



Enabling Grids for E-science

Grid Middleware

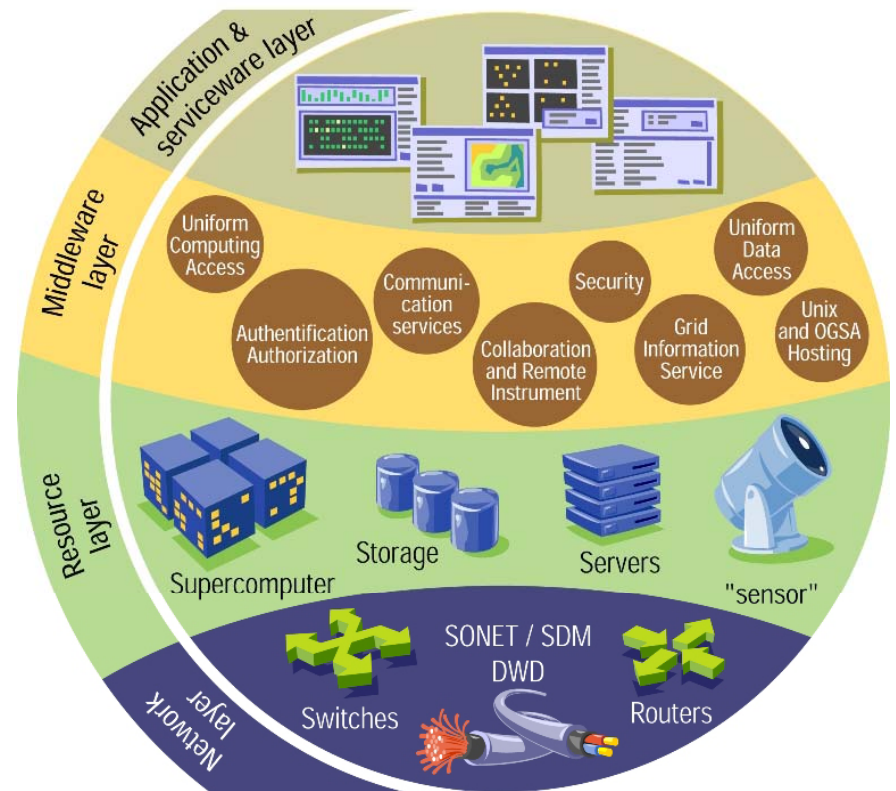
Andreas Unterkircher
gLite Release Manager
CERN

Openlab Summer Student Lecture 2009

www.eu-egee.org



- Grid Computing
- Grid Usage
- The gLite middleware
 - <http://glite.org/>
 - <http://www.eu-egee.org/>
- Future development



- Let's ask Les Robertson (Head of WLCG until end of 2008):
- “Computing wasn't included in the original costs of the LHC. [...] We clearly required computing, but the original idea was that it could be handled by other people”.

Around 2000, these “other people” had not stepped forward.

“There was no funding at CERN or elsewhere. A single organization could never find the money to do it. We realized the system would have to be distributed.

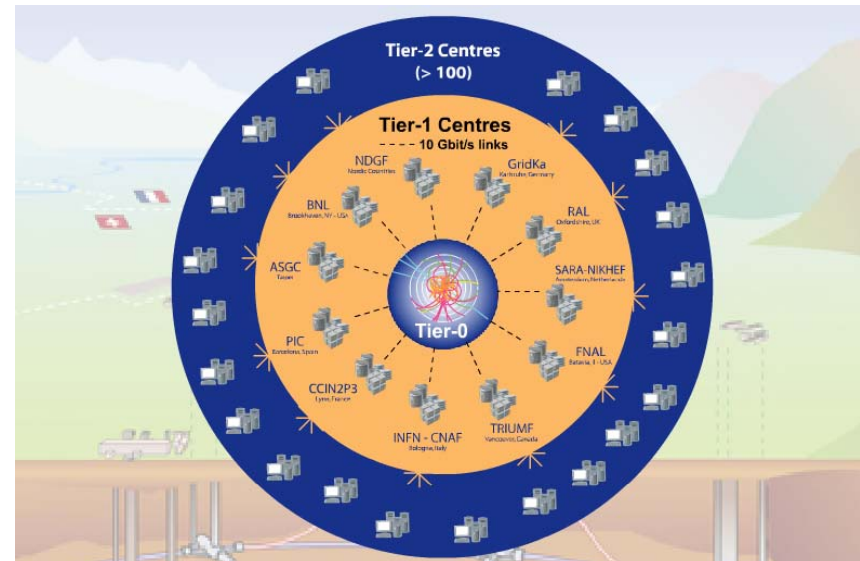
(Interview with Les Robertson, iSGTW Feb. 08)

- **Different computing centers provide resources, but**
 - A single user has to ask for access individually at every site
 - Usage rights, disk quotas, compute shares etc. are different at every site
 - Different batch systems (different CLI/API)
 - How do you get your data to every site?

At the end of the 90ies the HEP community initiated the MONARC project to do distributed computing for LHC.

At the same time “Grid Computing” emerged.

- **All about collaboration and sharing resources**
 - Distributed groups can pool their computing resources
 - Large and small computing centers can contribute
 - Users everywhere can get equal access to data and computing power
 - Exploit funding wherever it is provided



- Many Grid infrastructures (using different middleware) have been built, e.g.:

- Open Science Grid (USA)
- Naregi (Japan)
- DEISA (European Supercomputer Grid)
- Nordugrid (North European Grid)
- TeraGrid (USA)

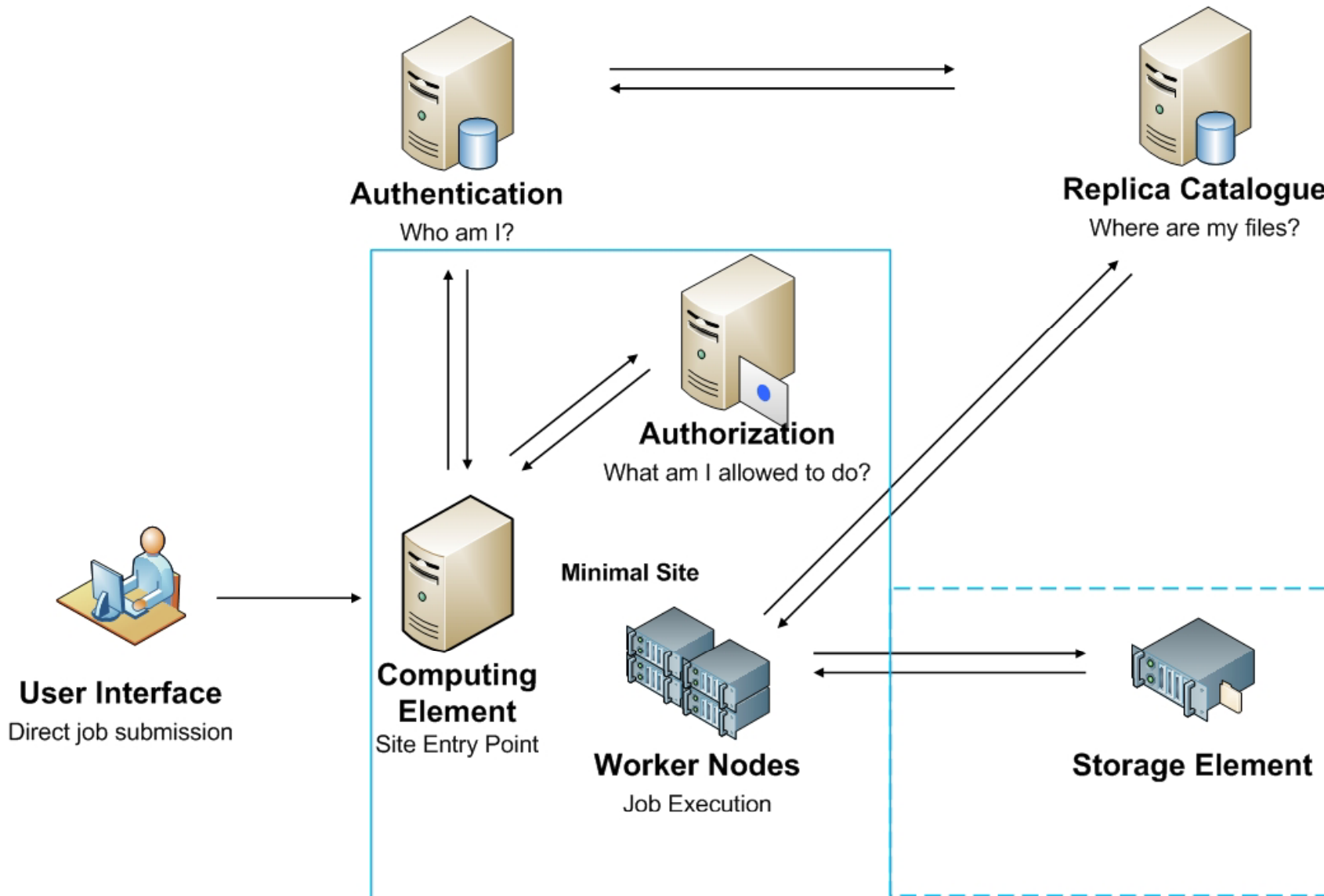
- EGEE
 - World's largest grid infrastructure (June 09 numbers)
 - 17 000 users
 - 139 000 LCPUs
 - 25Pb disk, 39Pb tape
 - 12 million jobs/month
 - 268 sites
 - 48 countries
 - 162 VO's



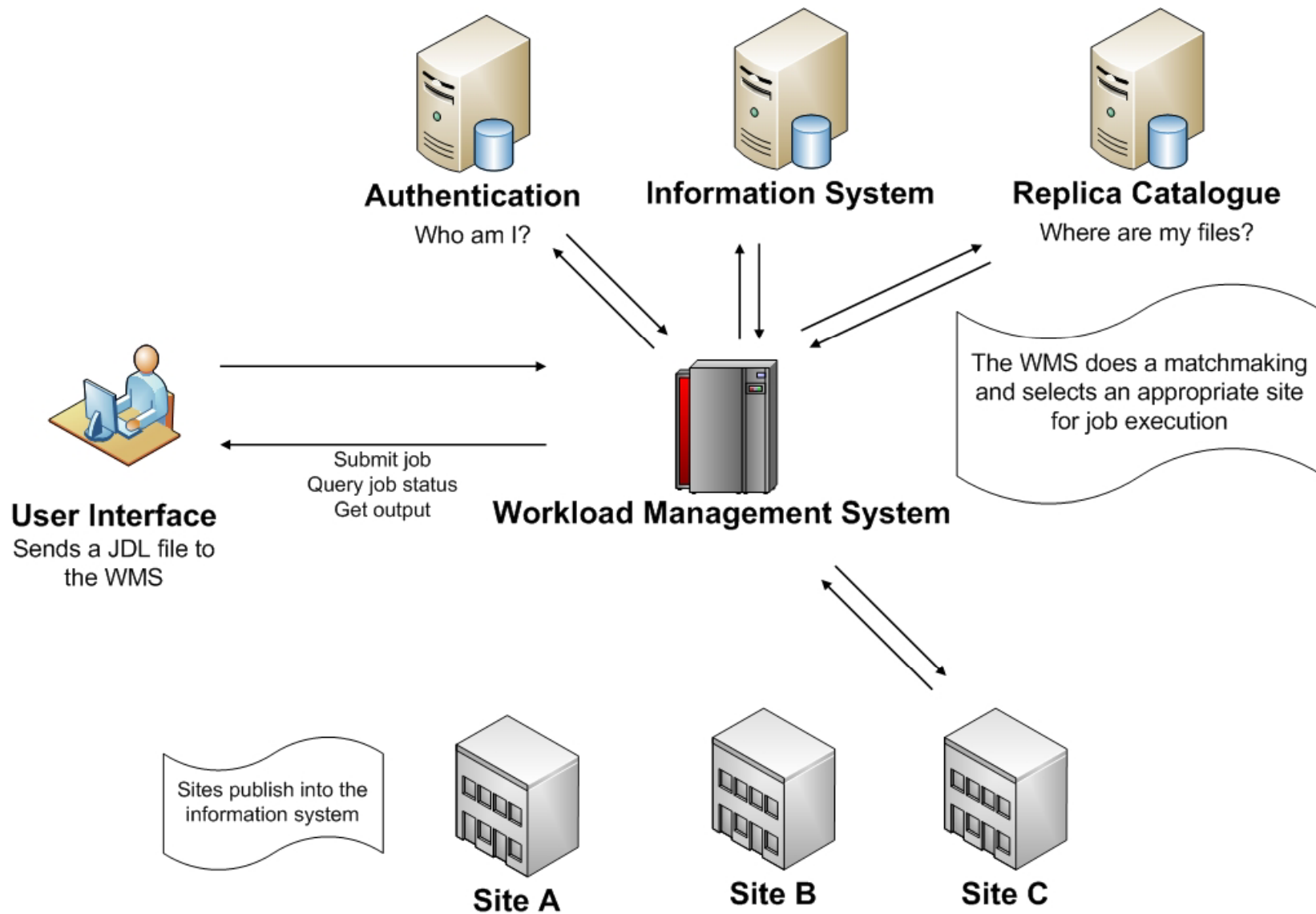
- **Virtual Organizations (VO)**
 - Researchers form VOs to collaborate, share resources and access common data.
 - Typically a long-lived group of researchers with a common purpose.
 - Each LHC experiment is managed by a single VO (large VOs ~ 2000 members)
 - Single VO can contain several subgroups that collaborate on specific topics but actually have little interaction with one another (e.g. biomedicine)
 - EGEE supports >150 VOs

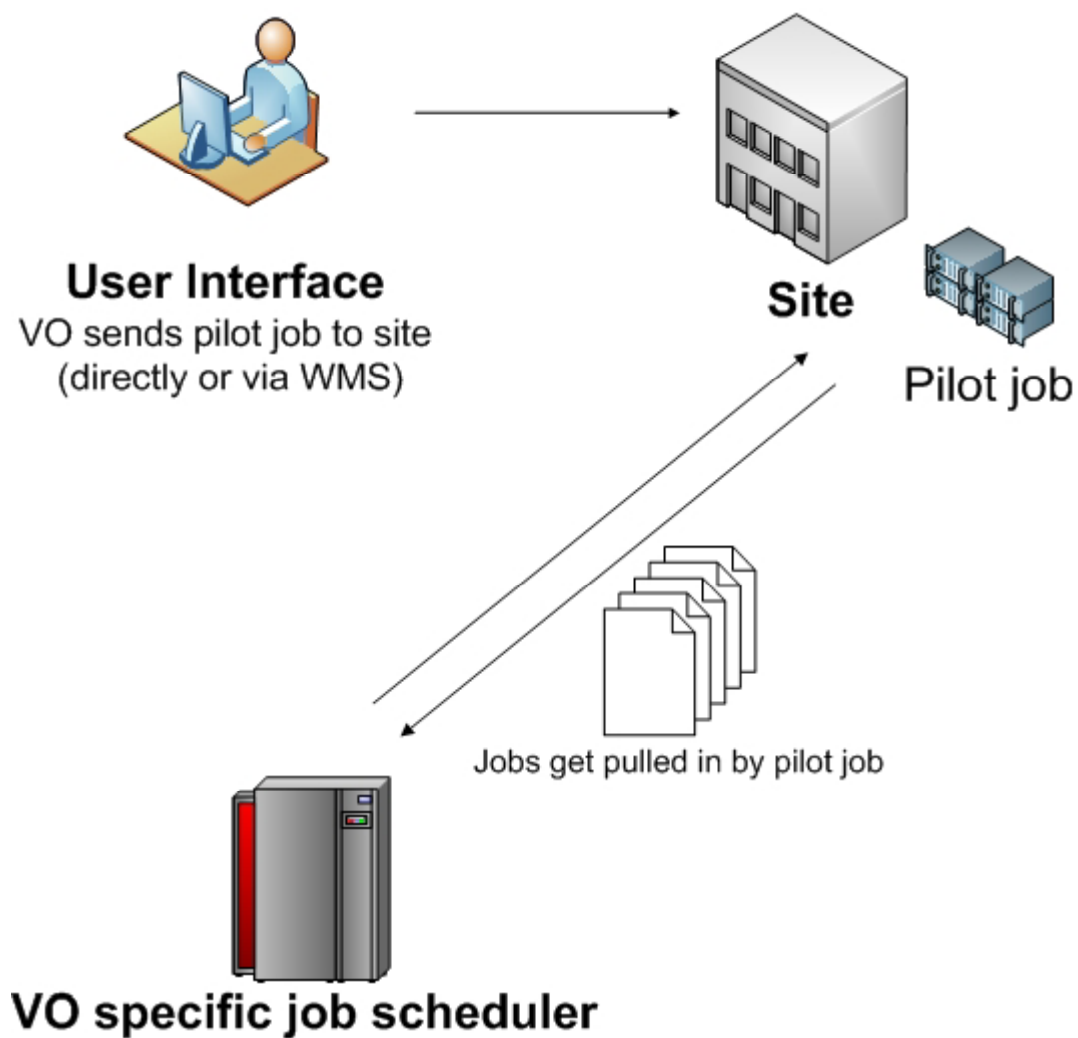
- **Grids are designed to handle large sets of limited duration jobs that produce or use huge amount of data**
- **How to submit a job to the Grid?**
 - Direct job submission
 - Job submission via the Workload Management System (WMS)
 - Pilot job submission

Direct job submission

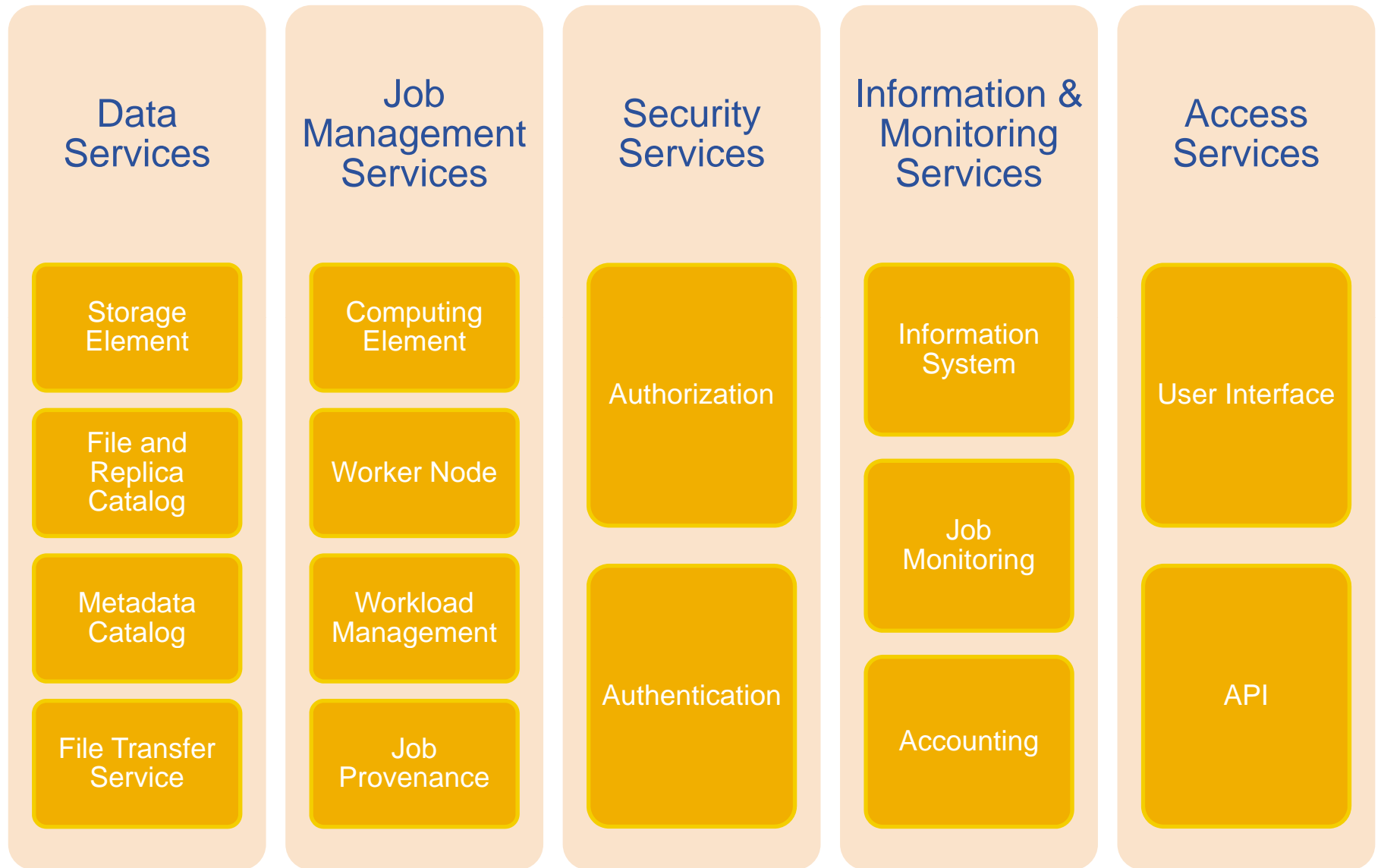


Job submission via WMS



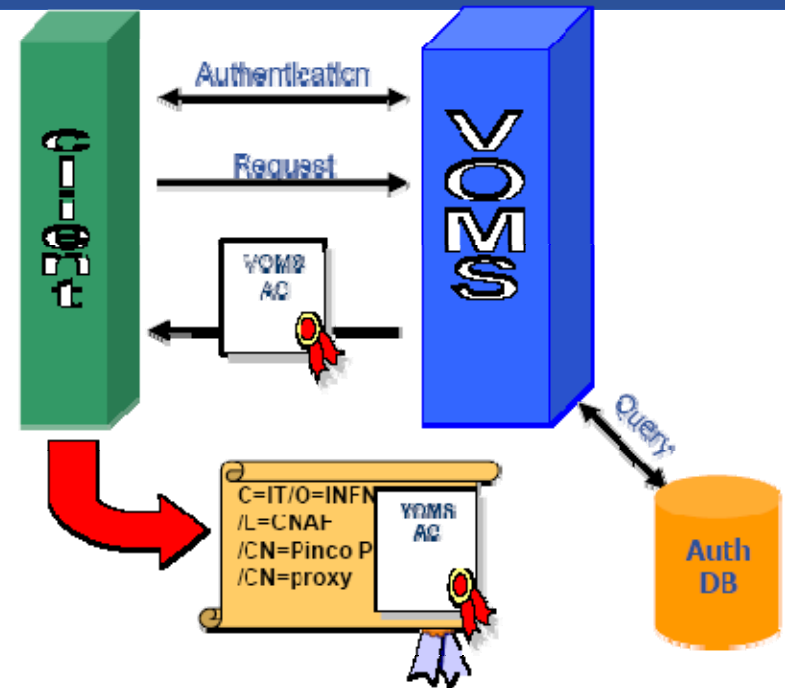


- WN reserved by pilot job is a first class resource (no need to wait for batch system or WMS)
- Just in time scheduling, VO policies implemented at the central VO queue
- Needs mechanism for identity switch for accounting (pilot job submitted by different user than pulled in jobs)



- **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**

- Virtual Organization Membership Service
- Central repository for user authorization information, providing support organizing users into a general group hierarchy, keeping track of their roles etc.
- **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



- **Based on their VOMS proxies users are mapped to a local account on the site**
- **Site Access Control (SAC) components**
 - Local Centre Authorization Service (LCAS)
 - Makes yes/no authorization decisions
 - Local Credential Mapping Service (LCMAPS)
 - Translates VOMS proxies to UNIX accounts
 - Site Central Authorization Service (SCAS)
 - Central administration point, contacted by clients (e.g. LCMAPS)
 - gLExec
 - Executes a binary with different uid (sudo); uses LCMAPS to get the uid. Used for pilot jobs
- **ARGUS**
 - New authorization service - more flexible, better support to specify policies, improved fault tolerance

- **Entry point to the site**
- **Takes grid job and gives it to the local batch system for execution**
- **lcg-CE: to be phased out**
- **CREAM (Compute Resource Execution And Management)**
 - Basic job operation (cancel, suspend etc.)
 - Proxy renewal & delegation
 - CLI

- **Worker Node**
 - A batch system node where the actual computation takes place
 - Has clients for data management, authorization service and information system
- **User Interface**
 - Contains all the clients needed to interact with the Grid: Job submission, data management, authorization, information system
 - To be installed on the user's PC

- **Disk Pool Manager (DPM)**

- Manages storage on disk servers
- Nearly 200 instances in use
- Can manage up to 360TB

- **dCache**

- Distributed with gLite but not developed within EGEE
- Can also handle tape storage

- **CASTOR**

- Handles tape storage
- Not distributed with gLite

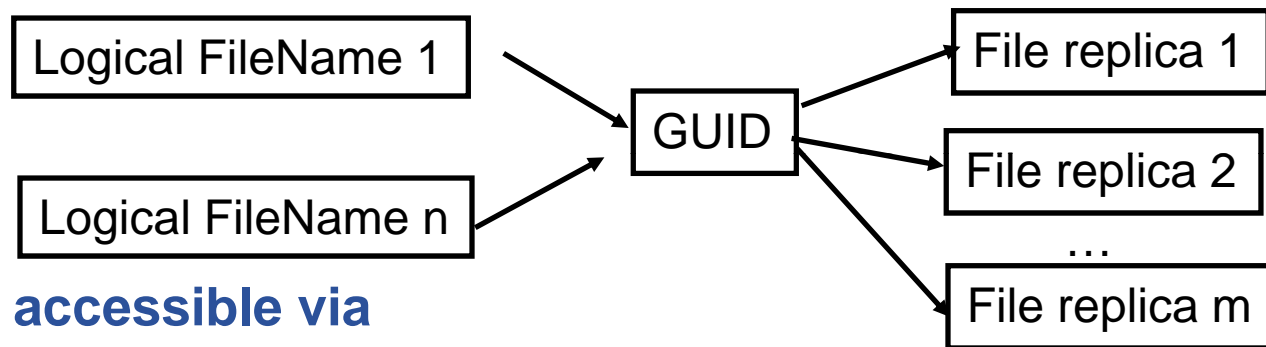
These SEs have their own File Access Protocols but also implement the SRM interface

- **STORM**

- Disk based
- Not distributed with gLite

- **Storage Resource Manager (SRM)**
 - Middleware component to provide dynamic space allocation and file management on shared storage components.
 - SRM v2.2 interface specified. Targets the specific needs in Grid Computing.
 - Storage Classes
 - Manage and reserve storage space
 - Filesystem-like operations
 - Volatile vs. permanent data
 - Transparent, automatic or forced migration to tertiary storage
 - Mechanisms for locating data
 - Namespaces
- **Implemented by all SEs (DPM, CASTOR, dCache, STORM)**

- The LFC stores mappings between
 - Users’ file names
 - File locations on the Grid

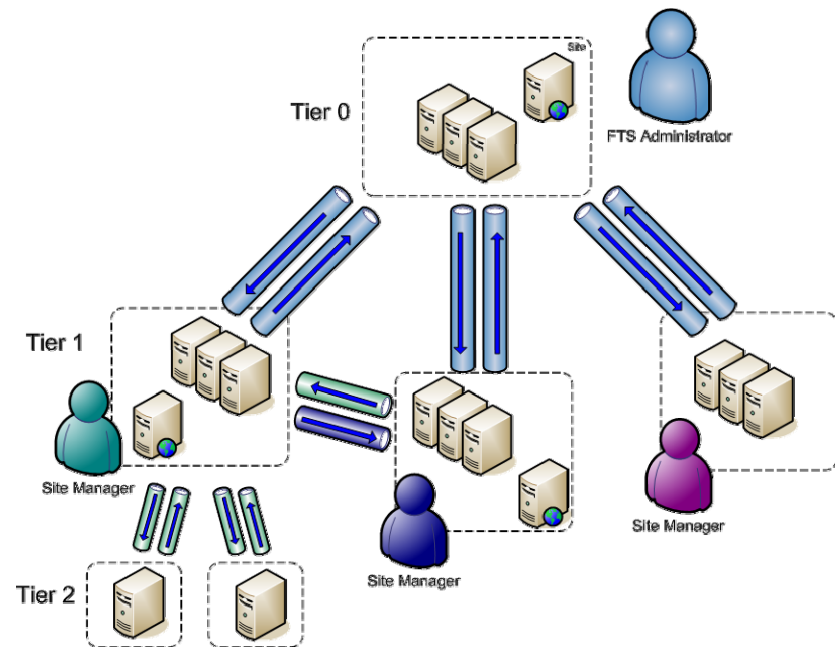


- The LFC is accessible via
 - CLI, C API, Python interface, Perl interface
 - Supports sessions and bulk operations
 - Data Location Interface (DLI)
 - Web Service used for match making:
 - *given a GUID, returns physical file location*
- 46 instances are in use

- **Hide the complexity from the user**
 - Interact with information system, file catalogues and SRM storage elements
- **GFAL**
 - POSIX-like I/O functions (open(), read() etc.)
 - SRM abstraction layer
 - C, Python APIs and CLI
- **lcg_util**
 - Cover most common use cases (file creation, registration, replication, deletion etc.)
 - C, Python APIs and CLI

- **Reliable File Transfer Service**

- Bulk data transfer between SRM compliant SEs (batch system for file transfers)
- Multi-VO service to balance network/SE utilization
- Prevents overloading network/SE resources
- Service monitoring and statistics



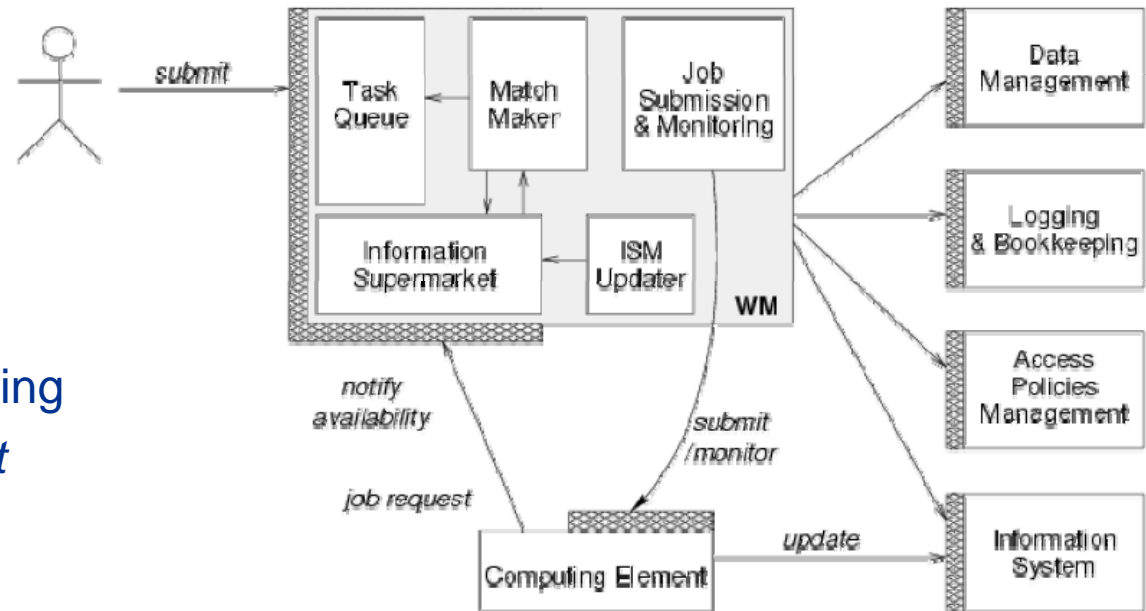
Workload Management System

- **WMS: Resource brokering, workflow management, I/O data management**
 - **Web Service interface: WMPProxy**
 - Task Queue: keep non matched jobs
 - Information SuperMarket: optimized cache of information system
 - Match Maker: assigns jobs to resources according to user requirements (possibly including data location)
 - Job submission & monitoring

→ **Condor-G**

→ **ICE (to CREAM)**

- External interactions:
 - Information System
 - Data Catalogs
 - Logging&Bookkeeping
 - *Policy Management systems*



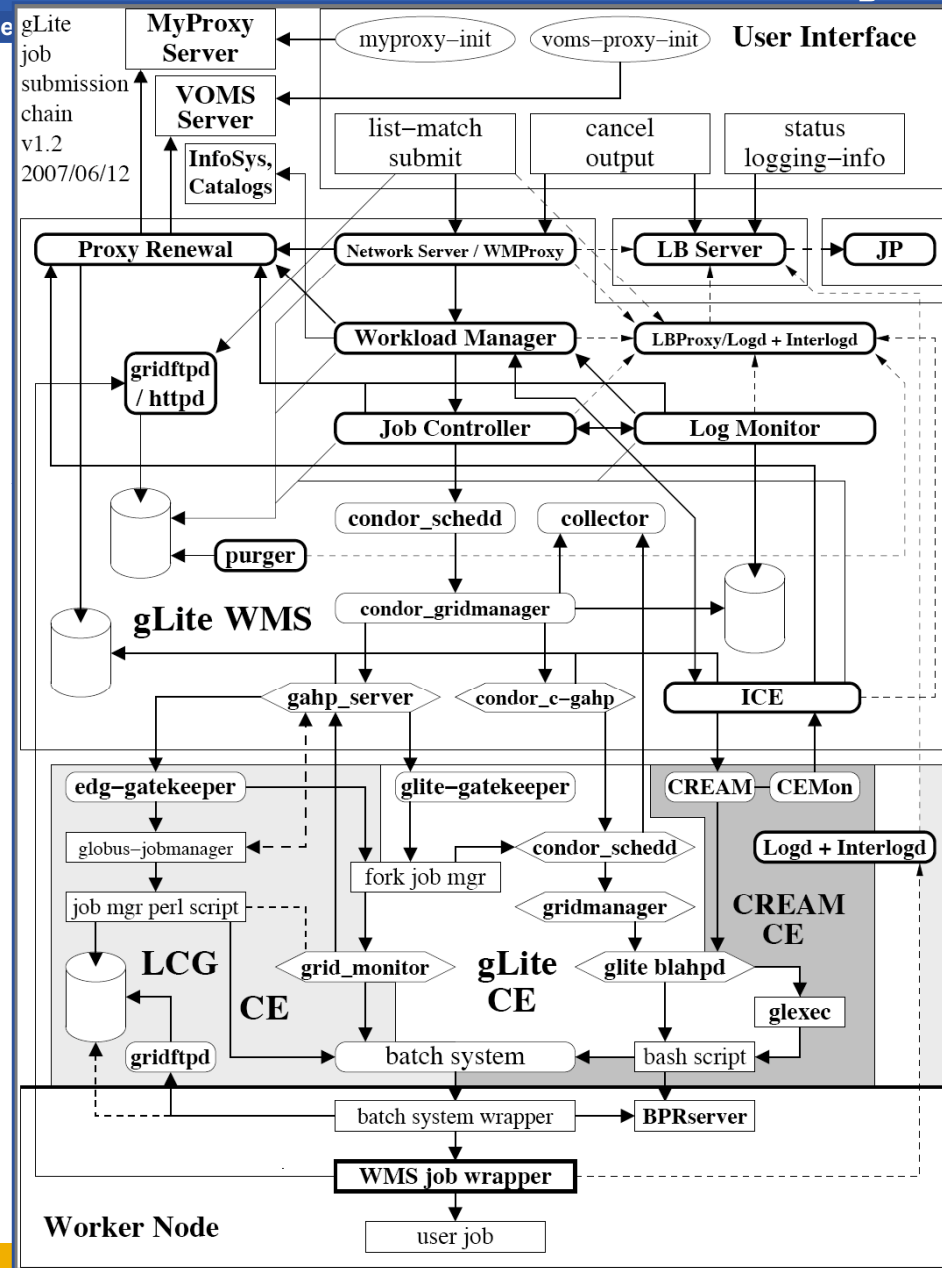
- **Describes (aggregates of) jobs and their characteristics and constraints**
- **Based on the Condor ClassAds**
- **Accepted by WMS and CREAM**

Example:

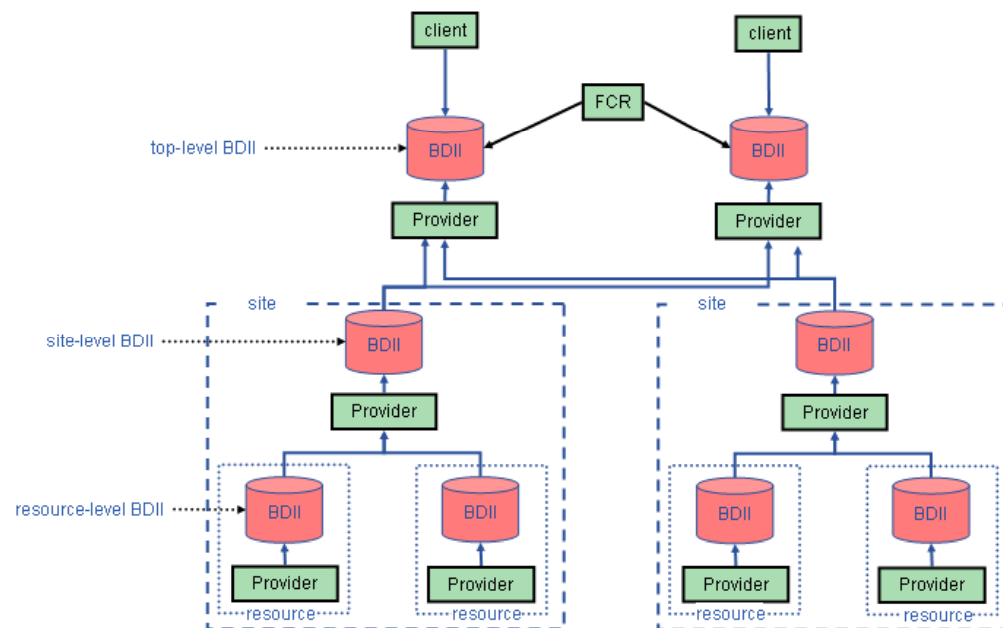
```
Executable = "test.sh";
Arguments = "fileA";
StdOutput = "std.out";
StdError = "std.err";
InputSandbox = {"test.sh", "fileA"};
OutputSandbox = {"std.out", "std.err"};
Requirements = Member("VO-dteam-SW-v2-01", other.GlueHostApplicationSoftwareRunTimeEnvironment);
```


What it really does

- Thanks to Maarten we know:
 - Simplified view



- **Berkeley Database Information Index (BDII)**
 - Uses a standard OpenLDAP server as supplied by the OS
 - With the Berkeley database backend
 - Updated by a Python process
 - Configuration file containing LDAP URLs for the sites
 - Use ldapsearch command used as it is stable

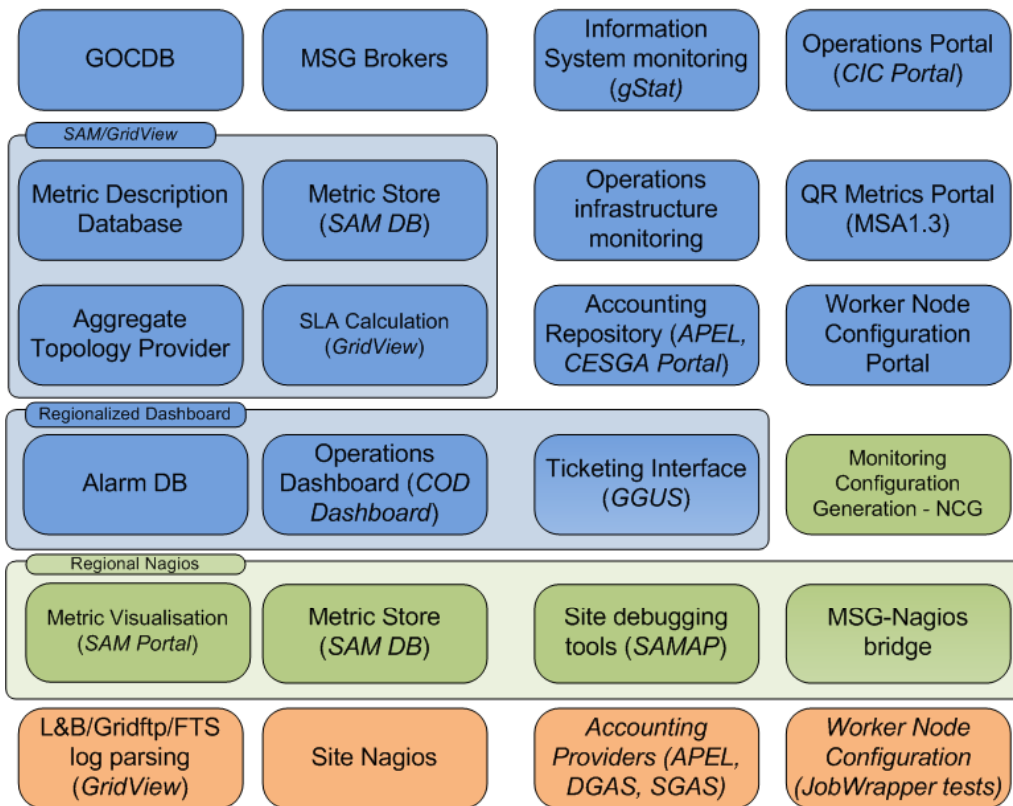


- **Generic Information Provider (GIP)**
 - Framework for information providers (plugins)
 - Populate BDII with information and also read information from other BDII (resource, site, top level hierarchy)
 - Get information and format it as LDIF
- **Freedom of Choice for Resources (FCR)**
 - FCR portal has list of services available to a VO
 - VO manager can black list sites
 - Information is propagated to VO specific top level BDII
- **GLUE Schema**
 - Schema that describes grid resources (computing, data, storage, services)
 - Collaborative effort between different grid projects, now within OGF

- **Logging and Bookkeeping (LB)**
 - Tracks jobs in terms of events (submission, starting execution etc.) gathered from the WMS and CE.
 - Events are stored and can be queried.
- **HYDRA**
 - Encrypted storage on SEs
 - Encryption keys are split and stored in >2 keystores
- **AMGA**
 - Metadata catalogue
 - Metadata Usually lives in relational databases
 - Why not accessing DBs directly on the Grid? Possible but
 - *Authentication (VOMS)*
 - *Connection pooling*
 - *Data replication*

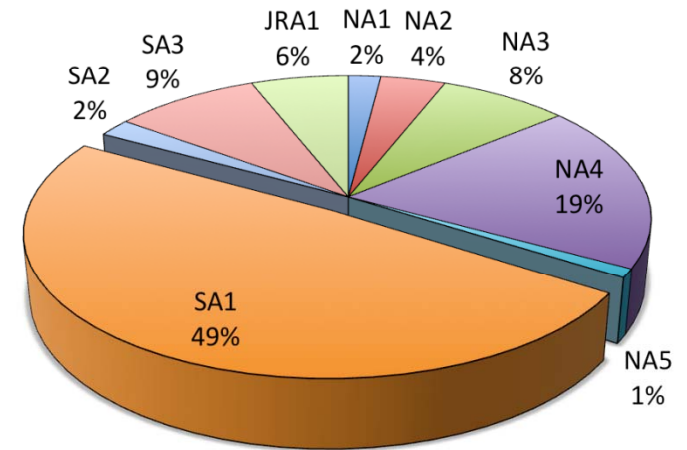
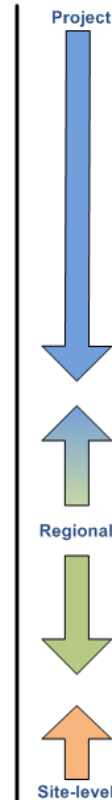
- Not part of gLite but indispensable for running the grid

Components in multi-level monitoring by deployment location*



Deployment location : ■ : Site-level ■ : Regional ■ : Project V1.7, James Casey, 18th Feb 2009

* Items in (..) refer to similar components that exist and are deployed centrally at this moment.



- **Middleware developed by geographically distributed teams (mostly at research institutes and Universities).**
- **A team focuses on a particular service.**
- **Teams are quite independent**
 - Coding conventions
 - Documentation
 - Naming conventions
 - ...
- **Currently ~ 20 FTEs (we are in EGEE III, manpower was more than double in the previous phase).**
- **More than 80 people in 12 different institutions involved.**

- **The project has a technical director but no single architect.**
- **Decisions are being taken in a collaborative, consensus based process.**
 - Bi-weekly phone conference to discuss short term priorities (bug fixing etc.).
 - Monthly Technical Management Board to discuss more strategic issues (new features etc.). Takes also input from the user community.
 - 2-3 all hands meetings per year.

- Distributed under an open source license.
- Main platform is Scientific Linux (recompiled RH EL).
- Many 3rd party dependencies (tomcat, log4*,gSOAP , ldap etc.). Pulled in from public repositories

Total Physical Source Lines of Code (SLOC)

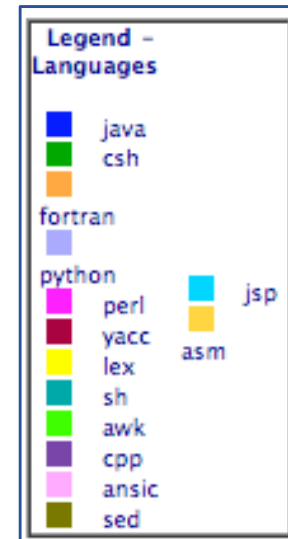
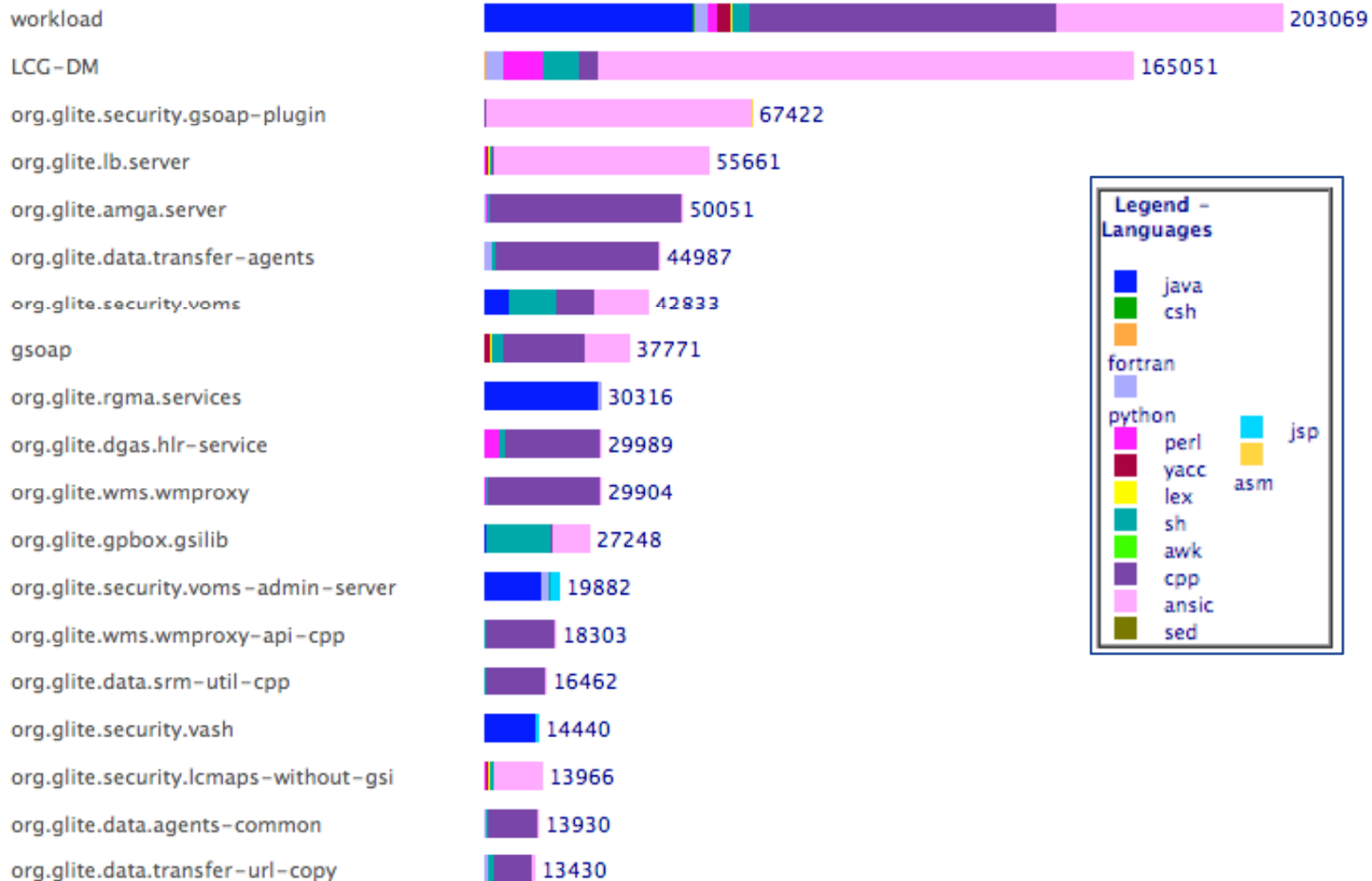
SLOC = 1622714

Total SLOC grouped by language (dominant language first)

Language	Total SLOC
ansic	578598 (35%)
cpp	491801 (30%)
java	251382 (15%)
sh	191798 (11%)
python	54510 (3%)
perl	39258 (2%)
yacc	7445 (0%)
jsp	4444 (0%)
lex	2274 (0%)
cs	701 (0%)
awk	307 (0%)
fortran	124 (0%)
sed	68 (0%)
asm	4 (0%)



SLOC by language for all modules



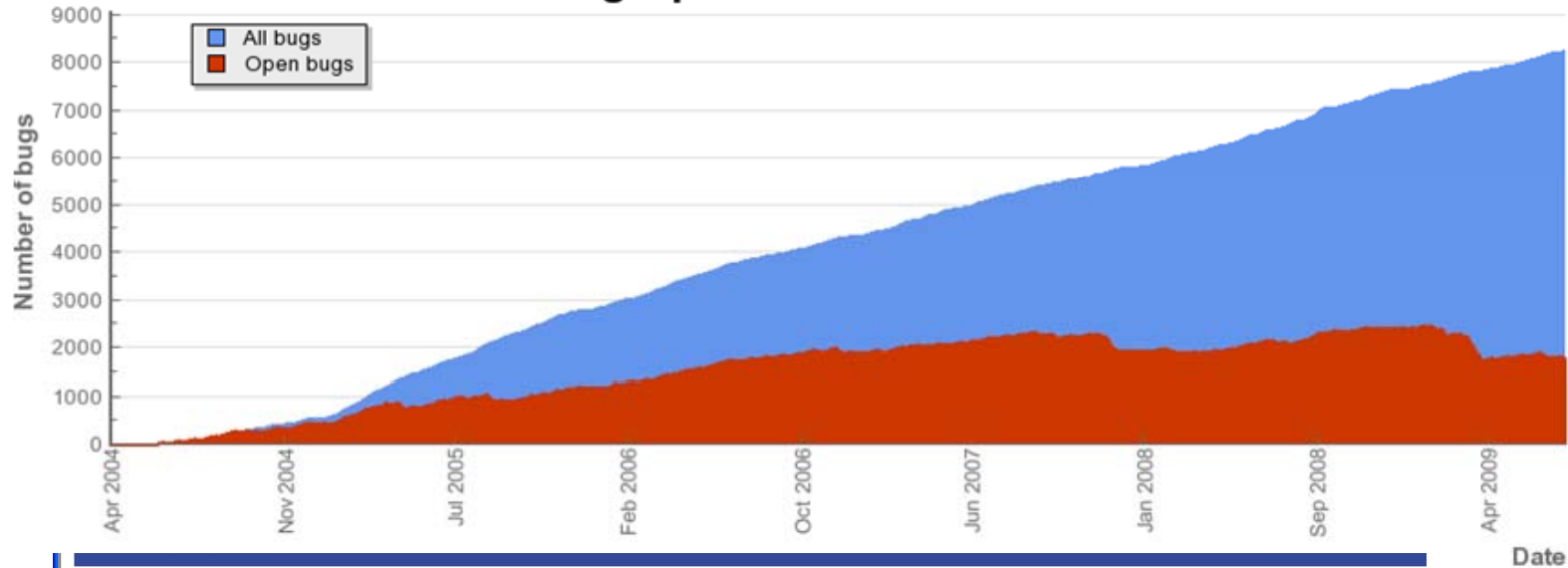


Enabling Grids for E-science

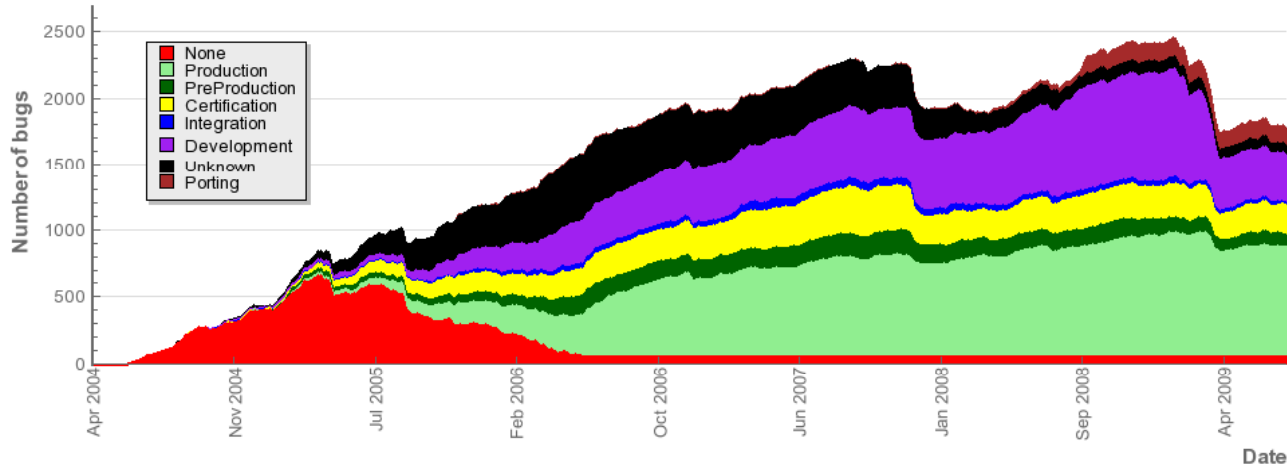
gLite code details

org.glite.security.obsolete	1342	org.glite.security.encryptid-storage-script	3633	org.glite.wms.common	12982
org.glite.security.cmaps-plugins-afs	1284	tdp-mgrimap	3602	fcgi	12879
org.glite.security.voms-mysql	1221	org.glite.gma.servicetool	3552	org.glite.wms.ice	12641
org.glite.service-discovery-api-c	1209	org.glite.gma.gin	3473	org.glite.data.transfer-fts	12578
org.glite.e2emontliperf	1191	org.glite.security.proxyrenewal	3458	org.glite.wms.jobsubmission	12209
org.glite.security.cmaps-plugins-gums-executable	1179	org.glite.security.glxcc	3445	org.glite.jdl-api-cpp	12067
org.glite.security.cas-plugins-check-executable	1179	org.glite.deployment.config	3395	org.glite.wms.client	12039
org.glite.security.cmaps-nfs-parser	1159	org.glite.jp.index	3339	org.glite.security.kmaps	11739
org.glite.gpobox.aajava	1144	org.glite.wms-utils.ds	3332	org.gridite.core	11393
org.glite.data.delegation-api-c	1122	org.glite.gma.api-python	3311	org.glite.ce.cream	11156
org.glite.dgas.pa-clients	1116	org.glite.ce.monitor-api-java	3272	org.glite.wms.manager	10708
org.glite.deployment.voms-server	1108	org.glite.security.auth2-framework-java	3135	org.glite.lib.client	10547
org.glite.security.gan-c	1073	org.glite.data.transfer-scripts	3101	org.glite.lib.common	10346
org.glite.wms-utils.exception	1072	org.glite.security.trustmanager	3074	org.glite.er.hahp	10110
org.glite.e2emontliperf	1017	org.glite.security.cmaps-plugins-jborep	2990	org.glite.gma.api-cpp	10017
org.glite.security.ad-parser	1005	org.glite.wms-ii-wrap-python	2964	org.glite.dgas.hlr-clients	10001
lgs-mon-tidout	985	org.glite.wms.brokerinfo	2910	org.glite.gpobox.pcp	9985
org.glite.data.dpm-httpd-col	980	org.glite.amga.api-java	2713	tdy-gluflur-client	9940
org.glite.service-discovery.gma-java	977	org.glite.wms.wmproxy-api-python	2702	org.glite.security.encrypted-storage-csp	9820
org.glite.lib.sshver-bones	906	org.glite.security.delegation-java	2692	org.glite.data.gfal	9708
org.glite.wms.configuration	875	org.glite.security.csi-gsmap	2608	lgs-sam-client	9113
org.glite.ce.job-plugin	758	org.glite.security.lcas-plugins-basic	2552	org.glite.sics.ui	8614
org.glite.yaim.clients	749	org.glite.data.hydra-service	2508	org.glite.security.gatekeeper	8074
org.glite.ce.csg-ce-plugin	703	org.glite.jdl.api-java	2408	org.glite.wms.thirdparty-bypass	7846
org.glite.e2emontliperf	695	org.glite.ce.cream-api-java	2426	org.glite.ce.monitor	7272
org.glite.data.srm-ii-rperl	675	org.glite.data.transfer-proxyrenewal	2287	org.glite.ce.cream-cli	6845
org.glite.lib.sshver-common	668	org.glite.service-discovery.file-c	2274	org.glite.e2emontliperf-update	6680
org.glite.security.cas-interface	659	org.glite.data.srm-cli	2285	org.glite.gma.api-c	6637
org.glite.wms-ui.configuration	654	org.glite.sics.common	2197	org.glite.wms.helper	6388
org.glite.apel.sf	633	org.glite.security.cmaps-plugins-verify	2155	org.glite.gpobox.server	6380
org.glite.lib.utils	629	org.glite.security.voms-secure	2132	org.glite.dgas.common	6359
org.glite.deployment.security-utils	601	org.glite.service-discovery.html-c	2016	org.glite.data.hydra-cli	6309
org.glite.apel.connector	599	org.glite.gma.command-line	2004	org.glite.data.dpm-xrootd	6245
org.glite.security.voms-abi	594	org.glite.yaim.core	1989	org.glite.lib.logger	6213
org.glite.apel.aps	589	org.glite.wms.matchmaking	1938	org.glite.gpobox.admin	6158
org.glite.apel.sge	520	org.glite.service-discovery.gma-c	1913	org.glite.jp.common	6015
org.glite.security.delegation-service-java	518	org.glite.datautil-c	1893	org.glite.data.dfm-util	6012
org.glite.ce.plugin	517	org.glite.security.lcas-kmaps-oid-interface	1851	org.glite.sics.server	5774
glite-info-provider-service	411	org.glite.wms.purger	1836	org.glite.apel.sure	5711
org.glite.data.dpm-httpd-mosd_keyauth	389	org.glite.security.cmaps-plugins-voms-attr	1834	org.glite.wms-ui.dli-python	5545
org.glite.lib.wa-interfaces	362	org.glite.security.ssss	1788	org.glite.rgrna.stubs-sewlet-java	5464
org.glite.data.dpm-httpd-mosd_dmpout	347	org.glite.wms-utils.pbic	1719	lgs-sam-client-sensors	5456
org.glite.data.dpm-httpd-shell	334	org.glite.data.common	1699	org.glite.gma.server.sevlet	5049
lgs-mon-logfile-common	315	org.glite.data.srm-api-c	1663	org.glite.amga.api-python	4919
org.glite.data.dpm-httpd-service	293	org.glite.wms.broker	1657	org.glite.security.kmaps-plugins-basic	4692
lgs-ManagementTag	291	org.glite.data.delegation-cli	1644	org.glite.wms.wmproxy-api-java	4685
lgs-nags	289	org.glite.service-discovery.cdi	1632	org.glite.ce.cream-client-api-c	4616
lgs-mon-tools	282	org.glite.wms-libs.asd	1621	org.glite.data.config-service	4585
org.glite.data.transfer-interface	277	org.glite.wms.rtsconf_plugin	1577	org.glite.security.lcas-plugins-voms	4507
org.glite.security.voms.config	272	org.glite.data.srm2-api-c	1536	org.glite.gma.base	4486
org.glite.gma.flexible-archiver	268	org.glite.jp.client	1494	org.glite.security.lcas-plugins-voms	4440
org.glite.data.catalog-interface	258	org.glite.e2emontliperf	1485	org.glite.ce.monitor-client-api-c	4322
org.glite.data.build-common-cpp	221	org.glite.wms.brokerinfo.access	1419	org.glite.security.kmaps-interface-api-without-gsi	4321
cleanup-grid-accounts	215	DPM-DSI	1420	org.glite.security.lcas	3951
lgs-mon-wn	207	org.glite.gma.api-java	1415	org.glite.security.util-java	3804
lgs-dgas-tools	195	org.glite.lib.proxy	1411	org.glite.dgas.pa-service	3683
org.glite.gma.gite-archiver	186	org.glite.lib.client-interface	1375	org.glite.ce.common-java	1644
lgs-explorer-rtmupdir	152	org.glite.data.transfer-api-perf	1374		
lcs-python-rtmuploader	142	org.glite.data.transfer-api-c	1344		
lgs-mon-gridftp	127	org.glite.apel.aubliher			

Bug Open/Closed over time



Bugs Detection Area over Time



- **Several releases per month**
- **Release model: pull model (rpm repository, tarballs to download).**
 - Sites pick up updates if they like to.
 - Multiple versions of services are in production.
 - Retirement of old versions is a lengthy process.
- **Release model: phased. We release updates to individual node types when needed, no big bang.**

- **EGEE project will terminate in April 2010**
- **European Grid Initiative (EGI)**
 - To establish a sustainable grid infrastructure in Europe
 - <http://web.eu-egi.eu/>
 - Handle middleware maintenance, integration, testing and release within EGI
 - To produce the Unified Middleware Distribution (UMD)
 - Convergence of ARC, gLite and Unicore
 - <http://knowledge.eu-egi.eu/knowledge/index.php/UMD>