fs - Bug #13777

Ceph file system is not freeing space

11/12/2015 02:50 AM - Eric Eastman

Status:	Resolved	**Start date:**	11/12/2015
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Source:	Community (user)	**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:	infernalis	**Component(FS):**	MDS
Regression:	No	**Labels (FS):**	
Severity:	3 - minor	**Pull request ID:**	
Reviewed:			

Description

I have a Ceph file system that is not freeing space. Using Ceph 9.1.0 I created a file system with snapshots enabled, filled up the file system over days while taking snapshots hourly. I then deleted all files and all snapshots, but Ceph is not returning the space. I left the cluster sit for two days to see if the cleanup process was being done in the background and it still has not freed the space. I tried rebooting the cluster and clients and the space is still not returned.

```
The file system was created with the command:
# ceph fs new cephfs cephfs_metadata cephfs_data

# getfattr -d -m ceph.dir.* /cephfs/
getfattr: Removing leading '/' from absolute path names
# file: cephfs/
ceph.dir.entries="0"
ceph.dir.files="0"
ceph.dir.rbytes="0"
ceph.dir.rctime="1447033469.0920991041"
ceph.dir.rentries="4"
ceph.dir.rfiles="1"
ceph.dir.rsubdirs="3"
ceph.dir.subdirs="0"

ls -l /cephfs/
total 0

# ls -l /cephfs/.snap
total 0

# grep ceph /proc/mounts
ceph-fuse /cephfs fuse.ceph-fuse rw,noatime,user_id=0,group_id=0,default_permissions,allow_other 0 0

# df /cephfs/
Filesystem     1K-blocks      Used Available Use% Mounted on
ceph-fuse      276090880 194162688  81928192  71% /cephfs

# df -i /cephfs/
Filesystem      Inodes IUsed IFree IUse% Mounted on
ceph-fuse      2501946     -     -     - /cephfs

# ceph df detail
GLOBAL:
    SIZE     AVAIL      RAW USED     %RAW USED      OBJECTS
    263G     80009M         181G         68.78        2443k
```

```
POOLS:
    NAME                ID     CATEGORY     USED         %USED      MAX AVAIL     OBJECTS     DIRTY
   READ      WRITE
    rbd                 0      -                   0           0       27826M           0         0
      0           0
    cephfs_data         1      -              76846M       28.50       27826M     2501672     2443k
   345k      32797k
    cephfs_metadata     2      -              34868k        0.01       27826M         259       259
   480k      23327k
    kSAFEbackup         3      -                108M        0.04       27826M          15        15
      0          49

Dumping the statistics shows lots of strays:
  "mds_cache": {
        "num_strays": 16389,
        "num_strays_purging": 0,
        "num_strays_delayed": 0,
        "num_purge_ops": 0,
        "strays_created": 17066,
        "strays_purged": 677,
        "strays_reintegrated": 0,
        "strays_migrated": 0,
        "num_recovering_processing": 0,
        "num_recovering_enqueued": 0,
        "num_recovering_prioritized": 0,
        "recovery_started": 0,
        "recovery_completed": 0
    },

The whole cluster and client systems are running on Trusty with a 4.3.0 kernel and Ceph version 9.
1.0
# ceph -v
ceph version 9.1.0 (3be81ae6cf17fcf689cd6f187c4615249fea4f61)
# uname -a
Linux ede-c2-adm01 4.3.0-040300-generic #201511020949 SMP Mon Nov 2 14:50:44 UTC 2015 x86_64 x86_6
4 x86_64 GNU/Linux
```

I am attaching the output of ceph mds tell * dumpcache /tmp/dumpcache.txt and the MDS log, with debug mds = 20, from startup to when the MDS went active.

Additional information is in the following list posts:
http://thread.gmane.org/gmane.comp.file-systems.ceph.user/25212

Related issues:

Related to fs - Bug #13782: Snapshotted files not properly purged	**Resolved**	**11/12/2015**
Copied to fs - Backport #14067: infernalis : Ceph file system is not freeing ...	**Resolved**	

Associated revisions

Revision 460c74a0 - 11/12/2015 01:57 PM - Yan, Zheng

mds: properly set STATE_STRAY/STATE_ORPHAN for stray dentry/inode

Fixes: #13777
Signed-off-by: Yan, Zheng <zyan@redhat.com>

Revision 0b474c52 - 11/27/2015 04:28 PM - John Spray

mon: don't require OSD W for MRemoveSnaps

Use ability to execute "osd pool rmsnap" command
as a signal that the client should be permitted
to send MRemoveSnaps too.

Note that we don't also require the W ability,
unlike Monitor::_allowed_command -- this is slightly
more permissive handling, but anyone crafting caps
that explicitly permit "osd pool rmsnap" needs to
know what they are doing.

Fixes: #13777
Signed-off-by: John Spray <john.spray@redhat.com>

Revision 5f54671e - 01/29/2016 10:19 AM - John Spray

mon: don't require OSD W for MRemoveSnaps

Use ability to execute "osd pool rmsnap" command
as a signal that the client should be permitted
to send MRemoveSnaps too.

Note that we don't also require the W ability,
unlike Monitor::_allowed_command -- this is slightly
more permissive handling, but anyone crafting caps
that explicitly permit "osd pool rmsnap" needs to
know what they are doing.

Fixes: #13777
Signed-off-by: John Spray <john.spray@redhat.com>
(cherry picked from commit 0b474c52abd3d528c041544f73b1d27d7d1b1320)

Revision 29d30ecd - 01/29/2016 10:22 AM - Yan, Zheng

mds: properly set STATE_STRAY/STATE_ORPHAN for stray dentry/inode

Fixes: #13777
Signed-off-by: Yan, Zheng <zyan@redhat.com>
(cherry picked from commit 460c74a0b872336a7279f0b40b17ed672b6e15a1)

History

#1 - 11/12/2015 02:59 AM - Eric Eastman

I noticed the MDS log file did not get attached. It is about 7MB, so maybe too big to upload? You can grab it at:

wget ftp://ftp.keepertech.com/outgoing/eric/ceph-mds.ede-c2-mds03.log.debug-20.bz2

#2 - 11/12/2015 12:51 PM - John Spray

So it looks like some portion of the strays are getting purged during this log, but you mentioned that there were many thousands of strays, and only a few hundred purges are happening here before it goes quiet.

I suspect there is some dependency between strays that we are failing to take account of: something is initially skipped because it is referenced by another stray, but when that other stray is purged we aren't coming back and kicking off the purge of the first guy.

It's probably worth trying restarting the MDS a few times and watching the perf stats to see it purge some portion of the strays each time.

however when I went a wrote a new test for strays vs. snapshots it seems there is more broken here, so you may find that even when the strays are purged you still have lingering trash in the data pool from the snapshots, if the files were updated since being snapped (see http://tracker.ceph.com/issues/13782)

#3 - 11/12/2015 12:52 PM - John Spray

- Related to Bug #13782: Snapshotted files not properly purged added

#4 - 11/12/2015 12:52 PM - John Spray

- Category set to 47

#5 - 11/12/2015 01:33 PM - Zheng Yan

- Status changed from New to Verified

seems like we forget to set CDentry::STATE_STRAY/CInode::STATE_ORPHAN for stray dentry/inode.

Eric Eastman, could your try running "ceph daemon mds.xxx flush journal", then restart the MDS

#6 - 11/12/2015 02:00 PM - Zheng Yan

- Status changed from Verified to Need Review

https://github.com/ceph/ceph/pull/6553

#7 - 11/12/2015 03:23 PM - Eric Eastman

- File dumpcache.2 added

- File perf.2 added

I ran the command **ceph daemon mds.ede-c2-mds03 flush journal** and restarted the MDS. It did not seem to free up any space on the mounted file system, but did clean up the strays and shortened the dumpcache output. I was running default logging and here the log and other outputs:

```

# ceph daemon mds.ede-c2-mds03  flush journal
{
    "message": "",
    "return_code": 0
}

# restart ceph-all

# cat ceph-mds.ede-c2-mds03.log
2015-11-12 09:39:00.333966 7f5205448700  1 mds.ede-c2-mds03 asok_command: flush journal (starting...)
2015-11-12 09:39:01.113283 7f5205448700  1 mds.ede-c2-mds03 asok_command: flush journal (complete)
2015-11-12 09:39:10.787832 7f520033d700 -1 mds.ede-c2-mds03 *** got signal Terminated ***
2015-11-12 09:39:10.787890 7f520033d700  1 mds.ede-c2-mds03 suicide.  wanted state up:active
2015-11-12 09:39:10.790872 7f520033d700  1 mds.0.11 shutdown: shutting down rank 0
2015-11-12 09:39:10.932678 7f7a24cce800  0 ceph version 9.1.0 (3be81ae6cf17fcf689cd6f187c4615249fea4f61), process ceph-mds, pid 10380
2015-11-12 09:39:15.826816 7f7a1da15700  1 mds.ede-c2-mds03 handle_mds_map standby
2015-11-12 09:39:15.830436 7f7a1da15700  1 mds.0.12 handle_mds_map i am now mds.0.12
2015-11-12 09:39:15.830452 7f7a1da15700  1 mds.0.12 handle_mds_map state change up:boot --> up:replay
```

```
2015-11-12 09:39:15.830474 7f7a1da15700  1 mds.0.12 replay_start
2015-11-12 09:39:15.830482 7f7a1da15700  1 mds.0.12  recovery set is
2015-11-12 09:39:15.838384 7f7a18708700  0 mds.0.cache creating system inode with ino:100
2015-11-12 09:39:15.838764 7f7a18708700  0 mds.0.cache creating system inode with ino:1
2015-11-12 09:39:15.843760 7f7a16afa700  1 mds.0.12 replay_done
2015-11-12 09:39:15.843775 7f7a16afa700  1 mds.0.12 making mds journal writeable
2015-11-12 09:39:16.831820 7f7a1da15700  1 mds.0.12 handle_mds_map i am now mds.0.12
2015-11-12 09:39:16.831828 7f7a1da15700  1 mds.0.12 handle_mds_map state change up:replay --> up:reconnect
2015-11-12 09:39:16.831846 7f7a1da15700  1 mds.0.12 reconnect_start
2015-11-12 09:39:16.831849 7f7a1da15700  1 mds.0.12 reopen_log
2015-11-12 09:39:16.831855 7f7a1da15700  1 mds.0.server reconnect_clients -- 1 sessions
2015-11-12 09:39:16.833059 7f7a162f9700  1 mds.client.cephfs ms_verify_authorizer: cannot decode auth caps bl 
of length 0
2015-11-12 09:39:16.833792 7f7a1da15700  0 log_channel(cluster) log [DBG] : reconnect by client.134402 10.15.2
.120:0/3437754152 after 0.001832
2015-11-12 09:39:16.833894 7f7a1da15700  1 mds.0.12 reconnect_done
2015-11-12 09:39:17.836633 7f7a1da15700  1 mds.0.12 handle_mds_map i am now mds.0.12
2015-11-12 09:39:17.836640 7f7a1da15700  1 mds.0.12 handle_mds_map state change up:reconnect --> up:rejoin
2015-11-12 09:39:17.836657 7f7a1da15700  1 mds.0.12 rejoin_start
2015-11-12 09:39:17.836690 7f7a1da15700  1 mds.0.12 rejoin_joint_start
2015-11-12 09:39:17.836768 7f7a1da15700  1 mds.0.12 rejoin_done
2015-11-12 09:39:18.838381 7f7a1da15700  1 mds.0.12 handle_mds_map i am now mds.0.12
2015-11-12 09:39:18.838389 7f7a1da15700  1 mds.0.12 handle_mds_map state change up:rejoin --> up:active
2015-11-12 09:39:18.838404 7f7a1da15700  1 mds.0.12 recovery_done -- successful recovery!
2015-11-12 09:39:18.838660 7f7a1da15700  1 mds.0.12 active_start
2015-11-12 09:39:18.838813 7f7a1da15700  1 mds.0.12 cluster recovered.
```

```
# ceph daemon mds.ede-c2-mds03 perf dump | grep stray
        "num_strays": 0,
        "num_strays_purging": 0,
        "num_strays_delayed": 0,
        "strays_created": 0,
        "strays_purged": 0,
        "strays_reintegrated": 0,
        "strays_migrated": 0,
```

```
# df /cephfs/
Filesystem     1K-blocks      Used Available Use% Mounted on
ceph-fuse      276090880 193851392  82239488  71% /cephfs
```

```
# df -i /cephfs/
Filesystem      Inodes IUsed IFree IUse% Mounted on
ceph-fuse      2501709     -     -     - /cephfs
```

```
# ceph df detail
GLOBAL:
    SIZE     AVAIL      RAW USED     %RAW USED     OBJECTS 
    263G     80312M         180G         68.67       2443k 
POOLS:
    NAME                ID     CATEGORY     USED       %USED     MAX AVAIL     OBJECTS     DIRTY     READ      
WRITE  
    rbd                 0      -                 0         0        28021M           0         0         0     
     0 
    cephfs_data         1      -            76846M     28.50        28021M     2501672     2443k      345k     
32797k 
    cephfs_metadata     2      -             2637k         0        28021M          20        20      481k     
23329k 
    kSAFEbackup         3      -              129M      0.05        28021M          17        17         0     
    58 
```

I tried unmounting and remounting the file system a couple times and still no change in used space. I tried running additional flush journals and restarting the MDS and that did not free up any space. I have attached the new dumpcache output as dumpcache.2 and the perf output as perf.2.

#8 - 11/13/2015 04:01 AM - Zheng Yan

It seems MDS has deleted all object? what is the output of "rados -p cephfs_data ls". If there are leftover objects. please check if these objects have 'parent' xattr and if these objects have snaps.

#9 - 11/13/2015 04:21 AM - Eric Eastman

The output of "rados -p cephfs_data ls" returns no objects. I also ran a rados df:

```
# rados -p cephfs_data ls
# rados df
pool name                 KB      objects       clones     degraded      unfound           rd        rd KB
       wr        wr KB
cephfs_data         78691119      2501672      2501672            0            0       353998       531850
 33584853    226152412
cephfs_metadata         2639           20            0            0            0       492642      8950254
 23889220    131338233
kSAFEbackup           132228           17            0            0            0            0            0
       58       132359
rbd                        0            0            0            0            0            0            0
        0            0
  total used       189579304      2501709
  total avail       82247384
  total space      276093280
```

#10 - 11/13/2015 07:57 AM - Zheng Yan

```
2015-11-11 21:10:40.949981 7f3f8ea44700 10 mds.0.snap check_osd_map need_to_purge={1=2,3,4,5,6,7,8,9,a,b,c,d,e
,f,10,11,12,13,14,15,16,17,18,19,1a,1b,1c,1d,1e,1f,20,21,22,23,24,25,26,27,28,29,2a,2b,2c,2d,2e,2f,30,31,32,33
,34,35,36,37,38,39,3a,3b,3c,3d,3e,3f,40,41,42,43,44,45,46,47,48,49,4a,4b,4c,4d,4f,50,51,52,53,54,55,56,57,58,5
9,5a,5b,5c,5d,5e,5f,60,61,62,63,64,65,66,67,68,69,6a,6b,6c,6d,6e,6f,70,71,72,73,74,75,76,77,78,79,7a,7b,7c,7d,
7e,7f,80,81,82,83,84,85,86,87,88,89,8a,8b,8c,8d,8e,8f,90,91,92,93,94,95,96,97,98,99,9a}
2015-11-11 21:10:40.950064 7f3f8ea44700 10 mds.0.snap requesting removal of {1=[2,3,4,5,6,7,8,9,a,b,c,d,e,f,10
,11,12,13,14,15,16,17,18,19,1a,1b,1c,1d,1e,1f,20,21,22,23,24,25,26,27,28,29,2a,2b,2c,2d,2e,2f,30,31,32,33,34,3
5,36,37,38,39,3a,3b,3c,3d,3e,3f,40,41,42,43,44,45,46,47,48,49,4a,4b,4c,4d,4f,50,51,52,53,54,55,56,57,58,59,5a,
5b,5c,5d,5e,5f,60,61,62,63,64,65,66,67,68,69,6a,6b,6c,6d,6e,6f,70,71,72,73,74,75,76,77,78,79,7a,7b,7c,7d,7e,7f
,80,81,82,83,84,85,86,87,88,89,8a,8b,8c,8d,8e,8f,90,91,92,93,94,95,96,97,98,99,9a]}
```

It seems no snap data were purged. please check if MDS has write capability to moniter. If MDS has, please set debug_osd=10 and set debug_mon=10, restart monitor and osds. check lines contain 'snap_trimq now' in osd log, and check lines contain 'preprocess_remove_snaps' in mon log.

#11 - 11/14/2015 03:47 AM - Eric Eastman

Here is how the MDS caps looked:

```
mds.ede-c2-mds03
    key: AQBcRjpWlOCgDxAAyvH2o4+BXUom50+D7ZvI2w==
    caps: [mds] allow
    caps: [mon] allow profile mds
    caps: [osd] allow rwx
```

Running with debug_mon=10 and grep for preprocess_remove_snaps on the monitor showed:

```
/var/log/ceph/ceph-mon.ede-c2-mon03.log:2015-11-13 18:43:34.219148 7fbee3c23700  0 mon.ede-c2-mon03@2(peon).osd e366 got preprocess_remove_snaps from entity with insufficient caps allow profile mds
```

Changing the caps to:

```
mds.ede-c2-mds03
    key: AQBcRjpWlOCgDxAAyvH2o4+BXUom50+D7ZvI2w==
    caps: [mds] allow
    caps: [mon] allow rwx
    caps: [osd] allow rwx
```

Restarting the MDS, MON and OSD started freeing space.

Looks like “mon ‘allow profile mds’” is not open enough. I think I got the information for using “mon 'allow profile mds’” from the ansible file at:

https://github.com/ceph/ceph-ansible/blob/master/roles/ceph-mds/tasks/pre_requisite.yml

Which shows:

```
- name: create mds keyring
  command: ceph --cluster ceph --name client.bootstrap-mds --keyring /var/lib/ceph/bootstrap-mds/ceph.keyring auth get-or-create mds.{{ ansible_hostname }} osd 'allow rwx' mds 'allow' mon 'allow profile mds' -o /var/lib/ceph/mds/ceph-{{ ansible_hostname }}/keyring
  args:
    creates: /var/lib/ceph/mds/ceph-{{ ansible_hostname }}/keyring
  changed_when: false
  when: cephx
```

#12 - 11/16/2015 11:43 AM - John Spray

Patch here for the capabilities issue:
https://github.com/ceph/ceph/pull/6601

#13 - 11/26/2015 08:45 PM - Eric Eastman

Can both pull requests in this ticket be back ported to the next Infernalis release?
Thanks
Eric

#14 - 11/26/2015 09:28 PM - Nathan Cutler

- Backport set to infernalis

#15 - 12/12/2015 01:51 PM - Abhishek Varshney

- Status changed from Need Review to Pending Backport

#16 - 12/12/2015 01:51 PM - Abhishek Varshney

- Copied to Backport #14067: infernalis : Ceph file system is not freeing space added

#17 - 02/11/2016 04:37 AM - Loic Dachary

- Status changed from Pending Backport to Resolved

#18 - 07/13/2016 05:54 AM - Greg Farnum

- Component(FS) MDS added

Files

dumpcache.txt.bz2	111 KB	11/12/2015	Eric Eastman
dumpcache.2	7.2 KB	11/12/2015	Eric Eastman
perf.2	4.97 KB	11/12/2015	Eric Eastman