User experience with monitoring of the distributed production (ATLAS)

Xavier Espinal (PIC/IFAE)

Generalitat de Catalunya
Departament d'Innovació, Universitats i Empresa

MINISTERIO DE CIENCIA E INNOVACIÓN

Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

UAB
Universitat Autònoma de Barcelona

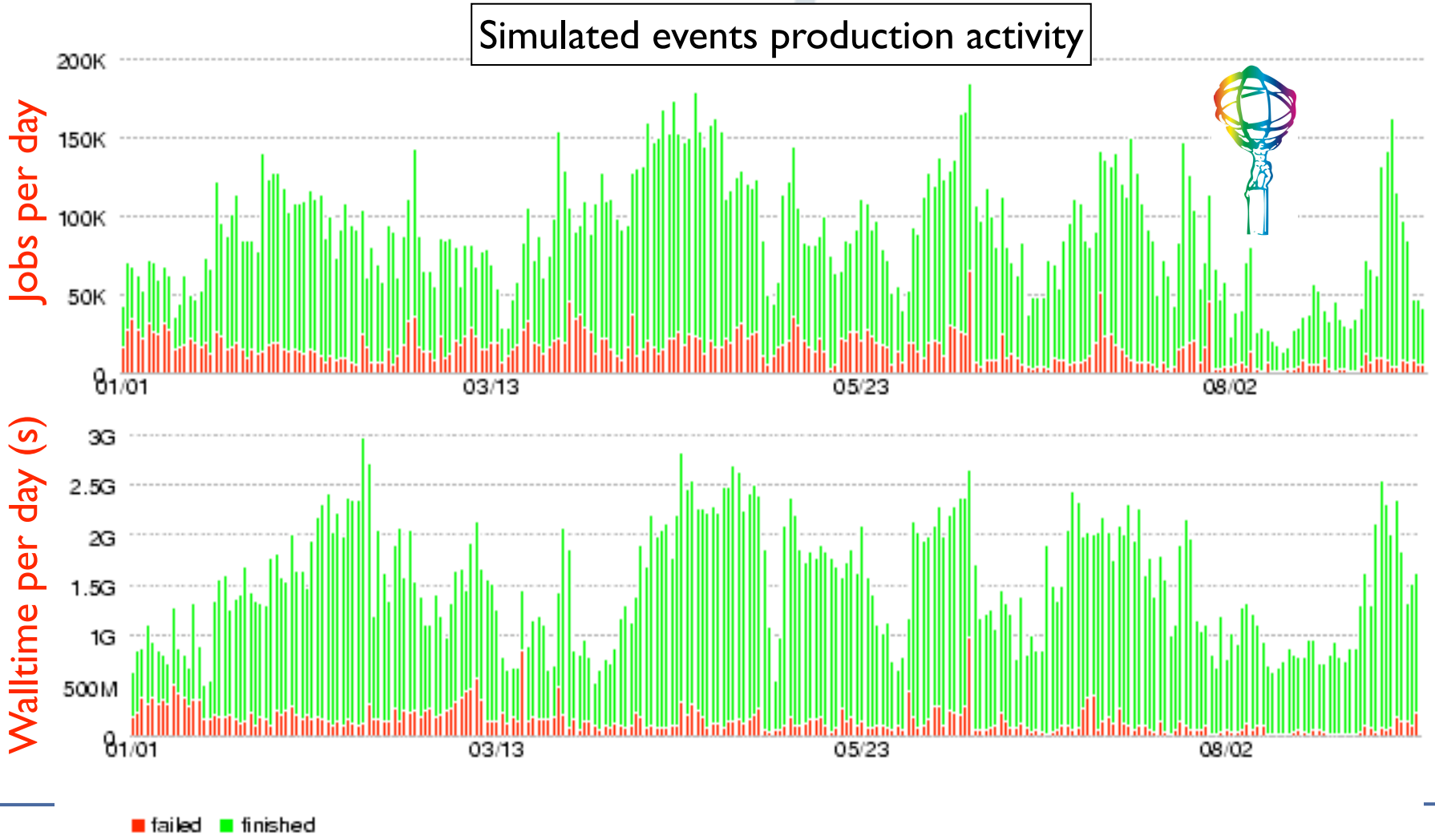
IFAE

Summary

- ATLAS simulated event production
  - Why simulated event production is so intensive and important to monitor ?
- ATLAS shift team
  - Which kind of monitoring is needed ?
- Monitoring the simulated production
  - Chasing sites
  - Chasing tasks
  - Production system functional tests
- Interaction with the dashboard team
- Summary and conclusions

ATLAS simulated events production

- Due to LHC's high SNR (10^{-9}) ATLAS needs a huge amount of simulated data
  - Normally running ~20-30k simultaneous jobs
  - Ending ~100k jobs per day involving ~50 different sites and a large number of different tasks
    - Stable shift teams are in place for monitoring event production since 2006
      - ➡ Efficient and robust monitoring is a must !

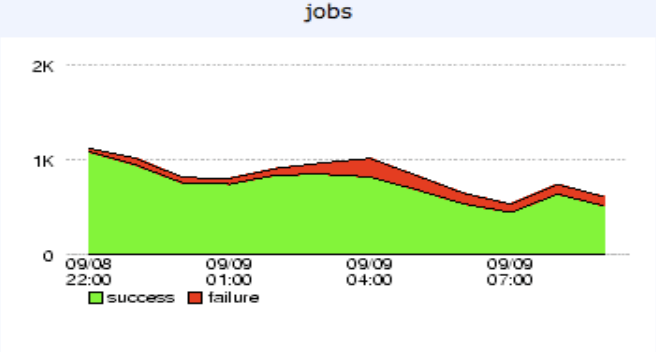


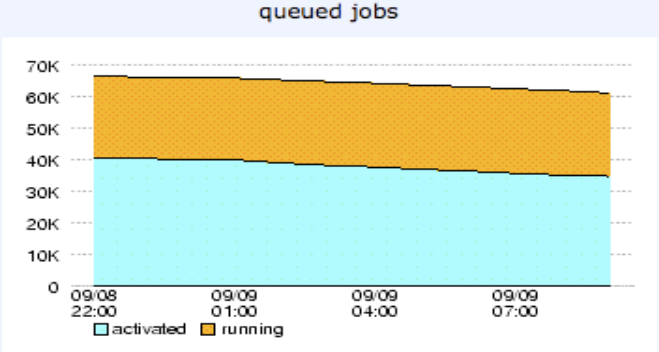
ATLAS shift teams

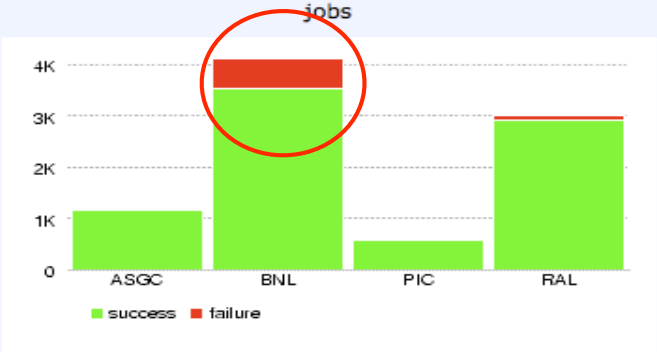
- ATLAS Distributed Computing Operations shifts team provides QoS to the ATLAS community.
- One of the main shifter duties is to track down the simulated event production world-wide
  - 50 people from different time-zones are involved providing 24/6 support
- Shifters do need very visual and easy to navigate monitoring pages
  - Lot of activities, try to minimize time browsing and getting to the relevant informations
- ATLAS production dashboard provides relevant visual information to quickly spot the main problems:
  - Clouds: LFC broken, FTS instabilities, T1 storage in troubles, etc.
  - Sites: failed access to input data, worker nodes misconfigurations, SW area not accessible, etc.
  - Tasks: wrongly defined tasks
  - Central services: Catalogues, PanDa, etc.
- High level views are needed
  - And obviously possibility to increase granularity and quickly get to the source of the problem
    - Heavy failures in a cloud => spot the site causing problems => get the CE/queue => get the job ID (either in ProdSys or in the local BS) => read the pilot/Athena/BS log file.
      - ➡ The root of the problem is found following top => bottom workflow
- ATLAS production dashboard is providing this

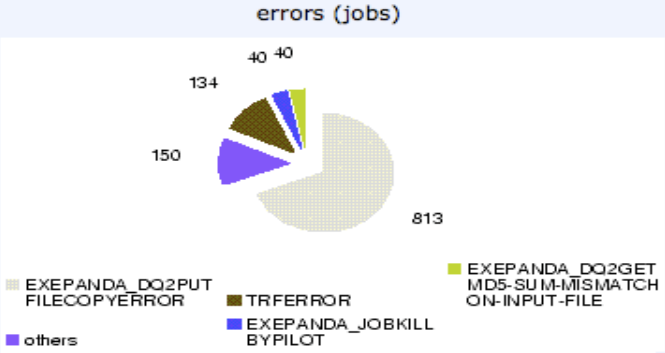
Monitoring production: chasing sites

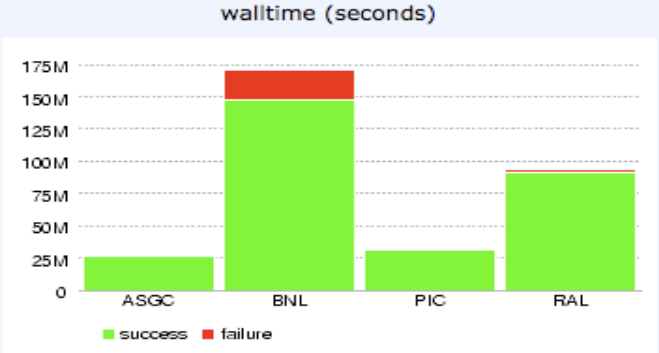
2009-09-08 22:00:00 — 2009-09-09 10:59:59











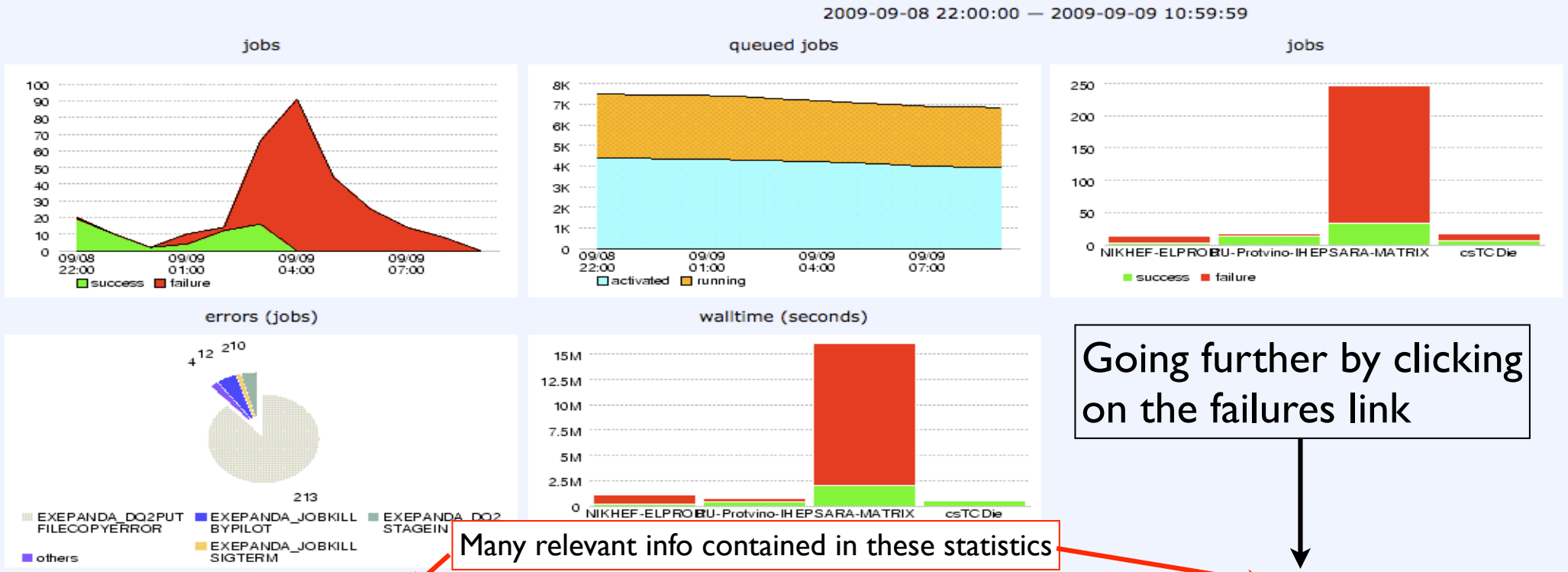
Click on the relevant cloud to get more info

cloud				activated	running	holding	transferring	success	failure	efficiency
BNL	0	93	0	13836	7664	1119	3334	3530	591	85.7%
RAL	0	0	0	3265	5186	391	1543	2919	99	96.7%
ASGC	0	0	0	16	266	430	57	1143	30	97.4%
PIC	0	0	0	1028	2208	321	2276	575	3	99.5%
SARA	0	1286	0	3928	2907	637	1597	63	241	20.7%
None	103	46	37	9215	693	353	922	183	15	92.4%
NDGF	0	0	0	0	2655	91	24	26	83	23.9%
TRIUMF	0	0	0	63	721	447	496	88	13	87.1%
CNAF	0	0	0	417	1072	39	1936	82	9	90.1%
FZK	0	0	0	2664	2863	211	9109	47	34	58%
LYON	0	0	0	0	94	613	50	49	11	81.7%
CERN	0	10	0	51	13	3	0	18	30	37.5%
total	*103*	*1439*	*37*	*34483*	*26342*	*4655*	*21344*	*8723*	*1159*	*88.3%*

SE in SD

CRITICAL | WARNING | NORMAL | GOOD | NO_ACTIVITY

Monitoring production: chasing sites



Going further by clicking on the failures link

Many relevant info contained in these statistics

site	defined	assigned	waiting	activated	running	holding	transferring	success	failure	efficiency
× SARA-MATRIX	0	1284	0	2538	798	21	0	33	213	13.4%

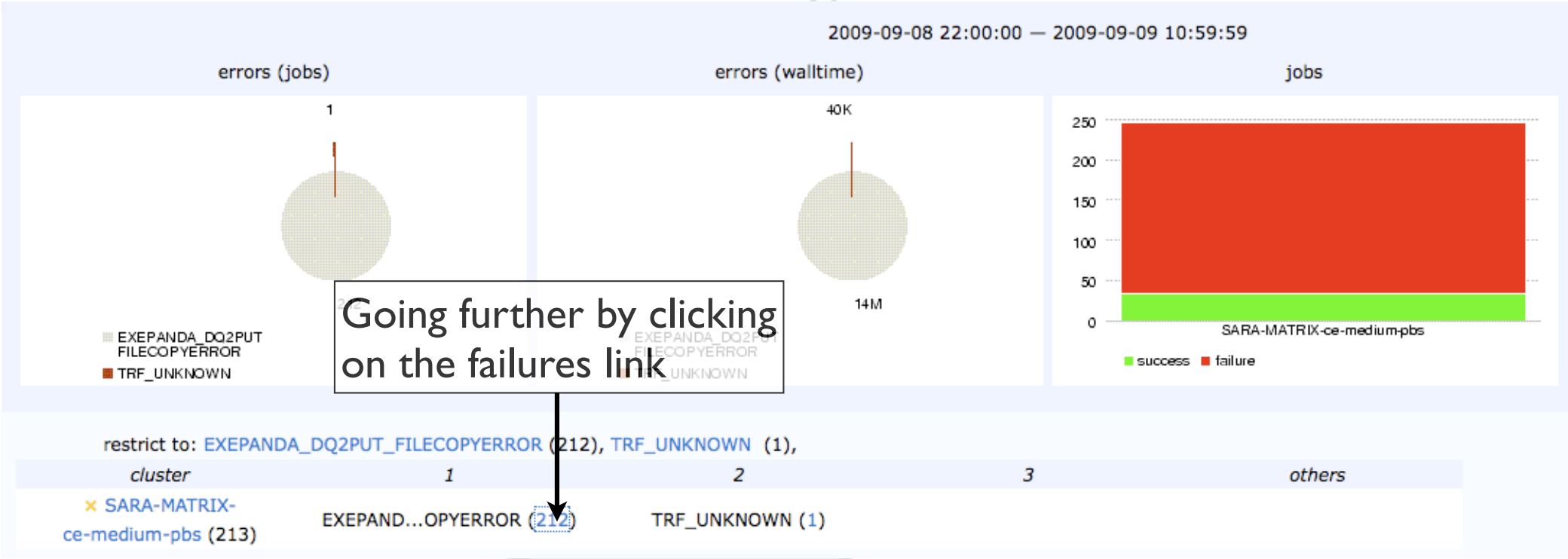
type	endpoint	status (SAM)	time
CE	ce.gina.sara.nl	ok	2009-09-08 12:24:20 UTC
CE	celisa.grid.sara.nl	na	2009-09-07 13:58:50 UTC

SAM CE tests results

× csTCDie	0	0	0	812	633	15	797	7	11	38.9%
× RU-Protvino-IHEP	0	0	0	110	127	1	158	14	3	82.4%
× NIKHEF-ELPROD	0	0	0	0	738	547	75	2	11	15.4%
× JINR-LCG2	0	0	0	346	232	3	131	6	3	66.7%
× TR-10-ULAKBIM	0	0	0	22	13	0	7	1	0	100%
× ru-PNPI	0	0	0	0	0	3	0	0	0	-
× RRC-KI	0	1	0	77	347	44	443	0	0	-
× ITEP	0	0	0	0	0	3	0	0	0	-
× ru-Moscow-SINP-LCG2	0	1	0	0	5	3	0	0	0	-
total	*0*	*1286*	*0*	*3905*	*2893*	*640*	*1611*	*63*	*241*	*20.7%*

Monitoring production: chasing sites

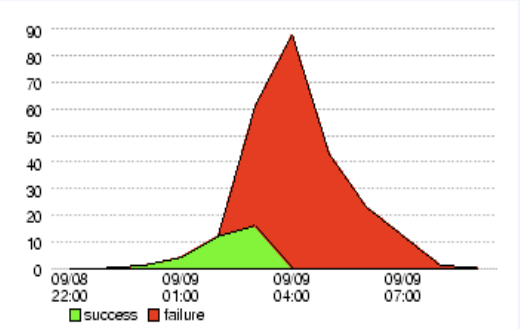
- From the following screenshot it is directly inferred that job cannot get the input files
  - Data cannot be copied from the WN to the SE
    - Error messages are well know by shifters (*DQ2PUT_FILECOPYERROR*)



- Now, shifters need more detailed info about the job. Several things could happen:
  - Failing stage-in/out: SE at the site is experiencing problems
    - Temporary outage/overload, single pool problem or big issue with the SE ?
  - Failing stage-in: data is not available at the site
    - Either input data is lost, job has been wrongly routed or there's problem at job definition
      - ➡ Luckily shifters can go further using the dashboard by clicking on the #errors

Monitoring production: chasing sites

this error (jobs)



most common error messages

message (click to expand)	*jobs*
Put error: Copy command returned error code 34304 and output: /opt/lcg/bin/lcg-cr lcg_util-1.7.6-1 GFAL-client-1.11.8-1	192
Put error: Copy command returned error code 256 and output: /opt/lcg/bin/lcg-cr lcg_util-1.7.6-1 GFAL-client-1.11.8-1 U	19
Copy command returned error code 256 and output: /opt/lcg/bin/lcg-cr lcg_util-1.7.6-1 GFAL-client-1.11.8-1 Using grid c	1

text/csv

COPYERROR

jobs 50 to 100

jobexeid	*jobdeffk*	*taskfk*	*jobname*	*error*	*message*
51718691	44019140	80192	mc09_valid.105013.J4_pythia_jetjet.simul.e344_s586_tid080192._001824.job	EXEPANDA_DQ2PUT_FILECOPYERROR	Put error: Copy command retur

error text:
Put error: Copy command returned error code 34304 and output: /opt/lcg/bin/lcg-cr lcg_util-1.7.6-1 GFAL-client-1.11.8-1 Using grid
catalog type: lfc Using grid catalog : lfc-atlas.grid.sara.nl Checksum type: None SE type: SRMv2 Destination SURL : srm
jobexeid: 51718691
supervisor:
infoexecutor:
creationtime:
attemptnr: 1
errorcode:
partnr:
endtime: 2009-09-09 03:36:37+00:00
modificationtime: 2009-09-09 03:40:12+00:00
nevents:
starttime: 2009-09-08 09:48:42+00:00
execluster: SARA-MATRIX-ce-medium-pbs
processing host: v31-21.gina.sara.nl
facilityid: 1020872490 (click for logs)
software: 15.1.0

Job execution details

Access to full job details

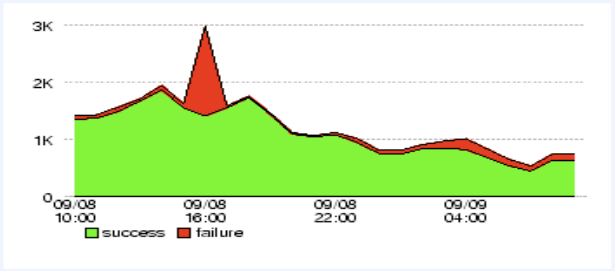
51718695	44019136	80192	mc09_valid.105013.J4_pythia_jetjet.simul.e344_s586_tid080192._001820.job	EXEPANDA_DQ2PUT_FILECOPYERROR	Put error: Copy command retur
51718657	44018671	80192	mc09_valid.105013.J4_pythia_jetjet.simul.e344_s586_tid080192._001355.job	EXEPANDA_DQ2PUT_FILECOPYERROR	Put error: Copy command retur

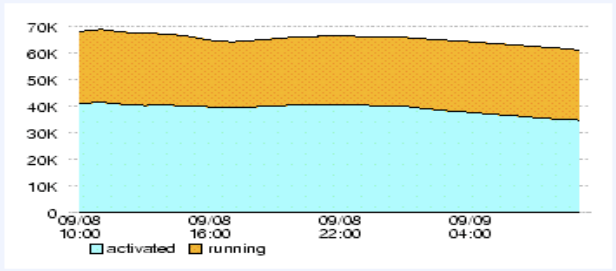
Monitoring production: chasing sites

- Once accessed to job's final information shifter has enough tools to triage the error and perform the required action:
  - Try to access the data (lcg-cp), try to store data (lcg-cr)
    - Disentangle transient vs. continuous problem
  - Browse the LFC
    - Disentangle lost vs. missing data at the site
- Then the problem is understood and shifters can report the problem:
  - Site problems: GGUS
  - Experiment related problems: Savannah

Monitoring production: chasing tasks/jobs

- Similar workflow is used by monitoring the problematic tasks
  - Monitoring duality: site efficiency vs. task efficiency
    - Snapshot of task progress and problematic jobs per cloud

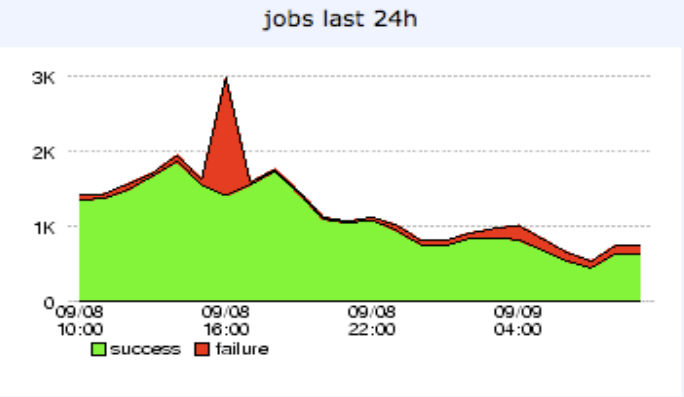


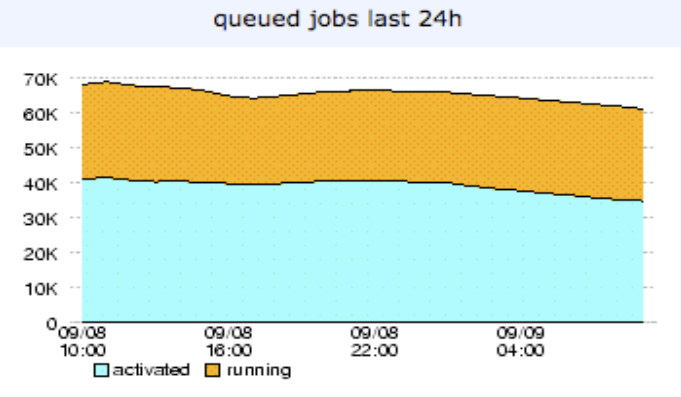


	prio	type	pickedup	submit	pending	running	finished	failed	failedpp	stuck	done	tobedone	aborted	progress
CA	999 950 920 600 300 100	pil sim rec	0	0	0	287	0	0	0	1	46466	0	1	-
CERN	999 950 600 500	mer sim rec	0	0	0	3	0	0	0	10	2111	44	0	-
DE	950 600 500 300 100	sim rec pil	0	0	0	202	0	0	0	3	108212	0	123	-
ES	950 600 300 100	sim rec	0	0	0	1582	0	0	0	0	23166	2148	6	-
FR	950 600 300 100	sim rec	0	0	0	110	0	0	0	1	115006	0	6	-
IT	950 600 300 100	sim rec	0	0	0	252	0	0	0	1	29443	0	4	-
ND	999 950 920 600 300 100	pil sim rec	0	0	0	110	0	0	0	18	65712	0	121	-
NL	950 600 300 100	sim rec	0	0	0	2242	0	0	0	0	51015	19037	15	-
TW	950 600 100	sim rec	0	0	0	274	0	0	0	0	25595	0	1	-
UK	950 600 500 300 200 100	sim rec pil	0	0	0	6856	0	0	0	1	76730	0	5	-
US	950 920 600 500 300 200 100 50 10	sim pil rec dig	0	0	0	11404	0	0	0	7	470633	969358	119	-

Monitoring production: chasing tasks/jobs

- Possibility to have a task breakdown per cloud with progress and problematic jobs (stuck)





task	prio	type	pickedup	submit	pending	running	finished	failed	failedpp	stuck	done	tobedone	aborted	progress
CA	999 950 920 600 300 100	pil sim rec	0	0	0	287	0	0	0	1	46466	0	1	-
80550	999	pile												0%
80541 •	950	simul				65					28			28%
80375 ×	920	simul				132					68			34%
80423 ×	920	pile				3				1	196			98%
80429 ×	920	pile				2					198			99%
80378 ×	920	simul				8					32			80%
70383 •	600	simul				19					3762		1	99.5%
80344 •	600	reco									180			97.3%
79325 •	300	simul				54					2006			97.4%
79934	100	simul				4					39996			100%
CERN	999 950 600 500	mer sim rec	0	0	0	3	0	0	0	10	2111	44	0	-
DE	950 600 500 300 100	sim rec pil	0	0	0	202	0	0	0	3	108212	0	123	-
ES	950 600 300 100	sim rec	0	0	0	1582	0	0	0	0	23166	2148	6	-
FR	950 600 300 100	sim rec	0	0	0	110	0	0	0	1	115006	0	6	-

Click to get detailed job info

Monitoring production: chasing tasks/jobs

- Shifter arrive to a nice summary of job attempt history
  - Job failed in several sites (ironing out site problems)
    - Shifter quickly know there must be something wrong with the job definition
      - ➡ Open bug to relevant people

44684056 — valid1.105200.T1_McAtNlo_Jimmy.recon.e380_s577_r774_tid080423._000056.job

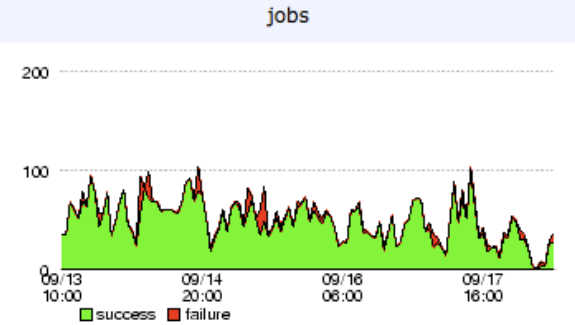
1	TORONTO-LCG2	TRFERROR	deprecated transexit code 10 / unknown...
2	TRIUMF-LCG2	TRF_UNKNOWN	Unchecked StatusCode in MuTagIMO::execute() from lib /opt/exp_software/atlas/p...
3	TORONTO-LCG2	TRF_UNKNOWN	Unchecked StatusCode in MuTagIMO::execute() from lib /opt/exp_software/atlas/p...
4	TRIUMF-LCG2	EXEPANDA_ATHENA_RAN-OUT-OF-MEMORY	Athena ran out of memory...
5	TRIUMF-LCG2	TRF_UNKNOWN	Unchecked StatusCode in MuTagIMO::execute() from lib /opt/exp_software/atlas/p...
6	TORONTO-LCG2	TRF_UNKNOWN	Unchecked StatusCode in MuTagIMO::execute() from lib /opt/exp_software/atlas/p...
7	TRIUMF-LCG2	TRF_UNKNOWN	Unchecked StatusCode in MuTagIMO::execute() from lib /opt/exp_software/atlas/p...

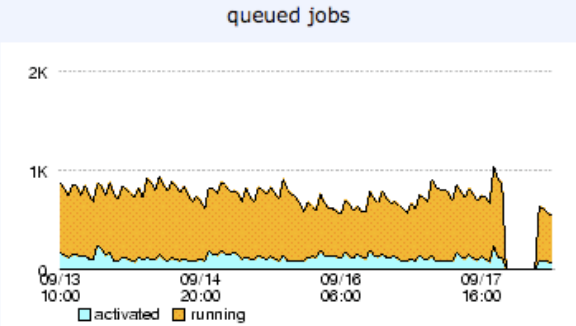
Monitoring production: production system functional tests

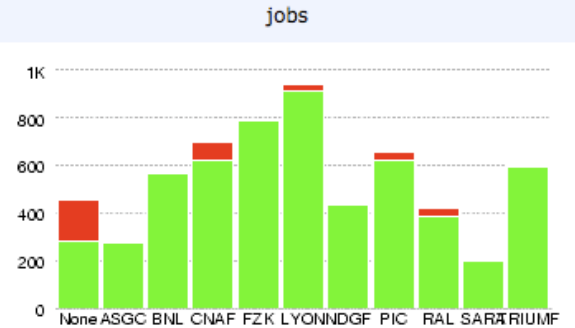
- ATLAS has a special group of well known jobs continuously running in the background
  - "well-known-jobs" means that problems with the SW, task/jobs definition, memory, etc. won't occur
    - Shifter can spot site issues immediately when those jobs fail
      - ➡ Usually it takes some time to disentangle site vs. SW issues...
- Production dashboard provides a HLV of the ATLAS Production Functional Tests

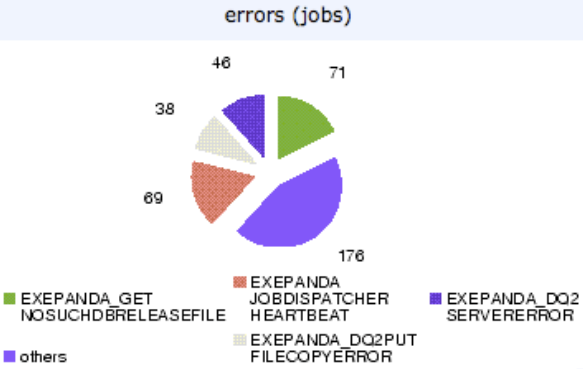
Monitoring production: production system functional tests

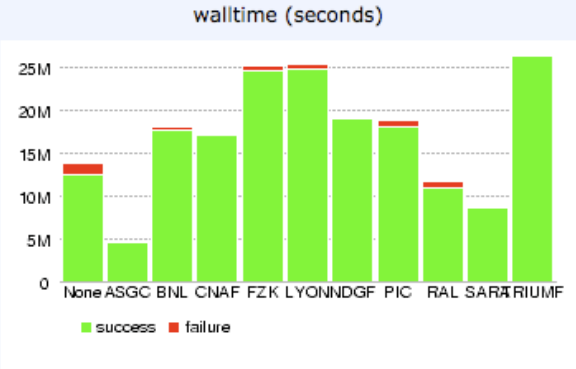
2009-09-13 10:00:00 — 2009-09-18 10:59:59











cloud	defined	assigned	waiting	activated	running	holding	transferring	success	failure	efficiency
LYON	0	0	0	0	3	0	44	909	31	96.7%
FZK	0	0	0	1	72	1	254	789	3	99.6%
CNAF	0	0	0	0	6	0	334	623	77	89%
PIC	0	0	0	0	53	8	178	618	40	93.9%
TRIUMF	0	0	0	0	139	8	241	591	12	98%
BNL	0	0	0	0	3	2	42	563	10	98.3%
None	0	0	0	36	84	3	28	279	175	61.5%
NDGF	0	0	0	0	88	0	0	434	1	99.8%
RAL	0	0	0	0	3	2	25	383	39	90.8%
ASGC	0	0	0	0	0	3	0	276	2	99.3%
SARA	0	0	0	8	4	0	75	200	1	99.5%
CERN	0	0	0	0	10	0	0	20	9	69%
total	*0*	*0*	*0*	*45*	*465*	*27*	*1221*	*5685*	*400*	*93.4%*

CRITICAL | WARNING | NORMAL | GOOD | NO_ACTIVITY

Providing feedback to dashboard team

- Quite nice interaction among shifters and dashboard developers
- Usually shifters submit bugs/suggestions using the dashboard itself
- Intensive work since the beginning and almost continuous improvement
  - Thanks to Ricardo and Ben for his work !

- User guide also in place for newcomers:

Home Next

Search

Arda Dashboard Production System Monitoring

Guide For Atlas Prodsys users

Benjamin Gaidioz

<benjamin.gaidioz@cern.ch>

ARDA Dashboard Production System Monitoring User's Guide

The Dashboard project is developed within the ARDA group and is part of the Enabling Grids For E-SciencE (EGEE) project. The software it provides is able to collect information regarding grid jobs and transfers from different sources and locations, and exposes to the grid end user a processed view of this data. This guide covers production systems and describe a end user view over the monitoring information.

Table of Contents

List of Figures

Summary and conclusions

- Large number of production jobs running: O(100k) per day in ATLAS
- Large number of sites ~50
- Strong need for shift team...
  - ...which needing an efficient monitoring!
- Production dashboard is providing the required level of monitoring and features that shift teams need
  - And obviously is constantly improving and adapting to the new requirements/services
- Production dashboard is a catch-all entry point holding all relevant info for simulated production:
  - Gathers info from: cic (SD), SAM, Production System database and PanDa
    - And is collecting and publishing this information in an structured way so it's useful for shifters
      - ➡ Reduce the "clicking" and browsing of different portals
- Currently having about 1000 page views per day and 600 unique visitors
  - And increasing