Benchmarking of storage solutions in cloud environment: the Cloud@CNAF experience

Costantini Alessandro, Diego Michelotto, Doina Cristina Duma, Antonio Falabella

Workshop sul Calcolo nell'INFN, Paestum, 23-27 Maggio 2022

Overview

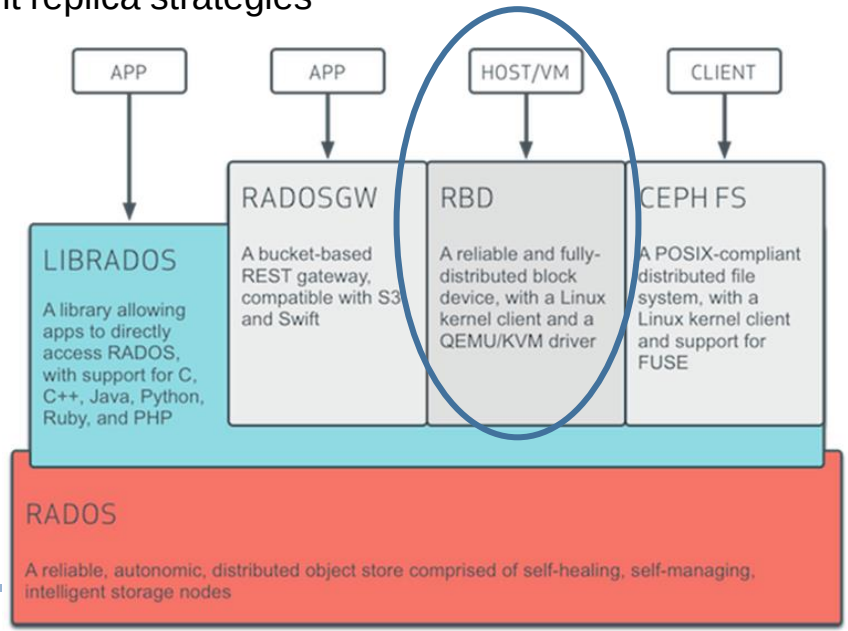
- Motivation
- CEPH in Cloud@CNAF
- Resources and configuration
- Performance and tests
- A cloud application over CEPH: MinIO
- Conclusions & Future activities

Motivation

- Old storage HW to be replaced
- Reduce cost of both HW and SW
- Using open-source solutions
- Scalable
- CEPH as reliable distributed file system
  - Already choose as a storage solution around us
  - Supported by big communities

CEPH Architecture

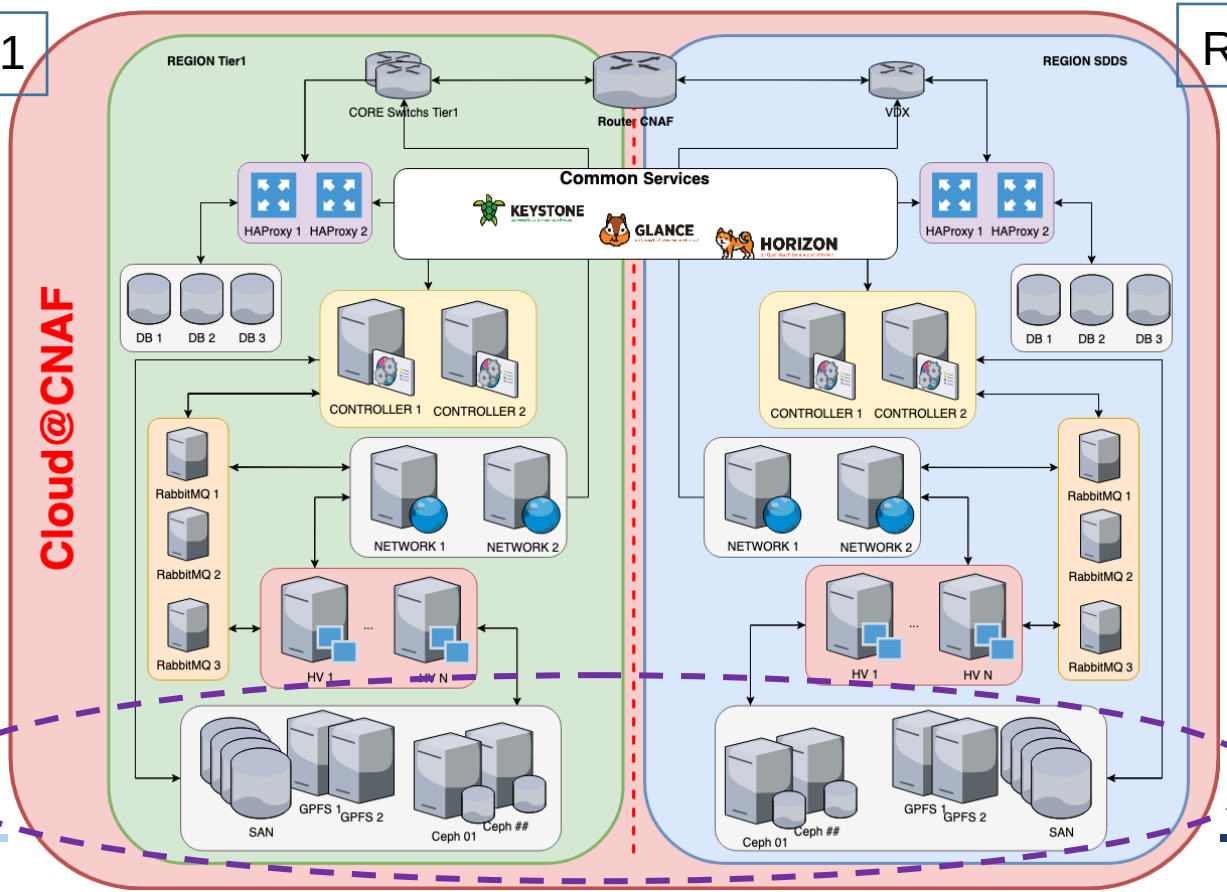
- Open-source software-defined storage platform
- Implements object storage on a single distributed cluster
- Provides interfaces for object-, block- and file-level storage.
- https://docs.ceph.com/
- Tested versions: **15** (Octopus), **16** (Pacific)
- Adopts different replica strategies



Cloud@CNAF Configuration

Region T1

Region SDDS



Storage Backend

Hardware

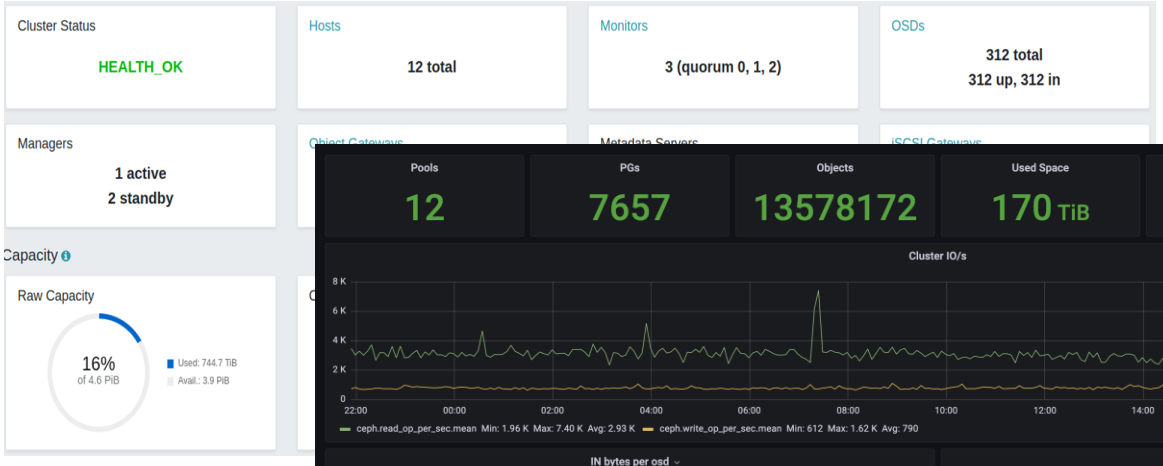
- **Region SDDS**
  - 8 servers
    - **2 x 4TB NVMe osd**
    - 2 x 2TB NVMe for OS (HOT SWAP)
      - Wal e db per osd
    - **24 x 16TB SAS disks** (HOT SWAP)
    - 1 x 2 port 25Gbit/s Ethernet
  - 2 servers
    - **2 x 2TB SSD wal e db per osd**
    - 2 x 240GB SSD for OS (HOT SWAP)
    - **42 x 14TB SAS disks** (HOT SWAP)
    - 1 x 2 port 10Gbit/s Ethernet
  - 2 x switch 48x25GbE + 8X100GbE
  - **3.8 PB RAW**
  - **Rete unica: Public/Cluster**
- **Region T1**
  - 12 servers
    - **2 x NVMe 1920GB**
    - 2 x 1TB for OS (HOT SWAP)
    - **24 x 18TB SAS disks** (HOT SWAP)
    - 1 x 4 port 25Gbit/s Ethernet
  - 2 x switch 48x25GbE + 8X100GbE
  - MGM Switch
  - **5.4 PB raw**
  - **Reti Public/Cluster separate**

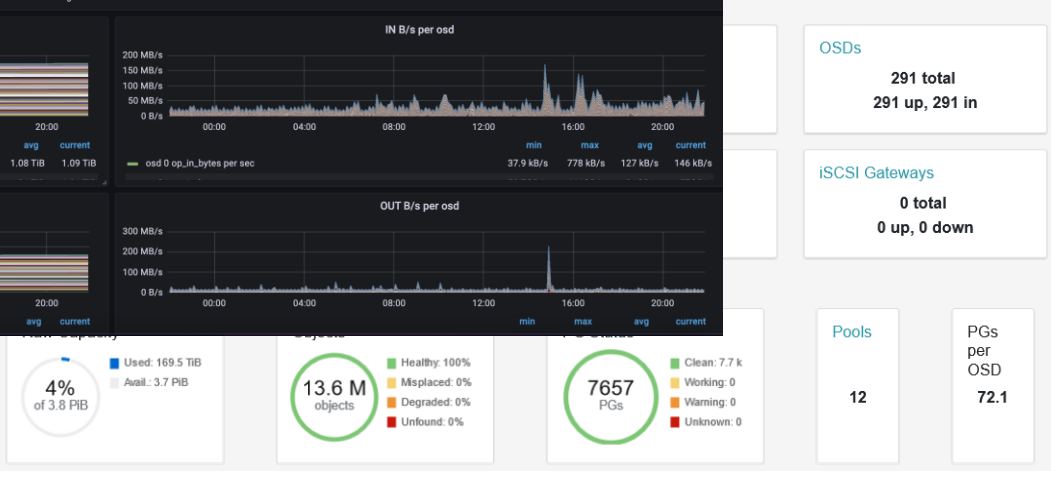
Ceph Configuration in Cloud@CNAF

- Region SDDS
  - Deployment using official CEPH 15 (Octopus) packages
  - 3 monitor nodes
  - 3 manager nodes
  - 291 OSD (over 8+2 nodes)
    - 275 rotational disk (Wal and DB on NVMe)
    - 16 NVMe (+ Wal and DB)
  - Monitoring: **sensu + grafana**
- Region T1
  - Deployment using official CEPH 16 (Pacific) packages
  - 3 monitor nodes
  - 2 manager nodes
  - 312 OSD services
    - 1 OSD per disk
    - Wal and DB on NVMe
  - Monitoring: **prometheus + grafana**

Cluster Monitoring

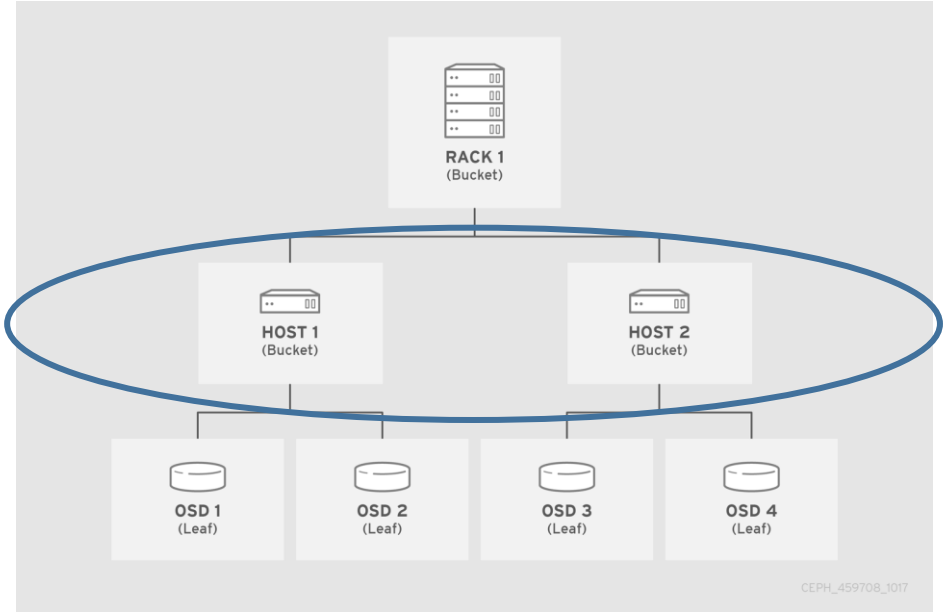
- CEPH dashboard
- Grafana





Service levels for Cloud (Region SDDS)

- R3 HDD (vms, volumes)
- R3 SSD (volumes-ssd)
- EC 6+2 (HDD)
  - volumes-ec (SSD-R3)
  - Volumes-ec-data
- EC 6+2 (SSD)
  - volumes-ec-ssd (SSD-R3)
  - Volumes-ec-ssd-data
- Failure domain: **Host**

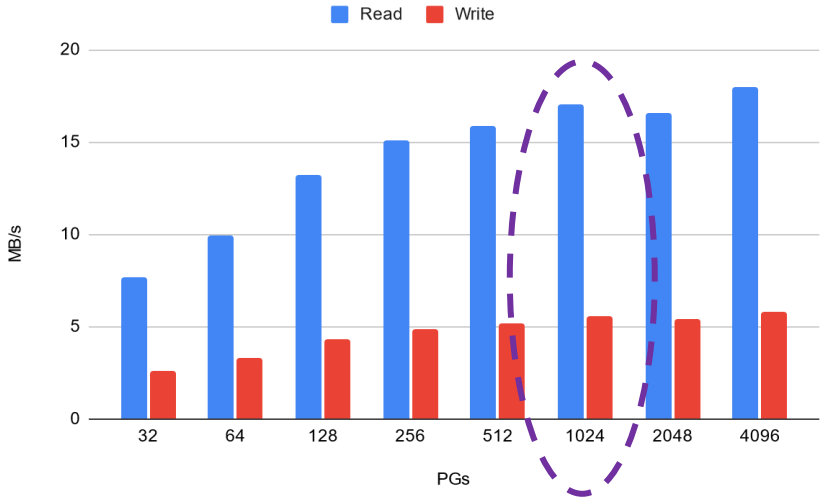


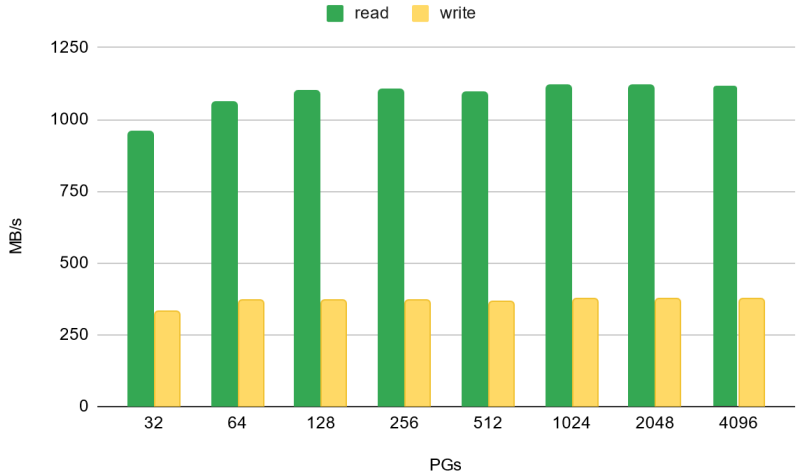
- Benckmark&test with FIO (Flexible I/O tester)
  - random R/W, 75% read
  - BS: 4k e 4M

PG optimization

- R3 HDD (vms, volumes)

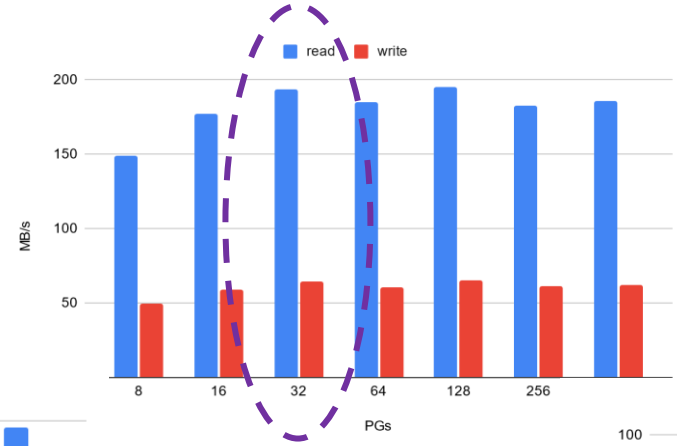
	Block size 4k				Block size 4M			
PG	Read MB/s	Write MB/s	iops read	iops write	Read MB/s	Write MB/s	iops read	iops write
32	7,7	2,6	1874	626	958	328	228	78
64	9,9	3,3	2412	806	1060	367	252	86
128	13,2	4,3	3148	1051	1098	367	261	87
256	15,1	4,9	3611	1206	1104	369	263	88
512	15,9	5,2	3775	1261	1092	365	260	87
1024	17,0	5,6	4052	1353	1117	374	266	89
2048	16,6	5,4	3961	1323	1116	374	266	89
4096	18,0	5,9	4289	1432	1118	374	266	89

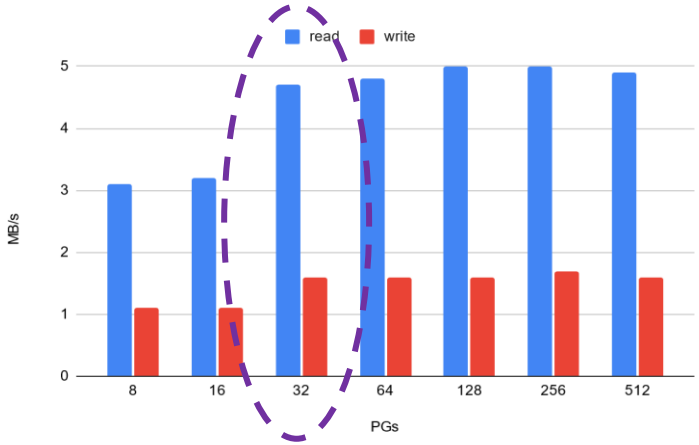




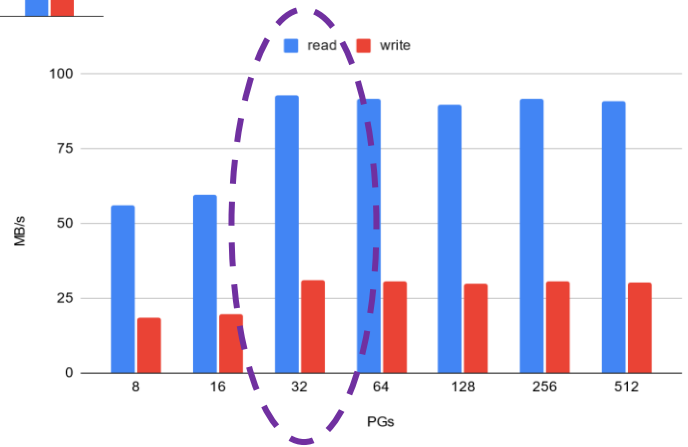
PG optimization

- EC 6+2 (HDD)
  - volumes-ec (SSD-R3)
  - Volumes-ec-data (HDD)





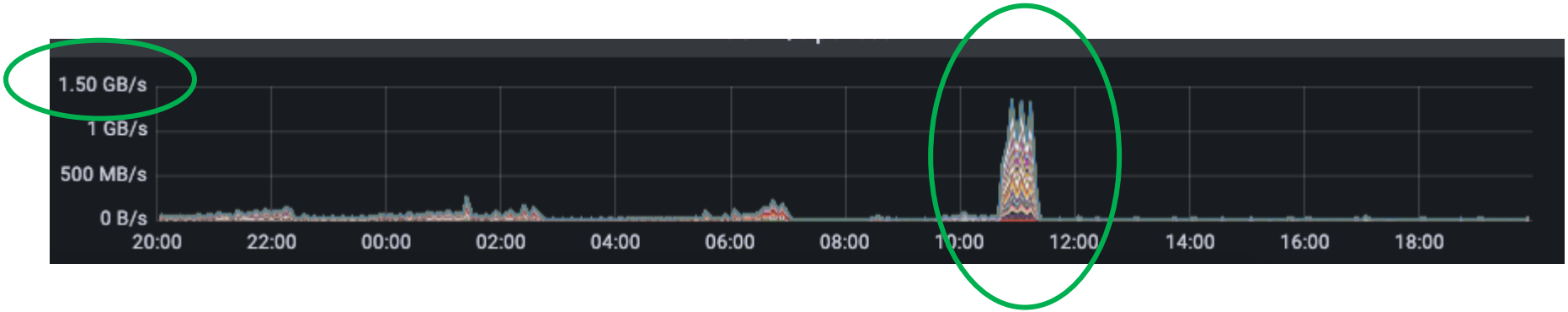
- EC 6+2 (SSD)
  - volumes-ec-ssd (SSD-R3)
  - Volumes-ec-ssd-data (SSD)



Performance with FIO

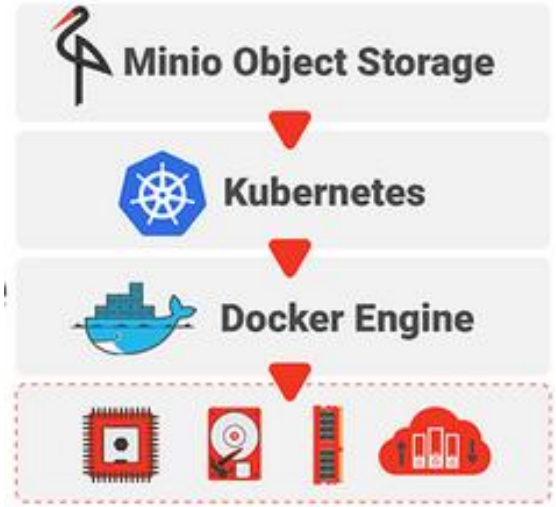
Region	Volume Type	Rand RW mix 75%-25% - BS 4M				Rand RW mix 75%-25% - BS 4K			
		IOPS		BW		IOPS		BW	
		R	W	R MB/s	W MB/s	R	W	R MB/s	W MB/s
SDDS	GPFS	29	11	**116**	**44**	940	110	**3.7**	**1.2**
	LSD/NVME	736	240	**2947**	**962**	66500	22200	**260**	**87**
	LSD/SSD	177	68	**710**	**372**	57000	19000	**224**	**75**
	LSD/HDD	17	6	**70**	**26**	240	80	**0.9**	**0.3**
	Ceph-HDD	266	89	**1118**	**374**	4289	1432	**17.59**	**6.13**
	Ceph-SSD	241	82	**1015**	**347**	47300	15800	**194**	**64.8**
	Ceph-HDD-EC	203	69	**855**	**293**	1142	318	**4.6**	**1.5**
	Ceph-SSD-EC	248	84	**1040**	**356**	22700	7572	**92.8**	**31**
Tier1	GPFS HV 1Gb/s	29	10	**114**	**42**	6500	2200	**25**	**8**
	GPFS HV 10Gb/s	171	55	**684**	**232**	27400	9159	**107**	**37.5**
	Ceph-HDD PG 64	226	73	**906**	**296**	1450	484	**5.8**	**1.9**
	Ceph-HDD PG 1024	227	73	**989**	**293**	8924	2981	**34.9**	**11.6**

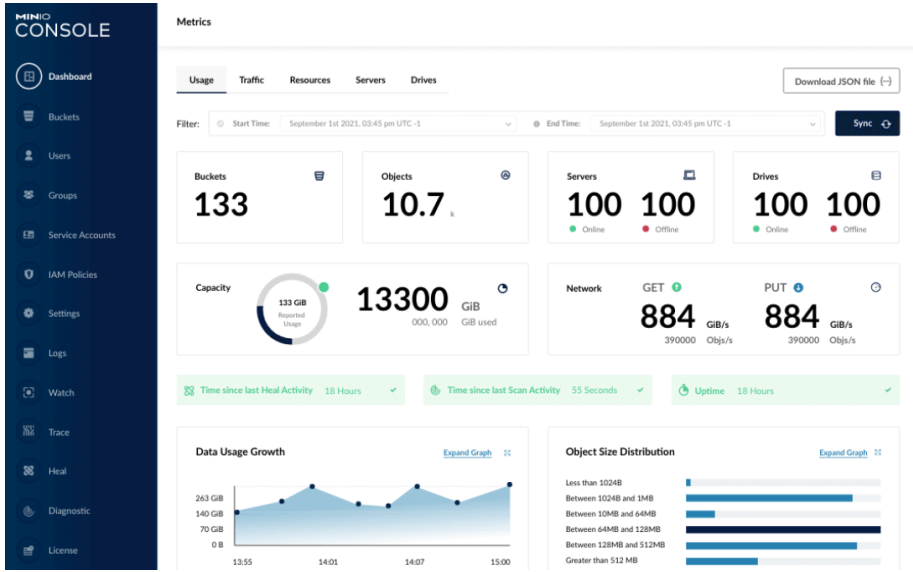
CEPH @ SDDS


1.50 GB/s
1 GB/s
500 MB/s
0 B/s
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00

Running an application: MinIO object storage

- S3 compatible object storage
- Native to Kubernetes
- Open source under GNU AGPL v3





MinIO deployment over Cloud

- **Server**
  - VM (8 vCPUs, 8GB RAM)
  - Volume R3-HDD (200 GB)
  - Volume R3-SSD (200 GB)

- **Client**
  - VM (8 vCPUs, 8GB RAM)
  - WARP S3 benchmarking tool
    - (1, 2, 4, 8, 16 **concurrent jobs**)

https://github.com/minio/warp

R3 HDD

WARP	4K				4M			
	Over Tenant		Intra Tenant		Over Tenant		Intra Tenant	
Concurrence	GET (MiB/s)	PUT (MiB/s)	GET (MiB/s)	PUT (MiB/s)	GET (MiB/s)	PUT (MiB/s)	GET (MiB/s)	PUT (MiB/s)
1	0.32	0.11	0.33	0.11	95.58	31.96	103.45	34.51
2	0.51	0.17	0.50	0.17	143.49	47.79	158.24	52.83
4	0.64	0.21	0.66	0.22	200.32	66.68	215.82	71.88
8	1.08	0.31	1.11	0.37	272.64	90.79	286.59	95.51
16	1.68	0.56	1.68	0.56	356.87	119.03	361.74	120.71

R3 SSD

WARP	4K				4M			
	Over Tenant		Intra Tenant		Over Tenant		Intra Tenant	
concurrence	GET (MiB/s)	PUT (MiB/s)	GET (MiB/s)	PUT (MiB/s)	GET (MiB/s)	PUT (MiB/s)	GET (MiB/s)	PUT (MiB/s)
1	0,59	0,20	0,69	0,23	137,25	46,00	147,00	49,06
2	1,19	0,40	1,32	0,44	244,98	81,68	261,81	87,44
4	2,27	0,76	2,45	0,82	387,99	129,35	426,84	142,44
8	4.52	1,51	5,08	1,70	534,00	178,26	628,82	209,79
16	6,67	2,22	7,33	2,44	567,14	188,89	791,34	263,64

Conclusions

- Test performed in Cloud using different volume types
  - Using CEPH backend
- Improved knowledge how to fine tune the CEPH configuration to accomodate different requirements
- Increase the number of performance tests
  - By leveraging concurrency of multiple clients
- Planning for CEPH upgrade (Quincy, 17)