Red Hat

Scale Testing RHCS with 10,000,000,000+ Objects

Karan Singh

Sr. Solution Architect

Cloud Storage & Data Services BU

Red Hat

Rare View cluster with 10B Objects

```
[root@rgw-5 ~]# ceph -s
  cluster:
    id:     795180c6-70c9-4cf1-a7a0-f0140add689f
    health: HEALTH_WARN
            13 nearfull osd(s)
            7 pool(s) nearfull
            BlueFS spillover detected on 296 OSD(s)
            19494 pgs not deep-scrubbed in time
            21070 pgs not scrubbed in time

  services:
    mon: 3 daemons, quorum rgw-1,rgw-2,rgw-3 (age 22h)
    mgr: rgw-4(active, since 5d), standbys: rgw-5, rgw-6
    osd: 318 osds: 318 up (since 5d), 318 in (since 5d)
    rgw: 12 daemons active (rgw-1.rgw0, rgw-1.rgw1, rgw-2.rgw0, rgw-2.rgw1, rgw-3.rgw0, rgw-3.rgw1, rgw-4.rgw0, rgw-4.rgw1, rgw-5.rgw0, rgw-5.r
gw1, rgw-6.rgw0, rgw-6.rgw1)

  task status:

  data:
    pools:   7 pools, 21760 p
    objects: 10.00G objects, 582 TiB
    usage:   3.7 PiB used, 1.0 PiB / 4.8 PiB avail
    pgs:     21732 active+clean
             28    active+clean+scrubbing+deep

  io:
    client:   83 MiB/s rd, 876 MiB/s wr, 84.84k op/s rd, 166.37k op/s wr

[root@rgw-5 ~]#
```

Why 10 Billion ? Motivations

- RHT tested 1 Billion Objects in Feb 2020 !! (What's Next ?)
  - https://www.redhat.com/en/blog/scaling-ceph-billion-objects-and-beyond
- Other Object Storage Systems **aspire** to scale to Billions of objects **one day**
  - Ceph can do it today, but can we Test ?
- Object Storage is getting popular for for Data Lake use cases
- Educate and Motivate Communities, Customers and Partners

Executive Summary

*"RHCS delivered **Deterministic Performance** at scale for both Small and Large object size workloads"*

Red Hat

Defining Scale

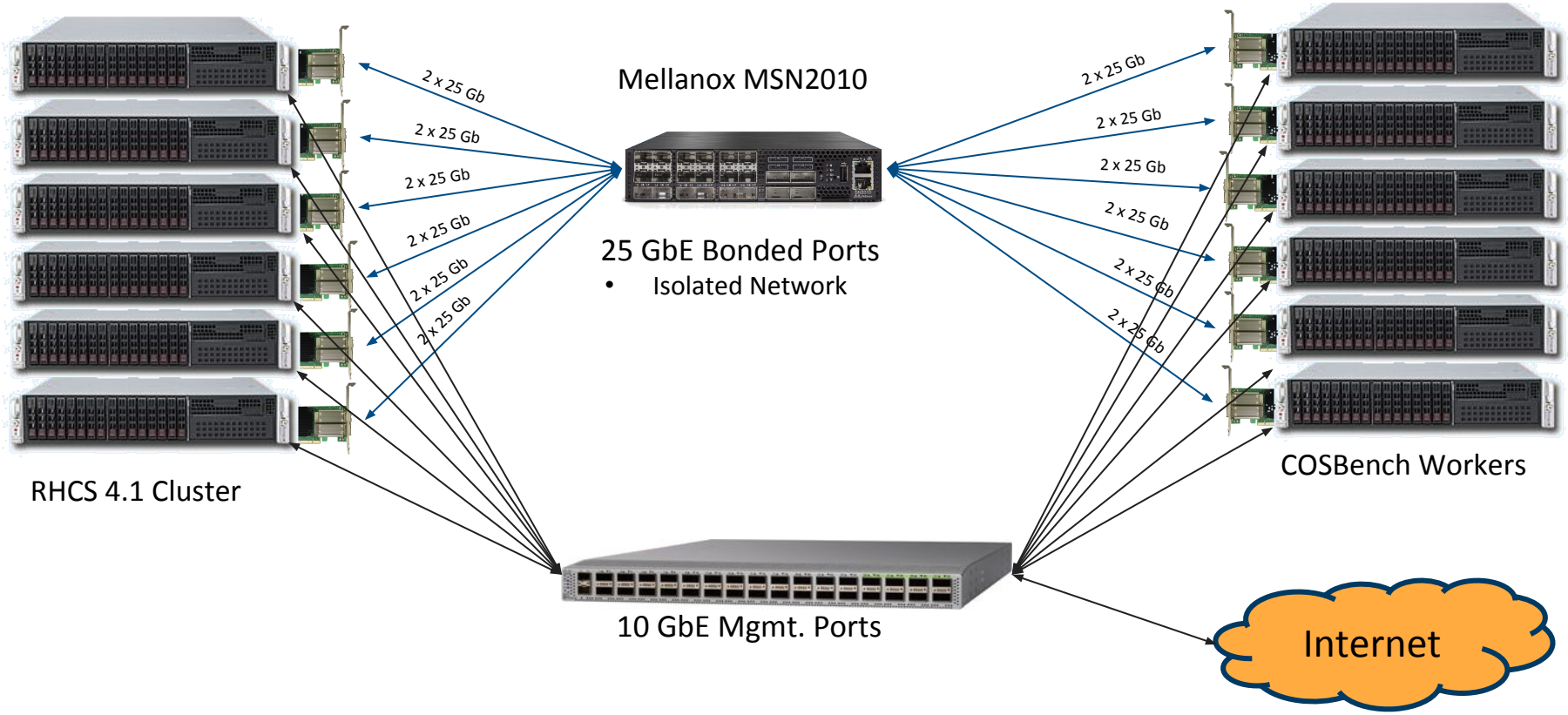
- 10,000,000,000+ Objects Ingested (*and retrieved*)
- 100,000+ Buckets
- 100,000 Objects / Bucket
- 318 HDDs / 36 NVMe devices
- 5.0 PB RAW capacity
- ~500 Test Runs

HW & SW Inventory

- 6 x RHCS Nodes
  - 53 x 16TB HDDs
    - Seagate Exos E 4U106
  - 6 x Intel QLC 7.6 TB
  - 2 x Intel Xeon Gold 6152
  - 256GB
  - 2 x 25GbE
- 6 x Client Nodes
  - 2 x 25GbE

- RHEL 8.1
- RHCS 4.1
  - Containerized Deployment
  - 2 x RGWs per RHCS node
  - EC 4+2
  - S3 Access Mode
  - 100K Objects / Bucket
- COSBench for workload generation
  - 6 x Drivers
  - 12 x Workers
    - 64 x Threads each

Test Lab Architecture



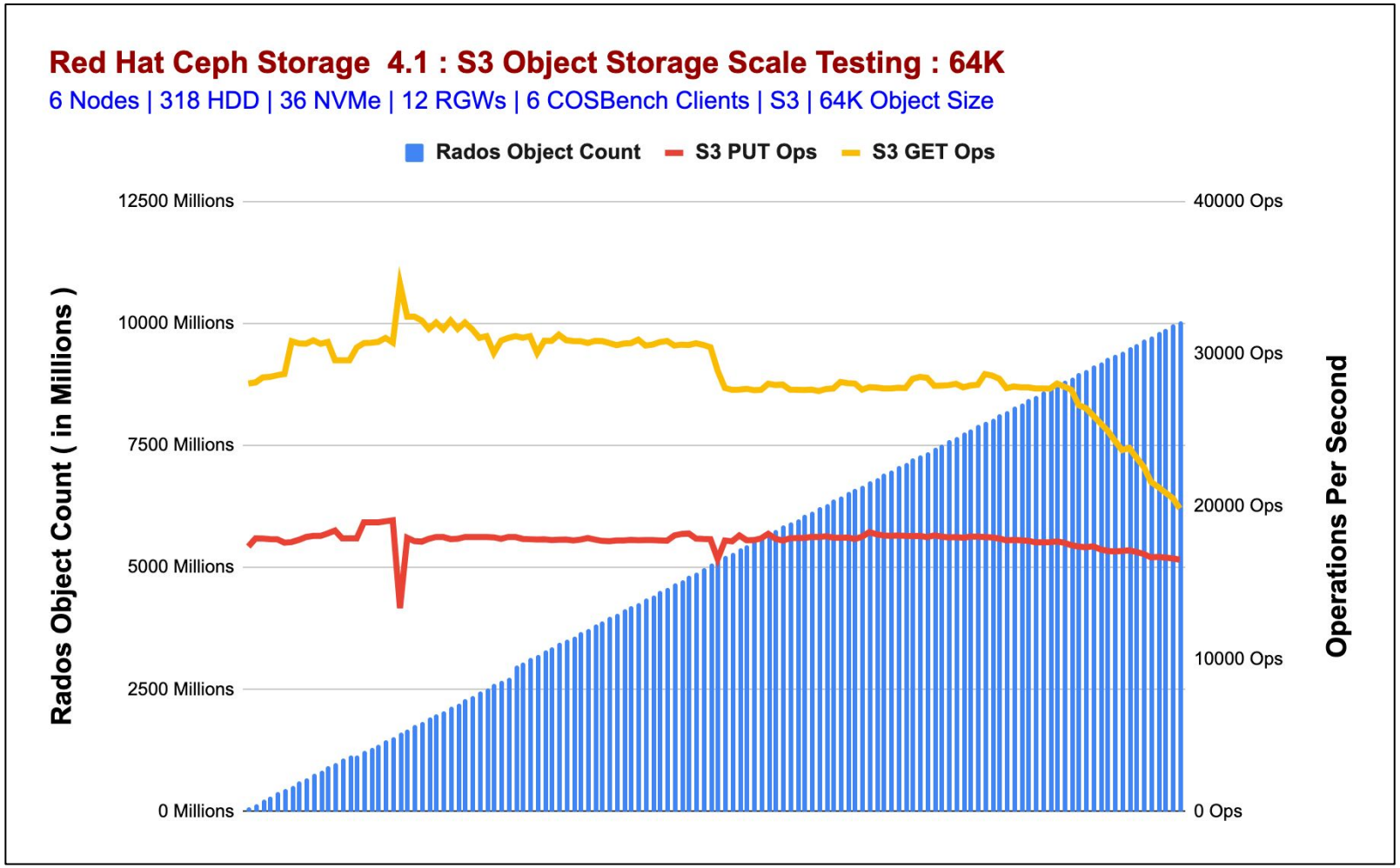
Workload Selection

- Object Sizes
  - 64KB (Small Objects)
  - 128MB (Large Objects)
- Access Pattern
  - 100% PUT
  - 100% GET
  - 70% GET, 20% PUT, 5% LIST, 5% Delete
- Degraded State Simulation
  - 1 x HDD Down
  - 6 x HDDs Down
  - 53 x HDDs Down (1 Node Failure)

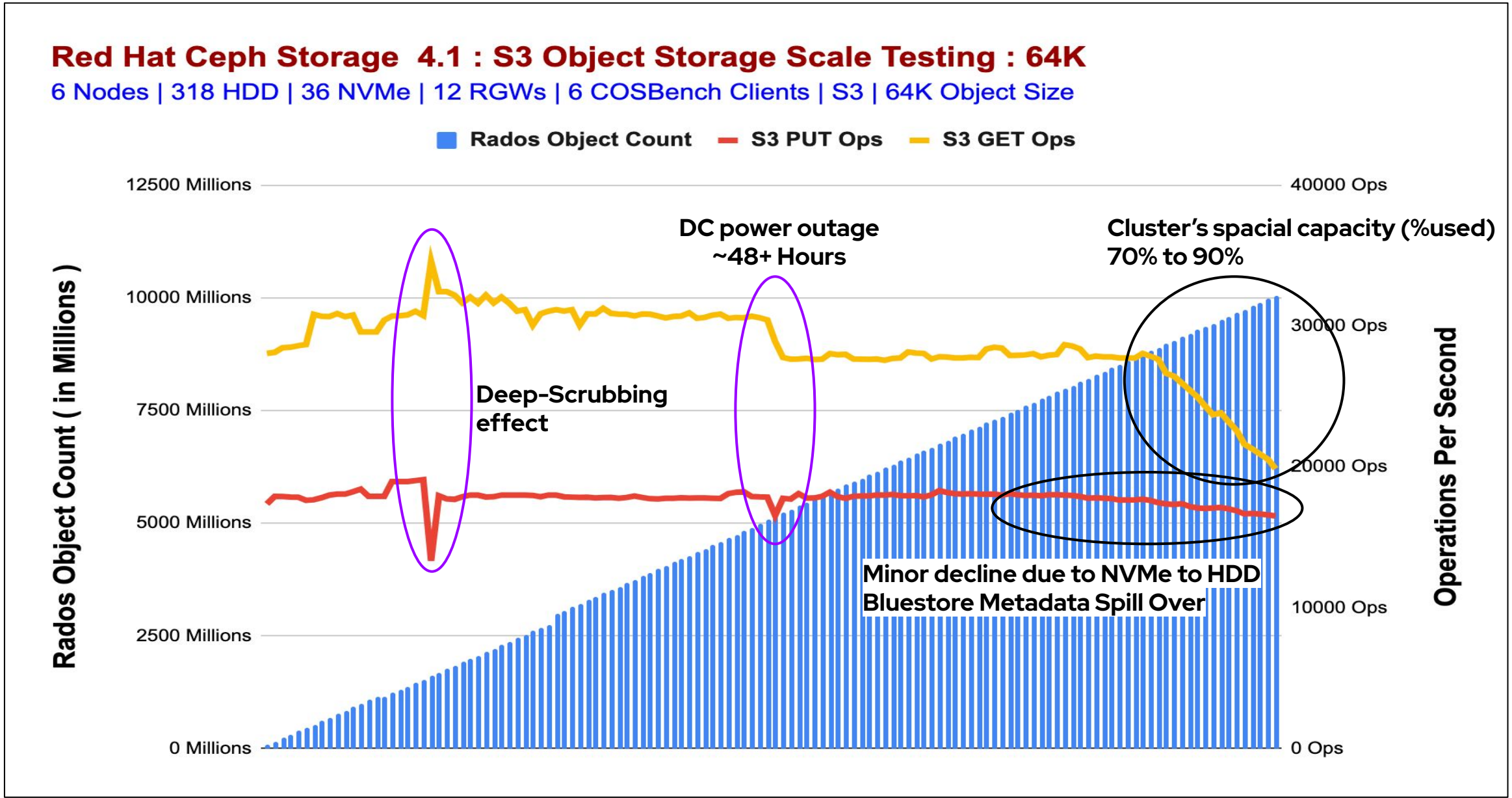
Small Object Performance : Operations Per Sec

- Average Cluster Performance
  - ~17,800 S3 PUT Ops
  - ~28,800 S3 GET Ops

- Avg Single HDD OSD Perf.
  - 60 S3 PUT Ops
  - 90 S3 GET Ops

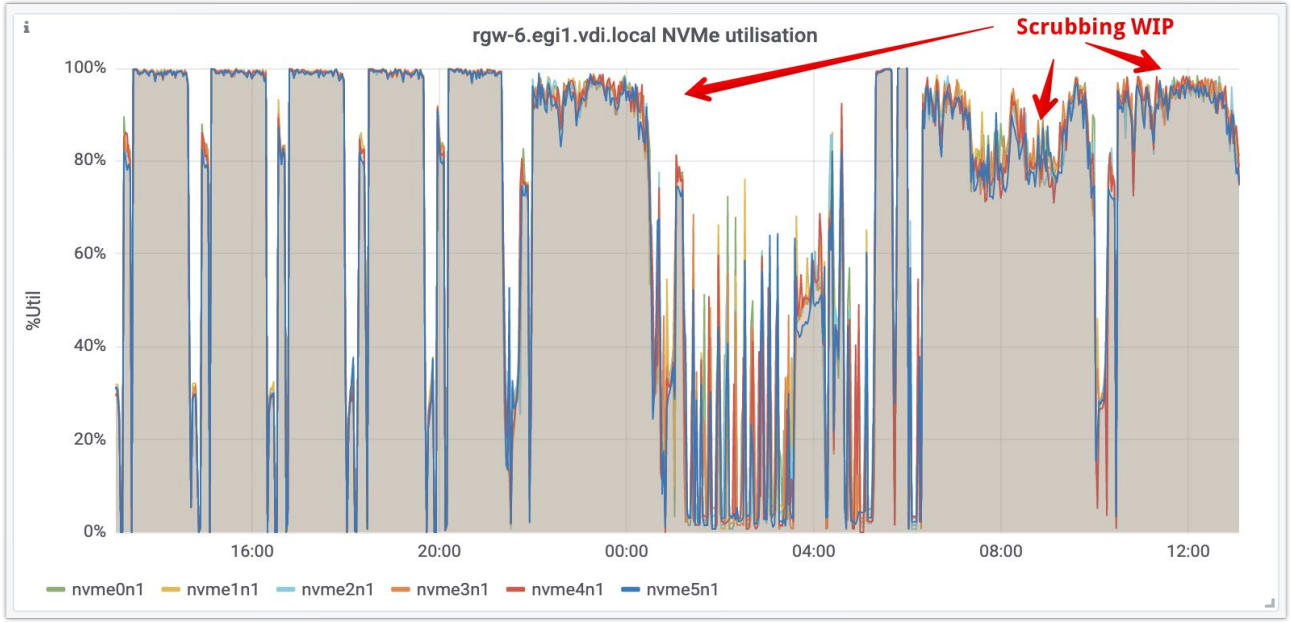


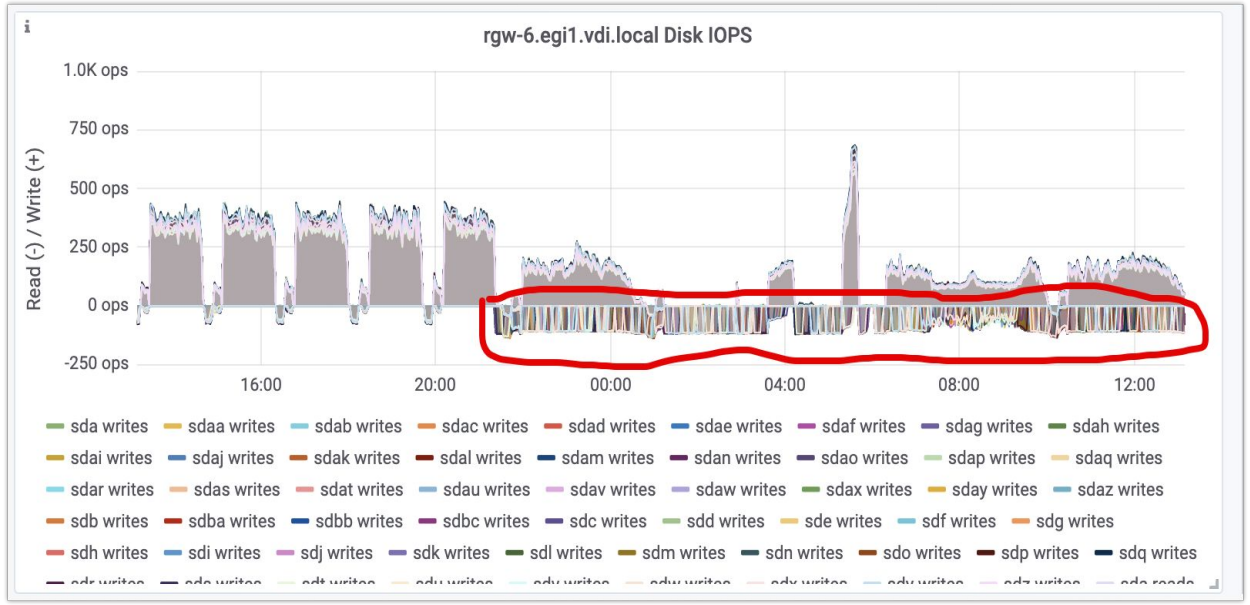
Small Object Performance *Dissection*



Small Object Performance *Dissection*

Deep-Scrubbing Affirmations



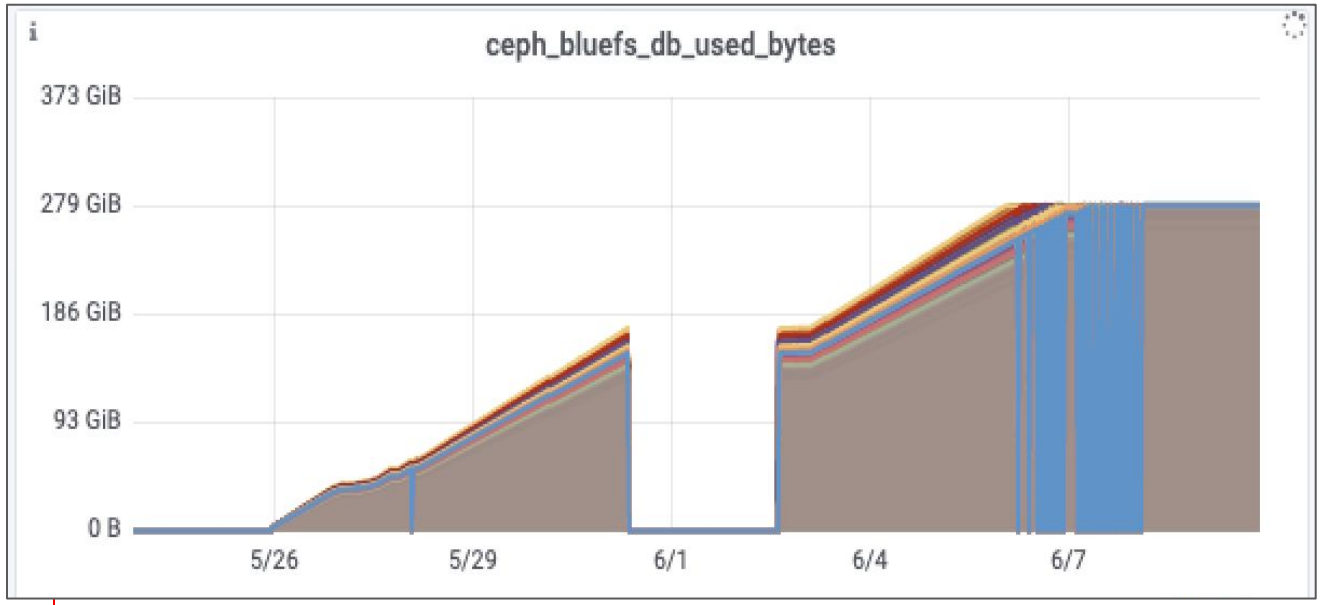


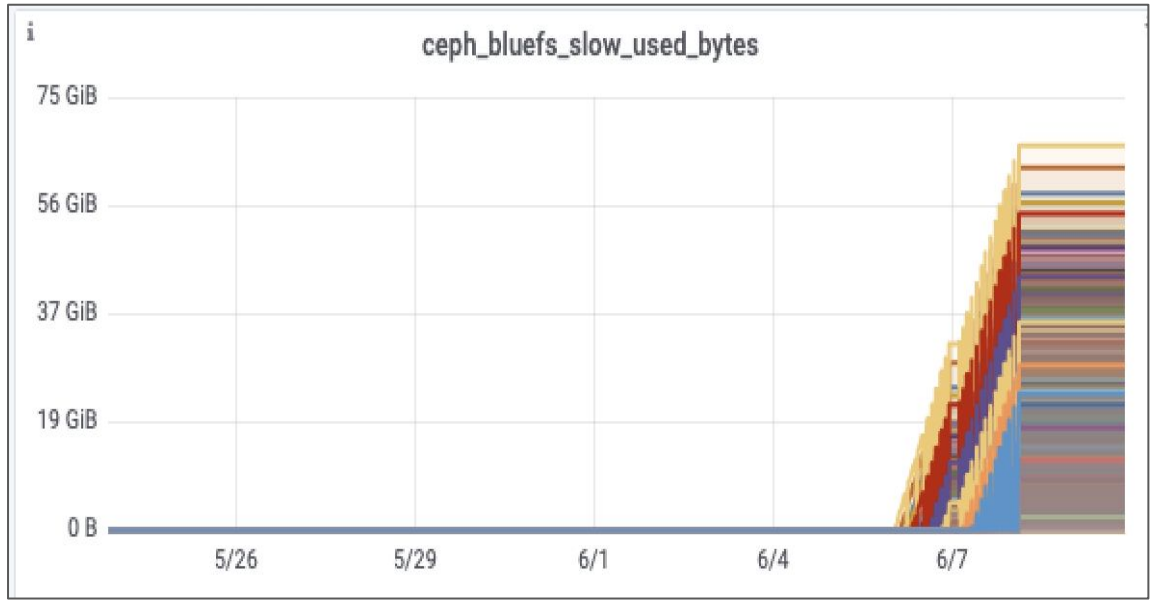
Red Hat

Small Object Performance *Dissection*

- Bluestore uses RocksDB
- RocksDB uses Level Style Compaction
  - L0: in memory
  - L1: 256MB
  - L2: 2.56 GB
  - L3: 25.6 GB
  - L4: 256 GB
  - L5: 2.56 TB ← L5 could not fit in Flash, hence spilled over to HDD
  - L6: 25.6 TB

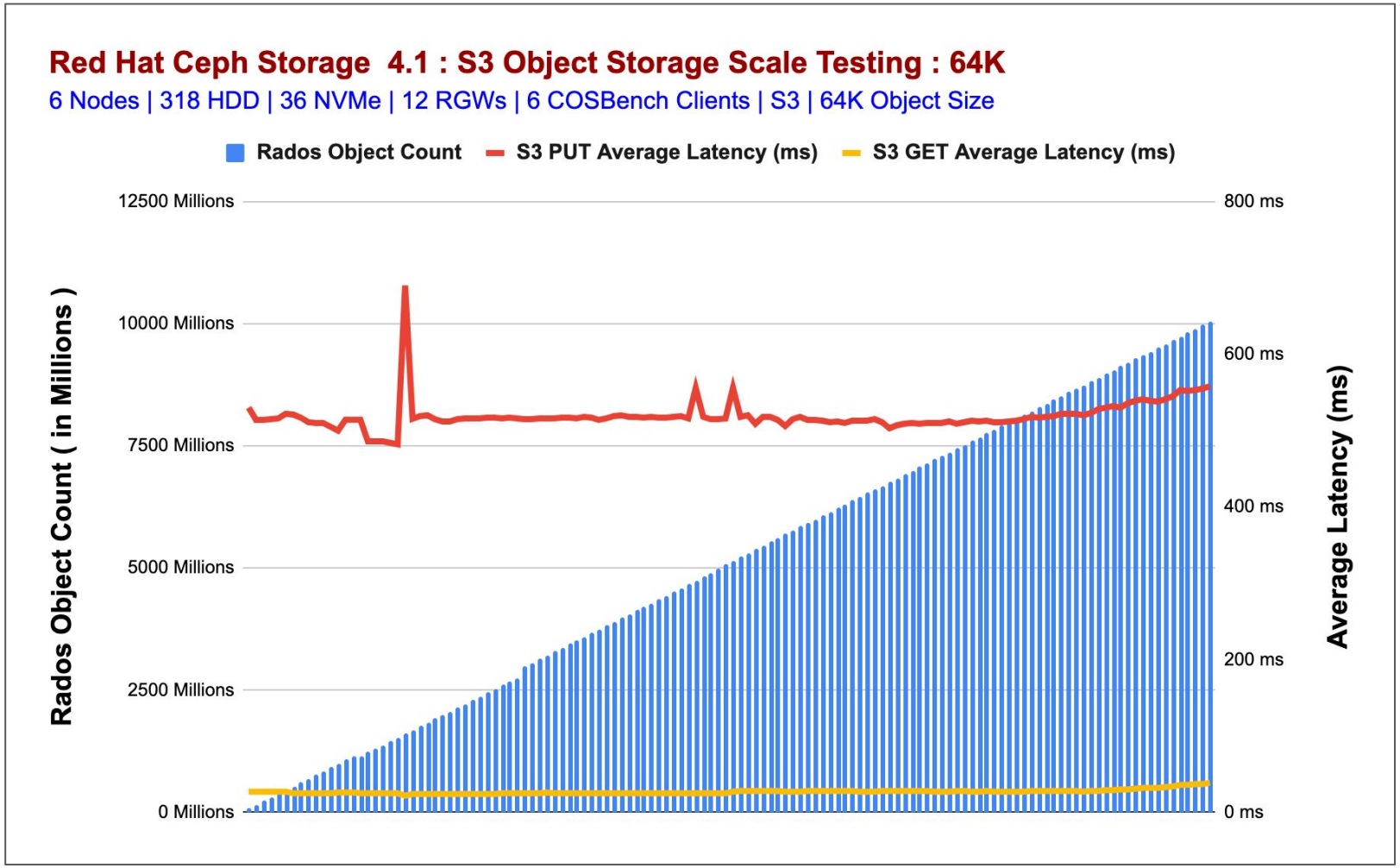
Bluestore and RocksDB Details
https://www.redhat.com/en/blog/scaling-ceph-billion-objects-and-beyond





Small Object Performance : Latency

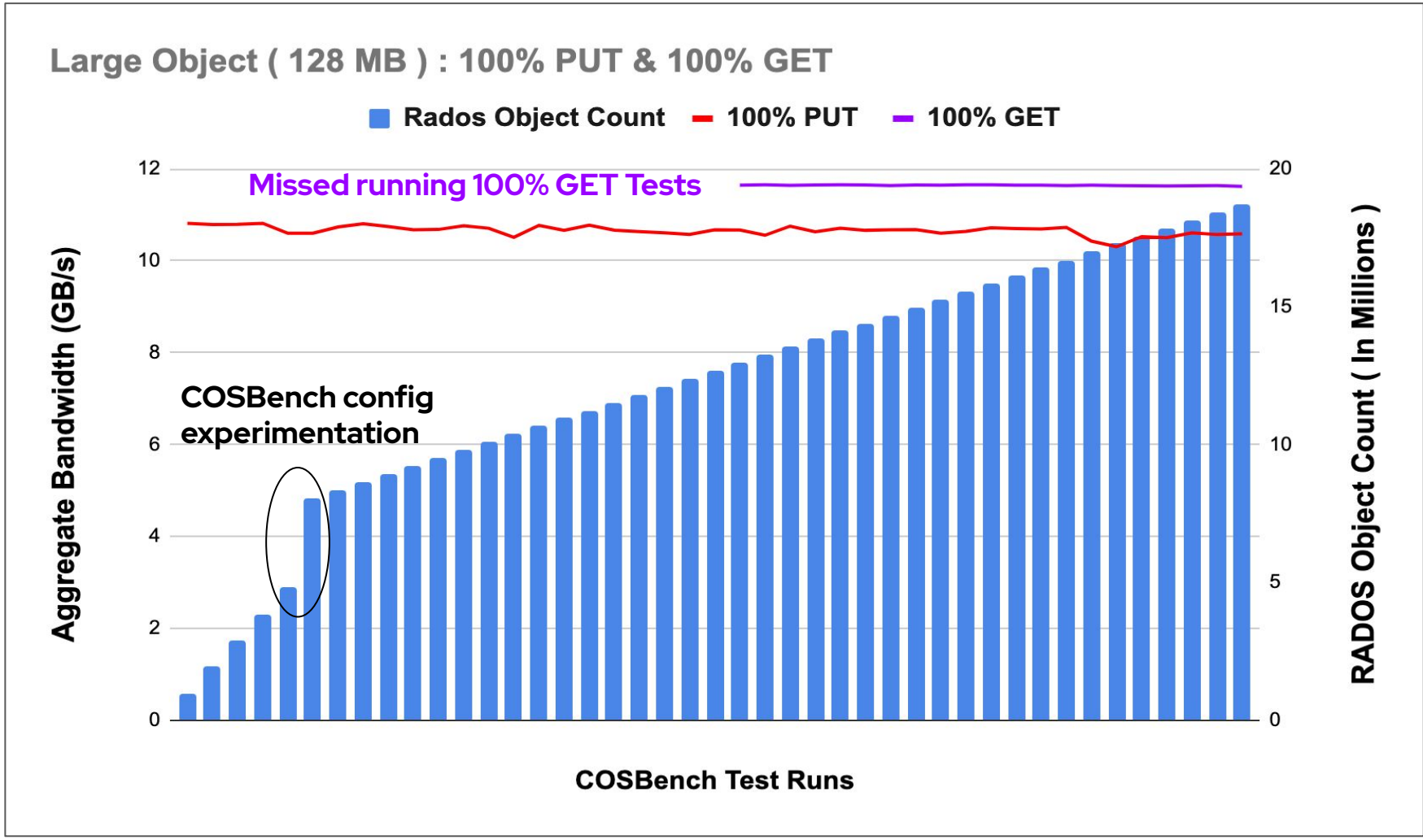
- Average Cluster Latency
  - 510 ms S3 PUT Latency
  - 27 ms S3 GET Latency



Large Object Performance : Bandwidth

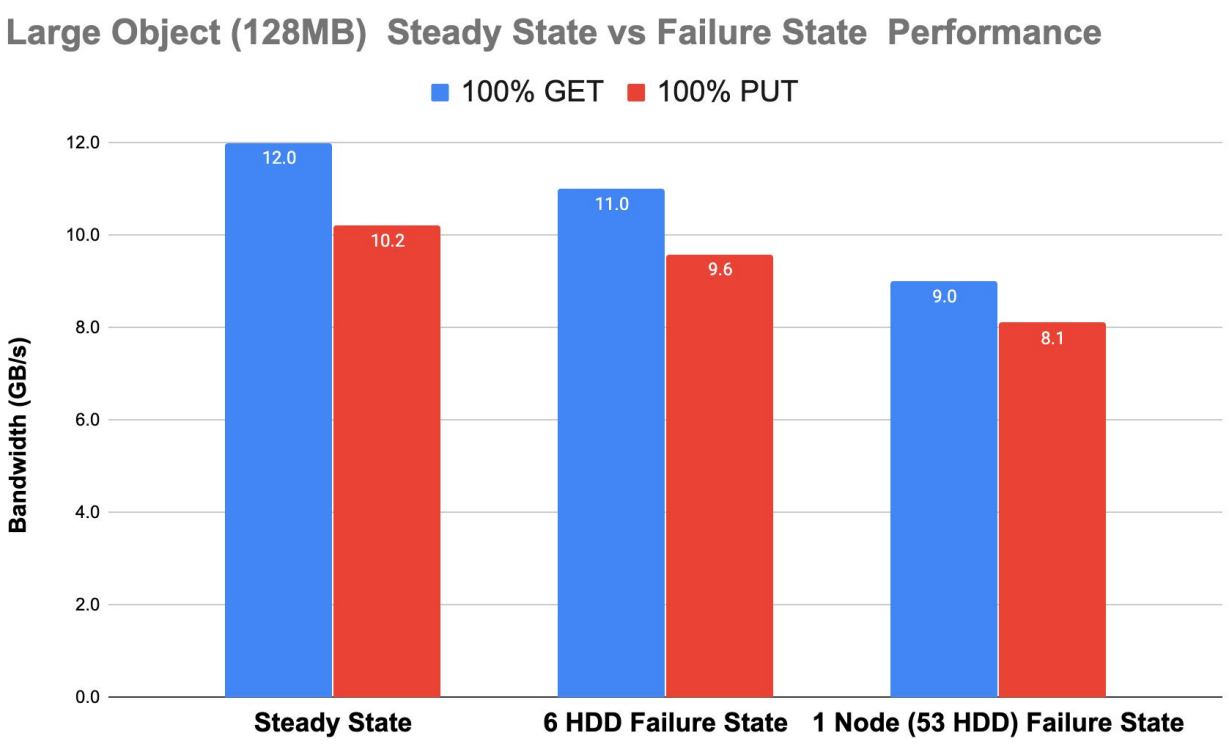
- Average Cluster Performance
  - ~10.7 GB/s S3 PUT BW
  - ~11.6 GB/s S3 GET BW

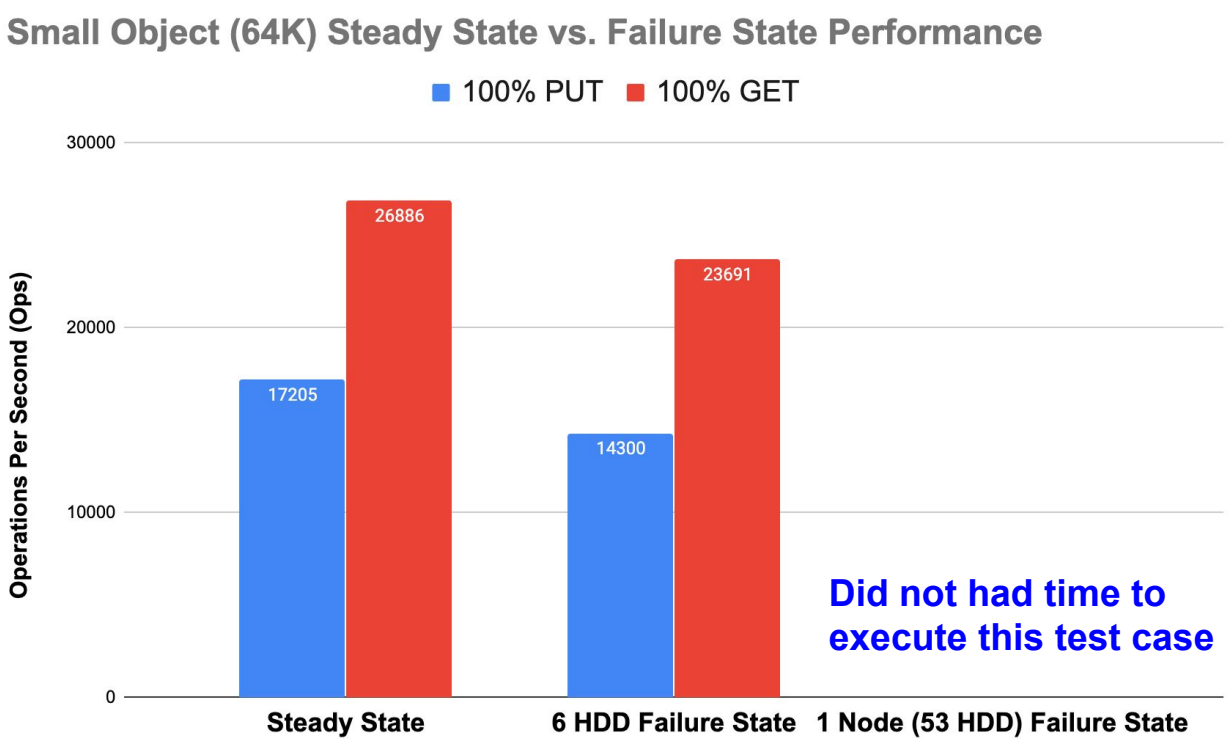
- Avg Single HDD OSD Perf.
  - 34 MBps S3 PUT BW
  - 37 MBps S3 GET BW



Performance during Degraded State

Total 318 HDDs	Storage Failure (%)	PUT Perf Drop (%)	GET Perf Drop (%)
6 HDDs Failed	2	6	8
53 HDDs Failed	17	21	25





Sizing Guidance

- I needs X Ops and Y GBps for S3 workload ? How to Size ?
  - Not a silver bullet, but can give you a ballpark number

Single HDD OSD Performance (with 4% Flash for Bluestore)		
S3 Access	**100% PUT**	**100% GET**
Small Object (64K)	60 Ops	90 Ops
Large Object (128M)	34 MBps	37 MBps

- Use 2 RGWs Instances per Ceph Node
- RHT recommendation of 4% for Bluestore is good at scale as well
  - Increase “max_bytes_for_level_base” (default 256MB) such that you can get most of your 4% Bluestore Flash allocation
- Embrace Co-located & Containerized Storage Demons
- Go big on osd_memory_target if you can (8-10 GB is good to have)

Summary

- Our testing showed RHCS achieving deterministic performance at scale for both Small and Large Object sizes, PUT and GET operations, before hitting resource saturation, capacity limits
- Performance during failure scenarios found to be acceptable
- Undoubtedly RHCS can scale a lot more than what we tested
  - 10 Billion objects are just Tested Maximum, This is NOT A LIMIT

Download the full performance report http://red.ht/10billion

Download the full performance report at
http://red.ht/10billion

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat