open storage network

Made possible by grants from the NSF and the Schmidt Foundation

Open Storage Network Seminar Series

Ceph

January 21, 2021

NSF grants
1747552
1747493
1747507
1747490
1747483

What is Ceph?

Open source distributed object storage service

- **S3 Compatible Object Storage**
- Block Storage
- POSIX-Compatible File System

Ink Tank released first stable version of Ceph in 2012

Red Hat purchased Ink Tank in 2014

- Red Hat Ceph Storage (RHCS) - Licensed
- **Community Supported - Free**

Ceph at the OSN

- Currently running the **community supported** version
- Only the S3 compatible **Object Storage** service
- Running version 14.2 (**Nautilus**)
  - Plans to upgrade to 15.2 (Octopus) in 2021
- Using **ceph-ansible** for configuration management
- All Ceph services will be migrated to run in **Docker Containers** by end of Q1 2021

NAUTILUS

Supported S3 Features

Ceph supports a RESTful API that is compatible with the basic data access model of the Amazon S3 API.

- List Buckets
- Delete Bucket
- Create Bucket
- Bucket Lifecycle
- Policy (Buckets, Objects)
- Bucket Website
- Bucket ACLs (Get, Put)
- Bucket Location
- Bucket Notification
- Bucket Object Versions
- Get Bucket Info (HEAD)
- Bucket Request Payment
- Put Object
- Delete Object
- Get Object
- Object ACLs (Get, Put)
- Get Object Info (HEAD)
- POST Object
- Copy Object
- Multipart Uploads
- Object Tagging
- Bucket Tagging
- Storage Class

S3 Compatible

More Details at https://docs.ceph.com/en/latest/radosgw/s3/

Erasure Code

Current Pods are using a 3+1 EC profile config

- Ceph Nautilus requires k+m+1 servers for an EC profile, with only 5 servers the OSN was limited to using 3+1 or 2+2 EC profiles.

Future Pods will support EC profiles with more redundancy

- With the new 2021 Pod configuration we can support 4+2 or 8+3
- EC profiles can be set on per Pod basis depending on the use case
- Ceph Octopus removes the +1 requirement in k+m+1

OSN Pod Durability

Erasure Code Profile	Durability
3+1 (OSN First Gen)	99.9999% (6 9s)
4+2 (OSN Future)	99.99999999% (10 9s)
AWS S3 Standard	*99.999999999% (11 9s)*
8+3 (OSN Future)	99.999999999999% (14 9s)

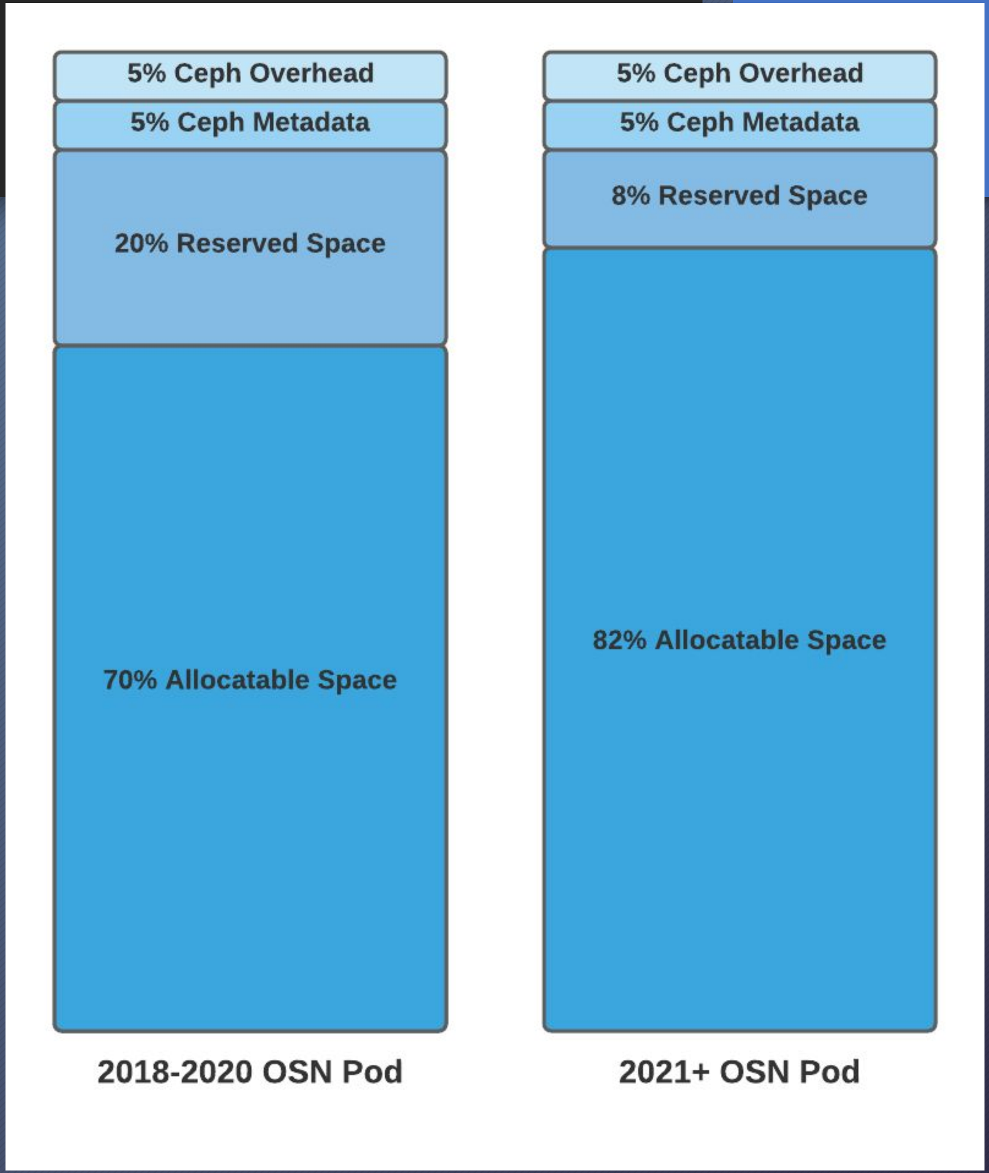
Assumptions:

- 0.44% Annual Failure Rate Reported by Seagate
- Failed disk removed within 1 day of failure

https://github.com/Backblaze/erasure-coding-durability

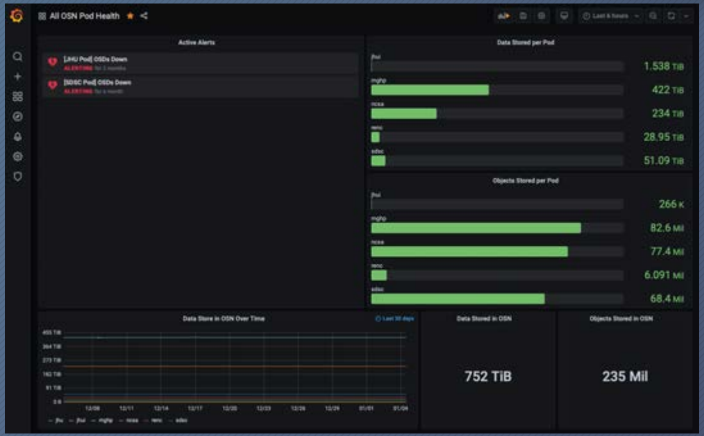
OSN Pod Space Reservations

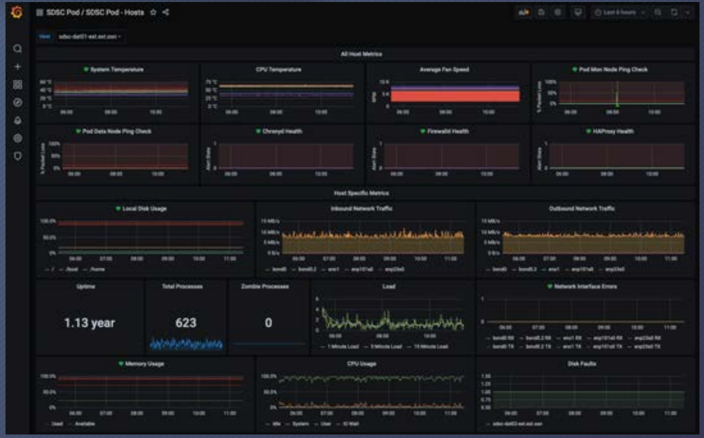
- **Ceph Overhead** - Disks won’t fill past 95%.
- **Ceph Metadata** - Space reserved for information about objects and accounts.
- **Reserved Space** - Space reserved for node failure. This space is only allocated to projects if the project cannot be allocated on another Pod.
- **Allocatable Space** - Space available to projects.

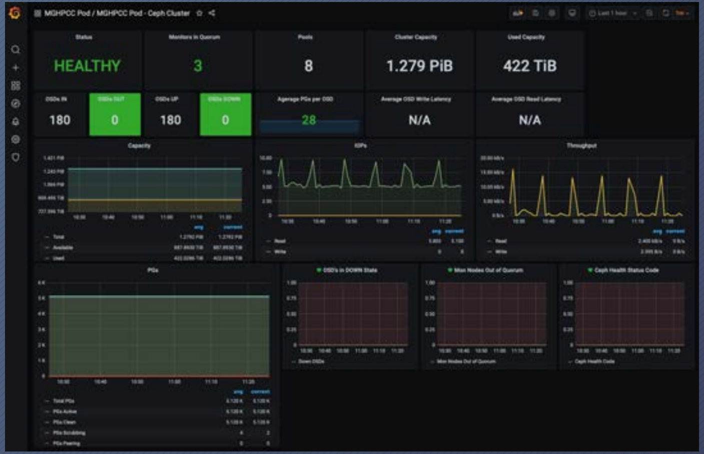


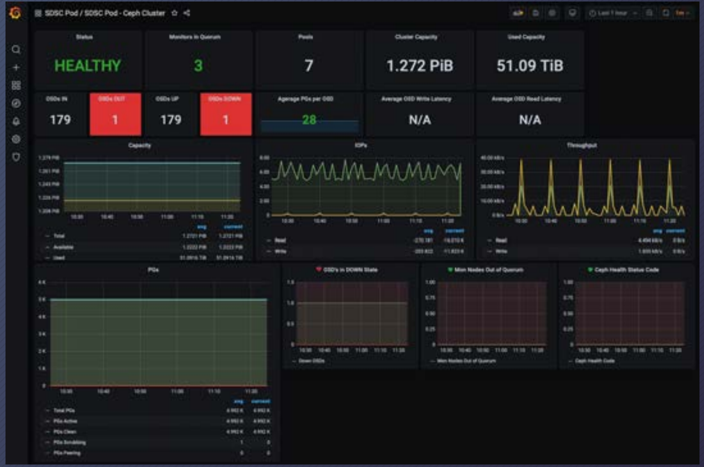
Monitoring

- Ceph statics and health information are sent from the MGR service to Telegraf and visualized using Grafana.
- Alerts are sent to a Slack channel monitored by the implementers team.
- Critical issues are triaged immediately, all other issues are dealt with at the weekly implementers team meeting.


All OSN Pod Health
Active Alerts
[JHU Pod] OSDs Down
[SDSC Pod] OSDs Down
Data Stored per Pod
1.538 TiB
422 TiB
234 TiB
28.95 TiB
51.09 TiB
Objects Stored per Pod
266 K
82.6 Mil
77.4 Mil
6.091 Mil
68.4 Mil
Data Store in OSN Over Time
Data Stored in OSN
752 TiB
Objects Stored in OSN
235 Mil


SDSC Pod / SDSC Pod - Hosts
All Host Metrics
System Temperature
CPU Temperature
Average Fan Speed
Pod Mon Node Ping Check
Pod Data Node Ping Check
Chronyd Health
Firewalld Health
HAProxy Health
Host Specific Metrics
Local Disk Usage
Inbound Network Traffic
Outbound Network Traffic
Uptime
1.13 year
Total Processes
623
Zombie Processes
0
Load
Network Interface Errors
Memory Usage
CPU Usage
Disk Faults


MGHPCC Pod / MGHPCC Pod - Ceph Cluster
Status
HEALTHY
Monitors in Quorum
3
Pools
8
Cluster Capacity
1.279 PiB
Used Capacity
422 TiB
OSDs IN
180
OSDs OUT
0
OSDs UP
180
OSDs DOWN
0
Average PGs per OSD
28
Average OSD Write Latency
N/A
Average OSD Read Latency
N/A
Capacity
IOPs
Throughput
PGs
OSDs in DOWN State
Mon Nodes Out of Quorum
Ceph Health Status Code


SDSC Pod / SDSC Pod - Ceph Cluster
Status
HEALTHY
Monitors in Quorum
3
Pools
7
Cluster Capacity
1.272 PiB
Used Capacity
51.09 TiB
OSDs IN
179
OSDs OUT
1
OSDs UP
179
OSDs DOWN
1
Average PGs per OSD
28
Average OSD Write Latency
N/A
Average OSD Read Latency
N/A
Capacity
IOPs
Throughput
PGs
OSDs in DOWN State
Mon Nodes Out of Quorum
Ceph Health Status Code