Ceph - Bug #735

Manual drive pull testing hangs filesystem

01/24/2011 12:31 PM - Brian Chrisman

Status:	Resolved	**% Done:**	0%
Priority:	High	**Spent time:**	0.00 hour
Assignee:	Colin McCabe		
Category:	OSD		
Target version:	v0.25		
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

It appears that drive failure problems in my configuration are not making their way up through the stack to kill off OSDs.
Setup:
- journal and data were two partition on same drive (for error isolation)
- 3 node cluster
- 4 SATA disks per node, cosd-per-disk config
- one partition from each disk in md (raid 1) root filesystem
- I/O generated continuously throughout testing
- kernel client running alongside daemons on all nodes
- running code is a 0.24rc with 2.6.37rc8 kernel
Symptoms:
- cosd of pulled drive reported journal errors on raw device journal
- md root filesystem recognized failure and responded properly
- cosd servicing pulled drive did not die and began inflating memory usage
- ceph filesystem unresponsive (waited >> 10 minutes for ls response on client)
- with same setup, if cosd is killed soon after drive pull, no problems at all
My Theory(ies):
- drive fail not being converted to cosd I/O error via btrfs, or I/O error ignored by cosd
- cosd memory inflation doesn't really matter, as cosd is expected to exit on error to allow re-peering

I can provide detailed hardware specs if it will help.

History

#1 - 01/25/2011 08:29 AM - Sage Weil

- Category set to OSD

- Priority changed from Normal to High

- Target version set to v0.25

Yep, this is a problem. The errors are causing btrfs operations to hang instead of return error codes.

What should the OSD do in this case? There should probably be a (long) timeout that will trigger a shutdown if the underlying fs becomes (very) unresponsive.

#2 - 01/25/2011 08:56 AM - Greg Farnum

Shouldn't btrfs be able to detect that the disk is gone and return appropriate error codes itself, rather than hanging?

#3 - 01/25/2011 12:57 PM - Brian Chrisman

I have a Quarch box in the lab that I was just pointed to. It has an ssh interface to power cycle drives for failure testing (little shims that sit between

the drive's sata/power and the system chassis, so I could automate some additional testing if we come up with a solution.
seems difficult to solve well without mandating an underlying filesystem. If btrfs is required for drive failure tolerance, then I imagine we can fix the btrfs handling of drive pull.

The only other way I can think of, is to have osds be informed of what actual devices they are using underneath, and then watching the sysfs entries for state changes on those drives. This would not detect/deal with filesystem level problems that aren't the result of a disk problem. For an integrated system/appliance, the sysfs monitoring would probably work. But for general use, something else needs to happen.

To keep the tradition of allowing any underlying filesystem, a load monitor/timeout should work. A large timeout would allow the system to become slow due to load without triggering an OSD exit. But it seems like a better solution would be to monitor activity and crash the osd only if there's no I/O completions **and** a request times out. I didn't notice the cosd's going into uninterruptible sleep (they were killable), so it'll probably work fairly cleanly.

#4 - 01/26/2011 12:50 PM - Colin McCabe

- Assignee set to Colin McCabe

#5 - 01/27/2011 10:23 AM - Colin McCabe

We need to be ready to handle unresponsive FileStores in general. Even if the underlying filesystem is 100% perfect (ha ha), the hardware itself may have problems which cause unresponsiveness. So we need to handle it in our code.

#6 - 01/31/2011 03:29 PM - Colin McCabe

- Status changed from New to 7

The ioctl timeout is now implemented in the ostimeo branch (2a266bd09d0db3b8d8c4f33a101229de1a4301a3)

#7 - 01/31/2011 04:23 PM - Sage Weil

merged by 6dc8994b750631c15e88553fd4fabdd9e4907989

#8 - 02/28/2011 09:23 AM - Colin McCabe

- Status changed from 7 to Resolved