Ceph New Cluster Deployment

Antonio Falabella - Andrea Prosperini

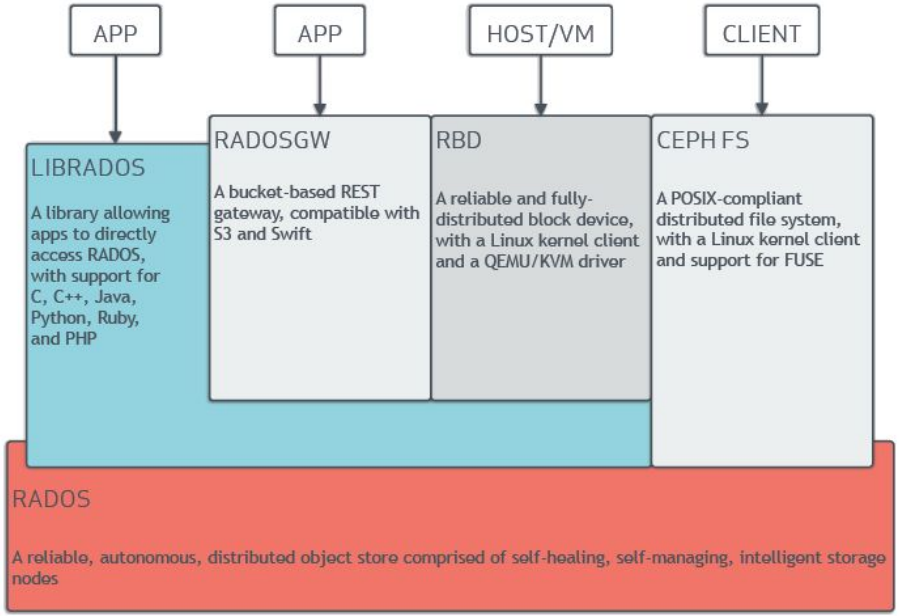
CCR Paestum 23-27 Maggio 2022

Introduction

- ~49 PB of disk space installed at CNAF
- Currently all of the pledged space is handled with IBM Spectrum Scale (formerly GPFS)
- Main feature is POSIX access → FS
- Technology scouting and investigation lead us to test other FS solutions with the following constraints:
    - **POSIX Compliance**
    - **Extended attributes**
    - **Quota support**

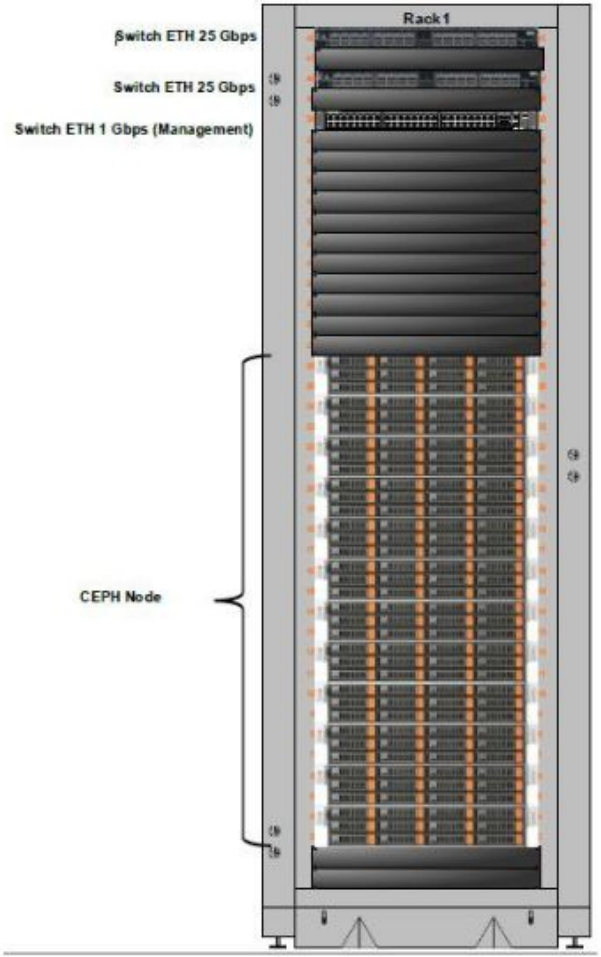
Ceph

- Clustered, Distributed and Network filesystem (open source - LGPL v2.1)
- https://docs.ceph.com/
- Offer **Object Storage, Block Storage and filesystem**
- Current version 17 (Quincy)
- Tested at CNAF 14, 15 and **16** (Nautilus, Octopus and Pacific)



Hardware Components

- 12 servers
  - 2U
  - 2 X 34305 Xeon 24-Core 5220R 2.2Ghz 35.75MB
  - 24 bays SATA/SAS
  - 384GB RAM (12*32)
  - 1 SAS HBA
  - 1 RAID controller (4 port)
  - **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
- **~90 euro/TB (3473 TB Netti)**



Servers Front View

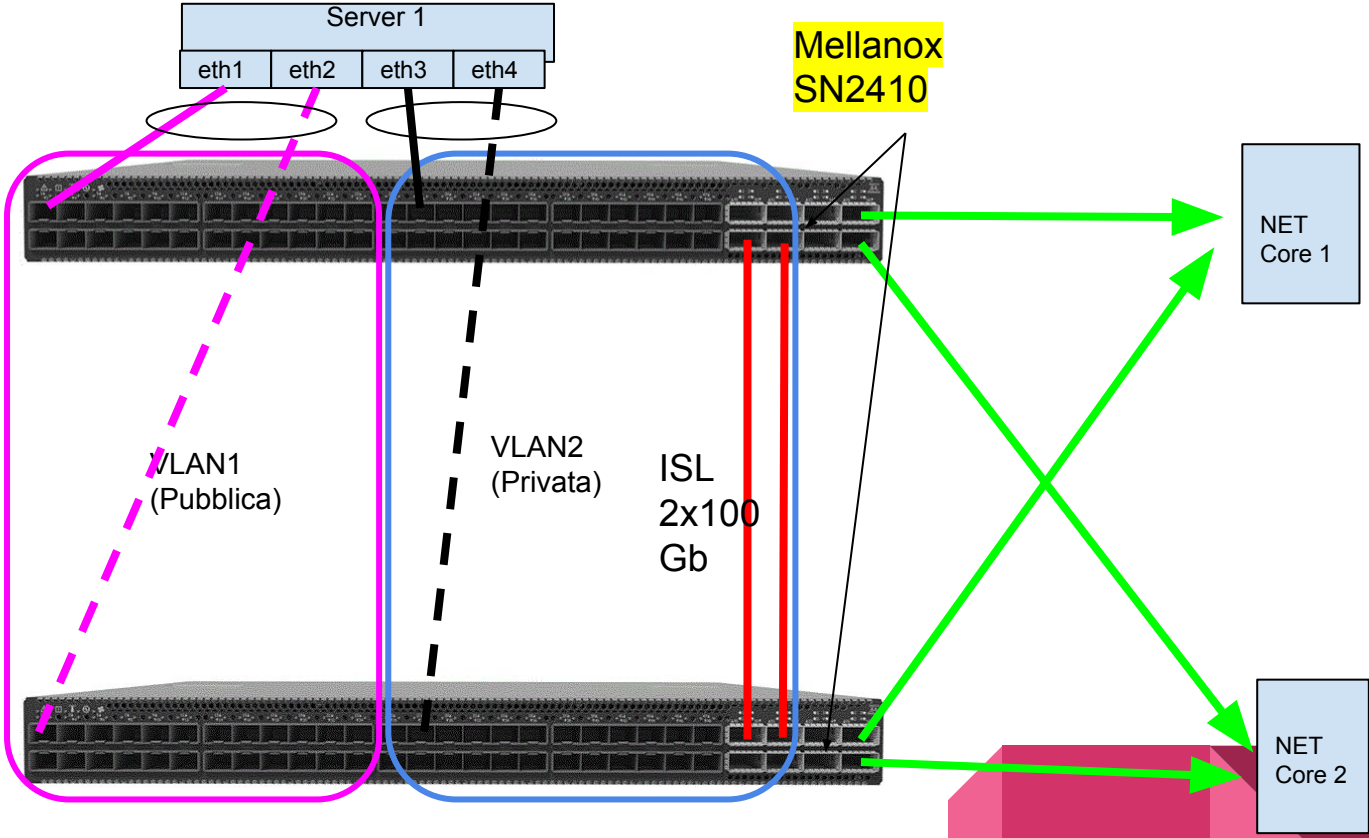
SuperStorage SSG-6029P-E1CR24L

(Front View – System)

Location	Description
0 ~ 11	12x 3.5" Hot-swap SAS3/SATA3 Drive, Front Bays
12 ~ 19	8x 3.5" Hot-swap SAS3/SATA3 Drive, Raiser Bays
20 ~ 23	4x 2.5 Hot-swap SAS3/SATA3/NVMe Hybrid Bays

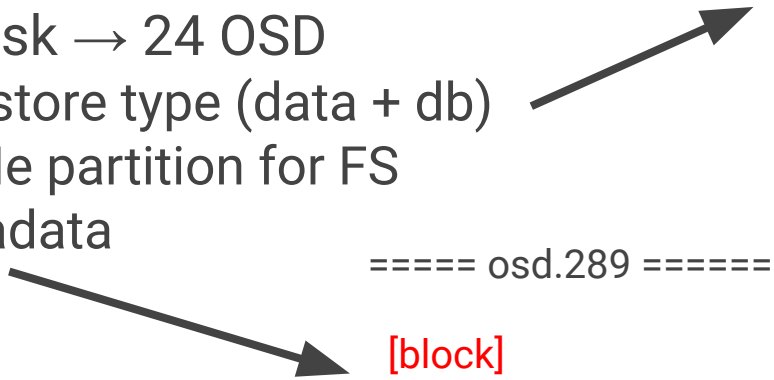
Network

- Link aggregation between different switch possible
- Maximum bandwidth on single stream not possible
- Aggregated throughput between streams from different OSDs would be maximized



Ceph Installation

- Deployment using official Ceph 16 (pacific) packages
- LVM partitions on disks
- 1 OSD per disk → 24 OSD
  - bluestore type (data + db)
  - NVMe partition for FS metadata

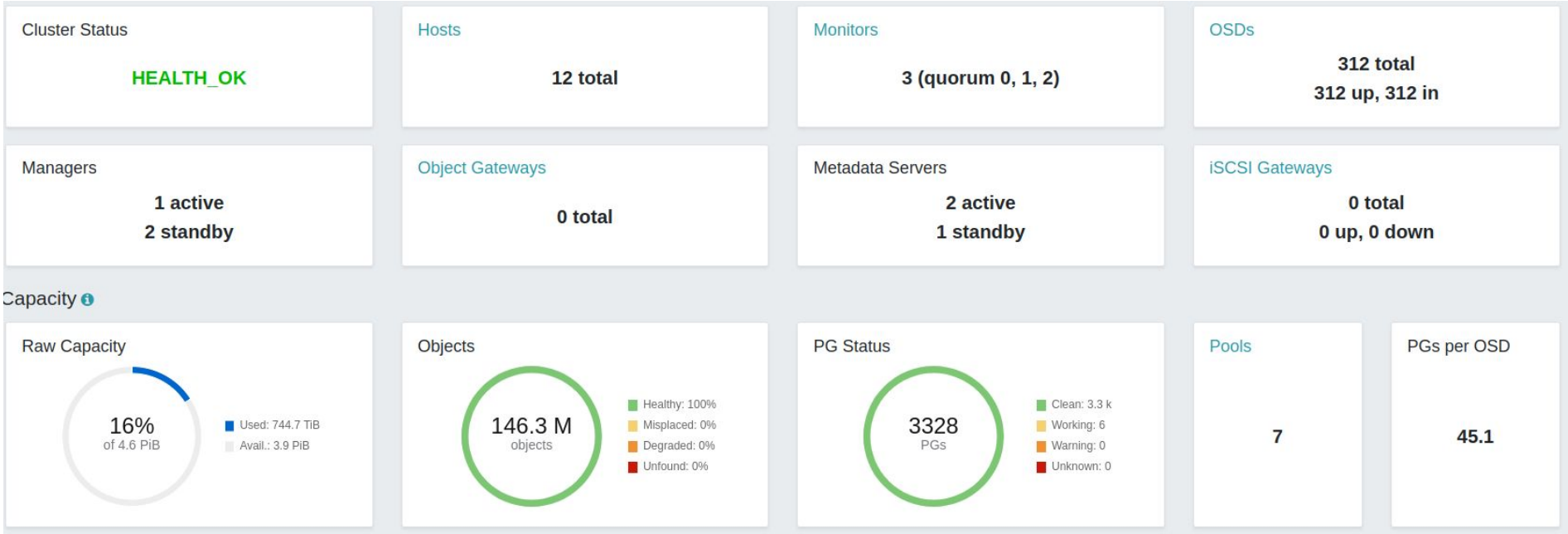


Cluster configuration

Cluster configuration:

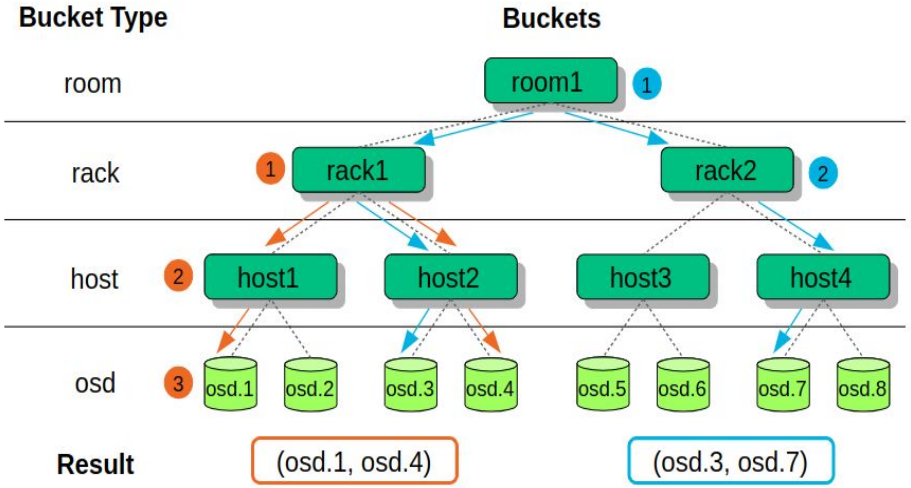
- 3 monitor nodes
- 2 manager nodes
  - coexisting with monitor services
- 3 mds server (1 active - 2 standby)
- 312 OSD services
  - 288 only rotational disk
  - 24 only SSD
- Monitoring and collection using **prometheus + grafana**
- clients -> linux kernel

Cluster Monitoring



Pools

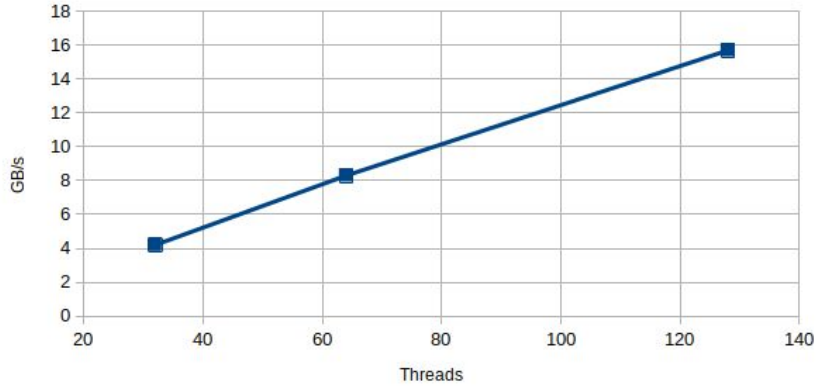
- 1 pool EC 8+4 for CephFS
  - **~67% of RAW space**
  - failure domain **host**
- 1 pool for metadata replica 3 (NVMe →Crush map)
- 3 pool for block storage for cloud services (replica 3)
  - VM
  - volumes
  - images
- 1 pool for default metrics



Benchmarking with IOZONE

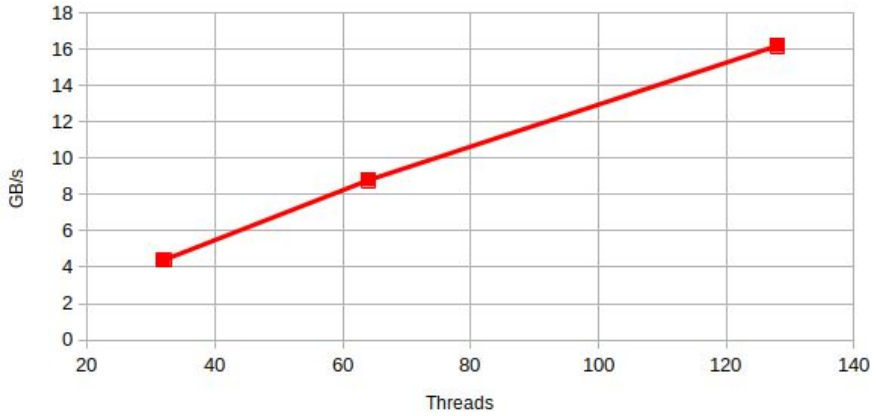
Block size	Threads	Throughput [GB/s] **WRITE**
1M	32	4.2
1M	64	8.3
1M	128	15.7

Write Performance



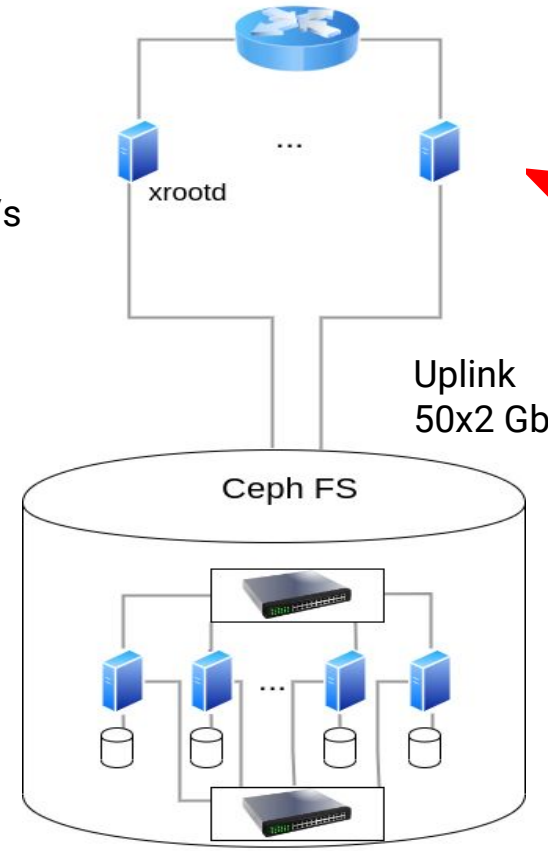
Block size	Threads	Throughput [GB/s] **READ**
1M	32	4.4
1M	64	8.8
1M	128	16.2

Read Performance



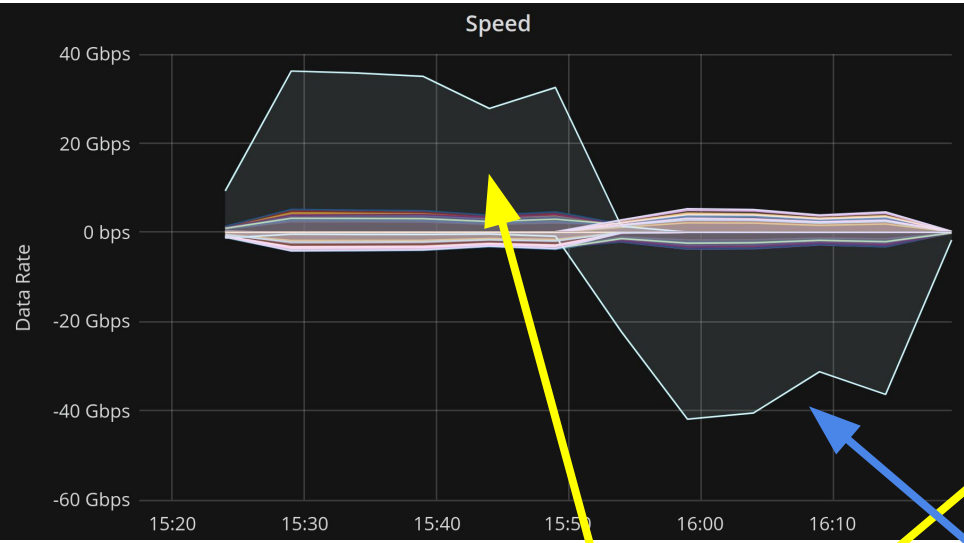
Pledged Space for Alice

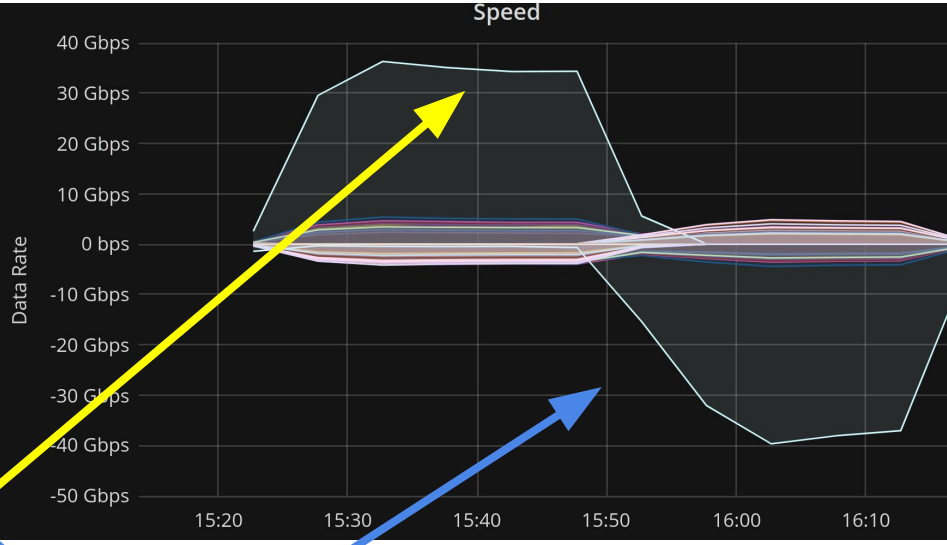
6 server
10x2 Gb/s



- 1.6PB Pledged space
- Special thanks to **Francesco Noferini** and **Latchezar Betev**
- 6 xrootd servers + redirectors

Benchmark on Alice FS

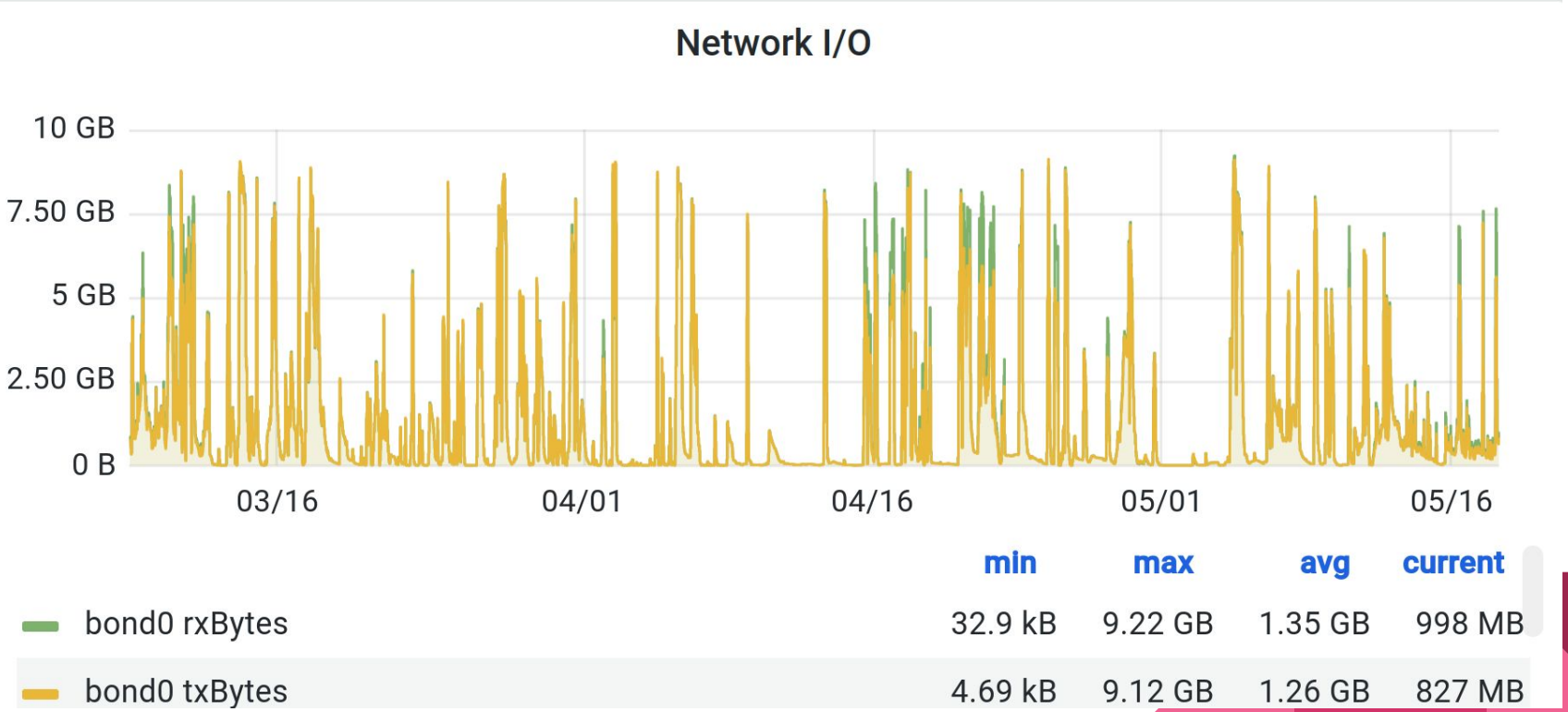




Uplink - Out

Uplink - In

Pledged Space for Alice



Conclusions

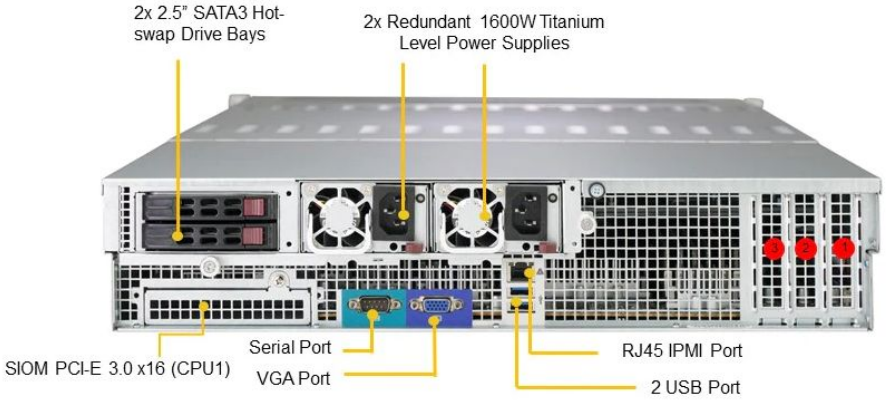
- In the last 2 years technology investigations on distributed fs lead us to study and deploy a dedicated Ceph cluster
- Ceph provide a lot of features that make it a viable solution for several use cases
- Deployed a fraction of the Alice pledge successfully in terms of performance and stability
- Network Access OK
- Massive client local access (worker nodes) not fully exploited
- Tape backend not available

Backup

Servers Rear View

SuperStorage SSG-6029P-E1CR24L

(Rear View – System)



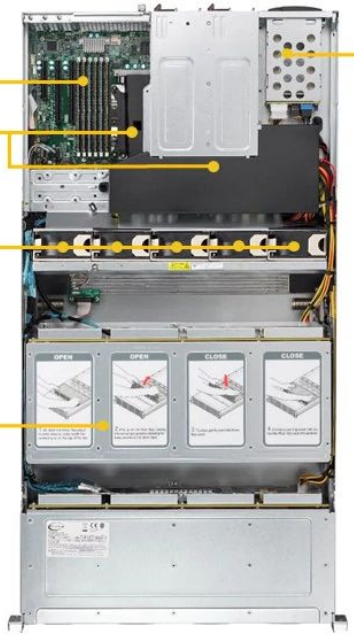
Location	Description
1	SLOT 1 PCI-E 3.0 x 8 (CPU2)
2	SLOT 2 PCI-E 3.0 x16 (CPU2)
3	SLOT 3 PCI-E 3.0 x16 (CPU2)

SUPERMICRO © Super Micro Computer, Inc. Information in this document is subject to change without notice.

Servers Top View

SuperStorage SSG-6029P-E1CR24L

(Top View – System)



© Super Micro Computer, Inc. Information in this document is subject to change without notice.

Benchmarking with IOZONE

Block size	Threads	Throughput [GB/s] **WRITE**
1M	32	4.2
1M	64	8.3
1M	128	15.7

Block size	Threads	Throughput [GB/s] **READ**
1M	32	4.4
1M	64	8.8
1M	128	16.2

Benchmarking - Mixed

Block size	Threads	Throughput [GB/s]
1M	32	5.2 **READ**
1M	64	8.9 **WRITE**

Costs comparisons

- RAW Space = 18TB * 288 → 5184TB
- Net Space with EC 8 + 4 → 3473TB → 71 euro /TB (87 IVA inclusa)
- Net Space Replica 2 → 2600TB→ 95 euro /TB (116 IVA inclusa)
- Net Space Replica 3 → 1728TB→ 143 euro /TB (175 IVA inclusa)