ITALY
OpenInfra Days

OPENINFRA DAYS
ITALY

Organized by

IRIDEOS

Binario Etico

Under the patronage of

AGID | Agenzia per l'Italia Digitale

Sponsored by

Mellanox TECHNOLOGIES

MESOSPHERE

gci SERVICE FACTORY

Stefano Stalio, Cristina Duma,
Alessandro Costantini, Marica Antonacci
on behalf INFN Corporate Cloud

alessandro.costantini@cnaf.infn.it

Roma, 3/10/2019

OpenStack storage solution

CEPH implementation in INFN Corporate Cloud

Summary

1. INFN & INFN-CC
2. Storage architecture in INFN-CC
3. Tests and results
4. Conclusions and future work

INFN & INFN-CC

INFN

INFN & Computing

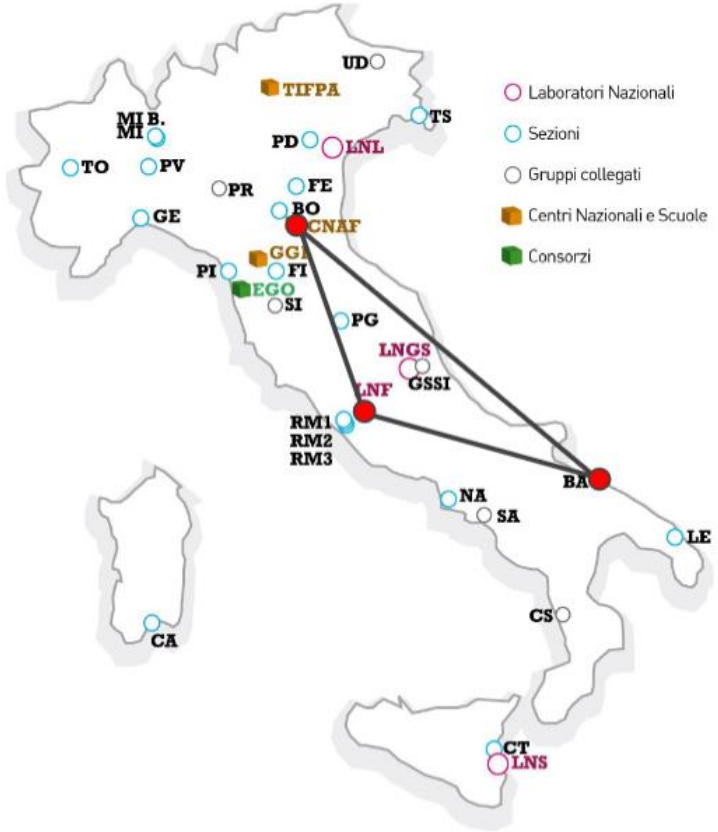
- A long tradition, from the first small clusters to GRID and Cloud based large scale computing
- INFN not interested in computing per-se, but as a mean for its research related activities
- In the last 10 years, this has principally meant supporting the Experiments @CERN (LHC)

- Currently, INFN operates:
  - 9 medium size centers (Tier-2s in the Worldwide LHC Computing GRID hierarchy)
  - 1 large WLCG Tier-1 center, at CNAF (Bologna)–certified ISO-27001

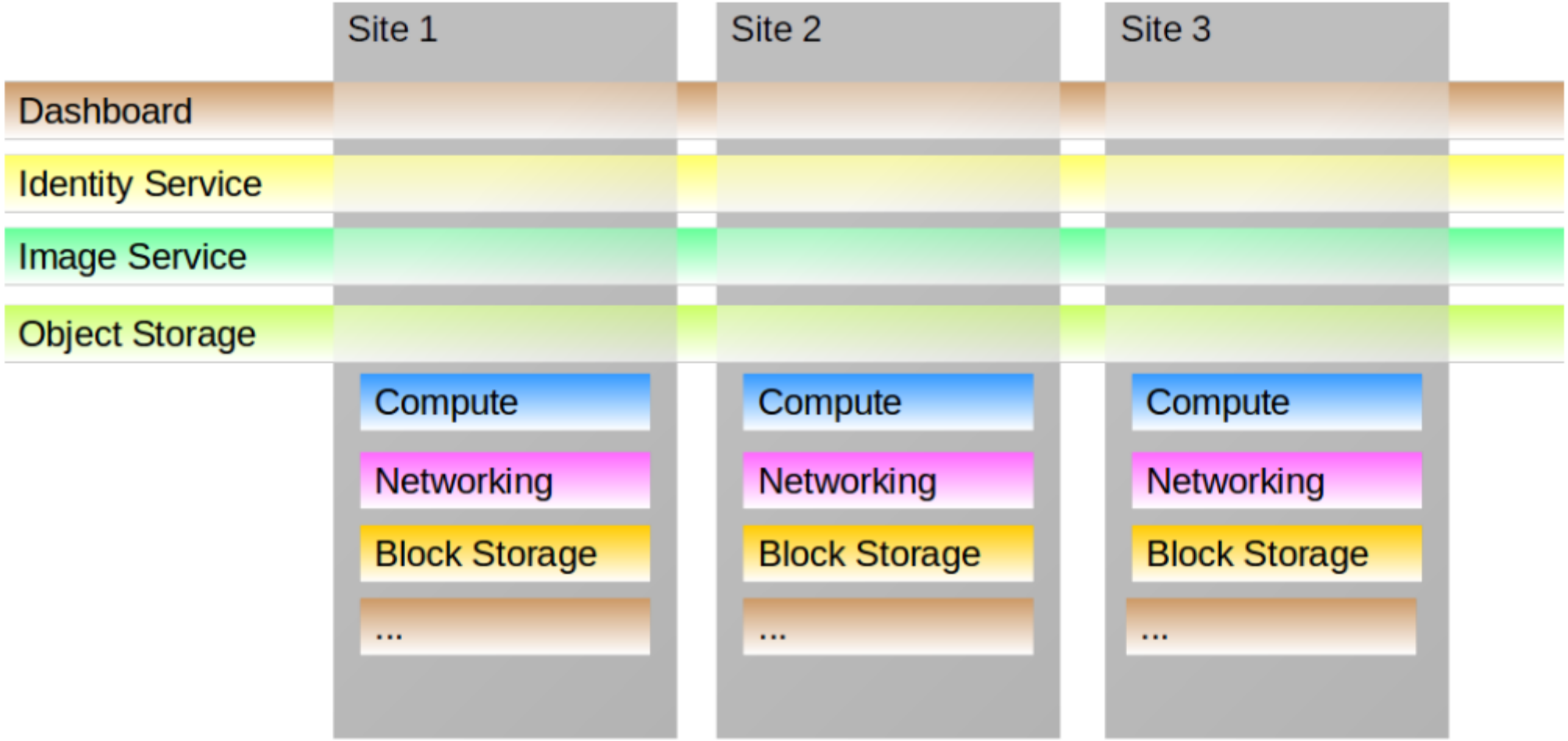
- All the centers are connected with at least 10 Gbit/s dedicated connections, currently being upgraded to 100 Gbit/s
- Collectively, our main centers have about 50,000 CPU cores, 50PB of enterprise-level disk space, 60PB of tape storage

INFN Corporate Cloud

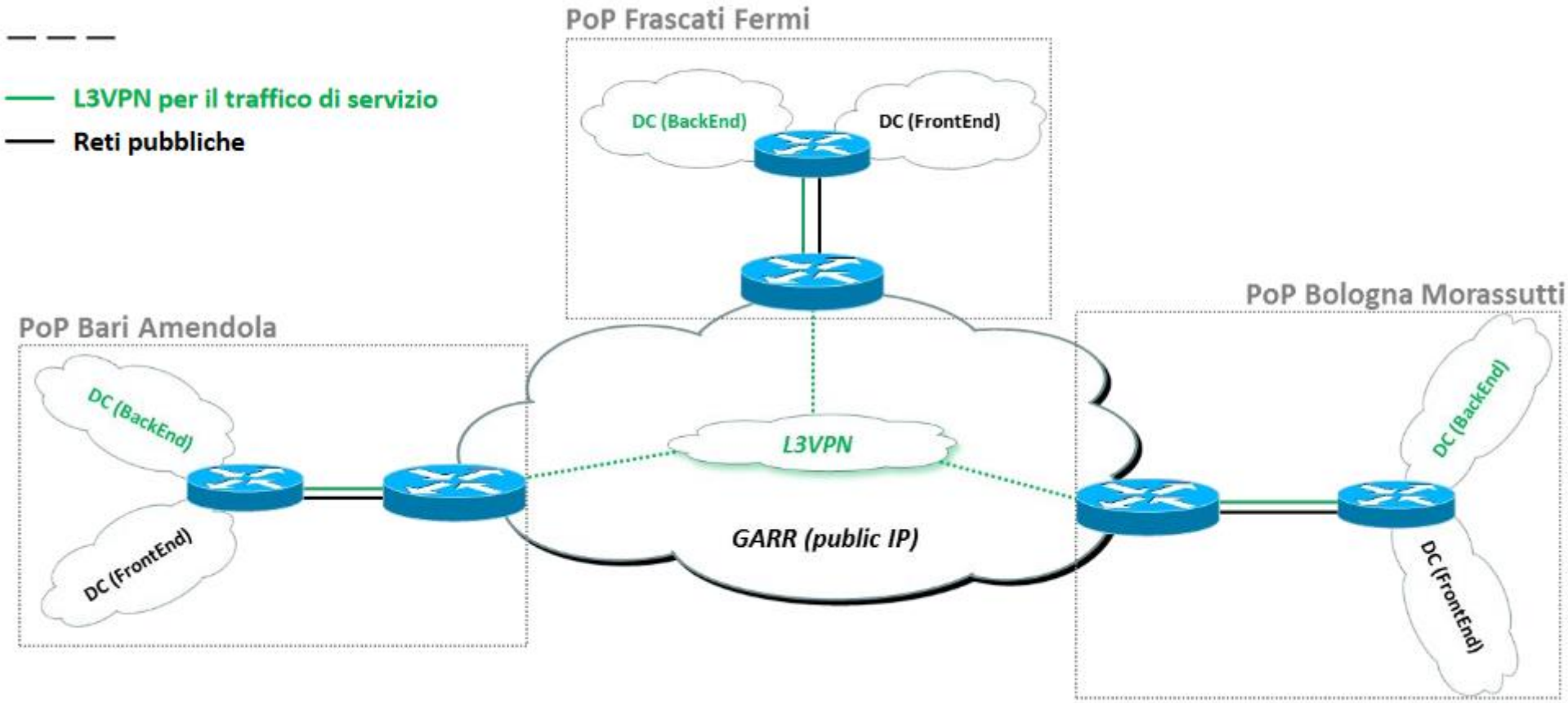
- Private cloud, on homogeneous resources
  - Based on OpenStack
- Distributed among 3 INFN sites
  - Bologna
  - Roma
  - Bari
- IaaS platform aimed at implementing
  - PaaS, SaaS
  - Cloud-oriented services for research



INFN CC - local and distributed services

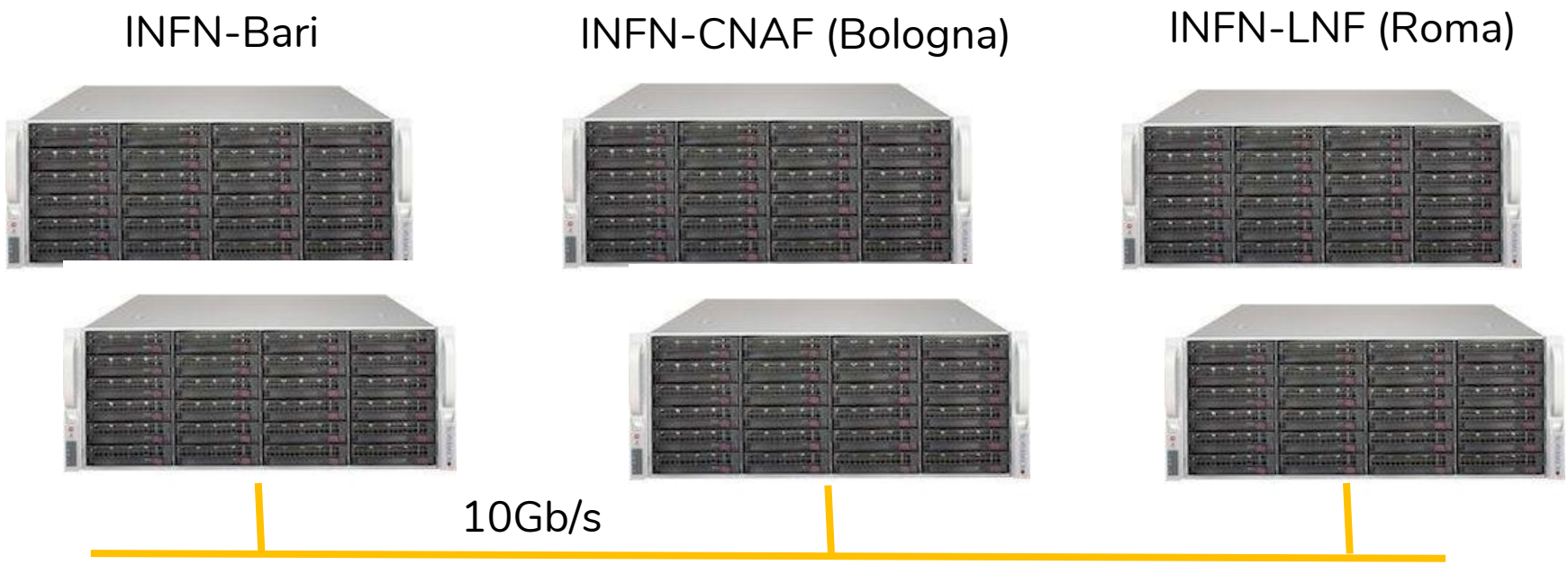


INFN CC - Network



Storage architecture in INFN-CC

Storage hardware layout



- 3 independent clusters, each with:
  - 2 Storage nodes
  - 12 HDD/server (6TB o 8TB)
  - 6 SDD/server (500GB)
  - & more….

Storage backend layer

CEPH

- Block device
  - Metadata on SSD
  - Data on HDD or SSD
  - Replica policy
    - Replicated and Erasure Coded

SWIFT

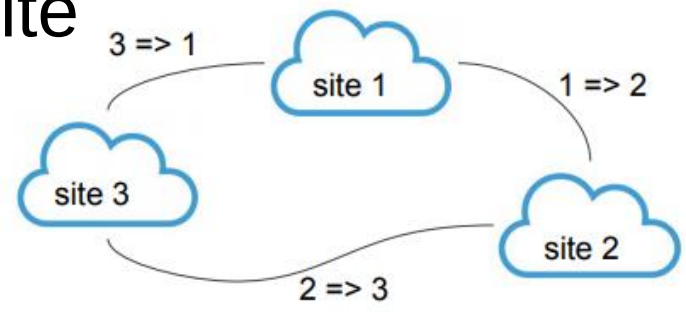
- Object storage
  - Deployed on the same hardware
  - Image and Object storage
  - Physical resources allocated on-demand

CEPH in INFN-CC

- Storage backend used for:
  - Nova
  - Cinder

- Different Pool types
  - HDD, SSD, EC-HDD, EC-SSD
  - accessed via VM through mounted volumes

- Some pools also replicated among site
  - RBD mirroring (asynchronous)
  - Mainly for Disaster Recovery



Replica strategies: EC vs Standard

- Replica 3 is a typical configuration
  - 33% of the total raw capacity is used
- CEPH Erasure Code
  - 66% in 4+2 OSD configuration, supports 2 OSD failure
  - uses more RAM and CPU than replication
  - slower in managing small files
- Evaluation of different configuration options (erasure coding VS standard replica) was performed
  - Comparing performances

Tests and results

Tools used

- Iperf
  - Network tests
    - CEPH node (Cluster Net) <-> Ceph node (Cluster Net)
    - Hypervisor <-> CEPH node (Public Net)
- Rados bench
  - CEPH performance
    - Read/Write for each selected Pool
- FIO
  - Read/Write from VM
  - Processes * File size = (total GB)
    - 8*256 (2GB), 8*512 (4GB), 8*1024 (8GB)
    - 8*1024 (8GB), 16*512 (8GB), 32*256 (8GB)

Iperf tests

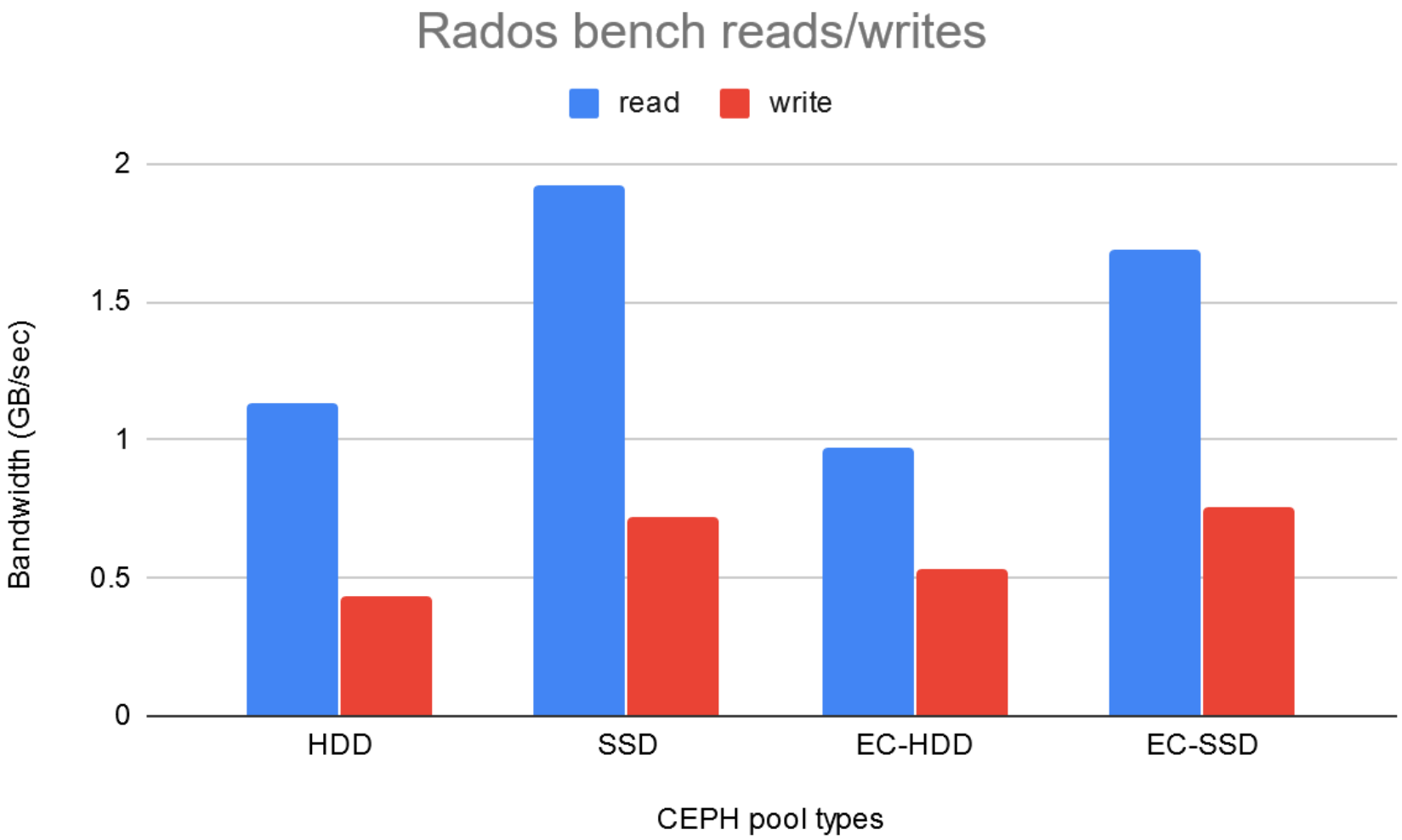
- CEPH node <-> CEPH node

```
# iperf -s
…
[ ID] Interval       Transfer     Bandwidth
…
[  4] 0.0-10.0 sec  10.9 GBytes 9.36 Gbits/sec
```

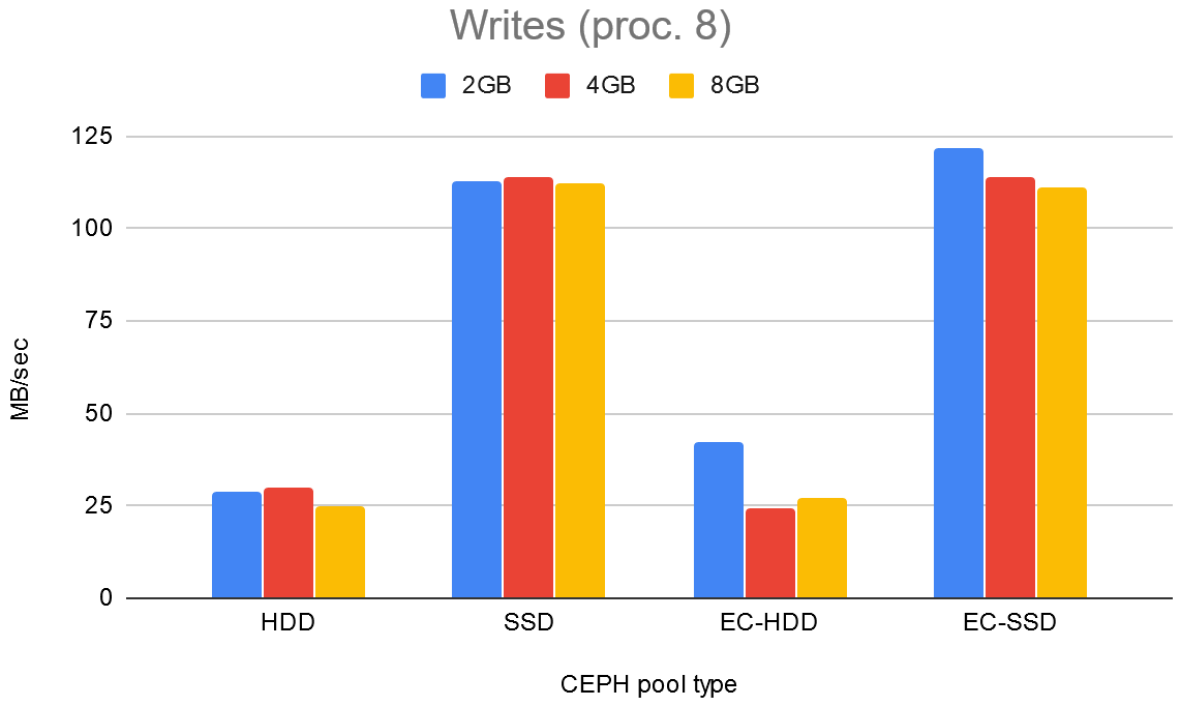
- CEPH node <-> Hypervisor

```
# iperf -s
…
[ ID] Interval       Transfer     Bandwidth
…
[  4] 0.0-10.0 sec  10.8 GBytes 9.30 Gbits/sec
```

Rados bench tests on CEPH nodes

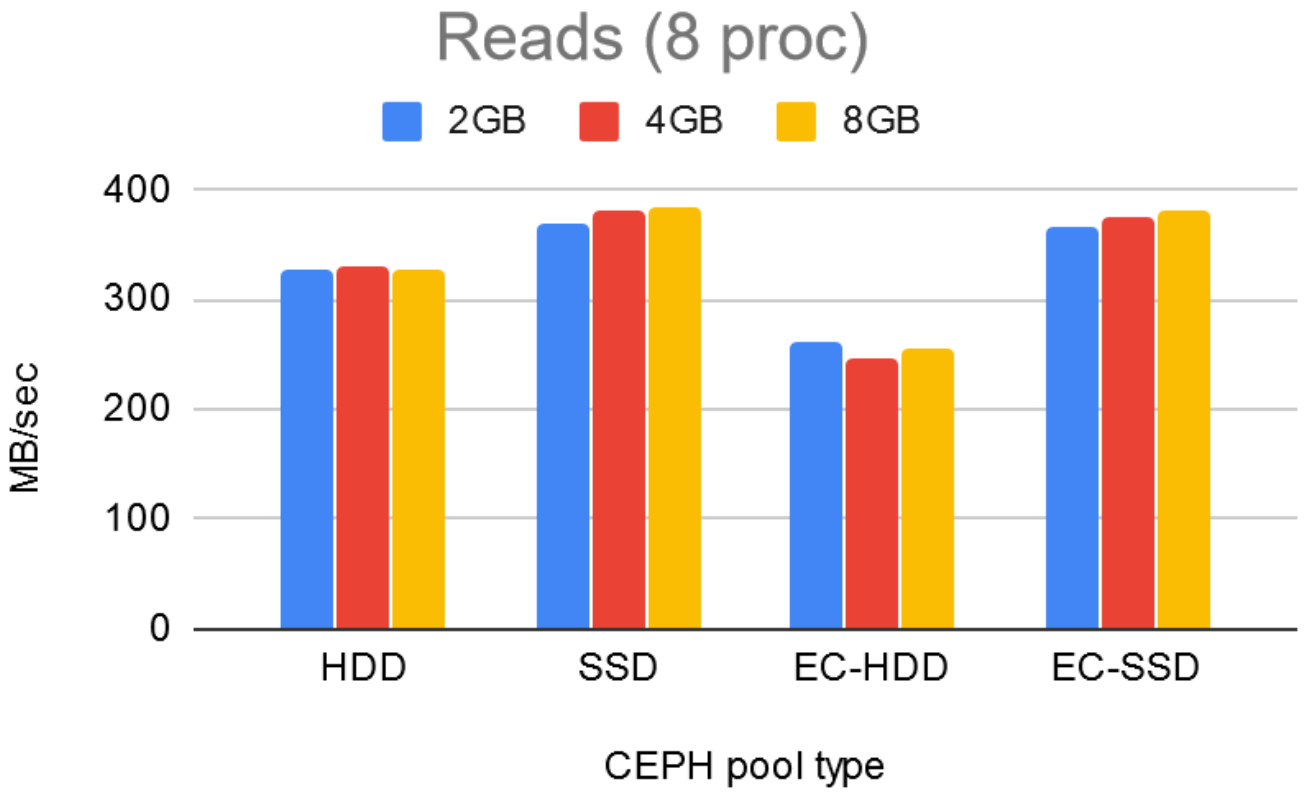


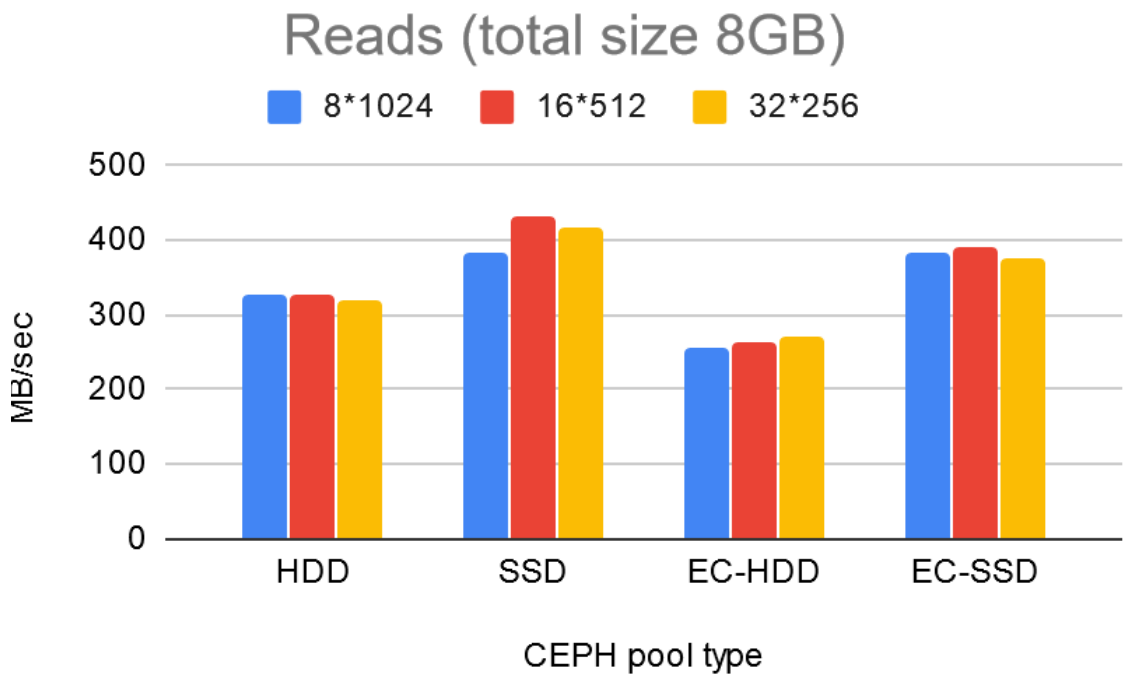
FIO - writes on VM



- EC-HDD: Metadata pool on HDD, data pool on HDD
- EC-SSD: Metadata pool on SSD, data pool on HDD

FIO - reads on VM





- EC-HDD: Metadata pool on HDD, data pool on HDD
- EC-SSD: Metadata pool on SSD, data pool on HDD

Findings

- The write speed on EC pools is highly dependent on
  - the device type used in the pool (SSD vs HDD)

- The read speed on EC pools is influenced by
  - the device type used in the pool (SSD vs HDD)
  - OpenStack software layers

- In general, read and write speeds from VM are comparable between the different replica strategies (EC vs replicated)

Special case: mirror RBD for remote copy

- RBD mirroring
  - Asynchronous process
  - Remote copy may take a long time
    - Depending on the written data size
- Easy to fill the metadata pool
  - Specially on SSD pools of limited size (SSD or EC-SSD)

Writes will fail

Conclusions and Future work

Conclusions

- Preliminary tests has been performed locally on each CEPH cluster of INFN-CC cloud infrastructure
  - 4 different types of pools
    - HDD replicated, SSD replicated
    - EC-HDD, EC-SSD
- Differences depends mainly from
  - Type of device used
  - Software layers
- EC pools can be adopted in place of standard (replicated) pools
  - read and write speeds from VM are comparable between the tested replica strategies (EC vs replicated)

Future work

- Perform other tests
  - different volume size
  - varying number concurrent processes
  - CEPH tier cache
- Goal
  - Evaluate the proper storage configuration to be adopted in production environments