RADOS - Bug #24361

auto compaction on rocksdb should kick in more often

05/31/2018 09:07 AM - Kefu Chai

Status:	Resolved	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Kefu Chai		
Category:	Performance/Resource Usage		
Target version:			
Source:		**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:	luminous, mimic	**Component(RADOS):**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature (v1):**	
Reviewed:		**Crash signature (v2):**	

Description

in rocksdb, by default, "max_bytes_for_level_base" is 256MB, "max_bytes_for_level_multiplier" is 10. so with this setting, the limit of each level of a rocksdb would look like

1. L0: in memory
2. L1: 256MB
3. L2: 2.56 GB
4. L3: 25.6 GB
5. L4: 256 GB
6. L5: 2.56 TB
7. L6: 25.6 TB

for monitor, 2.56 GB is relative large even for a large cluster. depending on the application of OSD, i'd say 2.56 GB is quite large for omap even taking the load of rgw into consideration.

in the case of monitor, if the cluster has been running for a long time in a large scale deployment, there is chance that the old and stale data could be migrated to L3. and new K/V data come in, the are written to lower level, like L0, L1. like

1. L1: 250MB
2. L2: 2 GB
3. L3: 25 GB
4. L4: 25 GB // stale data. non-user data

then we will be suffering from "space amplification". as the space amplification is (25 + 2 + 0.25 + 25) / (25 + 2 + 0.25) = 1.91 .

and the auto compaction does not help in this case, as none of sizes exceeds max_bytes limit. so a more flexible approach is to enable the dynamic level size for compaction.

---
[0] https://rocksdb.org/blog/2015/07/23/dynamic-level.html
[1] https://github.com/facebook/rocksdb/wiki/Leveled-Compaction

Related issues:

Copied to RADOS - Backport #24374: luminous: mon: auto compaction on rocksdb ...	**Resolved**
Copied to RADOS - Backport #24375: mimic: mon: auto compaction on rocksdb sho...	**Resolved**

History

#1 - 05/31/2018 09:16 AM - Kefu Chai

- Description updated

#2 - 05/31/2018 09:40 AM - Kefu Chai

- Status changed from New to Fix Under Review

https://github.com/ceph/ceph/pull/22337

#3 - 06/01/2018 06:08 AM - Kefu Chai

- Status changed from Fix Under Review to Pending Backport

#4 - 06/01/2018 06:08 AM - Kefu Chai

- Copied to Backport #24374: luminous: mon: auto compaction on rocksdb should kick in more often added

#5 - 06/01/2018 06:11 AM - Kefu Chai

- Copied to Backport #24375: mimic: mon: auto compaction on rocksdb should kick in more often added

#7 - 08/13/2019 12:55 AM - Neha Ojha

- Status changed from Pending Backport to Resolved