"LCG2 Operational Experience and Status"

Markus Schulz, IT-GD, CERN
markus.schulz@cern.ch

EGEE is a project funded by the European Union under contract IST-2003-508833

Outline

- Why LCG ???
- Building LCG-2
- Operating LCG
  - releases
  - changes for gLite
- Impact of Data Challenges on operations
- Problems
- Changes
  - operations
- Summary

Challenges for the LCG

http://lcg.web.cern.ch/lcg

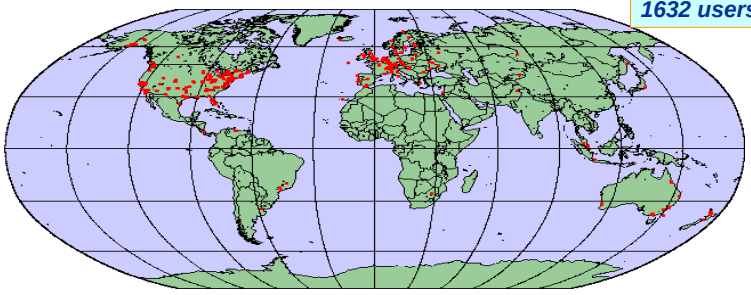
- LHC (Large Hadron Collider)
  - with 27 km of magnets the largest superconducting installation
  - proton beams collide at an energy of 14TeV
  - 40 million events per second from each of the 4 experiments
  - after triggers and filters 100-1000MBytes/second remain
  - every year **~15PetaByte** of data will be recorded
  - this data has to be reconstructed and analyzed by the users
  - in addition large computational effort to produce Monte Carlo data
  - O(100K) of current CPUs are needed for processing the data

User Community

- Collaborators
  - > 6000 users from 450 institutes

Europe:
267 institutes
4603 users

Elsewhere:
208 institutes
1632 users

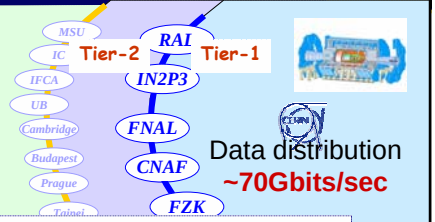


LCG Scale and Computing Model

- Tier-0
  - reconstruct (ESD)
  - record raw and ESD
  - distribute data to tier-1
- Tier-1
  - data heavy analysis
  - permanent, managed grid-enabled storage (raw, analysis, ESD), MSS
  - reprocessing
  - regional support
- Tier-2
  - managed disk storage
  - simulation
  - end user analysis
  - parallel interactive analysis

desktops portables

small centres



Data distribution ~70Gbits/sec

Current estimates of Computing Resources needed at Major LHC Centres

First full year of data - 2008

	*Processing M SI2000***	*Disk PetaBytes*	*Mass Storage PetaBytes*
CERN	20	5	20
Major data handling centres (Tier 1)	45	20	18
Other large centres (Tier 2)	40	12	5
Totals	**105**	**37**	**43**

*** Current fast processor ~1K SI2000*

LCG-2 software

- **Evolution through 2003/2004**
  - Focus has been on making these reliable and robust
    - Basic functionality and reliability rather than additional functionality
  - Respond to needs of users, admins, operators
- **The software stack is the following:**
  - Virtual Data Toolkit
    - Globus (2.4.x), Condor, etc
  - EDG developed higher-level components
    - Workload management (RB, L&B, etc)
    - Replica Location Service (single central catalog), replica management tools
    - R-GMA as accounting and monitoring framework
    - VOMS being deployed now
  - Operations team re-worked components:
    - Information system: MDS GRIS/GIIS → BDII
    - edg-rm tools replaced and augmented as lcg-utils
    - Developments on:
      - Disk pool managers (dCache, DPM)
      - Catalogue
  - Other tools as required:
    - e.g. GridIce - DataTag

- Maintenance agreements with:
  - VDT team (inc Globus support)
  - WLM, VOMS – Italy
  - DM – CERN

Experience

- Jan 2003 GDB agreed to take VDT and EDG components
- September 2003 LCG-1
  - **Extensive certification process**
  - **Integrated 32 sites ~300 CPUs first use for production**
- December 2003 LCG-2
  - **Deployed in January to 8 core sites**
  - **Introduced a pre-production service for the experiments**
  - **Alternative packaging (tool based and generic installation guides)**
- Mai 2004 -> now monthly incremental releases (not all distributed)
  - **Driven by the experiences from the data challenges**
  - **Balance between stable operation and improved versions (driven by users)**
  - **2-1-0, 2-1-1, 2-2-0, 2-3-0, (2-4-0)**
  - **(Production services RBs + BDIIs patched on demand)**
  - **> 110 sites (3-5 failed)**

Adding Sites

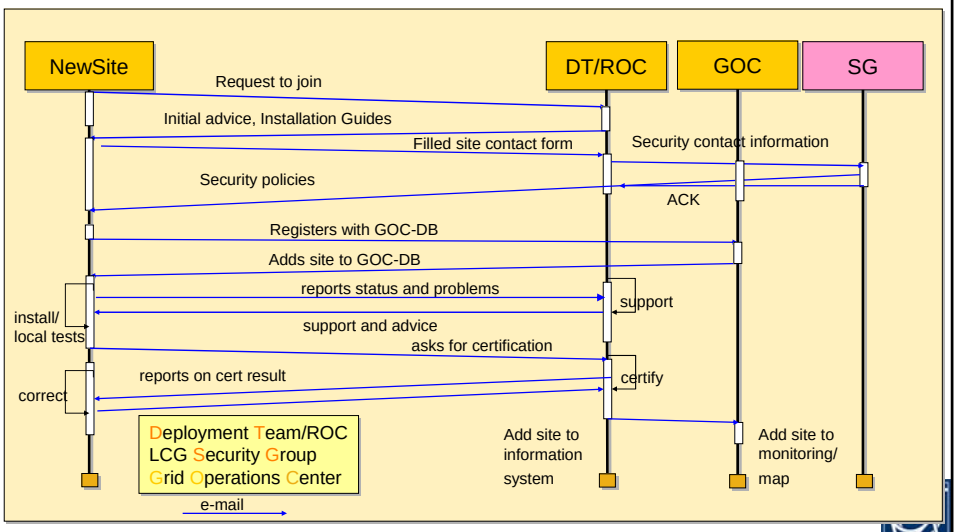
- Group **or** Regional Operation Center (ROC)
- Si page
- Sites tool based installation

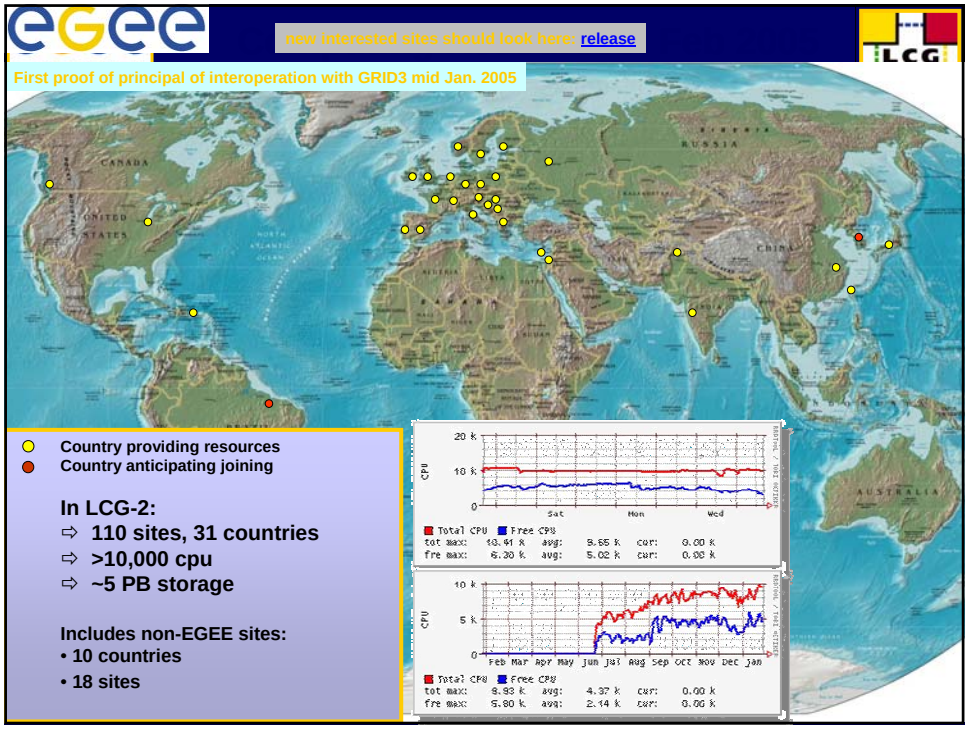
failed 3-5 times

work times

  - sites are daily re-certified and pr ANNAH
- Experiments install their software a ir IS
- **Adding new sites is now a quite smo**
- **Process fundamentally wrong:**
  - sites do not sign an agreement on "best practic
- New process: https://edms.cern.ch/docume 3/

Adding a Site





Preparing a Release

- Monthly process
  - Gathering of new material
  - Prioritization
  - Integration of items on list
  - Deployment on testbeds
  - First tests
    - feedback
  - Release to EIS testbed for experiment validation
  - Full testing (functional and stress)
    - feedback to patch/component providers
    - final list of new components
  - Internal release (LCFGng)
- On demand
  - Preparation/Update of release notes for LCFGng
  - Preparation/Update of manual install documentation
  - Test installations on GIS testbeds
  - Announcement on the LCG-Rollout list

C&T Certification & Testing

GDB Grid Deployment Board

EIS Experiment Integration Support

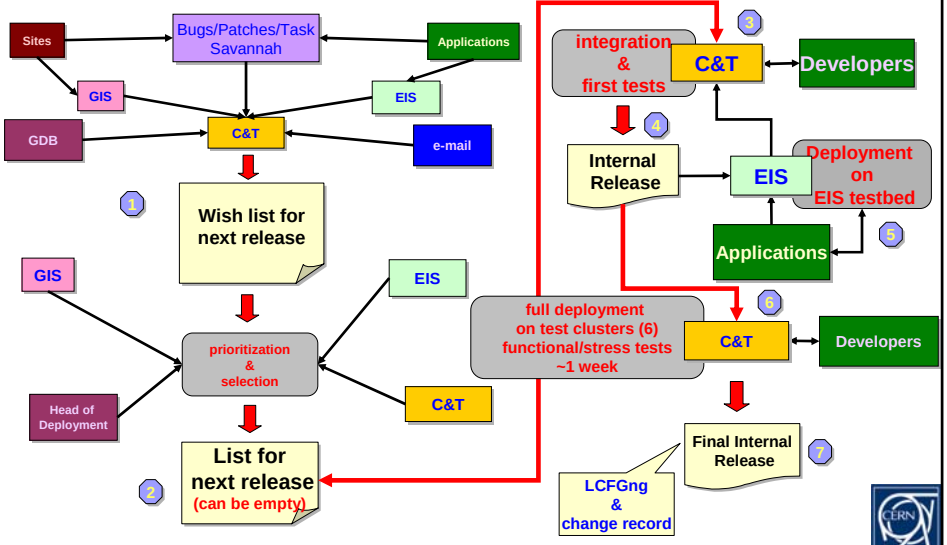
Applications

GIS Grid Infrastructure Support

ROCs/RCs Sites

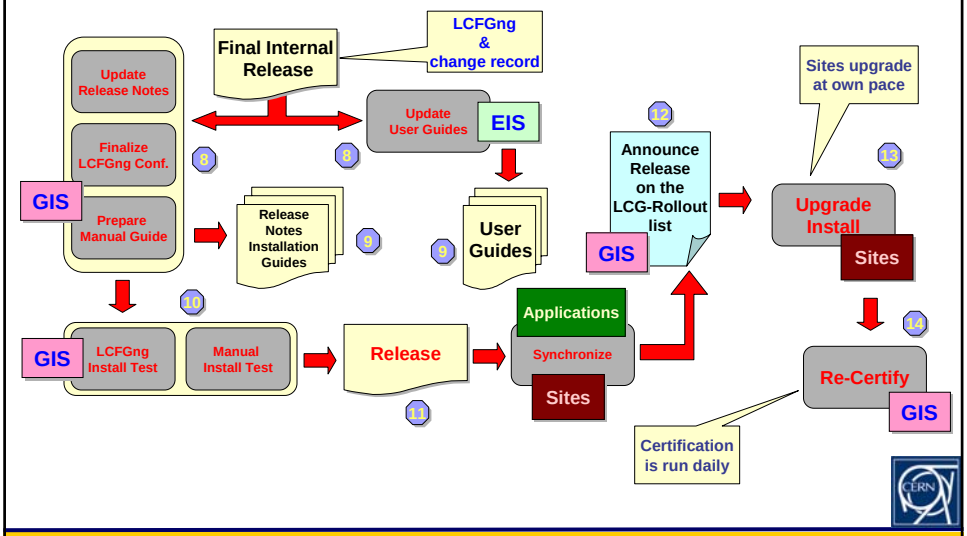
Preparing a Release

Initial List,Prioritization,Integration,EIS,StressTest



Preparing a Release

Preparations for Distribution, Upgrading



Experience

- Process was decisive to improve the middleware
- The process is time consuming (5 releases 2004)
  - Many sequential steps
  - Many different site layouts have to be tested
  - Format of internal and external releases differ
  - Multiple packaging formats (tool based, generic)
  - All components are treated equal
    - same level of testing for non vital and core components
    - new tools and tools in use by other projects are tested to the same level
- Process to include new components is not transparent
- Timing for releases difficult
  - users: now sites: scheduled
- Upgrades need a long time to cover all sites
  - some sites had problems to become functional after an upgrade

Additional Input

- Data Challenges
  - client libs need fast and frequent updates
  - core services need fast patches (functional/fixes)
  - applications need a transparent release preparation
  - many problems only become visible during full scale production
- Installation tool not available for new OS versions
- Configuration a major problem on smaller sites
- Operations Workshop
  - T2 sites can handle major upgrades only every 3 month
  - sites need to give input in the selection of new packages
- gLite releases need to be deployed
  - software already partially tested by JRA1
    - certification will need fewer iterations
  - preproduction service
    - replaces part of the certification process

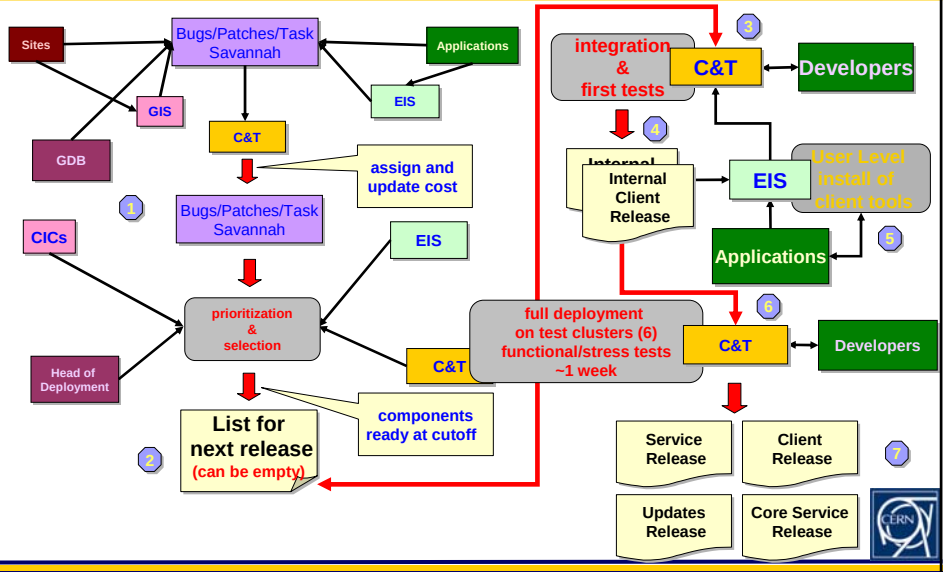
Changes I

- **Simple Installation/Configuration Scripts**
  - **YAIM (Yet Another Installation Method), semi automatic,simple configuration**
    - **all configuration for a site are kept in one file**
  - **APT (Debian) based installation of middleware RPMs**
    - **simple dependency management**
    - **updates (automatic or on demand)**
  - **Client libs packaged in addition as user space tar-ball**
    - **can be installed like application software**
- **Process (in development 2-4-0 is last release with old process)**
  - ***new process to gather and prioritize new packages***
    - ***formal***
    - ***tracking tool with priorities assigned to the packages***
    - ***cost to completion assigned (time of specific individual) at cut of day***
    - **selection process with participation of applications, sites and deployment**
    - **work will continue based on priority list between releases (rolling)**

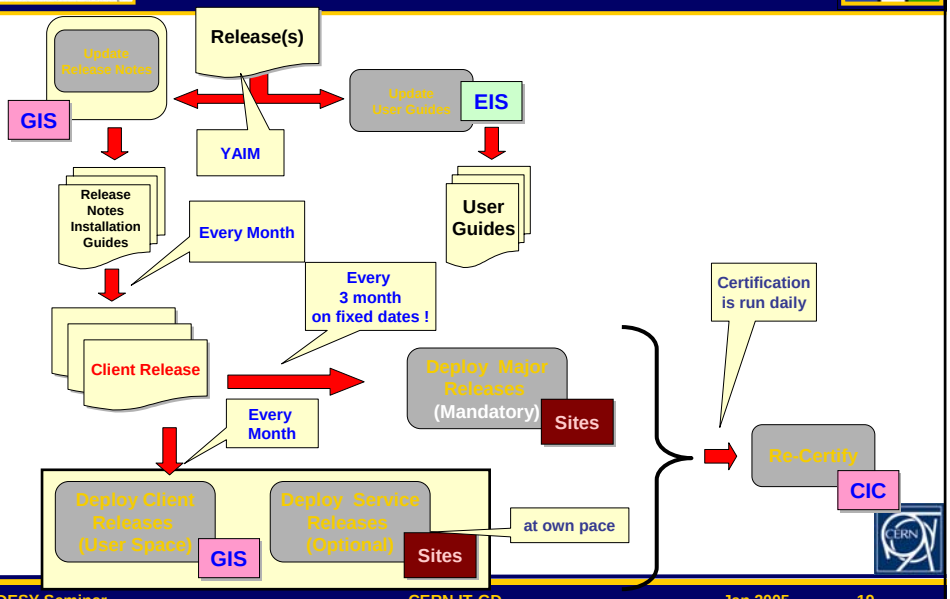
Changes II

- different release types
  - client libs
  - services (CE, SE)
  - core services (RB, BDII,..)
  - major releases (configuration changes, RPMs, new services)
  - updates (bug fixes) added any time to specific releases
  - non critical components will be made available with reduced testing
- Fixed release dates for major releases (allows planning)
  - every 3 month, sites have to upgrade within 3 weeks
- Minor releases every month
  - based on ranked components available at a specific date in the month
  - not mandatory for T2s to follow
    - client libs will be installed as application level software
  - early access to pre-releases of new software for applications
    - client libs. will be made available on selected sites
    - services with functional changes will be installed on EIS testbed
    - **early feedback from applications**

New Process (simplified)

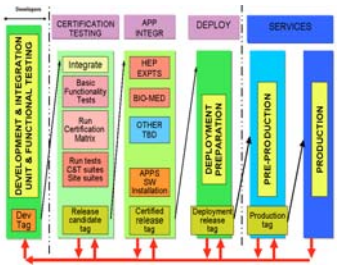


New Deployment



gLite

- Differences
  - unit and functional test already performed by JRA1
  - release cycle by JRA1
- New Sequence
  - Certification Testbed (CERN)
    - installation/config tests
    - rerun functional tests (to validate configuration)
    - synthetic stress tests
  - Preproduction Service
    - Sites
      - Krakow, FZK, IN2P3, CNAF, Padua, Bari, NIKHEF, SNIC, Protvino-IHEP, UOM ,LIP, PIC, RAL
      - sites test installation and configuration
    - Applications test their production software and give feedback on reliability and functionality
- Status
  - Documentation of process is in draft state
  - Certification Testbed
    - pre-release installed
  - Preproduction Service
    - sites have installed current LCG2 releases as a platform for the gLite components
  - Porting of tests and collection of existing tests (JRA1, NA4, ARDA) started



Summary of Release Process

- Certification of the middleware was the essential tool to improve its quality
- Early access to new releases was crucial for applications
- Process has to undergo evolutionary changes
  - software matures
    - certification becomes more complex (shift to applications)
  - scale (110 sites)
    - releases with radical changes become very hard to deploy
  - usage (production)
    - some uniformity and fast spread of fixes is expected by applications
- Preproduction Service for gLite
  - currently building up
    - new releases have to be introduced to subsets of the sites (staged)
  - feedback from applications essential to prioritize the work

Impact of Data Challenges

- Large scale production effort of the LHC experiments
  - test and validate the computing models
  - produce needed simulated data
  - test experiments production frame works and software
  - test the provided grid middleware
  - test the services provided by LCG-2
- All experiments used LCG-2 for part of their production

General Observations

- Real production is different from certification tests
  - Usage characteristics of real production can't be simulated
  - (Detailed) planning is of limited use
    - time better spend with small(er) scale pilot DCs
- Time matters
  - Delays in support and operation are deadly during DCs
  - Several iterations needed to get it right
  - Communication between sites, operations, and experiments matters
    - not all players handled DCs with same priority (communication problem)
- Middleware matures quickly during DCs
  - Scalability, robustness, functionality
  - Several of the experiments will do continues production from now on
- Probing concurrently multiple "dimensions" of the system can be confusing
  - Problems with different components mix
    - leads to confuse cause and effect
  - Dedicated "Service Challenges" will help to get a clear matrix

Problems during the data challenges

- All experiments encountered on LCG-2 similar problems
- LCG sites suffering from configuration and operational problems
  - not adequate resources on some sites (hardware, human..)
  - this is now the main source of failures
- **Load balancing between different sites is problematic**
  - jobs can be "attracted" to sites that have no adequate resources
  - modern batch systems are too complex and dynamic to summarize their behavior in a few values in the IS ----> Team working on evolution of the GLUE schema
- **Identification and location of problems in LCG-2 is difficult**
  - distributed environment, access to many logfiles needed.....
  - status of monitoring tools
- **Handling thousands of jobs is time consuming and tedious**
  - Support for bulk operation is not adequate
- **Performance and scalability of services**
  - storage (access and number of files)
  - job submission
  - information system
  - file catalogues
- **Services suffered from hardware problems**
  - (no fail over) ------> implemented workarounds for BDII and RBs

Outstanding Middleware Issues

- Collection: Outstanding Middleware Issues
  - Important: 1st systematic confrontation of required functionalities with capabilities of the existing middleware
    - Some can be patched, worked around,
    - Those related to fundamental problems with underlying models and architectures have to be input as essential requirements to future developments (EGEE)
- Middleware is now not perfect but quite stable
  - Much has been improved during DC's
    - A lot of effort still going into improvements and fixes
    - Easy to deploy and operate storage management still an issue
      - especially for Tier 2 sites
      - effort to adapt dCache
      - effort in simple Disk Pool Manager
- Production with Geant4 in December
  - 86% success for complex application

Operational issues (selection)

- **Slow response from sites**
  - **Upgrades, response to problems, etc.**
  - **Problems reported daily – some problems last for weeks**
- **Lack of staff available to fix problems**
  - **Vacation period, other high priority tasks**
- **Various mis-configurations (see next slide)**
- **Lack of configuration management – problems that are fixed re-appear**
- **Lack of fabric management (mostly smaller sites)**
  - **scratch space, single nodes drain queues, incomplete upgrades, ….**
- **Lack of understanding**
  - **Admins reformat disks of SE …**
- **Provided documentation often not (carefully) read**
- **new activity to simplify configuration and installation**
  - **simpler way to install middleware on farm nodes**
  - **opens ways to maintain middleware remotely in user space**
- **Firewall issues –**
  - **often less than optimal coordination between grid admins and firewall maintainers**
- **openPBS problems**
  - **Scalability, robustness (switching to torque helps)**

Site (mis) - configurations

- Site mis-configuration was responsible for **most** of the problems that occurred during the experiments Data Challenges. Here is a non-complete list of problems:

integrated all common small problems into

ONE

BIG PROBLEM

  - – Default user shell environment too big
  -
- **Only partly related to middleware complexity**

Operating Services for DCs

- Multiple instances of core services for each of the experiments
  - **separates problems, avoids interference between experiments**
  - **improves availability**
  - **allows experiments to maintain individual configuration**
  - **addresses scalability to some degree**
- Monitoring tools for services currently not adequate
  - **tools under development to implement control system**
- Access to storage via load balanced interfaces
  - **CASTOR**
  - **dCache**
- Services that carry “state” are problematic to restart on new nodes
  - **needed after hardware problems, or security problems**
  - **For RBs processing and storage of state has been separated**
- “State Transition” between partial usage and full usage of resources
  - **required change in queue configuration (faire share, individual queues/VO)**

Support during the DCs

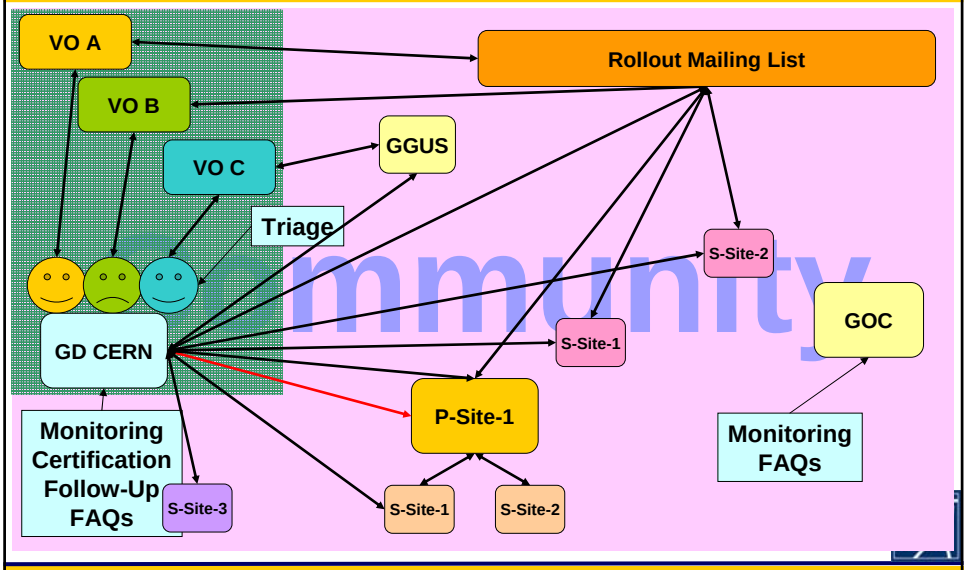
- User (Experiment) Support:
  - GD at CERN worked very close with the experiments production managers
  - Informal exchange (e-mail, meetings, phone)
    - "No Secrets" approach, GD people on experiments mail lists and vice versa
      - **ensured fast response**
    - tracking of problems tedious, but both sites have been patient
    - clear learning curve on **BOTH** sites
    - LCG GGUS (grid user support) at FZK became operational after start of the DCs
      - **due to the importance of the DCs the experiments switch slowly to the new service**
    - Very good end user documentation by GD-EIS
    - Dedicated testbed for experiments with next LCG-2 release
      - **rapid feedback, influenced what made it into the next release**
- Installation and site operations support:
  - GD prepared releases and supported sites (certification, re-certification)
  - Regional centres supported their local sites (some more, some less)
  - **Community style help via mailing list (high traffic!!)**
  - FAQ lists for trouble shooting and configuration issues: Taipei RAL

Support during the DCs

- Operations Service:
  - RAL (UK) is leading sub-project on developing operations services
  - Initial prototype http://www.grid-support.ac.uk/GOC/
    - Basic monitoring tools
    - Mail lists for problem resolution
    - Working on defining policies for operation, responsibilities (draft document)
    - Working on grid wide accounting
  - Monitoring:
    - GridICE (development of DataTag Nagios-based tools)
    - GridPP job submission monitoring
    - Information system monitoring and consistency check http://goc.grid.sinica.edu.tw/gstat/
  - CERN GD daily re-certification of sites (including history)
    - escalation procedure
    - tracing of site specific problems via problem tracking tool
    - tests core services and configuration

Problem Handling

Operation (most cases)



LCG Workshop on Operational Issues

- Motivation
  - LCG -> (LCG&EGEE) transition requires changes
  - Lessons learned need to be implemented
  - Many different activities need to be coordinated
- 02 - 04 November at CERN
  - >80 participants including from GRID3 and NorduGrid
  - Agenda: Here
  - 1.5 days of plenary sessions
    - describe status and stimulate discussion
  - 1 day parallel/joint working groups
    - very concrete work,
    - results into creation of task lists with names attached to items
  - 0.5 days of reports of the WG

LCG Workshop on Operational Issues
WGs I

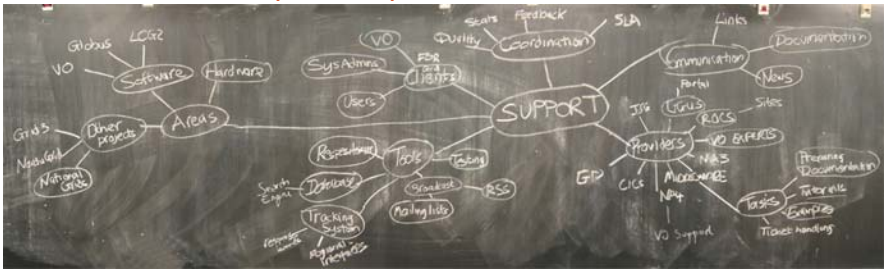
- Operational Security
  - Incident Handling Process
  - Variance in site support availability
  - Reporting Channels
  - Service Challenges
- Operational Support
  - Workflow for operations & security actions
  - What tools are needed to implement the model
  - “24X7” global support
    - sharing operational load (taking turns)
  - Communication
  - Problem Tracking System
  - Defining Responsibilities
    - problem follow-up
    - deployment of new releases
  - Interface to User Support

LCG Workshop on Operational Issues
WGs II

- Fabric Management
  - System installations
  - Batch/scheduling Systems
  - Fabric monitoring
  - Software installation
  - Representation of site status (load) in the Information System
- Software Management
  - Operations on and for VOs (add/remove/service discovery)
  - Fault tolerance, operations on running services (stop,upgrades, re-starts)
  - Link to developers
  - What level of intrusion can be tolerated on the WNs (farm nodes)
    - application (experiment) software installation
  - Removing/(re-adding) sites with (fixed)troubles
  - Multiple views in the information system (maintenance)

LCG Workshop on Operational Issues
WGs III

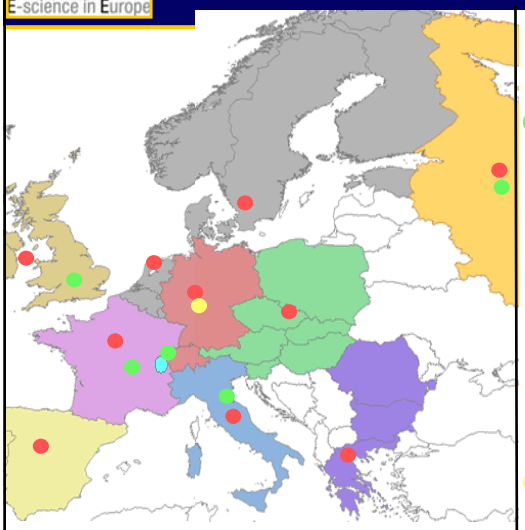
- User Support
  - Defining what "User Support" means
  - Models for implementing a working user support
    - need for a Central User Support Coordination Team (CUSC)
      - mandate and tasks
    - distributed/central (CUSC/RUSC)



LCG Workshop on Operational Issues
Summary

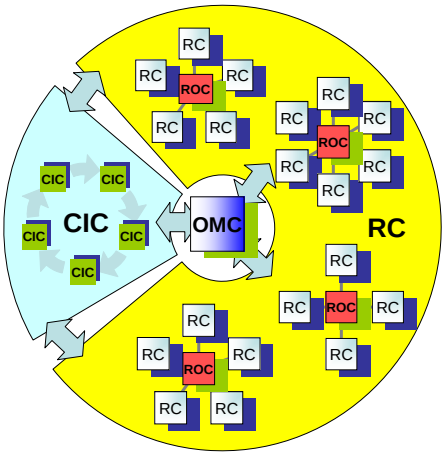
- **Very productive workshop**
- Partners (sites) assumed responsibility for tasks
- Discussions very much focused on **practical** matters
- Some problems ask for architectural changes
  - gLite has to address these
- It became clear that not all sites are created equal
- Removing troubled sites is inherently problematic
  - removing storage can have grid wide impact
- Key issues in all aspects is to define split between:
  - Local,Regional and Central control and responsibility
- All WGs discussed communication

SA1 – Operations Structure



- Operations Management Centre (OMC):
  - At CERN – coordination etc
- Core Infrastructure Centres (CIC)
  - Manage daily grid operations – oversight, troubleshooting
  - Run essential infrastructure services
  - Provide 2nd level support to ROCs
  - UK/I, Fr, It, CERN, + Russa (M12)
  - Taipei also run a CIC
- Regional Operations Centres (ROC)
  - Act as front-line support for user and operations issues
  - Provide local knowledge and adaptations
  - One in each region – many distributed
- User Support Centre (GGUS)
  - In FZK – manage PTS – provide single point of contact (service desk)
  - Not foreseen as such in TA, but need is clear

Grid Operations



• RC = Resource Centre

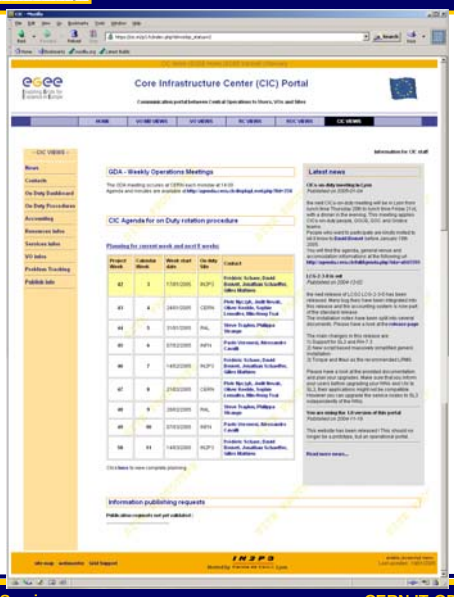
- The *grid* is flat, but
- *Hierarchy* of responsibility
  - Essential to scale the operation
- CICs act as a single Operations Centre
  - Operational oversight ***(grid operator)*** responsibility
  - rotates weekly between CICs
  - Report problems to ROC/RC
  - ROC is *responsible* for ensuring problem is resolved
  - ROC oversees regional RCs
- ROCs responsible for organising the operations in a region
  - Coordinate deployment of middleware, etc
- CERN coordinates sites not associated with a ROC

Grid Operations & Monitoring

- CIC-on-duty
  - Responsibility rotates through CIC's – one week at a time
  - Manage daily operations – oversee and ensure
    - Problems from all sources are tracked (entered into PTS)
    - Problems are followed up
    - CIC-on-duty hands over responsibility for problems
  - Hand-over in weekly operations meeting
- Daily operations:
  - Checklist
  - Various problem sources: monitors, maps, direct problem reports
- Next step:
  - Continue to develop tools to generate automated alarms and actions

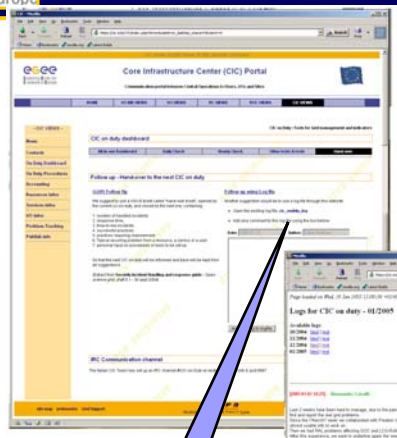
CIC Web Portal

http://cic.in2p3.fr/



- CIC portal:
- Centralized tool for egee operators
  - Communication tools
  - Link to monitoring tools
  - Summary of status
  - Provide a repository
    - Knowledge: configurations, published sites data, Faqs
    - Procedures and processes
    - Existing tools

CIC-on-duty (Operation Center)



Follow-up

Log file

Cic–on-duty agenda
weekly shifts – 8/5
GGUS, ROC User-support, mail
Monitor, diagnose troubles
Contact site administrators, ROC
Problem Tracking Tool

GDA meetings
COD meetings

CIC Operations

https://cic.in2p3.fr/index.php?id=cic&subid=cic_dash2

Cic operational procedure

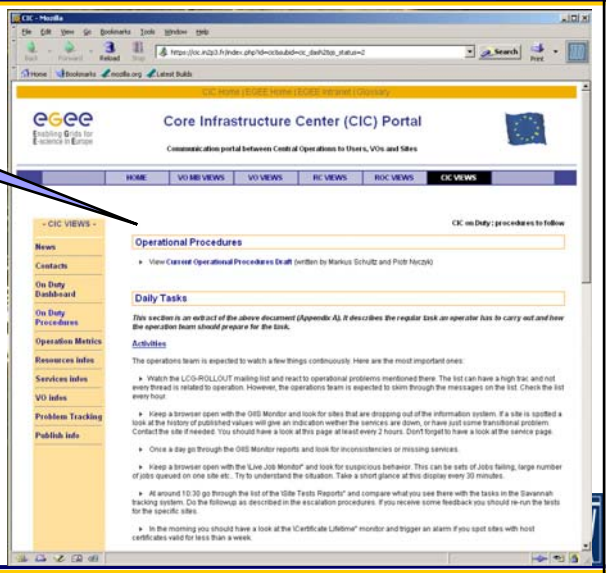
Ops Procedure

[This procedure is provided by OMC]

High level abstraction
of core tools results

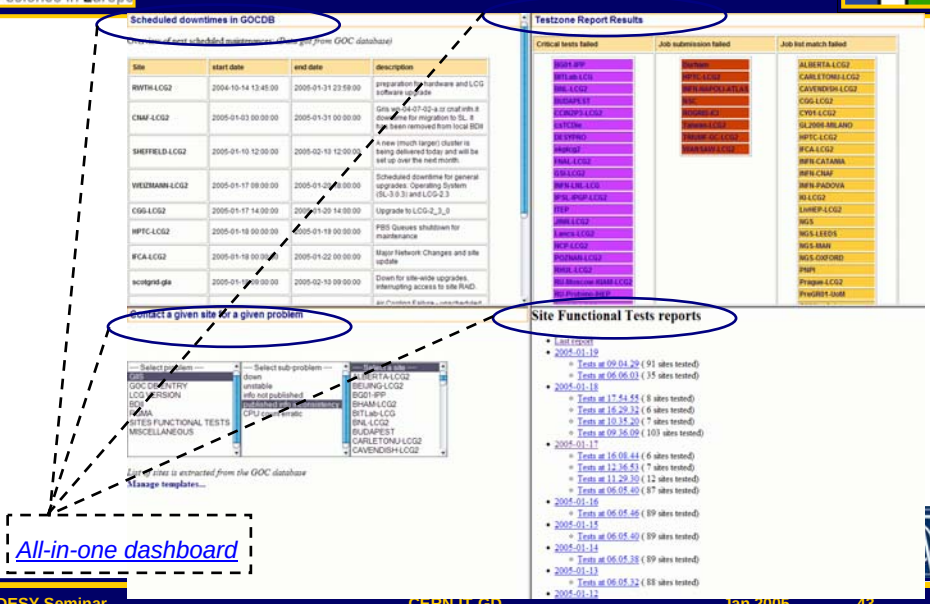
Link all existing tools

- monitoring
- diagnosis
- communication
  - mail
  - follow up log



Cic-on-duty Dashboard

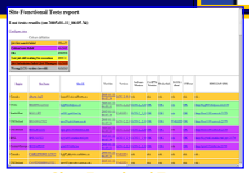
https://cic.in2p3.fr/pages/cic/framedashboard.html



Selection of Monitoring tools

GIIS Monitor

GIIS Monitor graphs


Sites Functional Tests

GOC Data Base

Scheduled Downtimes

Live Job Monitor

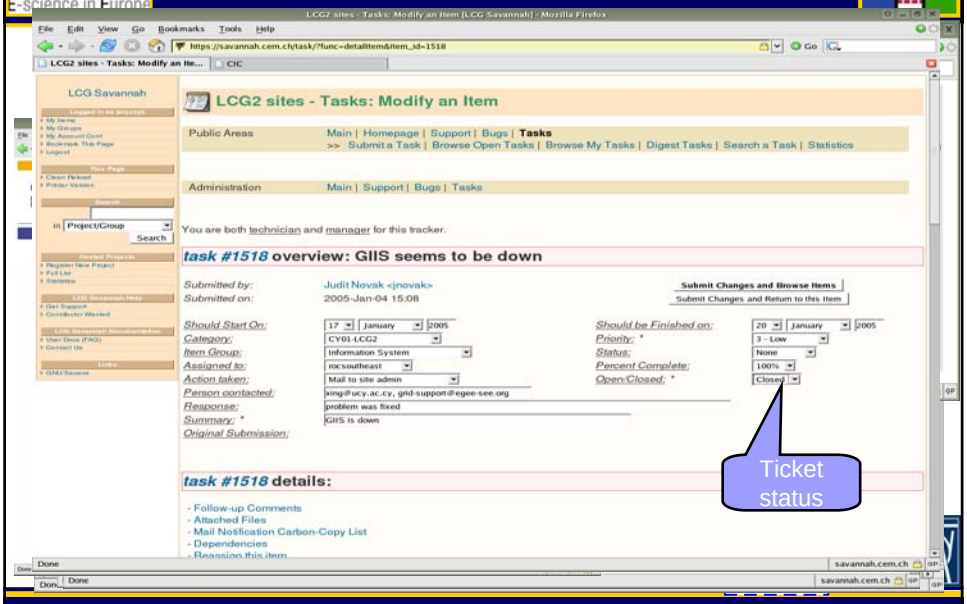
GridIce – VO view

GridIce – fabric view

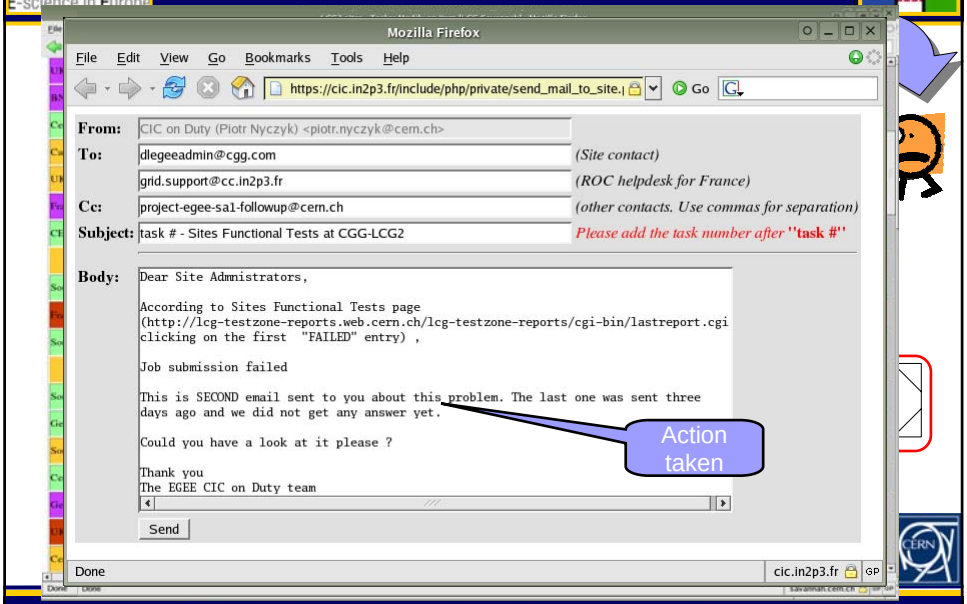
Certificate Lifetime Monitor

Note: Those thumbnails are links and are clickable

Scenarios of escalation : 1/2



Scenarios of escalation : 2/2



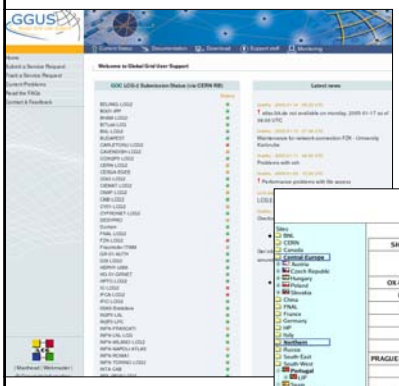
Incident closure

(6)

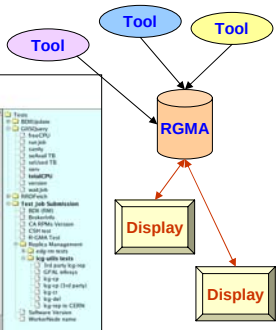
Monitoring tool

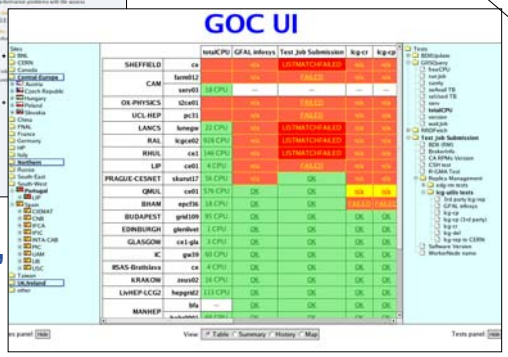
(1)

Next steps



Unification of information coming from various test tools: Common schema for R-GMA, allows complex alarms (sql)





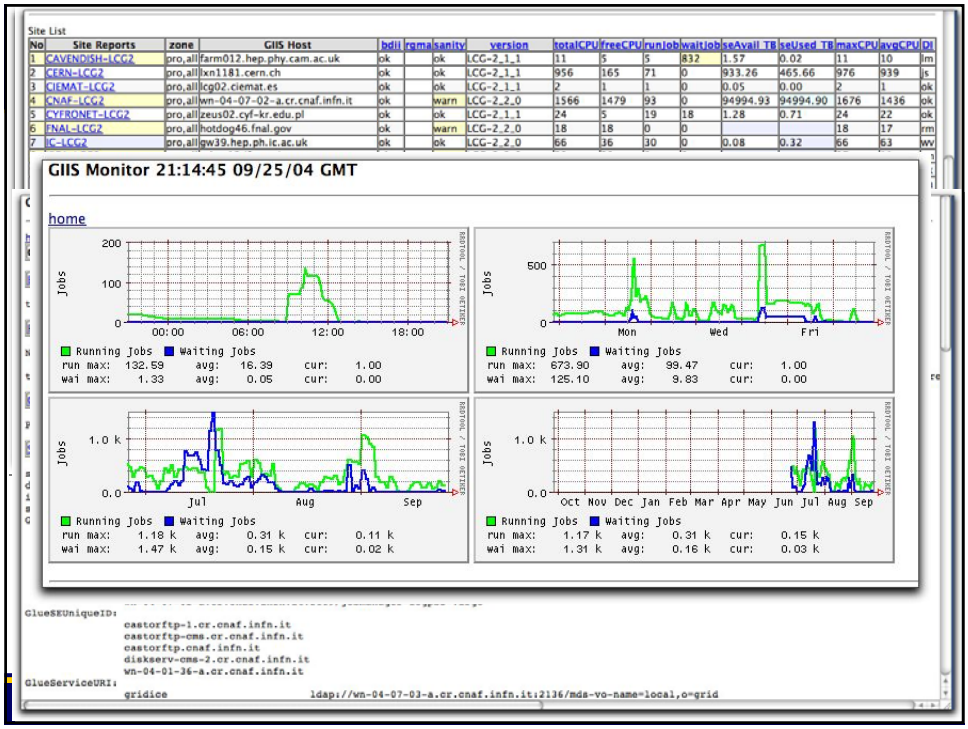
Move to GGUS, link with ROCs

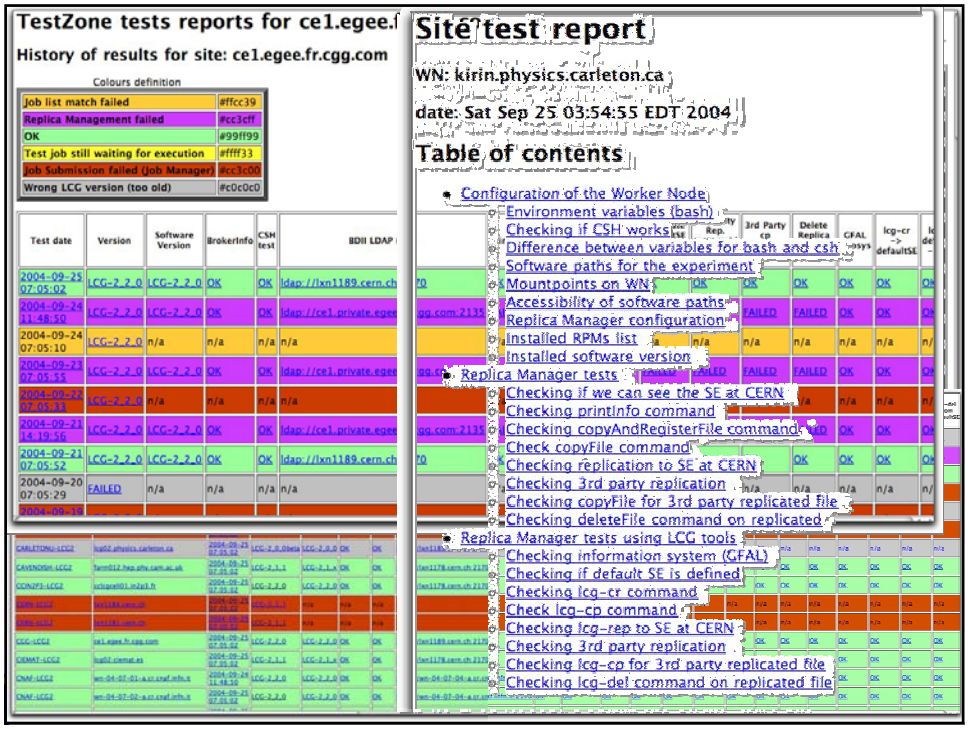
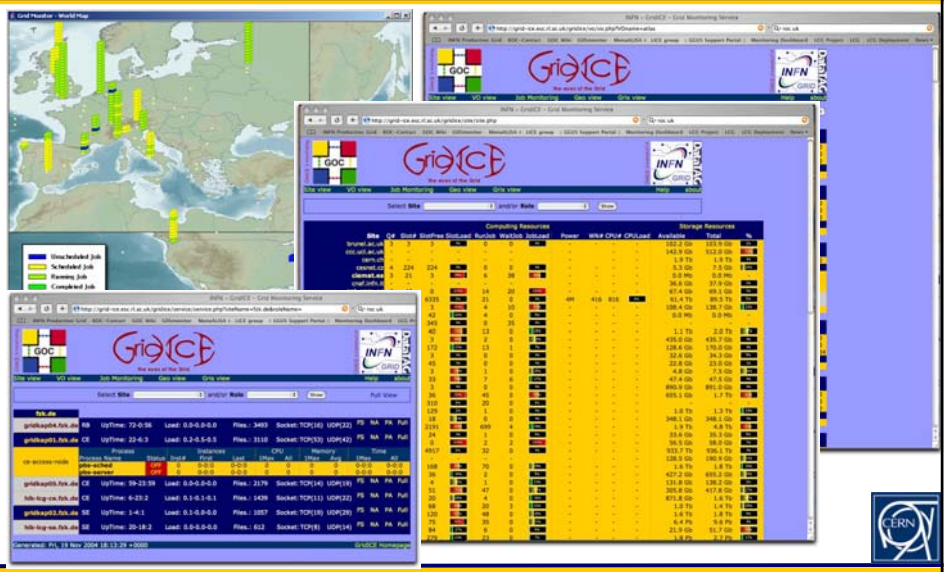
http://www.ggus.org

[This display based on RGMA is provided by the Russian federation]

Summary

- LCG-2 services have been supporting the data challenges
  - Many middleware problems have been found – many addressed
  - Middleware itself is reasonably stable (within the architecture)
- Biggest outstanding issues are related to providing and maintaining stable operations
- Future middleware has to take this into account:
  - Must be more manageable, trivial to configure and install
  - Management and monitoring must be built into services from the start on
- Operational Workshop has started many activities
- Operation moved to EGEE structure
- gLite and LCG2 will coexist for some time


Site List
GIIS Monitor 21:14:45 09/25/04 GMT
home
Running Jobs
Waiting Jobs
GlueSEUniqueID:
castorftp-1.cr.cnaf.infn.it
castorftp-cms.cr.cnaf.infn.it
castorftp.cnaf.infn.it
diskserv-cms-2.cr.cnaf.infn.it
wn-04-01-36-a.cr.cnaf.infn.it
GlueServiceURI:
gridice
ldap://wn-04-07-03-a.cr.cnaf.infn.it:2136/mds-vo-name=local,o=grid


TestZone tests reports for ce1.egee.f
History of results for site: ce1.egee.fr.cgg.com
Colours definition
Job list match failed
Replica Management failed
OK
Test job still waiting for execution
Job Submission failed (Job Manager)
Wrong LCG version (too old)
Site test report
WN: kirin.physics.carleton.ca
date: Sat Sep 25 03:54:55 EDT 2004
Table of contents
Configuration of the Worker Node
Environment variables (bash)
Checking if CSH works
Difference between variables for bash and csh
Software paths for the experiment
Mountpoints on WN
Accessibility of software paths
Replica Manager configuration
Installed RPMs list
Installed software version
Replica Manager tests
Checking if we can see the SE at CERN
Checking printInfo command
Checking copyAndRegisterFile command
Check copyFile command
Checking replication to SE at CERN
Checking 3rd party replication
Checking copyFile for 3rd party replicated file
Checking deleteFile command on replicated
Replica Manager tests using LCG tools
Checking information system (GFAL)
Checking if default SE is defined
Checking lcg-cr command
Check lcg-cp command
Checking lcg-rep to SE at CERN
Checking 3rd party replication
Checking lcg-cp for 3rd party replicated file
Checking lcg-del command on replicated file

Some More Monitoring