# Cep Performance Tools

Haomai Wang <haomai@xsky.com>
@yuyuyu101 at Github

# Haomai

- Join Ceph community since 2013
- Focus on performance mainly
- IO Stack under block interface
- Familiar with filesystem, database and cloud

# Agenda

- Performance Pain
- Tools
- Tools Types
    - Observability
    - Benchmarking
    - Tuning
- Case study

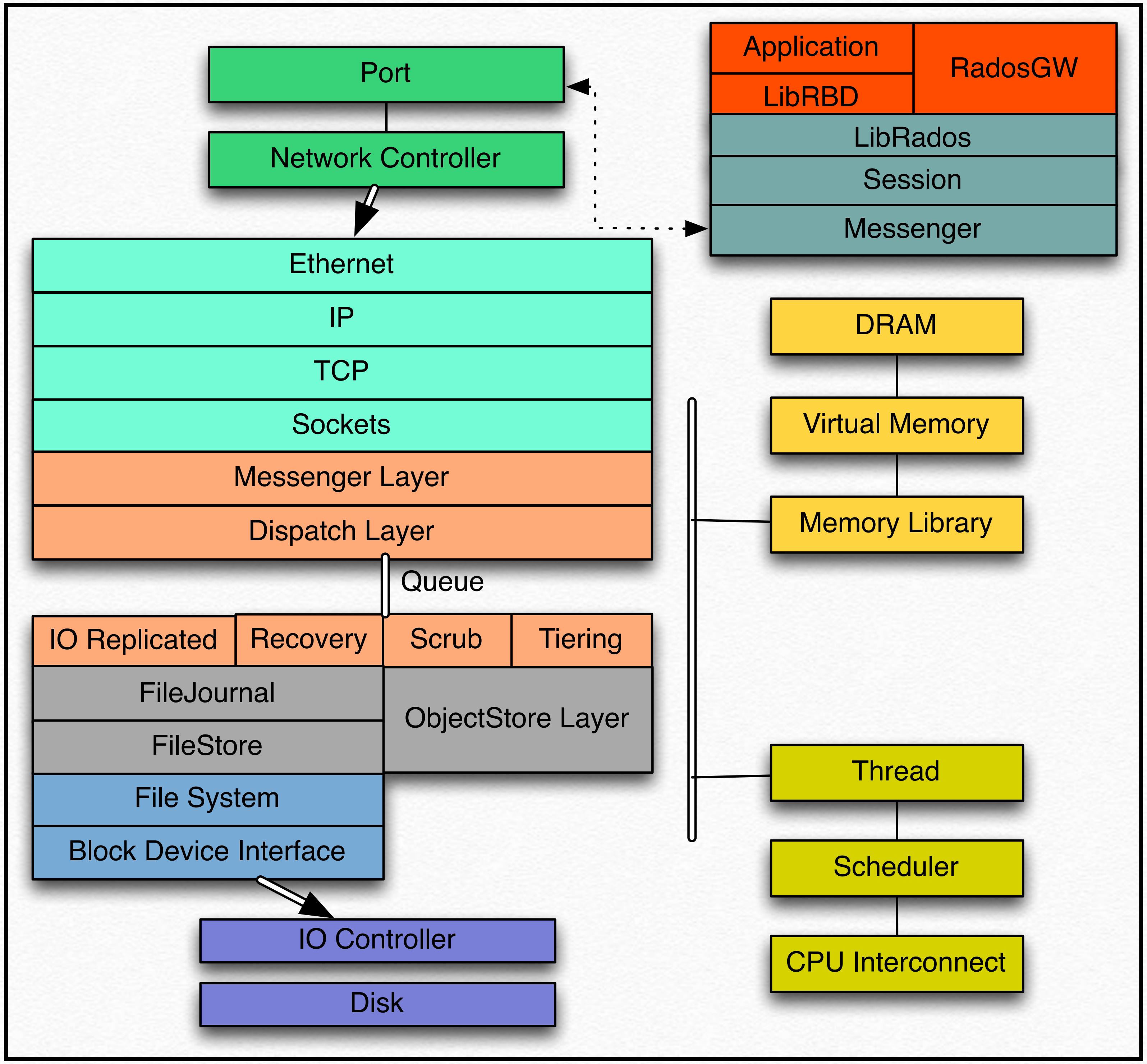
# My Ceph is slow…

- Questions on maillist:
  - Why I only get so little IOPS?
  - Did maximum performance reached?
  - Investigating my 100 IOPS limit?
- Potential Solutions:
  - Google “Ceph Performance Best Practice”
  - Take a chance to tune cep config value by experience
  - Random guess the problem then change things until it goes away

# Methodologies

- Performance tools for Linux, Ceph
- Problem statement
- Workload characterization
- Utilization, saturation check*
- Benchmarking*
- Tuning*

# How do we measure these?

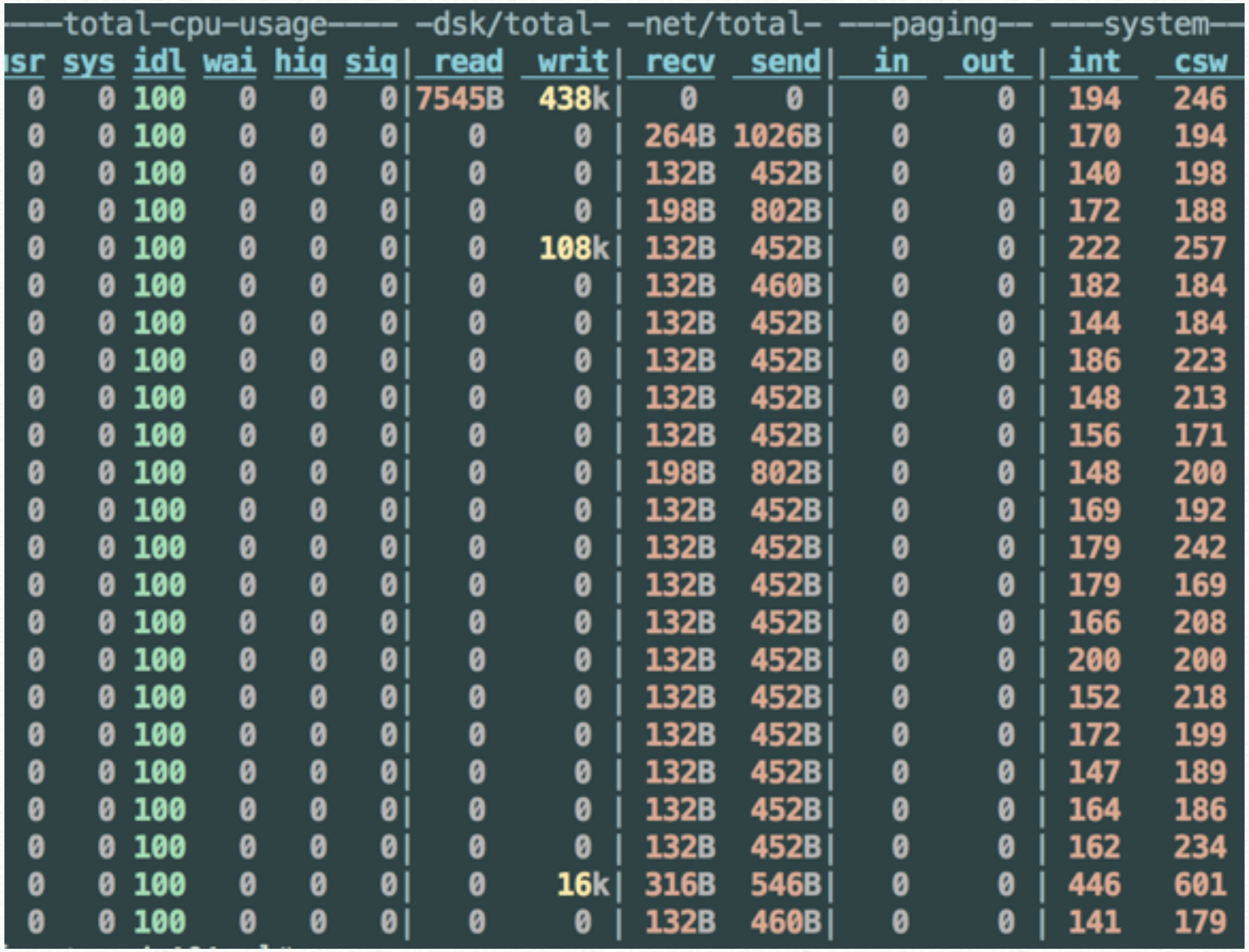


# Observability Tools

- Basic:
  - htop
  - dstat
  - iostat
  - iptraf
  - netstat
  - ceph -w
  - ceph tell osd.* heap stats

- Advance:
  - strace
  - blktrace
  - tcpdump
  - perf
  - systemtap/lttng
  - ceph daemon osd.* dump_historic_ops
  - ceph osd perf
  - ceph perf dump/ceph daemonperf(Infernalis release)

# dstat



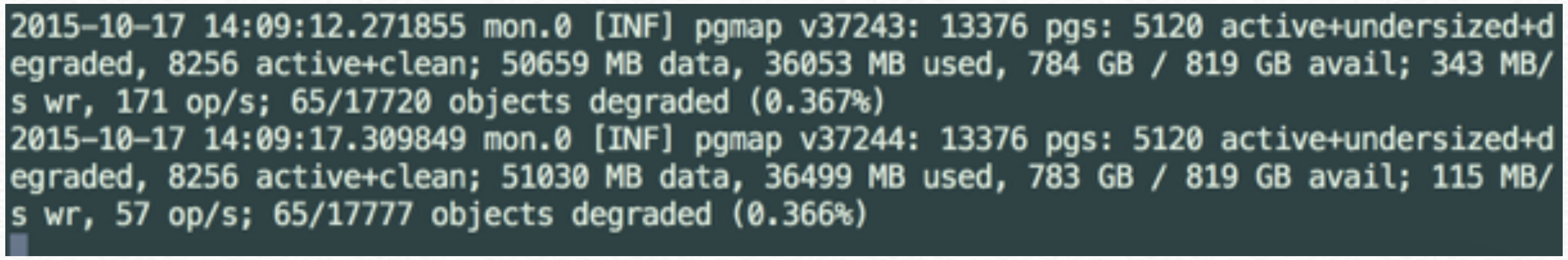
```
----total-cpu-usage---- -dsk/total- -net/total- ---paging-- ---system--
sr sys idl wai hiq siq| read  writ| recv  send|  in   out | int   csw
 0   0 100   0   0   0|7545B  438k|   0     0 |   0     0 | 194   246
 0   0 100   0   0   0|   0     0 | 264B 1026B|   0     0 | 170   194
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 140   198
 0   0 100   0   0   0|   0     0 | 198B  802B|   0     0 | 172   188
 0   0 100   0   0   0|   0   108k| 132B  452B|   0     0 | 222   257
 0   0 100   0   0   0|   0     0 | 132B  460B|   0     0 | 182   184
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 144   184
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 186   223
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 148   213
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 156   171
 0   0 100   0   0   0|   0     0 | 198B  802B|   0     0 | 148   200
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 169   192
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 179   242
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 179   169
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 166   208
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 200   200
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 152   218
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 172   199
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 147   189
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 164   186
 0   0 100   0   0   0|   0     0 | 132B  452B|   0     0 | 162   234
 0   0 100   0   0   0|   0    16k| 316B  546B|   0     0 | 446   601
 0   0 100   0   0   0|   0     0 | 132B  460B|   0     0 | 141   179
```

# iostat/iptraf

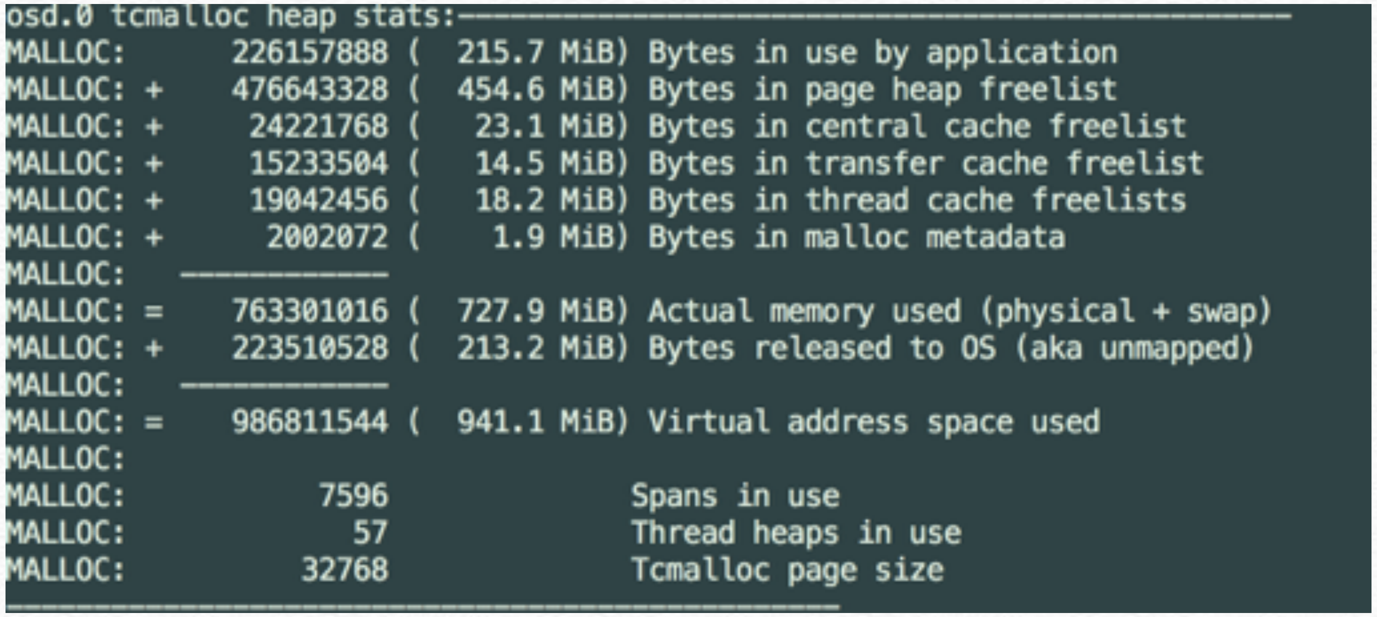
- iostat
  - await
  - util
- iptraf
  - Package size
  - Packages

# ceph -w/heap stats

- set “osd_op_complaint_time” to 1(or lower)
- ops
- slow requests

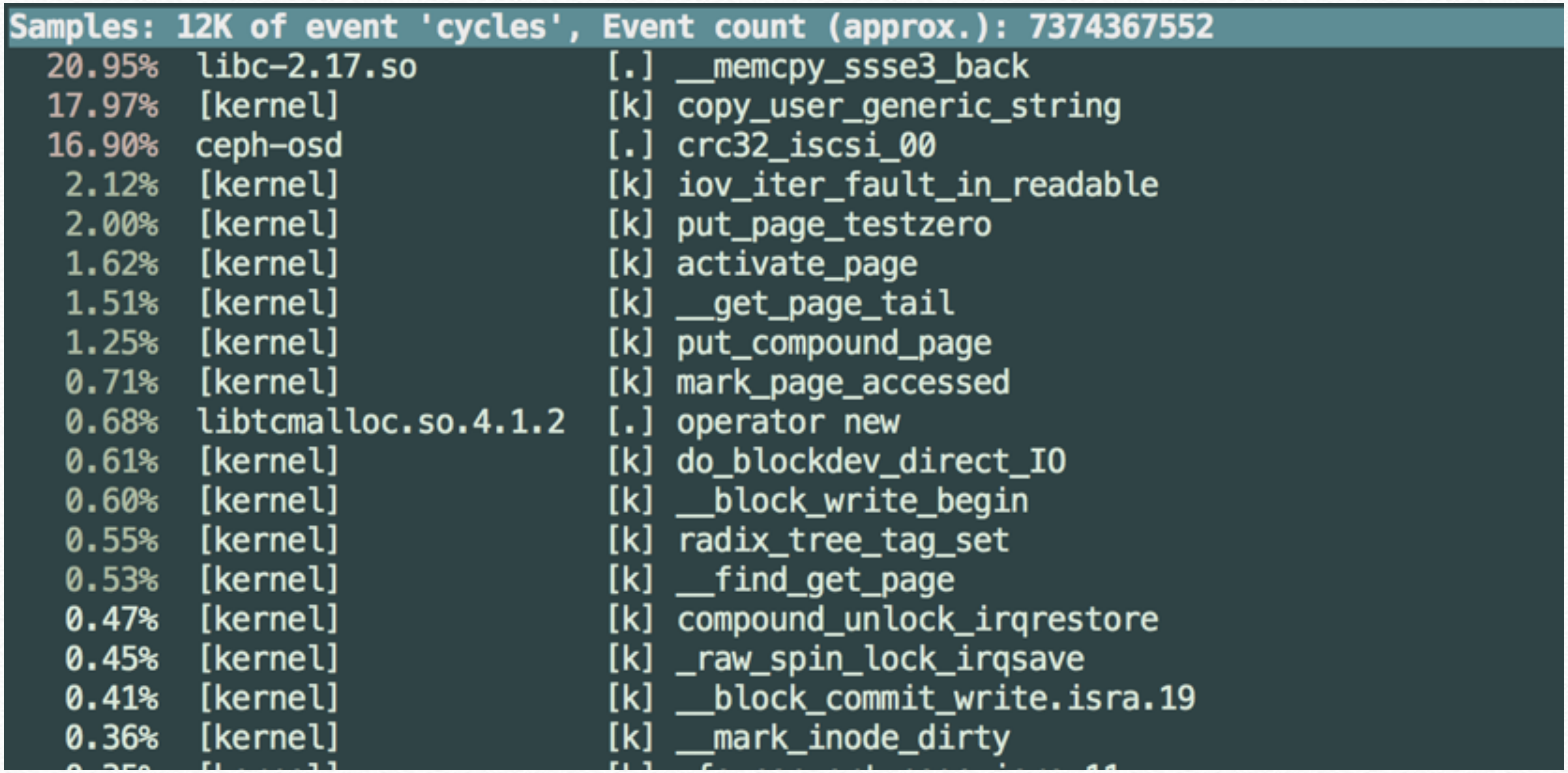


```
2015-10-17 14:09:12.271855 mon.0 [INF] pgmap v37243: 13376 pgs: 5120 active+undersized+d
egraded, 8256 active+clean; 50659 MB data, 36053 MB used, 784 GB / 819 GB avail; 343 MB/
s wr, 171 op/s; 65/17720 objects degraded (0.367%)
2015-10-17 14:09:17.309849 mon.0 [INF] pgmap v37244: 13376 pgs: 5120 active+undersized+d
egraded, 8256 active+clean; 51030 MB data, 36499 MB used, 783 GB / 819 GB avail; 115 MB/
s wr, 57 op/s; 65/17777 objects degraded (0.366%)
```



```
osd.0 tcmalloc heap stats:------------------------------------------------
MALLOC:      226157888 (  215.7 MiB) Bytes in use by application
MALLOC: +    476643328 (  454.6 MiB) Bytes in page heap freelist
MALLOC: +     24221768 (   23.1 MiB) Bytes in central cache freelist
MALLOC: +     15233504 (   14.5 MiB) Bytes in transfer cache freelist
MALLOC: +     19042456 (   18.2 MiB) Bytes in thread cache freelists
MALLOC: +      2002072 (    1.9 MiB) Bytes in malloc metadata
MALLOC:   ------------
MALLOC: =    763301016 (  727.9 MiB) Actual memory used (physical + swap)
MALLOC: +    223510528 (  213.2 MiB) Bytes released to OS (aka unmapped)
MALLOC:   ------------
MALLOC: =    986811544 (  941.1 MiB) Virtual address space used
MALLOC:
MALLOC:           7596              Spans in use
MALLOC:             57              Thread heaps in use
MALLOC:          32768              Tcmalloc page size
------------------------------------------------
```

# perf

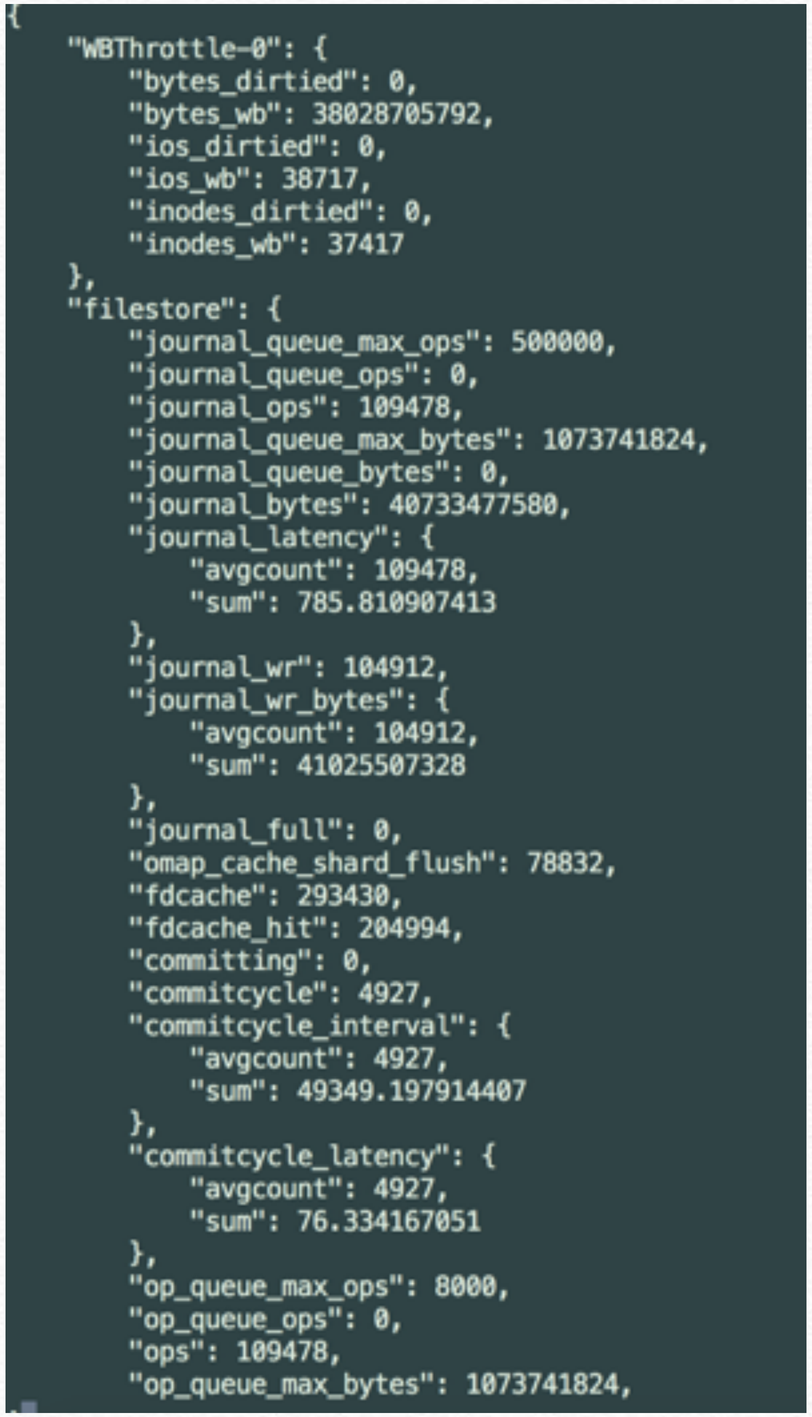


```
Samples: 12K of event 'cycles', Event count (approx.): 7374367552
  20.95%  libc-2.17.so          [.] __memcpy_ssse3_back
  17.97%  [kernel]              [k] copy_user_generic_string
  16.90%  ceph-osd              [.] crc32_iscsi_00
   2.12%  [kernel]              [k] iov_iter_fault_in_readable
   2.00%  [kernel]              [k] put_page_testzero
   1.62%  [kernel]              [k] activate_page
   1.51%  [kernel]              [k] __get_page_tail
   1.25%  [kernel]              [k] put_compound_page
   0.71%  [kernel]              [k] mark_page_accessed
   0.68%  libtcmalloc.so.4.1.2  [.] operator new
   0.61%  [kernel]              [k] do_blockdev_direct_IO
   0.60%  [kernel]              [k] __block_write_begin
   0.55%  [kernel]              [k] radix_tree_tag_set
   0.53%  [kernel]              [k] __find_get_page
   0.47%  [kernel]              [k] compound_unlock_irqrestore
   0.45%  [kernel]              [k] _raw_spin_lock_irqsave
   0.41%  [kernel]              [k] __block_commit_write.isra.19
   0.36%  [kernel]              [k] __mark_inode_dirty
```

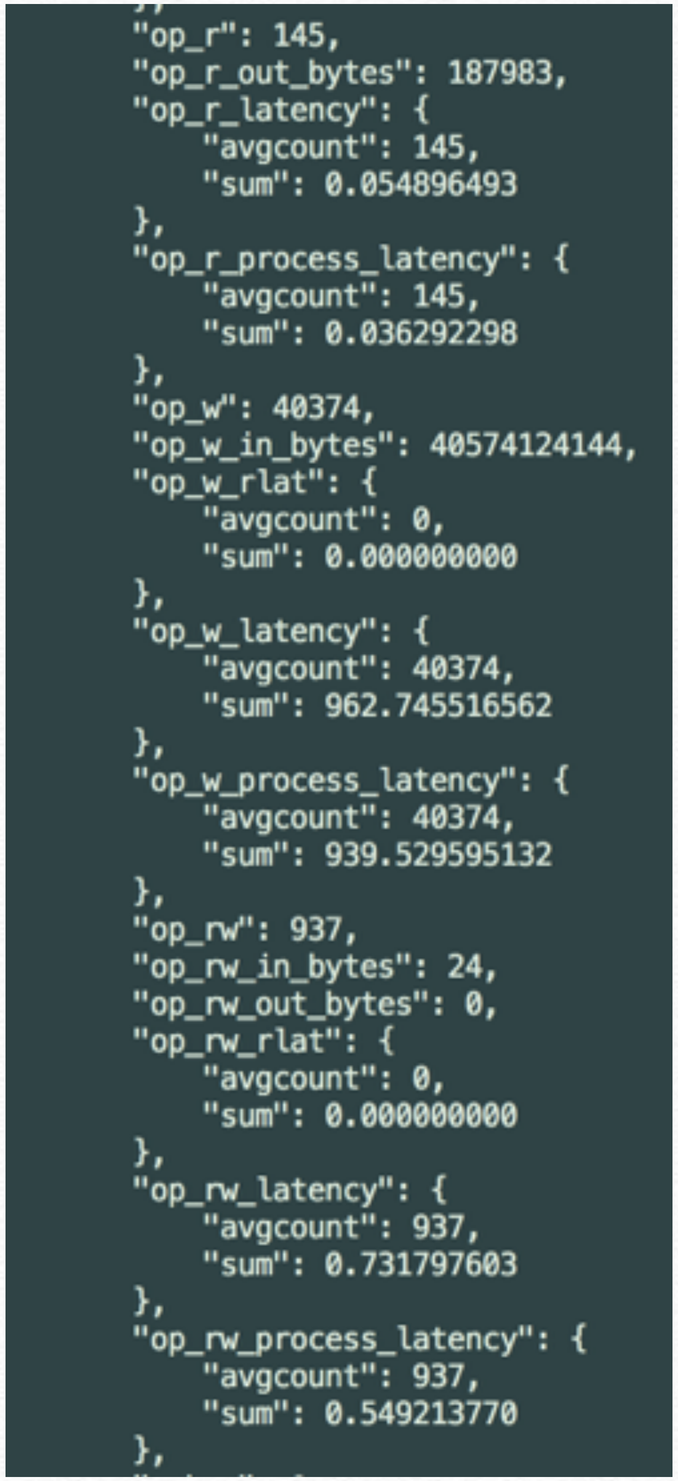
# systemtap/lttng

```
global do_op
probe process("/root/ceph-0.94.1/src/ceph-osd").function("do_op@osd/ReplicatedPG.cc").return
{
  do_op <<< gettimeofday_us() - @entry(gettimeofday_us())
}

global eval_repop
probe process("/root/ceph-0.94.1/src/ceph-osd").function("eval_repop@osd/ReplicatedPG.cc").return
{
  eval_repop <<< gettimeofday_us() - @entry(gettimeofday_us())
}

global submit_transaction
probe process("/root/ceph-0.94.1/src/ceph-osd").function("submit_transaction@osd/ReplicatedBackend.cc").return
{
  submit_transaction <<< gettimeofday_us() - @entry(gettimeofday_us())
}
```

# ceph perf dump

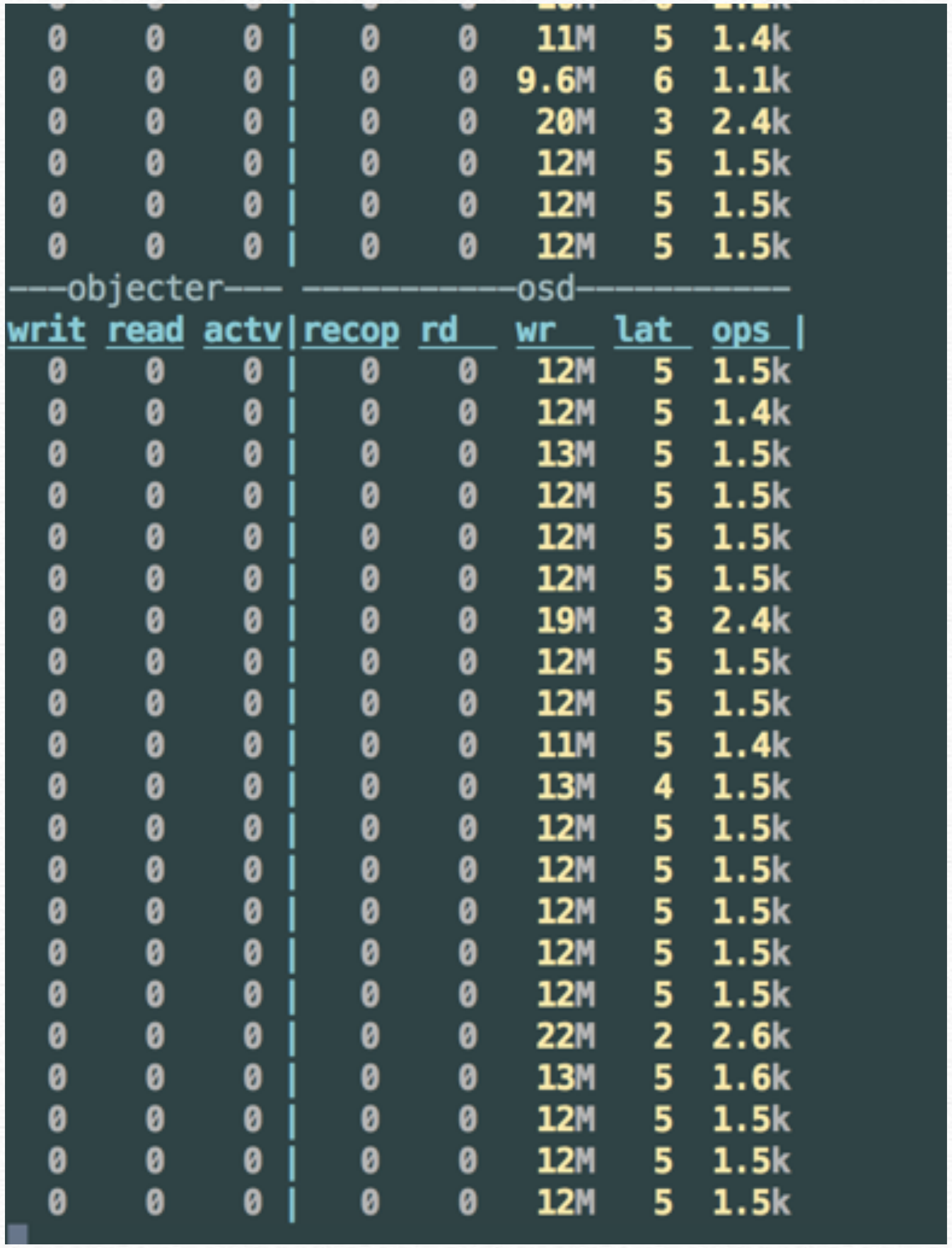


```
{
    "WBThrottle-0": {
        "bytes_dirtied": 0,
        "bytes_wb": 38028705792,
        "ios_dirtied": 0,
        "ios_wb": 38717,
        "inodes_dirtied": 0,
        "inodes_wb": 37417
    },
    "filestore": {
        "journal_queue_max_ops": 500000,
        "journal_queue_ops": 0,
        "journal_ops": 109478,
        "journal_queue_max_bytes": 1073741824,
        "journal_queue_bytes": 0,
        "journal_bytes": 40733477580,
        "journal_latency": {
            "avgcount": 109478,
            "sum": 785.810907413
        },
        "journal_wr": 104912,
        "journal_wr_bytes": {
            "avgcount": 104912,
            "sum": 41025507328
        },
        "journal_full": 0,
        "omap_cache_shard_flush": 78832,
        "fdcache": 293430,
        "fdcache_hit": 204994,
        "committing": 0,
        "commitcycle": 4927,
        "commitcycle_interval": {
            "avgcount": 4927,
            "sum": 49349.197914407
        },
        "commitcycle_latency": {
            "avgcount": 4927,
            "sum": 76.334167051
        },
        "op_queue_max_ops": 8000,
        "op_queue_ops": 0,
        "ops": 109478,
        "op_queue_max_bytes": 1073741824,
```



```
        },
        "op_r": 145,
        "op_r_out_bytes": 187983,
        "op_r_latency": {
            "avgcount": 145,
            "sum": 0.054896493
        },
        "op_r_process_latency": {
            "avgcount": 145,
            "sum": 0.036292298
        },
        "op_w": 40374,
        "op_w_in_bytes": 40574124144,
        "op_w_rlat": {
            "avgcount": 0,
            "sum": 0.000000000
        },
        "op_w_latency": {
            "avgcount": 40374,
            "sum": 962.745516562
        },
        "op_w_process_latency": {
            "avgcount": 40374,
            "sum": 939.529595132
        },
        "op_rw": 937,
        "op_rw_in_bytes": 24,
        "op_rw_out_bytes": 0,
        "op_rw_rlat": {
            "avgcount": 0,
            "sum": 0.000000000
        },
        "op_rw_latency": {
            "avgcount": 937,
            "sum": 0.731797603
        },
        "op_rw_process_latency": {
            "avgcount": 937,
            "sum": 0.549213770
        },
```

# ceph daemonperf



```
   0    0    0 |    0    0   11M    5  1.4k
   0    0    0 |    0    0  9.6M    6  1.1k
   0    0    0 |    0    0   20M    3  2.4k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
---objecter--- -----------osd-----------
writ read actv|recop rd    wr    lat  ops |
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.4k
   0    0    0 |    0    0   13M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   19M    3  2.4k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   11M    5  1.4k
   0    0    0 |    0    0   13M    4  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   22M    2  2.6k
   0    0    0 |    0    0   13M    5  1.6k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
   0    0    0 |    0    0   12M    5  1.5k
```

# Benchmarking

- Your workload
- Benchmark A but actually measure B
- Running benchmark with observability tools

# Benchmarking Tools

- Generic
  - Fio with librbd engine
  - Cosbench with S3
- Ceph specified
  - rbd-replay
  - rados/rbd bench
- Component:
  - Hardware/OS/Library: ceph_perf_local
  - Messenger: ceph_perf_msgr_client/ceph_perf_msgr_server
  - ObjectStore
    - Fio with objectstore engine
    - ceph_perf_objectstore
  - Erasure Code: ceph_erasure_code_benchmark

# Fio with librbd/objectstore

```
[global]
ioengine=rbd
clientname=admin
pool=rbd
rbdname=fio_test
invalidate=0 # mandatory
rw=randwrite
bs=4k

[rbd_iodepth32]
iodepth=32
```

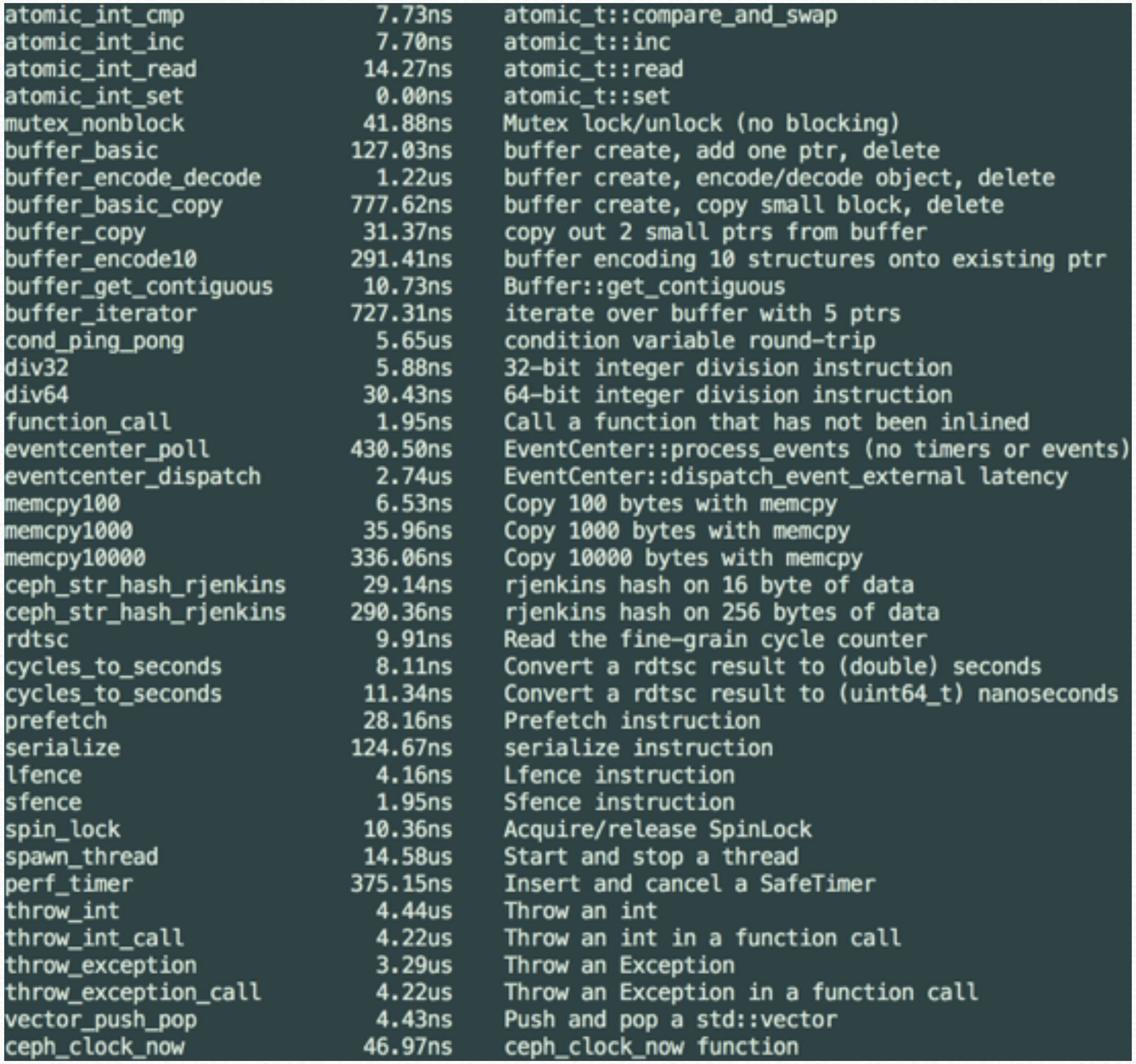
```
[global]
ioengine=libfio_ceph_objectstore.so
invalidate=0 # mandatory
rw=randwrite
size=1g
bs=4k

[ceph_objectstore]
iodepth=1
objectstore=filestore
#filestore_debug=20
directory=/mnt/fio_ceph_filestore
filestore_journal=/var/lib/ceph/osd/j
```

# rbd-replay

- Trace actual workload:
  - lttng create -o traces librbd
  - lttng enable-event -u 'librbd:*'
  - lttng add-context -u -t pthread_id
  - lttng start
  - ….
  - lttng stop
  - lttng view > trace.log
- rbd-replay-prep

- Replay
- lttng create && lttng enable-event -u 'librbd:*'
- lttng add-context -u -t pthread_id
- lttng start
- rbd-replay —conf=/etc/ceph/ceph.conf replay.bin "$@" | tee replay.log
- lttng stop
- lttng view > replay-trace.log

# ceph_perf_local



```
atomic_int_cmp                7.73ns    atomic_t::compare_and_swap
atomic_int_inc                7.70ns    atomic_t::inc
atomic_int_read              14.27ns    atomic_t::read
atomic_int_set                0.00ns    atomic_t::set
mutex_nonblock               41.88ns    Mutex lock/unlock (no blocking)
buffer_basic                127.03ns    buffer create, add one ptr, delete
buffer_encode_decode          1.22us    buffer create, encode/decode object, delete
buffer_basic_copy           777.62ns    buffer create, copy small block, delete
buffer_copy                  31.37ns    copy out 2 small ptrs from buffer
buffer_encode10             291.41ns    buffer encoding 10 structures onto existing ptr
buffer_get_contiguous        10.73ns    Buffer::get_contiguous
buffer_iterator             727.31ns    iterate over buffer with 5 ptrs
cond_ping_pong                5.65us    condition variable round-trip
div32                         5.88ns    32-bit integer division instruction
div64                        30.43ns    64-bit integer division instruction
function_call                 1.95ns    Call a function that has not been inlined
eventcenter_poll            430.50ns    EventCenter::process_events (no timers or events)
eventcenter_dispatch          2.74us    EventCenter::dispatch_event_external latency
memcpy100                     6.53ns    Copy 100 bytes with memcpy
memcpy1000                   35.96ns    Copy 1000 bytes with memcpy
memcpy10000                 336.06ns    Copy 10000 bytes with memcpy
ceph_str_hash_rjenkins       29.14ns    rjenkins hash on 16 byte of data
ceph_str_hash_rjenkins      290.36ns    rjenkins hash on 256 bytes of data
rdtsc                         9.91ns    Read the fine-grain cycle counter
cycles_to_seconds             8.11ns    Convert a rdtsc result to (double) seconds
cycles_to_seconds            11.34ns    Convert a rdtsc result to (uint64_t) nanoseconds
prefetch                     28.16ns    Prefetch instruction
serialize                   124.67ns    serialize instruction
lfence                        4.16ns    Lfence instruction
sfence                        1.95ns    Sfence instruction
spin_lock                    10.36ns    Acquire/release SpinLock
spawn_thread                 14.58us    Start and stop a thread
perf_timer                  375.15ns    Insert and cancel a SafeTimer
throw_int                     4.44us    Throw an int
throw_int_call                4.22us    Throw an int in a function call
throw_exception               3.29us    Throw an Exception
throw_exception_call          4.22us    Throw an Exception in a function call
vector_push_pop               4.43ns    Push and pop a std::vector
ceph_clock_now               46.97ns    ceph_clock_now function
```

# ceph_perf_msgr/ ceph_perf_client

```
#./ceph_perf_msgr_server 172.16.30.181:10001 0
using ms-type async
bind ip:port 172.16.30.181:10001
thinktime(us) 0

#./ceph_perf_msgr_client 172.16.30.181:10001 1 32 10000 10 4096
using ms-type async
server ip:port 172.16.30.181:10001
numjobs 1
concurrency 32
ios 10000
thinktime(us) 10
message data bytes 4096

Total op 10000 run time 852670us.
```

# Tuning tools

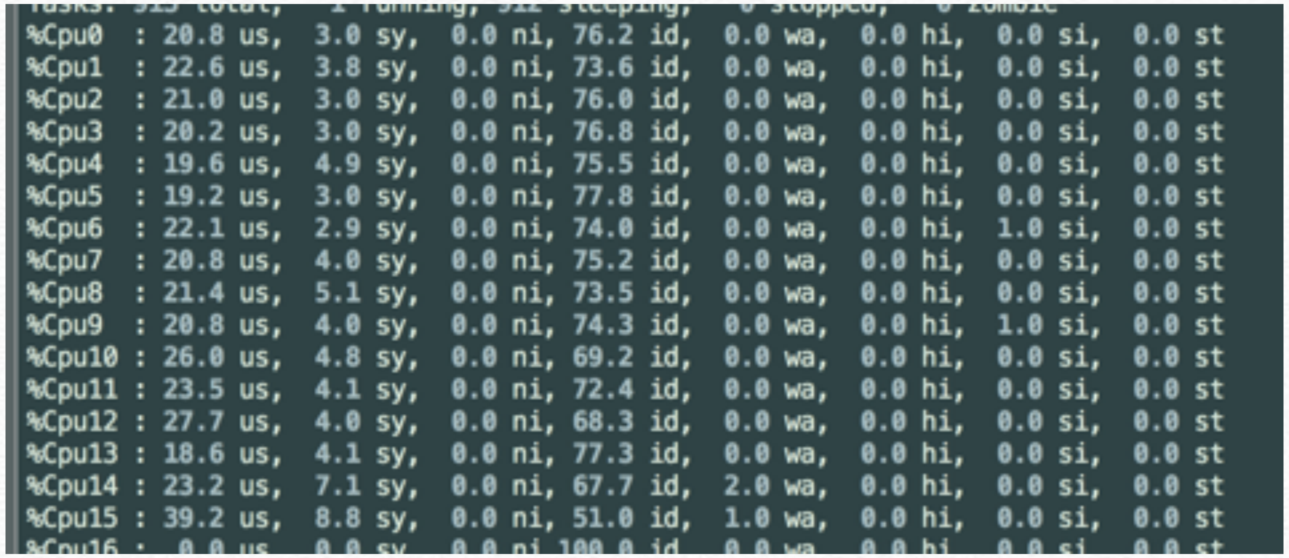
- OS
    - sysctl, /sys
    - cgroup/cpu frequency
    - mkfs/tune2fs
- Ceph
    - filestore
    - journal
    - osd
    - leveldb
    - throttle
- ceph daemon osd.* config set [field] [value](inject config value without restart)

# Case Study

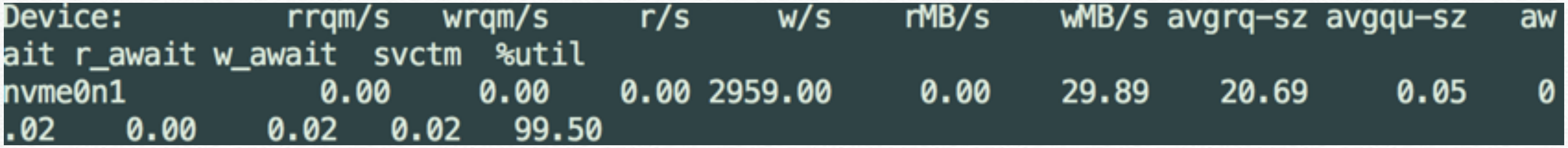
- My cluster is slow, only 3k iops(8k size) with three hosts, each host has one pcie ssd.
- Replicate size is 2
- each ssd has two partitions, one for journal, another for data

# Case Study

- Overview check firstly:
  - cpu: quite idle
  - memory: no paging
  - network: no dropping packages
  - io: high util



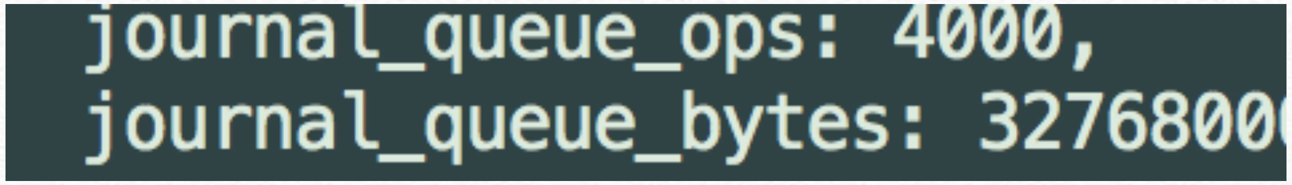
```
Tasks: 513 total,   1 running, 512 sleeping,   0 stopped,   0 zombie
%Cpu0  : 20.8 us,  3.0 sy,  0.0 ni, 76.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu1  : 22.6 us,  3.8 sy,  0.0 ni, 73.6 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu2  : 21.0 us,  3.0 sy,  0.0 ni, 76.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu3  : 20.2 us,  3.0 sy,  0.0 ni, 76.8 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu4  : 19.6 us,  4.9 sy,  0.0 ni, 75.5 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu5  : 19.2 us,  3.0 sy,  0.0 ni, 77.8 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu6  : 22.1 us,  2.9 sy,  0.0 ni, 74.0 id,  0.0 wa,  0.0 hi,  1.0 si,  0.0 st
%Cpu7  : 20.8 us,  4.0 sy,  0.0 ni, 75.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu8  : 21.4 us,  5.1 sy,  0.0 ni, 73.5 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu9  : 20.8 us,  4.0 sy,  0.0 ni, 74.3 id,  0.0 wa,  0.0 hi,  1.0 si,  0.0 st
%Cpu10 : 26.0 us,  4.8 sy,  0.0 ni, 69.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu11 : 23.5 us,  4.1 sy,  0.0 ni, 72.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu12 : 27.7 us,  4.0 sy,  0.0 ni, 68.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu13 : 18.6 us,  4.1 sy,  0.0 ni, 77.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu14 : 23.2 us,  7.1 sy,  0.0 ni, 67.7 id,  2.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu15 : 39.2 us,  8.8 sy,  0.0 ni, 51.0 id,  1.0 wa,  0.0 hi,  0.0 si,  0.0 st
%Cpu16 :  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
```



```
Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   aw
ait r_await w_await  svctm  %util
nvme0n1           0.00     0.00    0.00 2959.00     0.00    29.89    20.69     0.05    0
.02    0.00    0.02   0.02  99.50
```

# Case Study

- Ceph check queue/throttle: filejournal queue busy queue

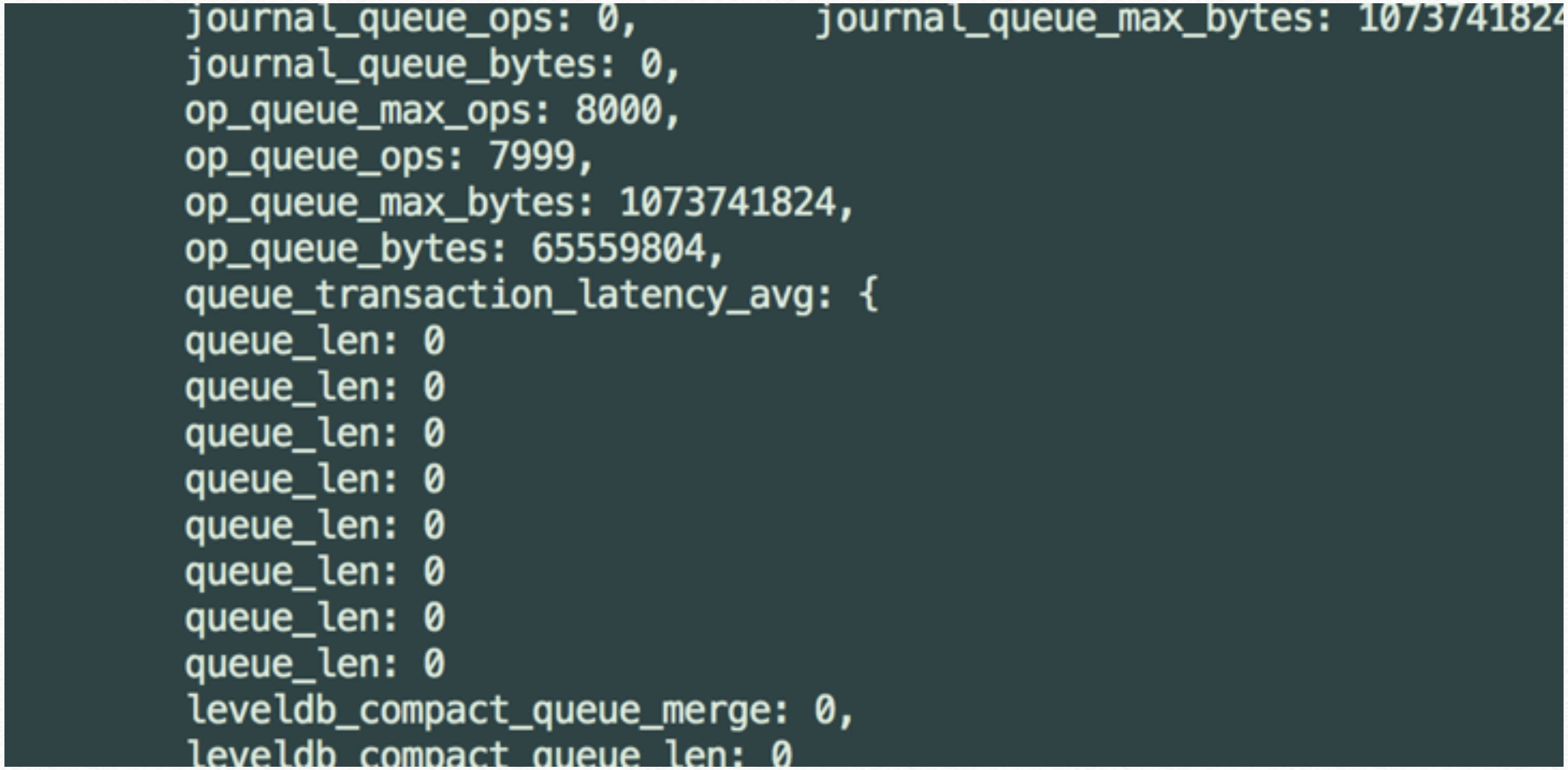


```
  journal_queue_ops: 4000,
  journal_queue_bytes: 3276800
```

- It must something wrong with journal
- Run fio with libaio with entire disk
  - High performance!
- Run fio with libaio with this journal partition
  - High utilization with low iops!
- fdisk found unaligned partition
  - Fix!
- But …

# Case Study

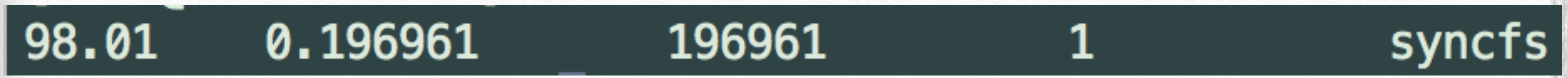
- Ceph check queue/throttle: filestore queue busy queue



```
        journal_queue_ops: 0,          journal_queue_max_bytes: 1073741824
        journal_queue_bytes: 0,
        op_queue_max_ops: 8000,
        op_queue_ops: 7999,
        op_queue_max_bytes: 1073741824,
        op_queue_bytes: 65559804,
        queue_transaction_latency_avg: {
        queue_len: 0
        queue_len: 0
        queue_len: 0
        queue_len: 0
        queue_len: 0
        queue_len: 0
        queue_len: 0
        queue_len: 0
        leveldb_compact_queue_merge: 0,
        leveldb_compact_queue_len: 0
```

# Case Study

- But the performance from fio with libaio engine in this ssd is well
- What's the difference with two workloads?
    - use "strace" to look for clues about io syscall
    - found high latency with "syncfs"



```
98.01    0.196961        196961           1            syncfs
```

- Run fio with sync engine, hit low iops!
    - then tested in different filesystem and raw block device
    - NVMe driver has bug with sync request under xfs(centos7) in vendor's firmware
- Follow vendor's instructions and downgrade NVMe driver, all is OK