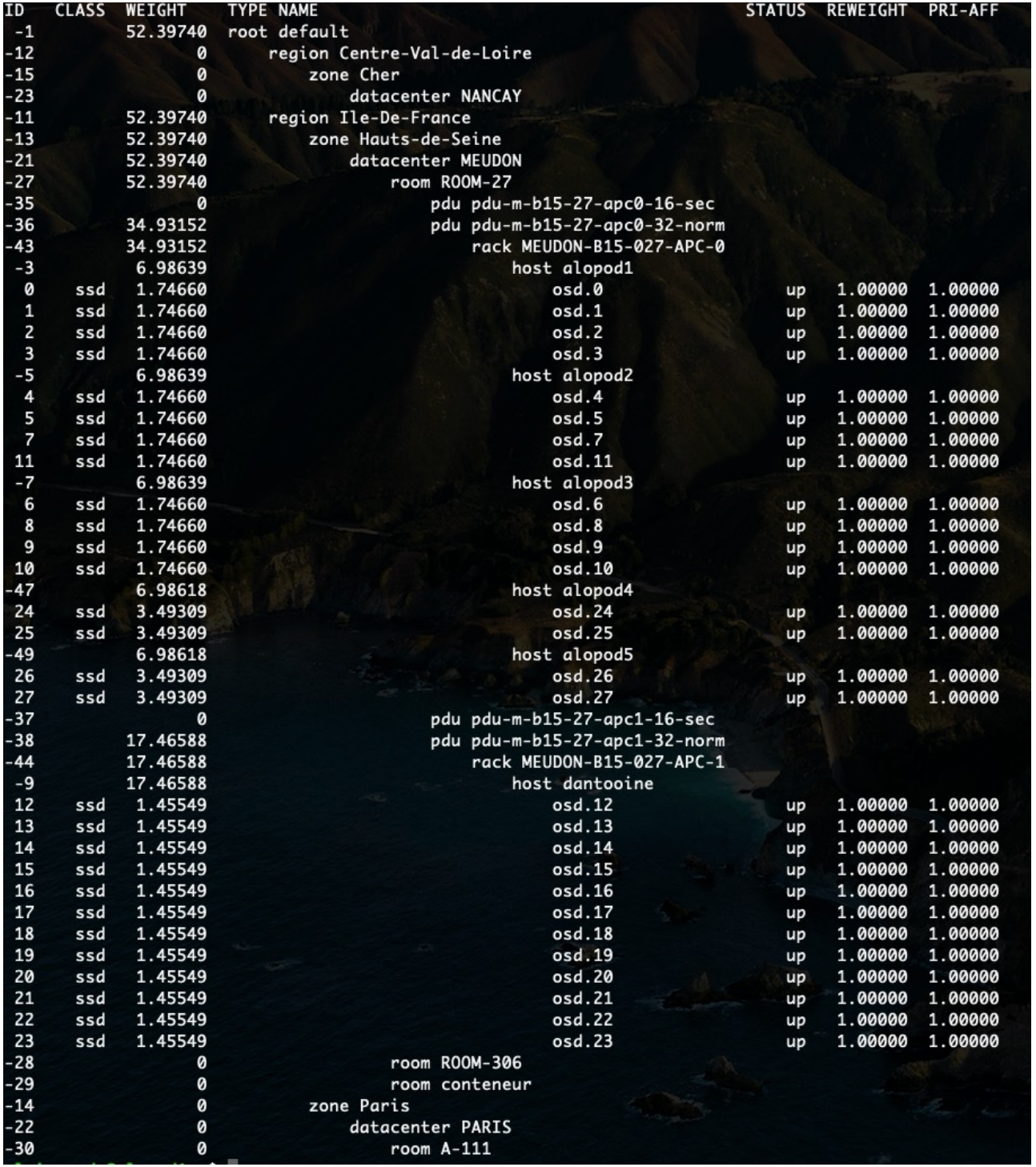
CEPH

Firefly	0.80	May 7, 2014	erasure coding, cache tiering, primary affinity, key/value OSD backend (experimental), standalone radosgw (experimental)
Giant	0.87	October 29, 2014	Redhat buys up Inktank around now
Hamme r	0.94	April 7, 2015	
Infernali s	9.2.0	November 6, 2015	Last release packaged for RHEL/Centos 6.x servers
Jewel	10.2.0	April 21, 2016	Stable CephFS, experimental RADOS backend named BlueStore
Kraken	11.2.0	January 20, 2017	Initial BlueFS support for OSD storage. Multiple active MDS support marked stable
Lumino us	12.2.0	August 29, 2017	BlueFS marked stable and becomes default store for new OSD servers
Mimic	13.2.0	June 1, 2018	CephFS snapshots with multiple MDS servers, RBD image deep-copy
Nautilus	14.2.0	March 19, 2019	Placement-group decreasing, v2 wire protocol, rbd image live-migration between pools, rbd image namespaces for fine-granular access rights
Octopus	15.2.0	March 23, 2020	new deployment tool, Scheduling of snapshots
Pacific	16.2.0	March 2021	

- Installation assez simple depuis la conteneurisation de la solution (*ceph-acile*)
- Utilise docker ou podman
- Très gourmand en ram : 1 disque = 1 docker
- 1 nœud admin + 1 en spare (veille) = 2 dockers
- Tous les nœuds sont moniteurs = *n* dockers
- J'ai ajouté une VM KVM comme moniteur supplémentaire : ca marche !
- Ne pas se louper sur le paramétrage lors de l'installation sinon recommence ! (*ceph-oireux*)
- **Crushmap !!!**
- La crushmap c'est l'équivalent d'une architecture base de données (tables + liens) : ne pas se louper dès le départ, sinon recommence



```
ID   CLASS  WEIGHT    TYPE NAME                                    STATUS  REWEIGHT  PRI-AFF
 -1         52.39740  root default
-12               0      region Centre-Val-de-Loire
-15               0          zone Cher
-23               0              datacenter NANCAY
-11         52.39740      region Ile-De-France
-13         52.39740          zone Hauts-de-Seine
-21         52.39740              datacenter MEUDON
-27         52.39740                  room ROOM-27
-35               0                      pdu pdu-m-b15-27-apc0-16-sec
-36         34.93152                      pdu pdu-m-b15-27-apc0-32-norm
-43         34.93152                          rack MEUDON-B15-027-APC-0
 -3          6.98639                              host alopod1
  0    ssd   1.74660                                  osd.0               up   1.00000  1.00000
  1    ssd   1.74660                                  osd.1               up   1.00000  1.00000
  2    ssd   1.74660                                  osd.2               up   1.00000  1.00000
  3    ssd   1.74660                                  osd.3               up   1.00000  1.00000
 -5          6.98639                              host alopod2
  4    ssd   1.74660                                  osd.4               up   1.00000  1.00000
  5    ssd   1.74660                                  osd.5               up   1.00000  1.00000
  7    ssd   1.74660                                  osd.7               up   1.00000  1.00000
 11    ssd   1.74660                                  osd.11              up   1.00000  1.00000
 -7          6.98639                              host alopod3
  6    ssd   1.74660                                  osd.6               up   1.00000  1.00000
  8    ssd   1.74660                                  osd.8               up   1.00000  1.00000
  9    ssd   1.74660                                  osd.9               up   1.00000  1.00000
 10    ssd   1.74660                                  osd.10              up   1.00000  1.00000
-47          6.98618                              host alopod4
 24    ssd   3.49309                                  osd.24              up   1.00000  1.00000
 25    ssd   3.49309                                  osd.25              up   1.00000  1.00000
-49          6.98618                              host alopod5
 26    ssd   3.49309                                  osd.26              up   1.00000  1.00000
 27    ssd   3.49309                                  osd.27              up   1.00000  1.00000
-37               0                      pdu pdu-m-b15-27-apc1-16-sec
-38         17.46588                      pdu pdu-m-b15-27-apc1-32-norm
-44         17.46588                          rack MEUDON-B15-027-APC-1
 -9         17.46588                              host dantooine
 12    ssd   1.45549                                  osd.12              up   1.00000  1.00000
 13    ssd   1.45549                                  osd.13              up   1.00000  1.00000
 14    ssd   1.45549                                  osd.14              up   1.00000  1.00000
 15    ssd   1.45549                                  osd.15              up   1.00000  1.00000
 16    ssd   1.45549                                  osd.16              up   1.00000  1.00000
 17    ssd   1.45549                                  osd.17              up   1.00000  1.00000
 18    ssd   1.45549                                  osd.18              up   1.00000  1.00000
 19    ssd   1.45549                                  osd.19              up   1.00000  1.00000
 20    ssd   1.45549                                  osd.20              up   1.00000  1.00000
 21    ssd   1.45549                                  osd.21              up   1.00000  1.00000
 22    ssd   1.45549                                  osd.22              up   1.00000  1.00000
 23    ssd   1.45549                                  osd.23              up   1.00000  1.00000
-28               0                  room ROOM-306
-29               0                  room conteneur
-14               0          zone Paris
-22               0              datacenter PARIS
-30               0                  room A-111
```

Cassage :

Perte d'un OSD :

1) Conditions simulées (on le sort du pool/on le remet)

   Ca marche très bien (*ceph-astoche*), peu chronophage en temps pour remettre dans le pool.

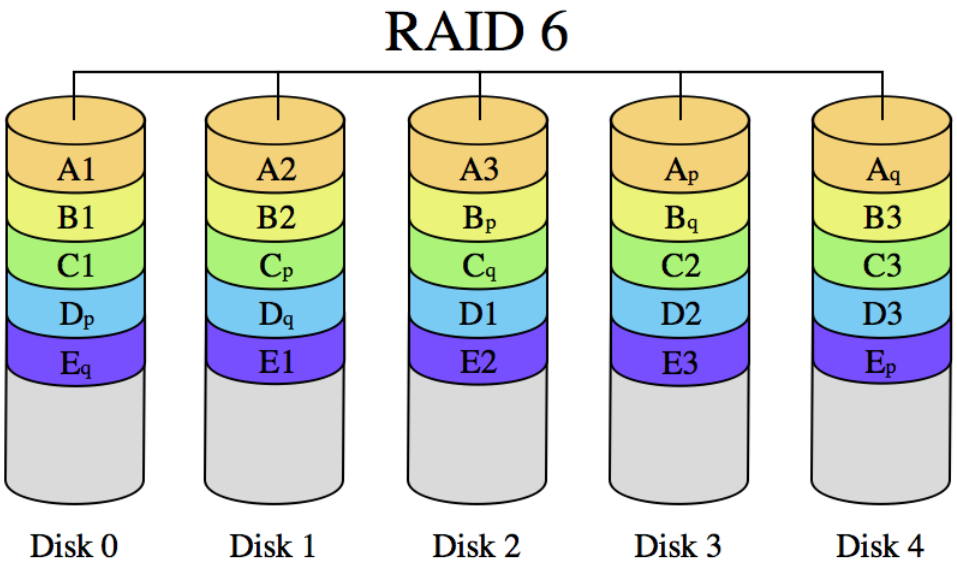
2) Conditions réelles (on l'arrache/on le remet)

   Ca marche très bien en apparence (*ceph-alacieux*), mais très nombreuses manips pour remettre le nouveau disque dans le pool (destruction de l'OSD, du docker, effacement disque, remise dans le pool, recréation du docker, etc.)

Perte d'un serveur entier

Pseudo testé avec 2 machines en même temps, 1h avant de partir en congés : ca à l'air de marcher puisqu'il a reconstruit l'intégralité des données en ~ 6h

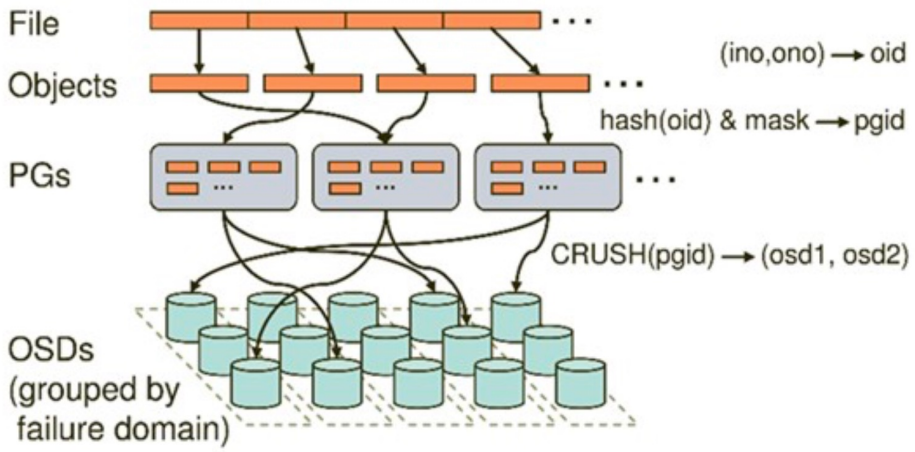
RAID vs CEPH



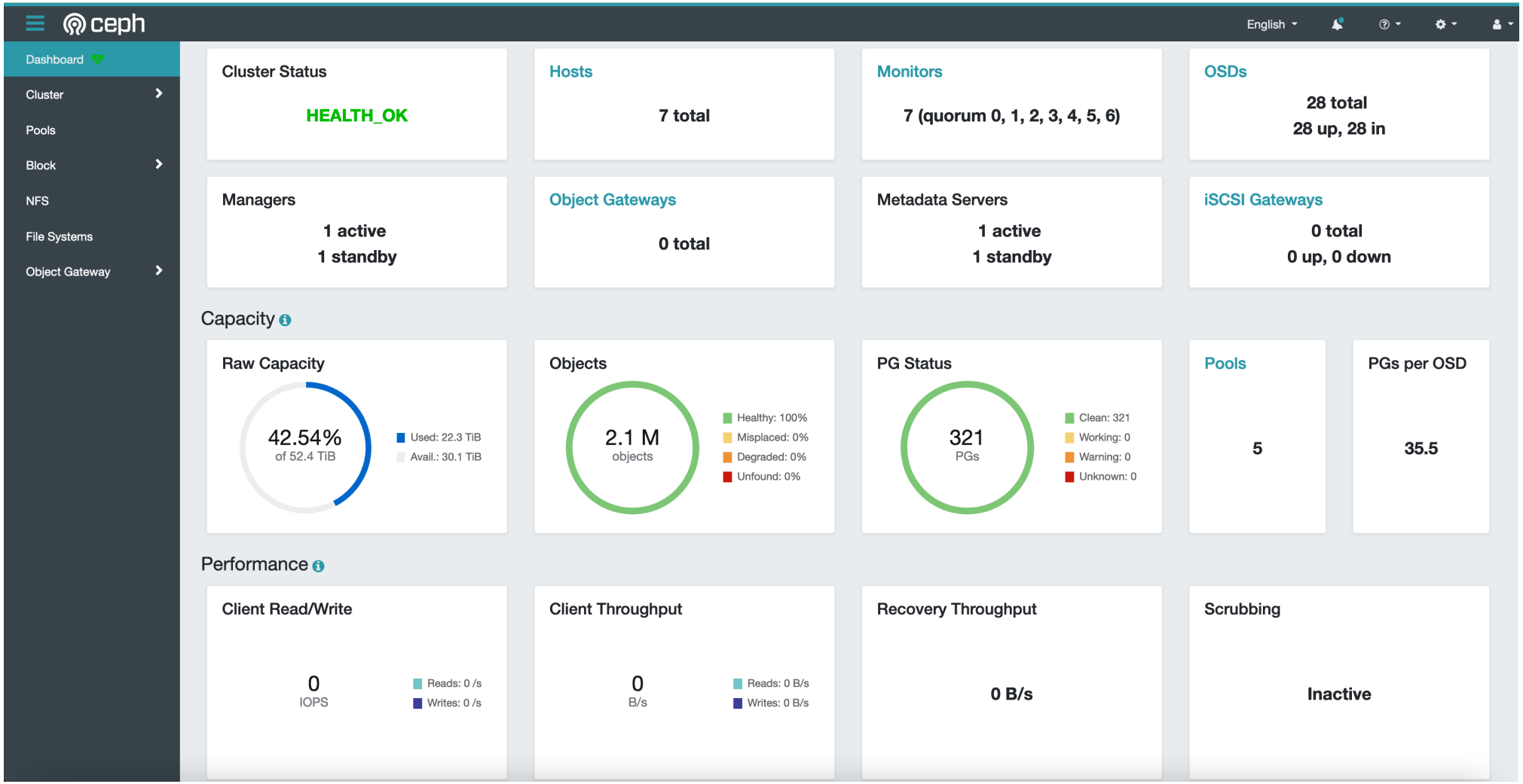
En cas de perte : reconstruit 100% du disque même l'espace vide

En cas de perte : reconstruit uniquement les PG impliqués (on ne sait pas ce qu'il se passe et on ne comprend pas trop, mais CEPH sait, lui …)

CEPH



GUI


ceph
English
Dashboard
Cluster
Pools
Block
NFS
File Systems
Object Gateway
Cluster Status
HEALTH_OK
Hosts
7 total
Monitors
7 (quorum 0, 1, 2, 3, 4, 5, 6)
OSDs
28 total
28 up, 28 in
Managers
1 active
1 standby
Object Gateways
0 total
Metadata Servers
1 active
1 standby
iSCSI Gateways
0 total
0 up, 0 down
Capacity
Raw Capacity
42.54%
of 52.4 TiB
Used: 22.3 TiB
Avail.: 30.1 TiB
Objects
2.1 M
objects
Healthy: 100%
Misplaced: 0%
Degraded: 0%
Unfound: 0%
PG Status
321
PGs
Clean: 321
Working: 0
Warning: 0
Unknown: 0
Pools
5
PGs per OSD
35.5
Performance
Client Read/Write
0
IOPS
Reads: 0 /s
Writes: 0 /s
Client Throughput
0
B/s
Reads: 0 B/s
Writes: 0 B/s
Recovery Throughput
0 B/s
Scrubbing
Inactive

GUI

Cluster » Hosts

Hosts List | Overall Performance

+ Add | 10

	Hostname	Services	Labels	Status	Model	CPUs	Cores	Total Memory	Raw Capacity	HDDs	Flash	NICs
›	alopod1	mds.aifritte.alopod1.mdiipn, mgr.alopod1.zdyjsq, mon.alopod1, osd.0, osd.1, osd.2, osd.3	_admin mon		ProLiant (ProLiant DL385 Gen10 Plus v2)	2	16	251.6 GiB	7.7 TiB	0	5	2
›	alopod2	mds.aifritte.alopod2.fkwqfp, mgr.alopod2.zozcaw, mon.alopod2, osd.11, osd.4, osd.5, osd.7	mon		ProLiant (ProLiant DL385 Gen10 Plus v2)	2	16	251.7 GiB	7.7 TiB	0	5	2
›	alopod3	mon.alopod3, osd.10, osd.6, osd.8, osd.9	mon		ProLiant (ProLiant DL385 Gen10 Plus v2)	2	16	251.7 GiB	7.7 TiB	0	5	2
›	alopod4	mon.alopod4, osd.24, osd.25	mon		ProLiant (ProLiant DL385 Gen10 Plus v2)	2	16	251.7 GiB	7.9 TiB	0	3	2
›	alopod5	mon.alopod5, osd.26, osd.27	mon		ProLiant (ProLiant DL385 Gen10 Plus v2)	2	16	251.7 GiB	7.9 TiB	0	3	2
›	dantooine	mon.dantooine, osd.12, osd.13, osd.14, osd.15, osd.16, osd.17, osd.18, osd.19, osd.20, osd.21, osd.22, osd.23	mon		PowerEdge (PowerEdge R740xd)	2	24	94 GiB	18.2 TiB	1	12	4
›	mon1-alopod	mon.mon1-alopod	mon		(Standard PC (Q35 + ICH9, 2009))	4	1	7.8 GiB	24 GiB	2	0	1

0 selected / 7 total

GUI

Cluster » OSDs

OSDs List | Overall Performance

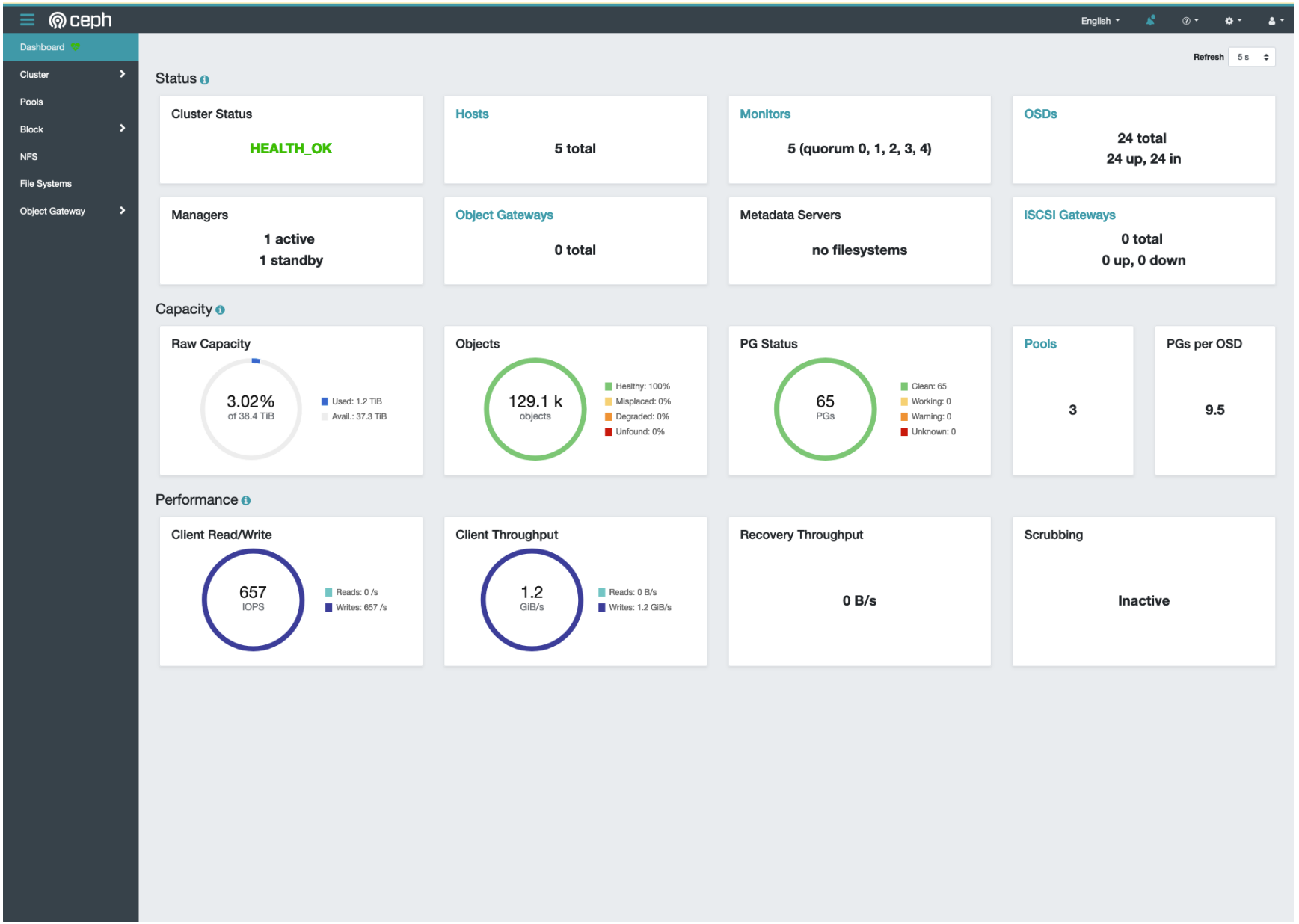
+ Create | Cluster-wide configuration | 10

		ID	Host	Status	Device class	PGs	Size	Flags	Usage	Read bytes	Write bytes	Read ops	Write ops
☐	›	0	alopod1	in up	ssd	43	1.7 TiB		51%			0 /s	0 /s
☐	›	1	alopod1	in up	ssd	44	1.7 TiB		57%			0 /s	0 /s
☐	›	2	alopod1	in up	ssd	43	1.7 TiB		54%			0 /s	0 /s
☐	›	3	alopod1	in up	ssd	33	1.7 TiB		46%			0 /s	0 /s
☐	›	4	alopod2	in up	ssd	41	1.7 TiB		51%			0 /s	0 /s
☐	›	5	alopod2	in up	ssd	33	1.7 TiB		37%			0 /s	0 /s
☐	›	6	alopod3	in up	ssd	39	1.7 TiB		51%			0 /s	0 /s
☐	›	7	alopod2	in up	ssd	31	1.7 TiB		41%			0 /s	0 /s
☐	›	8	alopod3	in up	ssd	35	1.7 TiB		43%			0 /s	0 /s
☐	›	9	alopod3	in up	ssd	34	1.7 TiB		46%			0 /s	0 /s

0 selected / 28 total

‹ « 1 2 3 » ›

Performances



Performances