Enabling Grids for E-sciencE

The gLite middleware

Lightweight Middleware for Grid Computing

Claudio Grandi – JRA1 Activity Manager

INFN and CERN

EGEE'06 Conference

Geneva, 25-29 September 2006

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

Information Society and Media

Outline

- Background and approach adopted
- Software process
- Status and highlights
- Plans
- Summary

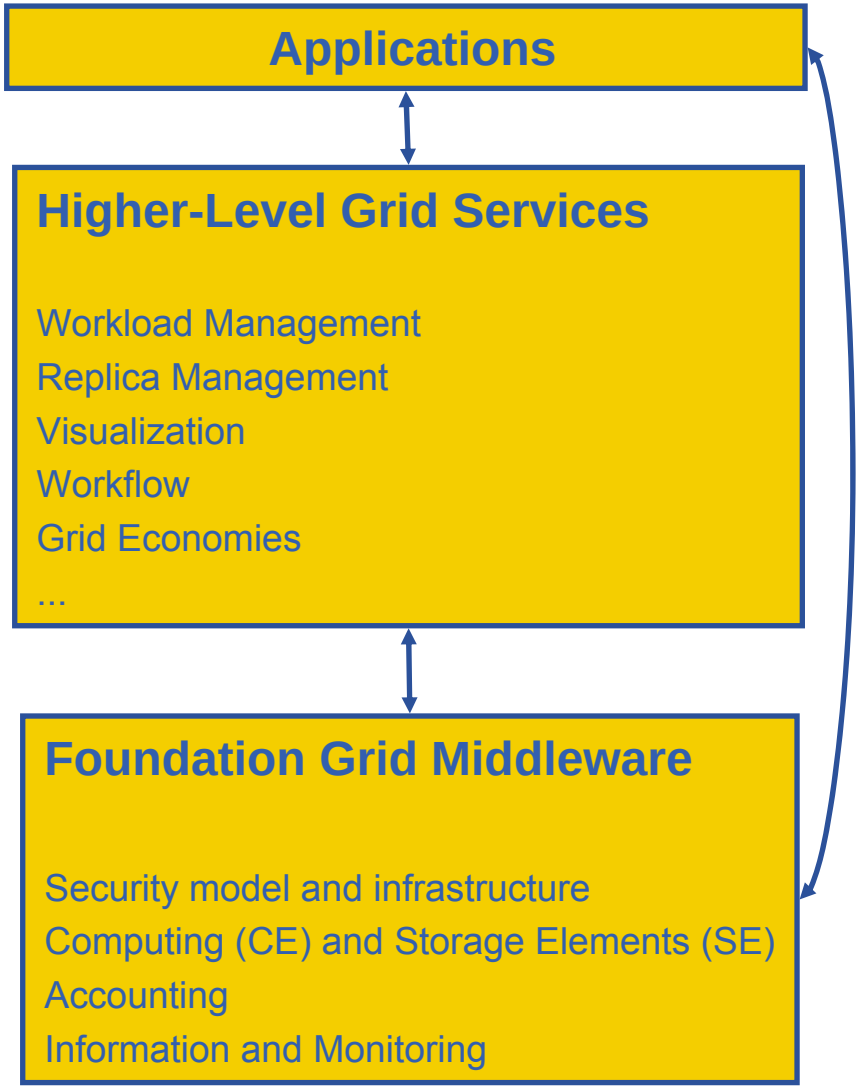
Background & Approach

- gLite
  - Exploit experience and existing components from VDT (Condor, Globus), EDG/LCG, and others
    - gLite is a distribution that combines components from many different providers!
  - Develop a lightweight stack of generic middleware useful to EGEE applications
    - Pluggable components
    - Follow SOA approach, WS-I compliant where possible
  - Focus is on re-engineering and hardening
  - Business friendly open source license
    - Plan to switch to Apache-2

Service Oriented Architecture

- **gLite follows a Service Oriented Architecture**
  - Facilitate interoperability among Grid services
  - Allow easier compliance with upcoming standards
  - The services work together in a concerted way but can also be deployed and used independently, allowing their exploitation in different contexts
- **Services communicate through the exchange of messages**
  - Slowly moving to WS-* interfaces
  - Still missing a real standard. Many WS-* specifications
  - Activity inside OGF-GIN

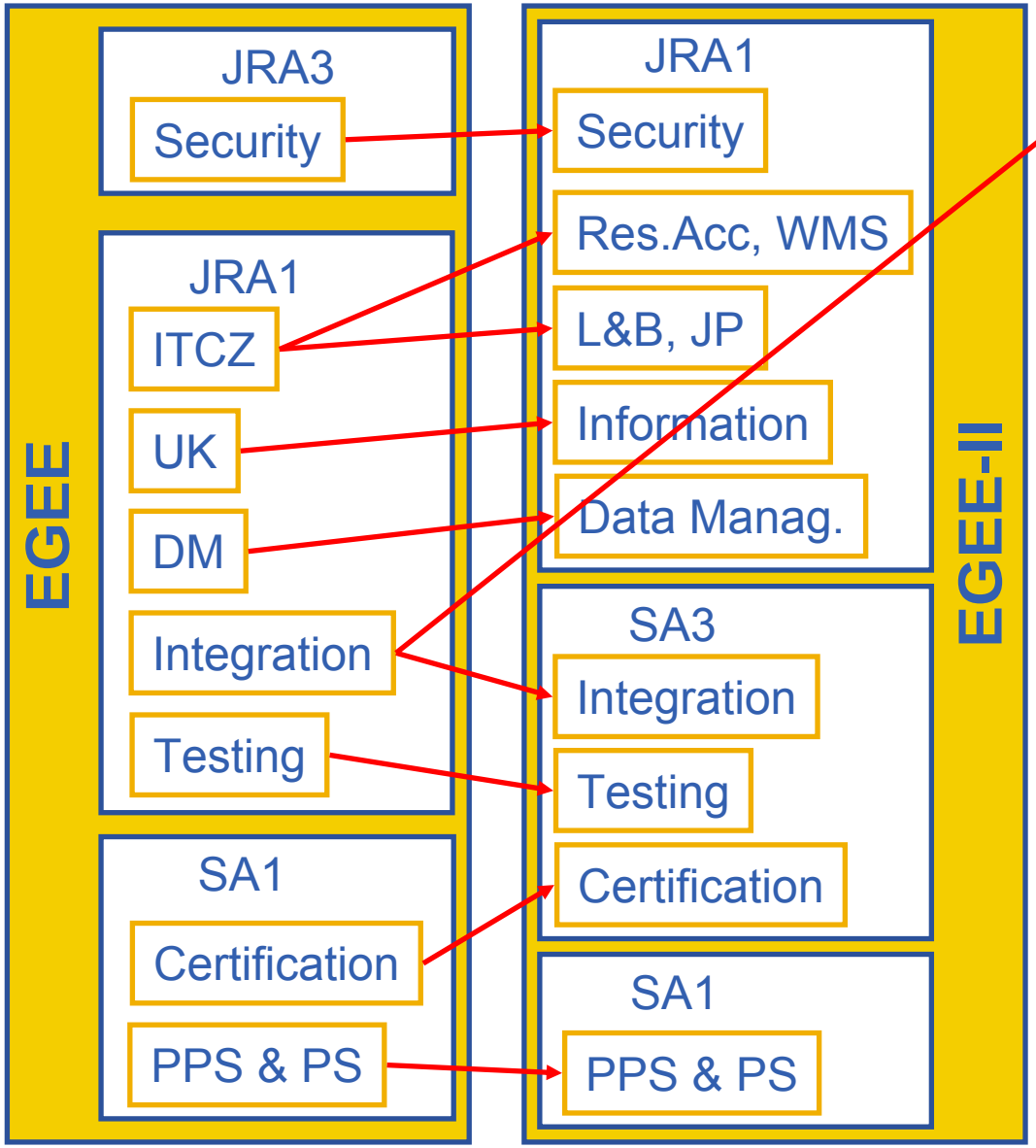
Middleware structure



- **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 development reorganization



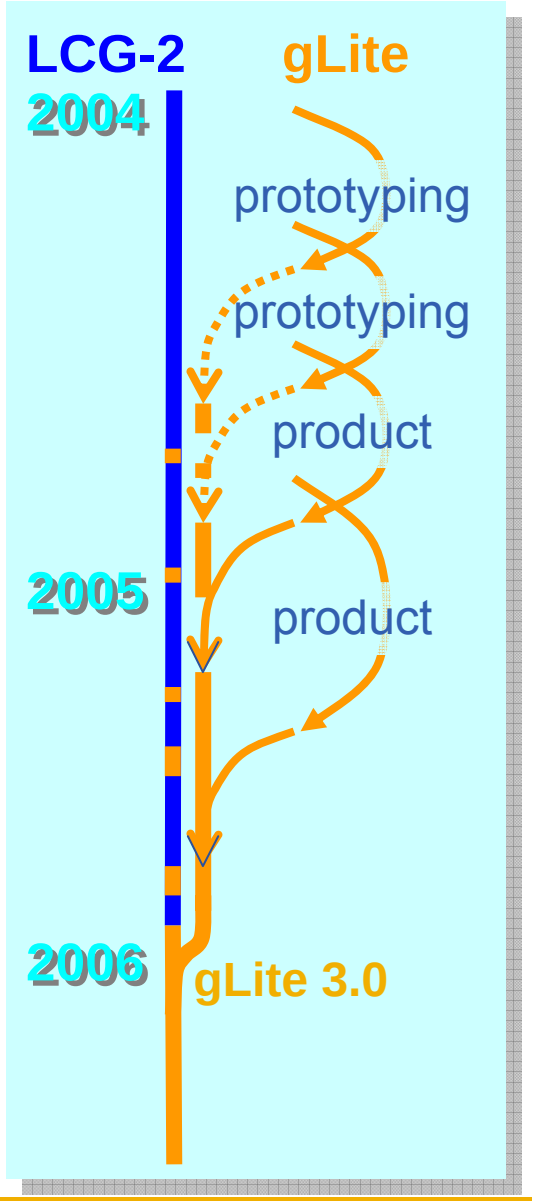
ETICS

Building & Testing Tools

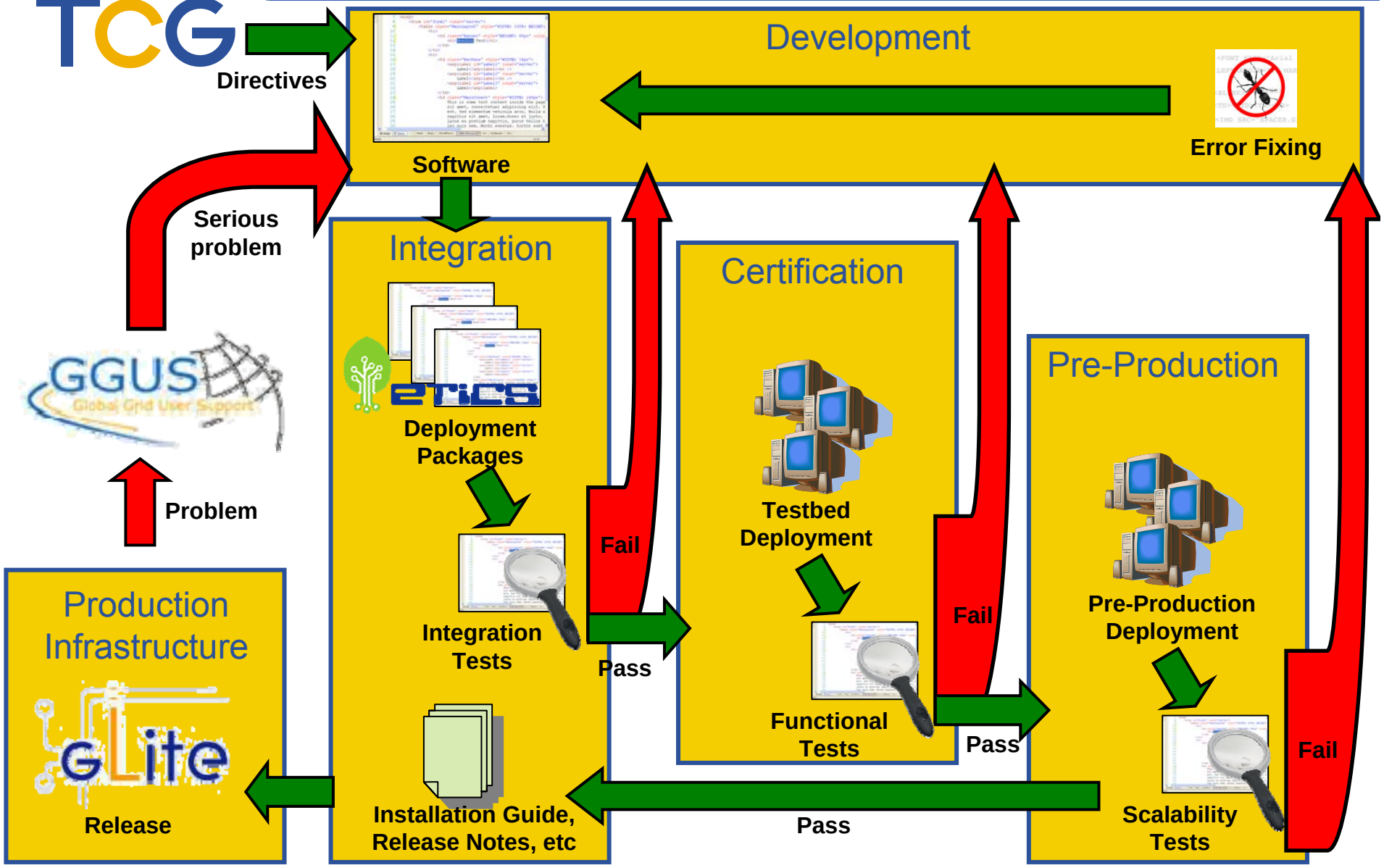
- **JRA1 is responsible for developing the middleware**
- **SA3 is responsible for integration, testing and certification, i.e. to produce the release**
- **SA1 runs the PPS and PS systems**
- **ETICS provides the tools for building and testing used by JRA1 and SA3**

The release of gLite 3.0

- **Convergence of LCG 2.7.0 and gLite 1.5.0 in spring 2006**
  - Continuity on the production infrastructure ensured usability by applications
  - Initial focus on the new Job Management
    - Thorough testing and optimization together with the applications
- **Migration to the ETICS build system**
  - ETICS project started in January
- **Reorganization of the work according to the new process**
  - EGEE Technical Coordination Group and Task Forces
  - Start of the EGEE SA3 Activity for integration and certification
  - "Continuous release process"
    - No big-bang releases!

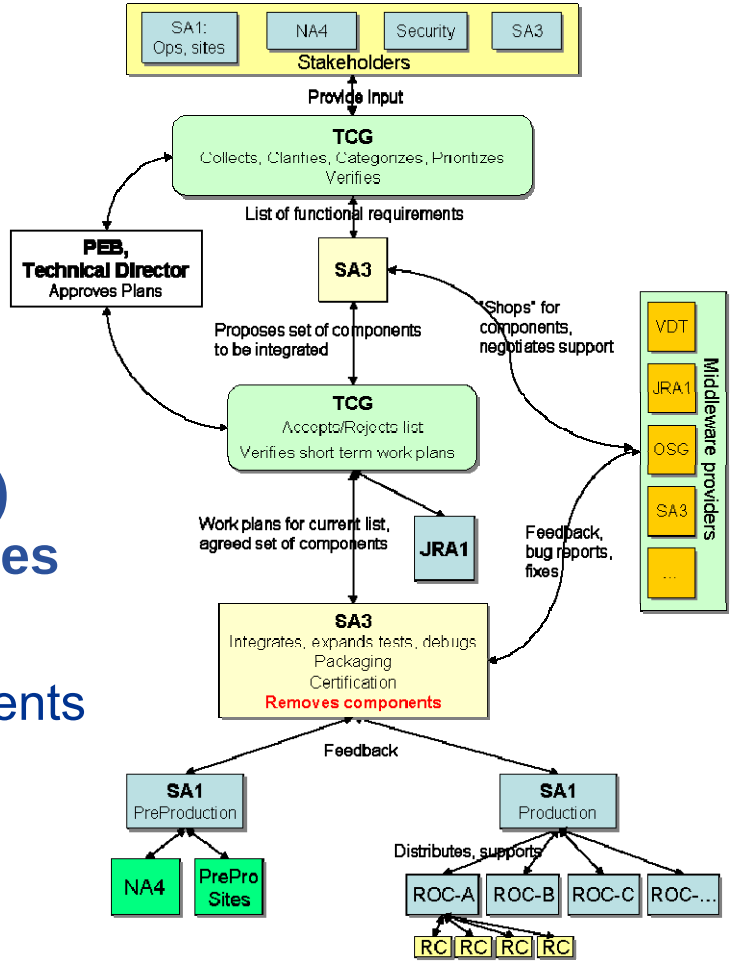


gLite Software Process



TCG and EMT

- **The EGEE Technical Coordination Group (TCG) defines the priorities for middleware development and certification**
  - Members from LHC experiments and other EGEE-NA4 applications, and form EGEE Technical activities
  - Collects requirements from the applications
    - Started from the LCG requirement list
    - Recently added JSPG and sites requests
  - Prioritizes the requirements
  - Approves the JRA1 and SA3 work plans
    - Focus on the short term
- **The Engineering Management Team (EMT) coordinates the production of gLite releases**
  - Members from SA3, JRA1, SA1 and VDT
  - Decides what and when to release components and patches
  - Follows critical bugs fixing individually
  - Works according to TCG directives



Main focus for the developers

- Give support on the production infrastructure (GGUS, 2nd line support)
- Fix bugs found on the production software

The above are estimated to take 50% of the resources!

- Support SL(C)4 and 64bit architectures (x86-64 first)
- Participate to Task Forces together with applications and site experts
- Improve robustness and usability (efficiency, error reporting, ...)
- Address requests for functionality improvements from users, site administrators, etc... (through the TCG)
- Improve adherence to international standards and interoperability with other infrastructures
- Deploy and expose to users new components on the preview test-bed

Preview test-bed

- **The SA3 integration and certification teams are focused on providing code for the production infrastructure**
  - Strong control over what is accepted, but slow process for the certification of the new components and of the improvements
- **JRA1 requested a test-bed to expose to users those components not yet considered for certification**
  - To get feedback from users and site managers
  - TCG and PEB acknowledged that this is needed, but no resources were foreseen for this activity in the EGEE-II proposal

The JRA1 partners which have also strong commitments in SA1 have been requested to provide resources (machines and manpower) for this activity <u>without compromising their commitment in SA1</u>

➔ **At present, only INFN and CESNET have committed significant resources**

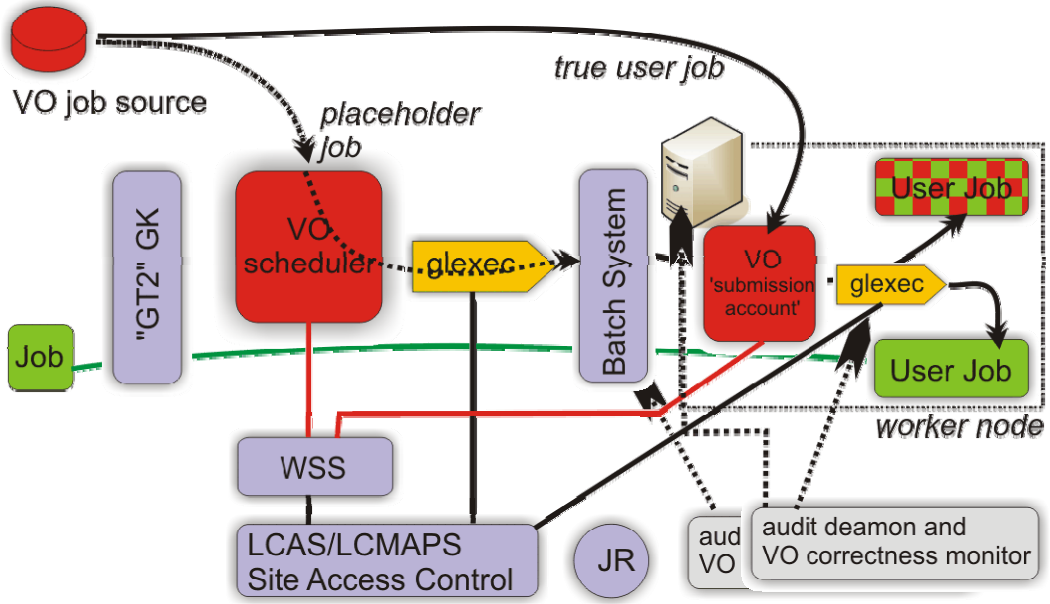
We are working to "stretch" it up to the Nordic countries!

Summary of current activities

- **Security**
  - Enabling glexec on Worker Nodes
  - Address user and security policy requirements in VOMS, VOMSAdmin
  - Proxy renewal library repackaged without WMS dependencies
  - Shibboleth short-lived credential service and interaction with VOMS
- **Job Management**
  - Improvement in functionality and performance on WMS and LB
  - Preparation for the deployment of the DGAS accounting system
  - Development and test on the preview test-bed of the new components
    - ICE-CREAM, G-PBox including LCAS/LCMAPS plug-ins, Job Provenance
- **Data Management *(mainly from LCG)***
  - Adding support for SRM v2.2 in DPM, GFAL and FTS
  - Working on new Encrypted Data Storage based on GFAL/LFC
  - Improvements in LFC distributed service
  - FTS proxy renewal and delegation
- **Information**
  - Improvements in R-GMA
  - Development for GLUE 1.3 *(from LCG)*

Highlights: glexec

- **glexec is used by the gLite Computing Elements to change the local uid as function of the user identity (DN & FQAN)**
- **Several VOs submit 'pilot' jobs with a single identity for all of the VO**
  - The pilot job gets user jobs in 'some' way and executes them with the placeholder's identity
  - The site does not 'see' the original submitter
- **Allowing the VO pilot job to run glexec on the WN could 'recover' the user identity and isolate the user job from the pilot job**



- **Issue: the sites do not like to run *sudo* code on the Worker Nodes**
- **A possibility is to run glexec in "null" mode:**
  - log the uid-change request but do not do it
  - The original user identity is recovered but there is no isolation of user and pilot

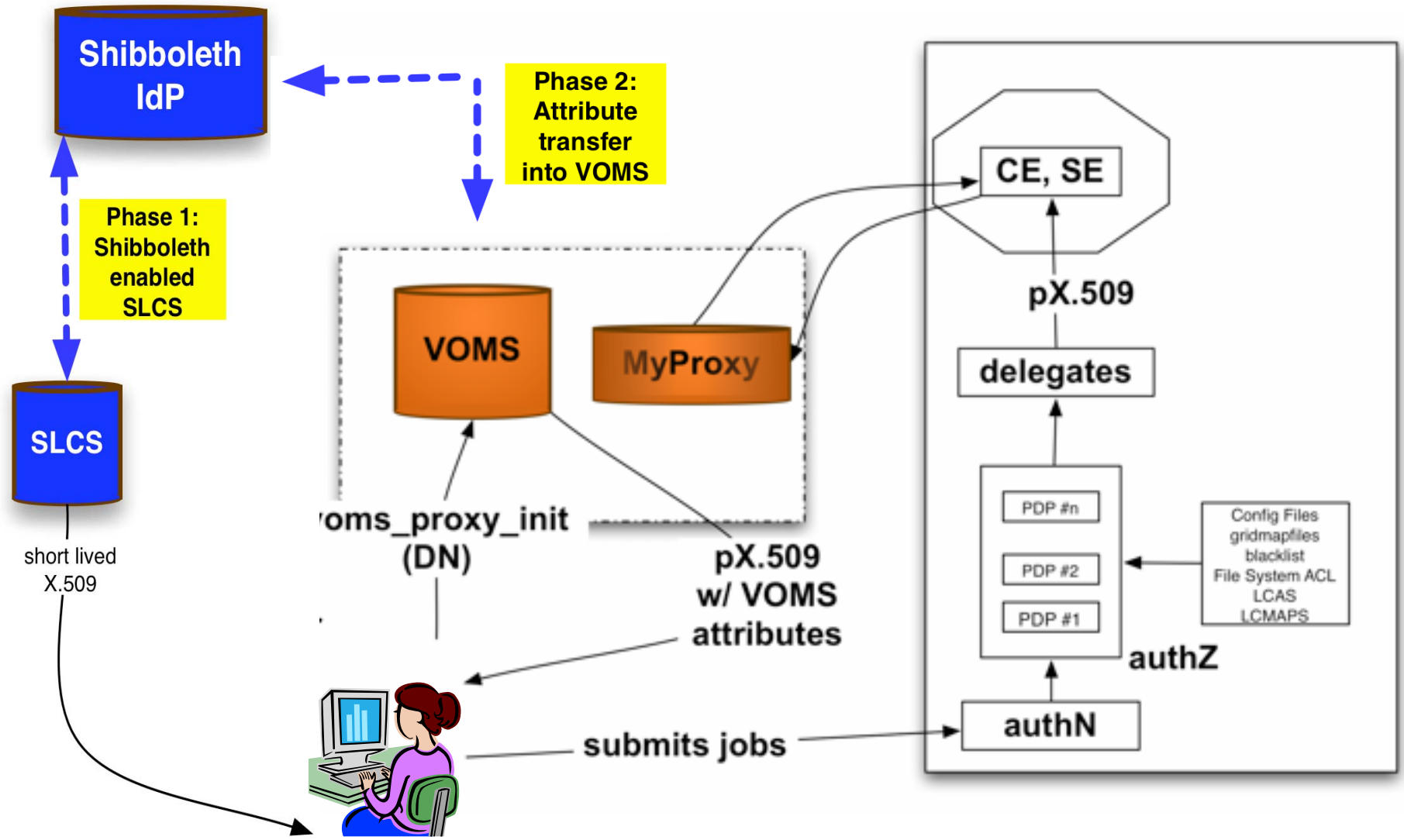
Highlights: Shibboleth

- **Shibboleth**
  - Federation of campus infrastructures
  - Developed by Internet2
  - Allows Single Sign On for web-based resources
  - Based on SAML (Security Assertion Markup Language)
- SWITCH
  - Manages an Authentication and Authorization Infrastructure (AAI) based on Shibboleth with about 160'000 users of the Swiss higher education sector
    - Activity started in 2002; in production since last summer
    - about 12'000 use SWITCHaai on a regular basis
- **Interoperability with gLite**
  - Specific for EGEE-2 infrastructure
    - NO replacement for X.509, VOMS, ...
  - Home institution of the user is the Identity Provider
  - Attributes both from home institution and the VO

Shibboleth: interoperability work plan

- **Three phases**
  - Two initial, shorter phases
    - Start small and hook up Shibboleth AAI to a gLite grid with minimum amount of changes (in particular no change at the CE)
    - Build up knowledge and expertise
    - April 06 → autumn/winter
  - A longer third phase
    - SAML support at the resource end
    - Design during phase 1 and 2
    - Implementation in 2007

Shibboleth: integration phases 1 & 2



Highlights: Accounting

- **Collect usage records for all jobs at sites**
  - *Local and global jobId, uid, DN, VOMS FQAN, system usage (cpuTime, ...), ...*
  - Information from log files by BLAH (gLiteCE, Cream) and LCG-CE
- **The information is collected from sites using APEL**
  - Currently insecure storage and transfer of accounting records via R-GMA.
    - Working to add an authorization layer; encryption is in certification
  - DGAS already provides proper management of privacy (records signed and encrypted) but doesn't have a proper interface for data visualization

- **Issues:**
  - Need to converge on a single accounting collection tool
    - Process for the merge of APEL and DGAS already started
  - Sensors have to be provided for all batch system AND grid infrastructures
    - Working with OSG to factorize local and grid information collection
  - The Condor local batch system in the gLiteCE bypasses BLAH
    - Working with the Condor team to get the needed information
    - Producing the BLAH plug-ins for Condor
  - Accounting for jobs executed via a VO pilot-job
    - Probably only VO-based accounting will be provided by sites for these jobs
    - User accounting will be provided by the VO software

Highlights: Job Priorities

- **Applications ask for the possibility to diversify the access to fast/slow queues depending on the user role/group inside the VO**
- **GPBOX is a tool that provides the possibility to define, store and propagate fine-grained VO policies**
  - based on VOMS groups and roles
  - enforcement of policies at sites: sites may accept/reject policies
  - Not yet certified. Certification will start when requested by the TCG.
- **Current plans: test job prioritization without GPBOX:**
  - Map VOMS groups to batch system shares (via GIDs?)
  - Publish info on the share in the CE GLUE 1.2 schema (VOView)
    - The gLite WMS has been modified to support GLUE 1.2
  - WMS match-making depending on submitter VOMS certificate
    - But no ranking of resources based on priority offered yet
  - Settings are not dynamic (via e-mail or CE updates)
- **If GPBOX is needed for LHC, tests must start now!**
    - Will be tested on the preview test-bed

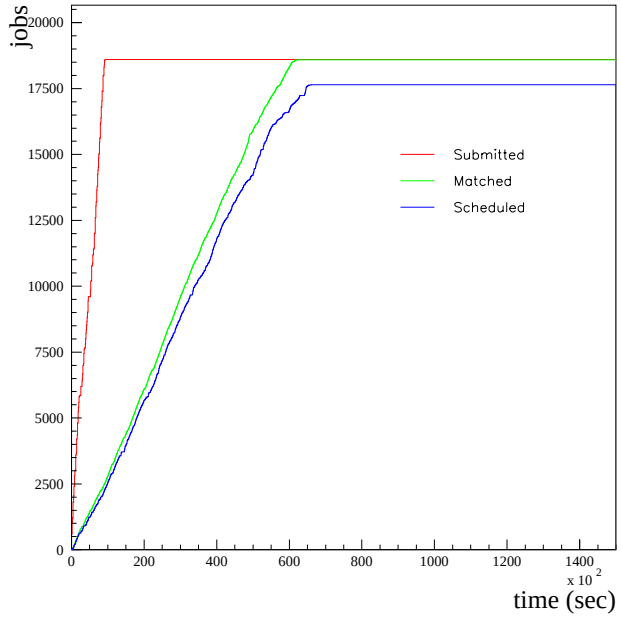
Highlights: gLite WMS and L&B

- **WMProxy: web interface to WMS**
  - decouples interaction with UI and internal procedures (logging to L&B, match-making, submission)
- **Support for compound jobs (Compound, Parametric, DAGs)**
  - One shot submission of a group of jobs
    - Submission time reduction (single call to WMProxy server)
    - Shared input sandboxes
    - Single Job Id to manage the group (single job ID still available)
- **Support for 'scattered' input/output sandboxes**
- **Support for *shallow resubmission***
  - Resubmission happens in case of failure only when the job didn't start

- **Issues:**
  - Needed fine tuning to work at the production scale
  - Difficulties in the management of DAGs
    - Will work to decouple Compound and Parametric jobs form DAGs
  - Implied a migration to Condor 6.7.19
    - Now need to test the new Condor also on the gLite-CE

Latest CMS Results

- **~20000 jobs submitted**
  - 3 parallel UIs
  - 33 Computing Elements
  - 200 jobs/collection
    - Bulk submission
- **Performances**
  - ~ 2.5 h to submit all jobs
    - 0.5 seconds/job
  - ~ 17 hours to transfer all jobs to a CE
    - 3 seconds/job
    - 26000 jobs/day
- **Job failures**
  - Negligible fraction of failures due to the gLite WMS
    - Either application errors or site problems

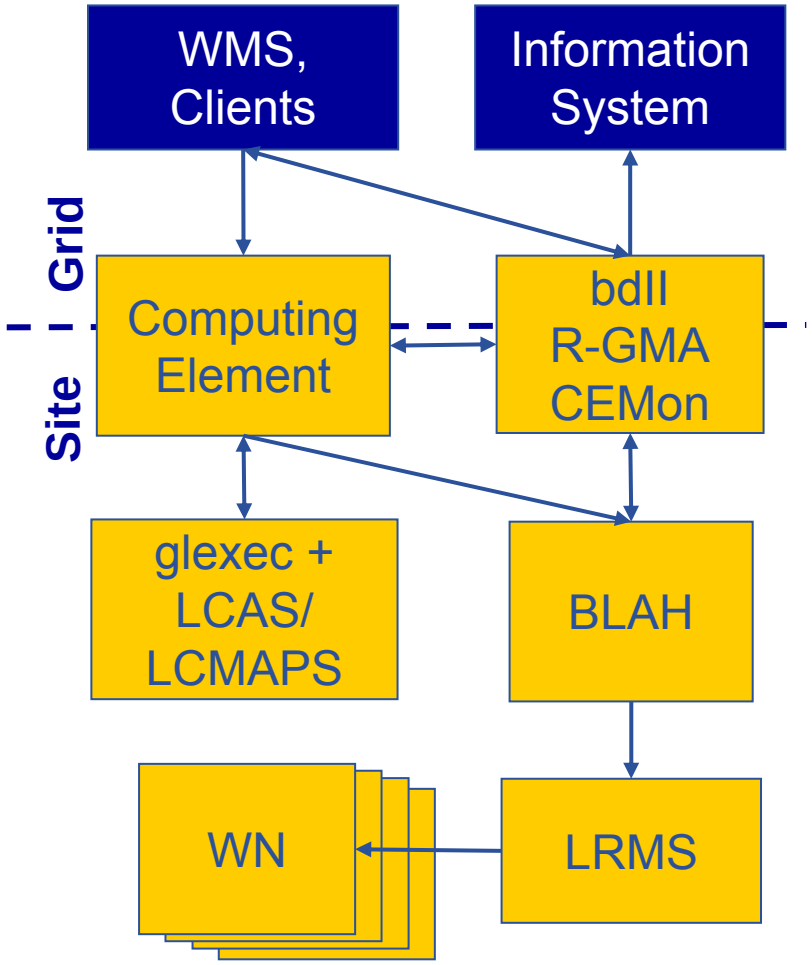


Failure reason	Job fraction (%)
Application error	28
Remote batch system	3.9
CRL expired	3.3
Worker Node problem	1.1
Gatekeeper down	0.2

By A.Sciabà

Highlights: Computing Element

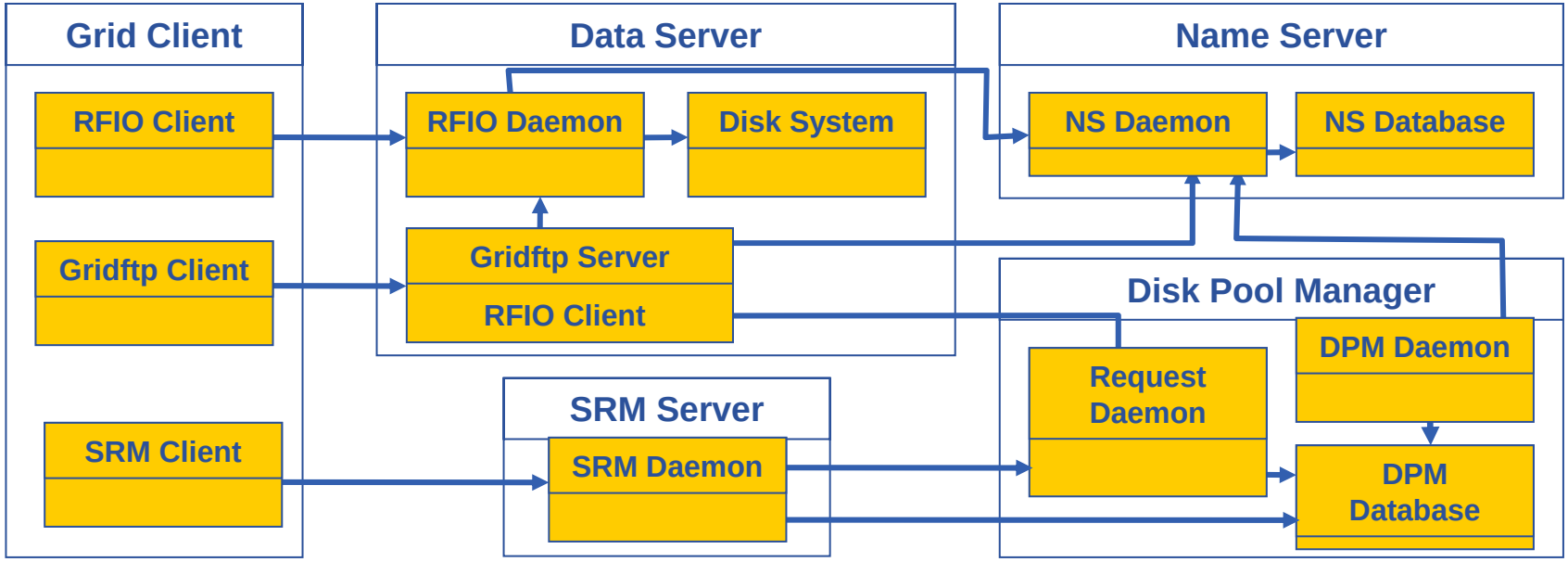
- **Three flavours available now:**
  - ➔ LCG-CE (GT2 GRAM)
    - in production now but will be phased-out by the end of the year
  - ➔ gLite-CE (GSI-enabled Condor-C)
    - already deployed but still needs thorough testing and tuning. Being done now
  - ➔ *CREAM (WS-I based interface)*
    - being deployed on the JRA1 preview test-bed now. After a first testing phase will be certified and deployed together with the gLite-CE
    - Our contribution to the OGF-BES group for a standard WS-I based CE interface



- **BLAH is the interface to the local resource manager (via plug-ins)**
  - CREAM and gLite-CE
  - *Information pass-through*: pass parameters to the LRMS to help job scheduling

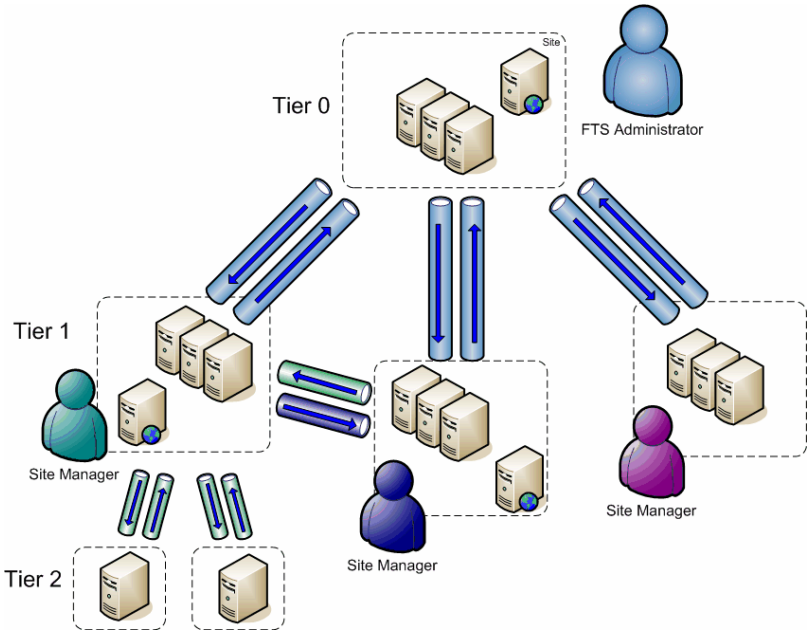
Highlights: Disk Pool Manager

- **Light-weight disk-based Storage Element**
  - Easy to install, configure, manage and to join or remove resources
  - Integrated security (authentication/authorization) based on VOMS groups and roles
    - All control and I/O services have security built-in: GSI or Kerberos 5
    - Problem of ACLs propagation during replication between SEs will be addressed in the first half of 2007
  - SRMv1 and SRMv2.1 interfaces. SRMv2.2 being added now



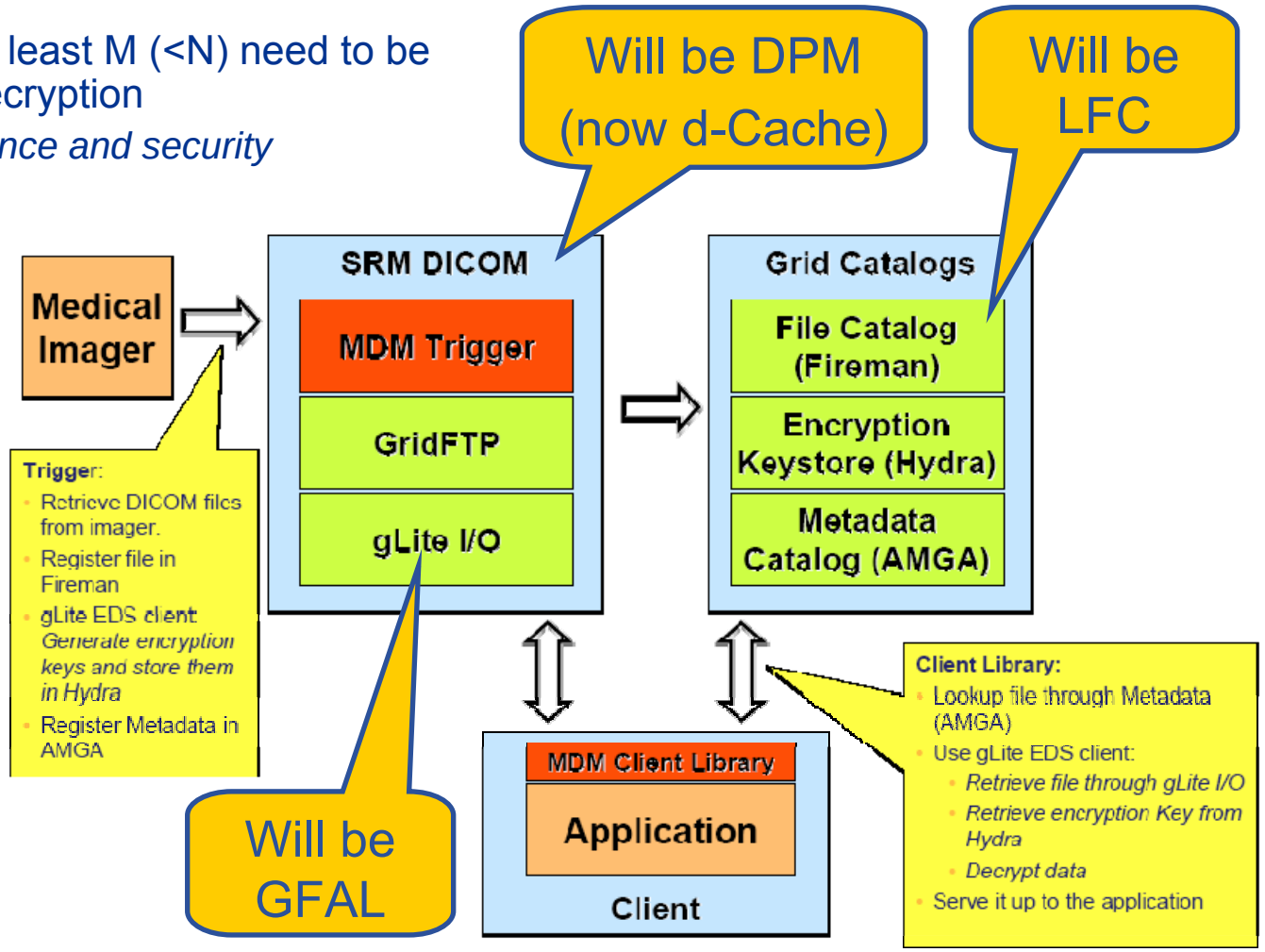
Highlights: FTS

- **Reliable and manageable File Transfer System for VOs**
- **Transfers are treated as jobs**
  - May be split onto multiple “channels”
  - Channels are point-to-point or “catch-all” (only one end fixed). More flexible channel definitions on the way...
- **New features that will be available in production soon:**
  - Cleaner error reporting and service monitoring interfaces
  - Proxy renewal and delegation
  - SRMv2.2 support
- **Longer term development:**
  - Optimized SRM interaction
    - split preparation from transfer
  - Better service manag. controls
  - Notification of finished jobs
  - Pre-staging tape support
  - Catalog & VO plug-ins framework
    - Allow catalog registration as part of transfer workflow



Highlights: EDS

- **Encrypted Data Storage**
  - encrypt and decrypt data on-the-fly
  - Key-store: Hydra
    - N instances: at least M (<N) need to be available for decryption
      - *fault tolerance and security*
  - Demonstrated with the SRM-DICOM demo at the EGEE Pisa Conf. (Oct'05)
  - Now porting to the deployed Data Management components (DPM, LFC, GFAL)



gLite development plans

- **Complete migration to VDT 1.3.X and support for SL(C)4 and 64-bit**
- **Complete migration to the ETICS build system**
- **Work according to work plans available at: https://twiki.cern.ch/twiki/bin/view/EGEE/EGEEgLiteWorkPlans**
- **In particular:**
  - Continue work on making all services VOMS-aware
    - Including job priorities
  - Improve error reporting and logging of services
  - Improve performances, in particular WMS and LB
  - Support for all batch systems in the production infrastructure on the CE
  - Use the *information pass-through* by BLAH to control job execution on CE
  - Complete support to SRM v2.2
  - Complete the new Encrypted Data Storage based on GFAL/LFC
  - Complete and test glexec on Worker Nodes
  - Standardization of usage records for accounting
- **Interoperation with other projects and adherence to standards**
- **Collaboration with EUChinaGrid on IPv6 compliance**

Summary

- **gLite 3.0 is an important milestone in EGEE program**
  - New components from gLite 1.X being deployed for the first time on the Production Infrastructure
    - Address requirements in terms of functionality and scalability
    - Components deployed for the first time need extensive testing!
  - New organization in EGEE II
    - New build and integration environment form ETICS
    - More controlled software process and certification
    - Development is client driven (TCG)
- **Development is continuing to provide increased robustness, usability and functionality**
- **Collaboration with other projects for interoperability and definition/adoption of international standards**

gLite

Lightweight Middleware for Grid Computing

www.glite.org