Enabling Grids for E-science in Europe

Grid developments and middleware components

Mike Mineter
EGEE Training team
mjm@nesc.ac.uk

http://egee-intranet.web.cern.ch

EGEE is a project funded by the European Union under contract IST-2003-508833

Acknowledgements

This presentation for the GGF Summer School, 2004 was prepared by the NeSC Edinburgh training team. It includes slides and information from many sources:

- Roberto Barbera (Slides on middleware are based on presentations given in Edinburgh, April 2004)
- Malcolm Atkinson and Ian Bird (*Sites in LCG-2/EGEE-0* at GGF-11)
- Other colleagues in EGEE (project overview slides)
- The European DataGrid training team
- Authors of the LCG-2 User Guide v. 2.0 : Antonio Delgado Peris, Patricia Méndez Lorenzo, Flavia Donno, Andrea Sciabà, Simone Campana, Roberto Santinelli
  https://edms.cern.ch/file/454439//LCG-2-UserGuide.html

Outline

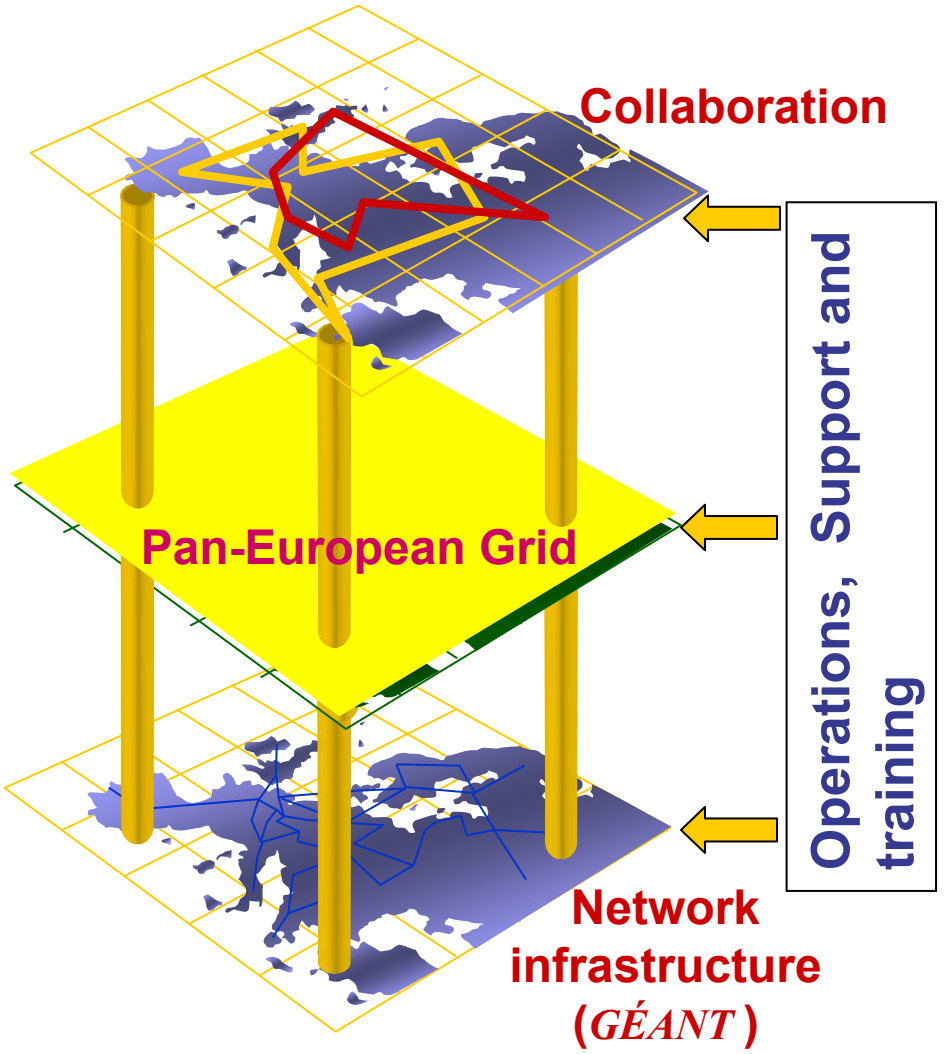
Enabling Grids for E-science in Europe

- Grid developments from an EGEE perspective:
  - Creating e-Infrastructure
  - Building on and with other Grid projects
  - Towards service-orientation
  - Establishing a “production Grid”
- Overview of the middleware of the current EGEE-0 system
  - Major components
  - Lifecycle of a job
- Summary

Towards a European e-Infrastructure

Enabling Grids for E-science in Europe

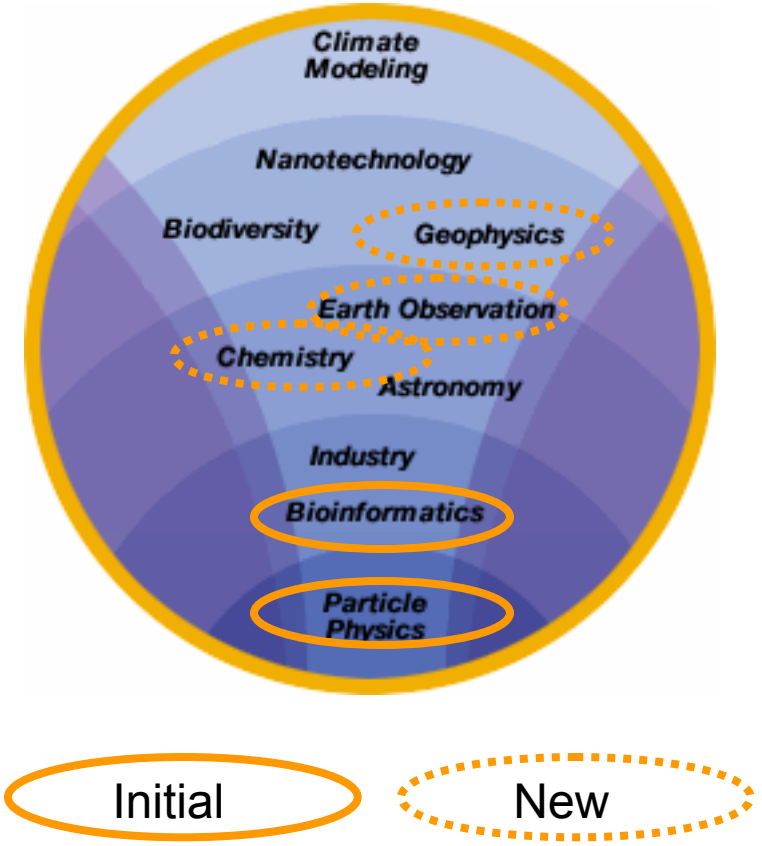
- To underpin European science and technology in the service of society
- To link with and build on
  - National, regional and international initiatives
  - Emerging technologies (e.g. fibre optic networks)
- To foster international cooperation
  - both in the creation and the use of the e-infrastructure



In 2 years EGEE will:

Enabling Grids for E-science in Europe

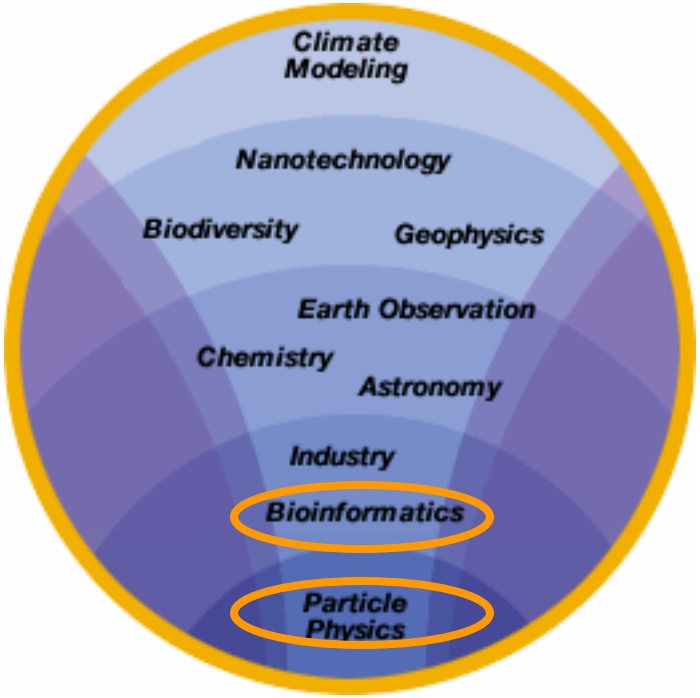
- **Establish production quality sustained Grid services**
  - 3000 users from at least 5 disciplines
  - over 20,000 CPU's, 50 sites
  - over 5 Petabytes (10^{15}) storage
- Demonstrate a viable general process to **bring other scientific communities on board**
- **Spend 32 Million Euros - started April 2004**
  - 70 institutions in 27 countries
- **Propose a second phase** in mid 2005 to take over EGEE in early 2006



EGEE will:

Enabling Grids for E-science in Europe

- **Establish production quality sustained Grid services**
  - 3000 users from at least 5 disciplines
  - over 20,000 CPU's
- Demonstrate a viable general process to **bring other scientific communities on board**
- **Spend 32 Million Euros over 2 years starting April 2004**
  - 70 institutions in 28 countries
- **Propose a second phase** in mid 2005 to take over EGEE in early 2006



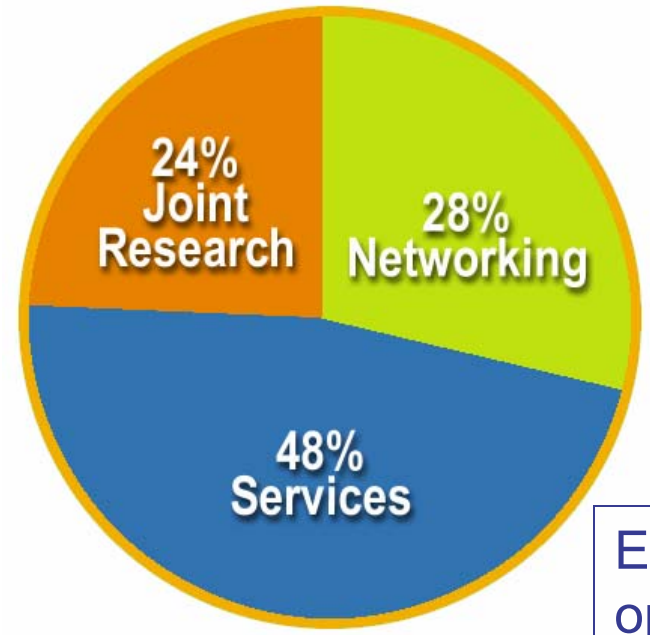
Initial domains

EGEE activity groups

Enabling Grids for E-science in Europe

24% Joint Research

1: Middleware Engineering and Integration
2: Quality Assurance
3: Security
4: Network Services Development



28% Networking

1: Management
2: Dissemination and Outreach
3: User Training and Education
4: Application Identification and Support
5: Policy and International Cooperation

48% Services

1: Grid Operations, Support and Management
2: Network Resource Provision

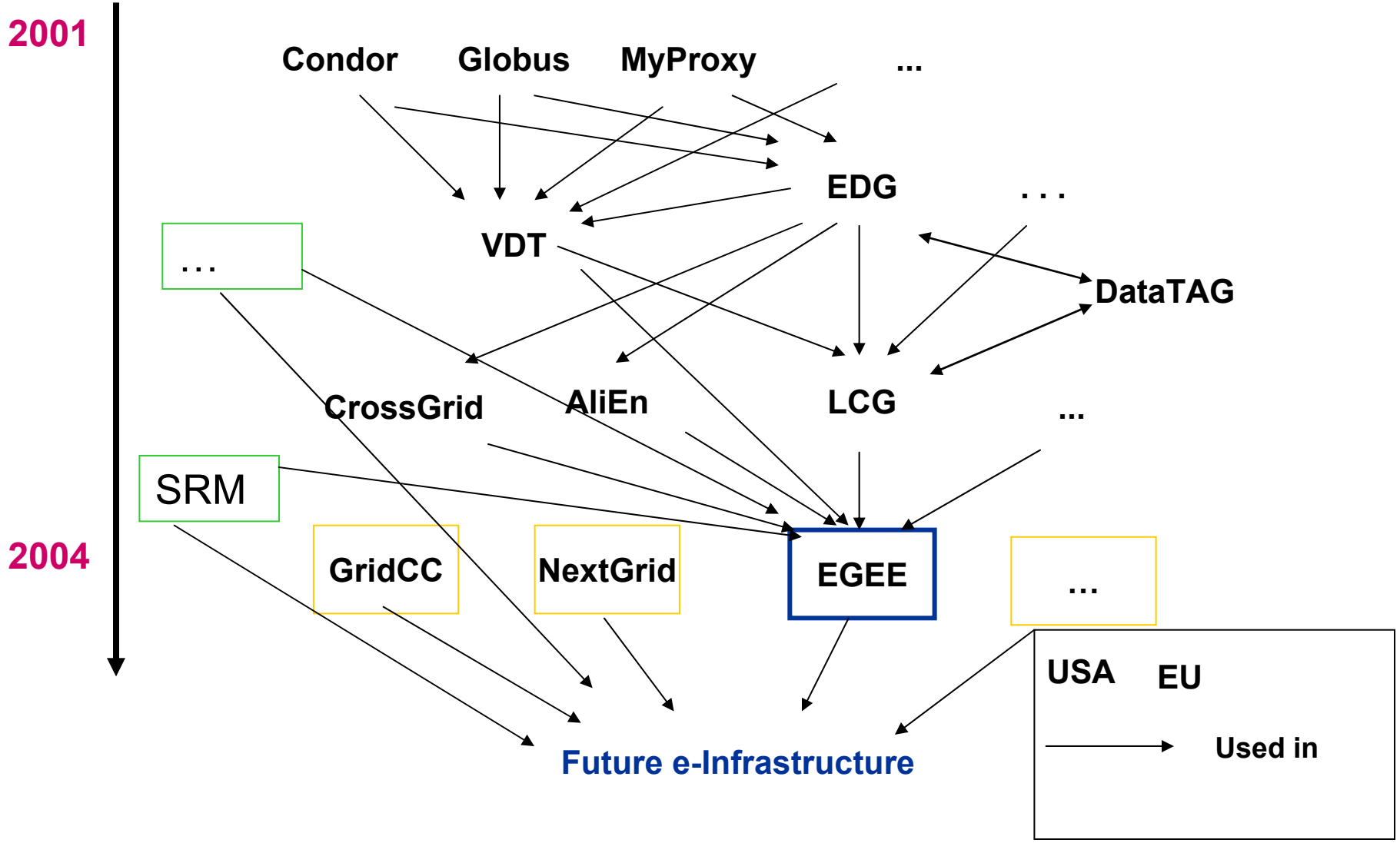
Emphasis in EGEE is on operating a production grid and supporting the end-users

Outline

Enabling Grids for E-science in Europe

- Grid developments from an EGEE perspective:
  - Creating e-Infrastructure
  - Building on and with other Grid projects
  - Towards service-orientation
  - Establishing a “production Grid”
- Overview of the middleware of the current EGEE-0 system
  - Major components
  - Lifecycle of a job
- Summary

EGEE view of history



Outline

Enabling Grids for E-science in Europe

- Grid developments from an EGEE perspective:
  - Creating e-Infrastructure
  - Building on and with other Grid projects
  - Towards service-orientation
  - Establishing a “production Grid”
- Overview of the middleware of the current EGEE-0 system
  - Major components
  - Lifecycle of a job
- Summary

Service orientation: building EGEE-1

Enabling Grids for E-science in Europe

- "gLite" - the new EGEE middleware (under test)
- Service oriented - components that are :
  - Loosely coupled (by messages – examples tomorrow)
  - Accessible across network; modular and self-contained; clean modes of failure
  - So can change implementation without changing interfaces
  - Can be developed in anticipation of new uses
- … and are based on standards. Opens EGEE to:
  - New middleware (plethora of tools now available)
  - Heterogeneous resources (storage, computation…)
  - Interact with other Grids (international, regional and national)

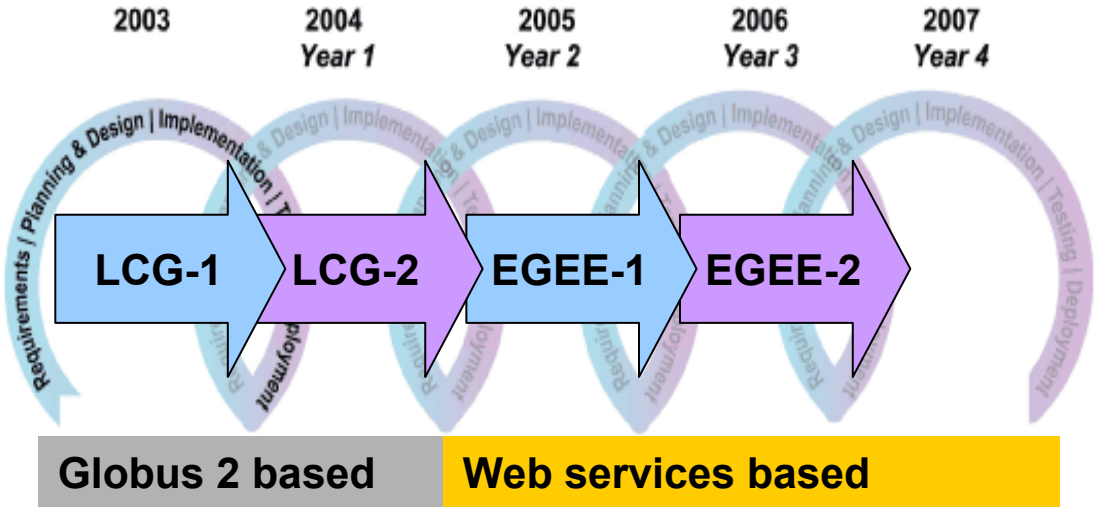
Outline

Enabling Grids for E-science in Europe

- Grid developments from an EGEE perspective:
  - Creating e-Infrastructure
  - Building on and with other Grid projects
  - Towards service-orientation
  - Establishing a “production Grid”
- Overview of the middleware of the current EGEE-0 system
  - Major components
  - Lifecycle of a job
- Summary

LCG and EGEE

- LCG: Large Hadron Collider Computing Grid
- LCG infrastructure running LCG-2 is “EGEE-0”
- In parallel producing new web-service-oriented middleware (“gLite”)
- Will replace LCG-2 as production facility in 2005
- New major releases each year



Sites in LCG-2/EGEE-0 : June 4 2004

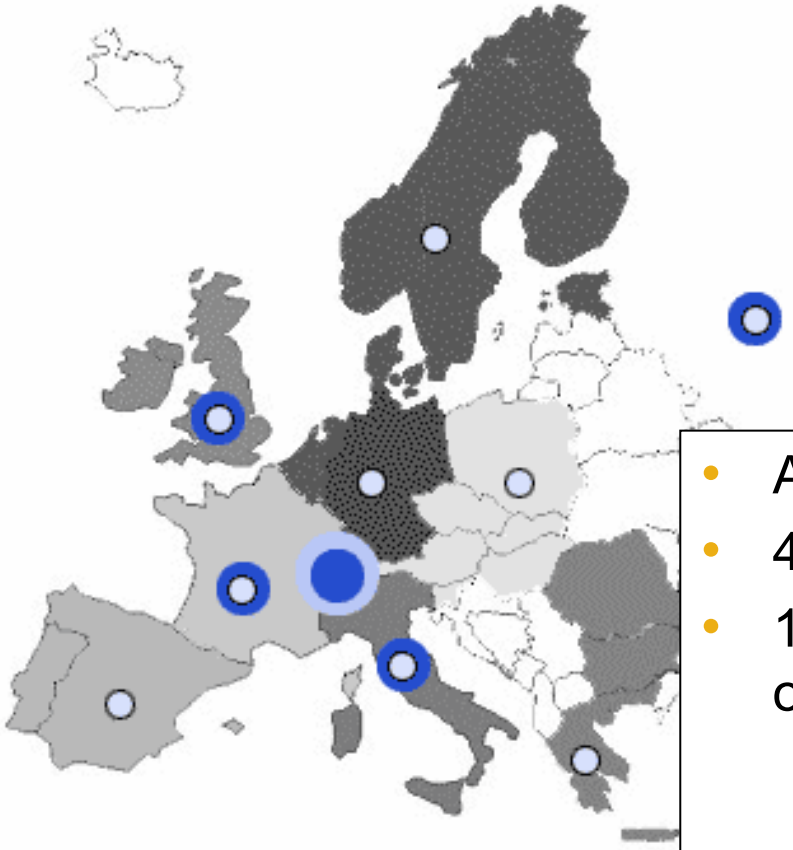
Enabling Grids for E-science in Europe

http://goc.grid-support.ac.uk/gppmonWorld/gppmon_maps/CERN_lxn1188.html

- 22 Countries
- 58 Sites (45 Europe, 2 US, 5 Canada, 5 Asia, 1 HP)
  - Coming: New Zealand, China, other HP (Brazil, Singapore)
- 3800 cpu

Operations Infrastructure

Enabling Grids for E-science in Europe



Operations Management Centre

Core Infrastructure Centre

Regional Operations Centre

- A lot more than middleware!!
- 40% of EGEE budget
- 10 ROCs: coordinate deployment & operation, tasks include:
  - First point of contact for all new sites, new users, and user support; Issue certificates
  - Negotiate policies with resource providers
- 5 CICs: tasks include provision of
  - VO services, core Grid services (RBs, UIs, database services, BDIIs)

EGEE: adding a VO

Enabling Grids for E-science in Europe

EGEE has a formal procedure for adding selected new user communities:

- Negotiation with one of the Regional Operations Centres
- Seek balance between the resources contributed by a VO and those that they consume.
- Resource allocation will be made at the VO level.
- Many resources need to be available to multiple VOs : shared use of resources is fundamental to a Grid

Story so far: themes illustrated by EGEE

Enabling Grids for E-science in Europe

- e-Infrastructure
  - Integrating networks, grids and emerging technologies
  - Based on standards
  - Underpinning research, industry, … the “knowledge economy”
- International, collaborative effort
- Moving to a Service Orientated Architecture
- Focus: Production grids for multiple VOs
  - Demands massive effort in organisation and administration:
    - Operations
    - Support
    - Training

1997- Present: Globus

Enabling Grids for E-science in Europe

- A software toolkit addressing certain technical problems in the development of Grid enabled tools, services, and applications
  - Offers a modular “bag of technologies”
  - Made available under liberal open source license
- *Not* turnkey solutions, but *building blocks* and *tools* for application developers and system integrators

Globus: Key components

Enabling Grids for E-science in Europe

- Grid Security Infrastructure (GSI)
  - X.509 authentication with delegates and single sign-on
- Grid Resource Allocation Mgmt (GRAM)
  - Remote allocation, monitoring of job, control of compute resources
- GridFTP protocol (FTP extensions)
  - High-performance data access & transport
- Grid Resource Information Service (GRIS) + Monitoring and Discovery Service (MDS)
  - Access to structure & state information
- XIO library
  - TCP, UDP, IP multicast, and file I/O
- Others…

Notes - VDT

Enabling Grids for E-science in Europe

- “The Virtual Data Toolkit (VDT) is an ensemble of grid middleware that can be easily installed and configured. In our experience, installing grid software is challenging and time consuming. The goal of the VDT is to make it as easy as possible for users to deploy, maintain and use grid middleware.” http://www.cs.wisc.edu/vdt/

Virtual Data Toolkit

Enabling Grids for E-science in Europe

- http://www.cs.wisc.edu/vdt/
- Condor Group
  - Condor/Condor-G
  - DAGMan
  - Fault Tolerant Shell
  - ClassAds
- Globus Alliance
  - Job submission (GRAM)
  - Information service (MDS)
  - Data transfer (GridFTP)
  - Replica Location (RLS)
- EDG & LCG
  - Make Gridmap
  - Certificate Revocation List Updater
  - GLUE Schema
- ISI & UC
  - Chimera & Pegasus
- NCSA
  - MyProxy
  - GSI OpenSSH
  - UberFTP
- LBL
  - PyGlobus
  - Netlogger
- Caltech
  - MonaLisa
- VDT
  - VDT System Profiler
  - Configuration software
- Others
  - KX509 (U. Mich.)

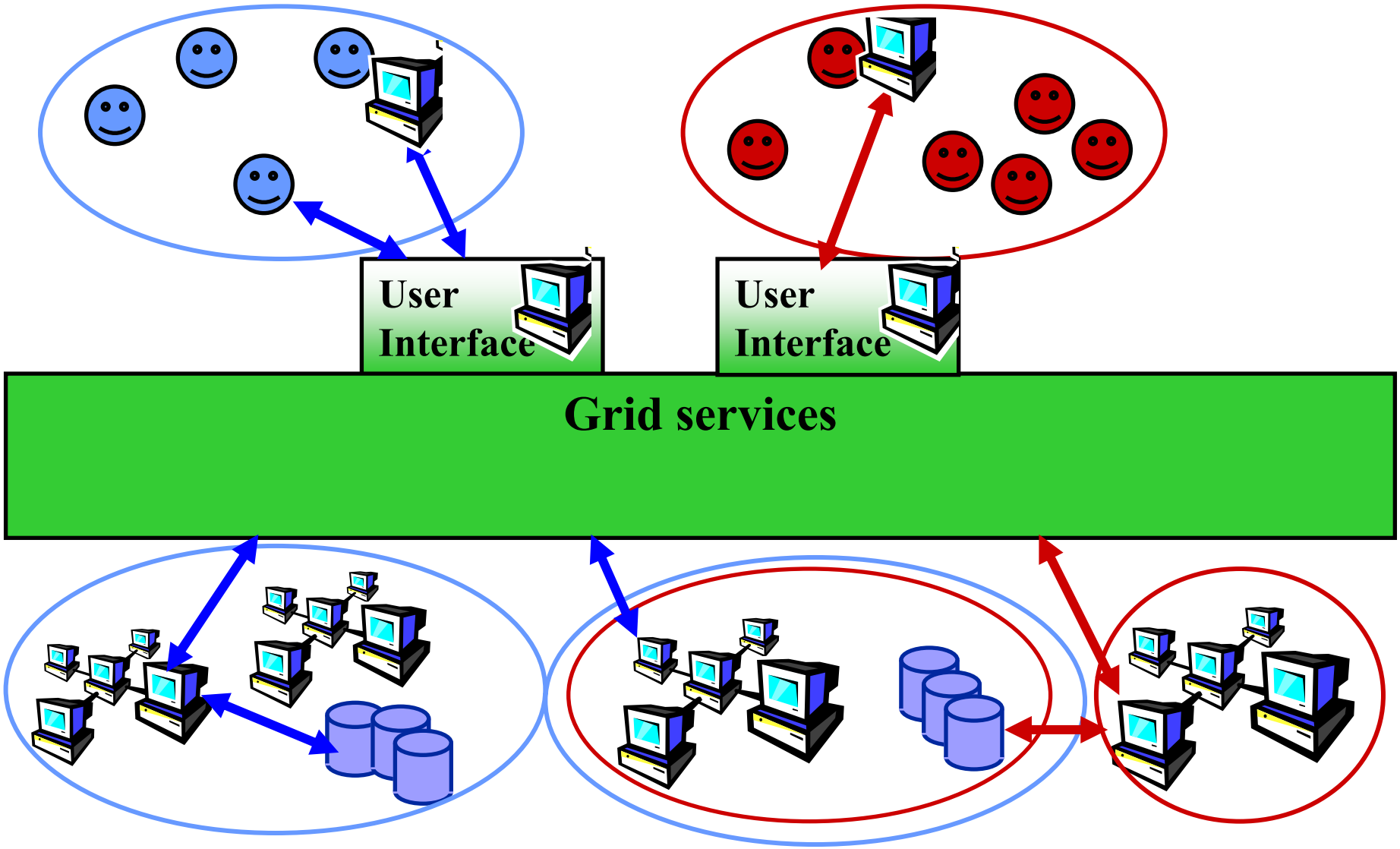
Outline

Enabling Grids for E-science in Europe

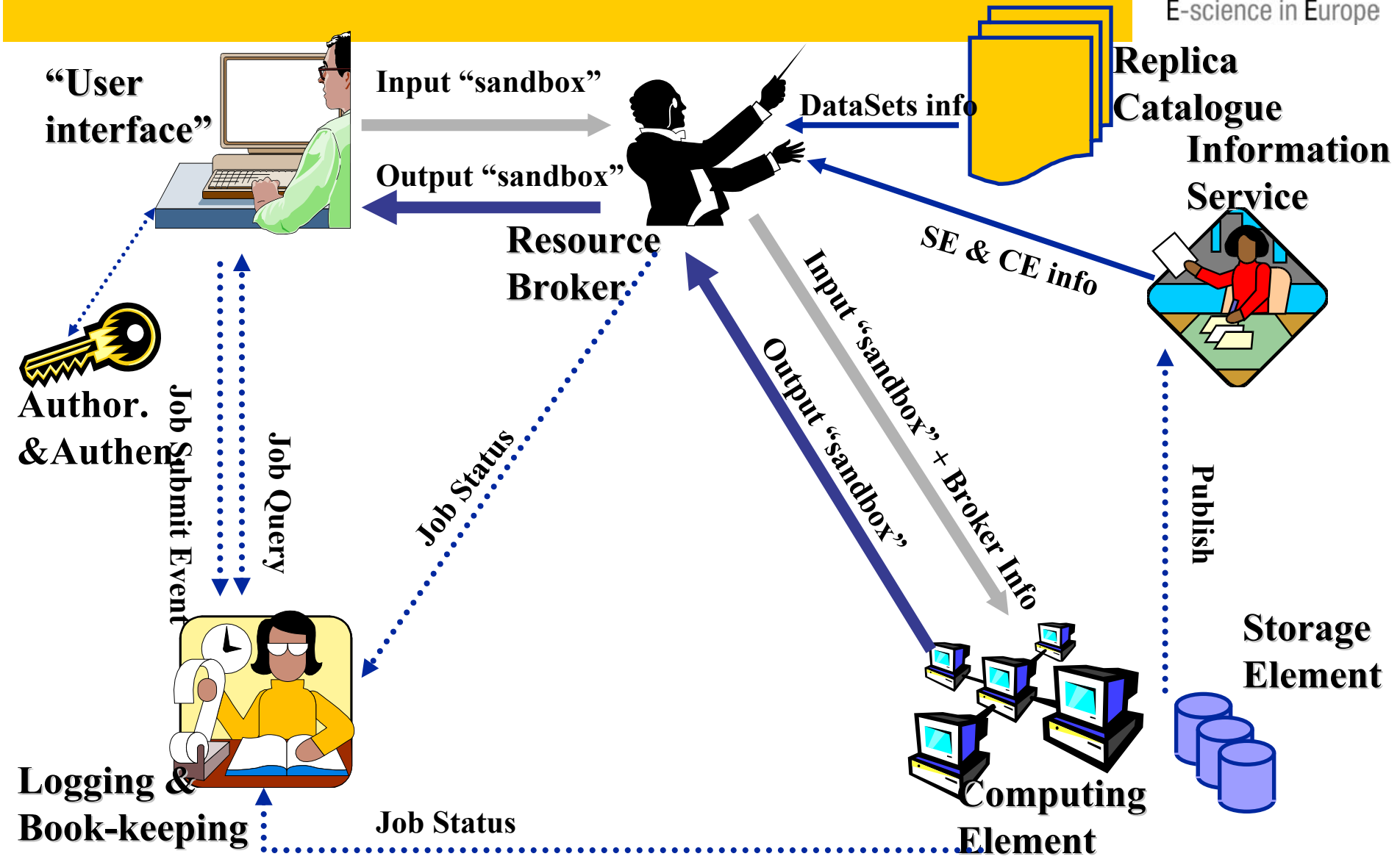
- Grid developments from an EGEE perspective:
  - Creating e-Infrastructure
  - Building on and with other Grid projects
  - Towards service-orientation
  - Establishing a “production Grid”
- Overview of the middleware of the current EGEE-0 system
  - Major components
  - Lifecycle of a job
- Summary

User-view of EGEE: a multi-VO Grid

Enabling Grids for E-science in Europe



Middleware components



Workload Management System (WMS)

Enabling Grids for E-science in Europe

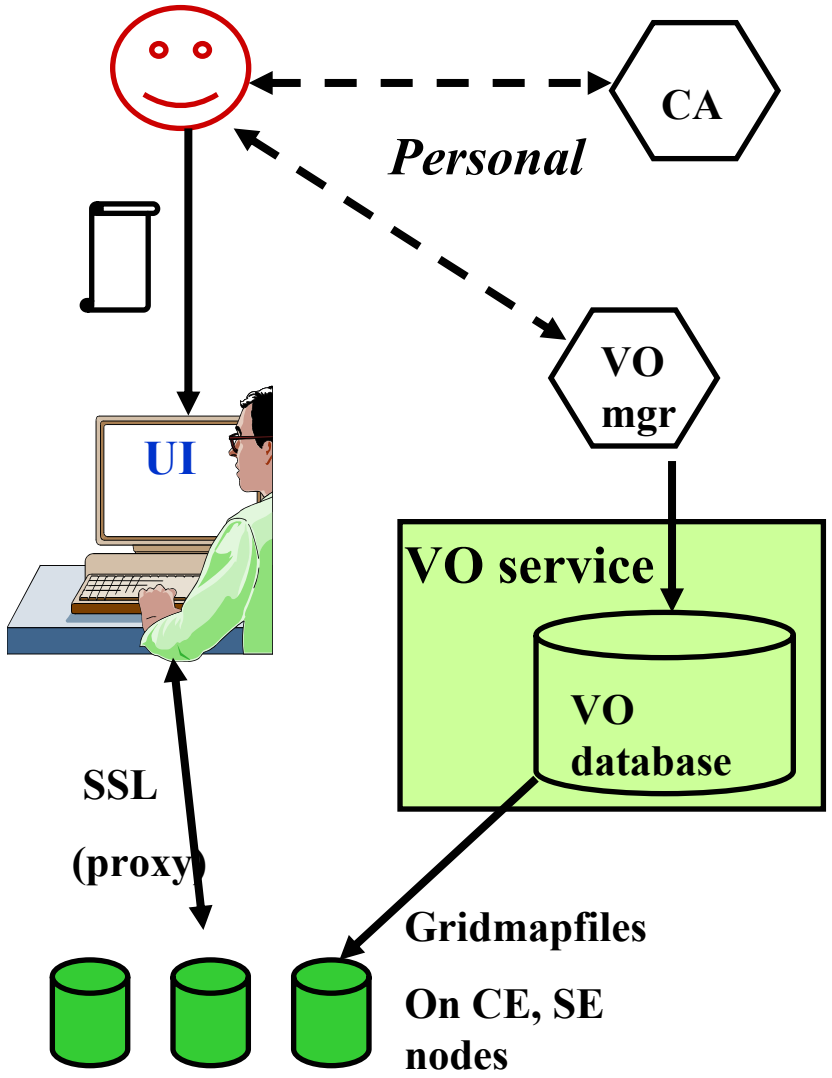
- Distributed scheduling
  - multiple UI's where you submit your job
  - multiple RB's from where the job is sent to a CE
  - multiple CE's where the job can be put in a queuing system
- Distributed resource management
  - multiple information systems that monitor the state of the grid
  - Information from SE, CE, sites

Authentication, Authorisation

- Authentication
  - User obtains certificate from CA
  - Connects to UI by ssh
  - Downloads certificate
  - Invokes Proxy server
  - Single logon – to UI - then Secure Socket Layer with proxy identifies user to other nodes

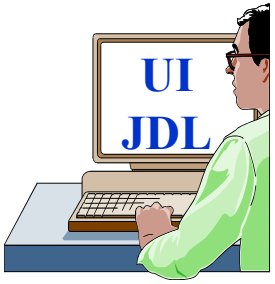
- Authorisation - currently
  - User joins Virtual Organisation
  - VO negotiates access to Grid nodes and resources (CE, SE)
  - Authorisation tested by CE, SE:

gridmapfile maps user to local account



User Interface node

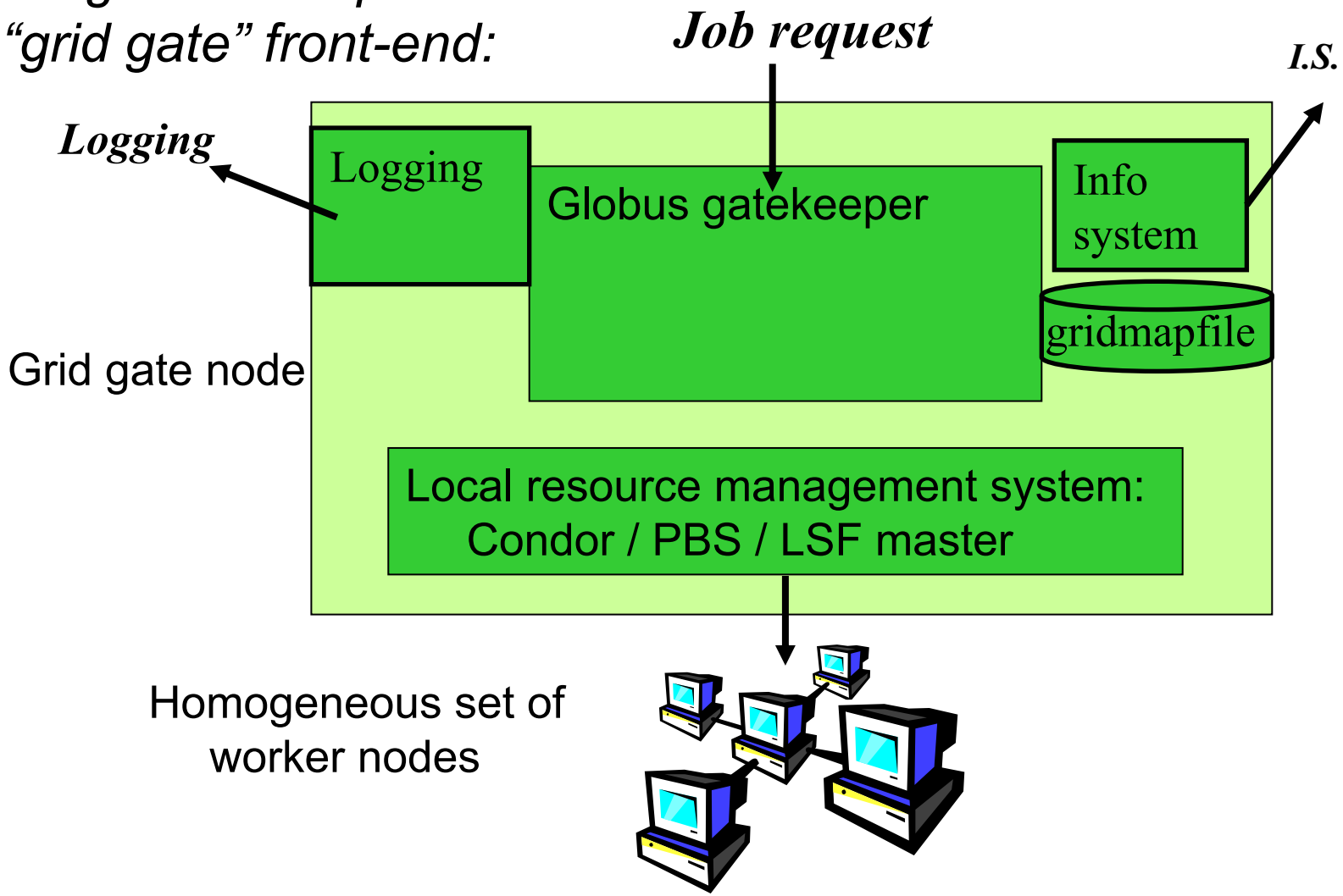
- The user's interface to the Grid
- Command-line interface to
  - Proxy server
  - Job operations
    - To submit a job
    - Monitor its status
    - Retrieve output
  - Data operations
    - Upload file to SE
    - Create replica
    - Discover replicas
  - Other grid services
- Also C++ and Java APIs



- To run a job user creates a JDL (Job Description Language) file

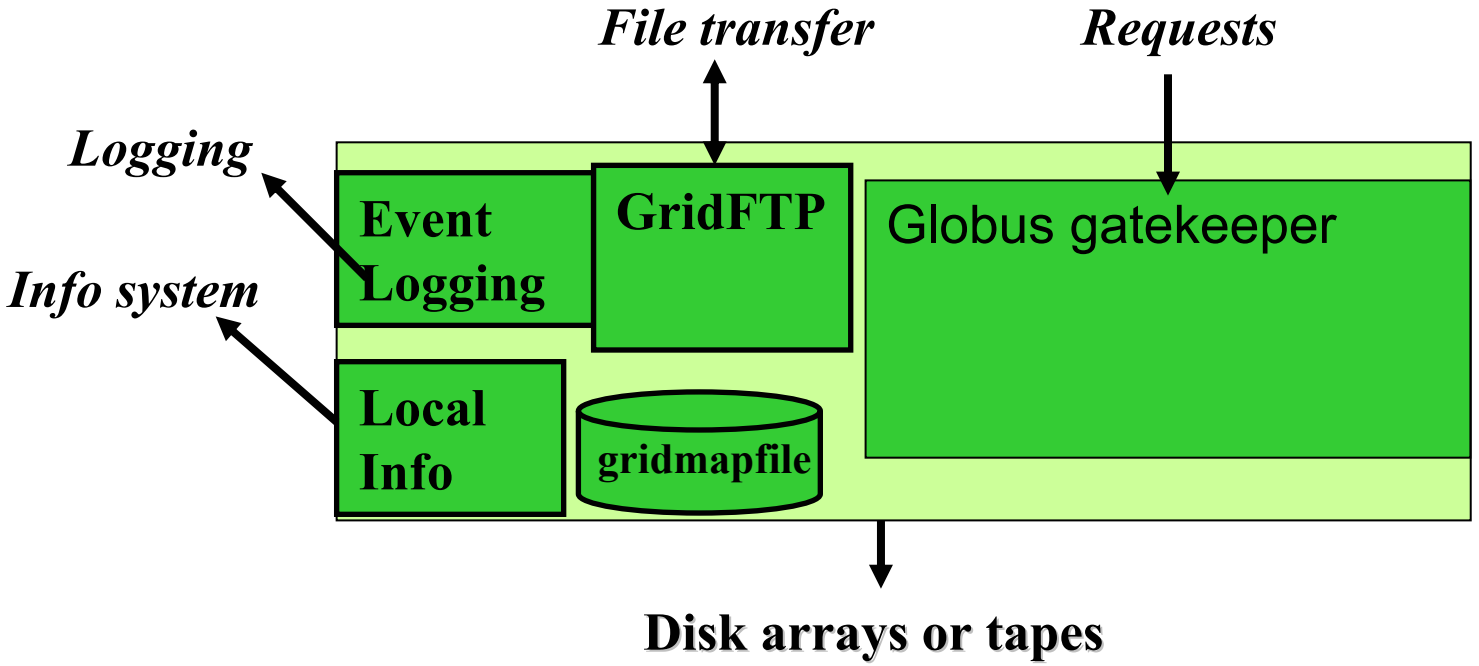
“Compute element” in LCG-2

A CE is a grid batch queue with a “grid gate” front-end:



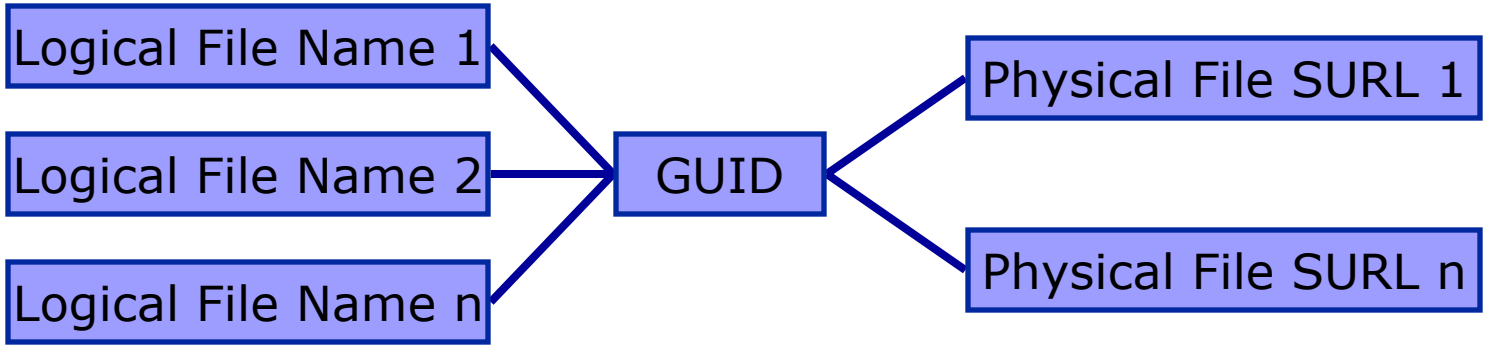
Storage elements and files

- Storage elements hold files: write once, read many
- Replica files can be held on different SE:
  - "close" to CE; share load on SE
- Replica Catalogue - what replicas exist for a file?
- Replica Location Service - where are they?



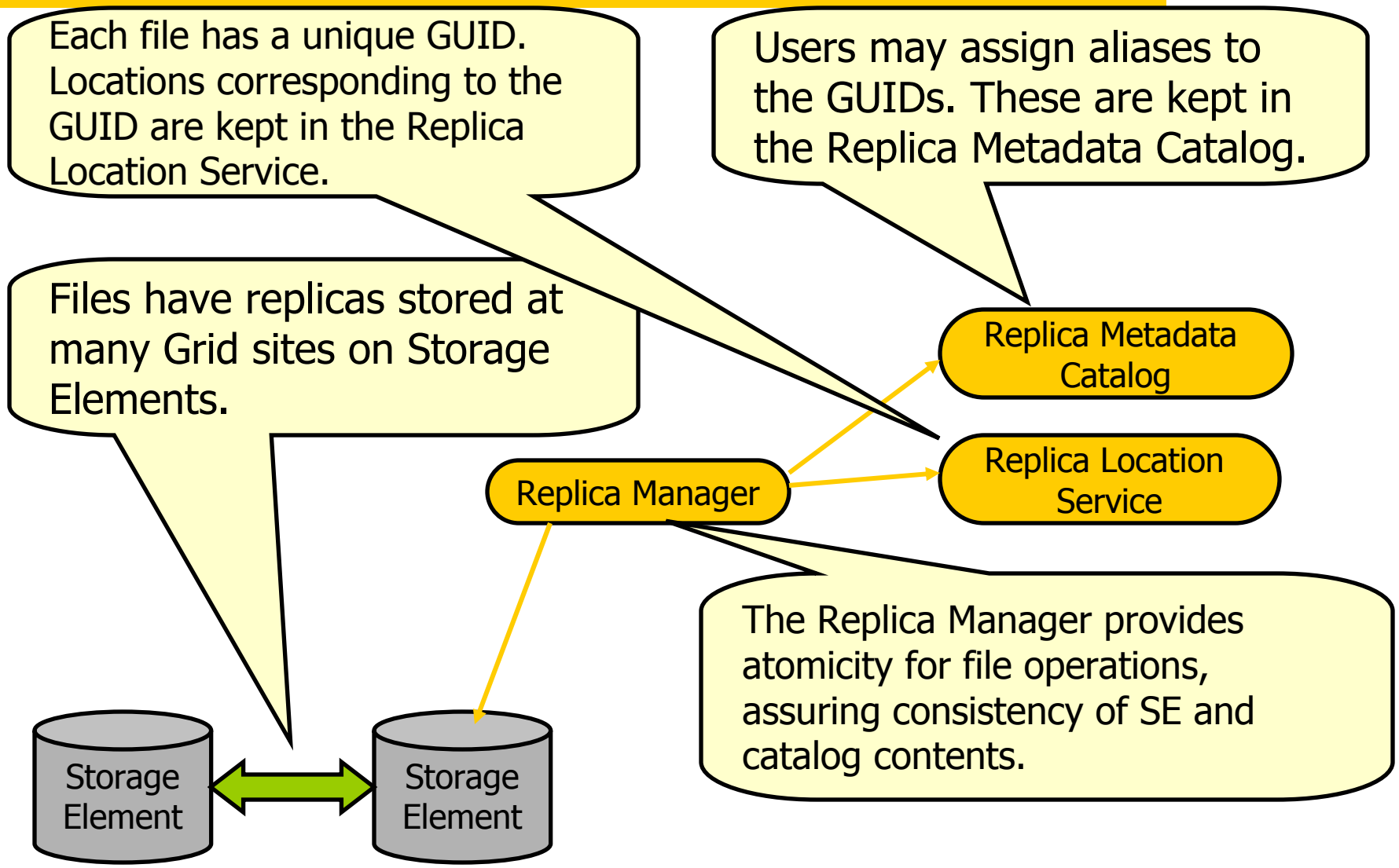
Naming Conventions

- Logical File Name (LFN)
  - An alias created by a user to refer to some item of data e.g. “lfn:cms/20030203/run2/track1”
- Site URL (SURL) (or Physical File Name (PFN))
  - The location of an actual piece of data on a storage system e.g. “srm://pcrd24.cern.ch/flatfiles/cms/output10_1”
- Globally Unique Identifier (GUID)
  - A non-human readable unique identifier for an item of data e.g. “guid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6”



Data Replication Services: Basic Functionality

Enabling Grids for E-science in Europe



Resource Broker nodes

Enabling Grids for E-science in Europe

- Run the Workload Management System
  - To accept job submissions
  - Dispatch jobs to appropriate Compute Element (CE)
  - Allow users
    - To get information about their status
    - To retrieve their output
- A configuration file on each UI node determines which RB node(s) will be used
- When a user submits a job, JDL options are to:
  - Specify CE
  - Allow RB to choose CE (using optional tags to define requirements)
  - Specify SE (then RB finds “nearest” appropriate CE, after interrogating Replica Location Service)

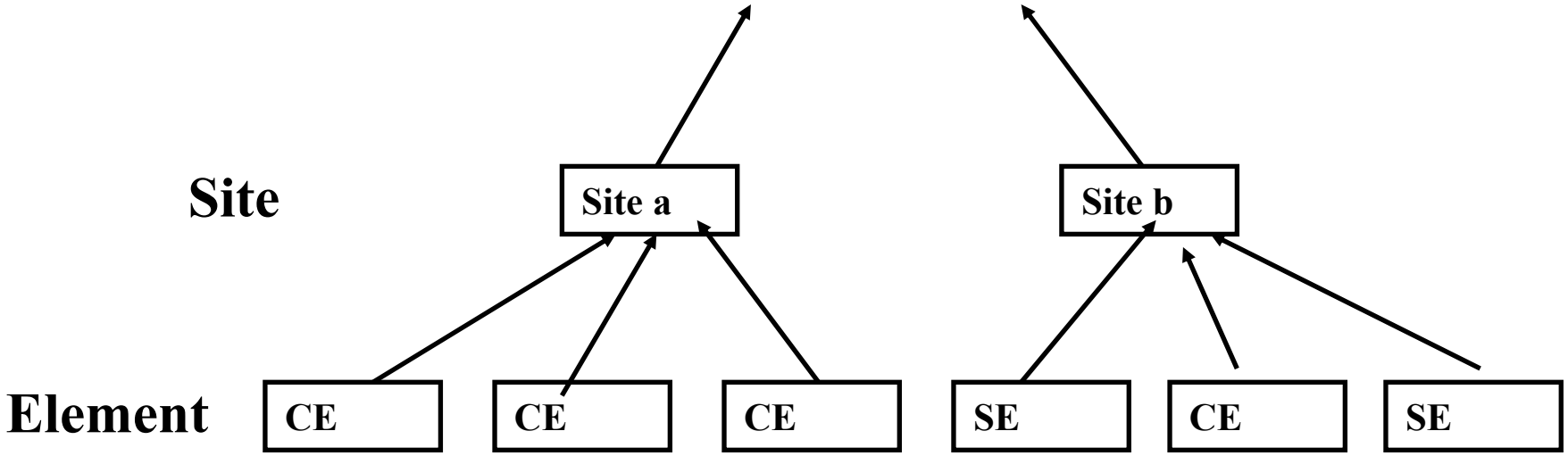
Logging and Book-keeping

Enabling Grids for E-science in Europe

- Who did what when??
- What's happening to my job?
- Usually runs on Resource Broker node
- See LCG-2 user guide for a bit more on this

Information System

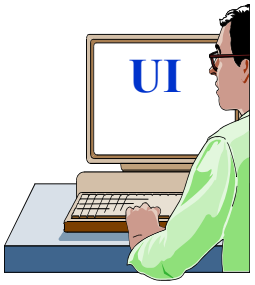
- Receives periodic (~5 minutes) updates from CE, SE
- Used by RB node to determine resources to be used by a job
- “Leaf/node” system: currently BDII is used



Outline

Enabling Grids for E-science in Europe

- Grid developments from an EGEE perspective:
  - Creating e-Infrastructure
  - Building on and with other Grid projects
  - Towards service-orientation
  - Establishing a “production Grid”
- Overview of the middleware of the current EGEE-0 system
  - Major components
  - Lifecycle of a job
- Summary



RB node

Network Server

Workload Manager

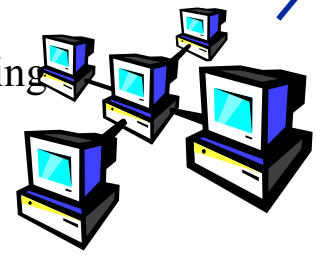
Job Contr. - CondorG

Replica Location Server

Inform. Service

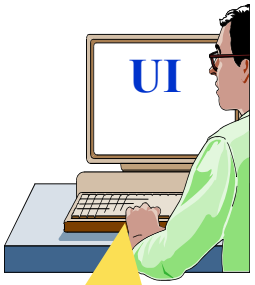
CE characts & status

SE characts & status



Computing Element

Storage Element



UI: allows users to access the functionalities of the WMS
(via command line, GUI, C++ and Java APIs)

RB node

Network Server

Workload Manager

Job Contr. - CondorG

Replica Location Server

Inform. Service

Job Status

submitted

CE characts & status

SE characts & status

Computing Element

Storage Element

edg-job-submit myjob.jdl

Myjob.jdl

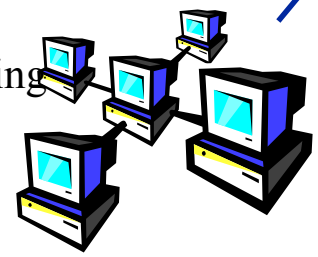
```
JobType = "Normal";
Executable     = "$(CMS)/exe/sum.exe";
InputSandbox = {"/home/user/WP1testC","/home/file*",
"/home/user/DATA/*"};
OutputSandbox = {"sim.err", "test.out", "sim.log"};
Requirements   = other. GlueHostOperatingSystemName ==
"linux" &&
other. GlueHostOperatingSystemRelease == "Red Hat 7.3"
&& other.GlueCEPolicyMaxCPUTime > 10000;
Rank  = other.GlueCEStateFreeCPUs;
```

Job Statu s

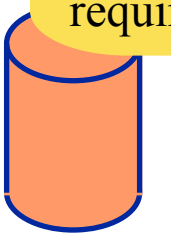
submitted

CE characts & status

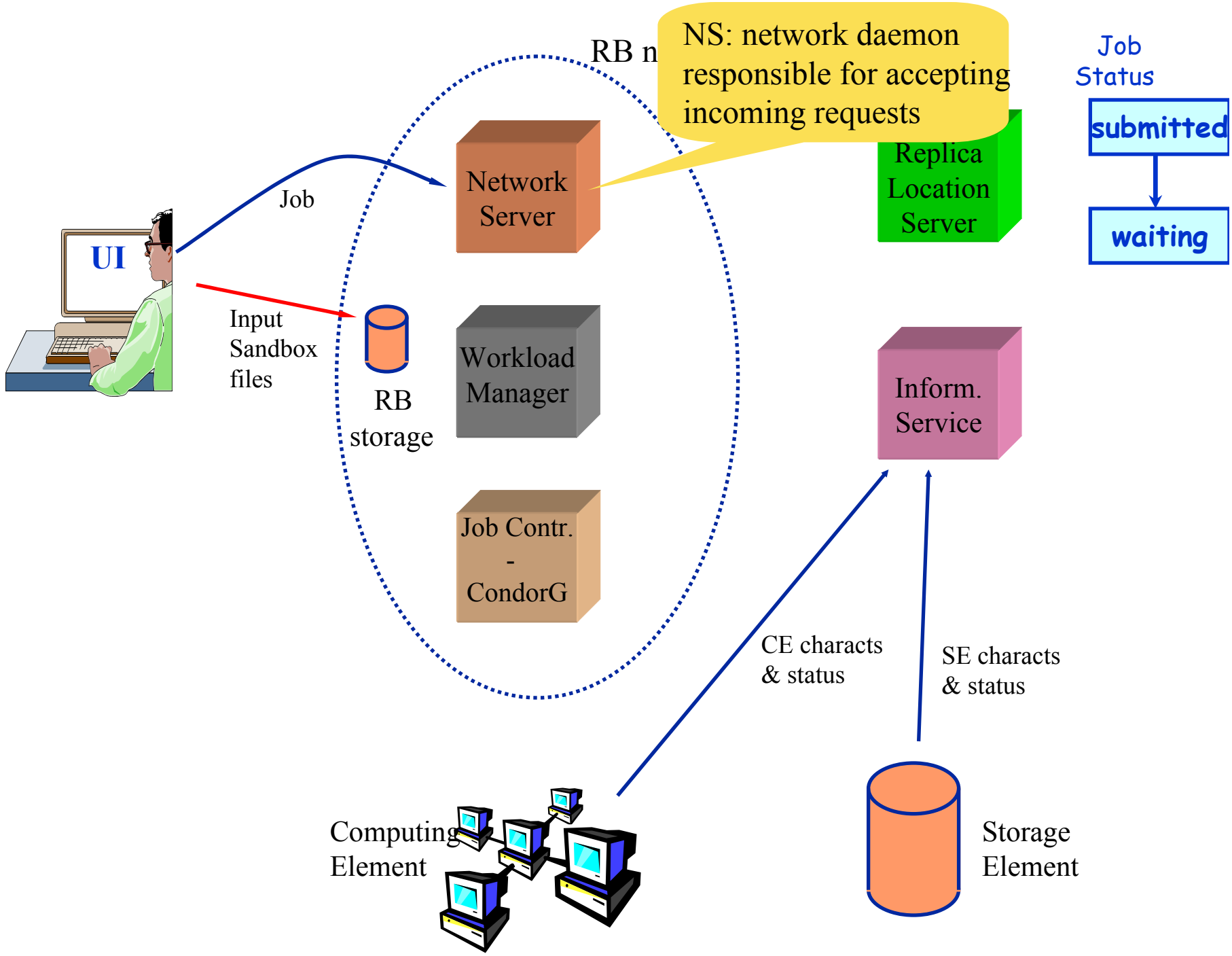
Job Description Languag (JDL) to specify job characteristics and requirements



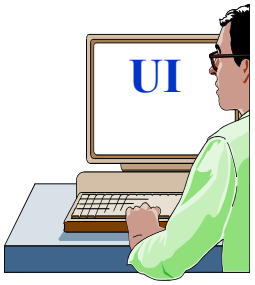
Computing Element



Storage Element


RB n
NS: network daemon responsible for accepting incoming requests
Job Status
submitted
waiting
Job
UI
Network Server
Replica Location Server
Input Sandbox files
RB storage
Workload Manager
Inform. Service
Job Contr. - CondorG
CE characts & status
SE characts & status
Computing Element
Storage Element

Job submission



RB node

Network Server

Job

Workload Manager

RB storage

Job Contr. - CondorG

WM: responsible to take the appropriate actions to satisfy the request

Replica Location Server

Inform. Service

Job Status

submitted

waiting

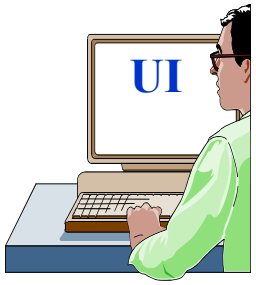
CE characts & status

SE characts & status

Computing Element

Storage Element

Job submission



RB node

Network Server

Match-Maker/ Broker

Replica Location Server

Job Status

submitted

waiting

RB storage

Workload Manager

Where must this job be executed ?

Inform. Service

Job Contr. - CondorG

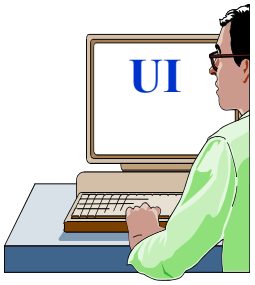
CE characts & status

SE characts & status

Computing Element

Storage Element

Job submission



RB node

Matchmaker: responsible to find the "best" CE where to submit a job

Match-Maker/ Broker

Workload Manager

RB storage

Job Contr. - CondorG

Replica Location Server

Inform. Service

Job Status

submitted

waiting

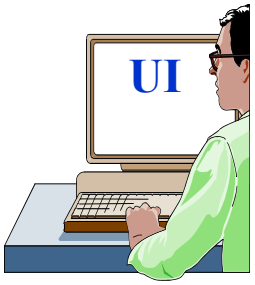
CE characts & status

SE characts & status

Computing Element

Storage Element

Job submission



RB node

Network Server

Workload Manager

Job Contr. - CondorG

RB storage

Match-Maker/ Broker

Where are (which SEs) the needed data ?

Replica Location Server

What is the status of the Grid ?

Inform. Service

Job Status

submitted

waiting

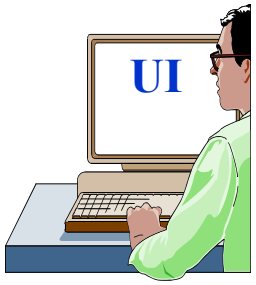
CE characts & status

SE characts & status

Computing Element

Storage Element

Job submission



RB node

Network Server

Match-Maker/ Broker

Workload Manager

CE choice

RB storage

Job Contr. - CondorG

Replica Location Server

Inform. Service

Job Status

submitted

waiting

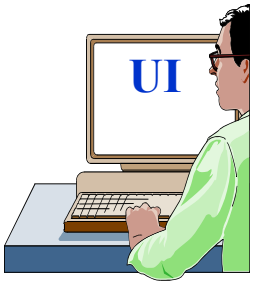
CE characts & status

SE characts & status

Computing Element

Storage Element

Job submission



RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Job Adapter

Replica Location Server

Inform. Service

Job Status

submitted

waiting

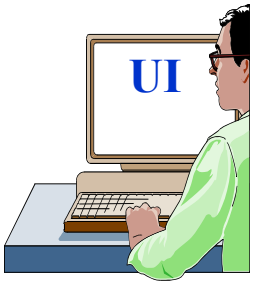
JA: responsible for the final "touches" to the job before performing submission (e.g. creation of wrapper script, etc.)

SE characts & status

Computing Element

Storage Element

Job submission



RB node

Network Server

RB storage

Workload Manager

Job

Job Contr. - CondorG

JC: responsible for the actual job management operations (done via CondorG)

Replica Location Server

Inform. Service

CE characts & status

SE characts & status

Computing Element

Storage Element

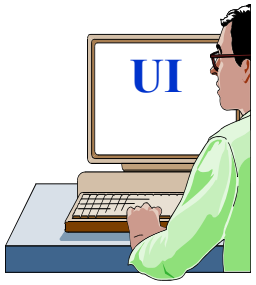
Job Status

submitted

waiting

ready

Job submission



RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Replica Location Server

Inform. Service

Job Status

submitted

waiting

ready

scheduled

Input Sandbox files

Job

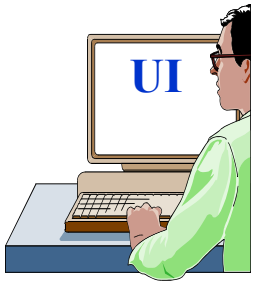
CE characts & status

SE characts & status

Computing Element

Storage Element

Job submission



RB node

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Input Sandbox

Replica Location Server

Inform. Service

Job Status

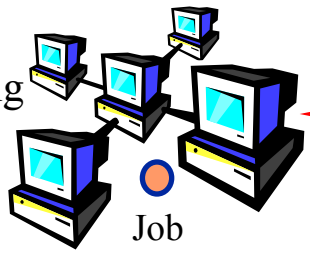
submitted

waiting

ready

scheduled

running



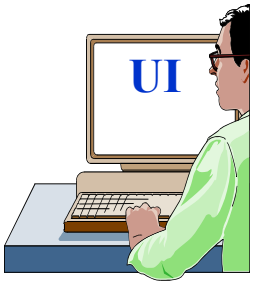
Computing Element

Job

"Grid enabled" data transfers/ accesses

Storage Element

Job submission



RB node

Network Server

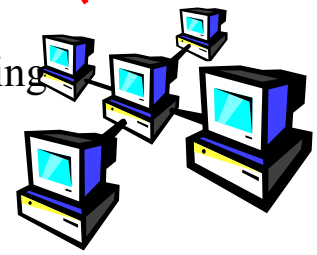
Workload Manager

RB storage

Job Contr. - CondorG

Output Sandbox files

Computing Element



Replica Location Server

Inform. Service

Storage Element

Job Status

submitted

waiting

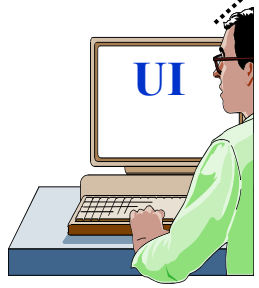
ready

scheduled

running

done

Job submis



RB node

edg-job-get-output <dg-job-id>

Network Server

Workload Manager

RB storage

Job Contr. - CondorG

Output Sandbox

Replica Location Server

Inform. Service

Job Status

submitted

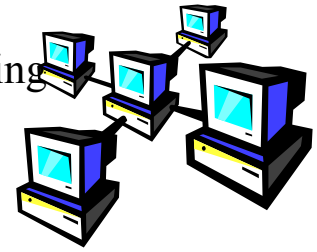
waiting

ready

scheduled

running

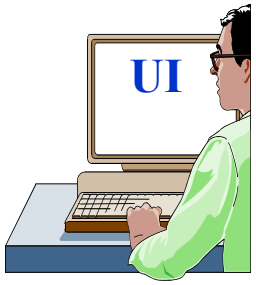
done



Computing Element

Storage Element

Job submission



Output Sandbox files

RB node

Network Server

Workload Manager

Job Contr. - CondorG

RB storage

Replica Location Server

Inform. Service

Job Status

submitted

waiting

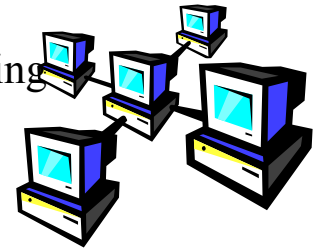
ready

scheduled

running

done

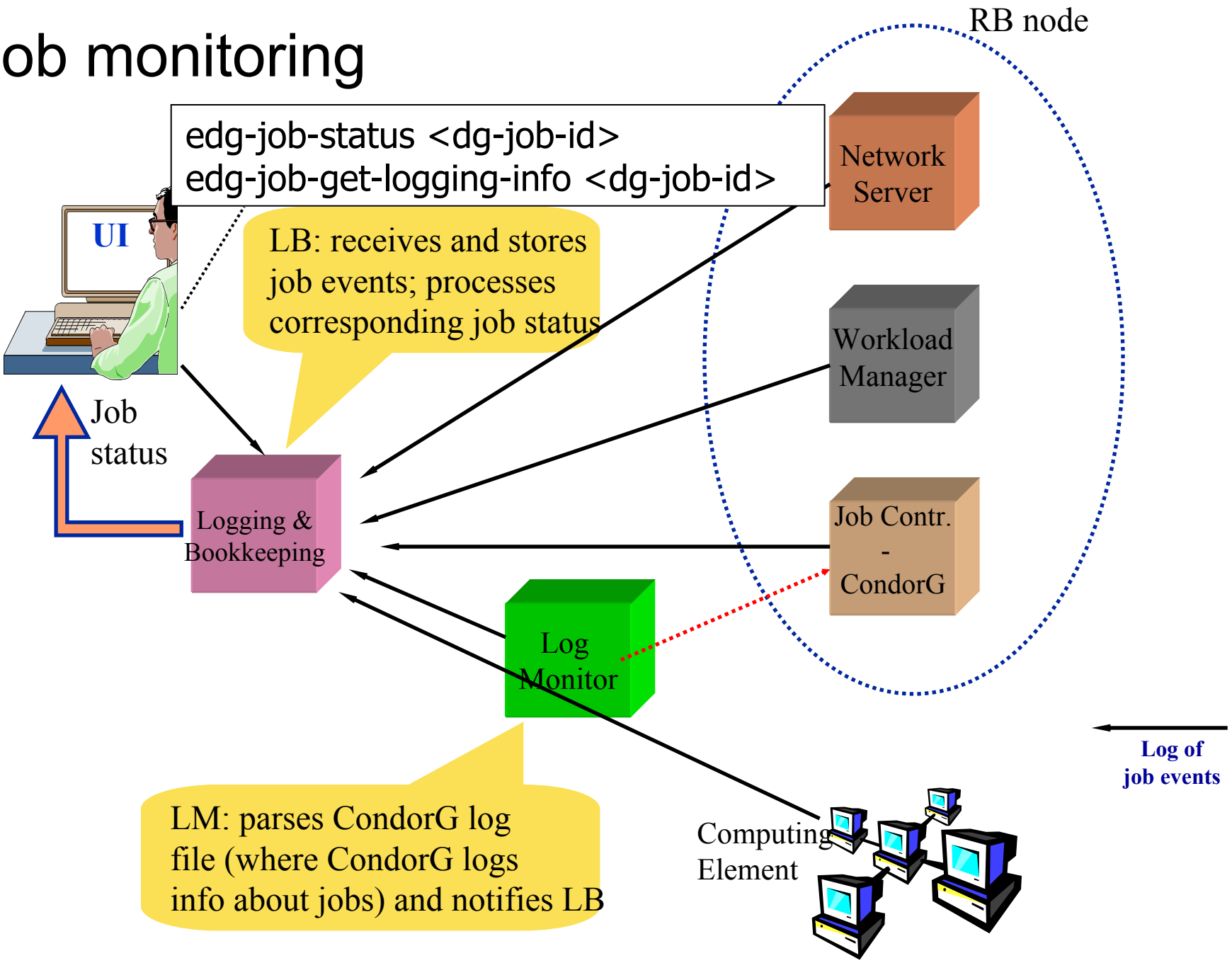
cleared



Computing Element

Storage Element

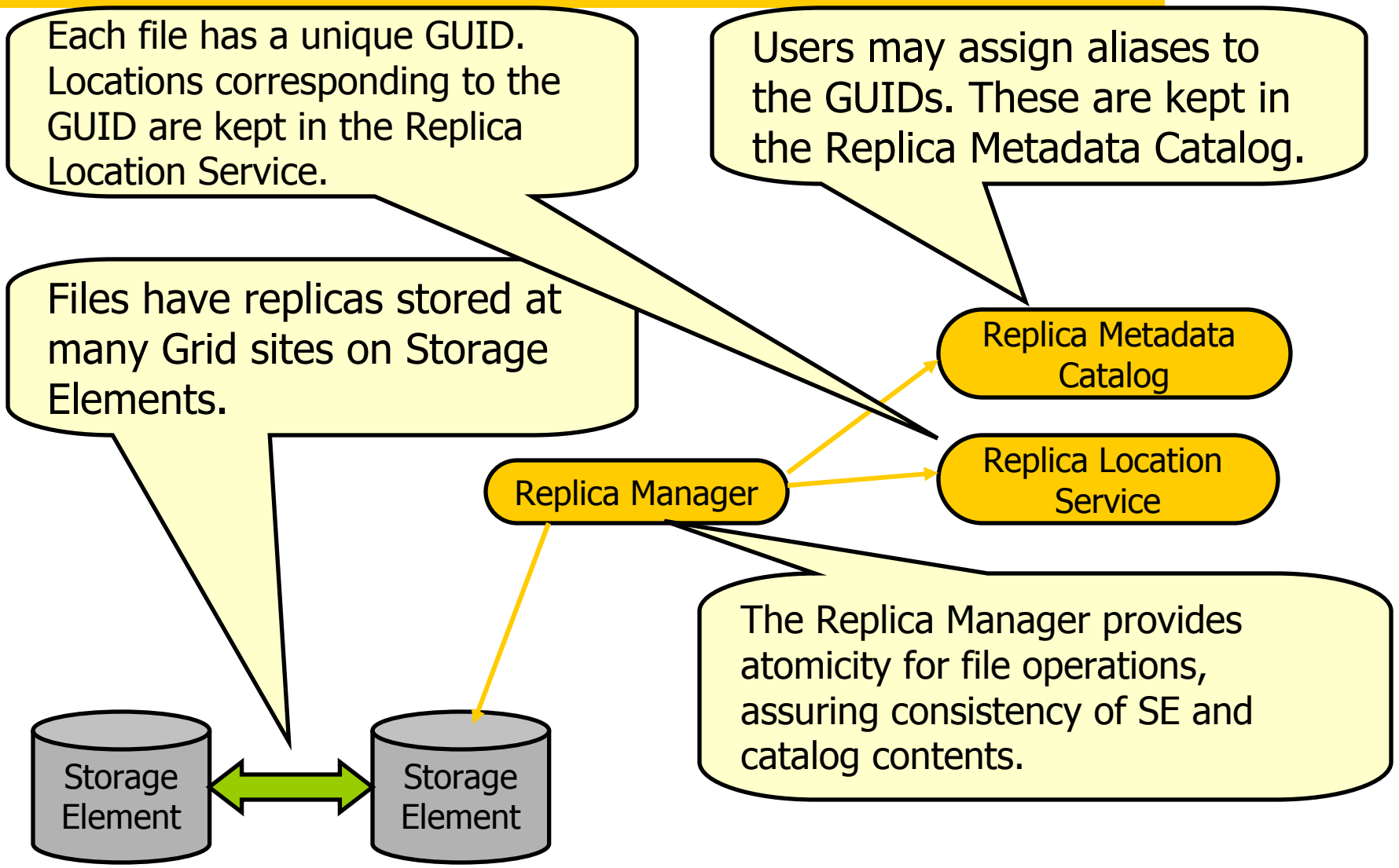
Job monitoring



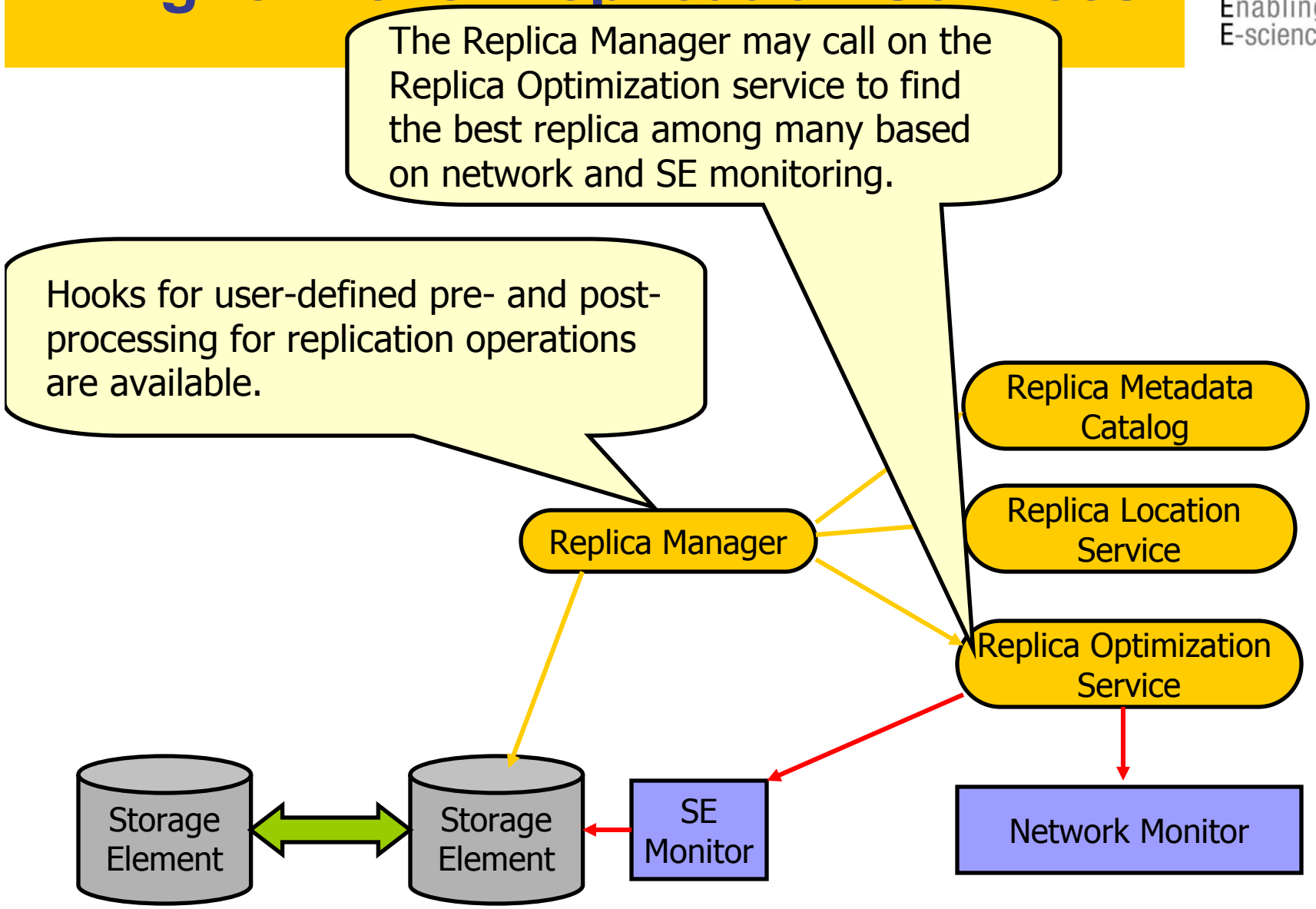
Other UI commands

- **`edg-job-list-match`**
  - Lists resources matching a job description
  - Performs the matchmaking without submitting the job
- **`edg-job-cancel`**
  - Cancels a given job
- **`edg-job-status`**
  - Displays the status of the job
- **`edg-job-get-output`**
  - Returns the job-output (the OutputSandbox files) to the user
- **`edg-job-get-logging-info`**
  - Displays logging information about submitted jobs (all the events “pushed” by the various components of the WMS)
  - Very useful for debug purposes

Replication Services: Basic Functionality

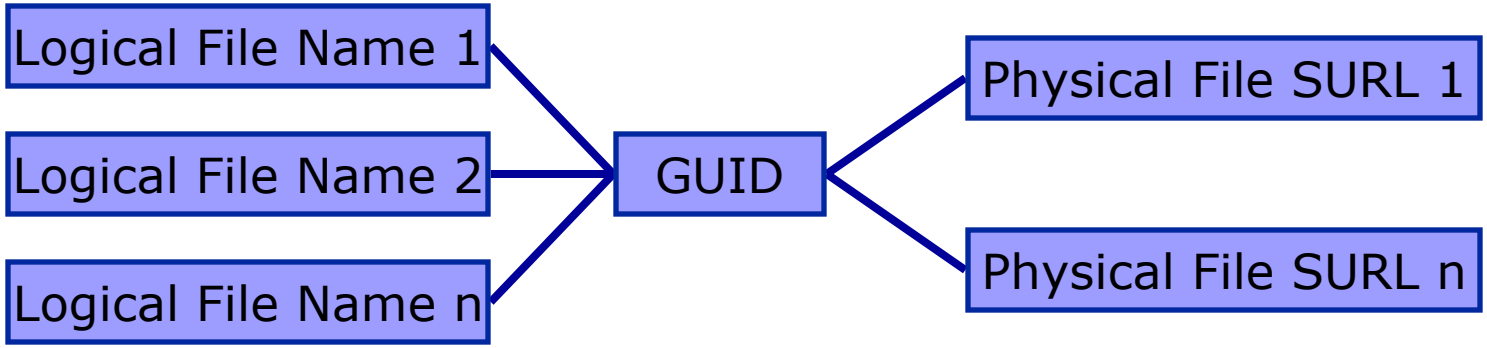


Higher Level Replication Services



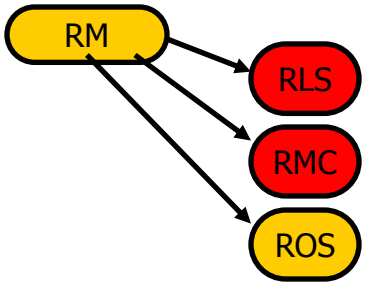
Naming Conventions

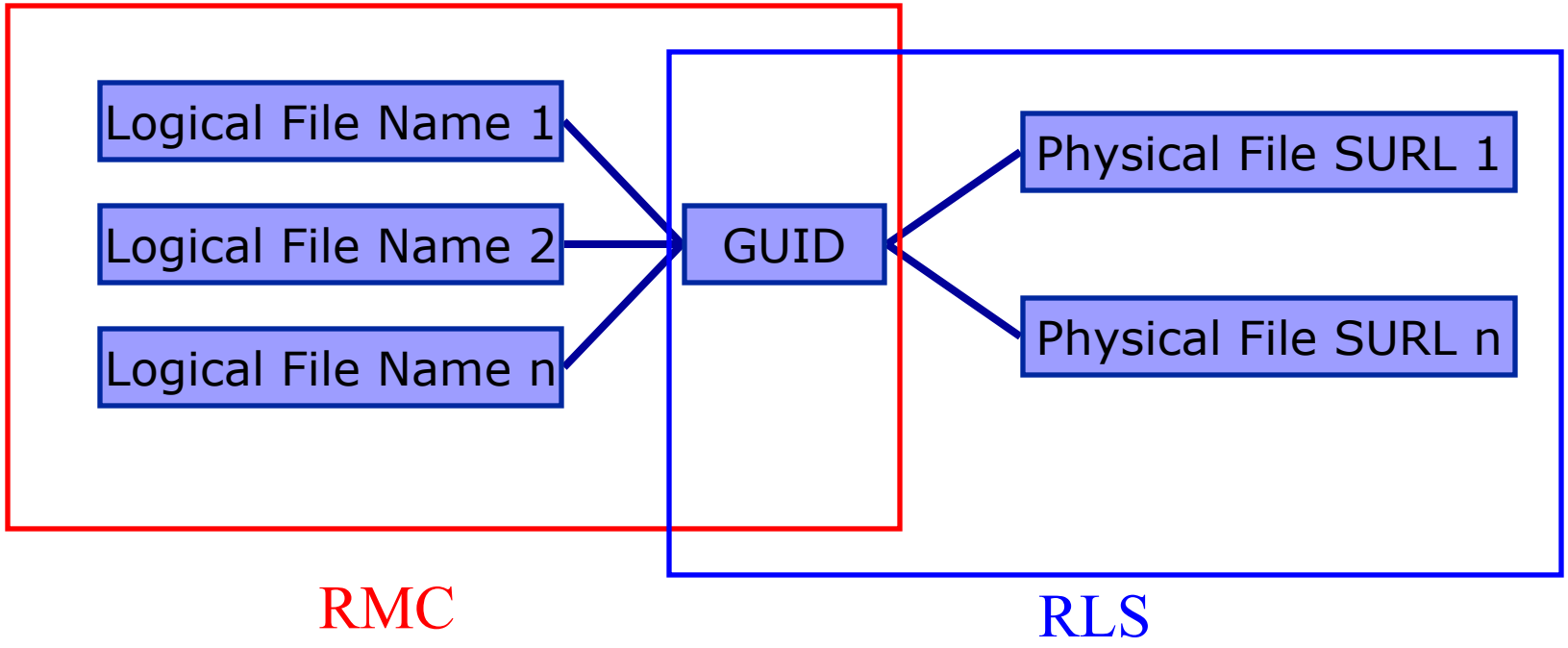
- Logical File Name (LFN)
  - An alias created by a user to refer to some item of data e.g. “lfn:cms/20030203/run2/track1”
- Site URL (SURL) (or Physical File Name (PFN))
  - The location of an actual piece of data on a storage system e.g. “srm://pcrd24.cern.ch/flatfiles/cms/output10_1”
- Globally Unique Identifier (GUID)
  - A non-human readable unique identifier for an item of data e.g. “guid:f81d4fae-7dec-11d0-a765-00a0c91e6bf6”



Replica Metadata Catalog (RMC) vs. Replica Location Service (RLS)

- RMC:
  - Stores LFN-GUID mappings
- RLS:
  - Stores GUID-SURL mappings

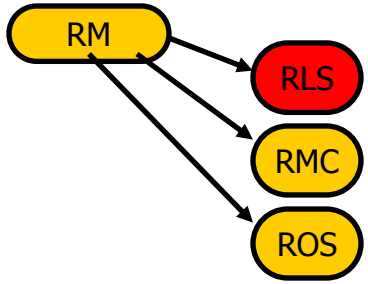




Replica Location Service (RLS)

Enabling Grids for E-science in Europe

- The Replica Location Service is a system that maintains and provides access to information about the physical location of copies of data files.
- It is a distributed service that stores mappings between globally unique identifiers of datafiles and the physical identifiers of all existing replicas of these datafiles.
- Design was a collaboration between Globus and EDG



Job submission

- **edg-job-submit [–r *<res_id>]* [-c *<config file>]* [-vo *<VO>]* [-o *<output file>]* *<job.jdl>***
  - -r the job is submitted directly to the computing element identified by *<res_id>*
  - -c the configuration file *<config file>* is pointed by the UI instead of the standard configuration file
  - -vo the Virtual Organization (if user is not happy with the one specified in the UI configuration file)
  - -o the generated edg_jobId is written in the *<output file>*
    - Useful for other commands, e.g.:
    - **edg-job-status –i** *<input file>* (or edg_jobId)

      -i the status information about edg_jobId contained in the *<input file>* are displayed

Job Definition Attributes

- **Executable** (mandatory)
  - The command name
- **Arguments** (optional)
  - Job command line arguments
- **StdInput, StdOutput, StdErr** (optional)
  - Standard input/output/error of the job
- **Environment** (optional)
  - List of environment settings
- **InputSandbox** (optional)
  - List of files on the UI local disk needed by the job for running
  - The listed files are staged from the UI to the remote CE
- **OutputSandbox** (optional)
  - List of files, generated by the job, which have to be retrieved

Resource Attributes

- **Requirements**
  - Job requirements on computing resources
  - Specified using attributes of resources published in the Information System
  - If not specified, default value defined in UI configuration file is considered
    - Default: other.GlueCEStateStatus == "Production" (the resource has to be in the Production grid)
- **Rank**
  - Expresses preference (how to rank resources that have already met the Requirements expression)
  - Specified using attributes of resources published in the Information Service
  - If not specified, default value defined in the UI configuration file is considered
    - Default: - other.GlueCEStateFreeCPUs (the highest number of free CPUs)

“Data” Attributes

- **InputData** (optional)
  - Refers to data used as input by the job: these data are published in the Replica Catalog and stored in the SEs)
  - PFNs and/or LFNs
- **DataAccessProtocol** (mandatory if InputData specified)
  - The protocol or the list of protocols which the application is able to speak with for accessing ***InputData*** on a given SE
- **OutputSE** (optional)
  - The hostname of the output SE
  - RB uses it to choose a CE that is compatible with the job and is close to SE
- **OutputData** (optional)
  - Output Data that will be registered at the end of the job

Example JDL File

```
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";
```

Summary… 1

Enabling Grids for E-science in Europe

- EGEE is creating a production-quality Grid as a step towards an emerging Europe-wide e-Infrastructure
  - Secure, reliable, sustainable
  - Wide spectrum of VOs
  - Integrating with national, regional, international grids and networks
- EGEE is reengineering middleware, with Service Orientation
- The LCG is providing a service now

- EGEE-0 components, Job submission and life-cycle have been described….

Summary -2: EGEE components