

Monitoring the ATLAS distributed production

*Benjamin Gaidioz (CERN),
Ricardo Rocha (CERN), Xavi Espinal (PIC),
Alex Read (Univ Oslo), Simone Campana (CERN)*

EGEE User Forum 2009

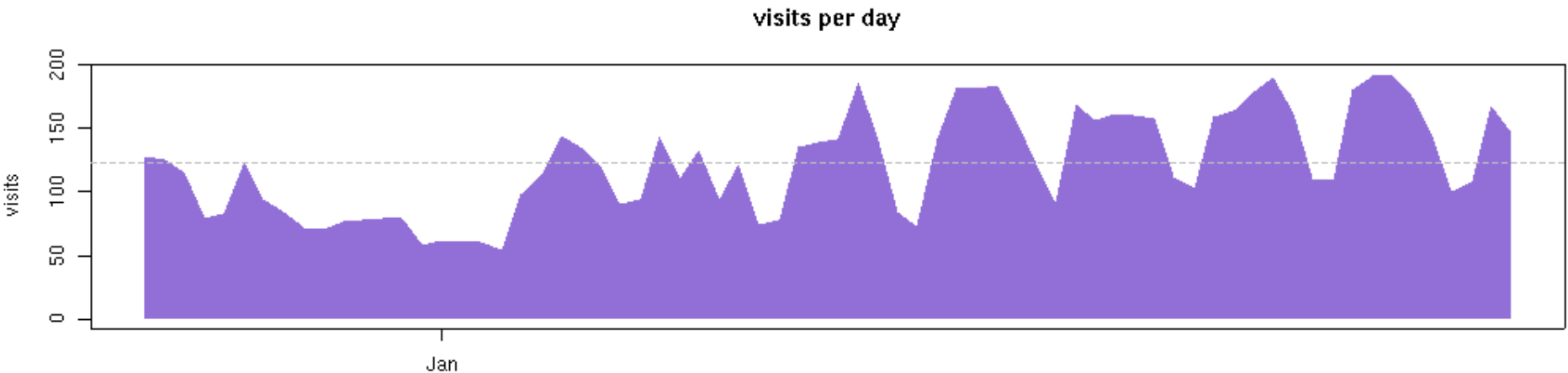
- **ATLAS production**

- runs on EGEE, OSG (panda) and Nordugrid (Dulcinea),
- user defined tasks:
 - automatically split and executed:
 - *one ATLAS task ~ 1000 subjobs*
 - *average: 60K jobs per day,*
 - *~ 800 active tasks per day, ~ 125 sites.*



- **ATLAS shifters**
 - spot failures, investigate and report:
 - task error → ATLAS software team
 - site error → GGUS
 - task or site (some errors don't say much)?
 - need:
 - an **immediate view** of the main problems,
 - a **practical interface** to get to the details.

- a python framework for implementing monitoring applications:
 - team based in IT/CERN
 - <http://dashboard.cern.ch>
- usage in ATLAS:
 - distributed data management monitoring,
 - production monitoring, analysis monitoring,
 - standard tool used by shifters, many tutorials given, etc.



- <http://dashb-atlas-prodsys.cern.ch/dashboard/request.py/overview>

File Edit View History Bookmarks Tools Help

http://dashb-atlas-prodsys.cern.ch/dashboard/request.py/shifters-display

Most Visited ATLAS production

ATLAS dashboard

Data: Tier 0 Data: Production Production

Tasks Grid jobs Summary Admin User Guide Feedback

noticeable failures computed automatically

We are trying to gather here a set of useful plots and links what it summarizes. This is just a beginning. Please, shifters, help us by telling us how it could improve. Thanks!

- EXEPANDA_DQ2PUT_FILECOPYERROR: LYON (1230)
- EXEPANDA_DQ2PUT_FILECOPYERROR: IN2P3-CC-T2 (1215)
- TRFERROR: pic (625)
- TRFERROR: PIC (625)
- TRF_UNKNOWN : pic (625)
- TRF_UNKNOWN : PIC (625)
- EXEPANDA_DQ2GET_LOCAL-INPUT-FILE-MISSING: MidwestT2 (598)
- EXEPANDA_DQ2PUT_FILECOPYERROR: 40917 (393)
- EXEPANDA_JOBEXPIRED_SIXDAYS: 40381 (316)
- EXEPANDA_DQ2_STAGEIN: 41563 (204)
- EXEPANDA_DQ2PUT_FILECOPYERROR: AGLT2 (199)
- EXEPANDA_JOBEXPIRED_SIXDAYS: 39727 (198)
- EXEPANDA_DQ2GET_LOCAL-INPUT-FILE-MISSING: 40896 (198)
- EXEPANDA_DQ2GET_LOCAL-INPUT-FILE-MISSING: 40710 (196)
- EXEPANDA_DQ2PUT_FILECOPYERROR: 41106 (195)
- EXEPANDA_JOBEXPIRED_SIXDAYS: csTCDie (166)
- EXEPANDA_JOBEXPIRED_SIXDAYS: SARA (166)
- EXEPANDA_JOBEXPIRED_SIXDAYS: (150)
- EXEPANDA_JOBEXPIRED_SIXDAYS: (150)
- EXEPANDA_DQ2PUT_FILECOPYERROR: 41401 (148)

view

- by grid
- by cloud
- by dest_cloud
- by executortype
- by executor
- by site
- by cluster
- by task

select cloud

TRFERROR errors

SWMISS errors

most common errors

most failing clouds

unknown cloud

LYON cloud

BNL cloud

ATLAS production

- IN2P3-CC-T2 has 1297 EXEPANDA_DQ2PUT...
- LYON has 1297 EXEPANDA_DQ2PUT_FILECO...
- pic has 641 TRFERROR
- PIC has 641 TRFERROR
- pic has 641 TRF_UNKNOWN
- PIC has 641 TRF_UNKNOWN
- MidwestT2 has 598 EXEPANDA_DQ2GET_LOC...
- 39727 has 477 EXEPANDA_JOBEXPIRED_SIX...
- 40381 has 429 EXEPANDA_JOBEXPIRED_SIX...
- 40917 has 393 EXEPANDA_DQ2PUT_FILECOP...
- csTCDie has 279 EXEPANDA_JOBEXPIRED_...
- SARA has 279 EXEPANDA_JOBEXPIRED_SIX...
- 41563 has 204 EXEPANDA_DQ2_STAGEIN
- AGLT2 has 204 EXEPANDA_DQ2PUT_FILECO...
- 40896 has 198 EXEPANDA_DQ2GET_LOCAL-I...
- 40710 has 196 EXEPANDA_DQ2GET_LOCAL-I...
- 41106 has 195 EXEPANDA_DQ2PUT_FILECOP...
- 41401 has 169 EXEPANDA_DQ2PUT_FILECOP...
- None has 150 EXEPANDA_JOBEXPIRED_SIX...
- None has 150 EXEPANDA_JOBEXPIRED_SIX...
- 41193 has 126 EXEPANDA_DQ2PUT_FILECOP...
- 41195 has 106 EXEPANDA_DQ2PUT_FILECOP...
- INFN-ROMA1 has 85 EXEPANDA_DQ2PUT_FI...
- 40546 has 42 EXEPANDA_DQ2PUT_FILECOP...

view

- by grid
- by cloud
- by dest_cloud
- by executortype
- by executor
- by site
- by cluster
- by task

select cloud

Open "ATLAS production"

Open All in Tabs

Jobs: Production Jobs: Analysis Panda: Production

Shifters Functional tests Admin User Guide Feedback

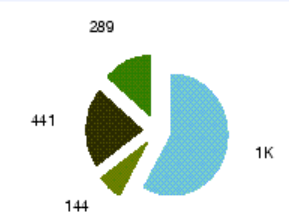
Plots for shifters. Each plot is clickable and brings you to the details of what it summarizes. This is just a beginning. Please, shifters, help us by [telling](#)

- IN2P3-CC-T2 (1230)
 - IN2P3-CC-T2 (1215)
 - AGLT2 (598)
 - 40896 (393)
 - 40710 (393)
 - 41106 (195)
 - 41401 (169)
 - None (150)
 - None (150)
 - 41193 (126)
 - 41195 (106)
 - INFN-ROMA1 (85)
 - 40546 (42)
- EXEPANDA_DQ2PUT_FILECOPYERROR: AGLT2 (199)
 - EXEPANDA_JOBEXPIRED_SIXDAYS: 39727 (198)
 - EXEPANDA_DQ2GET_LOCAL-INPUT-FILE-MISSING: 40896 (198)
 - EXEPANDA_DQ2GET_LOCAL-INPUT-FILE-MISSING: 40710 (196)
 - EXEPANDA_DQ2PUT_FILECOPYERROR: 41106 (195)
 - EXEPANDA_JOBEXPIRED_SIXDAYS: csTCDie (166)
 - EXEPANDA_JOBEXPIRED_SIXDAYS: SARA (166)
 - EXEPANDA_JOBEXPIRED_SIXDAYS: (150)
 - EXEPANDA_JOBEXPIRED_SIXDAYS: (150)
 - EXEPANDA_DQ2PUT_FILECOPYERROR: 41401 (148)

SWMISS errors

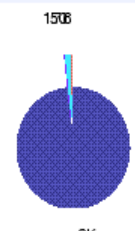


STAGEIN/STAGEOUT errors

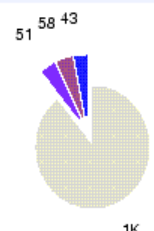


- TRF_UNKNOWN
- TRFERROR
- TRF_SVRINIT
- TRF_ATHENACRASH
- TRF_SEGFAULT
- None

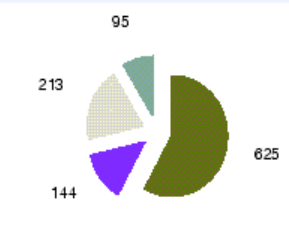
unknown cloud



LYON cloud



BNL cloud



find

view

- by grid
- by cloud
- by dest_cloud
- by executortype
- by executor
- by site
- by cluster
- by task

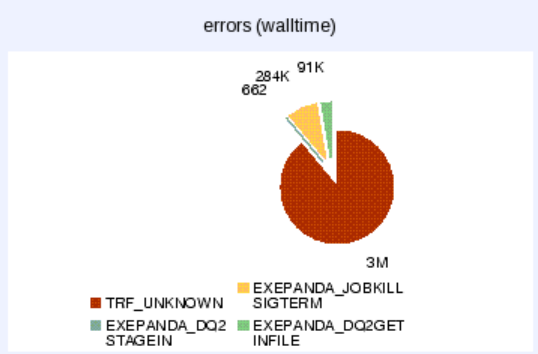
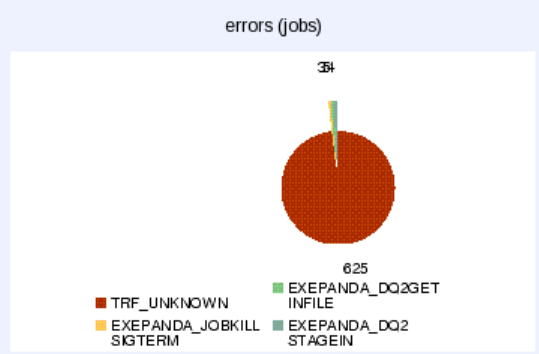
select site

pic

status

- failure
- site
- pic

2009-02-24 21:00:00 — 2009-02-25 09:59:59



restrict to: TRF_UNKNOWN (625), EXEPANDA_DQ2GET_INFIL (5), EXEPANDA_DQ2_STAGEIN (4),

site	1	2	3	others
pic (637)	TRF_UNKNOWN (625)	EXEPAND...T_INFIL (5)	EXEPAND..._STAGEIN (4)	others (3)

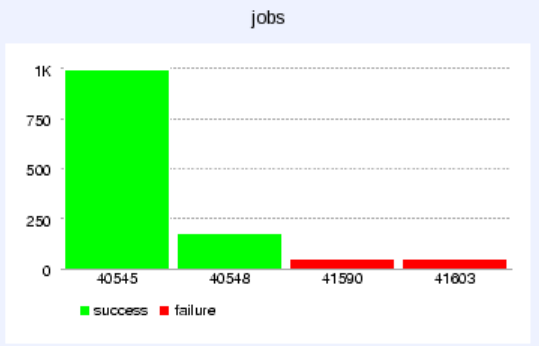
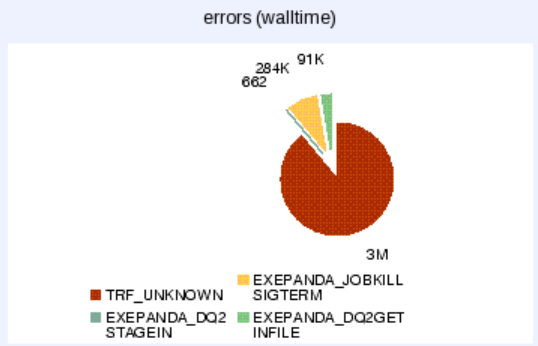
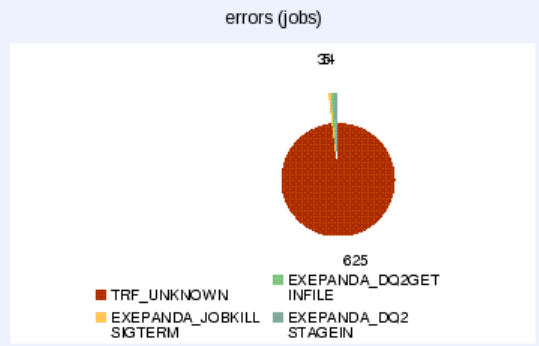
find

view

- by grid
- by cloud
- by dest_cloud
- by executortype
- by executor
- by site
- by cluster
- by task**

select task

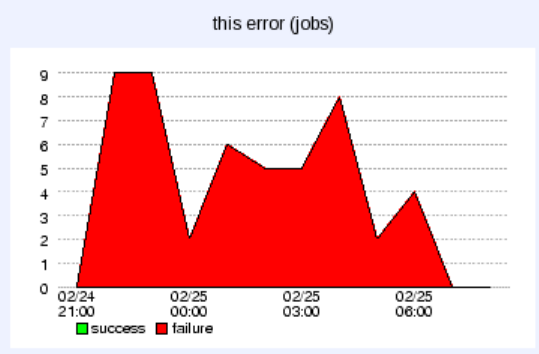
2009-02-24 21:00:00 — 2009-02-25 09:59:59



restrict to: TRF_UNKNOWN (625), EXEPANDA_DQ2GET_INFILE (5), EXEPANDA_DQ2_STAGEIN (4), EXEPANDA_JOBKILL_SIGTERM (3),

task	1	2	3	others
41582 (50) •	TRF_UNKNOWN (49)	EXEPAND..._STAGEIN (1)		
41590 (50) •	TRF_UNKNOWN (50)			
41598 (50) •	TRF_UNKNOWN (49)	EXEPAND..._STAGEIN (1)		
41603 (50) •	TRF_UNKNOWN (50)			
41604 (50) •	TRF_UNKNOWN (50)			
41614 (45) •	TRF_UNKNOWN (45)			
41759 (40) •	TRF_UNKNOWN (40)			
41552 (36) •	TRF_UNKNOWN (36)			
41626 (33) •	TRF_UNKNOWN (33)			
41550 (32) •	TRF_UNKNOWN (31)	EXEPAND..._STAGEIN (1)		
41686 (32) •	TRF_UNKNOWN (32)			
41696 (31) •	TRF_UNKNOWN (31)			
41693 (29) •	TRF_UNKNOWN (29)			
41551 (27) •	TRF_UNKNOWN (27)			
41549 (27) •	TRF_UNKNOWN (27)			
41023 (16) •	TRF_UNKNOWN (16)			
41399 (14) •	TRF_UNKNOWN (14)			
41023 (8) •	TRF_UNKNOWN (8)			
41399 (7) •	TRF_UNKNOWN (7)			
40545 (6) •	EXEPAND...T_INFILE (5)	EXEPAND..._STAGEIN (1)		
40812 (2) •	EXEPAND...SIGTERM (2)			

- view
- select task
- status
- failure
- task
- 41590
- site
- pic
- error
- TRF_UNKNOWN



most common error messages

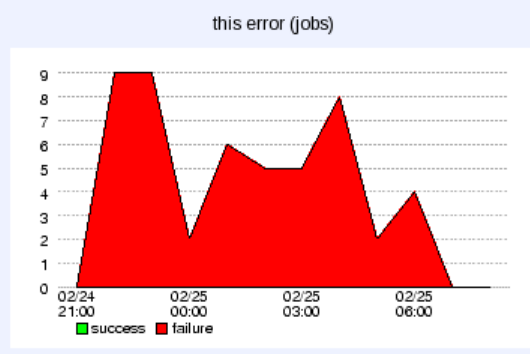
message (click to expand)	jobs
[Ermo 2] No such file or directory: 'ntuple_rdoesd.pmon.dat'	50

text/csv

jobs 50 to 50

jobexeid	jobdefk	taskfk	jobname	error	message
34014992	26367030	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10010.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34006857	26367030	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10010.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34000370	26367030	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10010.job	TRF_UNKNOWN	[Ermo 2] No such file or di
33990146	26367030	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10010.job	TRF_UNKNOWN	[Ermo 2] No such file or di
33982684	26367030	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10010.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34008887	26367029	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10009.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34001415	26367029	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10009.job	TRF_UNKNOWN	[Ermo 2] No such file or di
33995709	26367029	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10009.job	TRF_UNKNOWN	[Ermo 2] No such file or di
33982683	26367029	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10009.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34014991	26367028	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10008.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34006856	26367028	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10008.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34000373	26367028	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10008.job	TRF_UNKNOWN	[Ermo 2] No such file or di
33993471	26367028	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10008.job	TRF_UNKNOWN	[Ermo 2] No such file or di
33982682	26367028	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10008.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34006855	26367027	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10007.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34000372	26367027	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10007.job	TRF_UNKNOWN	[Ermo 2] No such file or di
33994811	26367027	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10007.job	TRF_UNKNOWN	[Ermo 2] No such file or di

- view
- select task
- status
- failure
- task
- 41590
- site
- pic
- error
- TRF_UNKNOWN



most common error messages

message (click to expand)	jobs
[Ermo 2] No such file or directory: 'ntuple_rdotsd.pmon.dat'	50

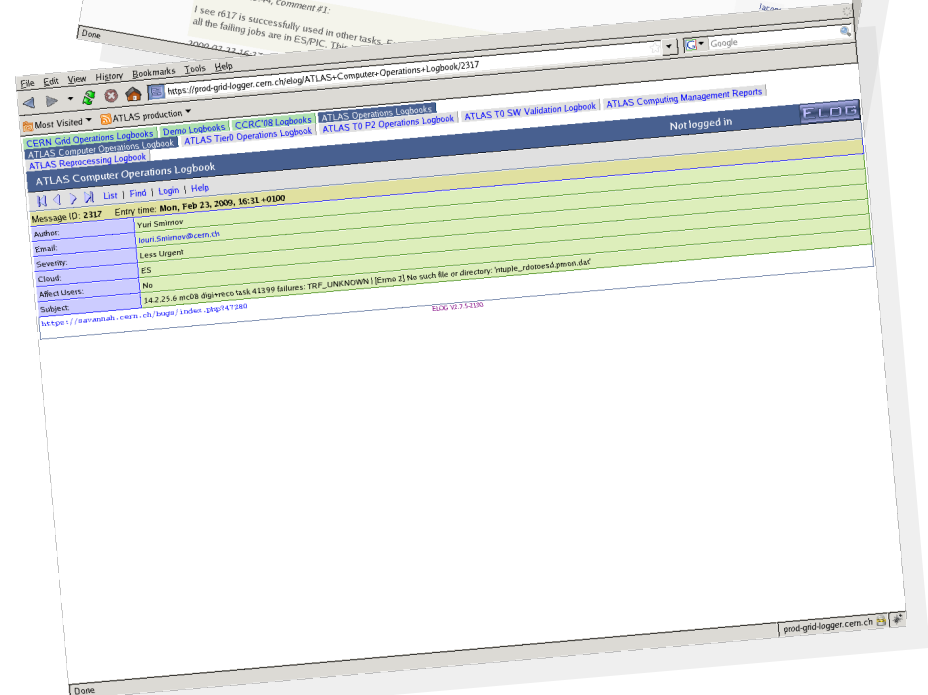
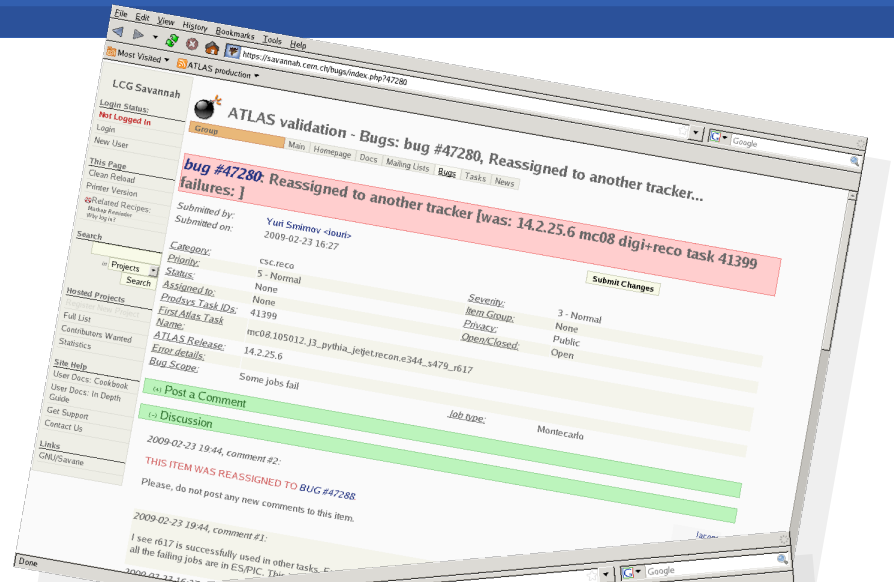
text/csv

jobs 50 to 50

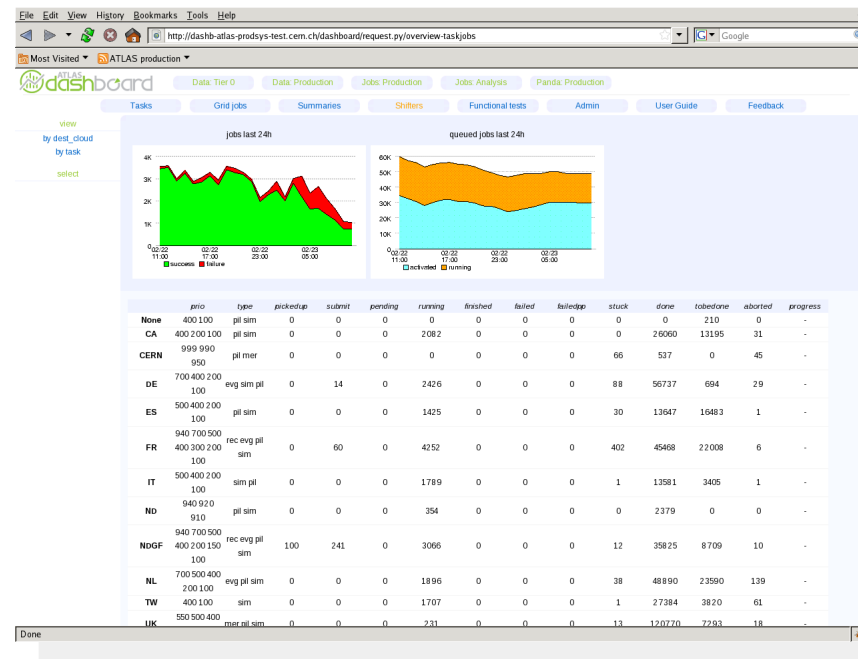
jobexeid	jobdefk	taskfk	jobname	error	message
34014992	26367030	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10010.job	TRF_UNKNOWN	[Ermo 2] No such file or di
34006857	26367030	41590	mc08.105805.filtered_minbias6.recon.e347_s462_r617_tid041590_10010.job	TRF_UNKNOWN	[Ermo 2] No such file or di

error text:
 [Ermo 2] No such file or directory: 'ntuple_rdotsd.pmon.dat'
 jobexeid: 34006857
 supervisor:
 infoexecutor:
 creationtime:
 attemptnr: 4
 errorcode:
 partnr:
 endtime: 2009-02-25 04:41:12
 modificationtime: 2009-02-25 04:54:35
 nevents:
 starttime: 2009-02-25 03:30:27
 exelcluster: pic-ce06-glong-lcgpbs
 processing host: td058.pic.es
 facilityid: [27248520](#) (click for logs)
 software: [14.2.25 Atlas-14.2.25](#)

- the shifter here submits a savannah bug,
- reports to the ATLAS eLog,
- bug fix will be logged as well,
- similar actions taken for site related errors (GGUS).

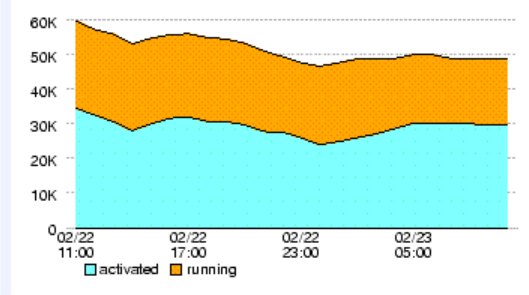
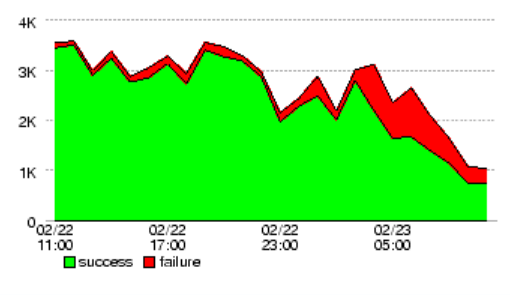


- **task display**
 - identification of stuck tasks
 - diagnostic page:
 - each stuck subjob:
 - *executed X times,*
 - *always failed.*
 - check common error patterns:
 - *same error in the same site,*
 - *same error in any site,*
 - *etc.*



by task

select



	prio	type	pickedup	submit	pending	running	finished	failed	failedpp	stuck	done	tobedone	aborted	progress
None	400 100	pil sim	0	0	0	0	0	0	0	0	0	210	0	-
CA	400 200 100	pil sim	0	0	0	2082	0	0	0	0	26060	13195	31	-
CERN	999 990 950	pil mer	0	0	0	0	0	0	0	66	537	0	45	-
DE	700 400 200 100	evg sim pil	0	14	0	2426	0	0	0	88	56737	694	29	-
ES	500 400 200 100	pil sim	0	0	0	1425	0	0	0	30	13647	16483	1	-
FR	940 700 500 400 300 200 100	rec evg pil sim	0	60	0	4252	0	0	0	402	45468	22008	6	-
IT	500 400 200 100	sim pil	0	0	0	1789	0	0	0	1	13581	3405	1	-
ND	940 920 910	pil sim	0	0	0	354	0	0	0	0	2379	0	0	-
NDGF	940 700 500 400 200 150 100	rec evg pil sim	100	241	0	3066	0	0	0	12	35825	8709	10	-
NL	700 500 400 200 100	evg pil sim	0	0	0	1896	0	0	0	38	48890	23590	139	-
40172	700	evgen				4				11	785			98.1%
40180	700	evgen				2				7		1		0%
38860	500	pile								1	11			91.7%
38861	500	pile								4	103		1	95.4%
41116	400	simul										10		0%
39223	400	pile				68				14	7766		133	97.3%
41031	400	simul				4					996			99.6%
41073	400	simul				1					99			99%

quick identification of failure patterns

84% of jobs failing with EXEPANDA_JOBKILL_SIGTERM: in RAL-LCG2

88% of jobs : in RAL-LCG2

100% of jobs failing with EXEPANDA_DQ2_STAGEIN: in RAL-LCG2

21040744 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00166.job [=> fail with EXEPANDA_JOBKILL_SIGTERM]

1	UKI-LT2-RHUL	EXEPANDA_JOBKILL_SIGTERM	Job killed by signal 15: Signal handler has set job result to FAILED, ec = 1201...
2	UKI-LT2-RHUL	EXEPANDA_JOBKILL_SIGTERM	Job killed by signal 15: Signal handler has set job result to FAILED, ec = 1201...
3	RAL-LCG2	EXEPANDA_JOBKILL_SIGTERM	Job killed by signal 15: Signal handler has set job result to FAILED, ec = 1201...
4	UKI-NORTHGRID-MAN-HEP	EXEPANDA_JOBKILL_SIGTERM	Job killed by signal 15: Signal handler has set job result to FAILED, ec = 1201...
5	RAL-LCG2	EXEPANDA_JOBKILL_SIGTERM	Job killed by signal 15: Signal handler has set job result to FAILED, ec = 1201...

21040747 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00169.job

21198885 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00185.job

21198889 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00189.job

21198894 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00194.job

21198906 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00206.job [=> fail with TRF_SEGFAULT]

21198995 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00295.job

21199002 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00302.job [=> fail with EXEPANDA_JOBKILL_SIGTERM]

21199004 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00304.job

21199010 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00310.job

21199022 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00322.job [=> fail with TRF_SEGFAULT]

21199023 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00323.job

21199062 - mc08.105017.J8_pythia_jetjet.recon.e344_s479_r563_tid028978._00362.job

action : manager cancels or reschedule jobs

- **Python API to access the monitoring info:**
 - Provided by the dashboard framework
 - Basic operations available (error breakdown, summary),
 - To automate tasks performed by shifters,
 - Not really successful so far... but I still hope.

```
File Edit View Terminal Tabs Help
bgaidioz@wynton: /home/bgaidioz  X bgaidioz@wynton: /home/bgaidioz/presentati... X
from datetime import datetime, timedelta

# import the production dashboard API
from dashboard.api.production.ProductionQuery import ProductionQuery

# search over three days
d2 = datetime.utcnow()
d1 = d2 - timedelta(days=3)

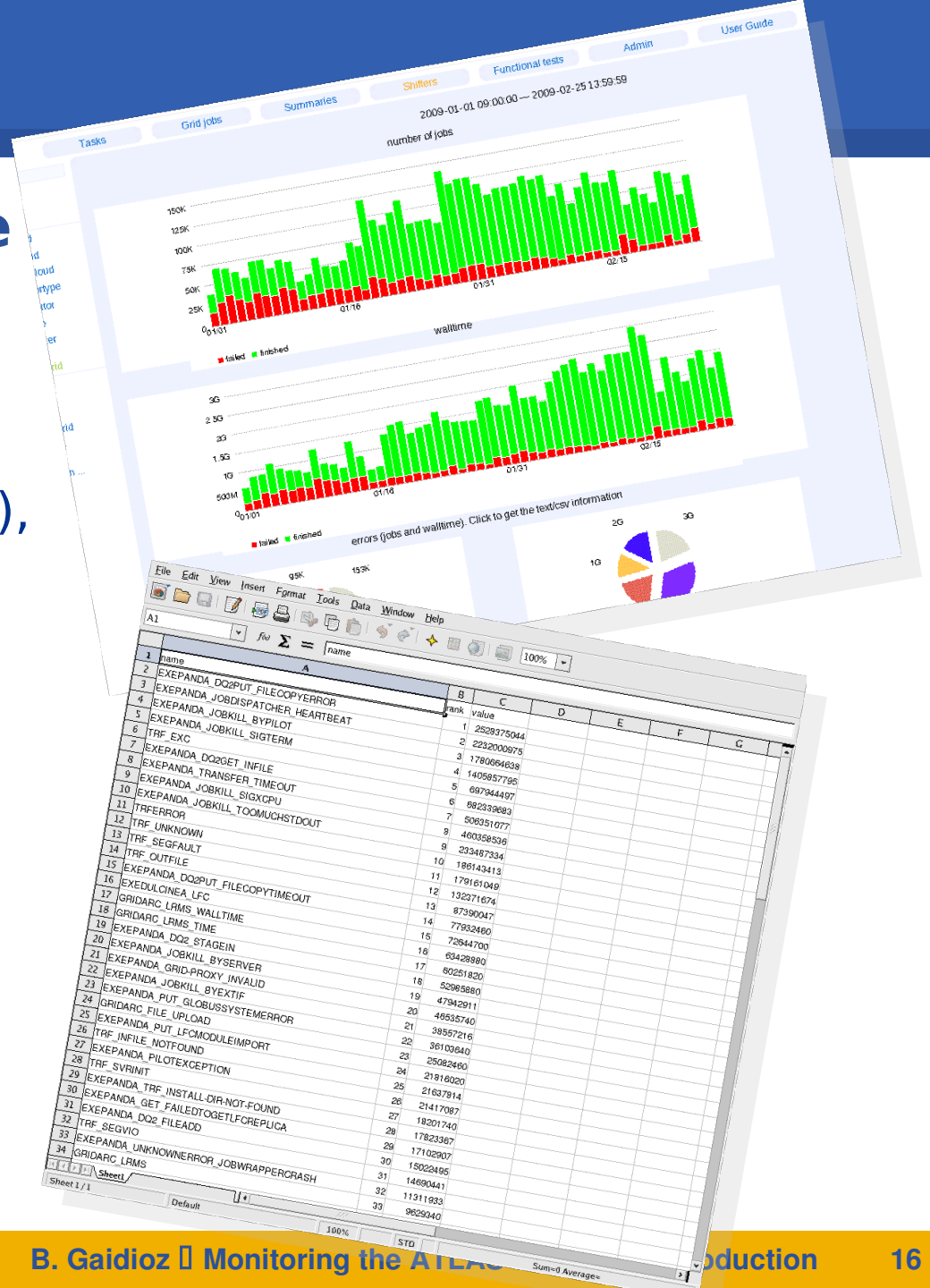
# instantiate the API
query = ProductionQuery('dashb-atlas-prodsys.cern.ch', 80)

# ask an error summary for a certain task (breakdown by site)
errors = query.errors(task=40420, grouping='site', startDate=d1,
                      endDate=d2)

for error in errors:
    # error contains the main three errors on a certain site
    # do something
~
~
~
~
```

19,18 All

- also available in the prodsys dashboard:
 - statistics over a long period
 - site admin page,
 - get data in CSV (Excel),
 - search ATLAS eLog.



- **alerts:**
 - use **data-mining** techniques
 - successful prototyping
 - see Gerhild Maier's work (was presented on Monday).
 - associate an error pattern to an “action to take”,
- **interface:**
 - include also **user analysis jobs** (prototype exists),
 - integrate fully with other ATLAS dashboards (DDM, AGIS),
 - assist the submission of **GGUS and Savannah** tickets.

- **dashboard tools for site admins will be presented on Friday:**
 - what are the VO running at my site?
 - what is the success rate of job processing or data transfer of the VO at my site?
 - is the VO happy with my site?
 - Friday at 9am. (room Leopardi.)