Introduction to the EGEE project and to the EGEE Grid

Peter Kacsuk and Gergely Sipos

MTA SZTAKI

www.lpds.sztaki.hu

Acknowledgement

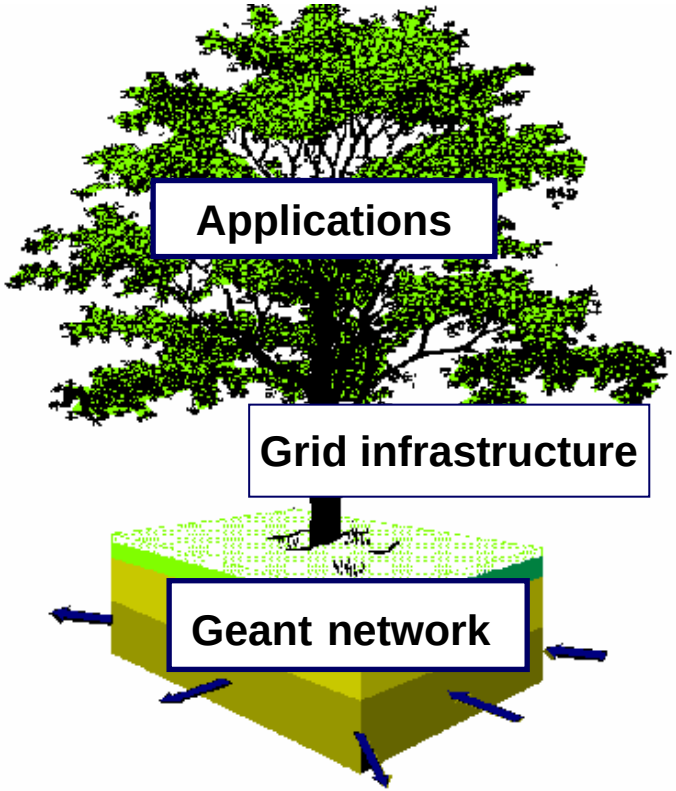
- **This tutorial is based on the work of many people:**
  - Fabrizio Gagliardi, Flavia Donno, Peter Kunszt
  - Riccardo Bruno, Marc-Elian Bégin, Martin Polak
  - the EDG developer team
  - the EDG training team
  - the NeSC training team
  - the SZTAKI training team

Contents

- Introduction to EGEE-I and EGEE-II
- Introduction to the EGEE middleware model
- Evolution of middleware
- Services in LCG
- Services in glite

What is EGEE ? (I)

- EGEE (Enabling Grids for E-sciencE) is a seamless **Grid infrastructure** for the support of scientific research, which:
  - Integrates current national, regional and thematic Grid efforts, especially in HEP (High Energy Physics)
  - Provides researchers in academia and industry with round-the-clock access to major computing resources, independent of geographic location



What is EGEE ? (II)

Main features of the EGEE-I project:

- 70 leading institutions in 27 countries, federated in regional Grids
- 32 M Euros EU funding (2004-5), O(100 M) total budget
- Aiming for a combined capacity of over 20’000 CPUs (the largest international Grid infrastructure ever assembled)
- ~ 300 dedicated staff

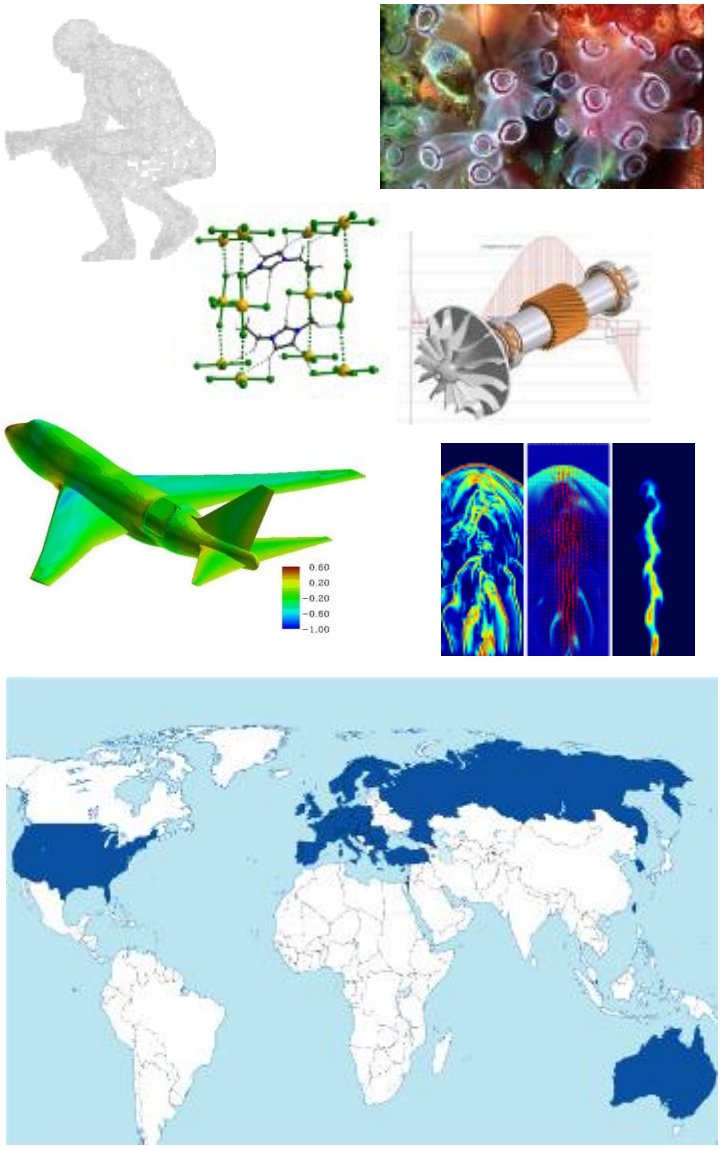
EGEE Community

Enabling Grids for E-sciencE

EGEE-II

- **EGEE-II proposal submitted to the EU**
  - On 8 September 2005
  - Started 1 April 2006

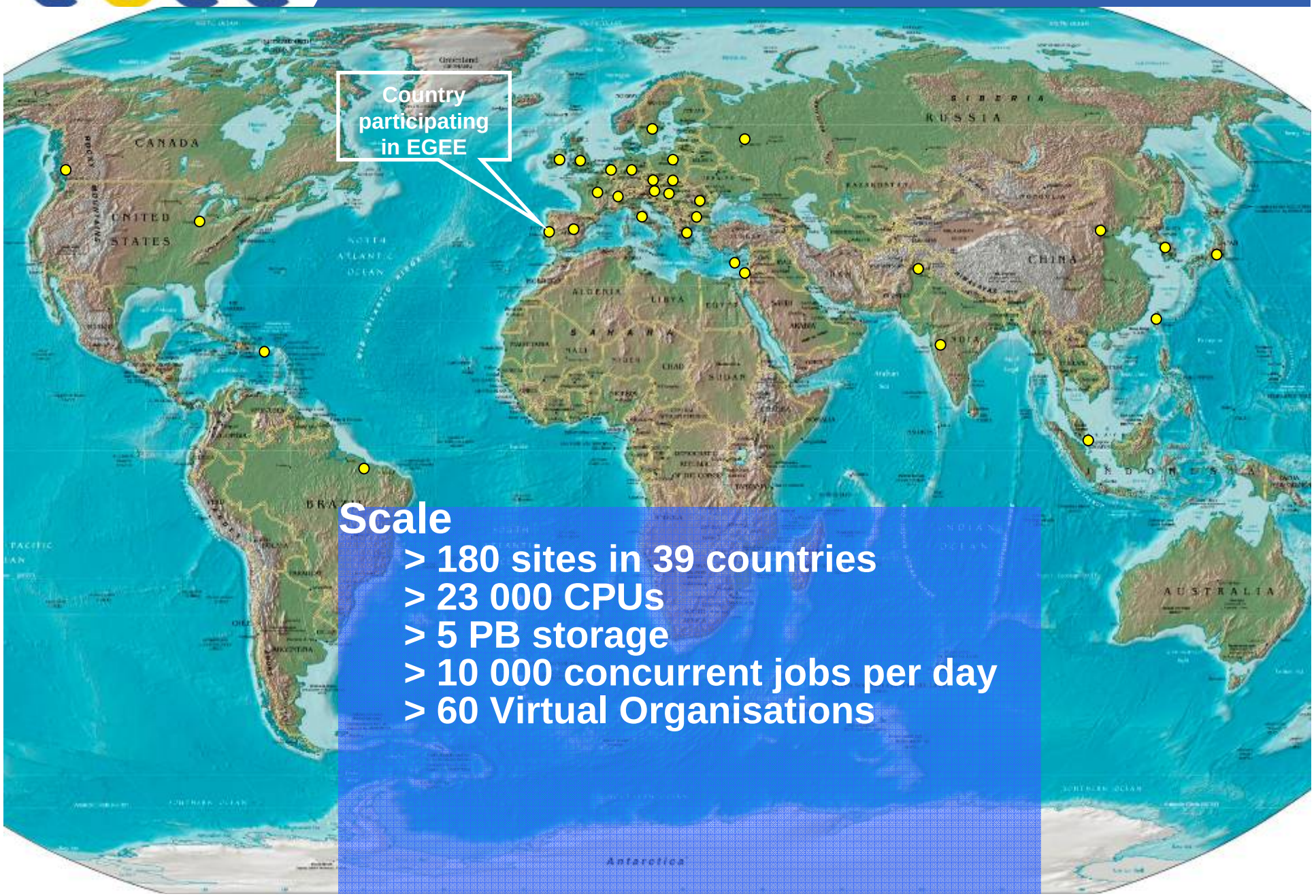
- **Natural continuation of EGEE**
  - Emphasis on providing an infrastructure for e-Science
    - → increased support for applications
    - → increased multidisciplinary Grid infrastructure
    - → more involvement from Industry
  - Expanded consortium
    - > 90 partners in 32 countries (Non-European partners in USA, Korea and Taiwan)
    - Related projects



→ **world-wide Grid infrastructure**

→ **increased international collaboration**

EGEE Infrastructure

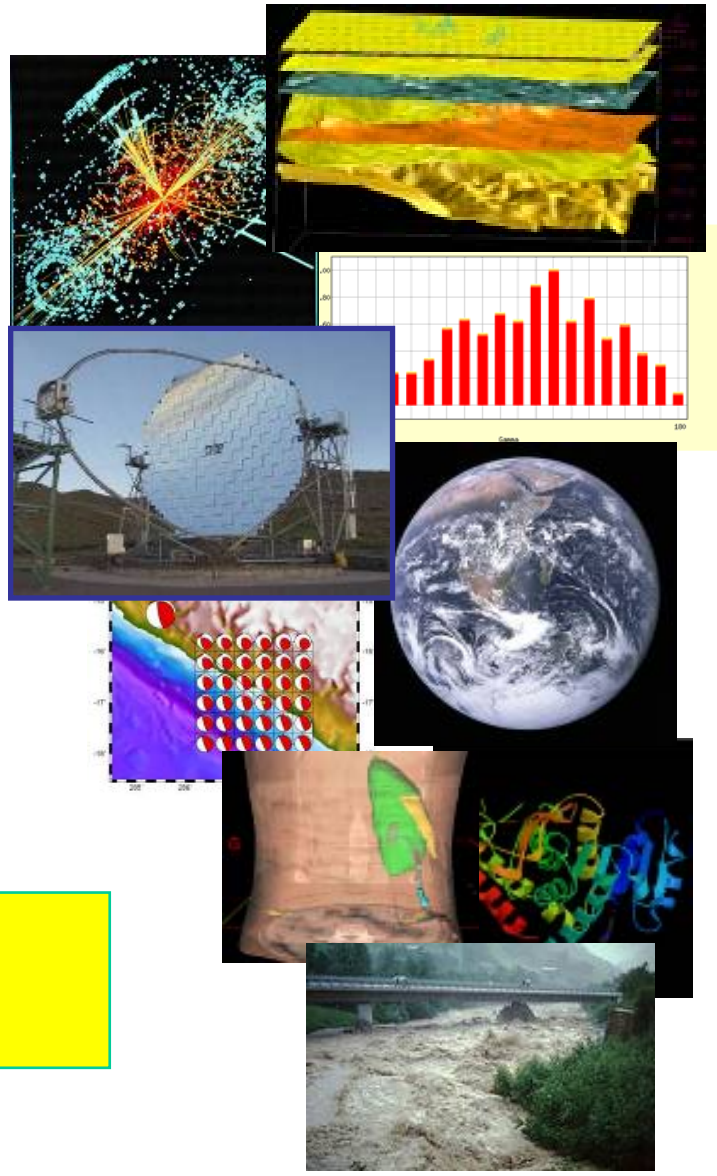


Scale

- > 180 sites in 39 countries
- > 23 000 CPUs
- > 5 PB storage
- > 10 000 concurrent jobs per day
- > 60 Virtual Organisations

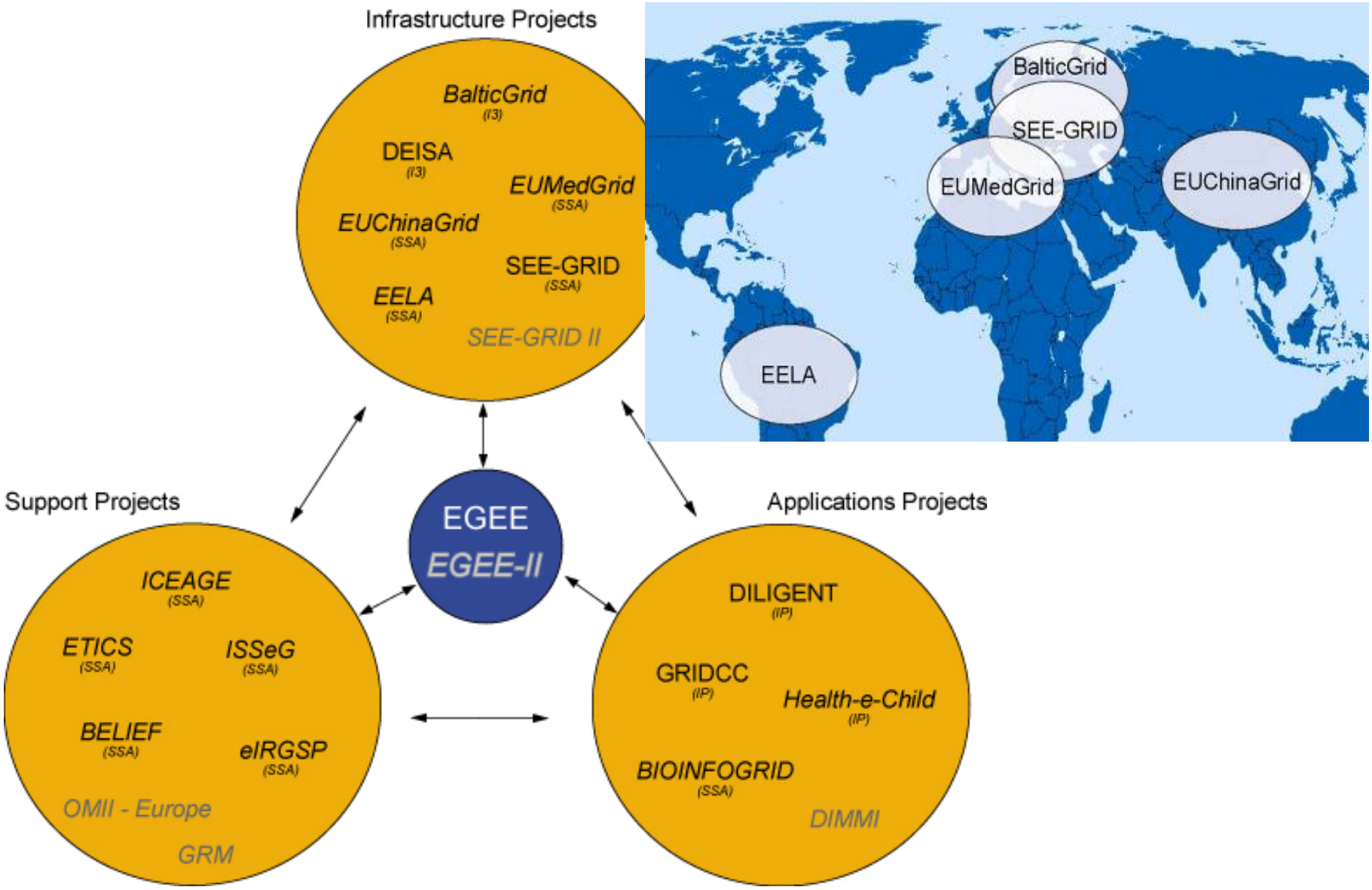
EGEE Applications

- **>20 applications from 7 domains**
  - High Energy Physics
  - Biomedicine
  - Earth Sciences
  - Computational Chemistry
  - Astronomy
  - Geo-Physics
  - Financial Simulation
- **Further applications in evaluation**

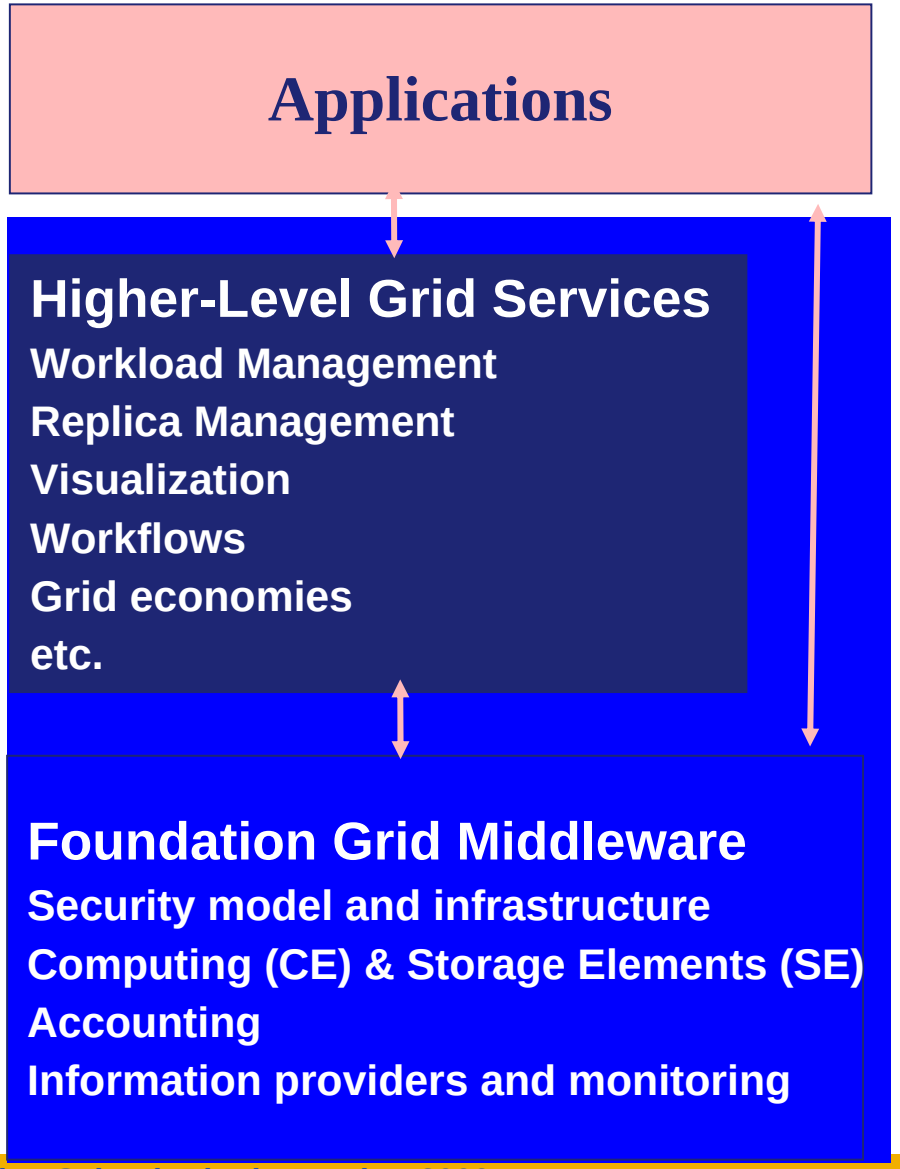


Applications now moving from testing to routine and daily usage

Related projects



Middleware in EGEE

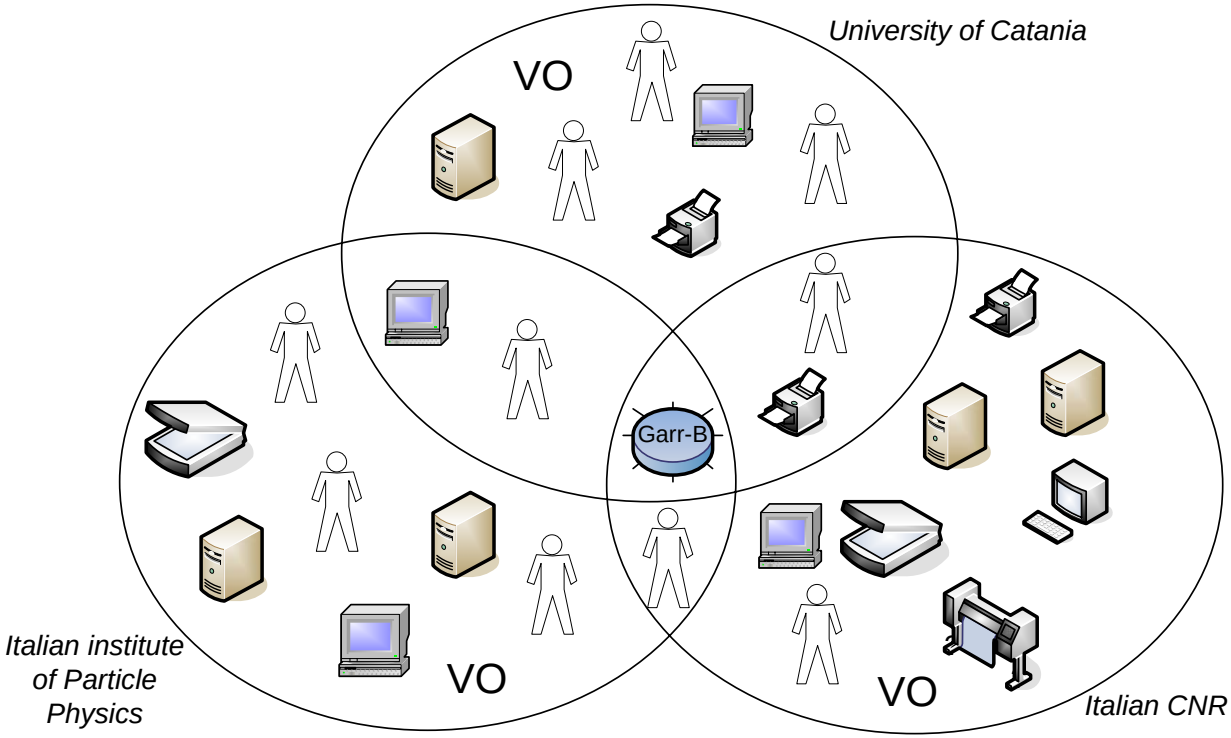


- Provide specific solutions for supported applications
- Host services from other projects
- More rapid changes than Foundation Grid Middleware
- Deployed as application software using procedure provided by grid operations

- Application independent
- Evaluate/adhere to new stds
- Emphasis on robustness/stability over new functionality
- Deployed as a software distribution by grid operations

gLite Middleware

- Many **VO**s need **sharing of resources** through services
  - **Accessing**
  - **Allocating**
  - **Monitoring**
  - **Accounting**



- **Grid Middleware** - Layer between services and physical resources
- **gLite** - Lightweight Middleware for Grid Computing

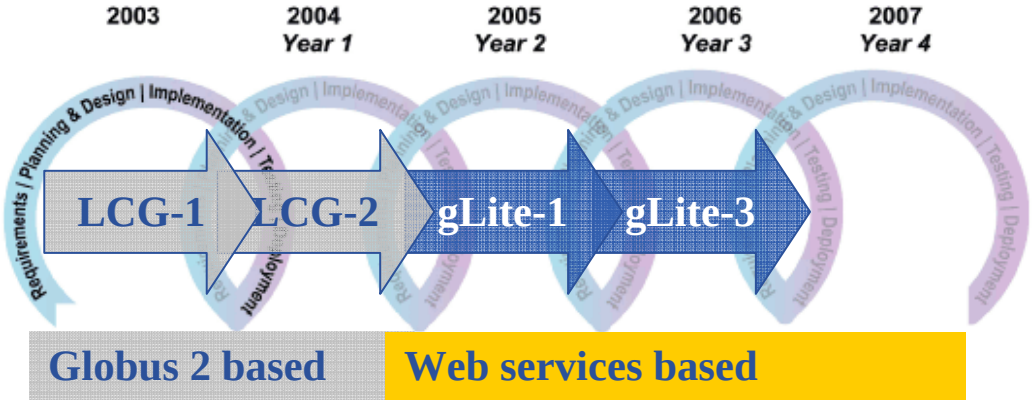
http://www.glite.org

gLite - Background

Other Grid Projects:

- Global Grid Forum - GGF
- Open Grid Services Architecture – OGSA
- EU DataGrid
- AliEn
- Globus
- Condor
- ...

LCG:



LCG-2

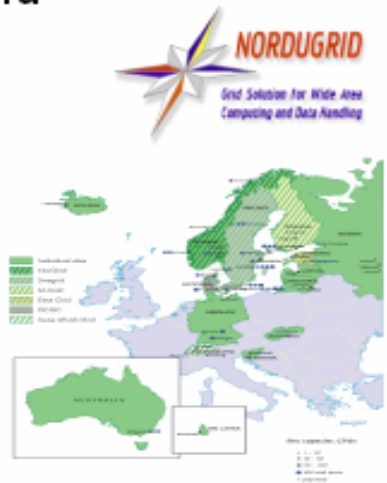
- **To understand the evolution of gLite**
  - Let's start with LCG and its terminology

What is LHC Grid (LCG)?

- LHC stands for Large Hadron Collider to be built by CERN

  http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/
- The LHC will be put in operation in 2007 with many experiments collecting 5-6 PetaB data per year
- The LHC Grid was built by CERN in order to provide storage and computing capacity for the process of this huge data set
- The LHC Grid current version is called **LCG-2**
- It was built based on the sw developed by the European DataGrid project and by the Gryphin US project
- LCG-2 was the first EGEE infrastructure

Grid Projects Collaborating in LHC Computing Grid



NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

Open Science Grid

ATLAS

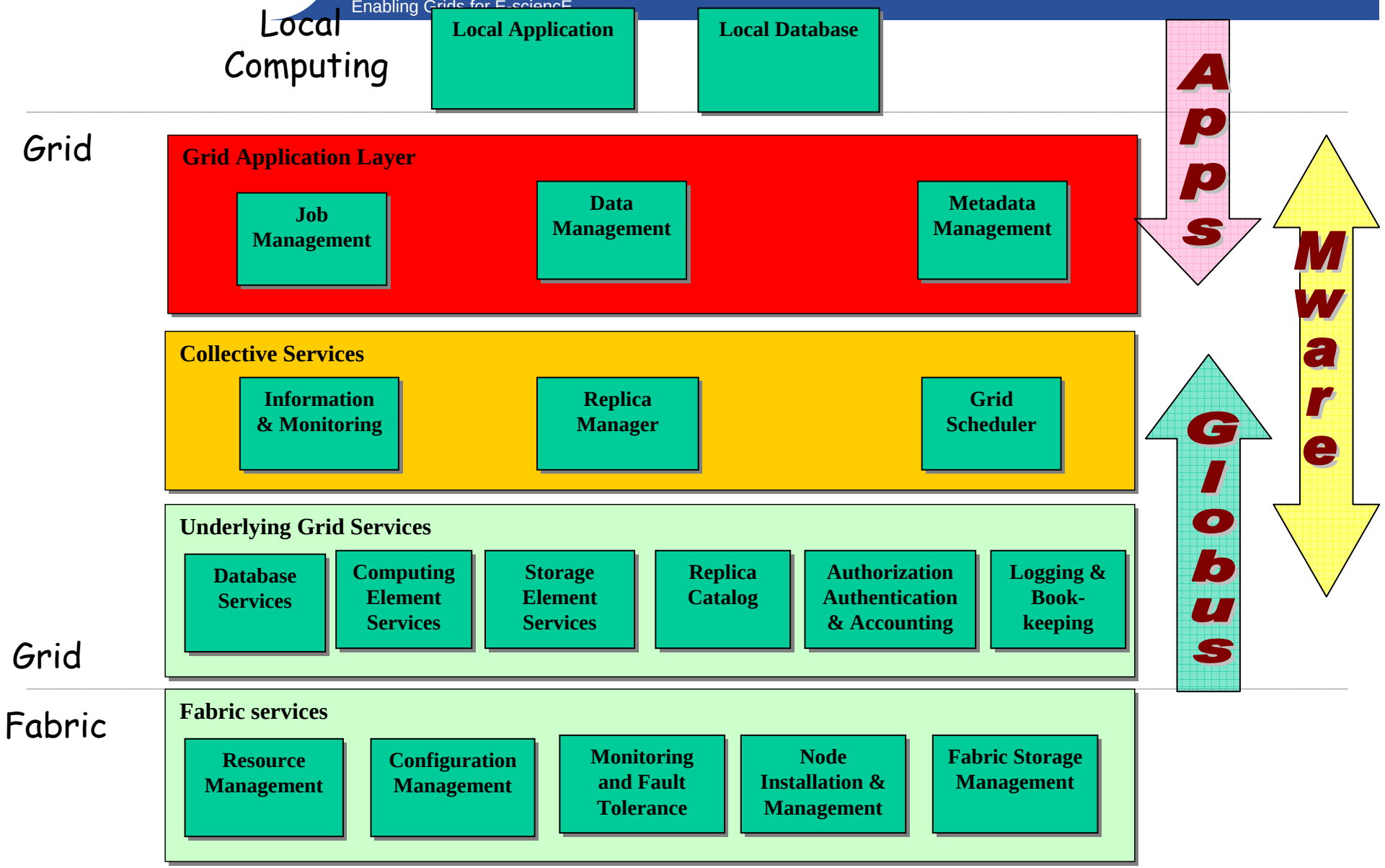
CMS

LHCb

EGEE Operations Information	
Active Sites	180
Available CPU	23317
Available Storage (TB)	102169

LastBuild:Mon Jun 26 09:16:01 BST 2006 GstatQuery:2006-06-26

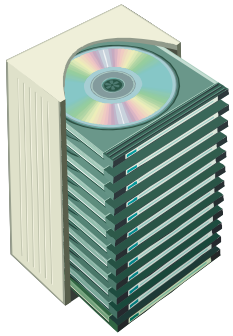
The LCG-2 Architecture



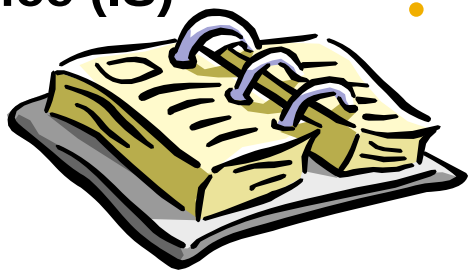
Main Logical Machine Types (Services) in LCG-2

- **User Interface (UI)**

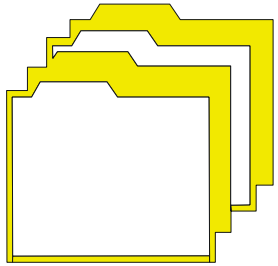
- **Storage Element (SE)**



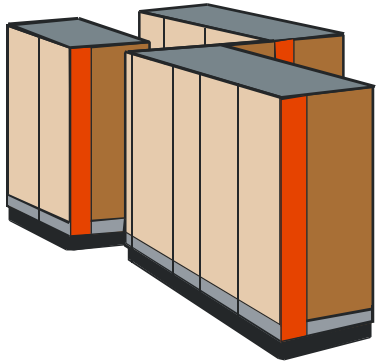
- **Information Service (IS)**



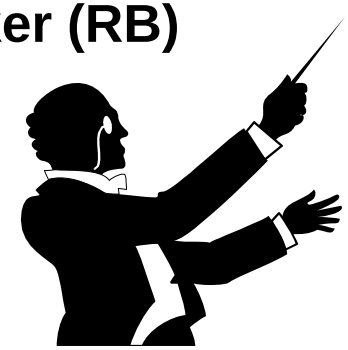
- **Replica Catalog (RC,RLS)**



- **Computing Element (CE)**
  - Frontend Node
  - Worker Nodes (WN)



- **Resource Broker (RB)**



User Interface

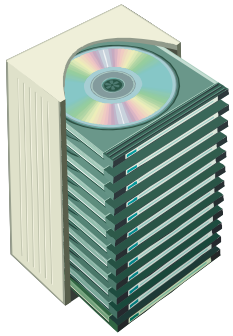
- **The initial point of access to the LCG-2 Grid is the User Interface**
- **This is a machine where**
  - LCG users have a personal account
  - The user's certificate is installed
- **The UI is the gateway to Grid services**
- **It provides a Command Line Interface to perform the following basic Grid operations:**
  - list all the resources suitable to execute a given job
  - replicate and copy files
  - submit a job for execution on a Computing Element
  - show the status of one or more submitted jobs
  - retrieve the output of one or more finished jobs
  - cancel one or more jobs

- **One or more UIs are available at each site part of the LCG-2 Grid**

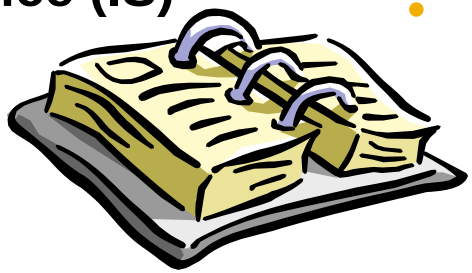
Main Logical Machine Types (Services) in LCG-2

- **User Interface (UI)**

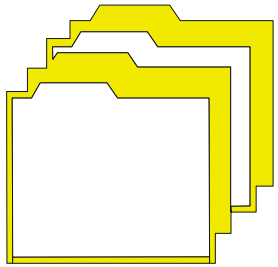
- **Storage Element (SE)**



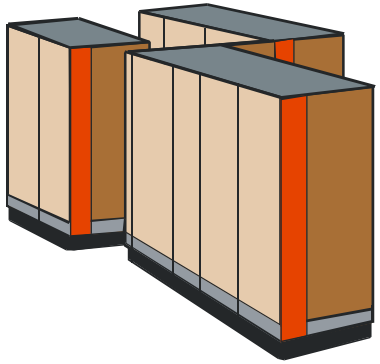
- **Information Service (IS)**



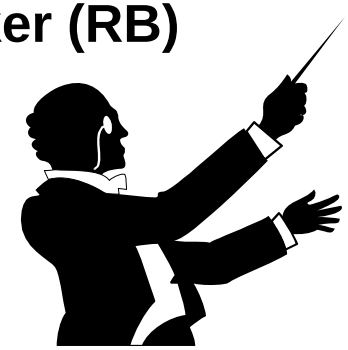
- **Replica Catalog (RC,RLS)**



- **Computing Element (CE)**
  - Frontend Node
  - Worker Nodes (WN)



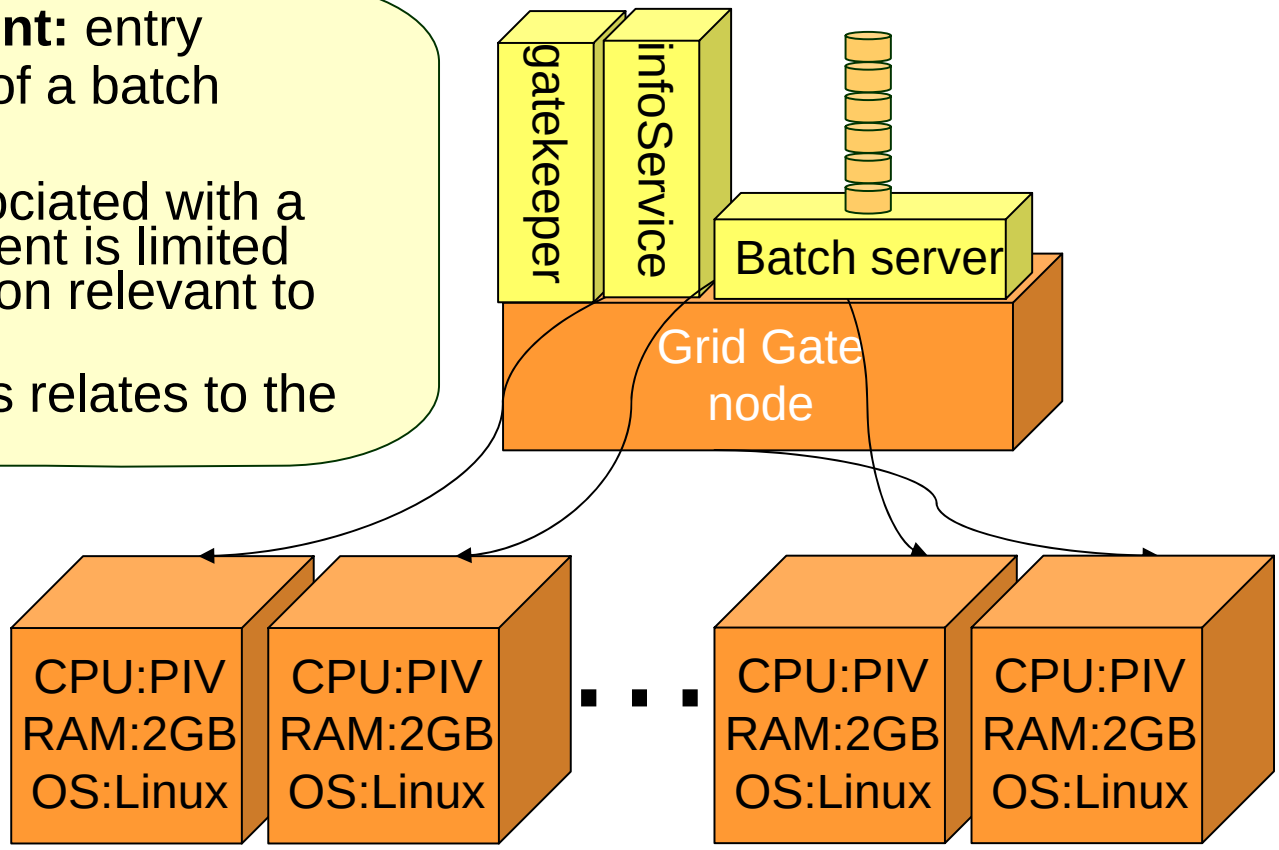
- **Resource Broker (RB)**



Computing Element

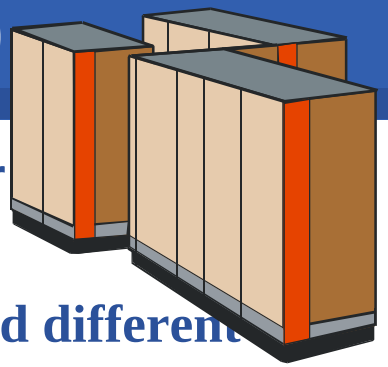
Computing Element: entry point into a queue of a batch system

- information associated with a computing element is limited only to information relevant to the queue
- Resource details relates to the system



A CE consist of homogeneous worker nodes

Computing Element (CE)



- **Defined as a Grid batch Queue and identified by a pair**

  <hostname>:<port>/<batch queue name>

- **Several queues defined for the same hostname are considered different. For example:**

  adc0015.cern.ch:2119/jobmanager-lcgpbs-long

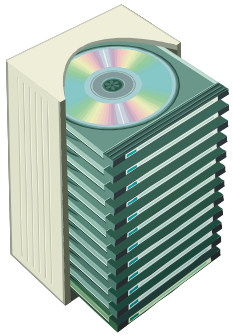
  adc0015.cern.ch:2119/jobmanager-lcgpbs-short

- **A Computing Element is built on a homogeneous farm of computing nodes (called Worker Nodes)**
- **One node acts as a *Grid Gate (GG)* or front-end to the Grid and runs:**
  - a Globus gatekeeper
  - the Globus GRAM (Globus Resource Allocation Manager)
  - the master server of a Local Resource Management System that can be:
    - PBS, LSF or Condor
  - a local Logging and Bookkeeping server
- **Each LCG-2 site runs at least one CE and a farm of WNs behind it.**

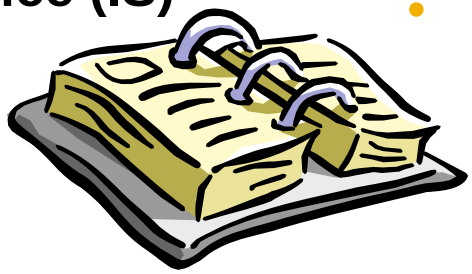
Main Logical Machine Types (Services) in LCG-2

- **User Interface (UI)**

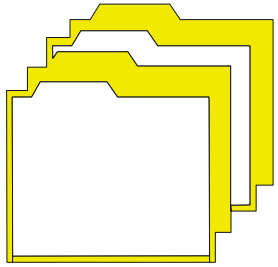
- **Storage Element (SE)**



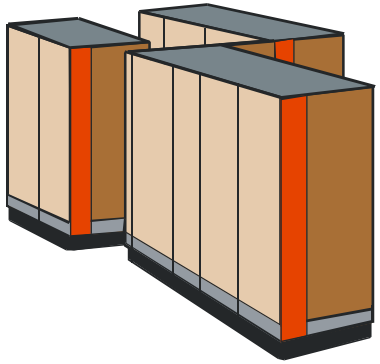
- **Information Service (IS)**



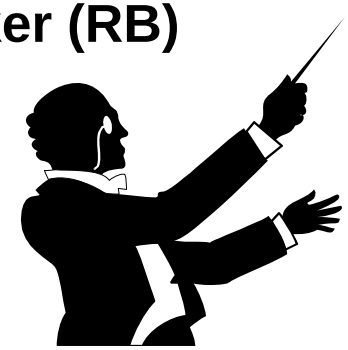
- **Replica Catalog (RC,RLS)**



- **Computing Element (CE)**
  - Frontend Node
  - Worker Nodes (WN)

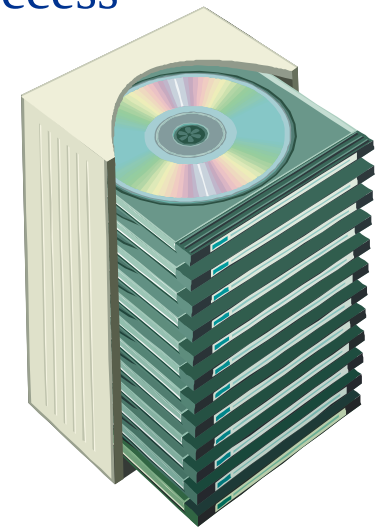


- **Resource Broker (RB)**



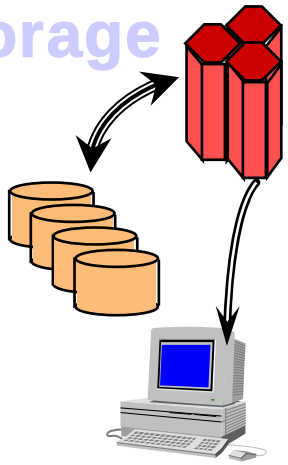
Storage Element (SE)

- **A *Storage Element (SE)* provides uniform access and services to large storage spaces.**
- **Each site includes at least one SE**
- **They use two protocols:**
  - ***GSIFTP*** for file transfer
  - ***Remote File Input/Output (RFIO)*** for file access



Storage Resource Management (SRM)

Data are stored on disk pool servers or Mass Storage Systems

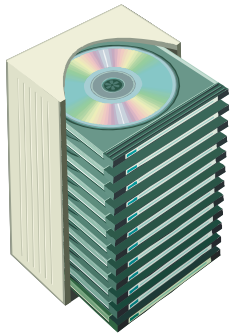


- storage resource management needs to take into account
  - Transparent access to files (migration to/from disk pool)
  - Space reservation
  - File status notification
  - Life time management
- SRM (Storage Resource Manager) takes care of all these details
  - SRM is a Grid Service that takes care of local storage interaction and provides a Grid interface to outside world

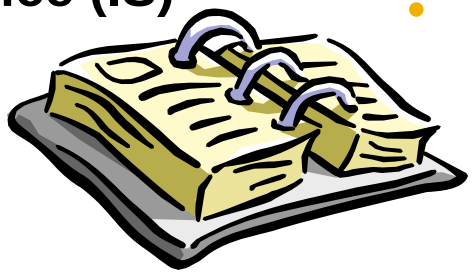
Main Logical Machine Types (Services) in LCG-2

- **User Interface (UI)**

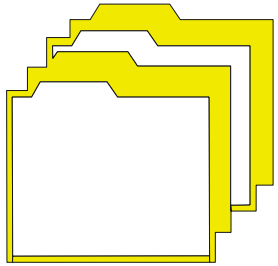
- **Storage Element (SE)**



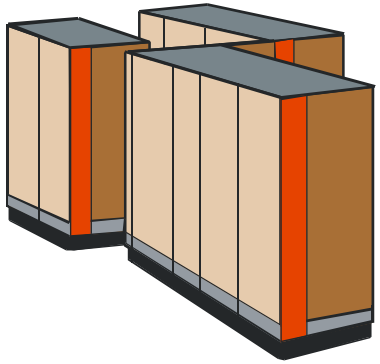
- **Information Service (IS)**



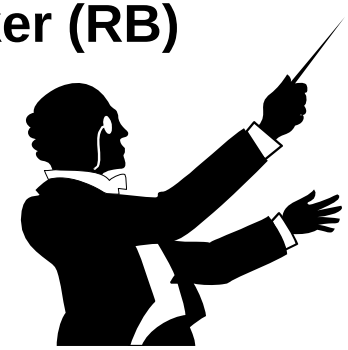
- **Replica Catalog (RC,RLS)**



- **Computing Element (CE)**
  - Frontend Node
  - Worker Nodes (WN)

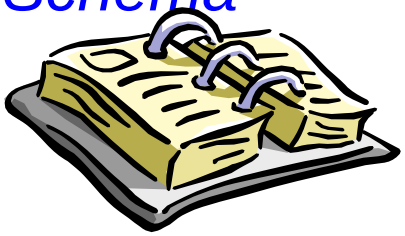


- **Resource Broker (RB)**



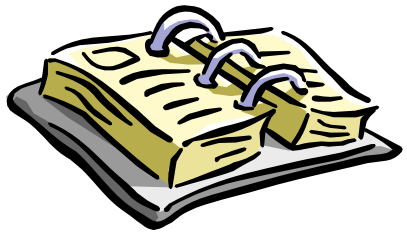
Information System (IS)

- **The Information System (IS) provides information about the LCG-2** Grid resources and their status
- **The current IS is based on** LDAP (Lightweight Directory Access Protocol)**: a directory service infrastructure which is a specialized database optimized for**
  - reading,
  - browsing and
  - searching information.
- **the LDAP schema used in LCG-2 implements the** *GLUE (Grid Laboratory for a Uniform Environment) Schema*

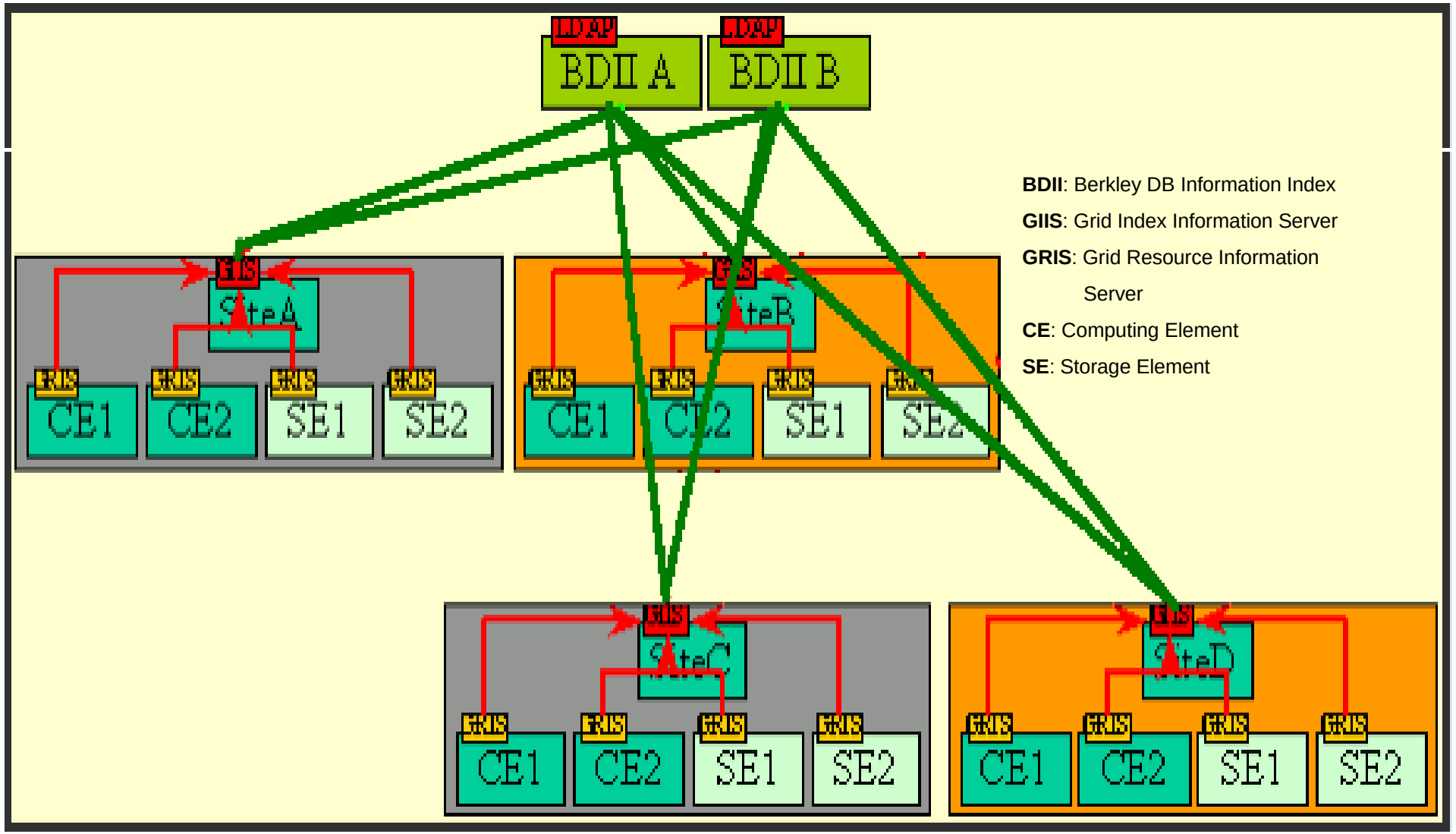


Information System (IS)

- **The IS is a hierarchical system with 3 levels from bottom up:**
  - **GRIS (*Grid Resource Information Servers)*** level (CE and SE level)
  - ***Grid Index Information Server (GIIS)*** level (site level)
  - Top, centralized level (Grid level)
- **the Globus *Monitoring and Discovery Service (MDS) mechanism* has been adopted at the GRIS level**
- **The other two levels use the *Berkeley DB Information Index (BDII) mechanism***

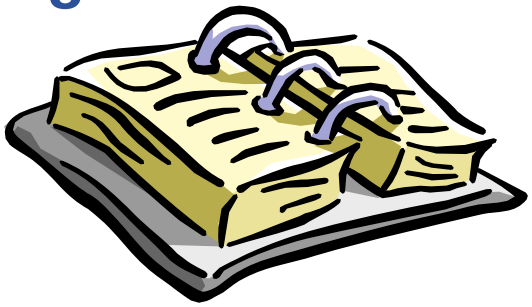


LCG-2 hierarchical Info system



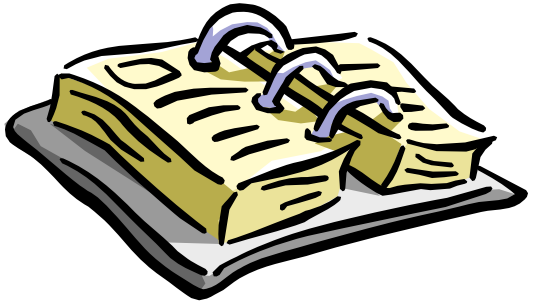
How to collect and store information?

- **All services are allowed to enter information into the IS**
- **The BDII at the top**
  - queries every GIIS in every 2 min and
  - acts as a cache storing information about the Grid status in its LDAP database
- **The BDII at the GIIS**
  - collects info from every GRIS in every 2 min and
  - acts as a cache storing information about the site status in its LDAP database
- **The GRIS updates information according to the MDS protocol**



How to obtain Information?

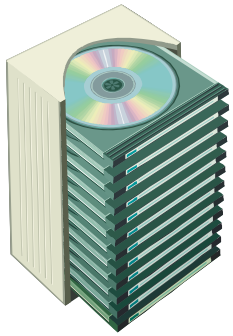
- **All users can browse the catalogues**
- **To obtain the information the client should:**
  - Ask BDII about possible GIIS/GRIS
  - Directly query GIIS/GRIS
  - Or use BDII cache
- **The IS scales to ~1000 sites (MDS much less: ~100)**



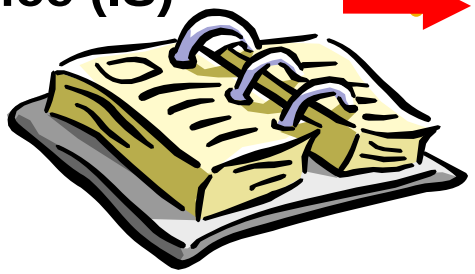
Main Logical Machine Types (Services) in LCG-2

- **User Interface (UI)**

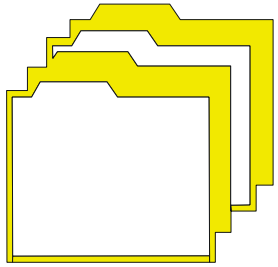
- **Storage Element (SE)**



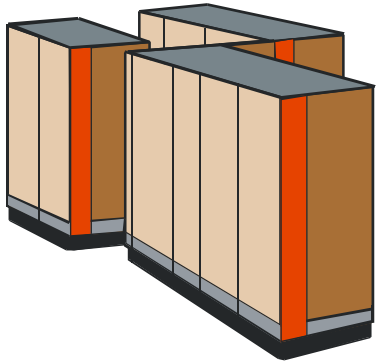
- **Information Service (IS)**



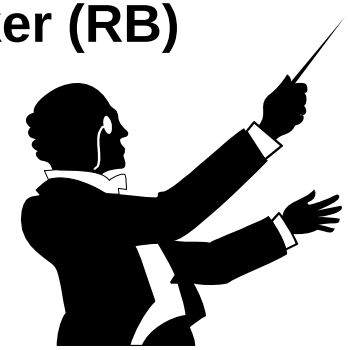
- **Replica Catalog (RC,RLS)**



- **Computing Element (CE)**
  - Frontend Node
  - Worker Nodes (WN)

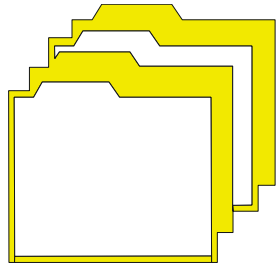


- **Resource Broker (RB)**



Data Management

- **In LCG, the data files are replicated:**
  - on a temporary basis,
  - to many different sites (depending on)
  - where the data is needed.

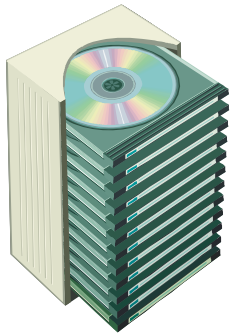


- **The users or applications do not need to know where the data is located, they use logical files names**
- **the Data Management services are responsible for locating and accessing the data.**

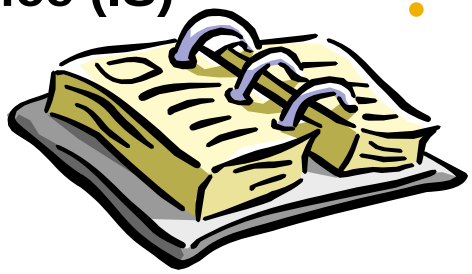
Main Logical Machine Types (Services) in LCG-2

- **User Interface (UI)**

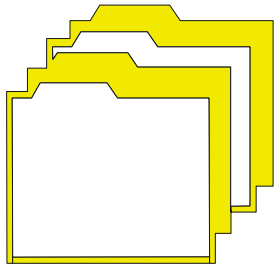
- **Storage Element (SE)**



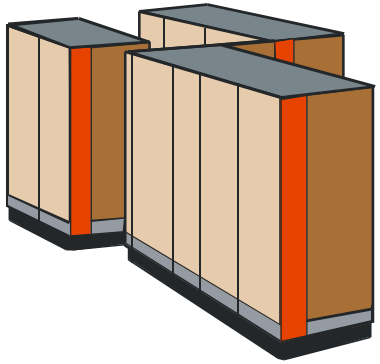
- **Information Service (IS)**



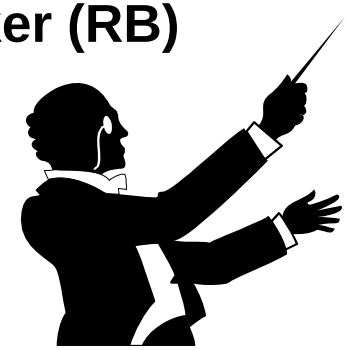
- **Replica Catalog (RC,RLS)**



- **Computing Element (CE)**
  - Frontend Node
  - Worker Nodes (WN)



- **Resource Broker (RB)**



Job Management

- **The user interacts with the Grid via a** Workload Management System (WMS)
- **The Goal of WMS is the** distributed scheduling and resource management in a Grid environment**.**
- **What does it allow Grid users to do?**
  - To submit their jobs
  - To execute them on the “best resources”
    - **The WMS tries to optimize the usage of resources**
  - To get information about their status
  - To retrieve their output

www

www.glite.org

gLite – SOA Services

Functions exposed as services with

- **Well-Defined**
- **Self-Contained**
- **Independent**
- **Message Based Interface**

Messaging

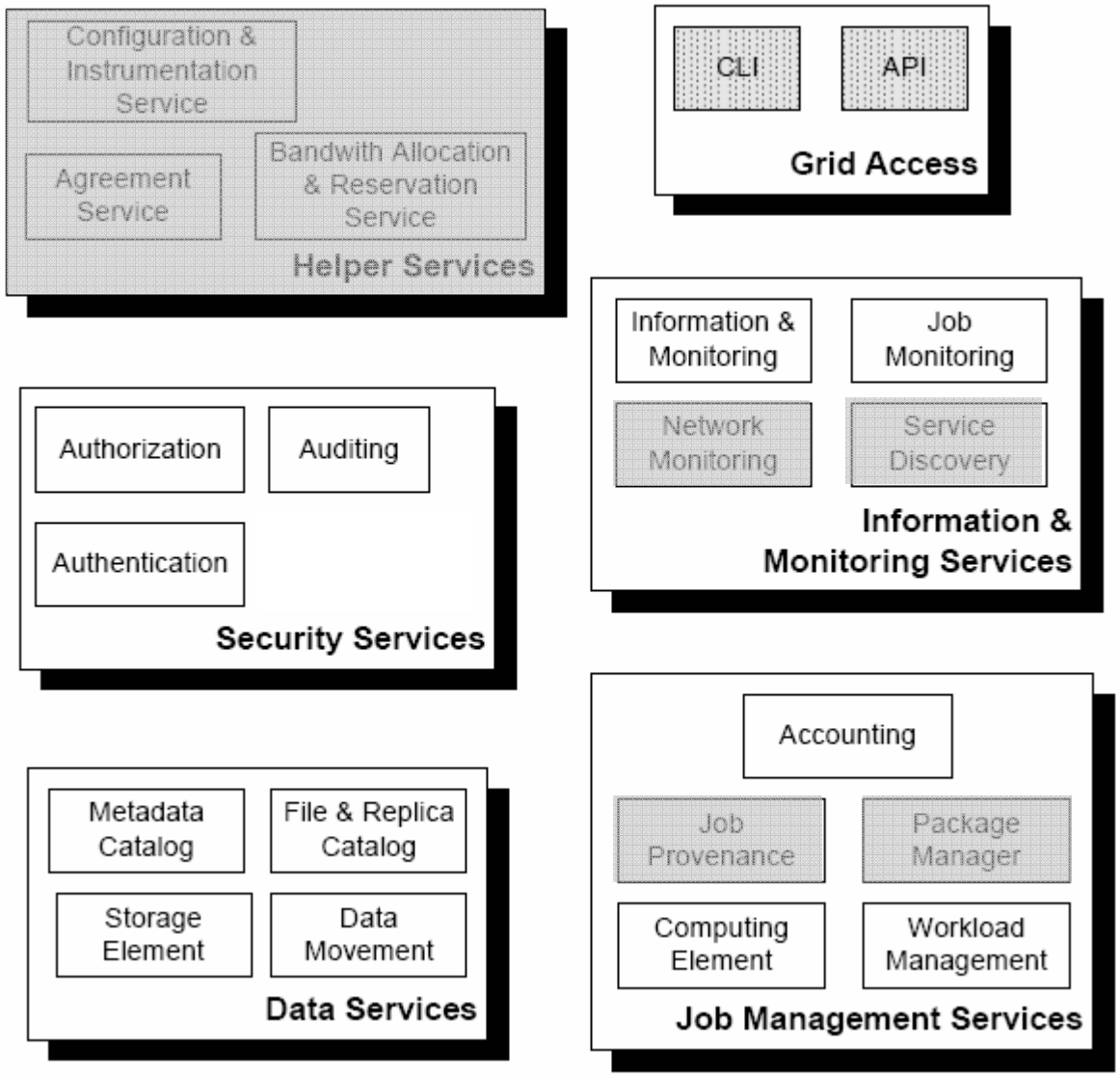
Service interaction by messages having a common messaging infrastructure

SOAP (Web Services) – Std Protocol to manage Messaging among Services

WSDL - A language that exposes the service interface.

gLite 3.0 – Service Decomposition

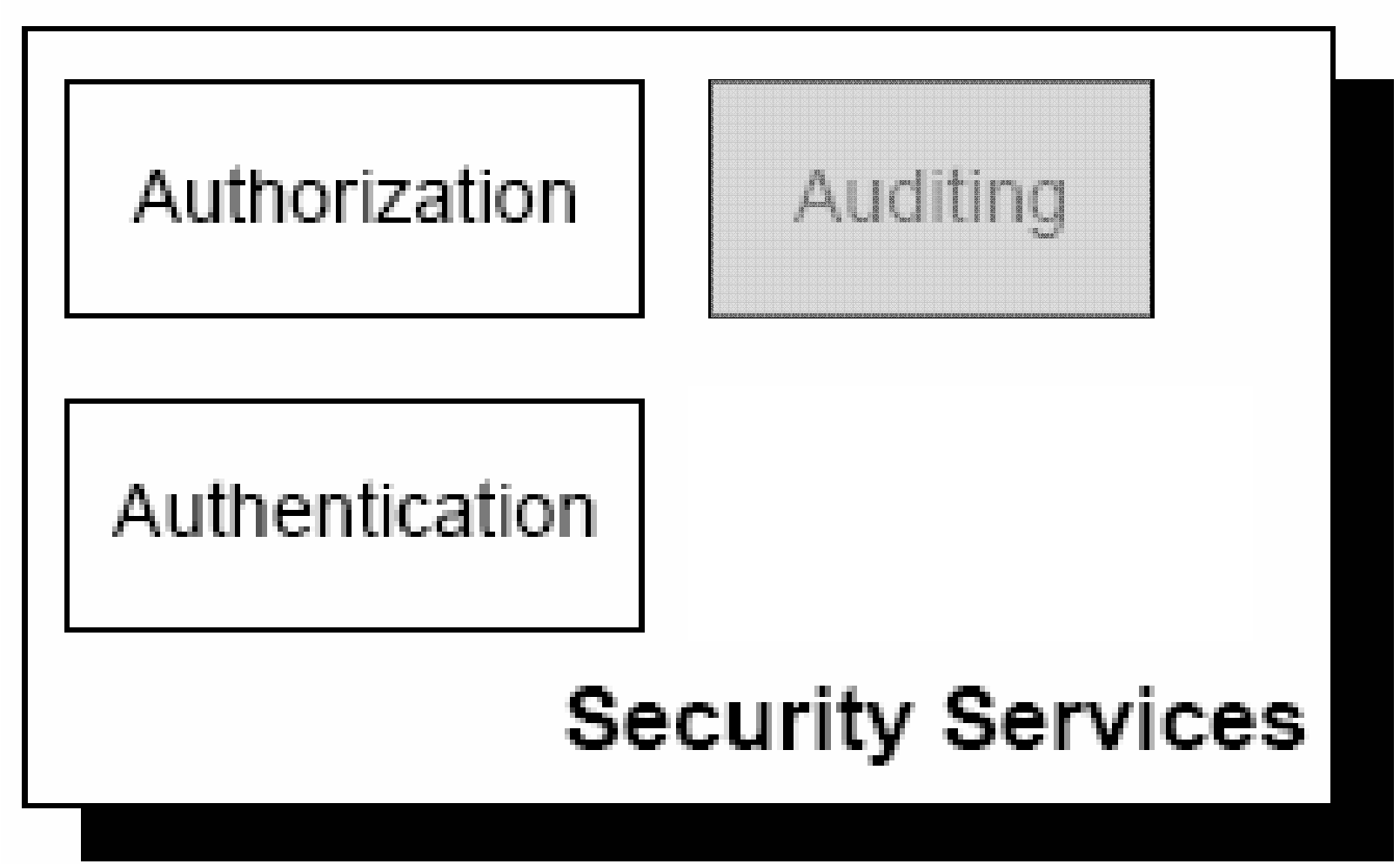
5 High level services
+ CLI & API



Legend:

- Available
- Soon Available

gLite – Security Services



Identify entities (users, systems and services) when establishing a context for message exchange (Who are you?).

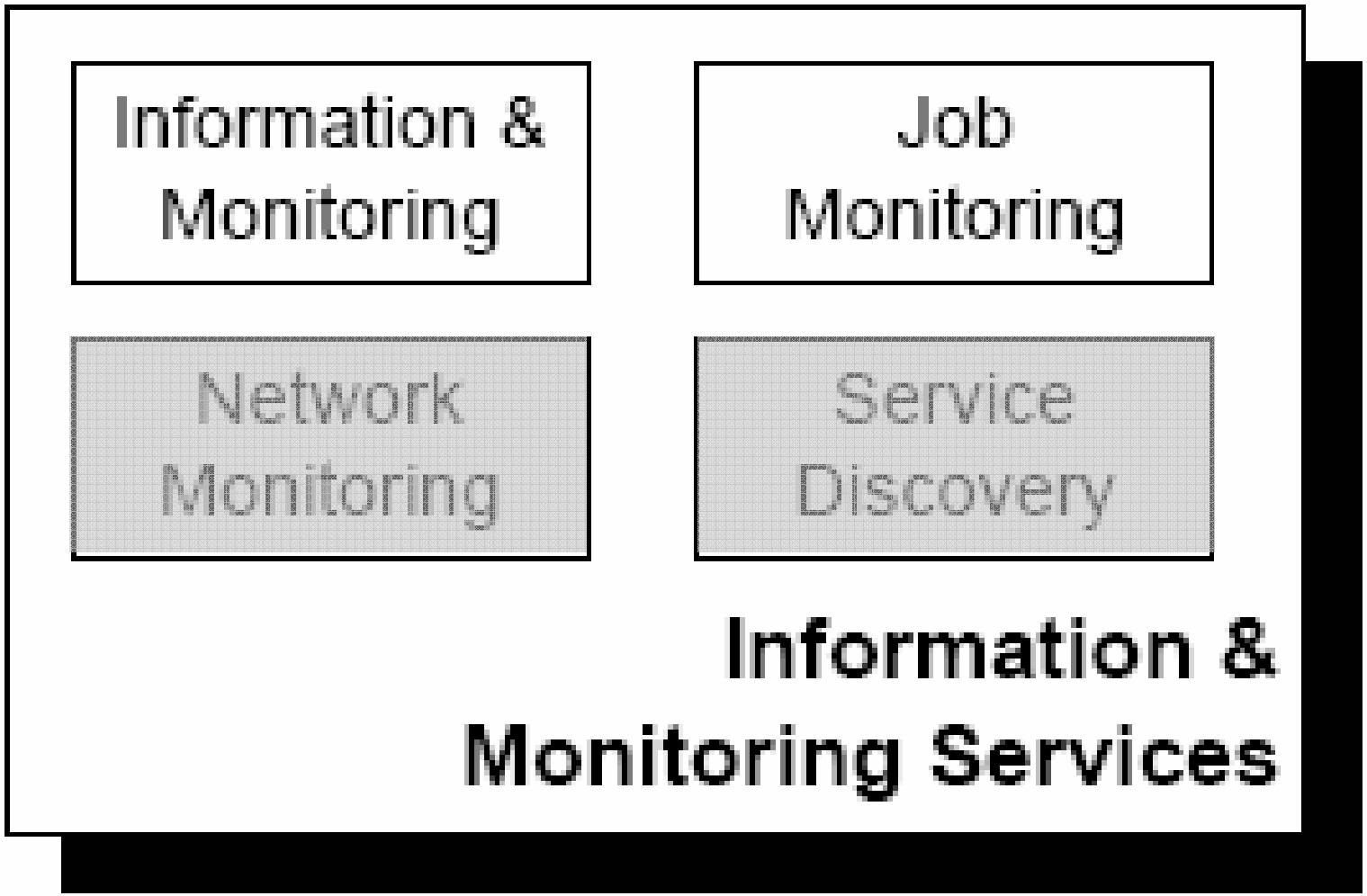
Aim - Provide a Credential having a universal value that works for many purposes across many infrastructures, communities, VOs and projects.

gLite uses: the **PKI** (X.509) infrastructure using **CA**s as thrusted third parties.

gLite uses**: *MyProxy* (**http://grid.ncsa.uiuc.edu/myproxy/**) extended by VOMS**.

Trust domain: The set of all EGEE CAs is our Trust Domain.

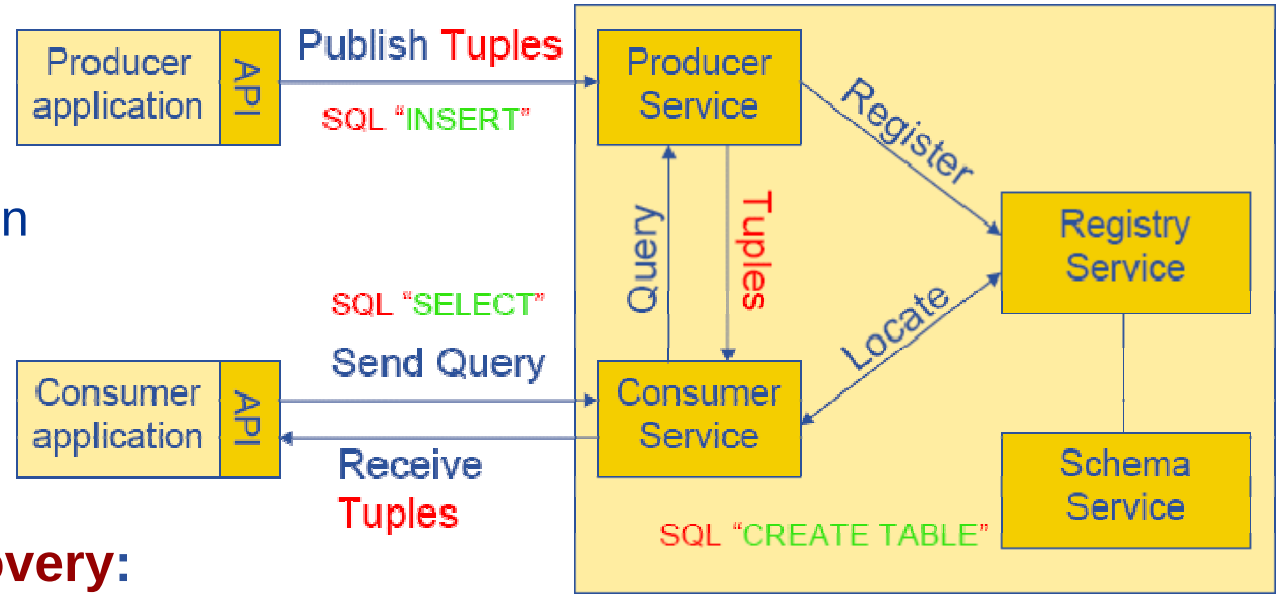
Information services are vital low level components of Grids.



gLite – Information and Monitoring Services

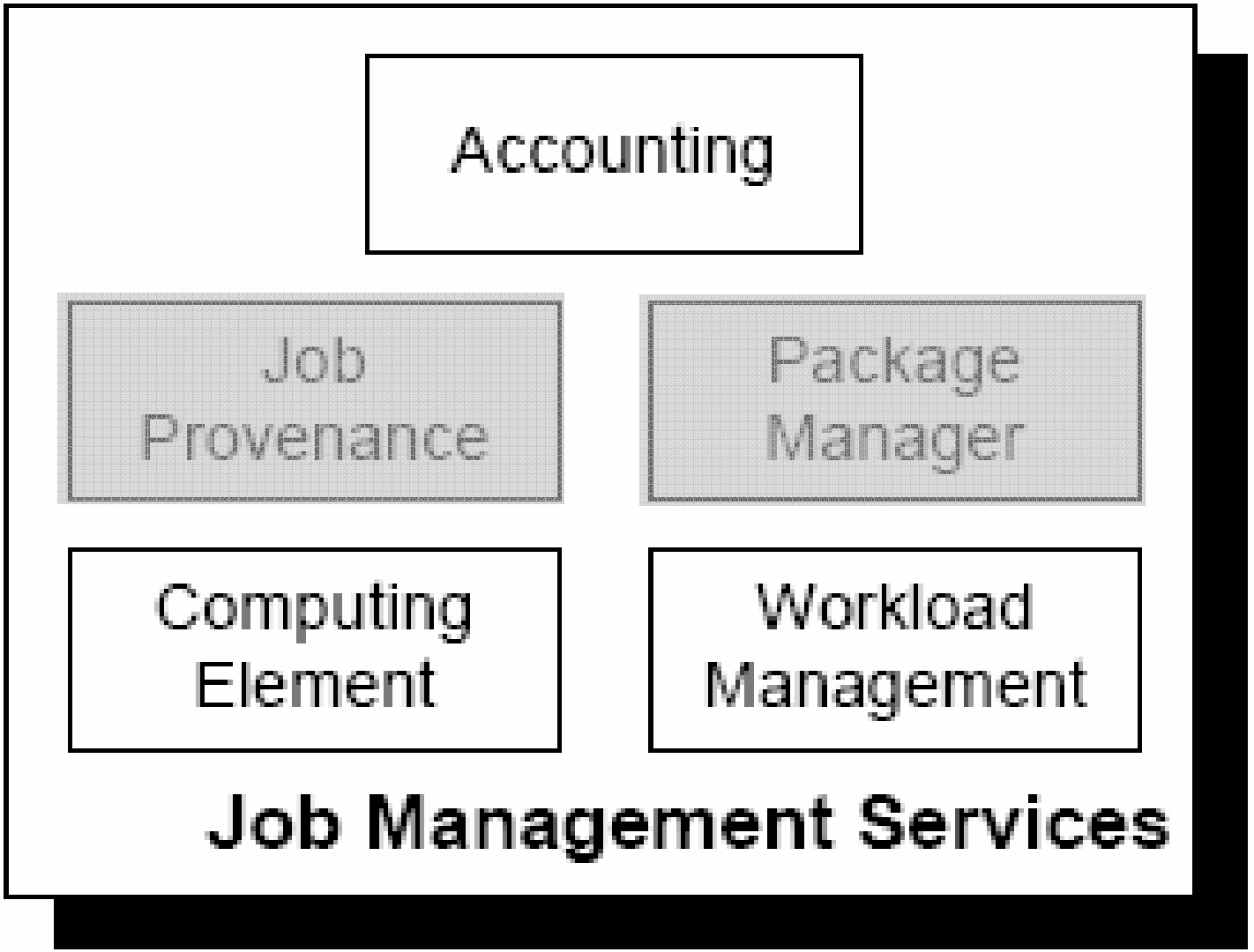
- **R-GMA: provides a uniform method to access and publish distributed information and monitoring data**
  - Used for job and infrastructure monitoring in gLite 3.0

Working to add authorization



- **Service Discovery:**
  - Provides a standard set of methods for locating Grid services
  - Currently supports R-GMA, BDII and XML files as backends
  - Will add local cache of information
  - Used by some DM and WMS components in gLite 3.0

gLite – Job Management Services



gLite – Job Management Services

Computing Element

- Service that represents the **computing resource** that is responsible of the job management: (submission, control, etc.)

- The CE may be used by a **Generic Client**:
  - an **end-user** interacting directly with the Computing Element,
  - or the **Workload Manager** that submits a given job to an appropriate CE found by a **matchmaking** process.

- Two job submission models (accordingly to user requests and site policies):

- **PUSH** (**Eager** *Scheduling*) (jobs pushed to CE),
- **PULL** (**Lazy** Scheduling) (jobs coming from WMS when CE has free slots)

- CE must also provide information describing itself.
- CE responsible to collect accounting information.

gLite – Job Management Services

Workload Management

- **WMS** is a set of middleware components responsible of **distribution** and **management** of **jobs** across Grid resources.

- There are two core components of the WMS:

  - **WM**: **accepts** and **satisfy requests** for job management.
    Matchmaking is the process of assigning the best available resource.

  - **L&B**: **keeps track** of job execution in terms of **events:**
    (Submitted, Running, Done,...)

gLite – Job Management Services

Workload Management

- WMS helps the user accessing computing resources
  - Resource brokering, management of job input/output, ...
- LCG-RB: GT2 + Condor-G
  - To be replaced when the gLite WMS proves reliability
- gLite WMS: Web service (WMProxy) + Condor-G
  - Management of complex workflows (DAGs) and compound jobs
    - bulk submission and shared input sandboxes
    - support for input files on different servers (scattered sandboxes)
  - Support for shallow resubmission of jobs
  - Supports collection of information from BDII, R-GMA and CEMon
  - Support for parallel jobs (MPI) when the home dir is not shared
  - Deployed for the first time with gLite 3.0

gLite – Job Management Services

Accounting

- Accumulates information about the **resource usage** done by users or groups of users (VOs).

- Information on Grid Services/Resources needs **sensors**
(Resource Metering, Metering Abstraction Layer, Usage Records).

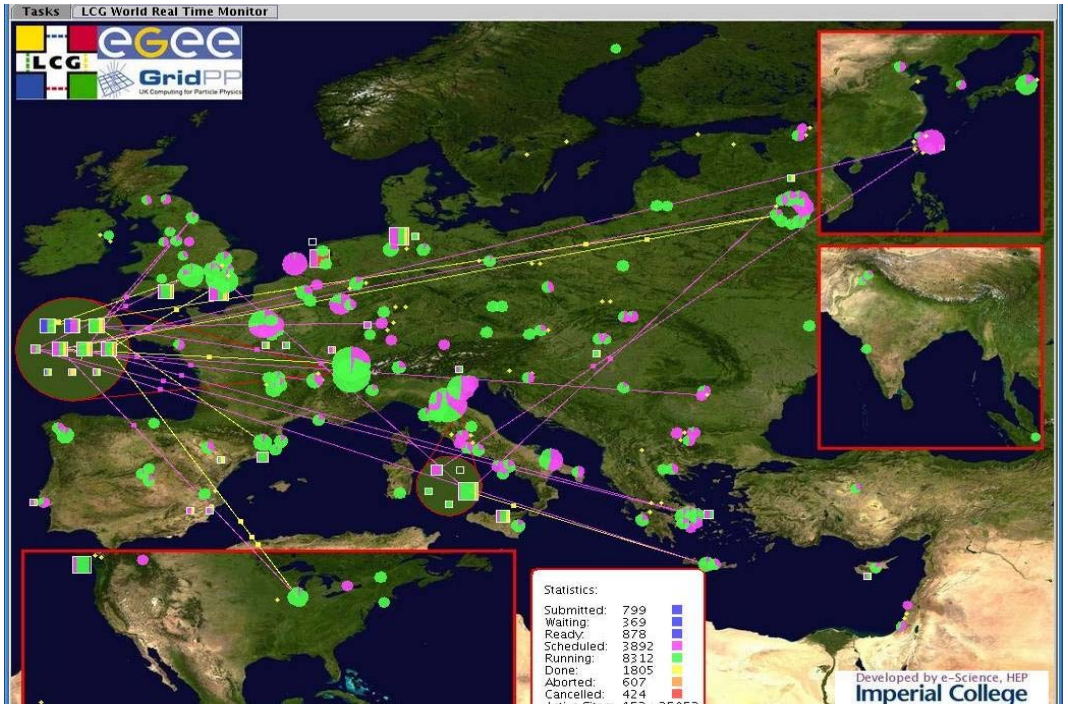
- Records are collected by the **Accounting System**
(Queries: Users, Groups, Resource)

- Grid services should register themselves with a pricing service when accounting is used for billing purposes.

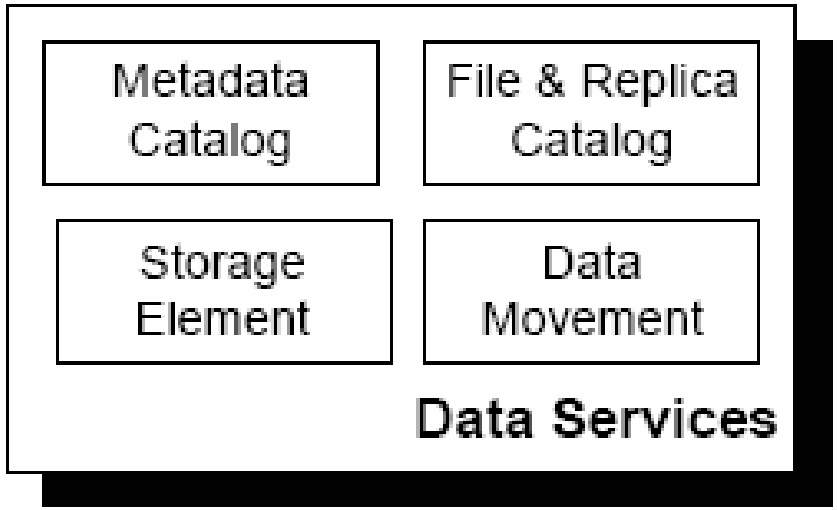
gLite – Job Management Services
Job Provenance

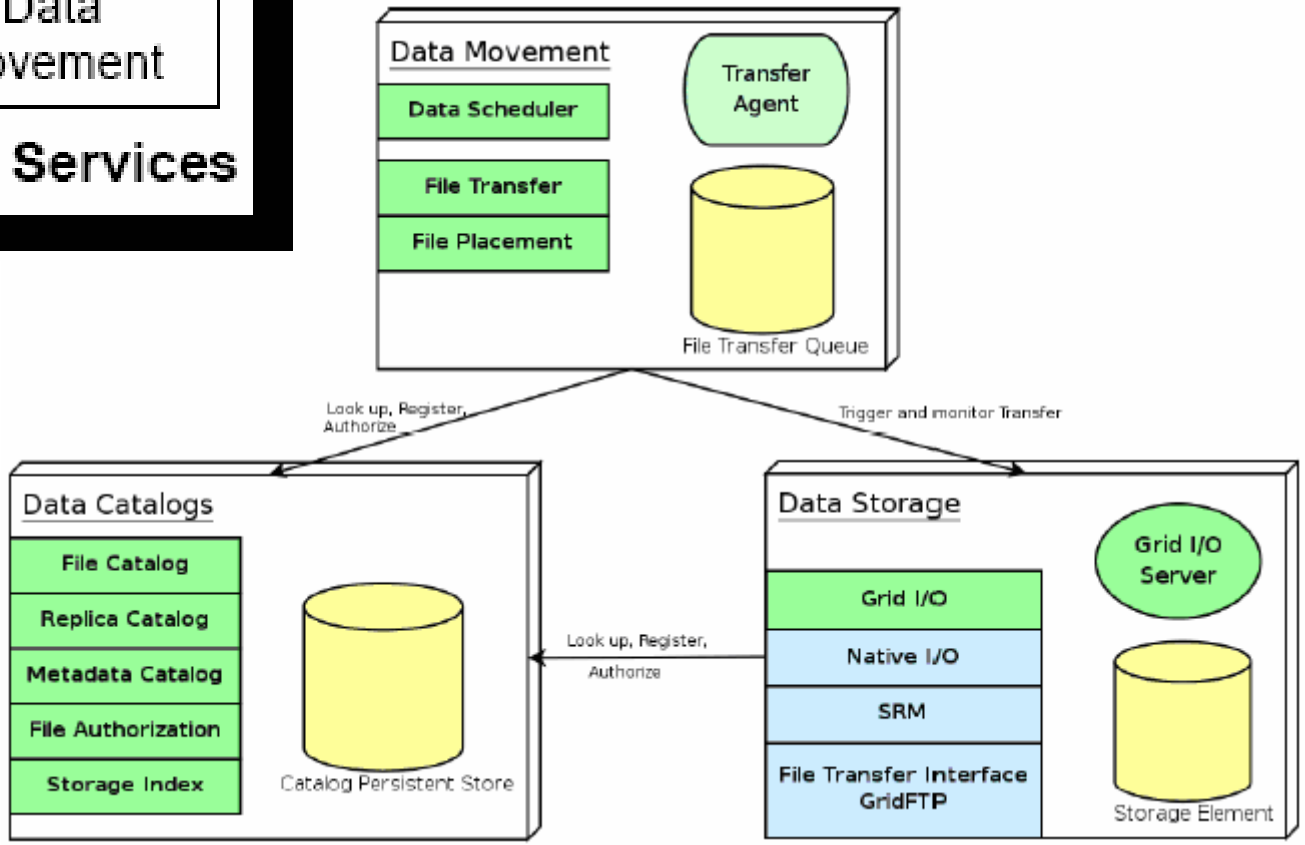
- **Logging and Bookkeeping service**
  - Tracks jobs during their lifetime (in terms of events)
  - LBProxy for fast access
  - L&B API and CLI to query jobs
  - Support for “CE reputability ranking“: maintains recent statistics of job failures at CE's and feeds back to WMS to aid planning

- **Job Provenance: stores long term job information**
  - Supports job rerun
  - If deployed will also help unloading the L&B
  - Not yet certified in gLite 3.0.



gLite – Data Services





Summary

- **EGEE-II will continue to build a large Grid infrastructure**
- **LCG-2 and gLite 3.0 are complete middleware stacks:**
  - security infrastructure
  - information system and monitoring
  - workload management
  - data management
- **gLite 3.0.0 is available on the production infrastructure**
- **Development is continuing to provide**
  - increased robustness,
  - usability and
  - functionality