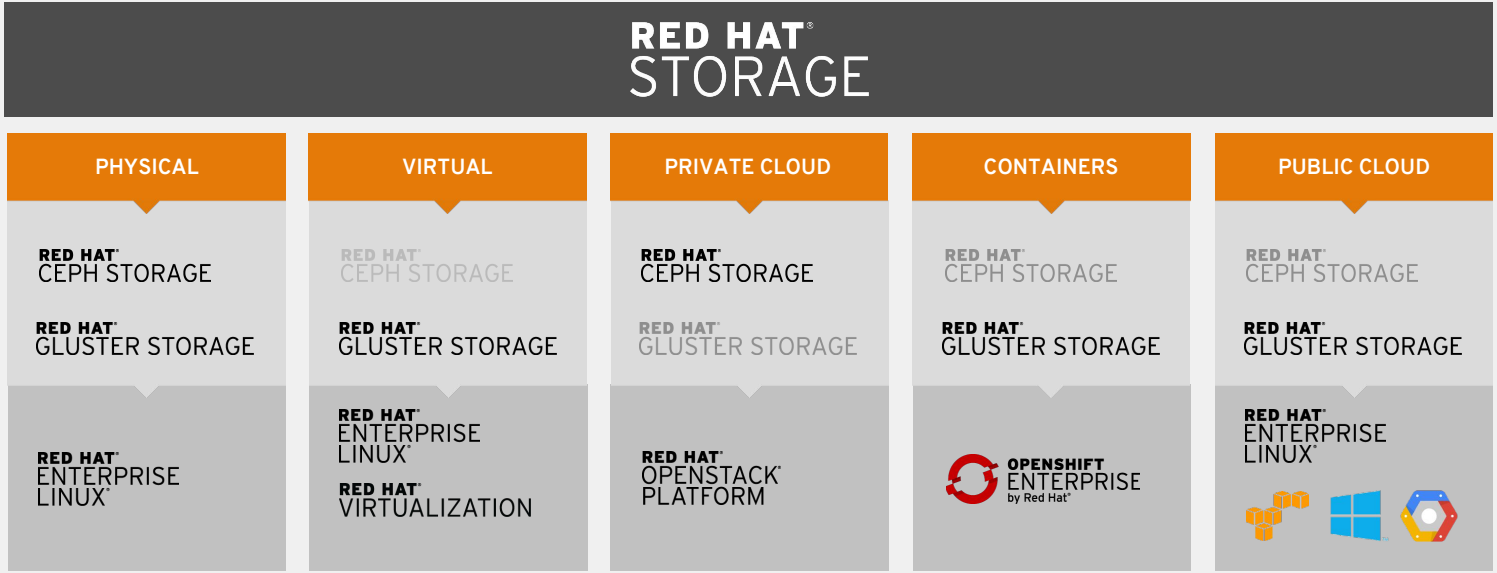
Как я перестал бояться и отказался от дедушкиных СХД

Red Hat Ceph Storage сегодня и завтра

RED HAT STORAGE IS DEEPLY INTEGRATED



RED HAT - A STORAGE VISIONARY

Strengths (Red Hat)

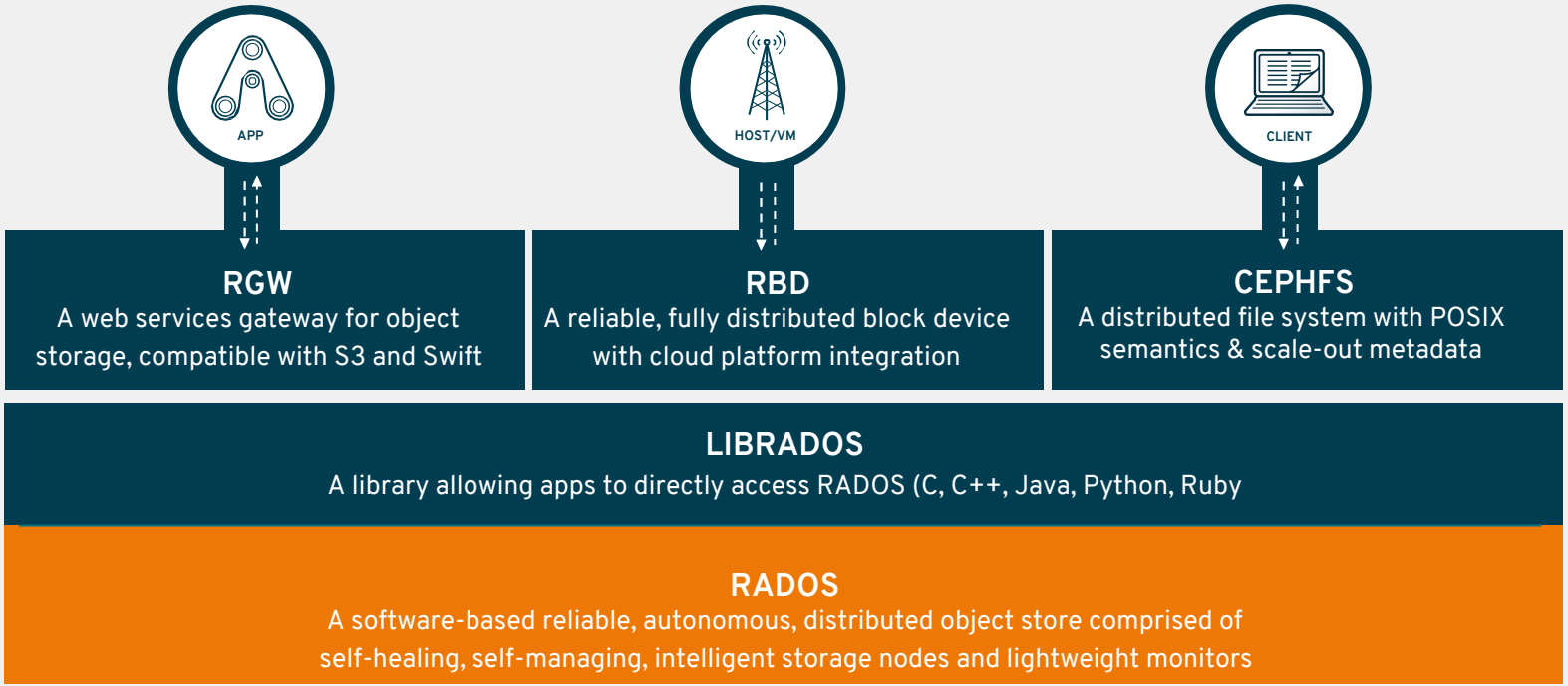
- Red Hat Ceph Storage is a versatile product that is increasingly being deployed as both a block and an object storage system, with new object storage capabilities such as multisite failover and close fidelity to the S3 API.
- Red Hat Gluster Storage is tightly integrated with both Docker and Kubernetes, enabling data persistence and protection for containerized workloads in either hyperconverged or disaggregated form factors.
- Both Red Hat Ceph and Gluster Storage are certified across a broad spectrum of server hardware with reference architectures available from leading server OEMs such as HPE, Cisco and Supermicro.

Gartner

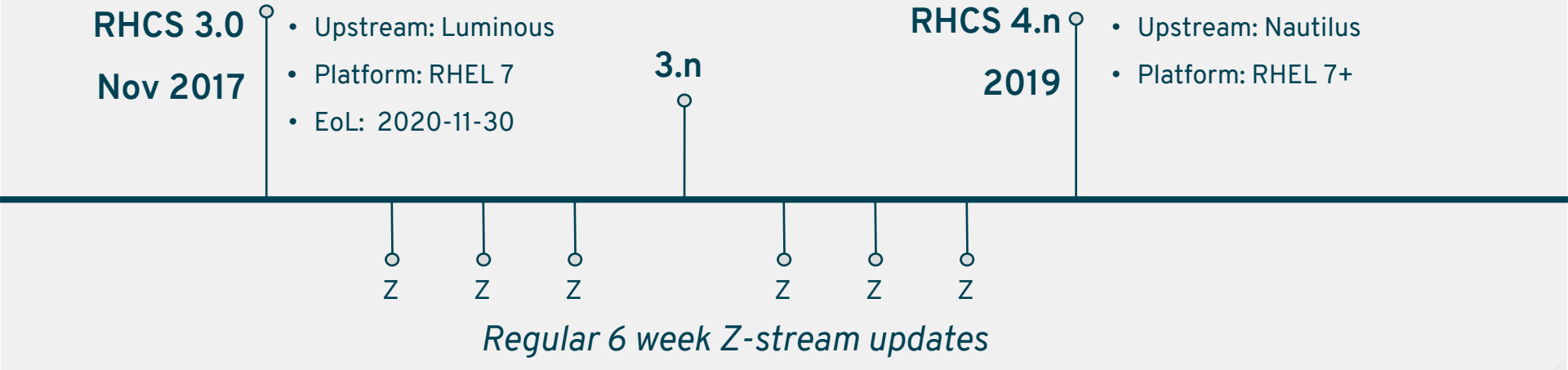
https://engage.redhat.com/gartnermagic-quadrant-storage-s-201710110441

redhat.

CEPH STORAGE OVERVIEW



RED HAT CEPH STORAGE VERSIONS



Regular 6 week Z-stream updates

SSD TRENDS AND IMPACTS

From Sage Weil "Future of Storage" presentation

- SSDs are getting faster
  - NVMe has replaced SATA/SAS
  - New devices incorporate DRAM or 3D XPoint
- Density is improving
  - 3D NAND
- Price continues to drop
  - Manufacturing capacity shortages are being addressed
- Software continues to adapt

Today

- Linux IO stack rewritten over last decade
- Red Hat Storage plays in SSD and HDD worlds

Future

- Not about total IOPS, but IOPS per CPU core
- Ceph project is reimplementing OSD
  - Seastar
    - ➢ run to completion framework for C++
    - ➢ confine request processing to CPU cores
  - Drivers in userspace
    - ➢ DPDK networking, SPDK NVMe storage
  - Designing a new NVMe-specific backend

RED HAT CEPH ROADMAP THEMES

THEME: USABILITY

Increase TB/admin and make common tasks simpler to perform

RHCS Dashboard for monitoring

Cleaner CLI

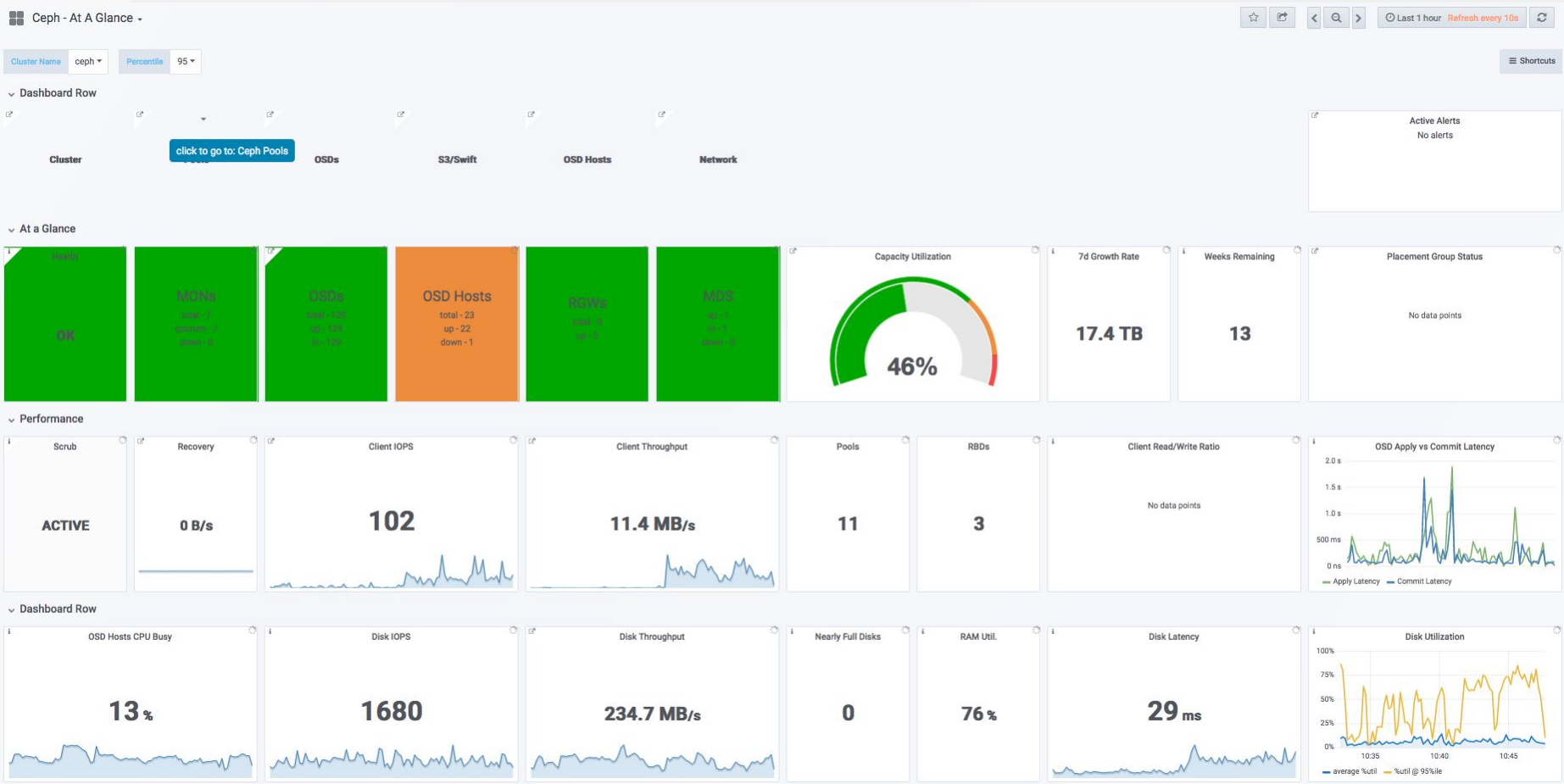
Improved Logging

RHCS Dashboard for management

Prometheus support (3.1)

Access Insights (4.0)

Automatic placement group management


Ceph - At A Glance
Last 1 hour
Refresh every 10s
Cluster Name
ceph
Percentile
95
Shortcuts
Dashboard Row
Cluster
click to go to: Ceph Pools
OSDs
S3/Swift
OSD Hosts
Network
Active Alerts
No alerts
At a Glance
Health
OK
MONs
OSDs
OSD Hosts
total - 23
up - 22
down - 1
RGWs
MDS
Capacity Utilization
46%
7d Growth Rate
17.4 TB
Weeks Remaining
13
Placement Group Status
No data points
Performance
Scrub
ACTIVE
Recovery
0 B/s
Client IOPS
102
Client Throughput
11.4 MB/s
Pools
11
RBDs
3
Client Read/Write Ratio
No data points
OSD Apply vs Commit Latency
2.0 s
1.5 s
1.0 s
500 ms
0 ns
Apply Latency
Commit Latency
Dashboard Row
OSD Hosts CPU Busy
13 %
Disk IOPS
1680
Disk Throughput
234.7 MB/s
Nearly Full Disks
0
RAM Util.
76 %
Disk Latency
29 ms
Disk Utilization
100%
75%
50%
25%
0%
10:35
10:40
10:45
average %util
%util @ 95%ile

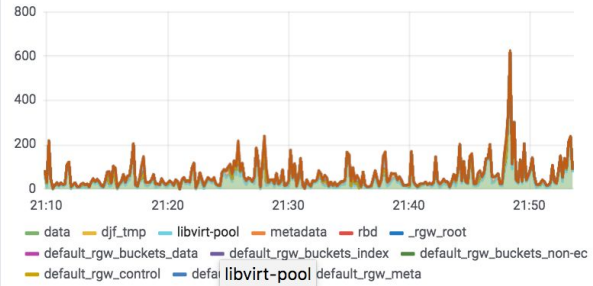
Ceph Pools

May 7, 2018 21:10:42 to May 7, 2018 21:52:35

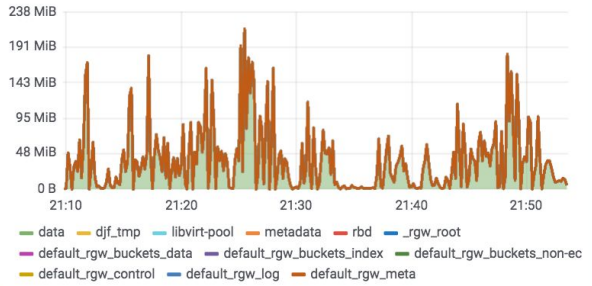
Cluster Name: ceph | Pool Name: All

Pool Overview : All

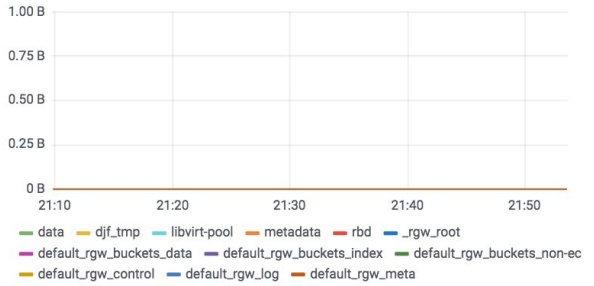
Client IOPS by Pool



Throughput by Pool



Recovery Workload by Pool



Top 5

Top 5 Pools by Client IOPS

Pool Name	IOPS
data	77
metadata	11
libvirt-pool	0
djf_tmp	0
default_rgw_log	0

Top 5 Pools by Throughput

Pool Name	Throughput
data	5.33 MiB
metadata	180.68 KiB
libvirt-pool	336.00 B
default_rgw_log	252.00 B
djf_tmp	0 B

Top 5 Pools by Capacity Used

Pool Name	Capacity Used
data	61.87%
default_rgw_buckets_data	2.07%
libvirt-pool	0.01%
metadata	0%
djf_tmp	0%

THEME: PERFORMANCE AND SCALE

Reduce $/IOPS and $/Gb

BlueStore (Tech Preview)

BlueStore (3.n)

Erasure Code for RBD (3.n)

OSD compression & dedup (3.n)

Consistent IO in recovery (4.n)

QoS

BLUESTORE

Micron test results

4k Random Block

Write:

- 18% higher IOPs
- 5% lower average latency
- Up to 70%+ reduced 99.99% latency

70% read / 30% write:

- 14% higher IOPs
- 80%+ lower read tail latency
- 70%+ lower write tail latency

4M Object

Write:

- 88% increase in throughput
- 47% decrease in average latency

70% read / 30% write:

- 64% increase in throughput
- 40% decrease in average latency

THEME: SECURITY

Meet compliance requirements

Per-Object encryption

Security Guidebook (3.n)

Kerberos integration with RADOS

On-The-Wire Encryption

THEME: OPENSTACK

Complete and tightly integrated storage for OpenStack

3.0

Manila support for CephFS (OSP 13)

Cinder encryption with RBD (OSP 13)

Hyperconverged deployment with Director (OSP 13)

Next

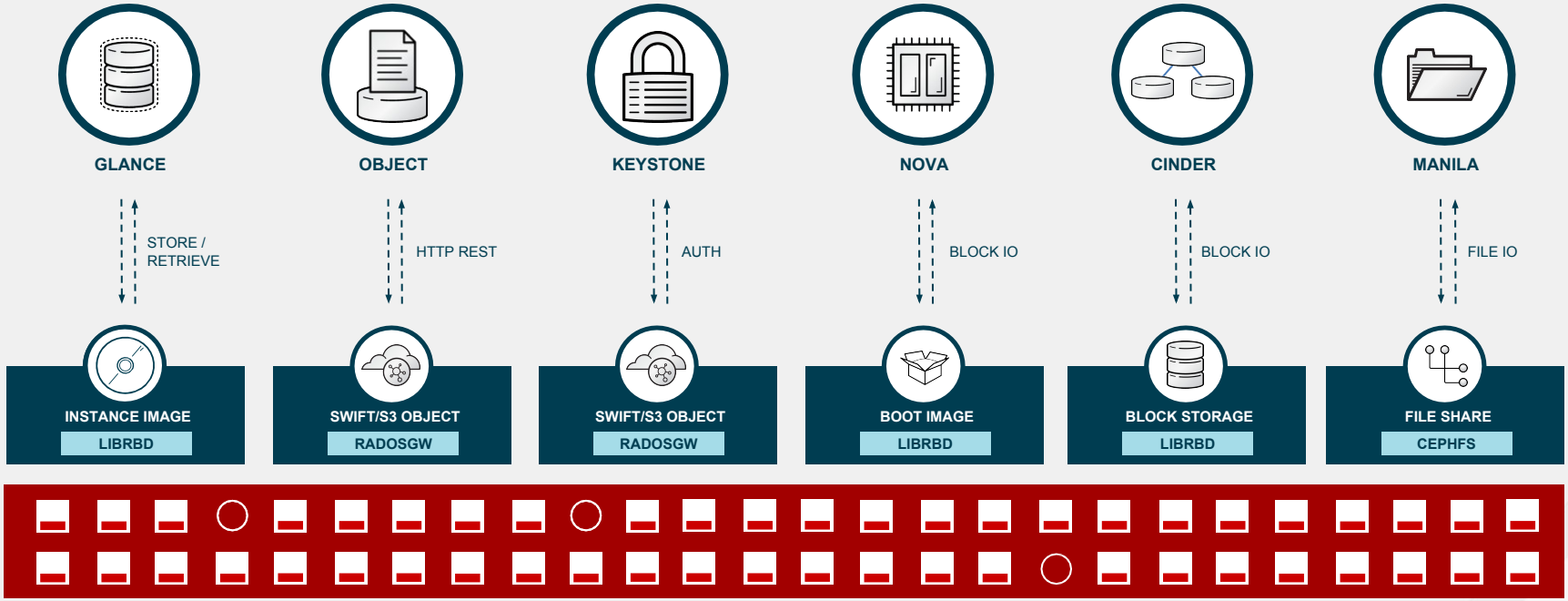
OCP Persistent volumes via Cinder/Manila (OSP 14)

Distributed compute with Director (OSP 14)

Future

Cinder QoS with RBD

THEME: OPENSTACK



CEPH - MAIN OPENSTACK STORAGE

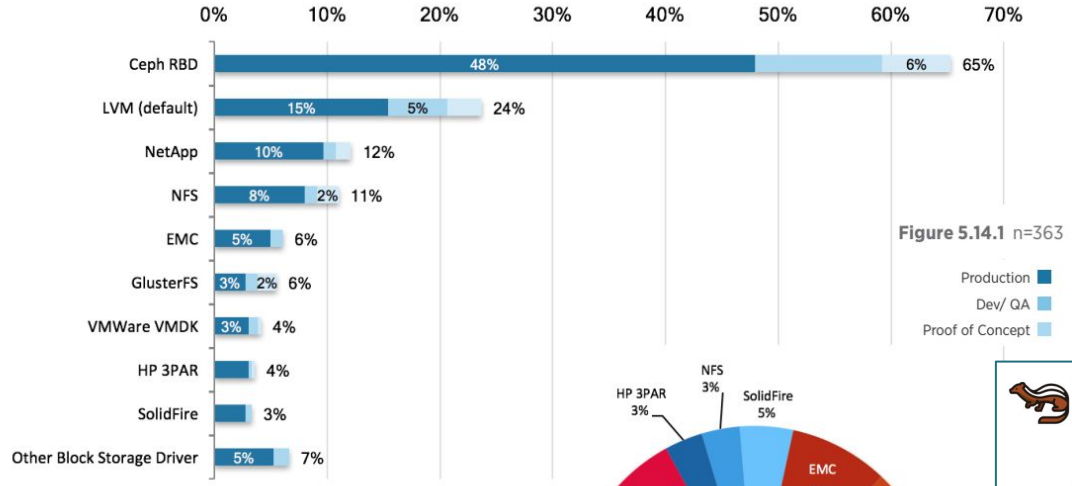


Figure 5.14.1 n=363

Among the largest clouds with 1,000 or more cores, Ceph RDB is still dominant, but not used by the majority, while other block storage drivers were also popular.

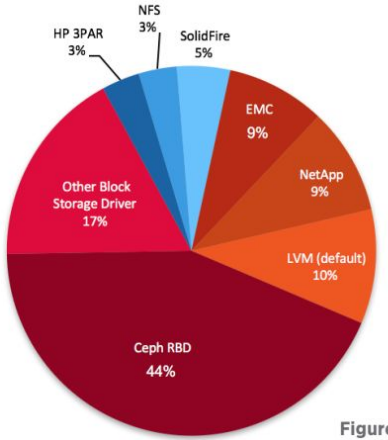


Figure 5.14.2 n=88

April 2017

OPENSTACK USER SURVEY

A snapshot of OpenStack users' perspectives and deployments

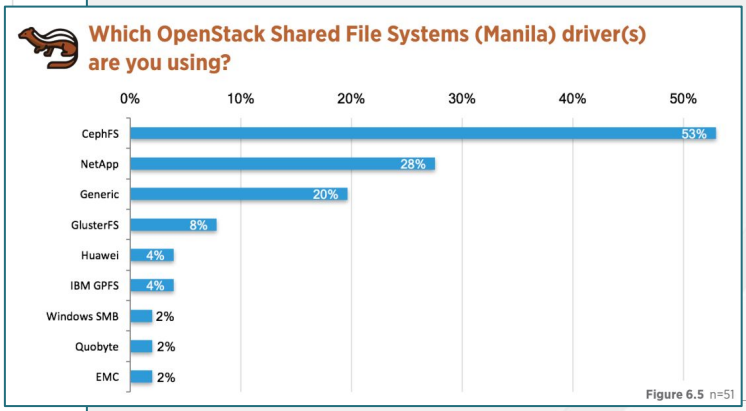


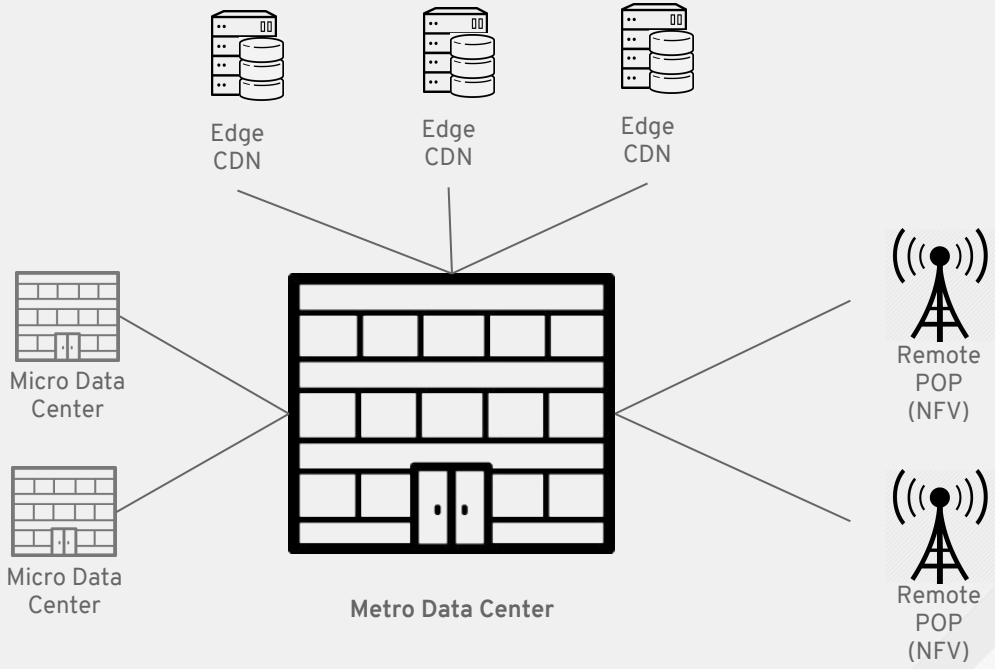
Figure 6.5 n=51

redhat.

USE-CASE: HYPERCONVERGED

With OpenStack: Red Hat Hyper Converged Infrastructure for Cloud (RHHI4C)

- **Edge sites hosting content for low-latency delivery**
- **Remote POPs running virtual network functions**
- **Micro data centers capturing IOT telemetry for real-time processing**



THEME: CONTAINERS

Simplify lifecycle operations and provide more flexible topologies

Containerized Service Daemons

Persistent Volumes via Cinder/Manila (OSP 14)

Kubernetes support (4.n)

Dynamic load balancing

Upstream work:

- **Rook (“Operator”)**
- **ceph-csi**

THEME: OBJECT STORAGE

Scale!

Dynamic Sharding of Bucket Indices

Per-object compression

Per-object encryption (SSE-C)

Ongoing S3 compatibility enhancements (3.n)

New RGW web server (4.n)

Cloud Sync

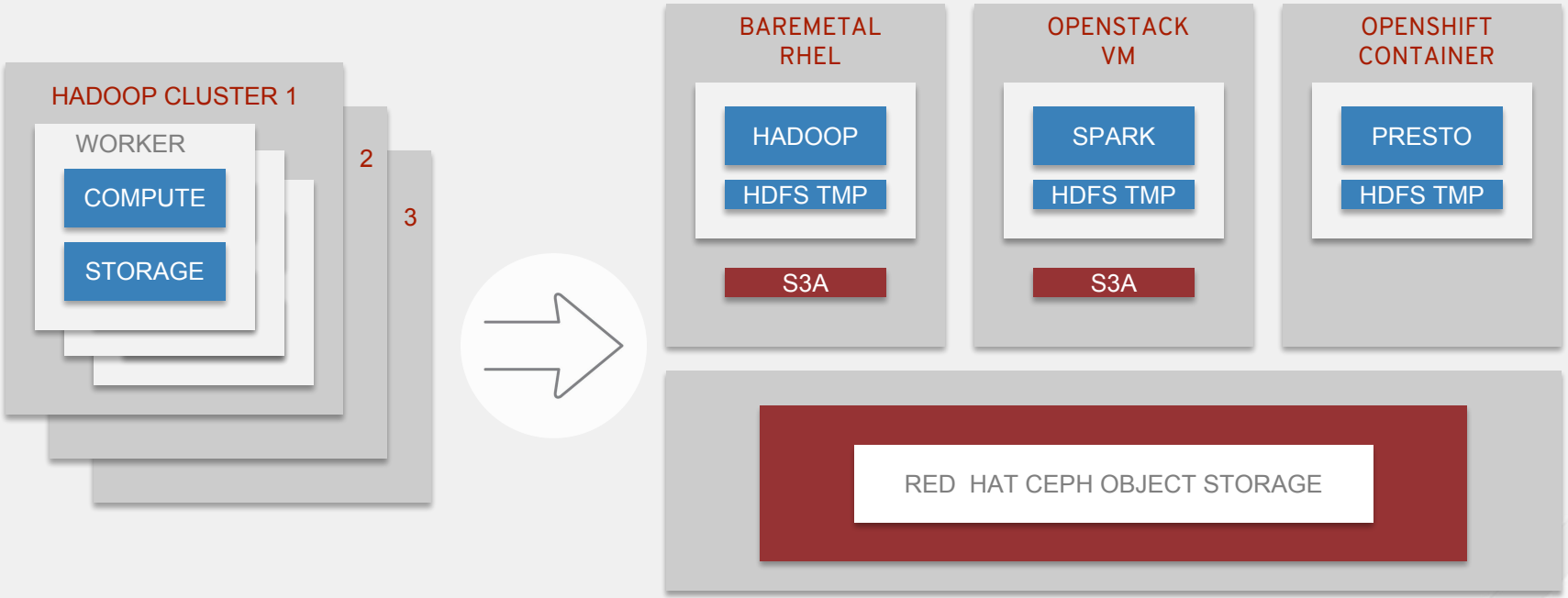
AWS Secure Token Service

THEME: OBJECT STORAGE

Modern backup infrastructure

THEME: OBJECT STORAGE

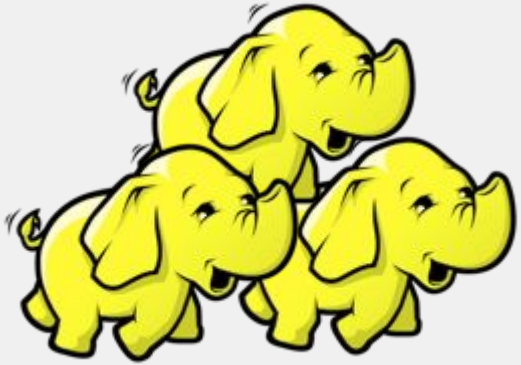
Data lakes with Ceph



https://redhatstorage.redhat.com/category/spark-2/

DATA LAKES: PROLIFERATION OF TOOLS

SPECIALIZED FOR DIFFERENT JOBS



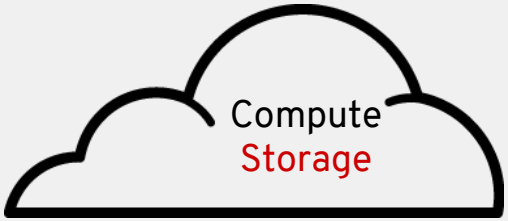
presto

APACHE Spark™

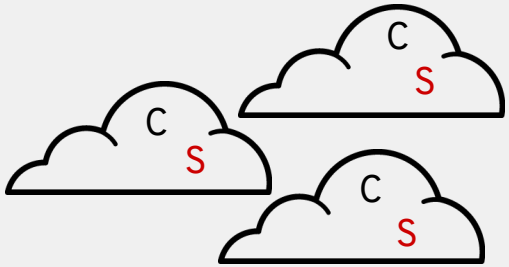
kafka

OPTIONS

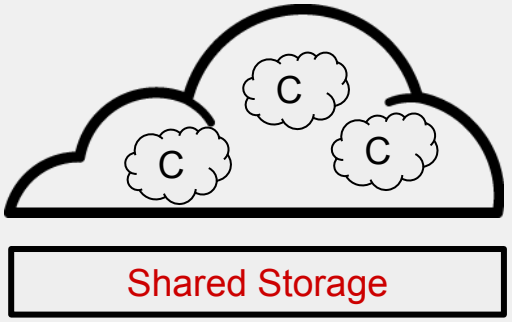
WHAT TO DO?



Larger mixed workload cluster



Workload specific clusters, PBs of duplicate data sets



Workload specific clusters, self service provisioning, shared data sets

HADOOP COMMON - S3A

Data lakes with Ceph

- Filesystem client that maps HDFS APIs to S3 API
- Allows interacting with S3 objects just like HDFS files**
- Included by default in Apache and vendor data platform products

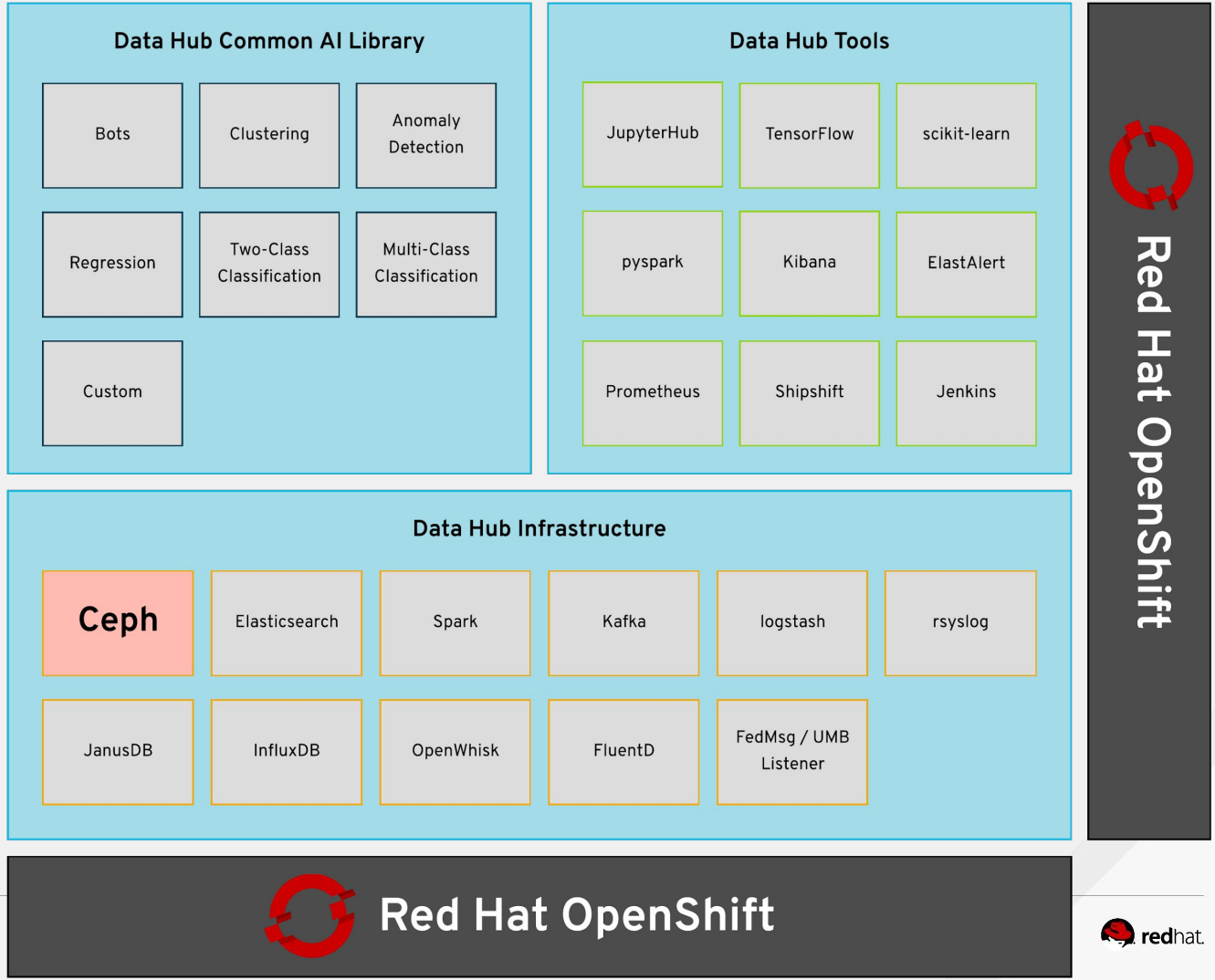
- Simple configuration
  - `fs.s3a.access.key`
  - `fs.s3a.secret.key`
  - `fs.s3a.endpoint`

- Can use in conjunction with Hive external tables
  - `create database mydb location 's3a://bucket/mydb';`

DATA HUB

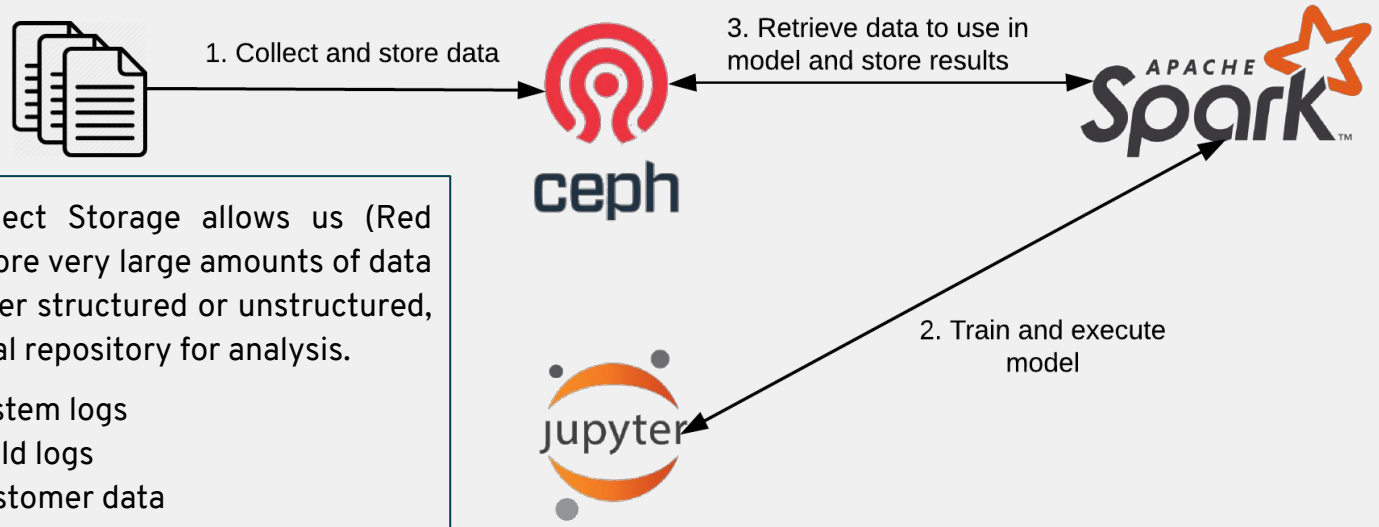
Designed for Data Science and Analysis

A collection of open source and cloud components packaged in a “machine learning-as-a-service” platform to solve internal business problems at Red Hat.



DATA HUB

Ceph Data Lake + Spark + Jupyter Notebook



Ceph Object Storage allows us (Red Hat) to store very large amounts of data as-is, either structured or unstructured, in a central repository for analysis.

- System logs
- Build logs
- Customer data
- System metrics
- Product operational data

THEME: PROTOCOLS

Broaden workload options

iSCSI -> RHV

CephFS -> OSP Manila

NFS v3/v4 Gateway for RGW

NFS for CephFS (4.n)

Ongoing enhancements


redhat.


You
Tube