rgw - Bug #23207

rgw: inefficient buffer usage for PUTs

03/04/2018 06:29 AM - Marcus Watts

Status:	Resolved	**Start date:**	03/04/2018
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:	luminous, jewel	**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor		

Description

At least in jewel using swift, radosgw is very inefficient about its buffering process. In RGWPutObj_ObjStore::get_data(), it allocates a buffer usually of size rgw_max_chunk_size, issues a single "mg_read()" call (which might fill about 1% of the buffer), then passes the result along. Eventually a list of buffers is condensed and given to write, but meanwhile there is a lot of wasted memory. With a chunk size of 1mb and a workload that wrote and deleted 262 objects using 64 parallel threads where no object was > 33m (avg size 7m), I was finding a peak allocation of 3.7g of ram, of which 99% was never being used.

I have an experimental patch that fixes this behavior. With that patch, using the same workload, I am now usually filling 1m buffers, the buffers do not need to be repacked to write them, and peak allocation went down to 138m of ram.

Related issues:

Related to rgw - Bug #23596: mg_read() call has wrong arguments	**Resolved**	**04/08/2018**
Copied to rgw - Backport #23347: luminous: rgw: inefficient buffer usage for ...	**Resolved**	
Copied to rgw - Backport #23348: jewel: rgw: inefficient buffer usage for PUTs	**Resolved**	

History

#1 - 03/04/2018 06:14 PM - Vikhyat Umrao

- Status changed from New to In Progress

https://github.com/ceph/ceph/pull/20698

#2 - 03/04/2018 10:59 PM - Marcus Watts

I updated my PR. I moved the fill loop down from just the put-object code path to just above mg_read(). I don't believe there was logic to handle short-reads anywhere else, and we've gotten reports of hard to reproduce signature verification failures that could be explained by short reads, if so that should be fixed with this updated patch.

It's likely main | luminous have the same problem but I have not yet looked to see if that's the case.

#3 - 03/05/2018 12:15 AM - Matt Benjamin

great work, thanks marcus!

Matt

#4 - 03/05/2018 08:47 PM - Marcus Watts

I checked a copy of master. It does not have the problem. I'll need to do more checking; this is a mutant master linked against openssl 1.1.

#5 - 03/05/2018 09:32 PM - Marcus Watts

Matt suggested that it might be best to apply this fix to master - that way it doesn't matter what behavior a particular version of civetweb has; our code will work with it. So I've put together,
https://github.com/ceph/ceph/pull/20724

#6 - 03/05/2018 09:47 PM - Nathan Cutler

- Backport set to luminous, jewel

#7 - 03/08/2018 07:26 PM - Yehuda Sadeh

- Status changed from In Progress to Need Review

#8 - 03/13/2018 04:02 PM - Casey Bodley

- Subject changed from rgw: (jewel) inefficient buffer usage for PUTs to rgw: inefficient buffer usage for PUTs

- Status changed from Need Review to Pending Backport

#9 - 03/13/2018 09:57 PM - Nathan Cutler

- Copied to Backport #23347: luminous: rgw: inefficient buffer usage for PUTs added

#10 - 03/13/2018 09:57 PM - Nathan Cutler

- Copied to Backport #23348: jewel: rgw: inefficient buffer usage for PUTs added

#11 - 04/08/2018 07:55 PM - Nathan Cutler

- Status changed from Pending Backport to Resolved

#12 - 04/08/2018 07:57 PM - Nathan Cutler

- Related to Bug #23596: mg_read() call has wrong arguments added