Unlocking The Performance Secrets of Ceph Object Storage

and

a Teaser for Shared Data Lake Solution

Karan Singh
Sr. Storage Architect
Storage Solution Architectures Team

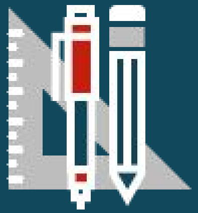
WHO DO WE HELP?

COMMUNITY

SOLUTIONS ARCHITECTS

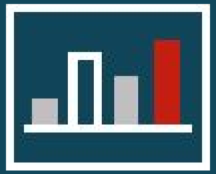
CUSTOMERS

HOW DO WE HELP?



ARCHITECTURE

OPTIMIZATIONS



PERFORMANCE

redhat.

OBJECT STORAGE

WHAT IS OBJECT STORAGE IS GOOD FOR ?

MULTI-PETABYTE!

DIGITAL MEDIA

DATA WAREHOUSE

ARCHIVE

BACKUP

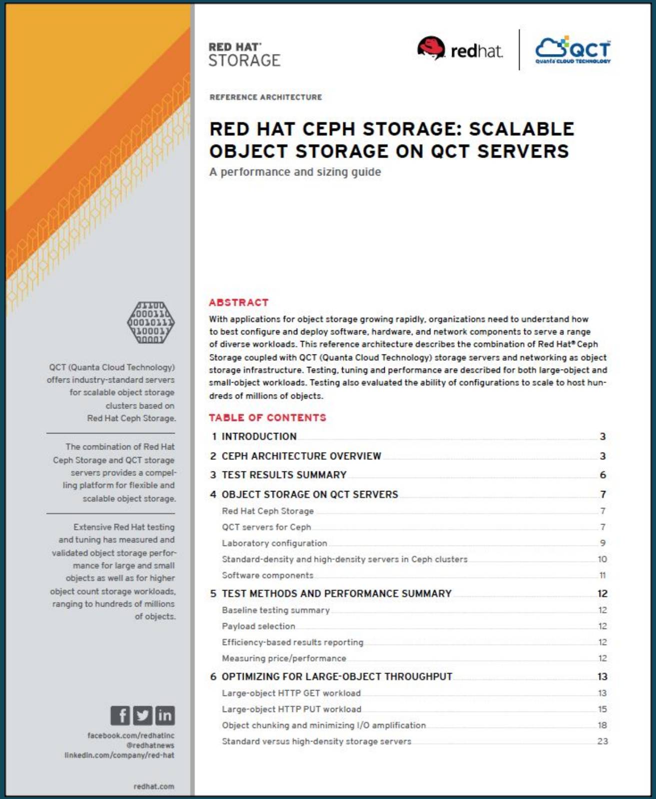
OBJECT STORAGE PERFORMANCE & SIZING ?

- Executed ~1500 unique tests
- Evaluated Small, Medium & Large Object sizes
- Storing millions of Objects (~130 Million to be precise)
  - 400 hours of testing for one of the scenario
- Have seen performance improvements as high as ~500%
- Testing was conducted on QCT sponsored LAB

QCT™
LAB PARTNER

redhat.

PERFORMANCE & SIZING GUIDE



RED HAT STORAGE

redhat. | QCT Quanta Cloud Technology

REFERENCE ARCHITECTURE

RED HAT CEPH STORAGE: SCALABLE OBJECT STORAGE ON QCT SERVERS

A performance and sizing guide

QCT (Quanta Cloud Technology) offers industry-standard servers for scalable object storage clusters based on Red Hat Ceph Storage.

The combination of Red Hat Ceph Storage and QCT storage servers provides a compelling platform for flexible and scalable object storage.

Extensive Red Hat testing and tuning has measured and validated object storage performance for large and small objects as well as for higher object count storage workloads, ranging to hundreds of millions of objects.

facebook.com/redhatinc
@redhatnews
linkedin.com/company/red-hat

redhat.com

ABSTRACT

With applications for object storage growing rapidly, organizations need to understand how to best configure and deploy software, hardware, and network components to serve a range of diverse workloads. This reference architecture describes the combination of Red Hat® Ceph Storage coupled with QCT (Quanta Cloud Technology) storage servers and networking as object storage infrastructure. Testing, tuning and performance are described for both large-object and small-object workloads. Testing also evaluated the ability of configurations to scale to host hundreds of millions of objects.

TABLE OF CONTENTS

Benchmarking Environment & Methodology

Hardware Configurations Tested

QCT LAB PARTNER

STANDARD DENSITY SERVERS

- 6x OSD Nodes
  - **12x HDD (7.2K rpm, 6TB SATA)**
  - **1x Intel P3700 800G NVMe**
  - 128G Memory
  - **1x Intel Xeon E5-2660 v3**
  - 1x 40GbE
- 4x RGW Nodes
  - 1x 40GbE
  - 96G Memory
  - 1x Intel Xeon E5-2670 v3
- 8x Client Nodes
  - 1x 10GbE
  - 96G Memory
  - 1x Intel Xeon E5-2670
- 1x Ceph MON (*)
  - 1x 10GbE
  - 96G Memory
  - 1x Intel Xeon E5-2670

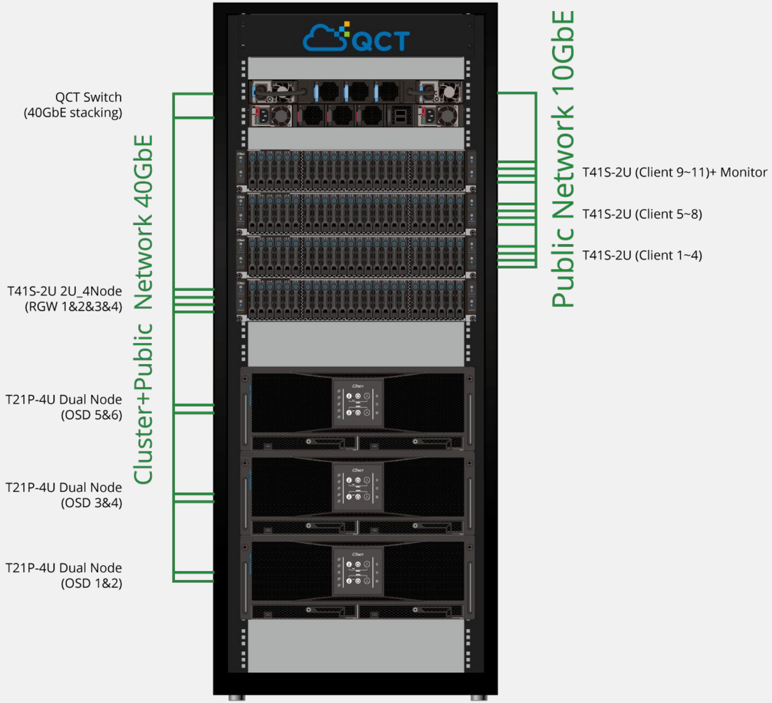
HIGH DENSITY SERVERS

- 6x OSD Nodes
  - **35x HDD (7.2K rpm, 6TB SATA)**
  - **2x Intel P3700 800G NVMe**
  - 128G Memory
  - **2x Intel Xeon E5-2660 v3**
  - 1x 40GbE
- 4x RGW Nodes
  - 1x 40GbE
  - 96G Memory
  - 1x Intel Xeon E5-2670 v3
- 8x Client Nodes
  - 1x 10GbE
  - 96G Memory
  - 1x Intel Xeon E5-2670
- 1x Ceph MON(*)
  - 1x 10GbE
  - 96G Memory
  - 1x Intel Xeon E5-2670

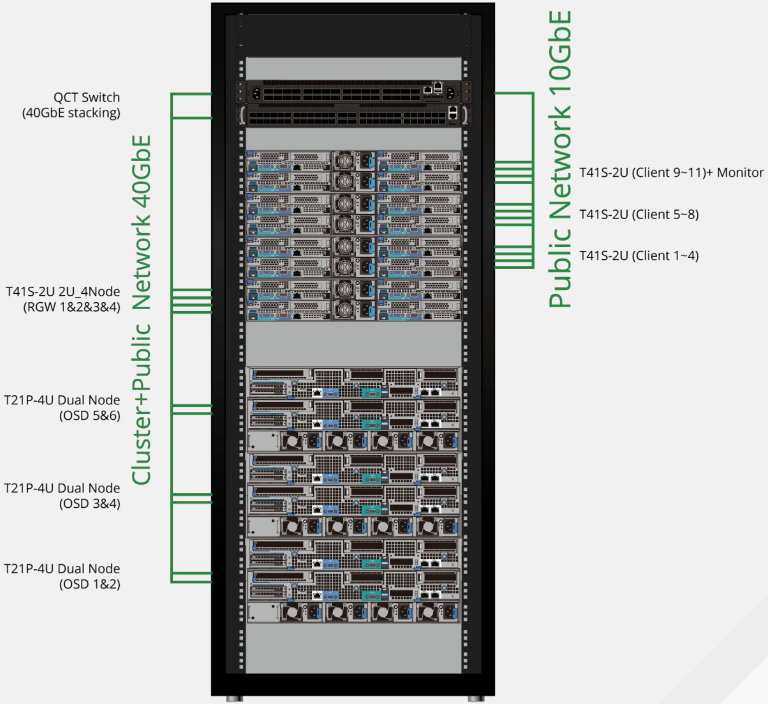
(*) For production usage use 3 MONs at least

Lab Setup

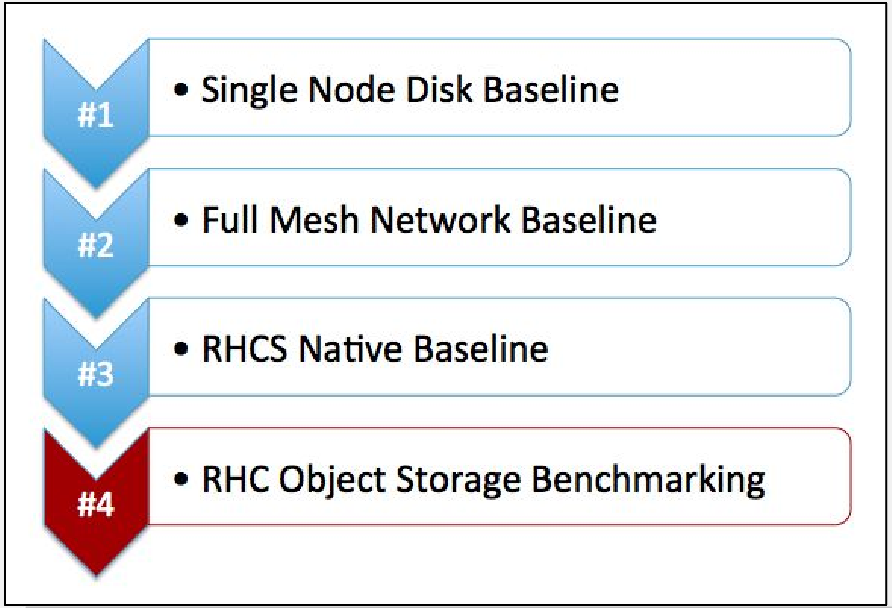
Front View



Rear View



Benchmarking Methodology & Tools



CEPH	Red Hat Ceph Storage 2.0 (Jewel)
OS	Red Hat Enterprise Linux 7.2
TOOLS	• Ceph Benchmarking Tool (CBT) • FIO 2.11-12 • iPerf3 • COSBench 0.4.2.c3 • Intel Cache Acceleration Software (CAS) 03.01.01

Payload Selection

- **64K** - Small size object (Thumbnail images, small files etc.)
- **1M**   - Medium size object (Images, text files etc.)
- **32M** - Large size object (Analytics Engines , HD images , log files, backup etc.)
- **64M** - Large size object (Analytics Engines , Videos etc.)

Optimizing for Small Object Operations (OPS)

Small Object (Read Ops)

- Client read ops scaled **linearly** while increasing RGW hosts (OSDs on standard density servers)
- Performance limited by **number of RGW hosts** available in our test environment
- Best observed perf. was 8900 Read OPS with **bucket index on Flash Media** on standard density servers

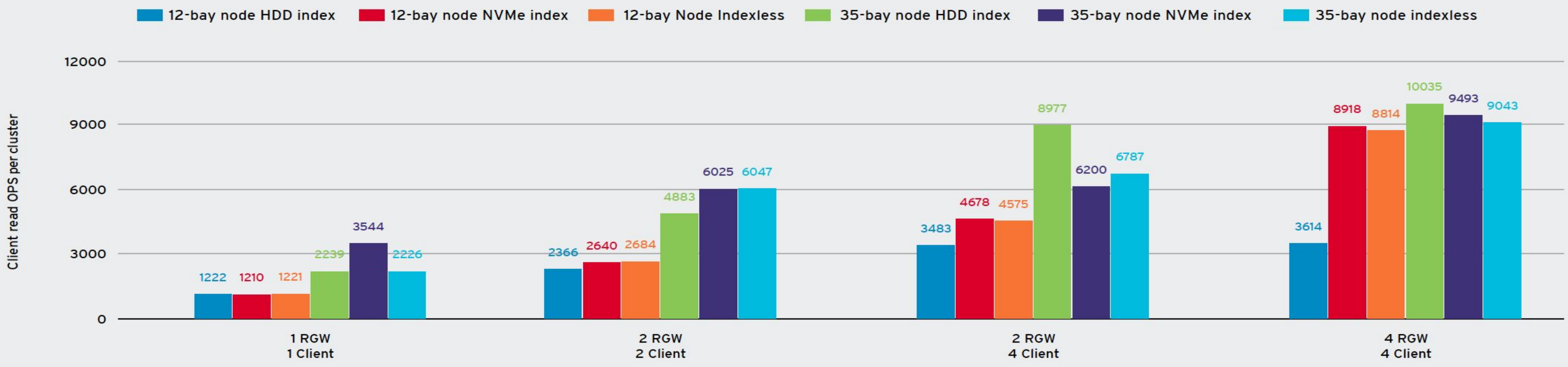


Figure 19. Aggregate small-object read (HTTP GET) performance per OSD across a variety of index configurations on standard-density versus high-density servers (64KB object size, higher is better).

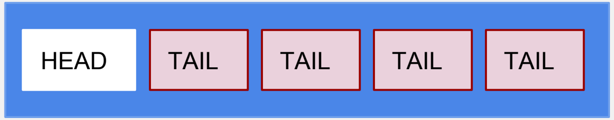
Read Operations are simple !!

… Let’s Talk About Write …

RGW Object Consists of

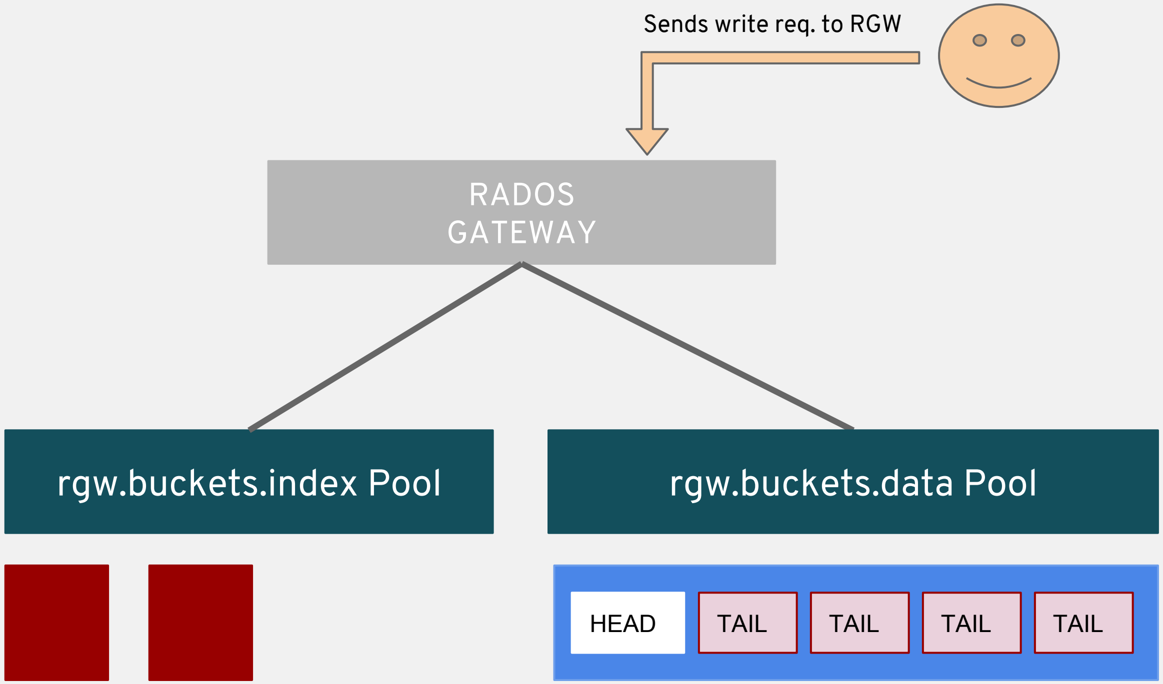
Object Index

Object Data



redhat.

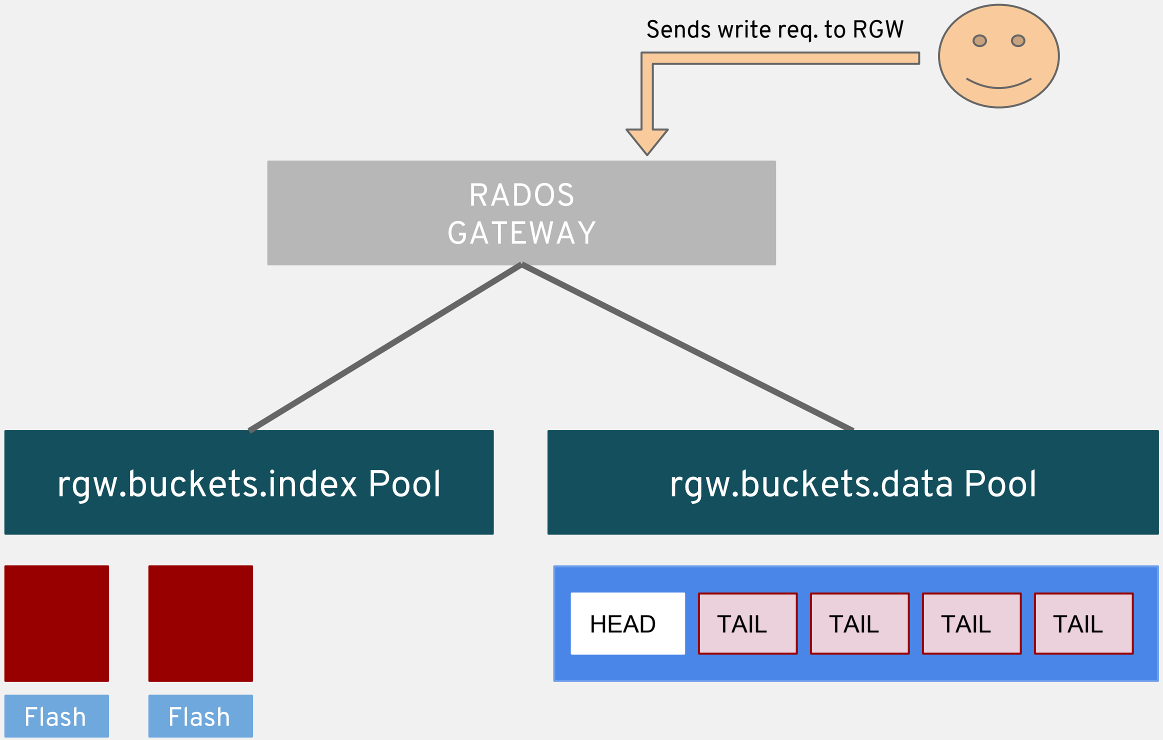
RGW Write Operation



Can we Improve This ?

... Yes We Can !!

Put bucket index on Flash Media



Small Object (Write Ops)

- Client write ops scaled **sub-linearly** while increasing RGW hosts
- Performance **limited by disk saturation** on Ceph OSD hosts
- Best observed Write OPS was 5000 on High Density Servers with **bucket index on flash media**

Higher write ops could have been achieved by adding more OSD hosts

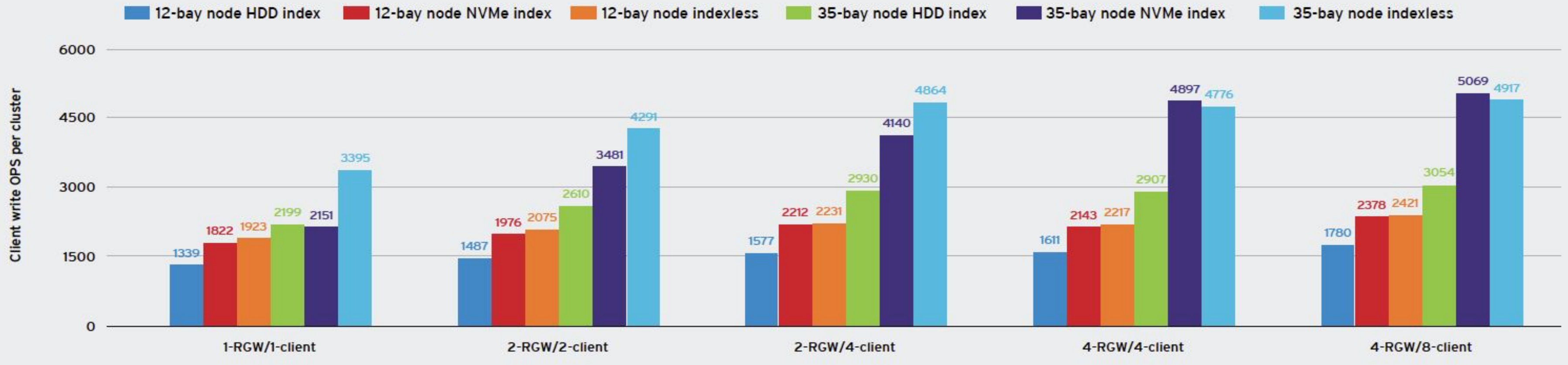


Figure 24. Small-object write (HTTP PUT) performance was greatly accelerated by placing bucket indices on NVMe (64KB object size, aggregate write performance per cluster).

Small Object (Write Ops) Saturating HDDs

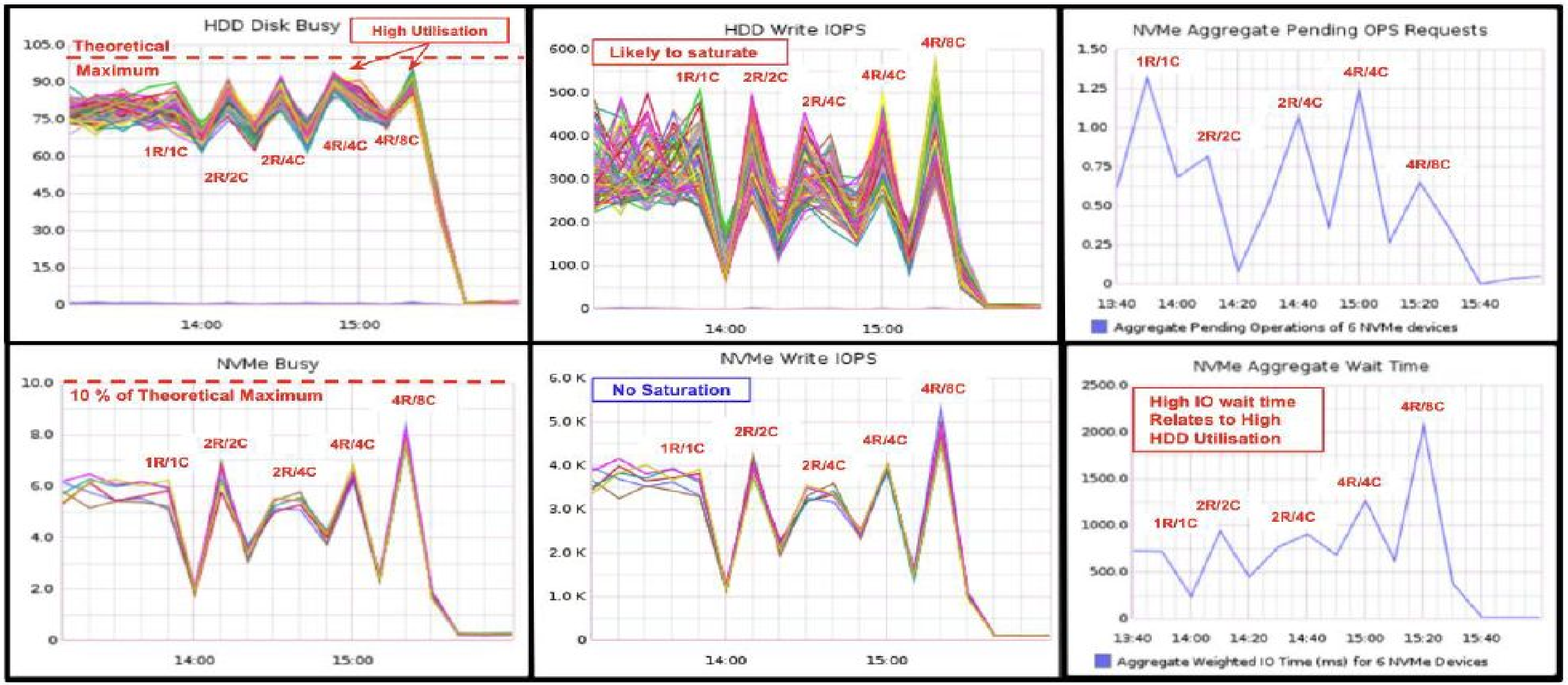


Figure 25. Small-object write (PUT) for standard-density servers was initially limited by HDD saturation, as evidenced by high NVMe aggregate wait time (#R/#C indicates number of RGW per number of clients).

Latency Improvements

Configuring Bucket Index on flash helped reduce:

- P99 Read latency reduced by ~149% (Standard density servers)
- P99 Write latency reduced by ~69% (High density servers)

SERVER CONFIGURATION	TOTAL OSDs IN CLUSTER	OBJECT SIZE	BUCKET INDEX CONFIGURATION	AVERAGE READ LATENCY (ms)	AVERAGE WRITE LATENCY (ms)
Standard Density	72	64 K	HDD Index	142	318
Standard Density	72	64 K	NVMe Index	**57**	**229**
High Density	210	64 K	HDD Index	51	176
High Density	210	64 K	NVMe Index	54	**104**

Optimizing for Large Object Throughput (MBps)

Large Object (Read MBps)

- Client read throughput scaled **near-linearly** while increasing RGW hosts
- Performance limited by **number of RGW hosts** available in our test environment
- Best observed read throughput was ~4GB/s with 32M object size on standard density servers

Higher read throughput could have been achieved by adding more RGW hosts

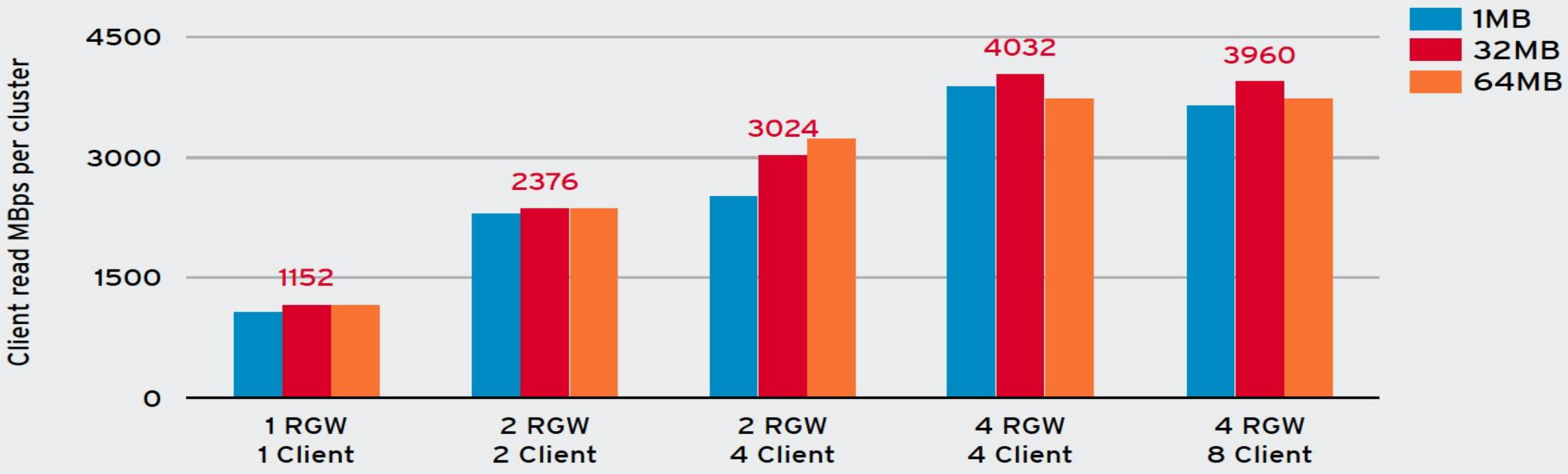
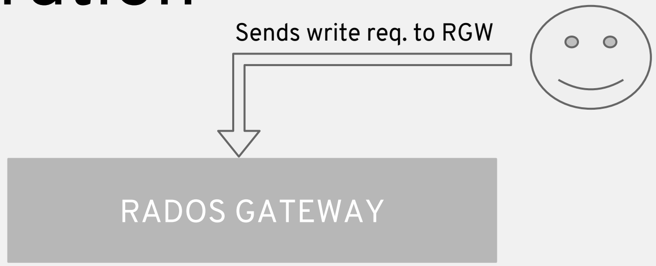


Figure 5. Large-object read test, cluster-wide aggregate performance, on standard-density servers.

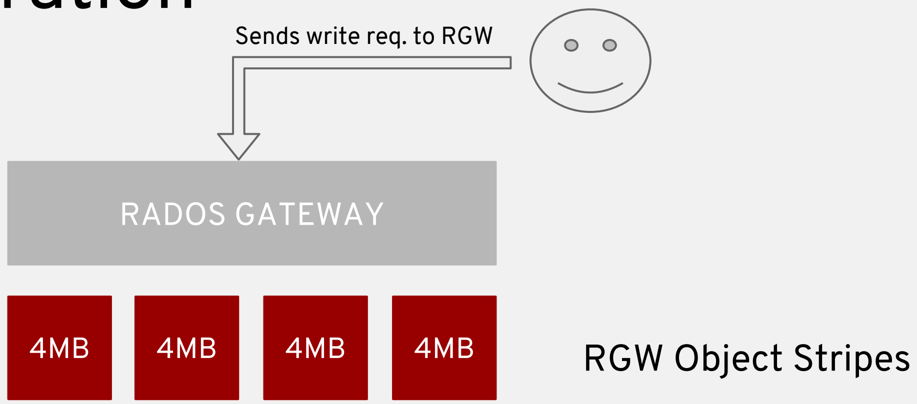
Read Operations are simple !!

… Let’s Talk About Write …

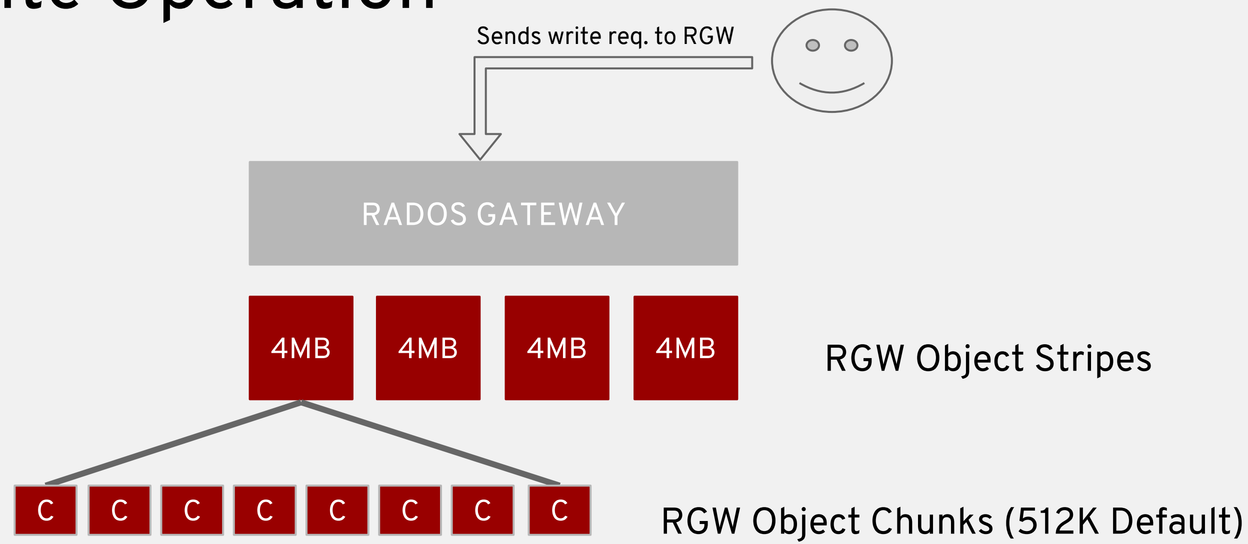
RGW Write Operation



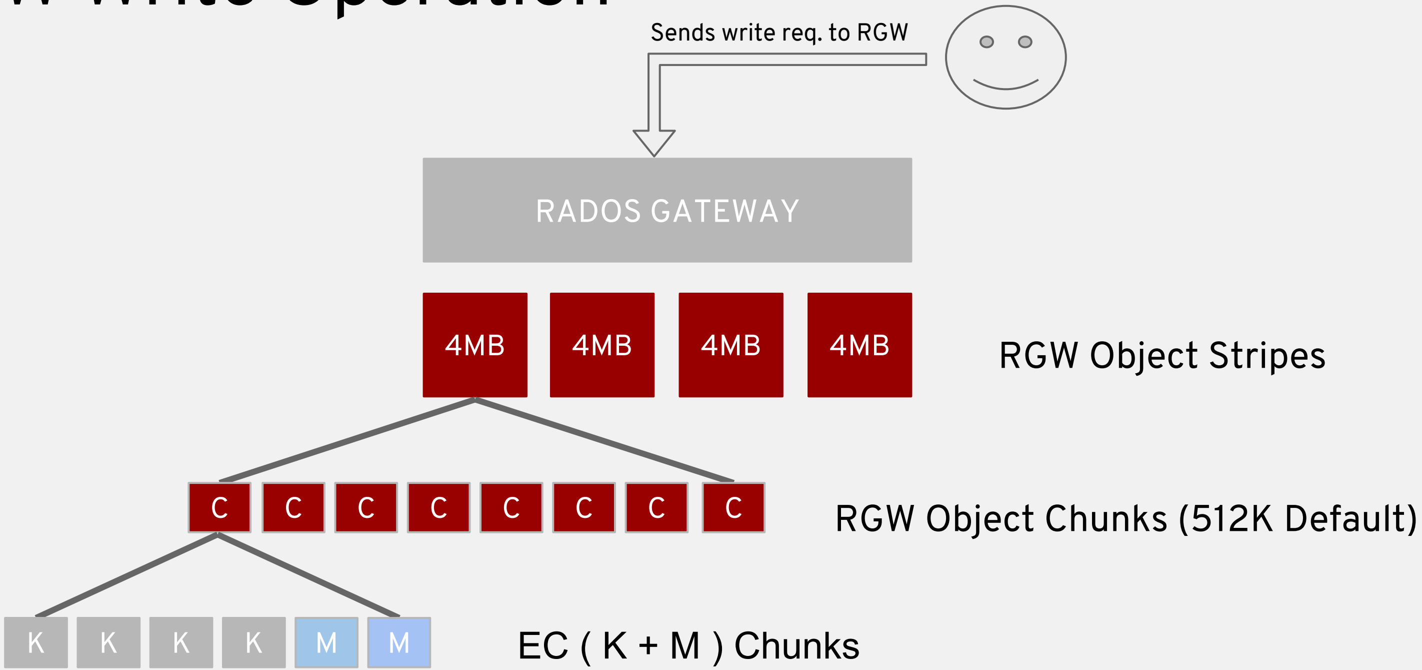
RGW Write Operation



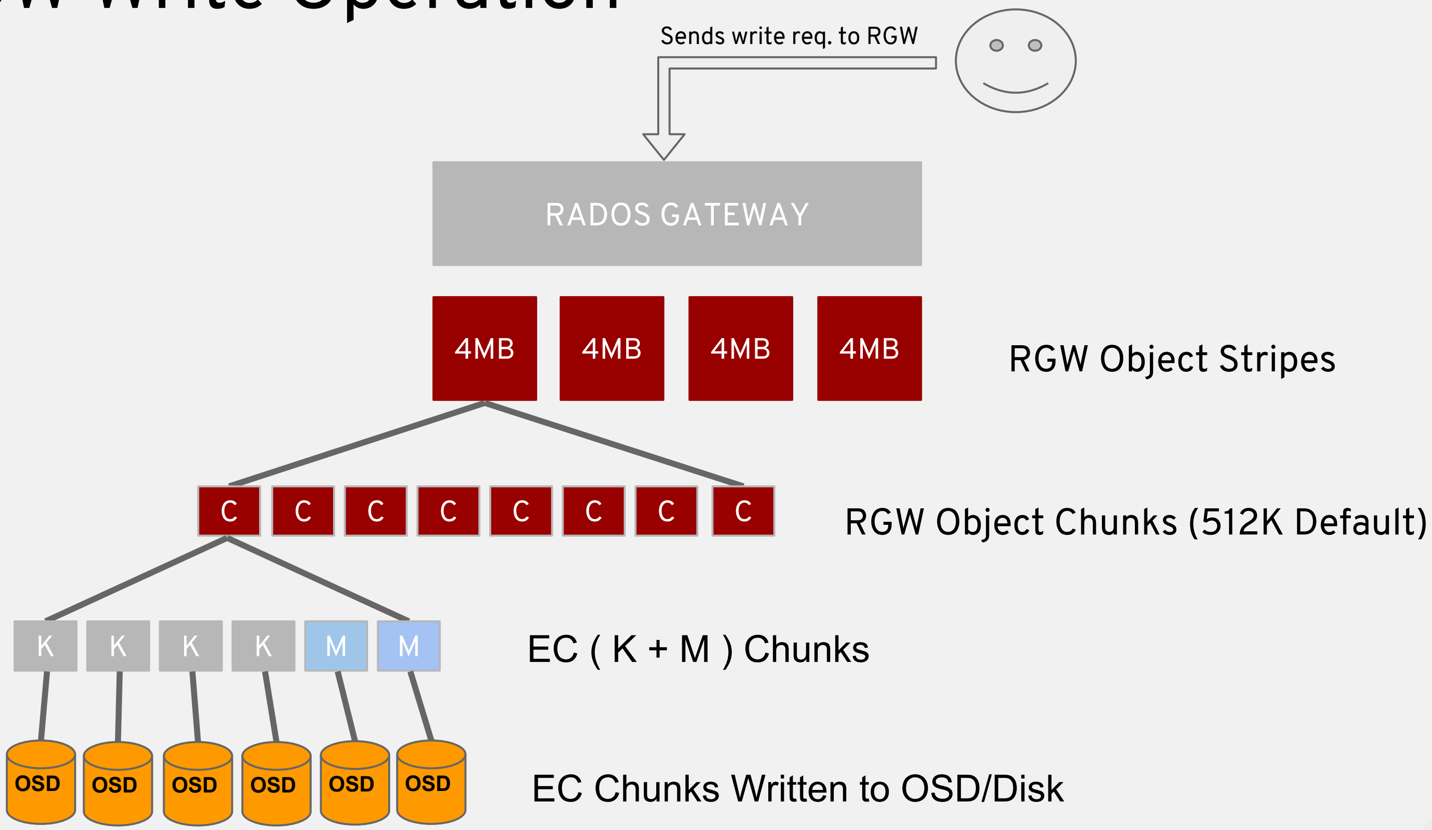
RGW Write Operation



RGW Write Operation



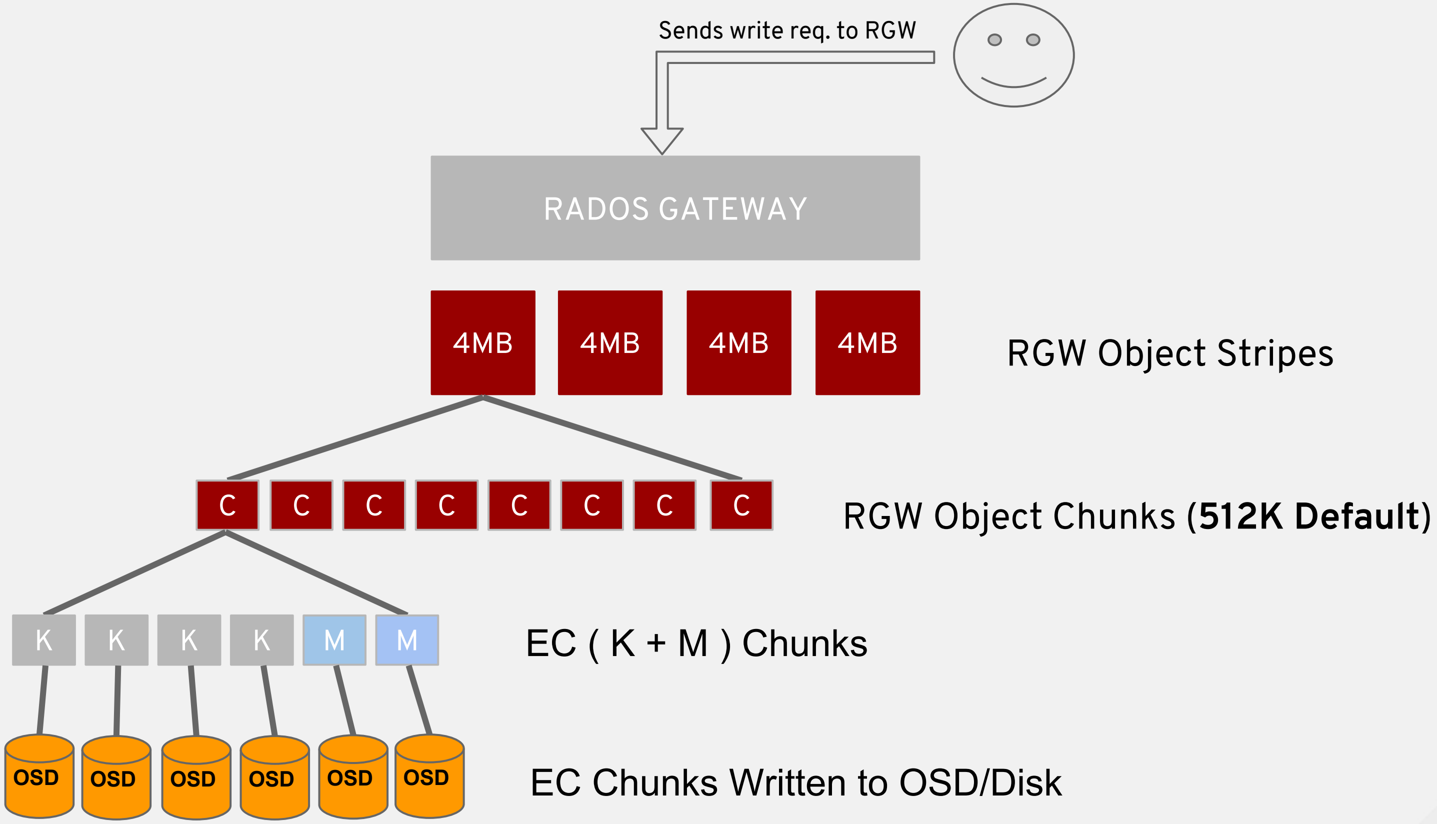
RGW Write Operation



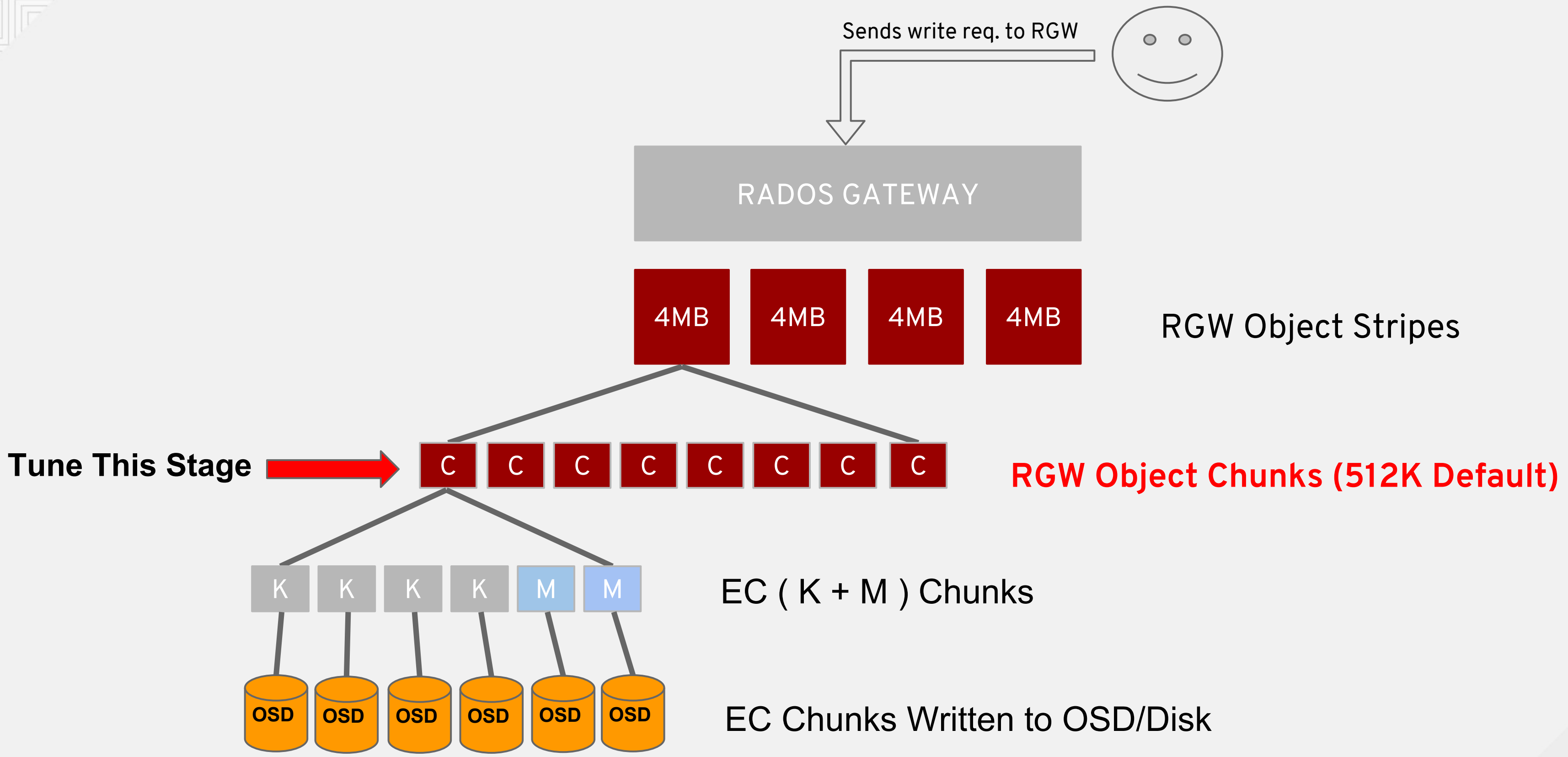
Can we Improve This ?

... Yes We Can !!

RGW Write Operation (Before Tuning)

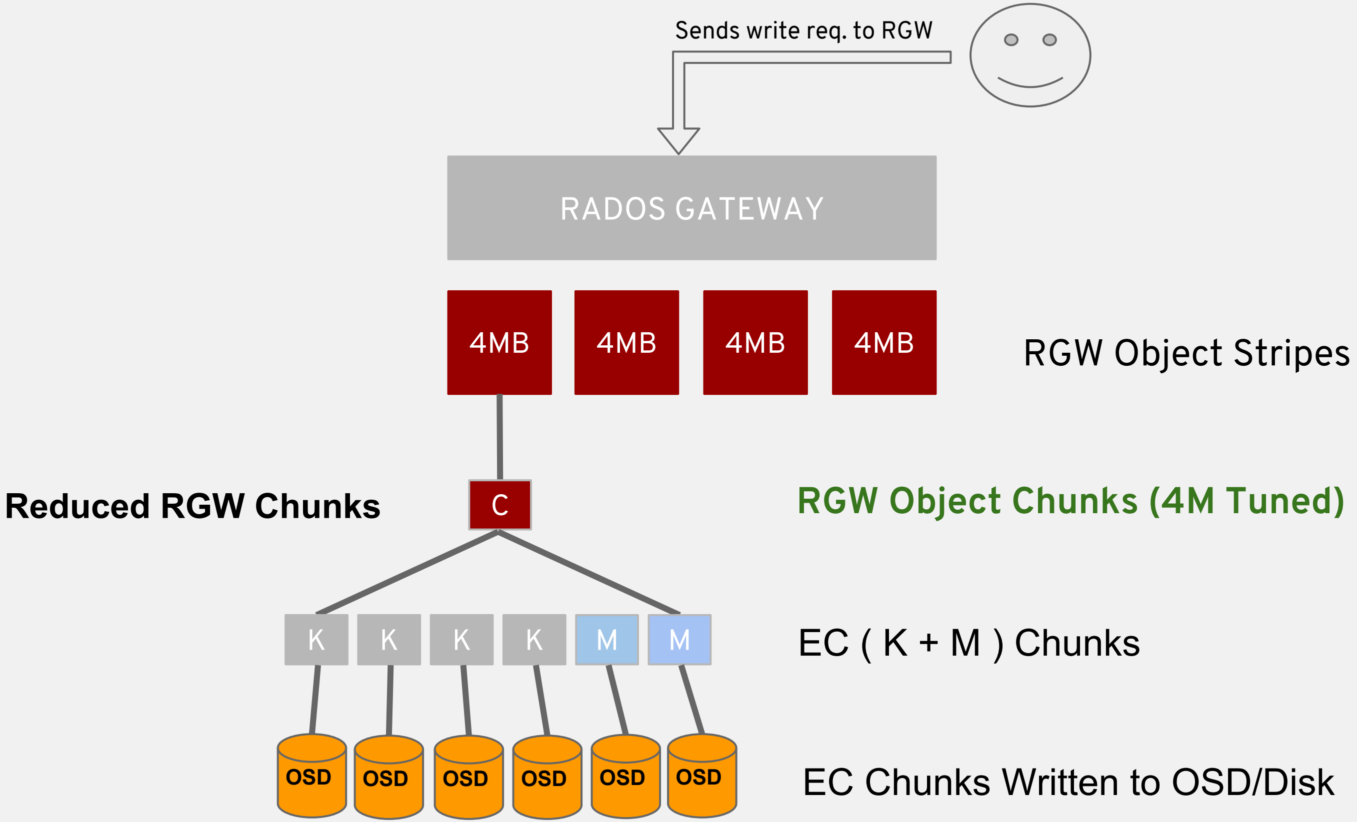


RGW Write Operation (Where to Tune)



redhat

RGW Write Operation (After Tuning)



Improved Large Object Write (MBps)

DEFAULT vs. TUNED

Tuning *rgw_max_chunk_size* to 4M , helped improve write throughput

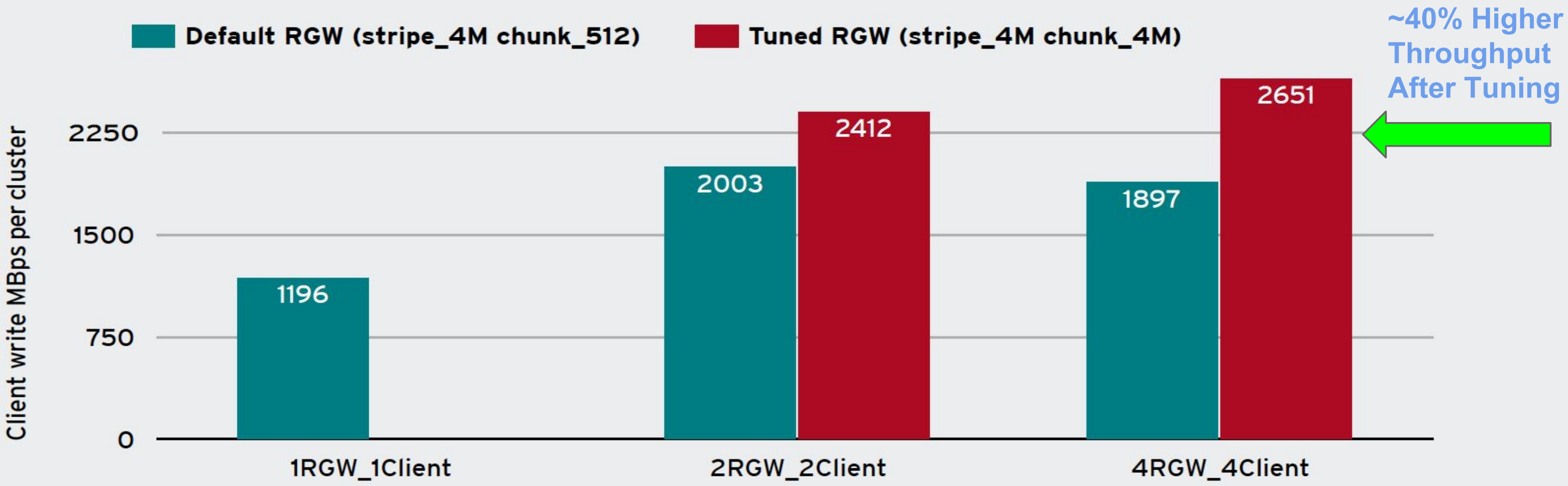


Figure 15.Tuning RGW object stripe size resulted in a roughly 40% improvement in write performance (high-density nodes, 32MB objects write workload, higher is better).

Large Object Write (MBps)

- Client write throughput scaled **near-linearly** while increasing RGW hosts
- Best observed (untuned) write throughput was 2.1 GB/s with 32M object size on high density servers

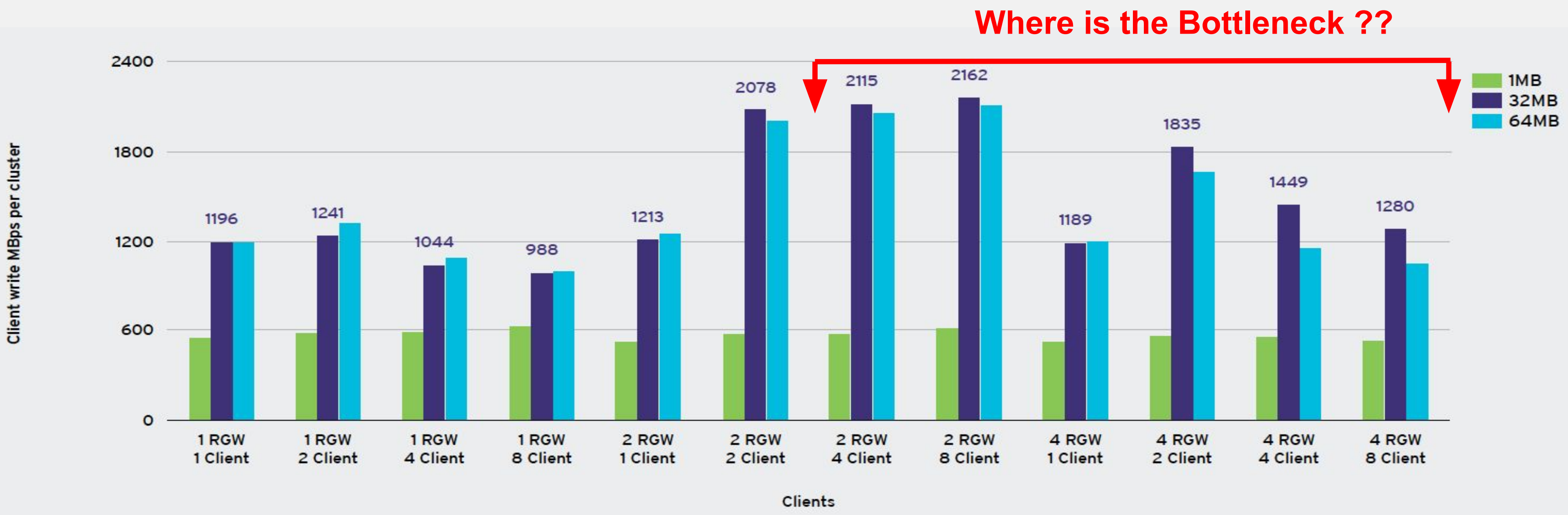


Figure 9. Large-object write test, cluster-wide aggregate performance on high-density servers.

Large Object Write

Bottleneck : DISK SATURATION

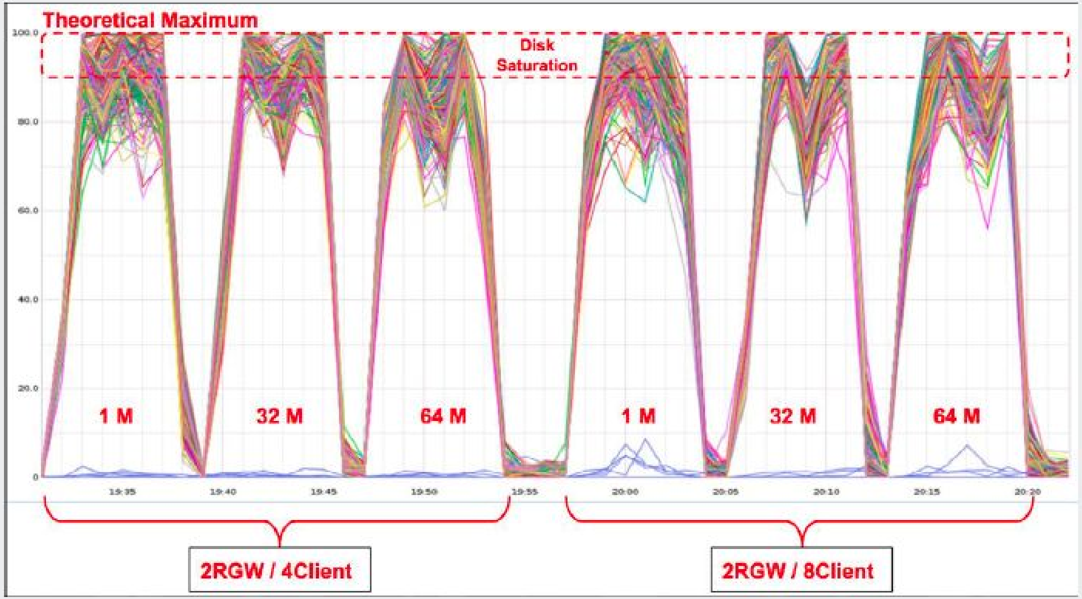


Figure 10. Ceph OSD disk utilization clearly showed disk saturation with only two RGWs and a variable numbers of clients.

Large Object Write (MBps)

- Performance **limited by disk saturation** on Ceph OSD hosts

Higher write throughput could have been achieved by adding more OSD hosts

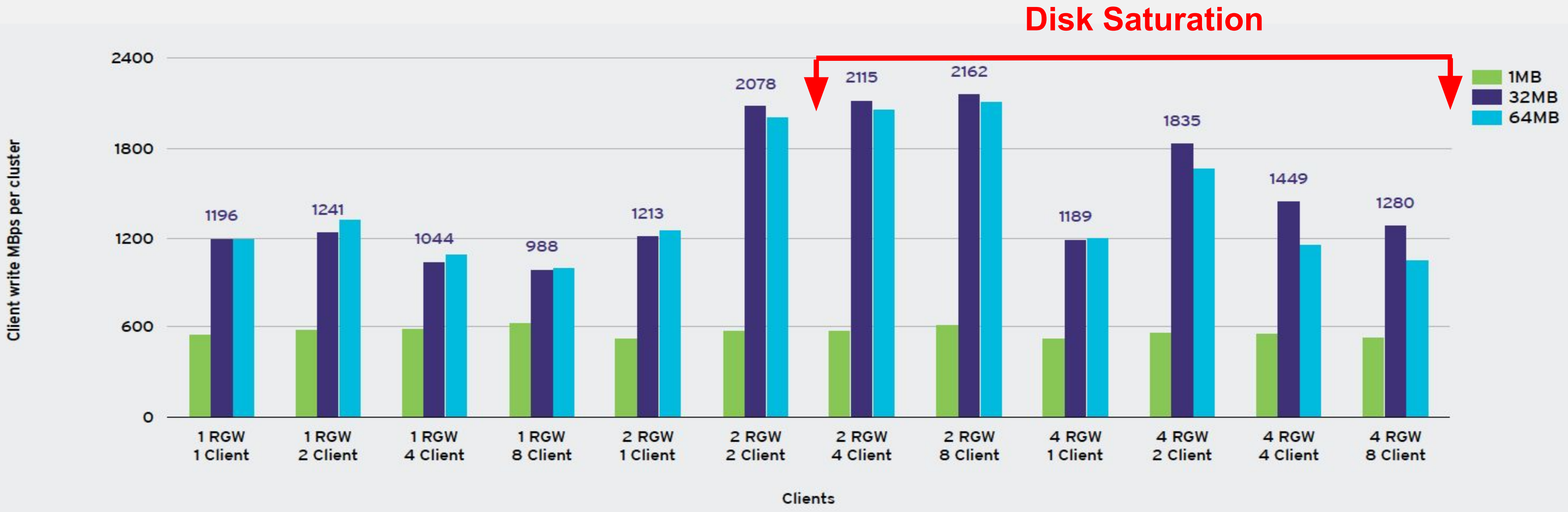


Figure 9. Large-object write test, cluster-wide aggregate performance on high-density servers.

Designing for
Higher Object Count (100M+ Objects)

Designing for Higher Object Count

100M+ OBJECTS

- **Tested Configurations**
  - Default OSD Filestore settings
  - Tuned OSD Filestore settings
  - Default OSD Filestore settings + Intel CAS (Metadata Caching)
  - Tuned OSD Filestore settings + Intel CAS (Metadata Caching)

- **Test Details**
  - High Density Servers
  - 64K object size
  - 50 Hours each test run (200 Hrs total testing)
  - Cluster filled up to ~130 Million Objects

Designing for Higher Object Count (Read)

Best Config: Default OSD Filestore settings + Intel CAS (Metadata caching)

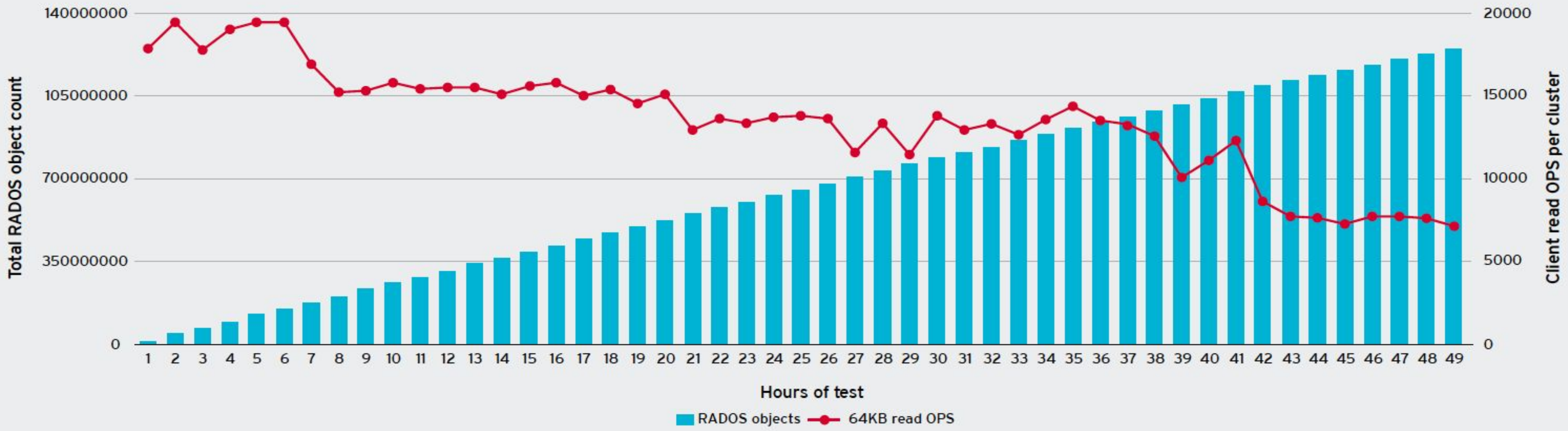


Figure 47. Default Ceph OSD filestore + Intel CAS (split:merge 2:10) 64KB object read OPS versus RADOS object count on high-density servers.

Designing for Higher Object Count (Read)

COMPARING DIFFERENT CONFIGURATIONS

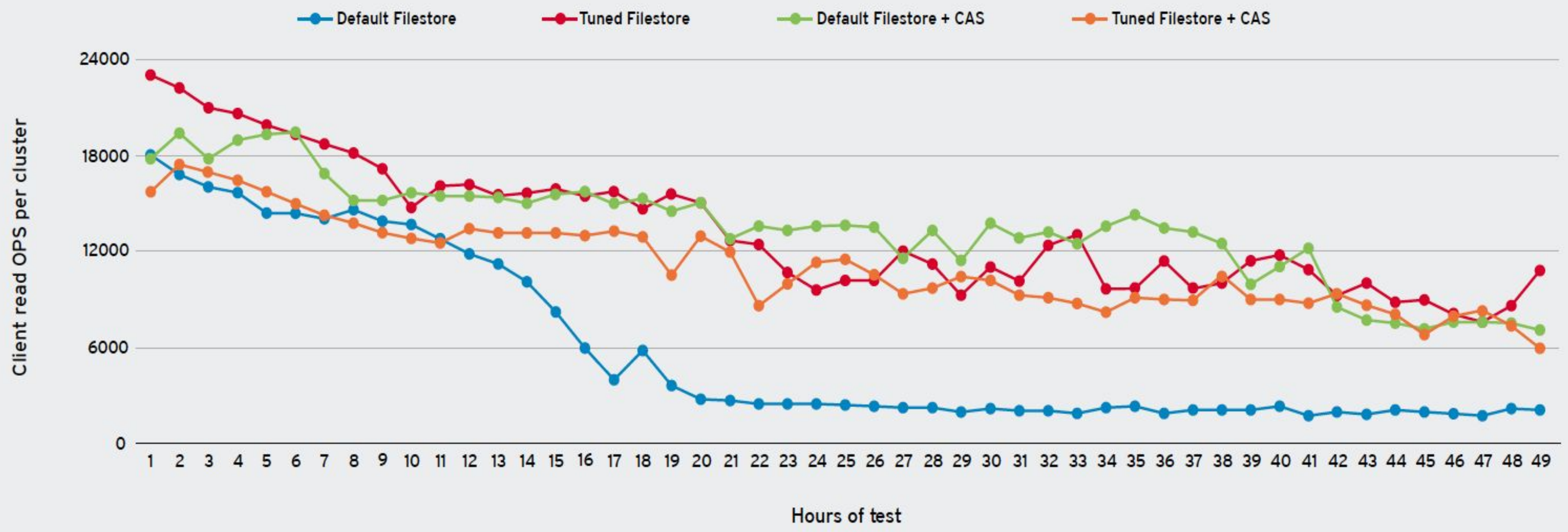


Figure 28. Read OPS per cluster for various filesystem configurations on high-density server-based cluster.

Designing for Higher Object Count (Write)

Best Config: Default OSD Filestore settings + Intel CAS (Metadata caching)

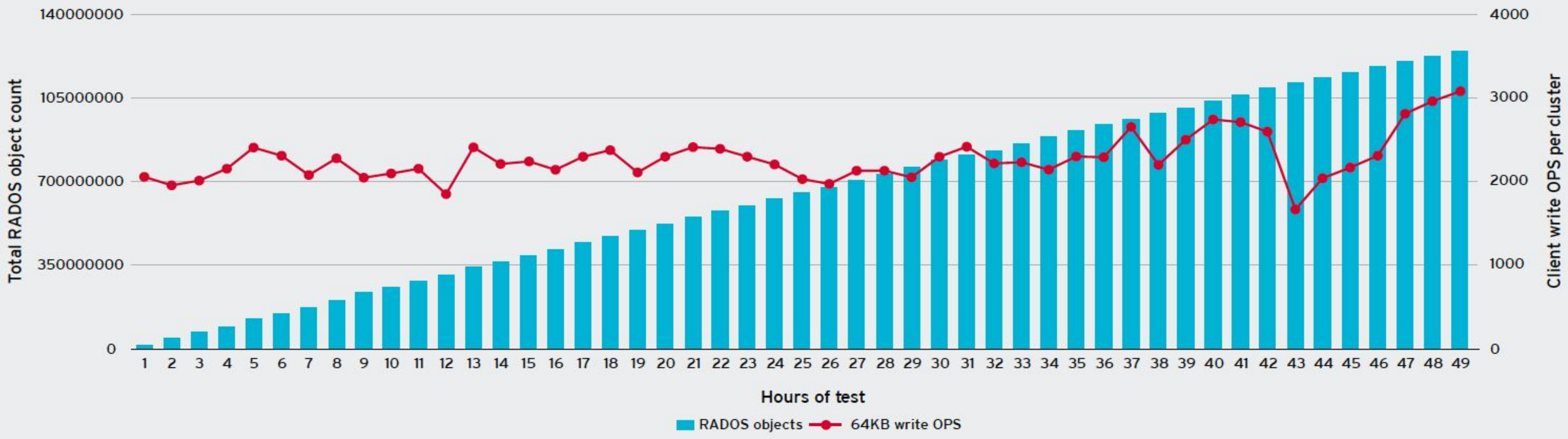


Figure 46. Default Ceph OSD filestore + Intel CAS (split:merge 2:10) 64KB object write OPS versus RADOS object count on high-density servers.

Designing for Higher Object Count (Write)

COMPARING DIFFERENT CONFIGURATIONS

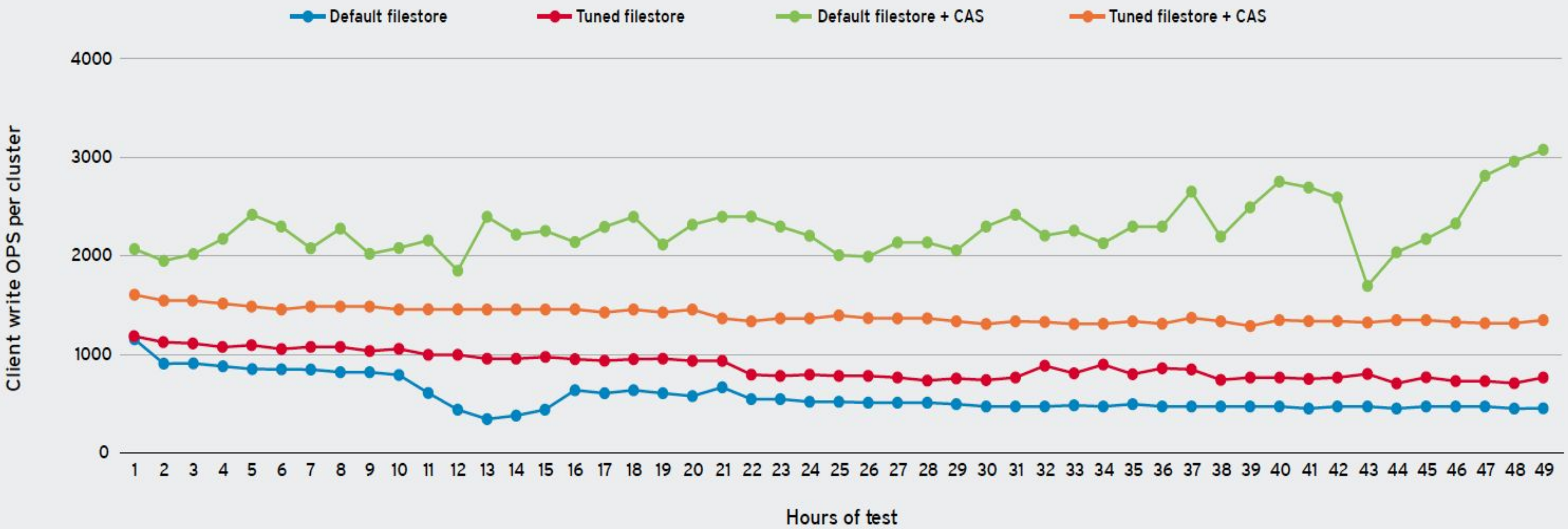


Figure 31. Write OPS per cluster for various filesystem configurations on high-density servers, 64KB objects.

Higher Object Count Test (Latency)

- Steady write latency with Intel CAS, as the cluster grew from 0 to 100M+ objects
- ~100% lower as compared to Default Ceph configuration

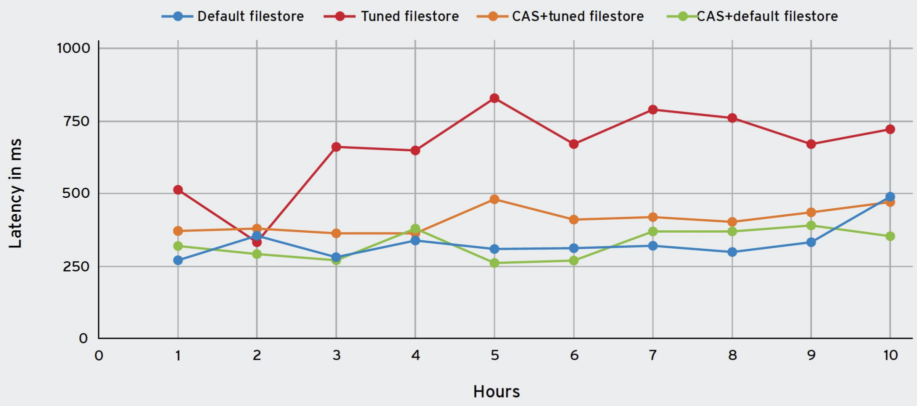


Figure 29. 99th percentile read latency for the first 10 hours of testing, high-density cluster, 64KB objects.

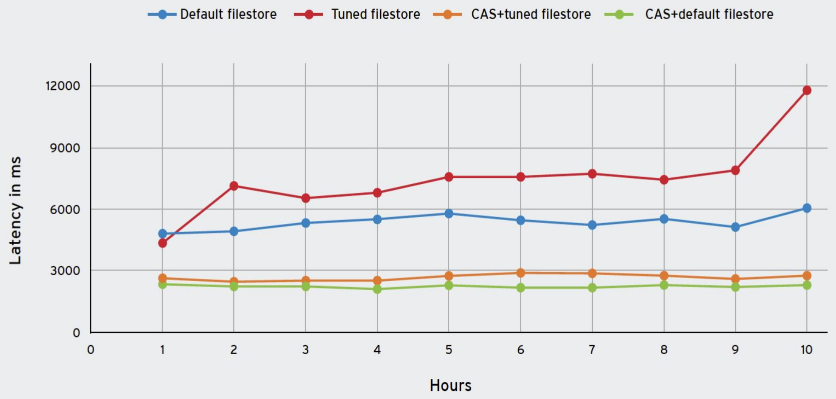


Figure 32. 99th percentile OPS latency for the first 10 hours of testing, high-density servers, 64K object size.

Secrets to
Object Storage Performance

(Key takeaways)

#1 Architectural Consideration for object storage cluster

- Object size
- Object count
- Server density
- Client : RGW : OSD Ratio
- Bucket index placement scheme
- Caching
- Data protection scheme
- Price/Performance

#2 Recommendations for Small Object Workloads

- 12 Bay OSD Hosts
- 10GbE RGW Hosts
- Bucket Indices on Flash Media

#3 Recommendations for Large Object Workloads

- 12 Bay OSD Hosts, 10GbE - Performance
- 35 Bay OSD Hosts, 40GbE - Price / Performance
- 10GbE RGW Hosts
- Tune *rgw_max_chunk_size* to 4M
- Bucket Indices on Flash Media

Caution : Do not increase rgw_max_chunk_size beyond 4M , this causes OSD slow requests and OSD flapping issues.

#4 Recommendations for High Object Count Workloads

- 35 Bay OSD Hosts, 40GbE (Price/Performance)
- Bucket Indices on Flash Media
- Intel Cache Acceleration Software (CAS)
    - Only Metadata Caching

#5 Recommendations for RGW Sizing

For Small Object (64K) 100% Write

- 1 RGW for every 50 OSDs (HDD)

For Large Object (32M) 100% Write

- 1 RGW for every 100 OSDs (HDD)

#6 RGW : 10GbE or 40GbE / Dedicated or Colo ?

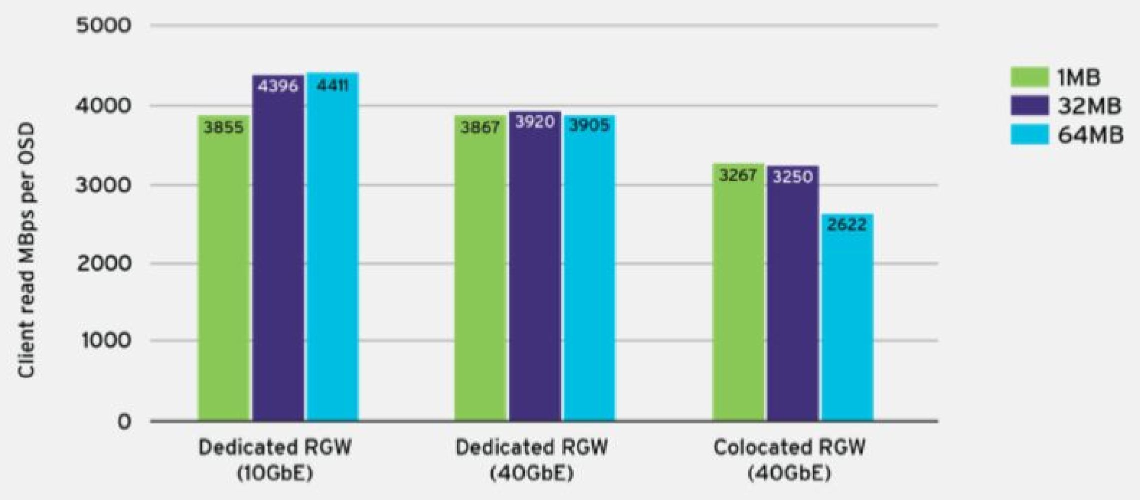


Figure 7. Dedicated RGWs with 10GbE interfaces provided better performance and price-performance (4x RGW, 8x client nodes, high-density servers).

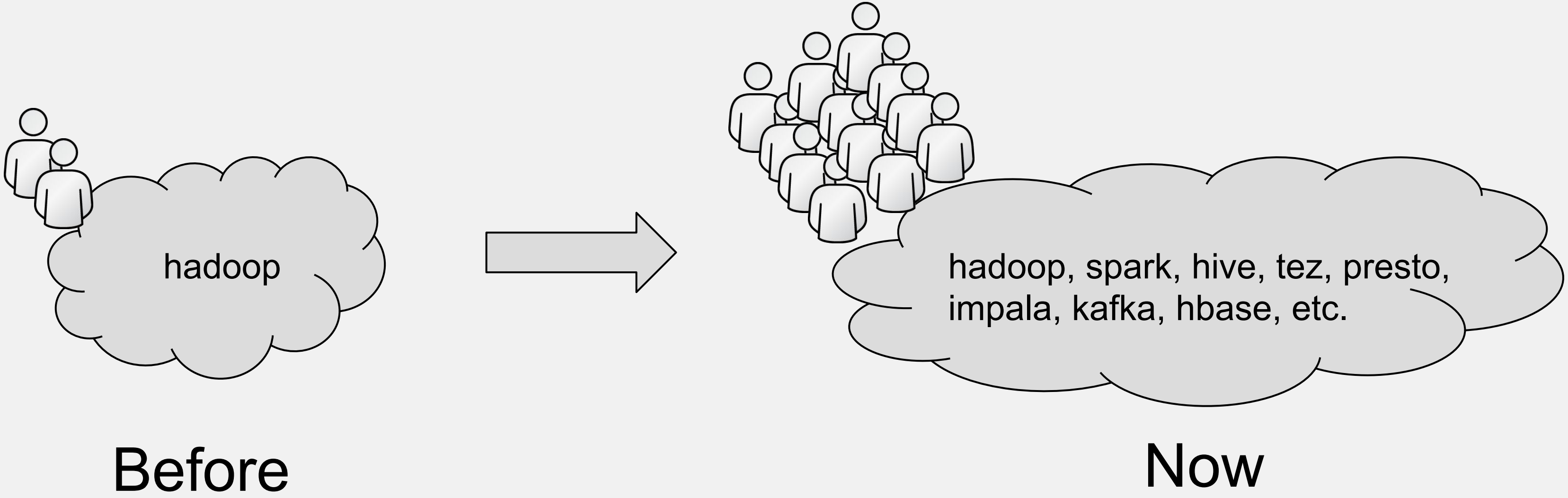
- **Take Away : Dedicated RGW node with 1 x 10GbE connectivity**

Shared Data Lake Solution Using Ceph Object Storage

(Teaser)

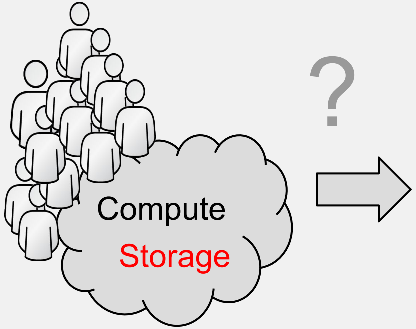
redhat.

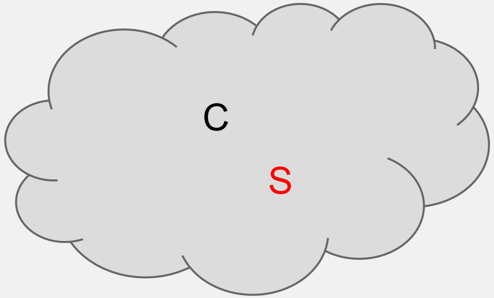
Discontinuity in Big Data Infrastructure - Why?



- **Congestion causing delay in response from analytics applications**
- **Multiple teams competing for the same big data resources**

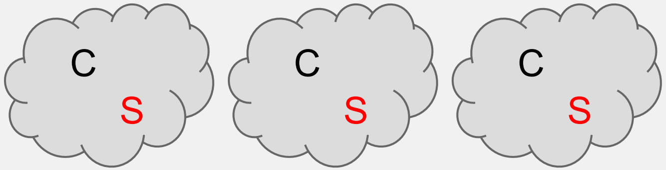
Discontinuity Presents Choice





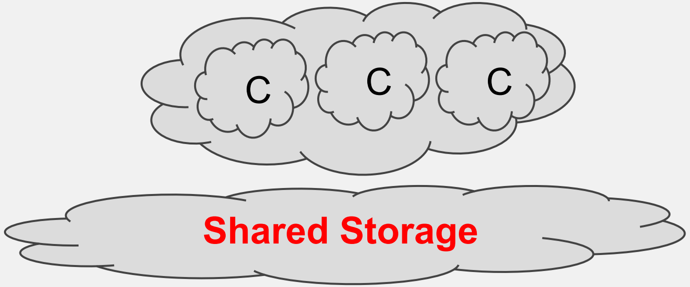
Get a bigger cluster

- Lacks isolation - still have noisy neighbors
- Lacks elasticity - rigid cluster size
- Can't scale compute/storage costs separately



Get more clusters

- Cost of duplicating big datasets
- Lacks on-demand provisioning
- Can't scale compute/storage costs separately

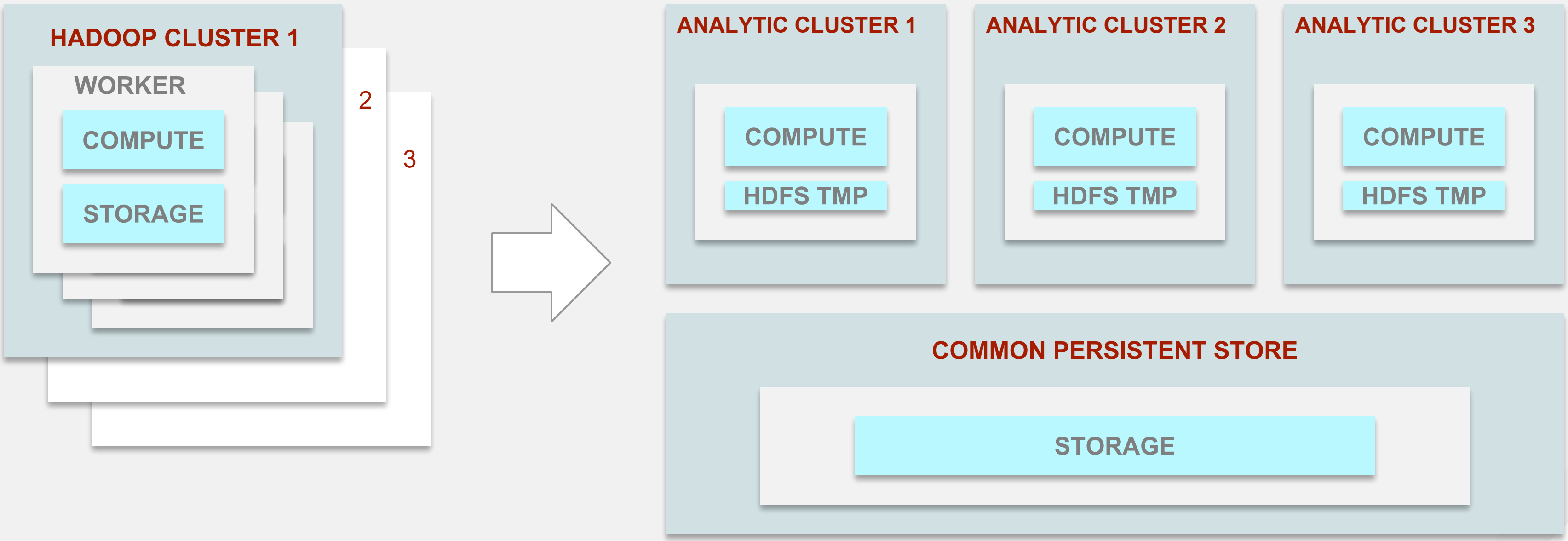


On-demand Compute and Storage pools

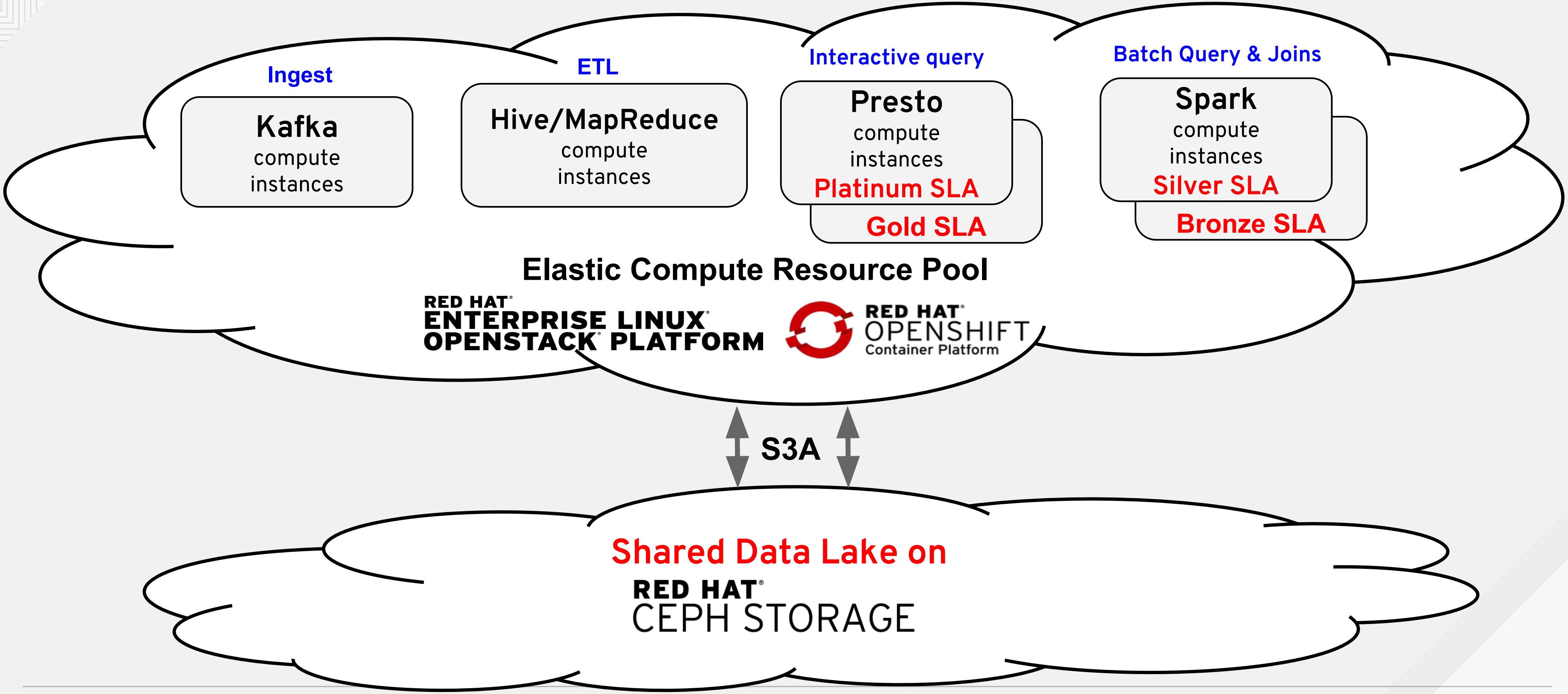
- Isolation of high-priority workloads
- Shared big datasets
- On-demand provisioning
- Compute/storage costs scale separately

Data Analytics Emerging Patterns

Multiple analytic clusters, provisioned on-demand, sourcing from a common object store



Red Hat Elastic Infrastructure for Analytics



Red Hat Elastic Infrastructure for Analytics

Analytics vendors focus on analytics. Red Hat on infrastructure.

Analytics vendors provide analytics software

cloudera
HIVE
HORTONWORKS
POWERING THE FUTURE OF DATA
APACHE Spark
MAPR
presto

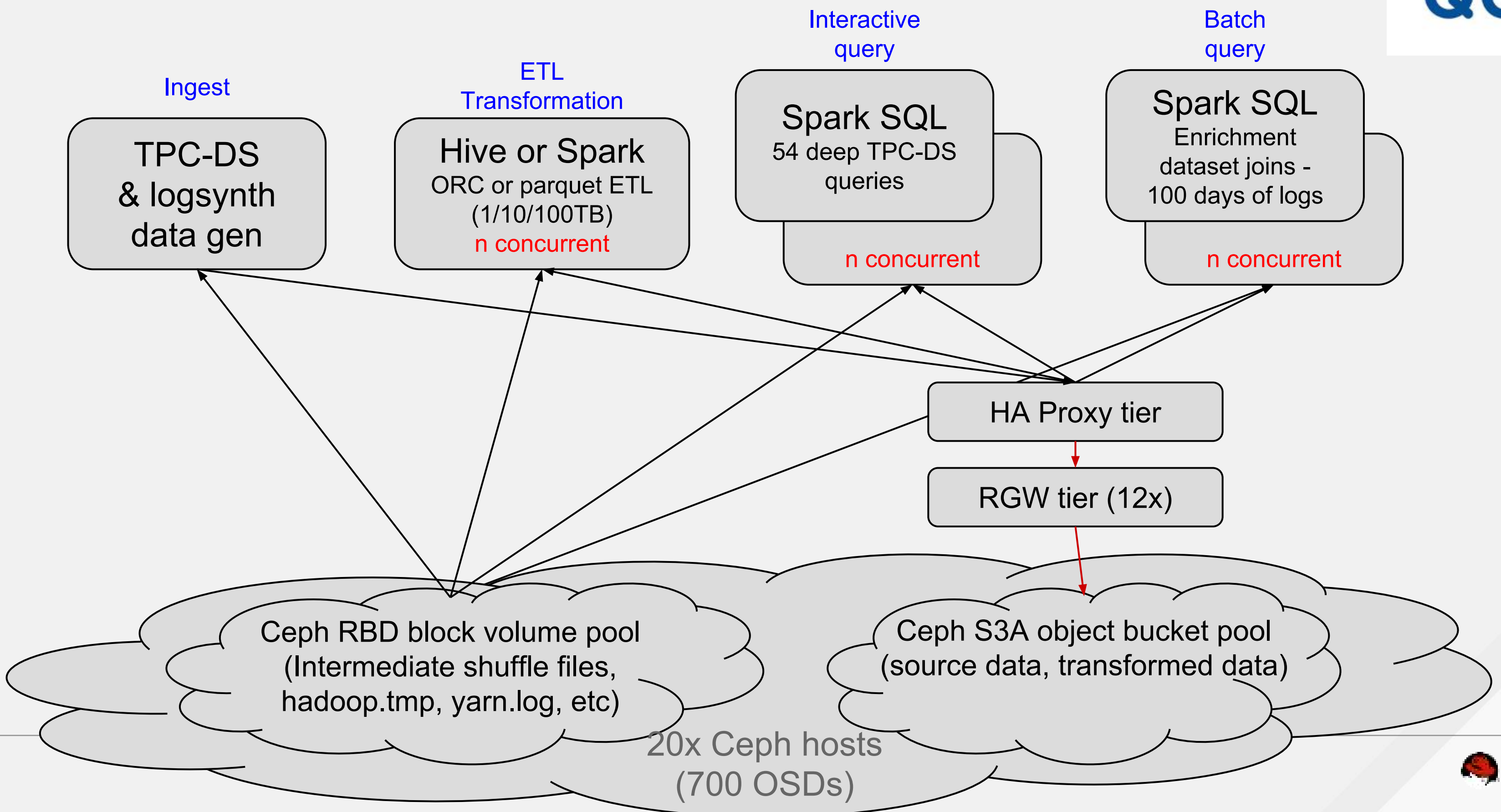
Analytics

Red Hat provides infrastructure software

redhat OpenStack or OpenShift Provisioned Compute Pool

redhat Shared Data Lake on Ceph Object Storage

Red Hat Solution Development Lab (at QCT)



Shared Data Lake on Ceph Object Storage Reference Architecture

Coming

This Fall ...

Get Ceph Object Storage P&S Guide : http://bit.ly/object-ra

redhat

THANK YOU

plus.google.com/+RedHat

facebook.com/redhatinc

linkedin.com/company/red-hat

twitter.com/RedHatNews

youtube.com/user/RedHatVideos