Ceph - Bug #1152

Mon getting killed by OOM killer

06/09/2011 10:29 AM - Wido den Hollander

Status:	Closed	**% Done:**	0%
Priority:	Normal	**Spent time:**	4.00 hours
Assignee:	Greg Farnum		
Category:	Monitor		
Target version:	v0.31		
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

I've been seeing this for the last few weeks, my single mon keeps eating more and more memory until it reaches 4G Res and gets killed by the OOM-killer.

Somewhere in late April my cluster (40 OSD's) started to crash and wouldn't recover since then. I've been pretty persistent not to wipe it clean, so I'm still trying to start it up again and see what happends.

So my OSD's are bouncing up and down, but in this process the mon keeps eating more and more memory.

```
root@monitor:~# dmesg
[4941536.986482] cmon invoked oom-killer: gfp_mask=0x201da, order=0, oom_adj=0
[4941536.986487] cmon cpuset=/ mems_allowed=0
[4941536.986490] Pid: 18311, comm: cmon Tainted: P           2.6.32-30-server #59-Ubuntu
[4941536.986492] Call Trace:
[4941536.986501]  [<ffffffff810b483d>] ? cpuset_print_task_mems_allowed+0x9d/0xb0
[4941536.986506]  [<ffffffff810f8bc4>] oom_kill_process+0xd4/0x2f0
[4941536.986510]  [<ffffffff810f9180>] ? select_bad_process+0xd0/0x110
[4941536.986513]  [<ffffffff810f9218>] __out_of_memory+0x58/0xc0
[4941536.986516]  [<ffffffff810f93ae>] out_of_memory+0x12e/0x1a0
[4941536.986521]  [<ffffffff8155e2ce>] ? _spin_lock+0xe/0x20
[4941536.986524]  [<ffffffff810fc511>] __alloc_pages_slowpath+0x571/0x590
[4941536.986528]  [<ffffffff810fc6a1>] __alloc_pages_nodemask+0x171/0x180
[4941536.986533]  [<ffffffff8112f777>] alloc_pages_current+0x87/0xd0
[4941536.986536]  [<ffffffff810f6387>] __page_cache_alloc+0x67/0x70
[4941536.986539]  [<ffffffff810fff39>] __do_page_cache_readahead+0xc9/0x210
[4941536.986543]  [<ffffffff811000a1>] ra_submit+0x21/0x30
[4941536.986546]  [<ffffffff810f7c5e>] filemap_fault+0x3fe/0x450
[4941536.986550]  [<ffffffff81114534>] __do_fault+0x54/0x500
[4941536.986554]  [<ffffffff81117ae8>] handle_mm_fault+0x1a8/0x3c0
[4941536.986558]  [<ffffffff811621c0>] ? mntput_no_expire+0x30/0x110
[4941536.986562]  [<ffffffff81560e45>] do_page_fault+0x125/0x3b0
[4941536.986565]  [<ffffffff8155e795>] page_fault+0x25/0x30
[4941536.986567] Mem-Info:
[4941536.986569] Node 0 DMA per-cpu:
[4941536.986571] CPU    0: hi:    0, btch:   1 usd:   0
[4941536.986574] CPU    1: hi:    0, btch:   1 usd:   0
[4941536.986576] CPU    2: hi:    0, btch:   1 usd:   0
[4941536.986578] CPU    3: hi:    0, btch:   1 usd:   0
[4941536.986580] Node 0 DMA32 per-cpu:
[4941536.986583] CPU    0: hi:  186, btch:  31 usd:  30
[4941536.986585] CPU    1: hi:  186, btch:  31 usd:   7
[4941536.986587] CPU    2: hi:  186, btch:  31 usd:  33
[4941536.986589] CPU    3: hi:  186, btch:  31 usd: 141
[4941536.986591] Node 0 Normal per-cpu:
[4941536.986594] CPU    0: hi:  186, btch:  31 usd:  42
```

```
[4941536.986596] CPU    1: hi:  186, btch:  31 usd:   0
[4941536.986598] CPU    2: hi:  186, btch:  31 usd:  30
[4941536.986601] CPU    3: hi:  186, btch:  31 usd: 124
[4941536.986605] active_anon:756721 inactive_anon:218519 isolated_anon:0
[4941536.986607]  active_file:184 inactive_file:244 isolated_file:0
[4941536.986608]  unevictable:0 dirty:0 writeback:3 unstable:0
[4941536.986609]  free:6710 slab_reclaimable:1166 slab_unreclaimable:2883
[4941536.986610]  mapped:234 shmem:2 pagetables:2893 bounce:0
[4941536.986612] Node 0 DMA free:15856kB min:28kB low:32kB high:40kB active_anon:0kB inactive_anon
:0kB active_file:0kB inactive_file:0kB unevictable:0kB isolated(anon):0kB isolated(file):0kB prese
nt:15276kB mlocked:0kB dirty:0kB writeback:0kB mapped:0kB shmem:0kB slab_reclaimable:0kB slab_unre
claimable:0kB kernel_stack:0kB pagetables:0kB unstable:0kB bounce:0kB writeback_tmp:0kB pages_scan
ned:0 all_unreclaimable? yes
[4941536.986621] lowmem_reserve[]: 0 3254 4012 4012
[4941536.986625] Node 0 DMA32 free:9472kB min:6560kB low:8200kB high:9840kB active_anon:2690984kB
inactive_anon:538168kB active_file:456kB inactive_file:620kB unevictable:0kB isolated(anon):0kB is
olated(file):0kB present:3332768kB mlocked:0kB dirty:4kB writeback:0kB mapped:360kB shmem:0kB slab
_reclaimable:1732kB slab_unreclaimable:1956kB kernel_stack:208kB pagetables:7396kB unstable:0kB bo
unce:0kB writeback_tmp:0kB pages_scanned:374 all_unreclaimable? yes
[4941536.986635] lowmem_reserve[]: 0 0 757 757
[4941536.986639] Node 0 Normal free:1512kB min:1524kB low:1904kB high:2284kB active_anon:335900kB
inactive_anon:335908kB active_file:280kB inactive_file:356kB unevictable:0kB isolated(anon):0kB is
olated(file):0kB present:775680kB mlocked:0kB dirty:0kB writeback:12kB mapped:576kB shmem:8kB slab
_reclaimable:2932kB slab_unreclaimable:9576kB kernel_stack:1552kB pagetables:4176kB unstable:0kB b
ounce:0kB writeback_tmp:0kB pages_scanned:1044 all_unreclaimable? no
[4941536.986650] lowmem_reserve[]: 0 0 0 0
[4941536.986653] Node 0 DMA: 2*4kB 1*8kB 2*16kB 2*32kB 2*64kB 2*128kB 0*256kB 0*512kB 1*1024kB 1*2
048kB 3*4096kB = 15856kB
[4941536.986664] Node 0 DMA32: 1386*4kB 0*8kB 6*16kB 6*32kB 5*64kB 3*128kB 1*256kB 1*512kB 0*1024k
B 1*2048kB 0*4096kB = 9352kB
[4941536.986673] Node 0 Normal: 378*4kB 0*8kB 0*16kB 0*32kB 0*64kB 0*128kB 0*256kB 0*512kB 0*1024k
B 0*2048kB 0*4096kB = 1512kB
[4941536.986683] 3825 total pagecache pages
[4941536.986685] 3355 pages in swap cache
[4941536.986687] Swap cache stats: add 11662870, delete 11659515, find 1323001/2079551
[4941536.986689] Free swap  = 0kB
[4941536.986691] Total swap = 1052664kB
[4941536.999070] 1048560 pages RAM
[4941536.999073] 34555 pages reserved
[4941536.999075] 956 pages shared
[4941536.999078] 1006238 pages non-shared
[4941536.999082] Out of memory: kill process 18303 (cmon) score 6569 or a child
[4941537.011910] Killed process 18303 (cmon)
```

My mon is running with 'debug mon = 20', but I don't know if this gives enough information.

I tried the [[http://ceph.newdream.net/wiki/Memory_Profiling|memory profiling]] as described in the Wiki, but these commands do not work with a monitor, but since a few weeks it's linked to tcmalloc, so in theory it should work, right?

I don't know how to track this down, but it seems like something serious to me.

Let me know what to try to hunt this one down.

History

#1 - 06/09/2011 11:14 AM - Greg Farnum

I've created #1154 to make those commands work and hope to get it done today, but if you're feeling dedicated you should be able to get profiling data dumps by starting up the monitors with the proper environment variables set. :)
http://google-perftools.googlecode.com/svn/trunk/doc/heapprofile.html (see the sections "Running the Code" and "Modifying Runtime Behavior").

#2 - 06/09/2011 02:09 PM - Sage Weil

- Target version set to v0.30

#3 - 06/09/2011 04:38 PM - Greg Farnum

Pushed it, let me know if it doesn't work for you.

#4 - 06/10/2011 10:22 AM - Wido den Hollander

I've started my mon with memory profiling enabled, but it will take some hours before it starts eating memory.

At which point should I dump the memory? And where will it be dumped? In the clog?

#5 - 06/10/2011 11:42 AM - Greg Farnum

It'll dump a summary in the clog and try to dump the analysis data into a file named something like osd.1.0001.heap. If you've got log_dir defined it will go there, otherwise (due to a bug) it will try to dump it to your root dir.
Pushed a fix for that to master just now so it will go into cwd instead.
(So, set a log dir via inject_args if it's not set already, or pull down the newest code.)

Anyway, once you've got the profiler started I believe it should dump all on its own every so often.

#6 - 06/10/2011 01:31 PM - Wido den Hollander

- File mon_heap.tar.gz added

Attached are the heap dumps from the mon process.

The log is a big, 1.4GB, so I didn't upload it. It can be found from 'ssh root@logger.ceph.widodh.nl' and from there on 'ssh monitor', you'll find the logs in /var/log/ceph

Hope this helps!

#7 - 06/15/2011 09:54 AM - Sage Weil

There are no symbols.. can you run a cmon that's build with -g (not from a stripped .deb)?

#8 - 06/15/2011 12:31 PM - Wido den Hollander

I installed both ceph and ceph-dbg on the machine, I build the debs with dpkg-buildpackage and install those.

The debug symbols should be available.

#9 - 06/16/2011 10:36 AM - Greg Farnum

- Status changed from New to In Progress

- Assignee set to Greg Farnum

#10 - 06/16/2011 10:47 AM - Greg Farnum

Unfortunately, it looks like the heap dumps we have are from after it grew too large and they don't capture the growth so there's nothing useful there. :(We're going to have to wait until we can reproduce this with the heap statistics already enabled.

#11 - 06/17/2011 11:48 AM - Wido den Hollander

I hit the bug again it seems. The OOM killer came around and killed my monitor again. I didn't touch the machine yet, so the heap dumps should match the cmon binary now.

I started the profiler yesterday when everything was fine, so I guess the current dumps should have some interesting information.

Fyi, the dumps are on monitor.ceph.widodh.nl, available through logger.ceph.widodh.nl

#12 - 06/17/2011 01:45 PM - Greg Farnum

Humm, these heap dumps are never larger than 0.2MB!

Looking at the total memory allocation over the lifetime of the heap dumping it's ~70GB, nearly all of which is in an open_memstream function which we don't call ourselves. But basically this stuff looks the way we expect it to.

Can you reproduce this with syslogging off? The best I can come up with is that maybe it's generating logging data more quickly than it can send it over the wire, and so all your memory is getting devoted to sockets which are billed to the monitor.

#13 - 06/20/2011 12:06 PM - Sage Weil

- Target version changed from v0.30 to v0.31

#14 - 06/21/2011 02:37 PM - Sage Weil

- translation missing: en.field_story_points set to 5

- translation missing: en.field_position set to 1

- translation missing: en.field_position changed from 1 to 694

#15 - 06/22/2011 08:46 AM - Wido den Hollander

- Status changed from In Progress to Closed

I'm going to have close this one for now. By accident I screwed up the data directory of my only monitor, which is now refusing to start... I'll be forced mkcephfs my cluster and start all over again I think.

My 'find' was a bit to aggressive with removing and deleted every file older then 7 days in my monitor data dir...

If it pops up again I'll re-open.

Files

mon_heap.tar.gz	3.01 MB	06/10/2011	Wido den Hollander