Ceph - Bug #12231

crush unable to generate 3 osds in teuthology run

07/07/2015 11:42 PM - Samuel Just

Status:	Resolved	**% Done:**	0%
Priority:	Urgent	**Spent time:**	0.00 hour
Assignee:	Samuel Just		
Category:			
Target version:			
Source:	other	**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:		**Pull request ID:**	
Regression:	No	**Crash signature (v1):**	
Severity:	3 - minor	**Crash signature (v2):**	
Reviewed:			

Description

Lately, wip-sam-testing (basically master) runs are reliably turning up a case 1 or 2 times per run where 3/6 osds are out and crush is unable to turn up more than 2 of the remaining 3 osds for at least one pg. I grabbed one of the osdmaps and found that on this one, the bad pg is pg 1.37

/home/sam/git-checkouts/ceph4/src/osdmaptool: osdmap file '/tmp/osdmap'
parsed '1.37' -> 1.37
1.37 raw ([4,1], p4) up ([1,4], p1) acting ([1,4], p1)

/home/sam/git-checkouts/ceph4/src/osdmaptool: osdmap file '/tmp/osdmap'
parsed '1.36' -> 1.36
1.36 raw ([4,1,3], p4) up ([4,1,3], p4) acting ([4,1,3], p4)

hashes (attached) has the draws for r=0 through 999999 on that pg and you'll see that indeed osd 3 does not win for the first time until between draws 50 and 60.

I see nothing new with the crush tunables. osdmaptool compiled on firefly agrees with the output, so it's not a change in crush. The straw weights appear to be 65535, so there is nothing wonky with the crush map construction. The two questions are:
1) Is this simply an indication that the hash is really bad and we need to begin switching it (possibly before jewel)?
2) Why has this not come up before? We started testing regularly with size 3 pools in teuthology in February. I haven't seen it yet in hammer runs either. Odd.

Associated revisions

Revision 042bd117 - 07/10/2015 12:03 AM - Samuel Just

3-size-2-min-size: keep 4 in during thrashing

Workaround for 12231.

Fixes: #12231
Signed-off-by: Samuel Just <sjust@redhat.com>

History

#1 - 07/07/2015 11:51 PM - Samuel Just

ubuntu@teuthology:/a/samuelj-2015-07-06_17:07:54-rados-wip-sam-testing-distro-basic-multi/963089

is the instance above.

#2 - 07/07/2015 11:53 PM - Samuel Just

- File osdmap2 added

For the other instance in that run (ubuntu@teuthology:/a/samuelj-2015-07-06_17:07:54-rados-wip-sam-testing-distro-basic-multi/962909)

we have a different set of 3 in osds, but still including 3. Thus, pg 1.37 once again has trouble:

~/git-checkouts/ceph4/src/osdmaptool --test-map-pg 1.37 /tmp/osdmap2
/home/sam/git-checkouts/ceph4/src/osdmaptool: osdmap file '/tmp/osdmap2'
parsed '1.37' -> 1.37
1.37 raw ([1,2], p1) up ([1,2], p1) acting ([1,2], p1)

Debugging has the same value for x.

#3 - 07/08/2015 12:06 AM - Samuel Just

I checked a few similar hammer runs and noticed that pgp_num didn't get high enough for 1.37 to have its own seed. Perhaps something in master is causing us to split more/faster in a single run?

#4 - 07/14/2015 02:24 PM - Samuel Just

- Status changed from New to Resolved

Updated ceph-qa-suite to keep 4 in for size 3 pools.

Files

osdmap	6.5 KB	07/07/2015	Samuel Just
osdmaptool_debug	61.9 KB	07/07/2015	Samuel Just
test_jenkins.c	462 Bytes	07/07/2015	Samuel Just
hashes	649 KB	07/07/2015	Samuel Just
osdmap2	5.38 KB	07/07/2015	Samuel Just