Ceph - Bug #46366

Octopus: Recovery and backfilling causes OSDs to crash after upgrading from nautilus to octopus

07/05/2020 12:36 PM - Wout van Heeswijk

Status:	Fix Under Review	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Kefu Chai		
Category:	OSD		
Target version:			
Source:	Community (user)	**Reviewed:**	
Tags:		**Affected Versions:**	v15.2.4
Backport:	octopus	**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	35952
Severity:	1 - critical	**Crash signature:**	

Description

A customer has upgraded the cluster from nautilus to octopus after experiencing issues with osds not being able to connect to each other, clients/mons/mgrs. The connectivity issues was related to the msgrV2 and require_osd_release setting not being set to nautilus. After fixing this the OSDs were restarted and all placement groups became active again.

After unsetting the norecover and nobackfill flag some OSDs started crashing every few minutes. The OSD log, even with high debug settings, don't seem to reveal anything, it just stops logging mid log line.

In the systemd journal there is the following message:

```
Jul 05 13:41:50 st0.r23.spod1.rtm0.transip.io ceph-osd[92605]: *** Caught signal (Segmentation fau
lt) **
Jul 05 13:41:50 st0.r23.spod1.rtm0.transip.io ceph-osd[92605]:  in thread 557dc6fb3510 thread_name
:tp_osd_tp
Jul 05 13:41:50 st0.r23.spod1.rtm0.transip.io ceph-osd[92605]: src/tcmalloc.cc:283] Attempt to fre
e invalid pointer 0x363bbb77000
Jul 05 13:41:50 st0.r23.spod1.rtm0.transip.io ceph-osd[92605]: *** Caught signal (Aborted) **
Jul 05 13:41:50 st0.r23.spod1.rtm0.transip.io ceph-osd[92605]:  in thread 557dc6fb3510 thread_name
:tp_osd_tp
Jul 05 13:41:50 st0.r23.spod1.rtm0.transip.io ceph-osd[92605]: src/tcmalloc.cc:283] Attempt to fre
e invalid pointer 0x363bbb77000
```

snippet of log from time around crash.

```
2020-07-05T06:31:33.547+0200 7f8860296700 -1 osd.127 1496224 heartbeat_check: no reply from 10.200
.19.17:6836 osd.111 since back 2020-07-05T06:28:30.776006+0200 front 2020-07-05T06:28:30.775261+02
00 (oldest deadline 2020-07-05T06:28:53.0
73588+0200)
2020-07-05T06:31:33.547+0200 7f8860296700 -1 osd.127 1496224 heartbeat_check: no reply from 10.200
.19.37:6901 osd.146 since back 2020-07-05T06:31:01.434299+0200 front 2020-07-05T06:31:01.434534+02
00 (oldest deadline 2020-07-05T06:31:27.2
33589+0200)
2020-07-05T06:31:33.547+0200 7f8860296700 -1 osd.127 1496224 heartbeat_check: no reply from 10.200
.19.38:6929 osd.180 since back 2020-07-05T06:28:18.971489+0200 front 2020-07-05T06:28:18.971597+02
00 (oldest deadline 2020-07-05T06:28:50.7
71298+0200)
2020-07-05T06:31:33.547+0200 7f8860296700 -1 osd.127 1496224 heartbeat_check: no reply from 10.200
.19.38:6891 osd.189 since back 2020-07-05T06:28:18.971678+0200 front 2020-07-05T06:28:18.971894+02
00 (oldest deadline 2020-07-05T06:28:44.8
69635+0200)
2020-07-05T06:31:33.547+0200 7f8860296700 -1 osd.127 1496224 heartbeat_check: no reply from 10.200
.19.48:6836 osd.229 since back 2020-07-05T06:31:07.237691+0200 front 2020-07-05T06:31:07.237226+02
```

```
00 (oldest deadline 2020-07-05T06:31:30.7
34951+0200)
2020-07-05T06:35:04.026+0200 7ff24a7e8d80  0 set uid:gid to 64045:64045 (ceph:ceph)
2020-07-05T06:35:04.026+0200 7ff24a7e8d80  0 ceph version 15.2.4 (7447c15c6ff58d7fce91843b705a268a
1917325c) octopus (stable), process ceph-osd, pid 1667604
2020-07-05T06:35:04.026+0200 7ff24a7e8d80  0 pidfile_write: ignore empty --pid-file
2020-07-05T06:35:04.026+0200 7ff24a7e8d80  1 bdev create path /var/lib/ceph/osd/ceph-127/block typ
e kernel
2020-07-05T06:35:04.026+0200 7ff24a7e8d80  1 bdev(0x55f03b8f6380 /var/lib/ceph/osd/ceph-127/block)
 open path /var/lib/ceph/osd/ceph-127/block
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f6380 /var/lib/ceph/osd/ceph-127/block)
 open size 12000134430720 (0xae9ffc00000, 11 TiB) block_size 4096 (4 KiB) rotational discard not s
upported
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bluestore(/var/lib/ceph/osd/ceph-127) _set_cache_size
s cache_size 1073741824 meta 0.4 kv 0.4 data 0.2
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev create path /var/lib/ceph/osd/ceph-127/block.db
type kernel
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f6a80 /var/lib/ceph/osd/ceph-127/block.
db) open path /var/lib/ceph/osd/ceph-127/block.db
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f6a80 /var/lib/ceph/osd/ceph-127/block.
db) open size 128849018880 (0x1e00000000, 120 GiB) block_size 4096 (4 KiB) non-rotational discard
supported
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bluefs add_block_device bdev 1 path /var/lib/ceph/osd
/ceph-127/block.db size 120 GiB
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev create path /var/lib/ceph/osd/ceph-127/block typ
e kernel
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f6e00 /var/lib/ceph/osd/ceph-127/block)
 open path /var/lib/ceph/osd/ceph-127/block
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f6e00 /var/lib/ceph/osd/ceph-127/block)
 open size 12000134430720 (0xae9ffc00000, 11 TiB) block_size 4096 (4 KiB) rotational discard not s
upported
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bluefs add_block_device bdev 2 path /var/lib/ceph/osd
/ceph-127/block size 11 TiB
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev create path /var/lib/ceph/osd/ceph-127/block.wal
 type kernel
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f7180 /var/lib/ceph/osd/ceph-127/block.
wal) open path /var/lib/ceph/osd/ceph-127/block.wal
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f7180 /var/lib/ceph/osd/ceph-127/block.
wal) open size 2147483648 (0x80000000, 2 GiB) block_size 4096 (4 KiB) non-rotational discard suppo
rted
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bluefs add_block_device bdev 0 path /var/lib/ceph/osd
/ceph-127/block.wal size 2 GiB
2020-07-05T06:35:04.030+0200 7ff24a7e8d80  1 bdev(0x55f03b8f7180 /var/lib/ceph/osd/ceph-127/block.
wal) close
```

A gdb backtrace is attached that reveals some more info.

History

#1 - 07/06/2020 02:16 PM - Kefu Chai

hi Wout,

what distro are you using? where did the ceph packages come from? could you install debug debugsymbls packages of ceph-osd and librados2 for a more readable backtrace?

#2 - 07/06/2020 02:26 PM - Wout van Heeswijk

Hi Kefu,

Distributor ID: Ubuntu
Description: Ubuntu 18.04.4 LTS
Release: 18.04

Codename: bionic
Source package: deb http://download.ceph.com/debian-octopus bionic main

Package: ceph-common
Status: install ok installed
Priority: optional
Section: admin
Installed-Size: 73640
Maintainer: Ceph Maintainers <ceph-maintainers@lists.ceph.com>
Architecture: amd64
Source: ceph
Version: 15.2.4-1bionic

ceph version 15.2.4 (7447c15c6ff58d7fce91843b705a268a1917325c) octopus (stable)

uname -a current kernel: 4.15.0-99-generic
uname -a backported kernel: 5.4.0-40-generic

I'm installing the debugsymbls packages as i'm typing this.

#3 - 07/06/2020 04:13 PM - Wout van Heeswijk

We've also tested with the docker container distribution. It crashes with the same segfault:

- Caught signal (Segmentation fault) **
  in thread 55918cd11a10 thread_name:tp_osd_tp

Using the cli below without the FSID.

/bin/docker run --rm --net=host --privileged --group-add=disk --name
ceph-[FSID-PREFIX]-osd.[OSD-ID] -e
CONTAINER_IMAGE=docker.io/ceph/ceph:v15 -e NODE_NAME=[HOSTNAME] -v
/var/run/ceph/[FSID-PREFIX]:/var/run/ceph:z -v
/var/log/ceph/[FSID-PREFIX]:/var/log/ceph:z -v
/var/lib/ceph/[FSID-PREFIX]/crash:/var/lib/ceph/crash:z -v
/var/lib/ceph/[FSID-PREFIX]/osd.[OSD-ID]:/var/lib/ceph/osd/ceph-3:z -v
/var/lib/ceph/[FSID-PREFIX]/osd.[OSD-ID]/config:/etc/ceph/ceph.conf:z -v
/dev:/dev -v /run/udev:/run/udev -v /sys:/sys -v /run/lvm:/run/lvm -v
/run/lock/lvm:/run/lock/lvm --entrypoint /usr/bin/ceph-osd
docker.io/ceph/ceph:v15 -n osd.[OSD-ID] -f --setuser ceph --setgroup
ceph --default-log-to-file=false --default-log-to-stderr=true
--default-log-stderr-prefix=debug

#4 - 07/06/2020 05:27 PM - Wout van Heeswijk

After reverting https://github.com/ceph/ceph/commit/74be30c8b7c the crashing ceph-osd's seem to be stable.

Related:
https://github.com/ceph/ceph/pull/30061
https://github.com/ceph/ceph/pull/29588

#5 - 07/06/2020 05:36 PM - Wout van Heeswijk

- File stacktrace.gz added

- File aio.patch added

Wout van Heeswijk wrote:

> After reverting https://github.com/ceph/ceph/commit/74be30c8b7c the crashing ceph-osd's seem to be stable.
>
> Related:
> https://github.com/ceph/ceph/pull/30061
> https://github.com/ceph/ceph/pull/29588

The above patch was found based on a backtrace that was executed with the attached patch applied.

#6 - 07/07/2020 05:28 AM - xie xingguo

Wout van Heeswijk wrote:

> Wout van Heeswijk wrote:
>
> > After reverting https://github.com/ceph/ceph/commit/74be30c8b7c the crashing ceph-osd's seem to be stable.
> >
> > Related:
> > https://github.com/ceph/ceph/pull/30061
> > https://github.com/ceph/ceph/pull/29588
>
> The above patch was found based on a backtrace that was executed with the attached patch applied.

Hi, Wout
I am wondering how could it be possible for an object to become so fragmented:(
Did you modify the default size of rados object? What kind of app (rbd, rgw, cephfs) are you using?

#7 - 07/07/2020 05:56 AM - Kefu Chai

- Status changed from New to Fix Under Review

- Assignee set to Kefu Chai

- Backport set to octopus

- Pull request ID set to 35952

hi, Wout,

thanks for your investigation and patch. i revised it to use boost::container::small_vector. see https://github.com/ceph/ceph/pull/35952

#8 - 07/07/2020 07:35 AM - Wout van Heeswijk

Kefu Chai wrote:

> hi, Wout,
>
> thanks for your investigation and patch. i revised it to use boost::container::small_vector. see https://github.com/ceph/ceph/pull/35952

No problem, we had a good bughunt session with our customer! Thank you for this quick fix Kefu!

Currently some jenkins tests are failing, but it is not related to the patch. It is because of space within the build environment. Other tests are also failing because of this.

#9 - 07/07/2020 10:26 AM - Wout van Heeswijk

xie xingguo wrote:

> Wout van Heeswijk wrote:
>
> > Wout van Heeswijk wrote:
> >
> > > After reverting https://github.com/ceph/ceph/commit/74be30c8b7c the crashing ceph-osd's seem to be stable.
> > >
> > > Related:
> > > https://github.com/ceph/ceph/pull/30061
> > > https://github.com/ceph/ceph/pull/29588
> >
> > The above patch was found based on a backtrace that was executed with the attached patch applied.
>
> Hi, Wout
> I am wondering how could it be possible for an object to become so fragmented:(
> Did you modify the default size of rados object? What kind of app (rbd, rgw, cephfs) are you using?

I've created logs with high debugging settings yesterday. I've asked they can be shared via another route with links here.

#10 - 07/07/2020 02:59 PM - Wout van Heeswijk

xie xingguo wrote:

> Wout van Heeswijk wrote:
>
> > Wout van Heeswijk wrote:
> >
> > > After reverting https://github.com/ceph/ceph/commit/74be30c8b7c the crashing ceph-osd's seem to be stable.
> > >
> > > Related:
> > > https://github.com/ceph/ceph/pull/30061
> > > https://github.com/ceph/ceph/pull/29588
> >
> > The above patch was found based on a backtrace that was executed with the attached patch applied.
>
> Hi, Wout
> I am wondering how could it be possible for an object to become so fragmented:(
> Did you modify the default size of rados object? What kind of app (rbd, rgw, cephfs) are you using?

Here are the rather large log files:
https://robing-disk.stackstorage.com/s/eEM0xaCuF6MBdkqg
https://robing-disk.stackstorage.com/s/KfFPQpAfrOgiCLF9

#11 - 12/17/2020 09:16 AM - Wout van Heeswijk

Kefu Chai wrote:

> hi, Wout,
>
> thanks for your investigation and patch. i revised it to use boost::container::small_vector. see https://github.com/ceph/ceph/pull/35952

Hi Kefu,

Do you have any thoughts on this ticket?

Files

backtrace.txt	4.7 KB	07/05/2020	Wout van Heeswijk
stacktrace.gz	6.04 KB	07/06/2020	Wout van Heeswijk
aio.patch	798 Bytes	07/06/2020	Wout van Heeswijk