openQA Infrastructure - action #96554

coordination # 96447 (Blocked): [epic] Failed systemd services and job age alerts

Mitigate on-going disk I/O alerts size:M

2021-08-04 09:31 - cdywan

Status:	Resolved	**Start date:**	2021-08-04
Priority:	Urgent	**Due date:**	
Assignee:	mkittler	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	Ready		

Description

Observation

- Alerts Disk I/O time for /dev/vdd (/results)

Suggestion

- Bump our thresholds
- Monitor the systemd journal (while the alert is running)
- Watch htop activity
- Observe team/squad channels

Related issues:

Related to openQA Project - action #96557: jobs run into MAX_SETUP_TIME, one ...	**Resolved**	**2021-08-04**	**2021-08-19**
Related to openQA Infrastructure - action #96807: Web UI is slow and Apache R...	**Resolved**	**2021-08-12**	**2021-10-01**
Copied to openQA Infrastructure - action #97409: Re-use existing filesystems ...	**New**		
Copied to openQA Infrastructure - action #97412: Reduce I/O load on OSD by us...	**New**		

History

#1 - 2021-08-04 09:34 - cdywan

- Status changed from Workable to In Progress

- Assignee set to mkittler

#2 - 2021-08-04 11:36 - mkittler

It is more likely that the high number of incompletes we recently saw is caused by the deployment http://mailman.suse.de/mlarch/SuSE/openqa/2021/openqa.2021.08/msg00001.html because the alert never lasted very long.

It occurred just now (2021-08-04T13:01 to 2021-08-04T13:04) again (for /dev/vdc) for 3 minutes. I wasn't fast enough to check htop but the journal doesn't contain anything helpful. The dmesg log also doesn't have anything special in it:

```
2021-08-04T12:58:47,908296+02:00 RDX: 0000000000000000 RSI: 0000000000000000 RDI: 0000000040000011
2021-08-04T12:58:47,910678+02:00 RBP: 00007ffebec14690 R08: 0000000000000003 R09: 00007ffebec14570
2021-08-04T12:58:47,913060+02:00 R10: 0000000000000000 R11: 0000000000000246 R12: 0000000000000001
2021-08-04T12:58:47,915242+02:00 R13: 0000000000000001 R14: 0000000000000000 R15: 0000000000000000
2021-08-04T13:04:48,554468+02:00 warn_alloc: 1 callbacks suppressed
2021-08-04T13:04:48,554475+02:00 vsftpd: page allocation failure: order:5, mode:0x40dc0(GFP_KERNEL|__GFP_COMP|
__GFP_ZERO), nodemask=(null),cpuset=/,mems_allowed=0
2021-08-04T13:04:48,561233+02:00 CPU: 3 PID: 4197 Comm: vsftpd Not tainted 5.3.18-lp152.84-default #1 openSUSE
 Leap 15.2
2021-08-04T13:04:48,563913+02:00 Hardware name: QEMU Standard PC (i440FX + PIIX, 1996), BIOS Bochs 01/01/2011
```

Btw, when checking htop while the alert is not active the I/O traffic is still constantly high. For instance seeing at least one Apache worker with ~60 M/s seems normal. Sometimes there's an occasional peak where an Apache worker shows ~3 G/s for an Apache worker but it is usually not for a very long time. The top openQA prefork workers usually take around 1 to 4 M/s.

#3 - 2021-08-05 04:10 - openqa_review

- Due date set to 2021-08-19

Setting due date based on mean cycle time of SUSE QE Tools

#4 - 2021-08-05 08:48 - cdywan

Seems to be getting worse. Disk I/O time for /dev/vdc (/assets) was alerting for 11 minutes again.

#5 - 2021-08-05 14:54 - mkittler

By the way, we're measuring the read_time/write_time here (and read_time seems to be the problem considering the alert history). From the documentation (https://github.com/influxdata/telegraf/tree/master/plugins/inputs/diskio#read_time--write_time):

> These values count the number of milliseconds that I/O requests have waited on this block device. If there are multiple I/O requests waiting, these values will increase at a rate greater than 1000/second; for example, if 60 read requests wait for an average of 30 ms, the read_time field will increase by 60*30 = 1800.

If /dev/vdc is too slow then asset downloads might time out and that's exactly what we see in issue #96557. This only raises the question why only a few workers were affected. Maybe it really takes a high number of worker slots and jobs with big assets for this to have an impact (in terms of jobs exceeding the setup timeout)? (At least the table #96557#note-19 would allow this conclusion.)

#6 - 2021-08-05 15:01 - mkittler

- Related to action #96557: jobs run into MAX_SETUP_TIME, one hour between 'Downloading' and 'Download processed' and no useful output in between auto_review:"timeout: setup exceeded MAX_SETUP_TIME":retry added

#7 - 2021-08-06 11:15 - mkittler

The comment #96557#note-24 contains some investigation regarding /dev/vdc. At least currently the performance doesn't look bad.

#8 - 2021-08-10 09:10 - cdywan

Disk I/O alerts 10.09-11.02, 9.45-10.01, 9.24-?, 8.43-9.03, 8.05-8.32, 7.44-7.46, 6.03-6.25, 5.43-5.45, 5.33-5.35, 5.13-5.17 (all CEST)

See also https://stats.openqa-monitor.qa.suse.de/d/WebuiDb/webui-summary?tab=alert&viewPanel=47&orgId=1

There was also CPU Load 6.09-6.29, 5.44-5.46, 5.36-5.37, 5.20-5.22 (all CEST) correlating □□□

#9 - 2021-08-10 09:55 - mkittler

This time the alert is still active. It is about the read performance of the device we're storing assets on. It doesn't have any impact *so far*.

#10 - 2021-08-10 10:53 - mkittler

> There was also CPU Load … correlating

Now the CPU load alert was even shortly firing.

Maybe the number of incompletes we're currently seeing (
https://monitor.qa.suse.de/d/nRDab3Jiz/openqa-jobs-test?viewPanel=14&orgId=1&from=1628546400000&to=now) are related (#96710).

#11 - 2021-08-11 10:00 - mkittler

The SR to increase the threshold for the assets disk has been merged. Let's see whether this was sufficient.

I've been creating another SR to increase the threshold for the results disk as we've seen an alert about it yesterday:
https://gitlab.suse.de/openqa/salt-states-openqa/-/merge_requests/545
It looks like the threshold there was actually accidentally quite low.

#12 - 2021-08-11 14:34 - mkittler

- Status changed from In Progress to Feedback

The 2nd SR has been merged as well.

#13 - 2021-08-12 09:17 - cdywan

https://github.com/os-autoinst/scripts/pull/99 which adds grep timeouts was merged

#14 - 2021-08-16 14:26 - mkittler

The alert happened again two times for the assets read I/O time last weekend. It could just be a symptom of having an unusual amount of scheduled

jobs due to other issues so I'm currently hesitant to bump the threshold again.

#15 - 2021-08-16 15:30 - okurz

please pause the alerts until we found mitigations.
I monitored data on monitor.qa.suse.de even though there was right now no alert triggering. "disk I/O time for /space-slow" was at 40s for nearly 5 minutes. disk i/o requests and bytes spiked at the same time so it was not like the storage stalled.
It coincides with an increase in average ping time of multiple workers and an increase in the number of running jobs so very likely just many jobs starting and requesting assets, nothing wrong with that. I am sure we could spread the load a bit if we manage to have less assets that need to be downloaded. One thing I could think of is to try to reuse the cache after reboot of workers instead of recreating the filesystem everytime and another idea is to increase the cache size, e.g. use all available space instead of the artifical limits of the cache directory.

#16 - 2021-08-17 08:57 - cdywan

- Related to action #96807: Web UI is slow and Apache Response Time alert got triggered added

#17 - 2021-08-18 12:43 - mkittler

Ok, I've been pausing the alert for /dev/vdd. I just kept in on because it was nice to see how the situation develops without having to check manually all the time.

> One thing I could think of is to try to reuse the cache after reboot of workers instead of recreating the filesystem everytime

That would make sense and would certainly help in cases when multiple workers are rebooted at the same time. Maybe that means we should migrate away from ext2, though.

> another idea is to increase the cache size, e.g. use all available space instead of the artifical limits of the cache directory

A good idea. The worker cache could just ensure a certain percentage of free disk space in the file system.

#18 - 2021-08-23 15:40 - mkittler

With https://gitlab.suse.de/openqa/salt-states-openqa/-/merge_requests/545 and https://gitlab.suse.de/openqa/salt-states-openqa/-/merge_requests/542 I made improvements for vdc and vdd so I'm resuming the alert for vdd again. (All other alerts aren't paused anyways.) I've also been creating a similar fix for vde: https://gitlab.suse.de/openqa/salt-states-openqa/-/merge_requests/562

#19 - 2021-08-23 15:44 - okurz

- Copied to action #97409: Re-use existing filesystems on workers after reboot if possible to prevent full worker asset cache re-syncing added

#20 - 2021-08-23 15:51 - okurz

- Copied to action #97412: Reduce I/O load on OSD by using more cache size on workers with using free disk space when available instead of hardcoded space added

#21 - 2021-08-23 16:17 - mkittler

- Status changed from Feedback to Resolved

https://gitlab.suse.de/openqa/salt-states-openqa/-/merge_requests/562 has been merged and deployed. okurz created tickets for further improvements. So I'm considering this ticket resolved.

#22 - 2021-08-27 09:36 - okurz

- Due date changed from 2021-08-19 to 2021-09-10

- Status changed from Resolved to Feedback

new alert from yesterday evening:
https://stats.openqa-monitor.qa.suse.de/d/WebuiDb/webui-summary?tab=alert&viewPanel=57&orgId=1&from=1630005418432&to=1630024386708

#23 - 2021-08-27 12:11 - mkittler

Not sure what to do except for increasing the threshold: https://gitlab.suse.de/openqa/salt-states-openqa/-/merge_requests/566

#24 - 2021-08-27 15:20 - okurz

- Due date deleted (2021-09-10)

- Status changed from Feedback to Resolved

MR merged. Currently there is no alert. Same as in before because we already have separate improvement tickets we can resolve.

#25 - 2021-10-19 09:37 - tinita

- Status changed from Resolved to Workable

- Assignee deleted (mkittler)

The alert was triggered again this morning:
https://stats.openqa-monitor.qa.suse.de/d/WebuiDb/webui-summary?tab=alert&viewPanel=48&orgId=1&from=1634589539385&to=1634626591378

#26 - 2021-10-21 08:42 - cdywan

There was an MR bumping the thresholds from 10 to 20s - maybe that already covers this?
https://gitlab.suse.de/openqa/salt-states-openqa/-/merge_requests/608/diffs

#27 - 2021-10-22 04:28 - okurz

Careful. This ticket was one of the things that we identified went wrong regarding qcow compress. So bumping alert thresholds without a good explanation why we think this is fine is what we decided we want to avoid

#28 - 2021-10-22 09:59 - okurz

- Assignee set to mkittler

mkittler as you requested, assigning to you again to crosscheck.

#29 - 2021-10-22 11:15 - mkittler

- Status changed from Workable to Resolved

Yes, the mentioned SR should fix this. It got only merged after the alert happened.