Ceph tuning practice on aarch64 full HDD servers

Ker Liu ker.liu@arm.com
Infrastructure Landing Team
Agenda

- Ceph Architecture
- Testing Cluster
- Ceph Tuning (Hardware, Software, OS, Ceph configuration)
- 4k Seq Write Tuning
Ceph Architecture
Testing Cluster

**Hardware:**
- 3 x ARM server
  - 64 x 2400 MHz cores
  - 128GB RAM
  - 10G NIC
  - 11 x 7200RPM 4TB HDD
  - 1 x Intel SSD DC3700 800GB

**Software:**
- CentOS 7.6.1810
- Ceph nautilus 14.2.2

**Benchmarking tool:**
- FIO 3.1
Testing Cluster – After installation

Ceph cluster from hardware view (One node)

A ssd is cut into 10 pieces as osd db.
Tuning Overview

- Hardware Tuning
- Ceph Software Tuning
- OS Tuning
- Ceph Configuration Tuning
Hardware Tuning

- HDD (SAS/SATA)
  - Latency: ~ 10ms
  - Bandwidth: ~6Gbps

- SSD (SATA)
  - Latency: ~ 0.1ms
  - Bandwidth: ~24Gbps

- SSD (NVME)
  - Latency: ~ 0.075ms
  - Bandwidth: ~24Gbps

Disk vs. Network

- Network
  - 10Gbps
  - 40Gbps
  - 100Gbps
Ceph Software Tuning

- Some fundamental algorithms are optimized with Neon and specific extensions
- CRC/SHA/AES optimizations
- ISA-L optimizations

Notes: ARM OSS team did lots of optimizations for above items.
# OS Tuning

<table>
<thead>
<tr>
<th>Disk</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I/O Scheduler</td>
</tr>
<tr>
<td></td>
<td>Swappiness</td>
</tr>
<tr>
<td></td>
<td>Size of read_ahead_kb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jumbo frames</td>
</tr>
<tr>
<td></td>
<td>Network related parameters</td>
</tr>
<tr>
<td></td>
<td>Interrupt</td>
</tr>
</tbody>
</table>
Ceph Configuration Tuning

**Bluestore**

- Bluestore cache
  Maximize bluestore cache within the system’s tolerance.

- Bluestore cache autotuning feature
  Example:
  - bluestore_cache_autotune = 0
  - bluestore_cache_kv_ratio = 0.2
  - bluestore_cache_meta_ratio = 0.8

- bluestore_rocksdb_options
Benchmarking Finding

- **Test**
  - 4k Seq Write
  - 4k Rand Write
  - 4M Seq/Rand Write

- **Statistics**
  - SSD:
    - I/O Scheduler
  - HDD:
    - NIC
  - Rocksdb

- **bottleneck**

- **Tuning**
4k Seq Write Tuning

While with SATA SSD as database device, Bluestore takes up most of the time.
4k Seq Write Tuning – Cont.

**Default**

bluestore rocksdb options =
- compression=kNoCompression,
- max_write_buffer_number=4,
- min_write_buffer_number_to_merge=1,
- recycle_log_file_num=4,
- write_buffer_size=256M,
- compaction_readahead_size=2M

**Tuned**

bluestore rocksdb options =
- compression=kNoCompression,
- max_write_buffer_number=32,
- min_write_buffer_number_to_merge=2,
- recycle_log_file_num=32,
- write_buffer_size=64M,
- compaction_readahead_size=2M

FIO Seq Write BS=4k

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB/s</td>
<td>~10%</td>
<td>~10%</td>
</tr>
</tbody>
</table>

~10%
Thank you

Join Linaro to accelerate deployment of your Arm-based solutions through collaboration

contactus@linaro.org