Ceph - Feature #3775

log: stop logging in statfs reports usage above some threshold

01/09/2013 04:34 PM - Anonymous

Status:	New	**% Done:**	0%
Priority:	Normal	**Spent time:**	5.00 hours
Assignee:			
Category:			
Target version:			
Source:	Q/A	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**Pull request ID:**	

Description

Add a 'log stop on utilization = .95' option that will make the log code print one last line like

--- suspending logging because disk utilization X > X 'log stop on utilization' threshold ---

and throw out events (as far as disk goes) until it drops down again.

Related issues:

Related to Ceph - Documentation #3804: Logging section recommends fairly high...	**Resolved**	**01/15/2013**
Related to Ceph - Feature #3805: log: detect dup messages	**New**	**01/15/2013**

History

#1 - 01/09/2013 04:35 PM - Anonymous

Deb Barba <deb.barba@inktank.com>
3:13 PM (1 hour ago)

to Dan
so, as I explained in chat.
i am again seeing issues with my centos system. dont want to wipe and move forward, cause i am here to make trouble and file bugs...

if i should stop bugging you and just file the bug, that is ok too

I am still getting one machine runnign out of / space, so started investigating, since one machine and not all means it is suspicious.

I found that my 3rd node is spewing to the /var/log/ceph/ceph-osd* logs. so fast you can't read them.

I had the debug set to 5, and was writing about 1g every 15min... i think...

so i turned off 2 of the osd devices on this box (centos3) and left one running
truncated the two stopped osd log files so i would have room
then turned up the debugging to 20 on the remaining osd

when the debug was set to 5, i got this.
2013-01-09 14:36:18.875442 7fb71ffff700 1 -- 10.1.10.125:6800/2336 <== mon.0 10.1.10.123:6789/0 6840045 ==== auth_reply(proto 2 0 Success) v1 ==== 194+0+0 (4185112793 0 0) 0x7fb7102b8c10 con 0x5888070

2013-01-09 14:36:18.875530 7fb71ffff700 1 -- 10.1.10.125:6800/2336 --> 10.1.10.123:6789/0 -- auth(proto 2 2 bytes epoch 0) v1 -- ?+0 0x5626e90 con 0x5888070

2013-01-09 14:36:18.875925 7fb71ffff700 1 -- 10.1.10.125:6800/2336 <== mon.0 10.1.10.123:6789/0 6840046 ==== auth_reply(proto 2 0 Success) v1 ==== 194+0+0 (4185112793 0 0) 0x7fb71039fe40 con 0x5888070

2013-01-09 14:36:18.876083 7fb71ffff700 1 -- 10.1.10.125:6800/2336 --> 10.1.10.123:6789/0 -- auth(proto 2 2 bytes epoch 0) v1 -- ?+0 0x5626e90 con 0x5888070

2013-01-09 14:36:18.876276 7fb71ffff700 1 -- 10.1.10.125:6800/2336 <== mon.0 10.1.10.123:6789/0 6840047 ==== auth_reply(proto 2 0 Success) v1 ==== 194+0+0 (4185112793 0 0) 0x7fb710443ff0 con 0x5888070

2013-01-09 14:36:18.876365 7fb71ffff700 1 -- 10.1.10.125:6800/2336 --> 10.1.10.123:6789/0 -- auth(proto 2 2 bytes epoch 0) v1 -- ?+0 0x5626e90 con 0x5888070

Now that the debug is set to 20, i see this spew
013-01-09 14:44:40.822037 7f8f843f9700 5 filestore(/var/lib/ceph/osd/ceph-6) queue_transactions existing osr(1.271 0x25852f0)/0x25852f0
2013-01-09 14:44:40.822040 7f8f843f9700 5 filestore(/var/lib/ceph/osd/ceph-6) queue_transactions (writeahead) 29776 0x7f8f705a5740
2013-01-09 14:44:40.822042 7f8f843f9700 5 journal submit_entry seq 29776 len 855 (0x7f8f703d2d70)
2013-01-09 14:44:40.822049 7f8f843f9700 10 osd.6 pg_epoch: 167 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] handle_advance_map [4]/[4]
2013-01-09 14:44:40.822057 7f8f843f9700 10 osd.6 pg_epoch: 168 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] state<Started>: Started advmap
2013-01-09 14:44:40.822063 7f8f843f9700 10 osd.6 pg_epoch: 168 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] handle_advance_map [4]/[4]
2013-01-09 14:44:40.822068 7f8f843f9700 10 osd.6 pg_epoch: 169 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] state<Started>: Started advmap
2013-01-09 14:44:40.822074 7f8f843f9700 10 osd.6 pg_epoch: 169 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] handle_advance_map [4]/[4,6]
2013-01-09 14:44:40.822079 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] state<Started>: Started advmap
2013-01-09 14:44:40.822082 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] acting_up_affected newup [4] newacting [4,6]
2013-01-09 14:44:40.822086 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] state<Started>: up or acting affected, transitioning to Reset
2013-01-09 14:44:40.822091 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] exit Started/Stray 3.090115 4 0.000070
2013-01-09 14:44:40.822096 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] exit Started 3.090143 0 0.000000
2013-01-09 14:44:40.822101 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] enter Reset
2013-01-09 14:44:40.822104 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=163 pi=144-162/3 inactive NOTIFY] set_last_peering_reset 170
2013-01-09 14:44:40.822108 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=170 pi=144-162/3 inactive NOTIFY] state<Reset>: Reset advmap
2013-01-09 14:44:40.822112 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=170 pi=144-162/3 inactive NOTIFY] _calc_past_interval_range: already have past intervals back to 154
2013-01-09 14:44:40.822115 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=170 pi=144-162/3 inactive NOTIFY] acting_up_affected newup [4] newacting [4,6]
2013-01-09 14:44:40.822119 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=170 pi=144-162/3 inactive NOTIFY] state<Reset>: up or acting affected, calling start_peering_interval again
2013-01-09 14:44:40.822123 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4] r=-1 lpr=170 pi=144-162/3 inactive NOTIFY] set_last_peering_reset 170
2013-01-09 14:44:40.822128 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/163/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] noting past interval(163-169 [4]/[4] maybe_went_rw)
2013-01-09 14:44:40.822132 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] up [4] ~~> [4], acting [4] -> [4,6], role -1 -> 1~~
~~2013-01-09 14:44:40.822137 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] on_change~~
~~2013-01-09 14:44:40.822141 7f8f843f9700 15 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] requeue_ops~~
~~2013-01-09 14:44:40.822153 7f8f843f9700 15 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] update_stats~~ not primary
2013-01-09 14:44:40.822157 7f8f843f9700 15 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] requeue_ops
2013-01-09 14:44:40.822161 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] remove_watchers
2013-01-09 14:44:40.822165 7f8f843f9700 15 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] requeue_ops
2013-01-09 14:44:40.822169 7f8f843f9700 15 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] requeue_ops
2013-01-09 14:44:40.822173 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] exit NotTrimming
2013-01-09 14:44:40.822178 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] enter NotTrimming
2013-01-09 14:44:40.822181 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] on_role_change
2013-01-09 14:44:40.822185 7f8f843f9700 15 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] requeue_ops
2013-01-09 14:44:40.822189 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] cancel_recovery
2013-01-09 14:44:40.822192 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] clear_recovery_state
2013-01-09 14:44:40.822197 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] handle_activate_map
2013-01-09 14:44:40.822220 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] update_heartbeat_peers unchanged
2013-01-09 14:44:40.822225 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] take_waiters
2013-01-09 14:44:40.822229 7f8f843f9700 15 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1

lpr=170 pi=144-169/4 remapped NOTIFY] requeue_ops
2013-01-09 14:44:40.822233 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] exit Reset 0.000133 1 0.000119
2013-01-09 14:44:40.822238 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] enter Started
2013-01-09 14:44:40.822245 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] enter Start
2013-01-09 14:44:40.822249 7f8f843f9700 1 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] state<Start>: transitioning to Stray
2013-01-09 14:44:40.822255 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] exit Start 0.000010 0 0.000000
2013-01-09 14:44:40.822260 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] enter Started/Stray
2013-01-09 14:44:40.822265 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] handle_peering_event: epoch_sent: 163 epoch_requested: 163 MQuery from 4 query_epoch 163 query: query(info 0'0)

2013-01-09 14:44:40.822269 7f8f843f9700 10 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] old_peering_msg reply_epoch 163 query_epoch 163 last_peering_reset 170
2013-01-09 14:44:40.822280 7f8f843f9700 20 osd.6 pg_epoch: 170 pg[1.275(empty local-les=154 n=0 ec=1 les/c 154/154 163/170/84) [4]/[4,6] r=1 lpr=170 pi=144-169/4 remapped NOTIFY] write_info bigbl 656
2013-01-09 14:44:40.822286 7f8f843f9700 5 filestore(/var/lib/ceph/osd/ceph-6) queue_transactions existing osr(1.275 0x25bcc10)/0x25bcc10
2013-01-09 14:44:40.822289 7f8f843f9700 5 filestore(/var/lib/ceph/osd/ceph-6) queue_transactions (writeahead) 29777 0x7f8f705ad510
2013-01-09 14:44:40.822291 7f8f843f9700 5 journal submit_entry seq 29777 len 890 (0x7f8f70d025b0)

then it finally switched over to the earlier auth errors that i was first seeing....

#2 - 01/10/2013 06:32 AM - Sam Lang

The easiest solution for this might be to adjust the default logrotate script (src/logrotate.conf) to use the size parameter. I've set this up on ceph deployments in the past. In fact, that logrotate script could probably be made to use a size value on deployment that is 20% percent of the free space on the root fs. For customers though, we probably want to set that up explicitly. From the logrotate man page:

```
size size
              Log files are rotated when they grow bigger than size bytes. If size is followed by k, the size is assumed to  be  in
              kilobytes.  If the M is used, the size is in megabytes, and if G is used, the size is in gigabytes. So size 100, size
              100k, size 100M and size 100G are all valid.
```

#3 - 01/10/2013 08:34 AM - Anonymous

Sam,

That is a cool idea. I will open a doc bug for that. Providing instructions for those with smaller root drives.

#4 - 01/10/2013 09:35 AM - Sage Weil

- Priority changed from Normal to High

#5 - 01/10/2013 12:50 PM - Ian Colle

- Assignee set to Samuel Just

#6 - 01/15/2013 10:35 AM - Samuel Just

- Assignee changed from Samuel Just to Dan Mick

#7 - 01/15/2013 11:39 AM - Dan Mick

So a couple of thoughts:

1) changing size in logrotate.conf doesn't help unless we also change frequency
2) with a small enough log partition, any logging is too much
3) Ceph logs can produce a lot of data. One has to be aware of that.

I'm not sure what compensatory measures we could or should take; duplication is rare enough that I don't think duplicate hysteresis would help much in the general case.

We could look for "failure to write log" and disable the log for...a few minutes?...which could lead to admins thinking the problem is gone and then logging starting again and filling the disk again. Or we could disable permanently, and require the admin to restart the daemon.

What other things could/should we do about this?

#8 - 01/15/2013 12:03 PM - Anonymous

so, a couple ideas of what can be done.

if we do set size and frequency (or inform the user how to), then it could be set to roll all copies off so that it never fills the disk

have a check that will send notification after a certain full, such as 90. Or inform the user that they need to set this up if they have a root disk smaller than Xsize.

I also really like the idea of grouping duplicate messages.
my mon.a log is from 2013-01-14 11:41:51.178312 to 2013-01-15 11:58:21.679958. about 24 hours. and in that time, I have 419884 messages about sessions. yes, my debug is set to 20, but my mds debug is set to 5, and i have 198092 messages in the 24 hr period. Very verbose. very repetitive.

looking at mon output

1. grep "ms_dispatch existing session MonSession" ceph-mon.a.log | awk '{print $5 $6 $7 $8 $9}'| wc -l
   419884
2. grep "ms_dispatch existing session MonSession" ceph-mon.a.log | awk '{print $5 $6 $7 $8 $9}'| grep electing | wc -l
   5
3. grep "ms_dispatch existing session MonSession" ceph-mon.a.log | awk '{print $5 $6 $7 $8 $9}'| grep probing | wc -l
   1
4. grep "ms_dispatch existing session MonSession" ceph-mon.a.log | awk '{print $5 $6 $7 $8 $9}'| grep leader | wc -l
   419590

This may be all documentation updates, rather than code changes.

but users need to be aware of it.

#9 - 01/15/2013 01:59 PM - Dan Mick

- Status changed from New to Need More Info

So I suggest we split this into two issues:

1) the documentation examples show an awfully-high logging value for some fairly-weird subsystems, and don't stress just how fast you can generate log messages as much as they could. I'll file a separate issue to clarify those two points.

2) there may be some specific messages that are pretty useless to repeat as often as we
do; if we can identify some of those, perhaps they can be changed to be less verbose/less
insistent.

The messages quoted above are from "debug mon = 20", which in my experience is pretty high, so may be appropriate at that level. A quick search shows \4 messages enabled by that level,
as compared to 97 for level 10, and 15 for level 1, 20 for 0 (i.e. always).

```
dout(20) << "get_global_paxos_version " << global_version << dendl;
    dout(20) << "have connection" << dendl;
      dout(20) << "ms_dispatch existing session " << s << " for " << s->inst << dendl;
    dout(20) << " caps " << s->caps.get_str() << dendl;
```

#10 - 01/15/2013 02:16 PM - Sage Weil

I agree. If there are lots of log messages at the default levels, that is the problem. I don't think there is much we can or should do if the admin cranks up logging to keep them from shooting themselves in teh foot.

Detecting dup messages would be nice. I'll open a feature request for that.

#11 - 02/22/2013 12:14 PM - Ian Colle

- Tracker changed from Bug to Feature

#12 - 02/27/2013 12:58 PM - Sage Weil

- Subject changed from ceph should not fill the root partition to 100% /var/log/ceph.... to log: stop logging in statfs reports usage above some threshold

- Description updated

- Status changed from Need More Info to 12

- Assignee deleted (Dan Mick)

- Priority changed from High to Normal

#13 - 12/05/2019 09:34 PM - Patrick Donnelly

- Status changed from 12 to New