



Enabling Grids for E-science

New job monitoring strategy on the WLCG scope

Julia Andreeva

CERN (IT/GS)

CHEP 2009, March 2009, Prague

www.eu-egee.org



- **Importance of job monitoring.**
- **Current status, with the main focus given to the LHC VOs.**
- **Looking forward - new job monitoring architecture. Ongoing development.**
- **Examples of new job monitoring applications.**
- **Summary.**

This work is carried out by a lot of people from different projects and institutes:

LB team, GridView team, Condor team, CERN IT-GS group, ICRTM team, EDS company collaborating with CERN via OpenLab, our colleagues from Russian institutions participating in the Dashboard development, our colleagues in the LHC experiments.

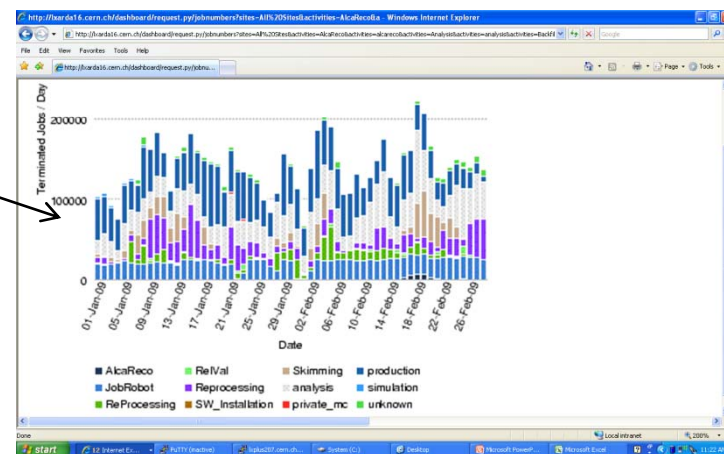
- **Data distribution and data processing are two main computing activities for the VOs running on WLCG infrastructure**
- **Quality of job processing to the large extent provides the estimation of the quality of the infrastructure in general and defines the overall success of the computing activities of the VOs**
- **On the other hand, detailed and reliable job monitoring helps to improve the computing models of the LHC VOs.**

- Very large scale.
Just CMS submits up to 200K jobs per day, and this number is steadily growing

- Infrastructure is not homogeneous.

Several middleware flavors are used.

- VOs are using various submission methods (via WMS, direct submission to CE)
- Multiple pilot systems are used by LHC VOs : Alien, Dirac, Panda, condor-glideins.
- **Therefore , currently there is no one single GRID service which can be instrumented in order to get information about all jobs submitted to the WLCG infrastructure.**



Complexity of the job monitoring task (estimation of efficiency)

- Currently two main categories are considered regarding job failure:
 - Grid aborts. Job was not successfully processed by the Grid through the job processing chain
submitted -> allocated to the site-> ran at the WN -> saved the output sandbox
 - Job was successfully processed by the GRID, but application exited with non 0 code. Normally considered as user failure

Complexity of the job monitoring task (estimation of efficiency)

- In reality when the job is aborted by the GRID this is not always problem of the GRID services.
 - *Examples: Error in the JDL file, expiration of user proxy*
- Even more often it happens that **application failure has nothing to do with the problem of application itself.**
 - *Examples: Job failed due to the problem of SE, catalogue,... while accessing input file or saving the output*
- Failure diagnostics both from the GRID sources and applications is very often incomplete, unclear or even misleading

ONLY a combination of GRID and what is considered application efficiency can give the estimation of the quality of the infrastructure.

But this implies proper decoupling of user errors from the problems caused by the GRID services or site misconfiguration.

- **ALICE and LHCb have central queue for VO users, most of jobs of these VOs are submitted via central queue.**
 - Single submission point → single point for collecting monitoring data.

Quite simple model regarding monitoring

- **For ATLAS and CMS situation is more complicated. Distributed submission systems, several middleware platforms are used, various submission methods and execution backends.**
 - Multiple solutions for job monitoring : PANDA monitoring, ProdAgent monitoring, Experiment Dashboard.

Rather complex task regarding monitoring

- **Information sources for job monitoring:**
 - **Job submission tools, jobs instrumented to report their status, GRID services keeping the track of status of jobs being processed like Logging and Bookkeeping system**
- **A variety of methods for information retrieval and transport protocols are used**
- **Regardless of organization of work load management systems of the experiments all LHC VOs need to query the GRID services keeping track of job status on regular basis**

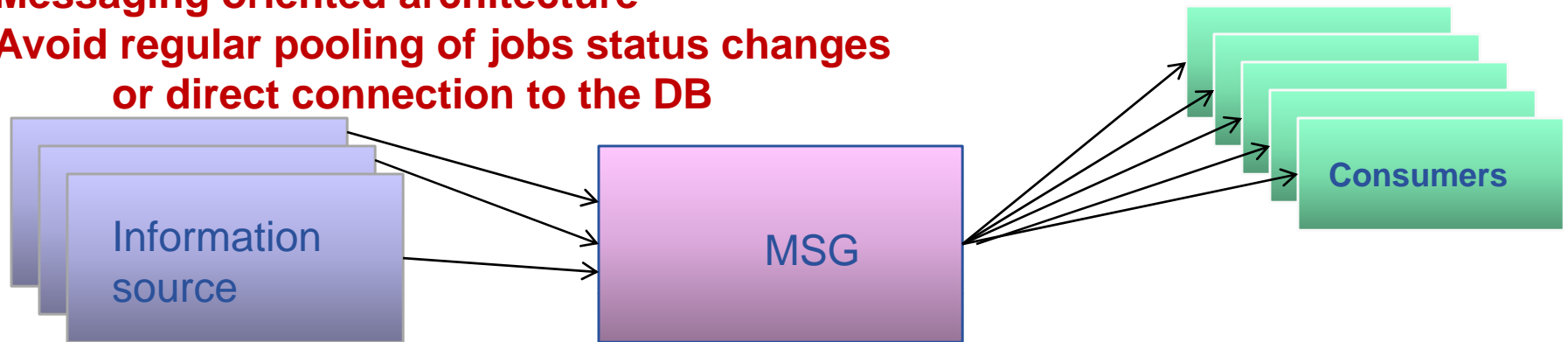
- Do we currently have a reliable overall runtime picture of job processing on the global WLCG scope?
- **We have to admit that situation is far of being ideal.**
- The only monitoring tool providing the overall view for all jobs (all VOs) running on the WLCG infrastructure is Imperial College Real Time Monitor (ICRTM).
- Recently the new instance of Dashboard Job Monitoring had been set up to show job processing of all VOs running on WLCG infrastructure. AS information source it is using xml files published by ICRTM :

<http://dashb-lcg-job.cern.ch/dashboard/request.py/jobsummary>

- **Currently ICRTM collects information via direct connection to Logging and Bookkeeping DB**
- **Only jobs submitted via WMS are recorded in LB and correspondingly are monitored by ICRTM**
- **Substantial fraction of jobs submitted via WMS escape ICRTM monitoring**

MAIN PRINCIPLES:

- 1). Messaging oriented architecture
- 2). Avoid regular pooling of jobs status changes or direct connection to the DB



LB, CREAM CE (via CEMon notification), Condor-g , jobs instrumented to report their progress, Job Submission Tools of the experiments

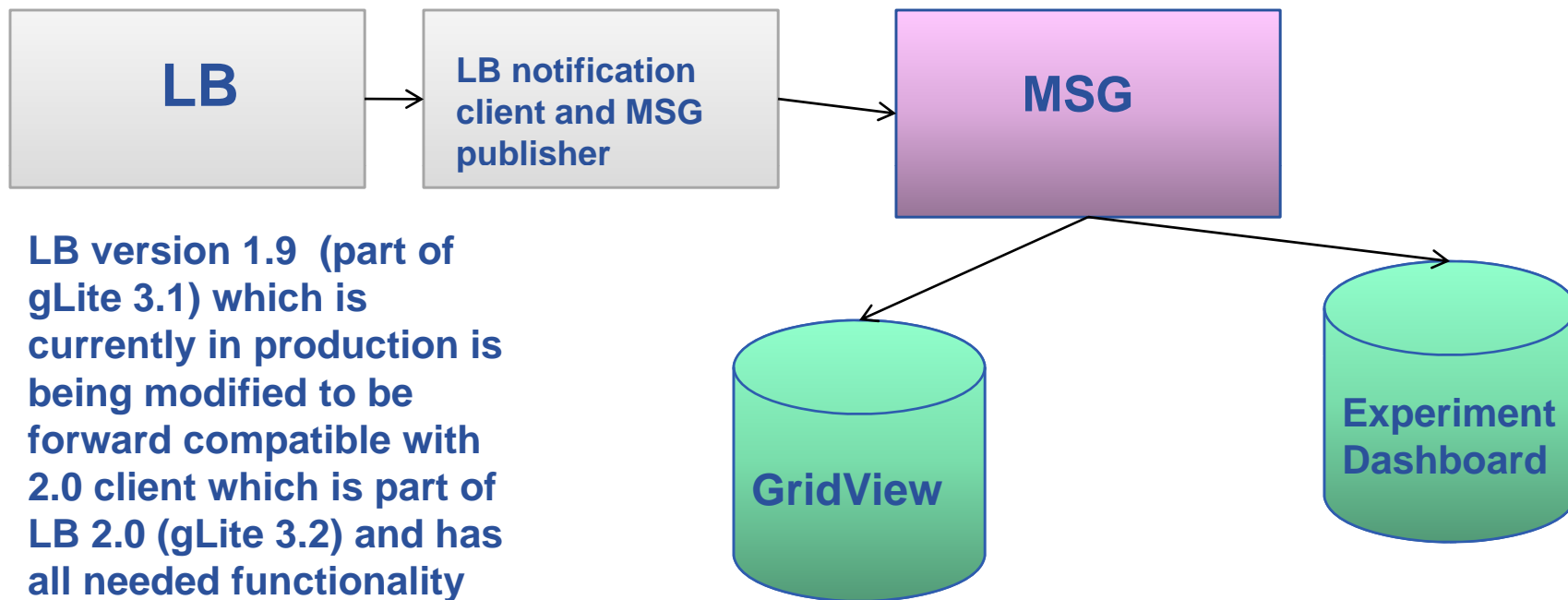
Messaging System for the Grids
Apache ActiveMQ implementation

Various clients of job monitoring information, like GridView, Dashboard, ICRTM, Dirac, CRAB server, etc...

Apache ActiveMQ had been evaluated as an appropriate solution for WLCG messaging system following the program of work defined by Grid Service Monitoring Working Group chaired by James Casey and Ian Neilson

- **Common way of publishing information by various information sources**
- **Common way of communication between different components of the WLCG infrastructure**
- **No need to connect to multiple instances of the information sources (LB DBs for example)**
- **Job monitoring information is publicly available for all possible interested parties**
- **Decreasing load on the Grid services caused by regular pooling of information about job status changes -> improving of their performance**

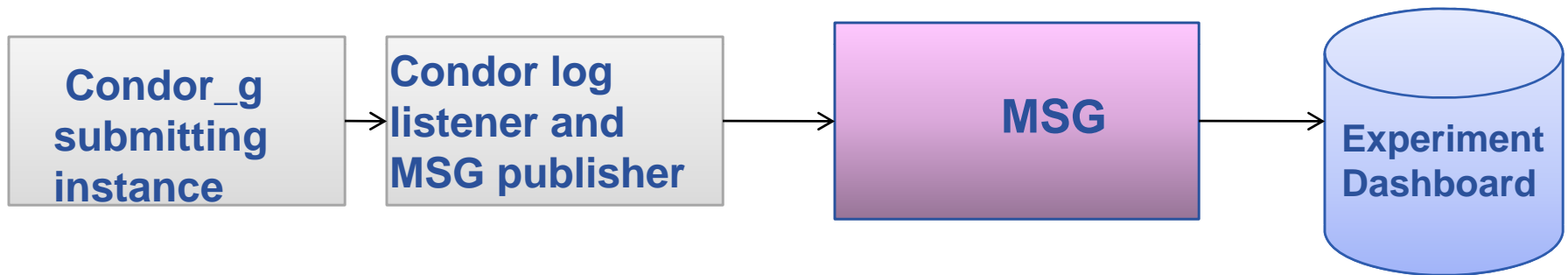
Collaboration of LB, GridView and Experiment Dashboard teams.



LB version 1.9 (part of gLite 3.1) which is currently in production is being modified to be forward compatible with 2.0 client which is part of LB 2.0 (gLite 3.2) and has all needed functionality for job monitoring. Should be ready for certification by the end of April.

For more details about MSG see poster of D. Rocha “MSG as a core part of the new WLCG monitoring infrastructure”

1). Collaboration of Condor and Dashboard teams
Instrumentation of condor_g for MSG reporting



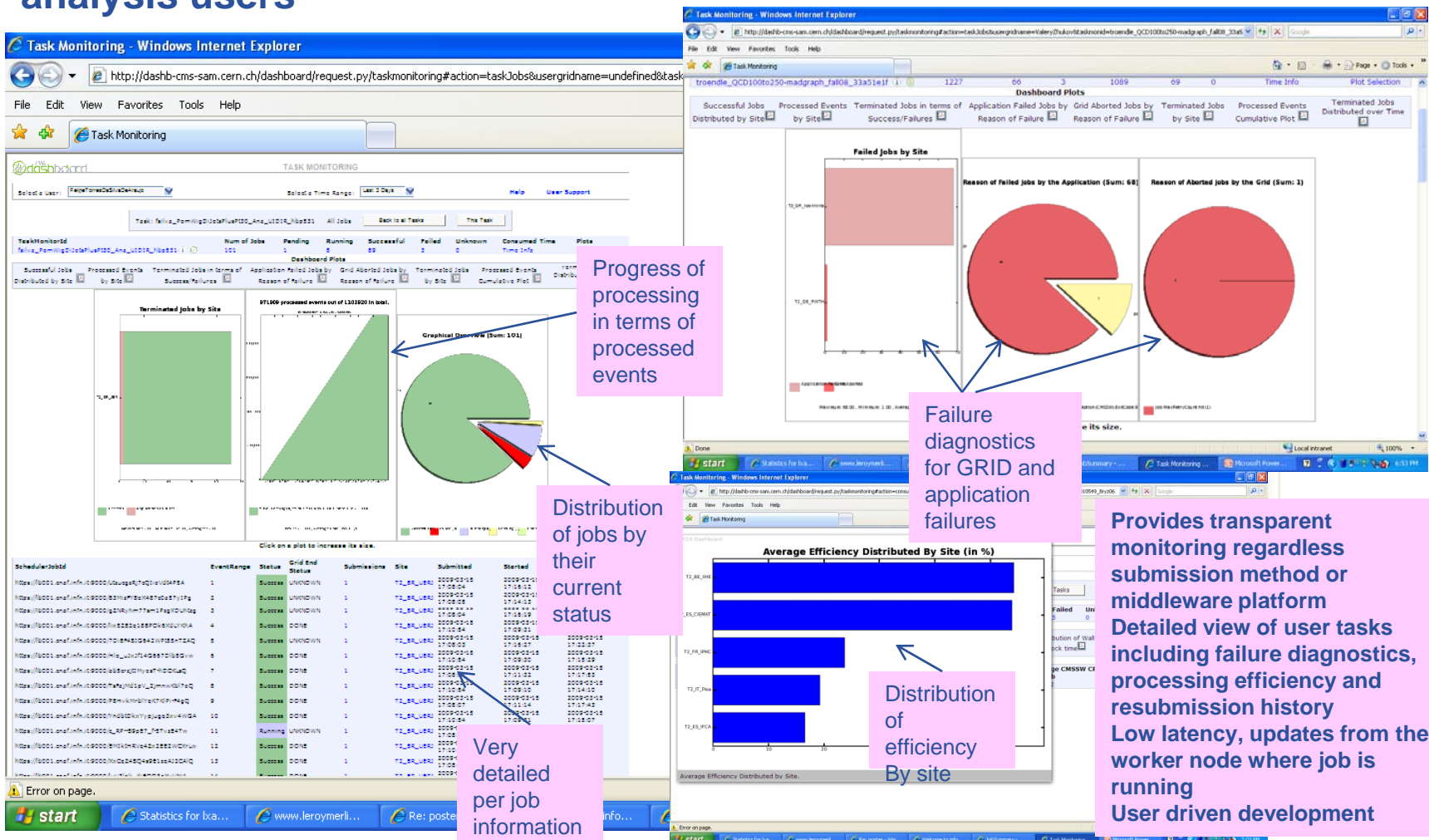
2). Collaboration of Dashboard team with LHC experiments
Instrumentation of Job Submission Tools of ATLAS and CMS for reporting of application level monitoring information via MSG

See another working example in the talk of U.Schwickerath “Monitoring the efficiency of the user jobs”

- Global view, performance of the infrastructure in general
(GridView, ICRTM, Experiment Dashboard, systems are in place, but need to improve reliability and completeness of provided data)
- VO view, whether VO can perform their tasks on the GRID
(Experiment specific monitoring systems like Dirac, MonAlisa for ALICE, Panda monitoring, Experiment Dashboard for ATLAS and CMS. Work quite well and provide reliable monitoring)
- Site view, whether my site is working well and satisfies the VO requirements
- User view, did my jobs run and produced needed data.
(Last two views in particular the one for sites are being addressed in the recent development, examples further in the talk)

As a rule the monitoring data repository keeps very detailed per job information. Variety of user interfaces is provided on top of central repository to satisfy different use cases (VO managers, production , operations , user support teams, users running jobs on the GRID)

CMS Task monitoring for analysis users



The screenshot shows the CMS Task Monitoring dashboard in a web browser. The interface includes a navigation menu, a task selection dropdown, and a main dashboard area with several charts and a table. Annotations in pink boxes highlight specific features:

- Progress of processing in terms of processed events:** Points to a green area chart showing the cumulative number of processed events over time.
- Failure diagnostics for GRID and application failures:** Points to two pie charts: 'Reason of Failed jobs by the Application (Sum: 68)' and 'Reason of Aborted jobs by the Grid (Sum: 3)'. A bar chart 'Failed Jobs by Site' is also visible.
- Provides transparent monitoring regardless submission method or middleware platform:** A general statement about the dashboard's transparency.
- Detailed view of user tasks including failure diagnostics, processing efficiency and resubmission history:** Points to a table listing individual tasks with columns for SchedulerJobId, EventRange, Status, Grid End Status, Submissions, Site, Submitted, and Started.
- Low latency, updates from the worker node where job is running:** A general statement about the dashboard's performance.
- User driven development:** A general statement about the dashboard's development philosophy.
- Distribution of jobs by their current status:** Points to a pie chart showing the distribution of jobs across different states.
- Very detailed per job information:** Points to the task table.
- Distribution of efficiency by site:** Points to a horizontal bar chart titled 'Average Efficiency Distributed By Site (in %)'.

See poster of E. Karavakis "CMS Dashboard Task Monitoring: A user-centric monitoring view."

GridMap - Windows Internet Explorer

http://dashb-siteview.cern.ch/generic/site-monitoring/test.html

Siteview GridMap Test Page

Site status

ALICE ATLAS CMS LHCb

job_processing

ALIC

job_processing/ATLAS

mc_production user_ana

data_transfer_in

data_transfer_out

mc_production

2009-03-15 18:37
Status: **critical**

| Name | Value | Status | Target | Go to |
|--|----------|-----------------|--------|---------------------|
| Running jobs [average over the last hour] | 1 jobs | unknown | -1 | URL |
| Completed jobs over the last hour | jobs | None | -1 | URL |
| Successfully completed jobs over the last hour | jobs | None | -1 | URL |
| Completed jobs over the last 24 hours | 173 jobs | critical | -1 | URL |
| Successfully completed jobs over the last 24 hours | 28 jobs | critical | -1 | URL |

OK Good Poor

Sites

Tier:

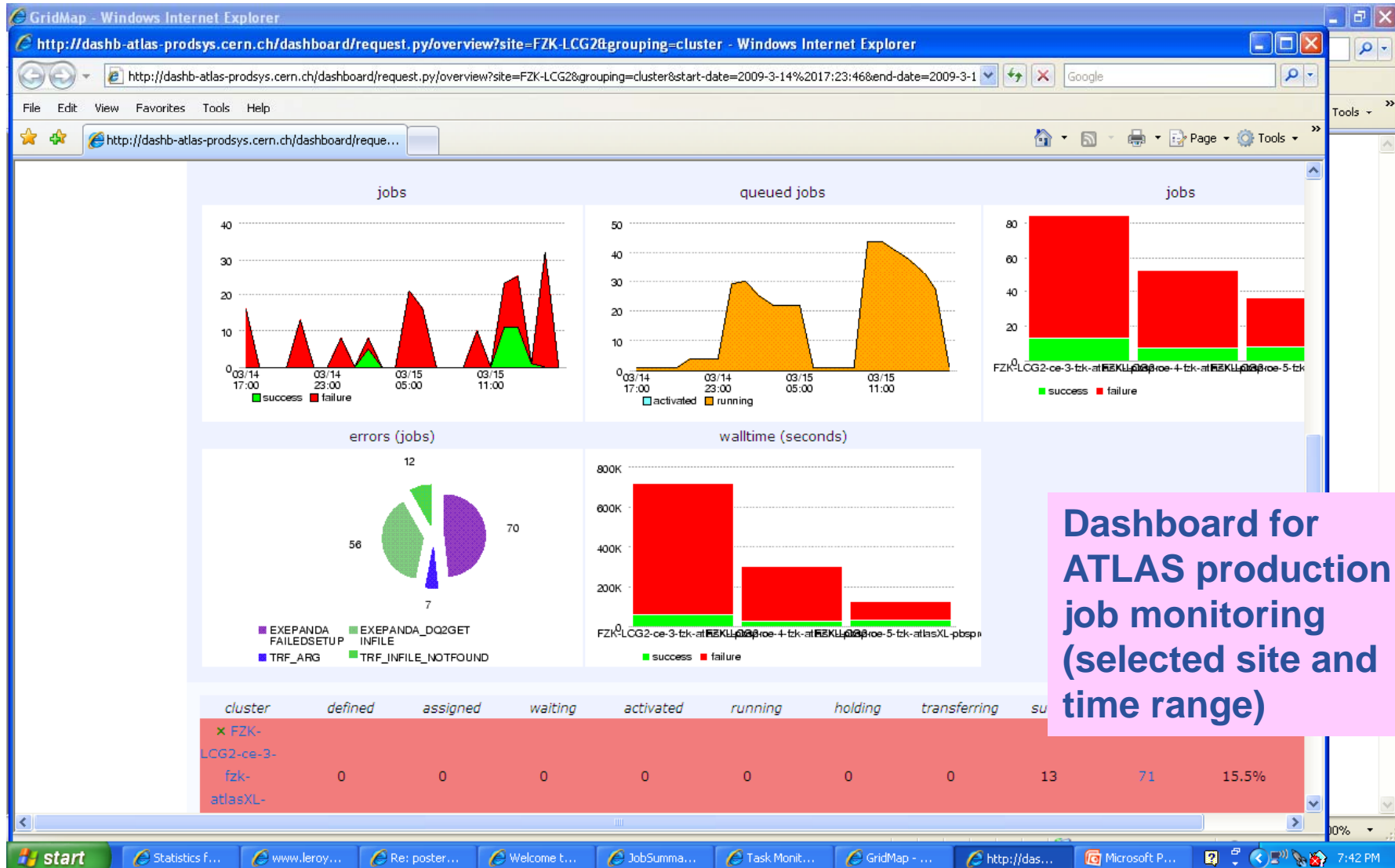
Site:

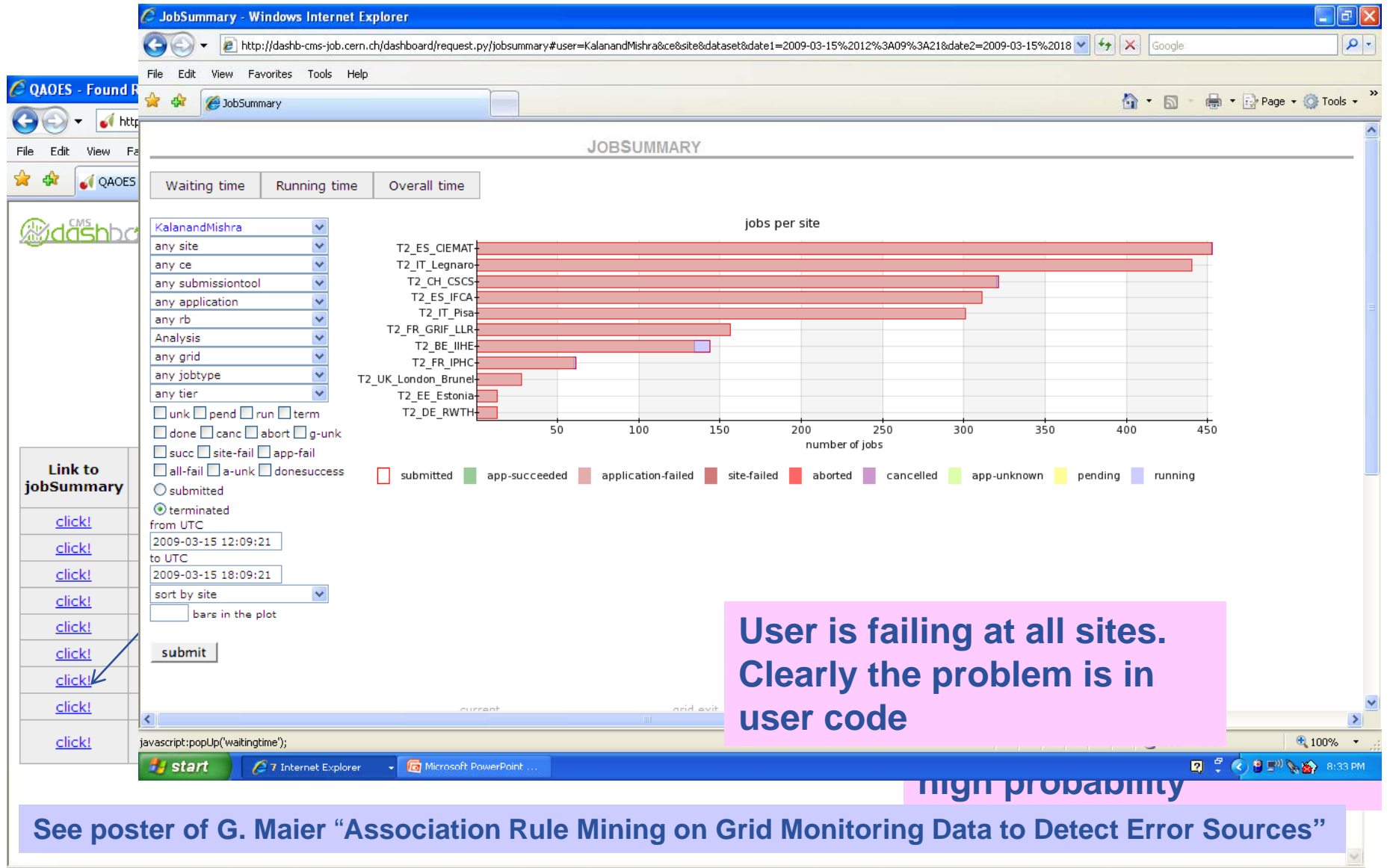
Done

Moving mouse over a case corresponding to a particular activity, results in opening a sub-map showing status and scale of sub-activity of a given activity

Provided URLs assist to navigate to the primary information sources

See poster of E. Lanciotti "High level view of the site performance from the LHC perspective"





JOBSUMMARY

Waiting time | Running time | Overall time

jobs per site

| Site | Number of Jobs | Status |
|---------------------|----------------|-----------|
| T2_ES_CIEMAT | 450 | submitted |
| T2_IT_Legnano | 440 | submitted |
| T2_CH_CSCS | 320 | submitted |
| T2_ES_IFCA | 310 | submitted |
| T2_IT_Pisa | 300 | submitted |
| T2_FR_GRIF_LLRC | 160 | submitted |
| T2_BE_IJHE | 140 | submitted |
| T2_FR_IPHC | 60 | submitted |
| T2_UK_London_Brunel | 30 | submitted |
| T2_EE_Estonia | 20 | submitted |
| T2_DE_RWTH | 20 | submitted |

Legend: submitted (red), app-succeeded (green), application-failed (light red), site-failed (dark red), aborted (orange), cancelled (purple), app-unknown (light green), pending (yellow), running (blue)

User: KalanandMishra

from UTC: 2009-03-15 12:09:21

to UTC: 2009-03-15 18:09:21

sort by site

submit

javascript:popup('waitingtime');

User is failing at all sites. Clearly the problem is in user code

high probability

See poster of G. Maier "Association Rule Mining on Grid Monitoring Data to Detect Error Sources"

- **Monitoring systems of the LHC VOs provide rather complete view of job processing on the WLCG infrastructure**
- **There is still a big room for improvements regarding job monitoring on the global WLCG scope**
- **Main principles for new job monitoring architecture had been defined. Implementation is ongoing.**
- **The monitoring systems of the LHC VOs as well as their work load management systems will benefit when the new system is in place**