System Analysis Working Group and Experiment Dashboard

Julia Andreeva CERN

Grid Operations Workshop – 2007

13-15 June , Stockholm

Goal of the System Analysis Working Group

- As stated in the mandate the goal is to gain understanding of application failures in the grid environment and to provide an application view of the state of the infrastructure
- Summarize experience gained by the LHC experiments in achieving this goal and provide input to grid service monitoring working group and WLCG management
- Propagate information related to the availability of the Grid infrastructure/services as they are seen/measured by the LHC VOs to the ROC managers and local fabrics monitoring systems so that eventual problems are fixed by people who can take actions

In practical terms

- We are not trying to introduce a new monitoring system
- Use what is already available (middleware itself, existing Grid monitoring tools, experiments work load management systems and data management systems, monitoring tools developed by the experiments, experiment dashboard).
- Analyze information flow - Identify information holes (if any) - understand how situation can be improved.
- Organize the application monitoring chain in a way which would ensure the improvements of the quality of the infrastructure and the efficiency of the use of the infrastructure by it's customers

What is the criteria for estimation of the quality of the Grid infrastructure?

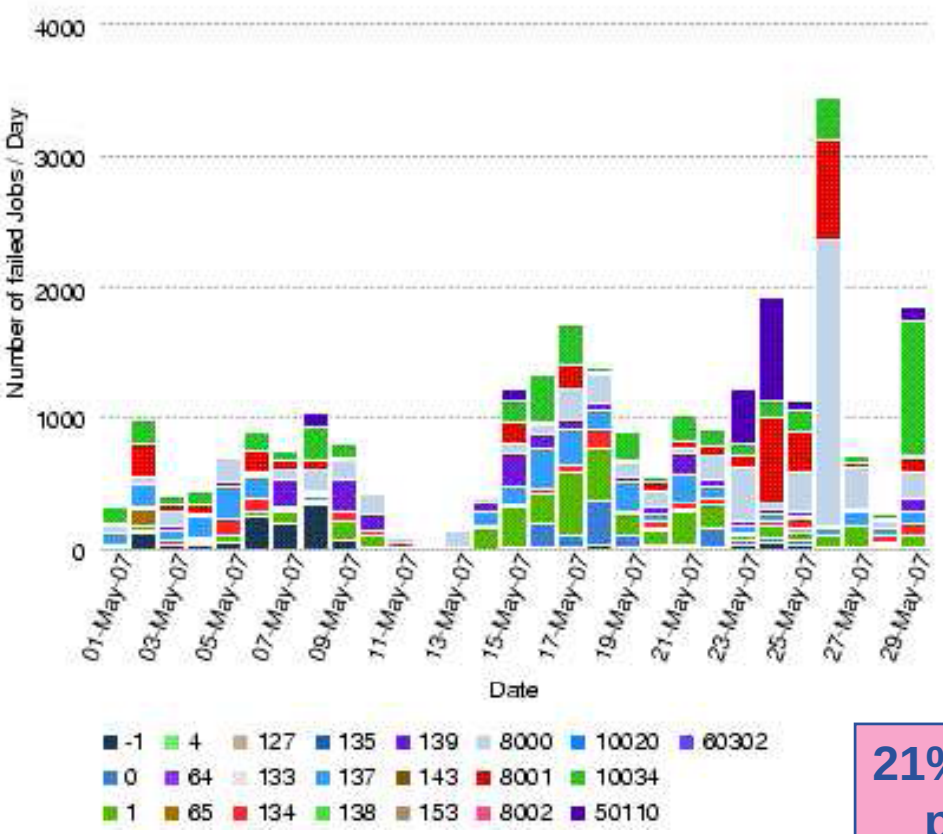
Can be many but the main judgment is of the users of the infrastructure:

- Can I run jobs?
- Can I transfer files?
- How much effort it requires?
- How quickly I get the results?

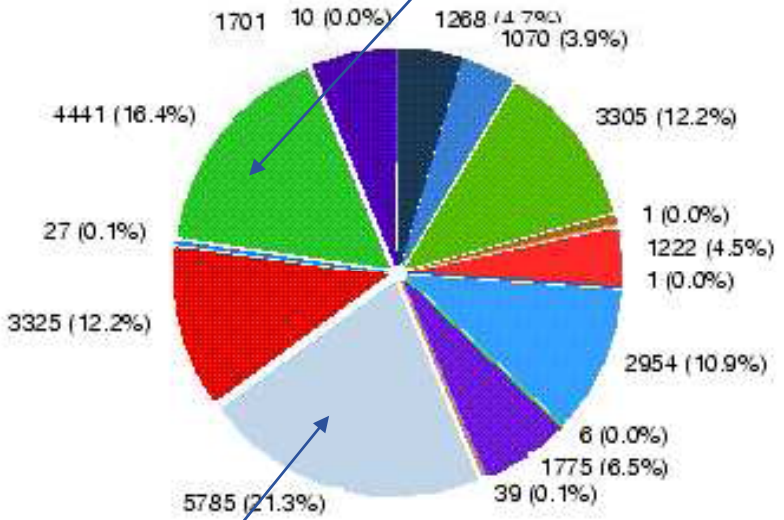
Analysis of the application failures User error or Service failure?

- Currently when we talk about Grid success rate we are ONLY PARTIALLY taking into account Grid services involved into job processing (RB, ,LB, CE, BDII, local batch system)
- Failures of the SEs, catalogues, access to the shared areas with software distributions, reading of the distributed databases are hidden in the application failures
- How to decouple these failures from the errors in the user code?

Application failures of the CMS analysis jobs over last month



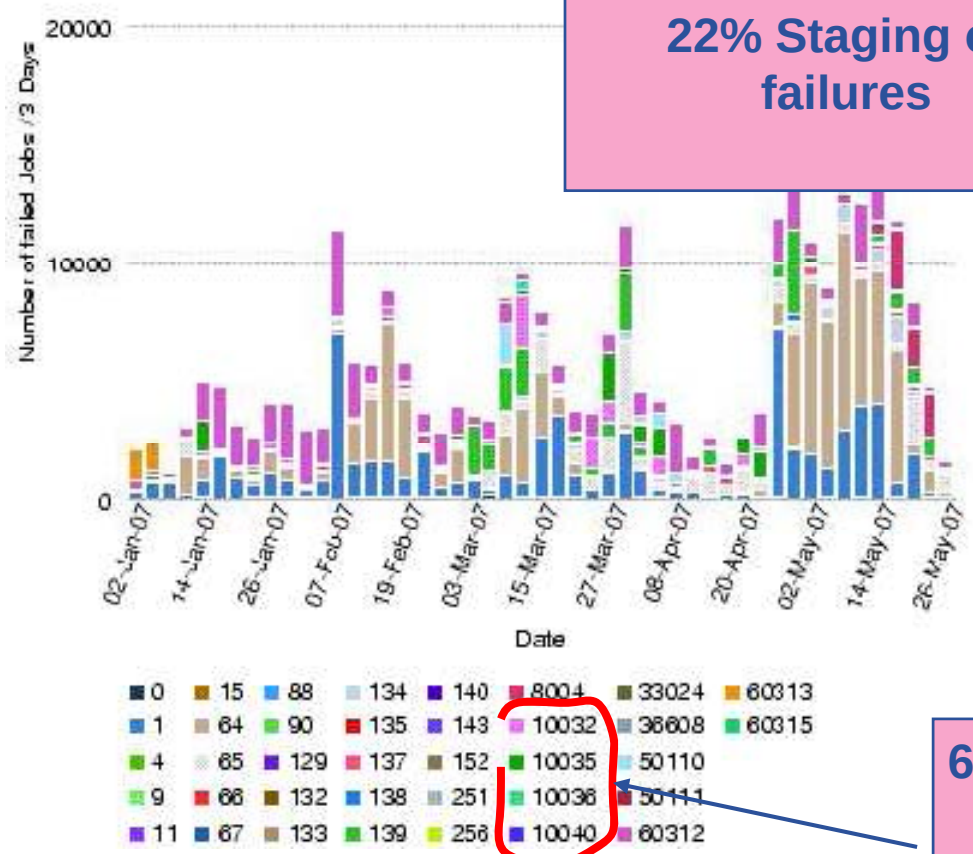
16% Required application version Is not found on the site

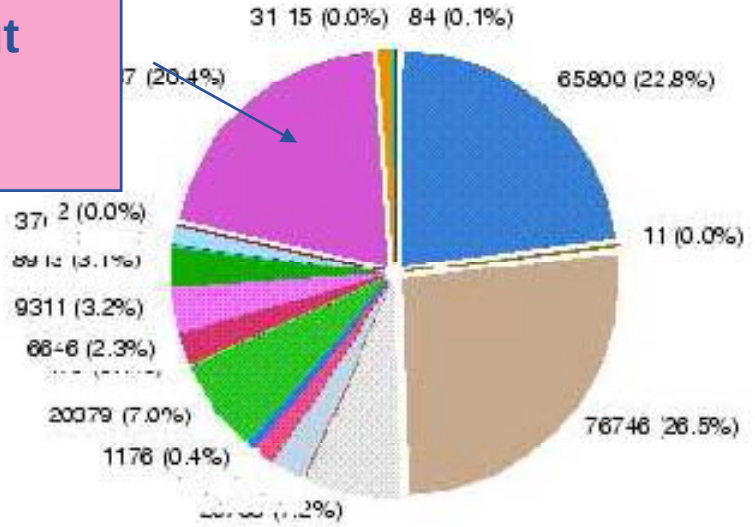


21% Failures caused by problems with data access

Application failures of the CMS production jobs since beginning of the year

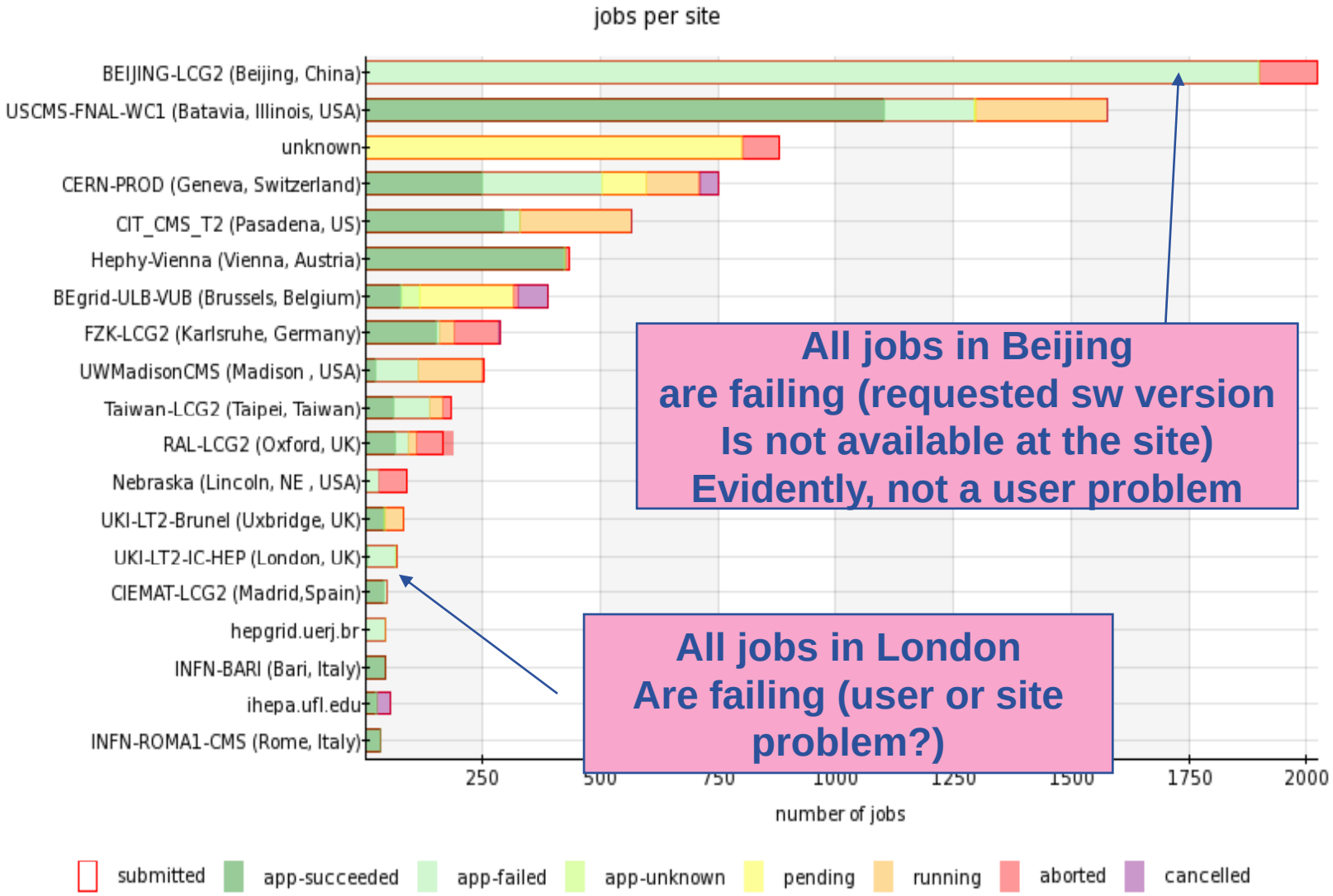
22% Staging out failures



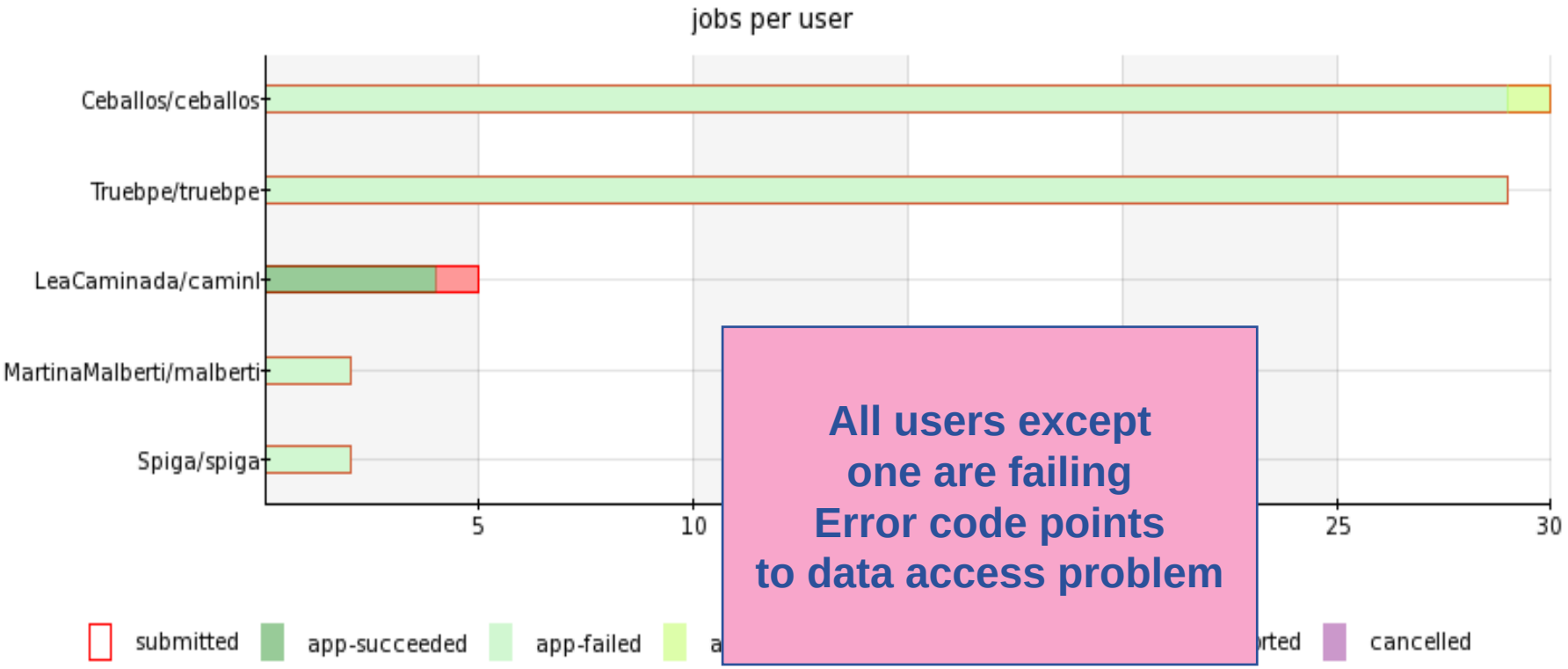


6.5% Failures relates to accessing/sourcing SW distribution At the site

One more example

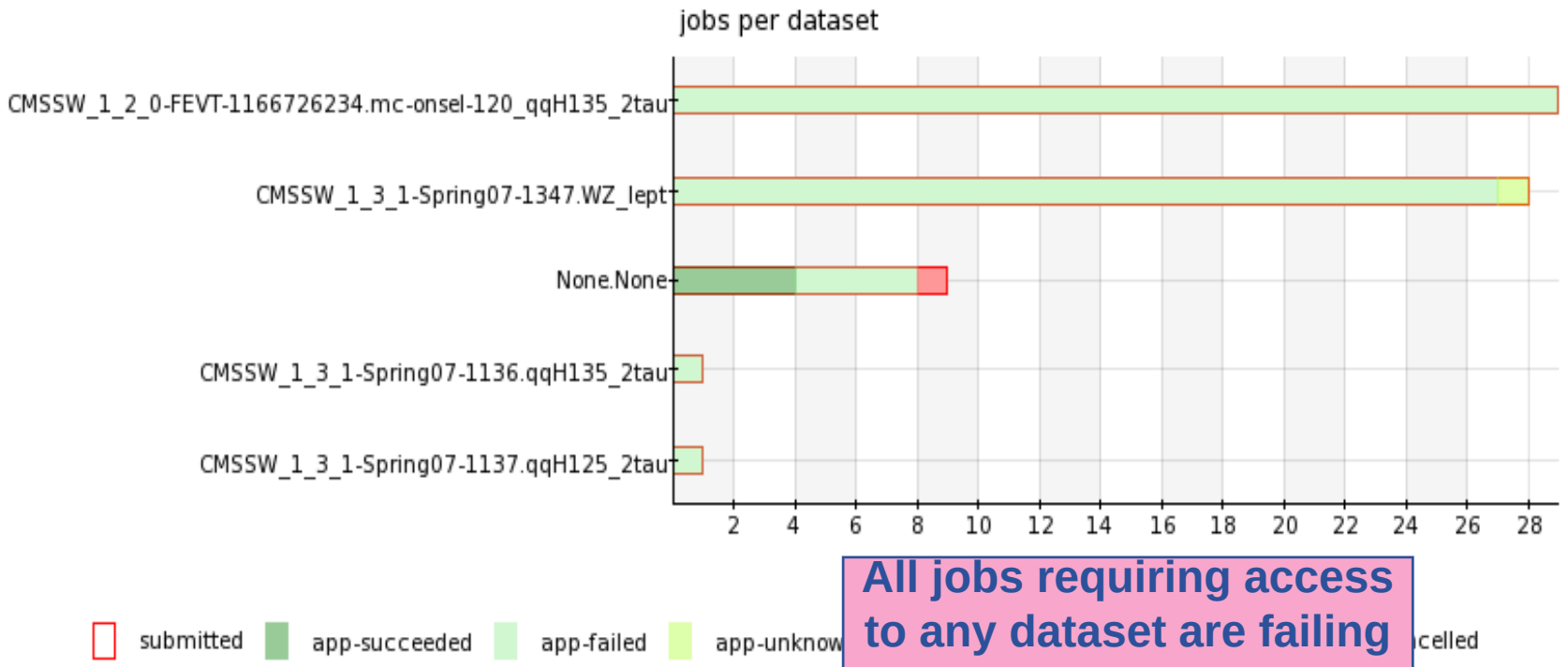


Sorted by user



All users except one are failing
Error code points to data access problem

Sorted by dataset



All jobs requiring access to any dataset are failing
Only jobs not using input succeeded
Rather site problem than the problem in the user code

Main reasons of the application failures

- Data access problems
- Problems related to saving output on the SE and registration in the catalog
- Experiments are not yet widely using distributed databases, but access to the DB from every individual job can cause problems in future

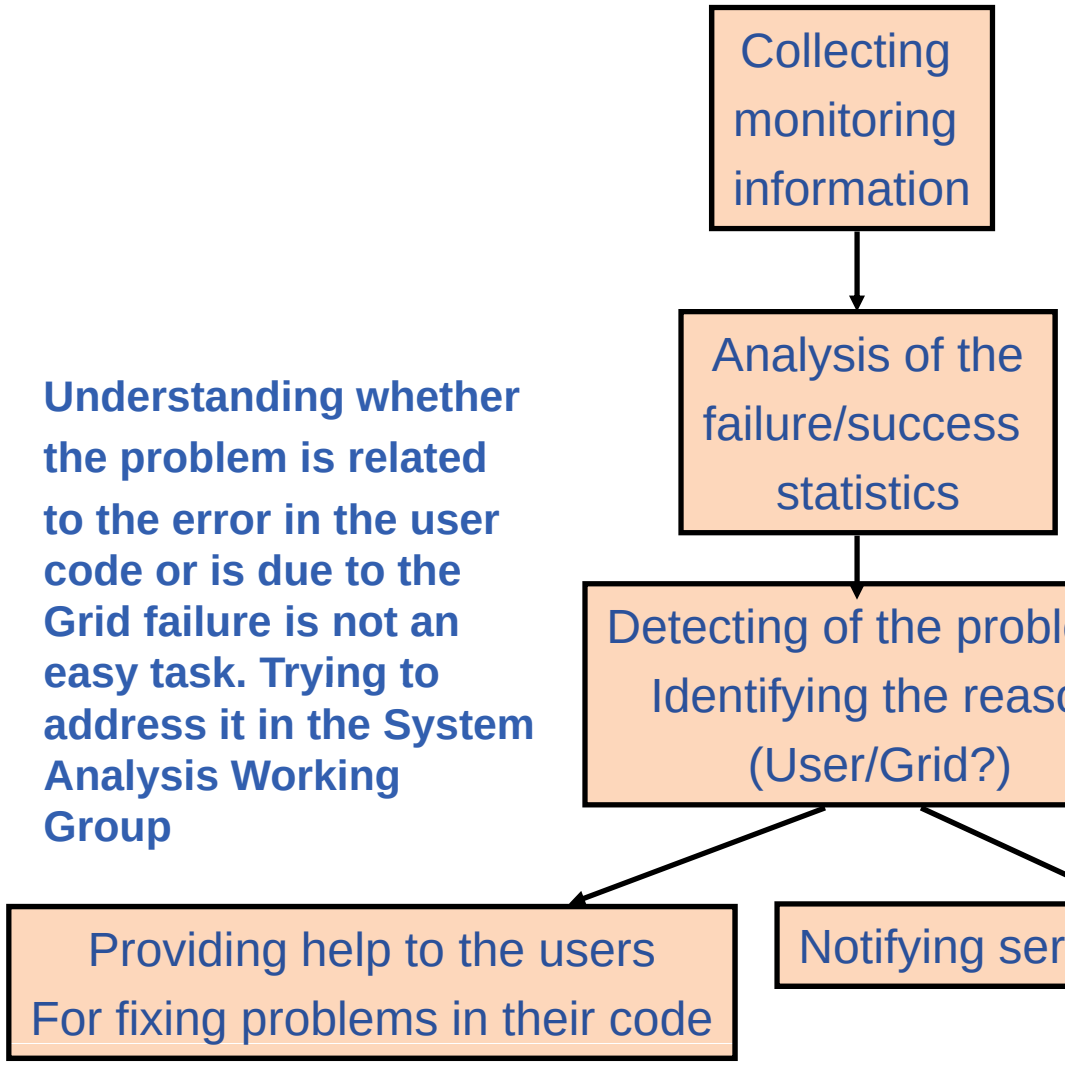
Raw estimation of the CMS application failures statistics shows that at least 30-40% of application failures are due to the failures/misconfiguration of the grid services or local sites

Still very often application failures which are not resulted to jobs aborted by the Grid are regarded as a VO problem

How experiments are currently trying to solve such problems?

- Very often the problem is first seen by the users/production teams, then reported to the experiment support lists and then either the ggus ticket is submitted or/and responsible people at the sites are contacted.
- Experiments are trying to put something in place to detect the problem as soon as possible (SAM tests, Job Robot in CMS , alarm system based on MonAlisa in Alice)
- Still the best indication of the fact that something is going wrong comes from of the massive job submission by the user community

Application monitoring chain



Done via Experiment Dashboard, Production monitoring, monitoring of the work load management systems of the experiments (Dirac, Alien)

Understanding whether the problem is related to the error in the user code or is due to the Grid failure is not an easy task. Trying to address it in the System Analysis Working Group

Communication channel for reporting problems related to the local services at the sites seen from the experiment applications has to be defined. Might follow the model of publishing back of the SAM results into the local fabric monitoring system prototyped by Grid Services Monitoring WG

Application failure diagnostics

- It is not easy to decouple application failures due to the user errors in the code or bad packaging of the user code from the failures caused by the problems of the Grid infrastructure itself

- Main problem - bad or inconsistent diagnostics from the application/job wrappers

  - developers of the application do not pay much attention to make error codes consistent and comprehensive

  - every job submission tool can have it's own set of exit codes from the job wrapper , which in the scope of the same experiment can easily overlap and have different meaning

- Solutions :

  - in the framework of SAWG work with the experiments trying to improve the situation with the exit codes and reporting in general from the physics applications and job wrappers

  - analysis of big statistics accumulated in the dashboard DB

  - SAM tests

Experiments monitoring systems

- In case of LHCb and ALICE there is one central submission queue, both for production and analysis ,

  all monitoring information can be easily collected in the database of the work load management systems
- In case of ATLAS and CMS there is no central queue. Multiple submission UIs. Collecting of the monitoring data in the experiment scope is much more complicated.
- This was one of the motivation for starting Experiment Dashboard project

Experiment Dashboard concept

Information sources

Monitoring systems (RGMA, GridIce, SAM, ICRTMDB, MonaAlisa, BDII, Apel…)

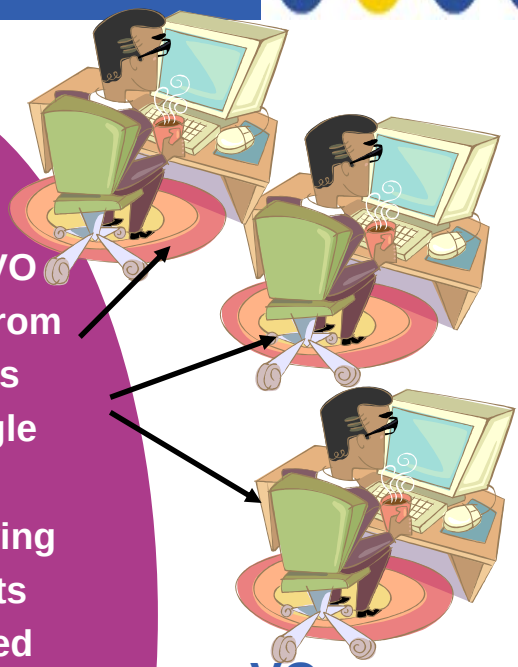
Generic Grid Services

Experiment specific services

Experiment work load management and data management systems

Jobs instrumented to report monitoring information

Collect data of VO interest coming from various sources
Store it in a single location
Provide UI following VO requirements
Analyze collected statistics
Define alarm Conditions
Propagate VO view of the infratructure To service providers



VO users with various roles

- Potentially other
- Clients:
- PANDA, ATLAS production
- <XML,CSV, image formats>

Dashboard status

- In production for all 4 LHC experiments
- Main applications:
  - Job monitoring (all experiments)
  - Data Management monitoring (ATLAS)
  - Data Transfer Monitoring (ALICE)
  - Site Reliability (all experiments)
  - Monitoring of the distributed DBs (implemented by 3D project) (ATLAS and LHCb)
- Recently the Job Monitoring and Site Reliability applications were installed by VLEMED VO outside the LHC community

Current development focused on the detection and reporting of the application failures

- Importing into Dashboard results of the SAM tests to apply them while analyzing of the user job failures
- Service availability application based on the results of the sanity check reports sent from the VO jobs (LHCb)
- Enabling generation of the application efficiency reports in the format consistent with the Grid Monitoring Data Exchange Format developed by the Grid Service Monitoring Working Group

Conclusisons

- Currently certain failures/inefficiencies of the Grid services cause the failure of the user application, though such jobs look like jobs properly handled by the Grid
- These problems have to be promptly detected, correctly identified and reported to the service providers / site administrators
- Work in the System Analysis WG and Experiment Dashboard in close collaboration with the LHC experiments and other monitoring working groups aims to address this issue.