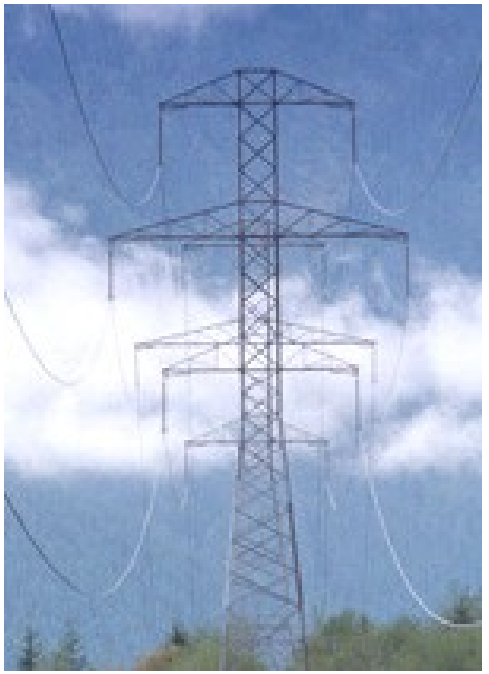
Grid computing

David Bouvet

Les Houches – September 19th, 2008

What is the grid? - Grid history

- Name “Grid” chosen by analogy with electric power grid (Foster and Kesselman 1997)
- Vision: plug-in computer for processing power just like plugging in toaster for electricity.
- Concept has been around for decades (distributed computing, metacomputing)
- Key difference with the Grid is to realise the vision on a global scale.

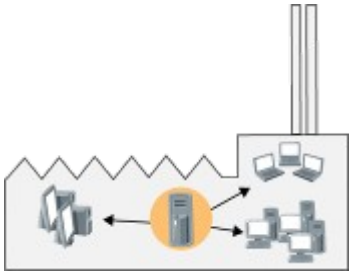


How does the Grid work?

- **It relies on advanced software, called** middleware**.**
- **Middleware automatically finds the** data **the scientist needs, and the** computing power **to analyse it.**
- **Middleware balances the load on different resources. It also handles** security, accounting, monitoring **and much more.**

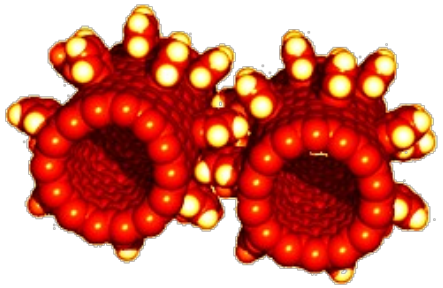
Different Grids for different needs

- **There is as yet no unified Grid (like there is a single web) rather** there are many Grids **for many applications.**
- **The word Grid is used to signify different types of distributed computing for example** Enterprise Grids **(within one company) and** public resource Grids **(volunteer your own PC).**
- **In this talk, focus is on** scientific Grids **that link together major computing centres in research labs and universities.**
- **Latest trend is to** federate national Grids **to achieve a global Grid infrastructure.**

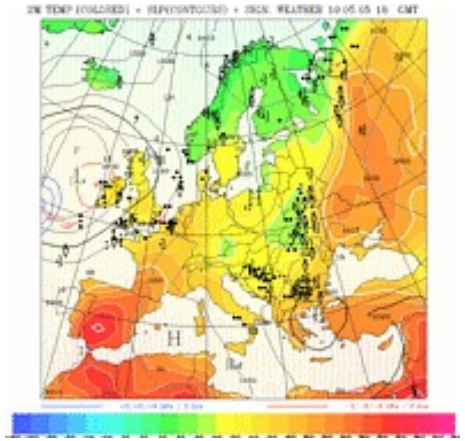


Why the grid? - Motivation

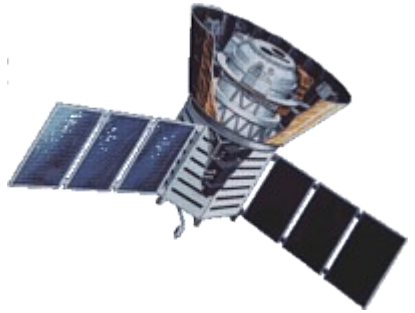
- **Science is becoming increasingly digital and needs to deal with increasing amounts of data and computational needs**



- **Simulation get ever more detailed**
  - Nanotechnology – design of new materials from the molecular scale
  - Modelling and predicting complex systems (weather forecast, river floods, earthquake)
  - Decoding the human genome

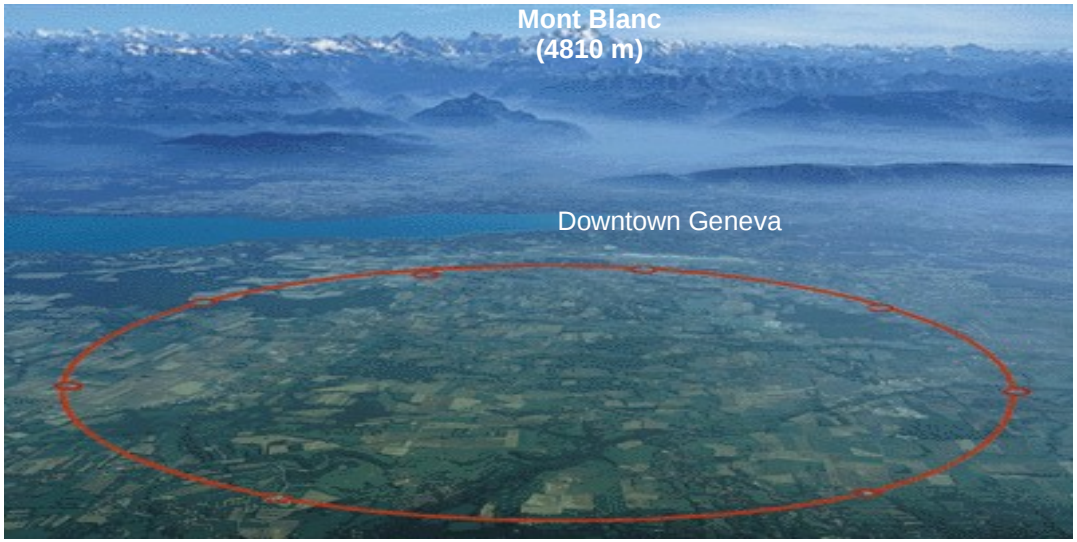


- **Experimental science uses ever more sophiticated sensors to make precies mesurements**
  - Need high statistics
  - Huge amount of data



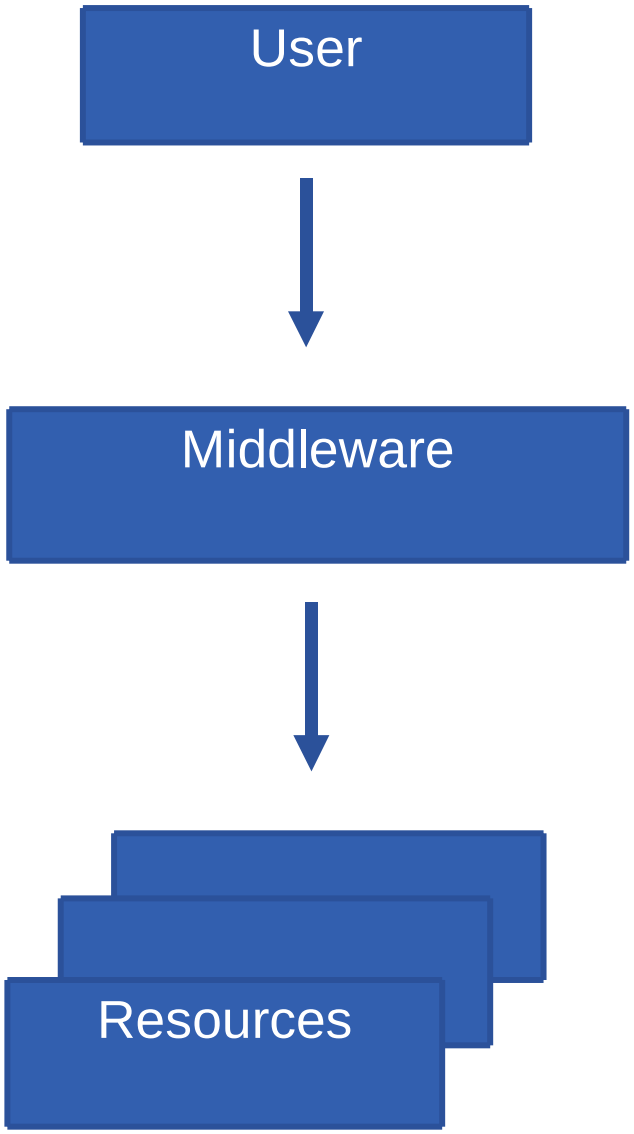
The need for grid in particle physics

- **Large Hadron Collider (LHC) at CERN**
  - One of the most powerful instruments ever build to investigate the matter
  - 4 experiments: ALICE, ATLAS, CMS, LHCb
  - 27 km circumference tunnel
- **LHC data**
  - 40 million collisions per second, 100 of interest after collision
  - 1 Mbytes of data for each collision ⇒ 0.1 GBytes/sec
  - ~15 petabytes/year (~10 Million GBytes)



Defining the grid

- **A grid is the combination if networked resources and the corresponding Grid middleware, which provides Grid services for the user.**
- **This interconnection of users, resources and services for jointly addressing dedicated tasks is called a virtual organization.**
- **Comparison between Grids and Networks:**
  - Network realize message exchange between endpoints
  - Grids realize services for the users.



Grid and virtualization

- **Virtual organizations (VO's) = Group of users, federating resources**
  - Heterogeneous: people from different organisations
  - Cooperation: common goals
  - For sharing: to solve problems by using common resources (computers, data files, scientific instruments...)
- **Virtualized shared computing and data resources**
  - Access to resources outside their institut for members of VO's
  - Resource providers negociate with VO not with individual members
- **Virtualization and sharing also possible for:**
  - Instruments, sensors, people, etc...

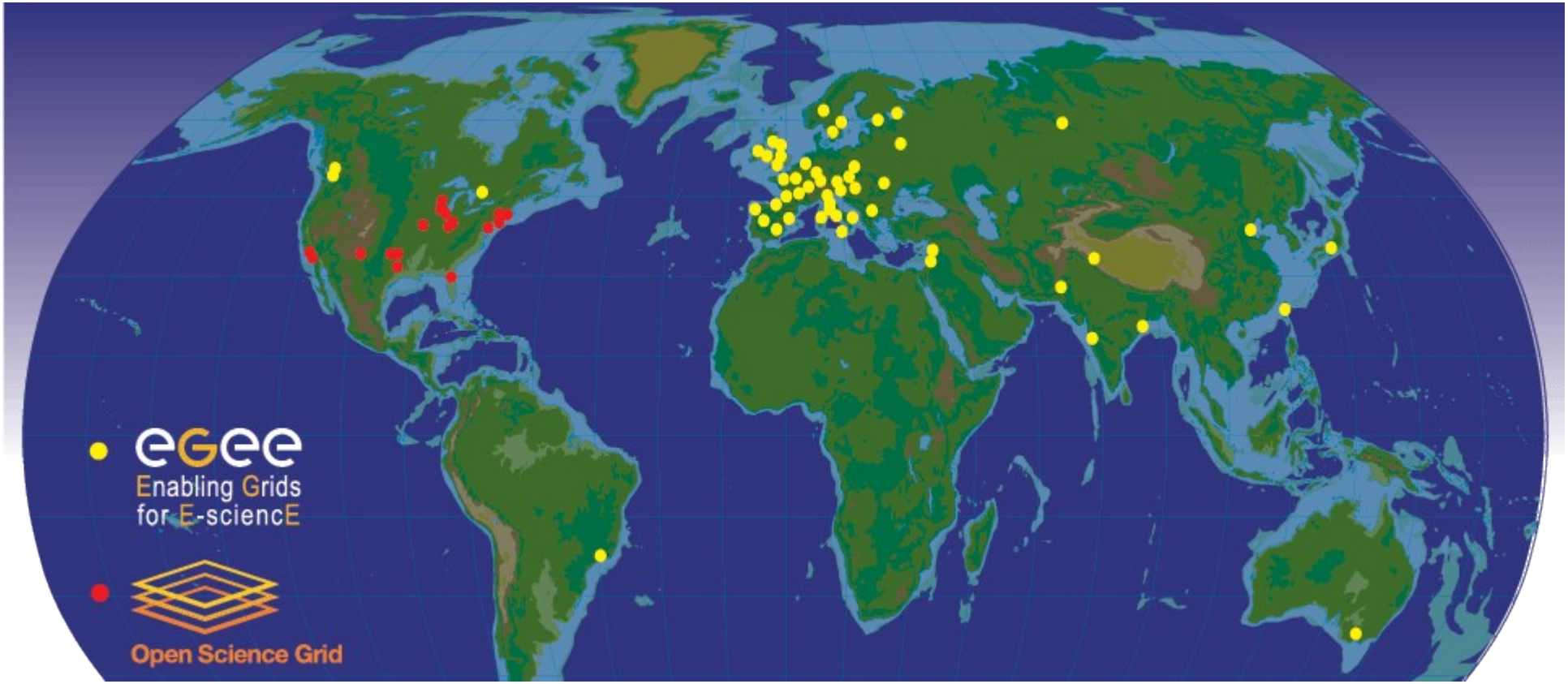
Virtualisation of resources is needed to hide their heterogeneity and present a simple interface to users

Worldwide LCG Collaboration

- **LHC Computing Grid**
  - Purpose: develop, build and maintain a distributed computing environment for the storage and processing of data for the 4 LHC experiments
  - Resources contributed by the countries participating in the experiments

WLCG: architecture

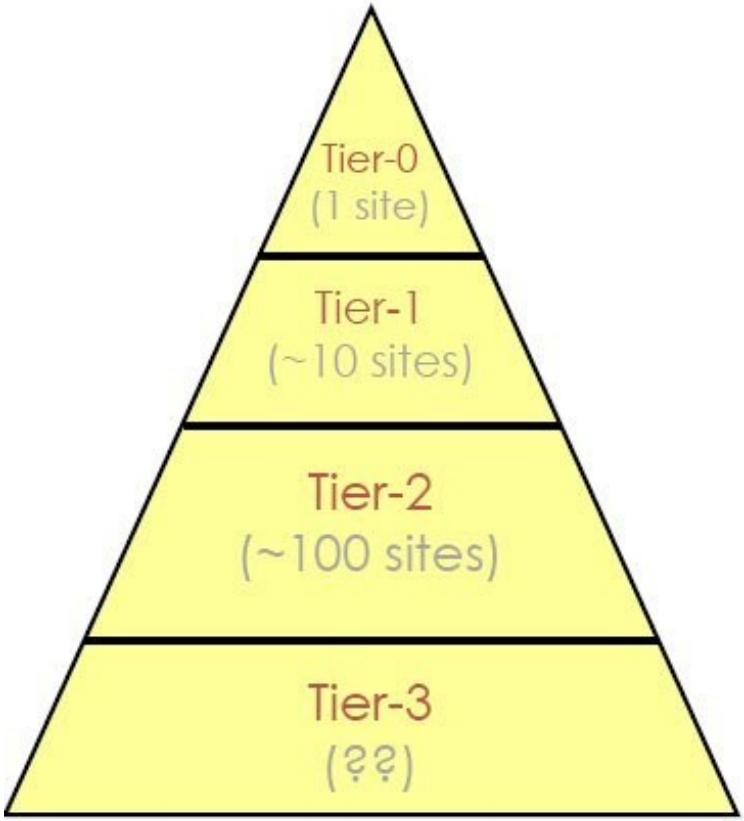
- **WLCG architecture**
  - Set of services and applications running on the grid infrastructures provided by EGEE (Europe and Asia Pacific), NorduGrid (nordic countries) and Open Science Grid (USA)



Inter-operation between Grid is working!

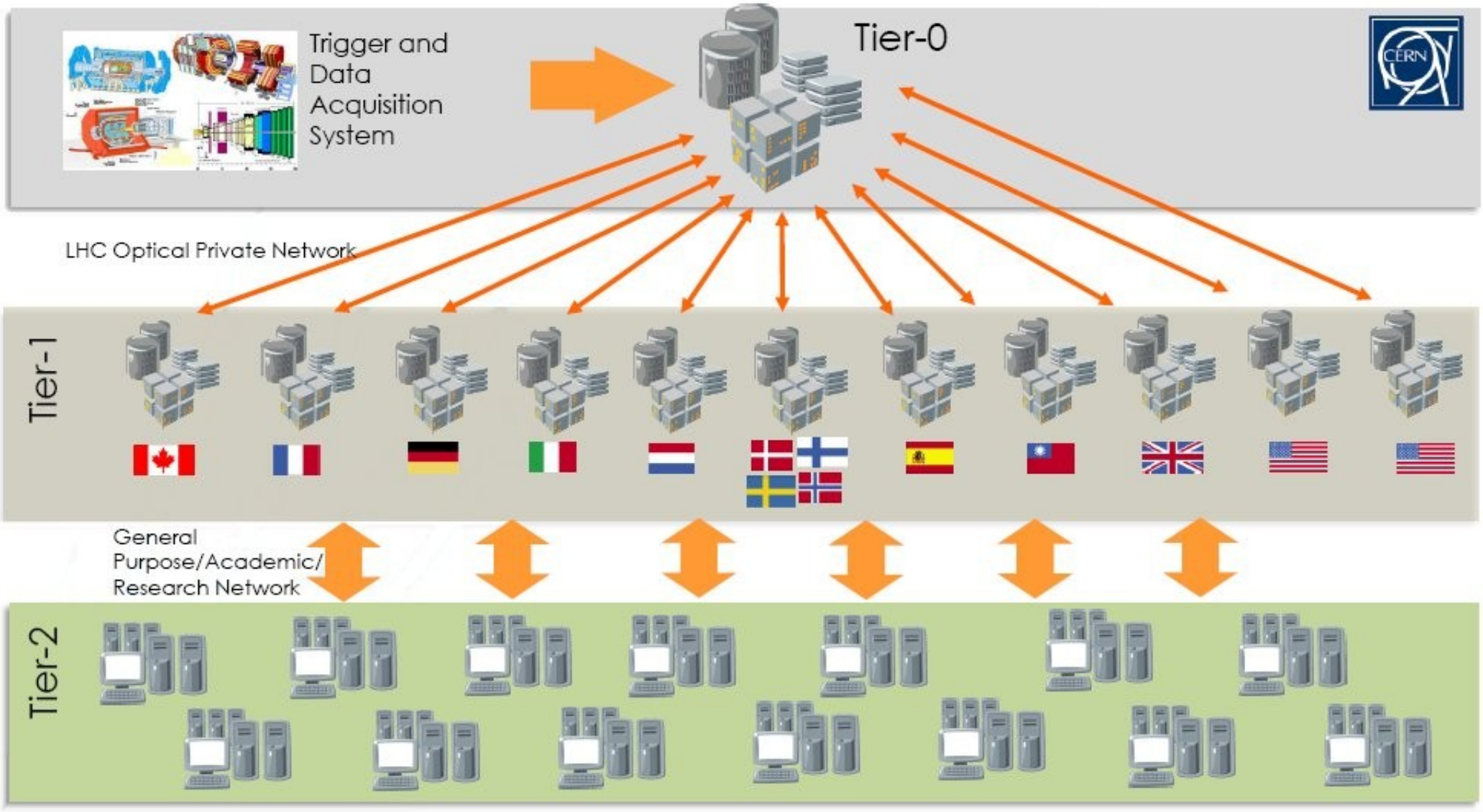
WLCG: architecture (cont.)

- **Tier-0: accelarator centre (CERN)**
  - Data acquisition and initial processing
  - Long-term data curation
  - Distribution of data to Tier-1 certres (near online)
- **Tier-1**
  - Viewed as online for the data acquisition process
    - High availability requied (24x7)
  - Grid-enabled data service (backed by a mass storage system)
  - Data-heavy analysis
  - 11 sites, 10 countries
- **Tier-2**
  - Simulation
  - Interactive and batch end-user data analysis
  - ~120 sites, ~35 countries
- **Tier-3**
  - End-user analysis



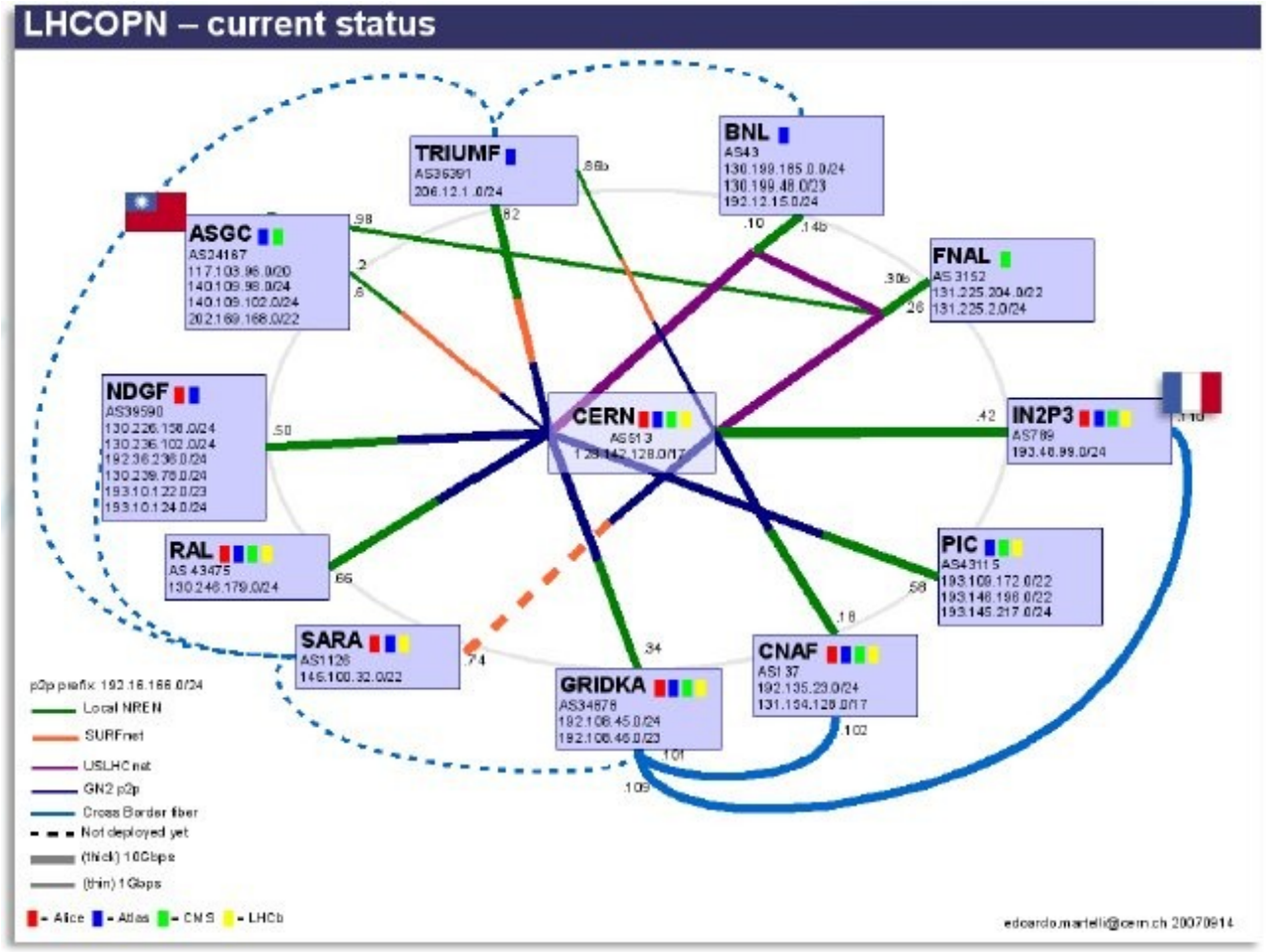
Four-tiered model

WLCG architecture (cont.)



WLCG architecture: network

- **Network is the spine cord of the grid:**
  - Without high performance network, no grid!
- **LHC Optical Private Network (LHCOPN)**
  - Reserved for carrying LHC data from T0 to T1s
    - And for exchanging data between T1s
  - Optical circuits provided and operated by national academic and research networks
  - Bandwidth: 10 Gbps per link
  - Designed for resilience



Source: https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

- **EGEE = Enabling Grids for E-sciencE**
  - Flagship Grid infrastructure project of the EU
  - Funded by European Commission
  - Primary objectives:
    - Operate a large scale, production quality (consistent, robust and secure) grid infrastructure for e-science
    - Improving and maintaining the middleware
    - Attracting new users from indudstry and science
  - Structure:
    - Over 250 leading institutions in more than 48 countries, federated in regional Grids
    - 68.000 CPUs
    - 5 PetaBytes storage

EGEE: what is delivered?

- **Infrastructure operation**
  - Sites distributed across many countries
    - Large quantity of CPUs and storage
    - Continuous monitoring of grid services & automated site configuration/management
    - Support multiple Virtual Organizations from diverse research disciplines
- **Middleware**
  - Production quality middleware distributed under business friendly open source licence
    - Implement a service-oriented architecture that virtualized resources
    - Adheres to recommendations on web service inter-operability and evolving towards emerging standards
- **User Support – *Managed process from first contact through to production usage***
  - Training
  - Expertise in grid-enabling applications
  - Online helpdesk
- **Networking events (user forum, conferences, etc.)**

Grid middleware: gLite

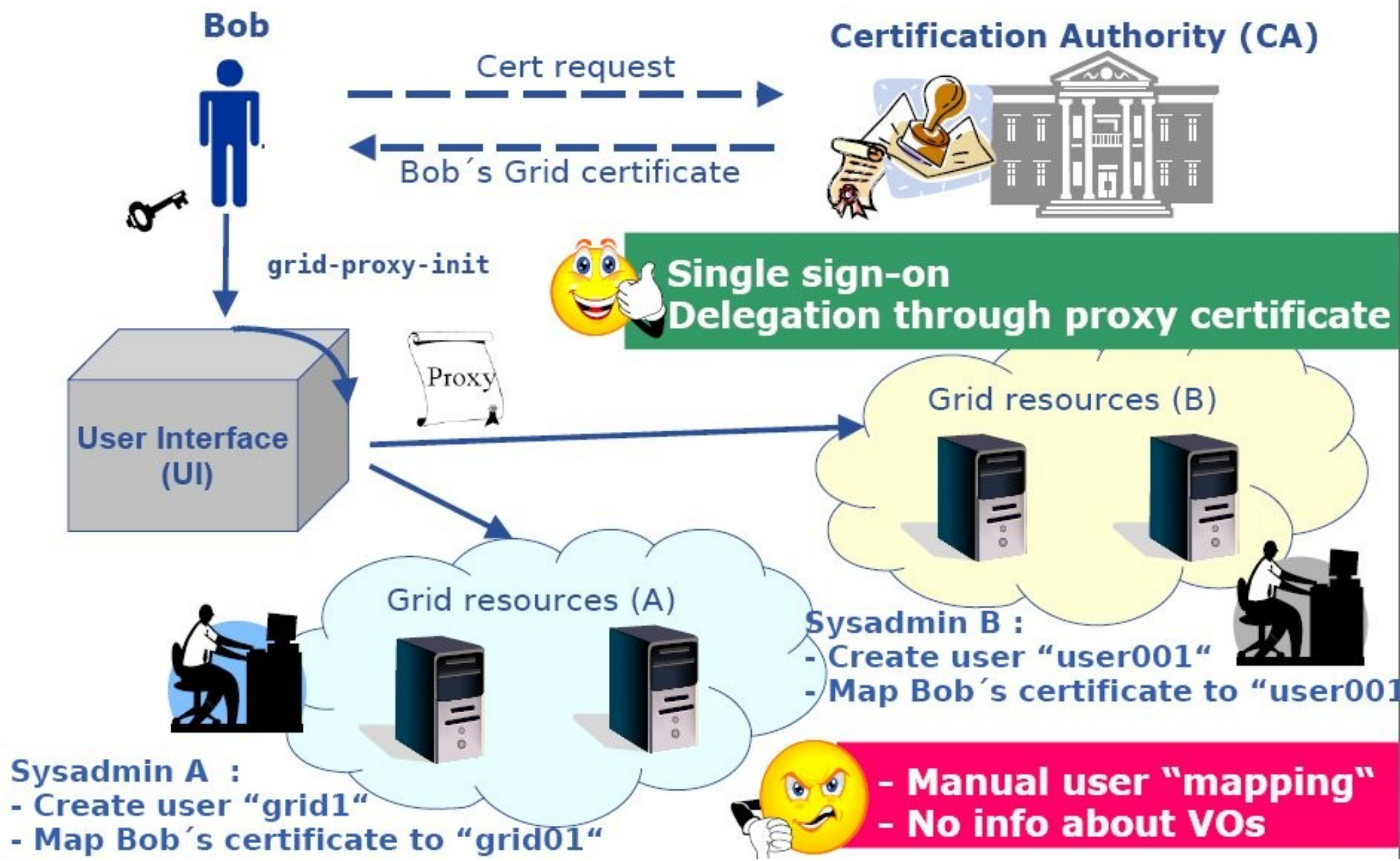
- **The Grid relies on advanced software – the middleware – which interfaces between resources and applications.**

- **The Grid middleware:**
  - Finds convenient places for the application to be executed
  - Optimizes use of ressources
  - Organises efficient access to data
  - Deals with authentication to the different sites that are used
  - Run the job and monitors progress
  - Transfers the results back to scientist

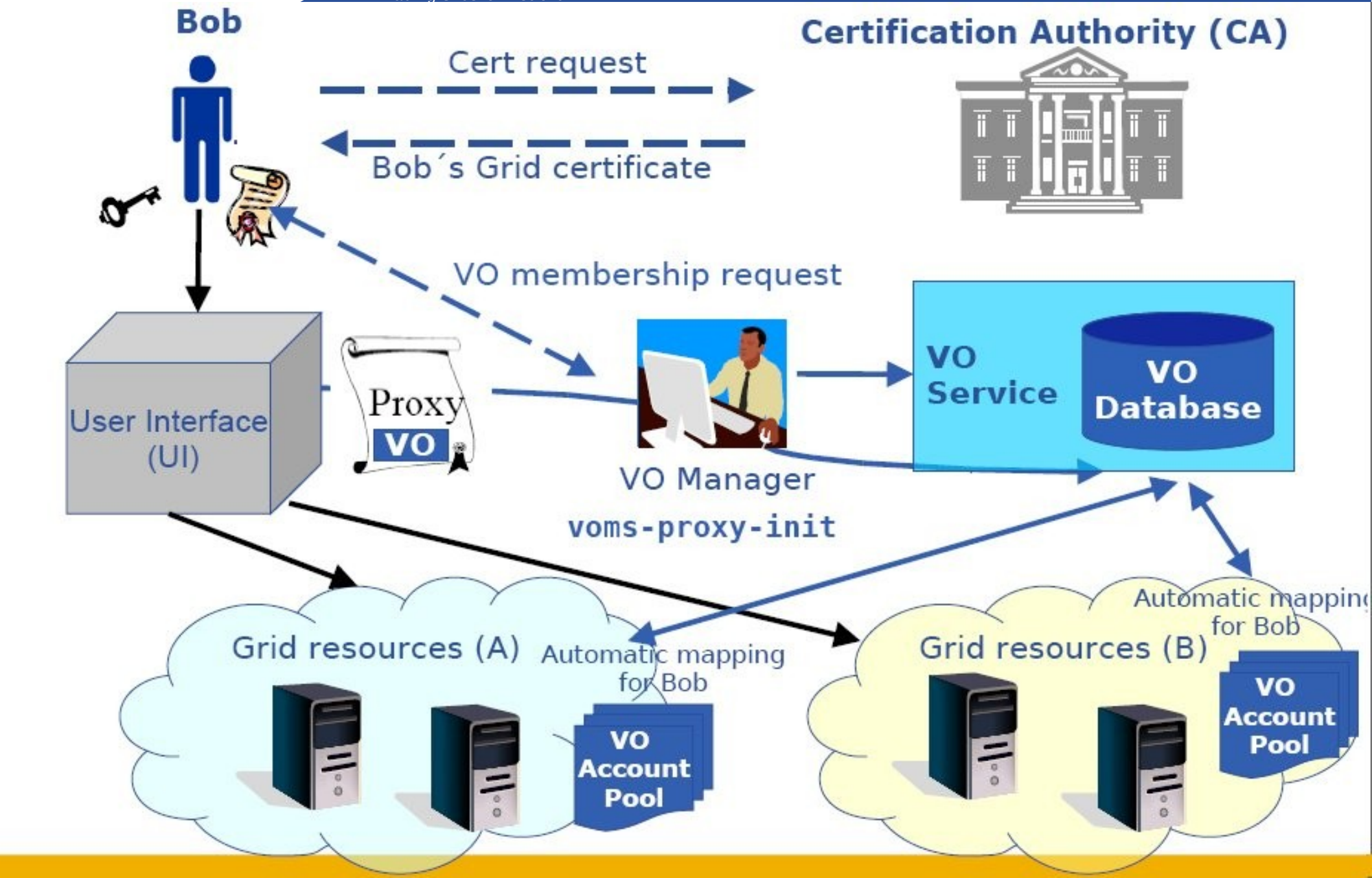
gLite: overview

- **gLite**
  - First release in 2005 (currently gLite 3.1)
  - Next generation middleware for grid computing
  - Intended to replace present middleware with production quality services
  - Developped from existing components (globus, condor,...)
  - Interoperability & co-existence with deployed infrastructure
  - Robust: performance & fault tolerance
  - Open source license
  - Platform: currently only Scientific Linux (RedHat like distro) supported

gLite: conventional grid security



gLite: enhanced security in gLite



gLite: services

- **User Interface (UI):**
  - Access point to the grid infrastructure
  - Any machine with a personal account and where the user certificate is installed
  - Provides a CLI tools to perform Grid operations
- **Computing Element (CE):**
  - Computing resource localized at site (i.e. cluster, computing farm)
  - includes:
    - A grid gate which act as a generic interface to the cluster
    - A Local Resource Management System (LRMS) - the batch system
    - The cluster itself: a collection of Worker Node (WN) on which the jobs run
- **Storage Element (SE):**
  - provides uniform access to data storage resources
  - may control simple disk servers, large disk arrays or tape-based Mass Storage Systems (MSS).

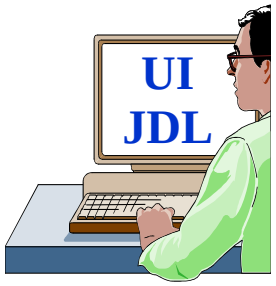
gLite: services

- **WMS (Workload Management System):**
  - accepts user jobs, assigns them to the most appropriate Computing Element
  - record their status and retrieve their output.
  - The **Resource Broker (RB)** is the machine where the WMS service runs.
    - The choice of CE to which the job is sent is made in a process called match-making, which first selects, among all available CEs, those which fulfil the requirements expressed by the user and which are close to specified input Grid files. It then chooses the CE with the highest rank.
- **Information Service/System (IS):**
  - provides information about the WLCG/EGEE Grid resources and their status.
  - This information is essential for the operation of the whole Grid, as it is via the IS that resources are discovered.
  - The published information is also used for monitoring and accounting purposes.
  - IS conforms to the GLUE Schema, which defines a common conceptual data model to be used for Grid resource monitoring and discovery.

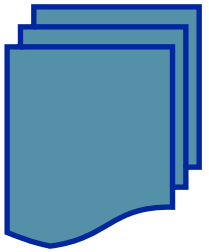
gLite: services

- **Virtual Organization Membership Service (VOMS):**
  - EGEE/gLite enhancement for VO management
  - Provides information on user's relationship with VO
    - Membership
    - Group membership
    - Roles of user
  -
  - Has an account database
    - Comparable to Kerberos server
    - Serving information in a special format (VOMS credentials)
    - Administration via CLI and web interface

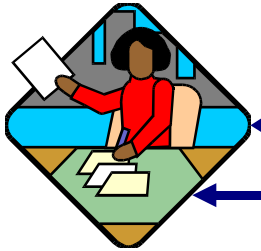
gLite: job flow



LCG File Catalog (LFC)



Information System (IS)

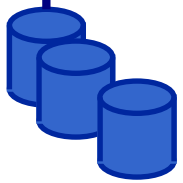


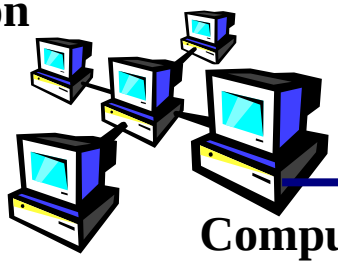
Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

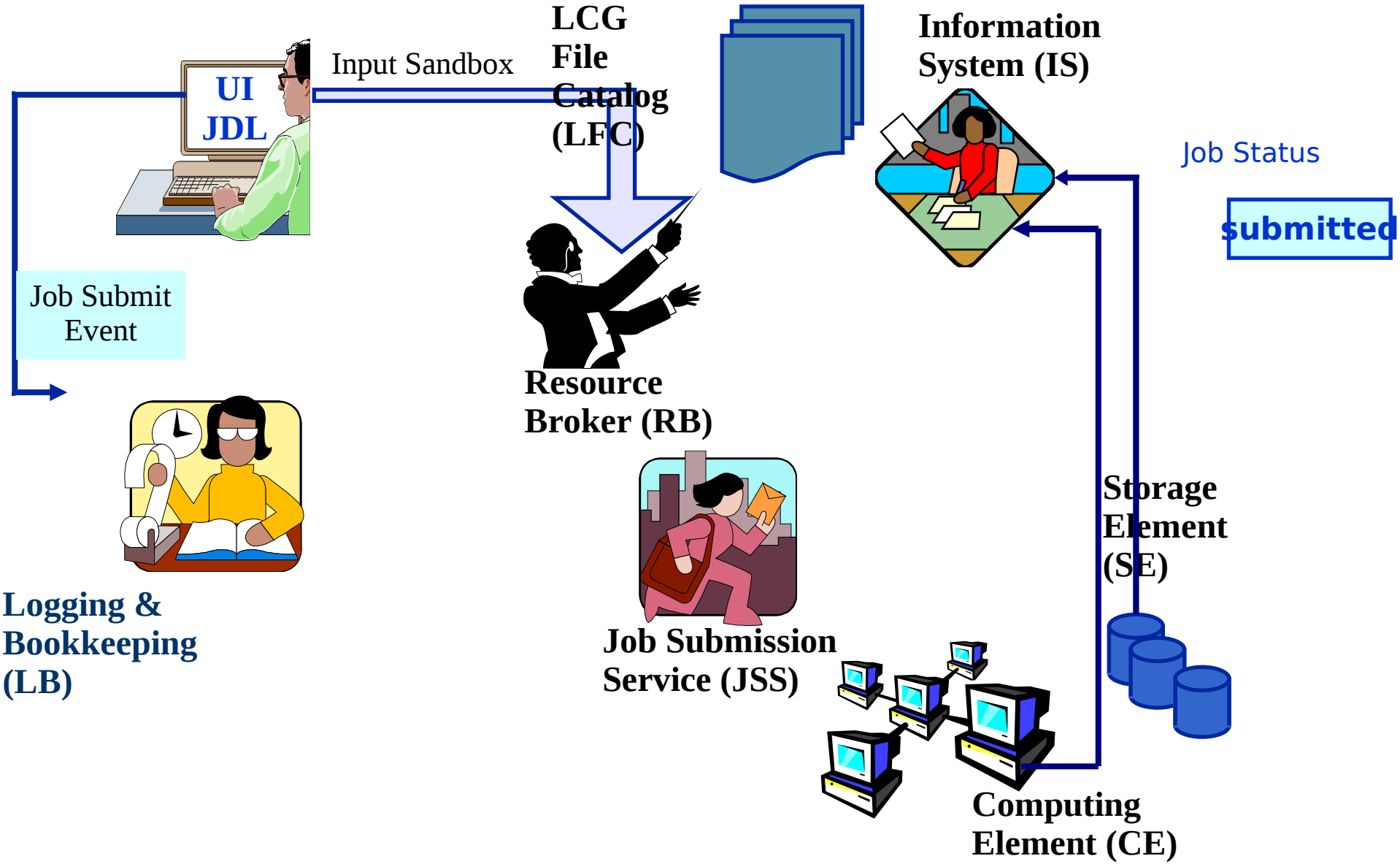
Storage Element (SE)



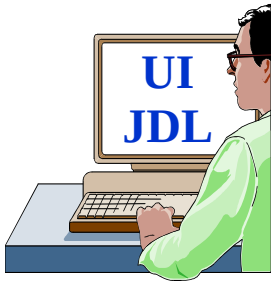


Computing Element (CE)

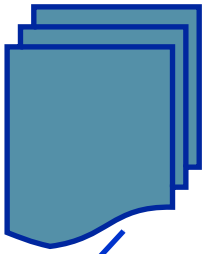
gLite: job flow



gLite: job flow



LCG File Catalog (LFC)



Information System (IS)

Job Status

submitted

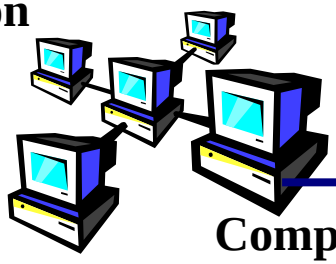
waiting

Resource Broker (RB)

Logging & Bookkeeping (LB)

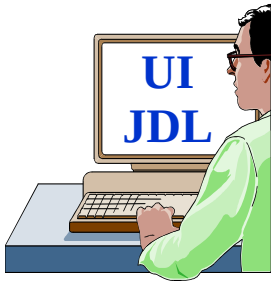
Job Submission Service (JSS)

Storage Element (SE)

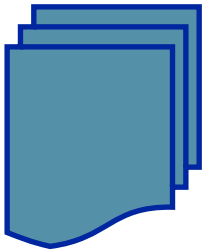


Computing Element (CE)

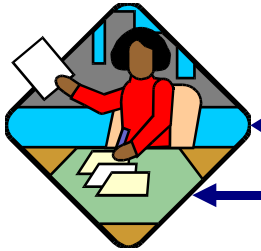
gLite: job flow



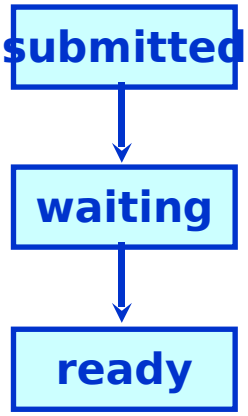
LCG File Catalog (LFC)

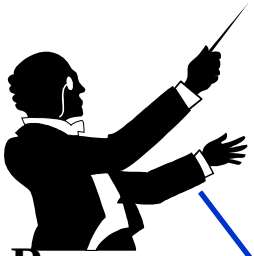


Information System (IS)



Job Status



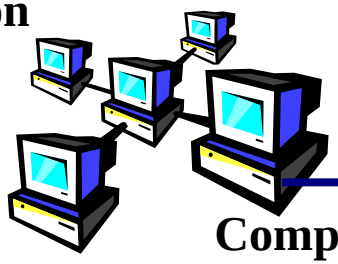


Resource Broker (RB)

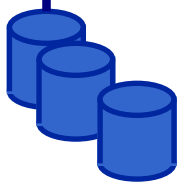
Job Status

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

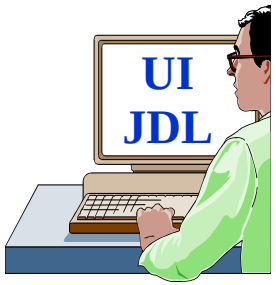


Computing Element (CE)

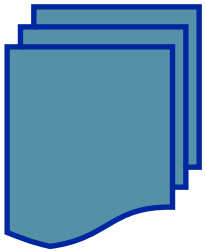


Storage Element (SE)

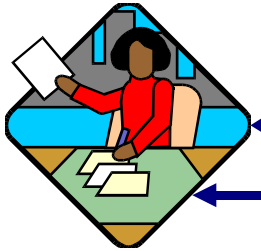
gLite: job flow



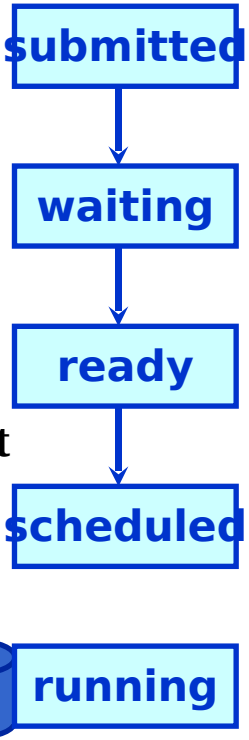
LCG File Catalog (LFC)



Information System (IS)



Job Status



Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Status

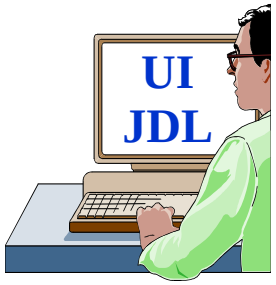
Job Submission Service (JSS)

BrokerInfo

Storage Element (SE)

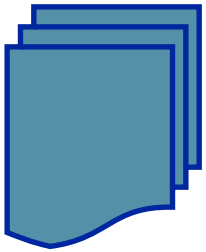
Computing Element (CE)

gLite: job flow

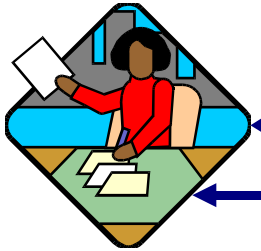


Logging & Bookkeeping (LB)

LCG File Catalog (LFC)



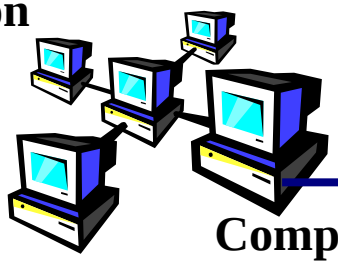
Information System (IS)



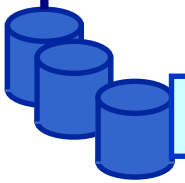
Resource Broker (RB)

Input Sandbox

Job Submission Service (JSS)

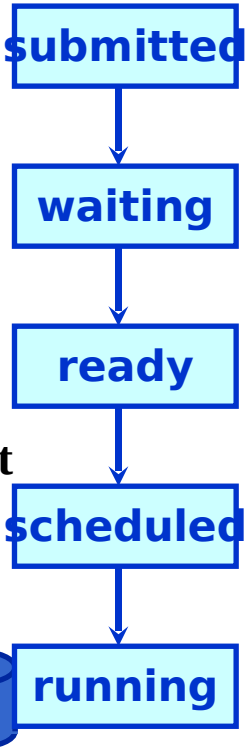


Computing Element (CE)

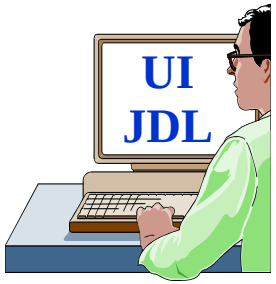


Storage Element (SE)

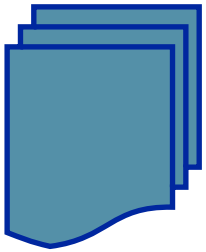
Job Status



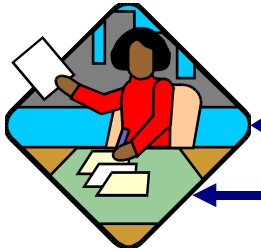
gLite: job flow



LCG File Catalog (LFC)



Information System (IS)

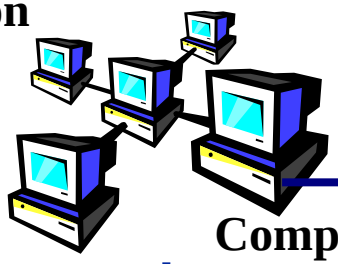


Job Status

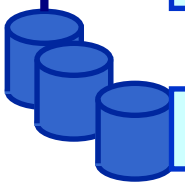
Resource Broker (RB)

Logging & Bookkeeping (LB)

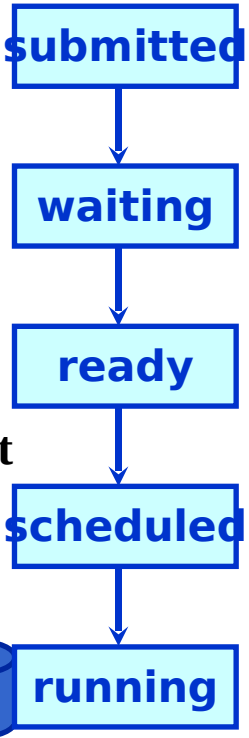
Job Submission Service (JSS)



Computing Element (CE)

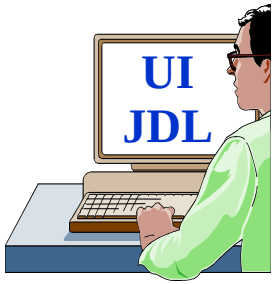


Storage Element (SE)

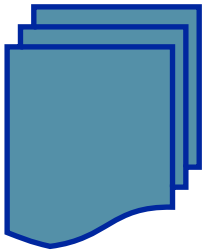


Job Status

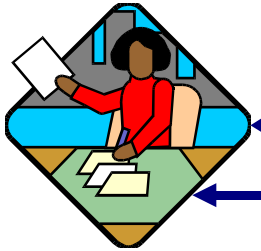
gLite: job flow



LCG File Catalog (LFC)



Information System (IS)

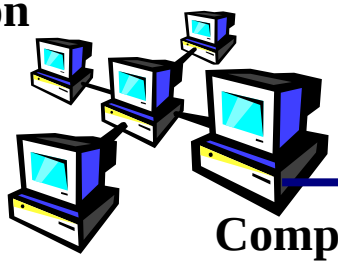


Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

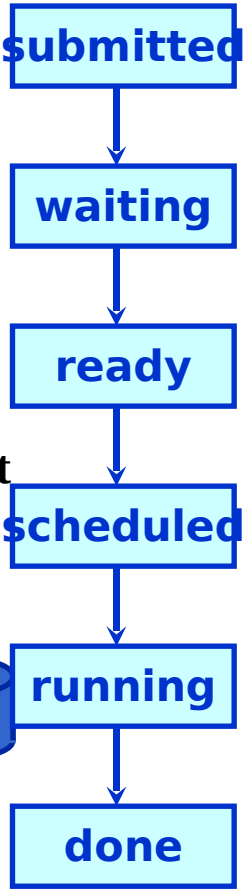
Storage Element (SE)



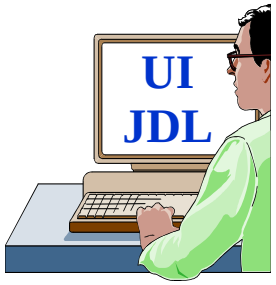
Computing Element (CE)

Job Status

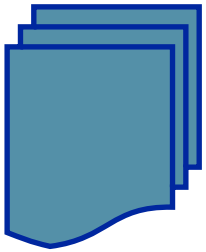
Job Status



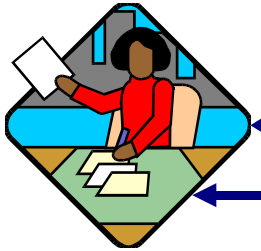
gLite: job flow



LCG File Catalog (LFC)



Information System (IS)

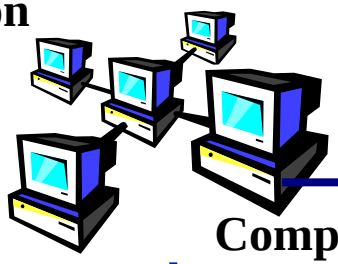


Resource Broker (RB)

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

Storage Element (SE)

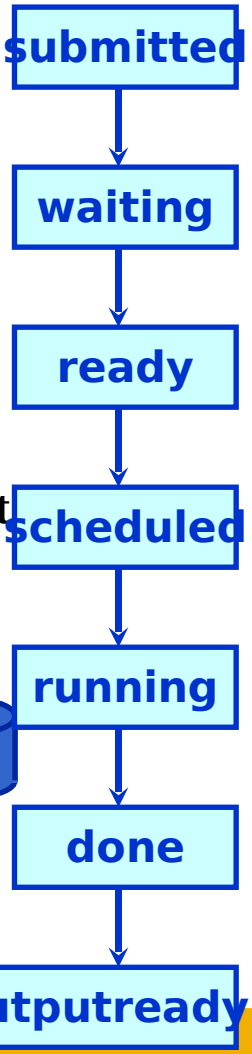


Computing Element (CE)

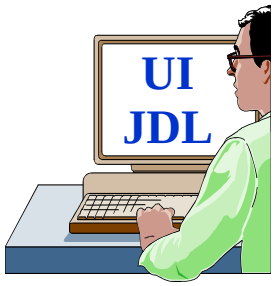
Output Sandbox

Job Status

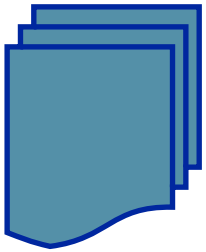
Job Status



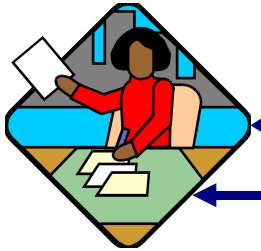
gLite: job flow



LCG File Catalog (LFC)



Information System (IS)



Job Status

- submitted
- waiting
- ready
- scheduled
- running
- done
- outputready
- cleared

Output Sandbox

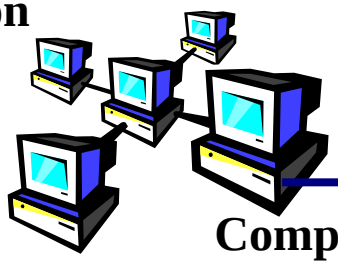
Resource Broker (RB)

Job Status

Logging & Bookkeeping (LB)

Job Submission Service (JSS)

Storage Element (SE)



Computing Element (CE)

Data management

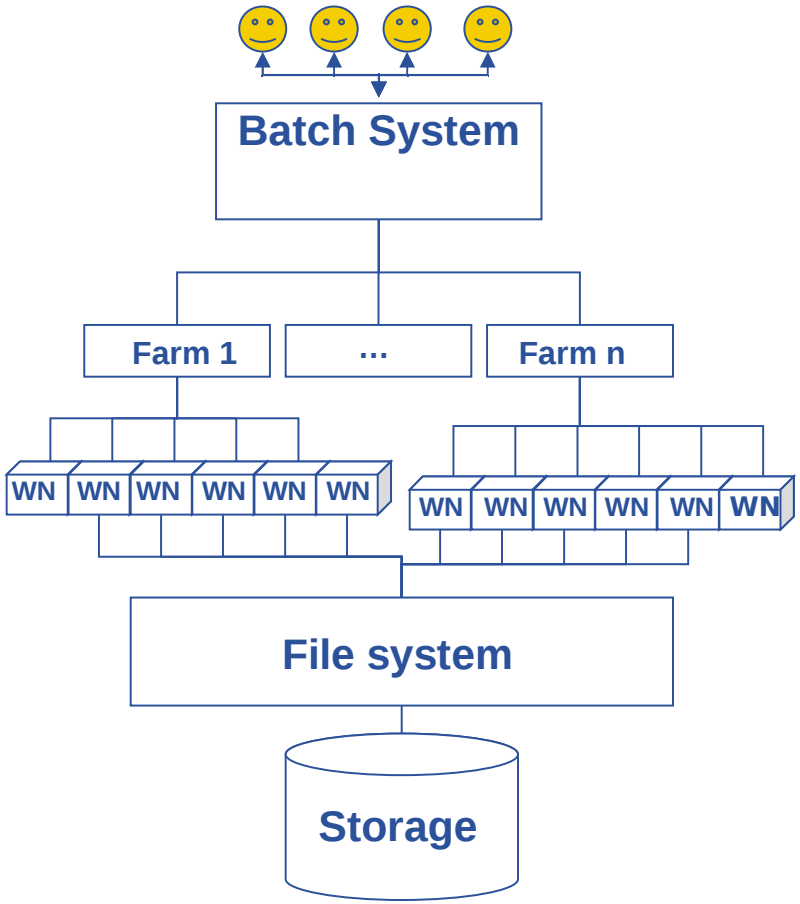
- **In a Grid environment, files can have replicas at many different sites. Because all replicas must be consistent, Grid files cannot be modified after creation, only read and deleted.**
- **Ideally, users do not need to know where a file is located, as they use logical names for the files that the Data Management services use to locate and access them.**
-
- **Files in the Grid can be referred to by different names:**
  - Grid Unique IDentifier (GUID)
  - Logical File Name (LFN)
  - Storage URL (SURL) and Transport URL (TURL)
    - While the GUIDs and LFNs identify a file irrespective of its location, the SURLs and TURLs contain information about where a physical replica is located, and how it can be accessed.

Grid sites: another level of resource mutualization

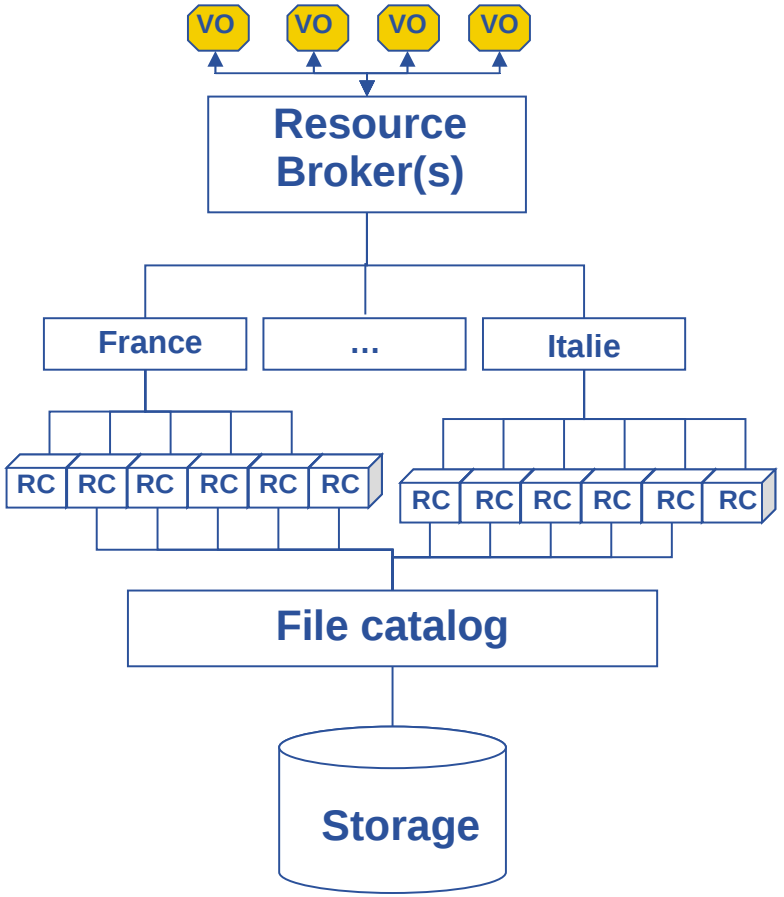
- **Computing centre:**
  - Supercomputer / computing farm
    - set of “Worker Nodes”
    - Heterogenuous machine (hardware, OS)
  - High storage capacity (disk, tape)
  - Management of user community through account (groups and users)
  - Management of production in an expert way
- **Grid computing:**
  - Mutualization of resource centres
  - Management of community through “Virtual Organization” concept
  - Minimization of latency for the users

Grid sites: another level of resource mutualization

- **Computing centre**

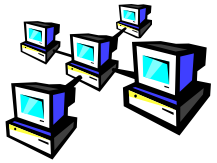


- **Computing centre**

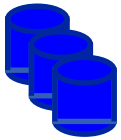


Grid sites: infrastructure

- **Resource centre**
  - Provides grid with
    - Computing resources

Computing Element 

    - Storage resources

Storage Element 

    - Eventually services global or per VO


RB

LFC

UI

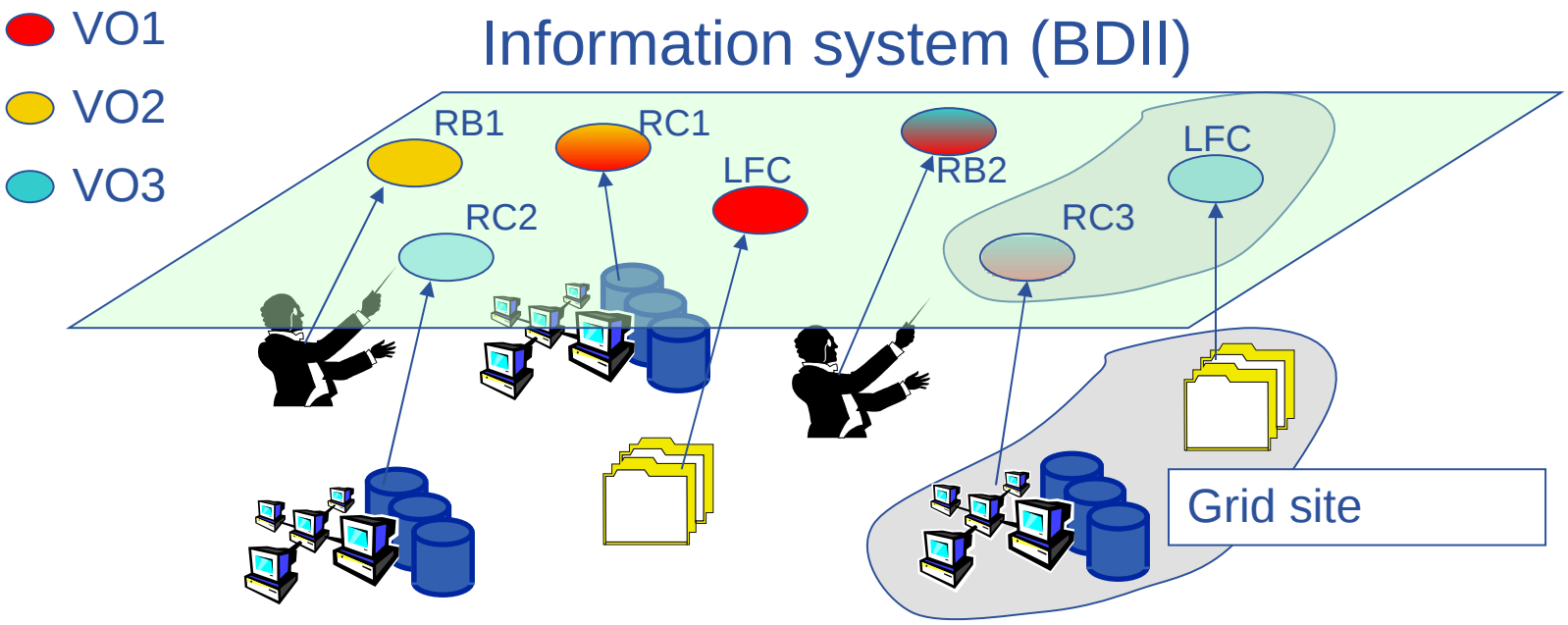
BDII

VOMS

FTS

Grid sites: functionnal aspects

- **Information published on the Grid:**
  - Each site publishes
    - A description of the resources/services provided per VO
    - Current state of resources (free cPUs, free storage space,...)
    - On the “Resource Centre” (RC), what a VO has installed (Tags)



Grid sites: functionnal aspects

- **Jobs management:**
  - Job submission
    - Processing of application on a grid WN
    - Data access the job needs
  - Job should give a description of its needs to select the correct environment of processing
    - VO name
    - Description of its needs through a JDL (Job Description Language)
      - *Program to run*
      - *Tag of the selected application (e.g.: the version number)*
      - *Estimated max. time for the processing*
      - *List of files needed*
      - *etc.*

Want to try?

- **Go to the GILDA testbed: https://gilda.ct.infn.it/**

- **GILDA: Grid Infn Laboratory for Dissemination Activities**
- **GILDA is a testbed for learning purpose only**
- **It is NOT the EGEE production infrastructure**
- **No guarantee for successful job execution**
  - Depending on the load, a “hello world” can take minutes/hours/...

Application in EGEE

- Particle physics
- Bioinformatics
- Industry
- Astronomy
- Chemistry
- Earth observation
- Geophysics
- Biodiversity
- Nanotechnology
- Climate modeling

Collaborating Cyber-Infrastructures

Enabling Grids for E-sciencE

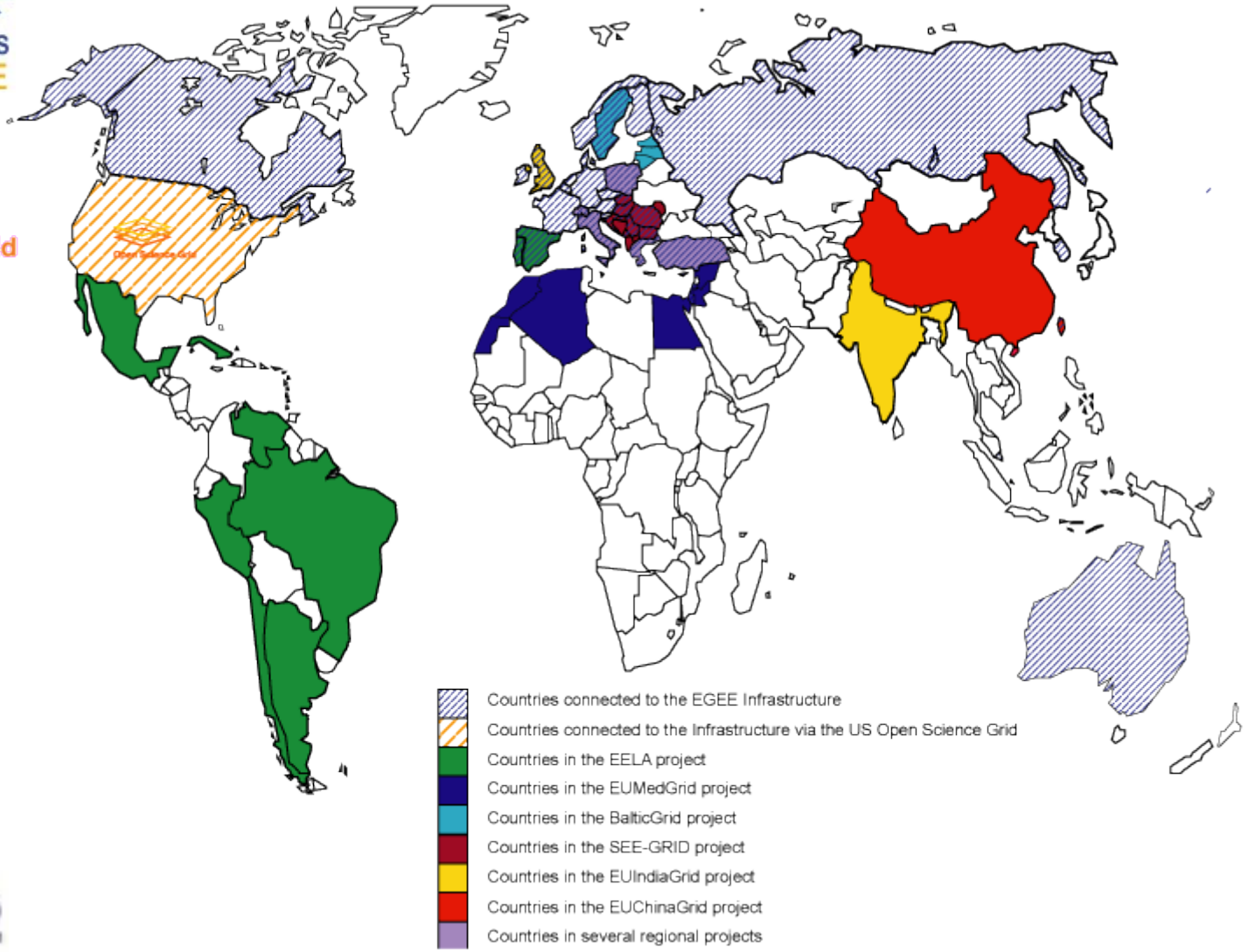
Open Science Grid

TeraGrid

Distributed European Infrastructure for Supercomputing Applications

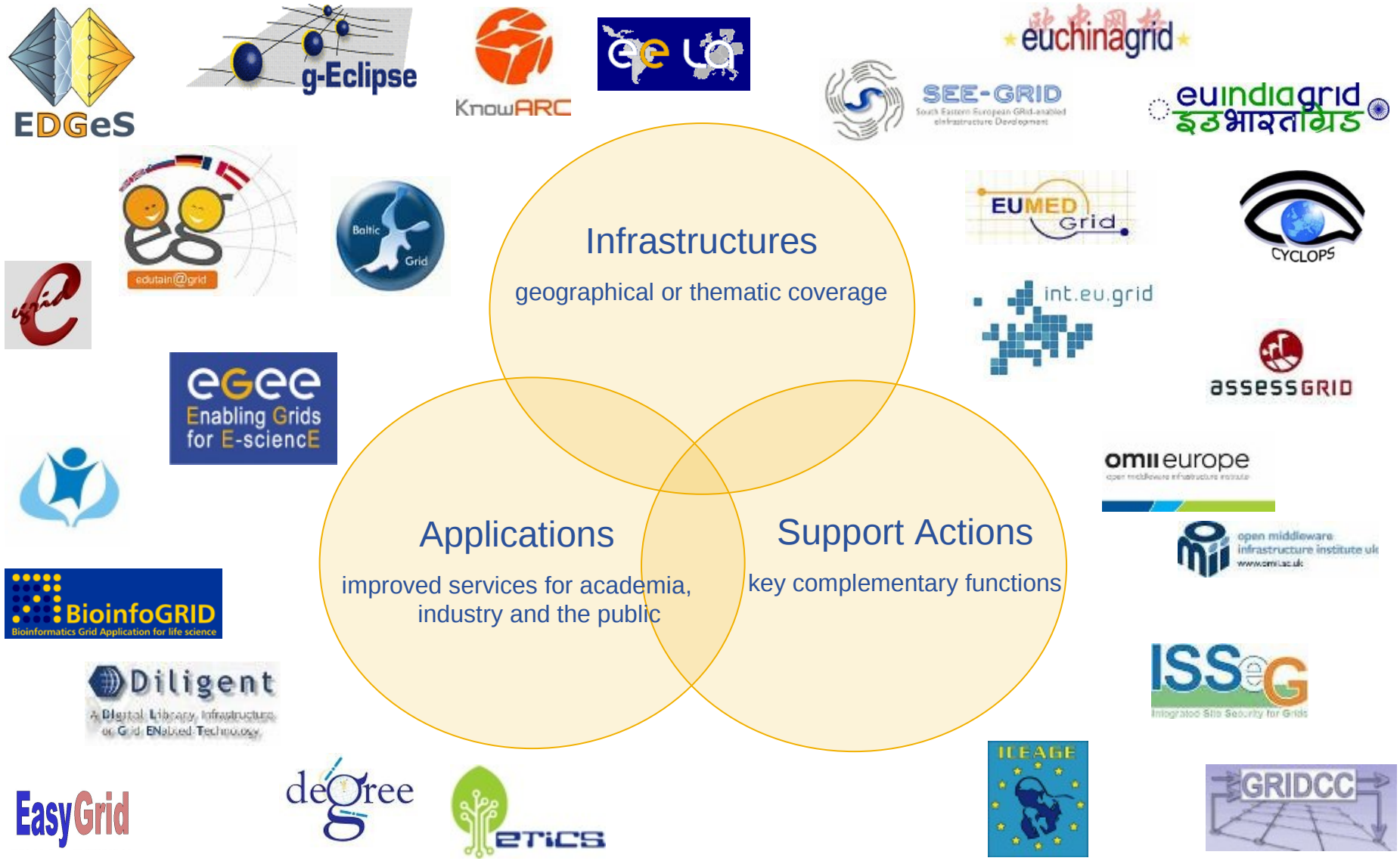
NAREGI

GÉANT2



Registered Collaborating Projects

27 projects have registered as on September 15, 2008: web page

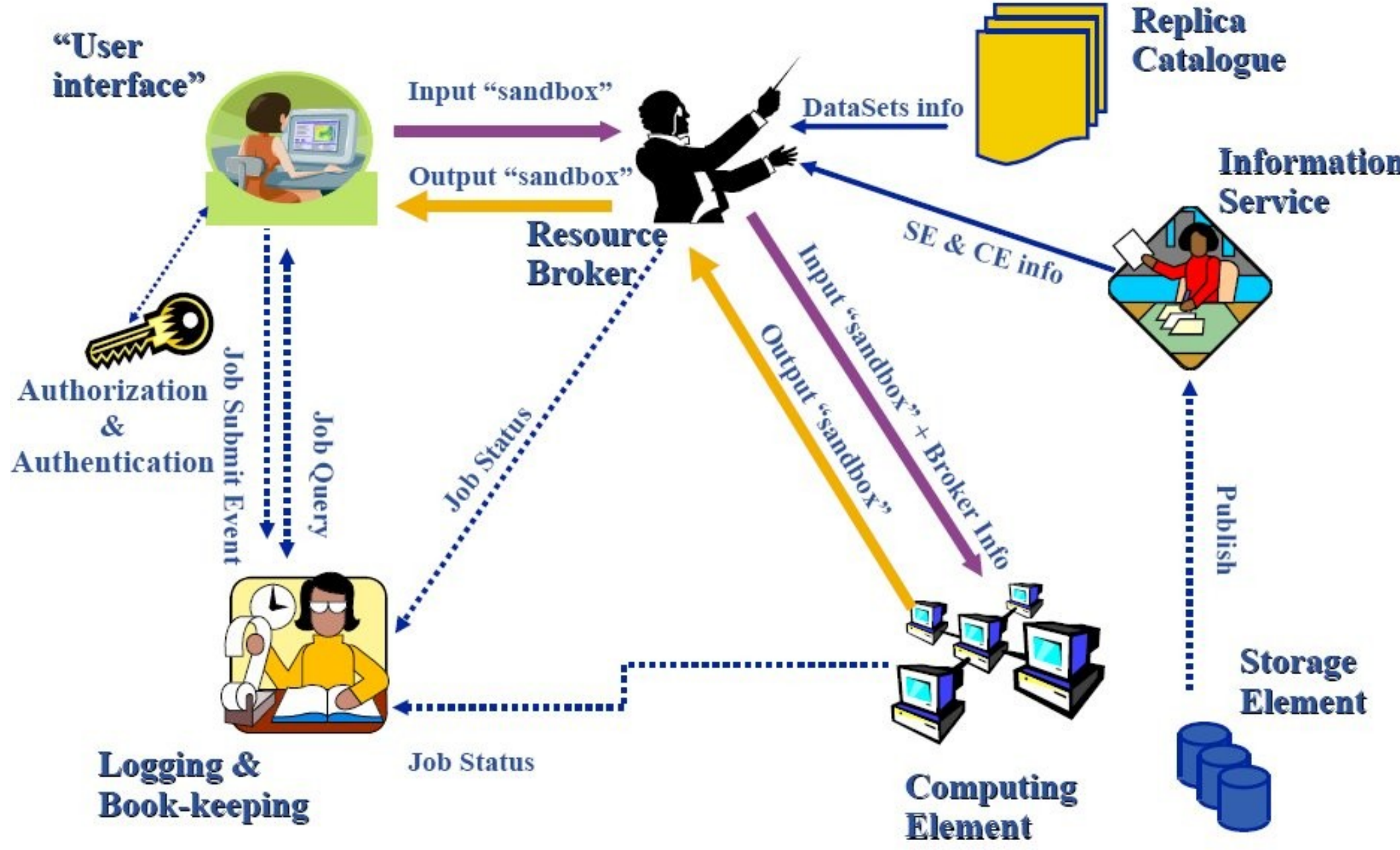


Conclusions

- **Grids are all about sharing: they are a means of working with groups around the world**
  - Today we have a window of opportunity to move grids from research prototypes to permanent production systems (as networks did a few years ago)
- **Interoperability is key to providing the level of support required for our user communities**
- **EGEE operates the world's largest multi-disciplinary grid infrastructure for scientific research**
  - In constant and significant production use
- **Need to prepare the long-term**
  - EGEE, collaborating projects, national grid initiatives and user communities are working to define a model for a sustainable grid infrastructure that is independent of short project cycles

Thanks for listening

Any questions?


"User interface"
Input "sandbox"
Output "sandbox"
Resource Broker
DataSets info
Replica Catalogue
Information Service
SE & CE info
Authorization & Authentication
Job Submit Event
Job Query
Job Status
Input "sandbox" + Broker Info
Output "sandbox"
Publish
Storage Element
Logging & Book-keeping
Job Status
Computing Element