

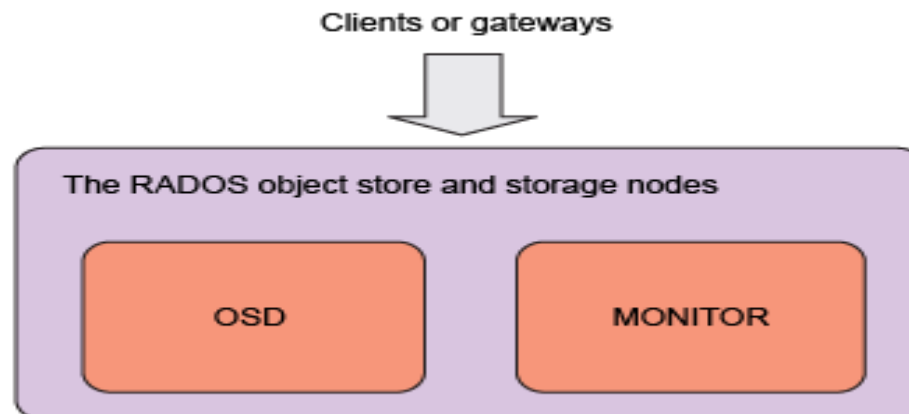


INFN CNAF

Matteo Favaro

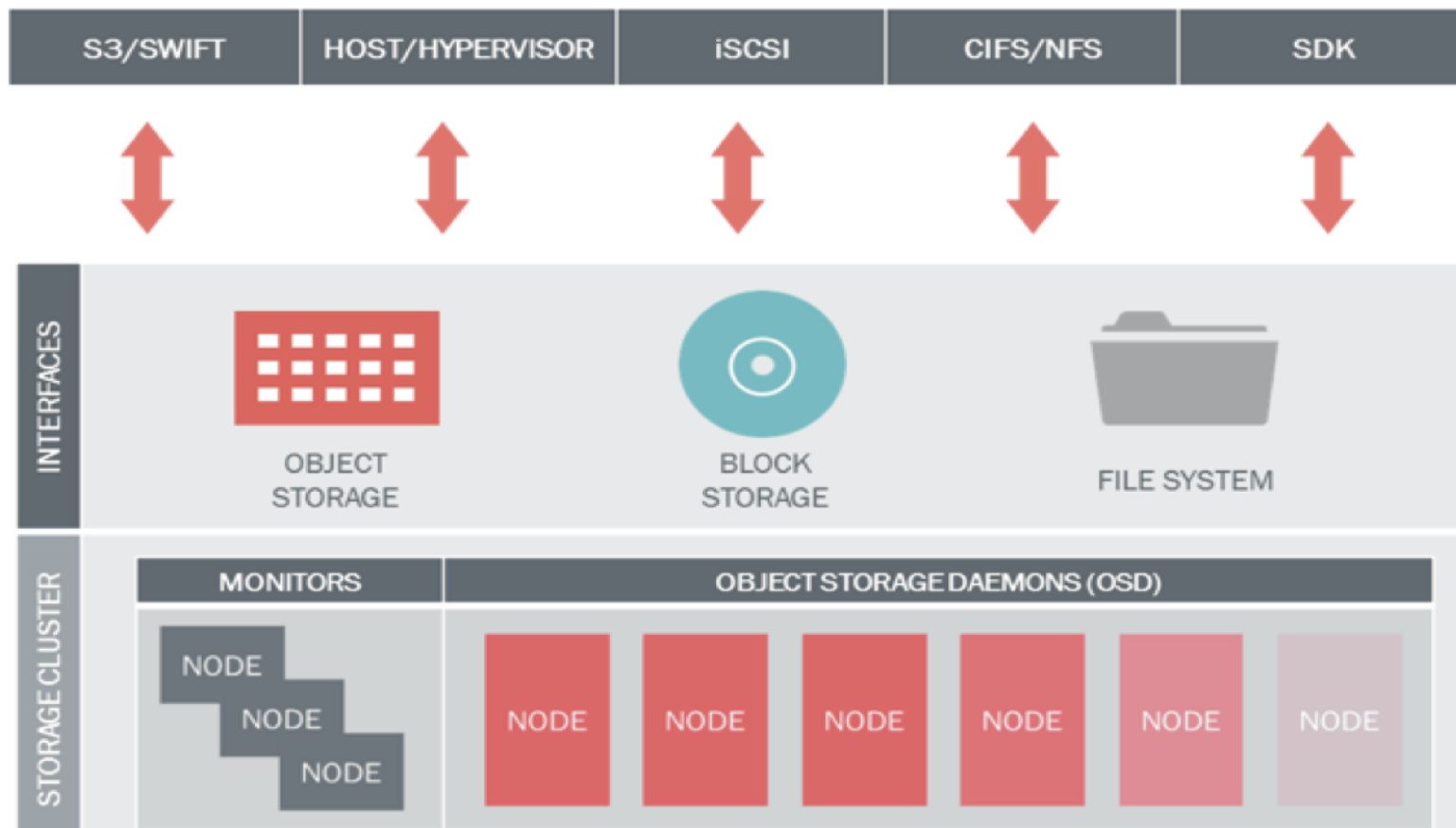
What is Ceph?

- Ceph is a distributed object store and file system designed to provide performance, reliability and scalability. It is open source and freely-available
- What does it provide?
 - Object Storage: access to the RADOS object-based storage system (RADOS = OSD + MON + MDS)
 - Librados: native api library
 - REST Gateway: cloud storage interface (S3 / openStack)



How is it accessed?

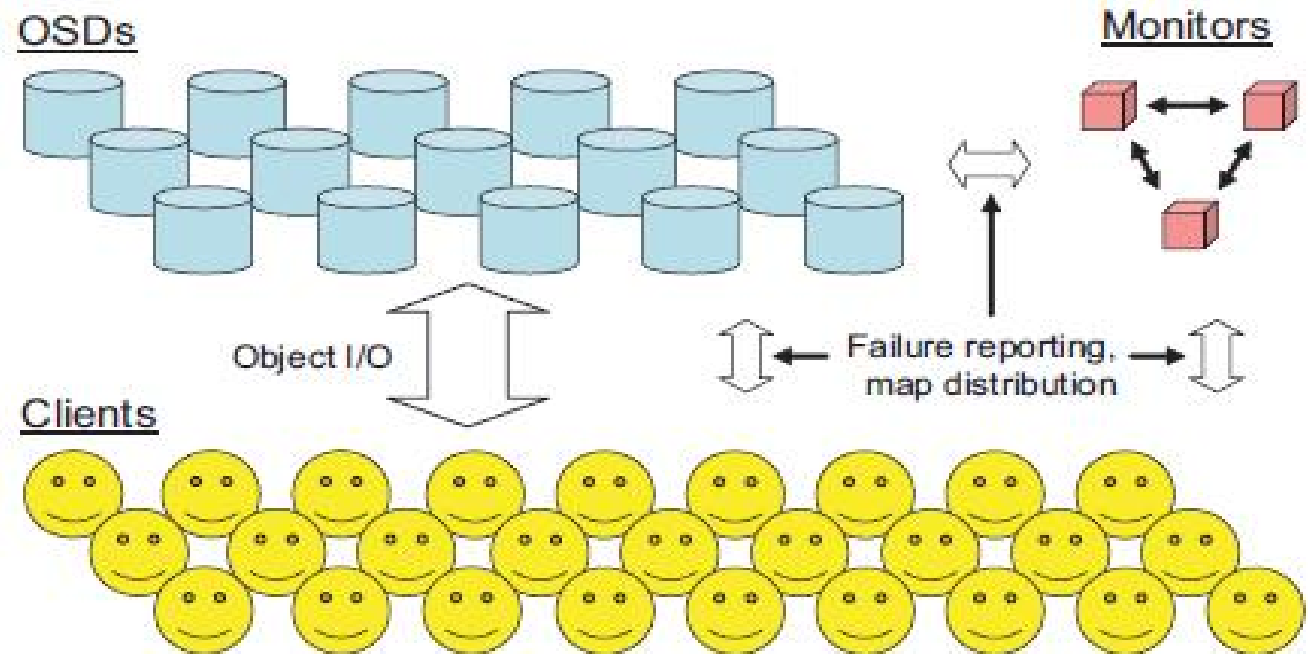
- native binding or RESTful APIs
- mount Ceph as a thinly provisioned block device
- Mount Ceph FS file system



Ceph system 1/2

Ceph has 3 base daemons that make it works

- Monitor
- OSD



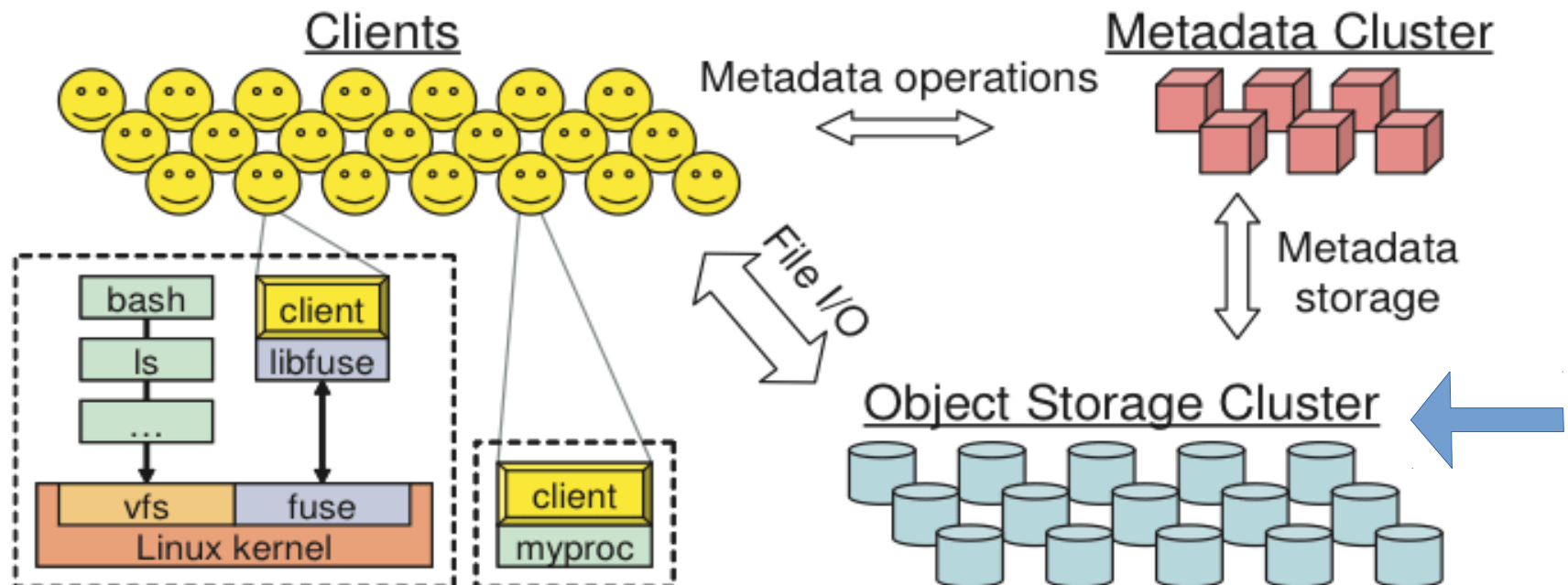
Ceph system 2/2

And...

- MDS

Kernel mount
only from kernel
version 3.0

Metadata
operation: cd,
ls, find



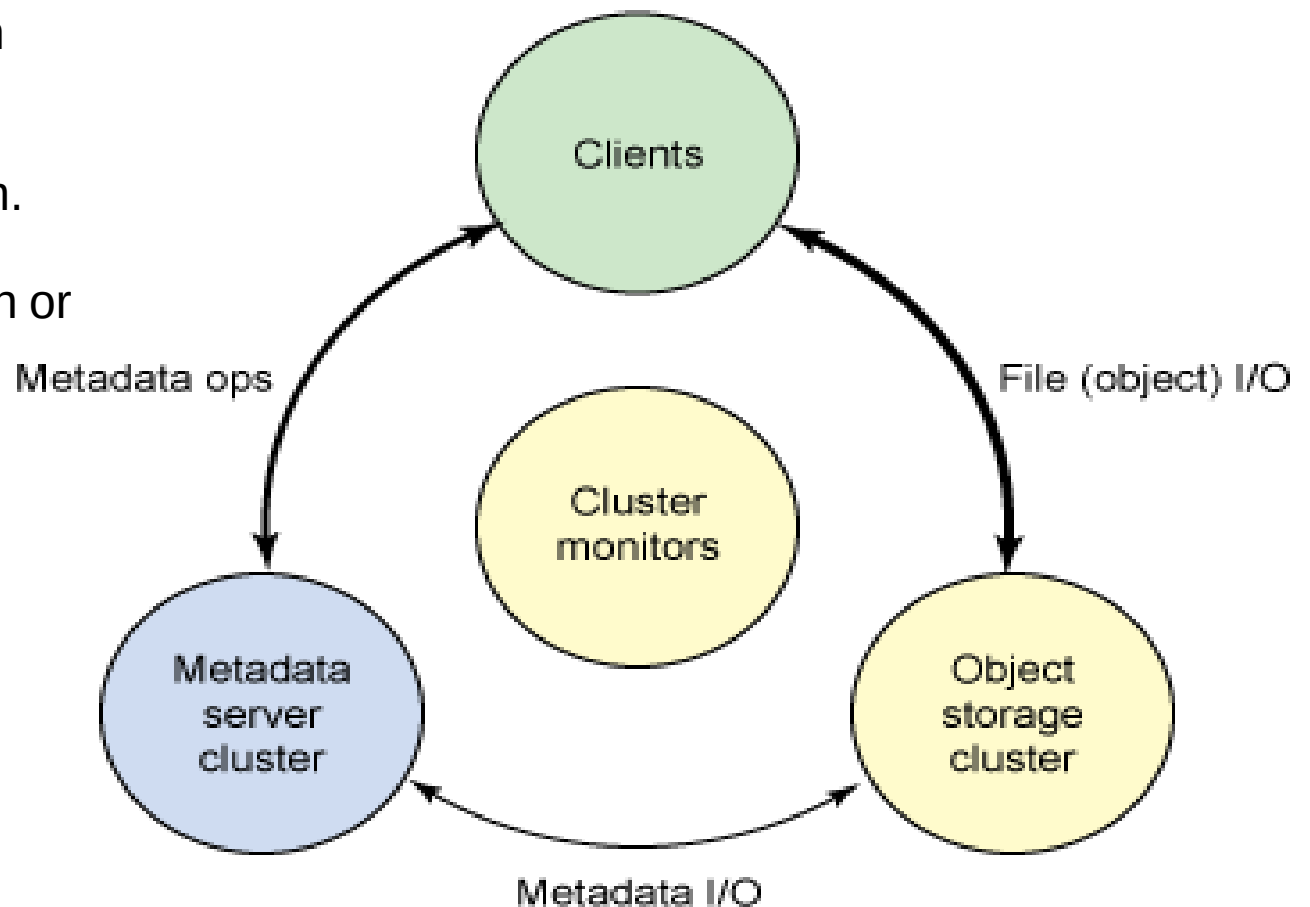
Ceph: Monitor

A Ceph Monitor maintains maps of the cluster state.

Monitor is the heart of a ceph cluster.

Monitor works with a quorum.

Monitor decides if an osd is in or out of the cluster.



No monitor no party...

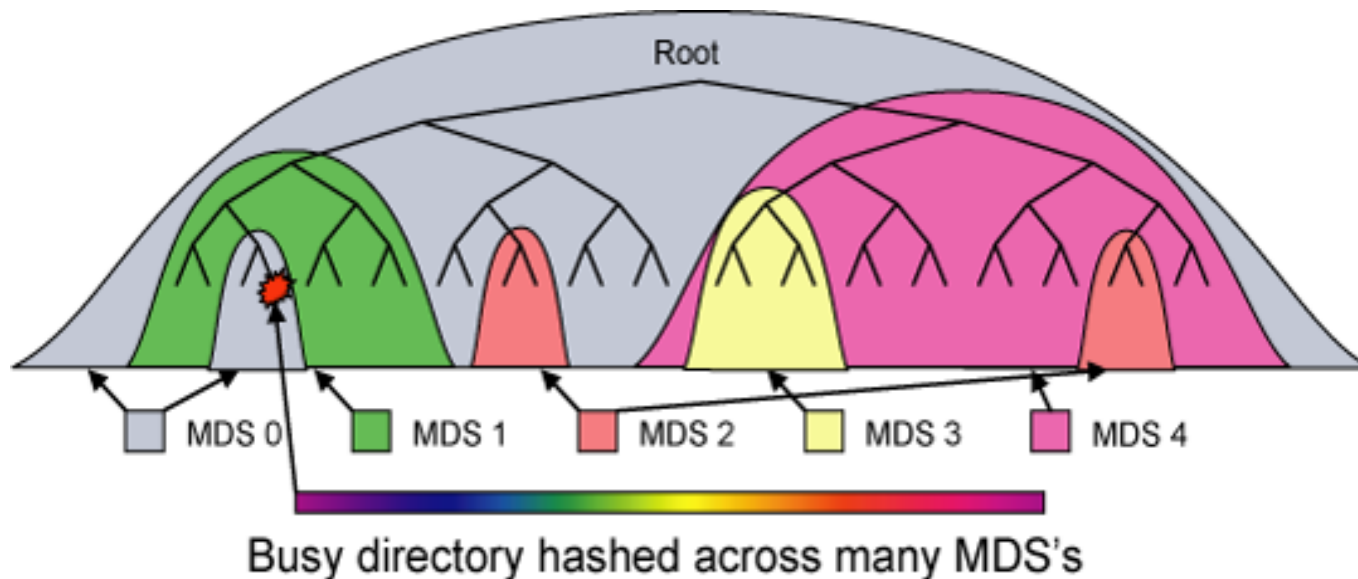
Ceph Maps

- the OSD map: how is deployed the OSDs
- Placement Group (PG): how the data is divided into the OSDs
- CRUSH map: define where put the data and gives priority to objects
- Monitor map: how the monitors are deployed
- MDS map: the structure of the tree of CEPHFS
- Ceph maintains a history (called an “epoch”) of each state change in the Ceph Monitors, Ceph OSD Daemons, and PGs.

```
[root@ds-07-01 ~]# ceph -w
cluster 1f0041f1-93d8-4da7-a859-8d7ae0531e4c
health HEALTH_OK
monmap e5: 3 mons at {ds-07-01=131.154.129.182:6789/0,ds-07-02=131.154.129.183:6789/0,ds-07-03=131.154.129.184:6789/0}, election epoch 20, quorum 0,1,2 ds-07-01,ds-07-02,ds-07-03
mdsmap e42728: 1/1/1 up {0=ds-07-05=up:active}
osdmap e247: 10 osds: 10 up, 10 in
pgmap v128406: 1536 pgs, 4 pools, 10366 GB data, 2649 kobjects
                20755 GB used, 52596 GB / 73352 GB avail
                1536 active+clean
```

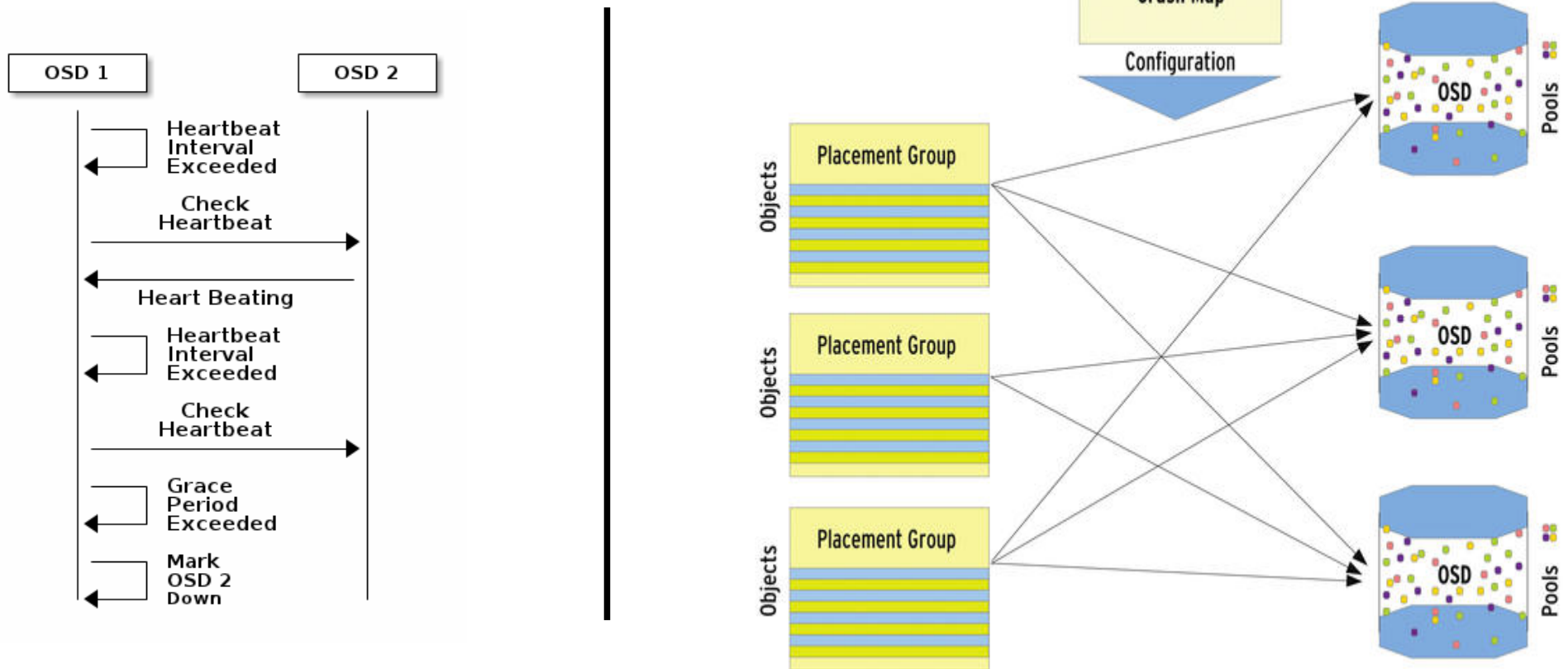
Ceph: MDSs

- A Ceph Metadata Server (MDS) stores metadata on behalf of the Ceph Filesystem (i.e., Ceph Block Devices and Ceph Object Storage do not use MDS). The MDS daemon permits to access CEPH FS via posix.

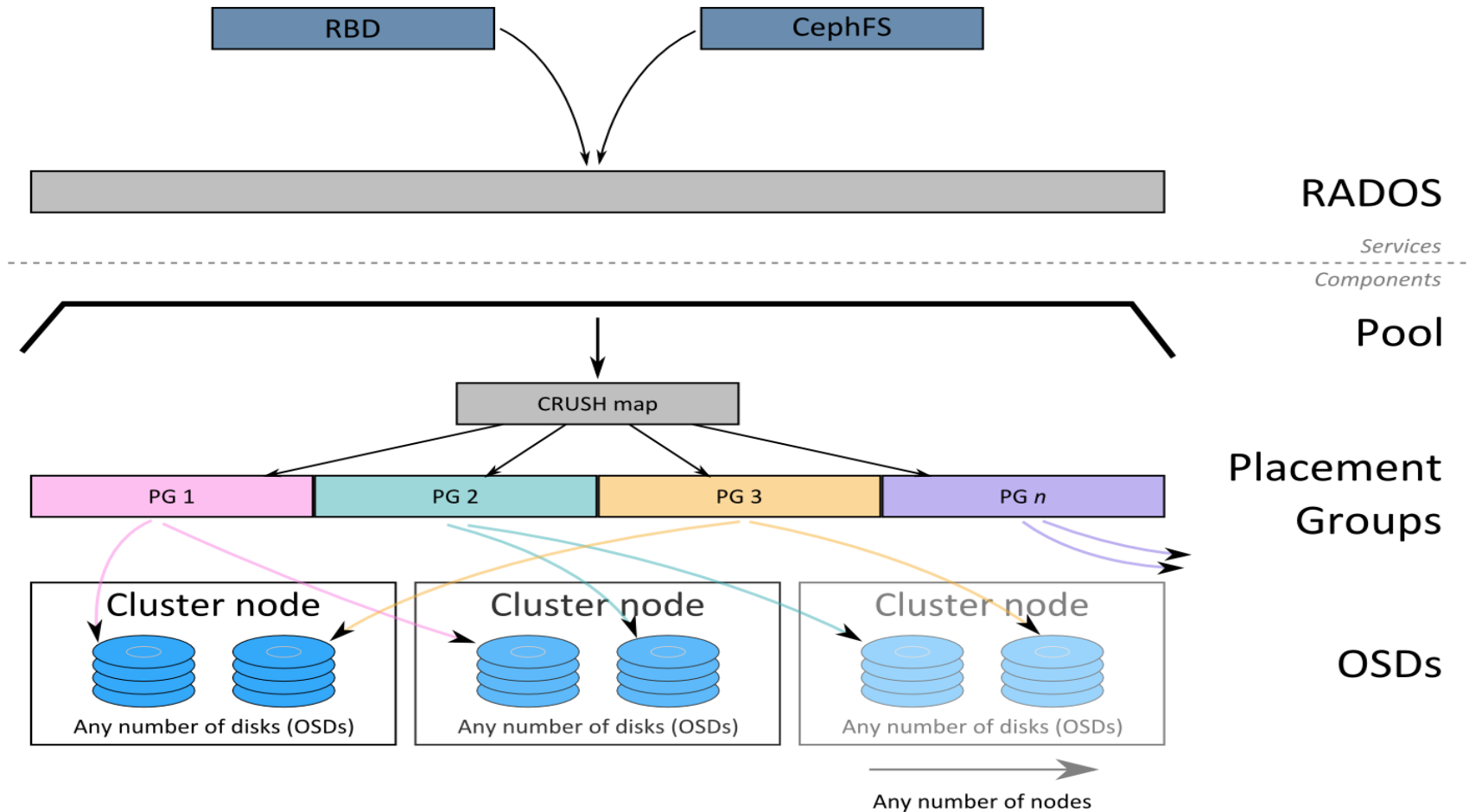


Ceph: OSD

- A Ceph OSD Daemon (Ceph OSD) stores data, handles data replication, recovery, backfilling, rebalancing, and provides information to Ceph Monitors by checking other Ceph OSD Daemons for a heartbeat.

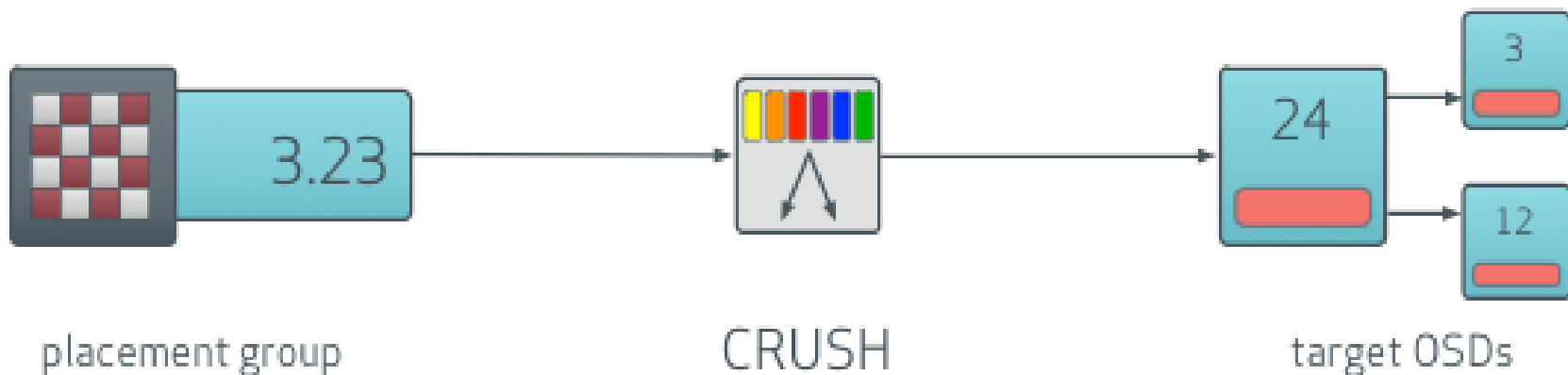
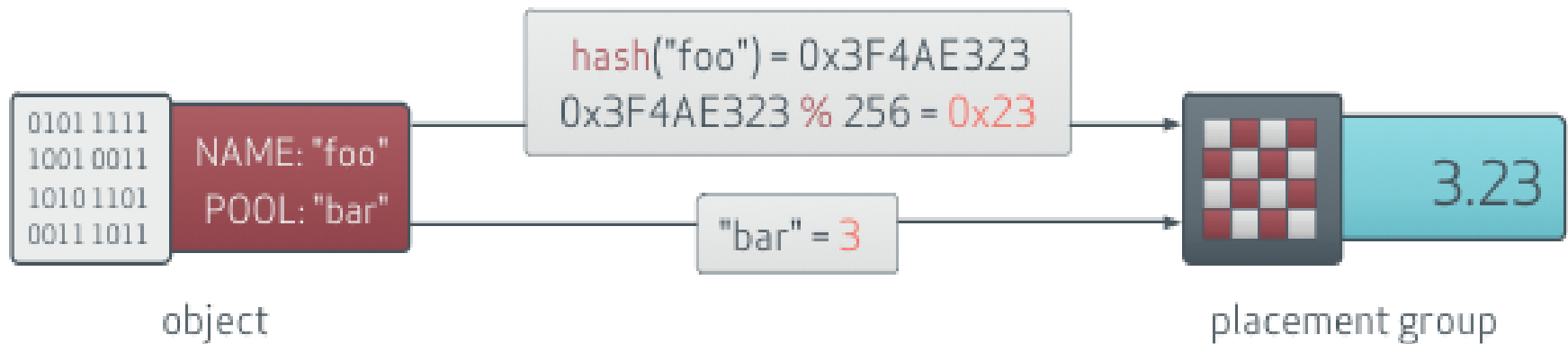


OSD vs pgs vs Pool



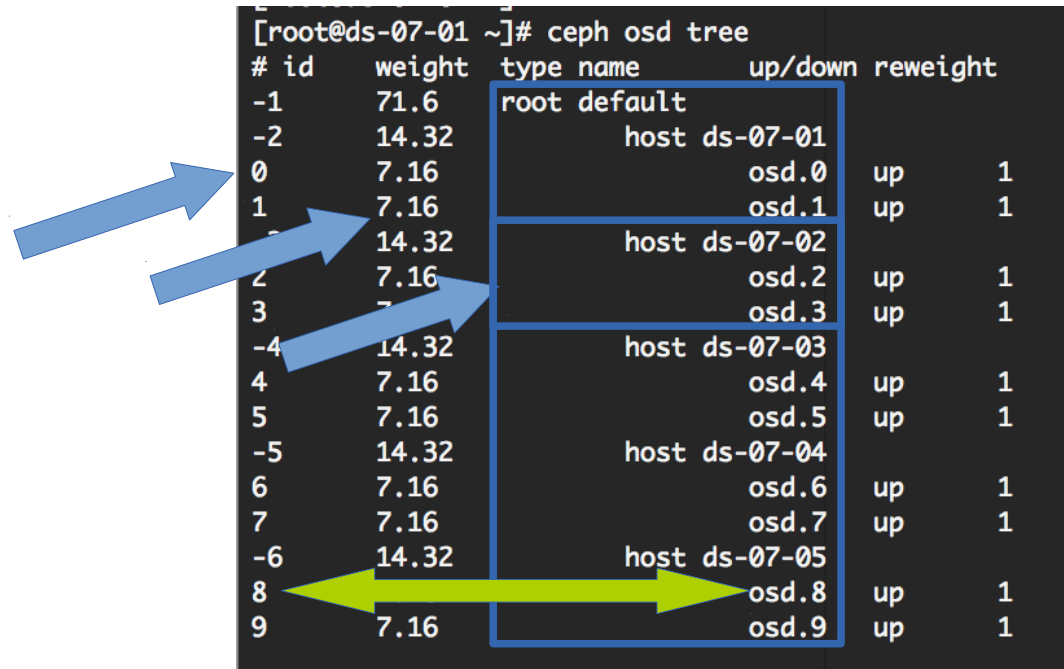
- A **pool** is like a partition
- The **Pgs** is a stripe of the data
- The **osd** is the container

How the "placement group position" is determined



Ceph common query and information

- Quick view osd distribution



The screenshot shows the output of the 'ceph osd tree' command. A blue box highlights the hierarchical structure of the OSD tree, showing the root, hosts, and individual OSDs. Three blue arrows point from the left towards the OSDs on host ds-07-01 (osd.0, osd.1), host ds-07-02 (osd.2, osd.3), and host ds-07-03 (osd.4, osd.5). A green double-headed arrow points from the left towards the OSDs on host ds-07-05 (osd.8, osd.9).

#	id	weight	type	name	up/down	reweight
-1		71.6	root	default		
-2		14.32	host	ds-07-01		
0		7.16		osd.0	up	1
1		7.16		osd.1	up	1
-3		14.32	host	ds-07-02		
2		7.16		osd.2	up	1
3		7.16		osd.3	up	1
-4		14.32	host	ds-07-03		
4		7.16		osd.4	up	1
5		7.16		osd.5	up	1
-5		14.32	host	ds-07-04		
6		7.16		osd.6	up	1
7		7.16		osd.7	up	1
-6		14.32	host	ds-07-05		
8		7.16		osd.8	up	1
9		7.16		osd.9	up	1

- Overall cluster information and health


```
[root@ds-07-01 ~]# ceph -w
cluster 1f0041f1-93d8-4da7-a859-8d7ae0531e4c
health HEALTH_OK
monmap e5: 3 mons at {ds-07-01=131.154.129.182:6789/0,ds-07-02=131.154.129.183:6789/0,ds-07-03=131.154.129.184:6789/0}, election epoch 20, quorum 0,1,2 ds-07-01,ds-07-02,ds-07-03
mdsmap e42728: 1/1/1 up {0=ds-07-05=up:active}
osdmap e247: 10 osds: 10 up, 10 in
pgmap v140294: 1536 pgs, 4 pools, 10366 GB data, 2649 kobjects
20755 GB used, 52596 GB / 73352 GB avail
1536 active+clean
```

Ceph common query and information



- Directory to look for: “/var/lib/ceph”
- Configuration File: “/etc/ceph/ceph.conf”
 - The ceph configuration file is:
 - Per client file
 - Used for configuration and building
 - Is sectioned

Filesystem space availability

- When we use the replica we have a usable space equal to $\text{totalspace} / \text{num_of_replica}$
 - We can view this with CephFS



```
[root@ds-07-06 ~]# df -h
Filesystem                                Size  Used Avail Use% Mounted on
/dev/sda3                                126G   2,8G  117G   3% /
tmpfs                                     7,9G     0   7,9G   0% /dev/shm
/dev/sda1                                 504M   141M  338M  30% /boot
131.154.129.182,131.154.129.183,131.154.129.184:/ 72T   21T   52T  29% /media/ceph
/dev/rbd0                                 9,4G    7,4G  1,4G  86% /mnt/testcephrbd
[root@ds-07-06 ~]#
```



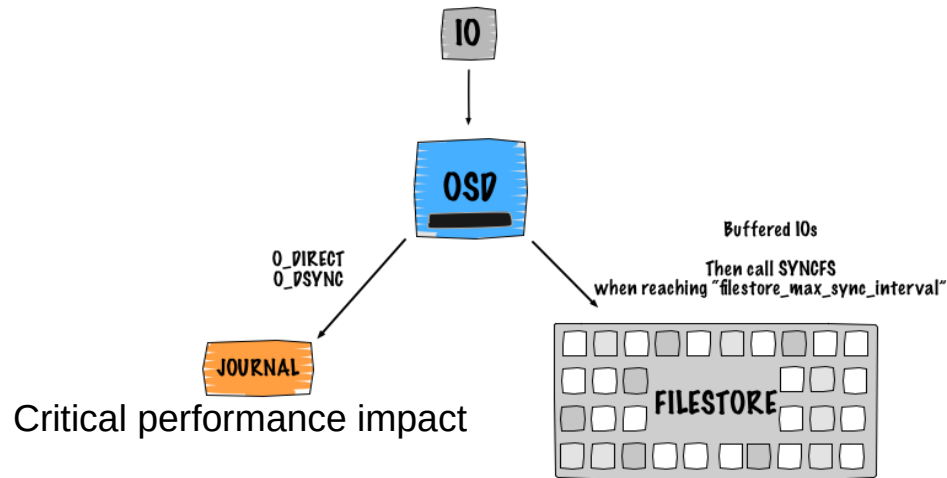
```
[root@ds-07-06 ~]# cd /media/ceph/
[root@ds-07-06 ceph]# pwd
/media/ceph
[root@ds-07-06 ceph]# du -sh
11T
[root@ds-07-06 ceph]#
```

Server replica configuration:

```
[root@ds-07-01 ~]# ceph osd pool get data size
size: 2
[root@ds-07-01 ~]#
```

Ceph critical point /1

- Journal mechanism



- OSD filesystem

```
[root@ds-07-01 ceph-0]# pwd
/var/lib/ceph/osd/ceph-0
[root@ds-07-01 ceph-0]# ll
total 48
-rw-r--r--  1 root root   37 Apr 18 14:01 ceph_fsid
drwxr-xr-x 291 root root 12288 May 15 10:21 current
-rw-r--r--  1 root root   37 Apr 18 14:01 fsid
lrwxrwxrwx  1 root root    9 Apr 30 16:51 journal -> /dev/ram0
lrwxrwxrwx  1 root root   15 Apr 18 14:02 journal.emc -> /dev/emcpower1
-rw-----  1 root root   56 Apr 18 14:01 keyring
-rw-r--r--  1 root root   21 Apr 18 14:01 magic
-rw-r--r--  1 root root    6 Apr 18 14:01 ready
-rw-r--r--  1 root root    4 Apr 18 14:01 store_version
-rw-r--r--  1 root root   42 Apr 18 14:01 superblock
-rw-r--r--  1 root root    2 Apr 18 14:01 whoami
[root@ds-07-01 ceph-0]#
```

Ceph critical point /2

- Multiple mds -> fixed last release (10 days ago)
- SL6 has kernel 2.6.32, CephFS requires 2.6.34 for cephFS
- Deployment -> ceph-deploy strange behaviour on deploy
 - With a manual deployment all work fine
- Raccomended minimum replica is 2 or 3 without replica the filesystem is extremely sensitive
 - Clock screw between osds and monitor
 - Recover osd (daemon died) bring always a lot of problems
 - A Ceph Storage Cluster requires at least two Ceph OSD Daemons to achieve an active + clean state when the cluster makes two copies of your data

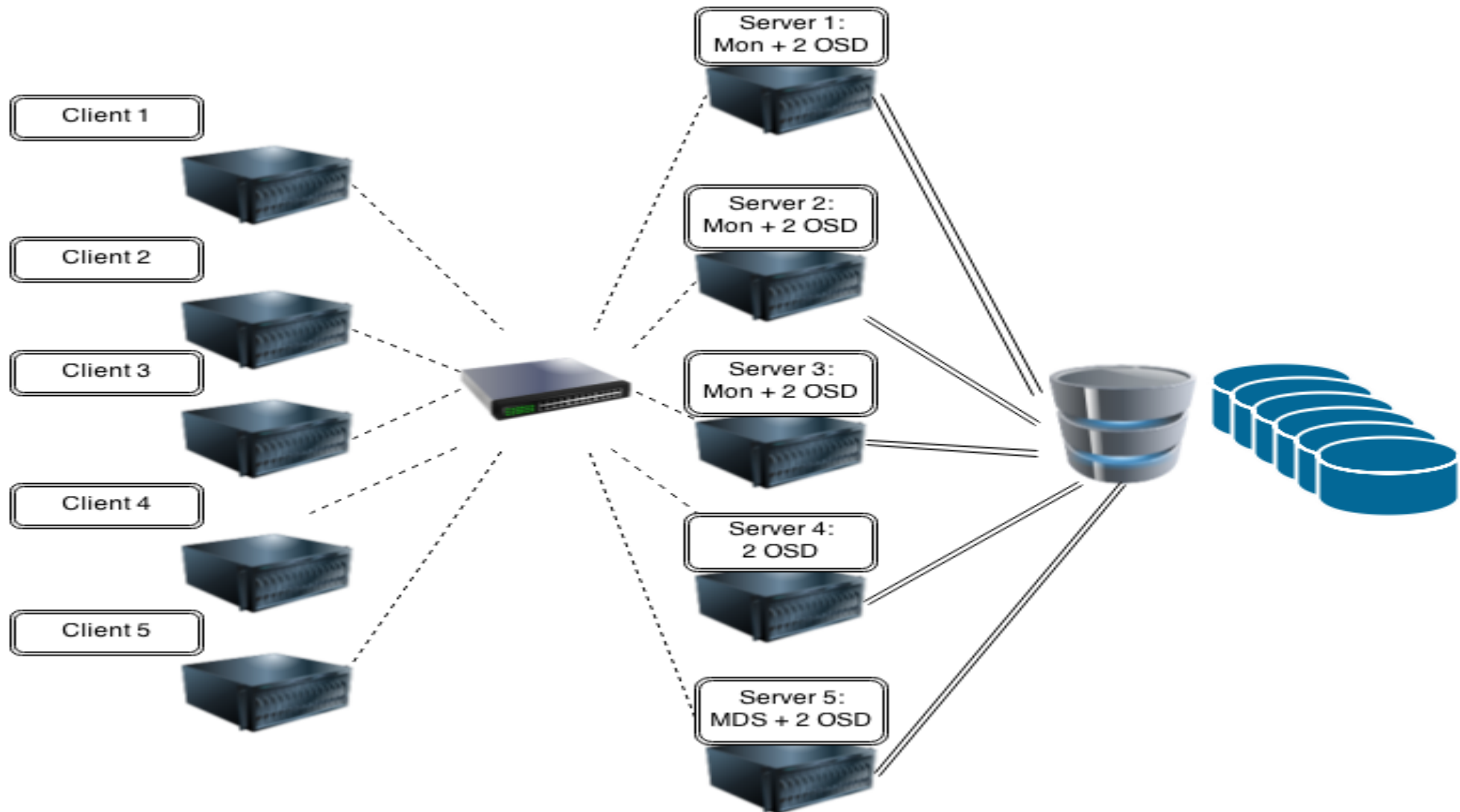
How to build a cluster

- Few steps:
 - Creating keys for cluster auth and start ONE initial Monitor
 - Add osds to cluster (here there is some preparation on the server in order to easily start the OSD daemon)
 - The ceph cluster is up and running
- In a second moment we can add OSDs or Monitors or MDSs
- The cluster automatically rebalance the data and increase the available space

Our test Environment

- Machines: 5 server and 5 client
- Every server (at the end) mount 2 disks and export as OSD
- Every server run a monitor daemon,
- One server run the MDS daemon
- The Machines are interconnected with a 1Gbps lan cable
- Disks are luns from Emc clariion CX3-380 (20 lun with 7.2 TB) with 8x4 Gbs fiber channel connection

Our Environment



Environment Write raw performance: each disk was mounted from All servers and all read/write concurrently

	Path	Write		Read	
1	emcpowerz	19,9	MB/s	20,5	MB/s
2	emcpowerq	19,9	MB/s	21,1	MB/s
3	emcpowerx	23,9	MB/s	46,6	MB/s
4	emcpowero	23,7	MB/s	52,6	MB/s
5	emcpowery	23,6	MB/s	47,4	MB/s
6	emcpowerl	23,9	MB/s	60,3	MB/s
7	emcpowerw	25,0	MB/s	43,2	MB/s
8	emcpowerk	24,4	MB/s	50,2	MB/s
9	emcpowerv	25,7	MB/s	46,9	MB/s
10	emcpowerj	24,2	MB/s	49,0	MB/s
11	emcpowers	21,2	MB/s	57,4	MB/s
12	emcpoweri	21,2	MB/s	58,2	MB/s
13	emcpoweru	23,5	MB/s	46,7	MB/s
14	emcpowerg	23,6	MB/s	45,6	MB/s
15	emcpowerr	23,6	MB/s	46,6	MB/s
16	emcpowerh	23,9	MB/s	44,1	MB/s
17	emcpowert	25,3	MB/s	42,3	MB/s
18	emcpowerd	25,2	MB/s	40,9	MB/s
19	emcpowerp	24,7	MB/s	47,9	MB/s
20	emcpowerf	24,1	MB/s	43,6	MB/s

Environment raw Write performance: each disk was mounted from one servers and tested this is a mean of each result, one write a time

1	emcpowerz	350 MB/s
2	emcpowerq	348 MB/s
3	emcpowerx	349 MB/s
4	emcpowero	354 MB/s
5	emcpowery	352 MB/s
6	emcpowerl	340 MB/s
7	emcpowerw	348 MB/s
8	emcpowerk	350 MB/s
9	emcpowerv	346 MB/s
10	emcpowerj	347 MB/s
11	emcpowers	351 MB/s
12	emcpoweri	346 MB/s
13	emcpoweru	349 MB/s
14	emcpowerg	349 MB/s
15	emcpowerr	350 MB/s
16	emcpowerh	350 MB/s
17	emcpowert	348 MB/s
18	emcpowerd	347 MB/s
19	emcpowerp	351 MB/s
20	emcpowerf	345 MB/s

Overall Raw score

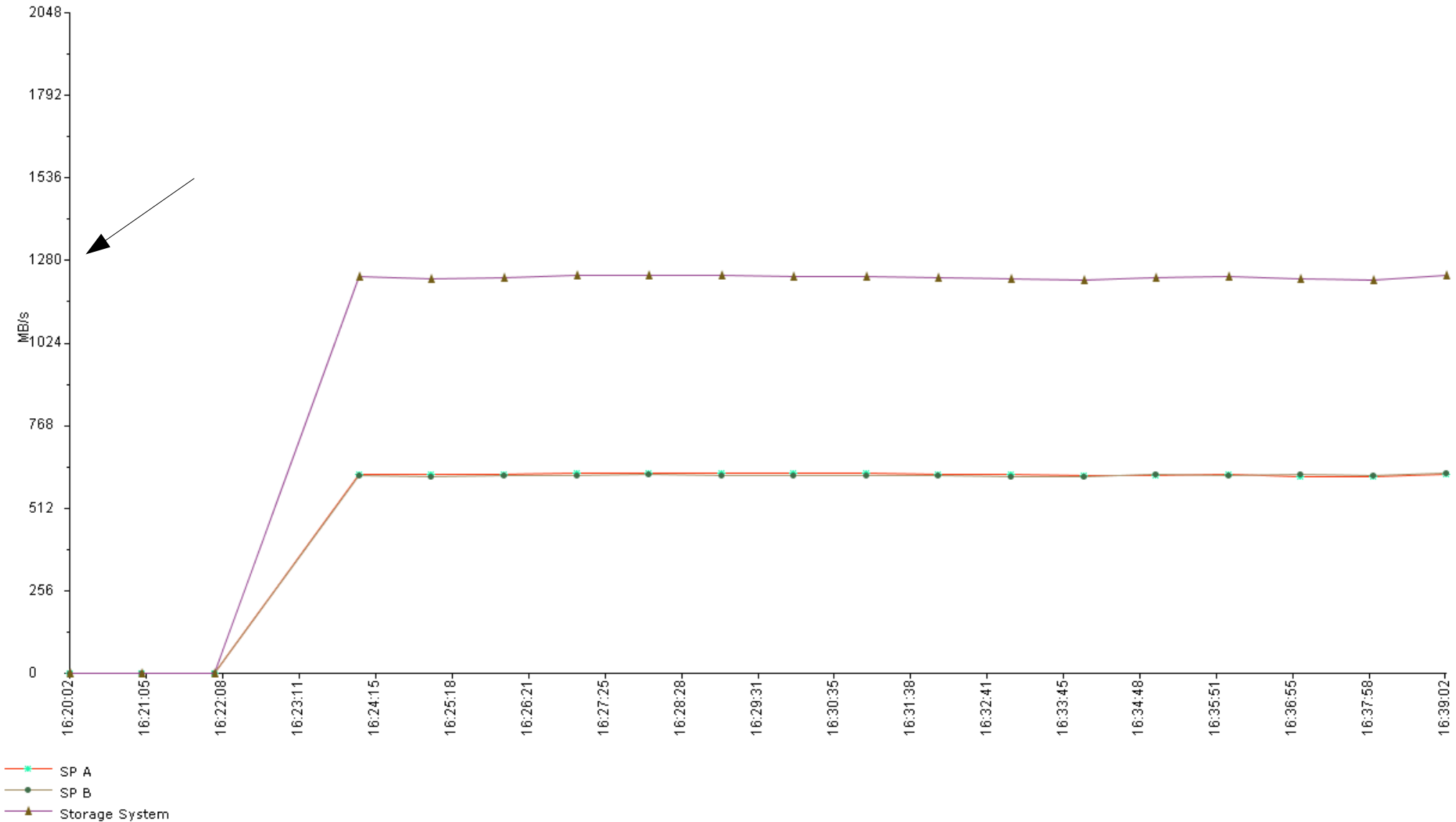
Total write bandwidth (MB/s):

607,6

Total read bandwidth (MB/s):

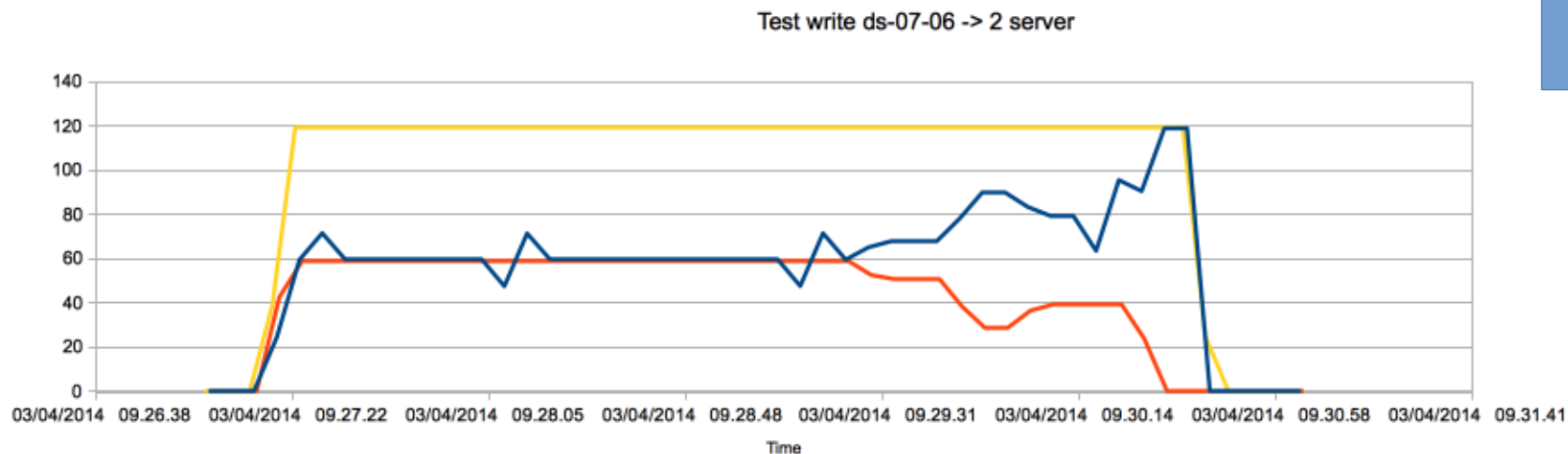
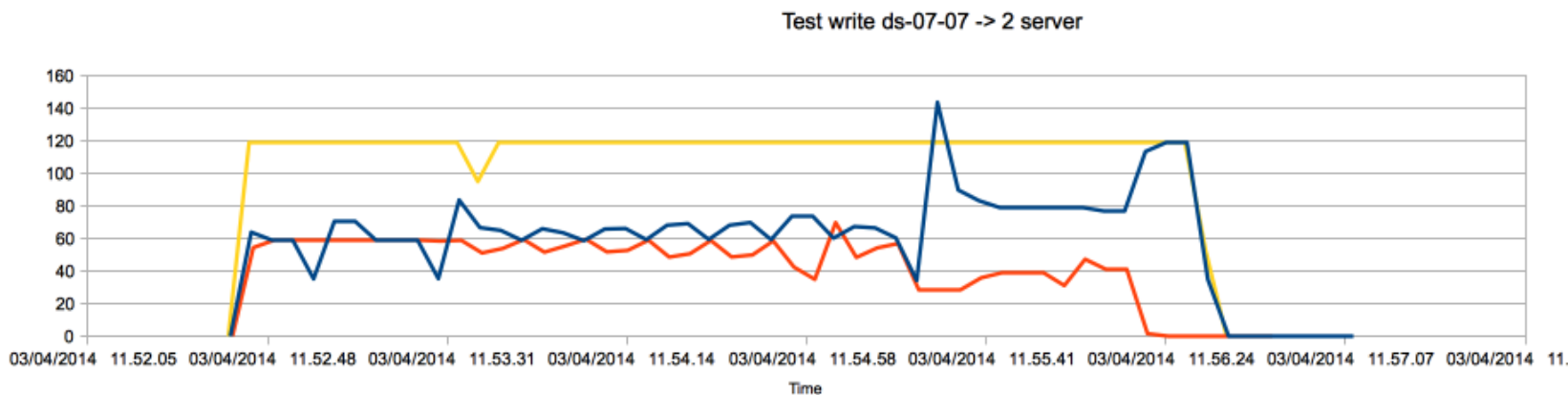
1196,2

Raw READ Performance graph



Test 1 With Ceph deployed

- 2 server and 1 client write. Journal are on the same osd disk

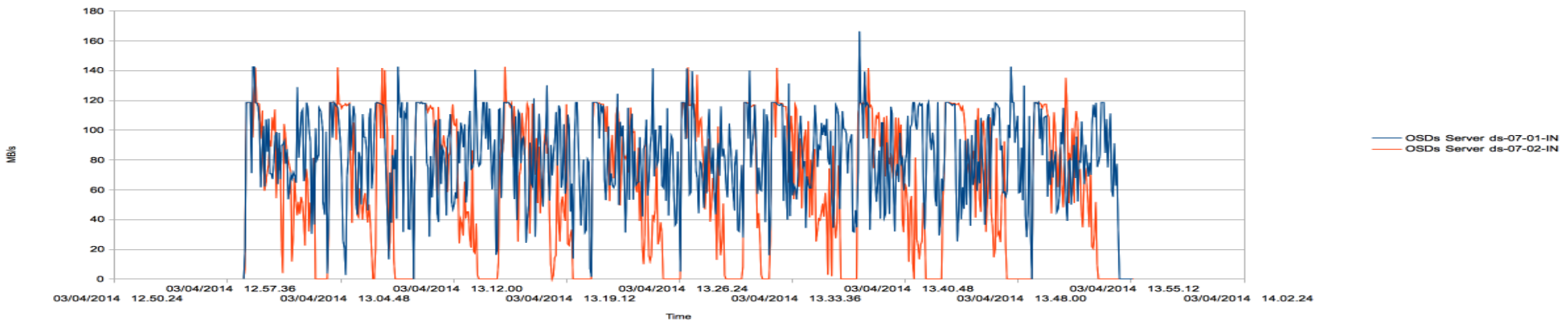


Ethernet
Link saturated

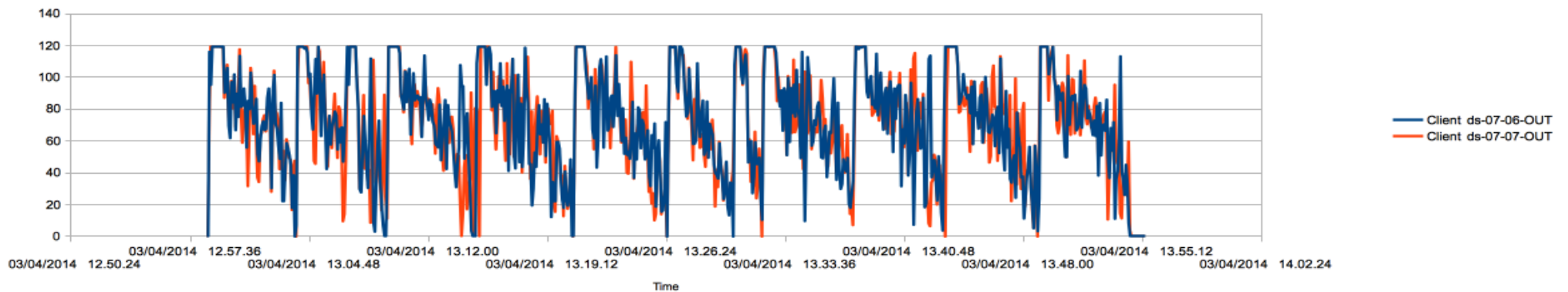
Test 2

- 2 server and 2 client write simultaneously. Journal on the same osd.

Test write ds-07-07 -> 2 server



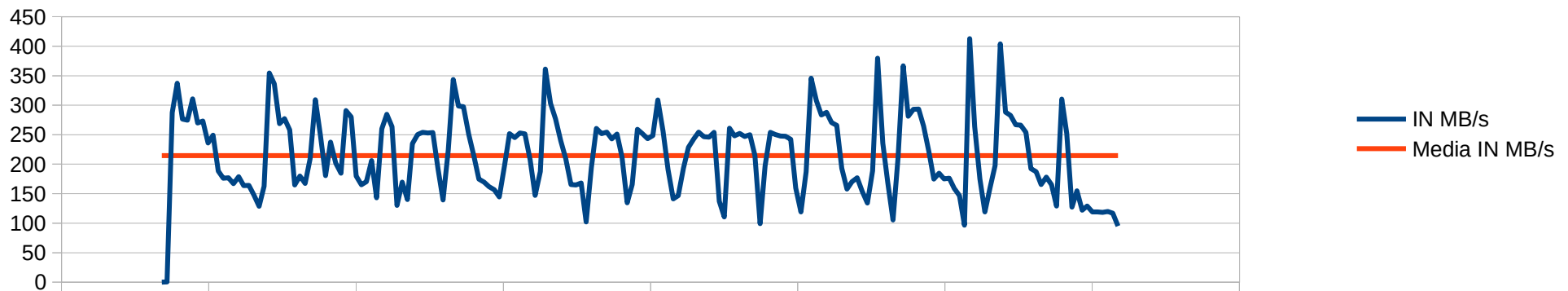
Client ds-07-06-OUT



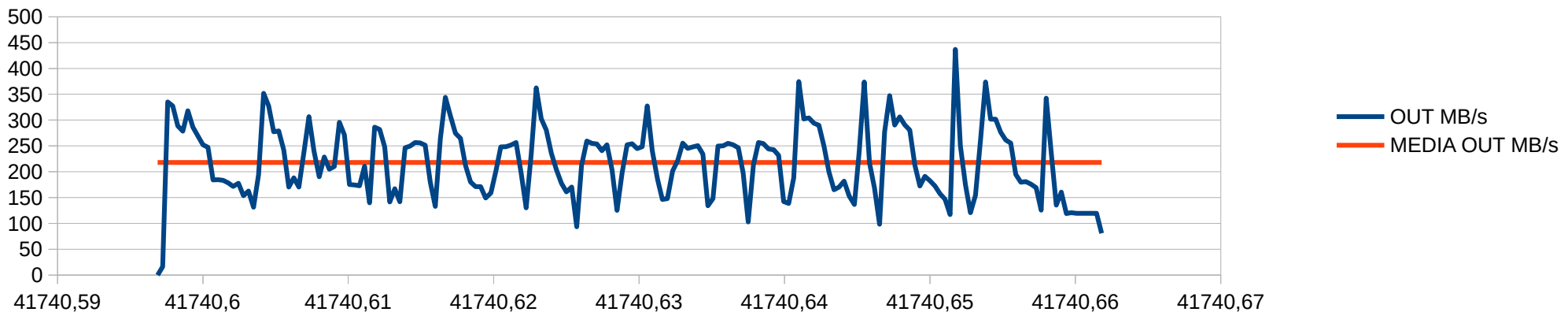
Test 4

- 5 server with 10 osd and 5 clients writting simultaneously one file each client

Server Aggregato IN



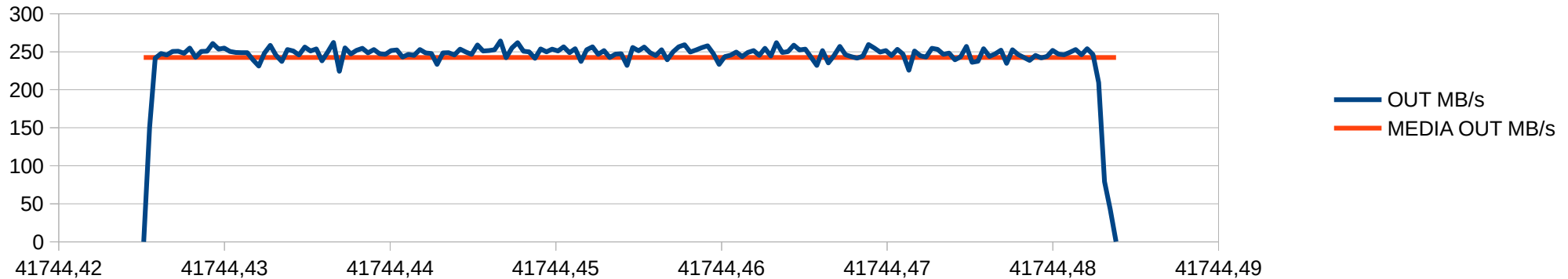
Client aggregato OUT



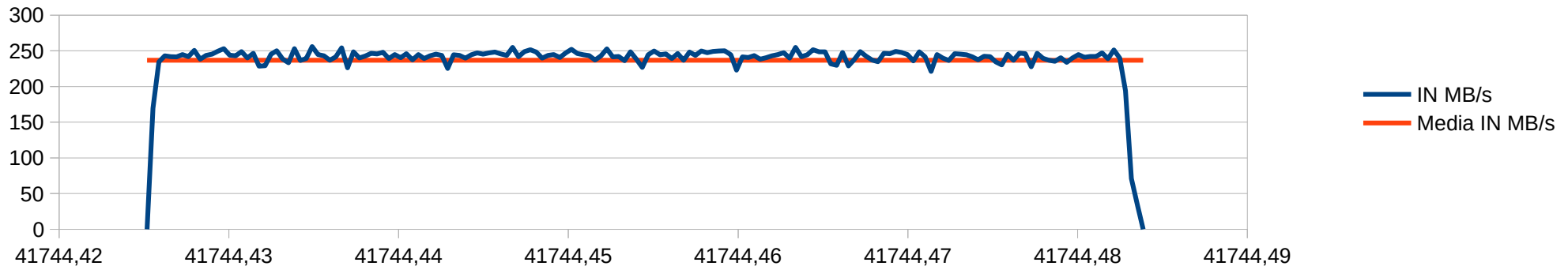
Test 5

- 5 server with 10 osd and 5 clients reading simultaneously one file each client

Server Aggregato OUT



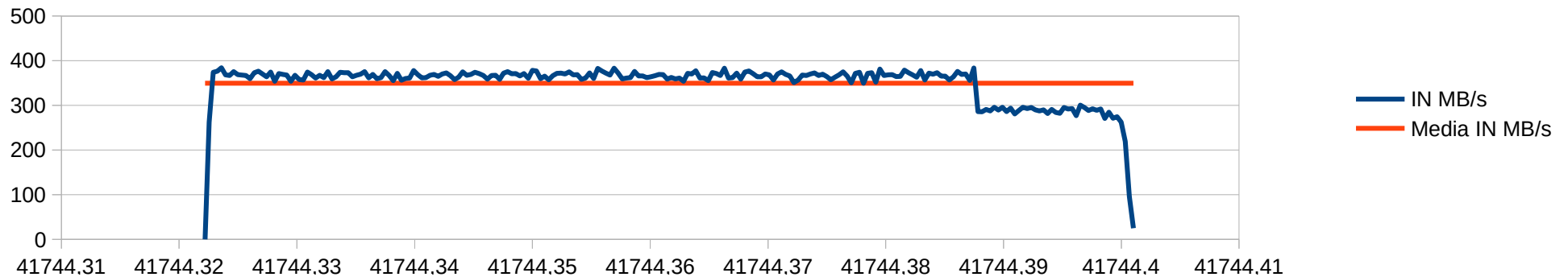
Client aggregato IN



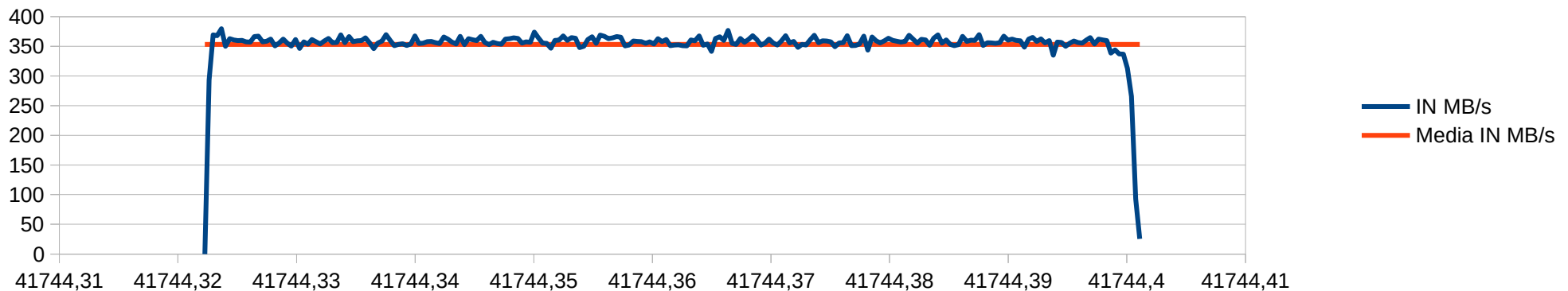
Test 6

- 5 server with 10 osd and 5 clients reading simultaneously TWO file each client, replica not active

Server Aggregato OUT



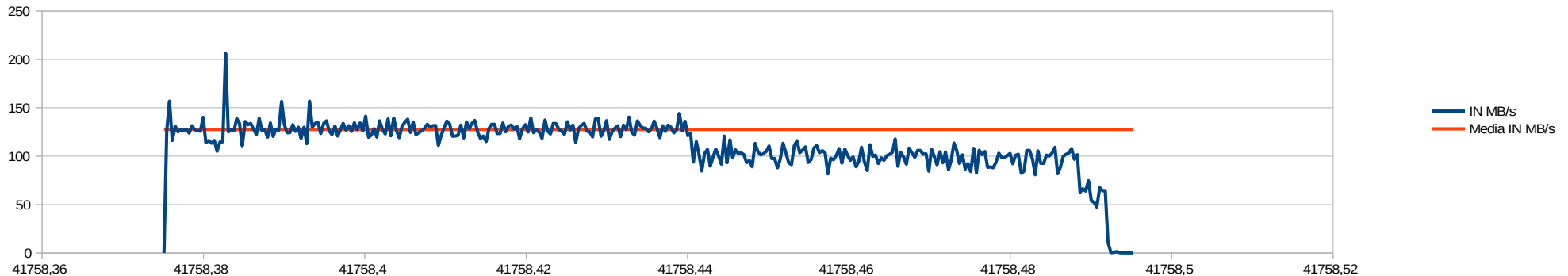
Client aggregato IN



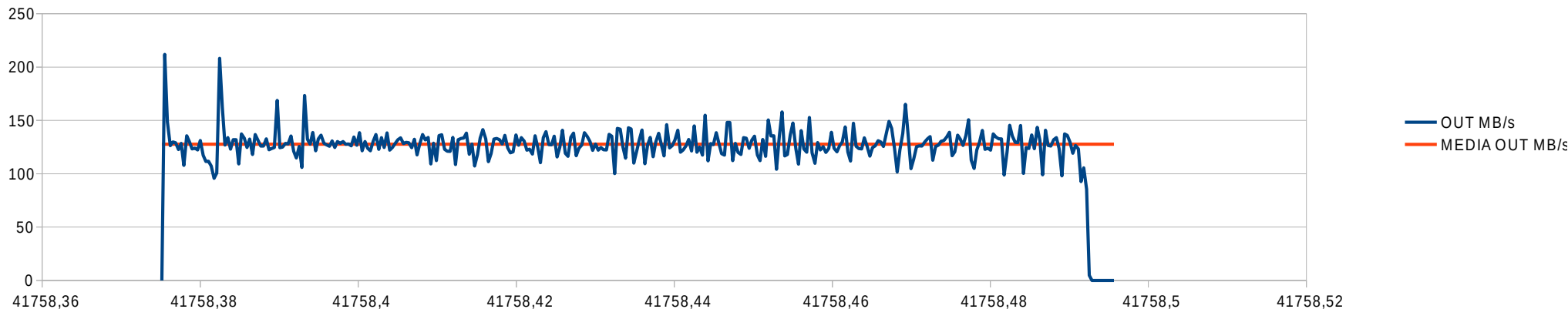
Test 7

- 5 server with 10 osd and 5 clients writting simultaneously one file each client. Journal partition moved on other disk, replica active

Server Agregato IN

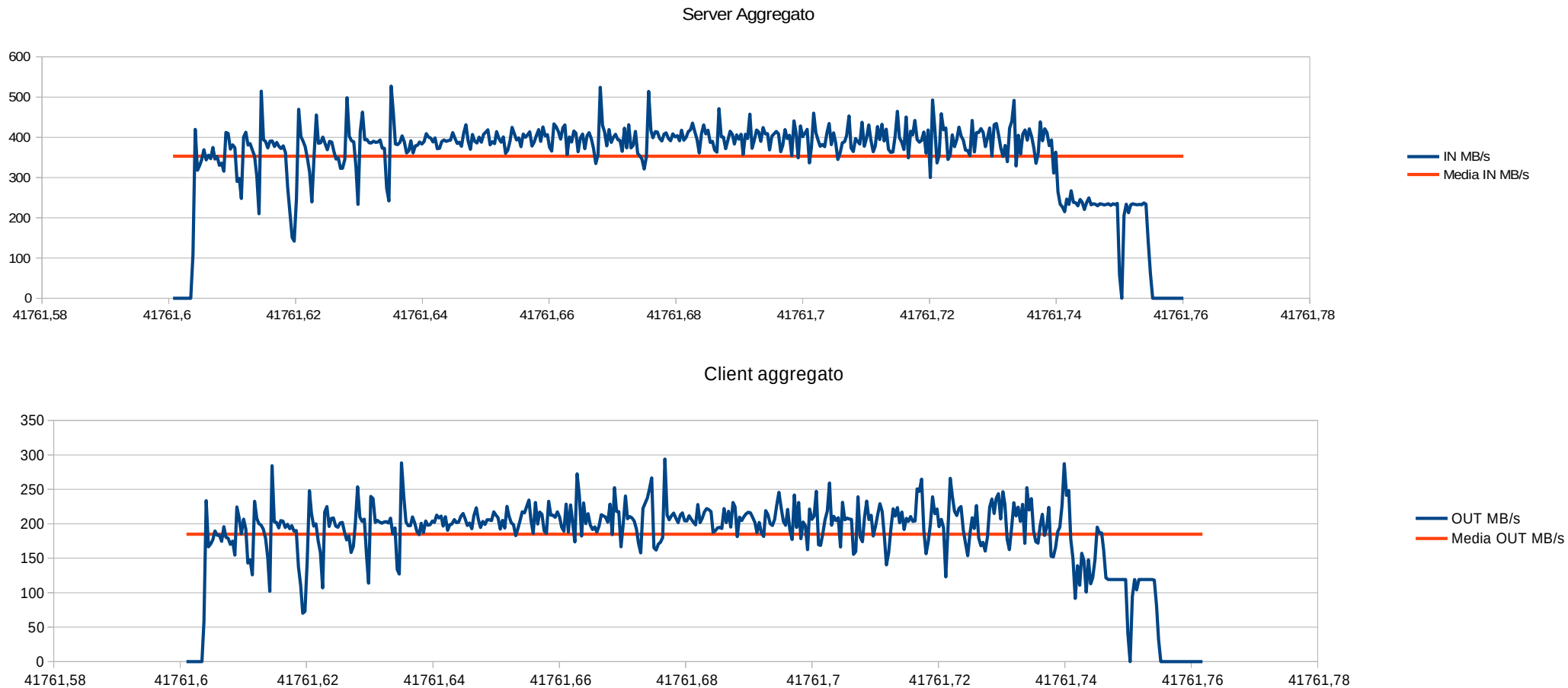


Client aggregato OUT



Test 8

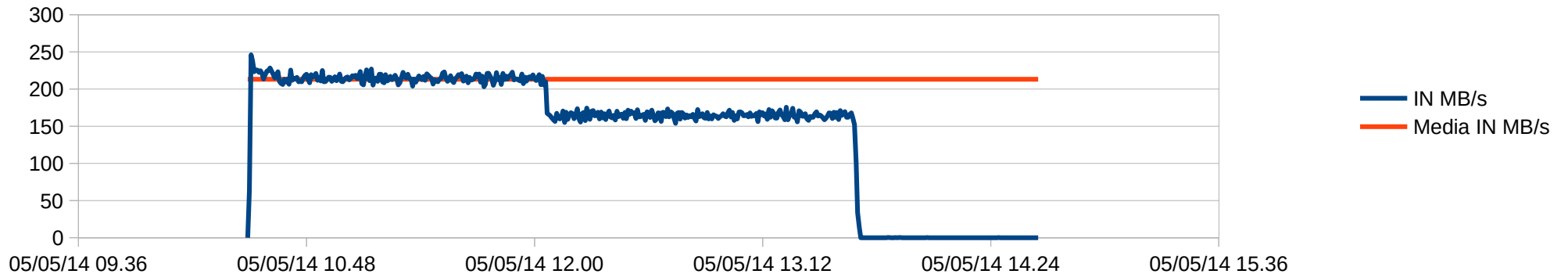
- 5 server with 10 osd and 5 clients writting simultaneously one file each client. Journal partition moved on ram disk. replica active



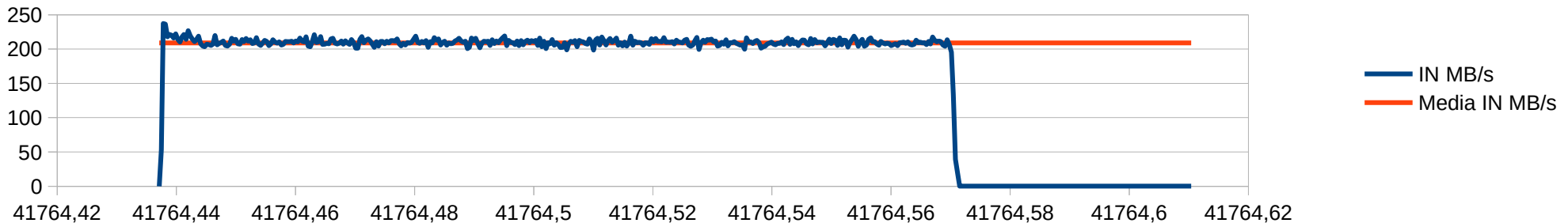
Test 9

- 5 server with 10 osd and 5 clients reading simultaneously one file each client. Journal partition moved on ram disk, replica active

Server Aggregato



Client aggregato



Conclusion

- Ceph has a lot of capability and is extremely reliable
- On our hardware it doesn't perform at maximum possible but there are TONS of tune factors maybe someone has been missed
- It has an extremely active and responsive mailing list where the developers answer very quickly
- CephFS has some problems but looking their roadmap, they want to improve this aspect in the near future
- The learning curve is not linear, in order to understand how this filesystem works and how to tune it in good way is necessary understand very well how the data are accessed and in which way ceph places the data. (PG – POC – CRUSH MAP – MON MAP etc)

Related Work

- We have read some other tests about ceph performed by INFN-Bari and Cern
- The works are being presented on last CHEP 2013 conference and Hepix 2014
- Every team reach our same conclusions:
 - Chep is very interesting as storage solution.
 - it has some performance problems

Cern related Work

- In particular the Cern's group tried this filesystem in a HUGE environment and they didn't reach a performance level in line with their test environment

Our 3PB Ceph Cluster

CERN IT
Department

48 OSD servers

Dual Intel Xeon E5-2650
32 threads incl. HT
Dual 10Gig-E NICs
Only one connected
24x 3TB Hitachi disks
Eco drive, ~5900 RPM
3x 2TB Hitachi system disks
Triple mirror
64GB RAM

5 monitors

Dual Intel Xeon L5640
24 threads incl. HT
Dual 1Gig-E NICs
Only one connected
3x 2TB Hitachi system disks
Triple mirror
48GB RAM

```
[root@p05151113471870 ~]# rados bench 30 -p test write -t 100
Total writes made:      7596
Write size:             4194304
Bandwidth (MB/sec):     997.560
Average Latency:        0.395118
[root@p05151113471870 ~]# rados bench 30 -p test seq -t 100
Total reads made:       7312
Read size:              4194304
Bandwidth (MB/sec):     962.649
Average Latency:        0.411129
```

- <http://indico.cern.ch/event/214784/session/6/contribution/68/material/slides/0.pdf>

INFN-BARI Related work

- At INFN-Bari site they has arrived at our same conclusions



CEPH pros&cons

- Complete storage solution (supports all the storage interfaces: posix, object, block)
- Great scalability
- Fault-tolerant solution
- Difficult to install and configure
- Performance issues
- Some instabilities while under heavy load

Ceph @ Hepix 2014 1/2

Ceph at the UK Tier 1

George Ryall (STFC)

James Adams (STFC), Alastair Dewhurst
(STFC), Rob Appleyard (STFC), Kenneth
Waegeman (UGent)

HEPiX Annecy-Le-Vieux, May 2014

- Ceph looks promising as a technology but currently has gaps in it's documentation lack of support for production use of CephFS is concerning
- Some inconsistencies in configuration. Pool numbers rather than names need to be specified, poor documentation on pools and assigning sections of file system to pools. Administrative interfaces often frustrating and not intuitive to use.

Ceph @Hepix 2014 2/2

Ceph @ CERN: one year on...

Dan van der Ster (daniel.vanderster@cern.ch)

Data and Storage Service Group | CERN IT Department

HEPIX 2014 @ LAPP, Annecy

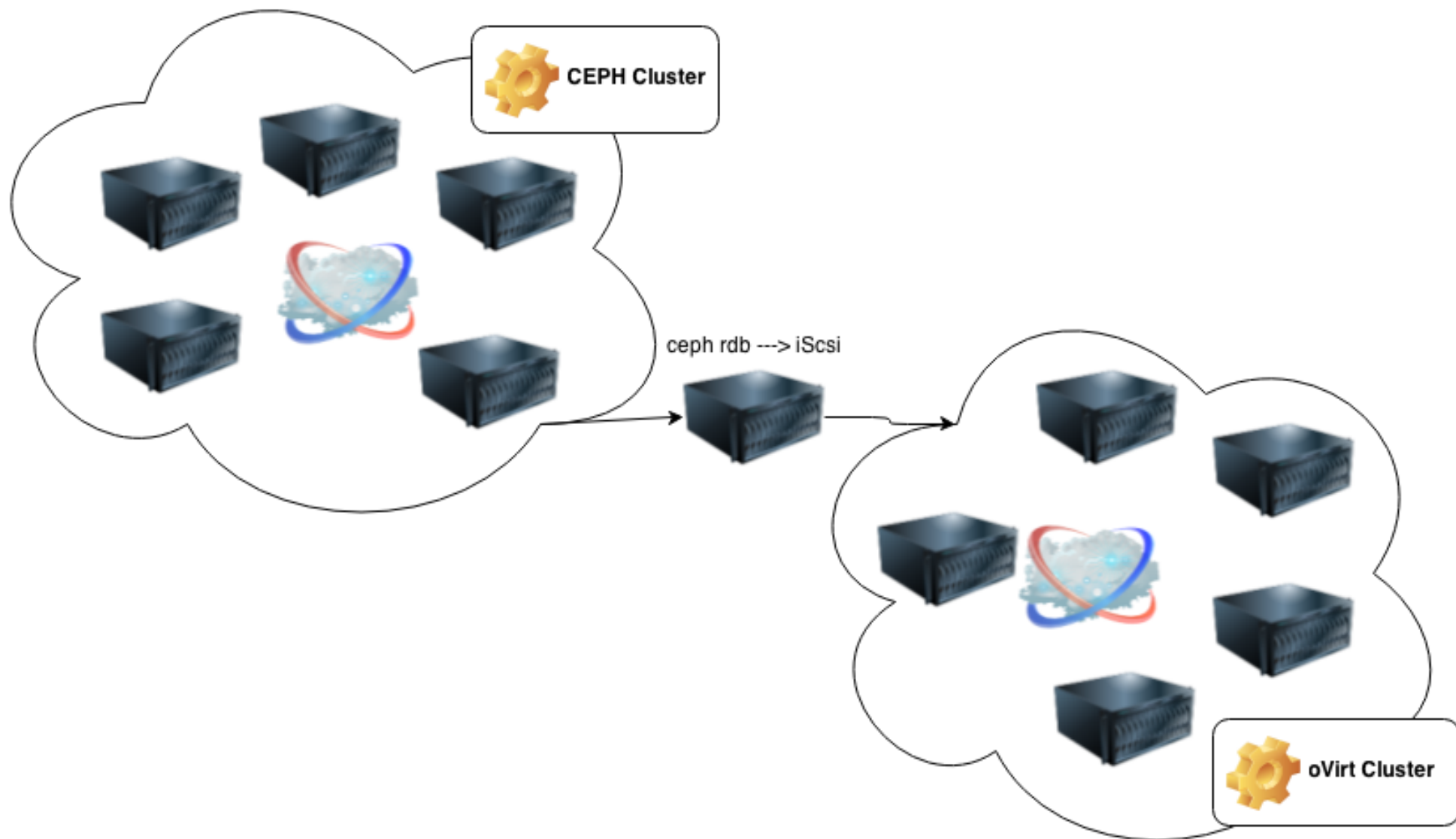
Same as CHEP conference with only some relevant adds:

- “For block storage, make sure you have SSD journals”
- “Still young, still a lot to learn, but seems promising.”

Use Case @ CNAF

- In collaboration with Servizi Nazionali
- We have attached a “pool” of ceph to oVirt infrastructure via an iScsi Export
- We have migrated some virtual machine on it

Structure Export



Questions?

Thanks