What's the next storage solution for OpenShift ?

Matthias Muench
EMEA Specialist Solutions Architect
mmuench@redhat.com

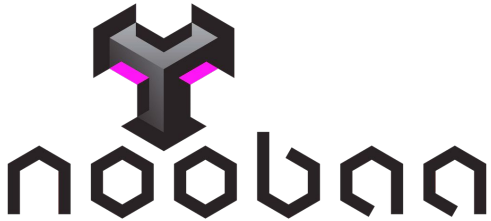
Red Hat

Agenda

- What and Why?
- Architecture
- Use cases
- Sizing
- Demo

Red Hat

What & Why



WHAT IS IT?

Add-On for OpenShift for running stateful apps

Highly scalable, production-grade persistent storage

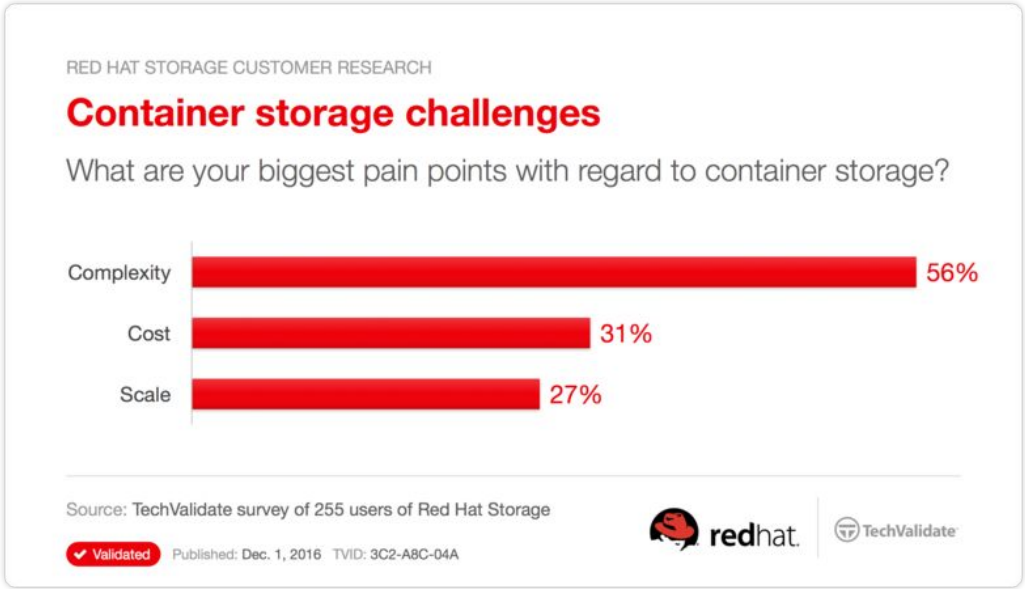
- For **stateful applications** running in Red Hat® OpenShift
- Optimized for Red Hat **OpenShift Infrastructure services**
- Developed, released and deployed in synch with Red Hat OpenShift
- Full stack supported by single vendor Red Hat
- Complete persistent storage fabric across hybrid cloud for OCP

WHY IS STORAGE IMPORTANT FOR CONTAINERS?

Complexity. Cost. Scale.

Top five challenges with container adoption

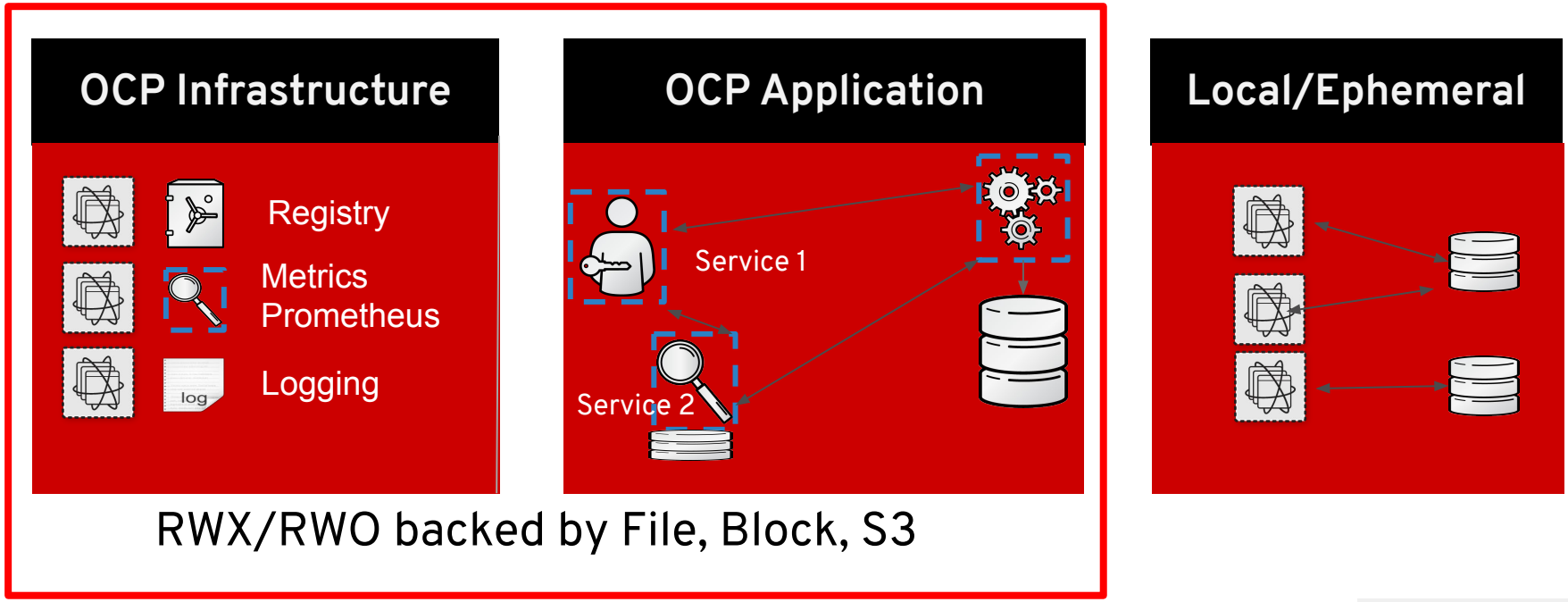
1. **Persistent storage**
2. Data management
3. Multi-cloud or cross-data center
4. Networking
5. Scalability



Red Hat

WHY DO YOU NEED PERSISTENT STORAGE?

For infrastructure and stateful applications



Possible Persistent Storage Providers (OpenShift 4.2)

Volume Plug-in	ReadWriteOnce	ReadOnlyMany	ReadWriteMany
AWS EBS	✅	-	-
Azure File	✅	✅	✅
Azure Disk	✅	-	-
Cinder	✅	-	-
Fibre Channel	✅	✅	-
GCE Persistent Disk	✅	-	-
HostPath	✅	-	-
iSCSI	✅	✅	-
Local volume	✅	-	-
NFS	✅	✅	✅
VMware vSphere	✅	-	-

Additional in-tree:

- FlexVolume
- Flocker
- Ceph RBD (Ceph Block Device)
- CephFS (tech preview)
- GlusterFS

Additional :

- through CSI (depends on driver)

Red Hat

Storage Provisioning in OpenShift

Static Provisioning:

- Storage admin creates storage volumes upfront
- OpenShift selects a predefined volume based upon claim, nearest available size
- No automated housekeeping - causing administrative burden
- Error prone due to increasing complexity and resulting administrative overhead

Dynamic Provisioning:

- OpenShift user requests for storage by persistent volume claim (PVC)
- Delivers the exact requested size and type of storage volume
- No administrative overhead and storage admin involvement upfront
- Automated housekeeping, better efficiency

Security of Provisioning:

- SElinux changes, custom security contexts

OpenShift Container Storage

What changes/d ?

OpenShift
- Transitions from OCP3 to OCP4 => migration needed

Deployments in OCP
- Everything operator based

Dynamic Provisioning :
- CSI provides unified interface

New storage needs :
- S3 is widely used

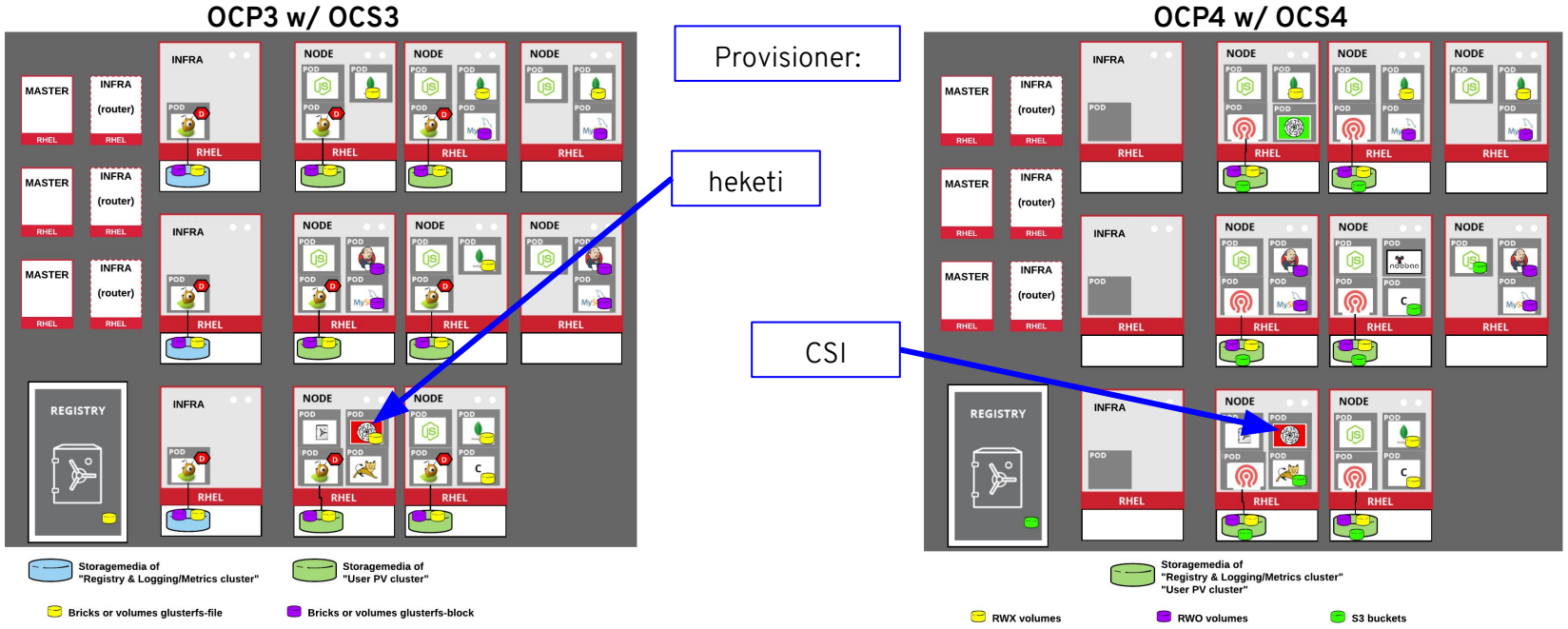
OpenShift Container Storage:
- OCS3 will not work in OCP4
- Real proven S3 stack needed

Red Hat

General structure: OCS3 vs OCS4

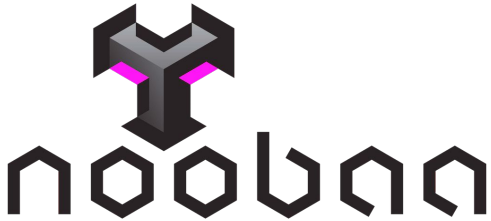
- Technology and its structure
  - Gluster => Ceph
  - Ansible => Rook (deployment)
  - Heketi => CSI/Rook (provisioning)
- RWX and RWO - how does it work ?
  - Similar to current OCS 3 using PVC
- Different: protocol changes
  - iSCSI => rbd
  - GlusterFS => CephFS
  - + S3 (Noobaa vs. standard external)

General structure: OCS3 vs OCS4



Red Hat

Architecture



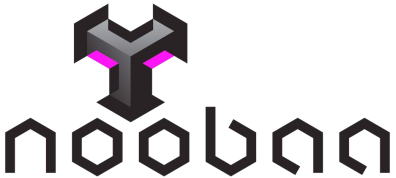
Red Hat

RED HAT OPENSHIFT CONTAINER STORAGE

aka RHOCS or OCS, v4.2 Technology Stack

+

+



- Orchestrator for Ceph storage services in OpenShift
- Responsible to simplify and automate the storage lifecycle
- Fully compliant with the new CSI kubernetes storage standard

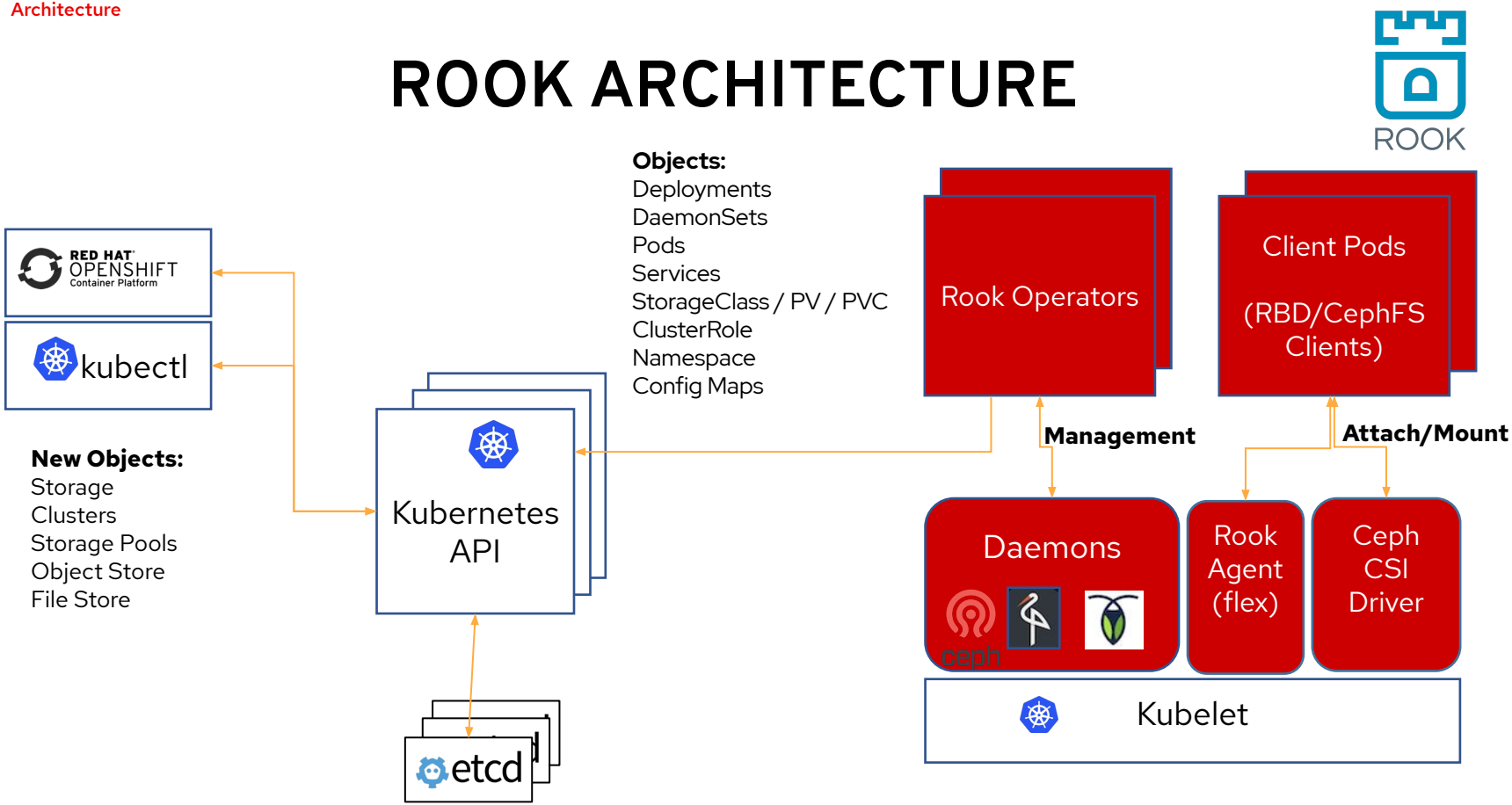
- Multiprotocol storage offers Block, File and Object interface
- Self-healing, self-management and rock solid technology
- Scale-Up and Scale-Out, performance and capacity at scale

- Multi Cloud Gateway enables S3 federation
- Provides elastic S3 data placement and improves security
- Multi-Cloud, Hybrid-cloud, Multi-Site Buckets

Rook Ceph Operator

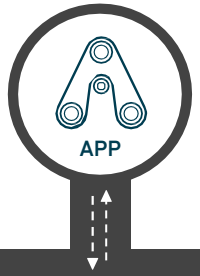
- The Operator leverages the full power of Kubernetes / OCP
  - Services, ReplicaSets, DaemonSets, Secrets, ...
- Contains all the logic to manage storage systems at scale
  - Handle stateful upgrades
  - Handle rebalancing the cluster
  - Handle health and monitoring tasks
- Not on the data path – can be offline for minutes

ROOK ARCHITECTURE



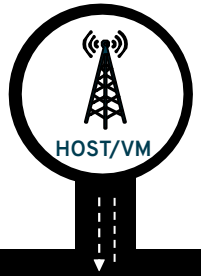
CEPH COMPONENTS

Storage services



RGW

A web services gateway for object storage, compatible with S3 and Swift



RBD

A reliable, fully distributed block device with cloud platform integration

CEPHFS

A distributed file system with POSIX semantics & scale-out metadata

LIBRADOS

A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby

RADOS

A software-based reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

ABOUT NOOBAA

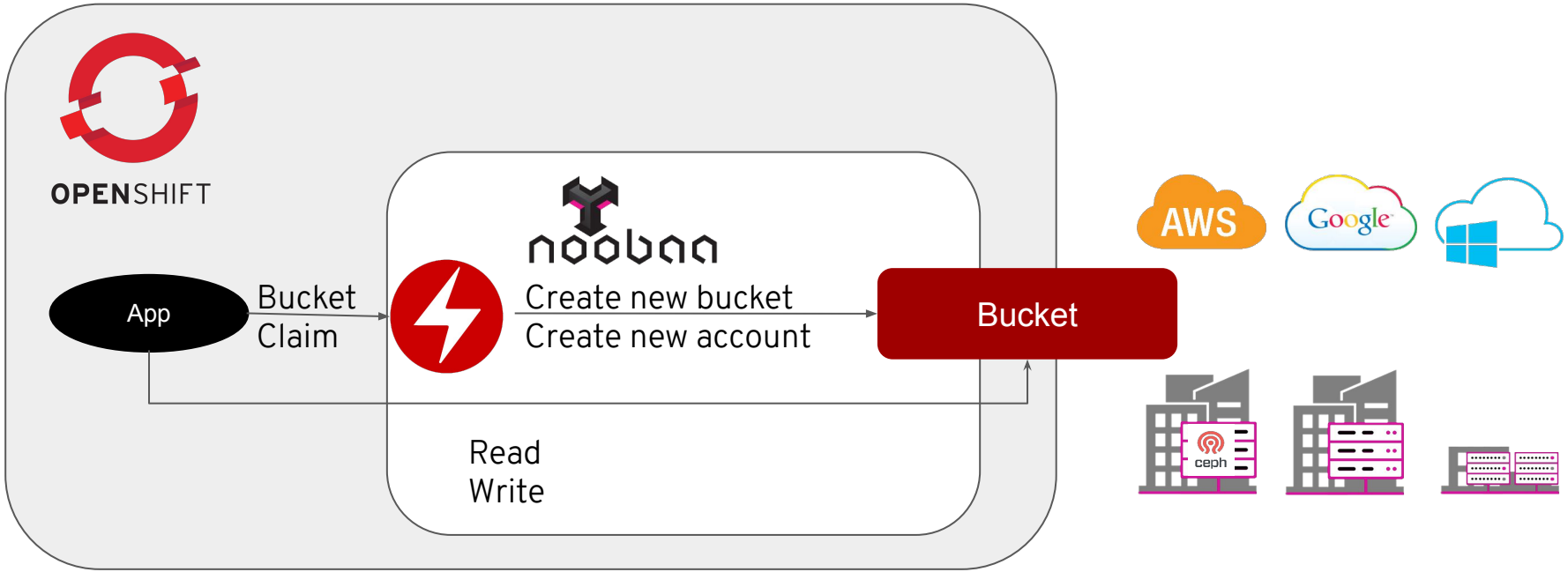
- **OCS MULTI CLOUD GATEWAY (NOOBAA)**
  NooBaa provides a consistent S3 endpoint across different infrastructures (AWS, Azure, GCP, Bare Metal, VMware)

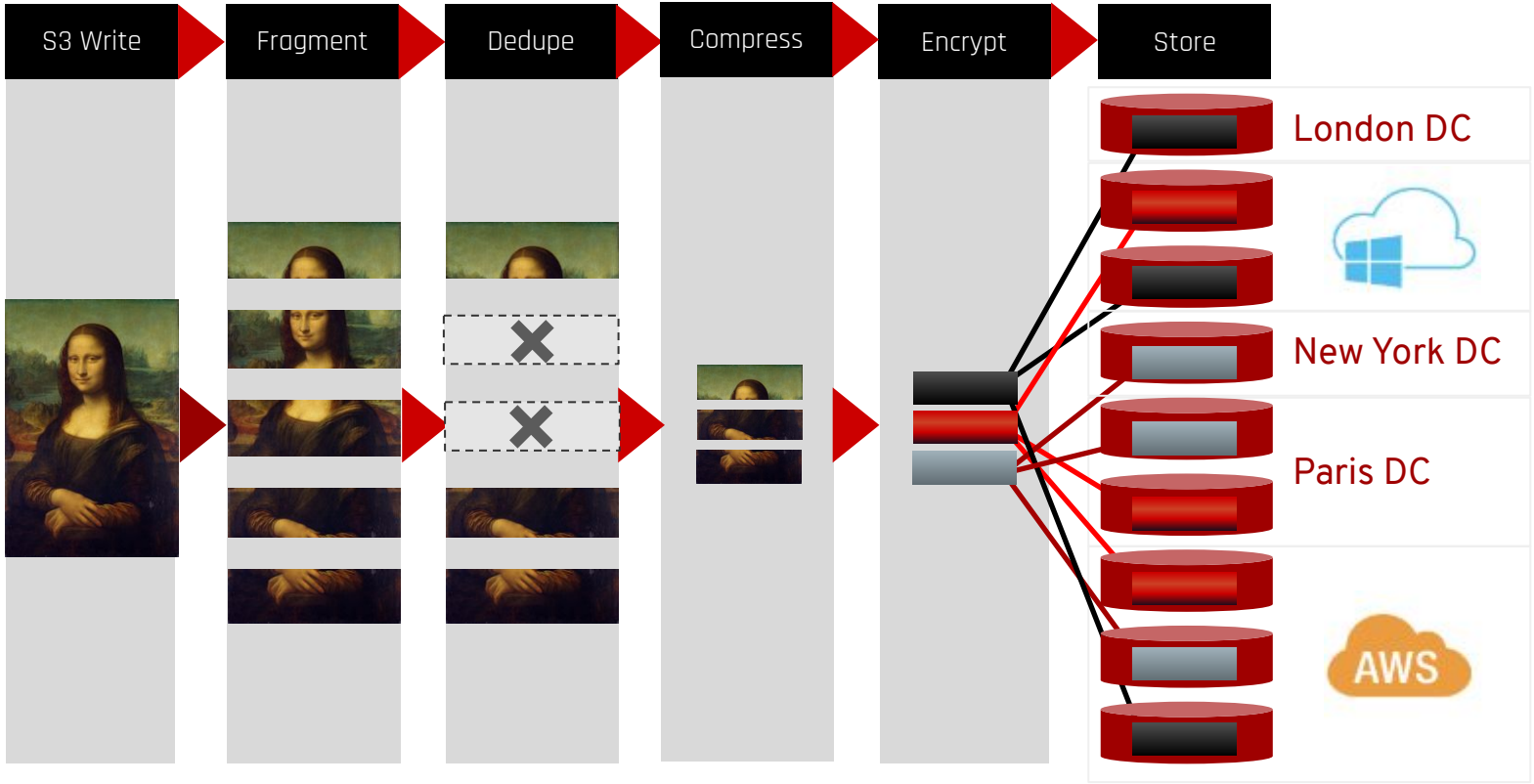
- **OCS MCG FUNCTIONALITY**
  Multi Cloud Object Gateway: Active/Active read/write across different clouds.

- **PRODUCTIZATION**
  productized as RHOCS Multi-Cloud Gateway, starting with OCS 4.2
  (NooBaa, is upstream only, downstream **OCS Multi-Cloud-Gateway**)

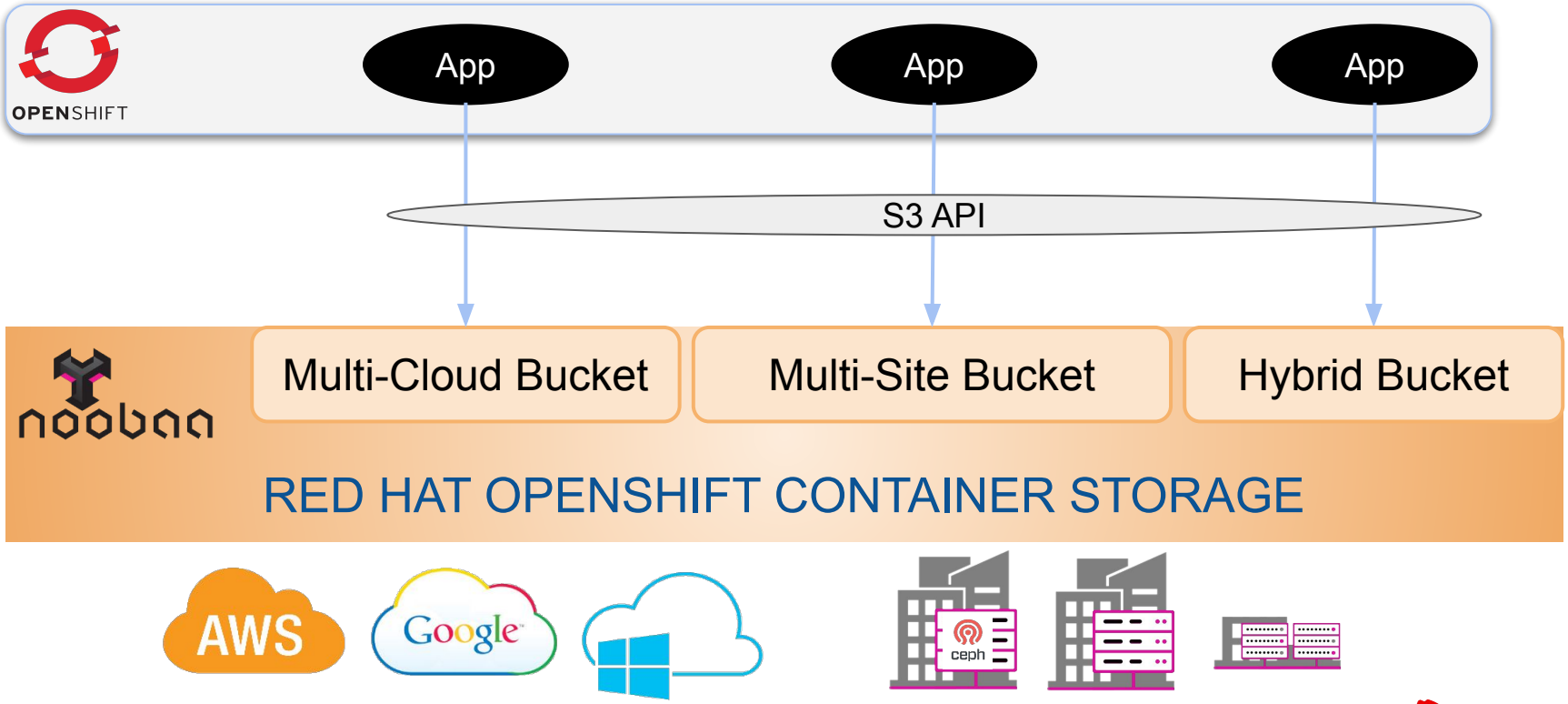
NOOBAA

S3 Federation with multi-cloud gateway




S3 Write
Fragment
Dedupe
Compress
Encrypt
Store
London DC
New York DC
Paris DC
AWS

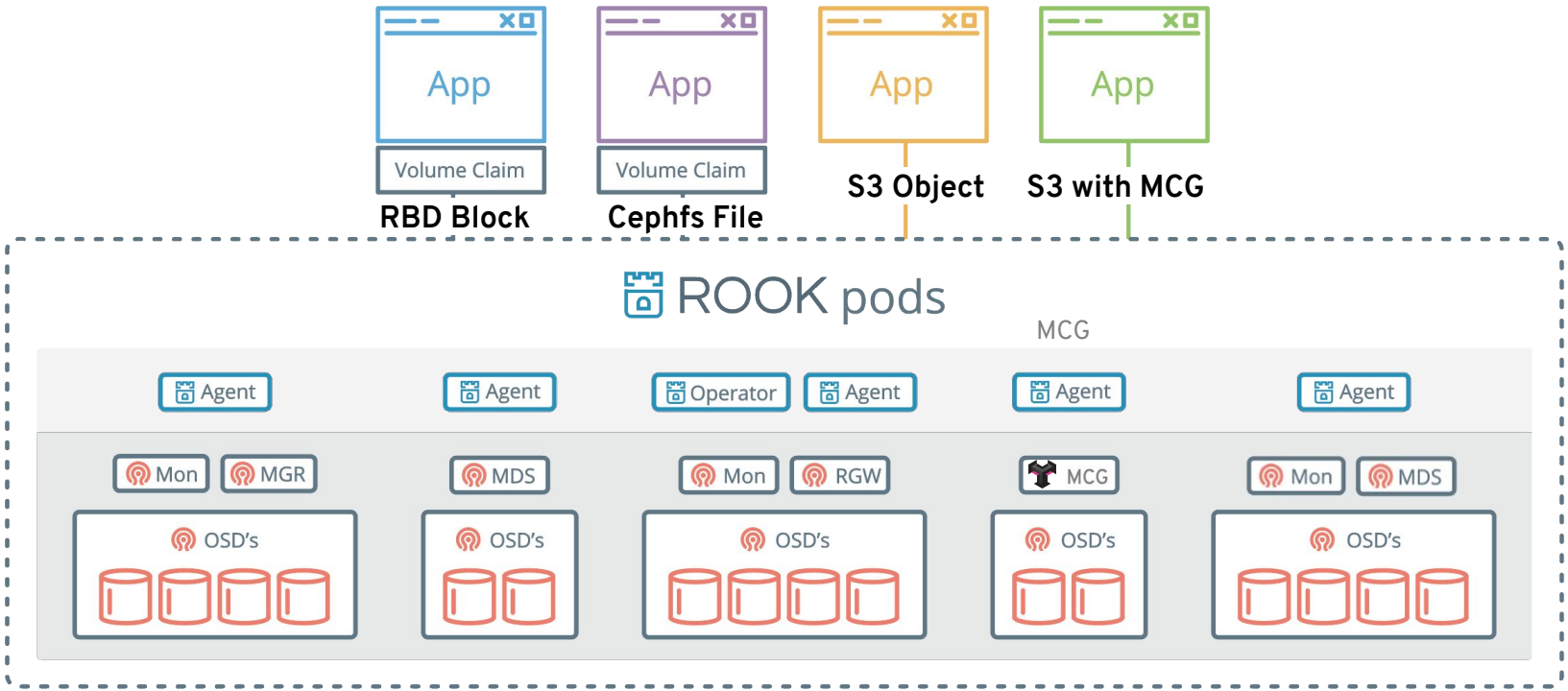
Multi-Cloud Object Gateway (NooBaa)



Red Hat

RHOCS ARCHITECTURE

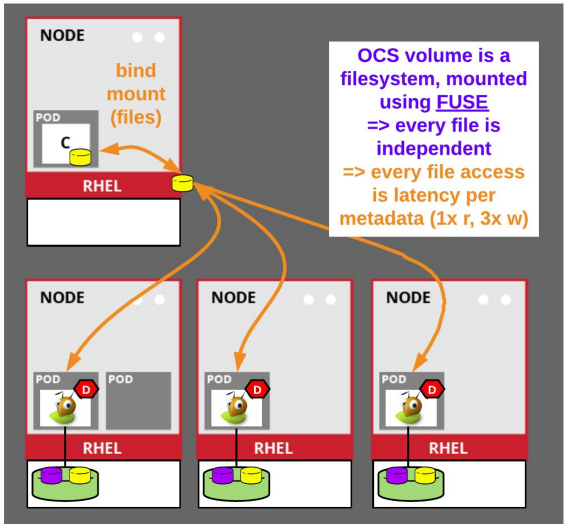
with Operator Lifecycle Manager (OLM)



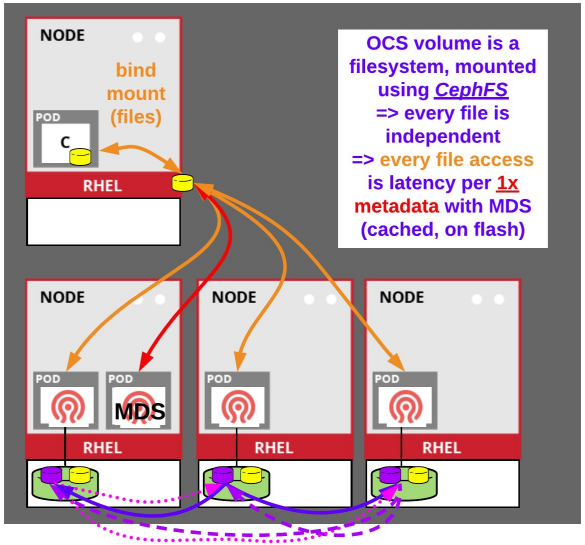
General structure: OCS3 vs OCS4

- RWX - how does it work ?

OCP3 w/ OCS3



OCP4 w/ OCS4

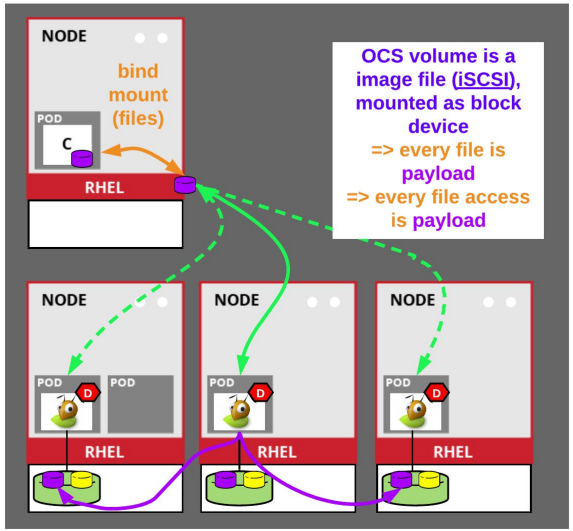


Red Hat

General structure: OCS3 vs OCS4

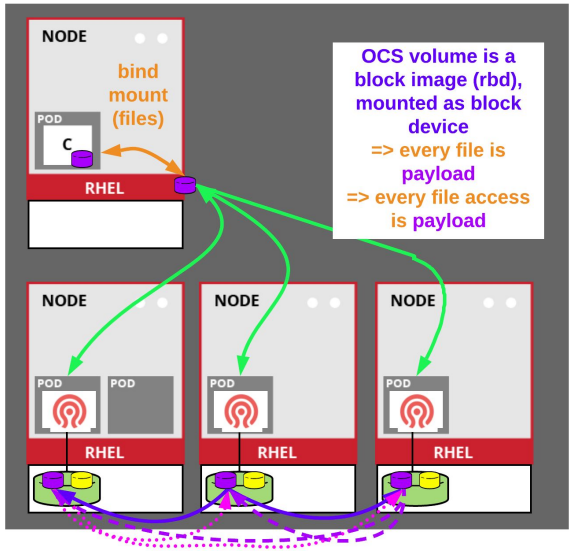
- RWO – how does it work ?

OCP3 w/ OCS3



- Single session, other failover only
- Same node replicates data

OCP4 w/ OCS4

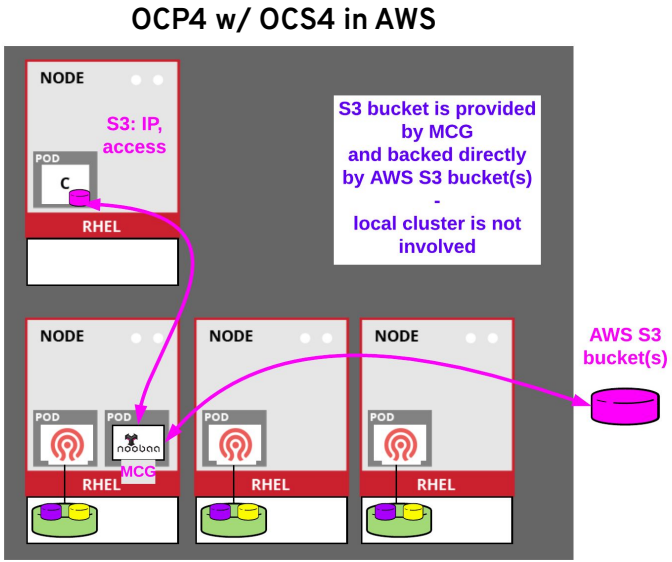


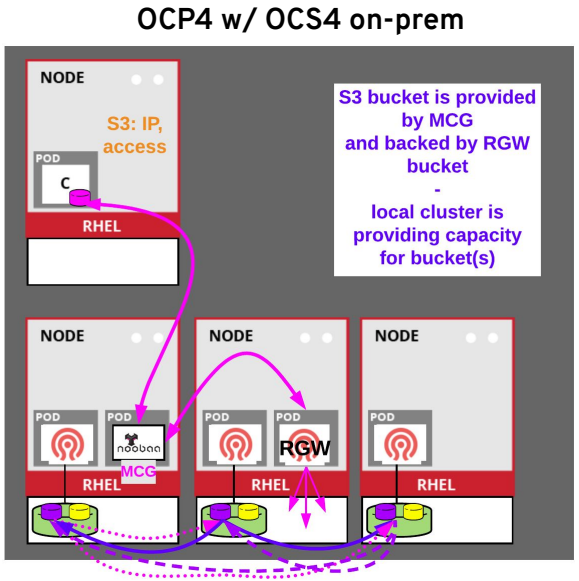
- All session
- All nodes replicate data (portions)

Red Hat

General structure: OCS3 vs OCS4

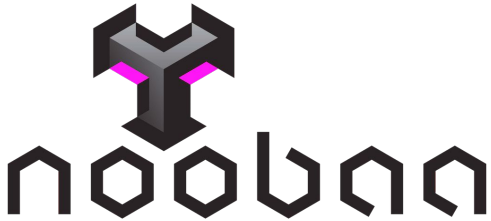
- S3 - how does it work ? --- OCS3 S3 is TP (still)





Red Hat

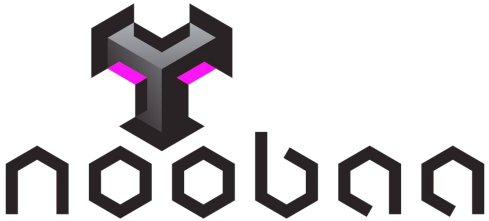
USE CASES



Use of OCS – What will change for applications ?

- Internal:
  - Registry is using CephFS PVs or S3 (instead of glusterfs PV)
  - Metrics and logging stay with block, but rbd based
- Apps:
  - Better latency and better throughput
  - Apps w/o need of scale-out for pods should go to block (rbd) > RWO
  - Apps w/ need of file haptic but w/o need of scale-out should go to block (rbd) > RWO
  - Apps w/o need of file haptic but w/ need of scale-out should go to S3 (MCG) > Object
  - Apps w/ need of file haptic and w/ need of scale-out should go to file (CephFS) > RWX

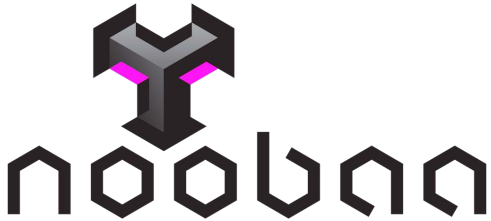
SIZING



SIZING GUIDELINES

- **MINIMUM NODES #** - The MINIMUM amount of storage nodes is **3**
- **REPLICA SIZE #** - Replica 3 (Erasure Coding planned on next releases)
- **PV SUPPORTED #** - Out-of-the-Box OCS4.2 supports up to **1500** PVs
- **ADDITIONAL NODES** - Each additional node enables for **+500** PVs
- **MAXIMUM NODES #** - The MAXIMUM number of nodes in a cluster is **10**
- **MAXIMUM PV #** - The MAXIMUM number of PVs can scale to **5000** PVs
- **OCS NODE CONFIG #** - MINIMUM OF 16vCPU AND 64GB RAM

DEMO



Red Hat

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