ROD and COD operational model

Marcin Radecki, Małgorzata Krakowian
EGI COD

ACC CYFRONET AGH

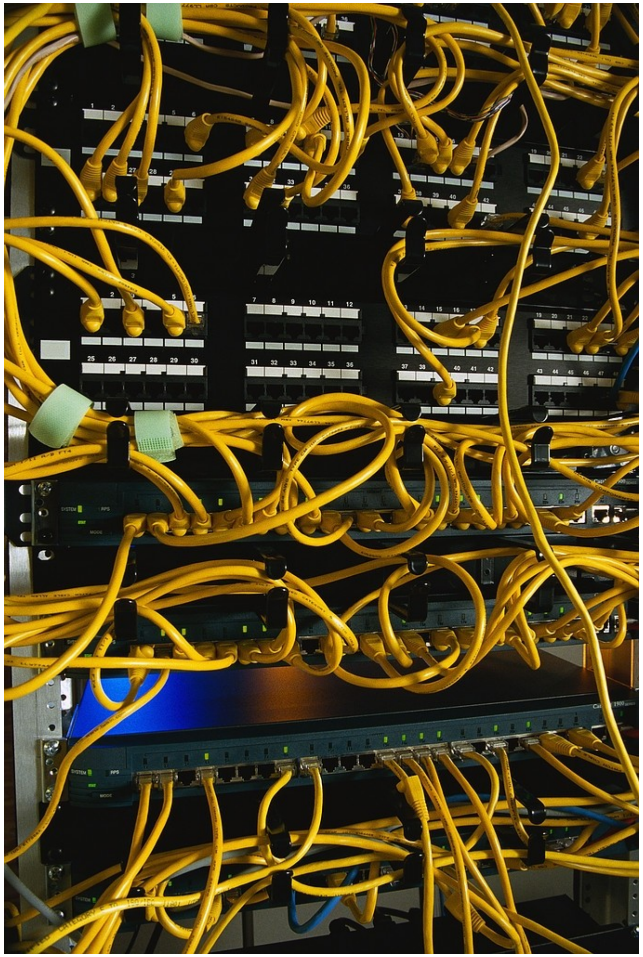
www.eu-egee.org

Agenda

- **Organizational structure of grid**
- **Highlights on what is important for keeping the infrastructure stable**
- **Operational model**
  - procedures
  - tools
- **Operational model metrics**

What do we deal with?

- ~330 sites from 59 countries
- almost 100k CPU
- tens of PB storage space managed by a variety of SM systems
- thousands of users
- tens of thousands of running jobs



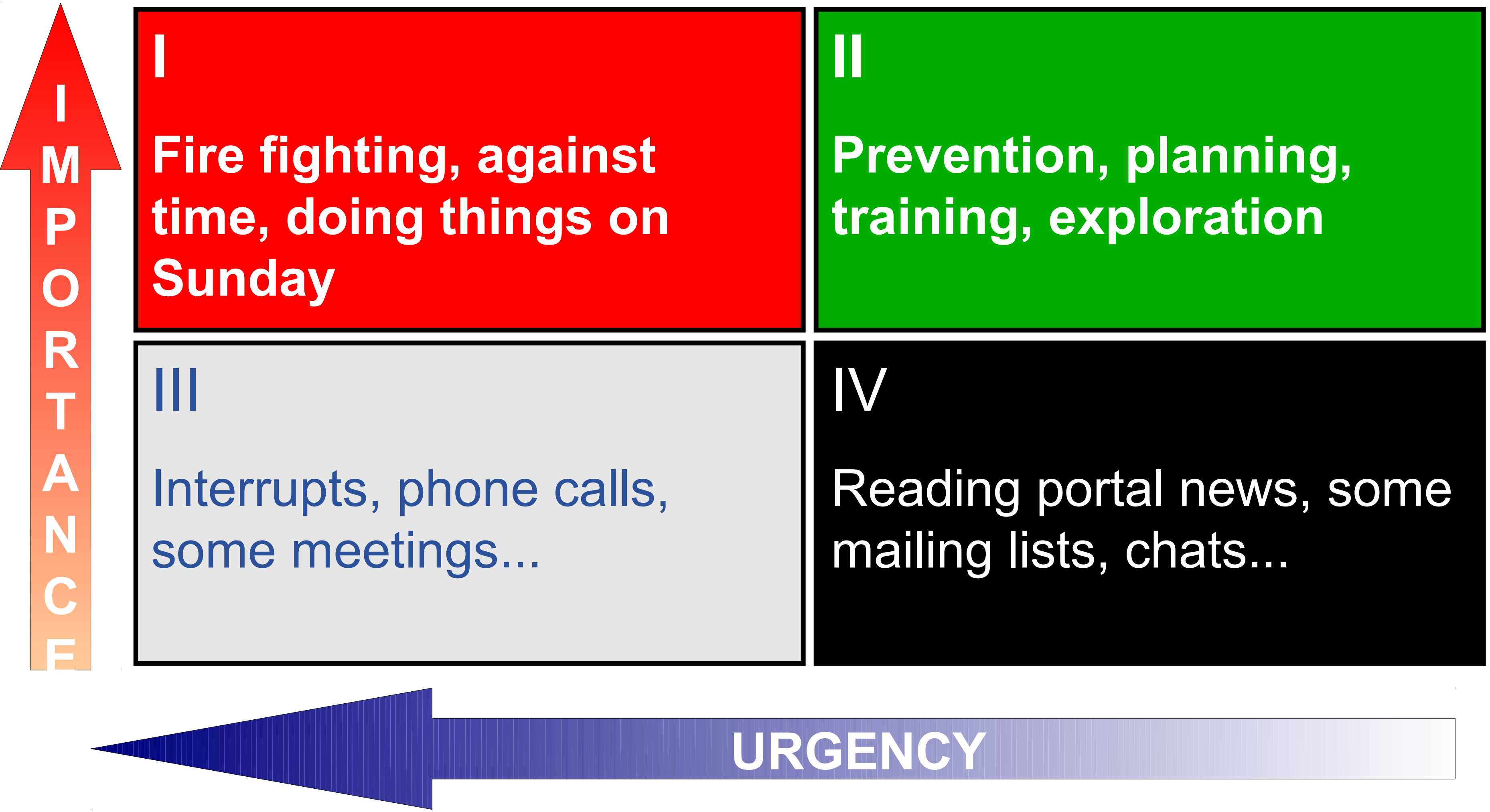
Grid is a complex system which requires staff and procedures in order to operate

Organizational structure

- **Hierarchical**
  - In EGEE
    - 1 Operations Coordination Centre
    - 11 instances of Regional Operations Centres
    - ~300 Grid Sites
  - In EGI
    - European Grid Initiative
    - **~40 NGIs**
    - ~300 Grid Sites
- **Role of NGI**
  - manage grid operations within its borders
  - provide helpdesk facility
  - provide operations support (ROD)
  - provide infrastructure monitoring
  - ...interface the above with EGI

- **ROCs were similar in terms of**
  - resources
  - responsibility
  - middleware
- **NGIs are different in many ways**
  - funding
  - resources
  - number of sites
  - internal organization
- **All this must be adapted to supply unified way of operations**
  - operational support
  - infrastructure monitoring
  - troube ticket processing

Principles of being effective



I

Fire fighting, against time, doing things on Sunday

II

Prevention, planning, training, exploration

III

Interrupts, phone calls, some meetings...

IV

Reading portal news, some mailing lists, chats...

Keeping infrastructure stable

- **notice a problem ASAP**
- **diagnose**
- **act precisely (without dead ends and U-turns)**
- **The above requires:**
  - tools (monitoring, dashboard)
  - well defined procedures
    - instruction on how to proceed in case of a failure
    - cover all aspects, details, nuances
  - collaboration
    - exchange experience, pass knowledge, get help on-line

Spotting a problem in Grid

- **Service availability monitoring in Grid**
  - Services are remote – impact of computer network
  - Complexity of Grid middleware
    - monitoring functionality for the user (replica management)
    - ...vs. monitor atomic functionality
    - middleware error messages: https://twiki.cern.ch/twiki/bin/view/LCG/BestErrorMessages
  - Nagios – a monitoring system aware of the dependencies between functional components
    - do not tests services on a host if the host is not reachable
    - also a source of issues during transition from SAM to nagios...

Diagnose problem

- **What is reported to site admin?**
  - command which returned an error
  - error message e.g. (top 4): “CGSI-gSOAP: Error reading token data: Success”
- **Experience is indispensable**
  - **...or support**
  - **documentation**
  - **knowledge base etc.**

Fix it! Site admin's checklist

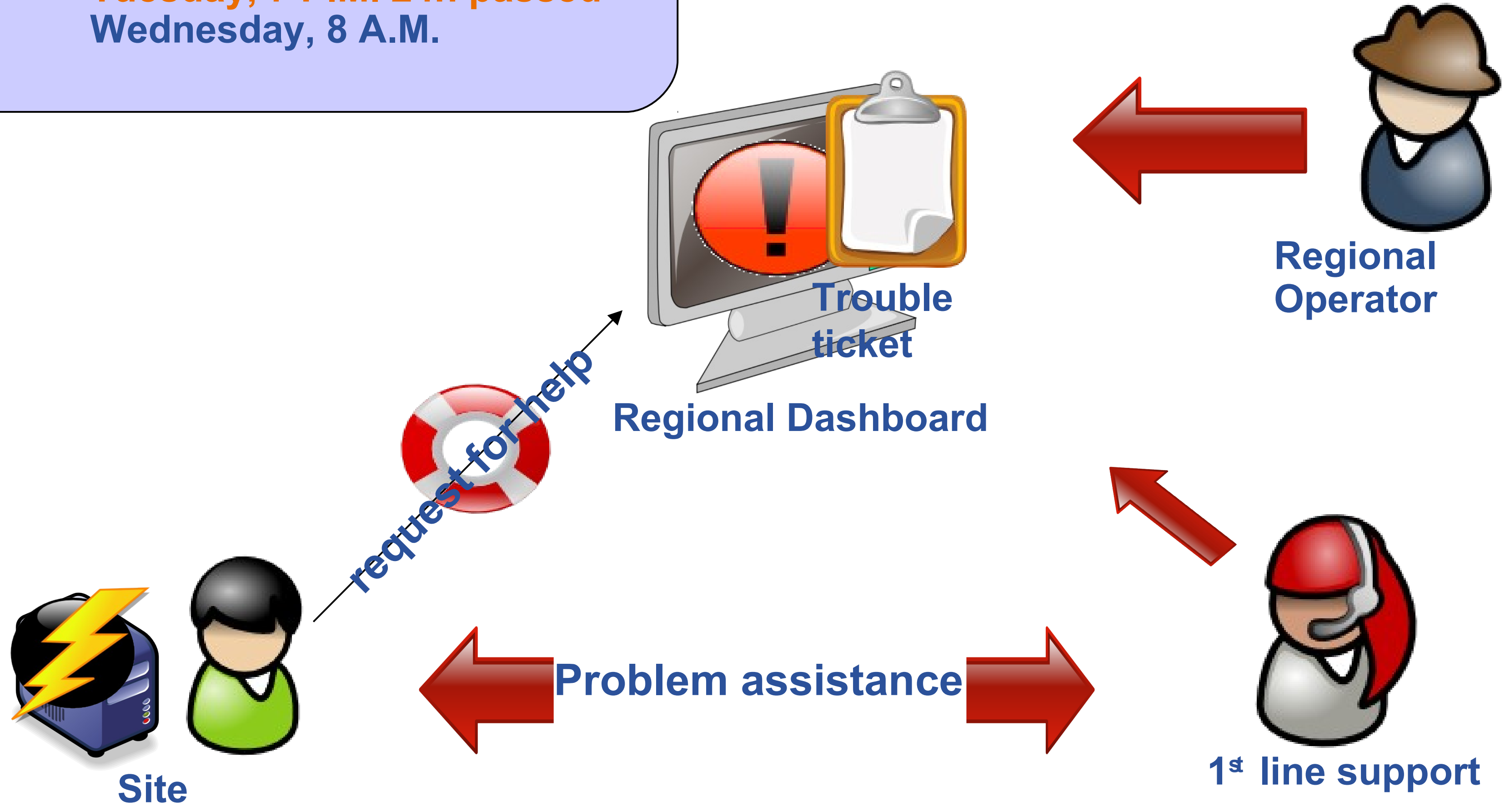
- Ideas that will not work
  - Search the error message and explanation in middleware manual
  - Ask the middleware developer for help
- Time consuming ideas
  - understand the software by yourself “Use the Source (code), Luke!”
- Practical, working (usually) solution
  - search the knowledge bases
    - http://goc.grid.sinica.edu.tw/gocwiki/SiteProblemsFollowUpFaq
    - https://weblog.plgrid.pl/baza-wiedzy/
    - some entries may be out of date
  - see if someone not stumbled already
    - in GGUS tickets – there is nice search engine, worse than knowledge base as may contain no solution
  - ask expert
    - your NGI 1st line support
    - post an e-mail to lcgrollout mailing list

Operations procedures

- **Indispensable for distributed systems**
  - collaboration principles must be defined
- **Define what to do in case of a service failure**
- **Actors**
  - Site Admin
  - ROD, Regional Operator on Duty
  - COD, Central Operator on Duty
- **Items to operate on**
  - alarm – problem reported by monitoring system. Contains info about time, localization of the failre. Appears in dashboard of ROD and COD.
  - (trouble) ticket – record of a problem handling. Is created when an alarm cannot be quickly turn off. Created in GGUS.

Handling operational emergencies

Monday, 7 P.M.
Tuesday, 8 A.M.
Tuesday, 9 A.M.
Tuesday, 7 P.M. 24h passed
Wednesday, 8 A.M.



Operations Support Model and Metrics

- **Model depends on timely actions**
  - first 24h – time for site & technical support team
  - [24,72) - time for ROD to clear the problem OR record it in GGUS
  - [72,∞) - model malfunction, COD comes into the game
  - ticket not handled on time (expiration date passed) → COD
  - ticket not solved in 30 days → COD
- **Metrics aim: indicate problems with operating model**
  - items not handled on time
  - items not handled according to procedures
  - assess workload on ROD & COD teams

COD workload

An „item” in the dashboard is either alarm or ticket that the relevant party (COD, ROD, 1st line) should take action upon.	
Description	Number of **items appearing in COD dashboard** indicates the amount of work that the operator has to deal with. It could also be **used to assess the quality of support process**. There should be no items in COD dashboard if the support process is working in a timely manner.
What is measured	Number of items in COD dashboard that needs immediate action, appearing on a given day. Items not done on a given day will be counted again the next day.
Purpose	To estimate the amount of daily work of COD operator and quality of support process.
Source of data	COD dashboard

ROD workload

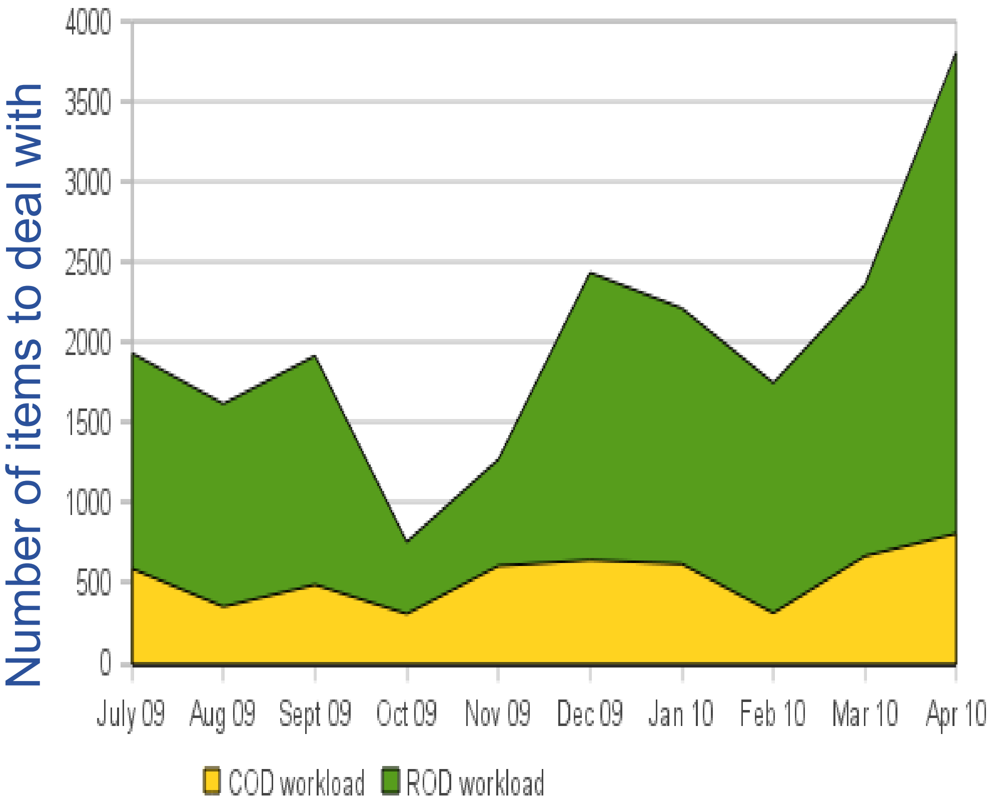
An „item” in the dashboard is either alarm or ticket that the relevant party (COD, ROD, 1st line) should take action upon.

Description	Number of **items appearing in ROD dashboard** indicates the amount of work that the operator has to deal with. In general it **cannot be used** to assess the quality of support process.
What is measured	Number of items in ROD dashboard that needs immediate action, appearing on a given day.
Purpose	To estimate the amount of daily work of ROD operator.
Source of data	Regional dashboard

Quality of regional support

Metric = (alarms_closed_with_OK/alarms_closed_in_total)	
Description	Regional ops. support staff can close an alarm if the actual state of the service is OK or some ERROR state. In general **they should fix problem** and close alarm only if the actual service state is OK.
What is measured	Fraction of alarms closed with OK status over some time period e.g. 1 month.
Purpose	Assess regional support quality, make sure model time rules are followed.
Source of data	Regional dashboard

Workload in General



- **Intermittent problems with operations tools in Sept.**

  EGEE'09

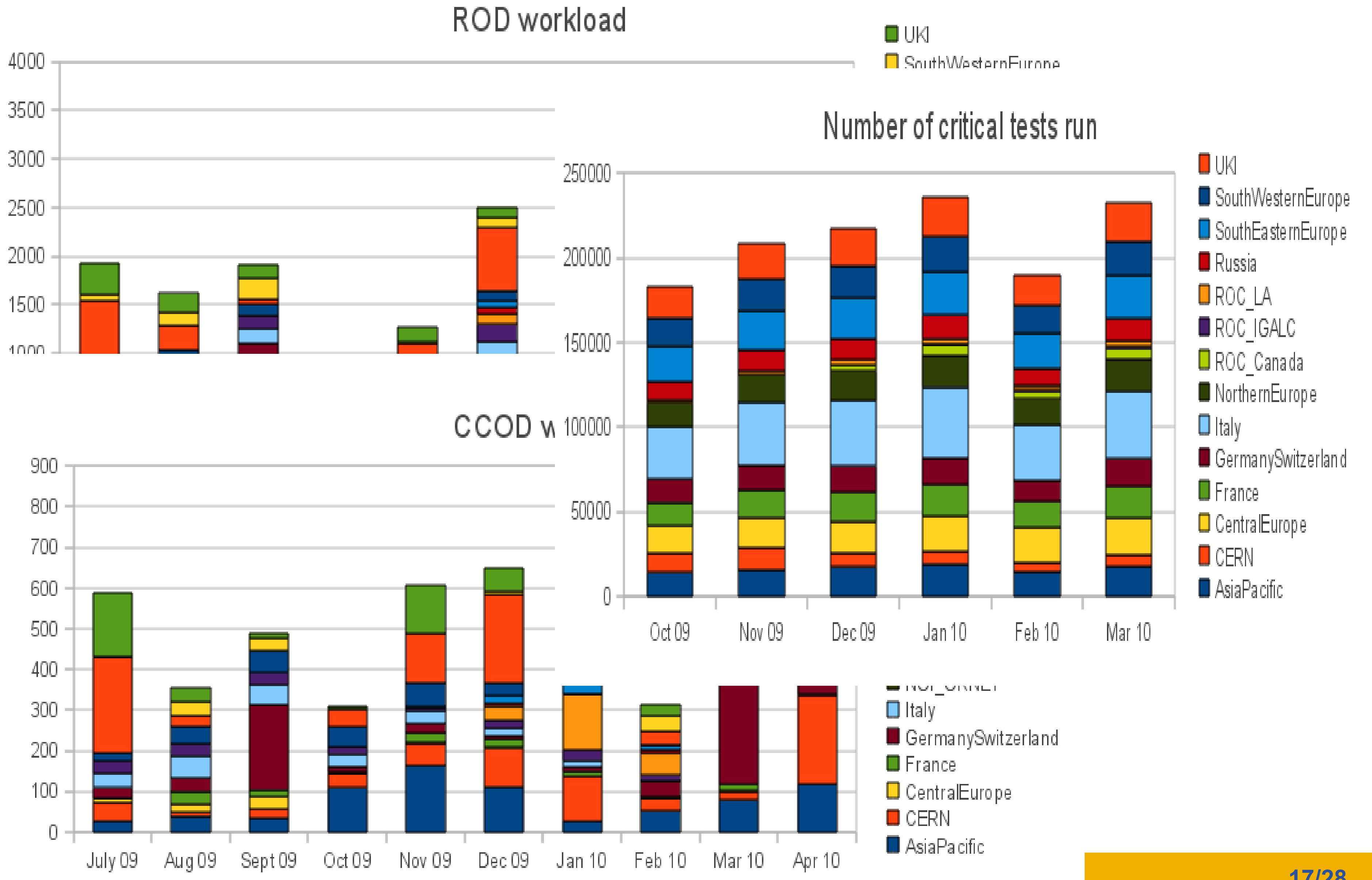
  Introduction of Cream-CE on 7.12.09

  Christmas period
  - less staffed
  - alarm ageing not sync. with

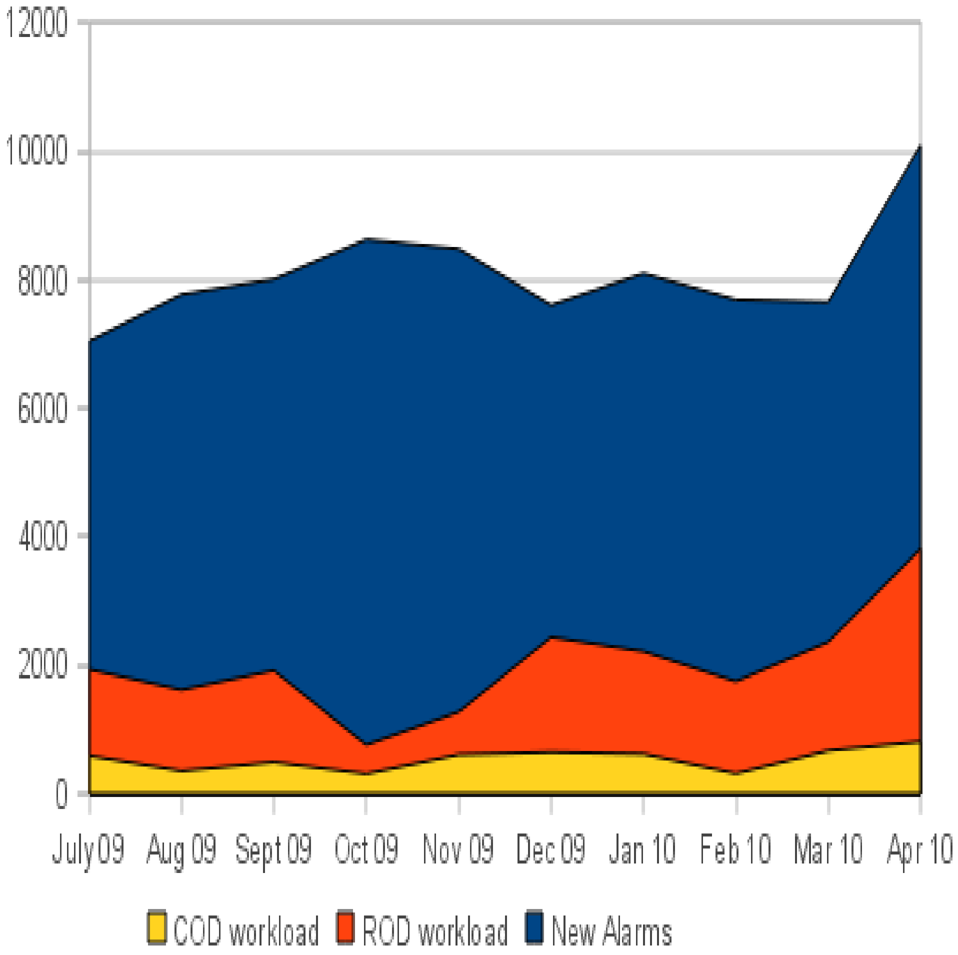
  March-April 2010
  - New monitoring system introduced
  - End of EGEE-III, staff change

  Conclusions
  - RODs do a lot of good job
  - Thanks that... COD workload is stable
  - Alarms should not age on bank holidays

Workload Origin

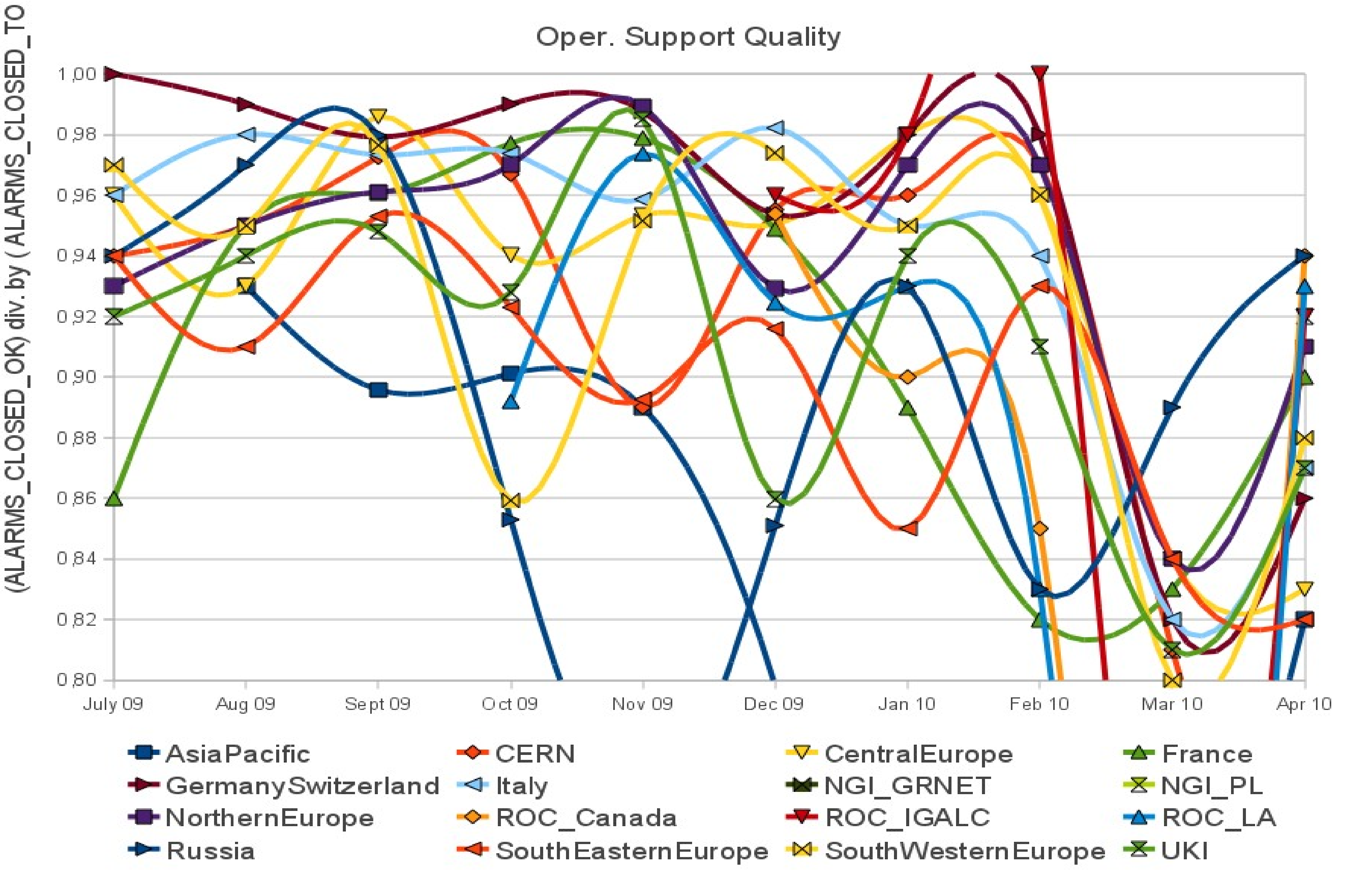


Operational Support Workload

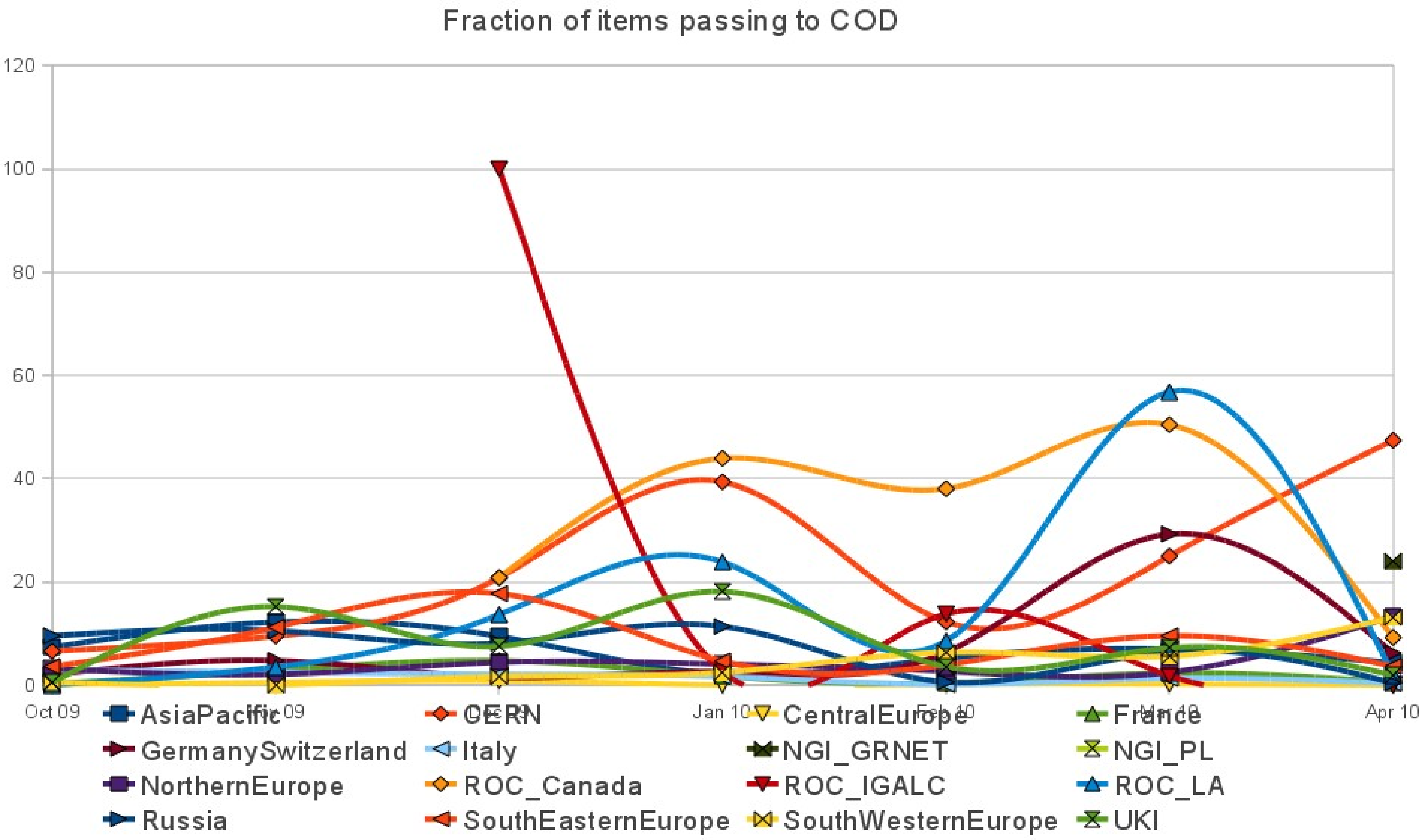


- **Note**
  - ROD/COD workload items are counted each day again until handled
  - Alarms (blue area) not cumulative
- **Making Cream-CE test critical**
  - 16.11.09 – request to add Cream-CE to critical tests
  - 7.12.09 – treshold of 75% passing, Cream-CE made critical
  - number of new alarms did not raise (April - ?)

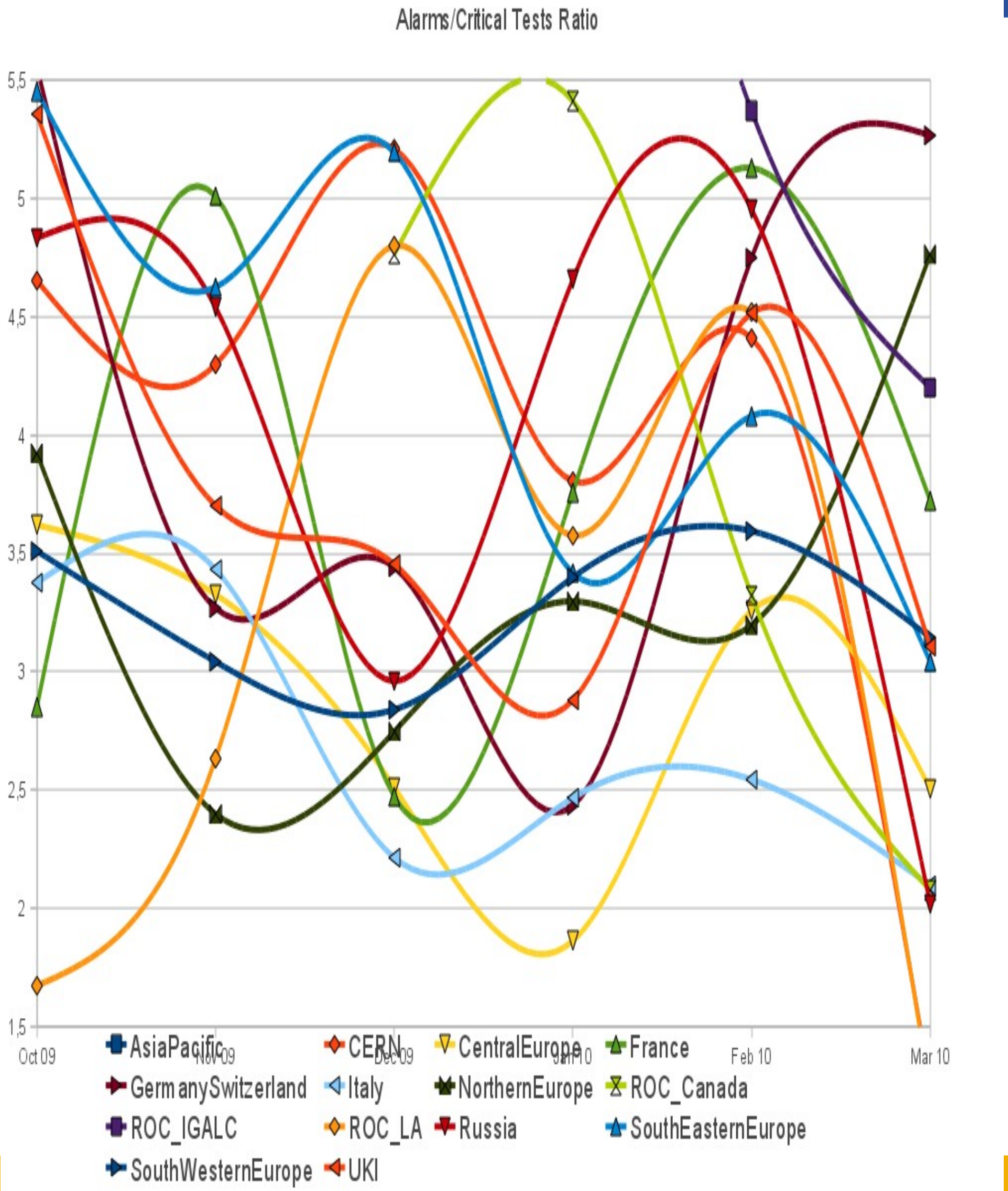
Regional Ops. Support Quality



Alarm: Trips from site to COD



Infrastructure Stability



Y axis

- (New_alarms/Number_of_Critical_tests)*100

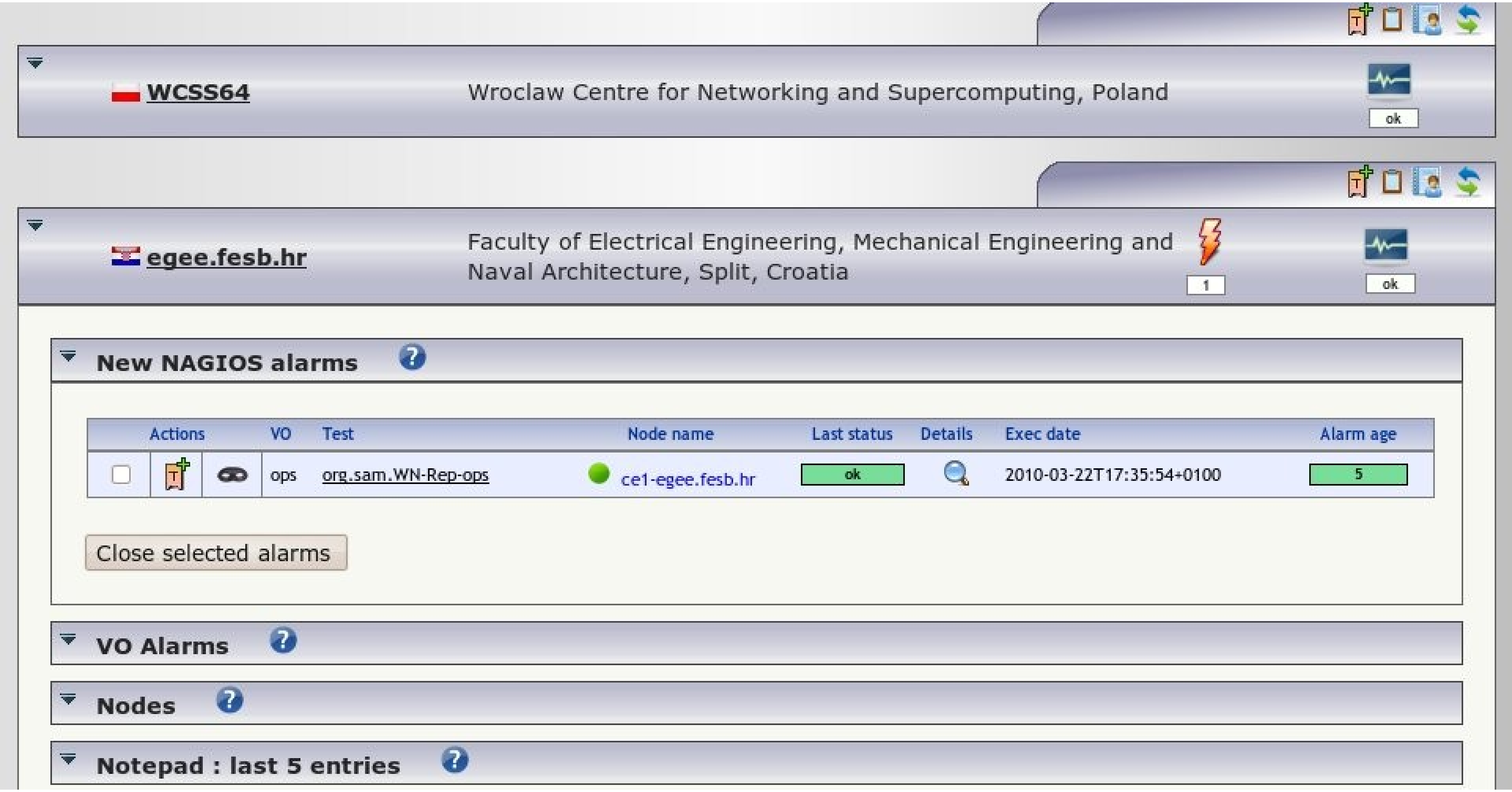
Interpretation

- how many alarms are generated from each 100 runs of critical test
- difference between 2,5 and 5 means that services fails 2 times more often

Sensitive for

- outages in monitoring system (less chances for new alarms)
- excessive use of SAMAP ;)

Operations dashboard



WCSS64 Wrocław Centre for Networking and Supercomputing, Poland

ok

egee.fesb.hr Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia

1

ok

New NAGIOS alarms

Actions			VO	Test	Node name	Last status	Details	Exec date	Alarm age
□			ops	org.sam.WN-Rep-ops	● ce1-egee.fesb.hr	ok	🔍	2010-03-22T17:35:54+0100	5

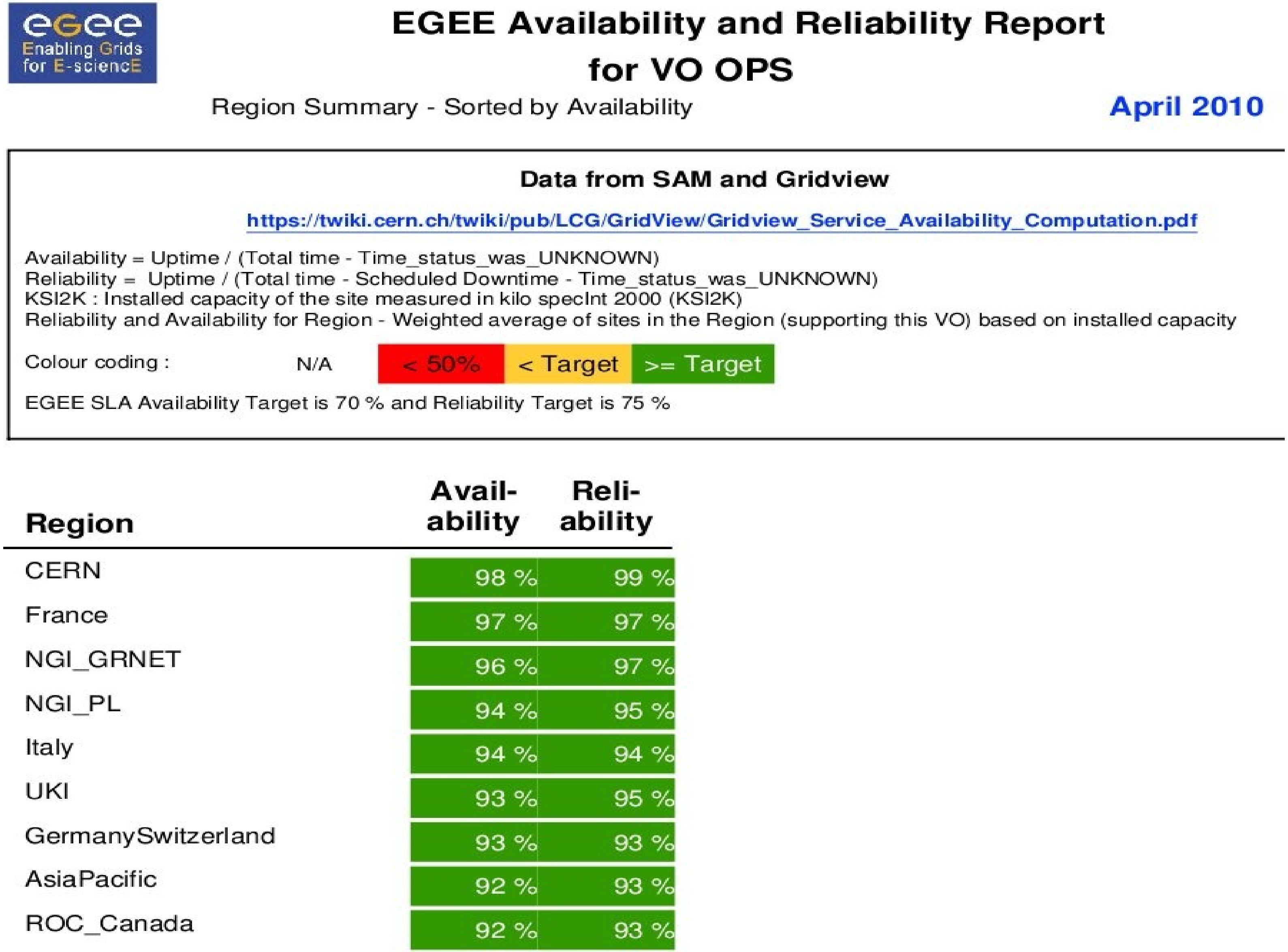
Close selected alarms

VO Alarms

Nodes

Notepad : last 5 entries

Why all this?



Enabling Grids for E-sciencE

EGEE Availability and Reliability Report for VO OPS

Region Summary - Sorted by Availability

April 2010

Data from SAM and Gridview

https://twiki.cern.ch/twiki/pub/LCG/GridView/Gridview_Service_Availability_Computation.pdf

Availability = Uptime / (Total time - Time_status_was_UNKNOWN)
Reliability = Uptime / (Total time - Scheduled Downtime - Time_status_was_UNKNOWN)
KSI2K : Installed capacity of the site measured in kilo specInt 2000 (KSI2K)
Reliability and Availability for Region - Weighted average of sites in the Region (supporting this VO) based on installed capacity

Colour coding : N/A 🟥 < 50% 🟨 < Target 🟩 >= Target

EGEE SLA Availability Target is 70 % and Reliability Target is 75 %

Region	Avail-ability	Reli-ability
CERN	98 %	99 %
France	97 %	97 %
NGI_GRNET	96 %	97 %
NGI_PL	94 %	95 %
Italy	94 %	94 %
UKI	93 %	95 %
GermanySwitzerland	93 %	93 %
AsiaPacific	92 %	93 %
ROC_Canada	92 %	93 %

Questions

very Day Is a New
Spelling List #24
1. cook
2. doctor
3. nurse
4. farmer
5. busdriver
6. teacher
7. police
8. painter
9. hairdr
10. prin
May 1
Challenge
these sentences
police officer
teacher
Is There