devops - Bug #9860

grub/os-prober launch kills most ceph OSD

10/22/2014 12:12 AM - Laurent GUERBY

Status:	Fix Under Review	**% Done:**	0%
Priority:	Normal		
Assignee:			
Category:			
Target version:			
Source:	Community (user)	**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:		**Pull request ID:**	
Regression:		**Crash signature (v1):**	
Severity:	3 - minor	**Crash signature (v2):**	
Reviewed:			

Description

Workaround

Disable os-probe with

```
GRUB_DISABLE_OS_PROBER=true
```

http://www.gnu.org/software/grub/manual/html_node/Simple-configuration.html

Description

This morning automatic debian jessie package upgrade on our running system:

libsigc++-2.0-0c2a,libssl1.0.0,man-db,libgtk2.0-common,libgtk2.0-bin,
libgtk2.0-0,openssh-sftp-server,openssh-server,
openssh-client,grub-pc,grub-pc-bin,grub2-common,grub-common,openssl,python-cryptography,python-pygraphviz

killed five OSD out of 15 on our ceph 0.80.6 cluster of 5 machines :

root@g2:/var/log/ceph# grep -E ^2014-10-22 ceph-osd.7.log
2014-10-22 07:41:36.783358 7f4d33d55700 -1 journal FileJournal::write_bl : write_fd failed: (1) Operation not permitted
2014-10-22 07:41:36.783617 7f4d33d55700 -1 journal FileJournal::do_write: write_bl(pos=793935872) failed
2014-10-22 07:41:36.800201 7f4d33d55700 -1 os/FileJournal.cc: In function 'void FileJournal::do_write(ceph::bufferlist&)' thread
7f4d33d55700 time 2014-10-22 07:41:36.783629
2014-10-22 07:41:36.847389 7f4d33d55700 -1 ** **Caught signal (Aborted) ***

root@n7:/var/log/ceph# grep -E ^2014-10-22 ceph-osd.10.log|cut -c-120
2014-10-22 07:42:18.169142 7f9b977df700 -1 journal FileJournal::write_bl : write_fd failed: (1) Operation not permitted

root@n7:/var/log/ceph# grep -E ^2014-10-22 ceph-osd.9.log|cut -c-120
2014-10-22 07:42:17.509579 7f6efa27b700 -1 osd.9 15390 heartbeat_check: no reply from osd.13 since back 2014-10-22 07:41
2014-10-22 07:42:17.509593 7f6efa27b700 -1 osd.9 15390 heartbeat_check: no reply from osd.14 since back 2014-10-22 07:41
2014-10-22 07:42:17.945433 7f6ef6a74700 -1 journal FileJournal::do_write: pwrite(fd=23, hbp.length=4096) failed :(1) Ope
2014-10-22 07:42:17.960678 7f6ef6a74700 -1 os/FileJournal.cc: In function 'void FileJournal::do_write(ceph::bufferlist&)

root@stri:/var/log/ceph# grep -E ^2014-10-22 ceph-osd.13.log
2014-10-22 00:42:01.140574 7fa929b8a700 -1 journal FileJournal::write_bl : write_fd failed: (1) Operation not permitted
2014-10-22 00:42:01.141439 7fa929b8a700 -1 journal FileJournal::do_write: write_bl(pos=3496448000) failed

root@stri:/var/log/ceph# grep E ^2014-10-22 ceph-osd.14.log
2014-10-22 00:41:54.828719 7f438eb45700 -1 osd.14 15388 heartbeat_check: no reply from osd.7 since back 2014-10-22 00:41:34.499777 front 2014-10-22 00:41:34.499777 (cutoff 2014-10-22 00:41:34.828717)
2014-10-22 00:41:55.241586 7f437217f700 0 - 192.168.99.246:6811/17136 >> 192.168.99.253:6806/25800 pipe(0x7f439f5fd900 sd=182 :6811 s=0 pgs=0 cs=0 l=0 c=0x7f43a71f1180).accept connect_seq 34 vs existing 33 state standby
2014-10-22 00:42:01.235014 7f438b33e700 -1 journal FileJournal::write_bl : write_fd failed: (1) Operation not permitted
2014-10-22 00:42:01.235032 7f438b33e700 -1 journal FileJournal::do_write: write_bl(pos=4626878464) failed

The OSD all died just after a run of os-prober according to the logs:

Oct 22 07:41:36 g2 os-prober: debug: running /usr/lib/os-probes/mounted/05efi on mounted /dev/sda1

os-prober likely did an operation on the journal partition causing the write, may be the OSD could be made more robust in this case.

Meanwhile we deactivated os-prober/grub updates.

History

#1 - 10/22/2014 08:25 AM - Laurent GUERBY

Adding more complete log lines with ASSERT references

<guerby> 2014-10-22 07:42:07.369785 7f6edf0b5700 0 -- 192.168.99.247:0/8971 >> 192.168.99.246:6812/17136 pipe(0x7f6f074c2f00 sd=79 :0 s=1 pgs=0 cs=0 l=1 c=0x7f6f158ed440).fault
<guerby> 2014-10-22 07:42:17.509579 7f6efa27b700 1 osd.9 15390 heartbeat_check: no reply from osd.13 since back 2014-10-22 07:41:56.985182 front 2014-10-22 07:41:56.985182 (cutoff 2014-10-22 07:41:57.509576)
<guerby> 2014-10-22 07:42:17.509593 7f6efa27b700 -1 osd.9 15390 heartbeat_check: no reply from osd.14 since back 2014-10-22 07:41:56.985182 front 2014-10-22 07:41:56.985182 (cutoff 2014-10-22 07:41:57.509576)
<guerby> 2014-10-22 07:42:17.945433 7f6cf6a74700 -1 journal FileJournal::do_write: pwrite(fd=23, hbp.length=4096) failed :(1) Operation not permitted
<guerby> 2014-10-22 07:42:17.960678 7f6cf6a74700 -1 os/FileJournal.cc: In function 'void FileJournal::do_write(ceph::bufferlist&)' thread 7f6cf6a74700 time 2014-10-22 07:42:17.945642
<guerby> os/FileJournal.cc: 1021: FAILED assert(0)
<guerby> ceph version 0.80.6 (f93610a4421cb670b08e974c6550ee715ae528ae)
<guerby> 1: (FileJournal::do_write(ceph::buffer::list&)+0xee4) [0x7f6f01b42e14]
<guerby> 2: (FileJournal::write_thread_entry()+0x7eb) [0x7f6f01b4674b]
<guerby> 3: (FileJournal::Writer::entry()+0xd) [0x7f6f01a51e8d]
<guerby> 4: (()+0x80a4) [0x7f6f00a0f0a4]
<guerby> sur osd 13 stri
<guerby> 2014-10-22 00:41:37.135046 7fa913bfd700 0 - 192.168.99.246:0/16346 >> 192.168.99.252:6807/4371 pipe(0x7fa93dbfe780 sd=201 :0 s=1 pgs=0 cs=0 l=1 c=0x7fa94c2d3b80).fault
<guerby> 2014-10-22 00:41:56.528719 7fa90ce14700 0 -- 192.168.99.246:6806/2016346 >> 192.168.99.246:6816/17923 pipe(0x7fa9553faf00 sd=272 :6806 s=0 pgs=0 cs=0 l=0 c=0x7fa940d61e40).accept connect_seq 2 vs existing 1 state standby
<guerby> 2014-10-22 00:42:01.140574 7fa929b8a700 -1 journal FileJournal::write_bl : write_fd failed: (1) Operation not permitted
<guerby> 2014-10-22 00:42:01.141439 7fa929b8a700 -1 journal FileJournal::do_write: write_bl(pos=3496448000) failed
<guerby> 2014-10-22 00:42:01.234485 7fa929b8a700 -1 os/FileJournal.cc: In function 'void FileJournal::do_write(ceph::bufferlist&)' thread 7fa929b8a700 time 2014-10-22 00:42:01.141447
<guerby> os/FileJournal.cc: 1028: FAILED assert(0)

#2 - 10/22/2014 08:42 AM - Loïc Dachary

- Description updated

- Status changed from New to 12

#3 - 10/22/2014 08:45 AM - Laurent GUERBY

Logs detailing what os-prober was doing when one of the OSD crashed, sda2 is the journal partition of osd.13 who got "write_fd failed: (1) Operation not permitted"

Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/50mounted-tests on /dev/sda2
Oct 22 07:42:01 stri kernel: [675159.805580] XFS (sda2): Invalid superblock magic number
Oct 22 07:42:01 stri kernel: [675159.815137] FAT-fs (sda2): utf8 is not a recommended IO charset for FAT filesystems, filesystem will be case sensitive!
Oct 22 07:42:01 stri kernel: [675159.831805] FAT-fs (sda2): Can't find a valid FAT filesystem
Oct 22 07:42:01 stri kernel: [675159.838616] FAT-fs (sda2): utf8 is not a recommended IO charset for FAT filesystems, filesystem will be case sensitive!
Oct 22 07:42:01 stri kernel: [675159.855214] FAT-fs (sda2): Can't find a valid FAT filesystem
Oct 22 07:42:01 stri kernel: [675159.865066] VFS: Can't find a Minix filesystem V1 | V2 | V3 on device sda2.
Oct 22 07:42:01 stri kernel: [675159.897277] hfsplus: unable to find HFS+ superblock
Oct 22 07:42:01 stri kernel: [675159.909261] You didn't specify the type of your ufs filesystem
Oct 22 07:42:01 stri kernel: [675159.909261]
Oct 22 07:42:01 stri kernel: [675159.909261] mount -t ufs -o ufstype=sun|sunx86|44bsd|ufs2|5xbsd|old|hp|nextstep|nextstep-cd|openstep ...
Oct 22 07:42:01 stri kernel: [675159.909261]
Oct 22 07:42:01 stri kernel: [675159.909261] >>>WARNING<<< Wrong ufstype may corrupt your filesystem, default is ufstype=old
Oct 22 07:42:01 stri kernel: [675159.936520] ufs_read_super: bad magic number
Oct 22 07:42:01 stri kernel: [675159.943478] hfs: can't find a HFS filesystem on dev sda2

May be one of these tests does something lock/read-write instead of ro

#4 - 10/22/2014 08:46 AM - Loïc Dachary

- Project changed from Ceph to devops

#5 - 10/22/2014 08:53 AM - Laurent GUERBY

And sda1 which is the ext4 mounted disj of osd.13

Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/05efi on mounted /dev/sda1
Oct 22 07:42:00 stri 05efi: debug: Not on UEFI platform
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/10freedos on mounted /dev/sda1
Oct 22 07:42:00 stri 10freedos: debug: /dev/sda1 is not a FAT partition: exiting
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/10qnx on mounted /dev/sda1
Oct 22 07:42:00 stri 10qnx: debug: /dev/sda1 is not a QNX4 partition: exiting
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/20macosx on mounted /dev/sda1
Oct 22 07:42:00 stri macosx-prober: debug: /dev/sda1 is not an HFS+ partition: exiting
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/20microsoft on mounted /dev/sda1
Oct 22 07:42:00 stri 20microsoft: debug: /dev/sda1 is not a MS partition: exiting
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/30utility on mounted /dev/sda1
Oct 22 07:42:00 stri 30utility: debug: /dev/sda1 is not a FAT partition: exiting
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/40lsb on mounted /dev/sda1
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/70hurd on mounted /dev/sda1

```
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/80minix on mounted /dev/sda1
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/83haiku on mounted /dev/sda1
Oct 22 07:42:00 stri 83haiku: debug: /dev/sda1 is not a BeFS partition: exiting
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/90linux-distro on mounted /dev/sda1
Oct 22 07:42:00 stri os-prober: debug: running /usr/lib/os-probes/mounted/90solaris on mounted /dev/sda1
```

#6 - 06/28/2016 04:11 PM - Tim Bishop

I hit this issue today, and I'll probably use the suggested workaround. Surely this should affect more people? And with the move to Bluestore I'd expect it to become a bigger issue.

An alternative solution might be to stop os-prober poking the Ceph disks. I tried the following, which appeared to work:

```
/usr/lib/os-probes/10ceph
```

```
#!/bin/sh
# Ceph tests
set -e
partition="$1"

. /usr/share/os-prober/common.sh

if [ -x /usr/sbin/ceph-disk ]; then
        /usr/sbin/ceph-disk list | grep "$partition" | grep -q ceph
        if [ $? = 0 ]; then
                debug "$1 is used by Ceph; skipping"
                exit 0
        fi
fi

exit 1
```

This runs before the main tests for unmounted disks and therefore avoids anything being checked. It's crude and could probably be made better though.

It might be worth adding something similar to the probes for mounted disks to avoid noise like this in the logs (which made me panic for a few minutes when I first saw them - corrupted disk?!):

```
Jun 28 03:30:33 johnson kernel: [292861.890032] JFS: nTxBlock = 8192, nTxLock = 65536
Jun 28 03:30:33 johnson kernel: [292861.912931] ntfs: driver 2.1.32 [Flags: R/O MODULE].
Jun 28 03:30:33 johnson kernel: [292861.944420] QNX4 filesystem 0.2.3 registered.
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/05efi on mounted /dev/sda1
Jun 28 03:30:33 johnson 05efi: debug: Not on UEFI platform
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/10freedos on mounted /dev/sda1
Jun 28 03:30:33 johnson 10freedos: debug: /dev/sda1 is not a FAT partition: exiting
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/10qnx on mounted /dev/sda1
Jun 28 03:30:33 johnson 10qnx: debug: /dev/sda1 is not a QNX4 partition: exiting
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/20macosx on mounted /dev/sda1
Jun 28 03:30:33 johnson macosx-prober: debug: /dev/sda1 is not an HFS+ partition: exiting
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/20microsoft on mounted /dev/sda1
Jun 28 03:30:33 johnson 20microsoft: debug: /dev/sda1 is not a MS partition: exiting
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/30utility on mounted /dev/sda1
Jun 28 03:30:33 johnson 30utility: debug: /dev/sda1 is not a FAT partition: exiting
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/40lsb on mounted /dev/sda1
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/70hurd on mounted /dev/sda1
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/80minix on mounted /dev/sda1
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/83haiku on mounted /dev/sda1
Jun 28 03:30:33 johnson 83haiku: debug: /dev/sda1 is not a BeFS partition: exiting
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/90linux-distro on mounted /dev/sd
a1
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/mounted/90solaris on mounted /dev/sda1
Jun 28 03:30:33 johnson os-prober: debug: running /usr/lib/os-probes/50mounted-tests on /dev/sda2
Jun 28 03:30:33 johnson kernel: [292862.118060] EXT4-fs (sda2): VFS: Can't find ext4 filesystem
Jun 28 03:30:33 johnson kernel: [292862.125952] EXT4-fs (sda2): VFS: Can't find ext4 filesystem
Jun 28 03:30:33 johnson kernel: [292862.133707] EXT4-fs (sda2): VFS: Can't find ext4 filesystem
```

```
Jun 28 03:30:33 johnson kernel: [292862.156358] FAT-fs (sda2): bogus number of reserved sectors
Jun 28 03:30:33 johnson kernel: [292862.162702] FAT-fs (sda2): Can't find a valid FAT filesystem
Jun 28 03:30:33 johnson kernel: [292862.185288] XFS (sda2): Invalid superblock magic number
Jun 28 03:30:33 johnson kernel: [292862.189496] FAT-fs (sda2): bogus number of reserved sectors
Jun 28 03:30:33 johnson kernel: [292862.195839] FAT-fs (sda2): Can't find a valid FAT filesystem
Jun 28 03:30:33 johnson kernel: [292862.208808] VFS: Can't find a Minix filesystem V1 | V2 | V3 on device sda2
.
Jun 28 03:30:33 johnson kernel: [292862.216387] hfsplus: unable to find HFS+ superblock
Jun 28 03:30:33 johnson kernel: [292862.222985] qnx4: no qnx4 filesystem (no root dir).
Jun 28 03:30:33 johnson kernel: [292862.231968] ufs: You didn't specify the type of your ufs filesystem
Jun 28 03:30:33 johnson kernel: [292862.231968]
Jun 28 03:30:33 johnson kernel: [292862.231968] mount -t ufs -o ufstype=sun|sunx86|44bsd|ufs2|5xbsd|old|hp|nex
tstep|nextstep-cd|openstep ...
Jun 28 03:30:33 johnson kernel: [292862.231968]
Jun 28 03:30:33 johnson kernel: [292862.231968] >>>WARNING<<< Wrong ufstype may corrupt your filesystem, defau
lt is ufstype=old
Jun 28 03:30:33 johnson kernel: [292862.263232] ufs: ufs_fill_super(): bad magic number
Jun 28 03:30:33 johnson kernel: [292862.272985] hfs: can't find a HFS filesystem on dev sda2
```

sda is a Ceph disk; sda1 data, sda2 journal.

#7 - 11/21/2016 04:23 PM - Kefu Chai

- Category deleted (OSD)

- Status changed from 12 to 15

see https://bugs.debian.org/cgi-bin/bugreport.cgi?bug=806273#30

#8 - 11/21/2016 04:23 PM - Kefu Chai

might want to document this...

#9 - 11/21/2016 04:34 PM - Nick Fisk

Ubuntu bug https://bugs.launchpad.net/ubuntu/+source/os-prober/+bug/1643614

#10 - 12/05/2019 09:45 PM - Patrick Donnelly

- Status changed from 15 to Fix Under Review