Monitoring for Support

Stefano for Crab2 Ops

We run a working system

- So far jobs mostly succeed
  - We count on this to keep being the case
- Yet at times something goes wrong
  - 80 sites, 500 users, there's always somebody with a problem
- Would like to be proactive, not complain-driven, in finding and fixing issues
  - largely unsuccessful so far
  - have oversold the line “will be better with Crab3”

How well do we fare ?

- How many jobs we run ? At which sites ?
- How's the distribution by user, dataset, site, exit code, running time... ?
- This is usually under control
- And not much interesting
-
- Nothing as boring as a system that works :-)
-

Problem stuff is more fun

- Users questions:
  - Why my jobs did not start running yet
  - Why not all my jobs have run yet
  - Why my jobs failed (aborted, removed, held …)
- Support questions:
  - Is some site not working, overloaded, failing ?
  - Is some user abusing of the system ?
  - Are pilots running ? Succeeding ?
  - Are we getting all the resources we should ?
  - Is data access working ? On WAN ?

Global view: sort of under control

- With a couple integrated weeks of coding effort times a few months of experience, James gave us a very good summary
- http://submit-3.t2.ucsd.edu/CSstoragePath/UserPrio/UserPrio-dev.html
- Continuously evolving
- The tricky part is not as much to spot problems there as to figure out what's going on and what to do
- Also this has likely reached the complexity limit for an “unstructured” tool

Overview: how much

CMS Analysis Monitoring for glideinWMS (remoteGlidein)
Updated every 20 minutes.
Last update Wed May 15 22:01:52 UTC 2013 (15:01:52 PDT)

SUMMARY OF QUEUED AND RUNNING USER JOBS FROM ALL SCHEDDS

Name	TotalRunningJobs	TotalIdleJobs	TotalHeldJobs
cmssrv168.fnal.gov	0	0	0
cmssubmit-r1.t2.ucsd	0	0	0
hcc-crabserver.unl.e	7643	10169	570
submit-2.t2.ucsd.edu	6689	7872	3947
submit-4.t2.ucsd.edu	7180	9915	4740
submit-5.t2.ucsd.edu	1	0	0
submit-6.t2.ucsd.edu	0	0	0
vocms20.cern.ch	32	0	0
vocms83.cern.ch	0	0	0
Total	21545	27956	9257

Overview: how well ?

SUMMARY OF CONDOR_HISTORY FOR 24 HOURS ENDING May 15 14:40 (PDT)

Number of Jobs	Final JobStatus	Previous JobStatus	WC Days	CPU Days	WC/Job(h)	CPU/WC(%)	ImageSize/Job (GB)	Starts/Job
75154	COMPLETED(4)	RUNNING(2)	19115	14911	6.1	78.0	1.3	1.1
39	REMOVED(3)	IDLE(1)	5	2	3.1	40.0	0.5	0.5
1071	REMOVED(3)	RUNNING(2)	945	748	21.2	79.2	1.2	2.3
454	REMOVED(3)	REMOVED(3)	36	7	1.9	19.4	1.5	1.1
0	REMOVED(3)	HELD(5)	0	0				
60	HELD<24h(5)	-	28	-	11.2	-	0.0	0.2

CRAB ExitCode	Number of Jobs	Percentage of Jobs
Success (rc=0)	56815	74.0569 %
Failed (rc!=0)	17974	23.4287 %
Unknown (rc=?)	1929	2.5144 %

Overview: where ?

DETAILS OF CLAIMED AND IDLE PILOTS BY SITE

Site	SE(s)	Claimed Pilots	Unclaimed Pilots
T1_US_FNAL	cmssrm.fnal.gov	3	
T2_AT_Vienna	hephyse.oeaw.ac.at	24	25
T2_BE_IIHE	maite.iihe.ac.be	906	4
T2_BE_UCL	ingrid-se02.cism.ucl.ac.be	498	47
T2_BR_SPRACE	osg-se.sprace.org.br	542	12
T2_CH_CERN	srm-eoscms.cern.ch	460	44
T2_CH_CSCS	storage01.lcg.cscs.ch	23	29
T2_CN_Beijing	srm.ihep.ac.cn	20	24
T2_DE_DESY	dcache-se-cms.desy.de	1434	17
T2_DE_RWTH	grid-srm.physik.rwth-aachen.de	1135	10
T2_EE_Estonia	ganymede.hep.kbfi.ee	903	13
T2_ES_CIEMAT	srm.ciemat.es	763	3
T2_ES_IFCA	storm.ifca.es,srm01.ifca.es	2	
T2_FI_HIP	madhatter.csc.fi	2	87
T2_FR_CCIN2P3	ccsrmt2.in2p3.fr	369	
T2_FR_GRIF_IRFU	node12.datagrid.cea.fr	125	2
T2_FR_GRIF_LLR	polgrid4.in2p3.fr	41	4
T2_FR_IPHC	sbgse1.in2p3.fr	252	25
T2_GR_Ioannina	grid02.physics.uoi.gr	1	5

Overview: who ?

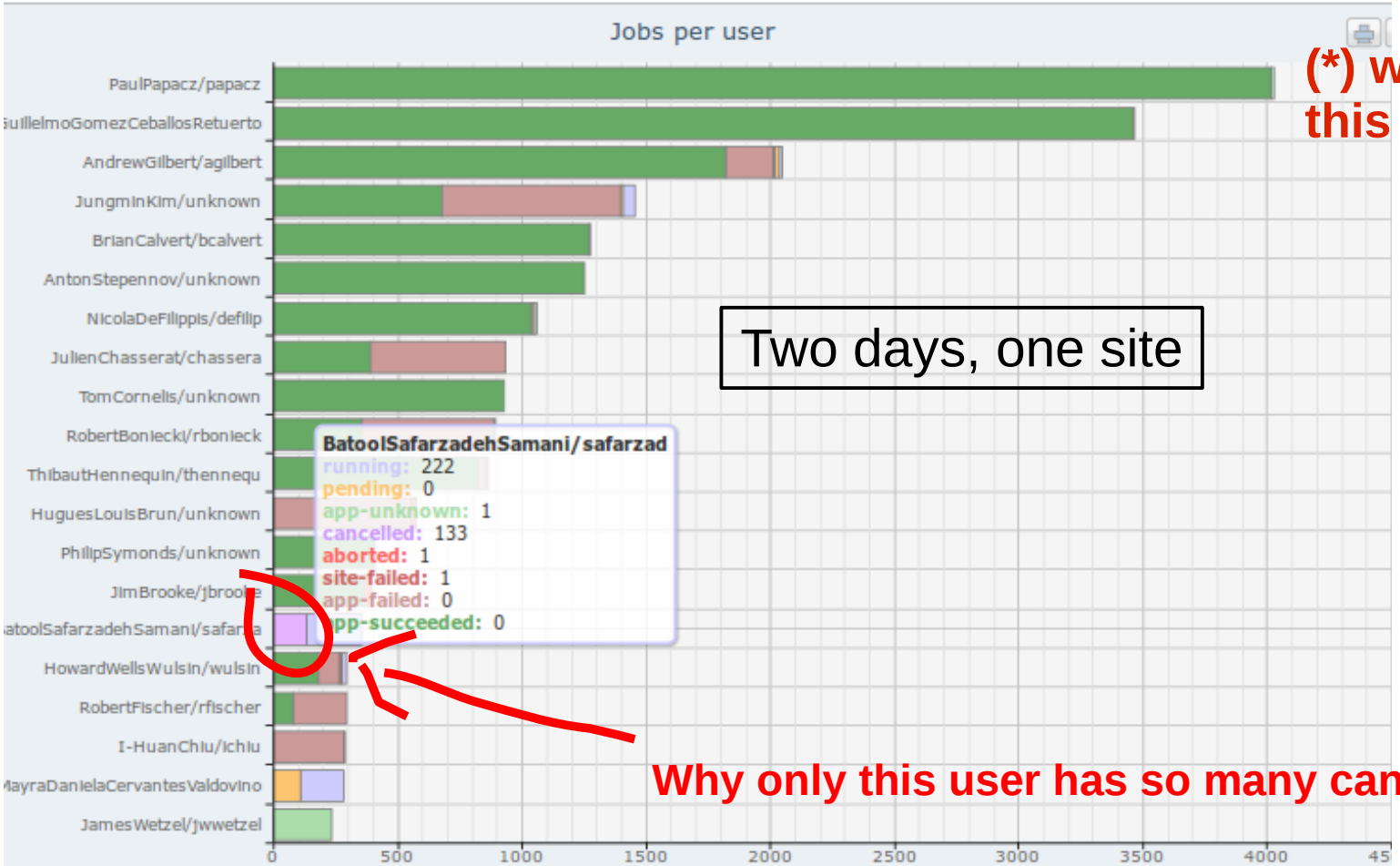
DETAILS OF USERS OVER 50% OF DISK QUOTA ON ANY OF THE SCHEDULERS:

uscms4149 (1) is using 65% of disk quota (100GB) on submit-2.
uscms3358 (nicola.defilippis@ba.infn.it) is using 54% of disk quota (100GB) on submit-4.

User Priorities and Running and Queued Job Summaries. Click links for details.

CMS Userid	DN	Priority	Running Jobs	Queued Jobs (Never Ran)	Avg. Time Queued Jobs (hours) ▲	Re-Queued Jobs (Ran, Failed, Re-queued)	Avg. Time Re-Queued Jobs (hours)	Unmatchable Job*	Held (Failed) Jobs
mkortela	/O=Grid/O=NorduGrid /OU=helsinki.fi/CN=Matti Kortelainen	63098.02	82	106	111	0	0	0	143
yeshaq	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=yeshaq /CN=696027/CN=Yossof Eshaq	1037547.62	622	4057	86	0	0	0	1
eaa	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=eaa /CN=666896/CN=Elif Asli Yetkin	1256752.62	396	2455	59	0	0	0	6
fgior8	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=fgior8 /CN=662088/CN=Ferdinando Giordano	259674.53	261	42	34	5	11	3	183
bbilin	/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=bbilin /CN=705521/CN=Bugra Bilin	267573.41	29	7	31	34	31	5	13

Drill down example (using another tool) (from an user who complained) (*)



(*) we did not spot this in our overview

Two days, one site

Why only this user has so many cancelled jobs ?

RemoveReason = "The system macro SYSTEM_PERIODIC_REMOVE expression '((NumJobStarts>9)=?=True)||((NumShadowStarts>19)=?=True)' evaluated to TRUE"

NumJobStarts = 10

(human time) expensive CLI usage needed to get this

Pilots help, but...

- When pilot fails to run the payload, diagnostic may be difficult
- e.g. glexec
  - “glexec failed to map, details are in syslog”
  - means logs local to the remote WN
  - mitigated (delayed retry to avoid black holes), not solved
  - believe that setting up dedicated monitoring is not worth, but it is a bet on future behavior of systems we (central support) do not control

Moving target

- Problems behind the questions often change
  - We fix one and discover the next ?
  - Stuff changes ?
  - We keep changing tools, patterns ?
- Distinct danger (and some example) of setting up monitoring for issues which are not relevant any more
-

High level specs (thx Igor)

- a) We need something that is flexible.
  - We both do not understand the system enough and the conditions change with time, so we cannot spec what we need in advance.
- b) We need intuitive visualization.
  - Getting the raw data is easy... just use the command line.
  - Displaying it in a way that problems jump to the eye is the real hard part.

c) need data mining/analytics

-
- The problems are usually hidden... it is very rare that a whole category of jobs fails at the same time... usually it is much more subtle.
-
- And it is tied to (a)(*), so pre-defining searches is not a comprehensive solution
  - (*) We both do not understand the system enough and the conditions change with time, so we cannot spec what we need in advance.

d) different levels of visualization

-
- We usually want to see the high level view of the system only.
-
- But when we do spot problems, we must be able to drill down to the smallest of the details, until we get to the root cause (if at all possible).

e) summary page

- We need a summary page for monitoring the known trouble issues.
- Basically, James' page equivalent must be easy to set up in the chosen monitoring tool, so we don't have to have a separate thing for it.