Red española de e-Ciencia

Meta Scheduling and Advanced Application Support on the Spanish NGI

Enol Fernández del Castillo (IFCA-CSIC)

OGF 25/EGEE User Forum

Catania, March 2nd 2009

Acción financiada por:

GOBIERNO
DE ESPAÑA

Entidad Coordinadora:

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Outline

Red Española de e - Ciencia

1. Spanish NGI
2. Metaschedulers
   - Grid-Way
   - CrossBroker
3. Advanced Application Support
   - Interactivity
   - MPI
4. Summary

Spanish NGI: NGI-ES

Red Española de e - Ciencia

- National Level entity which operates a general purpose e-science infrastructure
- Objectives:
  - Establish a collaboration framework between all participating institutions to foster a coordinated development of a Grid infrastructure in Spain
  - Propose a sustainable design of the Grid infrastructure that covers the ecosystem of different Grid projects, computing centers, grid infrastructures, etc…
  - Run central services to keep up the infrastructure

Components of NGI-ES

Red Española de
e - Ciencia

- **Core: Spanish institutions participating in Grid research and development projects:**
  - **EGEE, EUFORIA, DORII, EELA, i2g**, … with a common middleware based on gLite
  - The new infrastructure **GRID – CSIC http://www.grid.csic.es**
  - **Universities, computing centers** with Globus Toolkit 4 middleware
  - **RedIris** (Spanish NReN) support EUGRIDPMA certificates.
- Relation with the **Spanish Supercomputing Network (RES)**
  - This network comprises several Spanish research centers that operate a common infrastructure of supercomputing.
  - Analyze possibility of mixed workflows between both infrastructures
- Relation with the **Portuguese NGI in the framework of Ibergrid**

NGI-ES Interoperability

Red Española de e - Ciencia

- Interoperability is needed for the sustainability of the NGI
  - Key issue for the creation of the EGI infrastructure
  - Allows users to select the resources that better fit their necessities
  - Potential access to a significantly larger set of resources
- Reduced management overheads if only a single Grid middleware system needs deployment on each site

NGI-ES Architecture

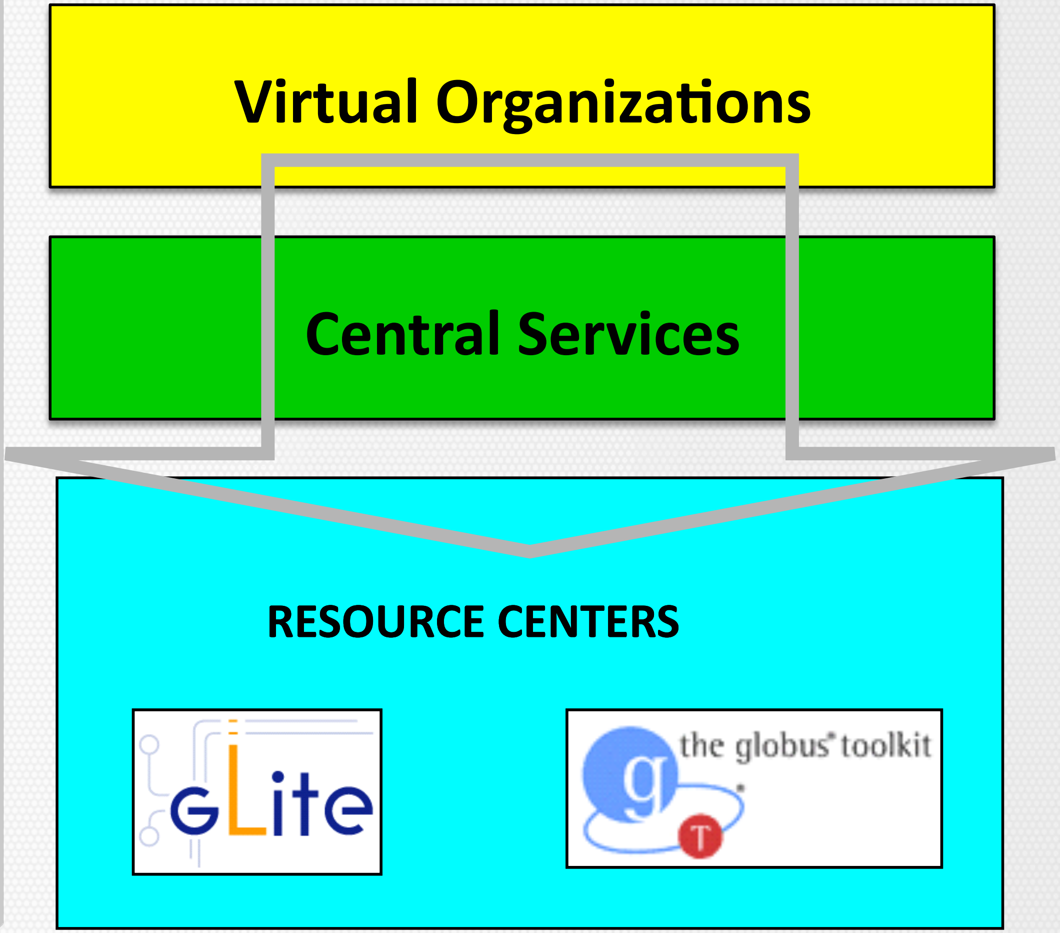
Red Española de
e - Ciencia

VO Oriented

The architecture of NGI-ES is oriented towards the support of Virtual Organizations

Key Issues

- ✓ Advanced VO Services
  - ✓ User Support
  - ✓Monitoring & Accounting
- ✓ Application porting and support
- ✓ Middleware driven by applications requirements

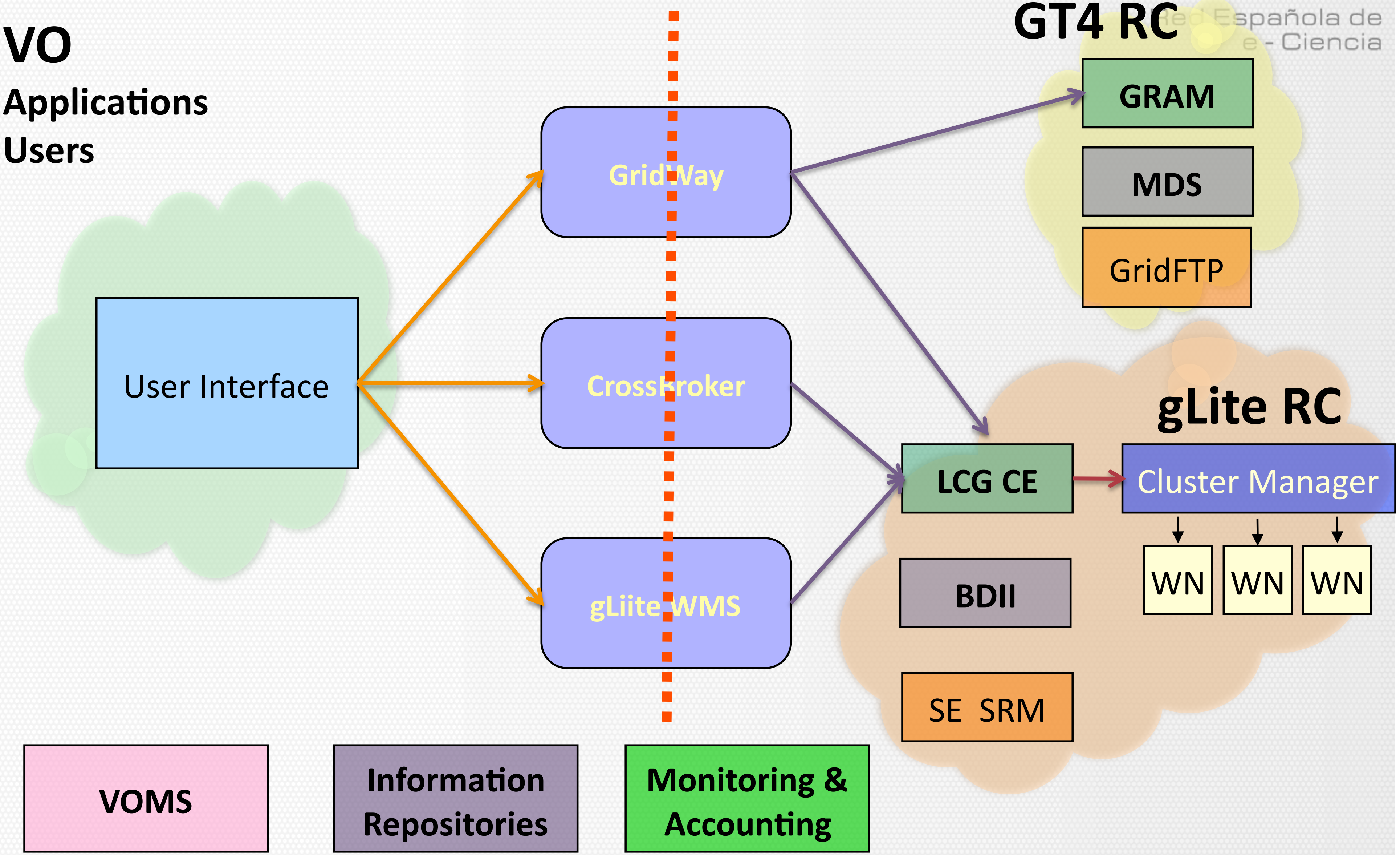


Virtual Organizations

Central Services

RESOURCE CENTERS

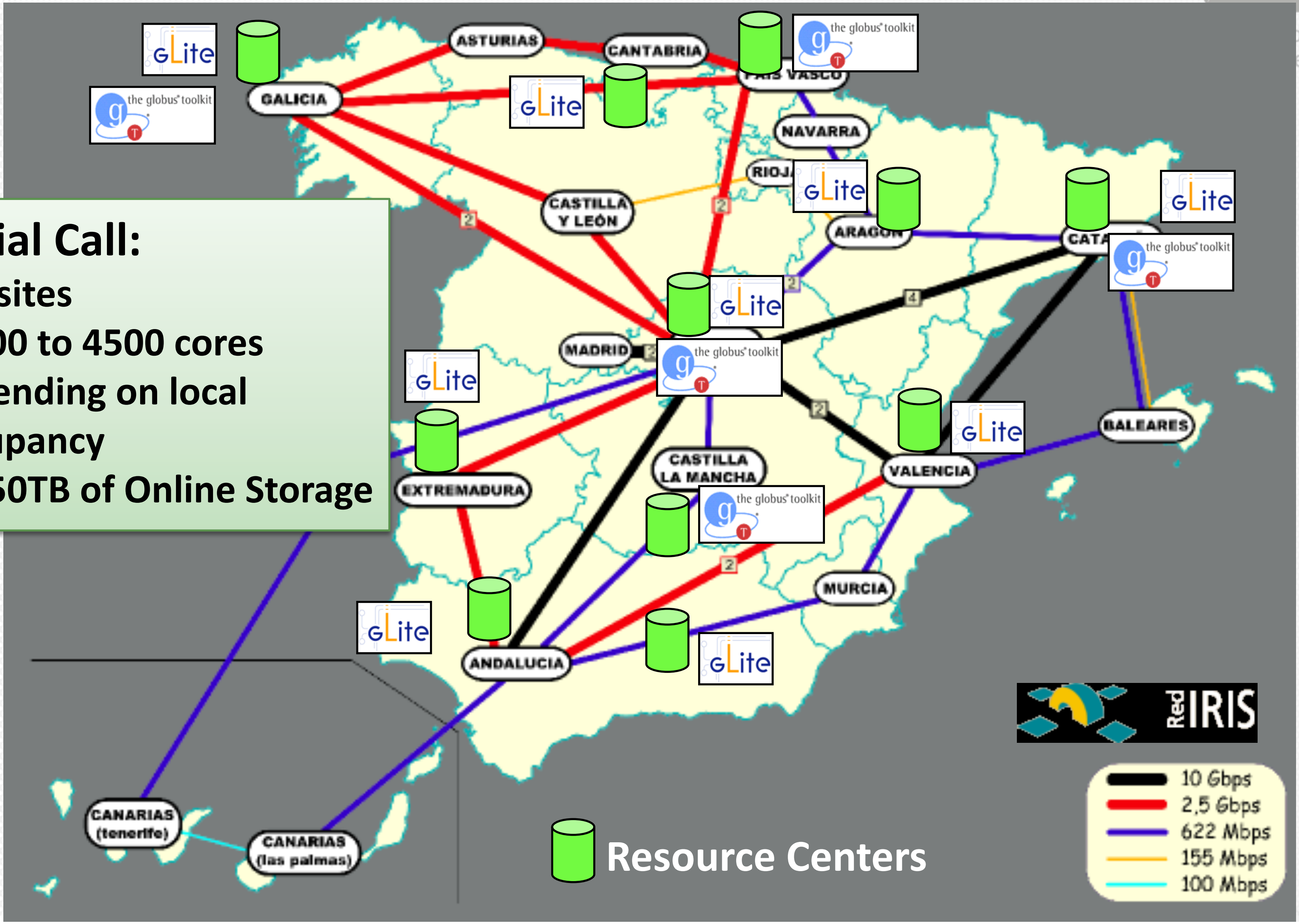
NGI-ES Architecture



Resource Centers Map

Initial Call:

- 18 sites
- 1300 to 4500 cores depending on local occupancy
- ~350TB of Online Storage



Resource Centers

Deployed Services

Red Española de e - Ciencia

- **Monitorization and accounting services at Cesga**
  - Accounting Portal: http://www.ngi.cesga.es/gridsite/accounting
  - Monitorization Portal: http://rnagios.ngi.cesga.es/nagios
- **Global Information Repositories:**
  - OpenLDAP + GlueSchema server at IFCA, integrated with NGI-PT
  - Global MDS server at RedIRIS for GT4 resources
- **Metaschedulers:**
  - ***GridWay*** for the NGI at RedIRIS
  - ***Crossbroker*** for gLite resources at IFCA
  - ***gLite WMS*** for bulk submission of serial jobs at IFIC
- **VOMS server** at IFCA
- **HelpDesk** at RedIRIS

NGI-ES Added Value

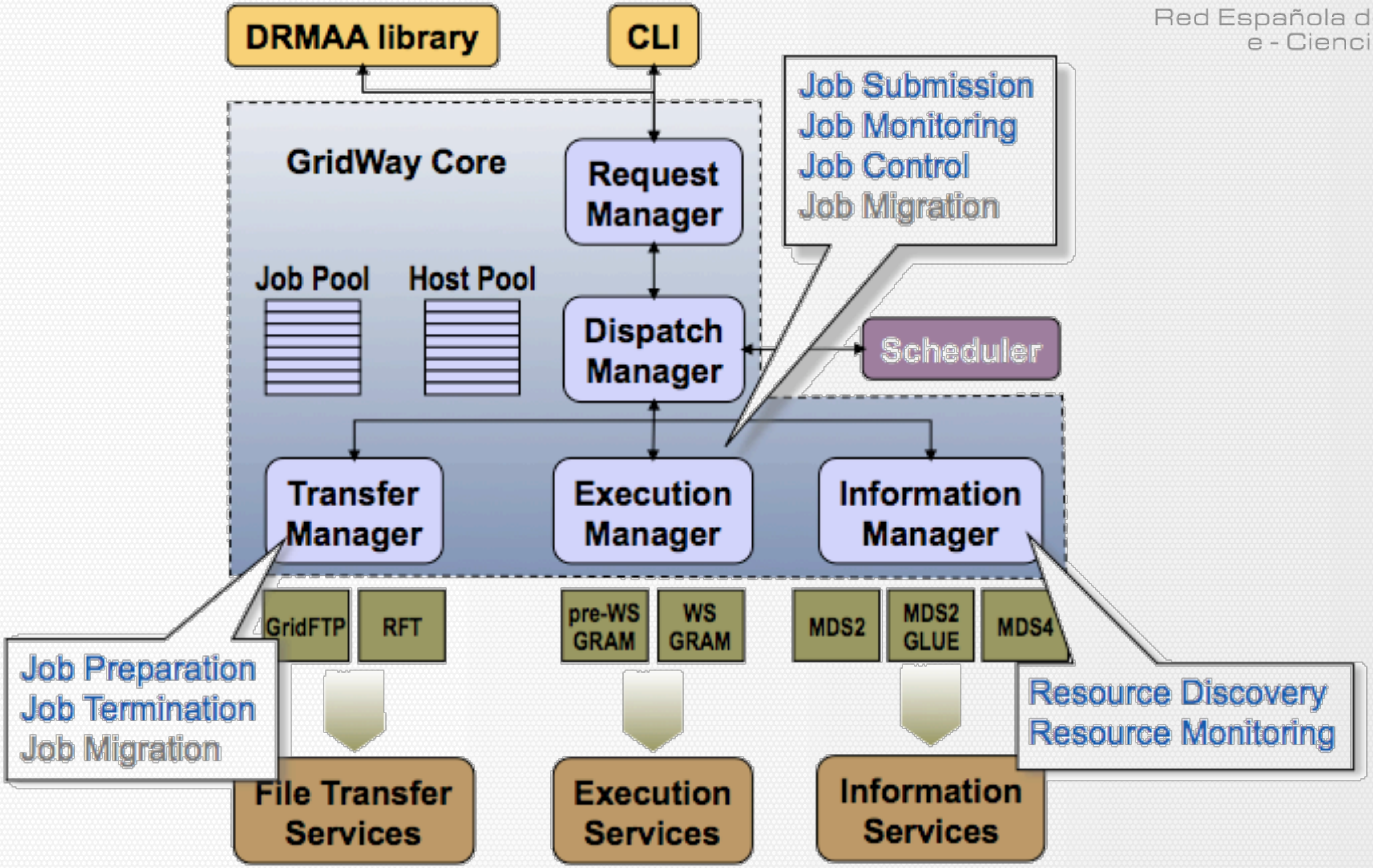
Red Española de
e - Ciencia

- Interoperability:
  - Resources from EU projects based on gLite (EGEE, I2G, EELA, WLCG, DORII)
  - Globus Toolkit 4 resources
- Metaschedulers developed by Spanish groups
  - GridWay
  - CrossBroker
- Advanced Application Support
  - Interactivity
  - MPI

Metaschedulers: GridWay

- GridWay allows the efficient use of computing resources of a Grid
  - Included in the Globus distribution
  - Support for both gLite and GT4 resource
- Used by several grid projects and initiatives worldwide
- Developed by UCM
- More information at: http://www.gridway.org

Metaschedulers: GridWay



Metaschedulers: CrossBroker

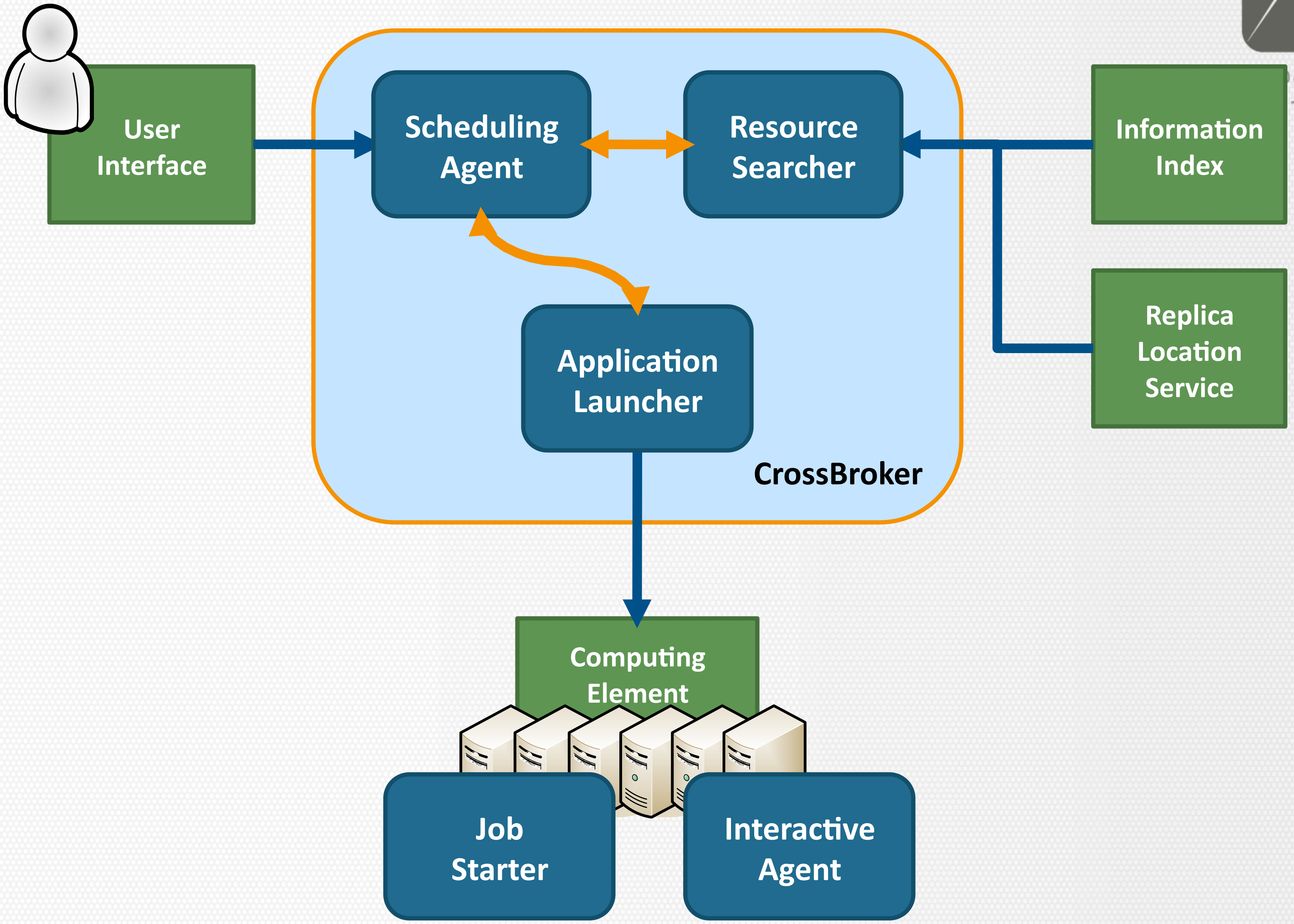
Red Española de e - Ciencia

- CrossBroker provides support for Parallel and Interactive jobs
  - High priority treatment for interactive jobs with the use of multiprogramming
  - Interoperable with EGEE, provides same services than gLite WMS

- Used in production environments
  - Used in EU CrossGrid, int.eu.grid and Euforia projects (12K – 55K jobs per month)

- Developed by UAB + CSIC
- More info: http://www.oliba.uab.es/crossbroker

Metaschedulers: CrossBroker



Interactivity Support

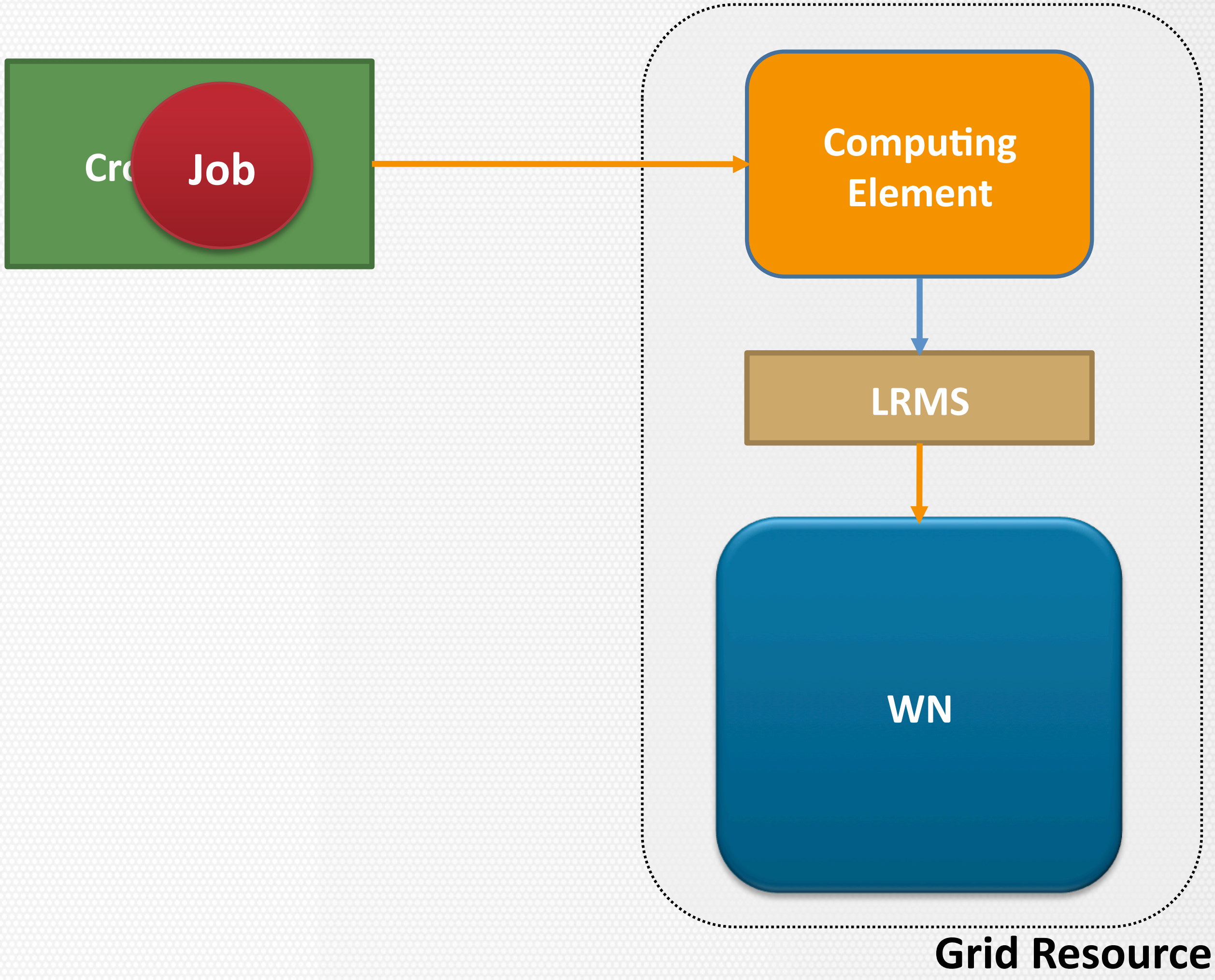
- Interactivity allows researchers to visualize results and obtain them faster
- Requirements:
  - Fast startup: the possibility of starting the application immediately, even in high occupancy scenarios
  - Online Input-Output streaming: the ability to have application input and output online.

Interactivity Support: Multiprogramming

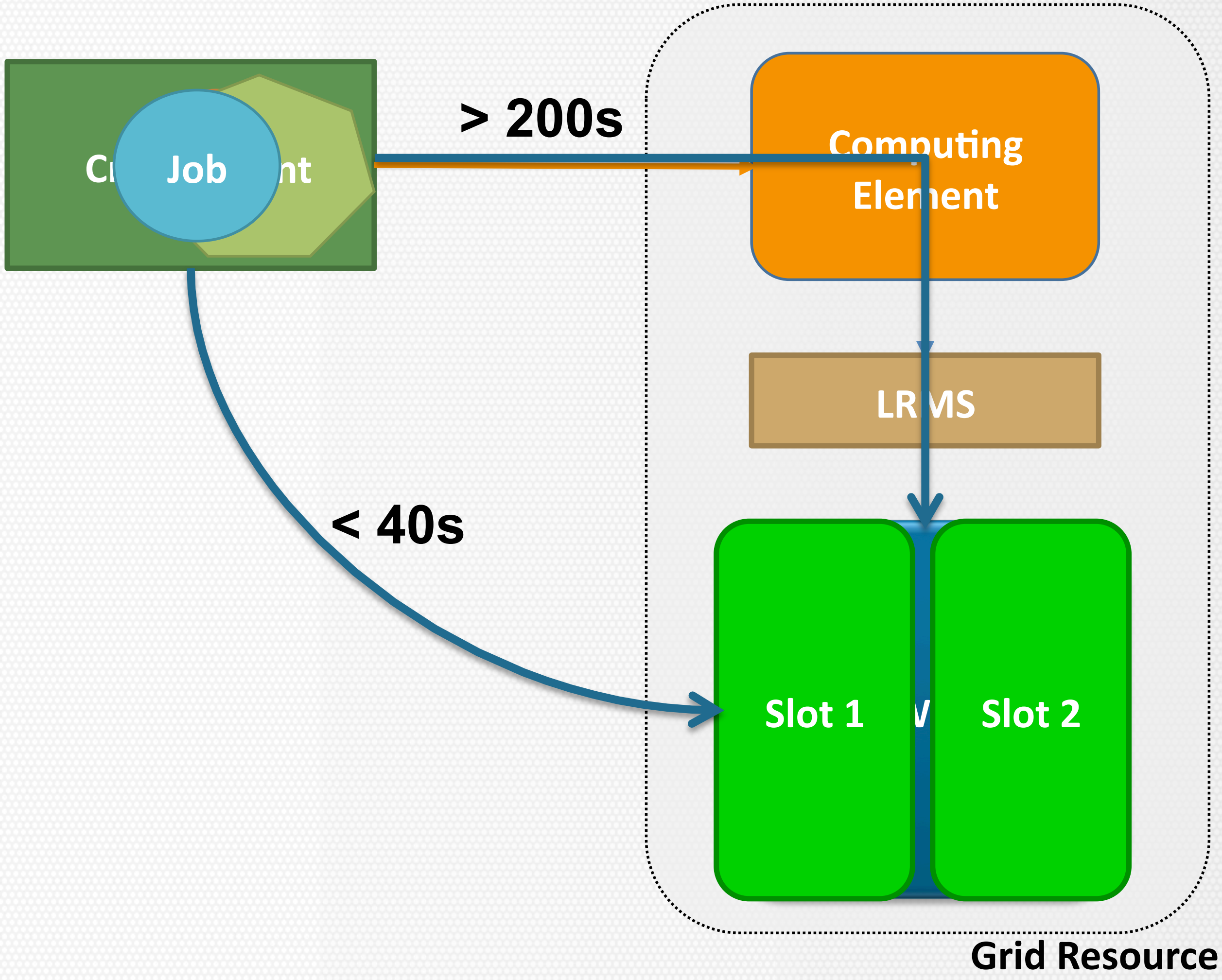
- The idea
  - Each job is encapsulated in an agent that takes control over the WN independently of its LRMS

- Lightweight "Virtual Machines"
  - Each Worker Node is divided in 2 execution slots
  - Each VM can execute jobs independently (e.g. batch and interactive)
  - NOT a full virtual machine (Xen, VMWare,…)
  - NO need for special priviledges in the WN

Interactivity Support: Multiprogramming

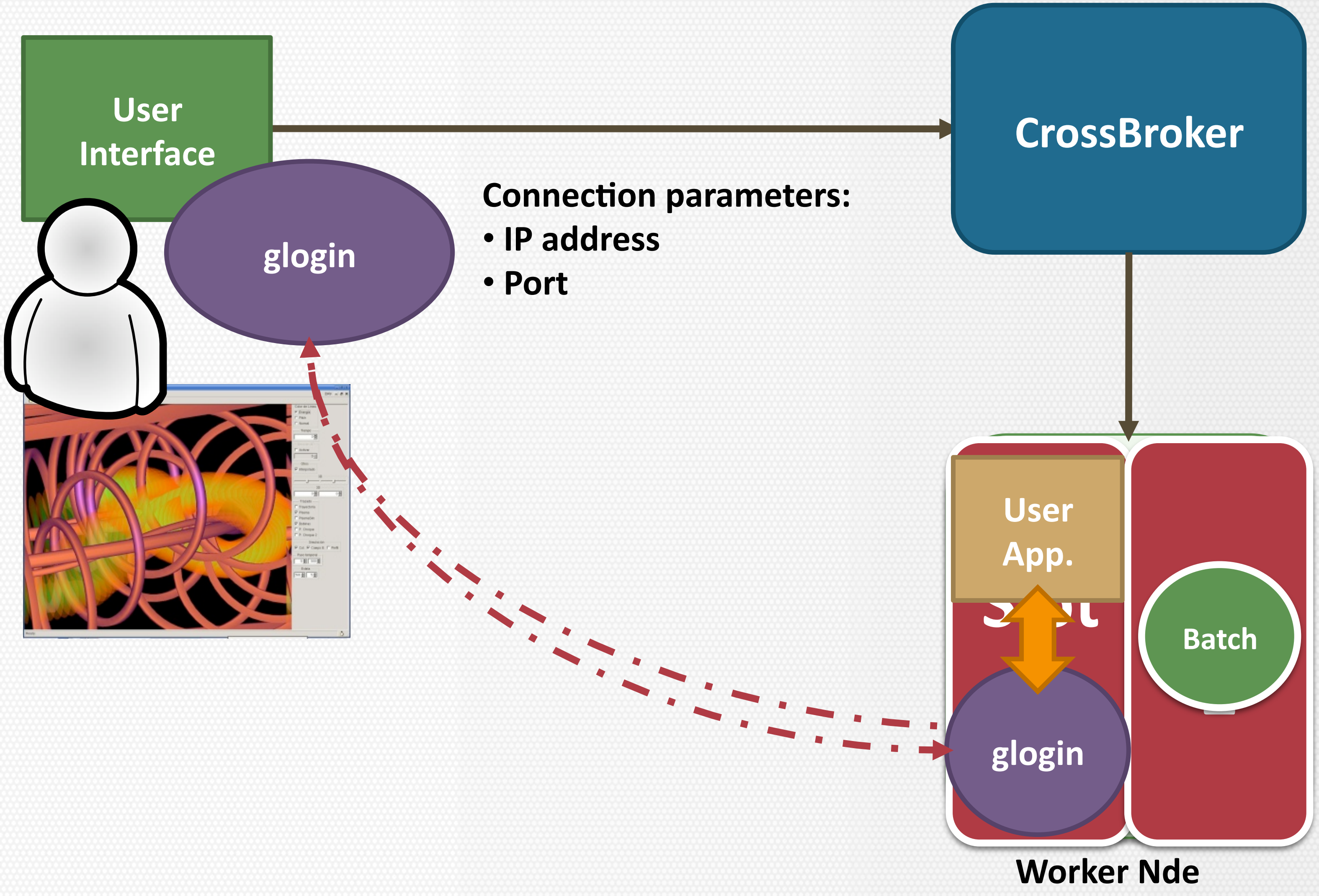


Interactivity Support: Multiprogramming



Interactivity Support: Interactive Agents

Red Española de e - Ciencia



Connection parameters:

- IP address
- Port

MPI Support

- Many application areas need MPI support
  - Earth Sciences, Biological sciences, Computational Chemistry, Nuclear Fusion,
  - Representative results can be obtained by using order of 10s-100s of CPUs
- Many clusters are MPI – ready
  - In local mode by direct submission
  - Shared filesystems with high performance intranet
- It is interesting to offer this capability when the user is working inside a Grid infrastructure
  - As an infrastructure on its own
  - As a testbed for small runs before executing on large HPC

MPI Support: Issues

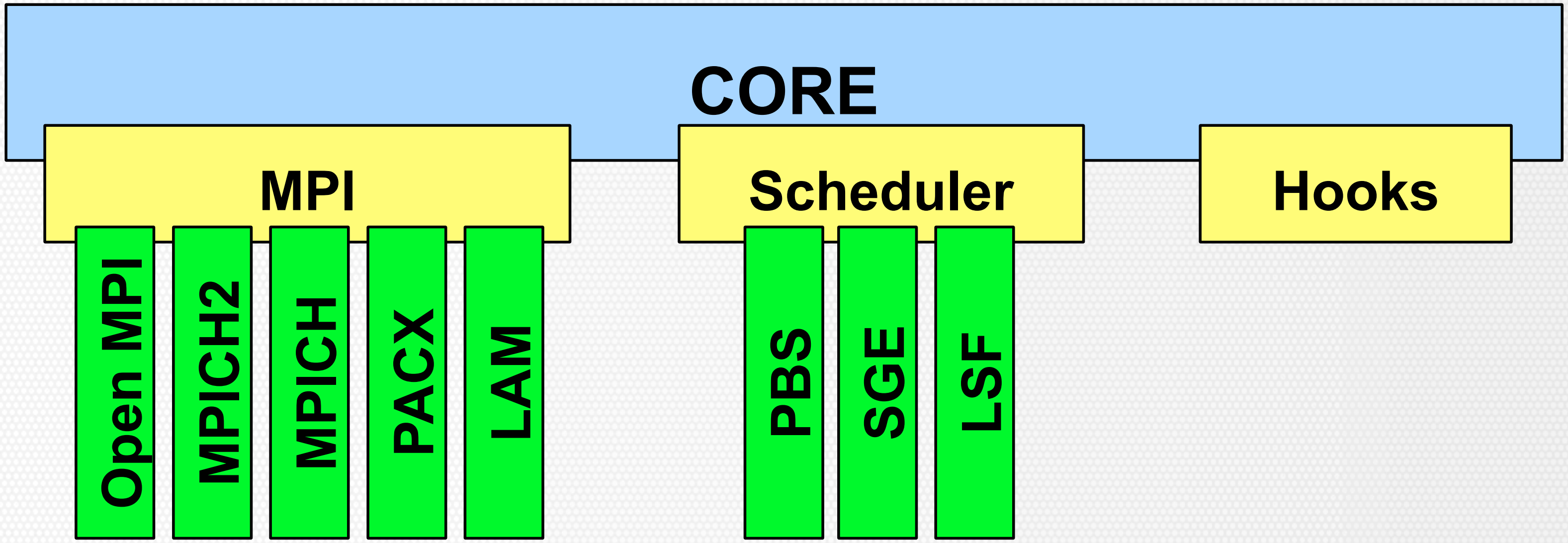
Red Española de
e - Ciencia

- There is no standard way of starting an MPI app
  - No common syntax for mpirun, mpiexec support optional
  - Schedulers (PBS, SGE, Condor…) handle machinefiles in different ways
  - Metascheduler services need to handle different implementations (OpenMPI, MPICH, LAM…) in a flexible and extensible way
  - Shared vs. Non-Shared filesystems

MPI Support: MPI-Start

Red Española de e - Ciencia

- Specifies a unique interface to the upper layer in the middleware to describe MPI jobs
- Support basic file distributions
- Implemented as portable shell scripts
- Extensible via user hooks and plugins at the site level



MPI Support: MPI-Start

- MPI-Start is used by the CrossBroker to support:
  - Intra-cluster apps with OpenMPI or MPICH
  - Inter-cluster apps using PACX-MPI or MPICH-G2
- User does not need to specify low level details of jobs:
  - MPI-Start copies all the input files to the WN (if not shared FS)
  - MPI-Start invokes the mpi program using the best configuration for the site.

Summary

Red Española de
e - Ciencia

- NGI Grid Infrastructure is well active!
  - – Deployment of NGI infrastructure going on
  - – Integrates EGEE like resources (i2g, EUFORIA, DORII, EELA, …) and GT 4 resources
- Added values of NGI-ES
  - – Metaschedulers middleware
  - – Support to MPI and Interactive jobs
- Interest in European Grid Infrastructure
- More info: www.e-ciencia.es