gLite middleware development

Claudio Grandi - JRA1 Activity Manager - INFN

LHCC review of LCG
CERN, 25 September 2006

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

Information Society and Media

Outline

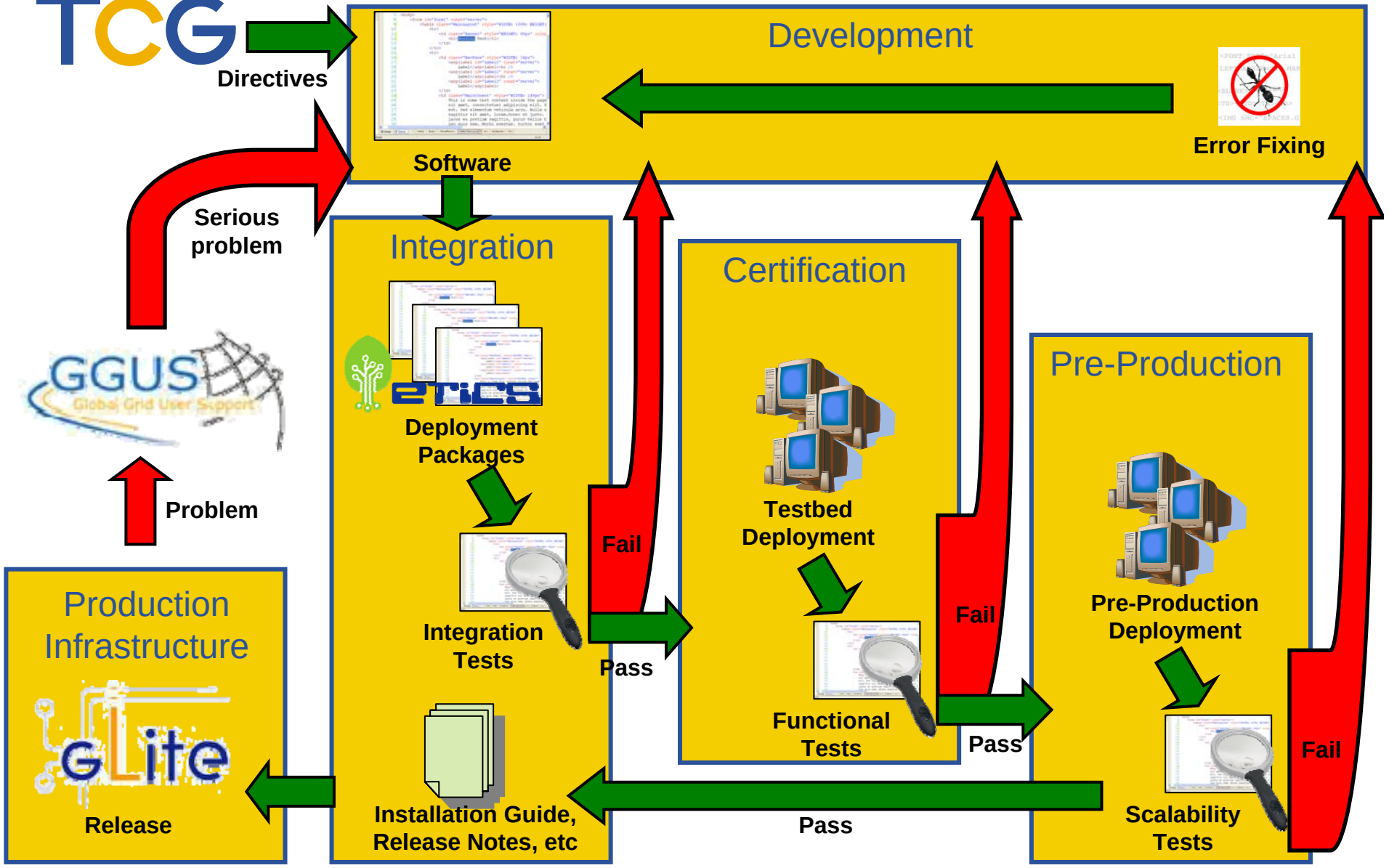
- Introduction
- The process and the role of the EGEE TCG
- Current activities
- Hot topics
- Plans for the future

Introduction

Main activities in 2006

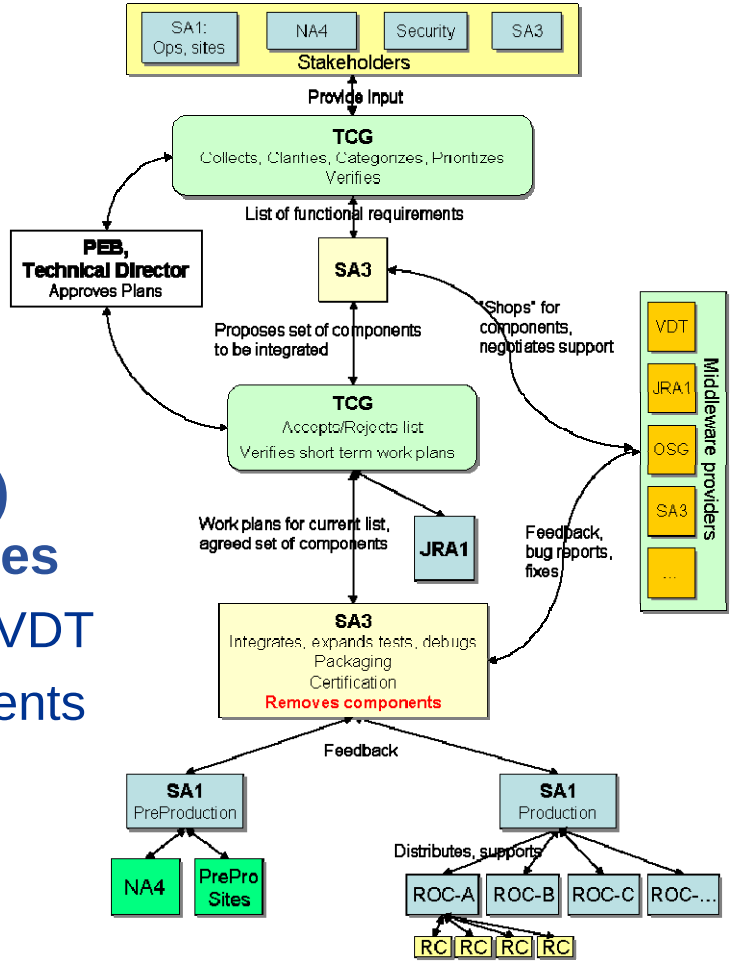
- **Started migration to the ETICS build system**
  - ETICS project started in January
- **Reorganization of the work according to the new process**
  - EGEE Technical Coordination Group and Task Forces
  - Start of the EGEE SA3 Activity for integration and certification
- **Convergence of gLite 1.5 and LCG 2.7.0 that was on the production infrastructure to a unique middleware stack**
  - Major effort
    - LCG 2.7.0 and gLite 1.5 were developed different environments
- **Release of gLite 3.0 (May)**
- **Tuning and patching of the new WMS**
  - On the Production Infrastructure, together with the experiments
- **Started porting to VDT 1.3.X (including GT4 pre-WS)**
  - Mandatory step to support Scientific Linux 4 and 64-bit

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 (PPS) 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!

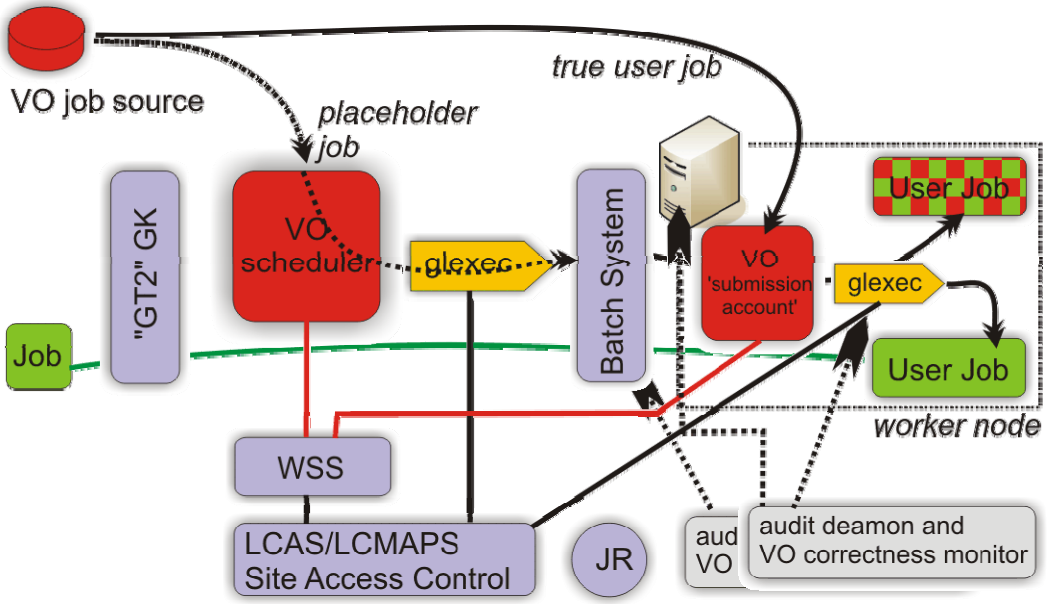
- Support SL(C)4 and 64bit architectures (x86-64 first)
- Participate to Task Forces together with applications and site experts
- Improve robustness and usability (efficiency, error reporting, ...)
- Address requests for functionality improvements from users, site administrators, etc... (through the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

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
  - Design of the Shibboleth-based short-lived credential service
- **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 middleware on the CE to change the local uid as function of the user identity (DN & FQAN)**
  - Developed for CREAM, will be used also by the gLite CE
- **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: Accounting

- **Collect usage records for all jobs at sites**
  - *Local and global jobId, uid, DN, VOMS FQAN, system usage (cpuTime, ...), ...*
  - Currently the information is taken from log files produced by BLAH (gLiteCE, Cream) and the LCG-CE
- **The information is collected from sites using APEL**
  - Currently insecure storage and transfer of accounting records via R-GMA. Working to add encryption and an authorization layer
  - DGAS already provides proper management of privacy (records signed and encrypted) but doesn't have a proper interface for data visualization
- **Issues:**
  - 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
  - Need to converge on a single accounting collection tool
    - Process for the merge of APEL and DGAS already started
  - 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:**
  1. Mapping of VOMS groups to batch system shares (via GIDs?)
  2. Two queues (long/short) for ATLAS & CMS
  3. Publish info on the share in the CE GLUE 1.2 schema (VOView)
     - The gLite WMS has been modified to support GLUE 1.2
     - Possibility to test with GPBOX if the “Service Class” is published
  4. WMS match-making depending on submitter VOMS certificate
     - But no ranking of resources based on priority offered yet
  5. Settings are not dynamic (via e-mail or CE updates)
- **If GPBOX is needed for LHC, tests must start now!**
     - 12 months are needed to bring it to production quality
     - EGEE JRA1 is setting up a small “preview test-bed” where we plan to expose new components (including GPBOX) to the users before certification

Future 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**

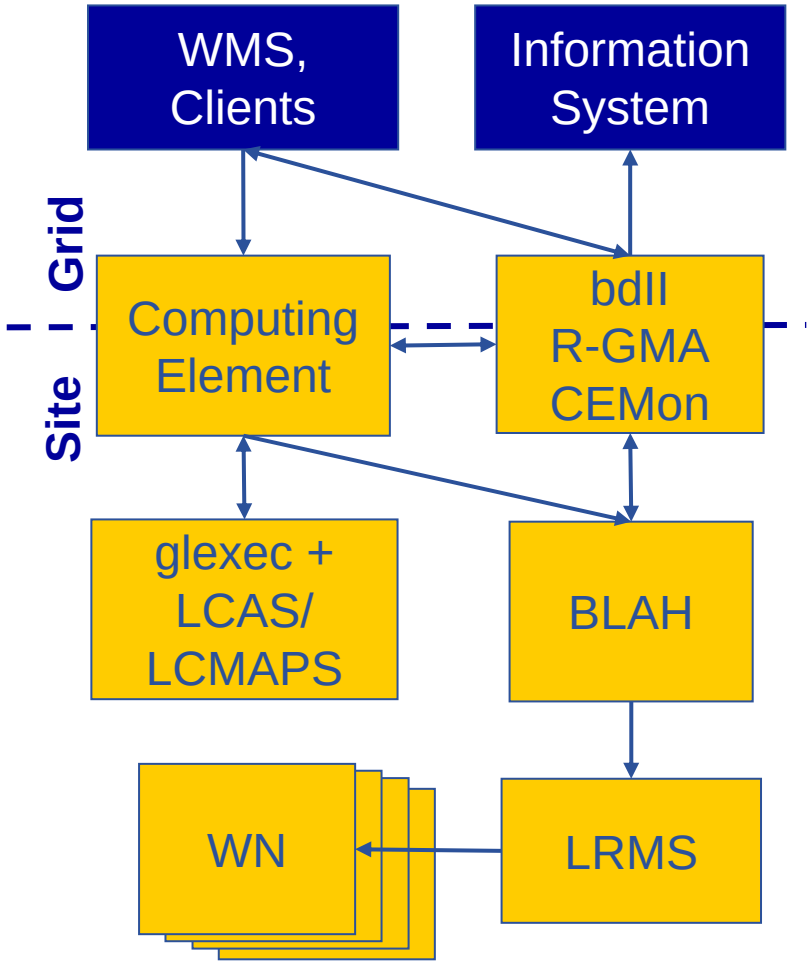
Backup slides

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)**
  - 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 WMProxy 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
- **Support for 'shared' and '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

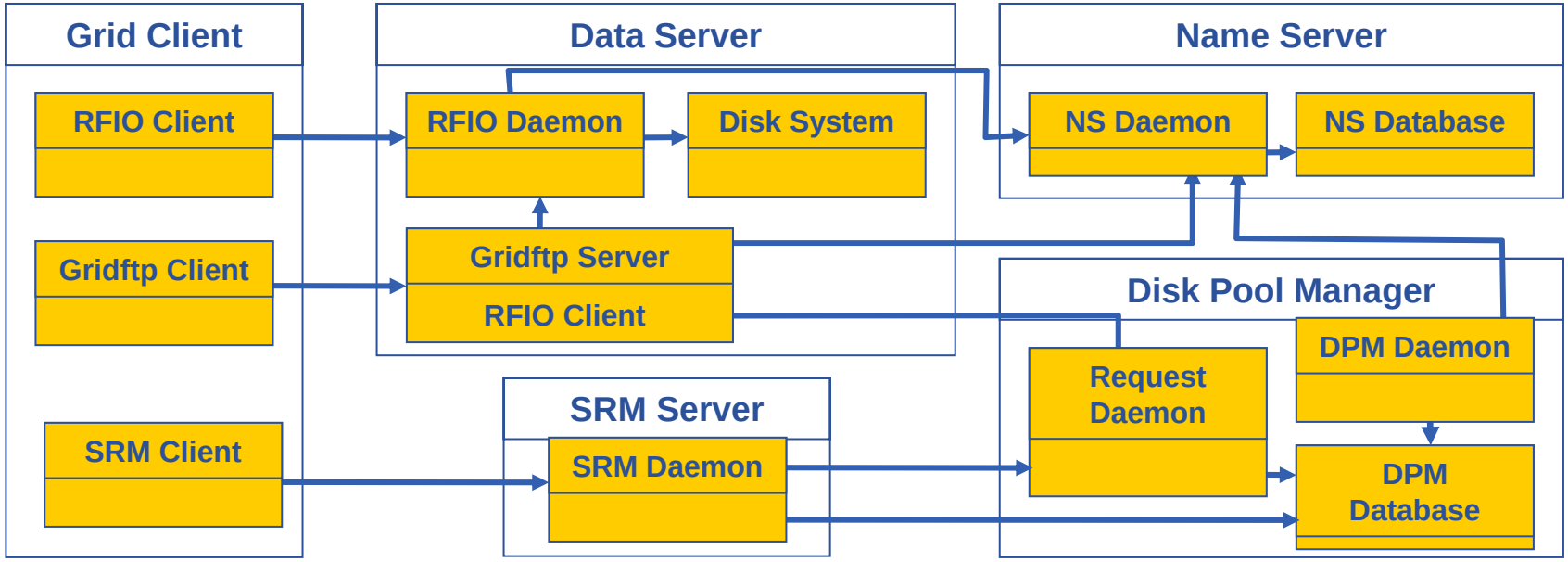
Highlights: Computing Element

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



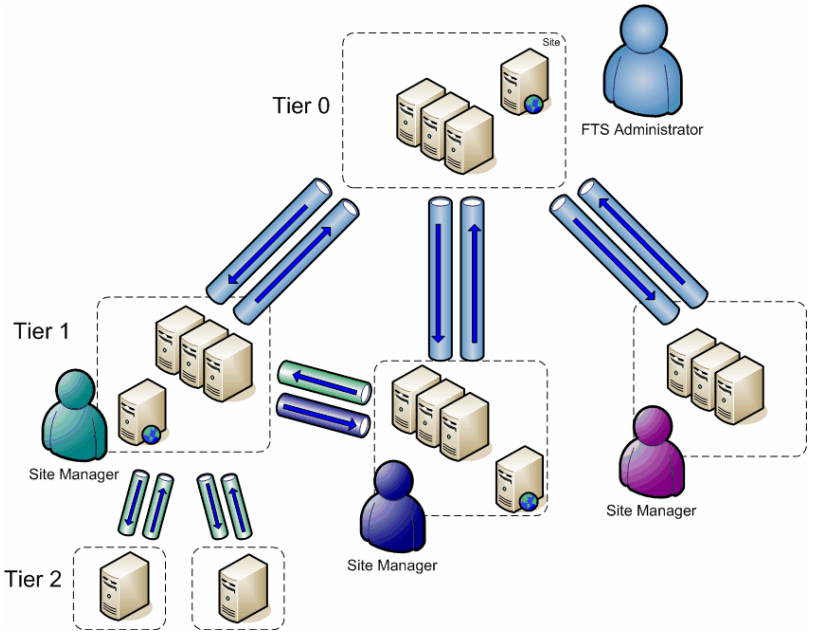
Highlights: Disk Pool Manager

- **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 being added now



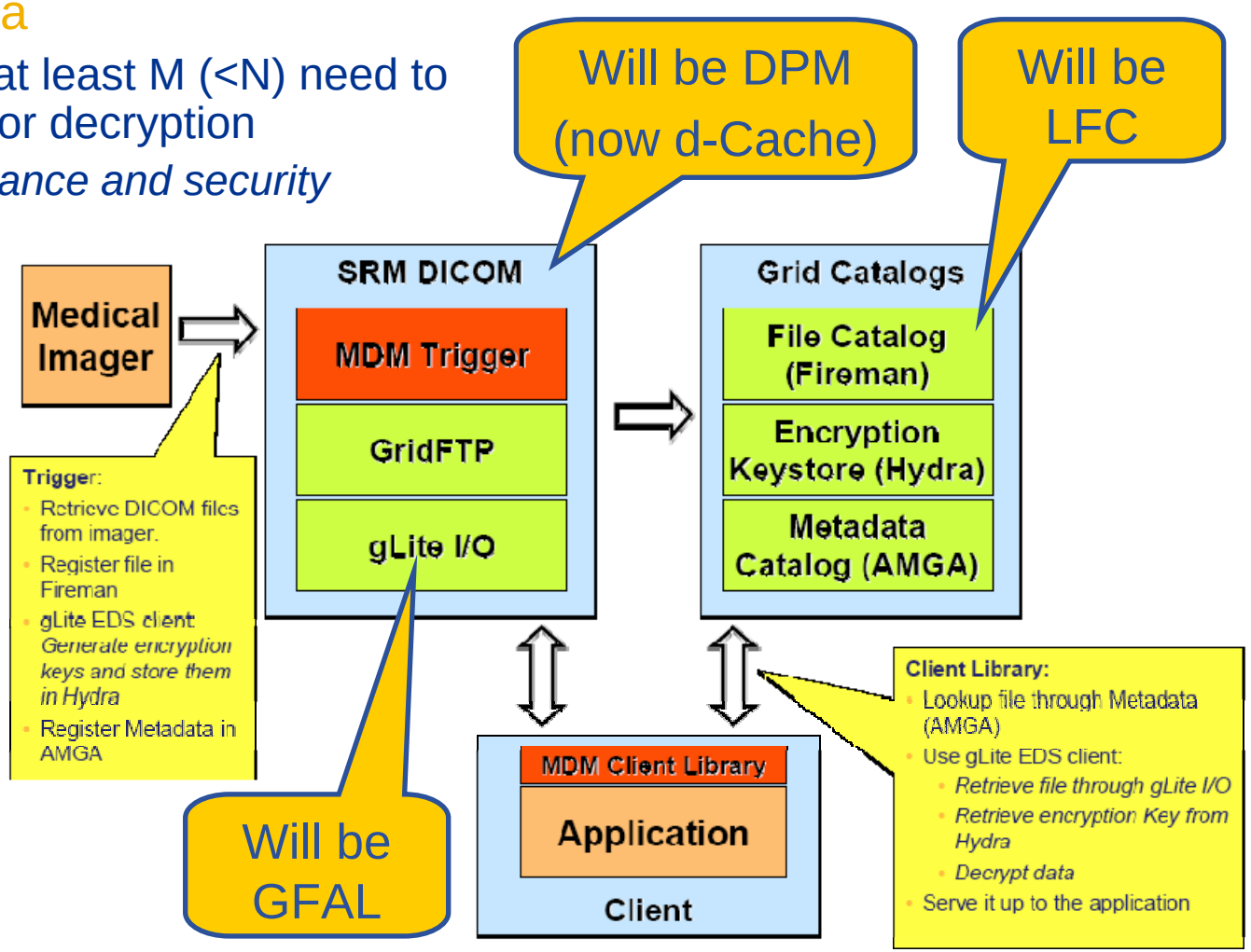
Highlights: FTS

- **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)



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 without compromising their commitment in SA1

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

We need more sites!!!