Ceph: A Journey to 1 TiB/s

Mark Nelson mark.nelson@clyso.com
03/28/2024

CLYSO

How it Started: A Great Customer

- ✓ Existing HDD cluster no longer met performance needs
- ✓ Wanted to expand cluster to 5-7PB Usable (680 15.36TB NVMe drives!)
- ✓ Extremely fast and well designed 100GbE Network infrastructure already in place
- ✓ Very open to hardware recommendations

CL'SO

Customer Constraints

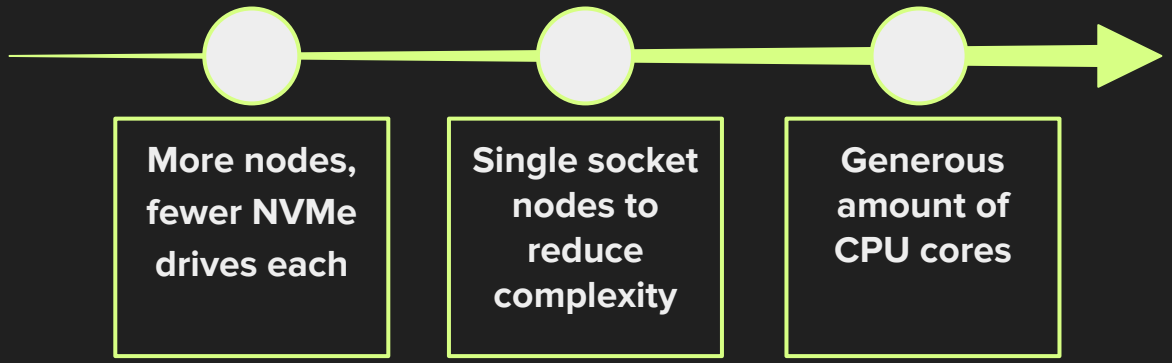
- ✓ Has to be Dell hardware.
- ✓ Limited to 4U per rack over 17 Racks.
- ✓ Per-rack power limited to roughly 1500 watts.
- ✓ Zero disruption during the hardware upgrade.

CLYSO

Hardware Design: K.I.S.S.

How to build fast Ceph nodes?

Don’t be too greedy with density. Sweet spot of 10-12 NVMe drives per node or 6-8 for maximum performance on EPYC processors.



CL'SO

Hardware Design: K.I.S.S.

Hidden constraint: memory bandwidth

Test	OSD Memory Bandwidth Amplification	FIO+Librbd Memory Bandwidth Amplification
4MB Reads	*5.7x*	*8.0x (why?)*
4MB Writes	*9.6x*	*2.5x*
4K Random Reads	*23.1x*	*13.2x*
4K Random Writes	*63.1x (yikes!)*	*5.2x*

Tested using: AMDuProfPcm -m memory -d 10

CL'SO

Hardware Design: K.I.S.S.

- ✓ To achieve 20 GB/s of 4MB read throughput, we need around 114 GB/s of OSD memory bandwidth.
- ✓ If clients are co-located on the OSD nodes, we need closer to 274 GB/s in aggregate!
- ✓ Writes and small random IO have worse amplification, but generally are limited by other constraints first.

Hardware Design: Final Config

Nodes	68 x Dell PowerEdge R6615
CPU	1 x AMD EPYC 9454P 48C/96T
Memory	192 GB DDR5 4800MT/s (12 Channels, 460GB/s)
Network	2 x 100GbE Mellanox ConnectX-6
NVMe	10 x Dell 15.36TB Enterprise NVMe Read Intensive AG
OS Version	Ubuntu 20.04.6 (Focal)
Ceph Version	Quincy v17.2.7 (Upstream Deb Packages)

CL'SO

The Plan

1. Upgrade existing cluster to Quincy.
2. Deploy new hardware and image nodes with customer environment. Run basic HW tests.
3. Install CBT and evaluate system and ceph configurations on new HW at various scales.
4. Re-deploy new hardware as OSDs on the existing cluster. Use upmap-remapped to migrate all data to the new OSDs.

CL'SO

Phases 1 & 2

- ✓ Tobias (Tobi!) at Clyso upgrades the existing cluster without issue. Cephadm works well!
- ✓ Customer deploys new hardware. Slight glitch in tooling puts OS on OSD drives. Initial tests done with 8 drives instead of 10.
- ✓ Basic tests look good. Network looks good, NVMe drives look good. What could go wrong?

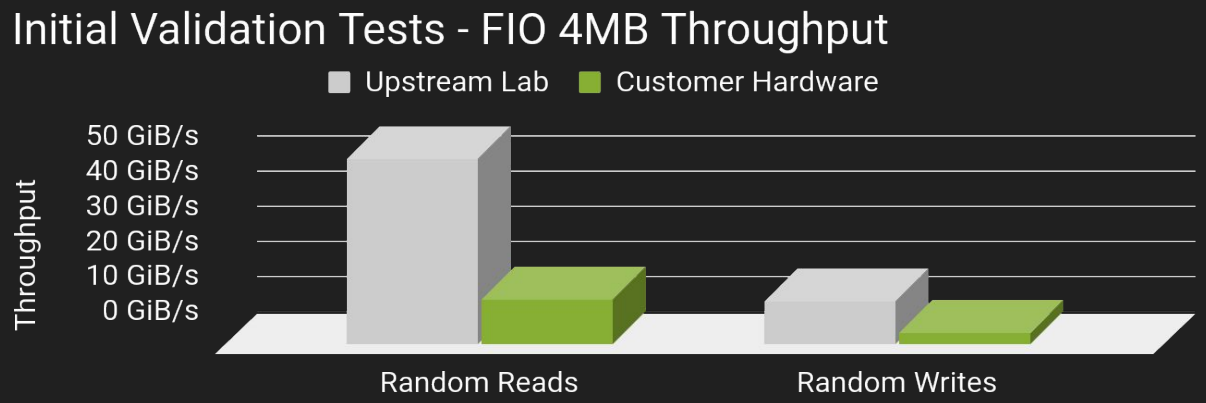
CLYSO

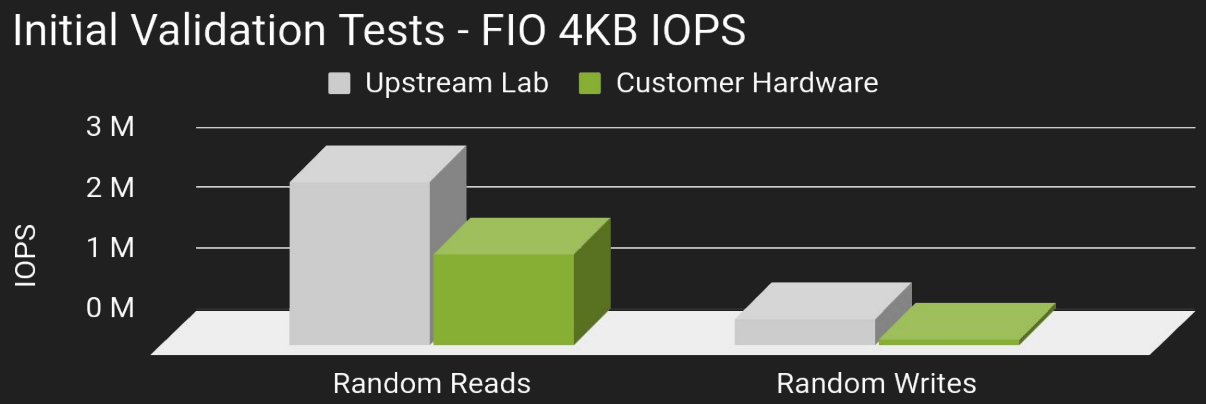
Phase 3A
Why “A”?

Initial Validation Tests vs. Published Upstream Lab Results		
	Upstream Lab	Customer Deployment
Nodes	5 x Dell PowerEdge R6515	3 x Dell PowerEdge R6615
CPU	1 x AMD EPYC 7742 64C/128T	1 x AMD EPYC 9454P 48C/96T
Memory	128 GB DDR4 3200MT/s (8 Channels, 204.8GB/s)	192 GB DDR5 4800MT/s (12 Channels, 460.8GB/s)
Network	1 x 100GbE	2 x 100GbE
NVMe	6 x Samsung 3.84TB PM983	8 x Dell 15.36TB PM1733a

CL'SO

Phase 3A (uh oh)





CL'SO

Phase 3B

- ✓ Pain and suffering! 2 Weeks of constant testing and analysis. See paper for details!
- ✓ Switched back to testing the raw hardware and single-OSD Ceph.
- ✓ Bizarre behavior! Inconsistent performance! No clear explanation what's going on.
- ✓ Several false leads. CPU thermal throttling looked hopeful for some time but was incorrect in the end.

CL`,`SO

Phase 3B

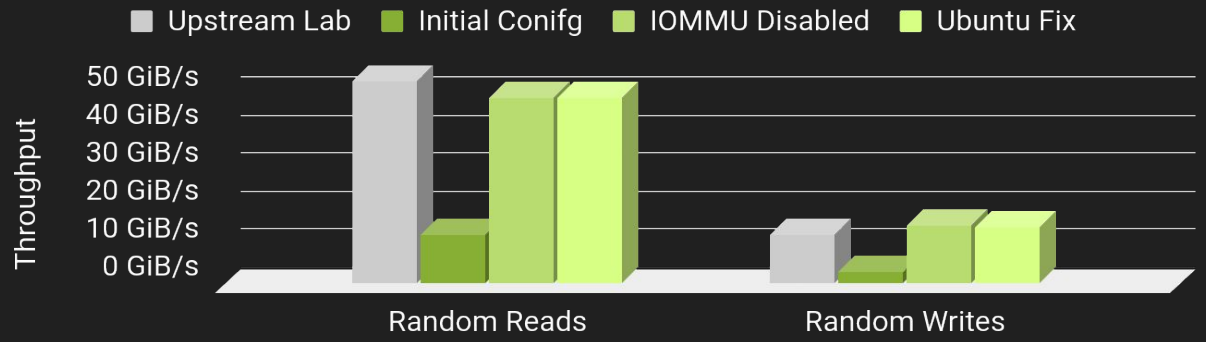
The 3 Fixes

1. Ensure the systems are running in maximum performance mode (Disable c-states).
2. Disable IOMMU (Customer figured this one out after testing with perf!)
3. Fix upstream Ceph Debian/Ubuntu builds to not compile RocksDB in debug mode.

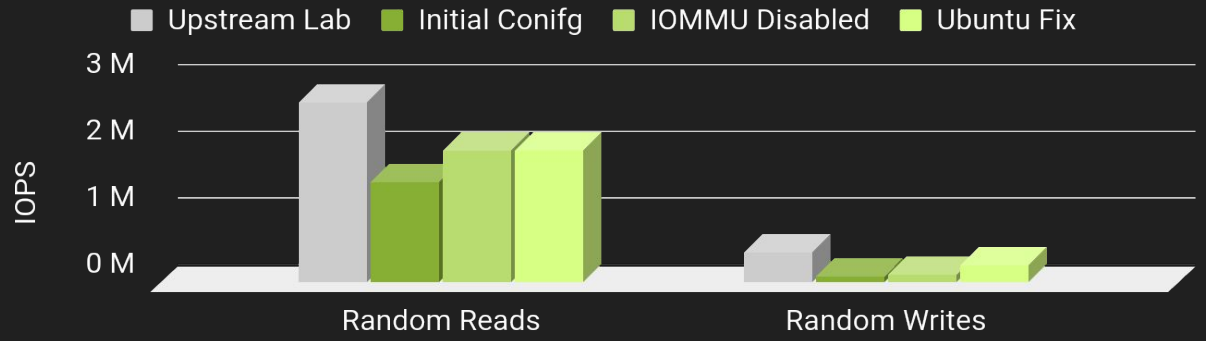
CL'SO

Phase 3B

All Fixes Validation Tests - FIO 4MB Throughput



All Fixes Validation Tests - FIO 4KB IOPS



Winter Break

- ✓ Plan was to have performance analysis done at this point, but we're back at the beginning (with fixes this time though!)
- ✓ Customer reimages all nodes with the fixes in place and also fixes the boot drive issue so we can use all 10 NVMe drives per node.
- ✓ Mark takes a much needed vacation to somewhere tropical and warm.

CLYSO

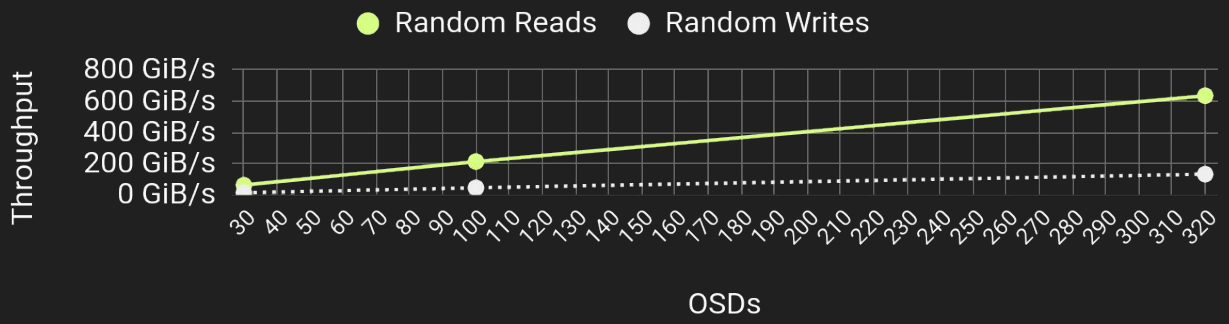
Phase 3C

- ✓ Back from winter break, recharged and ready to go!
- ✓ Only have one week to do all scaling tests before we need to move on to cluster integration and data migration.
- ✓ First four days of the new year occupied by critical outage at another customer site!
- ✓ Testing doesn’t start until Friday, but we got an extra day on Monday to wrap things up.

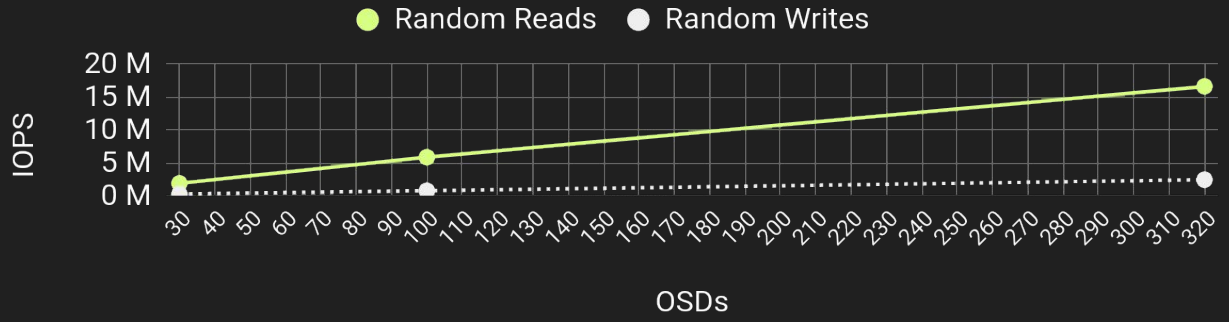
CLYSO

Phase 3C

Post-Fixes OSD Scaling - FIO 4MB Throughput



Post-Fixes OSD Scaling - FIO 4KB IOPS



CL'SO

Phase 3C

- ✓ After the fixes in Phase 2 performance scaled beautifully up to 320 OSDs.
- ✓ Hit peak throughput of over 650 GiB/s with less than half the total OSD nodes!
- ✓ 16.5M randread and 2.4M randwrite IOPS!
- ? But to test at larger scales, we need to put the fio clients on the OSD nodes and that uses more memory bandwidth...

CL'SO

Phase 3D

- ✓ Over the weekend we ran a large variety of tests to see how performance at this scale was affected by different settings.
- ? We observed significant performance issues with PGs going laggy at this scale, though this could be partially mitigated via various tunings.
- ✓ Did we hit our goal of 1 TiB/s? Yes!

Phase 3D

The best* numbers we got with 63 nodes:

Test	3X Replication	6+2 Erasure Coding
4MB Reads	*1025 GiB/s*	*547 GiB/s*
4MB Writes	*270 GiB/s*	*387 GiB/s*
4K Random Reads	*25.5M IOPS*	*3.4M IOPS*
4K Random Writes	*4.9M IOPS*	*936K IOPS*

** Not necessarily all with the same settings to work around the laggy PG Issue.*

CL'SO

Phase 3D

How did Ceph's per-node performance scale (3x Rep)?

Test	30 OSDs	100 OSDs	320 OSDs	630 OSDs*
Co-Located FIO	No	No	No	Yes
4MB Reads	*21.0 GiB/s*	*21.4 GiB/s*	*19.8 GiB/s*	*16.3 GiB/s***
4MB Writes	*4.9 GiB/s***	*4.6 GiB/s***	*4.2 GiB/s***	*4.3 GiB/s***
4K Random Reads	*621K IOPS*	*583K IOPS*	*518K IOPS*	*405K IOPS*
4K Random Writes	*82.7K IOPS***	*74.5K IOPS***	*75.4K IOPS***	*75.3K IOPS***

** Not necessarily all with the same settings to work around the laggy PG Issue.*

*** 3X Rep means 3 times the IO is actually hitting the OSDs versus what's shown here.*

**** Co-locating client process uses significantly more memory bandwidth!*

CL'SO

Phase 4

- ✓ Clyso handed the nodes back to the customer for final reimaging.
- ✓ Next, Tobi integrated all 68 of the new nodes into the existing cluster...
- ✓ ...And migrated all data to the new nodes in under two weeks...
- ✓ ...while the cluster was actively being used! (Helped by Dan's upmap-remapped script!)

CLYSO

Phase 4

- ✓ The customer indicated an immediate 10x performance increase with no additional tuning in their application.
- ✓ More application optimizations necessary to make use of the extra performance!
- ✓ Clyso expects to be deploying additional clusters like this in the future for many different use cases including HPC and AI workloads.

CLYSO

Thank You!

To see the original blog article visit:
https://ceph.io/en/news/blog/2024/ceph-a-journey-to-1tibps/
Or to download our PDF version:
https://clyso.com/static/Ceph-A-Journey-to-1TiBps.pdf

Contact: mark.nelson@clyso.com

CLYSO