openQA Infrastructure - action #19238

setup pool devices+mounts+folders with salt(was: ext2 on workers busted)

2017-05-19 04:48 - coolo

Status:	Resolved	**Start date:**	2017-05-19
Priority:	Normal	**Due date:**	
Assignee:	okurz	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	Ready		

Description

https://openqa.suse.de/tests/943176/file/autoinst-log.txt shows a broken job caused by inconsistent pool directory on worker5.

As we now set all file systems to not check on boot, we need a resetup of the pool directories on boot.

Related issues:

Related to openQA Infrastructure - action #49694: openqaworker7 lost one NVMe	**Resolved**	**2019-03-26**	
Related to openQA Project - action #46742: test incompletes trying to revert ...	**Resolved**	**2019-01-28**	**2020-02-18**

History

#1 - 2017-05-19 05:26 - coolo

This needs to be salted and systemded

```
#! /bin/sh

set -e

function _umount {
  if grep $1 /proc/mounts; then
     umount $1
  fi
}

POOL2="1 2 3 4 5 6 7 8 9 10 11 12"
POOL1="13 14 15 16 17 18 19 20"
_umount /var/lib/openqa/cache

for i in $POOL2; do
 _umount /var/lib/openqa/pool/$i
done

_umount /var/lib/openqa/pool2
_umount /var/lib/openqa/pool

mkfs.ext2 -F /dev/nvme0n1p1
mkfs.ext2 -F /dev/nvme1n1p1

mount /var/lib/openqa/pool
mount /var/lib/openqa/pool2

for i in $POOL2; do
  mkdir /var/lib/openqa/pool/$i
  mkdir /var/lib/openqa/pool2/$i
  chown _openqa-worker /var/lib/openqa/pool2/$i
  mount -o bind /var/lib/openqa/pool2/$i /var/lib/openqa/pool/$i
done

for i in $POOL1; do
  mkdir /var/lib/openqa/pool/$i
  chown _openqa-worker /var/lib/openqa/pool/$i
done

mkdir /var/lib/openqa/pool/cache
chown _openqa-worker /var/lib/openqa/pool/cache

mount -o bind /var/lib/openqa/pool/cache /var/lib/openqa/cache
```

#2 - 2017-05-26 20:23 - okurz

- Category set to 168

#3 - 2017-11-21 14:36 - coolo

- Subject changed from [tools] ext2 on workers busted to ext2 on workers busted

- Target version set to Ready

#4 - 2018-11-23 14:38 - mkittler

- Project changed from openQA Project to openQA Infrastructure

- Category deleted (168)

Seems to be an infra issue.

#5 - 2018-11-27 07:04 - nicksinger

- Status changed from New to Workable

#6 - 2019-09-24 19:07 - okurz

- Subject changed from ext2 on workers busted to setup pool devices+mounts+folders with salt(was: ext2 on workers busted)

By now we have the NVMe devices on the three arm workers setup with salt, see https://gitlab.suse.de/openqa/salt-states-openqa/tree/master/openqa/nvme_store . The caveat I saw there is that the file system is recreated on every reboot – as actually suggested here – but with the need to sync again especially the big test and needles repos the overall setup process takes rather long. I think we are able to find a way to re-use the existing partition and data with proper consistency checks and only repair what is necessary. Can you describe what was the original problem? Also, why ext2? I know, there is no journal but is it still the best approach?

EDIT: I tried on openqaworker10, mkfs.ext2 on an NVMe partition took 25s, mkfs.ext4 took 1s. As we are reformating on the arm workers on every reboot one more reason to use ext4.

http://www.ilsistemista.net/index.php/virtualization/47-zfs-btrfs-xfs-ext4-and-lvm-with-kvm-a-storage-performance-comparison.html has same info. https://www.phoronix.com/scan.php?page=article&item=linux-50-filesystems&num=2 indicates XFS might be good for us (by now) to run for the pool dir. Following https://wiki.archlinux.org/index.php/ext4#Improving_performance or https://www.thegeekdiary.com/what-are-the-mount-options-to-improve-ext4-filesystem-performance-in-linux/ I will try to use optimized settings for openqaworker10, see #32605 as well. Interesting enough, I could not easily proof that ext4 w/o journal is any better than ext2:

```
openqaworker10:/srv # time mkfs.ext2 -F /dev/nvme0n1p1
…

real    0m24.034s
openqaworker10:/srv # mount -o defaults /dev/nvme0n1p1 /var/lib/openqa/pool/
openqaworker10:/srv # mount | grep pool
/dev/nvme0n1p1 on /var/lib/openqa/pool type ext2 (rw,relatime,block_validity,barrier,user_xattr,acl)
openqaworker10:/srv # /tmp/avgtime -q -d -r 5 -h dd bs=4M count=1000 if=/dev/zero of=/var/lib/openqa/pool/test
.img
Avg time        : 7013.06
Std dev.        : 225.566
Minimum         : 6786.32
Maximum         : 7442.27
openqaworker10:/srv # /tmp/avgtime -d -r 5 -h dd bs=4M count=1000 if=/dev/zero of=/var/lib/openqa/pool/test.im
g
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 6.04282 s, 694 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 6.25296 s, 671 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 6.04532 s, 694 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 6.27314 s, 669 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 6.40667 s, 655 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 6.56258 s, 639 MB/s
…

Avg time        : 7090.44
Std dev.        : 171.836
Minimum         : 6789.99
Maximum         : 7304.62
openqaworker10:/srv # umount /var/lib/openqa/pool
openqaworker10:/srv # time mkfs.ext4 -O ^has_journal -F /dev/nvme0n1p1
…
real    0m0.757s
openqaworker10:/srv # mount -o defaults,noatime,barrier=0 /dev/nvme0n1p1 /var/lib/openqa/pool/
openqaworker10:/srv # /tmp/avgtime -d -r 5 -h dd bs=4M count=1000 if=/dev/zero of=/var/lib/openqa/pool/test.im
g
```

```
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 4.23314 s, 991 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 7.79238 s, 538 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 7.6331 s, 549 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 7.95202 s, 527 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 7.57801 s, 553 MB/s
4194304000 bytes (4.2 GB, 3.9 GiB) copied, 7.87948 s, 532 MB/s

Avg time       : 8476.28
Std dev.       : 163.676
Minimum        : 8273.7
Maximum        : 8641.14
openqaworker10:/srv #
```

I also conducted a test ext4+journal which was worse. However this is all still on openSUSE Leap 42.3 with Linux 4.4.159. I should redo this after upgrade (or reinstall).

EDIT: I checked all our current production workers and we have two nvme's on some, single nvme on openqaworker{9,10,13} and arm{1,2,3}:

```
$ sudo salt --no-color '*' cmd.run 'ls /dev/nvme?'
QA-Power8-4-kvm.qa.suse.de:
    ls: cannot access '/dev/nvme?': No such file or directory
QA-Power8-5-kvm.qa.suse.de:
    ls: cannot access '/dev/nvme?': No such file or directory
powerqaworker-qam-1:
    ls: cannot access '/dev/nvme?': No such file or directory
malbec.arch.suse.de:
    ls: cannot access '/dev/nvme?': No such file or directory
openqaworker2.suse.de:
    /dev/nvme0
    /dev/nvme1
openqaworker9.suse.de:
    /dev/nvme0
openqaworker8.suse.de:
    /dev/nvme0
openqaworker5.suse.de:
    /dev/nvme0
    /dev/nvme1
openqaworker3.suse.de:
    /dev/nvme0
    /dev/nvme1
openqaworker7.suse.de:
    /dev/nvme0
    /dev/nvme1
openqaworker6.suse.de:
    /dev/nvme0
    /dev/nvme1
grenache-1.qa.suse.de:
    ls: cannot access '/dev/nvme?': No such file or directory
openqa-monitor.qa.suse.de:
    ls: cannot access '/dev/nvme?': No such file or directory
openqa.suse.de:
    ls: cannot access '/dev/nvme?': No such file or directory
openqaworker10.suse.de:
    /dev/nvme0
openqaworker-arm-1.suse.de:
    /dev/nvme0
openqaworker13.suse.de:
    /dev/nvme0
openqaworker-arm-3.suse.de:
    /dev/nvme0
openqaworker-arm-2.suse.de:
    /dev/nvme0
ERROR: Minions returned with non-zero exit code
```

so we can either make the script dynamic to use 0-2 (or more) NVMe devices or we rely on the specific workers setup statically. Preferences or ideas?

#7 - 2019-10-18 06:30 - okurz

- Related to action #49694: openqaworker7 lost one NVMe added

#8 - 2020-01-14 12:53 - okurz

- Related to action #46742: test incompletes trying to revert to qemu snapshot auto_review:"Could not open backing file: Could not open ..qcow.*No such file or directory", likely premature deletion of files from cache added*

#9 - 2020-01-14 13:34 - okurz

Given that with #46742 we try to use efficient hard-links from cache to the pools I suggest to stripe all available NVMe's together, e.g. create a RAID0 of all NVMe's, mount as /var/lib/openqa as the latter on workers has commonly only three dirs, pool, cache and share which should be a mountpoint for NFS anyway. Based on http://www.fibrevillage.com/storage/429-performance-comparison-of-mdadm-raid0-and-lvm-striped-mapping I would choose mdadm RAID0 over striping LVM. We could configure volumes with https://docs.saltstack.com/en/latest/ref/states/all/salt.states.mdadm_raid.html but I doubt we can easily dynamically configure the number of devices based on what is present on the specific workers easily. I suggest to rework what we have in https://gitlab.suse.de/openqa/salt-states-openqa/tree/master/openqa/nvme_store which we currently only use for three ARM workers and apply that the same for all workers.

EDIT: I have an idea, maybe we *can* define it dynamically within salt based on grains, e.g.

```
sudo salt --no-color '*' grains.item SSDs
QA-Power8-4-kvm.qa.suse.de:
    ----------
    SSDs:
…
openqaworker10.suse.de:
    ----------
    SSDs:
        - nvme0n1
…
openqaworker-arm-2.suse.de:
    ----------
    SSDs:
        - sdb
        - sda
        - nvme0n1
```

see https://gitlab.suse.de/openqa/salt-states-openqa/merge_requests/250 for an idea how to do that.

#10 - 2020-01-31 22:24 - okurz

- Status changed from Workable to Feedback

- Assignee set to okurz

https://gitlab.suse.de/openqa/salt-states-openqa/merge_requests/250 merged. I am experimenting with openqaworker11 and openqaworker13 for reinstalls.

```
ipmitool -I lanplus -H openqaworker11-ipmi.suse.de -U $user -P $pass chassis bootdev pxe
ipmitool -I lanplus -H openqaworker11-ipmi.suse.de -U $user -P $pass power reset
sleep 3
ipmitool -I lanplus -H openqaworker11-ipmi.suse.de -U $user -P $pass sol activate
```

in PXE menu selecting openSUSE Leap 15.1, pressing "tab" for options, adding parameter autoyast=http://w3.nue.suse.com/~okurz/ay-openqa-worker.xml but so far I fail to see something useful on the screen so far, just after loading the kernel and initrd after some time, about a minute "àààààààààààààààüàààààüàààààff".

And now I'm lost.

EDIT: Trying with additional parameters console=ttyS1,115200 so the complete boot line in PXE:

```
/find/openSUSE-Leap-15.1-x86_64-DVD1/boot/x86_64/loader/linux initrd=/find/ope
nSUSE-Leap-15.1-x86_64-DVD1/boot/x86_64/loader/initrd install=http://dist.suse.d
e/netboot/find/openSUSE-Leap-15.1-x86_64-DVD1 splash=silent minmemory=128 ramdis
k_size=73728 vga=normal console=tty0 console=ttyS0,115200 sysrq_always_enabled l
inemode=1 panic=100 ignore_loglevel unknown_nmi_panic insecure=1 console=ttyS1,1
15200
```

EDIT: No luck

#11 - 2020-02-04 14:35 - nicksinger

okurz wrote:

> EDIT: Trying with additional parameters console=ttyS1,115200 so the complete boot line in PXE:
>
> ```
> /find/openSUSE-Leap-15.1-x86_64-DVD1/boot/x86_64/loader/linux initrd=/find/ope
> nSUSE-Leap-15.1-x86_64-DVD1/boot/x86_64/loader/initrd install=http://dist.suse.d
> e/netboot/find/openSUSE-Leap-15.1-x86_64-DVD1 splash=silent minmemory=128 ramdis
> k_size=73728 vga=normal console=tty0 console=ttyS0,115200 sysrq_always_enabled l
> inemode=1 panic=100 ignore_loglevel unknown_nmi_panic insecure=1 console=ttyS1,1
> 15200
> ```

EDIT: No luck

You where close. I realized it renders the black text on a black background. Changing the command-line to this fixes this and your AY profile starts up:

```
/find/openSUSE-Leap-15.1-x86_64-DVD1/boot/x86_64/loader/linux initrd=/find/openSUSE-Leap-15.1-x86_64-DVD1/boot
/x86_64/loader/initrd install=http://dist.suse.de/netboot/find/openSUSE-Leap-15.1-x86_64-DVD1 splash=silent co
nsole=ttyS1,115200 autoyast=http://w3.suse.de/~okurz/ay-openqa-worker.xml
```

I've really no clue why any of the existing vga, minmemory or ramdisk_size should cause this but just removing it worked fine. Be aware that the normal backspace does not work in PXE over SOL therefore you have to use ctrl+h instead.
Unfortunately your profile also fails quite early with:

```
salt-minion: The package is not available.
```

#12 - 2020-02-06 20:48 - okurz

trying now with

```
/find/openSUSE-Leap-15.1-x86_64-DVD1/boot/x86_64/loader/linux initrd=/find/openSUSE-Leap-15.1-x86_64-DVD1/boot
/x86_64/loader/initrd install=http://download.opensuse.org/distribution/leap/15.1/repo/oss/ autoyast=http://w3
.suse.de/~okurz/ay-openqa-worker.xml console=ttyS1,115200
```

but that somehow brought me into an installation summary screen, not what looks like autoyast, hm.

Anyway, I can also continue with manual migration which I need to do for o3 anyway:

- openqaworker4.o.o: migrated as one NVMe is broken and I experimented with the machine anyway, verified with jobs
- aarch64.o.o: migrated, openqa-clone-job --within-instance https://openqa.opensuse.org 1162608 _GROUP=0 BUILD=X TEST=okurz_poo19238 -> Created job #1165797: opensuse-Tumbleweed-DVD-aarch64-Build20200201-mediacheck@aarch64 -> https://openqa.opensuse.org/t1165797 -> passed
- power8.o.o: migrated, openqa-clone-job --within-instance https://openqa.opensuse.org 1164156 _GROUP=0 BUILD=X TEST=okurz_poo19238 -> Created job #1165798: opensuse-Tumbleweed-DVD-ppc64le-Build20200203-mediacheck@ppc64le -> https://openqa.opensuse.org/t1165798 -> passed
- openqaworker1.o.o: migrated, verified

for OSD I checked first again where needed (and where potentially not so many jobs running right now) salt -l error --no-color '*' cmd.run 'lsblk | grep -q nvme && ps auxf | grep -c isotovideo'

but then I realized that e.g. openqaworker2 has already /dev/md0 and /dev/md1. Actually the same we have on the o3 workers but I think we should use the same name on all workers regardless of existance of md0 or md1 so I created
https://gitlab.suse.de/openqa/salt-states-openqa/merge_requests/268
proposing /dev/md/openqa

Done:

- openqaworker-arm-1.suse.de: migrated, verified
- openqaworker-arm-2.suse.de: migrated, openqa-clone-job --skip-chained-deps --within-instance https://openqa.suse.de 3867058 _GROUP=0 BUILD=X TEST=mediacheck_okurz_poo19238 WORKER_CLASS=openqaworker-arm-2 -> Created job #3872316: sle-15-SP2-Online-aarch64-Build136.2-mediacheck@aarch64 -> https://openqa.suse.de/t3872316 -> passed
- openqaworker-arm-3.suse.de: migrated, verified
- openqaworker2.suse.de: migrated, verified
- openqaworker3.suse.de: migrated, verified
- openqaworker5.suse.de: migrated, verified
- openqaworker10.suse.de: migrated, verified
- openqaworker6.suse.de: migrated, verified
- openqaworker7.suse.de: migrated, verified
- openqaworker8.suse.de: same as for 9, needs manual handling or adaptions, verified
- openqaworker9.suse.de: has *only* NVMe, no other disks or SSDs, migrated manually, verified

created https://gitlab.suse.de/openqa/salt-states-openqa/merge_requests/269 to handle a single NVMe setup as for openqaworker8+9 automatically in the future.

#13 - 2020-02-21 13:15 - okurz

- Status changed from Feedback to In Progress

nicksinger ignored me in MR so merged myself ;)

I applied the state explicitly with salt -l error --no-color -C 'G@roles:worker' --state-output=changes state.apply openqa.nvme_store and the diff of changes looks fine.

https://gitlab.suse.de/openqa/salt-states-openqa/merge_requests/275 to apply by default.

#14 - 2020-02-21 14:06 - okurz

- Status changed from In Progress to Resolved

https://gitlab.suse.de/openqa/salt-states-openqa/merge_requests/275 merged and successfully applied