



Enabling Grids for E-science

EGEE Middleware





*Claudio Grandi
(INFN – Bologna)*

*EGEE-NAREGI meeting
March 20-22, 2006
CERN, Switzerland*

www.eu-egee.org
www.glite.org



-  **GLite Processes and Releases**
Lightweight Middleware for Grid Computing
-  **GLite Subsystems**
Lightweight Middleware for Grid Computing
 - Information System and Monitoring
 - Security Infrastructure
 - Workload Management
 - Data Management
- **Summary**

The logo for eGEE, with 'e' in blue, 'G' in yellow, and 'EE' in blue.

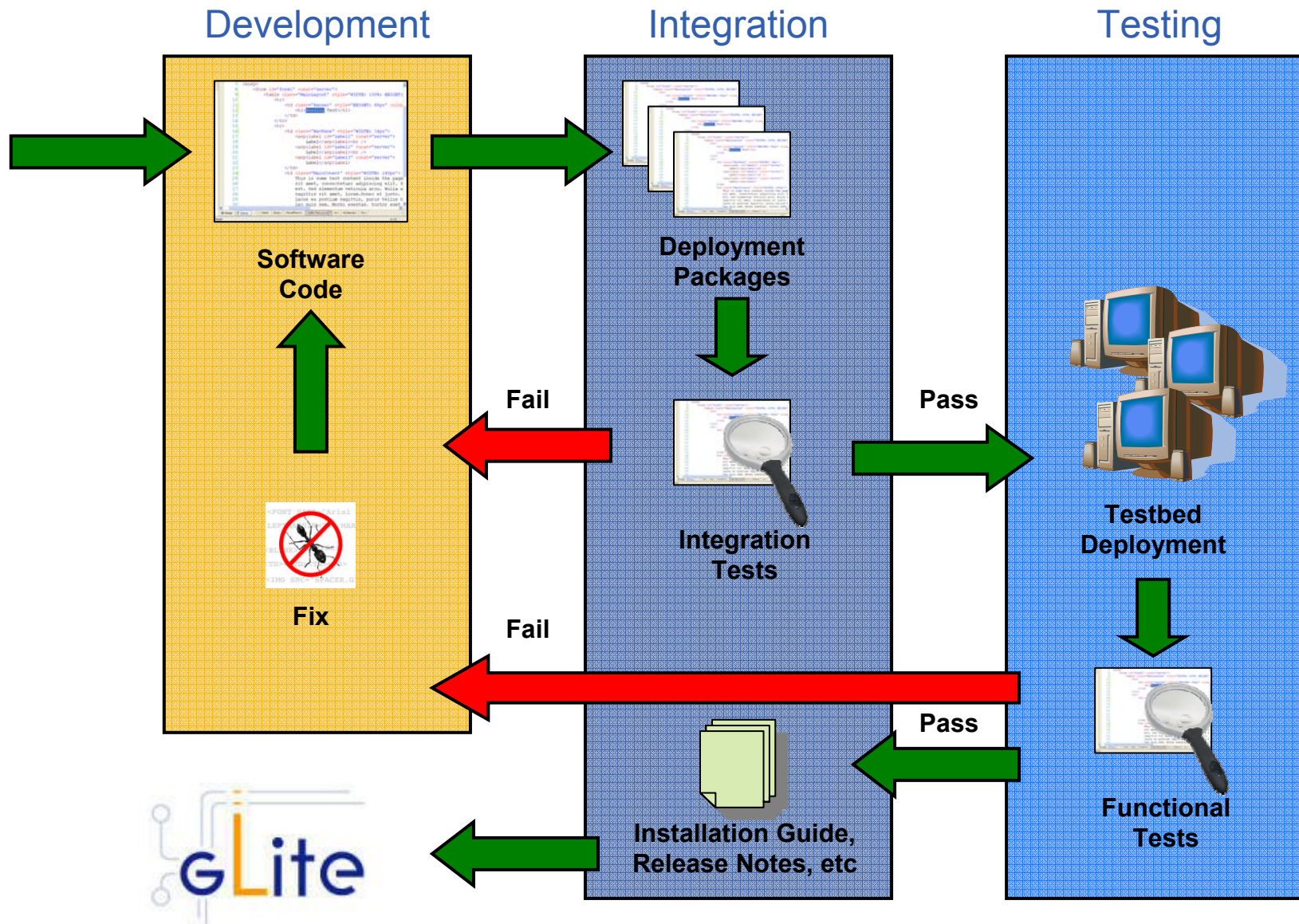
Enabling Grids for E-science

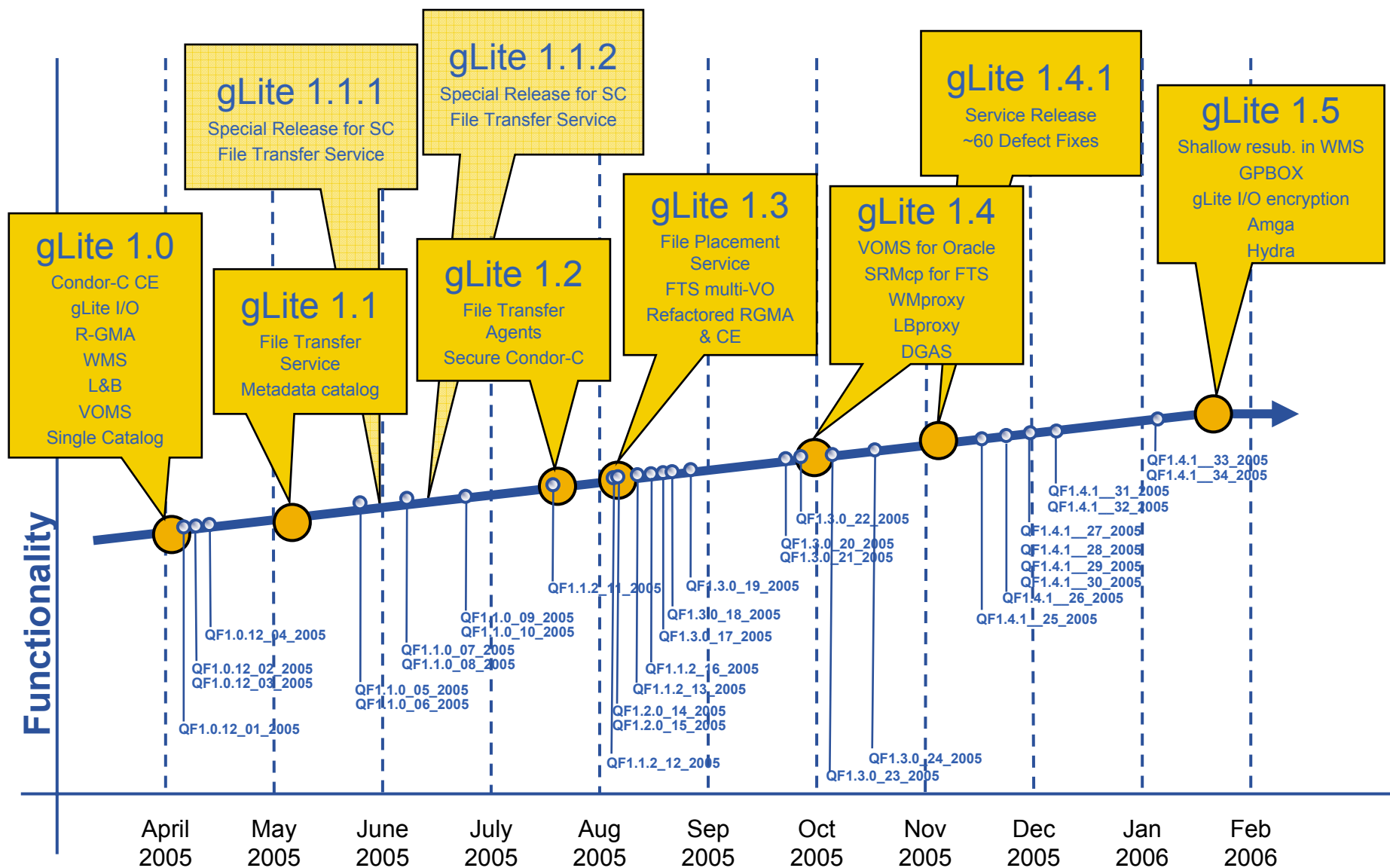


Process and Releases

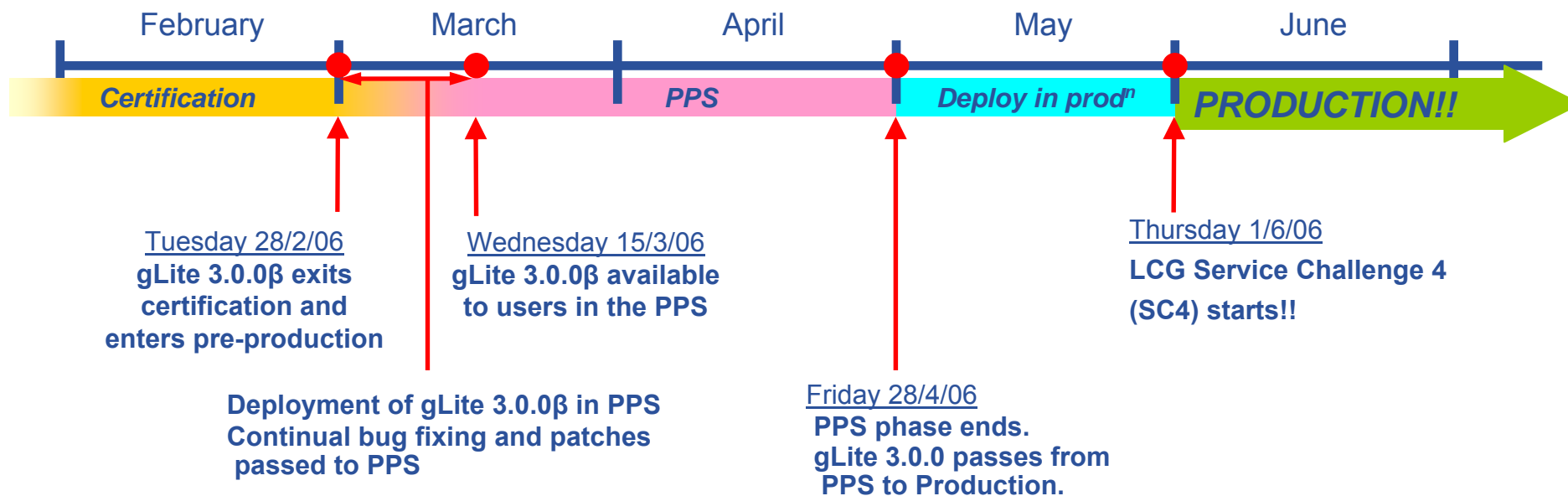
www.eu-egee.org
www.glite.org







- **Converge from LCG and gLite to a single middleware stack called gLite. The first version will be gLite 3.0.0**
 - Process controlled by the **Technical Coordination Group**
 - gLite 1.5.0 and LCG 2.7.0 have been the last independent releases
- **Components in gLite 3.0.0**
 - Certified:
 - All components already in LCG 2.7.0 plus upgrades
 - *this already includes new versions of VOMS, R-GMA and FTS*
 - The Workload Management System (with LB, CE, UI) of gLite 1.5.0
 - Tested to some degree and with limited deployment support:
 - The DGAS accounting system
 - Data management tools as needed by the Biomed community
 - *Hydra, AMGA, secure access to data*



After gLite 3.0.0:

- March 31st: code freeze for development release gLite 3.1.0
- April 30th: end of integration
- May 31st: end of certification. Deployment on PPS
- July 31st: release of production version gLite 3.2.0. Start deployment at sites
- September : gLite 3.2.0 installed at sites and usable.

The logo for eGEE, with 'e' in blue, 'G' in yellow, and 'EE' in blue.

Enabling Grids for E-science



Security Infrastructure

www.eu-egee.org
www.glite.org





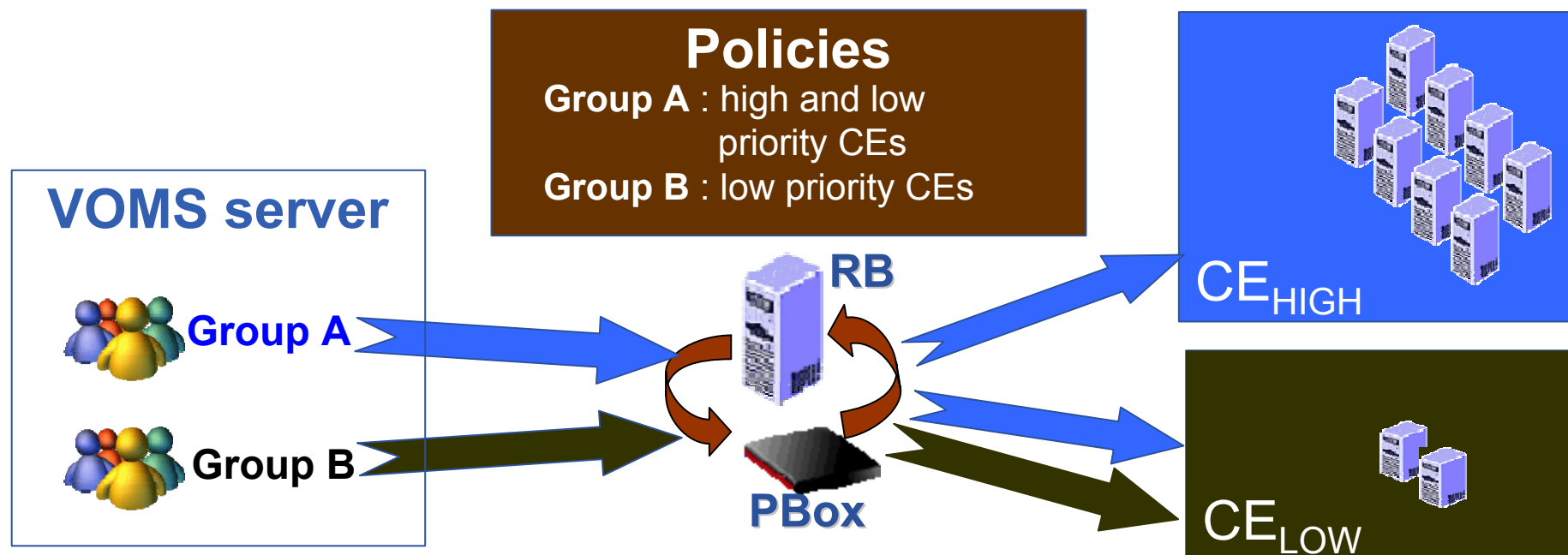
- **Based on trusted third parties (TTPs)**
 - Regional Certificate Authorities (CA)
- **X.509 PKI infrastructure**
 - Grid Security Infrastructure (from Globus 2.4.3, as in VDT-1.2.2)
- **Federations of CAs exists**
 - Europe EUGridPMA, Asia-Pacific Region APGridPMA, Americas TAGPMA
 - International Grid Trust Federation (IGTF) established in October 2005
- **Short-Lived Credential Services (SLCS) → proxies**
- **Site-integrated credential services (SICS)**
 - issue short-lived credentials to its local users (e.g. Kerberos at FNAL)
- **Proxy store: MyProxy (Version 1.14) from VDT 1.2.2**
 - Allows clean proxy renewal
 - working with VDT to provide VOMS-aware access to MyProxy

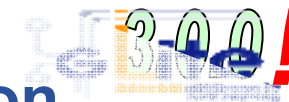


- **VO-specific attributes: VOMS**
 - VOMS issues Attribute Certificates that are attached to proxies and provide users with additional capabilities
 - are the base for Authorization process
- **Web services for delegation**
 - using portType (Java, for axis) and GridSite (C for Apache)
 - need to define a standard for interoperation
- **Authorization framework**
 - gLite Authorization Framework (compatible with XACML policies)
 - mainly used for Java-based applications
 - LCAS/LCMAPS
 - mainly used for C-based applications (e.g. GT2-GRAM, GridFTP server)
 - support for VOMS, blacklists, gridmap files

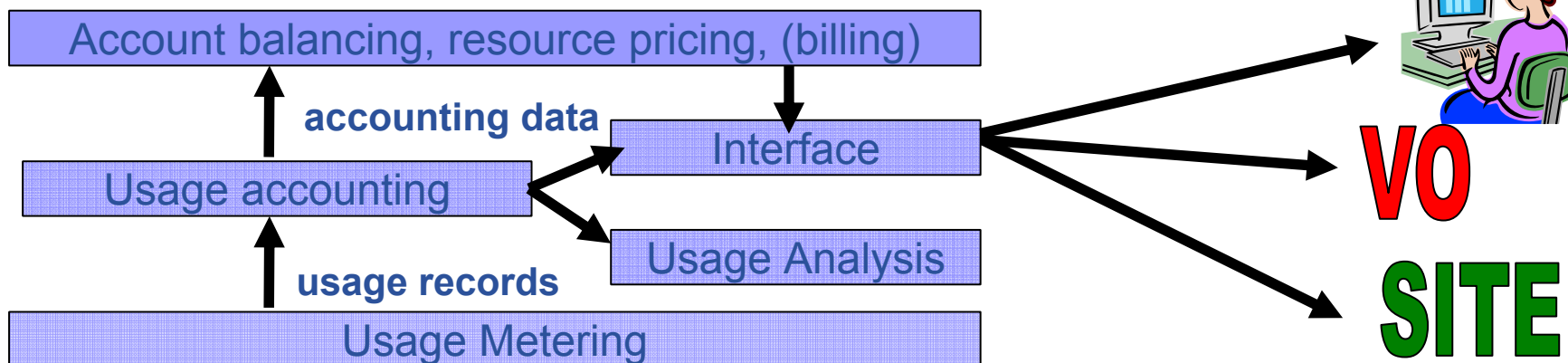


- **GPBOX: Distributed VO policy management tool**
 - Interface to define fine-grained VO policies
 - based on VOMS groups and roles
 - Store policies and propagate to sites
 - enforcement of policies done at sites – sites may accept/reject policies
 - May be interfaced to dynamic sources of information
 - e.g. an accounting system to provide fair share
 - Standards Compliant (RBAC, XACML, GSI)





- **DGAS: accumulates Grid accounting information**
 - *User, JobId, user VO, VOMS FQAN(role,capabilities), SI2K, SF2K, system usage (cpuTime, wallTime...),...*
 - allows billing and scheduling policies
 - levels of granularity: from single jobs to VO or grid aggregations
 - Privacy: only the user or VO manager can access information
 - site managers can keep accounting information available just for site internal analysis
 - Sites can substitute DGAS metering system with their own
- **Limited support in gLite 3.0.0**



The logo for eGEE, with 'e' in blue, 'G' in yellow, and 'EE' in blue.

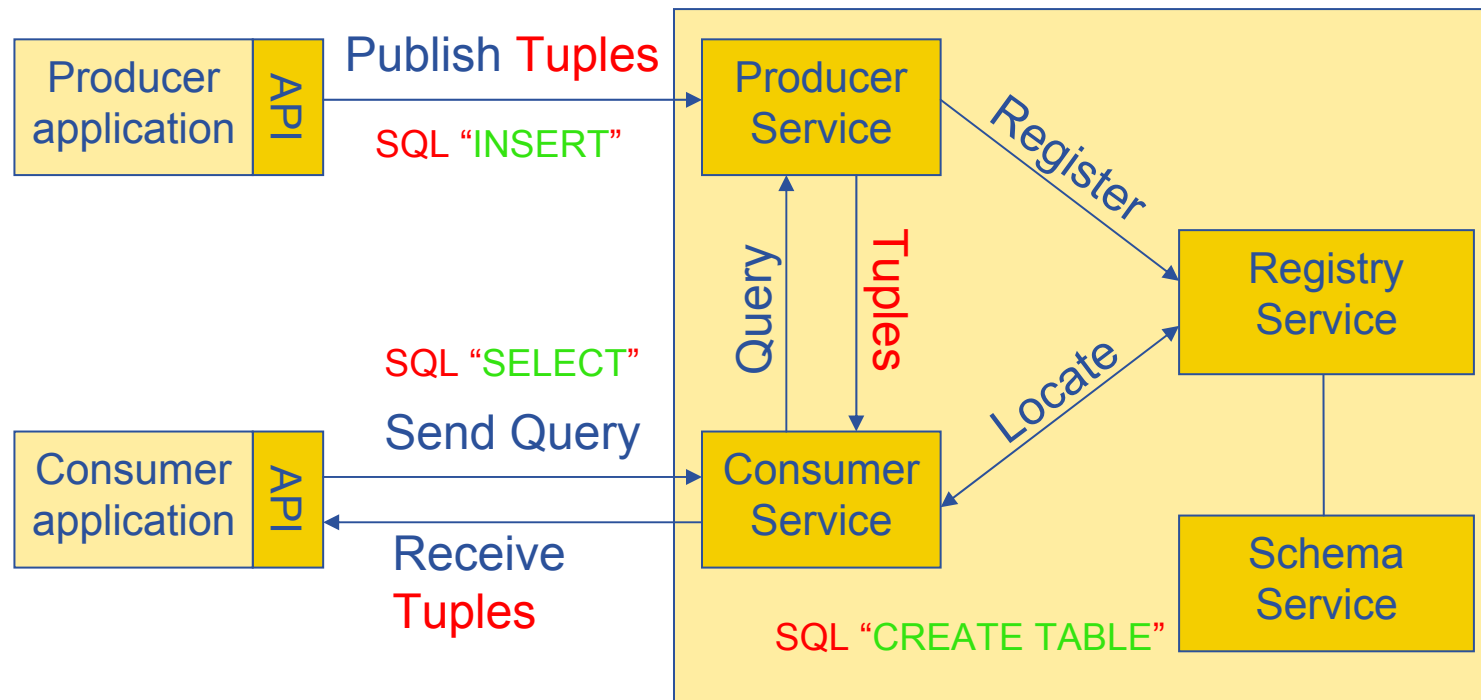
Enabling Grids for E-science

The GLite logo, featuring a stylized 'G' with circuit-like lines and the text 'Lite' in blue. Below it, the text 'Lightweight Middleware for Grid Computing' is written in a smaller font.

GLite Information System and Monitoring

www.eu-egee.org
www.glite.org





- The Relational Grid Monitoring Architecture (R-GMA) provides a uniform method to access and publish both information and monitoring data.
- From a user's perspective, an R-GMA installation currently appears similar to a single relational database.
- Relational implementation of the GGF's Grid Monitoring Architecture (GMA)



- **the gLite Service Discovery provides a standard set of methods for locating Grid services**
- **hides underlying information system**
- **plug-ins for R-GMA, BDII and XML files (others could be developed if required)**
- **API available for Java and C/C++**
- **command line version also available**
- **Used by WMS and Data Management clients**

- **Production Services still using BDII as the Information System**

The logo for eGEE, with 'e' in blue, 'G' in yellow, and 'EE' in blue.

Enabling Grids for E-science



Workload Management

www.eu-egee.org
www.glite.org

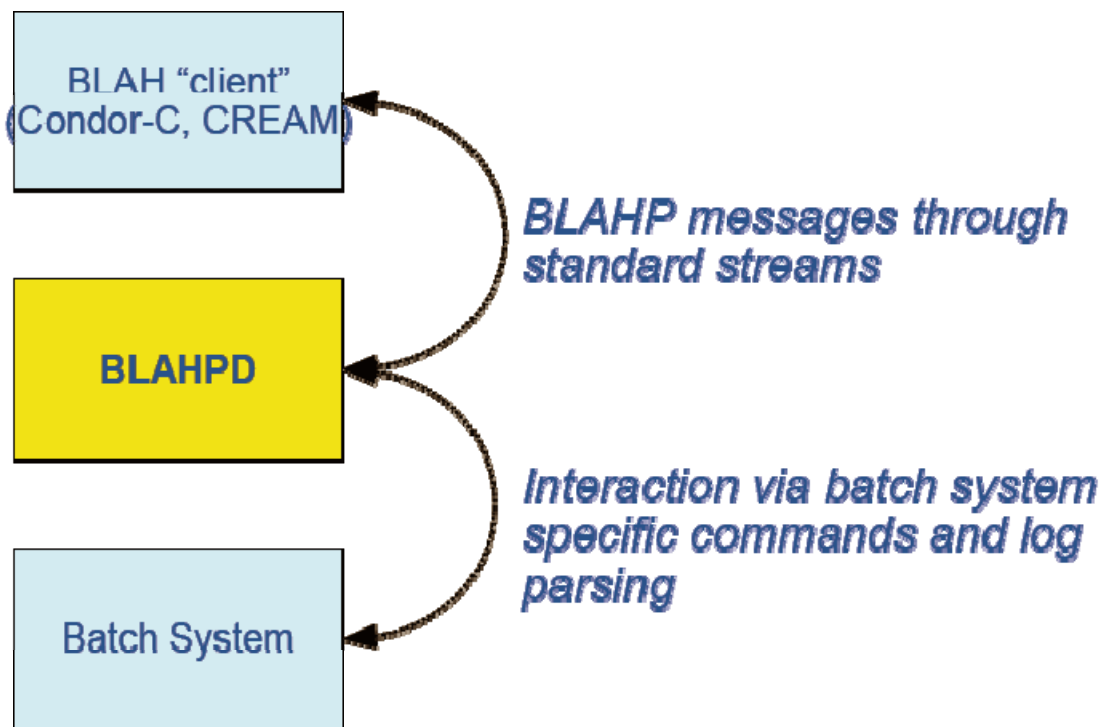


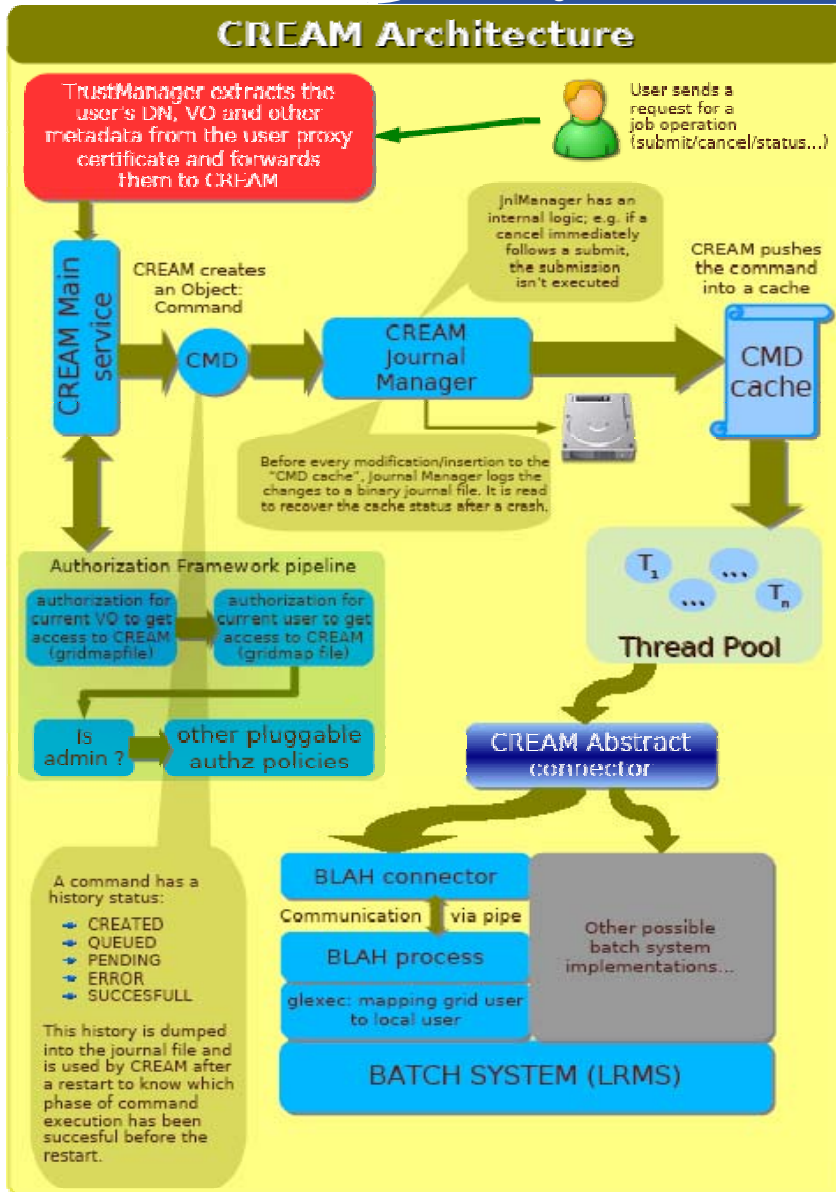


- **Service representing a computing resource**
- **Condor-C GSI enabled**
 - Based on Condor 6.7.10, will migrate to 6.7.18 in next release
- **Uses CEMon to publish information**
 - support for R-GMA and bdll
 - GLUE 1.1 (migration to GLUE 1.2 for next release)

- **Batch Local ASCII Helper (BLAH)**

- Interface between the CE and the Local Resource ManagerSystem
- submit, cancel and query
- Support for hold and resume
 - To be used to put a job on hold, waiting for e.g. the staging of the input data

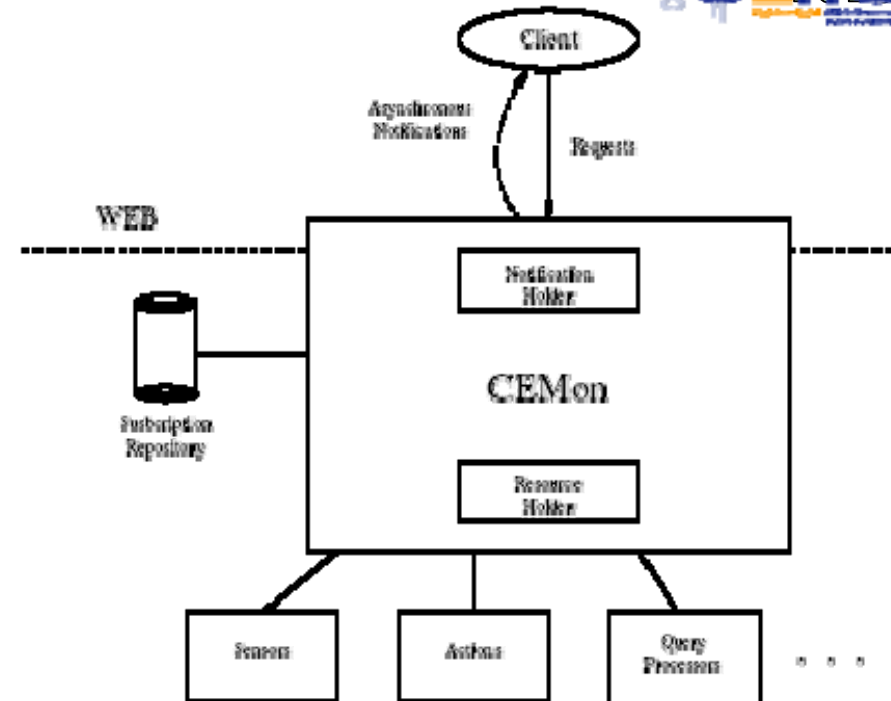




- **CREAM**
 - web service Computing Element
 - Cream WSDL allows defining custom user interface
 - C++ CLI interface allows direct submission
- **Lightweight**
- **Fast notification of job status changes**
 - via CEMon
- **Improved security**
 - no “fork-scheduler”
- **Will support for bulk jobs on the CE**
 - optimization of staging of input sandboxes for jobs with shared files
- **ICE (Interface to Cream Environment)**
 - being integrated in WMS for submissions to CREAM



- **Web service to publish status of a computing resource to clients**
 - WMS or individual clients
- **Synchronous queries or asynchronous notification of events**
- **Clients subscribe and are notified according to user defined policies**
 - on job status
 - on CE characteristics and status
 - May be used to pull jobs
- **Included in VDT and used in OSG for resource selection**
- **In gLite 3.0.0 will be available but the baseline is that the WMS queries the bdll**





- **Logging and Bookkeeping service**
 - Tracks jobs during their lifetime (in terms of events)
 - L&B Proxy provides faster, synchronous and more efficient access to L&B services to Workload Management Services
 - Support for “CE reputability ranking“
 - Maintains recent statistics of job failures at CE’s
 - Feeds back to WMS to aid planning
 - Working on inclusion of L&B in the VDT



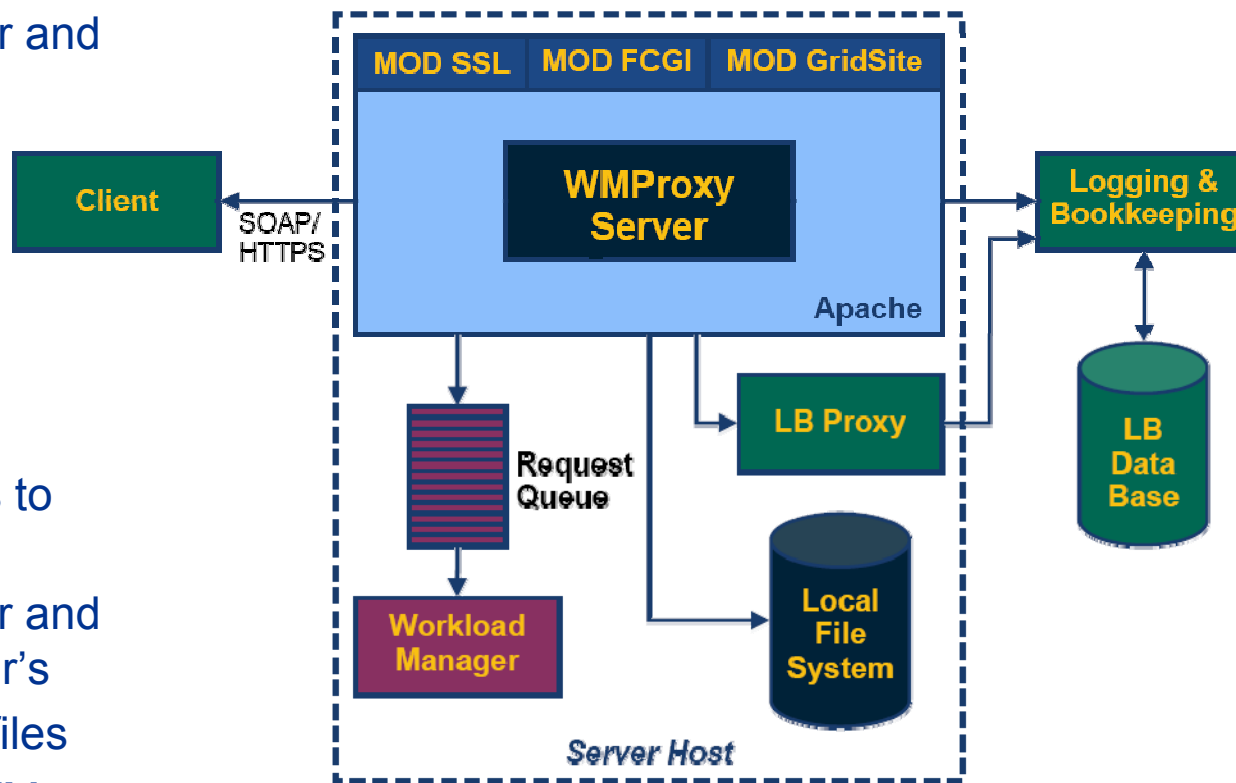
- **Job Provenance**
 - Long term job information storage
 - Useful for debugging, post-mortem analysis, comparison of job executions in different environments
 - Useful for statistical analysis



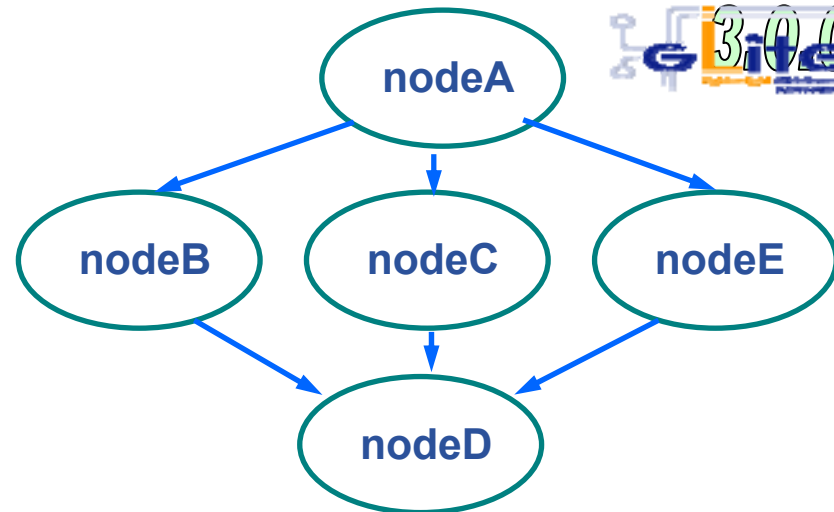
- **Helps the user accessing computing resources**
 - resource brokering
 - management of input and output
 - management of complex workflows
- **Job specification in JDL (based on Condor classad 0.9.6)**
 - Building a JDL ↔ JSDL translator
- **Backward compatible with LCG-2**
- **Support for *shallow resubmission***
 - Resubmission happens in case of failure only when the job didn't start
- **Support for MPI job even if the file system is not shared between CE and Worker Nodes (WN)**
- **Support collection of information from many sources**
 - CEMon, bdll, R-GMA
- **Support for Data management interfaces (DLI and StorageIndex)**
- **Support for execution of all DAG nodes within a single CE**
 - chosen by user or by the WMS matchmaker
- **Support for file peeking during job execution (Job File Perusal)**
- **Initial support for pilot job**
 - prepare the execution environment, then get and execute the user job



- WMPProxy is a SOAP Web service providing access to the Workload Management System (WMS)
- Job characteristics specified via JDL
 - jobRegister
 - create id
 - map to local user and create job dir
 - register to L&B
 - return id to user
 - input files transfer
 - GridFTP 1.12
 - jobStart
 - register sub-jobs to L&B
 - map to local user and create sub-job dir's
 - unpack sub-job files
 - deliver jobs to WM

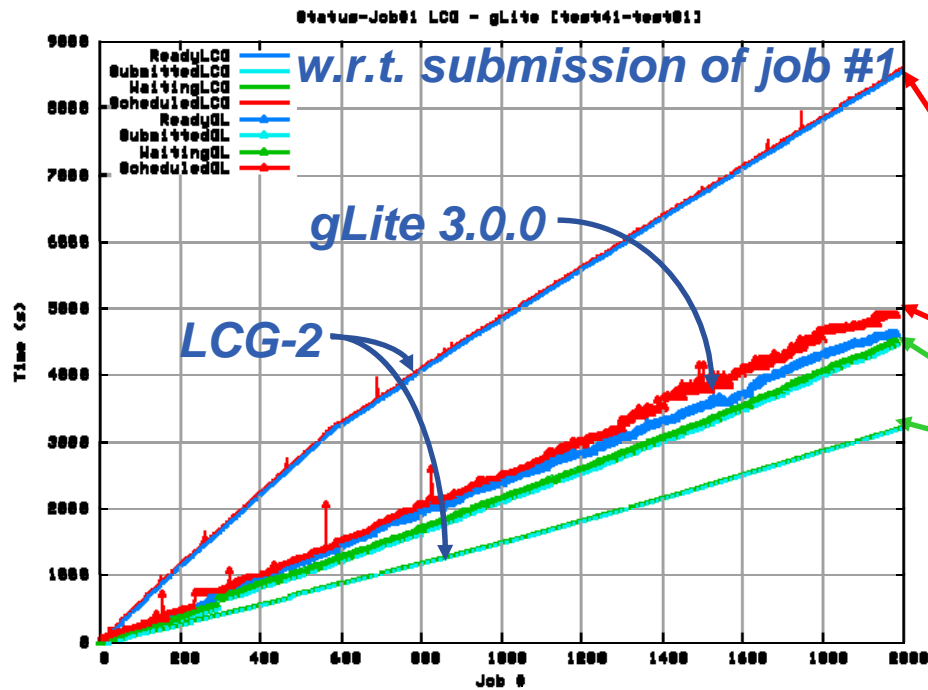


- **Direct Acyclic Graph (DAG)** is a set of jobs where the input, output, or execution of one or more jobs depends on one or more other jobs
- **A Collection** is a group of jobs with no dependencies
 - basically a collection of JDL's
- **A Parametric job** is a job having one or more attributes in the JDL that vary their values according to parameters
- **Using compound jobs** it is possible to have one shot submission of a (possibly very large, up to thousands) group of jobs
 - Submission time reduction
 - Single call to WMPProxy server
 - Single Authentication and Authorization process
 - Sharing of files between jobs
 - Availability of both a single Job Id to manage the group as a whole and an Id for each single job in the group





- **Shared Input sandbox**
 - Useful when submitting compound jobs
 - When sub-jobs input sandboxes contain instances of the same file, these are transferred only once and made available by WMPProxy to all involved sub-jobs
 - lower data size to transfer
 - minimize number of calls to file transfer service
- **Asynchronous job start will be available in next release**
 - Upon call to the jobStart operation, WMPProxy can complete processing of the request in background
 - Control is returned to the client soon after the request has been accepted by WMPProxy and the corresponding event has been logged to LB
 - All time-consuming actions needed to complete the processing of the request are performed “behind the scene” by the service
 - From user point of view makes submission time (almost) independent from the number of jobs



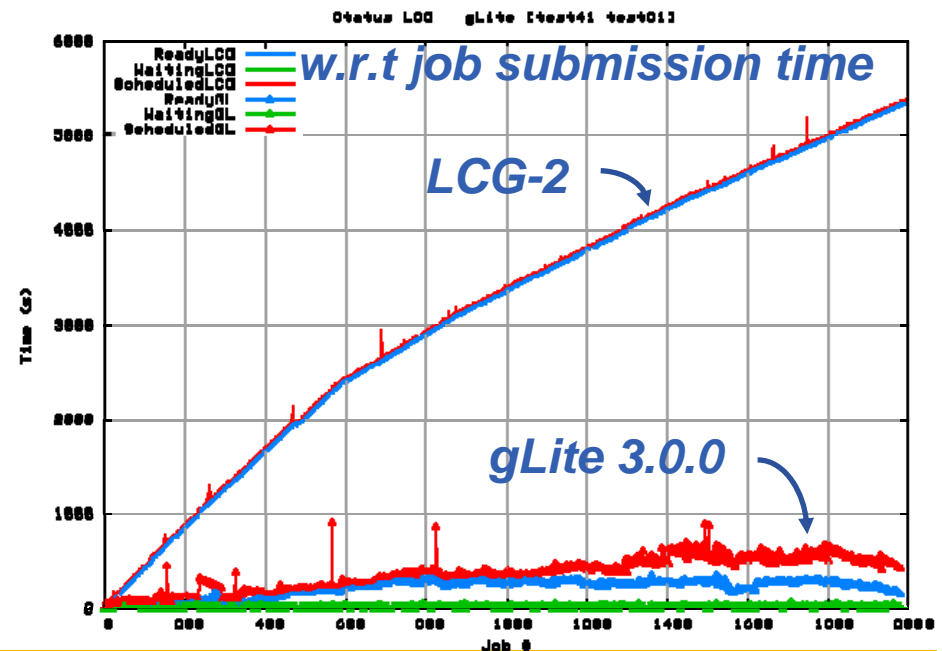
- Submission test with Network Server

- 10 threads of 200 jobs
- multithread active on WMS

All jobs scheduled

Prompt back

- Submission slower with gLite NS but after that all jobs are scheduled much faster
- But using the WMPProxy prompt back after a few seconds (almost independent of number of job!)



The logo for eGEE, with 'e' in blue, 'G' in yellow, and 'EE' in blue.

Enabling Grids for E-science



Data Management

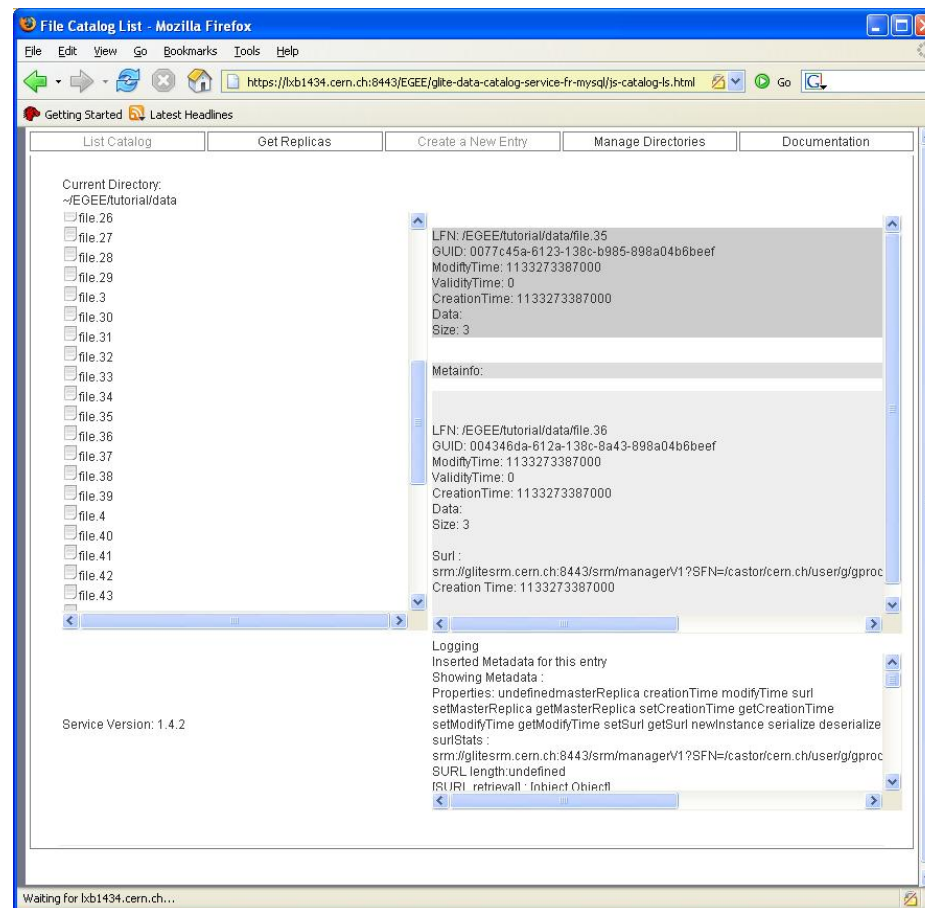
www.eu-egee.org
www.glite.org



- **A disk or tape storage resource**
- **Common interface: SRMv1**
 - negotiable transfer protocols (Gridftp, gsidcap, RFIO, ...)
 - work in progress to migrate to SRMv2
 - Various implementation from LCG and other external projects
 - disk-based: DPM, dCache
 - tape-based: Castor, dCache
 - Support for ACLs in DPM, in future in Castor and dCache
- **Posix-like file access:**
 - Grid File Access Layer (**GFAL**) by LCG. Provides also:
 - Abstractions: Storage Element, File Catalog, Information System
 - Support for ACL in the SRM layer
 - **gLite I/O**
 - Support for file ACL via the Fireman catalog
 - Interfaced to SRM Storage Elements (Castor, dCache and DPM)
 - Configuration using the common Service Discovery interfaces
 - Limited support in gLite 3.0.0 (replaced by GFAL)

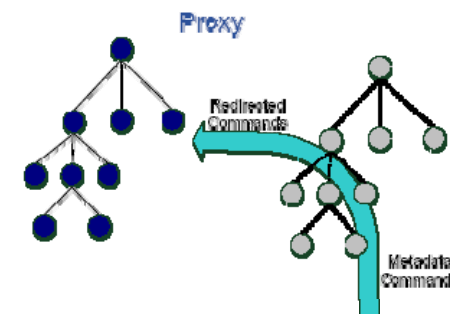
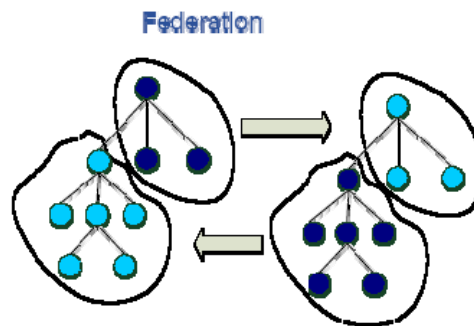
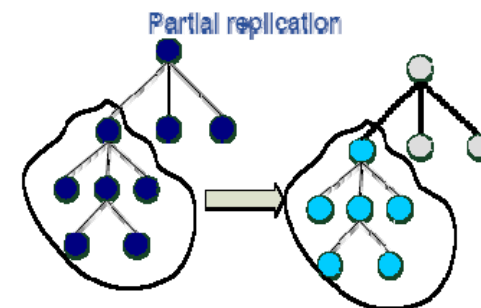
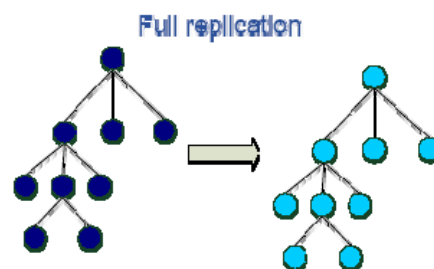


- Resolves logical file names (LFN) to physical location of files (URL understood by SRM) and storage elements
- Oracle and MySQL versions available
- Secure services, using VOMS groups, ACL support
- CLI and simple API for C/C++ wrapping a lot of the complexity for easy usage
- Attribute support
- Symbolic link support
- Exposing interfaces suitable for matchmaking (StorageIndex and DLI)
- Limited support in gLite 3.0.0 (replaced by LFC)



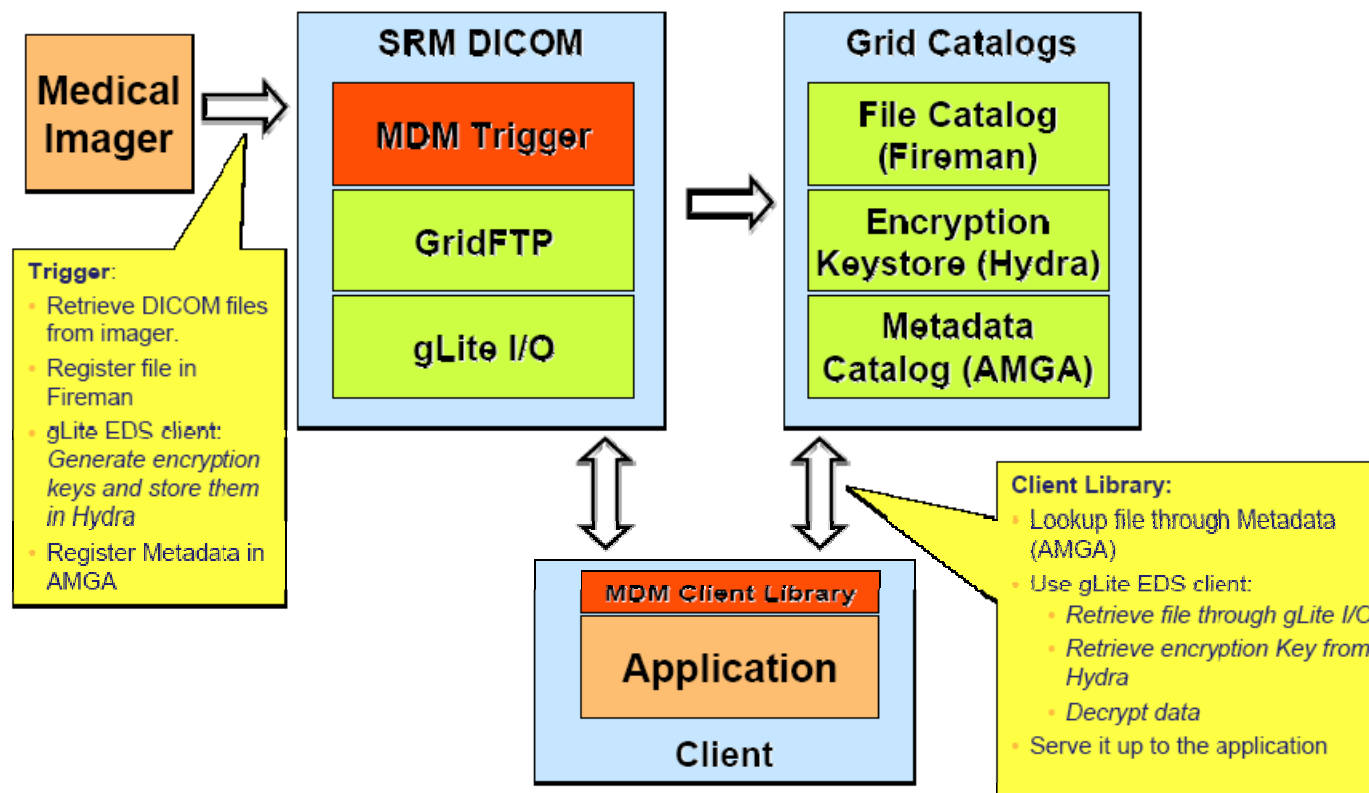


- Joint JRA1-NA4 (ARDA) development
- Authentication based on Password, X509 Cert, Grid Proxy
- Posix-ACLs and Unix permissions for entries and collections
 - not yet on individual attributes
- Built-in group-management like AFS or via VOMS
- SOAP and Text front-ends
- Streamed Bulk Operations
 - vital for WAN operations
- Scales well up to 100 concurrent clients
 - back-end limit
- Support for replication
 - also via LCG-3D for Oracle backend only
- Limited support in gLite 3.0.0





- **Hydra keystore**
 - store keys for data encryption
 - 3 instances: at least 2 need to be available for decryption
 - Limited support in gLite 3.0.0
- **Demonstrated with the SRM-DICOM demo at EGEE Pisa conference (Oct'05)**



- **Reliable, scalable and customizable file transfer**
reliable:

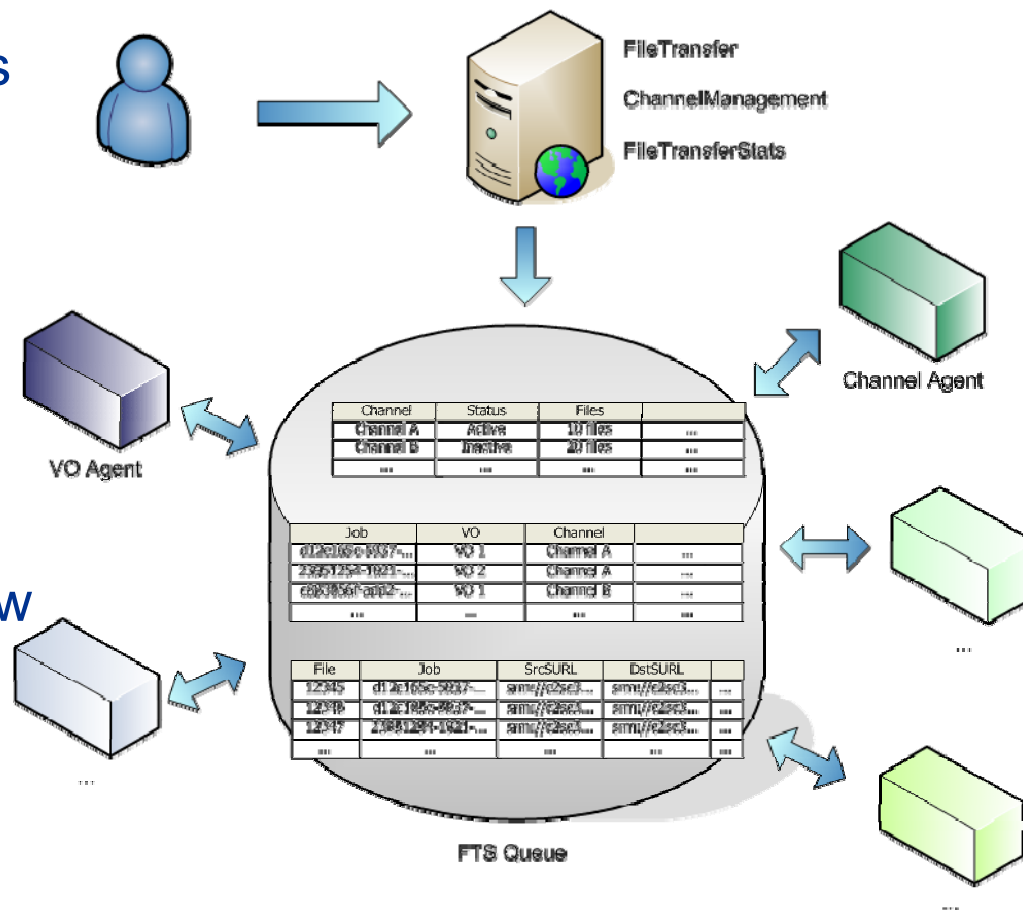
- transfer requests are jobs
 - in case of failure a transfer may be retried
 - subject to generic and VO-specific policies

scalable:

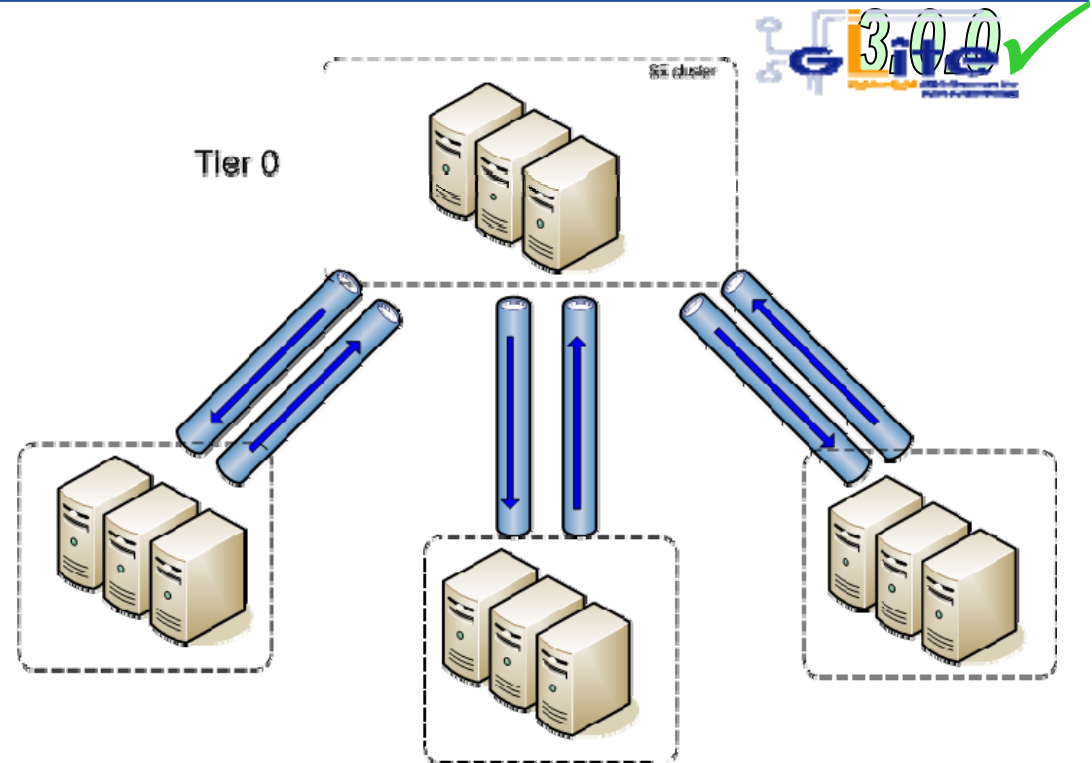
- more files in a single transfer request
- File Transfer queues allow inter-VO sharing

customizable:

- VO-specific File Transfer Agents



- **Logical unit of management**
 - Represent a mono-directional network pipe between two sites
- **Independently manageable**
 - State
 - Number of streams
 - Number of concurrent transfers
- **Production Channels**
 - Dedicated Tier0-Tier1, Tier1-Tier1 or Tier1-Tier2 transfers
 - efficient bulk distribution
- **Non-production Channels**
 - Open networks shared with other applications.





- **Client with SOAP API and CLI**

- Roles

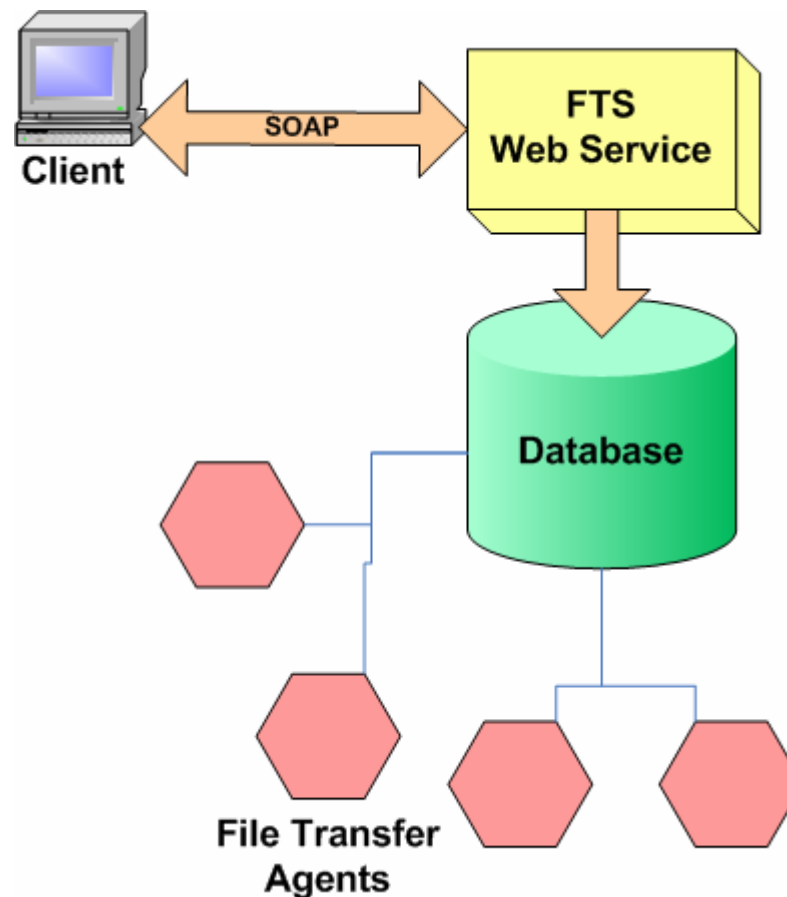
- Administrator of the service
 - Channel Manager
 - VO Manager
 - Submitter User
 - Regular User
 - Vetoed User

- **Web Service**

- **Database**

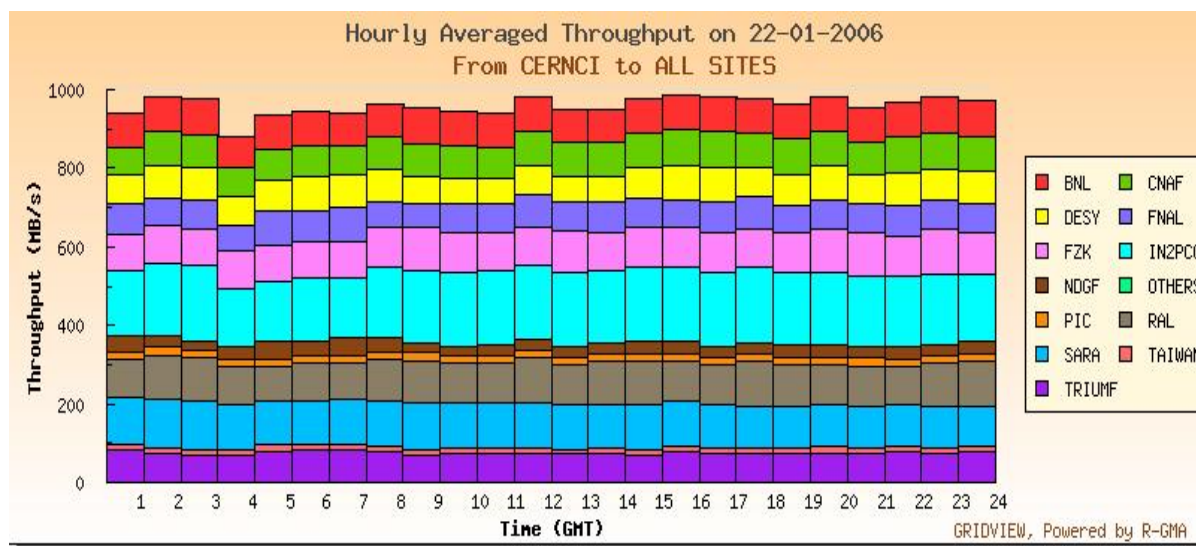
- Oracle and MySQL

- **Agents act on database conditions**





- **Service Challenge 3 Rerun (January 2006)**
 - All sites achieved target rate
 - 8/11 sites achieved nominal rate



- **Service Challenge 4 (start June 2006)**
 - increase stability, sustainability, robustness
 - support for SRMv2
 - integration with experiment frameworks

- **Complete middleware stack**
 - security infrastructure
 - information system and monitoring
 - workload management
 - data management
- **Developed according to a well defined process**
- **Controlled by the EGEE Technical Coordination Group**
- **gLite 3.0.0 will be available on the production infrastructure in less than 2 months**
- **Development is continuing to provide increased robustness, usability and functionality**



Lightweight Middleware for
Grid Computing

www.glite.org

