CL'SO

Understanding Ceph

One Performance Counter at a Time

Marcel Lauhoff | Staff Software Engineer
December, 2024

Scope

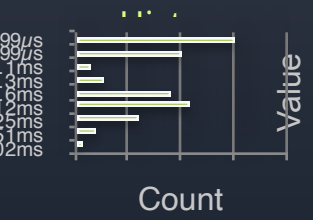
What can we learn by looking at Perf Counters?

Let's develop an intuition!

We won't go into analysis methods like the USE method

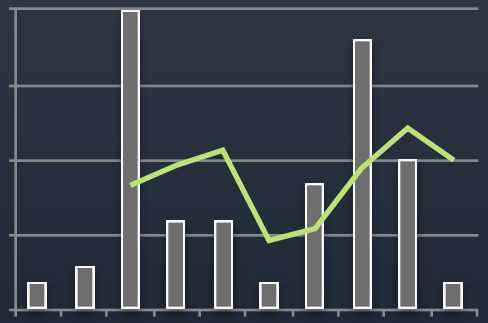
We look at *raw* performance counters.

No Prometheus, Grafana, etc.


Value
Count


Count
t1 t2 t3 t4 t5
t1 t3 t5 t7 t9


t1 t2 t3 t4 t5



What is a Perf Counter anyway?

A **metric** is a **measurement** of a service captured at runtime
(OpenTelemetry)

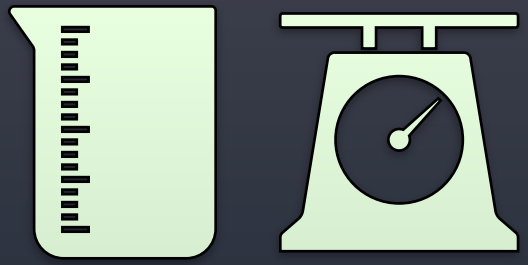
Values **captured** at **strategic** places
(me)

Examples:

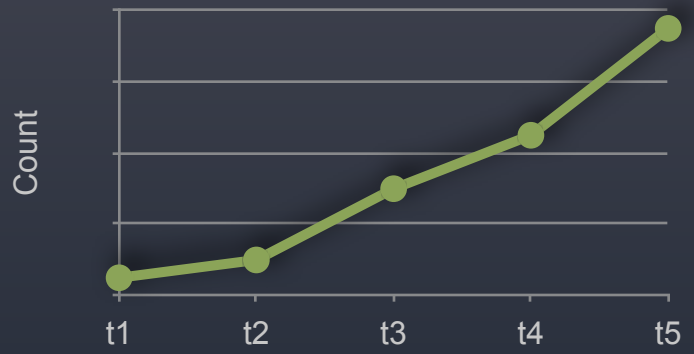
librados sends a write operation → increment osdop_write counter

OSD processed an operation → update op_latency

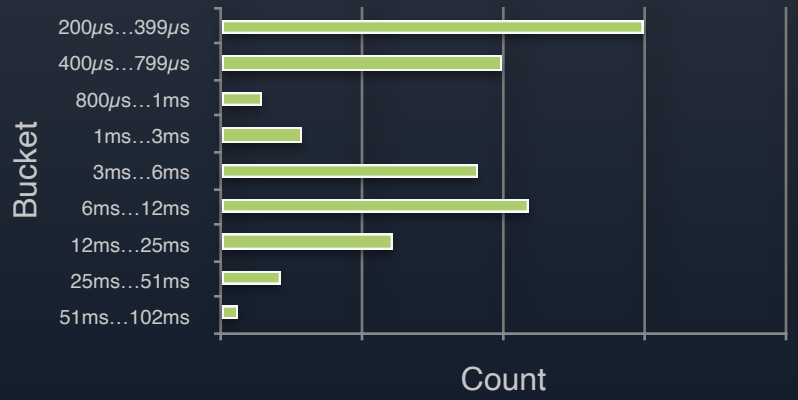
Gauge



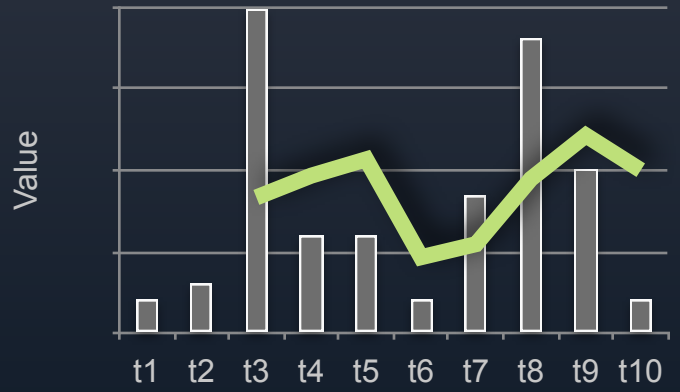
Counter



Histogram



Long Running Average



How to read Perf Counters?

```
# ceph daemon $out/radosgw.8000.asok perf dump
```

```
# ceph tell osd.0 perf dump
# ceph tell osd.0 perf schema
```

```
# ceph tell osd.0 counter dump
# ceph tell osd.0 counter schema
```

How to read performance counters?


```
{
  "AsyncMessenger::Worker-0": {
    "msgr_recv_messages": 1,
    "msgr_[illegible]
    "msgr_[illegible]55,
    "msgr_[illegible]
    "msgr_[illegible]
    "msgr_active_connections": 0,
    "msgr_running_total_time":
0.001140663,
    "msgr_running_send_time":
0.000978207,
    "msgr_running_recv_time":
0.000065659,
    "msgr_running_fast_dispatch_time":
0.000080838,
    "msgr_send_messages_queue_lat": {
      "avgcount": 1,
      "sum": 0.000080748,
      "avgtime": 0.000080748
    },
    "msgr_handle_ack_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "msgr_recv_encrypted_bytes": 0,
    "msgr_send_encrypted_bytes": 0
  },
  "AsyncMessenger::Worker-1": {
    "msgr_recv_messages": 0,
    "msgr_send_messages": 0,
    "msgr_recv_bytes": 0,
    "msgr_send_bytes": 0,
    "msgr_created_connections": 0,
    "msgr_active_connections": 0,
    "msgr_running_total_time":
0.000000000,
    "msgr_running_send_time":
0.000000000,
    "msgr_running_recv_time":
0.000000000,
    "msgr_running_fast_dispatch_time":
0.000000000,
    "msgr_send_messages_queue_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "msgr_handle_ack_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "msgr_recv_encrypted_bytes": 0,
    "msgr_send_encrypted_bytes": 0
  },
  "AsyncMessenger::Worker-2": {
    "msgr_recv_messages": 2,
    "msgr_send_messages": 2,
    "msgr_recv_bytes": 736,
    "msgr_send_bytes": 1523,
    "msgr_created_connections": 0,
    "msgr_active_connections": 0,
    "msgr_running_total_time":
0.000352540,
    "msgr_running_send_time":
0.000109558,
    "msgr_running_recv_time":
0.000089778,
    "msgr_running_fast_dispatch_time":
0.000135236,
    "msgr_send_messages_queue_lat": {
      "avgcount": 2,
      "sum": 0.000070978,
      "avgtime": 0.000035489
    },
    "msgr_handle_ack_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "msgr_recv_encrypted_bytes": 0,
    "msgr_send_encrypted_bytes": 0
  },
  "cct": {
    "total_workers": 0,
    "unhealthy_workers":
0
  },
  "mempool": {
    "bloom_filter_bytes": 0,
    "bloom_filter_items": 0,
    "bluestore_alloc_bytes": 0,
    "bluestore_alloc_items": 0,
    "bluestore_cache_data_bytes": 0,
    "bluestore_cache_data_items": 0,
    "bluestore_cache_onode_bytes": 0,
    "bluestore_cache_onode_items": 0,
    "bluestore_cache_meta_bytes": 0,
    "bluestore_cache_meta_items": 0,
    "bluestore_cache_other_bytes": 0,
    "bluestore_cache_other_items": 0,
    "bluestore_cache_buffer_bytes": 0,
    "bluestore_cache_buffer_items": 0,
    "bluestore_extent_bytes": 0,
    "bluestore_extent_items": 0,
    "bluestore_blob_bytes": 0,
    "bluestore_blob_items": 0,
    "bluestore_shared_blob_bytes": 0,
    "bluestore_shared_blob_items": 0,
    "bluestore_inline_bl_bytes": 0,
    "bluestore_inline_bl_items": 0,
    "bluestore_fsck_bytes": 0,
    "bluestore_fsck_items": 0,
    "bluestore_txc_bytes": 0,
    "bluestore_txc_items": 0,
    "bluestore_writing_deferred_bytes": 0,
    "bluestore_writing_deferred_items": 0,
    "bluestore_writing_bytes": 0,
    "bluestore_writing_items": 0,
    "bluefs_bytes": 0,
    "bluefs_items": 0,
    "bluefs_file_reader_bytes": 0,
    "bluefs_file_reader_items": 0,
    "bluefs_file_writer_bytes": 0,
    "bluefs_file_writer_items": 0,
    "buffer_anon_bytes": 359122,
    "buffer_anon_items": 121,
    "buffer_meta_bytes": 176,
    "buffer_meta_items": 2,
    "osd_bytes": 0,
    "osd_items": 0,
    "osd_mapbl_bytes": 0,
    "osd_mapbl_items": 0,
    "osd_pglog_bytes": 0,
    "osd_pglog_items": 0,
    "osdmap_bytes": 31920,
    "osdmap_items": 174,
    "osdmap_mapping_bytes": 0,
    "osdmap_mapping_items": 0,
    "pgmap_bytes": 0,
    "pgmap_items": 0,
    "mds_co_bytes": 0,
    "mds_co_items": 0,
    "unittest_1_bytes": 0,
    "unittest_1_items": 0,
    "unittest_2_bytes": 0,
    "unittest_2_items": 0
  },
  "objecter": {
    "op_active": 0,
    "op_laggy": 0,
    "op_send": 0,
    "op_send_bytes": 0,
    "op_resend": 0,
    "op_reply": 0,
    "op_latency": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "op_inflight": 0,
    "oplen_avg": {
      "avgcount": 0,
      "sum": 0
    },
    "op": 0,
    "op_r": 0,
    "op_w": 0,
    "op_rmw": 0,
    "op_pg": 0,
    "osdop_stat": 0,
    "osdop_create": 0,
    "osdop_read": 0,
    "osdop_write": 0,
    "osdop_writefull": 0,
    "osdop_writesame": 0,
    "osdop_append": 0,
    "osdop_zero": 0,
    "osdop_truncate": 0,
    "osdop_delete": 0,
    "osdop_mapext": 0,
    "osdop_sparse_read": 0,
    "osdop_clonerange": 0,
    "osdop_getxattr": 0,
    "osdop_setxattr": 0,
    "osdop_cmpxattr": 0,
    "osdop_rmxattr": 0,
    "osdop_resetxattrs": 0,
    "osdop_call": 0,
    "osdop_watch": 0,
    "osdop_notify": 0,
    "osdop_src_cmpxattr": 0,
    "osdop_pgls": 0,
    "osdop_pgls_filter": 0,
    "osdop_other": 0,
    "linger_active": 0,
    "linger_send": 0,
    "linger_resend": 0,
    "linger_ping": 0,
    "poolop_active": 0,
    "poolop_send": 0,
    "poolop_resend": 0,
    "poolstat_active": 0,
    "poolstat_send": 0,
    "poolstat_resend": 0,
    "statfs_active": 0,
    "statfs_send": 0,
    "statfs_resend": 0,
    "command_active": 0,
    "command_send": 0,
    "command_resend": 0,
    "map_epoch": 0,
    "map_full": 0,
    "map_inc": 0,
    "osd_sessions": 2,
    "osd_session_open": 0,
    "osd_session_close": 0,
    "osd_laggy": 0,
    "omap_wr": 0,
    "omap_rd": 0,
    "omap_del": 0,
    "replica_read_sent": 0,
    "replica_read_bounced": 0,
    "replica_read_completed": 0
  },
  "objecter-0x43dbad00": {
    "op_active": 0,
    "op_laggy": 0,
    "op_send": 0,
    "op_send_bytes": 0,
    "op_resend": 0,
    "op_reply": 0,
    "op_latency": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "op_inflight": 0,
    "oplen_avg": {
      "avgcount": 0,
      "sum": 0
    },
    "op": 0,
    "op_r": 0,
    "op_w": 0,
    "op_rmw": 0,
    "op_pg": 0,
    "osdop_stat": 0,
    "osdop_create": 0,
    "osdop_read": 0,
    "osdop_write": 0,
    "osdop_writefull": 0,
    "osdop_writesame": 0,
    "osdop_append": 0,
    "osdop_zero": 0,
    "osdop_truncate": 0,
    "osdop_delete": 0,
    "osdop_mapext": 0,
    "osdop_sparse_read": 0,
    "osdop_clonerange": 0,
    "osdop_getxattr": 0,
    "osdop_setxattr": 0,
    "osdop_cmpxattr": 0,
    "osdop_rmxattr": 0,
    "osdop_resetxattrs": 0,
    "osdop_call": 0,
    "osdop_watch": 0,
    "osdop_notify": 0,
    "osdop_src_cmpxattr": 0,
    "osdop_pgls": 0,
    "osdop_pgls_filter": 0,
    "osdop_other": 0,
    "linger_active": 0,
    "linger_send": 0,
    "linger_resend": 0,
    "linger_ping": 0,
    "poolop_active": 0,
    "poolop_send": 0,
    "poolop_resend": 0,
    "poolstat_active": 0,
    "poolstat_send": 0,
    "poolstat_resend": 0,
    "statfs_active": 0,
    "statfs_send": 0,
    "statfs_resend": 0,
    "command_active": 0,
    "command_send": 0,
    "command_resend": 0,
    "map_epoch": 0,
    "map_full": 0,
    "map_inc": 0,
    "osd_sessions": 0,
    "osd_session_open": 0,
    "osd_session_close": 0,
    "osd_laggy": 0,
    "omap_wr": 0,
    "omap_rd": 0,
    "omap_del": 0,
    "replica_read_sent": 0,
    "replica_read_bounced": 0,
    "replica_read_completed": 0
  },
  "objecter-0x43dbaf70": {
    "op_active": 0,
    "op_laggy": 0,
    "op_send": 3,
    "op_send_bytes": 4196000,
    "op_resend": 0,
    "op_reply": 3,
    "op_latency": {
      "avgcount": 3,
      "sum": 0.063001676,
      "avgtime": 0.021000558
    },
    "op_inflight": 0,
    "oplen_avg": {
      "avgcount": 3,
      "sum": 18
    },
    "op": 3,
    "op_r": 0,
    "op_w": 3,
    "op_rmw": 0,
    "op_pg": 0,
    "osdop_stat": 2,
    "osdop_create": 1,
    "osdop_read": 0,
    "osdop_write": 0,
    "osdop_writefull": 1,
    "osdop_writesame": 0,
    "osdop_append": 0,
    "osdop_zero": 0,
    "osdop_truncate": 0,
    "osdop_delete": 0,
    "osdop_mapext": 0,
    "osdop_sparse_read": 0,
    "osdop_clonerange": 0,
    "osdop_getxattr": 0,
    "osdop_setxattr": 9,
    "osdop_cmpxattr": 0,
    "osdop_rmxattr": 0,
    "osdop_resetxattrs": 0,
    "osdop_call": 5,
    "osdop_watch": 0,
    "osdop_notify": 0,
    "osdop_src_cmpxattr": 0,
    "osdop_pgls": 0,
    "osdop_pgls_filter": 0,
    "osdop_other": 0,
    "linger_active": 10,
    "linger_send": 0,
    "linger_resend": 0,
    "linger_ping": 0,
    "poolop_active": 0,
    "poolop_send": 0,
    "poolop_resend": 0,
    "poolstat_active": 0,
    "poolstat_send": 0,
    "poolstat_resend": 0,
    "statfs_active": 0,
    "statfs_send": 0,
    "statfs_resend": 0,
    "command_active": 0,
    "command_send": 0,
    "command_resend": 0,
    "map_epoch": 0,
    "map_full": 0,
    "map_inc": 0,
    "osd_sessions": 3,
    "osd_session_open": 0,
    "osd_session_close": 0,
    "osd_laggy": 0,
    "omap_wr": 0,
    "omap_rd": 0,
    "omap_del": 0,
    "replica_read_sent": 0,
    "replica_read_bounced": 0,
    "replica_read_completed": 0
  },
  "rgw": {
    "req": 1,
    "failed_req": 0,
    "qlen": 0,
    "qactive": 0,
    "cache_hit": 8,
    "cache_miss": 0,
    "keystone_token_cache_hit": 0,
    "keystone_token_cache_miss": 0,
    "gc_retire_object": 0,
    "lc_expire_current": 0,
    "lc_expire_noncurrent": 0,
    "lc_expire_dm": 0,
    "lc_transition_current": 0,
    "lc_transition_noncurrent": 0,
    "lc_abort_mpu": 0,
    "pubsub_event_triggered": 0,
    "pubsub_event_lost": 0,
    "pubsub_store_ok": 0,
    "pubsub_store_fail": 0,
    "pubsub_events": 0,
    "pubsub_push_ok": 0,
    "pubsub_push_failed": 0,
    "pubsub_push_pending": 0,
    "pubsub_missing_conf": 0,
    "lua_current_vms": 1,
    "lua_script_ok": 0,
    "lua_script_fail": 0
  },
  "rgw_op": {
    "put_obj_ops": 1,
    "put_obj_bytes": 4194304,
    "put_obj_lat": {
      "avgcount": 1,
      "sum": 0.063001675,
      "avgtime": 0.063001675
    },
    "get_obj_ops": 0,
    "get_obj_bytes": 0,
    "get_obj_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "del_obj_ops": 0,
    "del_obj_bytes": 0,
    "del_obj_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "del_bucket_ops": 0,
    "del_bucket_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "copy_obj_ops": 0,
    "copy_obj_bytes": 0,
    "copy_obj_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "list_obj_ops": 0,
    "list_obj_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    },
    "list_buckets_ops": 0,
    "list_buckets_lat": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "simple-throttler": {
    "throttle": 0,
    "waiting": 0,
    "max": 18446744073709551595
  },
  "throttle-msgr_dispatch_throttler-
radosclient": {
    "val": 0,
    "max": 104857600,
    "get_started": 0,
    "get": 3,
    "get_sum": 0,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 0,
    "take": 0,
    "take_sum": 0,
    "put": 3,
    "put_sum": 0,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-msgr_dispatch_throttler-
radosclient-0x43dbadd0": {
    "val": 0,
    "max": 104857600,
    "get_started": 0,
    "get": 3,
    "get_sum": 1211,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 3,
    "take": 0,
    "take_sum": 0,
    "put": 3,
    "put_sum": 1211,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-objecter_bytes": {
    "val": 0,
    "max": 104857600,
    "get_started": 0,
    "get": 0,
    "get_sum": 0,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 0,
    "take": 0,
    "take_sum": 0,
    "put": 0,
    "put_sum": 0,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-objecter_ops": {
    "val": 0,
    "max": 24576,
    "get_started": 0,
    "get": 3,
    "get_sum": 3,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 3,
    "take": 0,
    "take_sum": 0,
    "put": 3,
    "put_sum": 3,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-objecter_bytes-0x43d82820": {
    "val": 0,
    "max": 104857600,
    "get_started": 0,
    "get": 3,
    "get_sum": 4195297,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 3,
    "take": 0,
    "take_sum": 0,
    "put": 3,
    "put_sum": 4195297,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-objecter_bytes-0x43d83040": {
    "val": 0,
    "max": 104857600,
    "get_started": 0,
    "get": 0,
    "get_sum": 0,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 0,
    "take": 0,
    "take_sum": 0,
    "put": 0,
    "put_sum": 0,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-objecter_ops": {
    "val": 0,
    "max": 24576,
    "get_started": 0,
    "get": 0,
    "get_sum": 0,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 0,
    "take": 0,
    "take_sum": 0,
    "put": 0,
    "put_sum": 0,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-objecter_ops-0x43dbae40": {
    "val": 0,
    "max": 24576,
    "get_started": 0,
    "get": 3,
    "get_sum": 3,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 3,
    "take": 0,
    "take_sum": 0,
    "put": 3,
    "put_sum": 3,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  },
  "throttle-rgw_async_rados_ops": {
    "val": 0,
    "max": 64,
    "get_started": 0,
    "get": 0,
    "get_sum": 0,
    "get_or_fail_fail": 0,
    "get_or_fail_success": 0,
    "take": 0,
    "take_sum": 0,
    "put": 0,
    "put_sum": 0,
    "wait": {
      "avgcount": 0,
      "sum": 0.000000000,
      "avgtime": 0.000000000
    }
  }
}
```


17k JSON

CL'SO

PERF COUNTER BROWSER

Found 880 perf counters in 41 groups

Group	Name	Description	Type	Tags
async messenger	msgr_connection_ready_timeouts	Number of not yet ready connections declared as dead	integer	mds, mgr, mon, client, osd
async messenger	msgr_connection_idle_timeouts	Number of connections closed due to idleness	integer	mds, mgr, mon, client, osd
async messenger	msgr_recv_messages	Network received messages	integer	mds, mgr, mon, client, osd
async messenger	msgr_send_messages	Network sent messages	integer	mds, mgr, mon, client, osd
async messenger	msgr_recv_bytes	Network received bytes	integer	mds, mgr, mon, client, osd
async messenger	msgr_send_bytes	Network sent bytes	integer	mds, mgr, mon, client, osd
async messenger	msgr_created_connections	Created connection number	integer	mds, mgr, mon, client, osd
async messenger	msgr_active_connections	Active connection number	integer	mds, mgr, mon, client, osd
async messenger	msgr_running_total_time	The [illegible] hread running	real	mds, mgr, mon, [illegible]

Calendar

MON=3 OSD=3 MGR=1 RGW=1 MDS=2 NFS=0 vstart.sh

880 counters

in

41 groups

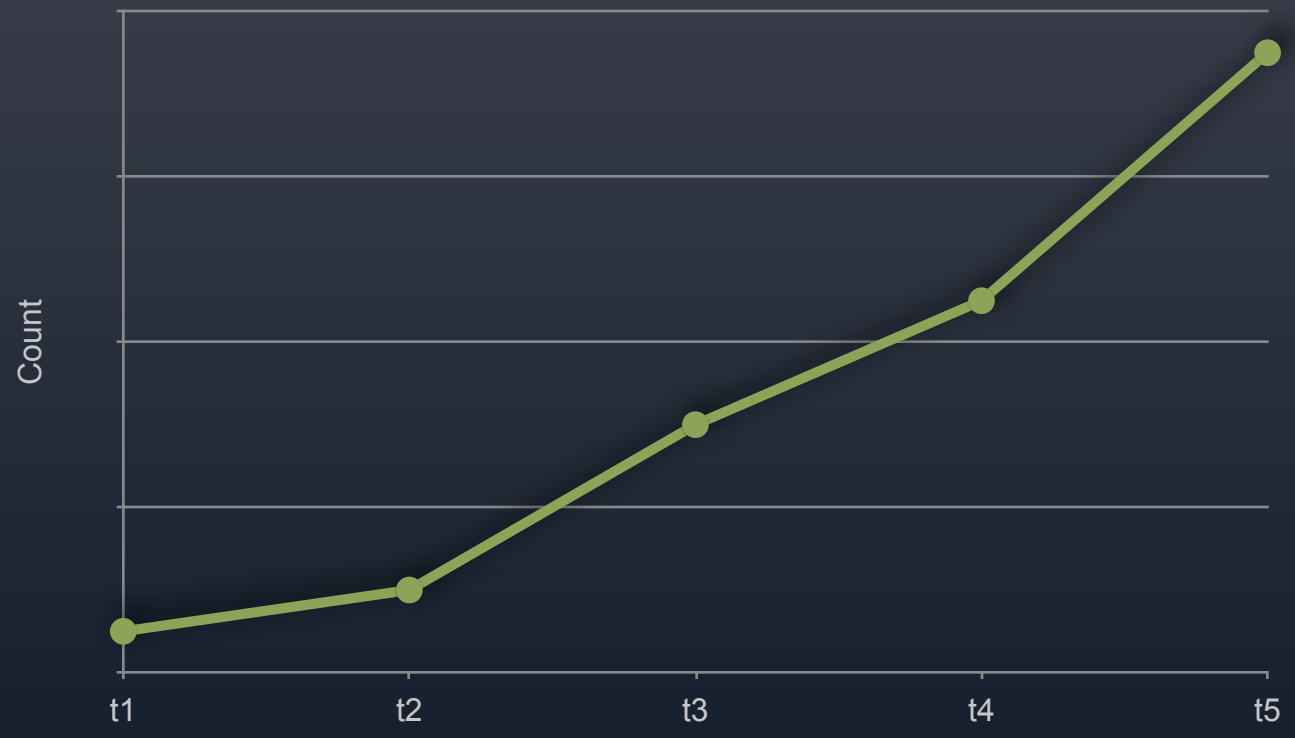
We focus on..

objecter (librados)

AsyncMessenger

osd

Counter



RGW: Single 4MB S3 PUT

2024-11-29T12:42:45.821+0100 7f0be538f6c0 1 beast: 0x7f0c53e61200: ::1 - testid [29/Nov/2024:12:42:45.758 +0100] "PUT /testbucket/13359 HTTP/1.1" 200 **4194304** - - - **latency=0.063001677s**

```
"rgw_op": {
    "put_obj_ops": 1,
    "put_obj_bytes": 4194304,
    ...
}
```

```
"rgw": {
    "req": 1,
    ...
}
```

CL'SO

RGW: Single 4MB S3 PUT

objecter.{osdop|omap}

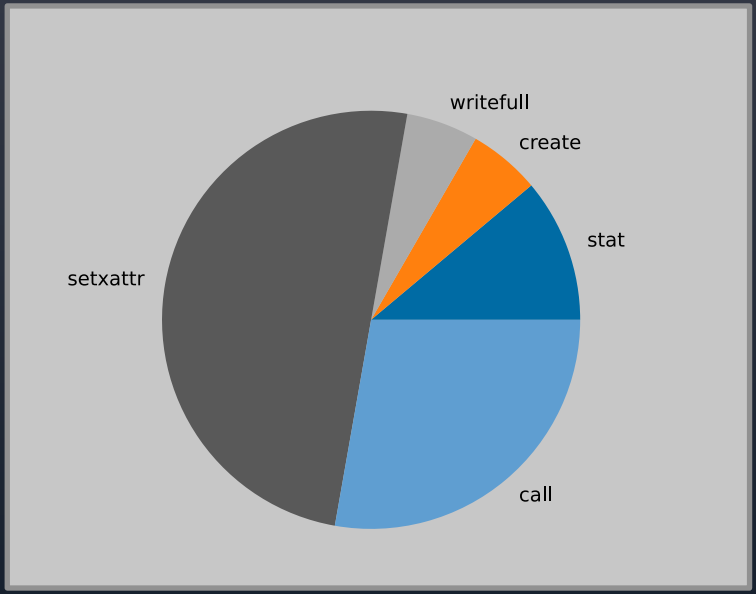
Table 1

"osdop_stat"	2
"osdop_create"	1
"osdop_read"	0
"osdop_write"	0
"osdop_writefull"	1
"osdop_writesame"	0
"osdop_append"	0
"osdop_zero"	0
"osdop_truncate"	0
"osdop_delete"	0
"osdop_mapext"	0
"osdop_sparse_read"	0
"osdop_clonerange"	0
"osdop_getxattr"	0
"osdop_setxattr"	9
"osdop_cmpxattr"	0
"osdop_rmxattr"	0
"osdop_resetxattrs"	0
"osdop_call"	5
"osdop_watch"	0
"osdop_notify"	0
"osdop_src_cmpxattr"	0
"osdop_pgls"	0
"osdop_pgls_filter"	0
"osdop_other"	0
“omap_wr"	0
“omap_rd"	0
“omap_del"	0

RGW: Single 4MB S3 PUT

objecter.{osdop|omap}_*

stat	2
create	1
writefull	1
setxattr	9
call	5



$\sum \ldots = 18$

RGW: Single 4MB S3 PUT

AsyncMessenger::Worker-*


```
{
  "msgr_recv_messages": 3,
  "msgr_send_messages": 3,
  "msgr_recv_bytes": 1481,
  "msgr_send_bytes": 4197637,
}
```

RGW: Single 4MB S3 PUT

$\sum$ (osdop_*, omap_*) = 18

$\sum$ msgr_send_messages = 3

objecter.op = 3

RGW: Single 4MB S3 PUT

$$\sum \text{msgr\_send\_bytes} - 4\text{MB}$$

$$\sim 4\text{k}$$

RGW: Single 4MB S3 PUT

osd.0


```
{
  "op": 6,
  "op_in_bytes": 0,
  "op_out_bytes": 224,
  "subop": 3,
  "subop_in_bytes": 4198251
}
```


osd.1


```
{
  "op": 4,
  "op_in_bytes": 0,
  "op_out_bytes": 0,
  "subop": 1,
  "subop_in_bytes": 4196187
}
```


osd.2


```
{
  "op": 4,
  "op_in_bytes": 4194304,
  "op_out_bytes": 0,
  "subop": 2,
  "subop_in_bytes": 2064
}
```

What does this teach us?

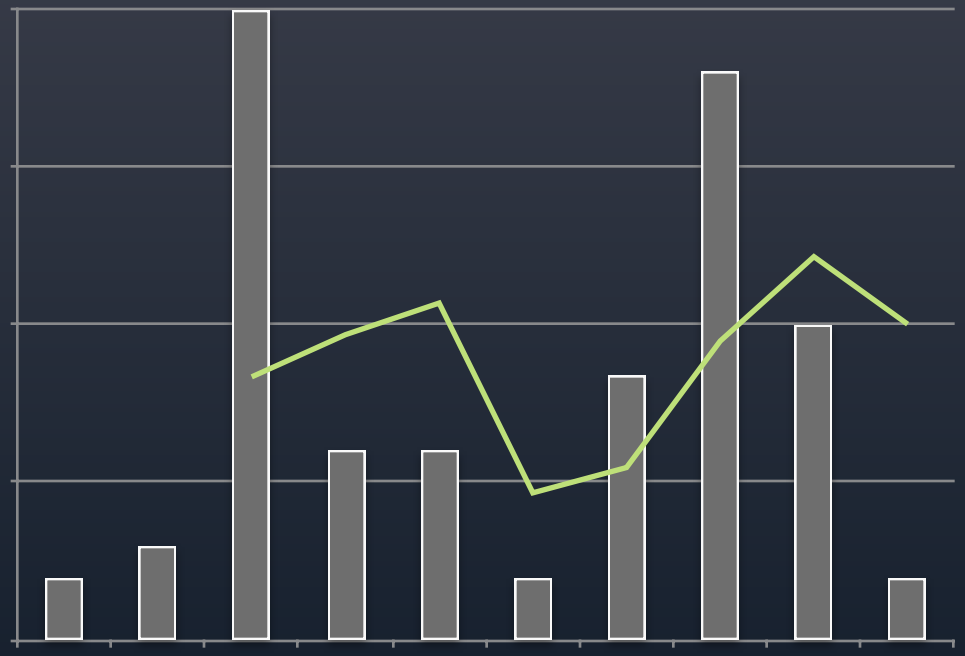
Op Mix

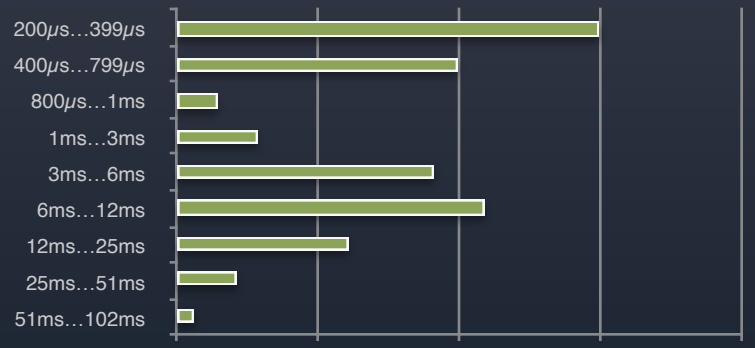
Op = [op, ..., op]

3 messages, 3 ops, 4k overhead

-> 18 queries and modifications to and from persistent data

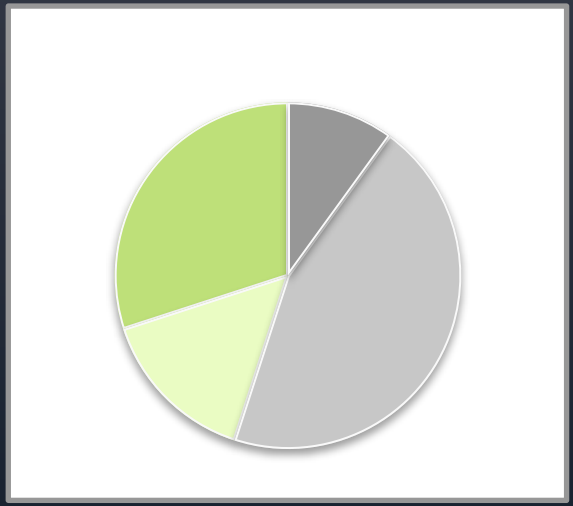
Latency

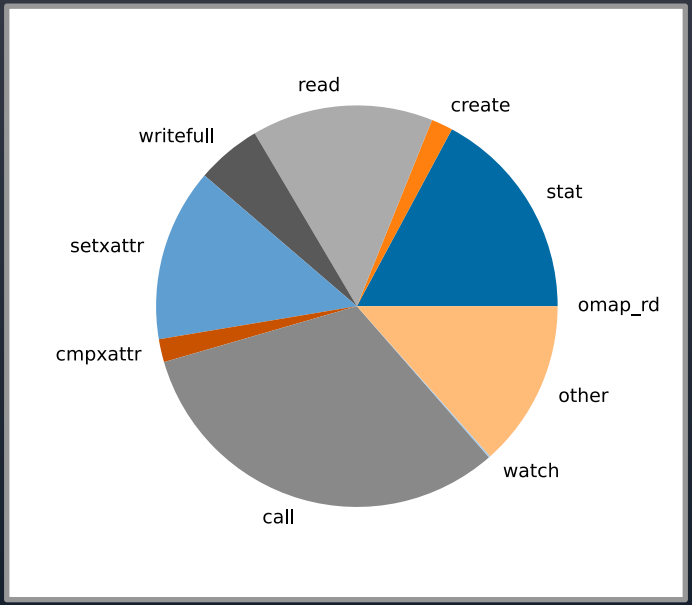



200μs…399μs
400μs…799μs
800μs…1ms
1ms…3ms
3ms…6ms
6ms…12ms
12ms…25ms
25ms…51ms
51ms…102ms

S3 Mixed Workload Benchmark

10% DELETE, 45% GET, 15% PUT, 30% STAT





stat	20947
create	2125
read	17836
writefull	6375
setxattr	17000
cmpxattr	2265
call	39004
watch	150
other	16417
omap$_{rd}$	2

S3 Mixed Workload Benchmark

Client / RGW:
objecter.op_latency 43ms

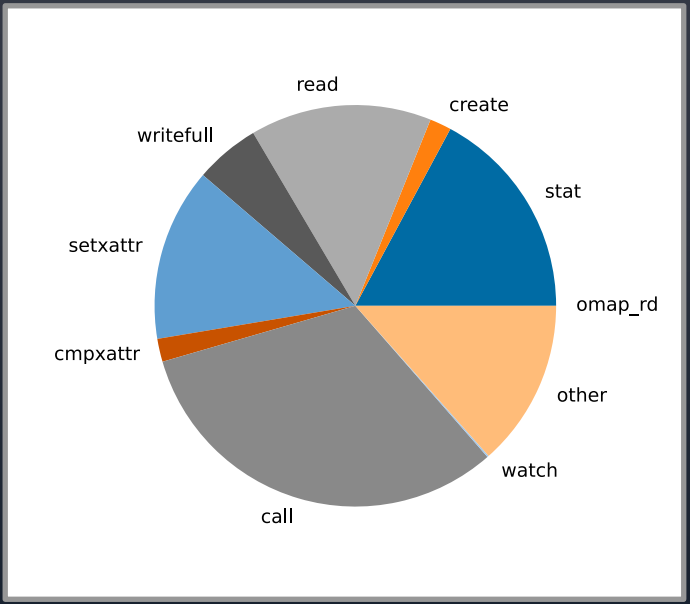
OSDs:
osd.op_latency

0: 41ms, 1: 44ms, 2: 39ms

S3 Mixed Workload Benchmark

~40 ms average latency

Is this actually meaningful?



S3 Mixed Workload Benchmark

osd.op_latency family:

osd.op_**w**_latency osd.op_**rw**_latency osd.op_**r**_latency

target	op_latency [ms]	op_r_latency [ms]	op_w_latency [ms]	op_rw_latency [ms]
osd.0	41	5	99	83
osd.1	44	5	98	82
osd.2	39	5	100	82

Histograms to the rescue

osd.0

41 ms average

5ms read

99ms write

83ms rw

Histograms to the rescue

op_latency captures every operation

op_r_latency_out_bytes_histogram	Histogram of operation latency (including queue time) + data read
op_w_latency_in_bytes_histogram	Histogram of operation latency (including queue time) + data written
op_rw_latency_in_bytes_histogram	Histogram of rw operation latency (including queue time) + data written
op_rw_latency_out_bytes_histogram	Histogram of rw operation latency (including queue time) + data read

CL'SO

Ceph 2D Histograms: latency x bytes

◪	0…511	512…1023	1K…2K	2K…4K	4K…8K	8K…16K	16K…32K	32K…64K	64K…128K	128K…256K	256K…512K	512K…1024K	1M…2M	2M…4M	4M…8M	8M…16M
<0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0ns…99µs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100µs…199µs	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
200µs…399µs	31	1421	0	0	0	0	0	0	0	0	0	0	0	333	32	0
400µs…799µs	45	413	9	1	0	0	0	0	0	0	0	0	0	185	567	0
800µs…1ms	17	39	26	3	0	0	0	0	0	0	0	0	0	59	104	0
1ms…3ms	3	21	5	0	0	0	0	0	0	0	0	0	0	382	308	0
3ms…6ms	1	33	0	0	0	0	0	0	0	0	0	0	0	530	1168	0
6ms…12ms	2	38	0	0	0	0	0	0	0	0	0	0	0	313	1020	0
12ms…25ms	3	26	0	0	0	0	0	0	0	0	0	0	0	135	466	0
25ms…51ms	0	21	0	0	0	0	0	0	0	0	0	0	0	35	120	0
51ms…102ms	0	8	0	0	0	0	0	0	0	0	0	0	0	10	26	0
102ms…204ms	0	3	0	0	0	0	0	0	0	0	0	0	0	2	7	0

osd.op_r_latency_out_bytes_histogram

CL'SO

Ceph 2D Histograms: latency x bytes

◪	0…511	512…1023	1K…2K	2K…4K	4K…8K	8K…16K	16K…32K	32K…64K	64K…128K	128K…256K	256K…512K	512K…1024K	1M…2M	2M…4M	4M…8M	8M…16M
<0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0ns…99µs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100µs…199µs	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
200µs…399µs	31	1421	0	0	0	0	0	0	0	0	0	0	0	333	32	0
400µs…799µs	45	413	9	1	0	0	0	0	0	0	0	0	0	185	567	0
800µs…1ms	17	39	26	3	0	0	0	0	0	0	0	0	0	59	104	0
1ms…3ms	3	21	5	0	0	0	0	0	0	0	0	0	0	382	308	0
3ms…6ms	1	33	0	0	0	0	0	0	0	0	0	0	0	530	1168	0
6ms…12ms	2	38	0	0	0	0	0	0	0	0	0	0	0	313	1020	0
12ms…25ms	3	26	0	0	0	0	0	0	0	0	0	0	0	135	466	0
25ms…51ms	0	21	0	0	0	0	0	0	0	0	0	0	0	35	120	0
51ms…102ms	0	8	0	0	0	0	0	0	0	0	0	0	0	10	26	0
102ms…204ms	0	3	0	0	0	0	0	0	0	0	0	0	0	2	7	0

osd.op_r_latency_out_bytes_histogram

CL'SO

Ceph 2D Histograms: latency x bytes

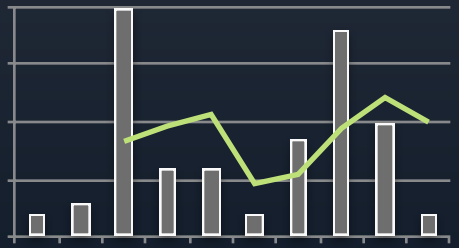
◪	0…511	512…1023	1K…2K	2K…4K	4K…8K	8K…16K	16K…32K	32K…64K	64K…128K	128K…256K	256K…512K	512K…1024K	1M…2M	2M…4M	4M…8M	8M…16M
<0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0ns…99µs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100µs…199µs	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
200µs…399µs	31	1421	0	0	0	0	0	0	0	0	0	0	0	333	32	0
400µs…799µs	45	413	9	1	0	0	0	0	0	0	0	0	0	185	567	0
800µs…1ms	17	39	26	3	0	0	0	0	0	0	0	0	0	59	104	0
1ms…3ms	3	21	5	0	0	0	0	0	0	0	0	0	0	382	308	0
3ms…6ms	1	33	0	0	0	0	0	0	0	0	0	0	0	530	1168	0
6ms…12ms	2	38	0	0	0	0	0	0	0	0	0	0	0	313	1020	0
12ms…25ms	3	26	0	0	0	0	0	0	0	0	0	0	0	135	466	0
25ms…51ms	0	21	0	0	0	0	0	0	0	0	0	0	0	35	120	0
51ms…102ms	0	8	0	0	0	0	0	0	0	0	0	0	0	10	26	0
102ms…204ms	0	3	0	0	0	0	0	0	0	0	0	0	0	2	7	0

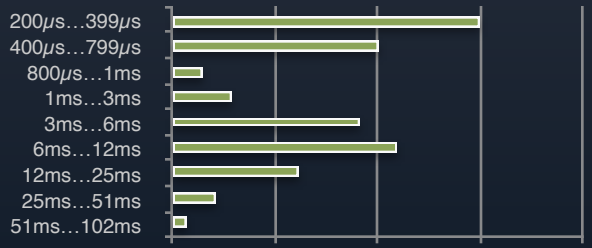
osd.op_r_latency_out_bytes_histogram

Recap

Counters - What is a S3 PUT?

Latency - Mixed S3 Workload





What's next?

Tracing (e.g bpftrace)

OSD Perf Queries

Collection at the Mgr: collection, Prometheus, etc

Bluestore metrics

Combining Ceph with system node exporter metrics

Messenger stats: cntop https://github.com/irq0/cntop

https://github.com/irq0/talk_ceph_perf_counters

CLYSO
Powered by ceph

Thank you!

Marcel Lauhoff <marcel.lauhoff@clyso.com>