# Whatever can go wrong, will go wrong - Rook/Ceph and storage failures


*Sagy Volkov, Red Hat*

# Agenda



- Storage Intro
- Resiliency (or storage terminology for developers)
- Ceph as a storage provider
- Rook as a storage orchestrator
- Failures
- ~~Live~~ Demo
- Questions

# Storage Intro (1)

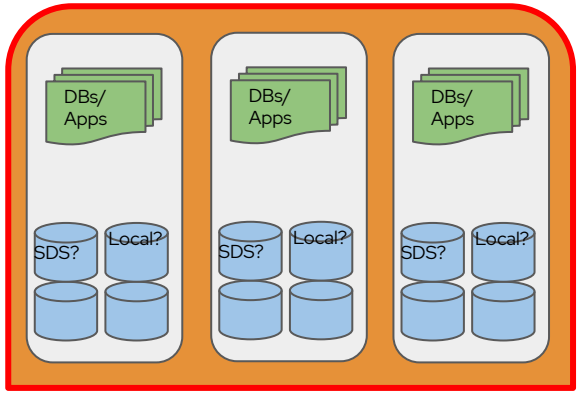


- We all need it and we all use it :)
- It might be ephemeral or persistent but we use it.
- Today’s discussion is on persistent storage.
- Persistent - anything we need to retain so we can use again

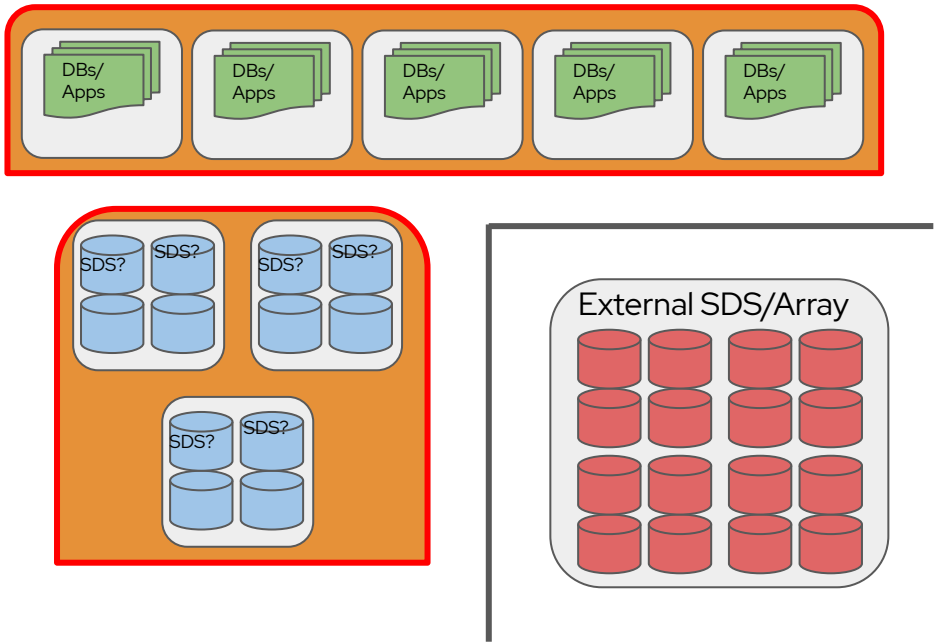
# Storage intro (2)



Converge




Non Converge / External storage



# Storage Intro (3)

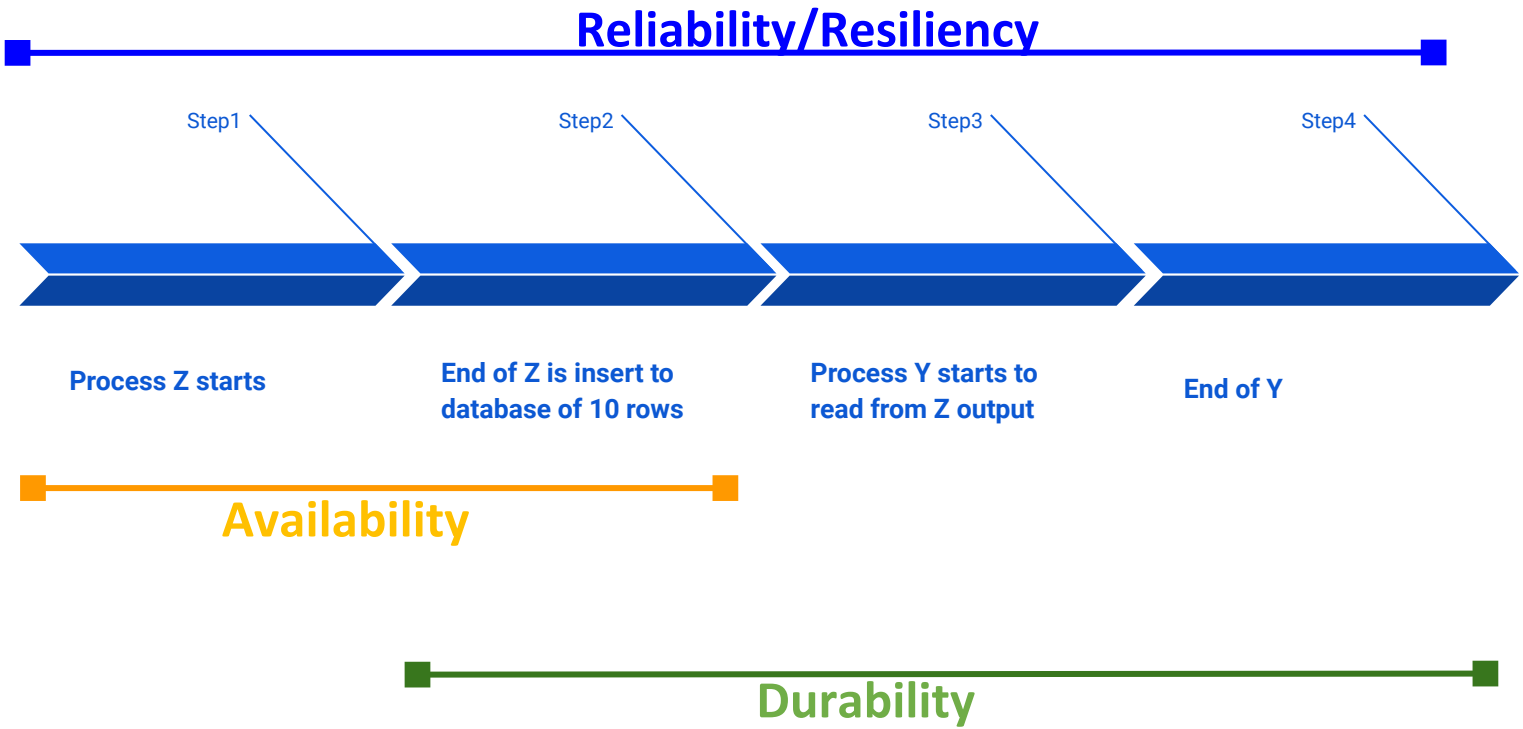


Some wide used terms:

- Availability - basically your storage uptime
- Durability - an existence of object/data/storage
- Reliability/Resiliency - the probability that your storage is working as planned/designed, the ability of a storage system to heal itself.

# Storage Intro (4)





# Resiliency (1)



## MTTR (Mean Time To Repair)

- How long it takes for you to fix a problem.
- In storage: drive, blade, array, switch, mostly hardware based.
- In SDS: Converge can be a concern. Mesh is a big help.
- Important for your SLA.

# Resiliency (2)



## MTBF (Mean Time Between Failures)

- It all comes down to the quality of all devices used for your storage.
- Basically the measurement from when your last failure occur until the next one. And they will happen...
- For comparison, Ent. grade drive have ~800,000 hours of MTBF, about 90 years.
- So… 90 drives = one failure every year,
  900 drives = every 5 weeks, and so on...

# Backup/Recovery

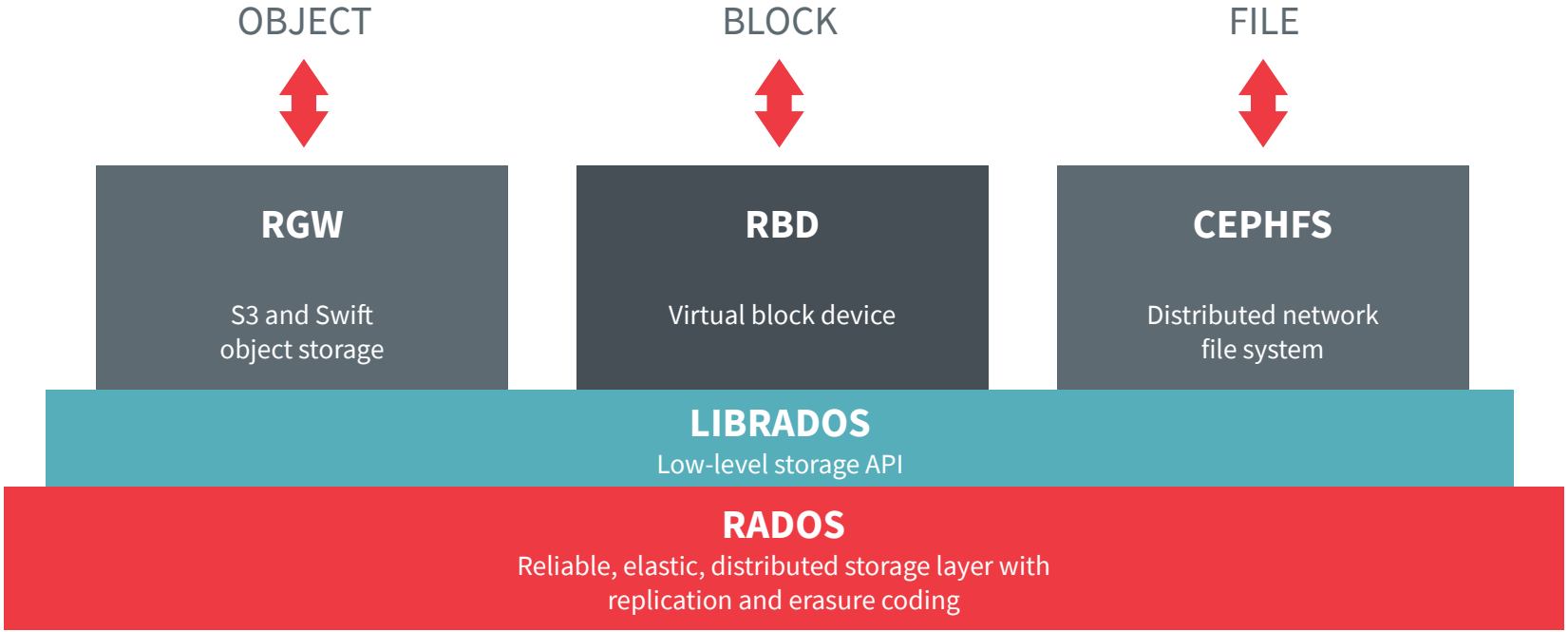


- RTO (Recovery Time Objective):
  - How long can your process, application, data center or company can survive without data access? However long it takes you to recover.
  - Usually done via tiers (1 to 3) with different SLAs.
- RPO (Recovery Point Objective):
  - Don’t trust your storage solution :) backup whenever you can.
  - In case of failure, how many backups, how often you run backups will determine your RPO.

# Ceph (1)



## Ceph is a unified storage system



# Ceph (2)




ceph-mon

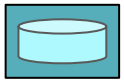
Monitor

- Central authority for authentication, data placement, policy
- Coordination point for all other cluster components
- Protect critical cluster state with Paxos
- 3-7 per cluster

ceph-mgr

Manager

- Aggregates real-time metrics (throughput, disk usage, etc.)
- Host for pluggable management functions
- 1 active, 1+ standby per cluster



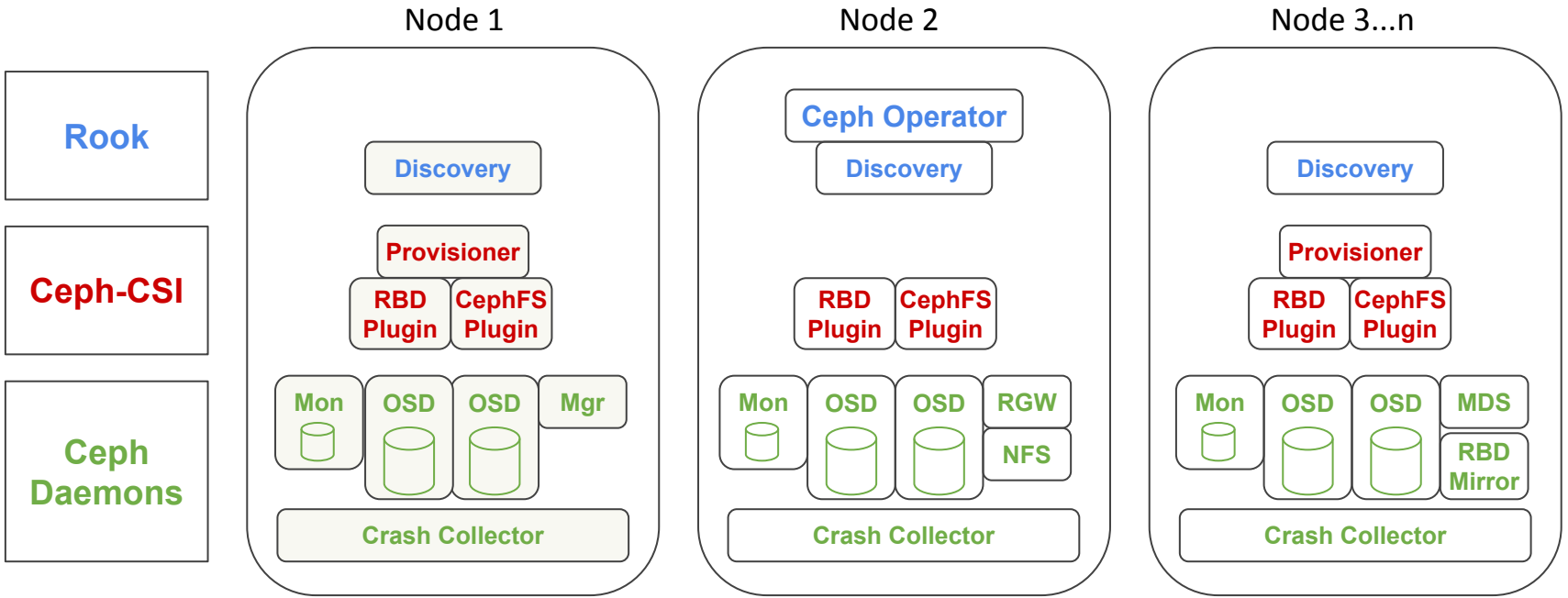
ceph-osd

OSD (Object Storage Daemon)

- Stores data on an HDD, SSD, NVMe, any block device
- Services client IO requests
- Cooperatively peers, replicates, rebalances data
- 10s-1000s per cluster

# Rook (1)




Rook
Ceph-CSI
Ceph Daemons
Node 1
Discovery
Provisioner
RBD Plugin
CephFS Plugin
Mon
OSD
OSD
Mgr
Crash Collector
Node 2
Ceph Operator
Discovery
RBD Plugin
CephFS Plugin
Mon
OSD
OSD
RGW
NFS
Crash Collector
Node 3...n
Discovery
Provisioner
RBD Plugin
CephFS Plugin
Mon
OSD
OSD
MDS
RBD Mirror
Crash Collector

# Rook/Ceph Resiliency (1)



- Every process is a pod - so what happens when a pod fails?
- MONs x 3
- MDSs x 2 (Active/Standby)
- None of the MONs or MGR processes/pods are in the data path.
- For replication/site mirroring:
  - Ceph RBD mirror
  - Ceph object multisite (GA in Ceph, experimental in Rook for now)

# Rook/Ceph Resiliency (2)



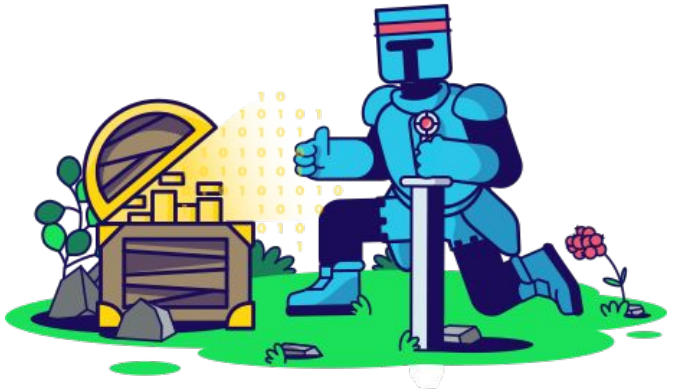
- Let's look at a few scenarios with OSD pod:
  - Delete an OSD pod
  - Reboot a node with OSDs
  - Device lost
- We will use the sherlock project to run database and create stress on the storage (rook/ceph).
- AWS was used for the demos.
- Workload software: https://github.com/sagyvolkov/sherlock

# Demo



# Thank you!

https://ceph.io
https://rook.io
https://github.com/sagyvolkov/sherlock




KEEP CLOUD NATIVE
EVERYWHERE
KubeCon
CloudNativeCon
North America 2020
Virtual
HELM
KV