Ceph - Bug #9927

RHEL: selinux-policy-targeted rpm update triggers slow requests

10/29/2014 03:35 AM - Dan van der Ster

Status:	Can't reproduce	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:			
Category:			
Target version:			
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

We observe slow requests while updating a server to RHEL6.6. The upgrade includes selinux-policy-targeted, which runs this during the update:

```
/sbin/restorecon -i -f - -R -p -e /sys -e /proc -e /dev -e /mnt -e /var/tmp -e /home -e /tmp -e /d
ev
```

restorecon is scanning every single file on the OSDs, e.g. from strace:

```
lstat("rbd\\udata.1b9d8d42be29bd3.000000000003e430__head_052DF076__4", {st_mode=S_IFREG|0644, st_s
ize=4194304, ...}) = 0
lstat("rbd\\udata.1c2064583a15ea.00000000000a8553__head_4B4DF076__4", {st_mode=S_IFREG|0644, st_si
ze=4194304, ...}) = 0
lstat("rbd\\udata.1c20d893e777ea0.000000000007ee23__head_2FDDF076__4", {st_mode=S_IFREG|0644, st_s
ize=4194304, ...}) = 0
lstat("rbd\\udata.1e02d691ddaefb.000000000000437c__head_1FADF076__4", {st_mode=S_IFREG|0644, st_si
ze=4194304, ...}) = 0
```

and it is using a default (be/4) io priority:

```
65567 be/4 root      768.61 K/s    0.00 B/s  0.00 %  0.00 % restorecon -i -f - -R -p -e /sys -e /p
roc -e /dev -e /mnt -e /var/tmp -e /home -e /tmp -e /dev
```

History

#1 - 10/29/2014 03:46 AM - Dan van der Ster

It is triggered by fixfiles -C /etc/selinux/targeted/contexts/files/file_contexts.pre restore

```
  |               `-yum,47342 /usr/bin/yum --skip-broken -x ceph* -x libceph* -x librados* -x librbd* -x kerne
l* -x ...
  |                   `-sh,51420 /var/tmp/rpm-tmp.GxmtU1 2
  |                       `-fixfiles,51822 /sbin/fixfiles -C /etc/selinux/targeted/contexts/files/file_context
s.pre restore
  |                           `-restorecon,51873 -i -f - -R -p -e /sys -e /proc -e /dev -e /mnt -e /var/tmp -e
 /home -e /tmp -e ...
```

#2 - 10/29/2014 05:25 AM - Dan van der Ster

Here's a solution:

```
echo "/var/lib/ceph/" >> /etc/selinux/fixfiles_exclude_dirs
```

#3 - 11/05/2014 08:30 PM - Wade Mealing

I would strongly reccomend limiting it to the subdirectories where large mounts are, not on the parent directory. This would probably solve the issue of having the relabeling eating into IO performance of those disks.

#4 - 05/28/2015 04:08 PM - Ken Dreyer

- Regression set to No

Milan, how can we implement a fix for this so it works out-of-the-box? Is selinux-policy-targeted the right place to fix this?

#5 - 05/28/2015 05:48 PM - Milan Broz

That "fix" can make things worse later... It is probably good for quick workaround for some particular case though.
IMHO the proper fix is to apply new selinux policy for ceph, we should start to test what we already have and will see if it is doable in some shorter term.

Anyway, it would be good to have this tracked in bugzilla for RHEL - that way RHEL selinux-policy maintainer can comment it.

#6 - 05/28/2015 06:38 PM - Boris Ranto

This is (well, will be) an intended behaviour, soon. We need to relabel the files for the SELinux policy to take effect (once it will be available).

That being said, the fixfiles script could probably be improved for better performance -- i.e. updating files based on policy changes or running single thread per single hdd/mount point to improve performance of the call.

btw: Do you see this behaviour also on rhel 7 and current range of fedoras? There might have been some performance improvements to this behaviour in later releases of selinux tools and these could technically get backported although it might already be too late for such a big change to rhel 6 environment in the rhel 6 release cycle.

#7 - 05/28/2015 07:33 PM - Ken Dreyer

Milan, one of the things Boris and I discussed is that Dan's strace shows a lot of stat() calls there. So even if the Ceph policy itself isn't changing, it still takes a while for restorecon to stat() everything under /var/lib/ceph/osd. I'm not sure what "a while" is, though (seconds, minutes, etc)?

#8 - 05/29/2015 07:04 AM - Dan van der Ster

Ken Dreyer wrote:

> I'm not sure what "a while" is, though (seconds, minutes, etc)?

It depends on the number of objects, of course. In our case it was taking 10's of minutes, hours in some cases. This was pretty nasty since the drives were pinned to 100% in iostat throughout the running of fixfiles and users were definitely suffering.

If this is a necessary operation, in the least fixfiles should run with a lower ionice priority, (e.g. -c2 -n7 like the mlocate daily cron), or best, ionice'd to

idle.

#9 - 05/29/2015 07:06 AM - Dan van der Ster

Boris Ranto wrote:

> btw: Do you see this behaviour also on rhel 7 and current range of fedoras?

I don't have a rhel7 ceph-osd server yet, so I can't comment.

#10 - 04/12/2017 04:37 PM - Greg Farnum

SELinux is used against Ceph now.

#11 - 04/12/2017 04:37 PM - Sage Weil

- Status changed from New to Can't reproduce

pls reopen if this is a problem on rhel7