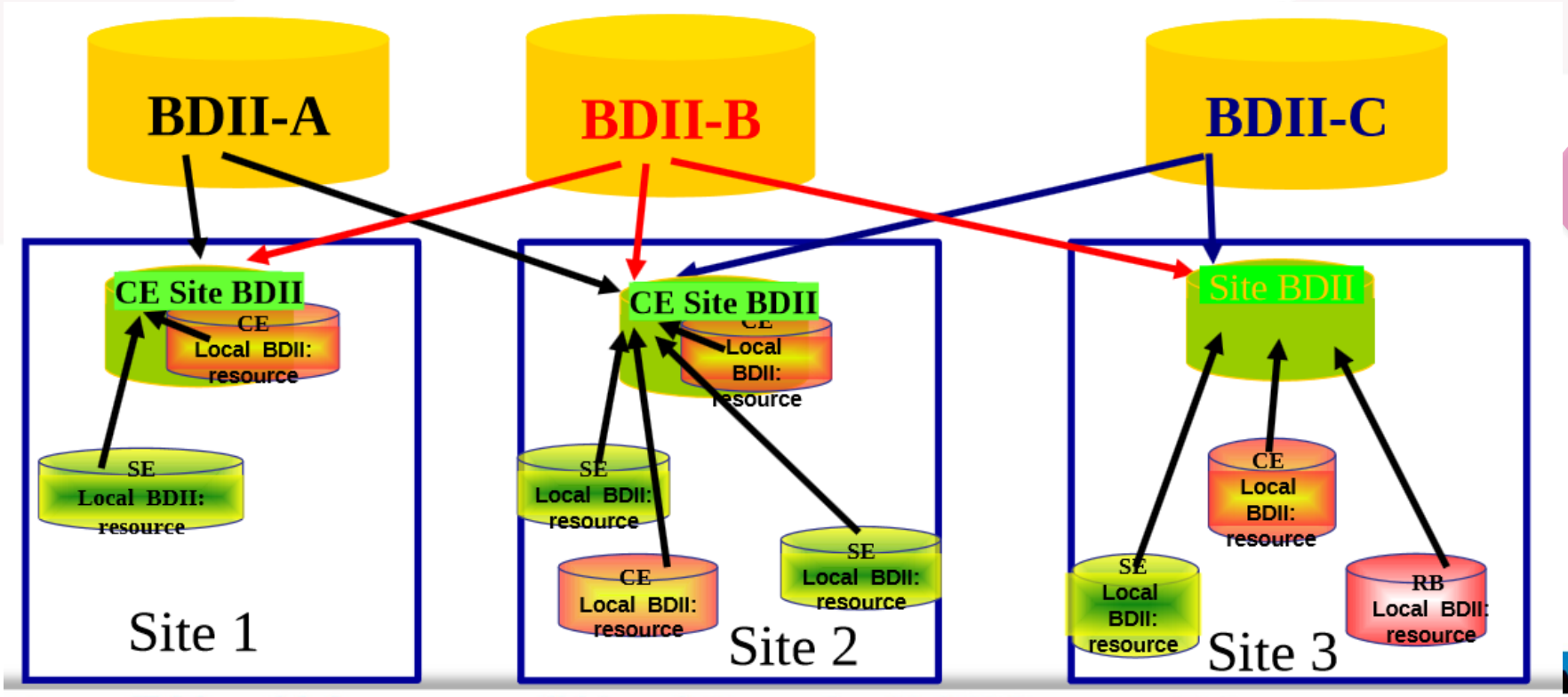
- **Utilisation de certificats robots par des portails/gateways**
  - Utilisateur générique (non personnel)
  - Pour cacher la complexité de l'utilisation de certificats
  - Certificat stocké sur la machine qui génère des proxies pour les utilisateurs loggés
  - "déporte" le mécanisme d'authentification

Système d'information

- **Publie les informations sur :**
  - La topologie (nature et propriétés statiques des services et ressources)
  - La configuration des ressources (job queues, VOs supportées, bibliothèques disponibles…)
  - Le statut des ressources (available, in maintenance, down, espace disponible, job slots disponibles)

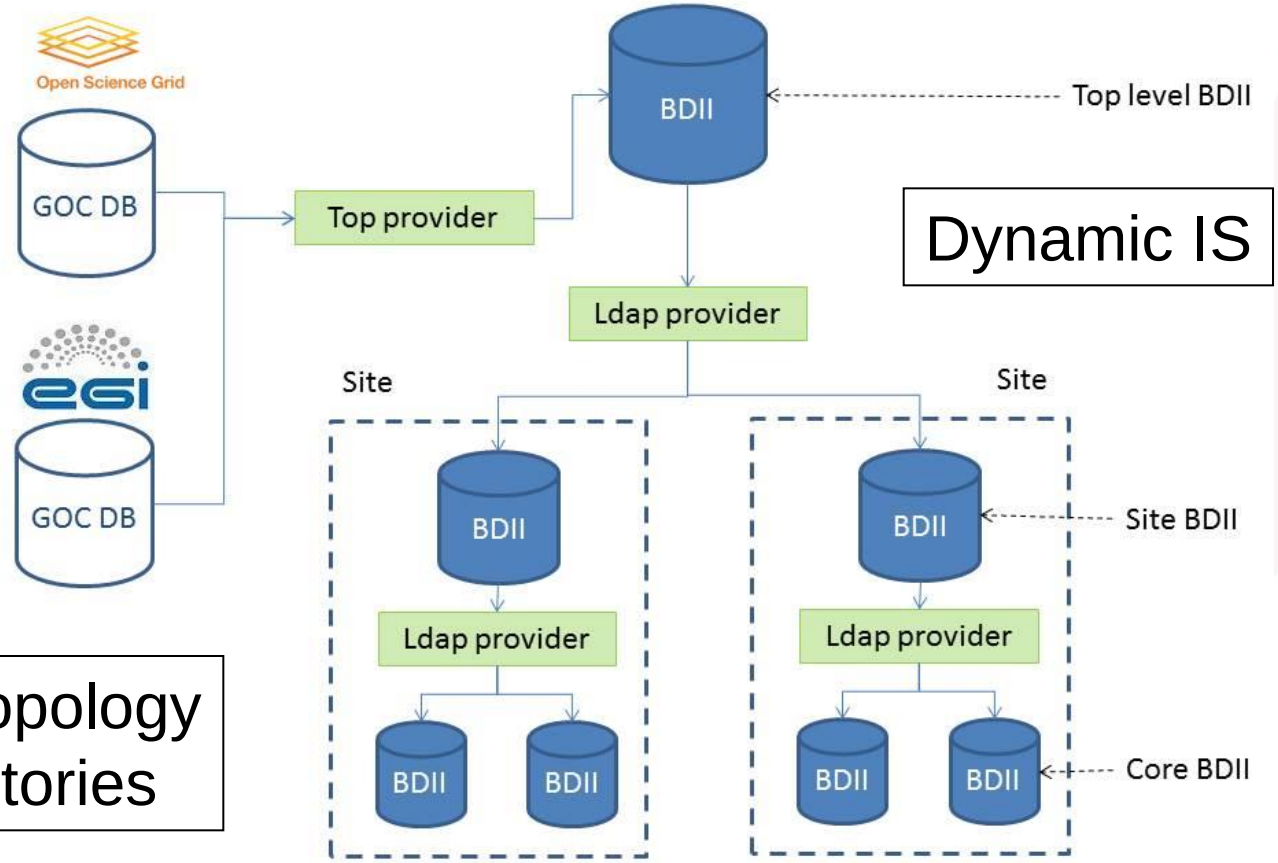
Système d'information : le BDII

- Dynamic IS: Berkeley DB Information Index (BDII)



13

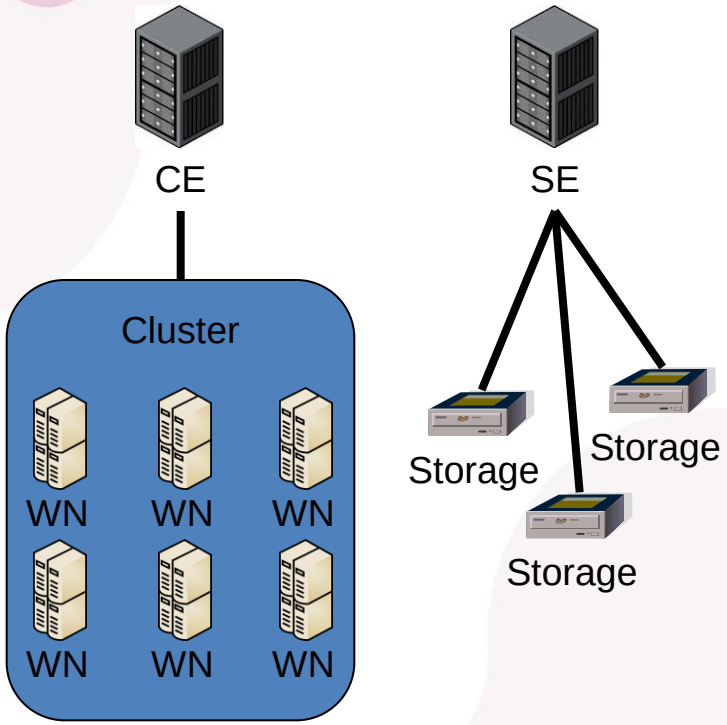
Information statique et dynamique



Dynamic IS

Static topology repositories

Un site grille "classique"



Minimum:
Nœuds de calcul
Nœud SI (site-BDII)

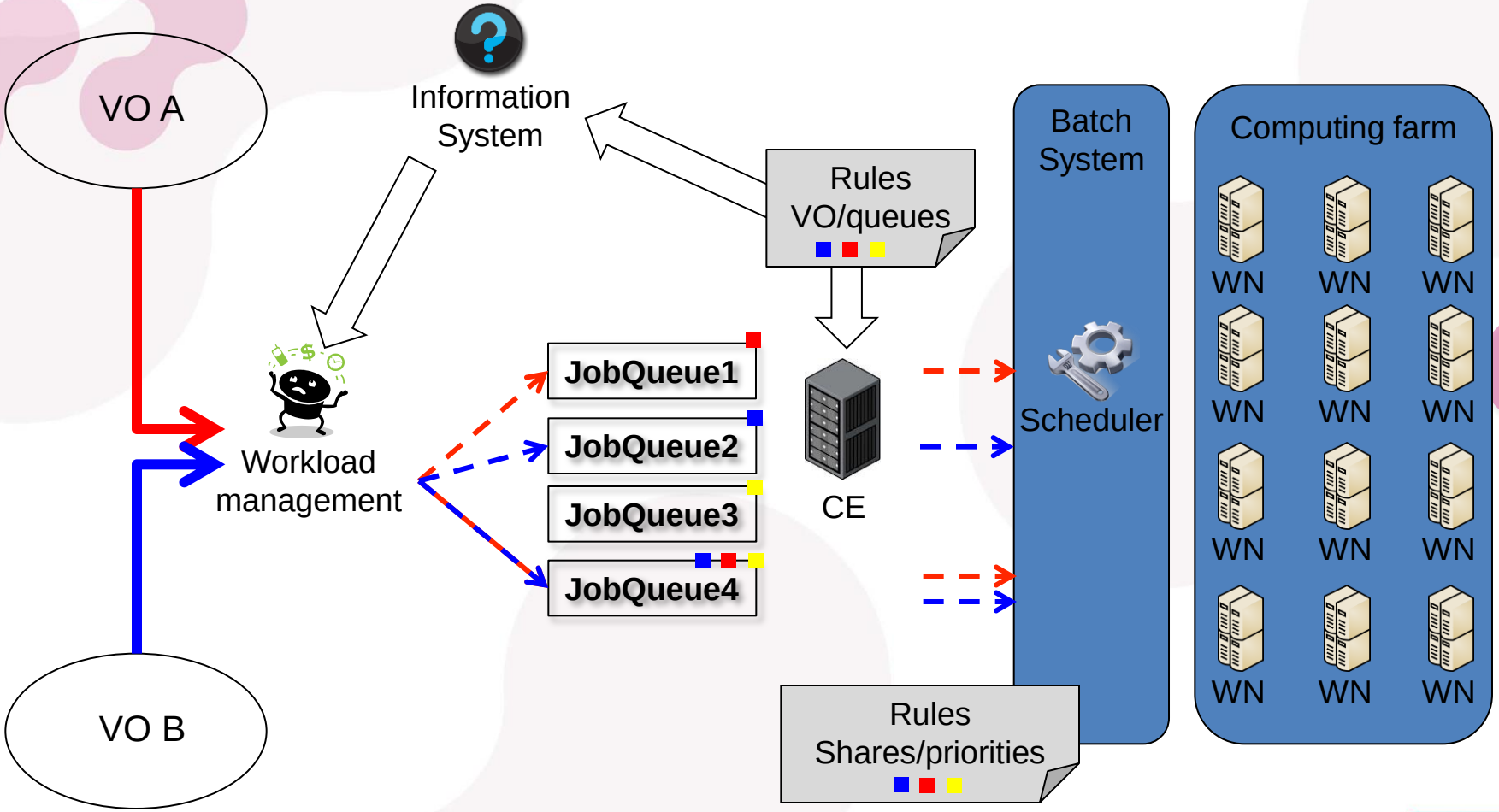
Optionnel:
Nœuds de stockage
Service UI

15

Les nœuds de calcul

- **Un élément de calcul (CE) est associé à plusieurs nœuds (Worker Nodes, WN)**
- **Le CE "gère".**
  - Définition des job queues
  - Gestion des logs, des régles de priorités, des Vos supportées, etc.
- **Les jobs tournent sur les WNs**
  - Ce sont les machines de base, configurées selon les besoins des utilisateurs concernés

Organisation des nœuds de calcul



17

Configuration des ressources

Queue=creamce01/Q_atlas

Queue=creamce01/Q_biomed

Queue=creamce02/Q_short

Queue=creamce02/Q_long

creamce01.site.fr

creamce02.site.fr

```
Q_atlas_VO = atlas
Q_atlas_job_size = all

Q_biomed_VO = biomed
Q_biomed_job_size = all

Q_short_VO = astro, esr
Q_short_job_size <= 21600s

Q_long_VO = astro, esr
Q_long_job_size >= 21600s
```

Queues config

Defines queues properties for user jobs submission

```
atlas_share = 80%
biomed_share = 15%
others.share =  5%
```

Fairshare config

Equilibrates resource usage in the long term

```
Short_jobs_priority = 1
Short_jobs_limit = 1000
Long_jobs_priority = 2
Long_jobs_limit = 3000
```

Priority/limit config

Defines submission rules to batch system

18

Construction d'un job

- **on spécifie au minimum dans le JDL :**
  - le programme et ses arguments
  - redirection des outputs et erreurs dans des fichiers
  - ce qu'on fait de la sortie (OutputSandbox)
- **cat HelloWorld.jdl**

```
Executable = "/bin/echo";
Arguments = "Hello World";
StdOutput = "message.txt";
StdError = "stderror";
OutputSandbox = {"message.txt", "stderror"};
```


JDL : Attributs

- **Attributs définissant le Job lui-même**
  - Executable, Arguments, Std(input/output/error), Environment, Inputsandbox, OutputSandbox
- **Attributs définissant les ressources nécessaires**
  - Informations disponibles dans le SI
- **Attributs définissant les données utilisées**
  - Points d'entrée vers le catalogue de données
  - Nom de fichier et protocole

JDL

- **Un exemple de JDL**

attribut job

attribut données

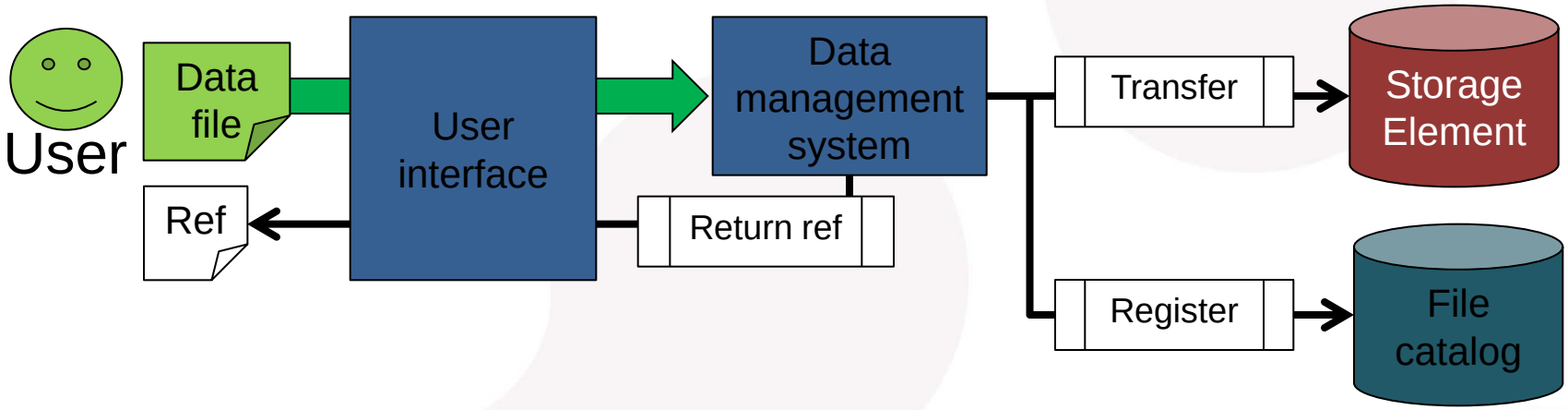
attributs ressources

```
Executable = "gridTest";
StdError = "stderr.log";
StdOutput = "stdout.log";
InputSandbox = {"/home/joda/test/gridTest"};
OutputSandbox = {"stderr.log", "stdout.log"};
InputData = "lfn:testbed0-00019";
DataAccessProtocol = "gridftp";
Requirements = other.Architecture=="INTEL" && \
               other.OpSys=="LINUX" && other.FreeCpus\
   >=4;
Rank = other.GlueHostBenchmarkSF00;
```

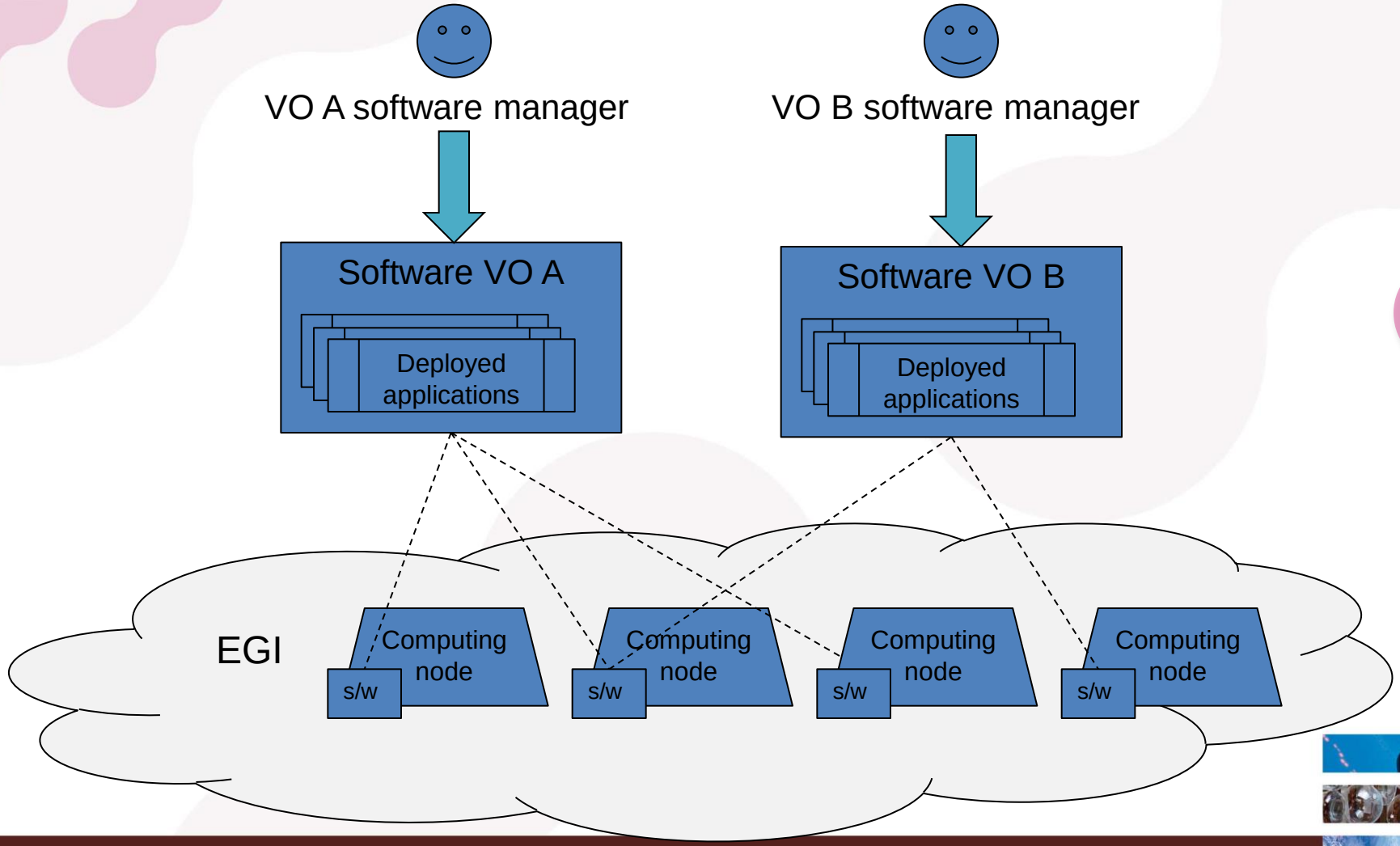

21

Gestion de données

- **Principe général**
  - L'utilisateur dépose ses données sur un SE
  - Elles sont référencées dans le catalogue
  - La référence est passée au job dans le JDL
  - Le job permet la manipulation de ces données

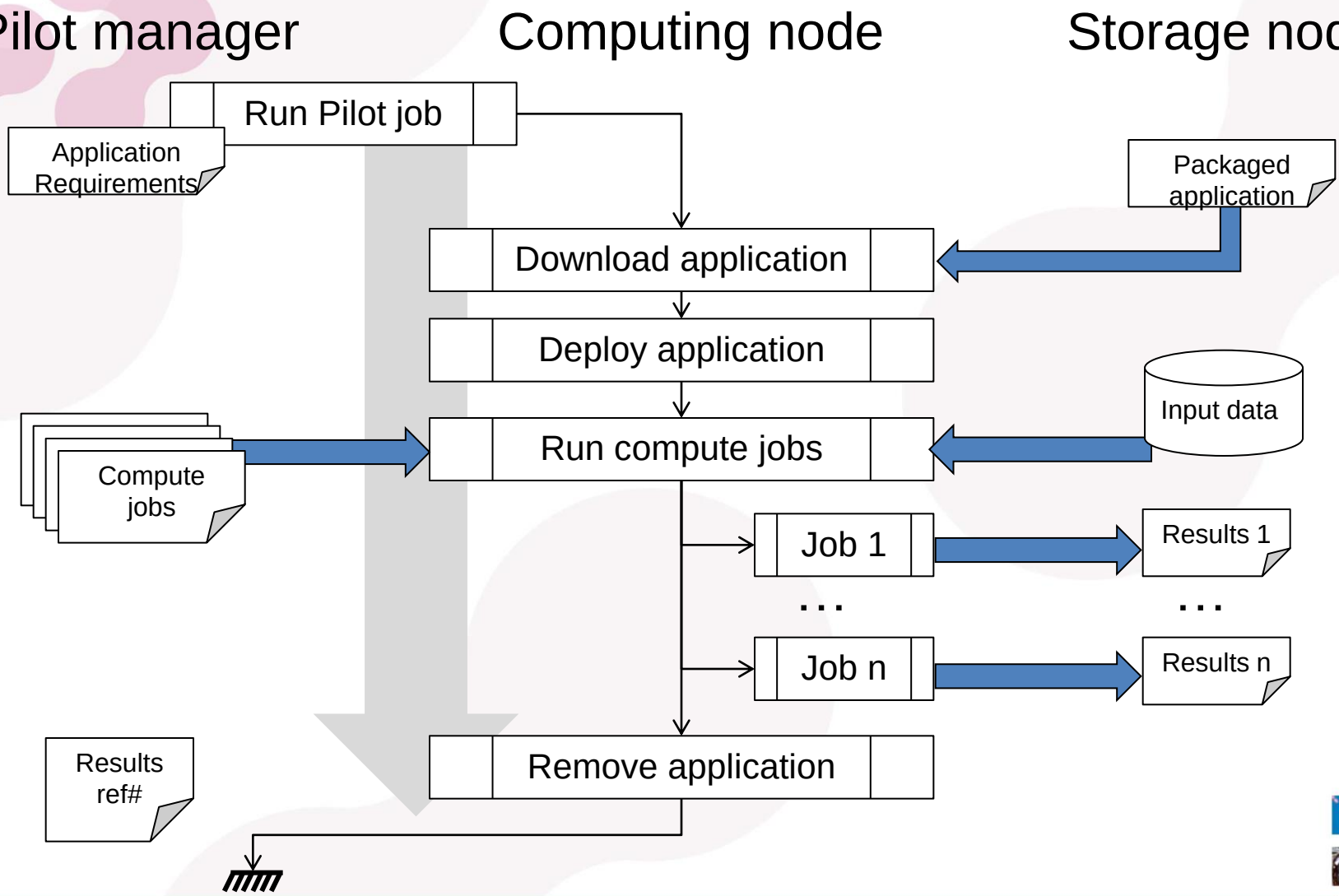


Application management: utilisation des software areas



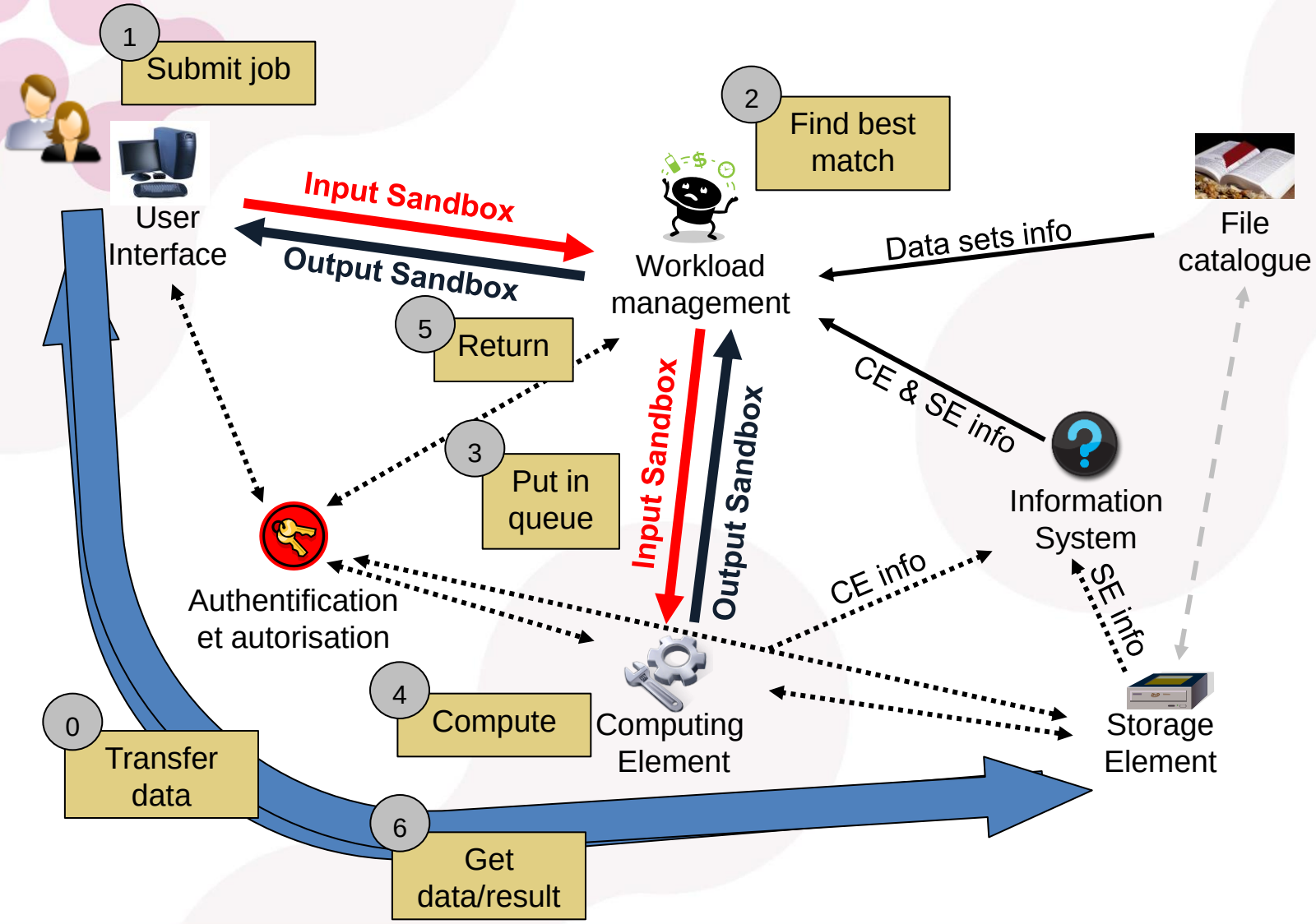
23

Application management : jobs pilotes



24

Pour tout résumer :



25

OPERATIONS

Opérer une grille

- **Opérer une grille de production nécessite de nombreux services :**
  - Monitoring (surveillance)
  - Accounting (comptabilité)
  - Autorité de Certification (authentification)
  - Sécurité (protection)
  - Outils collaboratifs (collaboration)

27

Opérations dans France Grilles

ORGANISATION DE LA GRILLE DE PRODUCTION EN FRANCE

OFFRES DE SERVICES VERS LES UTILISATEURS

- Catalogue France Grilles
- Catalogues des VOS

SERVICES OPERATIONNELS

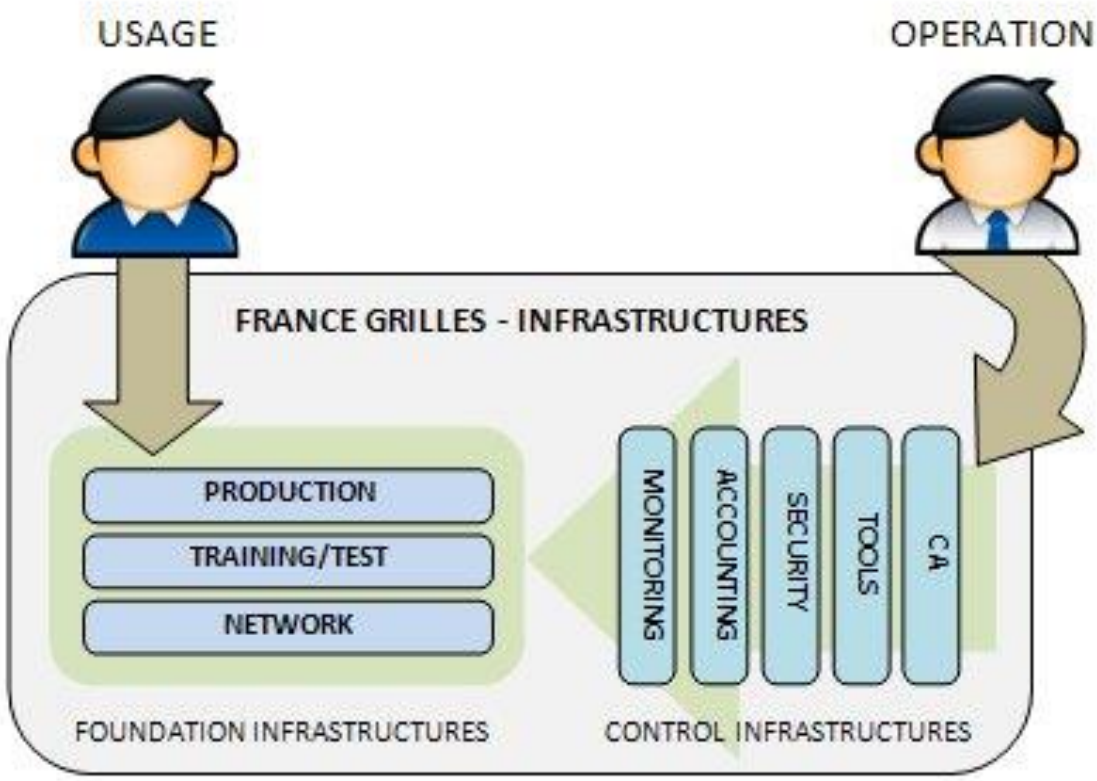
- Coordination
- Grille de production : Services centraux
- Grille de production : Services sites
- Services internes
- Support

RESSOURCES

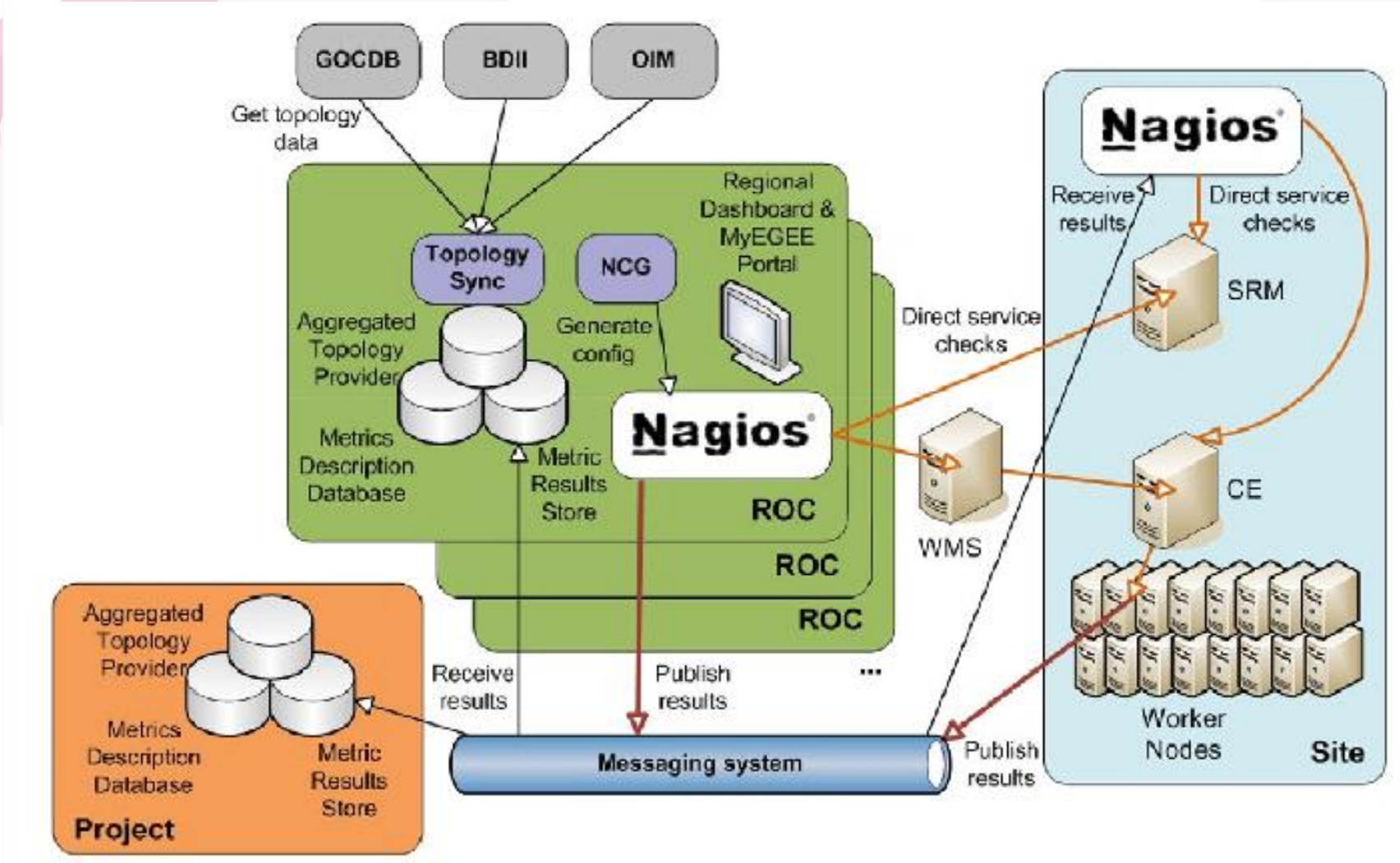
- Ressources humaines
- Infrastructures matérielles
- Réseaux

28

Opérations dans France Grilles

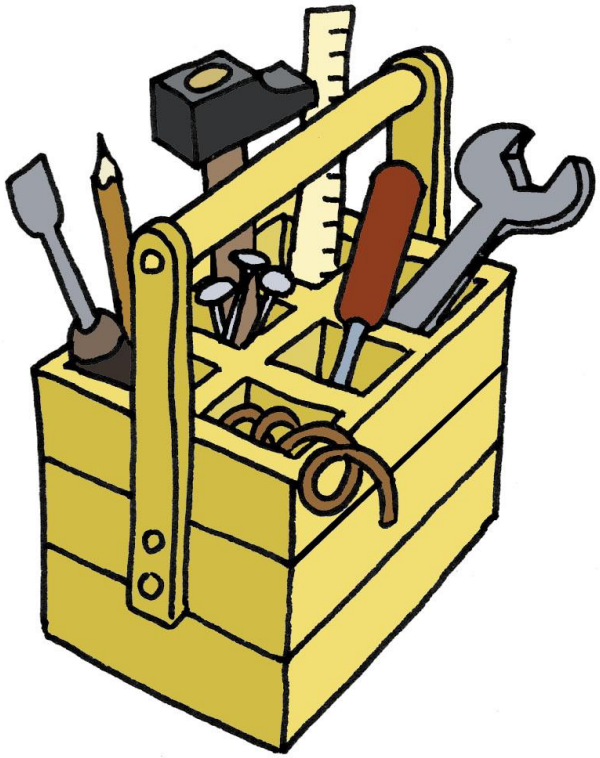


Exemple : monitoring EGI



30

OUTILS D'UTILISATION



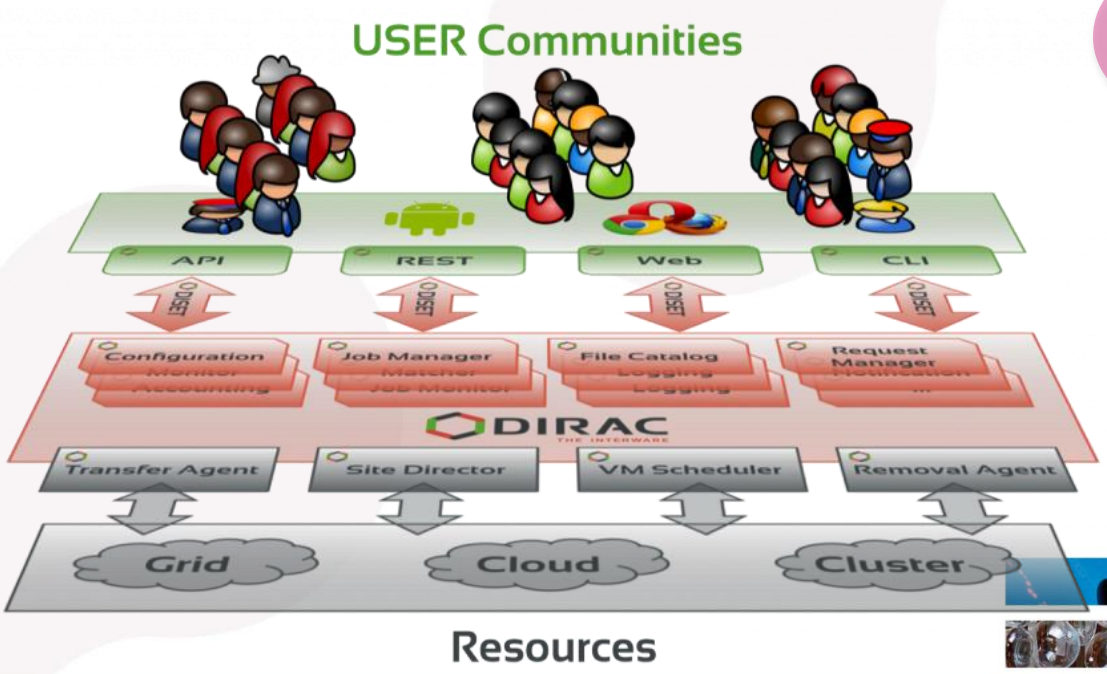
Tools and apps examples

- DIRAC: a general solution to access resources, manage computing and distributed data
- iRODS: a virtualised storage and data management system
- VIP (Virtual Imaging Platform): a successful example of Virtual Research Environment

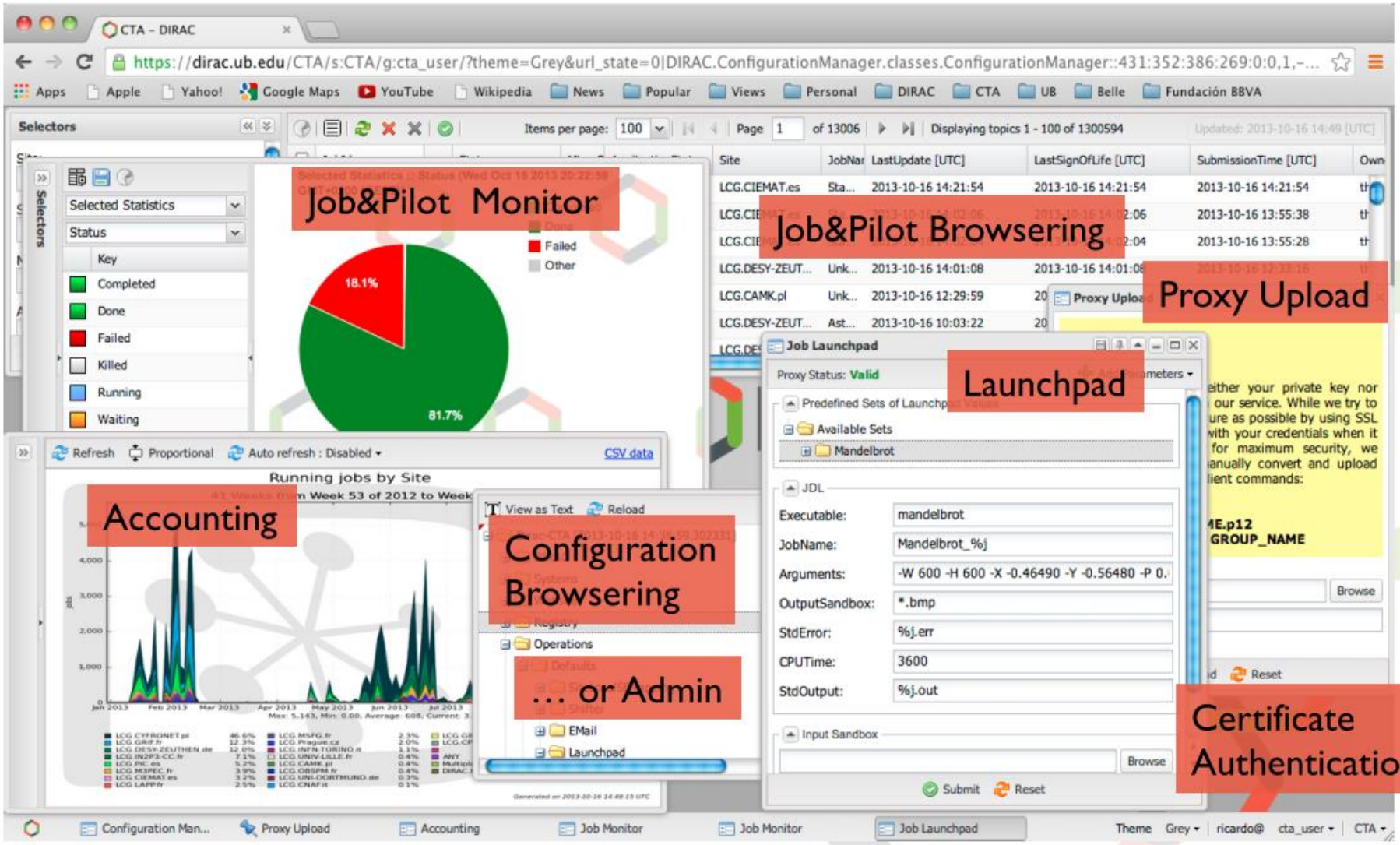
Tools and apps: DIRAC

- **An "interware" (interface between user and middleware)**
  - Efficient job management system
  - Facilitated access to the grid, clusters and clouds

France Grilles has deployed a national, multi-community instance of DIRAC

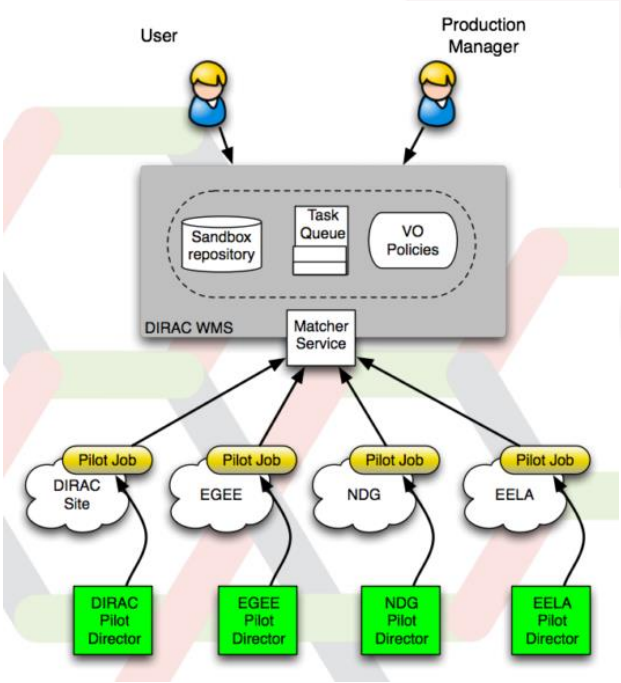


DIRAC web interface



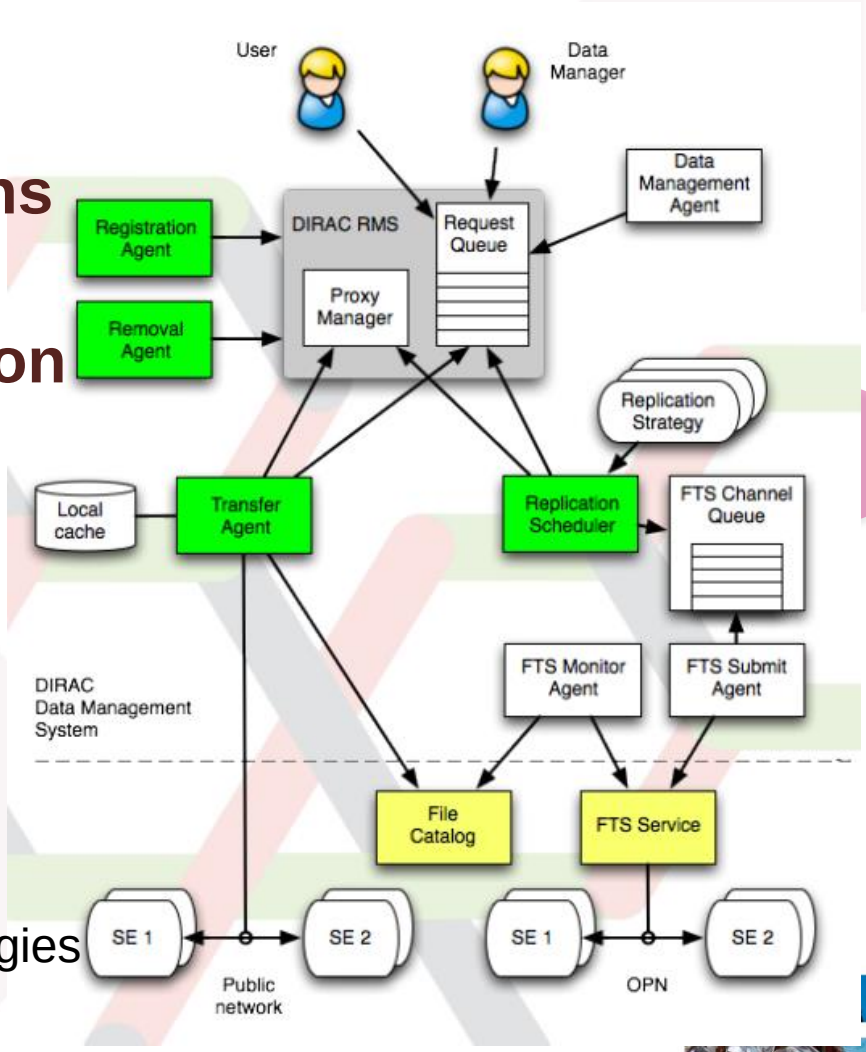
DIRAC workload management

- Jobs are submitted to the DIRAC Central Task Queue with credentials of their owner (VOMS proxy)
- Pilot Jobs are submitted by specific Directors to a Grid WMS with credentials of a user with a special Pilot role
- The Pilot Job fetches the user job and the job owner's proxy
- The User Job is executed with its owner's proxy used to access SE, catalogs, etc

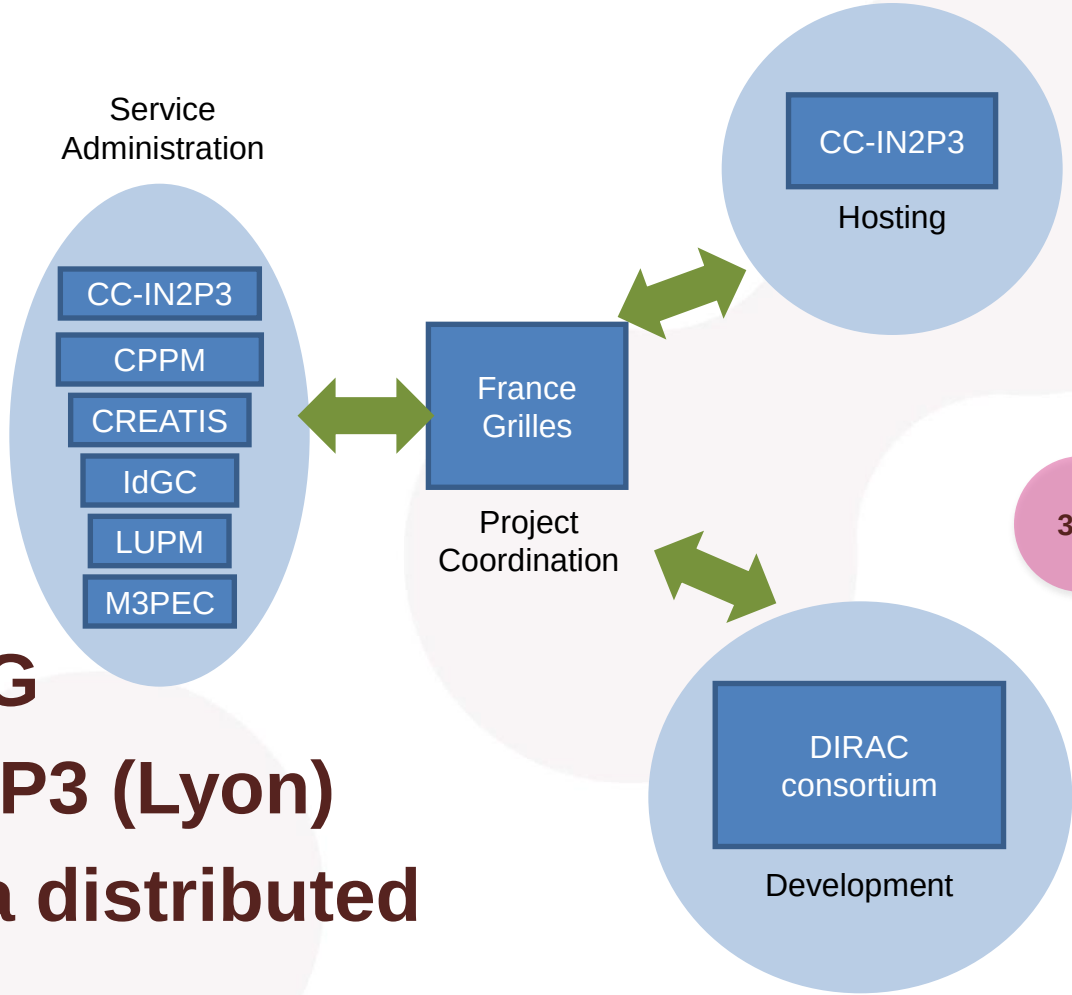


DIRAC data management

- **Based on the Request Management System**
- **Asynchronous data operations**
  - transfers, registration, removal
- **Two complementary replication mechanisms**
  - Transfer Agent
    - user data
    - public network
  - FTS service
    - Production data
    - Private FTS OPN network
    - Smart pluggable replication strategies



DIRAC national instance



- Coordinated by FG
- Hosted by CC-IN2P3 (Lyon)
- Administered by a distributed team (6 labs)
- Multi-communities

Principaux avantages de DIRAC

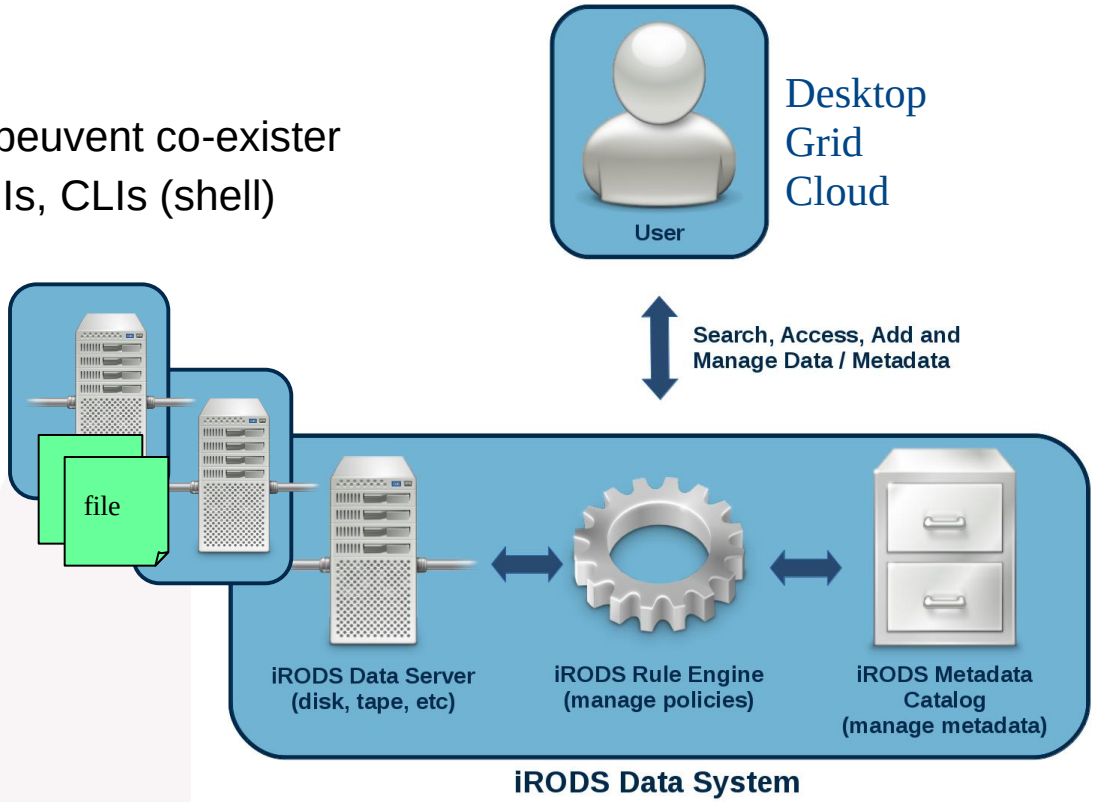
- **Gestion facilitée de productions**
  - Agrégation de ressources hétérogènes de manière transparente
  - Système de récupération des échecs
  - Soumission automatique
- **Gestion de données intégrée**
  - Distribution automatique, contrôles d'intégrité…
- **Ergonomie**
  - "The grid with a human face"…

Outils : iRODS

Éléments principaux définissant un système iRODS

- Base de données
- Moteur de règle
- Ressources

-Plusieurs ressources distantes peuvent co-exister
Interfaces utilisateurs : APIs, GUIs, CLIs (shell)



Authentification

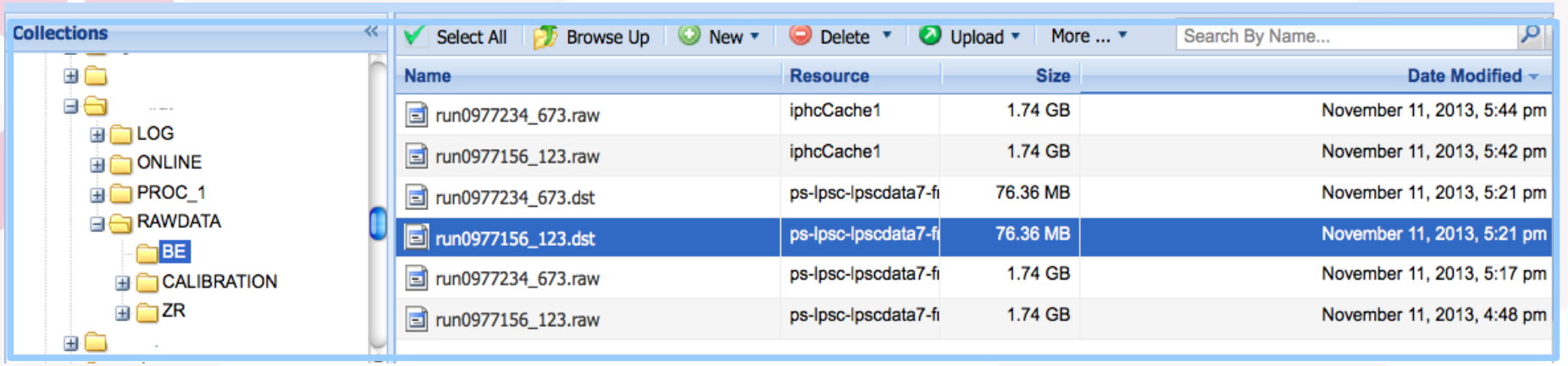
- Mot de passe
- Kerberos
- Certificat, ...

Gestion des données

- Flot des données prédéfini
et transparent pour l'utilisateur
- A la demande

39

iRODS user interface

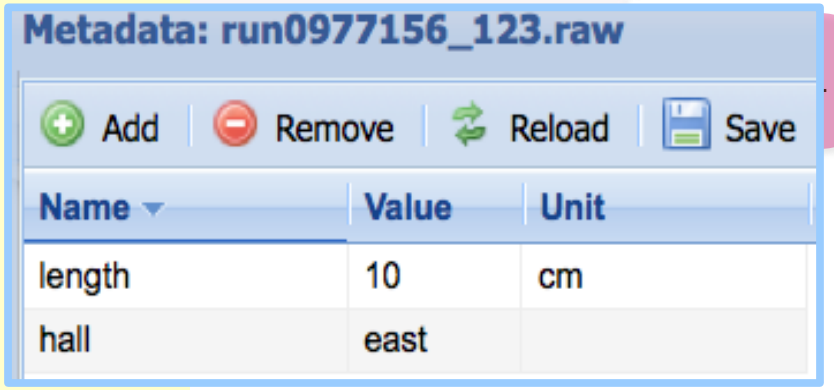



```
[user ~]$ ils
/frgrid/home/UNECOLLAB/RAWDATA:
 C- /frgrid/home/UNECOLLAB/RAWDATA/CALIBRATION
 C- /frgrid/home/UNECOLLAB/RAWDATA/BE
 C- /frgrid/home/UNECOLLAB/RAWDATA/ZR
[user ~]$ ils -l BE/
/frgrid/home/UNECOLLAB/RAWDATA/BE:
  owner   0 ps-lpsc-lpscdata7-fr   80072192 2013-11-11.16:21 & run0977156_123.dst
  owner   0 ps-lpsc-lpscdata7-fr 1748189011 2013-11-11.15:48 & run0977156_123.raw
  owner   1 iphcCache1           1748189011 2013-11-11.16:42 & run0977156_123.raw
  owner   0 ps-lpsc-lpscdata7-fr   80072192 2013-11-11.16:21 & run0977234_673.dst
...
```

Métadonnées iRODS

- Attachées à une collection, un fichier, un utilisateur
- Triplet: nom attribut unité

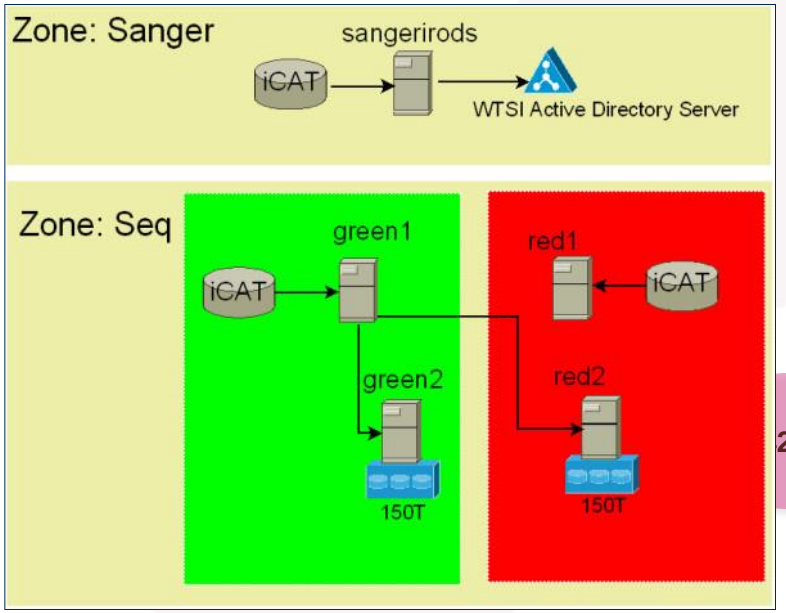
```
[user ~]$ imeta add -d run0977156_123.raw length 10 cm
[user ~]$ imeta add -d run0977156_123.raw hall east
[user ~]$ imeta ls -d run0977156_123.raw
AVUs defined for dataObj run0977156_123.raw:
attribute: length
value: 10
units: cm
----
attribute: hall
value: east
units:
[user ~]$ imeta -d qu hall east
collection: /frgrid/home/UNECOLLAB/RAWDATA/ZR
dataObj: run0977156_123.raw
----
collection: /frgrid/home/UNECOLLAB/RAWDATA/ZR
dataObj: run0817773_556.raw
```



iRODS: Genomic data management

WTSI Use Case:

- Managing and accessing sequencing Binary Alignment/Map (BAM) files
- 500 TB SAN Storage
- Integrated in the sequencing pipeline
- Fine-grained access control
- Data replication
- Metadata on alignment are automatically added
- Data federation with other research institutes



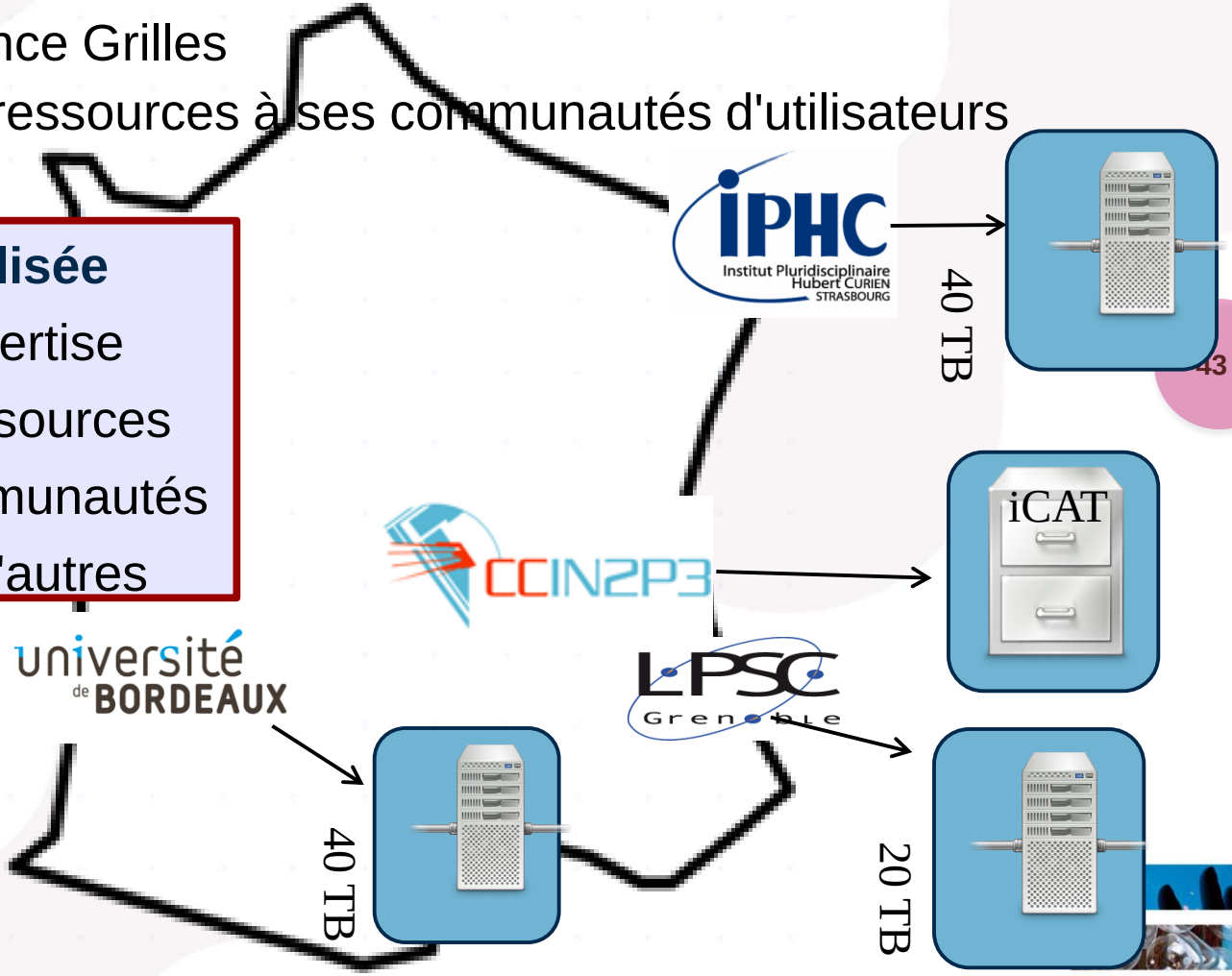
G.-T. Chiang, P. Clapham, G. Qi, K. Sale & G. Coates: Implementing a genomics data management system using iRODS in the Wellcome Trust Sanger Institute. BMC Bioinformatics 2011, 12, 361.

L'instance iRODS France Grilles

- Coordonnée par France Grilles
  - Faciliter l'accès aux ressources à ses communautés d'utilisateurs

Une instance mutualisée

- Mutualisation de l'expertise
- Mutualisation des ressources
- Aide aux petites communautés
- Terrain d'essai pour d'autres

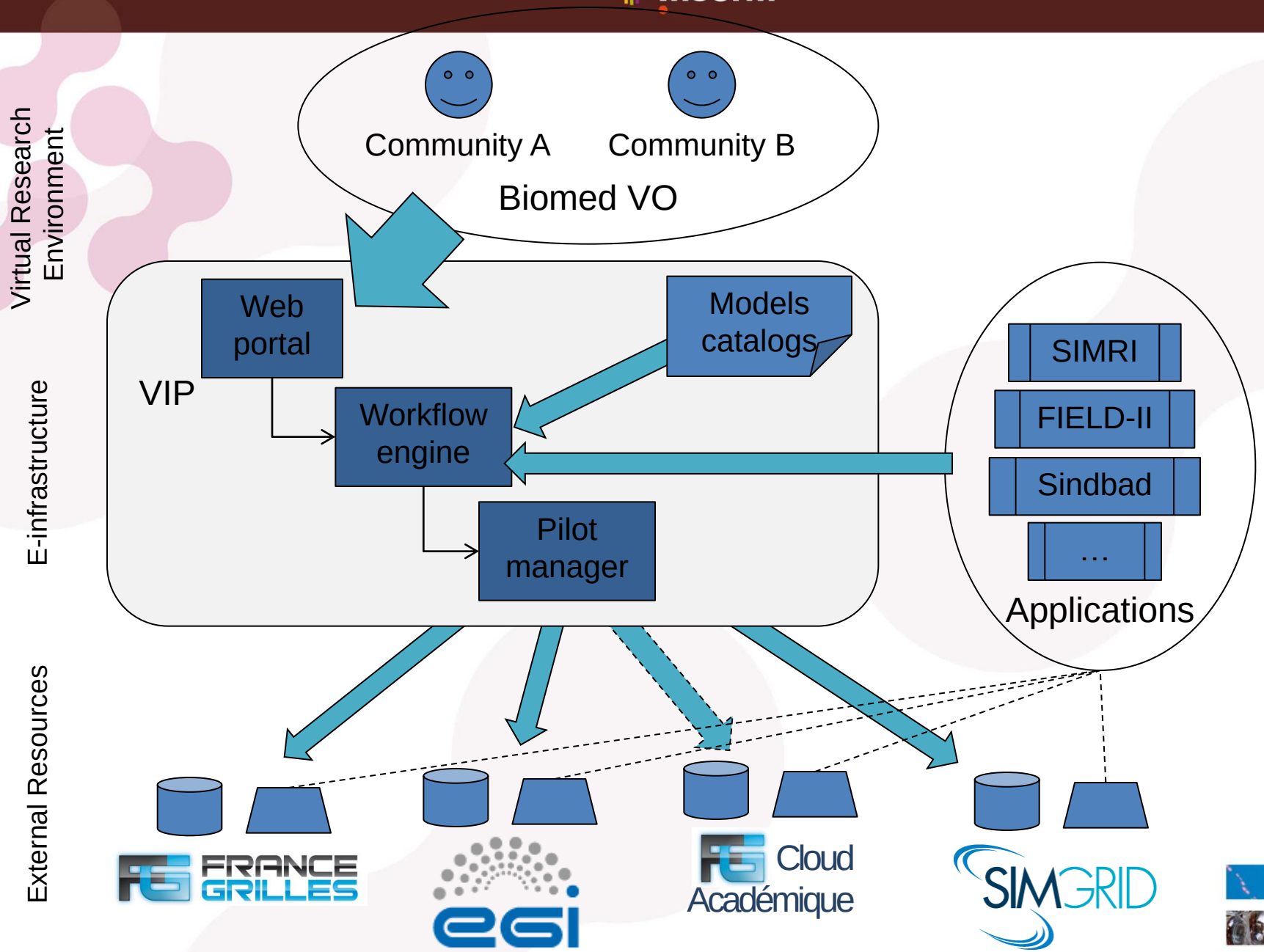


FRANCE GRILLES

Tools and apps: VIP

- **VIP = Virtual Imaging Platform**
  - A web platform to facilitate data sharing and access to computing resources for medical image simulation
  - Developed by CREATIS (Inserm, CNRS, INSA, Lyon University) - http://www.creatis.insa-lyon.fr/vip/
  - Recent article in EGI Newsletter (May15) : http://www.egi.eu/news-and-media/newsletters/Inspired_Issue_19/vip.html
- **A successful VRE example**

44


Virtual Research Environment
E-infrastructure
External Resources
Community A
Community B
Biomed VO
VIP
Web portal
Models catalogs
Workflow engine
Pilot manager
SIMRI
FIELD-II
Sindbad
…
Applications
FRANCE GRILLES
EGI
Cloud Académique
SIMGRID

45

VIP – "Applications as a service"

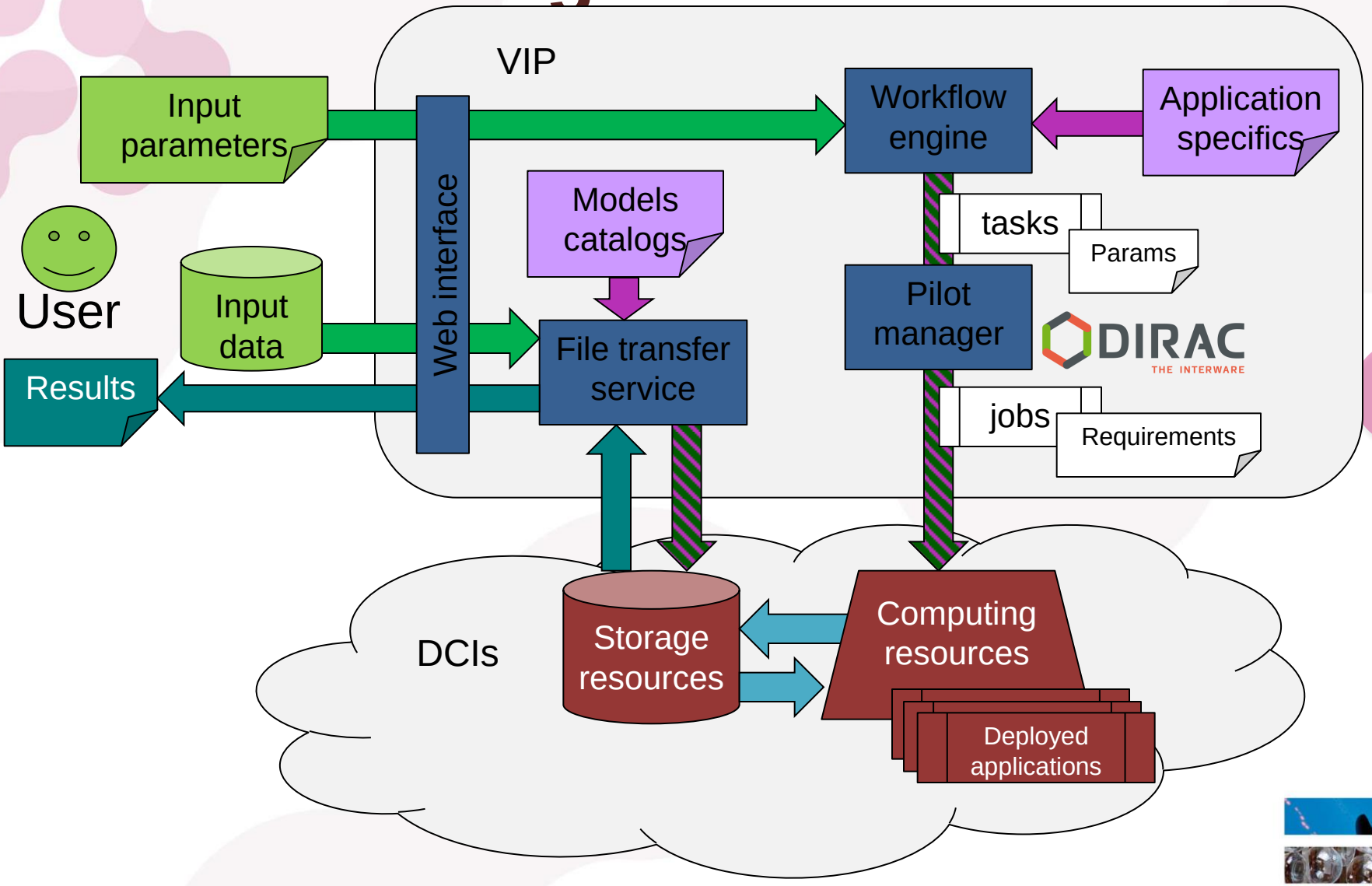
- **Simulators**
  - Gate/GateLab
  - Field-II
  - SIMRI
  - Sindbad
  - PET-Sorteo
- **Analysis tools**
  - Freesurfer
  - FSL
- **Other tools (file transfer, monitoring, documentation…)**

VIP – Resources usage strategy

- **Use of existing DCIs (EGI mainly)**
  - Application deployement through biomed VO
  - Opportunistic use of available resources

- **Optimization**
  - Concept of pilot job through DIRAC
  - Automatic task replication
  - Dynamic parallelisation (Gate only)
  - Merging strategies (Gate only)

VIP usage: details



48

Des exemples d'utilisation

- http://succes2013.sciencesconf.org/
- http://www.egi.eu/case-studies/medical/

Questions ?

50