Enabling Grids for E-sciencE

Advances in Grid Operations

Dr. Nicholas Thackray

CHEP 2009, Prague

www.eu-egee.org

e-infrastructure

But what is grid operations?

- **Infrastructures**
  - Production service
  - Pre-production service
  - Pilot services
- **Procedures**
  - Site SLAs
  - Middleware release process
  - Site registration
  - VO registration
  - etc.
- **Communications**
  - Daily, weekly, bi-weekly, meetings for all stakeholders
- **Grid security**
- **Interoperations**
  - At all levels
- **User support**
- **Operations tools**
  - Operations Portal
  - GOC database
  - Monitoring
    - Dashboards
    - SAM
    - Gstat
    - GridView
    - etc.
  - Trouble ticketing system (GGUS)
  - Accounting
- ***… and much more!***

In a nutshell

The goal of grid operations is the provisioning of a large-scale, production grid infrastructure that interoperates at many levels, offering reliable services to a wide range of applications

So, a 3 point check list

✓ Large-scale

- 265 sites
- 600 PB of storage
- 110,000 CPUs
- 50,000 concurrent jobs
- 270,000 jobs / day

All without any increase in the staffing levels of operations

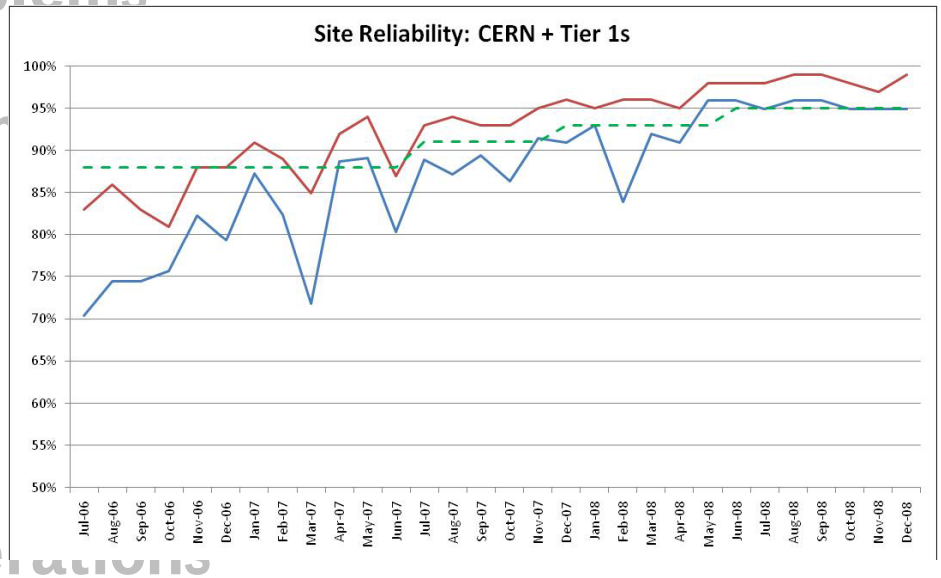
? Reliable services

Room for improvement?

End of 2007: where was there room for improvements?

- Reliability of “user services”
- Handling of grid problems
- Communications with users and sites
- Security
- Accounting
- Structure of grid operations

- **Reliability of "user services"**
- Handling of grid problems
- Communications with
- Security
- Accounting
- Structure of grid operations



Reliability of "User Services"

- **User Service: the whole chain, not just individual middleware services**
  - Encouraged improvements in site reliability
    - Calculate and publish site availability & reliability figures
    - Implemented site Service Level Agreement (SLA)
    - Improving the VO ID cards; so sites know what their VOs expect
  - Efforts to improve quality of the middleware reaching production
    - Constant, incremental improvements to middleware release process
    - Reformation of pre-production service (*PPS*)
      - *No more "2 week delay"*
      - *Make it useful for the VOs*

Reliability of "User Services"

- Requested features to reduce number of SPoFs and to survive glitches
  - Service discovery via Information System
  - Ability to put each middleware service on a cluster
  - Caching, retries, etc.
- Requests to middleware developers for improvements to make debugging easier (logging, error messages)
- Robustness of core middleware service instances
  - Ensuring high availability instances in all regions; e.g. top-level BDII
- Commissioned interoperability testbed (EGEE / OSG)
  - To catch middleware updates that break interoperability

Reliability of "User Services"

- **Improvements for the future**
  - Staged rollouts of middleware; on representative sites
  - Roll-back mechanism for failed middleware updates
  - Set the bar higher for availability and reliability
  - Add more peer grids to the interoperability testbed

- Reliability of “user s
- **Handling of grid problems**
- Communications with users and sites
- Security
- Accounting
- Structure of grid operations

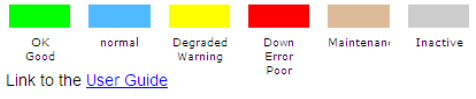
Handling of grid problems

- **Need to spot the problems as quickly as possible…**
  - Service Availability Monitoring (*SAM*) improvements:
    - Software release process for SAM improved
    - Monitoring of the monitoring
    - Monitoring of core grid services (WMS, top BDIIs, etc.)
  - Site level monitoring: Nagios
    - YAIM-installable package and auto-configures
    - Standard set of probes provided
    - SAM tests available to site Nagios through Active-MQ® message bus
    - Now: ~40 instances monitoring ~150 sites

Handling of grid problems

- Experiment dashboards

- Site dashboard



- GridMap

Handling of grid problems

- **...and act on them rapidly**
  - Improvements in the trouble ticketing tool (GGUS) and procedures
    - Ability for users to route their tickets directly to sites
    - Ability for users to escalate tickets
    - Alarm and team tickets
    - LHCOPN support
  - Regular reviews of metrics for 1st, 2nd & 3rd line support
  - Periodic testing of ALARM GGUS tickets to WLCG tier-1 sites
  - Escalation reports to catch stagnating tickets

Handling of grid problems

- **Improvements for the future**
  - *Multi-Level Monitoring* (MLM):

    An integrated, holistic approach to grid infrastructure monitoring
  - GStat 2
  - Grid Configuration Monitoring (GCM)
    - a.k.a. The job wrapper tests

- Reliability of “user services”
- Handling of grid problems
- **Communications with users and sites**
- Security
- Accounting
- Structure of grid operations

Communications

- **Introduced daily WLCG grid operations meetings**
- **Wiki entries created from the results of solved tickets**
- **GOCDB downtimes appear automatically in site Nagios instances**
- **Improved the site/service downtime notification**
  - ~~Spamming~~ ➔ user defined RSS feeds
- **GOC DB: Can now enter downtimes for grid operations tools**

Communications

- **Improvements for the future**
  - A graphical calendar view of site downtimes
    - *a la* ATLAS and LHCb
  - Automatic notification of changes to VO ID cards

- Reliability of “user services”
- Handling of grid problems
- Communications with users and sites
- **Security**
- Accounting
- Structure of grid operations

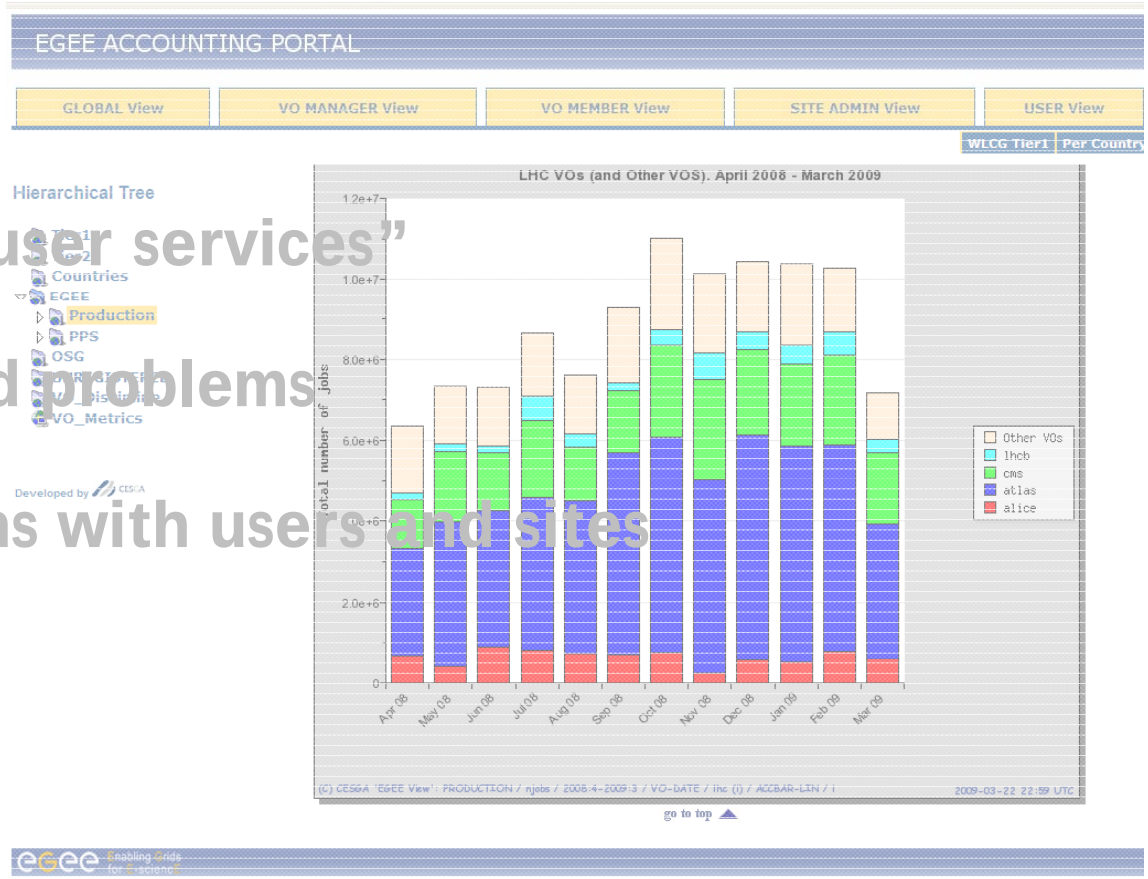
Security

- **Distributed security operations**
  - Team's activities also distributed and lead by the regions
- **Security drills**
  - Now showing improved results from the sites to respond to security incidents
- **Improved collaboration with peer grids and NRENs**
- **Comprehensive security training events organized during EGEE conferences**

Security

- **Improvements for the future**
  - More structured information flow to manage cross-grid security incidents
  - Integration of the NRENs in the incident response process
  - Implementing metrics to measure the quality of the response from the sites
  - Reviewing security training material, including a new website
  - Implementing additional security monitoring tools in collaboration with the Operations Automation Team (OAT)

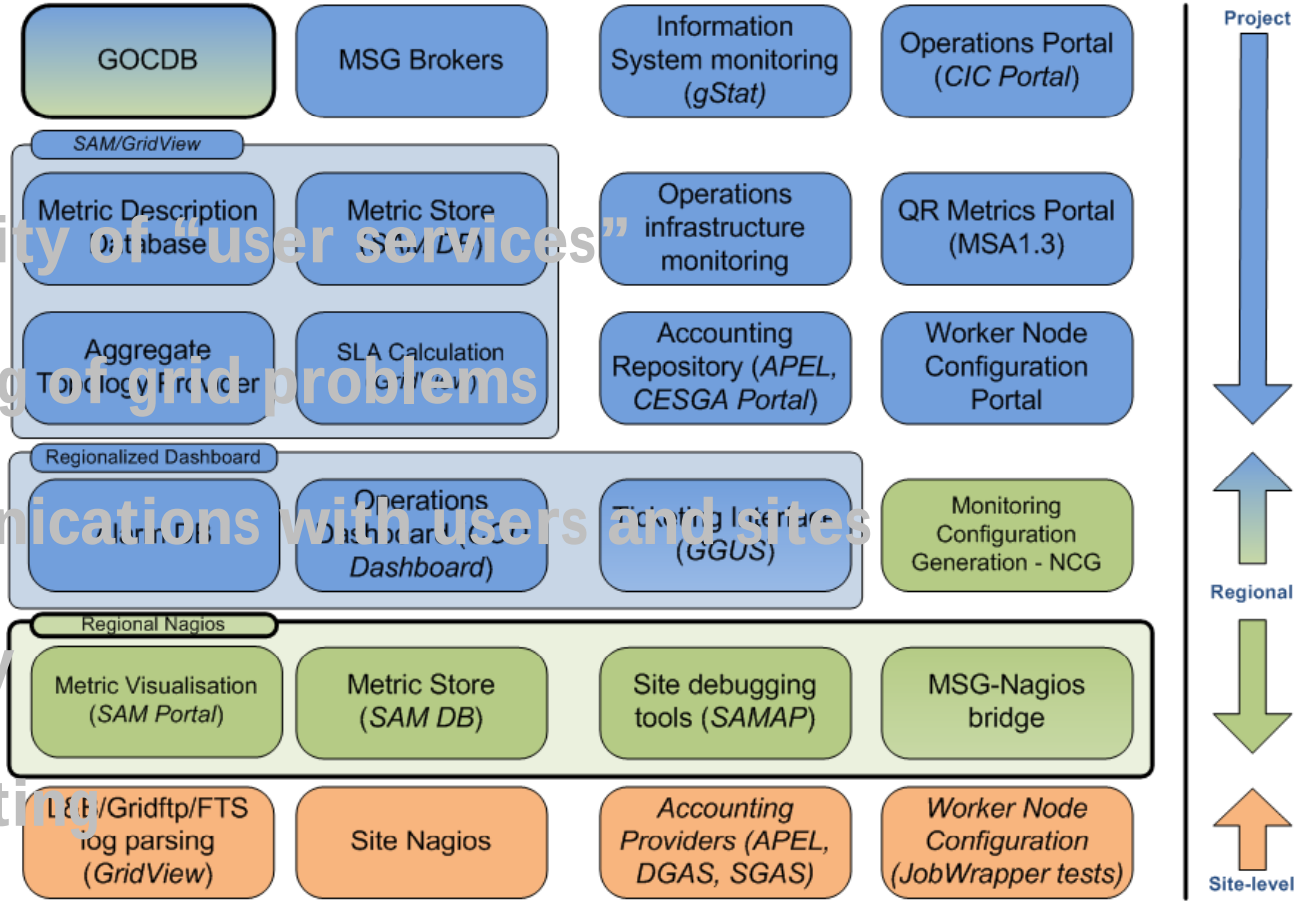
- Reliability of “user services”
- Handling of grid problems
- Communications with users and sites
- Security
- **Accounting**
- Structure of grid operations



Accounting

- **Recent improvements:**
  - Scalability: Allowing sites to publish backlogs in chunks so as not to overload memory
  - SAM critical tests used to highlight sites not publishing data
  - DGAS publishing changed from R-GMA to direct db insertion
  - SGAS started publishing accounts for NDGF tier-1
- **The publishing of FQAN of jobs was rolled out**
  - Developed in 2007 , 84% of sites now publishing
- **The publishing of UserDN was rolled out**
  - But not enabled due to lack of policy covering legal issues

- Reliability of "user services"
- Handling of grid problems
- Communications with users and sites
- Security
- Accounting
- **Structure of grid operations**



Structure of Grid Operations

- **Recent improvements**
  - "Regionalization": to make Regional Operations Centres (ROC) (and by further distribution NGIs) have independent infrastructures so that at the end of the EGEE III project they can function independently (April 2010)
- **Improvements for the future**
  - Completion of the regionalization
  - Much more automation of tools, monitoring, etc.
  - A common strategy and architecture for how the operations tools will work together
    - Introduction of Active-MQ® messaging system
  - This is all being coordinated by a dedicated working group: the *Operations Automation Team* (OAT)

Acknowledgements

- John Gordon (RAL)
- Romain Wartel (CERN)
- Gilles Mathieu (RAL)
- John Shade (CERN)
- Cyril L'Orphelin (IN2P3)
- James Casey (CERN)
- Jamie Shiers (CERN)
- Torsten Antoni (FZK)
- Maria Dimou (CERN)
- Diana Bosio (CERN)
- David Collados (CERN)
- Hélène Cordier (IN2P3)