bluestore - Bug #23333

bluestore: ENODATA on aio

03/13/2018 02:05 PM - Robert Sander

Status:	Resolved	**% Done:**	0%
Priority:	Normal		
Assignee:	Radoslaw Zarzynski		
Category:			
Target version:			
Source:	Community (user)	**Reviewed:**	
Tags:		**Affected Versions:**	v12.2.4
Backport:	luminous	**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

Since 3 days one of 18 BlueStore OSDs is constantly crashing:

2018-03-10 04:01:45.366202 mon.ceph01 mon.0 192.168.44.65:6789/0 3977 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.6)
2018-03-11 01:53:06.019025 mon.ceph01 mon.0 192.168.44.65:6789/0 8132 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.10)
2018-03-11 03:33:24.455042 mon.ceph01 mon.0 192.168.44.65:6789/0 8477 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.17)
2018-03-12 03:36:50.437396 mon.ceph01 mon.0 192.168.44.65:6789/0 12976 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.6)
2018-03-12 04:32:44.685349 mon.ceph01 mon.0 192.168.44.65:6789/0 13237 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-12 05:39:14.787894 mon.ceph01 mon.0 192.168.44.65:6789/0 13498 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.10)
2018-03-12 06:31:09.084123 mon.ceph01 mon.0 192.168.44.65:6789/0 13753 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-12 07:22:27.192456 mon.ceph01 mon.0 192.168.44.65:6789/0 13968 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-12 08:18:47.625200 mon.ceph01 mon.0 192.168.44.65:6789/0 14172 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.6)
2018-03-12 09:15:47.241273 mon.ceph01 mon.0 192.168.44.65:6789/0 14455 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.15)
2018-03-12 10:10:27.787580 mon.ceph01 mon.0 192.168.44.65:6789/0 14654 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.10)
2018-03-12 11:04:01.766949 mon.ceph01 mon.0 192.168.44.65:6789/0 14923 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.13)
2018-03-12 11:57:24.455534 mon.ceph01 mon.0 192.168.44.65:6789/0 15232 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.17)
2018-03-12 12:49:38.911668 mon.ceph01 mon.0 192.168.44.65:6789/0 15474 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-12 13:45:46.821370 mon.ceph01 mon.0 192.168.44.65:6789/0 15727 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.1)
2018-03-12 14:37:54.494178 mon.ceph01 mon.0 192.168.44.65:6789/0 15904 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-12 15:31:14.487131 mon.ceph01 mon.0 192.168.44.65:6789/0 16122 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.13)
2018-03-12 16:22:38.799025 mon.ceph01 mon.0 192.168.44.65:6789/0 16353 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.16)
2018-03-12 17:14:04.971014 mon.ceph01 mon.0 192.168.44.65:6789/0 16560 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-12 18:07:00.722340 mon.ceph01 mon.0 192.168.44.65:6789/0 16789 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.15)
2018-03-12 18:59:43.827737 mon.ceph01 mon.0 192.168.44.65:6789/0 17047 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.15)

2018-03-12 19:50:38.287630 mon.ceph01 mon.0 192.168.44.65:6789/0 17296 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.9)
2018-03-12 20:55:07.389005 mon.ceph01 mon.0 192.168.44.65:6789/0 17594 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.17)
2018-03-12 21:54:15.907260 mon.ceph01 mon.0 192.168.44.65:6789/0 17865 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.1)
2018-03-12 22:52:19.808324 mon.ceph01 mon.0 192.168.44.65:6789/0 18064 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.16)
2018-03-12 23:46:47.508301 mon.ceph01 mon.0 192.168.44.65:6789/0 18317 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.2)
2018-03-13 00:39:21.711974 mon.ceph01 mon.0 192.168.44.65:6789/0 18516 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.16)
2018-03-13 01:31:55.385104 mon.ceph01 mon.0 192.168.44.65:6789/0 18691 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.6)
2018-03-13 02:26:23.183917 mon.ceph01 mon.0 192.168.44.65:6789/0 18912 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.9)
2018-03-13 04:01:22.455514 mon.ceph01 mon.0 192.168.44.65:6789/0 19206 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.10)
2018-03-13 05:09:33.791115 mon.ceph01 mon.0 192.168.44.65:6789/0 19493 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.1)
2018-03-13 06:08:01.085420 mon.ceph01 mon.0 192.168.44.65:6789/0 19726 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.10)
2018-03-13 07:47:49.104523 mon.ceph01 mon.0 192.168.44.65:6789/0 20058 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-13 08:42:45.958985 mon.ceph01 mon.0 192.168.44.65:6789/0 20280 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-13 09:38:24.619503 mon.ceph01 mon.0 192.168.44.65:6789/0 20508 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.1)
2018-03-13 10:32:51.563617 mon.ceph01 mon.0 192.168.44.65:6789/0 20723 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.11)
2018-03-13 11:25:48.127581 mon.ceph01 mon.0 192.168.44.65:6789/0 20987 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.16)
2018-03-13 12:20:24.128393 mon.ceph01 mon.0 192.168.44.65:6789/0 21255 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.13)
2018-03-13 13:15:43.325775 mon.ceph01 mon.0 192.168.44.65:6789/0 21443 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.6)
2018-03-13 14:09:19.403727 mon.ceph01 mon.0 192.168.44.65:6789/0 21687 : cluster [INF] osd.12 failed (root=default,host=ceph02,ssd=sdh02) (connection refused reported by osd.15)

The latest logfile of that OSD is attached. At the time of the last crash it contains:

```
1> 2018-03-13 14:09:18.997826 7f9ef880a700 -1 bdev(0x558ab92fafc0 /var/lib/ceph/osd/ceph-12/block) _aio_thread got (61) No data available
     0> 2018-03-13 14:09:18.999284 7f9ef880a700 -1 /build/ceph-12.2.4/src/os/bluestore/KernelDevice.cc: In function 'void KernelDevice::_aio_thread()' thread 7f9ef880a700 time 2018-03
13 14:09:18.997856
/build/ceph-12.2.4/src/os/bluestore/KernelDevice.cc: 379: FAILED assert(0 == "got unexpected error from io_getevents")



ceph version 12.2.4 (52085d5249a80c5f5121a76d6288429f35e4e77b) luminous (stable)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x102) [0x558aae2fc872]
 2: (KernelDevice::_aio_thread()+0x16ca) [0x558aae29c6ea]
 3: (KernelDevice::AioCompletionThread::entry()+0xd) [0x558aae2a19cd]
 4: (()+0x76ba) [0x7f9f01b746ba]
 5: (clone()+0x6d) [0x7f9f00beb41d]
```

Related issues:

Related to bluestore - Bug #23426: aio thread got No space left on device	**Won't Fix**	**03/20/2018**
Copied to bluestore - Backport #23672: luminous: bluestore: ENODATA on aio	**Resolved**	

History

#1 - 03/14/2018 01:57 PM - Sage Weil

- Project changed from RADOS to bluestore

- Subject changed from One OSD constantly crashes to bluestore: ENODATA on aio

- Status changed from New to 12

#2 - 03/19/2018 01:50 PM - Radoslaw Zarzynski

- Status changed from 12 to Need More Info

- Assignee set to Radoslaw Zarzynski

This looks really interesting. The assertion failure came from the *io_getevents* (called by one of the *bstore_aio* threads) that finished with *ENODATA*. Surprisingly, the *man 2 io_getevents* doesn't enlist *ENODATA* as a possible error code. Maybe it's just undocumented, maybe originates in a lower layer (driver?). What is certain is that BlueStore currently is able to deal (in a way other than assertion) solely with the *EINTR*.

Could you please provide more information about used OS and hardware? Especially kernel version is vital as taking a look on the implementation details seems necessary.

#3 - 03/19/2018 03:01 PM - Robert Sander

Radoslaw Zarzynski wrote:

> Could you please provide more information about used OS and hardware? Especially kernel version is vital as taking a look on the implementation details seems necessary.

The Ceph OSDs run on Ubuntu 16.04.4 with kernel 4.13.0-36-generic on Dell PowerEdge R720xd with Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz CPUs. HDDs and SSDs are connect to an LSI Logic / Symbios Logic SAS2308 PCI-Express Fusion-MPT SAS-2 controller.

The BlueStore OSDs are on 4TB HDDs and configured to use a 10GB partition on SSD as WAL/DB. The size is not optimal but was used as this has been the FileStore journal before the migration to BlueStore.

Any other details you need?

#4 - 03/19/2018 05:17 PM - Radoslaw Zarzynski

Hi Robert,

thanks for providing the info! I've took a look on the implementation of *io_getevents* in your kernel but can't find any obvious way to get *ENODATA*. Maybe it can (?) be accidentally calculated by the loop in *aio_read_events_ring* (link to vanilla kernel) but this appears improbable to me. Before diving into improbabilities, let's make a quick cut between user and kernel spaces to exclude e.g. uspace error code corruption.

Could you try to run the OSD under e.g. *strace*? Something like *strace -f -e trace=io_getevents* would be really useful.

#5 - 03/19/2018 08:55 PM - Radoslaw Zarzynski

Errata, the issue is related to *aio_t::get_return_value*, not to *io_getevents*. Our debug message is misleading. I've just sent a pull request rectifying

that.

There is no business in getting output from *strace* at the moment. Instead, could you please:

- provide log with increased debug log levels (*debug_osd = 20* and *debug_bdev = 30*),
- verify the hardware for failure (*dmesg*, SMART etc.)?

Also, big thanks to Mr Marcin Gibula for the consultations.

#6 - 03/19/2018 09:22 PM - Robert Sander

Radoslaw Zarzynski wrote:

> Instead, could you please:
>
> - provide log with increased debug log levels (*debug_osd = 20* and *debug_bdev = 30*),
> - verify the hardware for failure (*dmesg*, SMART etc.)?

The issue only happened between March 12 6:31 and March 13 15:55 for 36 times, but not even once after that.
It's a little bit strange, but it seems to have healed itself.
I am not sure if I want to increase debug output on this production cluster.

#7 - 03/19/2018 09:25 PM - Robert Sander

Robert Sander wrote:

> It's a little bit strange, but it seems to have healed itself.

Looking around I just found entries in /var/log/kern.log related:

Mar 13 15:55:45 ceph02 kernel: [362540.919365] sd 0:0:4:0: [sde] tag#4 FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE
Mar 13 15:55:45 ceph02 kernel: [362540.919388] sd 0:0:4:0: [sde] tag#4 Sense Key : Medium Error [current]
Mar 13 15:55:45 ceph02 kernel: [362540.919394] sd 0:0:4:0: [sde] tag#4 Add. Sense: Read retries exhausted
Mar 13 15:55:45 ceph02 kernel: [362540.919401] sd 0:0:4:0: [sde] tag#4 CDB: Read(16) 88 00 00 00 00 01 38 af 66 f8 00 00 00 60 00 00
Mar 13 15:55:45 ceph02 kernel: [362540.919407] print_req_error: critical medium error, dev sde, sector 5245986552

Looking back it was always the same sector.

#8 - 03/20/2018 12:57 PM - Radoslaw Zarzynski

- Status changed from Need More Info to In Progress

> Mar 13 15:55:45 ceph02 kernel: [362540.919407] print_req_error: critical medium error, dev sde, sector 5245986552

This explains a lot. The associated *errno* for *BLK_STS_MEDIUM* is *ENODATA*. Thanks, Robert!

@Sage, do we want to alter our *allow_eio* policy? The kernel can set many other *E** treating *EIO* rather like a fallback when nothing suits better.

#9 - 03/20/2018 10:32 PM - Yuri Weinstein

- Related to Bug #23426: aio thread got No space left on device added

#10 - 04/09/2018 06:54 PM - Radoslaw Zarzynski

- Status changed from In Progress to Fix Under Review

PR: https://github.com/ceph/ceph/pull/21306.

#11 - 04/09/2018 06:54 PM - Radoslaw Zarzynski

- Status changed from Fix Under Review to 7

#12 - 04/11/2018 03:07 PM - Kefu Chai

- Status changed from 7 to Pending Backport

i believe this change should be backported.

#13 - 04/12/2018 01:38 AM - Nathan Cutler

- Backport set to luminous

#14 - 04/12/2018 01:39 AM - Nathan Cutler

- Copied to Backport #23672: luminous: bluestore: ENODATA on aio added

#15 - 05/23/2018 06:36 PM - Nathan Cutler

- Status changed from Pending Backport to Resolved

Files

ceph-osd.12.log.gz	306 KB	03/13/2018	Robert Sander