bluestore - Bug #23251

ceph daemon osd.NNN slow_used_bytes and slow_total_bytes wrong?

03/06/2018 07:52 PM - Ben England

Status:	Rejected	**Start date:**	03/06/2018
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	v12.2.4		
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	v12.2.1
Backport:		**ceph-qa-suite:**	
Regression:		**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

version: ceph-osd-12.2.1-34.el7cp.x86_64 = RHCS 3.0z1

In trying to understand ceph daemon osd.NNN perf dump counters, I came across this riddle. Bluestore is supposed to return the total amount of space available on the "slow" (data) device and the amount in use with:

```
[root@b10-h01-r620 bene]# ssh c05-h21-6048r ceph daemon osd.240 perf dump | grep slow
"slow_total_bytes": 79989571584,  (~80 GB)
"slow_used_bytes": 3394240512,    (~3 GB)
```

However, if I look at OSD-level counters:

```
[root@b10-h01-r620 bene]# ssh c05-h21-6048r ceph daemon osd.240 perf dump | grep stat_bytes
"stat_bytes": 2000811954176,  (~2 TB)
"stat_bytes_used": 1252726378496, (~1.25 TB)
"stat_bytes_avail": 748085575680,
```

stat_bytes matches what parted says about the partition size, so I believe this one and not the bluestore one. But how does the allocator know if it has free space when slow_total_bytes and slow_used_bytes are wrong?

Anyone else see this on newer versions?

History

#1 - 03/07/2018 09:31 AM - Igor Fedotov

"slow_total_bytes" and "slow_used_bytes" are under "BlueFS" section and denotes just a fraction of BlueStore block device space given/used by BlueFS (i.e. DB and/or WAL).
Hence that's not a bug IMO. Suggest to close.

#2 - 03/07/2018 01:24 PM - Ben England

Thanks for responding, I didn't realize that, thought from looking at code that it was used for data as well. You can close this tracker, but could someone tell me what it means if you have non-zero slow_total_bytes and slow_used_bytes when you have a dedicated partition/volume for the WAL and RocksDB? Does it mean that the OSD ran out of space in those dedicated partitions and had to resort to using the "slow" device space? This is important because it means that the partitions were sized wrong, am trying to learn how to size them. Maybe this is really a documentation bug? Because I didn't see this covered in Bluestore documentation.
-ben

#3 - 03/07/2018 01:45 PM - Igor Fedotov

- Status changed from New to Rejected

BlueStore has a BlueFS rebalance feature that dynamically reserves some amount of space for BlueFS at 'slow' device - current value is reported as slow_total_bytes. And yes - DB/WAL can use it when they lacks space at their major device(s) - slow_used_bytes tracks amount of data spilled over. So you'll have non-zero slow_total_bytes and zero slow_used_bytes in the normal state and non-zero slow_used_bytes in case of spillover. Don't know what documentation is saying on this topic...