Toolforge webservices are in the final stages of migrating to the toolforge.org domain.
Please help us clean up older documentation referring to tools.wmflabs.org!

Incident documentation/20160112-20160111-toollabs-SGE

< Incident documentation

Contents [hide]

Summary

From 20:00 UTC to 22:15 UTC, jobs submitted to SGE were queued but not executed, or executed with a large delay. Around 22:15 UTC, all queued jobs had started.

Timeline

See also Tools Server admin log entries during this time

Timeline 1: IRC

Yuvipanda and valhallasw are working on phab:T123270, which required changes to the SGE configuration. This change would limit the number of concurrent jobs per user.

- 19:59:06 YuviPanda sets `maxujobs` (max concurrent jobs per user) on gridengine to 128
- 20:02:11 valhallasw tests by submitting ~150 jobs. This doesn't cause an error, but some jobs do get into an error state. valhallasw qdels jobs.
- 20:10 valhallasw tries again, submitting ~150 jobs into the `cyberbot` queue (which only queues and never executes for non-cyberbot). No error.
- 20:10 we figure out we are using the wrong configuration setting.

an hour later

- 21:15:54 `<rohit-dua> my jstart jobs is in a state of qw..(including cron) is this a labs error?`

this takes a while to get noted,

- 21:30 valhallasw starts investigating
- 21:33:22 `<akoopal> hmm, can't get the webservice of the erwin85 tools restarted`
- valhallasw checks result of `qstat -j <jobid>` and notices a lot of queues in overload state because of the number of jobs recently started. Connects this to phab:T110994
- 21:37 yuvi is back
- 21:38 valhallasw suggests to reduce the "number of jobs started in the last X minutes" counter by setting load_adjustment_decay_time to something low and back to 7:30
- 21:40:33 valhallasw sets load_adjustment_decay_time to 0:0:1
- 21:41:12 valhallasw notes this doesn't help
- 21:41 *353 jobs in qw state*
- 21:42:44 valhallasw resets load_adjustment_decay_time to 0:7:30, as it's not having the intended effect
- 21:45:35 YuviPanda restarts gridengine master
- 21:47 *309 jobs in qw state*

Main page
Recent changes
Server admin log (Prod)
Server admin log (RelEng)
Deployments
SRE/Operations Help
Incident status

Cloud VPS & Toolforge
Cloud VPS documentation
Toolforge documentation
Request Cloud VPS project
Server admin log (Cloud VPS)

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Print/export
Create a book
Download as PDF
Printable version

- 21:50 *336 jobs in qw state*
- 21:53:29 valhallasw notices lighttpd queues are not overloaded, but lighttpd jobs are also not starting
- 21:55:59 valhallasw sets job_load_adjustments_decay_time = 0:0:0
  - this clears the list of overloaded queues, but jobs are still not scheduling
  - after resetting to 0:7:30, overload indicators return
- 21:58:06 YuviPanda notes some lighttpd queues are also overloaded
- 22:05:53 valhallasw reverts `maxujobs` config change, but this does not change anything
- 22:07:41 valhallasw removes all load adjustments by setting `job_load_adjustments` to `none` and `load_adjustment_decay_time` to 0:0:0. No effect.
- 22:12:30 YuviPanda restarts grid master again (maybe the configuration needs to be force-reloaded?)
- this drops the number of jobs in queue to 406, but then the number increases again
- 22:13 valhallasw opens the grid-master messages file, and notes:

```
01/11/2016 22:13:19|schedu|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
01/11/2016 22:13:19|worker|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
01/11/2016 22:13:19|worker|tools-grid-master|W|Skipping remaining 360 orders
```

- 22:14 valhallasw qdels 2221233 and all jobs in the queue are immediately submitted
- 22:19 valhallasw resets SGE configuration to pre-outage state [`maxujobs: 128` , `job_load_adjustments: np_load_avg=0.50` , `load_adjustment_decay_time: 0:7:30`]

Timeline 2: grid engine log file

from `/data/project/.system/gridengine/spool/qmaster/messages`

```
01/11/2016 19:36:04|worker|tools-grid-master|E|error writing object with key
"USER:tools.altobot" into berkeley database: (22) Invalid argument
01/11/2016 19:36:04|worker|tools-grid-master|W|aborting transaction (rollback)
01/11/2016 19:36:23| timer|tools-grid-master|E|Corrupted database detected.
Freeing all resources to prepare for a reconnect with recovery.
01/11/2016 19:36:23| timer|tools-grid-master|E|error checkpointing berkeley db:
(-30973) BDB0087 DB_RUNRECOVERY: Fatal error, run database recovery
01/11/2016 19:36:23| timer|tools-grid-master|E|trigger function of rule "default
rule" in context "berkeleydb spooling" failed
[...]
[valhallasw submits 150 jobs, triggering an avalanche of]
01/11/2016 20:01:19|worker|tools-grid-master|W|job 2219151.1 failed on host
tools-exec-1207.eqiad.wmflabs general searching requested shell because:
01/11/2016 20:01:18 [1092:23840]: execvlp(/var/spool/gridengine/execd/tools-
exec-1207/job_scripts/2219151, "/var/spool/gridengine/execd/tools-exec-
1207/job_scripts/2219151") failed: No such file or directory
01/11/2016 20:01:19|worker|tools-grid-master|W|rescheduling job 2219151.1
[... the last one of which is ...]
01/11/2016 20:02:31|worker|tools-grid-master|W|rescheduling job 2219318.1
[...]
[... every now and then we see a ...]
01/11/2016 20:12:23| timer|tools-grid-master|E|error checkpointing berkeley db:
(-30973) BDB0087 DB_RUNRECOVERY: Fatal error, run database recovery
[...]
01/11/2016 21:00:23|worker|tools-grid-master|E|Corrupted database detected.
Freeing all resources to prepare for a reconnect with recovery.
01/11/2016 21:00:23| timer|tools-grid-master|E|Corrupted database detected.
Freeing all resources to prepare for a reconnect with recovery.
01/11/2016 21:00:23| timer|tools-grid-master|E|error checkpointing berkeley db:
(-30973) BDB0087 DB_RUNRECOVERY: Fatal error, run database recovery
01/11/2016 21:00:23|worker|tools-grid-master|E|error starting a transaction: (-
30973) BDB0087 DB_RUNRECOVERY: Fatal error, run database recovery
01/11/2016 21:00:23|worker|tools-grid-master|W|rule "default rule" in spooling
context "berkeleydb spooling" failed writing an object
01/11/2016 21:00:23| timer|tools-grid-master|E|trigger function of rule "default
rule" in context "berkeleydb spooling" failed
01/11/2016 21:00:23|worker|tools-grid-master|E|error writing object "2221233.1"
to spooling database
01/11/2016 21:00:25|worker|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
```

```
2221233.1 twice
01/11/2016 21:00:25|worker|tools-grid-master|W|Skipping remaining 0 orders
01/11/2016 21:00:25|schedu|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
01/11/2016 21:00:25|worker|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
01/11/2016 21:00:25|worker|tools-grid-master|W|Skipping remaining 0 orders
[... repeats ...]
01/11/2016 21:00:29|worker|tools-grid-master|W|Skipping remaining 1 orders
[...]
01/11/2016 21:01:03|worker|tools-grid-master|W|Skipping remaining 10 orders
[...]
01/11/2016 21:11:03|worker|tools-grid-master|W|Skipping remaining 102 orders
[...]
01/11/2016 21:25:04|worker|tools-grid-master|W|Skipping remaining 203 orders
[...]
01/11/2016 21:45:07|worker|tools-grid-master|W|Skipping remaining 288 orders
01/11/2016 21:45:07|worker|tools-grid-master|W|Skipping remaining 248 orders
[this is probably Yuvis restart, but there's no actual restart message!]
[...]
01/11/2016 21:57:27|worker|tools-grid-master|W|Skipping remaining 312 orders
01/11/2016 21:57:34|worker|tools-grid-master|W|Skipping remaining 258 orders
[unclear why this happens -- this is at the same time
job_load_adjustments_decay_time is reset to 7:30]
[...]
01/11/2016 22:04:40|worker|tools-grid-master|W|Skipping remaining 409 orders
01/11/2016 22:04:41|worker|tools-grid-master|W|Skipping remaining 278 orders
[... again at the same time as a config change ...]
[...]
01/11/2016 22:07:37|worker|tools-grid-master|W|Skipping remaining 278 orders
01/11/2016 22:07:43|worker|tools-grid-master|W|Skipping remaining 424 orders
[... this is around when job_load_adjustments_decay_time is zeroed again]
[...]
01/11/2016 22:10:10|worker|tools-grid-master|W|Skipping remaining 439 orders
01/11/2016 22:10:14|worker|tools-grid-master|W|Skipping remaining 374 orders
[... grid master restart?]
[...]
01/11/2016 22:12:31|worker|tools-grid-master|W|Skipping remaining 404 orders
01/11/2016 22:12:35|worker|tools-grid-master|W|Skipping remaining 355 orders
[... or this one? ]
01/11/2016 22:13:45|schedu|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
01/11/2016 22:13:45|worker|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
01/11/2016 22:13:45|worker|tools-grid-master|W|Skipping remaining 362 orders
01/11/2016 22:13:45|schedu|tools-grid-master|E|scheduler tries to schedule job
2221233.1 twice
01/11/2016 22:13:48|worker|tools-grid-master|E|denied: job "2221233" does not
exist
[... this was valhallasw qdel'ing twice ...]
01/11/2016 22:14:58|worker|tools-grid-master|W|unable to find job 2221984 from
the scheduler order package
01/11/2016 22:14:58|worker|tools-grid-master|W|Skipping remaining 0 orders
[... and we're back! ...]
```

Job 2221233

```
valhallasw@tools-bastion-01:~$ qstat -j 2221233
==============================================================
job_number:                 2221233
exec_file:                  job_scripts/2221233
submission_time:            Mon Jan 11 21:00:22 2016
owner:                      tools.phetools
uid:                        52004
group:                      tools.phetools
gid:                        52004
sge_o_home:                 /data/project/phetools
sge_o_log_name:             tools.phetools
sge_o_path:                 /usr/bin:/bin
sge_o_shell:                /bin/sh
sge_o_workdir:              /data/project/phetools
sge_o_host:                 tools-submit
account:                    sge
cwd:                        /data/project/phetools/botpywi
stderr_path_list:
NONE:NONE:/data/project/phetools/log/sge/hocr_12564.err
hard resource_list:         release=trusty,h_vmem=2097152k
mail_list:                  tools.phetools@tools.wmflabs.org
notify:                     FALSE
job_name:                   hocr
stdout_path_list:
NONE:NONE:/data/project/phetools/log/sge/hocr_12564.out
jobshare:                   0
hard_queue_list:            task
env_list:                   LANG=en_US.UTF-8
job_args:                   '/data/project/phetools/phe/hocr/hocr.py','-
lang:bn','-book:হিতোপদেশঃ.pdf'
script_file:                /usr/bin/python2.7
usage    1:                 cpu=01:11:22, mem=673.40125 GBs, io=8.63248,
vmem=315.504M, maxvmem=325.859M
scheduling info:            queue instance "webgrid-lighttpd@tools-webgrid-
lighttpd-1201.eqiad.wmflabs" dropped because it is temporarily not available
                            queue instance "continuous@tools-exec-
1402.eqiad.wmflabs" dropped because it is disabled
                            queue instance "continuous@tools-exec-
1203.eqiad.wmflabs" dropped because it is disabled
                            queue instance "mailq@tools-exec-1402.eqiad.wmflabs"
dropped because it is disabled
                            queue instance "mailq@tools-exec-1203.eqiad.wmflabs"
dropped because it is disabled
                            queue instance "task@tools-exec-1402.eqiad.wmflabs"
dropped because it is disabled
                            queue instance "task@tools-exec-1203.eqiad.wmflabs"
dropped because it is disabled
```

Summary

- tl;dr: SGE database corrupted, scheduler is confused and stops scheduling
- solution was deleting the job the scheduler was confused about
- solving the issue took long because we chased a red herring (the testing we did earlier), and completely missed the information in the `messages` file

Conclusions

What weakness did we learn about and how can we address them?

- We are still dealing with the outfall from the Dec 30 bdb issue. We don't have enough SGE knowledge to effectively solve these issues, and learning by doing is a slow process.
- We don't have a protocol for solving an overloaded scheduler, causing us to chase that red herring for too long.

Actionables

Explicit next steps to prevent this from happening again as much as possible, with Phabricator tasks linked for every step.

- Status: 🟥 **Unresolved** Fix SGE database corruption issues (bug T122638)
- Status: 🟥 **Unresolved** Document solving 'queue overloaded due to # of recently started jobs' (bug T123411)

Category: Incident documentation

This page was last edited on 4 February 2016, at 04:48.

Privacy policy About Wikitech Disclaimers Code of Conduct Developers Statistics Cookie statement Mobile view