The gLite middleware

Lightweight Middleware for Grid Computing

Claudio Grandi
EGEE-II JRA1 Manager
Claudio.Grandi@cern.ch

www.eu-egee.org

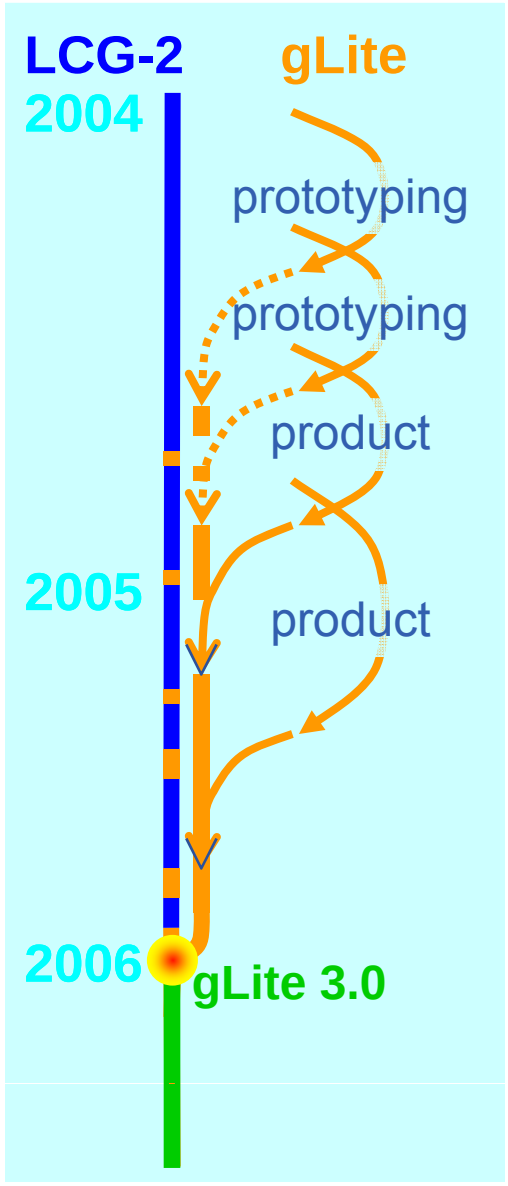
Information Society and Media

Outline

- gLite distributions
- Software process
- gLite restructuring
- Experimental Services
- Highlights
- Standards and interoperability

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 released in May 2006
- **Focus on providing a deployable MW distribution for EGEE production service**



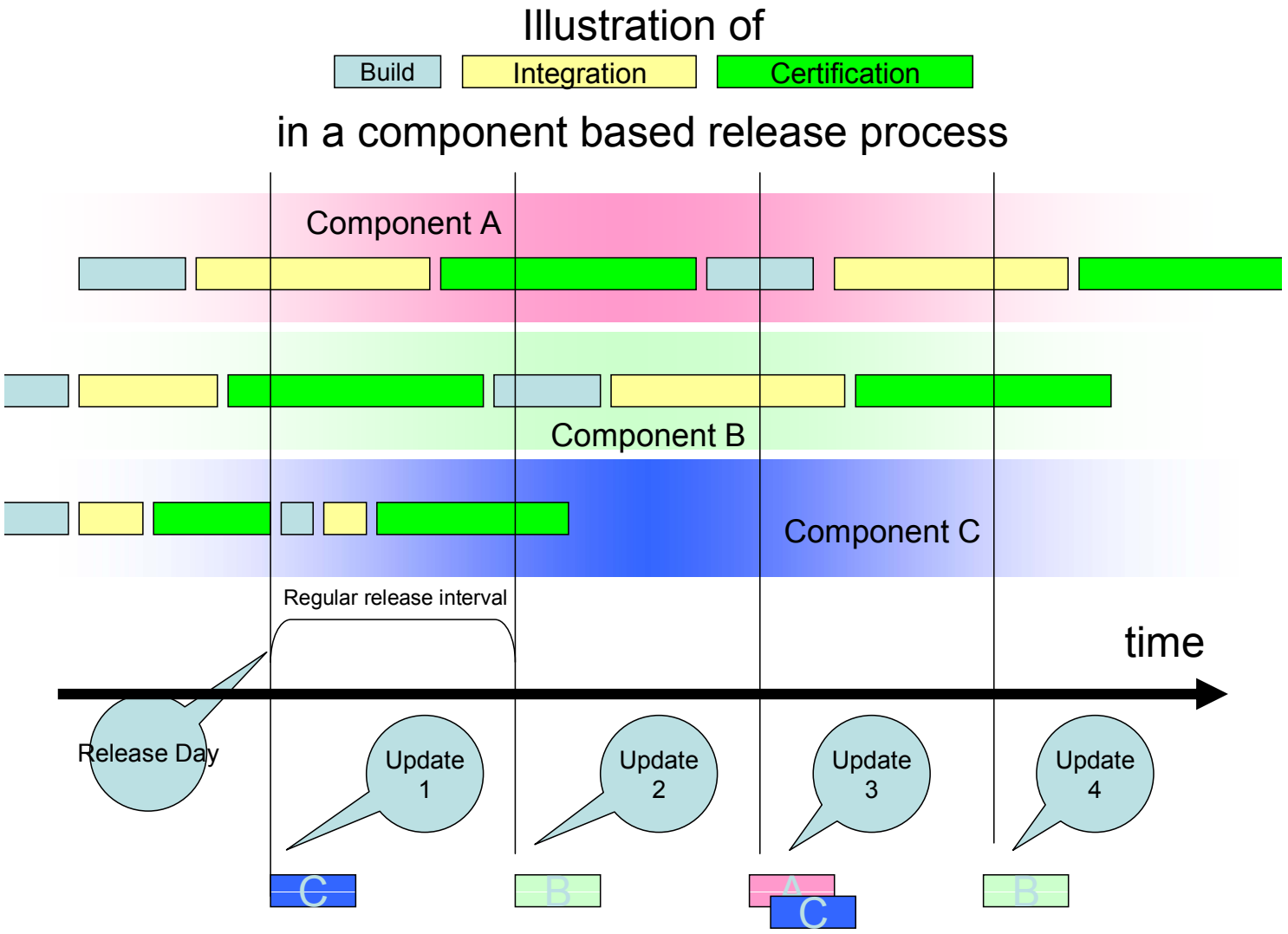
gLite 3.1

- **gLite 3.0 is working on Scientific Linux 3 – x86-32**
  - Release based on legacy ANT-based and LCG build systems
  - Using VDT 1.2.2, Globus 2
- **The new components part of the gLite 3.1 stack will initially be working on Scientific Linux 4 – x86-32**
  - Release fully based on ETICS build system
  - Using VDT 1.6.0, Globus 4
    - This complicated the process significantly
  - Support for other platforms, starting with SL4 – x86-64, will follow
- **gLite 3.1 components release status:**
  - All clients (Worker Node and User Interface): released
  - LCG-CE, BDII, DPM, LFC, VOMS/VOMS-Admin: end-October
  - WMS, FTS, VOBOX: before the end of 2007
  - CREAM, R-GMA: beginning of 2008

Software Process

- **Introduced new software lifecycle process**
  - Based on the gLite process and LCG-2 experience
  - Components are updated independently
  - Updates are delivered on a ***weekly*** basis to PPS
    - Each week either a gLite 3.1 or 3.0 update (if needed)
    - Move after ***2 weeks*** to production
  - Acceptance criteria for new components
  - Clear link between component ***versions***, ***Patches*** and ***Bugs***
    - Semi-automatic release notes
  - Clear prioritization by stakeholders
    - TCG for medium term (3-6 months) and EMT for short term goals
  - Clear definition of roles and responsibilities
  - Documented in MSA3.2
  - In use since July 2006

Component based process



Configuration Management

- **The configuration tool is YAIM**
  - Chosen by site administrators
  - Based on Key-Value pairs + bash
  - Easy to integrate with other fabric management tools
- **Now using YAIM 4.0**
  - Supports component based configuration
  - Enables software providers to maintain configuration independently
- **Installation tool**
  - APT for (semi) automatic RPM updates
    - Standard Debian tool, widely used
  - RPM lists for other tools
  - Tarballs for UIs and WNs

Certification

- **Formal process for Patch certification**
- **Extended test beds: 8 sites**
  - about 100 nodes to cover additional deployment scenarios
- **Extensive use of *virtualized* test beds**

	a	b	c	d	e	f	g	h	i	j	k	l
0		2033 BDII_top Configured	2020 PX Configured	1B	lxb2017 BDII_top Configured	2	3+sl4	1738 VOBOX Configured		Stable	Config	1774 VOBOX Configured
1				2032 WMS Configured		0744 WMS Configured	6122 WMS 03	1928 VOMS Configured				1912 WMS Configured
2	2057 UIcomb Configured			1765 UIcomb Configured			6119 UI Configured	0714 SAM Configured		1917 UI Configured		1778 UI Configured
3	2016 RB Configured			1762 RB Configured		0730 RB Configured				1794 RB Configured		
4	lxb1776 BDII_site Configured			lxb6121 gLiteCE Configured sl3.arch32.mw32		0743 gLiteCE Configured				1936 BDII_top Configured		1919 gLiteCE Configured
5	2018 CE Configured			2034 CE Configured		2035 CE Configured		LSF		1938 CE Configured		1779 CE Configured
6	0731 WNcomb Configured			6122 WNcomb Configured sl4.arch64.mw32		0741 WNcomb Configured		6115 CE Configured		1758 WNcomb Configured		0718 WNcomb Configured
7	1921 DPM_mysql Configured			lxb6125 WNcomb Configured sl4+arch32+mw32		1716 WNcomb Configured				1916 DPM_pool Configured		0734 WNreloc Configured
8				lxb6126 WNcomb Configured sl4+arch32+mw32		1720 WNcomb Configured	6127 DPM_mysql 03	gCondor	torqueTest	1751 dCache_mysql Configured		1777 dCache_mysql Configured
9				lxb1917 DPM_mysql Configured		0724 SEclassic Configured	6128 LFC_mysql 03	lxn1176 gLiteCE Configured	1920 CE Configured	1915 DPM_mysql Configured		1775 SEclassic Configured
10	2019 MON Configured					lxb2036 DPM_mysql Configured	6129 FTS 03	0738 WNcomb Configured	0739 WNcomb Configured			
11	1941 LFC_mysql Configured	lxb1909 FTS Configured	lxb1543 MySqlDB Configured sl4+arch32+mw32	1782 LFC_oracle Configured	2011 FTS Configured			lxshare0297 WNcomb 03				0727 DPM_pool Configured

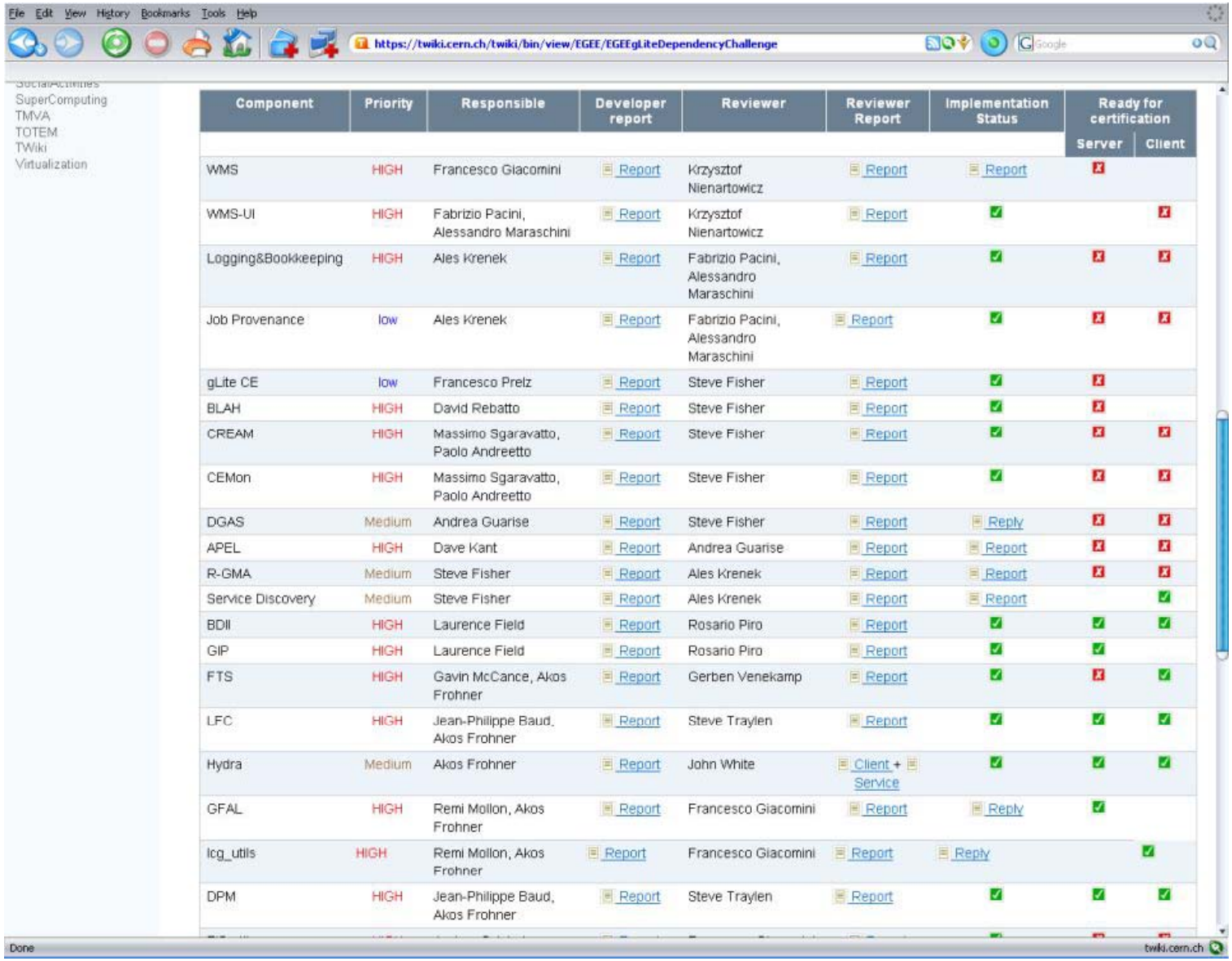
ETICS

- **The build tools now in use are provided by ETICS**
  - The ETICS Project started in January 2006
  - EGEE started using ETICS tools in Autumn 2006
  - Migration process to ETICS has taken longer than foreseen
- **The gLite 3.1 components are built exclusively with ETICS**
  - The legacy ANT-based and LCG systems are still active for updates to the gLite 3.0 version
  - We now have a better understanding and control of the package dependencies

gLite restructuring

- **Requirements on gLite are increasing due to:**
  - Increased number of sites, with different levels of gLite expertise
  - Increased number of applications, with diverse needs
  - Demand for more supported platforms
- **Long term maintenance of gLite needs to be assured**
  - Standard and/or commercial solutions not widely available, yet
  - Support of legacy components and dependencies on external packages made the gLite stack grow too complex
- **In January 2007 the project decided it was necessary to invest effort to address the long term sustainability of the middleware**
  - Cleanup of code-base and dependencies for gLite 3.1 is now being pursued with high priority at expense of adding new functionality
- **The gLite restructuring process started on May 28th and coexists with the current activities needed to support the applications on the production infrastructure**

Dependency Challenge



Component	Priority	Responsible	Developer report	Reviewer	Reviewer Report	Implementation Status	Ready for certification	
							Server	Client
WMS	HIGH	Francesco Giacomini	Report	Krzysztof Nienartowicz	Report	Report	✗	
WMS-UI	HIGH	Fabrizio Pacini, Alessandro Maraschini	Report	Krzysztof Nienartowicz	Report	✓		✗
Logging&Bookkeeping	HIGH	Ales Krenek	Report	Fabrizio Pacini, Alessandro Maraschini	Report	✓	✗	✗
Job Provenance	low	Ales Krenek	Report	Fabrizio Pacini, Alessandro Maraschini	Report	✓	✗	✗
gLite CE	low	Francesco Prelz	Report	Steve Fisher	Report	✓	✗	
BLAH	HIGH	David Rebatto	Report	Steve Fisher	Report	✓	✗	
CREAM	HIGH	Massimo Sgaravatto, Paolo Andreetto	Report	Steve Fisher	Report	✓	✗	✗
CEMon	HIGH	Massimo Sgaravatto, Paolo Andreetto	Report	Steve Fisher	Report	✓	✗	✗
DGAS	Medium	Andrea Guarise	Report	Steve Fisher	Report	Reply	✗	✗
APEL	HIGH	Dave Kant	Report	Andrea Guarise	Report	Report	✗	✗
R-GMA	Medium	Steve Fisher	Report	Ales Krenek	Report	Report	✗	✗
Service Discovery	Medium	Steve Fisher	Report	Ales Krenek	Report	Report		✓
BDII	HIGH	Laurence Field	Report	Rosario Piro	Report	✓	✓	✓
GIP	HIGH	Laurence Field	Report	Rosario Piro	Report	✓	✓	
FTS	HIGH	Gavin McCance, Akos Frohner	Report	Gerben Venekamp	Report	✓	✗	✓
LFC	HIGH	Jean-Philippe Baud, Akos Frohner	Report	Steve Traylen	Report	✓	✓	✓
Hydra	Medium	Akos Frohner	Report	John White	Client + Service	✓	✓	✓
GFAL	HIGH	Remi Mollon, Akos Frohner	Report	Francesco Giacomini	Report	Reply	✓	
lcg_utils	HIGH	Remi Mollon, Akos Frohner	Report	Francesco Giacomini	Report	Reply		✓
DPM	HIGH	Jean-Philippe Baud, Akos Frohner	Report	Steve Traylen	Report	✓	✓	✓

Experimental services

- **Performance of some critical components needed improvements**
  - Stability of the Workload Management System (**WMS**) and the Logging & Bookkeeping (**LB**)
  - Stability and performance of the gLite Computing Element (**gLiteCE**) and the test of the new **CREAM** Computing Element
- **The normal certification process could not help**
  - Needed testing at the production scale
- **Introduced the *Experimental Services***
  - Instances of the services attached to the production infrastructure
  - Maintained by SA1 and SA3
  - JRA1 patches are installed immediately (before the certification)
  - Testing done by selected application users
  - Process controlled by the EMT
- **Rapid improvement of the components**
  - **Improved WMS** and **LB** in production
  - **gLiteCE** and **CREAM** passed the acceptance tests

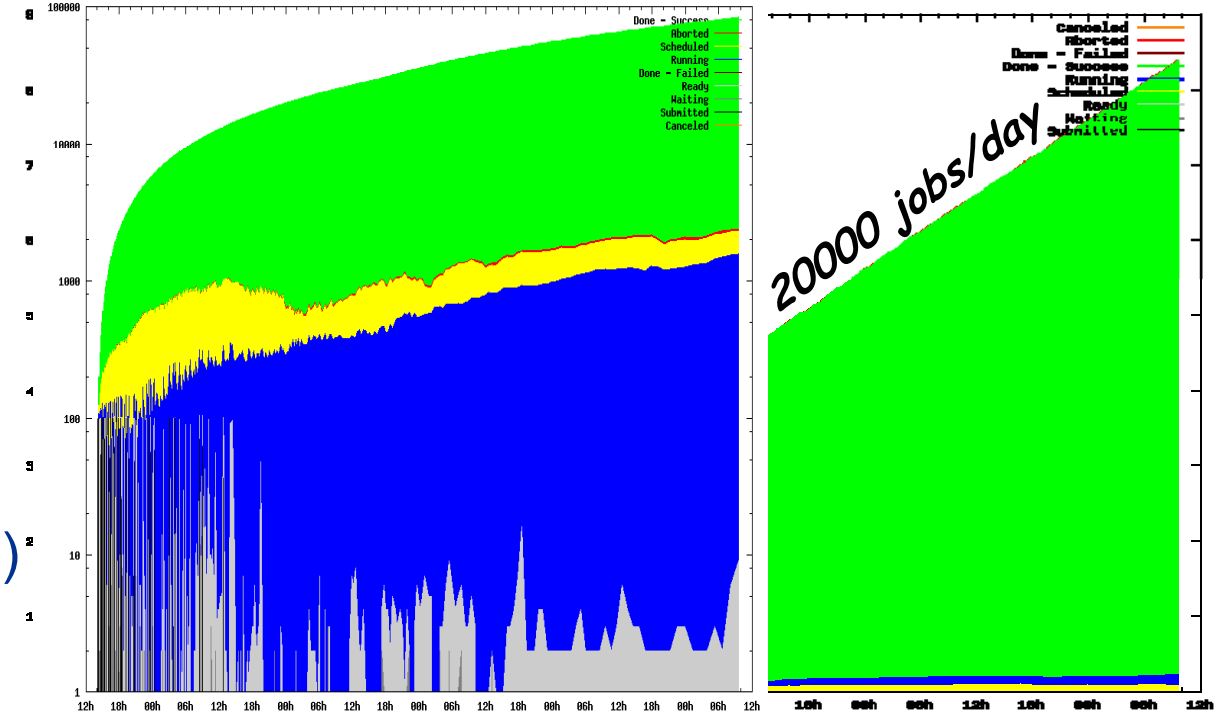
WMS/LB: acceptance criteria

- **Before being taken in certification:**
  - 5 days consecutive run without intervention
  - 10K jobs/day handled by a single instance of the WMS
  - number of jobs in non-final state < 0.5%
  - Proxy renewal must work at the 98% level

- **Note that the LB should be installed on a separate machine (that can serve multiple WMS instances at the same time)**

Status of WMS and L&B

- **Acceptance tests passed in Easter:**
  - run one week at 15000 jobs/day without manual intervention with 0.3% of jobs in non-final state
- **Stress test:**
  - 27000 jobs/day
- **SL3 version is now in production**
  - Patch #1251
- **Removed support:**
  - NS interface (better performance)
  - Checkpointable and partitionable jobs



- **Improved management of collections (without DAGMan)**
  - Bulk match-making
- **Support for sandbox file transfer via gsiftp and https**

CE: acceptance criteria

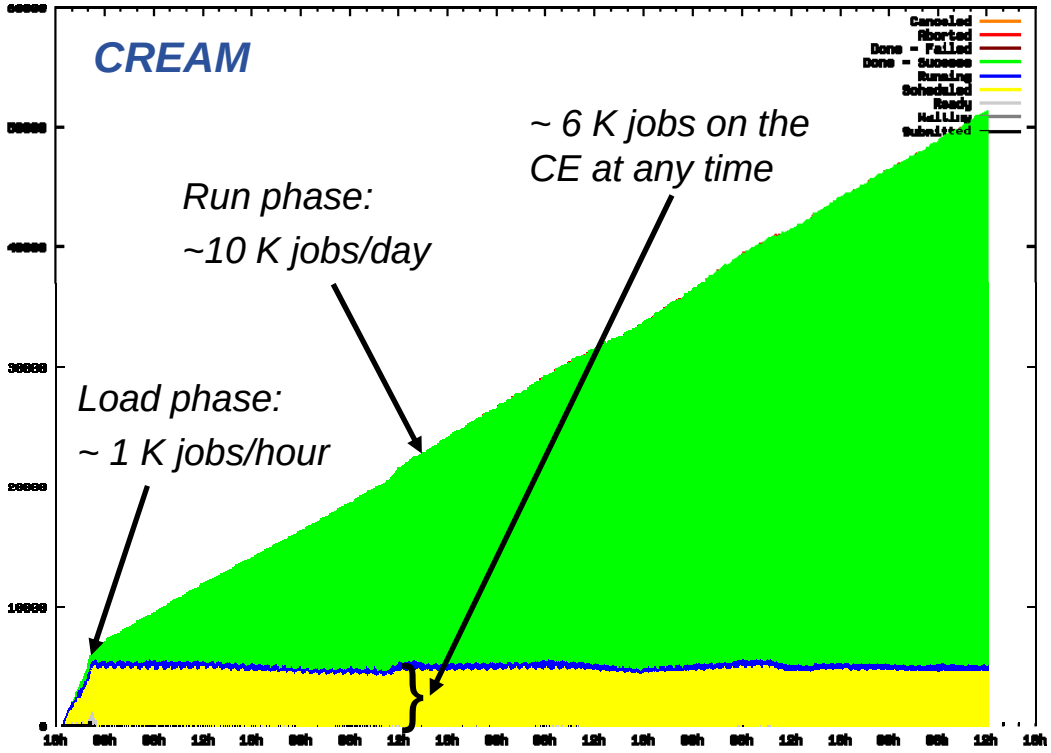
- **Before being taken in certification:**
  - 5000 simultaneous jobs per CE node
  - Job failure rates due to CE in normal operation: < 0.5%
  - Job failures due to restart of CE services or CE reboot <0.5%.
  - For 2007
    - 50 *user/role/submission_node* combinations (Condor_C instances) per CE node
    - 5 days unattended running with performance on day 5 equivalent to that on day 1
  - In the longer term
    - 1 CE node should support an unlimited number of user/role/submission node combinations, from at least 20 VOs
    - 1 month unattended running without performance degradation

CE: additional requirements

- **Support for commonly used batch systems**
  - High priority: LSF, PBS-Torque/Maui
  - Condor, SGE
- **Support for commonly used submission systems**
  - gLite WMS, Condor-G, pre-WS GRAM
  - In future BES
  - Note that by construction not all the submission systems may be implemented by the same product
    - may need more than one product at each site...
- **Support for services needed on the infrastructure**
  - GIP to publish on the bdII information system
  - Accounting data to the GOC (APEL, DGAS)
- **Interoperability with our major partner projects (OSG)**

Status of Compute Element 1/2

- **The gLiteCE and CREAM passed the accepatance tests**
  - gLite CE (GSI-enabled Condor-C) on SL3
    - Implements the identity change with glexec (called by BLAH)
    - Up to 8000 jobs in a day submitted with 40 users with no failures
    - https://edms.cern.ch/document/863275/1
  - CREAM on SL3
    - Uses the same BLAH of the gLite CE
    - > 90000 in 8 days with 50 users and 111 failures (LSF errors)
    - https://edms.cern.ch/document/863276/1
  - Same performances verified also on the SL4 version



Status of Compute Element 2/2

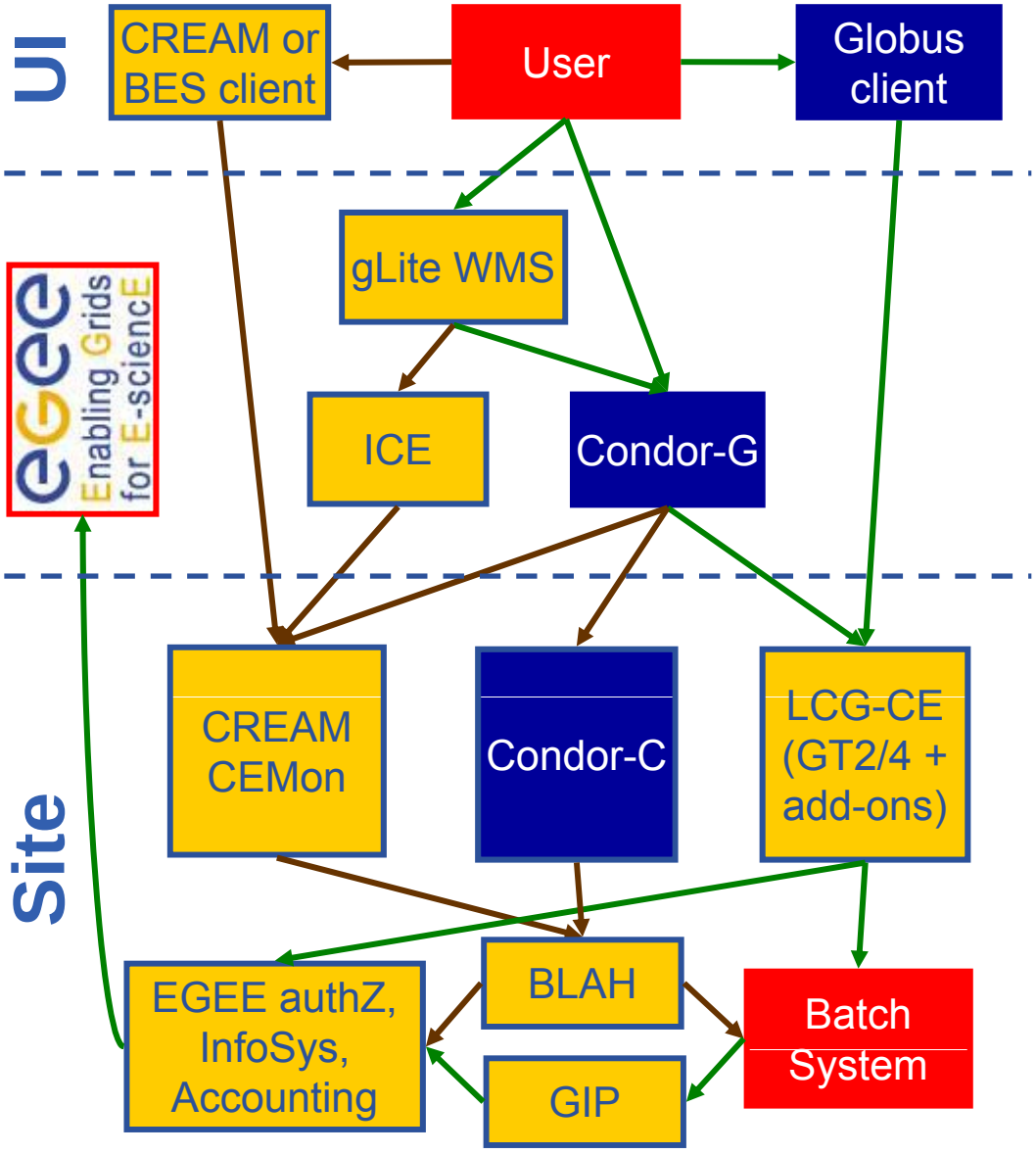
- **LCG-CE has been ported to SL4 / VDT 1.6.0**
  - GT4 pre-WS GRAM (tests @CERN)
    - Uses the globus call-out implemented in GT4 and the new LCAS/LCMAPS that was developed for the gLiteCE
    - Does not scale to the level of the acceptance tests
      - *ok up to about 10 concurrent users, but 21% success rate with 9000 jobs/day submitted by 15 users and high load on the machine*
    - https://twiki.cern.ch/twiki/bin/view/LCG/LCGCETest

CE: next steps

- **Assuming that the project cannot support more than one CE in the long term, the TCG decided that:**
  - The first priority is to deploy the LCG-CE on SL4 as it is
    - On the PPS in early October, on the PS by end of October
  - Full develop CREAM and make it available on the production infrastructure
    - Certification starts in December, in production in Spring 2008
    - Status: https://twiki.cern.ch/twiki/bin/view/EGEE/CECheckList
  - Sites supporting applications that use native globus interface for job submission will continue to deploy the LCG-CE
  - No effort will be spent on the gLite CE
    - We will continue to accept requirements on BLAH and glexec from Condor that will be included in the RedHat and Fedora distributions
  - TCG document:
    - https://twiki.cern.ch/twiki/bin/viewfile/EGEE/TCGHome?rev=1;filename=CE-proposal.doc

Compute Element

- **LCG-CE (GT4 pre-WS GRAM)**
  - Globus client, Condor-G
  - Globus code not modified
  - *Not passing acceptance tests*
- **CREAM (WS-I)**
  - CREAM client, ICE
  - Condor-G (collaboration with Condor team)
  - BES (collaboration with OMII-EU for new BES 1.0 support)
- **Condor-C (was in gLite-CE)**
  - Condor-G
  - Maintained by Condor
  - VDT includes BLAH & glexec



gLite component | non-gLite component | User / Resource

In production

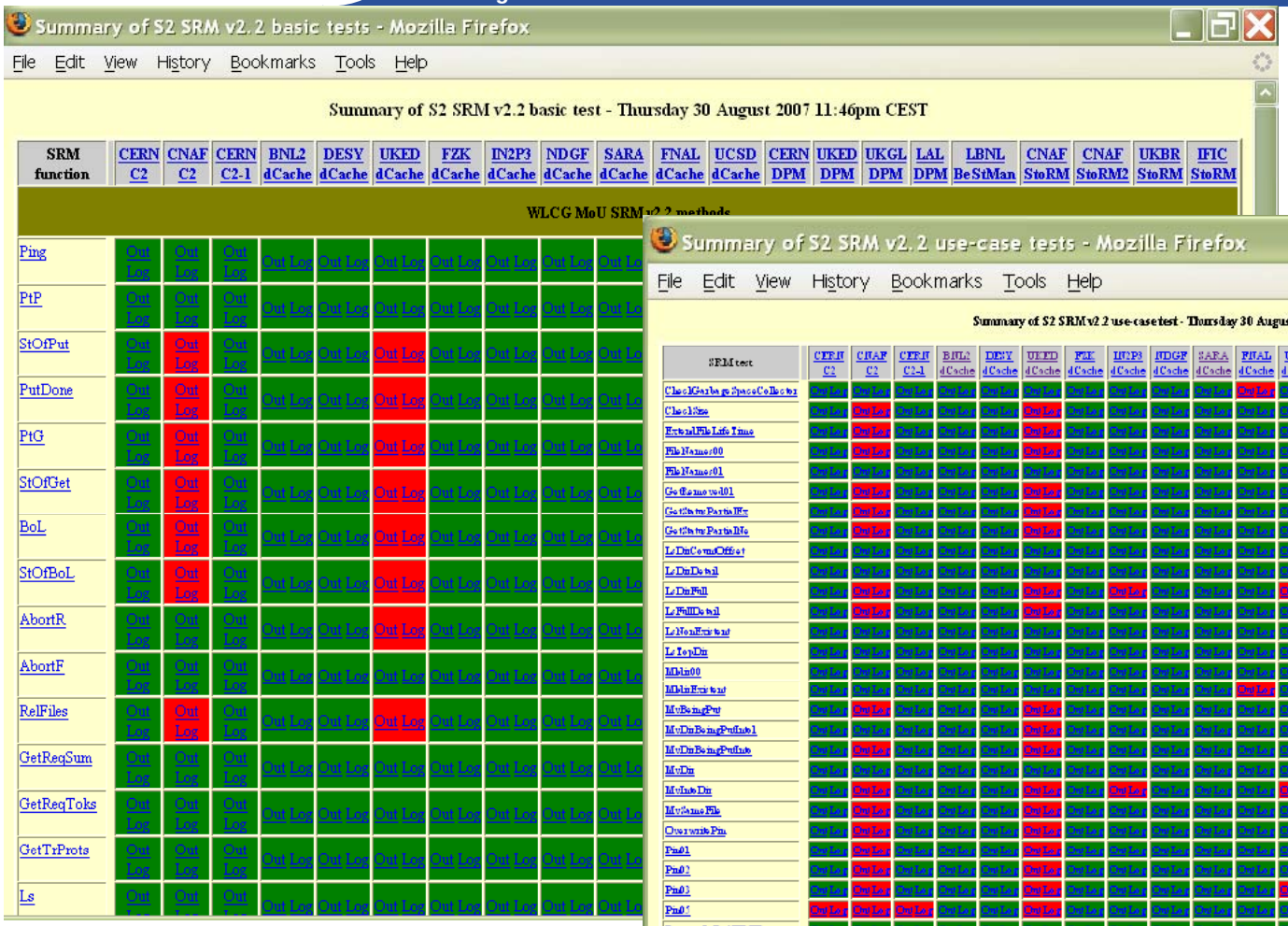
Existing prototype

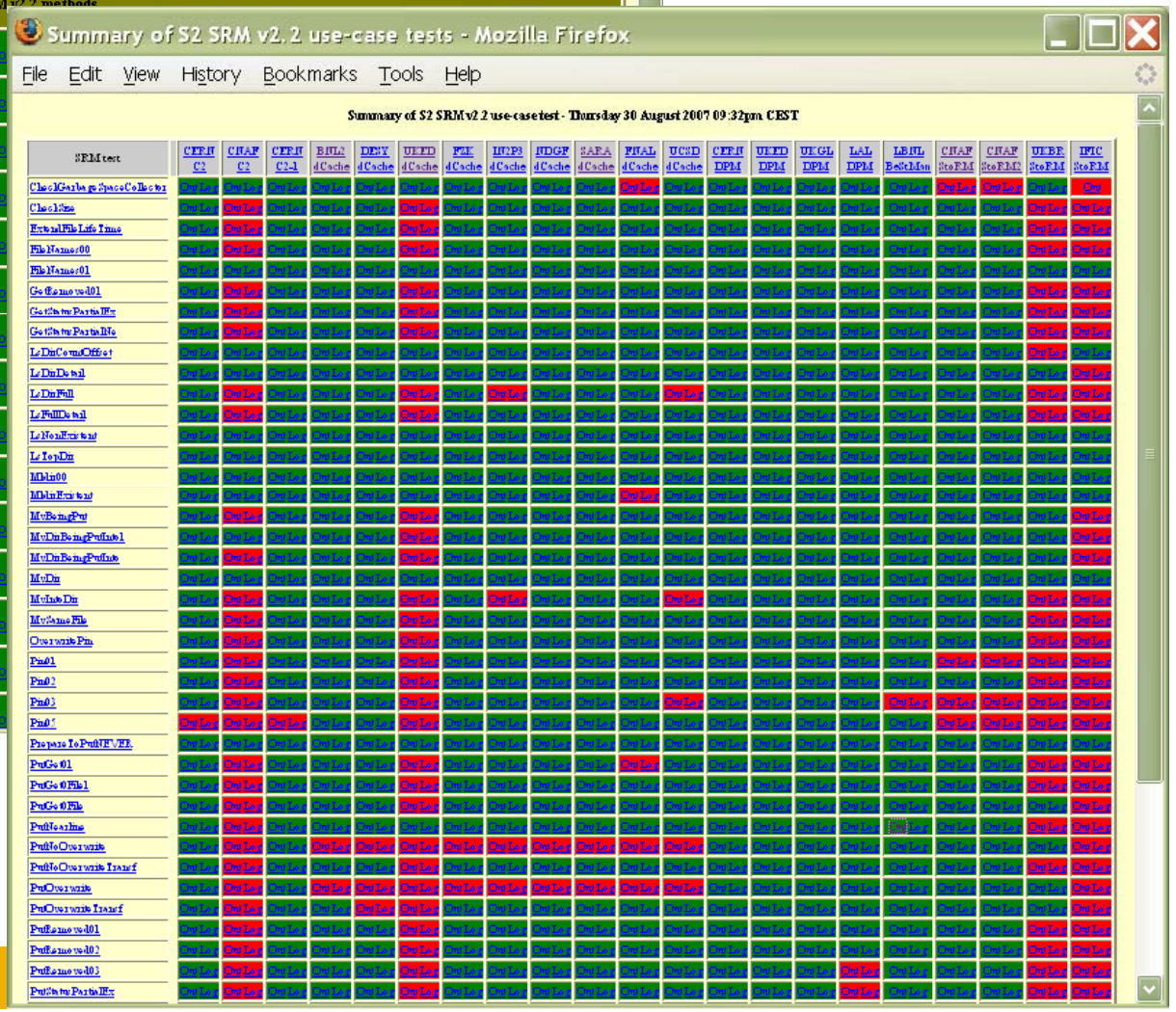
SRM 2.2

- **The Storage Resource Manager (SRM) interface is the basis for the gLite Storage Elements**
  - Hides the storage system implementation
  - Handles the authorization based on VOMS credentials
  - Posix-like access to SRM via **GFAL** (Grid File Access Layer)
- **Several implementations considered in EGEE:**
  - **CASTOR2** : Hierarchical Storage Server (HSS). Developed by CERN and RAL. SRM v2.2 support in v.2.1.4.
  - **dCache**: HSS developed by DESY and FNAL. SRM 2.2 support in v1.8.
  - **DPM**: disk-only developed by CERN. SRM v2.2 support in v1.6.5 in production.
  - **StoRM**: disk-only developed by INFN and ICTP. SRM v2.2 interface for many filesystems: GPFS, Lustre, XFS and POSIX generic filesystem. SRM v2.2 support in v1.3.15.
  - **BeStMan**: disk-based developed by LBNL. SRM v2.2 support in v2.2.0.0.

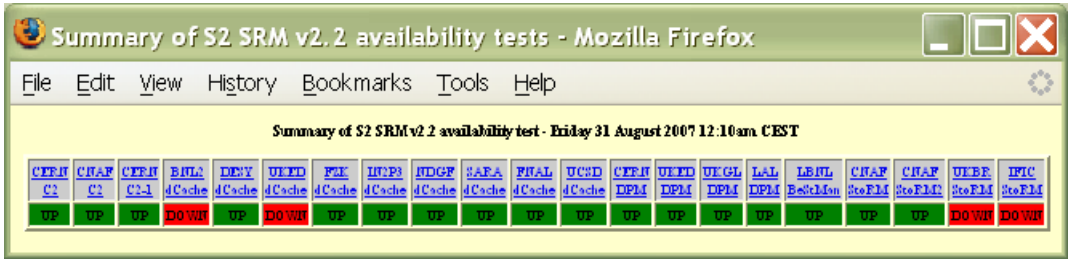
SRM basic and use cases tests

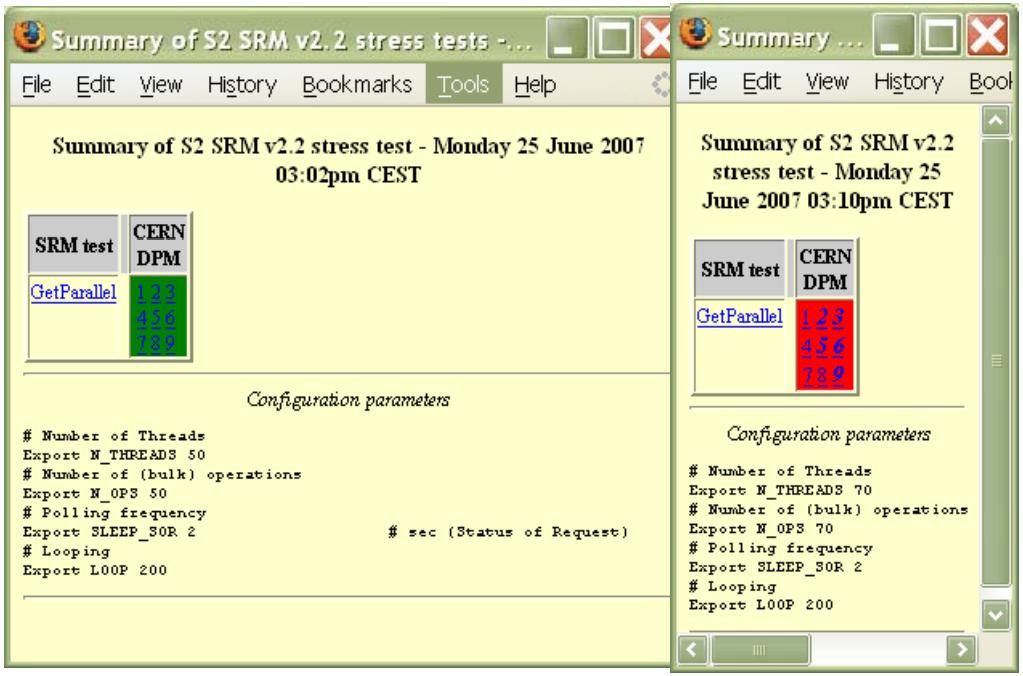
Enabling Grids for E-sciencE

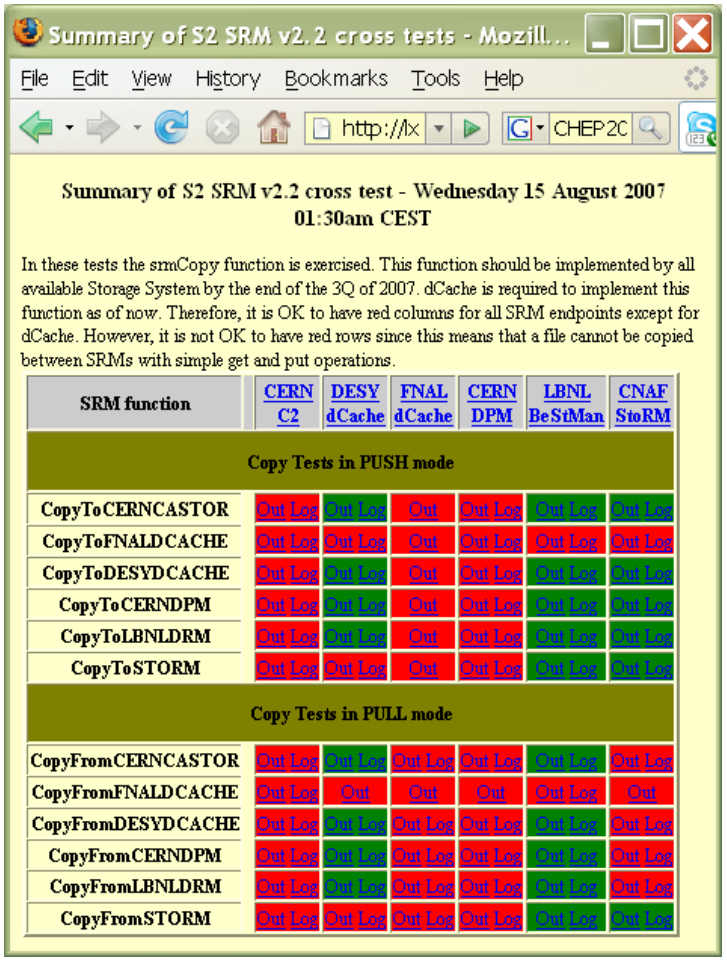




SRM: availability, interoperability and stress tests







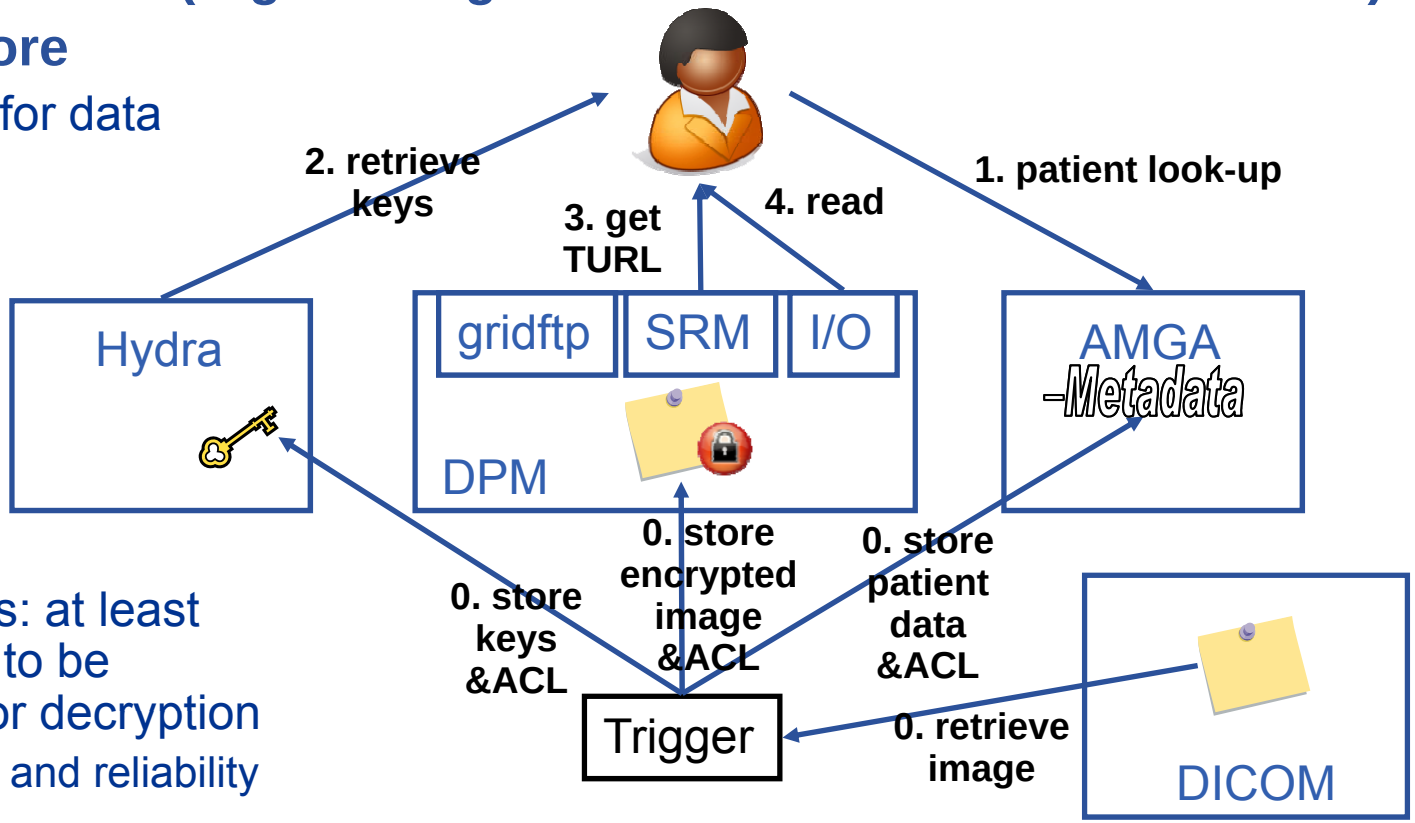
- **See Flavia Donno's presentation at CHEP'07:**
  http://indico.cern.ch/materialDisplay.py?contribId=78&sessionId=26&materialId=slides&confId=3580

Other Data Management tools

- **SRM 2.2 support has been added to all DM tools**
- **FTS 2.0 released**
  - Better security model
    - proper certificate delegation rather than retrieval from MyProxy
  - Better administration tools - improved monitoring capabilities
  - Better database model - improved performance and scalability
- **New version of LFC**
  - Introduced bulk operations → 20 times faster!
  - 110 LFCs in production (37 central, 73 local)
- **AMGA metadata catalogue (by NA4) being certified now**
- **Encrypted Data Storage**
  - Prototype based on DPM/LFC ready by the end of October
    - Will allow dismissing support for Fireman and gLiteI/O
    - Still to be done: support for ACL synchronization across replicas

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



Information System & Monitoring

- **Now using version 1.3 of the GLUE schema**
  - Aim to contribute and adopt the GLUE 2.0 schema as defined by the new GLUE Working Group at OGF
- **Access to the Information System via the Service Discovery**
  - gLite Service Discovery currently supports R-GMA, BDII and XML files back ends
  - Working on a **SAGA**-compliant interface
- **EGEE is using the BDII as Service Discovery back-end**
  - Based on an LDAP database
  - Adequate performance to address the infrastructure needs
    - Up to 2 million queries/day served (over 20 Hz)
- **gLite 3.1 R-GMA will have authorization, Virtual DB support and schema replication (beginning ’08)**

Authorization Services

- **VOMS is now a de-facto standard**
  - Attribute Certificates provide users with additional capabilities defined by the VO.
  - Basis for the authorization process
- **Authorization: via mapping to a local user on the resource**
  - glexec changes the local identity (based on suexec by Apache)
- **Designing an Authorization Service**
  - Uniform implementation of authorization in gLite components
  - Easier management at the sites
  - Compatible with SAML and XACML standards
  - Full design foreseen to be completed by Dec. 2007, implementation will start in 2008
- **Prototype being prepared now**
  - Common interface agreed with OSG
  - Needed to support identity change on the WN (pilot jobs)

Interoperability

- **gLite needs to interoperate with other infrastructures**
- **Use of standards is the way to go but in the meantime we need pragmatic approaches for interoperability**
- **Focus is on the Grid Foundation middleware**
  - Security
    - Certificates for AuthN and VOMS for AuthZ; SAML assertions in future
  - Information systems
    - GLUE schema (1.3 now 2.0 in future) and adapters to interoperate
  - Data Management
    - SRM 2.2 interface for data access and GridFTP for file transfers
  - Job Management
    - BES interface in future (CREAM). Legacy pre-WS GRAM deployed in parallel. gLite WMS, Condor-G and BLAH to build gateways.
    - OGF-RUS will be used for accounting

Summary

- The main goal for gLite is the support of the production infrastructure
- The new software process based on component release is now used routinely
- The activity to add support for SL4 showed the need to rationalize the software. This is now being pursued as an high priority activity
- Improvements have been made on many components with the aim to increase the performance and usability
- New activity starting aiming at the rationalization of the authorization mechanisms, including support for VO policies
- Significant effort for improving interoperability with other infrastructures and standardization

www.glite.org