OM performance test result (prototype)

Scope:

- Leader side execution
- Optimized flow for OBS create/commit (and involves overwrite of previous key)

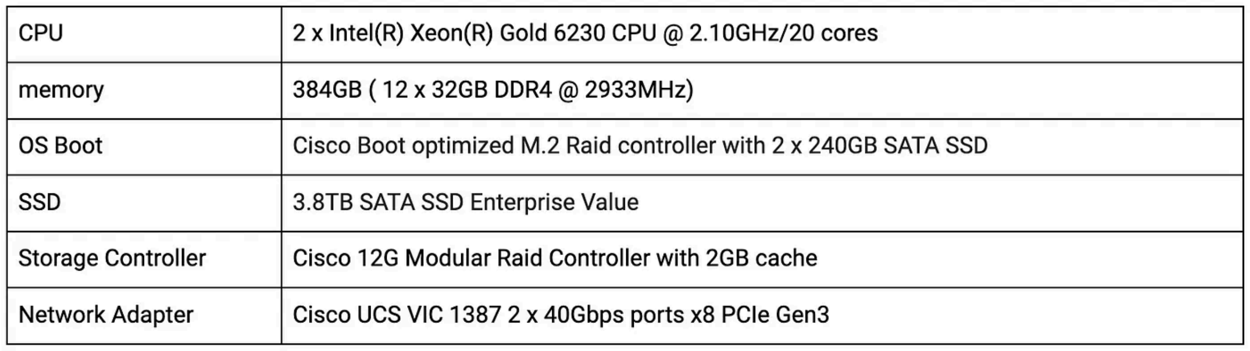
Optimization key points:

1. Parallel execution of different keys (Granular lock for key using stripped lock)
2. No Caching
3. Batching of ratis db update to flush to all nodes
4. Flow optimization removing redundant and un-ncecessary operation

Environment:

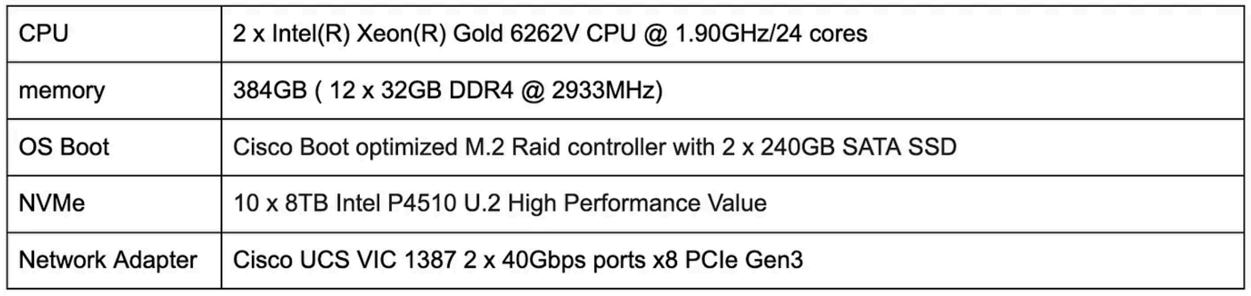
Cluster with 16 DNs, OM 3 node HA, SCM 3 node HA

Master nodes:



CPU	2 x Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz/20 cores
memory	384GB (12 x 32GB DDR4 @ 2933MHz)
OS Boot	Cisco Boot optimized M.2 Raid controller with 2 x 240GB SATA SSD
SSD	3.8TB SATA SSD Enterprise Value
Storage Controller	Cisco 12G Modular Raid Controller with 2GB cache
Network Adapter	Cisco UCS VIC 1387 2 x 40Gbps ports x8 PCIe Gen3

Datanodes:



CPU	2 x Intel(R) Xeon(R) Gold 6262V CPU @ 1.90GHz/24 cores
memory	384GB (12 x 32GB DDR4 @ 2933MHz)
OS Boot	Cisco Boot optimized M.2 Raid controller with 2 x 240GB SATA SSD
NVMe	10 x 8TB Intel P4510 U.2 High Performance Value
Network Adapter	Cisco UCS VIC 1387 2 x 40Gbps ports x8 PCIe Gen3

Comparison:

sno	Category	Old Flow	New Flow
1	Operation / Second (key create / commit)	12K+	40K+
2	Key Commit / Second	5.9K+	20K+ (3.3 times)
3	CPU Utilization Leader	16% (unable to increase load)	33%
4	CPU Utilization Follower	6% above	4% below
5			

Further Improvement

Further improvements can increase to **30K/Second**:
1. Caching of open Key for commit to avoid get from table
2. Codec Buffer encoding/decoding to be improved as per new flow
3. GetBlock caching at OM

Note: This is from the fact,
- Mocking getBlock and ranger reached capability to 40k+ key commit / second (almost double)

Few more area of improvement to increase caps:
1. Ranger Validation - to optimize / cached frequent call
2. Ratis upgrade to new version and further improvement

Command used for test

(thread in increased and multiple instance run to increase load under various test):

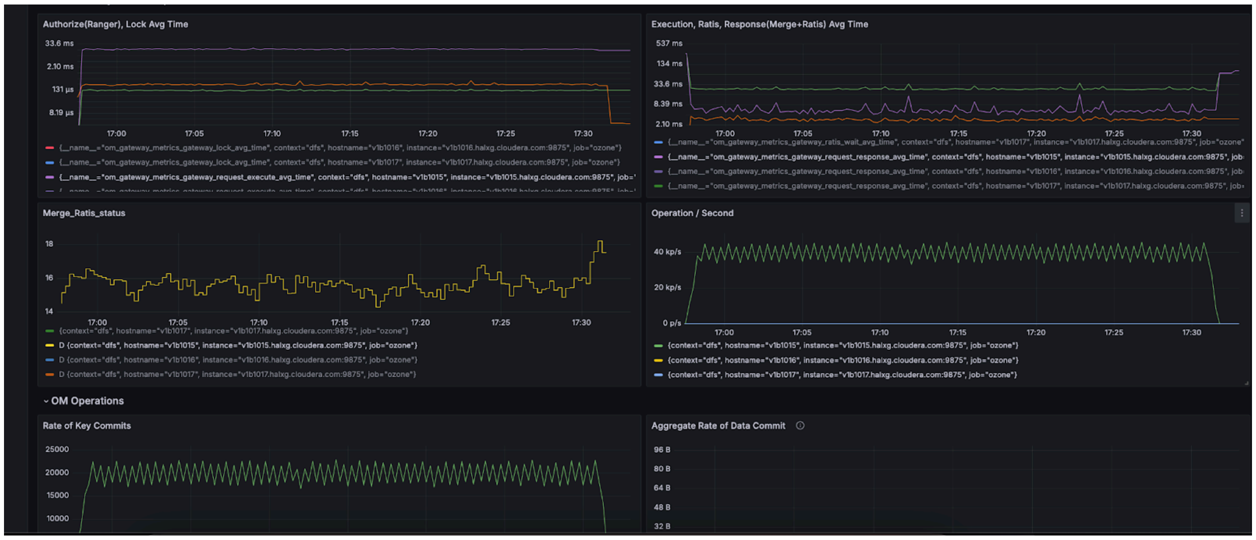
```
ozone freon ockrw -n 10000000 -t 100 --percentage-read 0 --size 0 -r 1000000 -v voltest -b
buckettest -p performanceTest
```

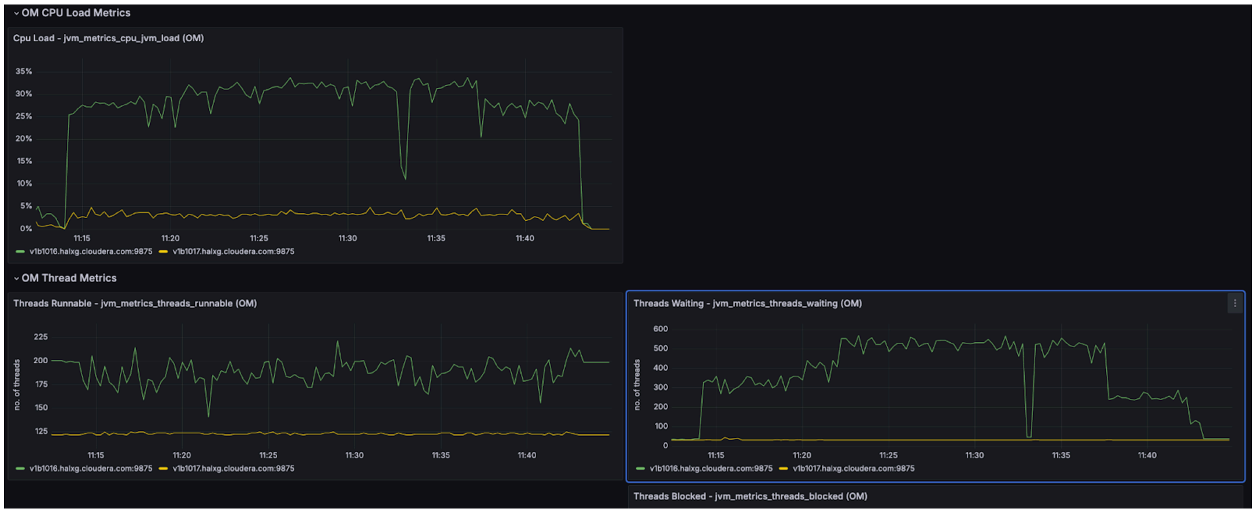
New Flow:

This is data with one freon running. When load increases with multiple freon, its crosses 21k easily.

```
key-read-write-list
             count = 10000000
         mean rate = 19565.78 calls/second
     1-minute rate = 19989.40 calls/second
     5-minute rate = 18552.83 calls/second
    15-minute rate = 16043.74 calls/second
               min = 3.02 milliseconds
               max = 25.84 milliseconds
              mean = 4.67 milliseconds
            stddev = 2.02 milliseconds
            median = 4.26 milliseconds
              75% <= 4.64 milliseconds
              95% <= 6.40 milliseconds
              98% <= 14.51 milliseconds
              99% <= 16.30 milliseconds
            99.9% <= 18.39 milliseconds


Total execution time (sec): 513
Failures: 0
Successful executions: 10000000
```

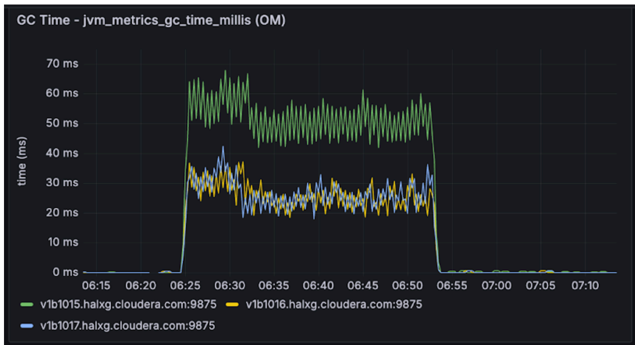

Authorize(Ranger), Lock Avg Time
33.6 ms
2.10 ms
131 µs
8.19 µs
17:00
17:05
17:10
17:15
17:20
17:25
17:30
{__name__="om_gateway_metrics_gateway_lock_avg_time", context="dfs", hostname="v1b1016", instance="v1b1016.halxg.cloudera.com:9875", job="ozone"}
{__name__="om_gateway_metrics_gateway_lock_avg_time", context="dfs", hostname="v1b1017", instance="v1b1017.halxg.cloudera.com:9875", job="ozone"}
{__name__="om_gateway_metrics_gateway_request_execute_avg_time", context="dfs", hostname="v1b1015", instance="v1b1015.halxg.cloudera.com:9875", job='
Execution, Ratis, Response(Merge+Ratis) Avg Time
537 ms
134 ms
33.6 ms
8.39 ms
2.10 ms
{__name__="om_gateway_metrics_gateway_ratis_wait_avg_time", context="dfs", hostname="v1b1017", instance="v1b1017.halxg.cloudera.com:9875", job="ozone"
{__name__="om_gateway_metrics_gateway_request_response_avg_time", context="dfs", hostname="v1b1015", instance="v1b1015.halxg.cloudera.com:9875", job
{__name__="om_gateway_metrics_gateway_request_response_avg_time", context="dfs", hostname="v1b1016", instance="v1b1016.halxg.cloudera.com:9875", job
{__name__="om_gateway_metrics_gateway_request_response_avg_time", context="dfs", hostname="v1b1017", instance="v1b1017.halxg.cloudera.com:9875", job
Merge_Ratis_status
18
16
14
{context="dfs", hostname="v1b1017", instance="v1b1017.halxg.cloudera.com:9875", job="ozone"}
D {context="dfs", hostname="v1b1015", instance="v1b1015.halxg.cloudera.com:9875", job="ozone"}
D {context="dfs", hostname="v1b1016", instance="v1b1016.halxg.cloudera.com:9875", job="ozone"}
D {context="dfs", hostname="v1b1017", instance="v1b1017.halxg.cloudera.com:9875", job="ozone"}
Operation / Second
40 kp/s
20 kp/s
0 p/s
{context="dfs", hostname="v1b1015", instance="v1b1015.halxg.cloudera.com:9875", job="ozone"}
{context="dfs", hostname="v1b1016", instance="v1b1016.halxg.cloudera.com:9875", job="ozone"}
{context="dfs", hostname="v1b1017", instance="v1b1017.halxg.cloudera.com:9875", job="ozone"}
OM Operations
Rate of Key Commits
25000
20000
15000
10000
Aggregate Rate of Data Commit
96 B
80 B
64 B
48 B
32 B

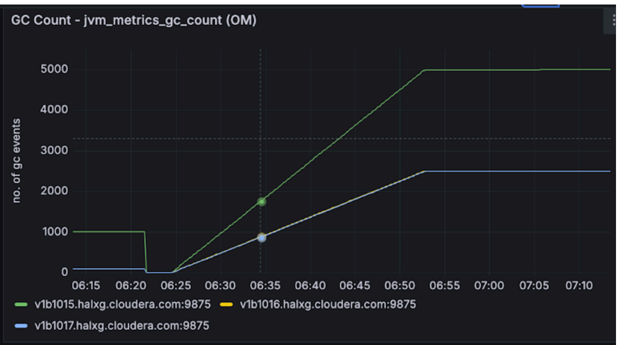


Old flow:

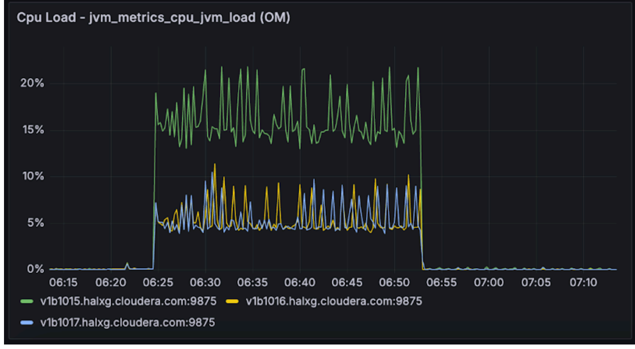
```
key-read-write-list
             count = 10000000
         mean rate = 5913.05 calls/second
     1-minute rate = 5822.04 calls/second
     5-minute rate = 5859.77 calls/second
    15-minute rate = 5567.46 calls/second
               min = 12.08 milliseconds
               max = 927.93 milliseconds
              mean = 17.38 milliseconds
            stddev = 27.01 milliseconds
            median = 14.94 milliseconds
              75% <= 16.25 milliseconds
              95% <= 32.03 milliseconds
              98% <= 35.81 milliseconds
              99% <= 39.04 milliseconds
            99.9% <= 76.52 milliseconds


Total execution time (sec): 1693  ⇒ 28 minute 29 second
```


GC Time - jvm_metrics_gc_time_millis (OM)


GC Count - jvm_metrics_gc_count (OM)

OM CPU Load Metrics


Cpu Load - jvm_metrics_cpu_jvm_load (OM)