rbd - Fix #4429

libceph: support a list of data items in a message

03/12/2013 02:57 PM - Alex Elder

Status:	Duplicate	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Alex Elder		
Category:			
Target version:	v0.61 - Cuttlefish		
Source:	Development	**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:		**Pull request ID:**	
Reviewed:		**Crash signature:**	

Description

This will require a number of patches to complete, but
it's a logical unit of work.

An abstracted data item is now used to represent a block
of data, either a list of pages, a page array, or a bio
list. What we will want for multiple ops is to allow
more than one such data item to be associated with a
message. That way, an osd request can build up a single
request message containing an array of ops, and each of
those ops can independently append a data item to be used
for outgoing or incoming data, in whatever form is required
(page list, bio, etc.).

This depends on completion of the work to abstract the
data item, and to use it for both read and write requests.

The general sequence I intend to follow is:
- collapse the separate structures representing data in
a message (now named "p", "l", and "b" for a page array,
page list, and bio list) into a single item. It turns
out only one of these is ever (currently) used for any
message (though that wasn't obvious before). They'll
all be represented by a structure "data" that will be
handled by the messenger according to its representation.
- Replace the single "data" structure with a pointer to such
a structure. Allocate that structure dynamically only
when a message will have a data portion.
- Implement a list of these structures, with the "data"
field in the message tracking the head (and tail) of
data items on that list.
- Have the callers that now "set" the data field instead
"add" it to the existing data in a message.
- Use this capability in the osd client for multiple ops
that carry data. (This may just be a contrived test
for now.)

This is related to http://tracker.ceph.com/issues/3761.

Related issues:

Related to rbd - Feature #3761: kernel messenger: need to support multiple op...	**Resolved**	**03/09/2013**

History

#1 - 03/12/2013 03:12 PM - Ian Colle

- Project changed from Ceph to rbd

- Target version set to v0.60

#2 - 03/12/2013 03:13 PM - Ian Colle

- Tracker changed from Bug to Fix

#3 - 03/12/2013 06:18 PM - Alex Elder

- Status changed from In Progress to Fix Under Review

The following patches have been posted for review:
[PATCH] libceph: collapse all data items into one
[PATCH] libceph: make message data be a pointer

#4 - 03/15/2013 11:39 AM - Ian Colle

- Target version changed from v0.60 to v0.61 - Cuttlefish

#5 - 03/26/2013 03:58 PM - Alex Elder

(The following really applies to this issue as well
as 4428, 4427, and 4426.)

The patch(es) for this were posted and reviewed by Josh.
Just before I went on vacation I was doing some final
testing of this before committing it and ran into some
strange crashes involving apparmor. Not having time
to investigate, I left things as they were.

Today I've updated the patches so they're now based on
the latest "testing" branch and have been testing with
that result. At this point I have seen no errors, but
I'm going to let my current testing run to completion,
and may run it through some additional tests.

If I see another crash while doing this I'll try to
get to the bottom of it. It may be a memory leak
or something (because I can't think of how apparmor
would be involved otherwise).

If I see no crash, I think I'll commit these changes
and hope for the best...

#6 - 03/27/2013 06:00 AM - Alex Elder

This ran these tests, along with a "full" xfstests run
overnight with no problems:
- kernel_untar_build.sh
- misc/trivial_sync.sh

- rbd/map-unmap.sh
- rbd/kernel.sh
- suites/blogbench.sh
- suites/dbench.sh
- suites/tiobench.sh
- suites/fsstress.sh

I'm starting the same tests up again, iterating xfstests 3 times and reducing the injected socket frequency by a factor of 10 (to get the testing done more quickly).

I'm prepared to check all this in but I'll wait to discuss it at the stand up meeting (by which point much of the above retesting will have been done).

#7 - 03/27/2013 07:05 AM - Alex Elder

I hit the problem described here http://tracker.ceph.com/issues/4450 while testing this. So before committing these changes I need to try to track down and fix the root cause of that problem.

#8 - 03/28/2013 07:55 PM - Alex Elder

- Status changed from Fix Under Review to Duplicate

We decided that there was no benefit to having both this asnd 3761, so I'm marking this as a duplicate.

http://tracker.ceph.com/issues/3761

#9 - 03/29/2013 03:19 PM - Alex Elder

Adding a final note that the following has been committed to the ceph-client "testing" branch:

c658410 libceph: make message data be a pointer

I mention it here because that commit refers back here...