openQA Project - action #39743

[o3][tools] o3 unusable, often responds with 504 Gateway Time-out

2018-08-15 04:48 - okurz

Status:	Resolved	**Start date:**	2018-08-15
Priority:	Urgent	**Due date:**	
Assignee:	okurz	**% Done:**	0%
Category:	Feature requests	**Estimated time:**	0.00 hour
Target version:	Done		
Difficulty:			

Description

Also reported by users in irc://chat.freenode.net o3 is unusable, returning with 504 near-always or always. Restarting the webui did not help. Checking /var/log/openqa reveals that openQA *is* working on jobs though.

Related issues:

Related to openQA Project - action #39068: Webui killed by out of memory in o...	**Rejected**	**2018-08-01**
Related to openSUSE admin - tickets #39866: o3 unable to send any mail, "Doma...	**Closed**	**2018-08-16**
Related to openQA Project - action #39905: Job trying to download worker loca...	**Resolved**	**2018-08-17**
Related to openQA Project - action #39833: [tools] When a worker is abruptly ...	**Resolved**	**2018-08-16**
Related to openSUSE admin - tickets #39932: updating issues on status.o.o doe...	**Closed**	**2018-08-17**
Related to openQA Project - action #40001: [negotiation:error] [pid 9953] [cl...	**Rejected**	**2018-08-20**
Related to openQA Project - action #40004: worker continues to work on job wh...	**Resolved**	**2018-08-20**
Related to openQA Project - action #40013: [functional][u] Failed to upload t...	**Rejected**	**2018-08-20**
Related to openQA Project - action #39881: Worker dies during file upload (Mo...	**Resolved**	**2018-08-16**
Related to openQA Project - action #40103: [o3] openqaworker4 not able to fin...	**Resolved**	**2018-08-22**
Related to openQA Infrastructure - action #40196: [monitoring] monitor intern...	**New**	**2018-08-23**
Related to openQA Project - coordination #40199: [EPIC] Better rollback capab...	**Resolved**	**2018-08-23**
Copied to openQA Project - action #39920: [o3][tools] Correct domain config o...	**Resolved**	**2018-08-15**

History

#1 - 2018-08-15 04:53 - okurz

/var/log/openqa reveals that a worker tries to connect with a mismatching timestamp.

```
okurz@ariel:~> sudo salt '*' cmd.run 'date'
power8.openqanet.opensuse.org:
    Wed Aug 15 04:44:44 UTC 2018
openqaworker4.openqanet.opensuse.org:
    Wed Aug 15 06:44:44 CEST 2018
openqaworker1.openqanet.opensuse.org:
    Wed Aug 15 06:44:20 CEST 2018
imagetester.openqanet.opensuse.org:
    Wed Aug 15 04:44:44 UTC 2018
openqa-aarch64:
    Wed Aug 15 06:44:44 CEST 2018
```

-> time on openqaworker1 is not in sync. The files /etc/ntp.conf have different content.

progress.infra.opensuse.org recently also had a time mismatch and ntp1.i.o.o was inactive. This was fixed by tampakrap. The revelant configuration part on progress.i.o.o is

```
server ntp1.infra.opensuse.org iburst
server ntp2.infra.opensuse.org iburst
server ntp3.infra.opensuse.org iburst
restrict ntp1.infra.opensuse.org
restrict ntp2.infra.opensuse.org
restrict ntp3.infra.opensuse.org
```

so I configured this on openqaworker1 as well now and brought time in sync. Does not seem to be the reason though. webui still unresponsive.

Still, connections with timestamp mismatch are reported for the IPv4 adresses of power8, openqaworker1 and openqaworker4, something else? Is it maybe that just worker services need to be restarted now?

I stopped worker instances on power8 and openqaworker1, this seemed to have helped, https://openqa.opensuse.org is reactive again. Retriggered latest incomplete openSUSE Tumbleweed and Leap tests.

Restarted worker instances on openqaworker4 as well which seems to have caused the webui to go unresponsive again. Stopped all and restarted only openqa-worker@{1..2}, will monitor for now.

#2 - 2018-08-15 06:04 - szarate

- Related to action #39068: Webui killed by out of memory in o3 (triggered by postgresql) added

#3 - 2018-08-15 07:06 - okurz

- Status changed from In Progress to Feedback

- Priority changed from Immediate to High

webui seems to be stable for now. Tuning "prio" down to "High". I will track stability with less workers. szarate you may take over in case you want to test a new version.

#4 - 2018-08-15 09:36 - okurz

- Subject changed from o3 unusable, often responds with 504 Gateway Time-out to [o3][tools] o3 unusable, often responds with 504 Gateway Time-out

#5 - 2018-08-16 06:12 - okurz

foursixnine deployed a new version and ramped up numbers of worker instances again.

but …

[16 Aug 2018 08:11:30] foursixnine: o3 gives 504, too many workers again?

I ran

```
sudo salt openqaworker1* cmd.run systemctl stop openqa-worker@{5..16}
```

and the webui is responsive again …

#6 - 2018-08-16 06:46 - szarate

- Status changed from Feedback to In Progress

- Priority changed from High to Urgent

So we're basically spamming ourselves apparently:

[Thu Aug 16 00:52:01 2018] TCP: request_sock_TCP: Possible SYN flooding on port 80. Sending cookies. Check SNMP counters.

#7 - 2018-08-16 08:18 - szarate

- Project changed from openQA Tests to openQA Project

- Category deleted (Infrastructure)

#8 - 2018-08-16 10:00 - szarate

- Category set to 168

- Assignee changed from okurz to szarate

- Target version set to Current Sprint

It looks like it's a problem with the HA proxy, as the ammount of 504's does not seem to add up.

I have asked already in the #opensuse-admin irc channel, but in deed we still need to tune our apache config.

#9 - 2018-08-16 13:44 - okurz

- Assignee changed from szarate to okurz

Current state after discussion with szarate and with his cooperation as well as crosschecking with openSUSE heroes: It seems that any recently deployed changes can not be the source of the problems but might make the symptoms more severe. szarate has identified no code parts that are revertable to the state of Monday except for a one-line change preventing an OOM condition.

The network involves an HA proxy which responds back with "504". However 504 on the proxy means that the backend of openQA itself did not serve

content in time so the actual problem according to tampakrap is that openQA is slow on serving the request.

We are investigating how the requests of certain openQA workers are not reaching the openQA webui itself which we assume can cause the webUI itself to timeout on requests to the users.

trying my best to support with investigation, symptom handling, expectation management and status updates. I also set an according status on status.opensuse.org, currently o3 should be able to serve content statically over the webUI and intermittently work on individual testing tasks so setting according priorities on the scheduled jobs might help

ETA for the situation to be resolved: best case: Friday, 2018-09-17, worst case: Next week

#10 - 2018-08-16 14:16 - okurz

A revert had been conducted as far as possible already. As the changes since the last deployment include a database migration and other backward incompatible test setting changes it was not feasible to do a full-revert without loosing test data for more than probably one week

#11 - 2018-08-16 15:10 - okurz

Currently we are running with the packages provided by https://build.opensuse.org/package/show/home:EDiGiacinto:branches:devel:openQA/openQA which is version 4.6.1534327299.034244df from https://github.com/mudler/openQA/tree/current_sched that is reverting the latest scheduler changes.

We are currently back to having most workers (100% on x86_64, less on ports) enabled which are executing tests seemingly fine and also the webui is responsive. /var/log/apache2/error_log mentions
errors like

```
[Thu Aug 16 14:47:42.316038 2018] [proxy_http:error] [pid 31070] (70007)The timeout specified has expired: [cl
ient XXX:55228] AH01110: error reading response, referer: https://openqa.opensuse.org/tests/734133
…
```

where XXX is the IPv4 adress of the openSUSE network proxy.

What we did in /etc/apache2/vhosts.d/openqa-common.inc is to disable the routing to the "liveviewhandler" which is linked to a recent feature for the "developer mode". It certainly involves "network traffic" so we disabled that for the time being as it is a non-critical component:

```
# pass websocket server to handle live view to port 9528
#ProxyPass "/liveviewhandler/" "ws://localhost:9528/liveviewhandler/" keepalive=On
```

We plan to monitor the system unchanged over the next hours and check performance figures unless we hit more severe problems again.

#12 - 2018-08-16 15:35 - szarate

On top of that, greping the apache logs, show that only 110 requests were actual 504

```
/var/log/apache2 # xzgrep -e ' 504 ' access_* | wc -l
110
```

#13 - 2018-08-17 05:23 - okurz

Confirmed. Based on much fewer 504 in the apache access logs than what we actually observe we assume that the HA proxy is involved in answering some requests as such. Will involve heroes.

Yesterday evening I observed again more 504 again, stopped some worker instances with

```
salt 'openqaworker*' cmd.run 'systemctl stop openqa-worker@{10..16}'
```

to reduce the traffic which seeminly fixed it for the time being again. Today in the morning ramped up to previous worker capacity again. It could be with more actions and requests over the usual European working day we again see more problems.

#14 - 2018-08-17 05:25 - okurz

Asked in IRC #opensuse-admin

#15 - 2018-08-17 07:47 - RBrownSUSE

How could a HA Proxy be responsible for these results?

https://openqa.opensuse.org/tests/overview?arch=&failed_modules=&todo=1&distri=kubic&distri=opensuse&version=Tumbleweed&build=20180815&groupid=1#

#16 - 2018-08-17 07:47 - RBrownSUSE

- Priority changed from Urgent to Immediate

#17 - 2018-08-17 08:00 - okurz

RBrownSUSE wrote:

> How could a HA Proxy be responsible for these results?
>
> https://openqa.opensuse.org/tests/overview?arch=&failed_modules=&todo=1&distri=kubic&distri=opensuse&version=Tumbleweed&build=20180815&groupid=1#

Probably not. More likely an outcome of investigation work gone wrong.

#18 - 2018-08-17 09:41 - okurz

Most of the failed tests and probably including incompletes are actually caused by https://bugzilla.opensuse.org/show_bug.cgi?id=1105181 about a font change. I am trying to handle this accordingly with either marking the tests with the bug, leaving the decision if the bug should be ignored by ttm to openSUSE RMs or applying soft-fail workaround needles to get the jobs further.

Currently monitoring the web UI, the workers, apache logs as well as trying with the help of admins to see if logs on the HA proxy help us.

Crosschecked with szarate that no unexpected traffic is routed over the openSUSE HA proxy, e.g. worker traffic. We monitor a lot of TCP (not HTTP) traffic that we can not pin point yet but are not coming from workers.

Still having the developer mode route disabled in apache routes to rule out impact from that component.

#19 - 2018-08-17 10:32 - szarate

We looked at the HA proxy logs: They were not there. Apparently the remote machine where the logs were being sent to, was not logging anymore.

I'll update when I have more information on this. For the time being, logging at the HA proxy level is working again

#20 - 2018-08-17 11:59 - tampakrap

three side-issues I found at your configs:
1) postfix configuration was wrong, I fixed it with Thorsten, see https://progress.opensuse.org/issues/39866
2) ntp configuration was wrong. ntp[1-3].infra.opensuse.org are unreachable by ariel, because they belong to the heroes vlan (vlan47, 192.168.47.0/24, infra.opensuse.org) while ariel belongs to the suse-dmz vlan (vlan42, suse-dmz.opensuse.org, 192.168.254.0/24). So I changed them to ntp[1-2].opensuse.org which are two NTP servers with public IPs (anybody on the internet can use them)
3) the domain in /etc/resolv.conf and the FQDN of the machine are wrong, they should be ariel.suse-dmz.opensuse.org. I don't want to change it though as it may create issues, I'd like to get the approval of somebody responsible for the service before I fix it

#21 - 2018-08-17 12:20 - okurz

On top of what tampakrap provided what we think has happened is that the HA proxy itself is more loaded than expected due to the remote logging backend being unavailable. This causes a staggered delay in processing which can exaggerate the effect of sporadic, eventual long-running requests to openQA to time out on the side of HA proxy leaving the response connection from openQA dangling on the side of the apache process on ariel. This in turn will prevent this apache process to act upon further requests as well as the HA proxy to shortcut further requests from the outside which are then answered with 504. This can also be seen in the fact that the returned 504 page is the standard one from process on the HA proxy machine and not the one we have on the apache instance on ariel. Another potential error source adding to this are the current internal network limitations regarding to core switches causing potential packet losses. With the remote logging expected to be repaired for the moment the same symptoms can still happen but probably less likely so.
Restarting the local apache instance fixes the clogged connections in a drastic way by terminating all connections and establishing new ones. We assume that stopping/restarting worker instances has the same effect: To free connections to the local apache instance which frees it to also act upon more requests to the outside even though the two different request types come over different sections of the network (workers internal, other requests from outside over the HA proxy).

What we think should/can be improved as of now:

- Workaround to check for the responsiveness on HTTP requests on port 80 locally on ariel. As a mitigation the local apache process can be restarted in case of unresponsiveness
- By configuration for the apache instance on ariel prevent to be stuck with requests leaving no apache worker available to react on new requests
- Monitor response times
- Prevent openQA itself from throwing error conditions which are seen as 502/503/504 on the outside
- Monitor the availability of remote logging backends
- Monitor external port 443 accessibility of o3
- Full rollback capability of openQA deployments (at least older versions of RPM files, btrfs snapshots, database dumps linked to each pre-deployment state)

#22 - 2018-08-17 12:23 - okurz

- Related to tickets #39866: o3 unable to send any mail, "Domain not found" added

#23 - 2018-08-17 12:33 - okurz

- Related to action #39905: Job trying to download worker local file "aavmf-aarch64-vars.bin" into cache and fails with 404 added

#24 - 2018-08-17 13:07 - okurz

- Copied to action #39920: [o3][tools] Correct domain config on o3 added

#25 - 2018-08-17 14:13 - okurz

It seems we have identitied multiple potential conditions leading to the observed problems. The system seems to be stable again with only a slightly reduced worker capacity for the ports workers (aarch64 and ppc64le). We are monitoring the current behaviour over time having a close look at the availability of the webUI, the performance of the workers as well as individual openQA test results. No further changes are planned for the time being but some points for improvements have been identified that we plan to work in the upcoming days and weeks.

#26 - 2018-08-17 14:43 - okurz

- Related to action #39833: [tools] When a worker is abruptly killed, jobs get blocked - CACHE: Being downloaded by another worker, sleeping added

#27 - 2018-08-17 14:47 - okurz

https://openqa.opensuse.org/admin/workers confirms that many workers are stuck on #39833 now.

Manual mitigation:

```
# Kill all workers in that state as they do not react to terminate (and therefore systemctl stop also not work
ing):
for i in $(systemctl status openqa-worker@* | grep -B 2 'Being downloaded by another worker' | grep -oP '(?<=i
nstance )[0-9]*') ; do pkill -9 -f "instance $i" ; done
# Edit the database looking for the blocking asset entry marked as downloading
sqlite3 /var/lib/openqa/cache/cache.sqlite 'delete from assets where downloading="1";'
# Start the killed workers again
for i in $(systemctl status openqa-worker@* | grep -B 2 'Being downloaded by another worker' | grep -oP '(?<=i
nstance )[0-9]*') ; do echo systemctl start openqa-worker@$i; done
```

#28 - 2018-08-17 18:55 - okurz

- Related to tickets #39932: updating issues on status.o.o does not work, causes internal server error added

#29 - 2018-08-17 20:55 - okurz

- Priority changed from Immediate to Urgent

Mostly workers are fine, have not observed an outage since about 10h, continuing to monitor

#30 - 2018-08-20 05:32 - okurz

Checking the state at 2018-08-18 0500 UTC revealed no problems. However shortly afterwards 504 again. This coincides with a slightly higher load on o3 at around 0500 UTC when "openqa-review" looks for all latest build test results.

We also observed in dmesg -T:

```
[Sat Aug 18 02:31:10 2018] Out of memory: Kill process 4670 (openqa) score 401 or sacrifice child
```

For the time being we have a script in place to monitor the local apache configuration and restart the local apache2 instance when necessary:

```
cat - > /usr/local/bin/monitor_and_restart_openqa_apache <<EOF
#!/bin/sh -e
while true; do echo "-- $(date +%F-%H-%M-%S): Checking openQA availability" && passed=0; for i in {1..3} ; do
echo "try $i" && timeout 30 curl -s http://localhost | grep -q '<h1>Welcome to openQA</h1>' && passed=1 && bre
ak ; done; if [ "$passed" != "1" ] ; then echo "-- $(date +%F-%H-%M-%S): openQA unresponsive, restarting apach
e proxy" && systemctl restart apache2 ; fi ; sleep 300 ; done
EOF
cat - > /etc/cron.d/openqa-monitor-apache-restart <<EOF
-*/10    * * * *   root    pgrep -f monitor_and_restart_openqa_apache >& /dev/null || flock /tmp/monitor_and_r
estart_openqa_apache.lock /usr/local/bin/monitor_and_restart_openqa_apache >& /var/log/monitor_and_restart_ope
nqa_apache.log
EOF
cat - >> /etc/logrotate.d/openqa <<EOF

/var/log/monitor_and_restart_openqa_apache.log {
    compress
    weekly
    notifempty
    missingok
    copytruncate
    compresscmd /usr/bin/xz
    uncompresscmd /usr/bin/xzdec
}
EOF
```

Over the course of the past two days the apache server is restarted around every 90 minutes. Additionally the number of worker instances has been reduced for the time being.

With this we seem to have contained the system well enough for the weekend. We have yet to see over the work week how the system behaves.

#31 - 2018-08-20 06:58 - szarate

There's also the suspicion that we're being also hit by: https://bz.apache.org/bugzilla/show_bug.cgi?id=58280 but it's hard to prove without logs, during one occurence of that 504, it was noted that there was no connections (at least on apache logs) comming from the proxy (not even showing in logs in the HAPorxy), for about 4 minutes, until apache was restarted.

#32 - 2018-08-20 08:55 - okurz

- Priority changed from Urgent to Immediate

Additional error reports from /var/log/apache2/error_log:

```
[Sun Aug 19 00:10:48.351551 2018] [negotiation:error] [pid 9964] [client <ip_of_openqaworker1>:45678] AH00690:
 no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var
[Sun Aug 19 00:10:48.437930 2018] [negotiation:error] [pid 9953] [client <ip_of_openqaworker4>:35634] AH00690:
 no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var
[Sun Aug 19 00:10:48.438179 2018] [negotiation:error] [pid 9915] [client <ip_of_openqaworker4>:35630] AH00690:
 no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var
[Sun Aug 19 00:10:48.438802 2018] [negotiation:error] [pid 9960] [client <ip_of_openqaworker4>:35620] AH00690:
 no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var
[Sun Aug 19 00:10:48.440593 2018] [negotiation:error] [pid 9963] [client <ip_of_openqaworker4>:35628] AH00690:
 no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var
[Sun Aug 19 00:10:48.454441 2018] [negotiation:error] [pid 9947] [client <ip_of_openqaworker4>:35632] AH00690:
 no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var
```

#33 - 2018-08-20 09:12 - okurz

- Related to action #40001: [negotiation:error] [pid 9953] [client <ip_of_openqaworker4>:35634] AH00690: no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var added

#34 - 2018-08-20 09:12 - okurz

- Priority changed from Immediate to Urgent

Should stay on "Urgent".

#35 - 2018-08-20 09:16 - okurz

504 has been observed by human observers. This is expected as we only contained the problem but did not yet fix it. I decreased the monitoring checking period on port 80 availability from 300 to 120 seconds to make it less likely for human reviewers to hit this problem ;)

o3 will receive more CPU cores assigned, up from the current 4 assigned. This will involve a restart of the VM scheduled for later today.

#36 - 2018-08-20 09:20 - okurz

- Related to action #40004: worker continues to work on job which he as well as the webui considers dead added

#37 - 2018-08-20 09:28 - okurz

For yet unknown reasons 16 worker instances have been started on openqa-aarch64 at "Mon 2018-08-20 11:19:02 CEST". This is certainly too much for the machine to handle in a stable way. To prevent this for the future I masked the additional worker instances with systemctl mask openqa-worker@{3..16} but for now have not stopped these.

#38 - 2018-08-20 10:53 - szarate

For the time being, apache configuration has been changed to, but still there's a lot of latency

Also, completely unrelated, scheduler logs have been set to a separate file: /var/log/openqa_scheduler with it's logrotate rule in place too, this will take action next time the scheduler is restarted, as it is not critical.

```
--- server-tuning.conf.old      2018-08-20 10:16:17.615437859 +0000
+++ server-tuning.conf  2018-08-20 10:50:13.615530849 +0000
@@ -11,22 +11,22 @@
<IfModule prefork.c>
# number of server processes to start
# http://httpd.apache.org/docs/2.4/mod/mpm_common.html#startservers
-       StartServers         5
+       StartServers         30
# minimum number of server processes which are kept spare
```

```
# http://httpd.apache.org/docs/2.4/mod/prefork.html#minspareservers
-       MinSpareServers      20
+       MinSpareServers      30
# maximum number of server processes which are kept spare
# http://httpd.apache.org/docs/2.4/mod/prefork.html#maxspareservers
MaxSpareServers     80
# highest possible MaxClients setting for the lifetime of the Apache process.
# http://httpd.apache.org/docs/2.4/mod/mpm_common.html#serverlimit
-       ServerLimit        250
+       ServerLimit        1500
# maximum number of server processes allowed to start
# http://httpd.apache.org/docs/2.4/mod/mpm_common.html#maxclients
-       MaxClients         250
+       MaxRequestWorkers         1500
# maximum number of requests a server process serves
# http://httpd.apache.org/docs/2.4/mod/mpm_common.html#maxrequestsperchild
-       MaxRequestsPerChild  10000
+       MaxConnectionsPerChild  10000
</IfModule>
# worker MPM
@@ -72,7 +72,9 @@
# KeepAliveTimeout: Number of seconds to wait for the next request from the
# same client on the same connection.
#
+
KeepAliveTimeout 15
+Timeout 120
#
# MaxRanges: Maximum number of Ranges in a request before
```

#39 - 2018-08-20 13:14 - okurz

o3 was shutdown, number of assigned cores bumped to 10, brought up again.

another change on o3 is required to update the internal domain configuration. This might help to resolve observed problems as well or optimize existing behavior. A temporary unavailability of o3 is expected

#40 - 2018-08-20 13:34 - okurz

tampakrap wrote:

> 3) the domain in /etc/resolv.conf and the FQDN of the machine are wrong, they should be ariel.suse-dmz.opensuse.org. I don't want to change it though as it may create issues, I'd like to get the approval of somebody responsible for the service before I fix it

This change had been done some minutes ago by tampakrap.

#41 - 2018-08-20 14:21 - szarate

Another change in the saga, workers were also hit by (webui is not):
https://github.com/kraih/mojo/commit/61f6cbf22c7bf8eb4787bd1014d91ee2416c73e7

```
Aug 20 15:34:53 openqaworker1 worker[18590]: Use of uninitialized value in hash element at /usr/share/openqa/s
cript/../lib/OpenQA/Worker/Common.pm line 411.

Aug 20 15:34:53 openqaworker1 worker[18590]: Use of uninitialized value in hash element at /usr/share/openqa/s
cript/../lib/OpenQA/Worker/Common.pm line 411.

Aug 20 15:34:53 openqaworker1 worker[18590]: Use of uninitialized value $loop in hash element at /usr/lib/perl
5/vendor_perl/5.18.2/Mojo/UserAgent.pm line 208.

Aug 20 15:34:53 openqaworker1 worker[18590]: Mojo::Reactor::Poll: Timer failed: Can't call method "remove" on
an undefined value at /usr/lib/perl5/vendor_perl/5.18.2/Mojo/UserAgent.pm line 264.

Aug 20 15:34:53 openqaworker1 worker[18590]: Use of uninitialized value in hash element at /usr/share/openqa/s
cript/../lib/OpenQA/Worker/Common.pm line 411.

Aug 20 15:34:53 openqaworker1 worker[18590]: Use of uninitialized value in hash element at /usr/share/openqa/s
cript/../lib/OpenQA/Worker/Common.pm line 411.

Aug 20 15:34:53 openqaworker1 worker[18590]: Use of uninitialized value $loop in hash element at /usr/lib/perl
5/vendor_perl/5.18.2/Mojo/UserAgent.pm line 208.

Aug 20 15:34:53 openqaworker1 worker[18590]: Mojo::Reactor::Poll: Timer failed: Can't call method "remove" on
an undefined value at /usr/lib/perl5/vendor_perl/5.18.2/Mojo/UserAgent.pm line 264.
```

```
Aug 20 15:34:57 openqaworker1 worker[18590]: [error] Unable to upgrade connection for host "openqa1-opensuse"
to WebSocket: [no code]. proxy_wstunnel enabled?

Aug 20 15:34:58 openqaworker1 worker[18590]: [error] Unable to upgrade connection for host "openqa1-opensuse"
to WebSocket: [no code]. proxy_wstunnel enabled?
```

Workers have been updated now to perl-Mojolicious 7.93-2.1, while webui is still running: perl-Mojolicious 7.81-2.1

#42 - 2018-08-20 14:50 - szarate

- Related to action #40013: [functional][u] Failed to upload test logs added

#43 - 2018-08-20 14:54 - szarate

All workers are now enabled, and openqa.opensuse.org should be operating at full capacity, after tuning apache, and upgrading Mojolicious version.

#44 - 2018-08-20 15:28 - szarate

Monitoring for the time being with the following command, for already known errors

```
while true; do salt '*' cmd.run 'export INSTANCES=`ps aux | grep -v grep | grep instance | wc -l`; for i in `s
eq 1 $INSTANCES`; do systemctl status openqa-worker@$i |  grep -e Mojo:: -e proxy_ -e sleeping && (echo $(host
name) openqa-worker@$i; date systemctl is-active openqa-worker@$i); done'; sleep 20m; done;
```

#45 - 2018-08-20 20:21 - okurz

Since the last change described by szarate there had been no further outages reported by our monitoring script, no delayed response to internal port 80, no forced restart of the internal apache instance. Also only a single report of "AH00690: no acceptable variant: /usr/share/apache2/error/HTTP_BAD_GATEWAY.html.var" in /var/log/apache2/error_log .

#46 - 2018-08-21 16:29 - szarate

- Status changed from In Progress to Feedback

I think we're already on the safe side, to sum it up here are the facts:

okurz: feel free to update/change anything if I'm wrong, as you were helping a lot :)

- On 14-08 In order to fix upgrade tests, os-autoinst had to be deployed on the workers (https://github.com/os-autoinst/os-autoinst/pull/1010), but this also required a bump in perl-Mojo-IOLoop-ReadWriteProcess due to flaky exit codes (https://github.com/os-autoinst/os-autoinst/pull/1011). With the update of perl-Mojo-IOLoop-ReadWriteProcess perl-Mojolicious was also updated on the workers.
- The same day the webUI was updated with the same HEAD we're running on OSD noted already here, as we also needed changes in the worker code
- Work was being carried on by openqa.opensuse.org completely fine, until about end of day that day where first report of 504 on o3 is given.
- After restarting apache for the first time, everything went back to normal.
- On the following days, when reports started to be more often, the team started to look closer, and noticed an increased ammount of incomming connections that were not closed in time (from HAproxy), which was already replying 504
- Upon further investigation, it was found out that the HAProxy was logging to a remote machine that was not logging anymore (therefore, no logs previous to the problem are available), this might have contributed to increase the problems
- Mojo version of the webUI was checked and it still runs 7.81-2.1, and since workers were taking jobs and uploading results, no further check was done at the time
- After checking again the HAProxy and why it was suddenly not reporting any traffic until apache was restarted, it was suggested to add more cores to the machine to be able to handle the load. Also the errors reported here by Theo, but these are unrelated
- Apache configuration was compared between internal openQA instance (which runs an aggresive configuration) and openqa.opensuse.org, which had rather conservative configuration that added up to the time for apache to reply to some requests, along with https://bz.apache.org/bugzilla/show_bug.cgi?id=58280 seems to be already present for us (just not too bad)
- More cores were added and Apache configuration was tuned, which seemed to alleviate the problem but still 504 were being seen by the monitoring and "recovery" tricks
- After more digging, workers were double checked again, and they were running Mojolicious version 7.88
- Worker hosts were updated and worker instances were sequentially updated after verifying that there were no problems.
- Monitoring the amount of open connections to the server, a decrease was noted (as we were seeing 1K at times from hosts, including the proxy), to ~30 or even down to 2, depending on the worker host, or if there are requests coming form the external network.

So as a result of all of this, in the end we got some more cores on o3 (10 now) and a bit more relaxed apache server along with some nice scripts that can help checking for stuff specific for openQA :)

I set now the ticket to feedback, with the side note that the oom killer is still waking up from time to time, something that doesn't happens on our internal instance, that has the same specs (memory and cores) as o3, but has more worker instances connected to it.

#47 - 2018-08-21 18:04 - szarate

- Related to action #39881: Worker dies during file upload (Mojo bug) added

#48 - 2018-08-22 11:05 - RBrownSUSE

- Status changed from Feedback to In Progress

https://openqa.opensuse.org/tests

https://openqa.opensuse.org/admin/workers/92

Jobs are failing at an very unacceptable rate

I do not consider this issue resolved

#49 - 2018-08-22 12:33 - okurz

- Related to action #40103: [o3] openqaworker4 not able to finish any jobs added

#50 - 2018-08-22 12:35 - okurz

- Status changed from In Progress to Feedback

RBrownSUSE wrote:

> https://openqa.opensuse.org/admin/workers/92
> Jobs are failing at an very unacceptable rate
> I do not consider this issue resolved

Will be handled in #40103 which I linked as related.

For the problem of general availability of the openQA webUI I have updated the incident on status.opensuse.org to "Fixed" as we have not observed any webUI outage since about 48 hours now.

#51 - 2018-08-22 22:13 - okurz

So to come to a conclusion from my side:

Root Cause

What we accept as the root cause for the main problems is the upgrade of perl-Mojolicious on the workers causing the web UI to become unresponsive in conjunction with the degraded HA proxy. As in this case the workers were upgraded and not the web UI even though problems were observed on the web UI while workers executed test jobs just fine any rollback of actions on o3 itself could not have helped. Please keep in mind that many things added up so no single root cause should be used to judge about what would need to be done to prevent a similar problem in the future.

Lessons Learned

On top of the various individual issues that have been found during investigation – many of them reported in individual tickets linked to here – we were confirmed in the following points for improvements which we have at least partially seen already in before:

- **Complete deployment rollbacks for the whole infrastructure would be nice** (including openQA packages, database and test settings, system packages on both web UI as well as workers) but there will always be factors which are changing outside our control
- **Expectation management is important** so always give an ETA to stakeholders. Anyone looking into the problems even before understanding the whole picture should have a better understanding than outside users so provide estimates. In this particular case I personally tried to focus on this point while szarate could focus on the root cause analysis involving many more persons on a technical level
- **Responsibles for openqa.opensuse.org must exist** otherwise without explicitly adressing who is responsible there will just be confusions and frustrated stakeholders because it is still assumed that someone *is* responsible. In the case of the current ticket we found it is not exactly clear what different parties see for responsibilities especially for QA tools team. This is adressed now on a higher management level and will hopefully make things more clear.
- **Monitoring can help**, if not permanent monitoring then at least what we did is to for example monitor log files closely over the days of investigation as well as for example number of established network connections from and to the various network connections

#52 - 2018-08-23 14:41 - szarate

Thanks a lot for the writeup okurz :) I think we can close this one, or would you like to keep it around for a while?

#53 - 2018-08-23 19:27 - okurz

- Related to action #40196: [monitoring] monitor internal port 9526, port 80, external port 443 accessibility of o3 and response times added

#54 - 2018-08-23 19:33 - okurz

- Related to coordination #40199: [EPIC] Better rollback capabilities of (worker) deployments added

#55 - 2018-08-23 19:35 - okurz

- Status changed from Feedback to Resolved

szarate wrote:

> Thanks a lot for the writeup okurz :) I think we can close this one, or would you like to keep it around for a while?

I wanted to keep it open until we have a bit more specific outcomes written down in tickets. I did that now :)

Ticket resolved.

#56 - 2018-09-05 13:48 - szarate

O3 stability issues and downtime post-mortem summary

Even though it has been already explainded [a][b], there's the request to express in a more higher level language

Root cause analysis

Root cause of the downtime of openqa.opensuse.org was the upgrade of Mojolicious to version 7.88, an openQA dependency that was affected by a bug [0], deployed on the workers causing the web UI to become unresponsive [1] and this was only checked few days after the whole problem was there [a].

Since the webUI was not having the affected version (It was running 7.81) and the workers were able to finish jobs, but still were affected by a bug that let open connections to the webUI, eventually starving the apache webserver, causing it to not be able to process any more requests; so acting as a de facto, self-induced DoS.

Solution

Workers have been updated now to perl-Mojolicious 7.93-2.1, while the web UI is still running: perl-Mojolicious 7.81-2.1. and there is no need (yet) to upgrade it.

Reasoning for deploying

On 14-08, In order to timely fix upgrade tests from SLE11 to Leap, as a result of a big rework of openQA's QEMU interface, os-autoinst and the worker's codebase had to be released and upgraded on the workers hosts [7], which in turn required a bump in perl-Mojo-IOLoop-ReadWriteProcess [8] and consequentially also Mojolicious was upgraded.

Consequences after deploying

In this time frame the version of the library which openQA depends on, made it's way to the workers and went unnoticed.

OpenQA currently does not make visible the versions of the dependencies which is relying on (and should not), instead, a proper configuration management system is prefered.

The configuration management which is in place on openqa.opensuse.org is not actively used, and the tool used for worker's administration does not have the necessary elements to fully address to this kind of problems (as in: There are no configuration/salt recipes).

In the moment, the openqa.opensuse.org webUI instance got updated to match the same library version that openqa.suse.de was running after a revert to previous scheduler changes.

openqa.opensuse.org never had changes to the scheduler, as a measure to keep o3 safe from any serious fallout that could affect testing of products.

Too many elements came into play at the same time and made it a bit more difficult to identify where the actual problem was.

- Degraded HA proxy [2][3]
- Outdated configuration for apache [4][5]
- Possible bugs in mod_proxy from apache [6]

Two reasons led to the deployment:

- Request to deploy changes that would fix the upgrade tests [7]
- Bugs that were already present in os-autoinst and worker code that had fixes already needed to be deployed on the workers.

Which resulted in this story.

Lessons Learned

- o3 needs an admin, not just people that look at it from time to time.
- Rolling out of new features should be more verbose, so that users are aware of what to expect.
- Better monitoring is needed.
- Better configuration management is needed overall.

- Awareness for potential problems in new Mojolicious versions (and other dependencies) should be improved. It can have a big impact and in this case it took quite a while before this was even considered.
- Salt recipes need to be prepared for o3.
- OpenQA's own test matrix needs a lot of improvement.
- Acceptance test cycles for openQA should include the full development and production stack (as in, similar scenarios, just in a very smaller scale).
- There are many underlying errors in the web UI and worker code as result of heavy rework that need to be addressed. Otherwise more fallout will come for the next big rework.

[a]https://progress.opensuse.org/issues/39743#note-46
[b]https://progress.opensuse.org/issues/39743#note-51
[0] https://github.com/kraih/mojo/commit/61f6cbf22c7bf8eb4787bd1014d91ee2416c73e7
[1]https://progress.opensuse.org/issues/39743#note-41
[2]https://progress.opensuse.org/issues/39743#note-19
[3]https://progress.opensuse.org/issues/39743#note-8
[4]https://progress.opensuse.org/issues/39743#note-38
[5]https://progress.opensuse.org/issues/39743#note-6
[6]https://progress.opensuse.org/issues/39743#note-31
[7]https://github.com/os-autoinst/os-autoinst/pull/1010
[8]https://github.com/os-autoinst/os-autoinst/pull/1011

#57 - 2018-09-10 17:12 - RBrownSUSE

Thanks for the analysis

I do have one question from that though.

If there was known changes made on 14-08, why were those changes not reverted on the workers when these problems first became apparent?

Your analysis makes it sound like the changes were only made to workers and the scheduler was untouched - as workers do not contain any stateful data, wouldn't one of the natural troubleshooting steps even in the muddied waters of the other issues mentioned have still been to revert the changes made to the workers on 14-08?

Why was that option overlooked?

#58 - 2018-09-12 05:28 - sebchlad

There are actually 2 things which makes me worry about the situation:
#1 apparent lack of responsibilities for o3.
#2 error-prone workflow which is getting more and more crucial (with more and more parties relying on it)

#1 as for this situation - as soon as it was brought up to my attention by RBrownSUSE - I assured we will work on finding resolution to the problem at hand. However bigger problem remains - clear responsibility for o3. I asked for support in clarifying this, which seems to be really happening, as some executives are really working now towards having this responsibility clarified. I also updated other teams inside Suse trying to make them aware of how crucial o3 and OSD are, so perhaps this can render some better structured and organized support for both openQA 'instances'. This is mainly Security and Maintenance Department. I hope to get a clear understanding within the company on how we would like to make sure openQA instances are up and running.

One remaining flow however is the communication. I understand certain discussions happened in daily 10 o'clock calls. This is good and I have no problem with who is talking to whom. However that meeting has very different definitions depending on the person you talk to, so some people do not see that meeting as a valid place to talk about openQA instances. Moreover there is apparent lack of connection between that call and so-called Tools team, so at times information might get lost.
While I was asking if I could join the call, I was rejected.

Once we clarify the responsibilities we must ensure good understanding around this, so any problem around openQA should be directed to the people responsible for fixing it/team responsible for o3 and osd.

#2 Another issue remains however. Even we have responsibilities clarified and communication improved, we will be still exposed to problems whenever we refactor some major parts of openQA.
So currently we are in the situation where whatever bigger change in the os-autoinst is needed for the purpose of some extended/additional testing, we risk major disruptions in our more and more important workflow.
I'm afraid this situation is possibly very damaging for overall openQA adaption within and outside of Suse, but might be easily overlooked in the daily routines focused on solving only apparent problems as they appear.

I will work towards ensuring some major stakeholders and sponsors understand this ramification but is beyond my responsibilities to find and make effective a resolution here.

#59 - 2018-09-14 12:18 - coolo

- Target version changed from Current Sprint to Done