Open Infrastructure - Task #8069

Investigate potential bottleneck on storage/CEPH at DCL

05/27/2020 10:54 AM - Timothée Floure

Status:	Closed	**Start date:**	05/27/2020
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
PM Check date:			

Description

History

#1 - 05/27/2020 10:59 AM - Timothée Floure

Our hardware:

- RAID controllers: perc h700, perc h800
  - Technical manual: https://www.dell.com/learn/us/en/04/shared-content~data-sheets/documents~perc-technical-guidebook.pdf
  - 2x4 ports, 6GB SAS 2.0, x8 PCIe 2.0
  - 512M to 1GB cache, 800MHz DDR2
  - IO load balacing on H800, not on h700 <- how does it work, is it significant?
- Each server has dual 10Gbps connectivity.
- Arista switches: 7050s
  - Datasheet: https://www.arista.com/assets/data/pdf/Datasheets/7050S_Datasheet.pdf
  - 52 x 1/10GbE SFP
  - 4GB RAM, Dual-core x86 CPU
  - '1.04 Tbps'
  - 9MB Dynamic Buffer Allocation
- Cables?
- Disks?

#2 - 05/27/2020 11:09 AM - Nico Schottelius

Some questions we should be able to answer:

Real scenarios

NOTE: assuming all disks running at 'full speed'.
NOTE: big big unknown here is how the cache of the RAID controller behave.
NOTE: unknown IOPS limitations on raid controllers. <--- TODO, more important than bandwith!

- The R710 server has 8 disks slots (supposedly with a h700 controller). Given that we fully populate the server, what is the maximum bandwidth available per OSD running on that machine?
  -> 4 GB/s from PCIe but 3GB/s for SATA -> 375 MB/s per disk modulo caching from RAID controller.

- The R815 has 6 disk slots (is that true? -> Balazs). Same question as above.
  -> SAS 6GB/s but 3GB/s for SATA -> 500 MB/s per disk, module caching from RAID controller.

- What about an R815 with an md array (12x 3.5" HDD via SAS cable attached to H800)
  -> 4 GB/s from PCIe connector (SAS supports 6GB/s) -> 333 MB/s per device.
  - Is the bottleneck likely a) the disk b) the controller c) the network of the server d) another component in the server
    - 10 Gbps = 1.25 GB/s = 104 MB/s per disk at full speed.
    - Controller PCIe limits at 500 MB/s per disk at full speed.
    -> Bottleneck likely to be on disk or network.

- Given an Arista 7050 and an imaginary bandwidth per disk of 50 MB/s, how many disks can we run on one 7050?
  - The Arista is supposed to handle 1.04 Tbps = 130000 MB/s = 2600 * 50 MB/s => not an issue.

- Is the PCI-E bus (it's not a bus anymore - afair it's point-to-point) on either server model a limitation?
  - It provides access to networking, disks and has an interconnect to the cpus

-> No worried, but TODO.

- We are using ceph bluestore (https://ceph.io/community/new-luminous-bluestore/)

- Does it make sense to switch our storage model to use 2 SSDs (f.i. 1TB) in a raid1 in front of HDDs and drop the distinction of HDD/SSD?
  - raid1 is needed as on the failure of the SSD all osds that have the rocksdb/bluefs on it fail

-> TODO

(skip answers if they are too far from what you can gather)

#3 - 05/27/2020 11:26 AM - Timothée Floure

Regarding the RAID controllers:

- RAID0 (striping - redundancy is handled by CEPH across physical servers).
- Some controllers are battery-backed:
  - Likely write-back cache.
- Some are not:
  - Likely write-though cache.
  - .. or forced WB via BIOS/firmware setting?
- Read cache defaults to 'Adaptive Read Ahead': When selected, the controller begins using Read-Ahead if the two most recent disk accesses occurred in sequential sectors.
  - Fairly useless for random reads.

#4 - 05/27/2020 11:46 AM - Timothée Floure

Regarding PCIe AND SAS/SATA:

- Controllers are connected on x8 PCIe 2.0 => 500 MB/s per-lane for PCIe 2.0 -> x8 = 4 GB/s
- 6 GB/s SAS 2.0 connectivity -> how is this split between disks? Should be fine anyway.
  - perc h700 supports SATA 3GB/s, perch800 does not support SATA.
- How are our network cards connected? Should be fine anyway. 10Gbpe = 1.25 GB/s -> even PCIe 2.0 4x is more than enough.

#5 - 05/27/2020 12:42 PM - Timothée Floure

I'll be AFK for a little while: the big pain point is the hardware RAID controller.

- Unknown effect on IOPS (needs more digging, not obvious).
  - The internet says (reddit, random wikis, CEPH mailing list) using RAID0 when passthrough is not supported is BAD (lower performance/IOPS, buggy firmware, some (unknown?) implication on cache, ...).
- Unknown effect from the cache.

#6 - 05/29/2020 09:02 AM - Timothée Floure

- Status changed from In Progress to Waiting

Waiting for @llnu to test a RAID controller with passthrough.

https://redmine.ungleich.ch/issues/8063?issue_count=4&issue_position=1&next_issue_id=8002#note-22

#7 - 01/03/2024 06:31 PM - Nico Schottelius

- Status changed from Waiting to Closed