Data Script | IT training center

Morning Talks 19.4. 2018

Jiri Suchora

Just another storage

Agenda

- Introduction of CEPH, releases
- Basic design of CEPH
- Usage, modes
- (dis)Advantages
- Interconnection with other platforms
- Features
- How to start
- Lessons Learned, Perf Tunning, Monitoring, Benchmarking
- Enterprise distribution with support, real workload examples
- **DEMO**
- Questions?

Introduction of CEPH

- OpenSource project
- 88850 commits, 16847 pull requests, 885 authors (as for 26.9.2017)
- History:
  - 2007 as doctoral dissertation
  - 2012 Inktank was created
  - 2014 Acquisition by Red Hat
  - 2015 Ceph Community Advisory Board was formed for Ceph Project
  - Made by: Red Hat, Intel, SUSE, Mirantis, Cisco, IBM, CERN, Fujitsu, SanDisk, big bunch of individuals and many others…

Introduction cont.

- Name "Ceph" is a common nickname given to pet octopuses and derives from cephalopods
- Pools (logical partitions for storing objects)
- Replicated Objects
- Objects in Placement groups
- Balanced thru cluster

Stable versions

- Jewel – 10.2.x (LTS, until June 2018)
- Luminous – 12.2.x (new LTS, no estimated retirement yet)
  - Recommended way for starting to play with

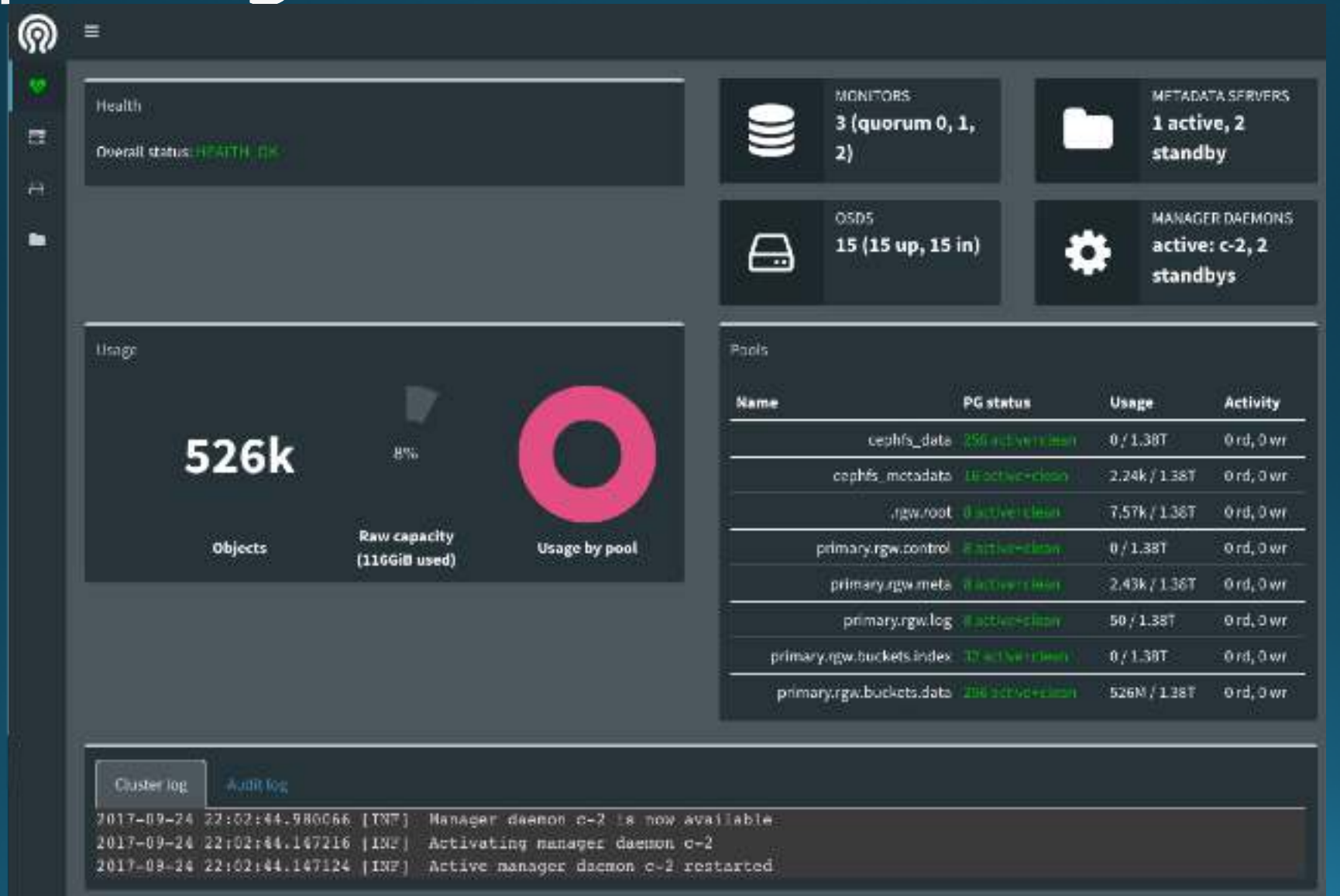
Design

- Cluster monitors (ceph-mon)
  - keep track of active and failed cluster nodes
- Metadata servers (ceph-mds)
  - store the metadata of inodes and directories
- Object storage device nodes (ceph-osd)
  - store the content of files in an XFS file system, or BlueStore
- RGW nodes (ceph-rgw)
  - RESTful gateways – S3 and OpenStack Swift APIs

Design cont.

- NFS Gateways
  - NFS access for RGW via NFS-Ganesha, still limited operations right now
- iSCSI Gateways
  - provide a Highly Available (HA) iSCSI target that exports RADOS Block Device (RBD) images as SCSI disks
- Ceph Manager Daemon (ceph-mgr)
  - provide WEB GUI Dashboard with additional monitoring and interfaces to external monitoring and management systems + bundle of modules (RESTful, Zabbix, Prometheus…)

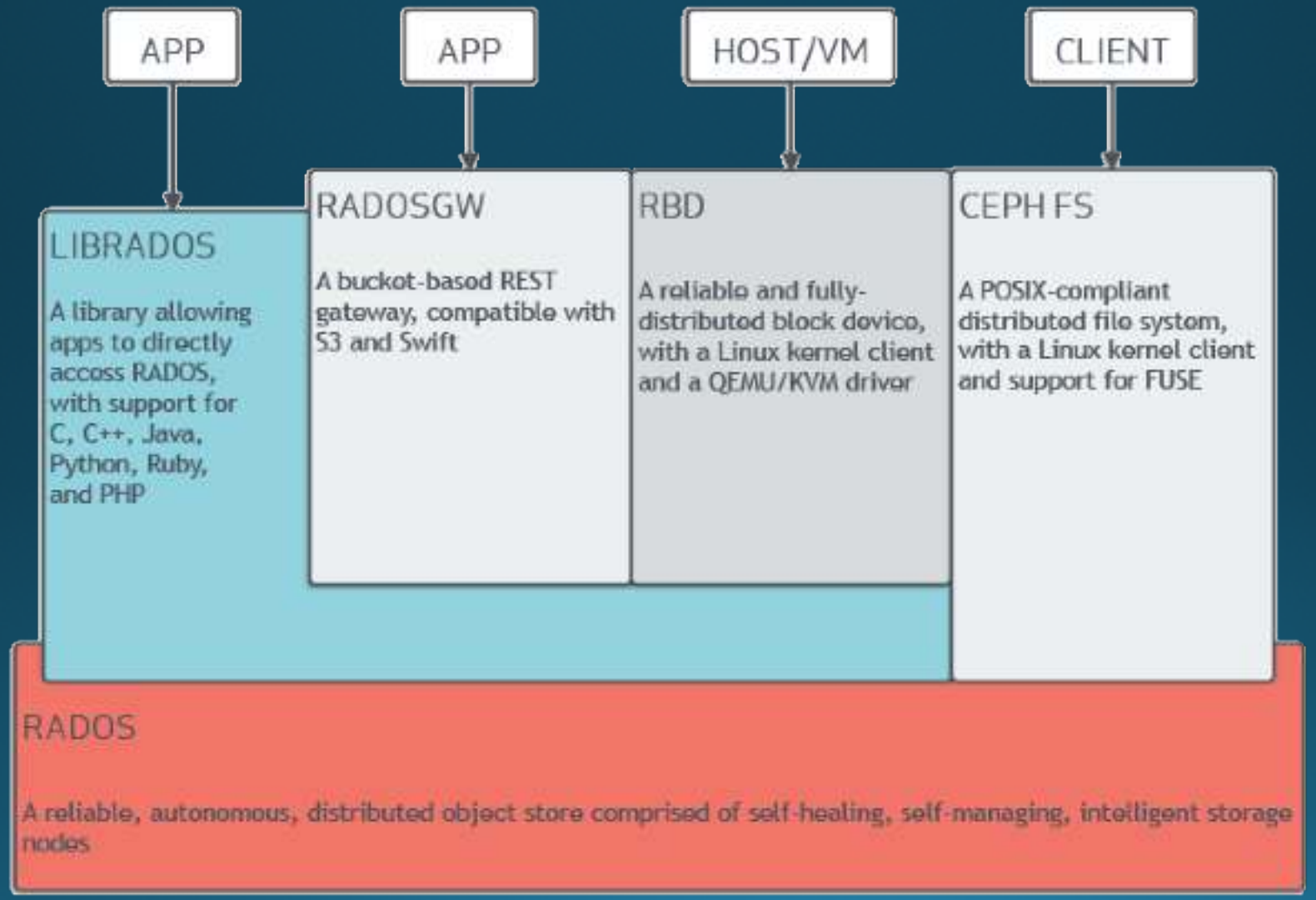
Ceph-mgr daemon



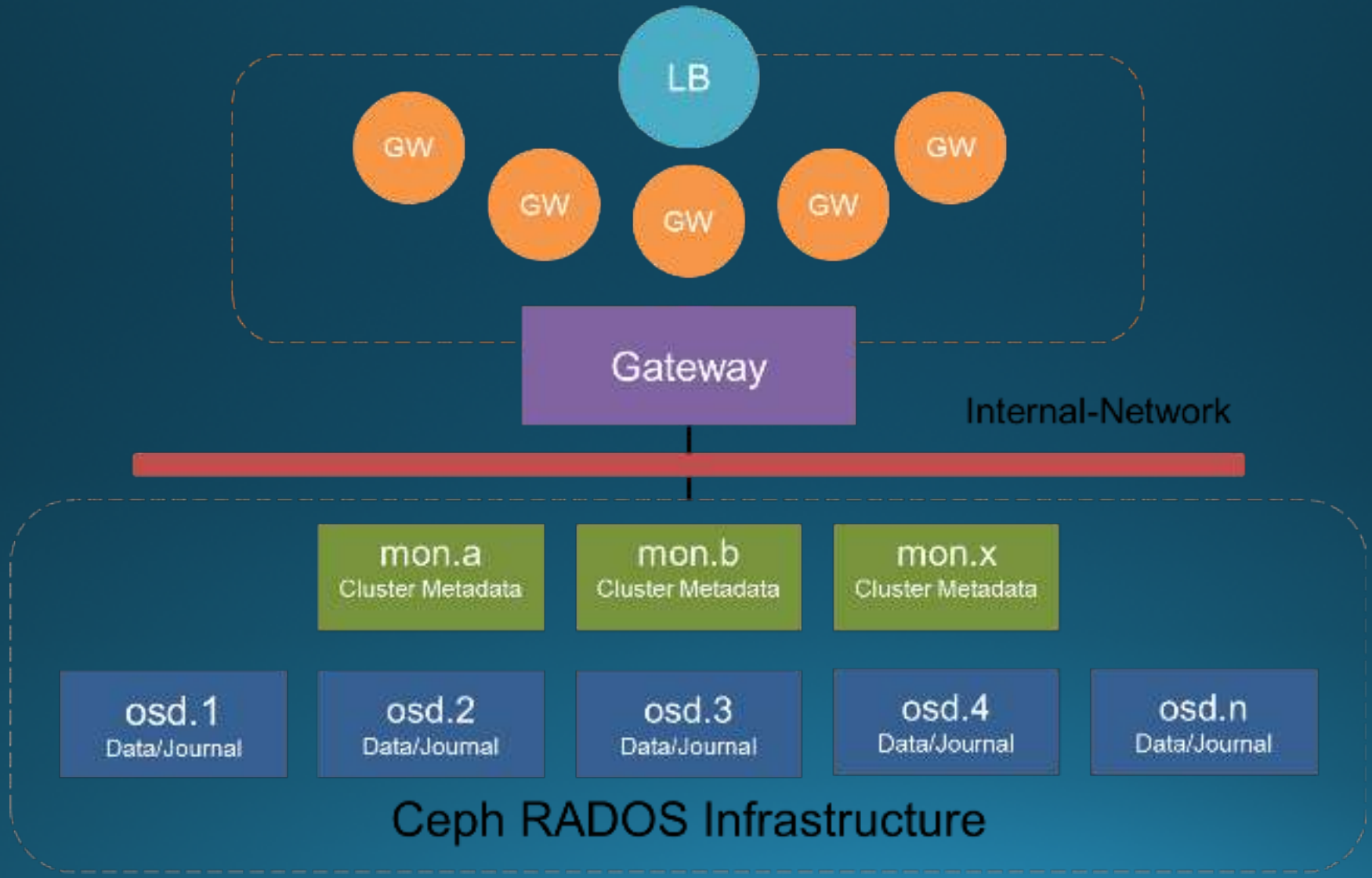
Usage, access to data

- Librados (library for direct access – support of Java, Ruby, Python, C,C++ ..)
- RBD (RADOS block device)
- S3 (AWS S3 API)
- Swift
- CephFS (POSIX compatible, mounts as filesystem)
- iSCSI(via iSCSI gateway)
- pNFS (ganesha)

Usage, modes


APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

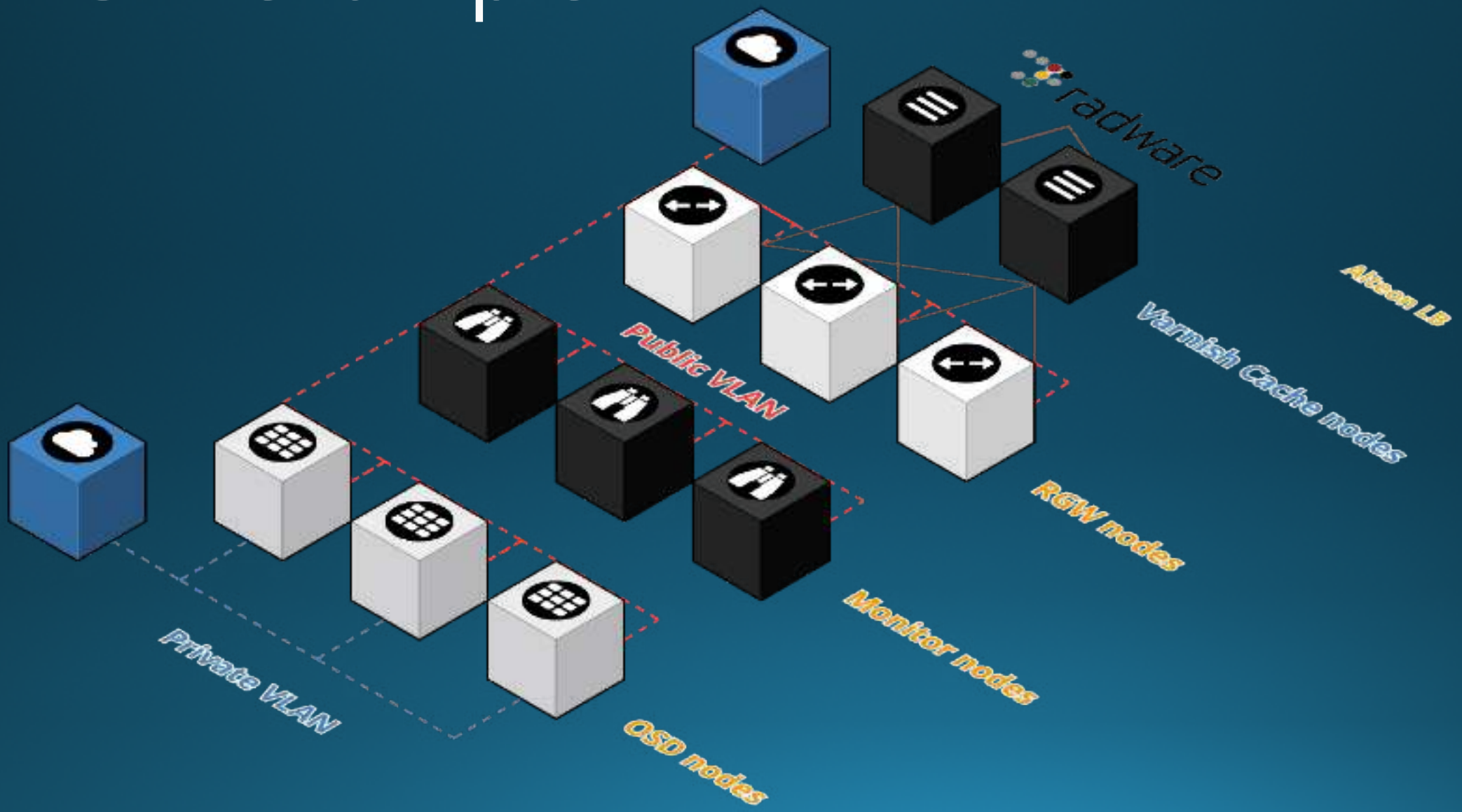
S3 access via RGW, LB appliance



S3 access via RGW, other ways?

- Haproxy instead of LB appliance (cost)
  - At least 2 nodes
  - Keepalive between them
- Varnish Cache
  - If RAM is not a problem
  - At least 2 nodes
  - LB in the front

RGW example


radware
Public VLAN
Private VLAN
Alteon LB
Varnish Cache nodes
RGW nodes
Monitor nodes
OSD nodes

(dis)Advantages

- Scale-out vs Scale-up
- Fast self healing
- Commodity hardware
- Prevent vendor lock-in
- Open Source where feasible
- Reduce Total Cost of Ownership (TCO)

- Has to be learned, adopted
- Needs change of mind (conservative thinking)
- Not cure for everything -really depends on case!

Scalability

CERN as example:

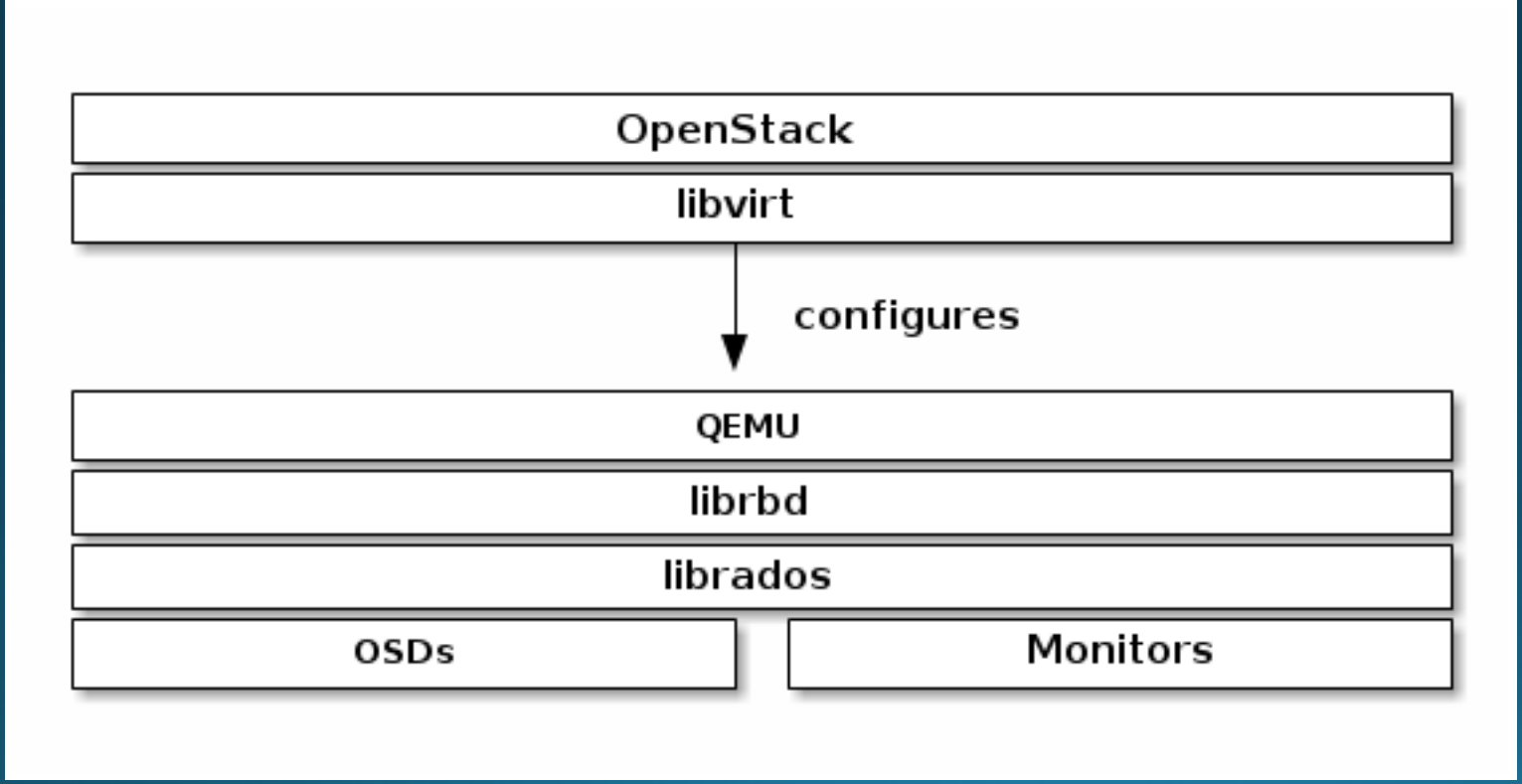
- 225 servers
- 48 6TB hard drives per server
- 10,800 OSDs in total, nearly 65 petabytes!
- 3 monitors and managers, collocated with OSDs

Interconnection with other platforms

- OpenShift, Kubernetes
- OpenStack (Cinder, Glance)
- RHEV/oVirt
- Apache CloudStack
- OpenNebula
- Ganeti
- Proxmox
- Dovecot
- Vmware (ESXi via iscsi)

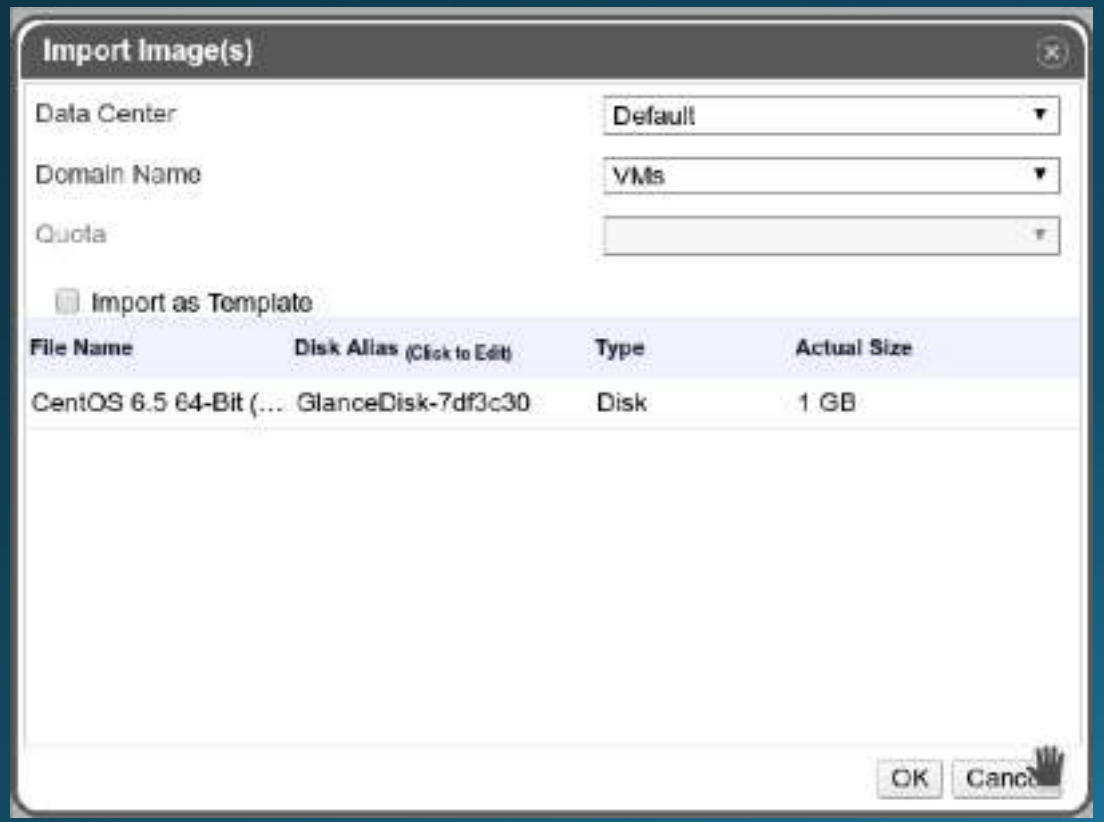
Ceph with OpenStack

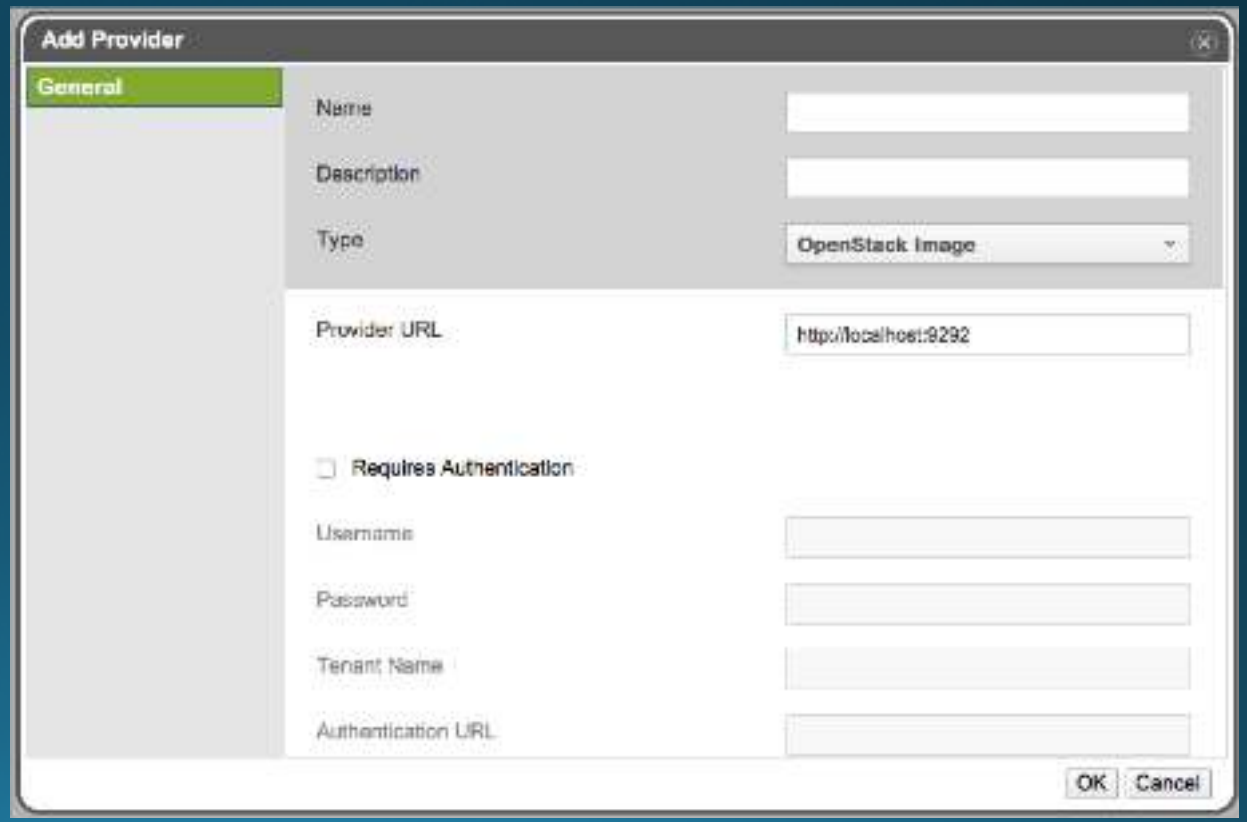
- Cinder, Glance, via libvirt/librbd



Ceph with RHEV/oVirt

- through libvirt, which configures the QEMU interface to librbd

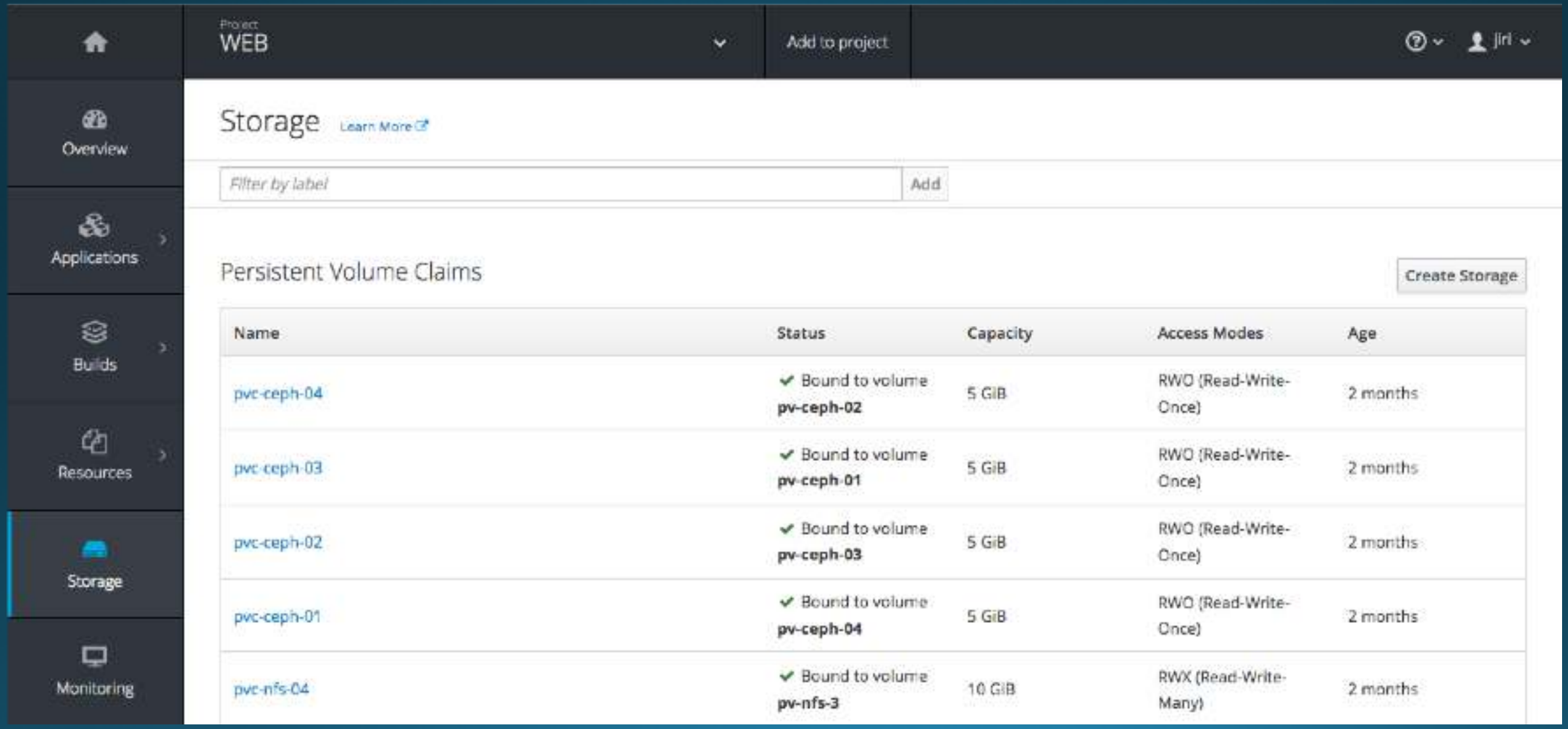




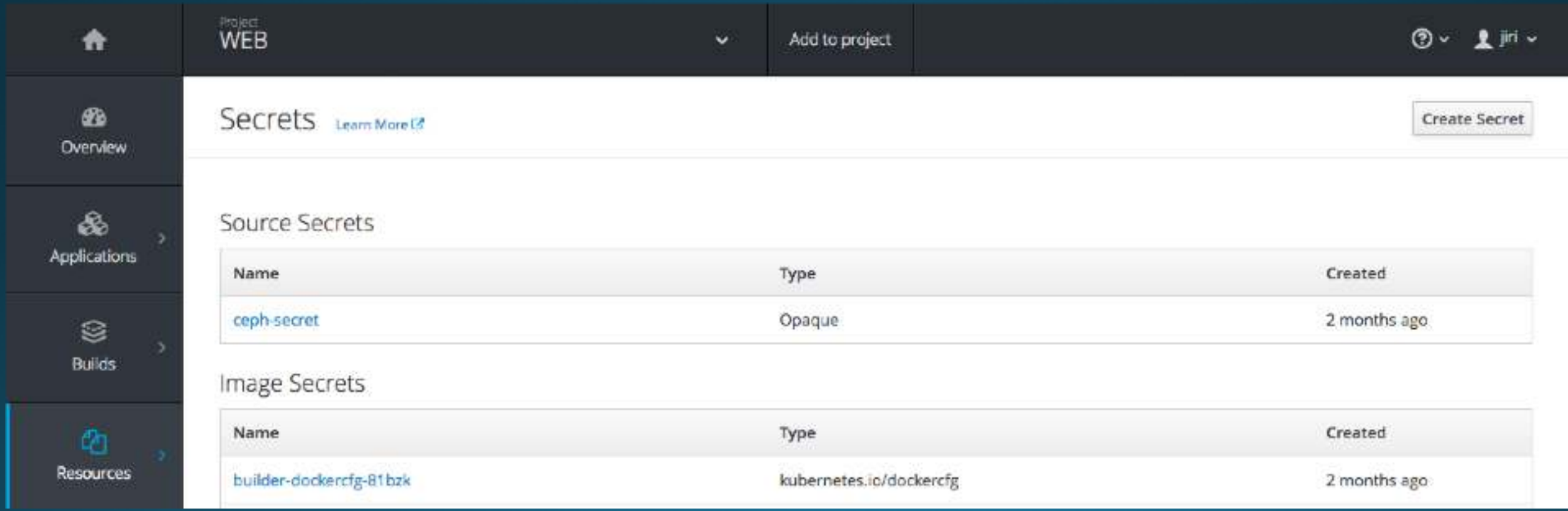
Ceph with OpenShift

- Persistent storage using Ceph RBD
- Persistent volumes (PVs) and persistent volume claims (PVCs) can share volumes across a single project
- Access via authorization key from CEPH
- Security (diff from NFS,GlusterFS) - user and group IDs defined in the pod definition or docker image are applied to the target physical storage

Ceph with OpenShift

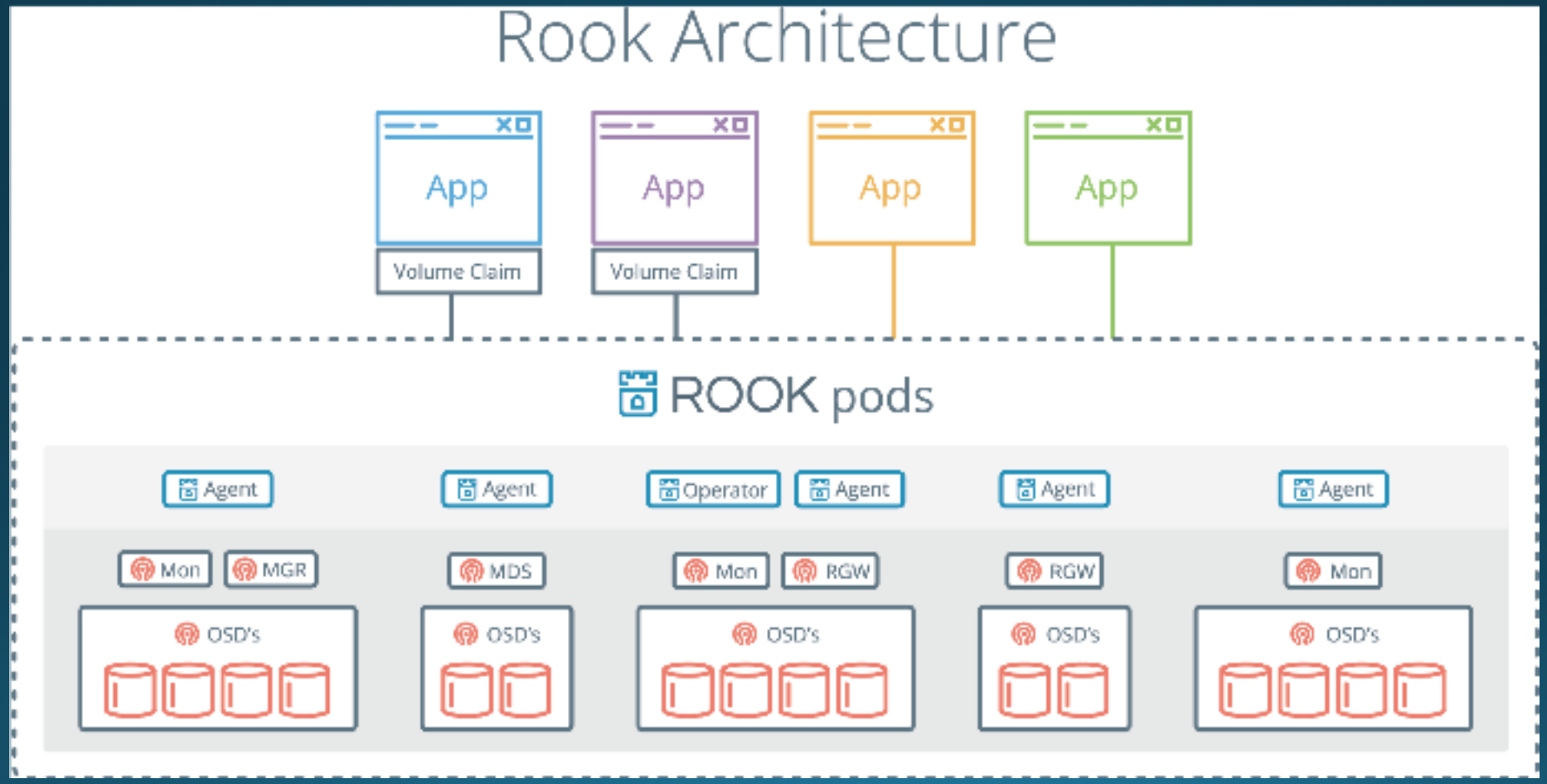


Ceph with OpenShift



Ceph with Kubernetes

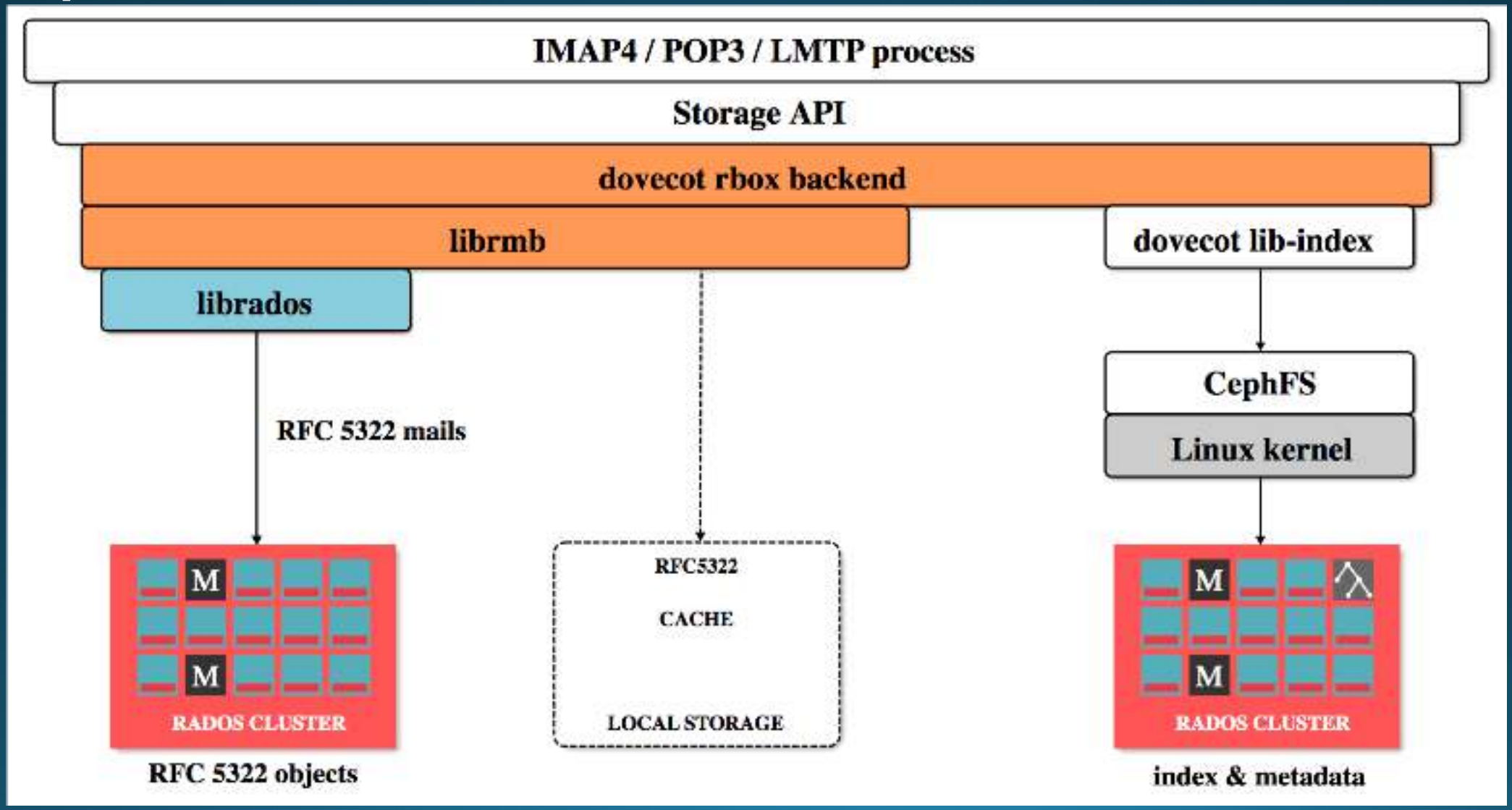
Ceph via Rook.io (managed by operator)



Ceph with Dovecot

- open-source IMAP and POP3
- 72% market share (openemailsurvey.org, 02/2017)
- Objectstore plugin obox (S3/swift), not opensource, Dovecot Pro
- Librmb (layer between dovecot backend and librados)
  - Deutsche Telekom, SUSE…
  - OpenSourced (https://github.com/ceph-dovecot/dovecot-ceph-plugin)
  - Being implemented in Detsche Telekom
  - 1.3 petabyte net storage
  - ~39 million accounts, 6,7 bilion emails

Ceph with Dovecot



Other features…

- Multisite Deployment (Realms, Zone groups, Zones), Disaster Recovery to Second site
- Snapshots (Pools, Objects)
- Quotas per:
  - CephFS directories,
  - RGW Buckets
  - RGW Users
- RGW Metadata Search – via Elasticsearch (from Luminous)
- Bluestore backend – better perf, inline compression…

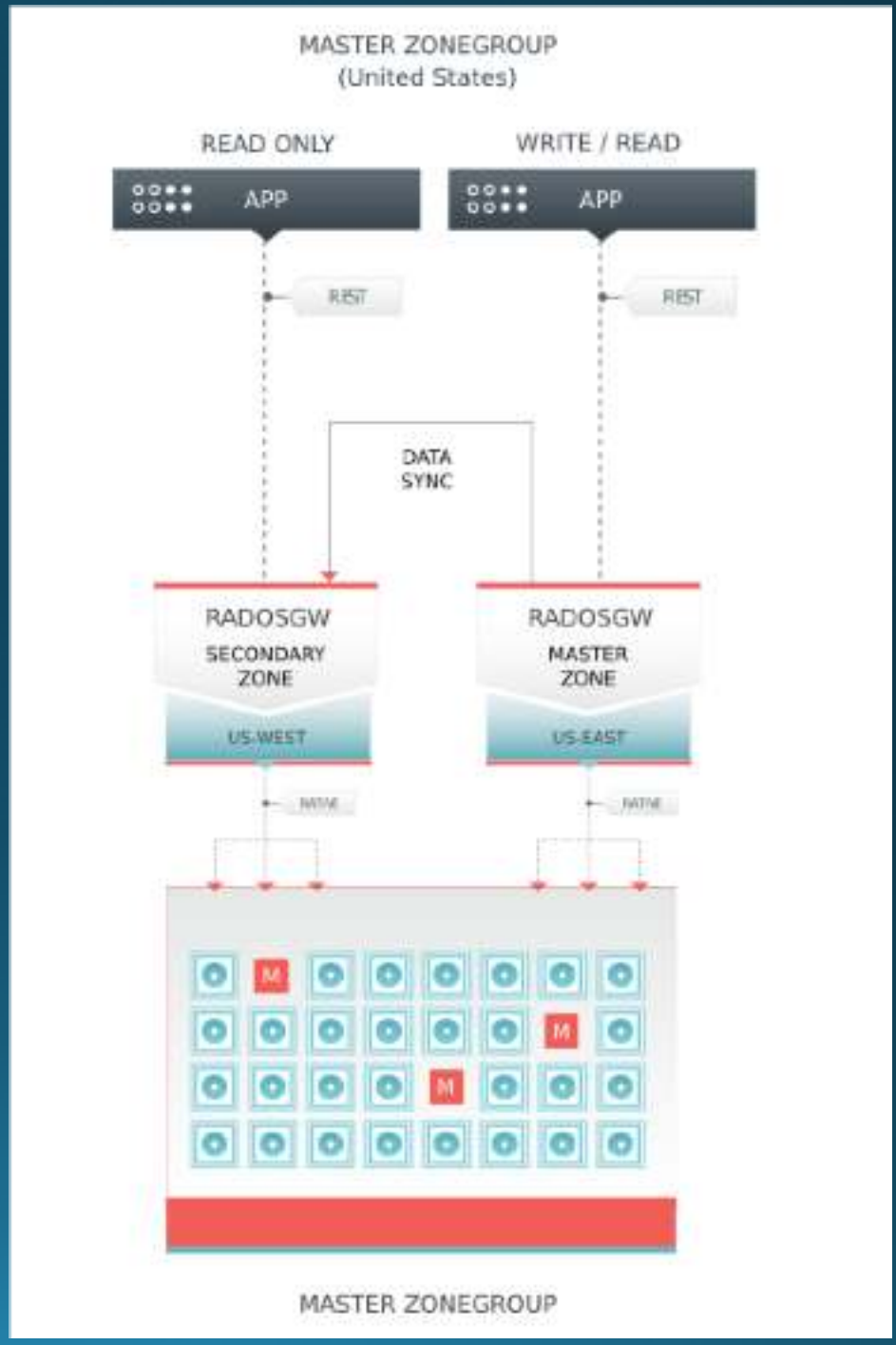
Other features

- Storage tiering (eg.)
  - fast: all flash
  - medium: disks accelerated by some flash journals
  - slow: archive disks with collocated journals
  - Tiered CRUSHmap (eg.)
    - Bigger than 6 TB → Archive drive
    - Between 1.6TB and 6TB → Disk with flash journal
    - Smaller than 1.6TB → SSD assumed

Multisite Deployment

- Zone: Logical Group of one or more RGW instances
- Zone group: one or more group, region
- Realm: container for one or more Zone groups

- Synchronous/Asynchronous
- Failover possible
- Migration from Single to Multisite possible



How to start

- How to deploy?
  - Ceph-ansible (https://github.com/ceph/ceph-ansible)
  - Ceph-deploy (https://github.com/ceph/ceph-deploy)
  - DeepSea and Salt (for SUSE Enterprise Storage)
  - Crowbar (http://crowbar.github.io)
  - Ceph-docker (https://github.com/ceph/ceph-docker)
  - Manually

How to start

- Single VM -> just for testing
- Ceph in docker – all in one -> just for testing
  https://hub.docker.com/r/ceph/demo/
- At least 3-3-3 for production

Lessons learned

- **Definitely not** VMs for production!!! (at least for OSD nodes)
- Local storage, **not** SAN/NAS (esp. with own logic, tiering)!!!
- SSD, SSD, SSD if possible ☺ (at least for journals)
- Perf. Testing before production – try to simulate workload as much as possible
- Tuning – there is no default recipe for everything, listen your CEPH cluster and tune it as you go!

Enterprise, HW, Alternatives

- Red Hat Ceph Storage (RHEL, Ubuntu)
- SUSE Enterprise Storage (SLES)
- OpenStack distributions (Mirantis, Canonical, …)
- Minio S3
- PetaSAN.org (ceph via iscsi)

- EMC ECS
- Hitachi HCP
- Cisco Object Storage

Where (not) to run Ceph?

- OpenShift, oVirt , small clusters? (vs. glusterfs)
- Vmware/ESXi, via iscsi?
- Windows via iscsi?
- Enterprise vs. Free
- Software vs. HW

Monitoring, Benchmarking

MONITORING

- Zabbix (native plugin for MGR module from Luminous)
- Nagios
- Grafana+Prometheus (native plugin for MGR module from Luminous)
- ELK

BENCHMARKING

- COSBench (https://github.com/intel-cloud/cosbench)
- CBT (https://github.com/ceph/cbt) - FIO on steroids
- AB (ApacheBench), WRK (https://github.com/wg/wrk)

COSBench tool by Intel

General Report

Op-Type	Op-Count	Byte-Count	Avg-ResTime	Avg-ProcTime	Throughput	Bandwidth	Succ-Ratio
op1: read	563 ops	563 KB	24,3 ms	24,28 ms	112,02 op/s	112,02 KB/S	99,82%
op2: write	62 ops	62 KB	22,02 ms	22,02 ms	12,35 op/s	12,35 KB/S	100%

The snapshot was taken at 21:21:57 with version 62.

Stages

Current Stage	Stages completed	Stages remaining	Start Time	End Time	Time Remaining

ID	Name	Works	Workers	Op-Info	State	Link
w4-s1-init	init	1 wks	7 wkrs	init	completed	view details
w4-s2-main	main	1 wks	3 wkrs	read, write	running	view details

Performance Graph

Where Ceph is running?

- CERN
- Dropbox
- Bloomberg
- Deutsche Telekom
- Cisco
- …
- Openstack Clouds (by: Mirantis, Canonical Cloud,…)

DEMO

- MGMT Web Dashboard
- COSBench (web controller, XML examples)
- RBD device for linux client
- CephFS for linux client
- Persistent storage for Kubernetes cluster
- S3

Location of this PDF...

- https://github.com/yrjo/ceph-tweaks/blob/master/mtalks_19042018.pdf

Do you need help?

- Trainings?
  - www.datascript.cz - OpenSource Trainings
  - Contact at: info [AT] datascript [DOT] cz
- Help needed?
  - PoC
  - Consultations
  - Design, build, support
- Contact me:


jiri [AT] noibit [DOT] cz


https://www.linkedin.com/in/jirisuchora/

Questions?

Thanks!