1. Introduction

This effort sought to test the effectiveness of dm-cache [1] used as a writeback cache for a Ceph RBD device for random-access (database-oriented) workflows.

2. Test hardware

Servers from the Ceph Community Test Lab were configured as a Ceph cluster to serve as backing store for these tests. Each node was equipped with

* Two 10-core Intel Xeon E5-2650 processors, 20 virtual cores total
* 64 GB main memory
* 8 SATA 1TB Seagate Constellation.2 7200 RPM disk drives
  (7 used as OSDs on servers)
* Intel XL710 Ethernet adapter (40Gbit/sec)
* 4 Intel DC P3700 SSD (PCIe NVMe device, 1 used as dm-cache on client)

3 nodes were configured as servers with each non-OS disk serving OSD data (21 OSDs total). 1 node was configured as the client, as well as running the Ceph monitor.

3. Test software

The servers used the CentOS 7.1 kernel (3.10.0-229.7.2.el7). The client used Linux 4.2.0.

The cluster was configured with Ceph 0.94.3 ("Hammer") using Yves Trudeau's fork [2] of the Ceph Benchmark Tool (CBT, cf. [3]). Minor modifications of the tool were required to fit the test environment.

Yves' fork of CBT allows it to drive the SysBench [4] benchmark. SysBench 0.5 was used for these tests. Specifically, the LUA-based OLTP benchmark (oltp.lua) was used here. Detailed information about the specific benchmark is available at [5].

oltp.lua was run on top of MariaDB 5.55.44 [6].

4. Test methodology

CBT was configured to set up a Ceph cluster as described above, create an XFS-backed RADOS cluster, and instantiate one 8192MB RBD, used as the test target. dm-cache was configured on the client in writeback mode with one NVMe device namespace defined as the cache device, the RBD as the origin device. A separate NVMe namespace (on another device) was used for metadata. dm-cache's stochastic multiqueue (smq) policy was used.

To control for lingering cache effects, the dm-cache device was created freshly for each test.

For cache-tiering tests, the Ceph cluster was configured using one NVMe device namespace per server node as an additional OSD. These OSDs were configured as a cache tier overlaying the spinning disk-based OSDs.

To control for lingering cache effects, the cache tier was manually flushed between tests.

Tests were run for 10 minutes and the transaction rate (TPS) and 95% latency values were noted. The default database size of 10000 was used for base cases,

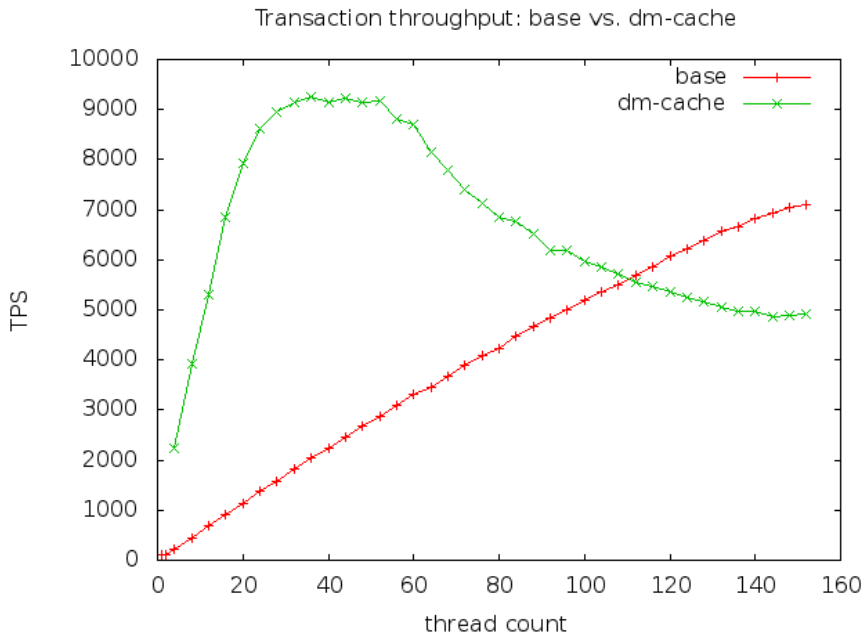
results for larger database sizes are noted below. oltp.lua warms the database before the test period.

5. Test results

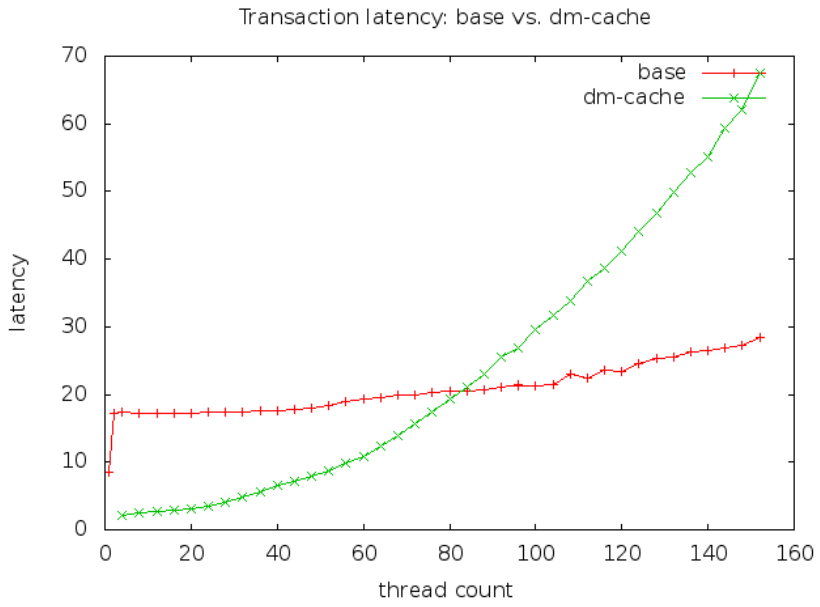
5.1. No caching vs. dm-cache vs. Ceph cache tiering

5.1.1. Results

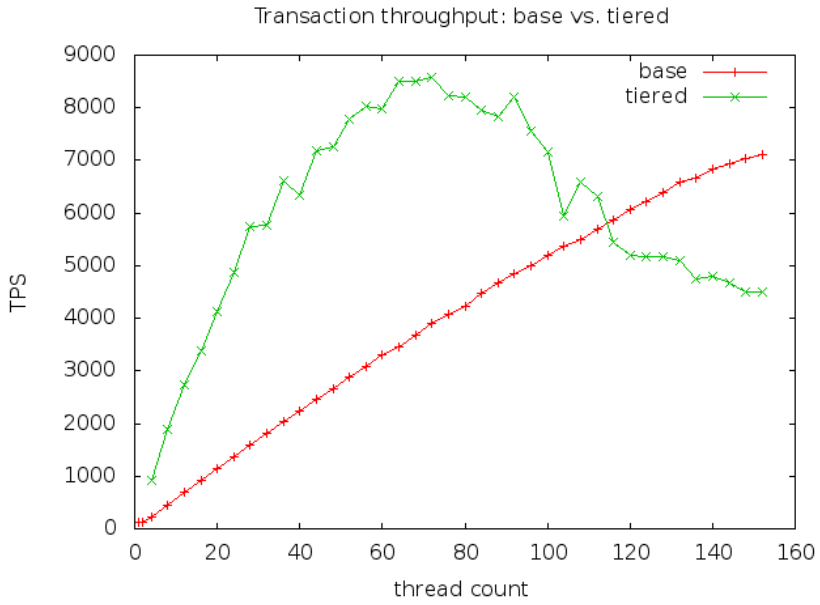
Compared to baseline (test results without dm-cache), dm-cache significantly improved throughput at thread counts below 116. From 116 to 152 threads, throughput was greater without dm-cache. This may be due to false sharing of cache blocks between threads.



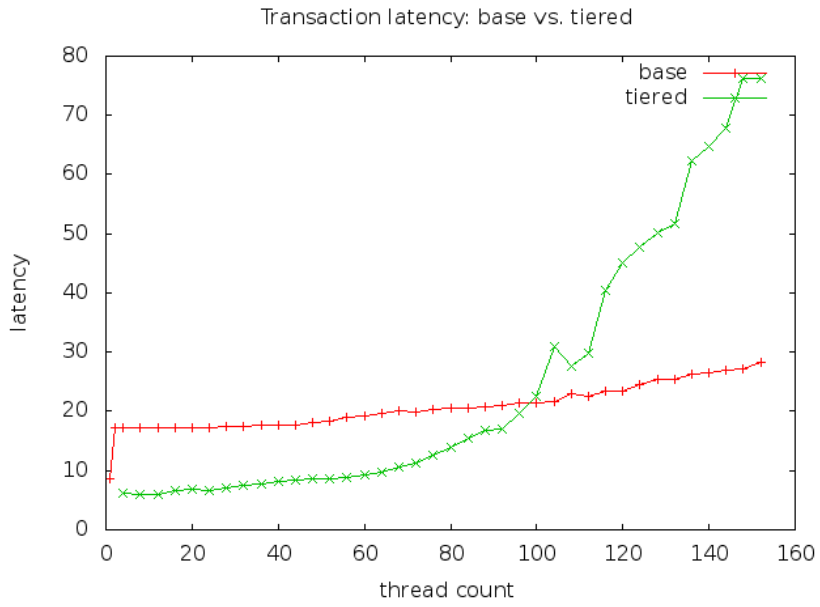
dm-cache improved latency slightly at thread counts below 80. Above 80 threads, latency rose sharply and increased more rapidly than without dm-cache.



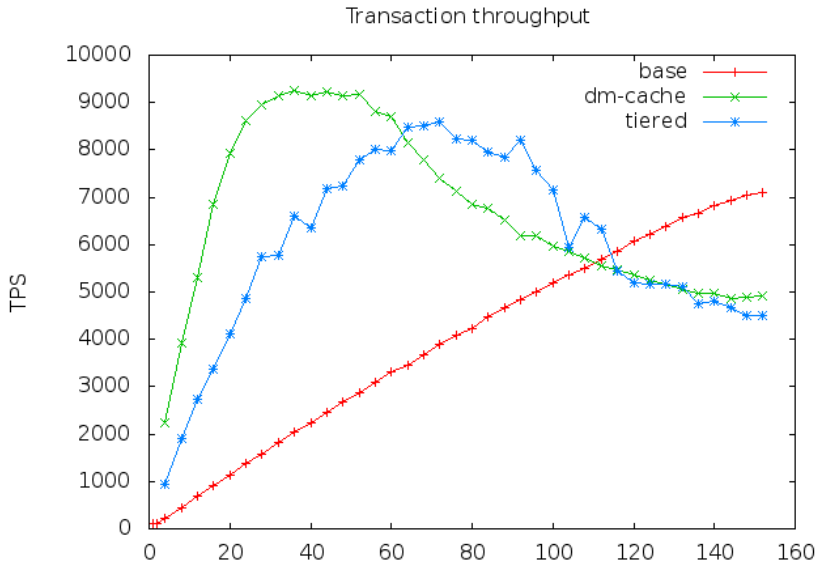
Cache tiering significantly improved throughput over baseline to 116 threads.

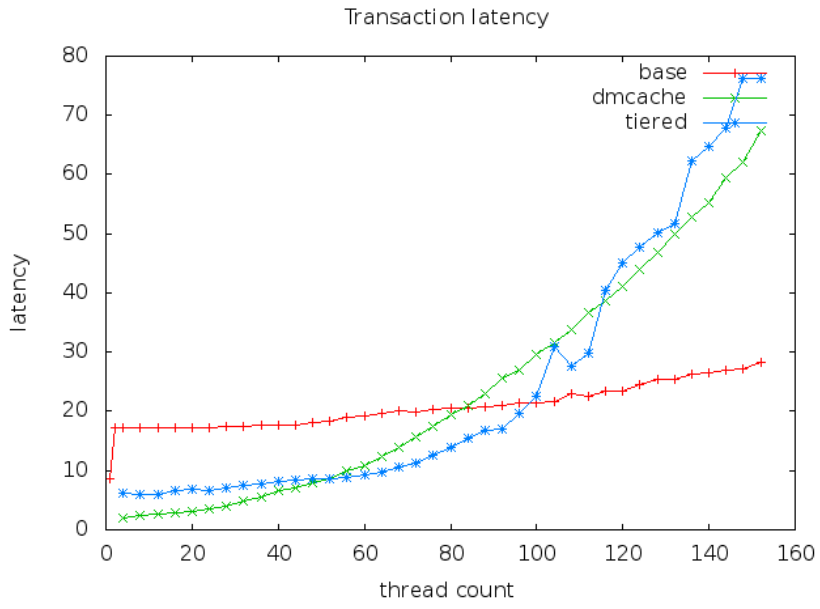


Latency was significantly improved under 100 threads.



Aggregate results are presented below.





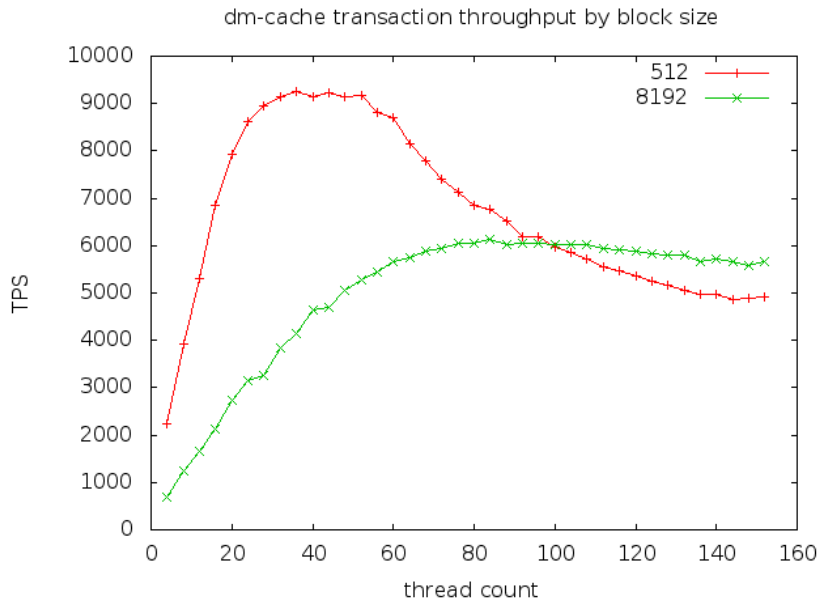
5.1.2. Discussion

Both cache strategies improve latency at lower thread counts until a breakdown point at which latency increases exponentially. This likely occurs once overhead from locking contention overwhelms the advantage provided by the storage media.

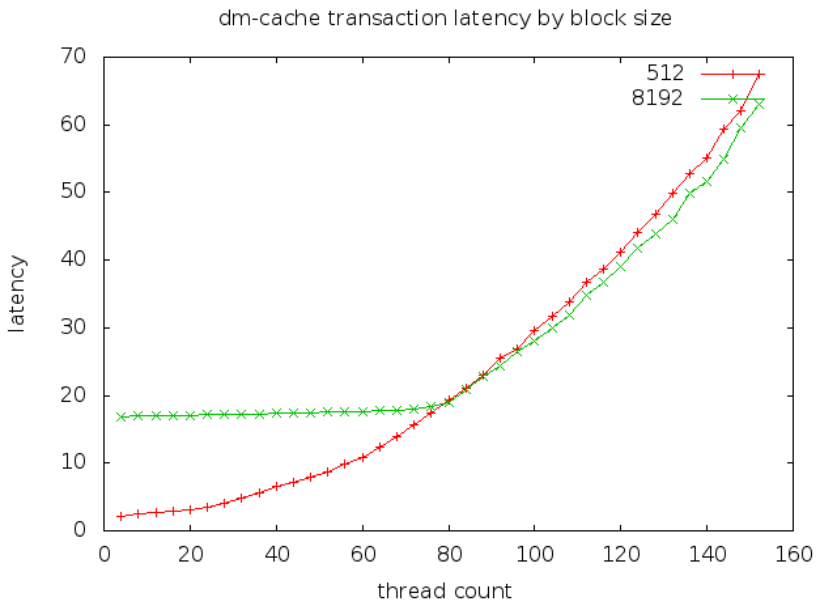
The relatively small size of the database (< 1GB) in these tests allows the entire dataset to fit into the cache. Indeed, hit rates over 90% were observed. If a warm database is assumed, then the long time required for promotion of RBD objects into a cache tier is largely mitigated. Also, cache flush times were very long for the RBD cache tier. This may have an impact in real-world scenarios in which the cache tier is shared with other workloads. For situations in which the cache tier is ideal (the entire dataset fits in the cache tier and demotion of objects is unnecessary), it is unclear why the backing store is necessary. That is, if data can always be resident on the cache tier, then that media could be productively used as direct OSD storage for the relevant dataset.

5.2. dm-cache tuning

The effect of different block sizes for dm-cache was studied to determine if performance could be improved by optimizing them. Smaller block sizes (down to 512 bytes) exhibited better performance before thread saturation, though more memory is required for the relevant data structures. This is likely due to a reduction in sharing and locking contention.

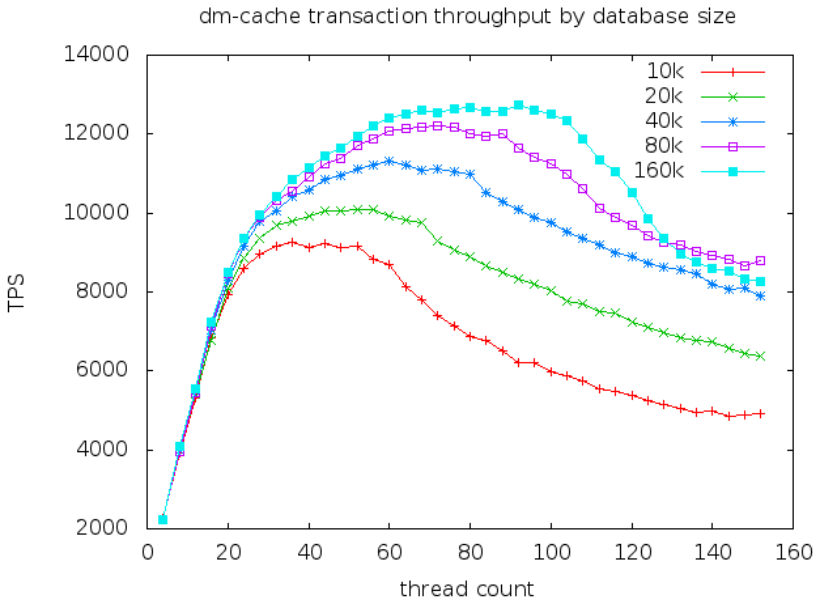


Following thread saturation, larger block sizes exhibited improved throughput, supporting this conclusion.

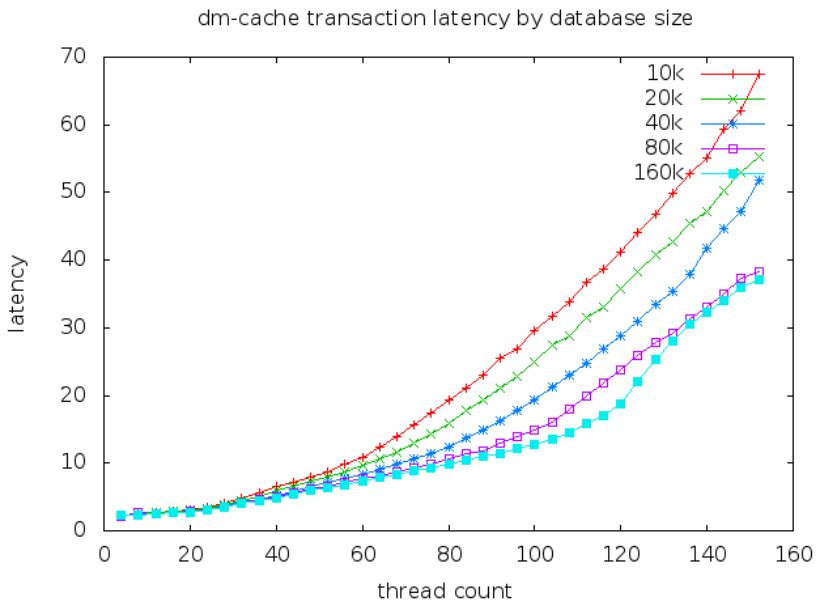


5.3. Database Size

The SysBench MySQL benchmark's default database of 10000 is relatively modest. Throughput on larger databases was measured,



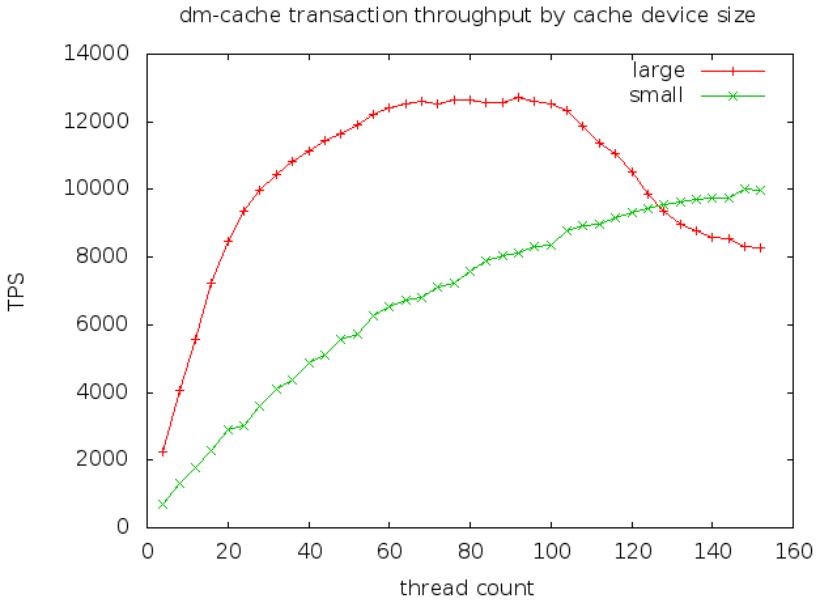
as was latency.



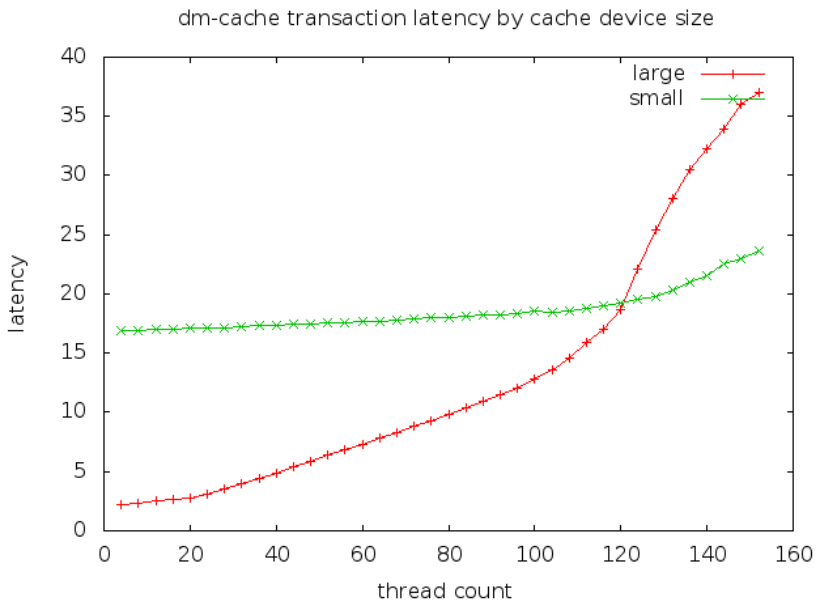
Overall, more throughput was possible with larger databases. Larger databases also exhibited longer latency tails, further supporting the assertion that thread contention is the major performance bottleneck in this scenario.

5.4. Inducing cache misses

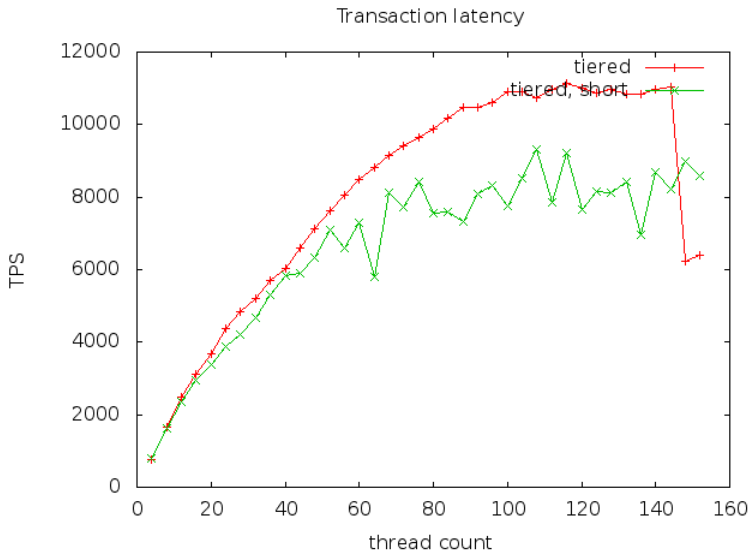
For the cases above, ideal cache conditions were observed (cf. 5.1.2). To observe the effect of reduced hit rates, a test run with a smaller cache device (512 MB) was conducted using a database size of 160000. Hit rates were still relatively high (75-96%), but additional evictions were necessary and throughput dropped significantly.



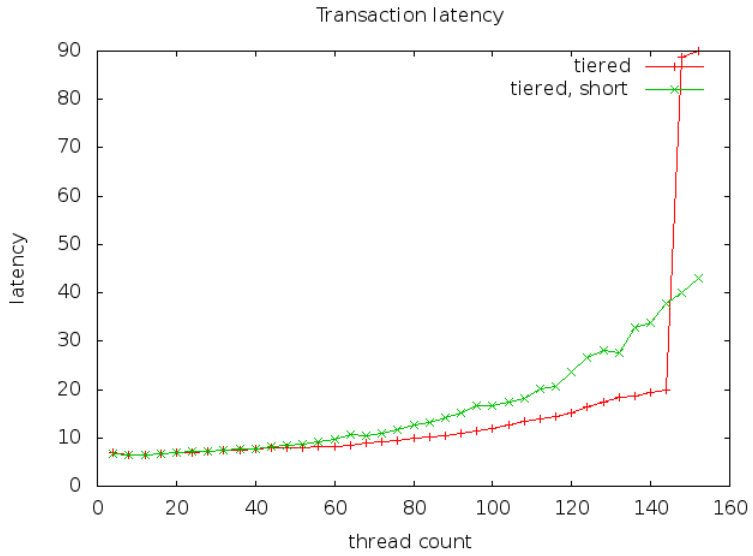
Latency was higher, yet steadier.



Cache misses were induced in the cache tier by reducing the tier pool's target_max_bytes value to 512MB. Like the dm-cache case, there appeared to be few misses, having somewhat less impact on performance.



Latency was similarly affected.



6. Conclusions and future work

For workloads where cache devices are practical and large enough to fit the dataset, both dm-cache and Ceph cache tiering can have significant benefits. Given the cache conditions required for significant performance gains, the benefit of using these fast devices as cache (as opposed to base storage) may be limited. This of course depends on the particular workload and storage conditions under consideration.

The test conditions, as of late 2015, used small amounts of high-performance hardware (late-model Xeon processors, 40Gbps Ethernet, and NVMe storage devices). This hardware configuration is probably not yet typical in the field, but is likely to become more realistic as time progresses.

[1] https://www.kernel.org/doc/Documentation/device-mapper/cache.txt
[2] https://github.com/y-trudeau/cbt
[3] https://github.com/ceph/cbt
[4] https://github.com/akopytov/sysbench
[5] [sic] https://www.percona.com/docs/wiki/benchmark:sysbench:olpt.lua
[6] https://mariadb.com/

I. Configuration files:

dm-cache constructor:

```
dmsetup create rbd-cached --table '0 2147483648 cache /dev/nvme3n1
/dev/nvme2n1p1 /dev/rbd0 512 1 writeback default 0'
```

ceph.conf (default):

```
[global]
        osd pool default size = 1

        osd crush chooseleaf type = 0

        keyring = /home/dfuller/tmp/cbt/ceph/keyring
        osd pg bits = 8
        osd pgp bits = 8
        auth supported = none
        log to syslog = false
        log file = /home/dfuller/tmp/cbt/ceph/log/$name.log
        filestore xattr use omap = true
        auth cluster required = none
        auth service required = none
        auth client required = none

        public network = 10.0.10.0/24
        cluster network = 10.0.10.0/24
        rbd cache = true
        rbd cache writethrough until flush = false
        osd scrub load threshold = 0.01
        osd scrub min interval = 137438953472
        osd scrub max interval = 137438953472
        osd deep scrub interval = 137438953472
        osd max scrubs = 16

        filestore merge threshold = 40
        filestore split multiple = 8
        osd op threads = 8
        osd max throughput = 3500


#        debug_lockdep = "0/0"
#        debug_context = "0/0"
#        debug_crush = "0/0"
#        debug_mds = "0/0"
#        debug_mds_balancer = "0/0"
#        debug_mds_locker = "0/0"
#        debug_mds_log = "0/0"
#        debug_mds_log_expire = "0/0"
#        debug_mds_migrator = "0/0"
#        debug_buffer = "0/0"
#        debug_timer = "0/0"
#        debug_filer = "0/0"
         debug objecter = 1"
#        debug_rados = "0/0"
#        debug_rbd = "0/0"
#        debug_journaler = "0/0"
#        debug_objectcacher = "0/0"
#        debug_client = "0/0"
```

```
         debug osd = 1
#        debug_optracker = "0/0"
#        debug_objclass = "0/0"
#        debug_filestore = "0/0"
#        debug_journal = "0/0"
#        debug_ms = "0/0"
#        debug_mon = "0/0"
#        debug_monc = "0/0"
#        debug_paxos = "0/0"
#        debug_tp = "0/0"
#        debug_auth = "0/0"
#        debug_finisher = "0/0"
#        debug_heartbeatmap = "0/0"
#        debug_perfcounter = "0/0"
#        debug_rgw = "0/0"
#        debug_hadoop = "0/0"
#        debug_asok = "0/0"
#        debug_throttle = "0/0"

        mon pg warn max object skew = 100000
        mon pg warn min per osd = 0
        mon pg warn max per osd = 32768

[client]
      client slo iops reserve = 1000
      client slo iops prop = 1000
      client slo iops limit = 0


[mon]
      mon data = /home/dfuller/tmp/cbt/ceph/mon.$id

[mon.a]
      host = incerta05.front.sepia.ceph.com
        mon addr = 10.0.10.105:6789


[osd.0]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-0-data
        osd journal = /dev/disk/by-partlabel/osd-device-0-journal

[osd.1]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-1-data
        osd journal = /dev/disk/by-partlabel/osd-device-1-journal

[osd.2]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-2-data
        osd journal = /dev/disk/by-partlabel/osd-device-2-journal

[osd.3]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-3-data
        osd journal = /dev/disk/by-partlabel/osd-device-3-journal

[osd.4]
        host = incerta06.front.sepia.ceph.com
```

```
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-4-data
        osd journal = /dev/disk/by-partlabel/osd-device-4-journal

[osd.5]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-5-data
        osd journal = /dev/disk/by-partlabel/osd-device-5-journal

[osd.6]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-6-data
        osd journal = /dev/disk/by-partlabel/osd-device-6-journal

[osd.7]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-0-data
        osd journal = /dev/disk/by-partlabel/osd-device-0-journal

[osd.8]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-1-data
        osd journal = /dev/disk/by-partlabel/osd-device-1-journal

[osd.9]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-2-data
        osd journal = /dev/disk/by-partlabel/osd-device-2-journal

[osd.10]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-3-data
        osd journal = /dev/disk/by-partlabel/osd-device-3-journal

[osd.11]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-4-data
        osd journal = /dev/disk/by-partlabel/osd-device-4-journal

[osd.12]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-5-data
        osd journal = /dev/disk/by-partlabel/osd-device-5-journal

[osd.13]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-6-data
        osd journal = /dev/disk/by-partlabel/osd-device-6-journal

[osd.14]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-0-data
        osd journal = /dev/disk/by-partlabel/osd-device-0-journal

[osd.15]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-1-data
        osd journal = /dev/disk/by-partlabel/osd-device-1-journal

[osd.16]
```

```
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-2-data
        osd journal = /dev/disk/by-partlabel/osd-device-2-journal

[osd.17]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-3-data
        osd journal = /dev/disk/by-partlabel/osd-device-3-journal

[osd.18]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-4-data
        osd journal = /dev/disk/by-partlabel/osd-device-4-journal

[osd.19]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-5-data
        osd journal = /dev/disk/by-partlabel/osd-device-5-journal

[osd.20]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-6-data
        osd journal = /dev/disk/by-partlabel/osd-device-6-journal


ceph.conf (with cache tier)

[global]
        osd pool default size = 1

        osd crush chooseleaf type = 0

        keyring = /home/dfuller/tmp/cbt/ceph/keyring
        osd pg bits = 8
        osd pgp bits = 8
        auth supported = none
        log to syslog = false
        log file = /home/dfuller/tmp/cbt/ceph/log/$name.log
        filestore xattr use omap = true
        auth cluster required = none
        auth service required = none
        auth client required = none

        public network = 10.0.10.0/24
        cluster network = 10.0.10.0/24
        rbd cache = true
        rbd cache writethrough until flush = false
        osd scrub load threshold = 0.01
        osd scrub min interval = 137438953472
        osd scrub max interval = 137438953472
        osd deep scrub interval = 137438953472
        osd max scrubs = 16

        filestore merge threshold = 40
        filestore split multiple = 8
        osd op threads = 8
        osd max throughput = 3500
```

```
#        debug_lockdep = "0/0"
#        debug_context = "0/0"
#        debug_crush = "0/0"
#        debug_mds = "0/0"
#        debug_mds_balancer = "0/0"
#        debug_mds_locker = "0/0"
#        debug_mds_log = "0/0"
#        debug_mds_log_expire = "0/0"
#        debug_mds_migrator = "0/0"
#        debug_buffer = "0/0"
#        debug_timer = "0/0"
#        debug_filer = "0/0"
         debug objecter = 1"
#        debug_rados = "0/0"
#        debug_rbd = "0/0"
#        debug_journaler = "0/0"
#        debug_objectcacher = "0/0"
#        debug_client = "0/0"
         debug osd = 1
#        debug_optracker = "0/0"
#        debug_objclass = "0/0"
#        debug_filestore = "0/0"
#        debug_journal = "0/0"
#        debug_ms = "0/0"
#        debug_mon = "0/0"
#        debug_monc = "0/0"
#        debug_paxos = "0/0"
#        debug_tp = "0/0"
#        debug_auth = "0/0"
#        debug_finisher = "0/0"
#        debug_heartbeatmap = "0/0"
#        debug_perfcounter = "0/0"
#        debug_rgw = "0/0"
#        debug_hadoop = "0/0"
#        debug_asok = "0/0"
#        debug_throttle = "0/0"

        mon pg warn max object skew = 100000
        mon pg warn min per osd = 0
        mon pg warn max per osd = 32768

[client]
      client slo iops reserve = 1000
      client slo iops prop = 1000
      client slo iops limit = 0


[mon]
      mon data = /home/dfuller/tmp/cbt/ceph/mon.$id

[mon.a]
      host = incerta05.front.sepia.ceph.com
        mon addr = 10.0.10.105:6789


[osd.0]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-0-data
        osd journal = /dev/disk/by-partlabel/osd-device-0-journal
```

```
[osd.1]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-1-data
        osd journal = /dev/disk/by-partlabel/osd-device-1-journal

[osd.2]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-2-data
        osd journal = /dev/disk/by-partlabel/osd-device-2-journal

[osd.3]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-3-data
        osd journal = /dev/disk/by-partlabel/osd-device-3-journal

[osd.4]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-4-data
        osd journal = /dev/disk/by-partlabel/osd-device-4-journal

[osd.5]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-5-data
        osd journal = /dev/disk/by-partlabel/osd-device-5-journal

[osd.6]
        host = incerta06.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-6-data
        osd journal = /dev/disk/by-partlabel/osd-device-6-journal

[osd.7]
      host = incerta06.front.sepia.ceph.com
      osd data = /home/dfuller/tmp/cbt/mnt/osd-device-7-data
      osd journal = /dev/disk/by-partlabel/osd-device-7-journal

[osd.8]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-0-data
        osd journal = /dev/disk/by-partlabel/osd-device-0-journal

[osd.9]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-1-data
        osd journal = /dev/disk/by-partlabel/osd-device-1-journal

[osd.10]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-2-data
        osd journal = /dev/disk/by-partlabel/osd-device-2-journal

[osd.11]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-3-data
        osd journal = /dev/disk/by-partlabel/osd-device-3-journal

[osd.12]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-4-data
```

```
        osd journal = /dev/disk/by-partlabel/osd-device-4-journal

[osd.13]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-5-data
        osd journal = /dev/disk/by-partlabel/osd-device-5-journal

[osd.14]
        host = incerta07.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-6-data
        osd journal = /dev/disk/by-partlabel/osd-device-6-journal

[osd.15]
      host = incerta07.front.sepia.ceph.com
      osd data = /home/dfuller/tmp/cbt/mnt/osd-device-7-data
      osd journal = /dev/disk/by-partlabel/osd-device-7-journal

[osd.16]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-0-data
        osd journal = /dev/disk/by-partlabel/osd-device-0-journal

[osd.17]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-1-data
        osd journal = /dev/disk/by-partlabel/osd-device-1-journal

[osd.18]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-2-data
        osd journal = /dev/disk/by-partlabel/osd-device-2-journal

[osd.19]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-3-data
        osd journal = /dev/disk/by-partlabel/osd-device-3-journal

[osd.20]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-4-data
        osd journal = /dev/disk/by-partlabel/osd-device-4-journal

[osd.21]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-5-data
        osd journal = /dev/disk/by-partlabel/osd-device-5-journal

[osd.22]
        host = incerta08.front.sepia.ceph.com
        osd data = /home/dfuller/tmp/cbt/mnt/osd-device-6-data
        osd journal = /dev/disk/by-partlabel/osd-device-6-journal

[osd.23]
      host = incerta08.front.sepia.ceph.com
      osd data = /home/dfuller/tmp/cbt/mnt/osd-device-7-data
      osd journal = /dev/disk/by-partlabel/osd-device-7-journal

CRUSH map used for cache tier:
```

```
# begin crush map
tunable choose_local_tries 0
tunable choose_local_fallback_tries 0
tunable choose_total_tries 50
tunable chooseleaf_descend_once 1
tunable straw_calc_version 1

# devices
device 0 osd.0
device 1 osd.1
device 2 osd.2
device 3 osd.3
device 4 osd.4
device 5 osd.5
device 6 osd.6
device 7 osd.7
device 8 osd.8
device 9 osd.9
device 10 osd.10
device 11 osd.11
device 12 osd.12
device 13 osd.13
device 14 osd.14
device 15 osd.15
device 16 osd.16
device 17 osd.17
device 18 osd.18
device 19 osd.19
device 20 osd.20
device 21 osd.21
device 22 osd.22
device 23 osd.23

# types
type 0 osd
type 1 host
type 2 chassis
type 3 rack
type 4 row
type 5 pdu
type 6 pod
type 7 room
type 8 datacenter
type 9 region
type 10 root

# buckets
host incerta06.front.sepia.ceph.com {
      id -2       # do not change unnecessarily
      # weight 8.000
      alg straw
      hash 0      # rjenkins1
      item osd.0 weight 1.000
      item osd.1 weight 1.000
      item osd.2 weight 1.000
      item osd.3 weight 1.000
      item osd.4 weight 1.000
      item osd.5 weight 1.000
      item osd.6 weight 1.000
```

```
}

host incerta06-ssd {
        id -6
       alg straw
       item osd.7 weight 1.000
}

host incerta07.front.sepia.ceph.com {
       id -4       # do not change unnecessarily
       # weight 8.000
       alg straw
       hash 0      # rjenkins1
       item osd.8 weight 1.000
       item osd.9 weight 1.000
       item osd.10 weight 1.000
       item osd.11 weight 1.000
       item osd.12 weight 1.000
       item osd.13 weight 1.000
       item osd.14 weight 1.000
}

host incerta07-ssd {
        id -7
       alg straw
       item osd.15 weight 1.000
}

host incerta08.front.sepia.ceph.com {
       id -5       # do not change unnecessarily
       # weight 8.000
       alg straw
       hash 0      # rjenkins1
       item osd.16 weight 1.000
       item osd.17 weight 1.000
       item osd.18 weight 1.000
       item osd.19 weight 1.000
       item osd.20 weight 1.000
       item osd.21 weight 1.000
       item osd.22 weight 1.000
}

host incerta08-ssd {
        id -8
        alg straw
       item osd.23 weight 1.000
}

rack localrack {
       id -3       # do not change unnecessarily
       # weight 24.000
       alg straw
       hash 0      # rjenkins1
       item incerta06.front.sepia.ceph.com weight 8.000
       item incerta07.front.sepia.ceph.com weight 8.000
       item incerta08.front.sepia.ceph.com weight 8.000
}

rack ssd-rack {
```

```
        id -9
      alg straw
      hash 0
      item incerta06-ssd
      item incerta07-ssd
      item incerta08-ssd
}

root default {
      id -1       # do not change unnecessarily
      # weight 24.000
      alg straw
      hash 0      # rjenkins1
      item localrack weight 24.000
}

root ssd {
        id -10
      alg straw
      hash 0
      item ssd-rack weight 3.00
}

# rules
rule replicated_ruleset {
      ruleset 0
      type replicated
      min_size 1
      max_size 10
      step take default
      step choose firstn 0 type osd
      step emit
}

rule ssd_ruleset {
        ruleset 1
      type replicated
      min_size 0
      max_size 4
      step take ssd
      step choose firstn 0 type osd
      step emit
}

# end crush map

ceph benchmark tool configuration sample (num-threads and oltp-table-size varied)

cluster:
  user: 'dfuller'
  head: "incerta05.front.sepia.ceph.com"
  clients: ["incerta05.front.sepia.ceph.com"]
  osds: ["incerta06.front.sepia.ceph.com", "incerta07.front.sepia.ceph.com",
"incerta08.front.sepia.ceph.com"]
  mons:
    incerta05.front.sepia.ceph.com:
      a: "10.0.10.105:6789"
  osds_per_node: 7
  fs: 'xfs'
```

```
  mkfs_opts: '-f -i size=2048 -n size=64k'
  mount_opts: '-o inode64,noatime,logbsize=256k'
  conf_file: '/home/dfuller/incerta_conf/ceph.conf'
  iterations: 1
#  use_existing: True
  clusterid: "ceph"
  tmp_dir: "/home/dfuller/tmp/cbt"
#  ceph-osd_cmd: "env -i TCMALLOC_MAX_TOTAL_THREAD_CACHE_BYTES=134217728
/usr/bin/ceph-osd"
  ceph-osd_cmd: "/usr/bin/ceph-osd"
  ceph-mon_cmd: "/usr/bin/ceph-mon"
  ceph-run_cmd: "/usr/bin/ceph-run"
  pool_profiles:
#    radosbench:
#      pg_size: 1024
#      pgp_size: 1024
#      replication: 3
#      replication: 'erasure'
#      erasure_profile: 'ec62'
    rbd:
      pg_size: 8192
      pgp_size: 8192
      replication: 3
  erasure_profiles:
    ec62:
      erasure_k: 6
      erasure_m: 2

benchmarks:
  mysqlsysbench:
    num-threads: 32
    cmd_path: "/home/dfuller/packages/sysbench/sysbench/sysbench"
    test-path: "/home/dfuller/packages/sysbench/sysbench/tests/db/oltp.lua"
    prepare-path:
"/home/dfuller/packages/sysbench/sysbench/tests/db/parallel_prepare.lua"
#    use-local-path: "/mnt/rbd-cached/datadir"
    oltp-table-size: 20000
    max-time: 600

#   nullbench:
#     none:
#  radosbench:
#    op_size: [4194304, 131072, 4096]
#    write_only: False
#    time: 300
#    concurrent_ops: [32]
#    concurrent_procs: 4
#    osd_ra: [4096]
#    pool_profile: 'radosbench'
#  librbdfio:
#    time: 300
#    vol_size: 131072
#    mode: ['read', 'write', 'randread', 'randwrite', 'rw', 'randrw']
#    rwmixread: 50
#    op_size: [4194304, 131072, 4096]
#    procs_per_volume: [1]
#    volumes_per_client: [2]
#    iodepth: [32]
#    osd_ra: [4096]
```

```
#    cmd_path: '/home/ubuntu/src/fio/fio'
#    pool_profile: 'rbd'
#    log_avg_msec: 100
```