Enabling Grids for E-sciencE

EGEE middleware: gLite

Claudio Grandi - INFN

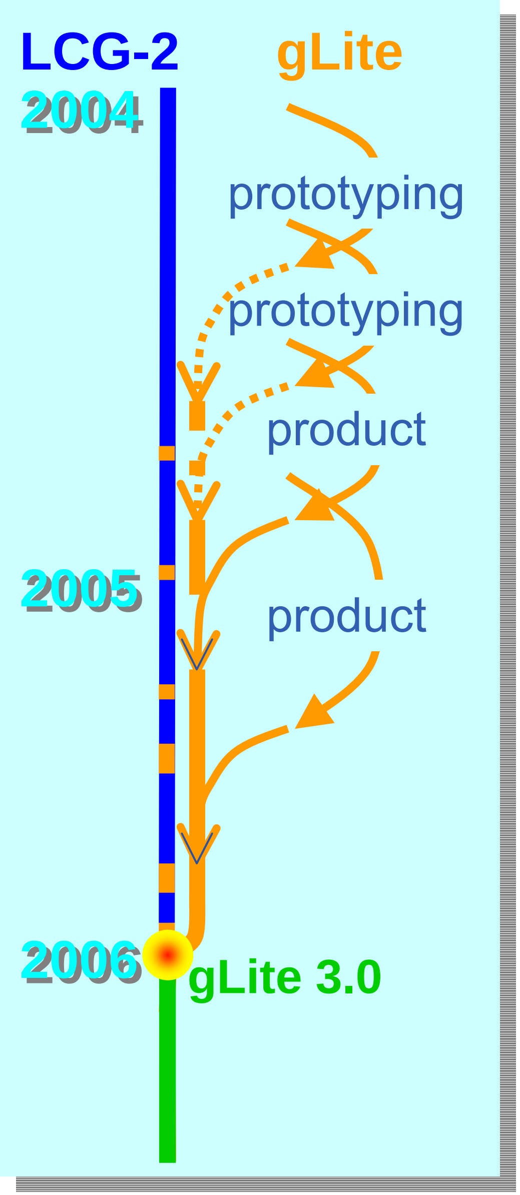
5th NRENs and Grids Workshop

Paris, 11-12 June 2007

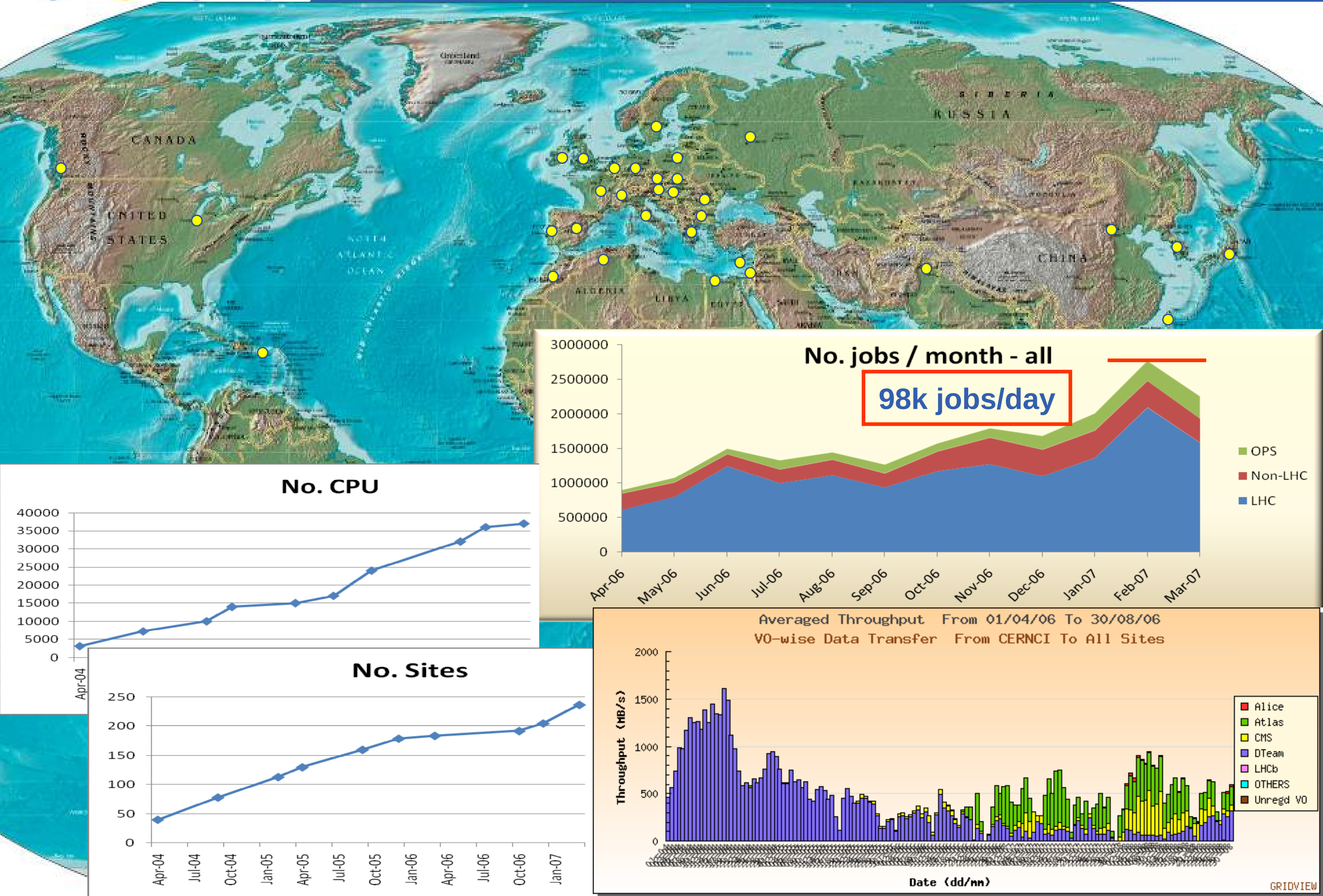
www.eu-egee.org
www.glite.org

gLite Middleware Distribution

- **Combines components from different providers**
  - Condor and Globus (via VDT)
  - LCG
  - EDG/EGEE
  - Others
- **After prototyping phases in 2004 and 2005 convergence with LCG-2 distribution reached in May 2006**
  - gLite 3.0
- **Focus on providing a deployable MW distribution for EGEE production service**

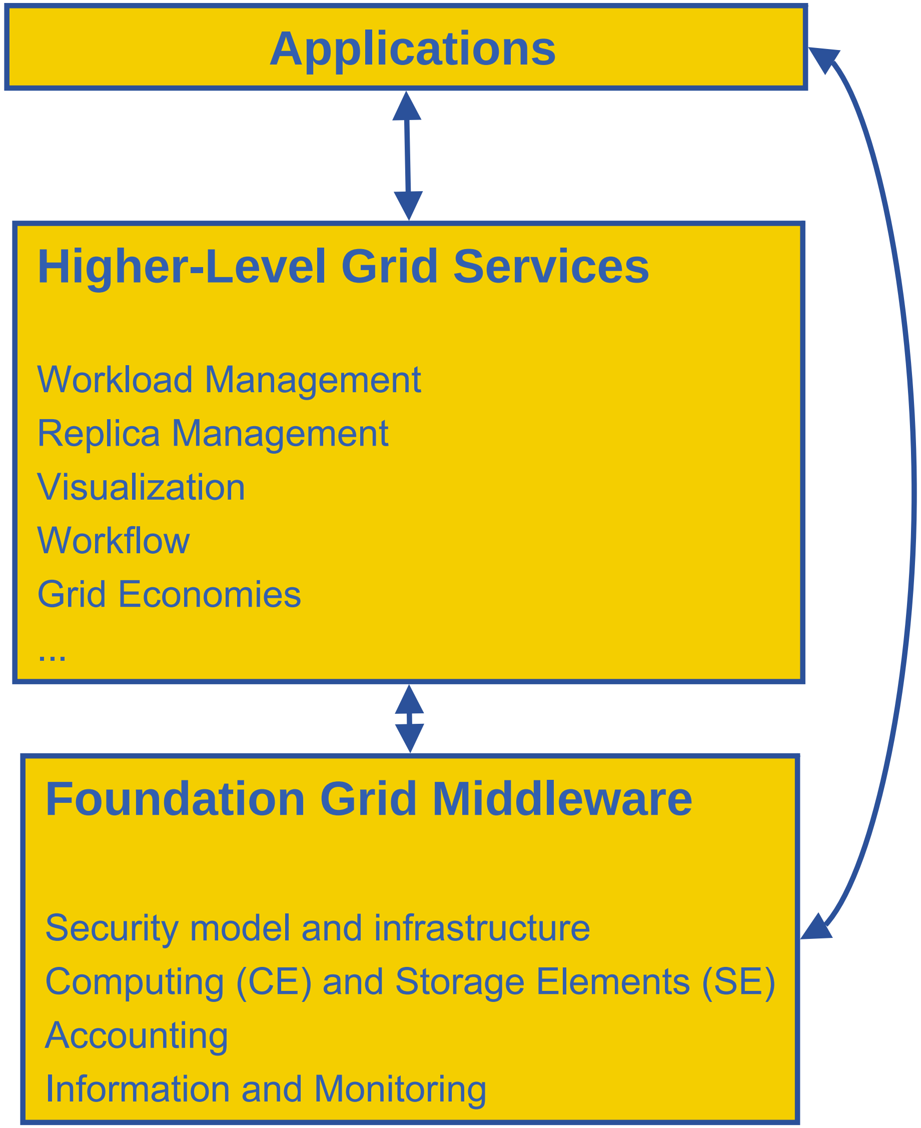


EGEE Production Service



98k jobs/day

Middleware structure



Applications

Higher-Level Grid Services

Workload Management
Replica Management
Visualization
Workflow
Grid Economies
...

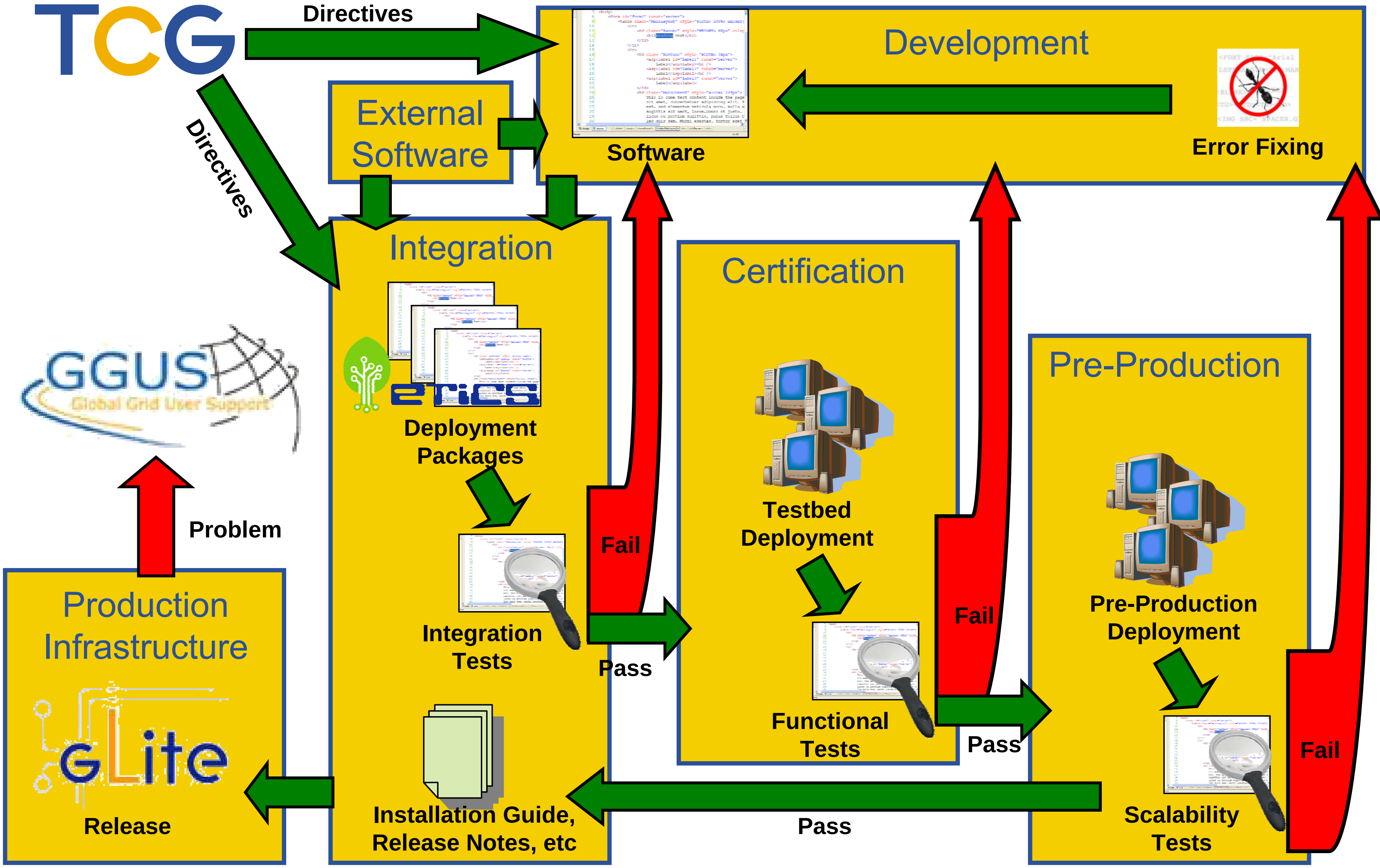
Foundation Grid Middleware

Security model and infrastructure
Computing (CE) and Storage Elements (SE)
Accounting
Information and Monitoring

- **Applications have access both to Higher-level Grid Services and to Foundation Grid Middleware**
- **Higher-Level Grid Services are supposed to help the users building their computing infrastructure but should not be mandatory**
- **Foundation Grid Middleware will be deployed on the EGEE infrastructure**
  - Must be complete and robust
  - Should allow interoperation with other major grid infrastructures
  - Should not assume the use of Higher-Level Grid Services

Overview paper http://doc.cern.ch//archive/electronic/egee/tr/egee-tr-2006-001.pdf

gLite Process



TCG

Directives

Directives

External Software

Development

Software

Error Fixing

Integration

Deployment Packages

Integration Tests

Installation Guide, Release Notes, etc

Fail

Pass

Certification

Testbed Deployment

Functional Tests

Fail

Pass

Pre-Production

Pre-Production Deployment

Scalability Tests

Fail

Pass

Production Infrastructure

Release

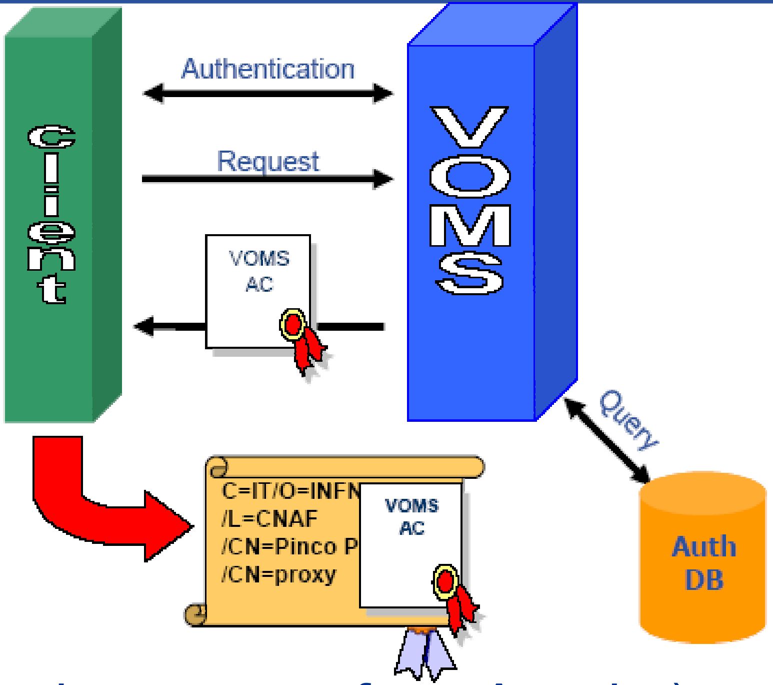
Problem

Authentication

- **Authentication is based on X.509 PKI infrastructure**
  - Certificate Authorities (CA) issue (long lived) certificates identifying individuals (much like a passport)
    - Commonly used in web browsers to authenticate to sites
  - Trust between CAs and sites is established (offline)
  - In order to reduce vulnerability, on the Grid user identification is done by using (short lived) proxies of their certificates
- **Short-Lived Credential Services (SLCS)**
  - issue short lived certificates or proxies to its local users
    - e.g. from Kerberos or from Shibboleth credentials (new in EGEE II)
- **Proxies can**
  - Be delegated to a service such that it can act on the user's behalf
  - Be stored in an external proxy store (MyProxy)
  - Be renewed (in case they are about to expire)
  - Include additional attributes

Authorization

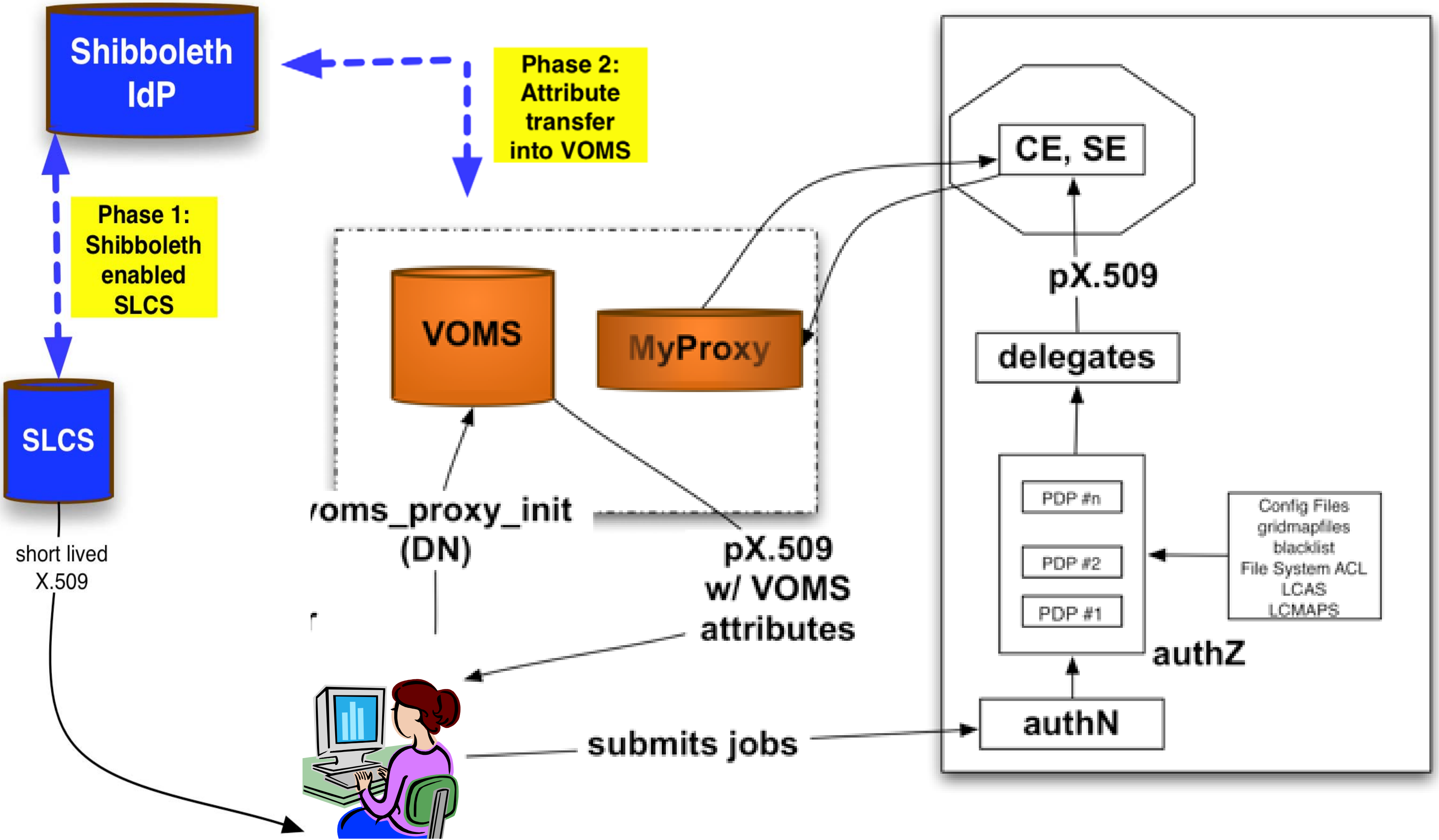
- **VOMS service issues Attribute Certificates that are attached to certificate proxies**
  - Provide users with additional capabilities defined by the Virtual Organization
  - Base for the Authorization process
- **Authorization: via mapping to a local user on the resource**



  - glexec changes the local identity (based on suexec from Apache)
  - LCAS/LCMAPS use different plug-ins to determine if and how to map a grid user to a local user
    - mainly used for C-based applications
  - gLite Java Authorization Framework (XACML-compatible)
    - mainly used for Java-based applications
  - Compatible with the future *G-PBox* policy management system

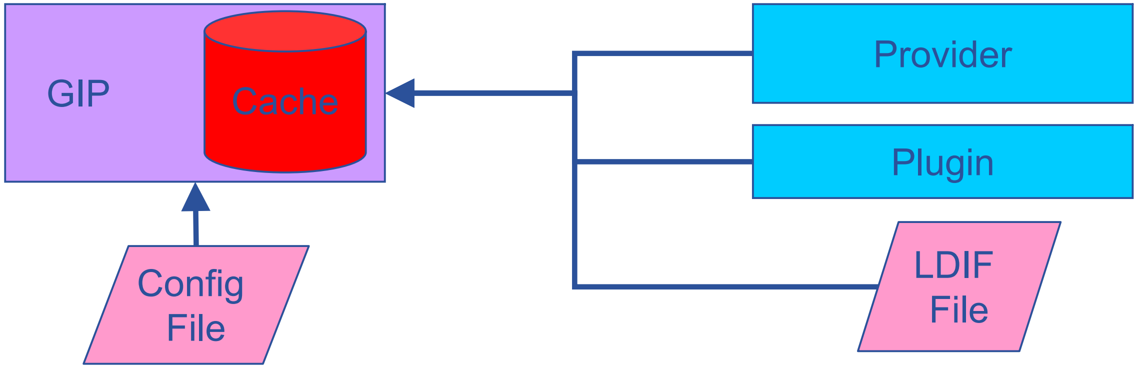
Coming: Shibboleth SLCS

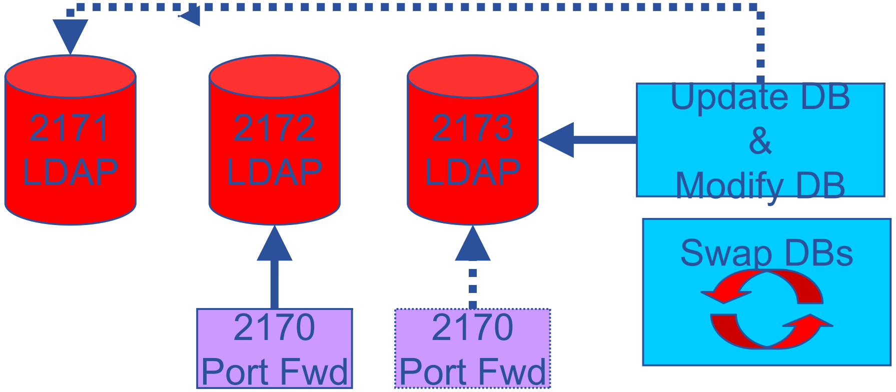
Long lived certificates may be replaced by short lived certificates provided by a Shibboleth identity Provider



Information Systems

- **Generic Information Provider (GIP)**
  - Provides information about a grid service in accordance to the GLUE Schema
- **BDII: Information system**
  - LDAP database that is updated by a process
  - More than one DBs is used separate read and write
  - A port forwarder is used internally to select the correct DB

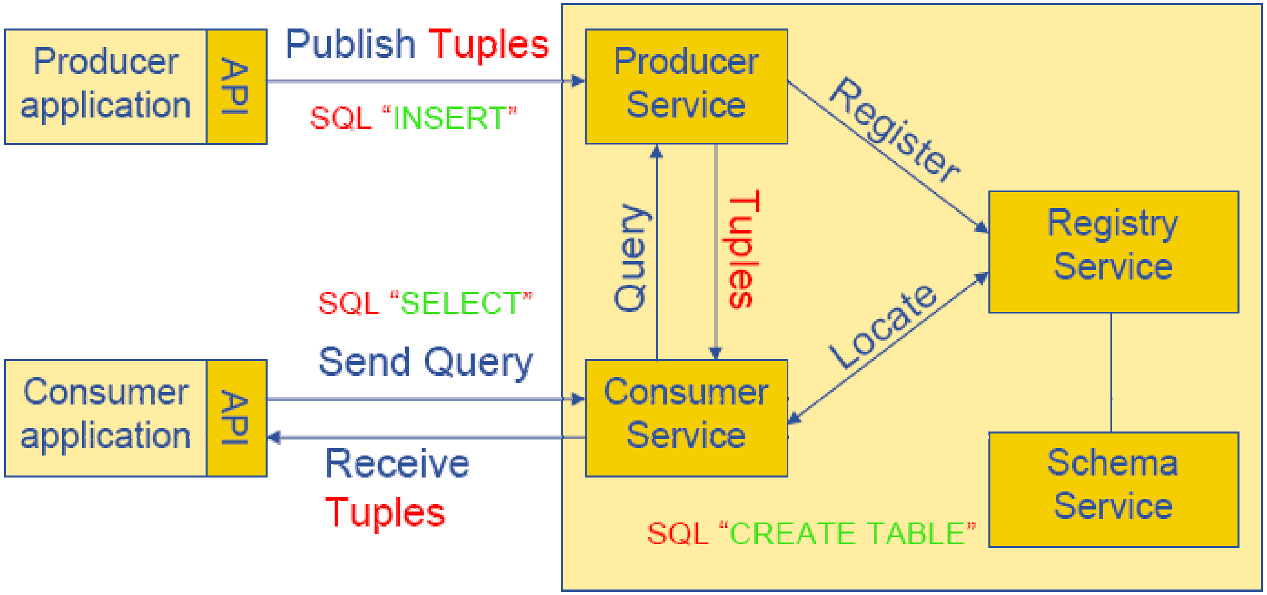




- **Freedom of choice portal: VOs can white- or black-list resources so that BDII DBs are updated accordingly**
- **Sites failing Site Functional Tests may also be excluded**
- **Up to 2 million queries per day served (over 20 Hz)**

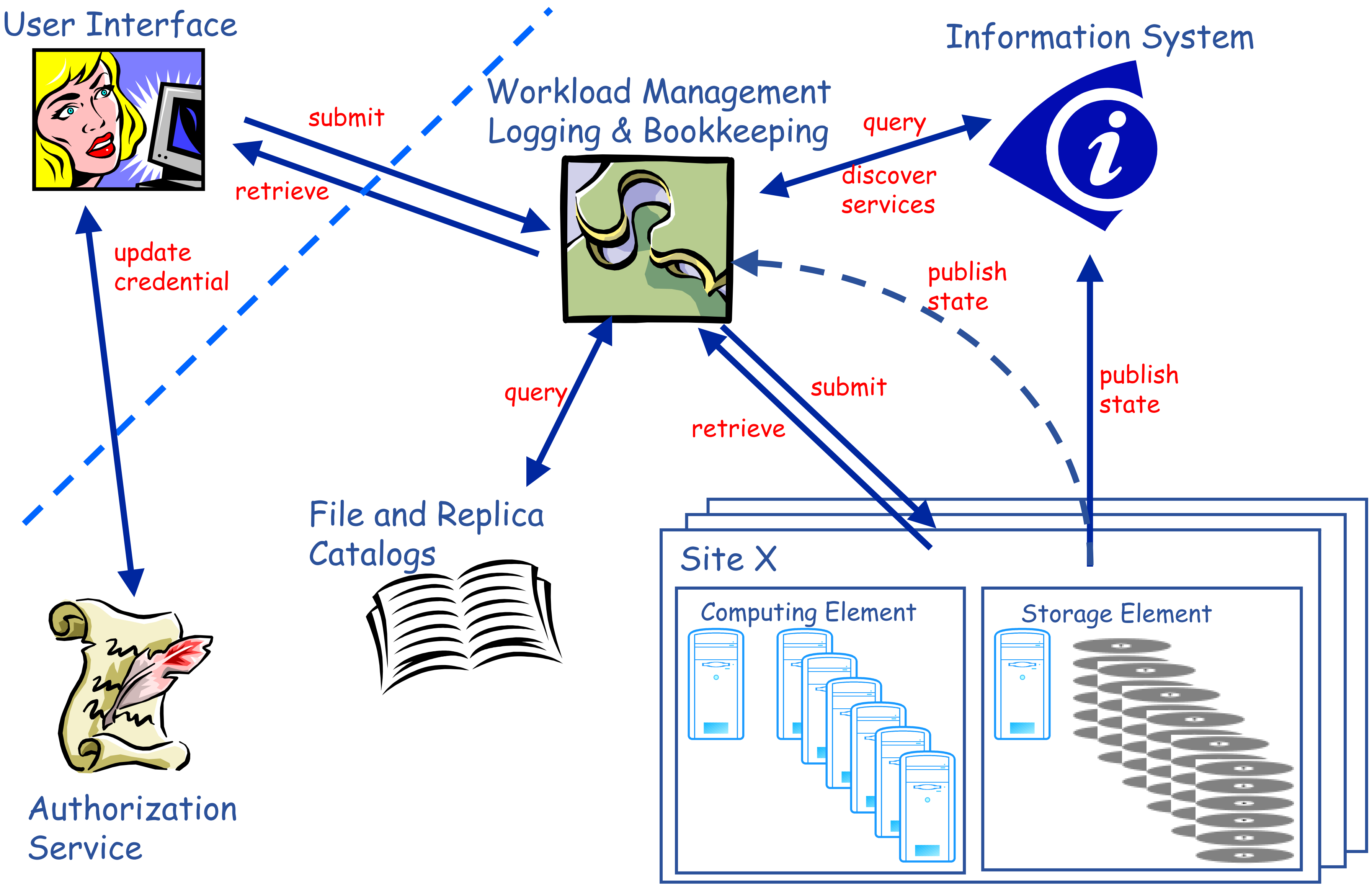
Monitoring

- **R-GMA: provides a uniform method to access and publish distributed information and monitoring data**
  - Backbone of EGEE job and infrastructure monitoring
  - 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

Job Management Services



Resource Access in EGEE

- ➔ **LCG-CE (GT2 GRAM)**
  - Not ported to GT4. To be dismissed
- ➔ **gLite-CE (Condor-C+GSI)**
  - Deployed (GT2 version) but still needs tuning
- ➔ ***CREAM (WS-I)***
  - Prototype. OGF-BES (see demo at SC'06)
- **Possible developments:**
  - GT4 → BLAH submissions?

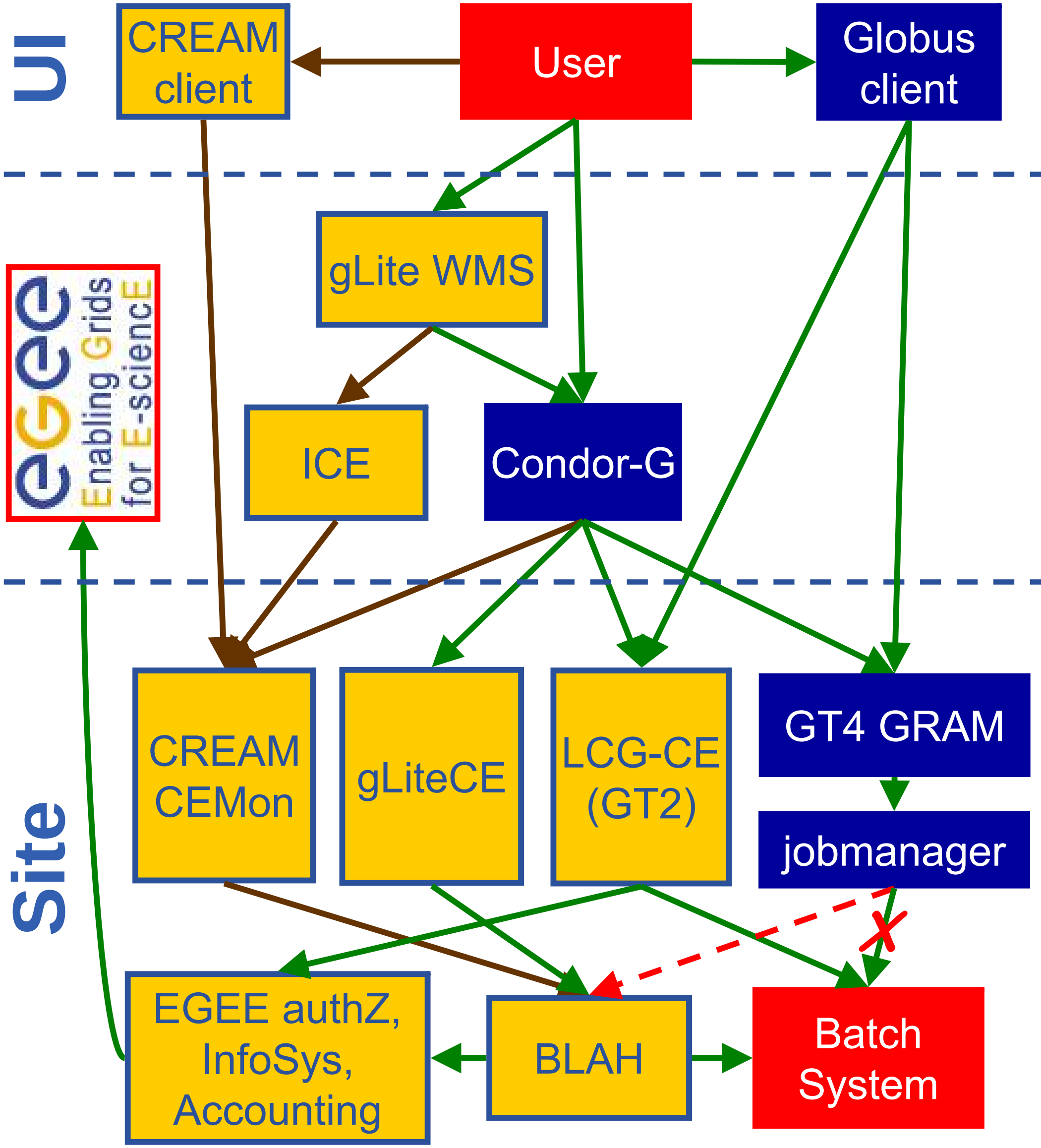
Choose your preferred path to the Batch System!

gLite component	non-gLite component	User / Resource

→ In production

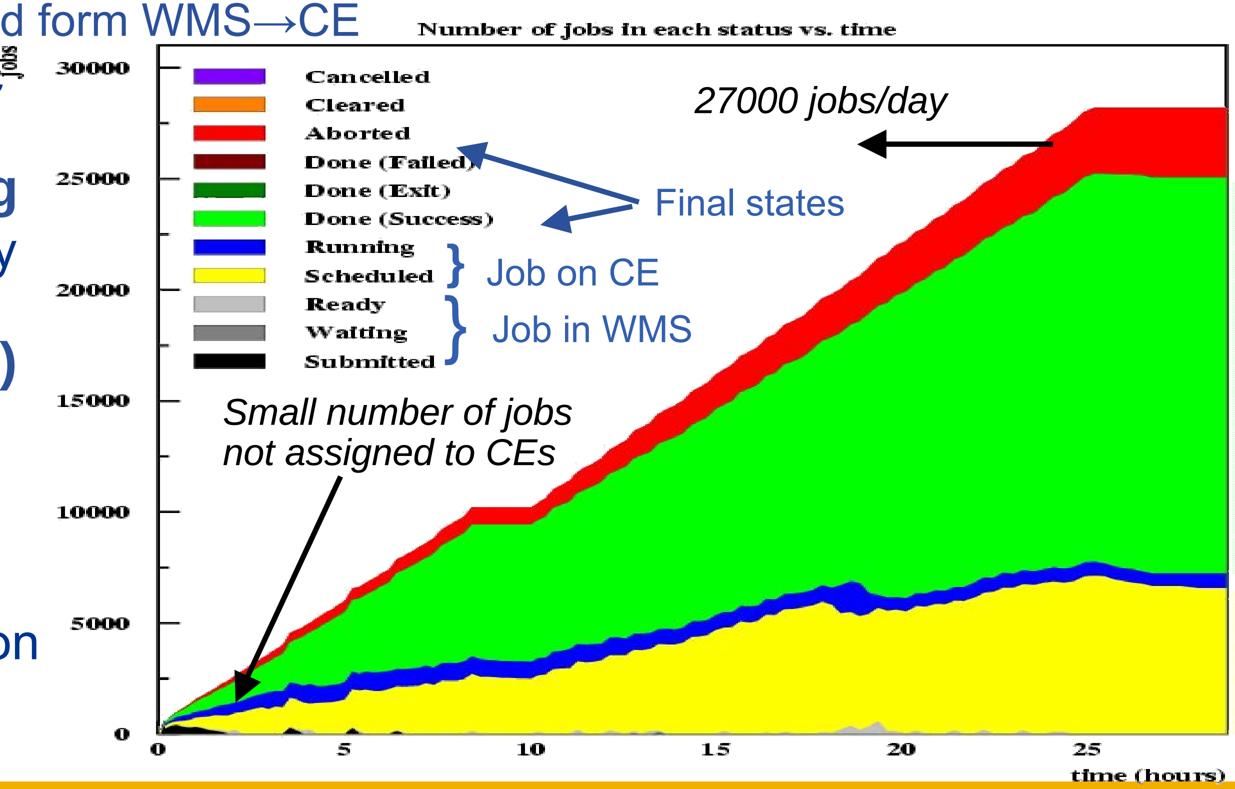
→ Existing prototype

- - → Possible development



Workload Management System

- **Workload Management System**
  - Assigns jobs to resources according to user requirements
    - possibly including data location and user defined ranking of resources
  - Handles I/O data (*input* and *output sandboxes*)
  - Support for compound jobs and workflows (Direct Acyclic Graphs)
    - One shot submission of a group of jobs, shared *input sandbox*
  - Web Service interface: WMProxy
    - UI→WMS decoupled form WMS→CE
  - Support for automatic re-submissions
- **Logging&Bookkeeping**
  - Tracks jobs while they are running
- **Job Provenance (new!)**
  - Store and retain data on finished jobs
  - Provides data mining capabilities
  - Allows job re-execution



27000 jobs/day

Final states

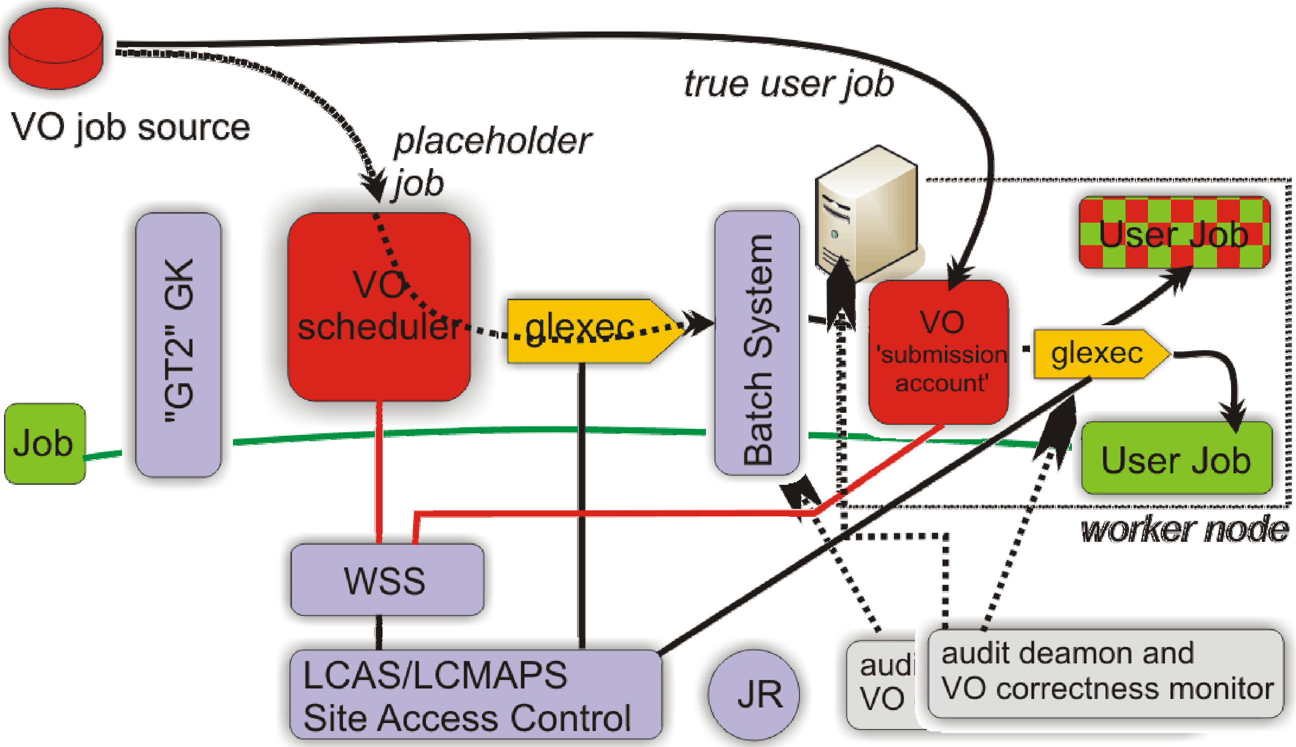
Job on CE

Job in WMS

Small number of jobs not assigned to CEs

Coming: support for pilot jobs

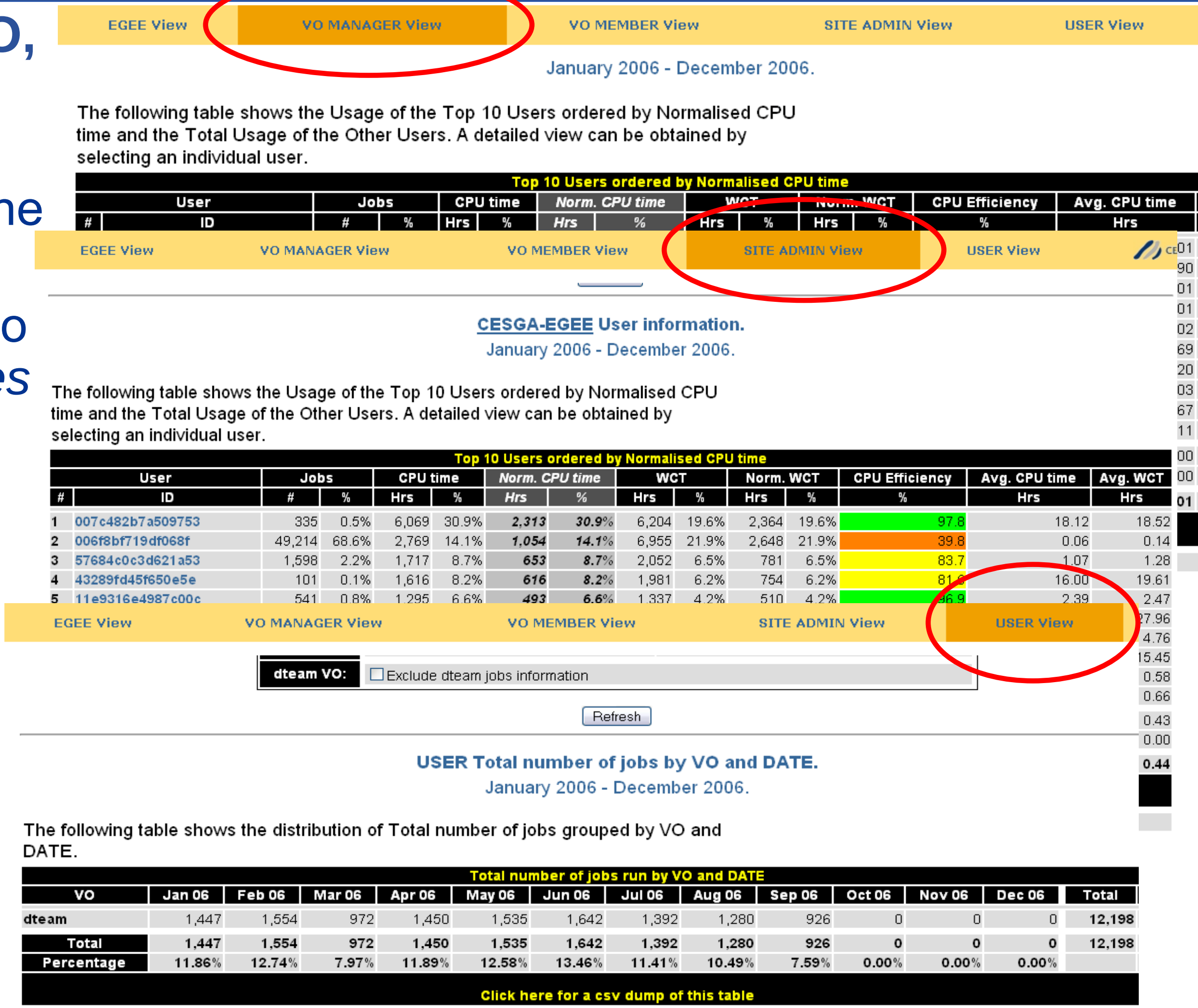
- **Several VOs submit *pilot jobs* with a single identity for all of the VO**
  - The pilot job gets the user job when it arrives on the WN and executes it
    - Just-in-time scheduling. VO policies implemented at the central queue



- **Use the same mechanism for changing the identity on the Computing Element also on the Worker Nodes (glexec)**
  - The site may know the identity of the real user

Job accounting

- **Resource usage by VO, group or single user**
  - Sensors running on resources to determine usage
  - It would be possible to enable *Pricing policies* associate a cost to resource usage
    - market-based resource brokering
  - privacy: access to accounting data granted only to authorized people (user, provider, VO manager)



January 2006 - December 2006.

The following table shows the Usage of the Top 10 Users ordered by Normalised CPU time and the Total Usage of the Other Users. A detailed view can be obtained by selecting an individual user.

CESGA-EGEE User information.
January 2006 - December 2006.

The following table shows the Usage of the Top 10 Users ordered by Normalised CPU time and the Total Usage of the Other Users. A detailed view can be obtained by selecting an individual user.

Top 10 Users ordered by Normalised CPU time

User		Jobs		CPU time		Norm. CPU time		WCT		Norm. WCT		CPU Efficiency	Avg. CPU time	Avg. WCT
#	ID	#	%	Hrs	%	Hrs	%	Hrs	%	Hrs	%	%	Hrs	Hrs
1	007c482b7a509753	335	0.5%	6,069	30.9%	2,313	30.9%	6,204	19.6%	2,364	19.6%	97.8	18.12	18.52
2	006f8bf719df068f	49,214	68.6%	2,769	14.1%	1,054	14.1%	6,955	21.9%	2,648	21.9%	39.8	0.06	0.14
3	57684c0c3d621a53	1,598	2.2%	1,717	8.7%	653	8.7%	2,052	6.5%	781	6.5%	83.7	1.07	1.28
4	43289fd45f650e5e	101	0.1%	1,616	8.2%	616	8.2%	1,981	6.2%	754	6.2%	81.9	16.00	19.61
5	11e9316e4987c00c	541	0.8%	1,295	6.6%	493	6.6%	1,337	4.2%	510	4.2%	96.9	2.39	2.47

dteam VO: □ Exclude dteam jobs information

USER Total number of jobs by VO and DATE.
January 2006 - December 2006.

The following table shows the distribution of Total number of jobs grouped by VO and DATE.

Total number of jobs run by VO and DATE

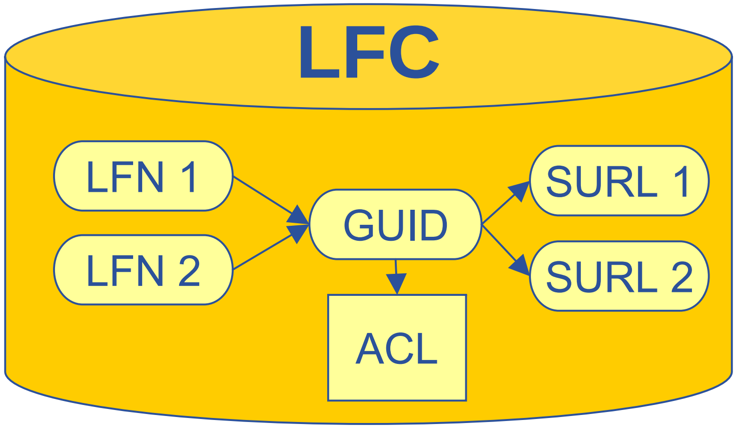
VO	Jan 06	Feb 06	Mar 06	Apr 06	May 06	Jun 06	Jul 06	Aug 06	Sep 06	Oct 06	Nov 06	Dec 06	Total
dteam	1,447	1,554	972	1,450	1,535	1,642	1,392	1,280	926	0	0	0	12,198
Total	1,447	1,554	972	1,450	1,535	1,642	1,392	1,280	926	0	0	0	12,198
Percentage	11.86%	12.74%	7.97%	11.89%	12.58%	13.46%	11.41%	10.49%	7.59%	0.00%	0.00%	0.00%	

Click here for a csv dump of this table

- **Information collected at the Grid Operations Centre (GOC)**
- **Basic functionality in APEL, full functionality in DGAS**

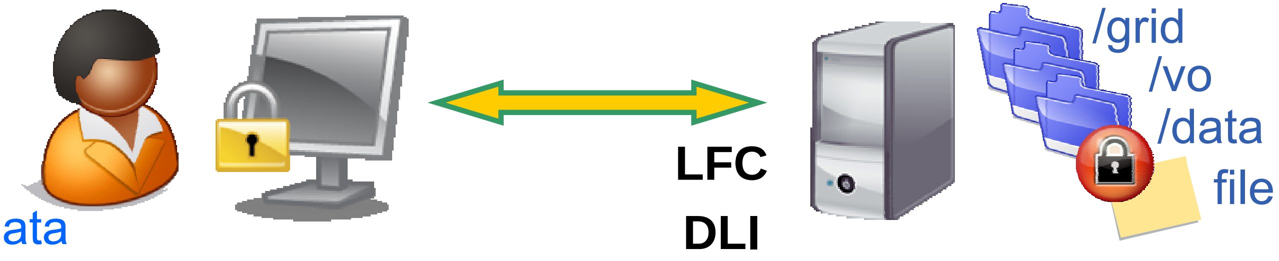
LCG File Catalog

- **LFC maps LFNs to SURLs**
  - *Logical File Name* (LFN): user file name
    - in VO namespace, aliases supported
  - *Glbally Unique IDentifier* (GUID)
    - unique string assigned by the system to the file
  - *Site URL* (SURL): identifies a replica
  - A Storage Element and the logical name of the file inside it



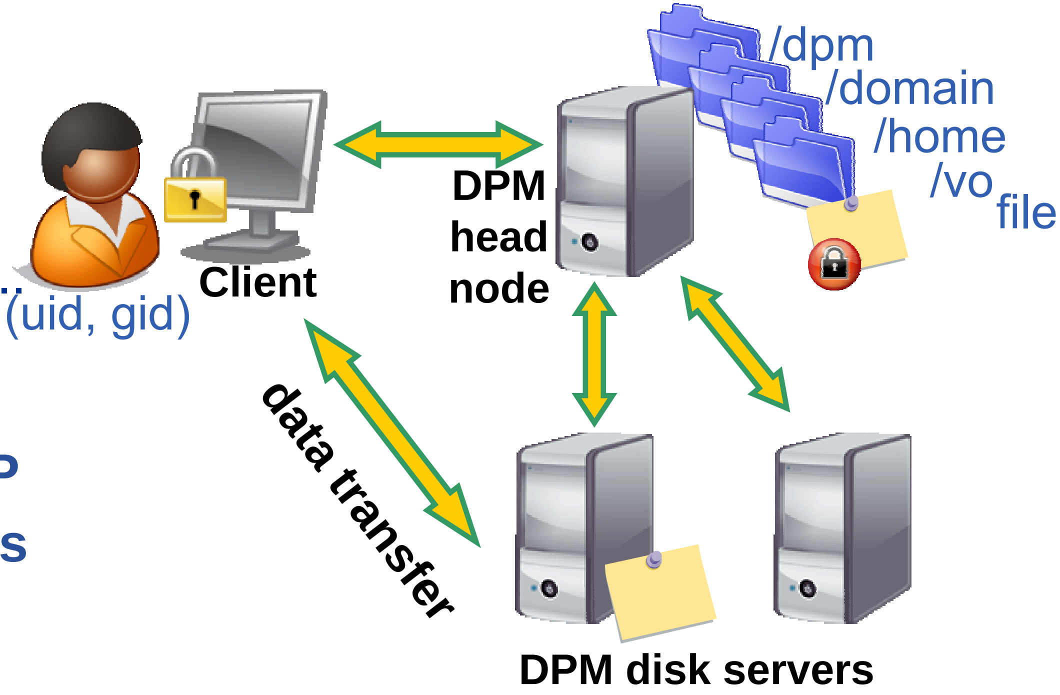
- **GSI security: ACLs (based on VOMS)**
  - To each VOMS group/role corresponds a virtual group identifier
  - Support for secondary groups
- **Web Service query interface: Data Location Interface (DLI)**
- **Hierarchical Namespace**
- **Supports sessions and bulk operations**

```
lfc-ls –l /grid/vo/
lfc-getacl /grid/vo/data
```



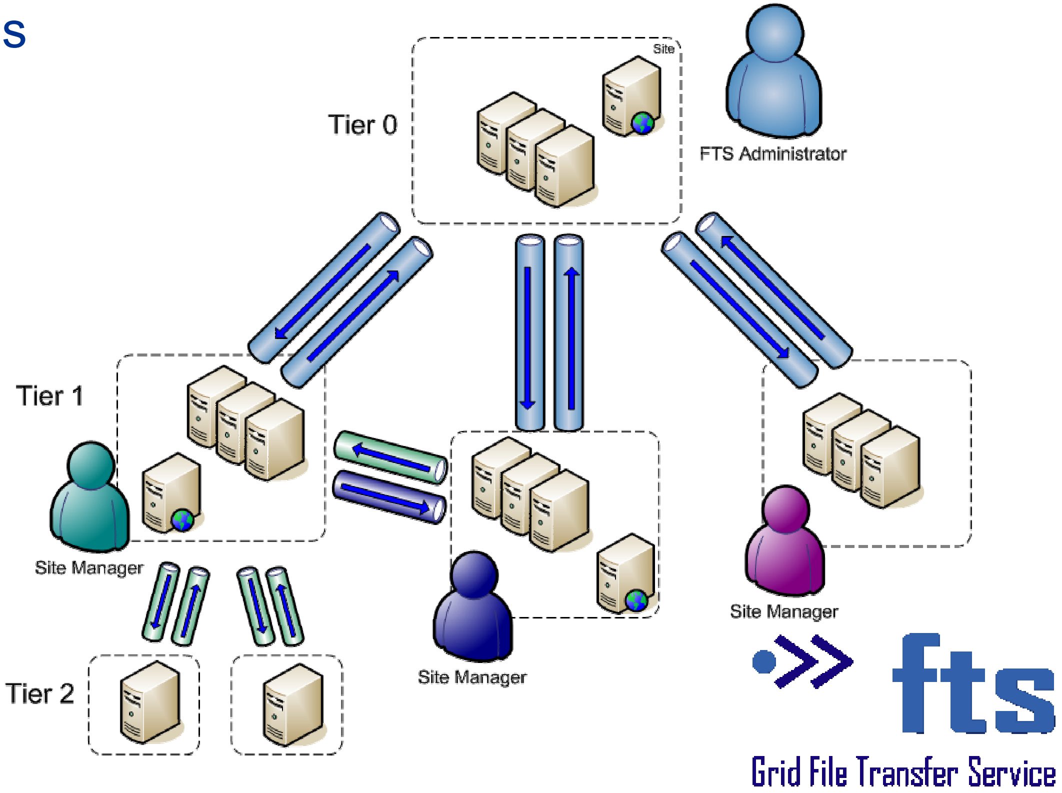
Storage Element - DPM

- **Storage Resource Manager (SRM): translates SURLs to TURLs**
  - *Transfer URL* (TURL): allows direct access to the file
  - Interface that hides the storage system implementation
  - Handles the authorization based on VOMS credentials
- **Disk-based: DPM, dCache,+; tape-based: Castor, dCache**
- **File I/O: *posix-like* access from local nodes or the grid**
  - GFAL (Grid File Access Layer)
- **Disk Pool Manager (DPM)**
  - Manages storage on disk servers
- **Uses LFS as local catalog**
  - Same features for ACLs, etc...
- **Direct data transfer from/to disk server (no bottleneck)**
- **External transfers via gridFTP**
- **Target: small to medium sites**
  - One or more disk servers



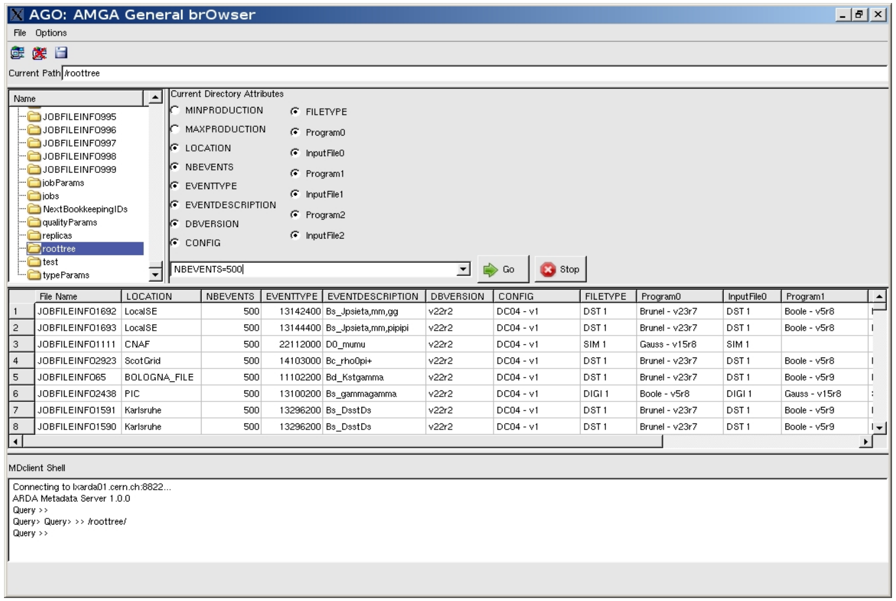
File transfer Service

- **FTS: Reliable, scalable and customizable file transfer**
  - Multi-VO service, used to balance usage of site resources according to the SLAs agreed between a site and the VOs it supports
  - WS interface, support for different user and administrative roles (VOMS)
  - Manages transfers through channels
    - mono-directional network pipes between two sites
  - File transfers handled as jobs
    - Prioritization
    - Retries in case of failures
  - Automatic discovery of services
- **Designed to scale up to the transfer needs of very data intensive applications**
  - Demonstrated about 1 GB/s sustained
  - Over 9 petabytes transferred in the last 6 months (> 10 million files)



AMGA Metadata Catalog

- **AMGA is a general purpose metadata catalog**
  - Keeps information about data stored in files
  - Used by several application domains
  - SOAP interface
  - VOMS authorization
  - Shell-like client
  - Graphical Browser (Python)
- **Performance comparable to direct DB access**
  - C++, TCP streaming protocol, very fast SSL sessions
- **LHCb (HEP VO use case)**
  - 100 Million entrie
  - 150GB data
  - 10^5 entries/day insert rate
  - 10 entries/sec read-rate



Current Path /roottree

Name
- JOBFILEINFO0995
- JOBFILEINFO0996
- JOBFILEINFO0997
- JOBFILEINFO0998
- JOBFILEINFO0999
- jobParams
- jobs
- NextBookkeepingIDs
- qualityParams
- replicas
- roottree
- test
- typeParams

Current Directory Attributes

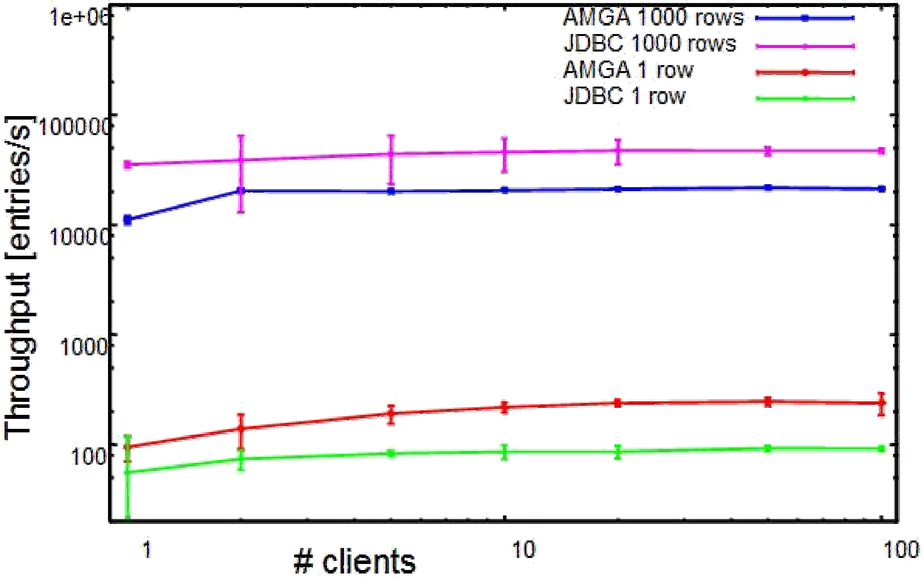
○ MINPRODUCTION
○ MAXPRODUCTION
● LOCATION
● NBEVENTS
● EVENTTYPE
● EVENTDESCRIPTION
● DBVERSION
● CONFIG
● FILETYPE
● Program0
● InputFile0
● Program1
● InputFile1
● Program2
● InputFile2

NBEVENTS=500

	File Name	LOCATION	NBEVENTS	EVENTTYPE	EVENTDESCRIPTION	DBVERSION	CONFIG	FILETYPE	Program0	InputFile0	Program1
1	JOBFILEINFO1692	LocalSE	500	13142400	Bs_Jpsieta,mm,gg	v22r2	DC04 - v1	DST 1	Brunel - v23r7	DST 1	Boole - v5r8
2	JOBFILEINFO1693	LocalSE	500	13144400	Bs_Jpsieta,mm,pipipi	v22r2	DC04 - v1	DST 1	Brunel - v23r7	DST 1	Boole - v5r8
3	JOBFILEINFO1111	CNAF	500	22112000	D0_mumu	v22r2	DC04 - v1	SIM 1	Gauss - v15r8	SIM 1	
4	JOBFILEINFO2923	ScotGrid	500	14103000	Bc_rho0pi+	v22r2	DC04 - v1	DST 1	Brunel - v23r7	DST 1	Boole - v5r8
5	JOBFILEINFO65	BOLOGNA_FILE	500	11102200	Bd_Kstgamma	v22r2	DC04 - v1	DST 1	Brunel - v23r7	DST 1	Boole - v5r9
6	JOBFILEINFO2438	PIC	500	13100200	Bs_gammagamma	v22r2	DC04 - v1	DIGI 1	Boole - v5r8	DIGI 1	Gauss - v15r8
7	JOBFILEINFO1591	Karlsruhe	500	13296200	Bs_DsstDs	v22r2	DC04 - v1	DST 1	Brunel - v23r7	DST 1	Boole - v5r9
8	JOBFILEINFO1590	Karlsruhe	500	13296200	Bs_DsstDs	v22r2	DC04 - v1	DST 1	Brunel - v23r7	DST 1	Boole - v5r9

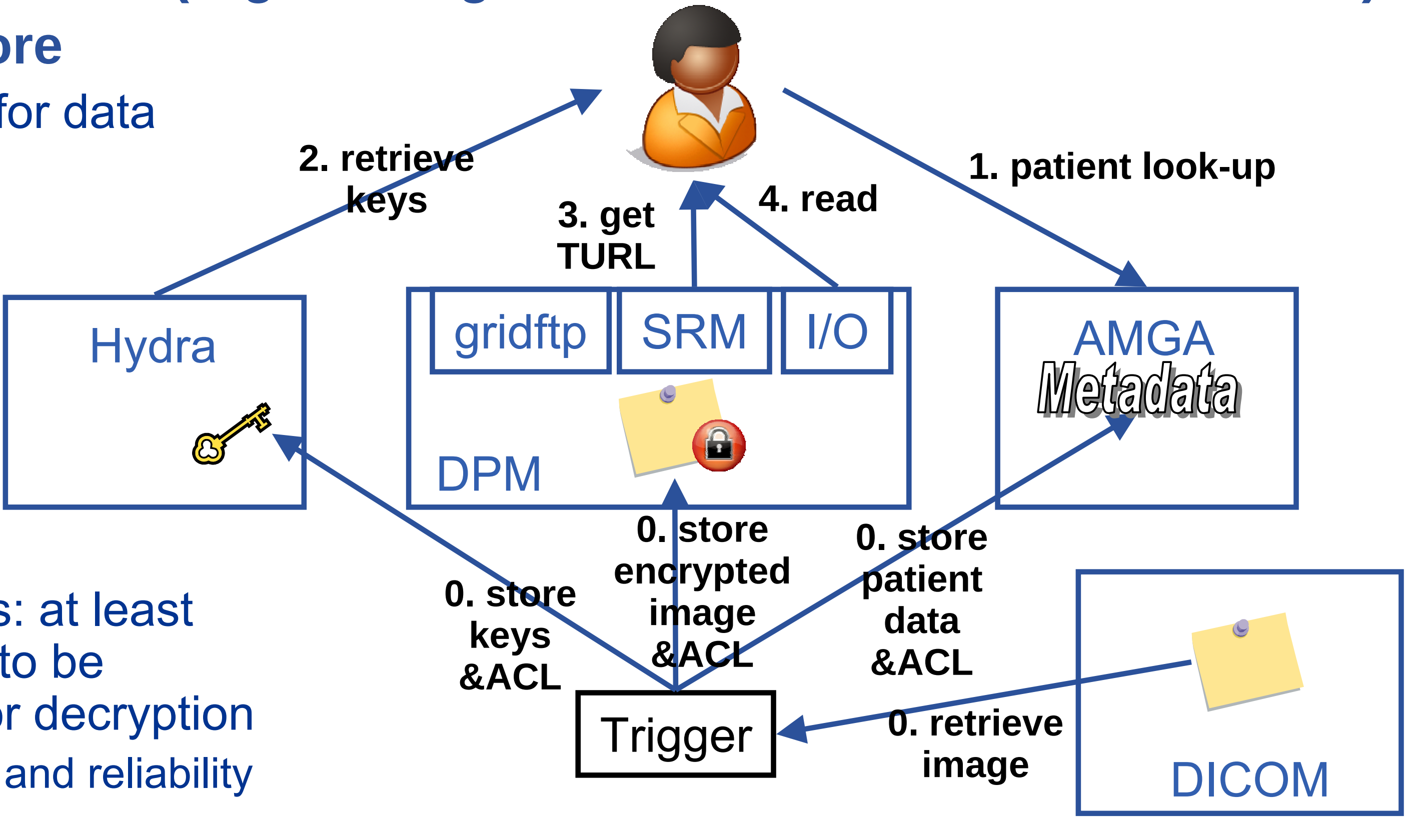
MDclient Shell

```
Connecting to lxarda01.cern.ch:8822...
ARDA Metadata Server 1.0.0
Query >>
Query> Query> >> /roottree/
Query >>
```



Encrypted Data Storage

- **Intended for VO’s with very strong security requirements**
  - e.g. medical community
    - anonymity (patient data is separate)
    - fine grained access control (only selected individuals)
    - privacy (even storage administrator cannot read)
- **Interface to DICOM (Digital Image and COmmunication in Medicine)**
- **Hydra keystore**
  - store keys for data encryption
  - N instances: at least M<N need to be available for decryption
    - security and reliability



How to contribute/influence

- **gLite process driven by application and operational requirements**
  - New components added based on their requests and overall importance
- **RESPECT – Program to collect useful tools that work with gLite**
  - See EGEE application portal: http://egeena4.lal.in2p3.fr/index.php
  - Under construction