Ceph Hammer OSD Shard Tuning

PCIe SSD #1

Mark Nelson

mnelson@redhat.com

2/25/2015

Table of Contents

INTRODUCTION

The Ceph Giant release included several major performance improvements. One particularly interesting improvement in the OSD is a sharded optracker implementation by Somnath Roy from Sandisk. It can dramatically improve OSD performance by increasing the number of operations that the OSD can concurrently process. Benchmarking was performed as part of Somnath's work to determine the default number of shards and number of threads per shard. This document will explore how changing those two settings affects performance in the hammer release of Ceph on an OSD backed by a high performance PCIe SSD. This is done both as a validation of the original findings, and also to determine if there are any benefits to changing the default settings. In a future article, performance across several consumer grade SATA SSDs may also be examined to provide insight into whether different hardware devices require different tuning parameters.

HARDWARE SETUP

A relatively simple system configuration was created. A single test node was employed with a single OSD backed by one PCIe SSD. No replication was used, and RADOS bench was configured to run on the same host as the OSD to limit the effect of network latency. A full description of the hardware follows:

Device	Model
Chassis	Supermicro SC847A
Motherboard	Supermicro X9DRH-7F
Disk Controller	Integrated + LSI SAS9207-8I
CPUS	2 X Intel XEON E5-2630L (2.0GHz, 6-core)
RAM	8 X 4GB Supermicro ECC Registered DDR 1333
NIC	Intel X520-DA2 10GbE (bonded configuration)
PCIe SSD #1	930GB Consumer/Enterprise grade PCIe SSD. Capable of 110K write and 155K read IOPS. Has power failure protection.

GNU parted was used to create a 10GB journal partition at the beginning of the drive with the remainder dedicated to a data storage partition. The “optimal” alignment setting was used to ensure that partition boundaries were well aligned.

SOFTWARE SETUP

A recent pull of Ceph Hammer (from the master branch) was utilized for these tests. Several settings, including disabling in-memory logging, were utilized to reduce overhead and improve performance. In January of 2015, Stephen Blinick from Intel noted that authentication appears to have a large effect on SSD performance in recent versions of Ceph which was verified in subsequent testing. For this reason authentication was disabled during these tests. A list of the software utilized for these tests follows:

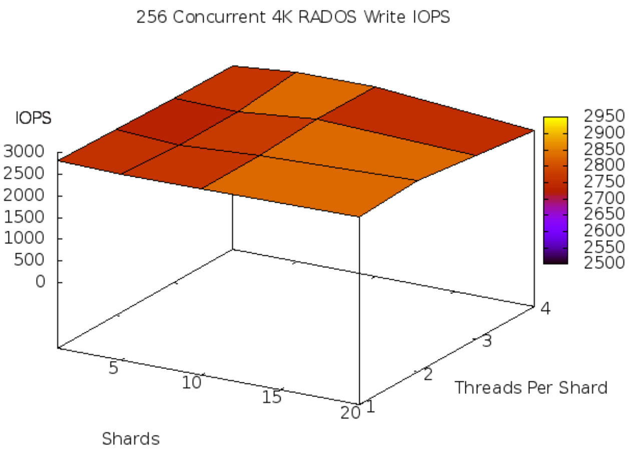
Software	Version
OS	Fedora Core 20
Kernel	3.17.4-200 from source
Ceph Hammer (master)	0.89-465-gb2ca2e2
GNU parted	3.1
CBT	Latest Master

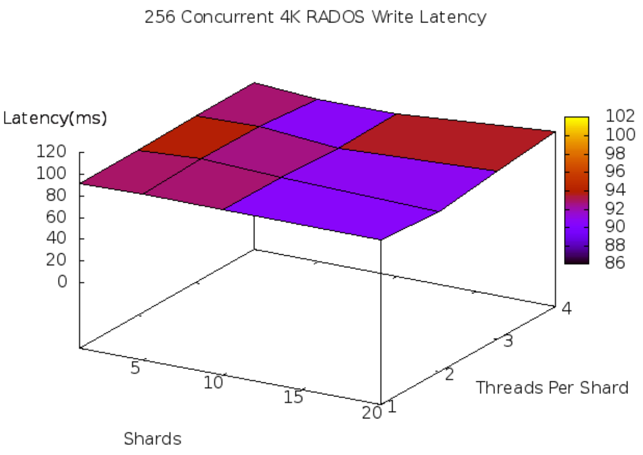
Two Ceph configuration settings were modified during these tests with values as follows:

Setting	Values
osd _op_num_shards	1, 5 (default), 10, 20
osd_op_num_threads_per_shard	1, 2 (default), 3, 4

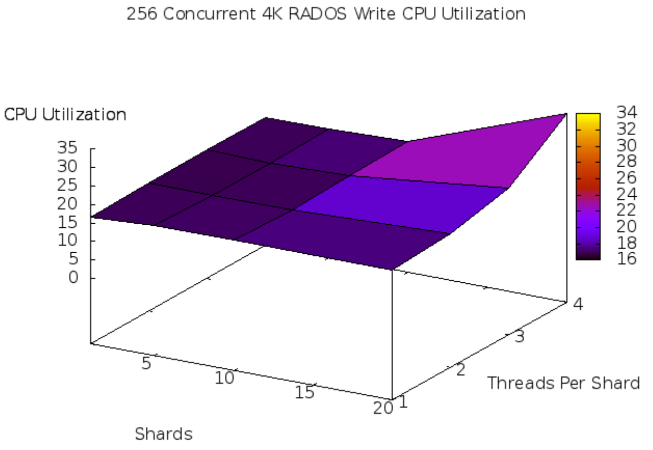
A python script called makecephconf.py in the CBT distribution was used to create ceph.conf files covering the entire parameter space. A bash script was also generated that repeatedly called CBT with the each associated ceph.conf file. See Appendices A and B for details regarding the CBT, and makecephconf.py configuration files used to generate and run these tests.

256 CONCURRENT 4K WRITE RESULTS

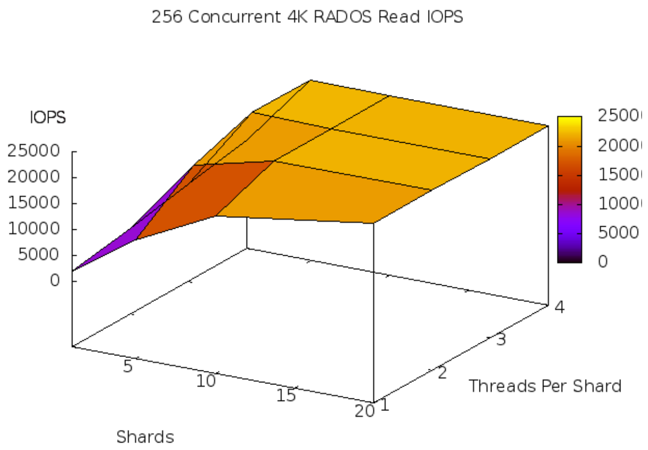


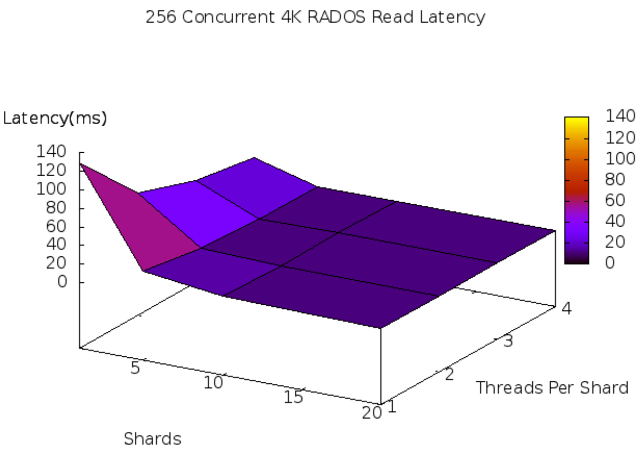


At least for the purposes of concurrent 4K writes, the two OSD shard tuning parameters that were examined in this test do not have a significant performance impact. Other bottlenecks such as lock serialization are limiting performance. CPU Utilization remains fairly constant until high combinations of threads/shards where CPU usage increases accompanied by potentially a minor performance loss.

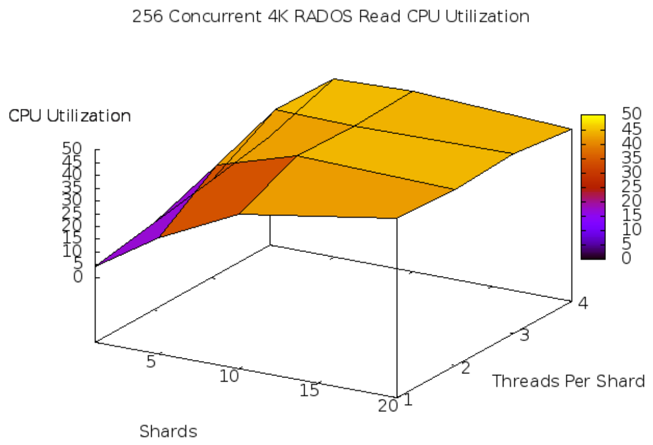


256 CONCURRENT 4K READ RESULTS

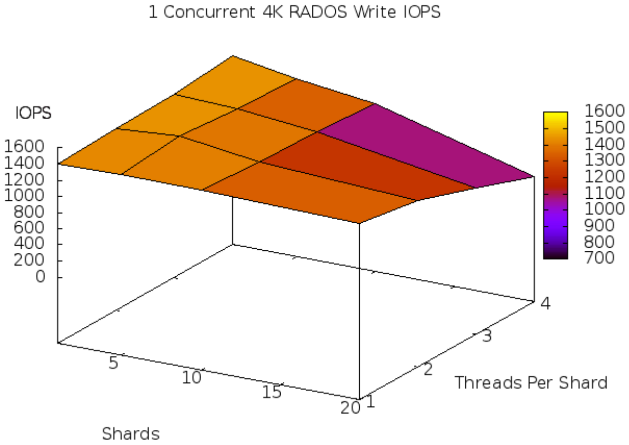


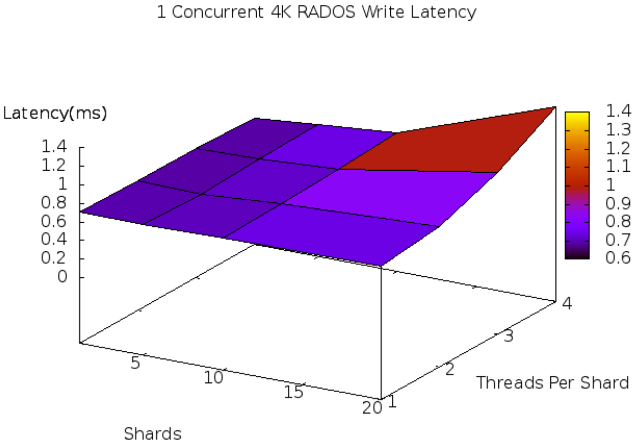


Concurrent read tests on the other hand see a dramatic improvement as the number of shards and threads per shard increase up to around 22K IOPS. At least on this hardware, there appears to be roughly a 10% performance improvement when increasing either the number of shards or the number of threads per shard over the default values. There does not appear to any noticeable performance drop at high shard or threads per shard combinations. CPU Usage during 4K reads neared 50% across all cores. A manual inspection of the CPU utilization across all cores (not shown) indicates that some cores are occasionally getting close to 90-95% utilization, though no cores appear to be consistently oversubscribed.

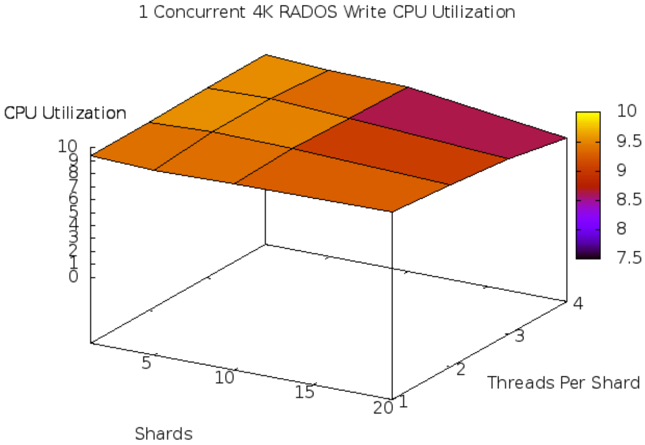


SINGLE OP 4K WRITE RESULTS

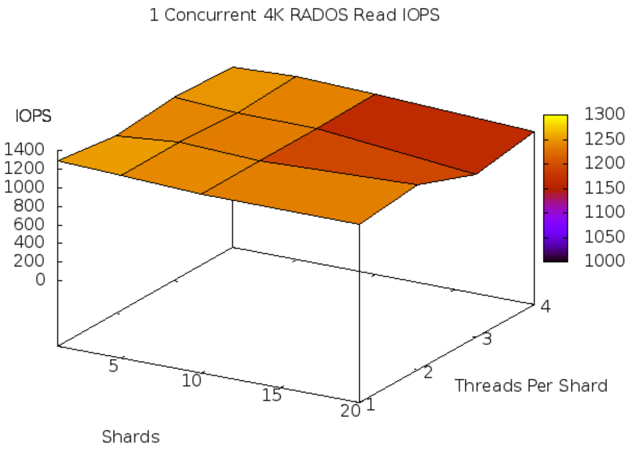


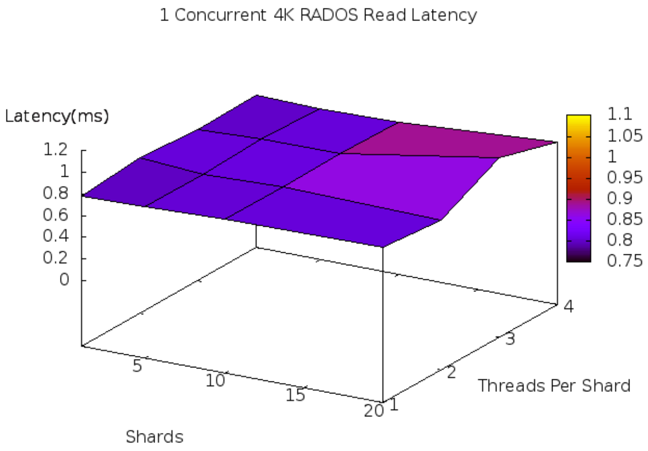


Considerable overhead was observed at high shard/thread counts when only a single write is issued. The default value of 5 shards and 2 threads per shard appears to be a good compromise between highly concurrent workloads and single op workloads. While latency increased with increased shard/thread counts, CPU utilization decreased.

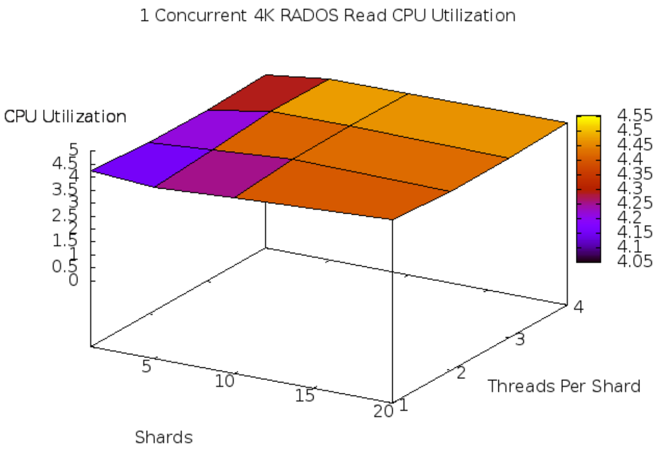


SINGLE OP 4K READ RESULTS





Single read tests showed a slight performance degradation at high shard and threads per shard counts. CPU Utilization may have increased slightly with high shard/thread counts though the effect was minor.



CONCLUSION

On this hardware configuration, the default OSD shard and threads per shard tuning parameters appear to be well chosen. A 10% improvement in concurrent read performance may be gained by increasing the number of shards or threads per shard, though potentially at the expense of higher single operation write latency. This is especially true when these settings are configured to be significantly higher than default. Lowering the default values potentially can dramatically decrease concurrent read performance. The node used in this testing has 12 physical cores and it may be that simply matching the total number of shards/threads (across all OSDs) to the number of cores tends to produce the best overall results.

APPENDIX A. CBT CONFIGURATION

```
cluster:
  user: 'nhm'
  head: "burnupiX"
  clients: ["burnupiX"]
  osds: ["burnupiX"]
  mons:
    burnupiX:
      a: "192.168.10.1:6789"
  osds_per_node: 1
  fs: 'xfs'
  mkfs_opts: '-f -i size=2048 -n size=64k -K'
  mount_opts: '-o inode64,noatime,logbsize=256k'
  conf_file: '/home/nhm/src/ceph-tools/regression/test/memstore/ceph.conf'
  iterations: 3
  use_existing: False
  clusterid: "ceph"
  pool_profiles:
    radosbench:
      pg_size: 1024
      pgp_size: 1024
      replication: 1

benchmarks:
  radosbench:
    op_size: [4096]
    write_only: False
    time: 30
    concurrent_ops: [1, 256]
    concurrent_procs: 1
    osd_ra: [4096]
    pool_profile: 'radosbench'
```

APPENDIX B. MAKECEPHCONF.PY CONFIGURATION

```
settings:
    osd_servers: [burnupiX]
    osds_per_server: 1

    outdir: "/home/nhm/data/sharding"
    runtests_exec: "/home/nhm/src/ceph-tools/cbt/cbt.py"
    runtests_conf:
        xfs: "/home/nhm/src/ceph-tools/cbt/sharding/runtests.xfs.yaml"


default:
  global:
    osd_crush_update_on_start: "false"
    osd_crush_chooseleaf_type: "0"
    osd_pg_bits: "10"
    osd_pgp_bits: "10"
    osd_pool_default_size: "1"
    auth_cluster_required: "none"
    auth_service_required: "none"
    auth_client_required: "none"
    keyring: "/tmp/cbt/ceph/log/$name.log"
    log_to_syslog: "false"
    log_file: "/tmp/cbt/ceph/log/$name.log"
    rbd_cache: "true"
    filestore_merge_threshold: "40"
    filestore_split_threshold: "8"
    osd_op_threads: "8"
    mon_pg_warn_max_object_skew: "100000"
    mon_pg_warn_min_per_osd: "0"
    mon_pg_warn_max_per_osd: "32768"
    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: "0/0"

    debug_rados: "0/0"

    debug_rbd: "0/0"

    debug_journaler: "0/0"

    debug_objectcacher: "0/0"

    debug_client: "0/0"

    debug_osd: "0/0"

    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.a:
        host: "burnupiX"
        mon_addr: "192.168.10.1:6789"
        mon_data: "/tmp/cbt/ceph/mon.$id"

parametric:
  osd_op_num_shards: [1, 5, 10, 20]
  osd_op_num_threads_per_shard: [1, 2, 3, 4]
```