# Block Storage Service: Status and Performance

*Dan van der Ster, IT-DSS, 6 June 2014*

## Summary

This memo summarizes the current status of the Ceph block storage service as it is used for OpenStack Cinder Volumes and Glance Images as of May 2014. We present the block storage activity on the current cluster, measuring IOPS and latencies, and present a cost/benefit analysis of using SSDs to optimize the cost and performance efficiency of the service. During tests in collaboration with IT-CF, we have concluded that by adding SSDs as the synchronous write journals (used to guarantee data durability), we are able increase the IOPS capacity by 4-5 times, at a cost of decreasing the available volume by 20%.

Further, the testing has shown that the Ceph implementation is able to operate at the limit of the hardware performance; software-induced performance limitations were not yet observed in either the spinning disk or SSD configurations. In addition, we believe that increasing small write performance with SSDs is applicable only to the block storage use-case; high-bandwidth use-cases such as physics data storage should not require SSDs.

## Configuration

The current production block storage cluster is running the latest Inktank Ceph Enterprise version 1.1 (equivalent to Ceph 0.67.9). It is composed of 48 disk servers (hardware described below[1]) resulting in a total raw capacity of roughly 3PB. Seven of the 48 servers are not in production; they are used for tests/preprod work, and one of the servers was DoA and not yet repaired.

Servers are grouped logically by racks, with data being replicated across 4 racks; this policy allows for service continuity in the case of the simultaneous disk or host failure in any 3 racks. Note that we have calculated that 3x replication would provide adequate data durability[2], so in the near future that change will be implemented.

Ceph guarantees data consistency using write-ahead journalling[3]. The current configuration co-locates the journal on the same drive as the XFS filestore. The best practices documentation recommends using SSDs for the write journals, but until now we have not explored this option in production.

---

[1] 2x Xeon E5-2650, 64GB RAM, 1x LSI SAS2008 HBA, 24x Hitachi HUA5C303 3TB drives, 3x Hitachi HUA72302 2TB system drives.

[2] The Ceph reliability calculator computes 9-nines for 3x RADOS replication and 12-nines for 4x RADOS replication.

[3] In short, Ceph acknowledges writes only after they have written synchronously to a journal (i.e. a direct IO file or block device), followed by a buffered write to the filestore (i.e. XFS) which is flushed periodically.

The original plan for our block storage service was to evaluate the usage of our standard EOS disk server hardware. We now have enough experience with this hardware type to understand its limitations with this new use-case and can recommend a way forward to unlock increased performance.

On paper, the current hardware has known IOPS limitations. Assuming 130 IOPS per OSD drive[4], with 40 servers the cluster supports 130 IOPS * 40 * 24 = 124'800 IOPS. With 4x replication this drops to 31'200 writes/sec, and with each disk being used for both the journal and the filestore, the final capacity is at worst roughly 15'000 writes/sec. Assuming a 50% read/write ratio in the cluster, and with the knowledge that reads are much less IO intensive on the OSDs than writes[5], the total IOPS capacity of the cluster is estimated at between 20'000 and 30'000. (With 3x replication is roughly 41'600 writes/sec to handle the journal and filestore writes -- the other numbers would scale up accordingly).

OpenStack qemu-kvm hypervisors connect to the service via *librbd*, a user-mode plugin for the RADOS Block Device (RBD) component of Ceph. Librbd can be configured with a "disk-like" write-back cache -- currently the hypervisors use a 32MB writeback cache. Qemu-kvm also supports the throttling of block device IOs -- the configuration up until 20.05.2014 allowed 200 write iops, 400 read iops, 40MB/s written, and 80MB/s read, all per attached block device. From 20.05.2014 volumes are throttled to 100 IOPS read and write and 80 MB/s read and write.

## Status

Ceph has been used for the Glance image service since fall 2013, and the Cinder volume service was gradually offered to beta testers starting in late 2013, then to our IT colleagues in February 2014, and finally becoming general available in March.

The cluster is instrumented and a variety of metrics are displayed in SLS[6]. The metrics show linear growth towards the current status of more than 400 Cinder volumes with 140TB provisioned space, which is stored in Ceph as 55TB of written data, occupying 220TB on disk.

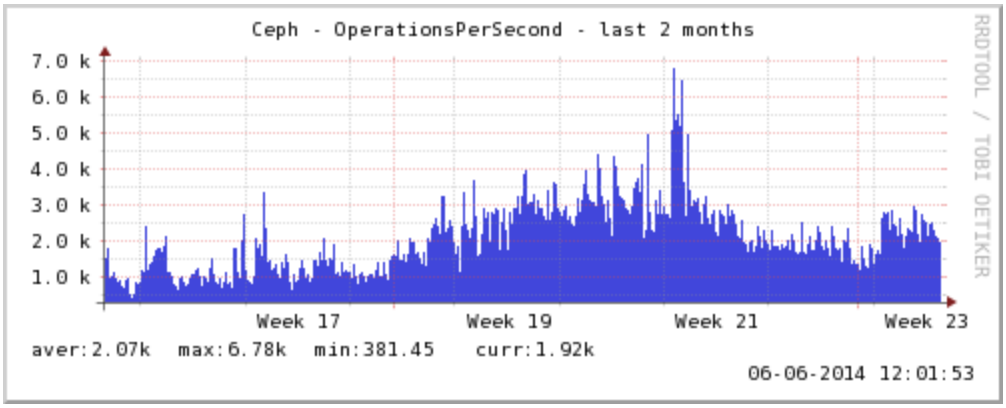
In addition, we monitor all reads and writes to the cluster and can summarize the access patterns to learn that out of roughly 320 million writes per day, around 75% are 4096 bytes in length -- hence the cluster is confirmed to be dominated by small writes.

## The Problem: Limited IOPS Capacity and Increasing Latencies

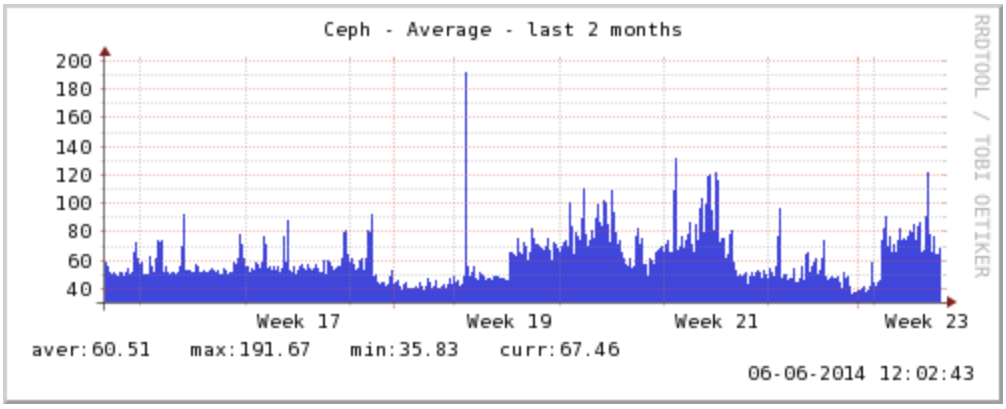
One notable SLS metric shows the increasing operations (read + write) per second as reported by the Ceph monitors.

---

[4] Measured by fio with 4kB random writes, ioengine=libaio, and direct=1.
[5] Reads are sent to only one of the replicas, and are often already cached.
[6] http://sls.cern.ch/sls/history.php?id=Ceph&more=ALL&period=24h




During the past two months, increased usage of the service has resulted in peaks of up to 7 to 8000 IOPS reported by Ceph. Now note how these increased IOPS lead to increased latency (mean 64k write latency in ms):




When the cluster was not heavily used, the mean latency was 50ms. A configuration change in week 18 allowed for 40ms, but heavy usage of the cluster since week 19 has resulted in >60ms latency, with peaks up to 100ms.

Looking forward, we expect that this usage will continue to grow in the coming weeks/months, the IOPS capacity of the cluster, and in particular the small writes capacity, will soon limit its growth and performance may start to suffer. And since we are already observing a high rate of small writes with an occupied space of ~200TB, it is clear that the IOPS capacity will be exceeded well before the full 3PB of storage can be significantly occupied.

## Solution: SSD Journals

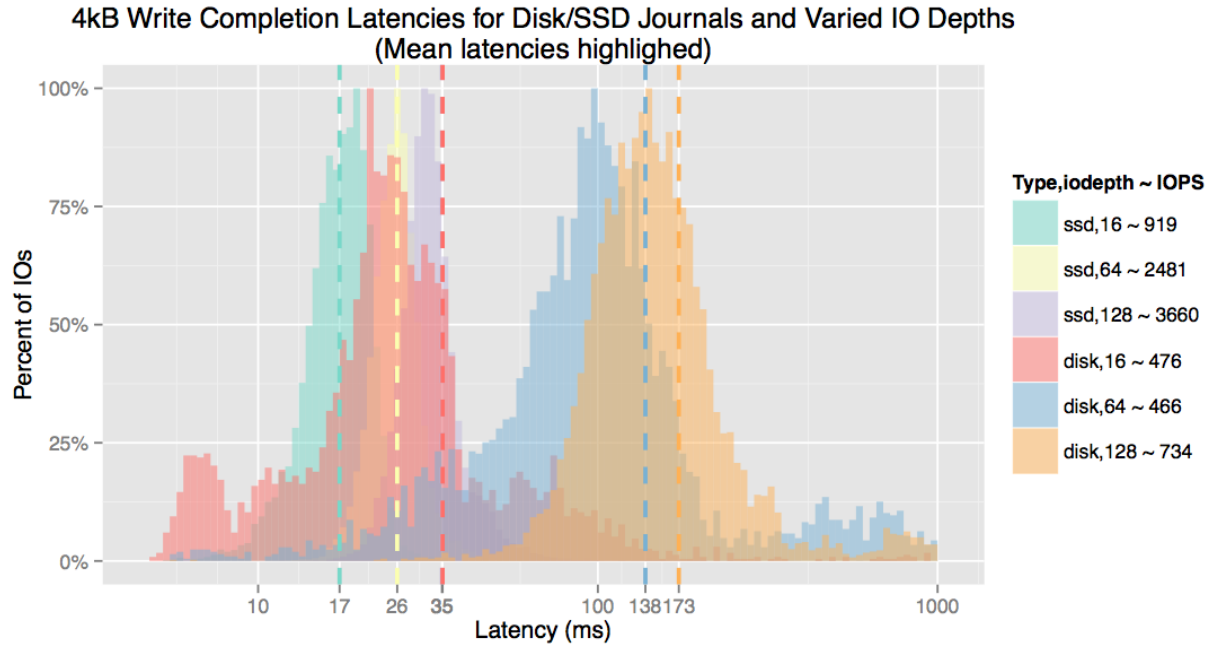
The best practices recommended in the Inktank Hardware Suggestion Guide[7] suggest the usage of SSDs for the write journal at a ratio of up to 5 OSDs to 1 SSD. The Intel DC S3700 or equivalent is recommended for its performance and long-term durability. Since the write

[7] http://www.inktank.com/resources/inktank-hardware-selection-guide/

journal uses synchronous writes, a fast journalling device can help with the IOPS performance. (The write journal is analogous to an NVRAM write cache in a NetApp filer, or the ZIL in a ZFS system, to take two examples). When using an SSD journal, we should expect at least a 100% increase in IOPS performance (due to halving the spinning disk writes), however larger short term burst increase may be possible. We have studied the expected IOPS per server with sample units provided by IT-CF.

With IT-CF we have built one server with 4x SSDs and 20x 3TB drives to compare its per-server performance versus the current disk-only configuration. Using fio with its new RBD driver[8] we have measured the 4kB write IOPS capacity and related latencies for various IO depths to the sample OSD servers. (For example, with iodepth=1, fio will keep a single IO in the libaio queue; this will not heavily load the OSD server. But with iodepth=128, fio will keep the queue full with 128 outstanding writes, which creates more parallelism on the OSDs and better evaluates the IOPS capacity and latencies). In all cases, we use 3x replication across OSDs but within the same physical server, and client-side write back caching is disabled.

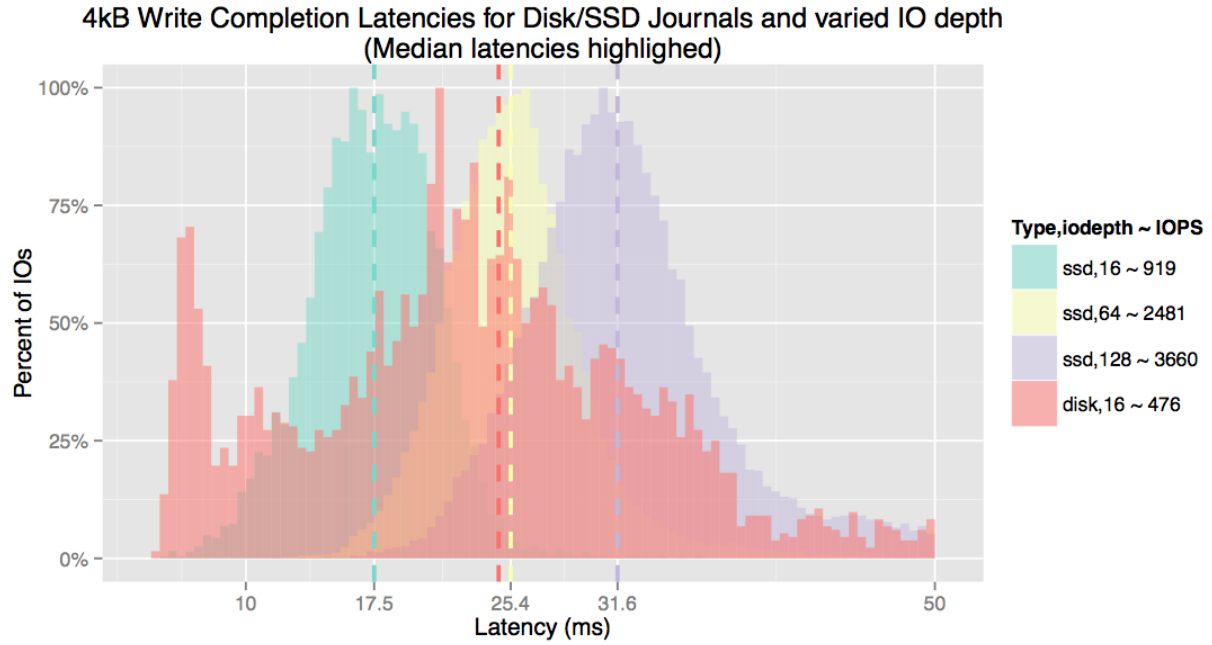
Below we highlight the main results of these tests.



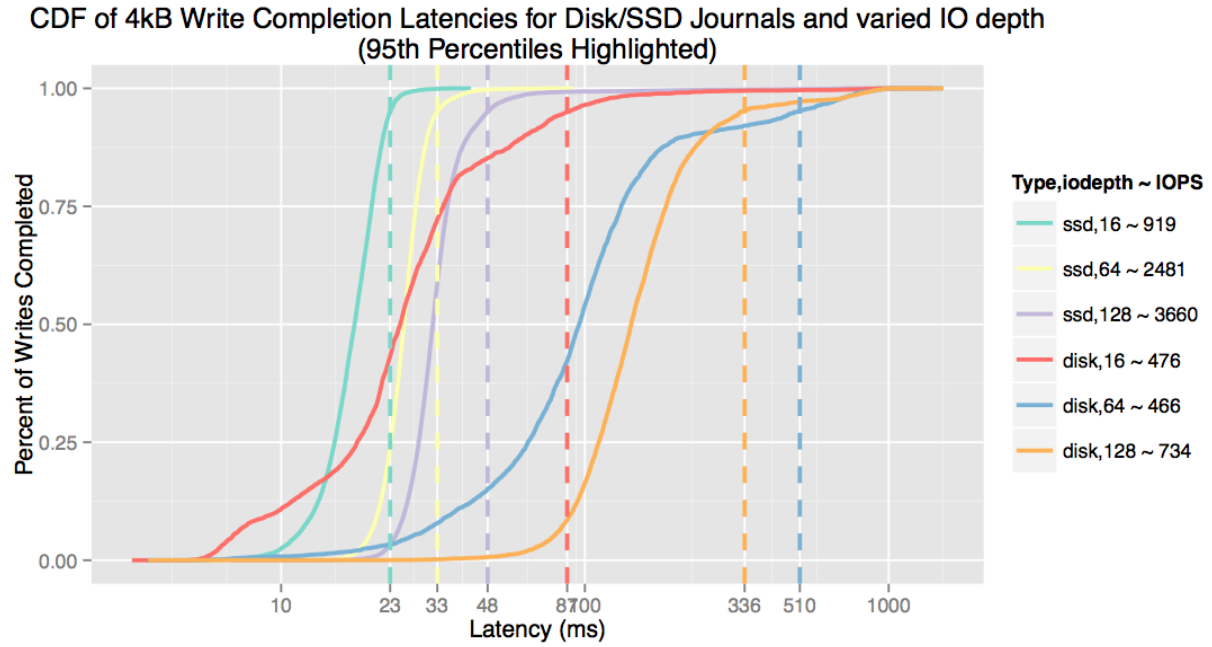

Above, we have shown the completion latencies for varied IO depths to an OSD servers having SSD or spinning disk journals. With a disk-only OSD server, fio achieves 476 IOPS with a mean latency of 35 ms when using iodepth=16. With SSD-based journals, that same latency is achieved while writing 3660 IOPS. Writing more than ~500 IOPS is not practical for the disk-only server; when writing at 734 IOPS, the mean latency was 173 ms.

[8] http://telekomcloud.github.io/ceph/2014/02/26/ceph-performance-analysis_fio_rbd.html

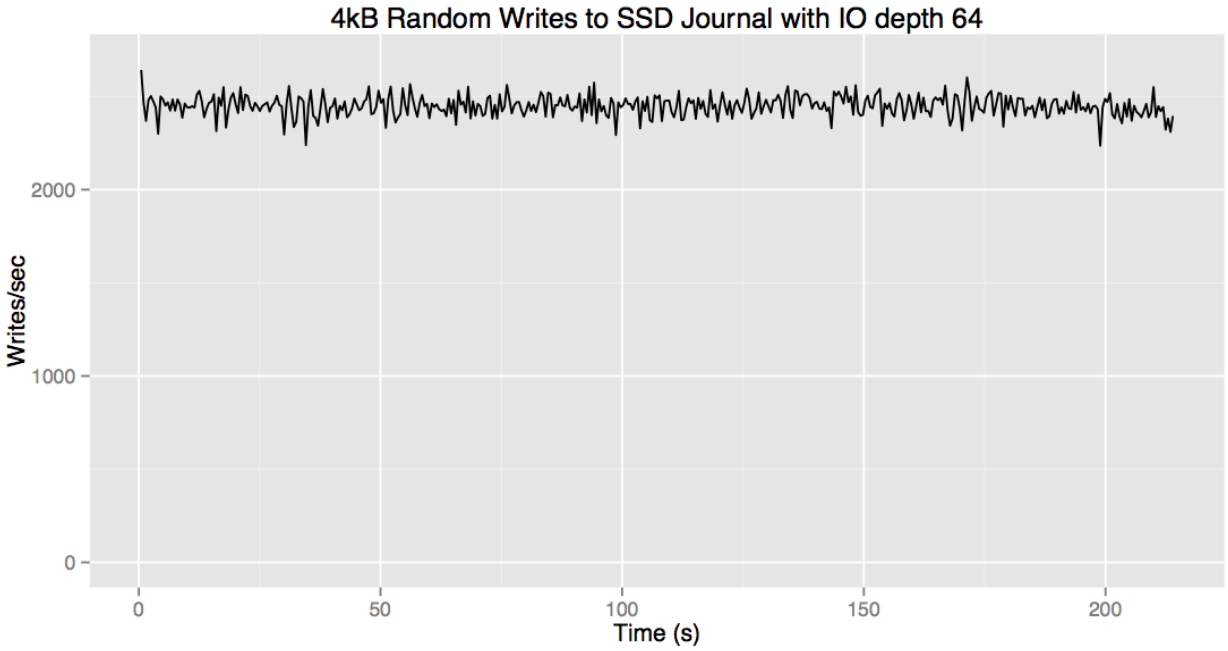
We look closer at the configurations having mean latency less than 50 ms. Here we see that the SSD configurations have narrow distributions, while the disk configuration is wide. The next plot highlights this point.




In the CDF plot below we have highlighted the 95th percentile of write completions. The SSD configurations complete 95% of the writes in under 23 ms, 33 ms, and 48 ms, respectively. The *disk,16* configuration, despite having a mean latency similar to the SSDs (though, at lower IOPS) requires 87 ms to complete 95% of the writes.



Finally, we plot a longer test of the ssd,64 configuration to check for slowdowns due to OSD flushes or fsync barriers. We can confirm that (at least in this configuration) 2400 writes/sec can be sustained for more than 200s, which exceeds the Linux kernel and Ceph flush intervals.




## Concluding Discussion and Costs

In the tests above we observe that the SSD journal servers can deliver 2000-3000 writes/sec without a large increase in latency. Scaled out to the entire cluster (40 servers in production) this would allow bursts of up to 80'000 to 120'000 writes/sec across the cluster, a factor of 4-6 above the current configuration. Note that one limitation of these tests was that long term (many hour) tests were not performed. Effects which only become apparent at those time scales (e.g. interference with background scrubbing, SMART, or other disk intensive activities) may decrease the overall IOPS capacity.

The Intel DC S3700 200GB has a retail price of 450CHF[9]. Four SSDs per server would increase the cost by 1800CHF (minus the cost of 4 spinning disks) while decreasing the raw capacity by 12TB (16.7%). The 100GB Intel DC S3700 (with half the cost) would have adequate volume to be used for this purpose, however these have a write limitation of 200MB/s, which would prove to be a bottleneck in the throughput to the servers.

[9] http://www.toppreise.ch/prod_301951.html

In future, we may also consider providing pools for high-IOPS use-cases, using SSD-only configurations[10]. To test this scenario, I propose to outfit all 48 Ceph servers with the four SSDs; then the eight preprod servers may then be temporarily configured without spinning disk OSDs for benchmarking a “provisioned IOPS”-like service. The total cost of this operation is therefore estimated at 450CHF * 48 * 4 = 86’400CHF minus taxes and volume discounts.

One unknown in the future expansion of the block storage cluster is how it would perform with single pools having mixed resources: some SSD/disk servers and some disk-only servers. Without running tests, we can predict that IOs to objects with at least one (out of three/four) replicas on the disk-only server would be penalized, with the synchronous write to the disk-only replica becoming a bottleneck. Thus, in a cluster with 90% SSD/disk and 10% disk-only servers, we estimate that up to 30% of IOs would be affected by the disk-only IOPS limitation (affecting 100% of Cinder Volumes).

[10] This would be similar to the Amazon EBS “Provisioned IOPS” service, where users can pay extra for 1000/2000/3000 IOPS volumes