nffa-di

Piano Nazionale di Ripresa e Resilienza

AREA SCIENCE PARK

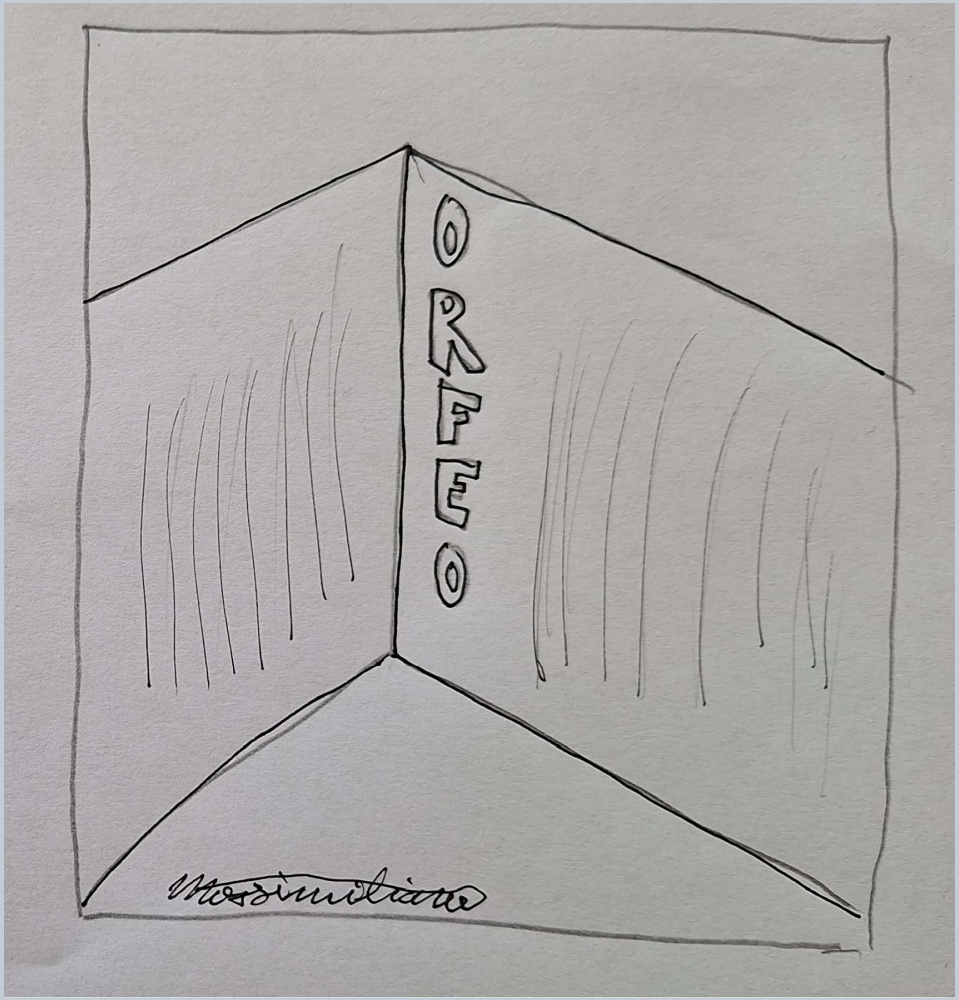
Data Infrastructure

Lecture 4: Case Study: ORFEO

Federica Bazzocchi
24/4/2026

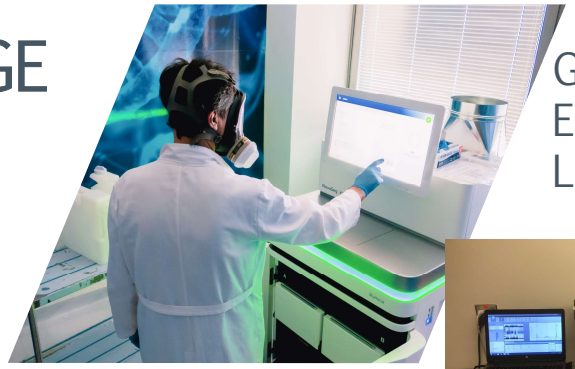
LECTURE 4 OUTLINE:

- ORFEO: the origin and the future
- ORFEO: the storage facility
- ORFEO: the computing facility
- ORFEO: the network
- ORFEO: benchmarking

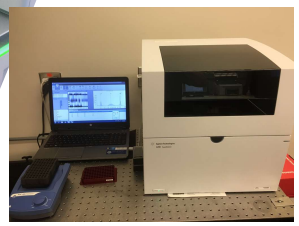


Open
Research
Facility for
Epigenomic and
Other

LAGE



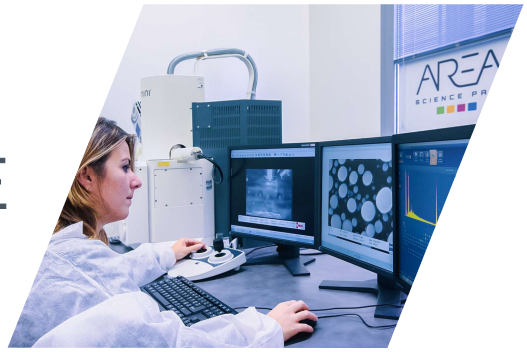
GENOMICS AND EPIGENOMICS LABORATORY



LADE

DATA ENGINEERING LABORATORY

LAME

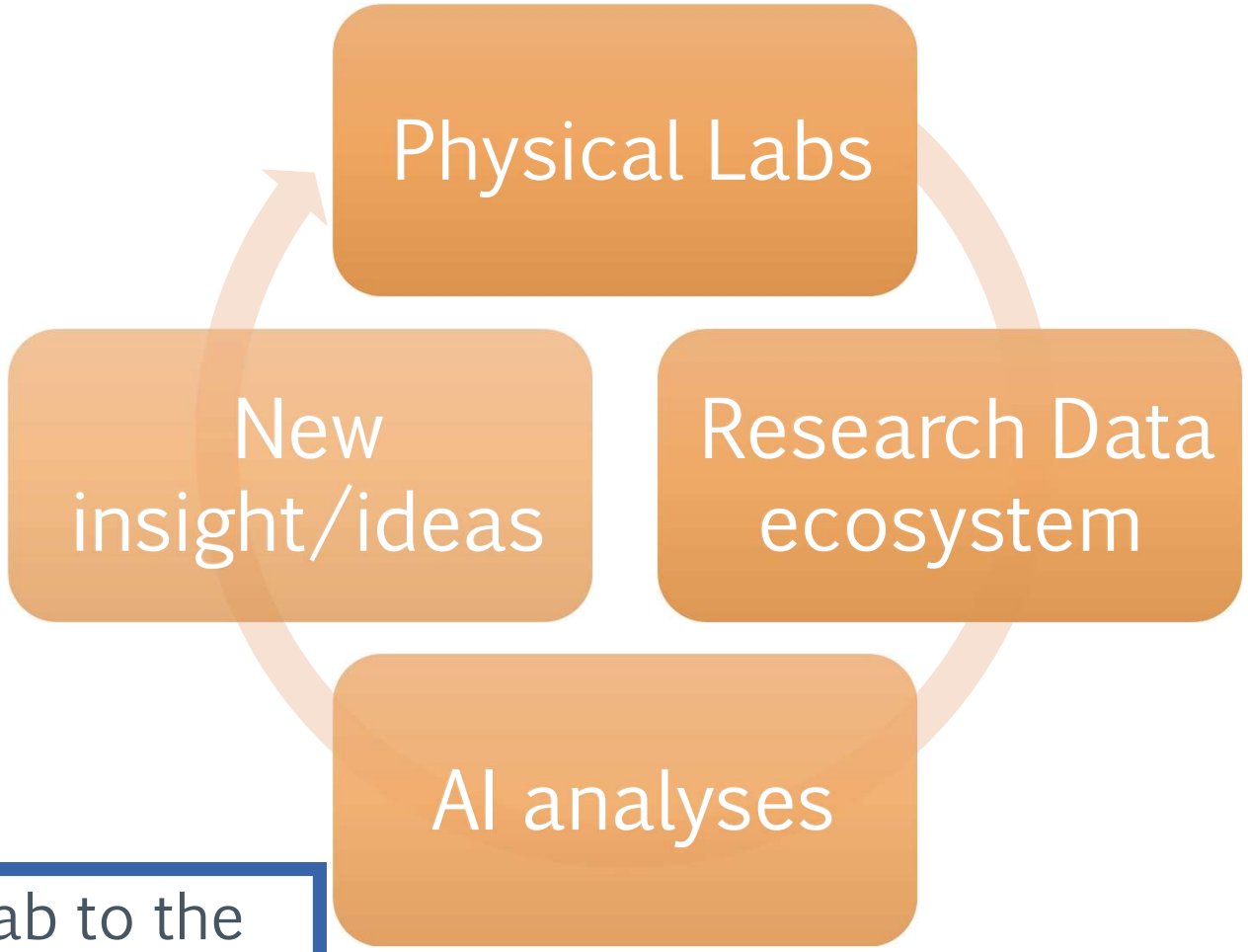


ELECTRON MICROSCOPY LABORATORY

SERVICES for RESEARCH INFRASTRUCTURES

SERVICES for OTHER STAKEHOLDERS

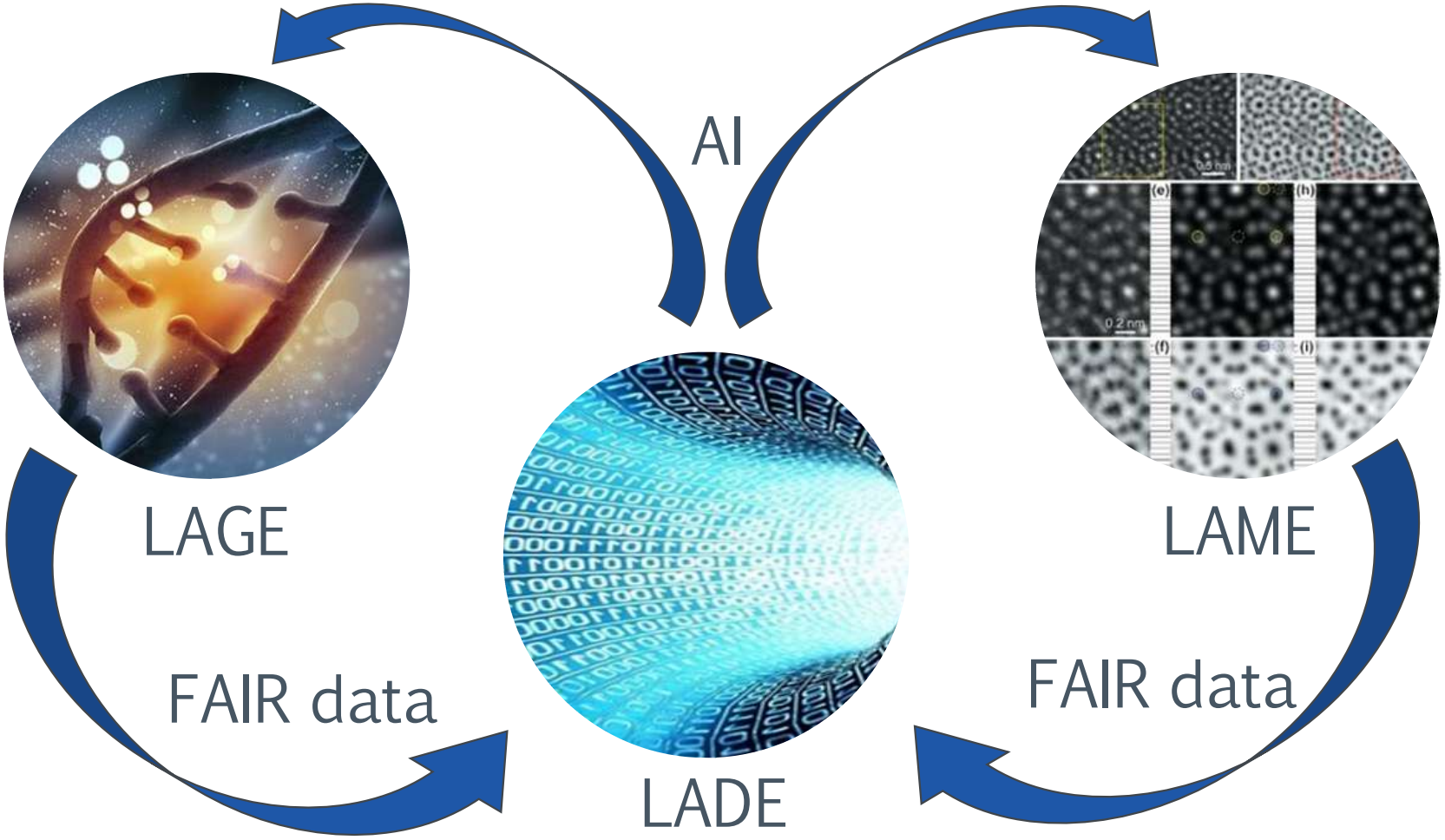
DATA ECOSYSTEM



From the lab to the lab

RIT ECOSYSTEM

FAIR-by-design Laboratory workflows



ORFEO

Thought from the beginning to be

flexible scalable

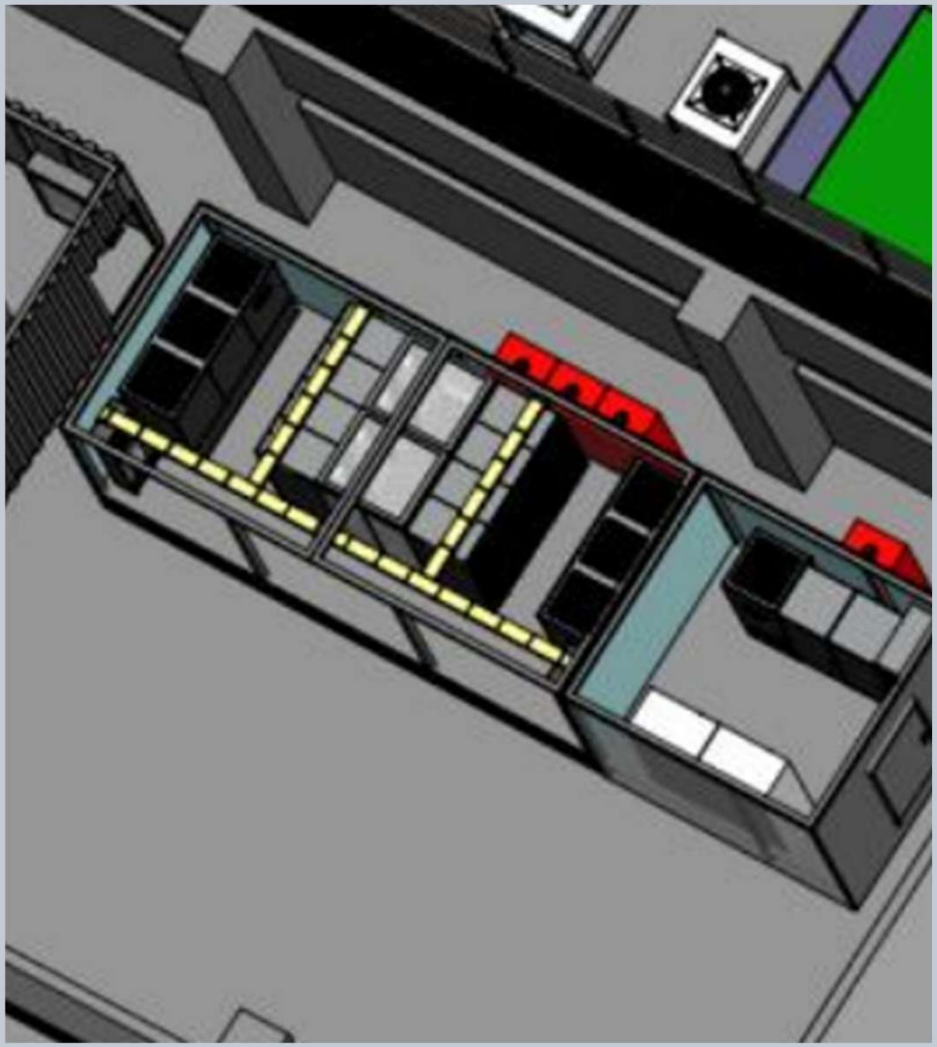
Storage facility as well as computing one

ORFEO BEFORE THE PNNR

- **20 regular nodes** for a total of **80 TFLOPs** computations
- **6 GPU nodes** with Nvidia **V100** and **A100** accelerators for 400 TFLOPs for accelerated data analysis
- **9 nodes** for **cloud** services
- **100 Gbit/s** low latency **InfiniBand** connection
- Redundant 25 Gbit/s connection for **high availability**
- **12 storage nodes** for a distributed CEPH parallel filesystem,
  Total raw space of **1.6 PiBs** on regular **HDD** and 280 TiBs on fast SSDs
- **1 SAN appliance** for **2.8 PiBs** of long **term preservation storage**

ORFEO THANKS TO THE PNNR

- A **new server room** capable of serving 125 kW to appliances with **reduced carbon footprint** and OPEX thanks to the chosen refrigeration technologies
- **+ 13 regular nodes** for an additional computation capacity of more than **50 TFLOPs**
- **+ 3 GPU nodes** with 8 accelerators each (Nvidia **H100**) More than 1.5 PFLOPs for GPU accelerated data analysis.
- **+ 200 Gbit/s** low latency **InfiniBand** connection
- **+ 12 storage nodes** for an additional **2.9PiBs** of raw **HDD** storage and 340 TiBs of NVMe storage



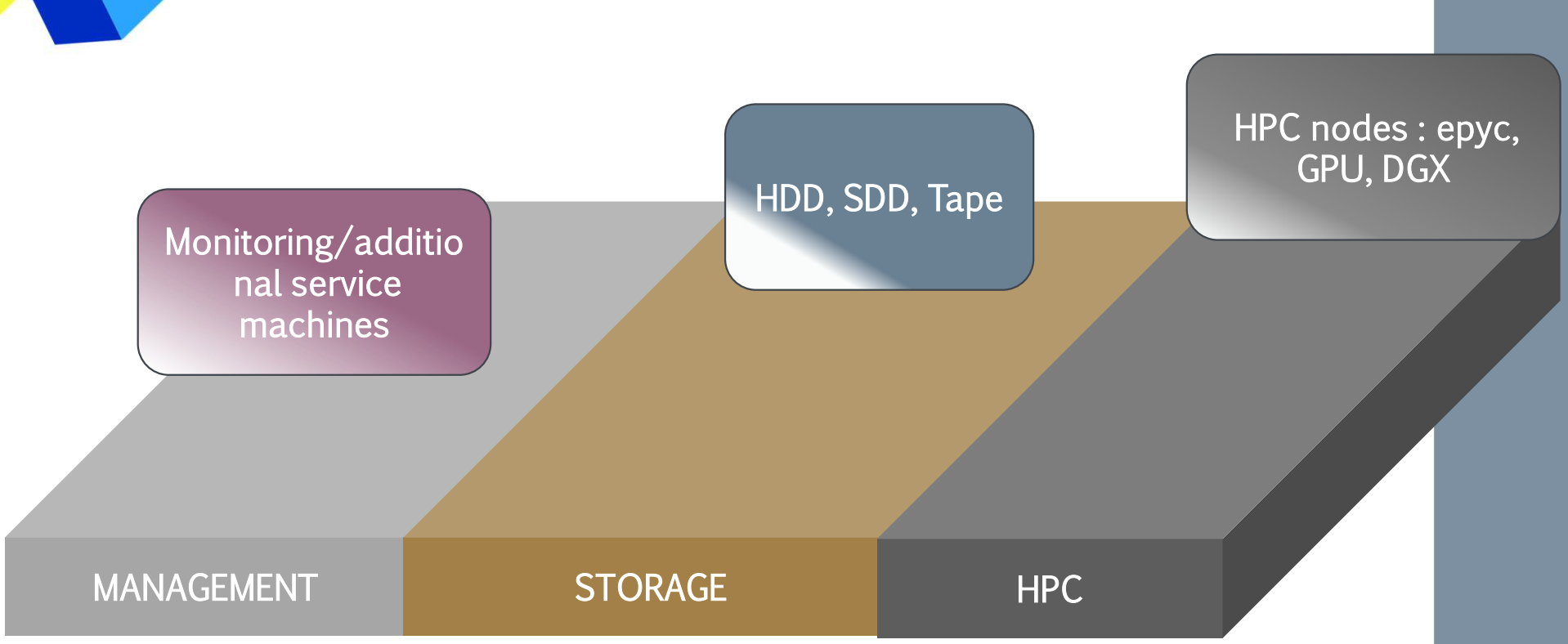
ORFEO today !


IT
Module
Power
module
Open
Research
Facility for
Epigenomics and
Others
IT
Module
(new)

Core Hardware Components

- Storage Systems: HDD, SSD, NAS, SAN
- Servers: Compute power for running applications and processing data
- Networking: Switches, routers, firewalls, load balancers
- Data Centers: Physical facilities for computing and storage
- Additional Systems: Backup and Disaster Recovery Systems

Hardware Component



ORFEO: THE STORAGE FACILITY

I/O SYSTEM ON ORFEO

The storage is provided mainly by:

- A CEPH cluster consisting of 23 nodes equipped with HDD, SSD, and NVMe technologies (distributed)
- Dell PowerVault SAN, connected via Fibre Channel to an NFS server (not distributed)
- Soon it will be deployed a Dell ML3 Tape Library to archive data on LTO-9 cartridges, each offering up to 45 TB of capacity.

ORFEO storage after PNRR

	Standard storage (HDD)	Fast storage (SSD)	Super Fast storage (NVME)	Long Term Storage
Node capacity	(Gen1) 212 TB (Gen2) 294 TB (Gen3) 294 TB	84 TB	2 x 1.6 TB 2 x 14 TB (Gen3 nodes)	2962.5 TB
# of Nodes	6 x Gen1 + 2 x Gen2 + 11 x Gen3	4 nodes	2 devices each node, **Gen3 nodes have 2x 14 TB !**	Served via NFS server with a SAN.
Storage provider	CEPH	CEPH	CEPH	Network FS (NFS)
Increment wrt before PNRR	+ 130 % more HDD storage		+ 1340 % more NVME storage	
RAW storage	~ 5684 TB	~ 340 TB	~ 340 TB	~ 2962 TB

CEPHFS

LTS

~ 9,2 PB

```
[fbazzocchi@login02 orfeo]$ ls -lh
total 0
drwxr-xr-x 13 root root 167 Oct 25 10:34 LTS
drwxrwxr-x  7 root root   5 Feb 27 14:51 cephfs
lrwxrwxrwx  1 root root  18 Nov 18 17:10 fast -> /orfeo/cephfs/fast
lrwxrwxrwx  1 root root  17 Aug 12  2024 opt -> /orfeo/cephfs/opt
lrwxrwxrwx  1 root root  21 Aug 12  2024 scratch -> /orfeo/cephfs/scratch
```

Remember from lecture 3

```
[fbazzocchilogin02 orfeo]$ ls
LTS  cephfs  fast  opt  scratch

[fbazzocchi@login02 orfeo]$ stat  f LTS
  File: "LTS"
    ID: fd0000000000 Namelen: 255     Type: xfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 3915776    Free: 1820163    Available: 1820163
Inodes: Total: 7864320    Free: 7656589
[fbazzocchi@login02 orfeo]$ stat  f cephfs/
  File: "cephfs/"
    ID: b92f1c82ffffffff Namelen: 255     Type: ceph
Block size: 4194304    Fundamental block size: 4194304
Blocks: Total: 1290859070 Free: 1033832946 Available: 1033832946
Inodes: Total: 463971472  Free:  1@
```

Different file system, different choices of block size

Ceph filesystem on ORFEO:

- /orfeo/cephfs/home
  - Host the user's home.
  - A quota of 200GB and 10^6 files is enforced.
  - It runs on **HDD with replica 3 data protection.**

- /orfeo/cephfs/scratch
  - It is large area intended to be used to store data that need to be elaborated.
  - It is also physically located on ceph large FS and exported via infiniband to all the computational nodes.
  - It runs on **HDD** and **6+2 EC data protection**

- /orfeo/cephfs/fast
  - is a fast space available for each user, on all the computing nodes.
  - is intended to be a **fast scratch area** for data intensive application.
  - It runs on **SSD with replica 3 data protection.**

- replica 3 data protection.
  - Stores three full copies of each data block
  - High redundancy and fast recovery
  - Consumes 3x storage capacity
  - Simple to implement and manage

Useful for services that cannot permit long breakdown

- 6+2 EC (Erasure Coding) data protection
  - Splits data into 6 chunks and adds 2 parity chunks
  - Any 6 of the 8 chunks are enough to recover data
  - Offers fault tolerance with ~1.33x storage overhead
  - More space-efficient but more CPU-intensive for recovery

Remember RAID from lecture 2

CAP theorem (Brewer theorem)

Parenthesis

Consistency-Availability-Partition Tolerance

Trade-offs between three key properties in distributed system

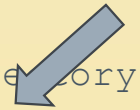
Consistency: all nodes in the system have the same data view

Availability: guartantees that every request receives a response. Even if some nodes fail, system remains operational and responsive

Partition Tolerance: system ability to continue functioning despite network partitions or communication breakdowns between nodes

The CAP theorem states that a distributed system can only guarantee two of these three properties at any given time, not all three simultaneously.

```
[fbazzocchi@login02 cephfs]$ stat fast
  File: fast
  Size: 17              Blocks: 0          IO Block: 65536  directory
Device: 0,55    Inode: 1100038035299  Links: 18
Access: (0755/drwxr-xr-x)  Uid: (    0/    root)   Gid: ( 1002/    area)
Access: 2024-11-08 14:17:39.301684093 +0100
Modify: 2024-11-08 10:28:40.615166140 +0100
Change: 2025-02-26 09:40:57.565780618 +0100
 Birth: 2024-11-08 14:17:39.301684093 +0100

[fbazzocchi@login02 cephfs]$ stat scratch
  File: scratch
  Size: 17              Blocks: 0          IO Block: 65536  directory
Device: 0,55    Inode: 1099511627777  Links: 19
Access: (0755/drwxr-xr-x)  Uid: (    0/    root)   Gid: (    0/    root)
Access: 2023-06-01 14:43:48.660963613 +0200
Modify: 2024-10-25 12:38:07.445574151 +0200
Change: 2024-10-25 12:38:07.445574151 +0200
 Birth: 2023-06-01 14:43:48.660963613 +0200
```



ORFEO: Long Term Storage (LTS) I/II

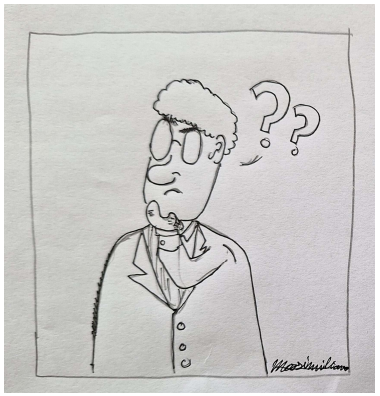
- A NAS for ORFEO Cluster (Network Attached Storage)
- Internally: entry level SAN (Storage Area Network iSCSI on Fibre Channel)
- Raw capacity 3 PB
- Served via NFS (Network File System) with 2 x 25Gbit link
- It is intended for long-term storage of final processed dataset

ORFEO: Long Term Storage (LTS) II/II

- Volumes' file systems can be set up as:
  - XFS: fast but limited to 128 TB maximum size per volume.
  - XFS+LVM: slightly slower, can stripe data to multiple PowerVault volumes.
  - ZFS: slower than XFS+LVM, can stripe data to multiple volumes, performs extensive checksums on data and metadata, CoW (Copy on Write) allows instant read-only snapshots. Ideal for archives.

What is PowerVault?

- Dell PowerVault is a line of affordable, entry-level data storage systems
- Offers both DAS (Direct Attached Storage) and SAN configurations
- Common features include:
  - High-capacity storage
  - Redundancy with RAID support
  - Easy-to-use management interfaces
  - Compatibility with Dell servers and backup software



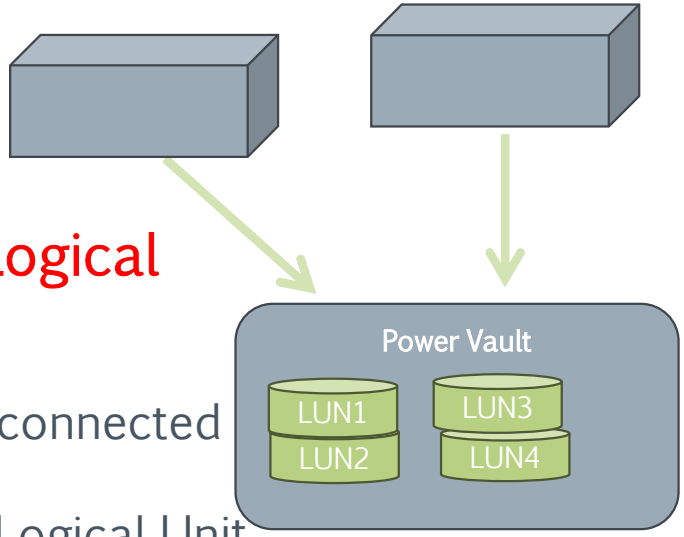
SAN (Storage Area Network)	NAS (Network Attached Server)
Use fibre channel or SCSI	Use internet to transport data
Provides block-level storage access	Provides file-level storage access
Server mantains SAN FS	Has its own FS (NFS)
Used for high-performance enterprise applications	Common in SMBs (Server Message Block) and for file sharing

SMB (Server Message Block) is a client-server protocol that regulates access to files and entire directories, and network resources such as printers, routers, and network-shared interfaces. Additionally, the SMB protocol can manage the exchange of information between different processes within a system

SMB: in cosa consiste il Server Message Block? - IONOS

3 tiers storage stracture:

- SAN
- FS
- NFS



The SAN Layer: Connecting Initiators to Logical Units

- Suppose to have 2 Physical Servers (Initiators) connected central PowerVault SAN (Target).
- The SAN does not see "files." It manages LUNs (Logical Unit Numbers), which are chunks of blocks.
- Mapping Example:
  - Server A: Maps LUN_01 and LUN_02.
  - Server B: Maps LUN_03 and LUN_04.
- Isolation & Masking
  - Through "LUN Masking," the SAN ensures Server A cannot see or corrupt the blocks belonging to Server B.
  - Redundancy: If Server A fails, the SAN can "remap" LUN_01 and LUN_02 to a backup server instantly.

The PowerVault (The Target) : it is a specialized hardware entity dedicated to data persistence.

- ❑ It houses physical drives (HDDs or SSDs), redundant power supplies, and storage sontrollers.
- ❑ The controllers aggregate physical disks into **RAID groups** (for redundancy) and then carve them into logical slices called LUNs (Logical Unit Numbers).
- ❑ It "presents" or "exposes" these LUNs to the SAN network, waiting for an authorized server to request access to the raw blocks.

The Servers (The Initiators): manage two distinct types of storage resources:

A. **Local Disks (Internal to the Server):** their role is limited

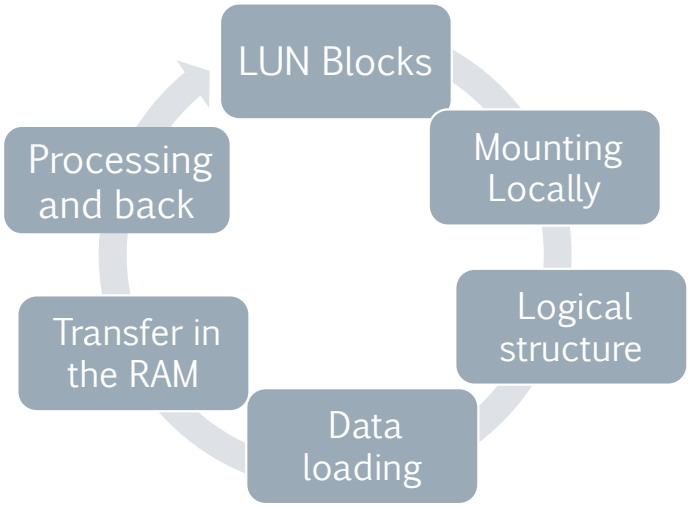
- ❑ They store the Operating System (e.g., Linux, Windows Server).
- ❑ They allow the server to turn on and load the drivers needed to "talk" to the SAN.
- ❑ Local Logs & Swap: Used for the server's internal "diaries" and temporary memory overflow.
- ❑ **Safety:** If the SAN network fails, the server remains powered on and manageable because the OS is stored locally.

B. **System RAM (The Workspace):** it is the server's high-speed "workbench."

- ❑ When a server requests data from the PowerVault, the blocks travel across the SAN and are loaded into the RAM.
- ❑ The CPU can only process data that is currently residing in the RAM.

The Operational Flow: From Block to RAM

1. **LUN Management**: PowerVault defines a set of blocks as **LUN_01**.
2. **Mounting**: The Server (Initiator) connects to the SAN, recognizes **LUN_01**, and "mounts" it as if it were a local drive.
3. The Server writes a logical structure (like XFS or ZFS) onto that LUN.
4. **Data Loading**: An application requests a file. The Server's Kernel translates this into: *"Give me blocks 10 through 20 from LUN_01."*
5. **Transfer**: The PowerVault sends the blocks. They pass through the server's HBA (Host Bus Adapter) and land directly in the **RAM**.
6. **Processing**: The CPU processes the data from the RAM. Any changes are sent back across the SAN to be written permanently on the PowerVault.



But who is the "brain" behind all the operations among server and PV?

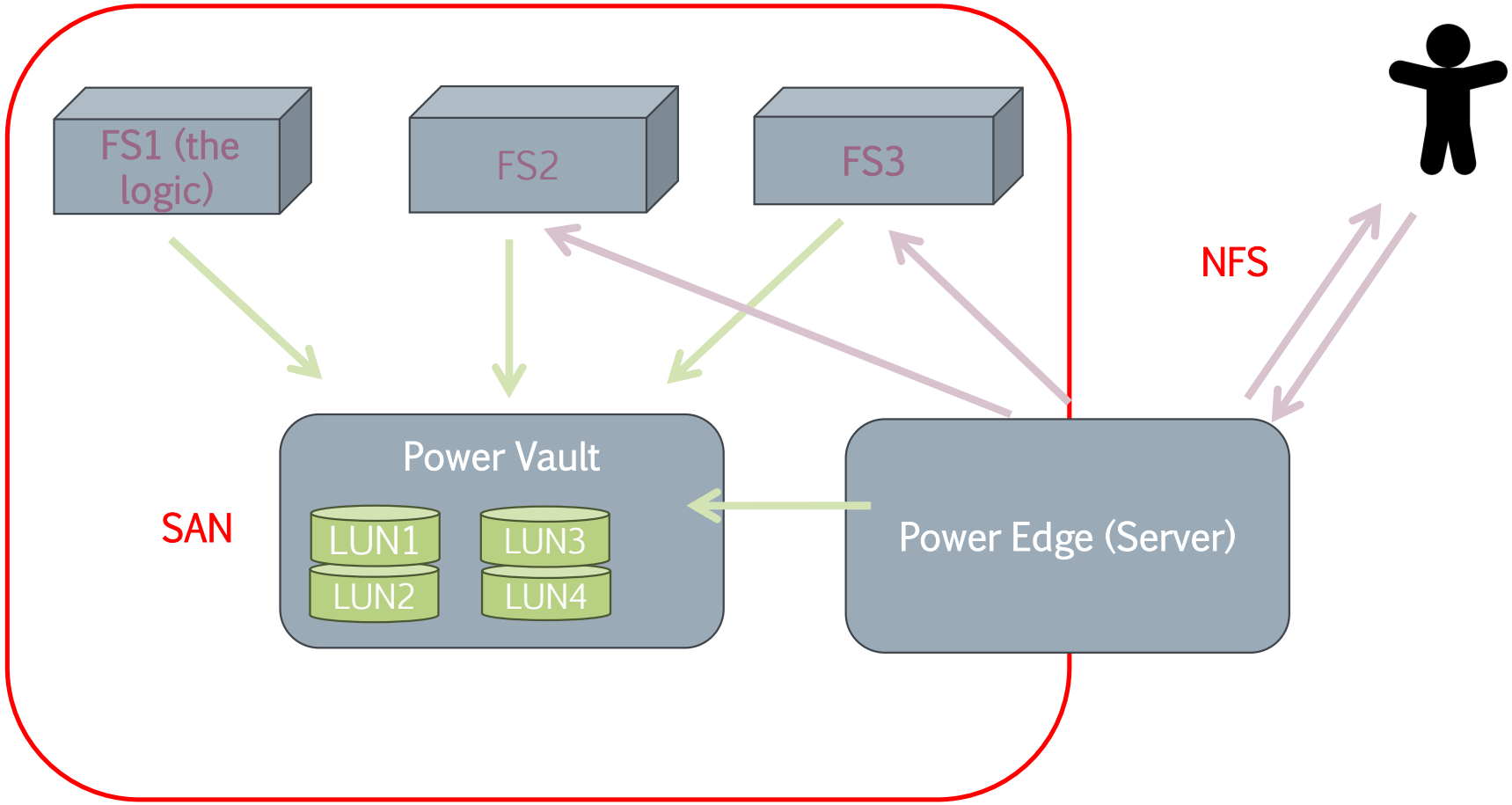
The Storage Head Server

1. The "Head": While the PowerVault is the physical container of disks (the "Body"), the Storage Head Server is the “brain” layer.

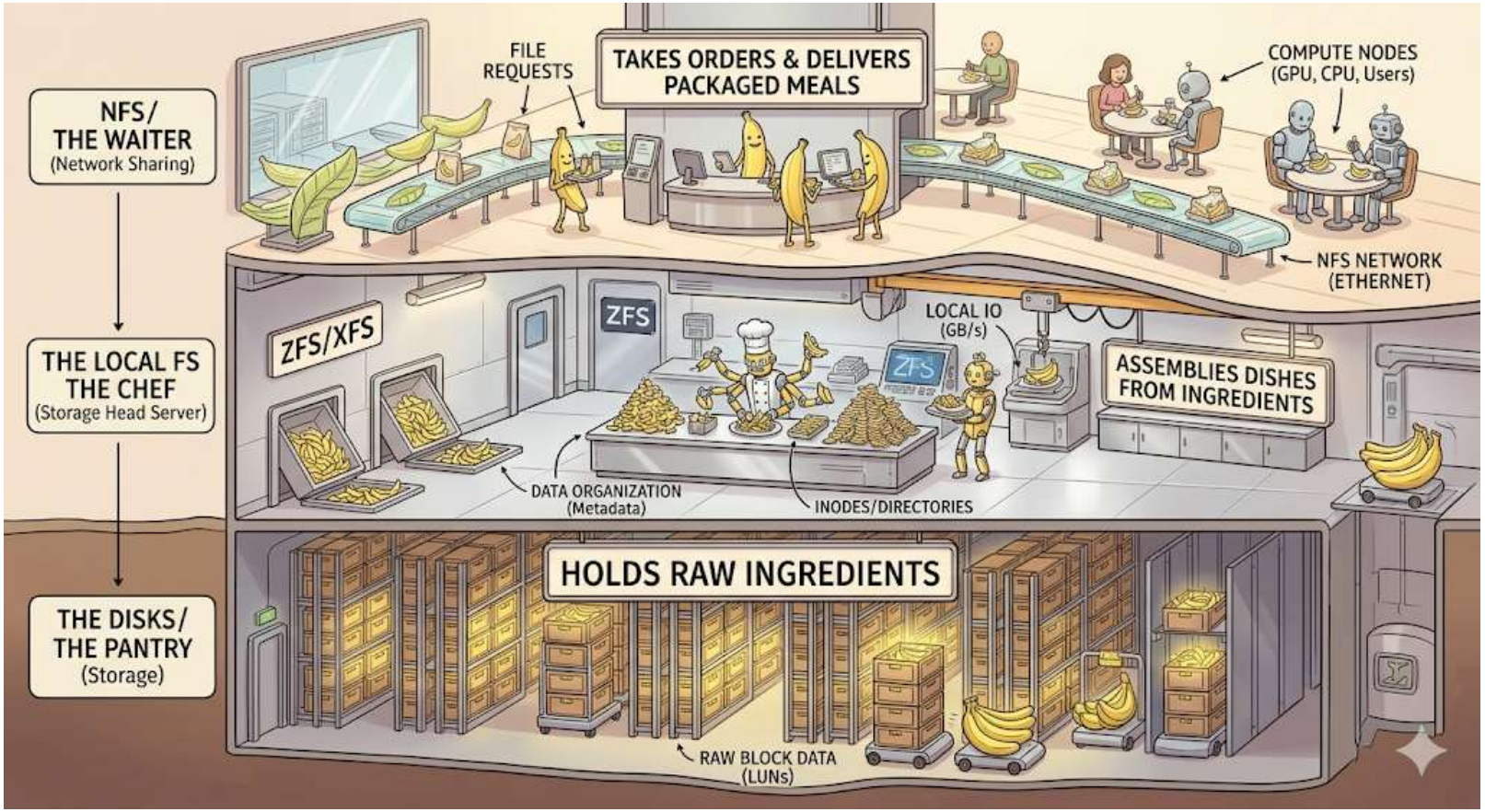
- ❑ Physical Nature: It is a high-performance server (e.g., a Dell PowerEdge or similar) equipped with specialized HBA cards to talk to the SAN and High-speed NICs (Network Interface Cards) to talk to the users.
- ❑ Role in the SAN: It acts as the primary Initiator. It is the only machine that "claims ownership" of the LUNs coming from the PowerVault.
- ❑ Role in the LAN/HPC Network: It acts as the NFS Server. It takes the storage it has mounted and "serves" it to the users

Feature	PowerVault (The Array)	Storage Head Server (The Head)
Data Visibility	Sees Blocks (Sector 0 to 1,000,000).	Sees Files (/data/experiment_01.csv).
Intelligence	Manages RAID and Disk Health.	Manages the File System (ZFS, XFS) and Permissions (ACLs).
Communication	Talks SCSI / Fibre Channel (Block protocols).	Talks NFS / SMB (File protocols).

3 tiers storage stack

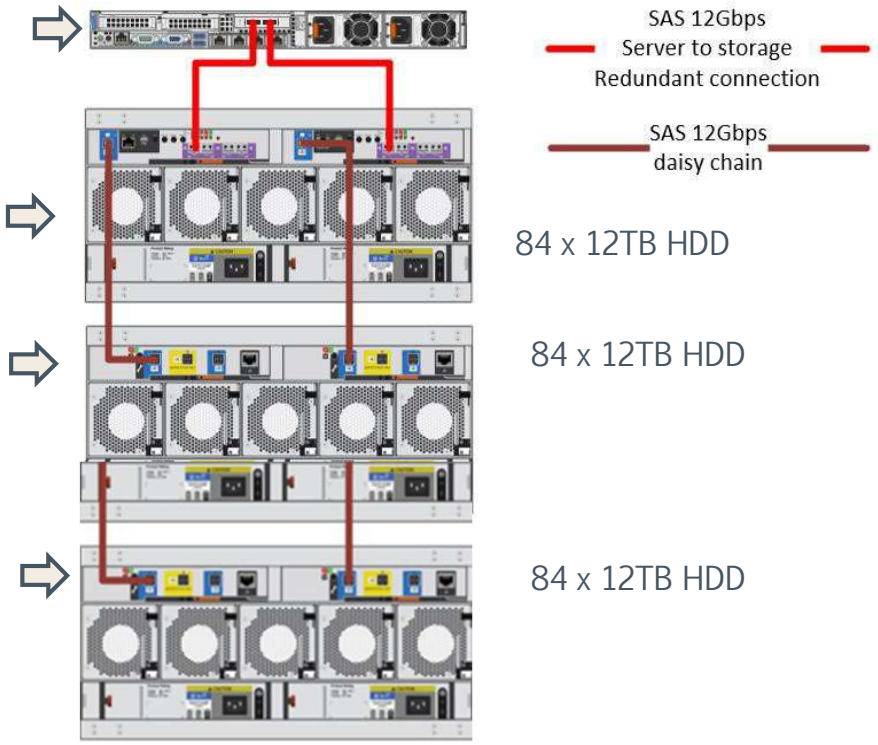


An analogy: the restaurant



LTS hardware

- Dell EMC PowerEdge R640
- Dell EMC PowerVault ME4084
- Dell EMC PowerVault ME4084
- Dell EMC PowerVault ME4084



Comparison: XFS vs XFS + LVM vs ZFS

XFS	High-performance 64-bit journaling file system	Fast for large files and streaming workloads	No native volume management or snapshots
XFS + LVM	Combines XFS with Logical Volume Manager	Adds flexibility: resizing volumes, snapshot support	More complexity, separate layers to manage
ZFS	Integrated file system and volume manager	Built-in RAID, snapshots, checksums, compression	Higher memory usage, more advanced features

LVM (Logical Volume Manager)

storage management tool primarily used in Linux operating systems. It allows managing disk space more flexibly compared to traditional partitions. With LVM it is possible creating, resizing and moving logical volumes without modifying physical partitions.

Main Components

1. Physical Volumes (PV): These are the physical disks or partitions added to LVM.
2. Volume Groups (VG): A group of PVs treated as a single storage unit.
3. Logical Volumes (LV): These are the virtual partitions created within a VG. These volumes can be resized and moved easily.

Key Features

- Flexibility: resize logical volumes without interrupting the system.
- Space Management: aggregate space from multiple disks into a single volume.
- Snapshots: create snapshots of logical volumes for backup and recovery.

[fbazzocchi@login02 ~]$ ssh 10.128.2.231

nfs01

Welcome to new NFS01

Long term storage data are located in /srv/
nf
Last login: Wed Apr 9 11:48:16 2025 from 10.128.2.81

[fbazzocchi@nfs01 ~]$ ls -lh
total 237M

```
lrwxrwxrwx.  1 fbazzocchi area            30 Jul  2  2021 local_scratch -> /local_scratch/area/fbazzocchi
-rw-r--r--.  1 fbazzocchi area          148M Aug  4  2022 log_CBM.txt
-rw-r--r--.  1 fbazzocchi area          2.1K May 10  2023 monitoring.py
lrwxrwxrwx.  1 fbazzocchi area            24 Jul  2  2021 scratch -> /scratch/area/fbazzocchi
drwxr-xr-x.  2 fbazzocchi area             0 Apr  4  2022 seaborn-data
-rw-r--r--   1 fbazzocchi fbazzocchi 8.9K Jan 12  2024 simple_example_write1.hdf5
-rw-r--r--   1 fbazzocchi fbazzocchi   13K Jan 12  2024 simple_example_write2.hdf5
-rw-r--r--   1 fbazzocchi fbazzocchi 8.1K May 28  2024 slurm-434546.out
-rw-r--r--   1 fbazzocchi fbazzocchi   512 Apr  5 20:48 small-write
drwxr-xr-x.  4 fbazzocchi area             2 Nov 25  2022 snakemake
-r-xr-xr-x.  1 fbazzocchi area          4.4K Jul 28  2022 ss.csv
lrwxrwxrwx.  1 fbazzocchi area            24 Jul  2  2021 storage -> /storage/area/fbazzocchi
```

In the root?

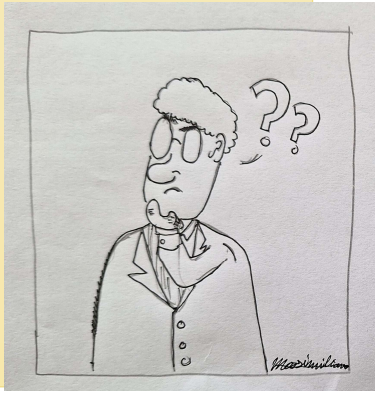
```
[fbazzocchi@nfs01 /]$ ls -lh
total 32K
dr-xr-xr-x.      2 root root      6 Oct   2   2024 afs
lrwxrwxrwx.      1 root root      7 Oct   2   2024 bin -> usr/bin
dr-xr-xr-x.      5 root root 4.0K Dec 19 17:00 boot
drwxr-xr-x      28 root root 7.4K Apr   7 13:20 dev
drwxr-xr-x.    103 root root 8.0K Apr   4 10:26 etc
drwxr-xr-x.      2 root root      6 Jul 17   2024 fast
drwxr-xr-x.      2 root root      6 Oct   2   2024 home
lrwxrwxrwx.      1 root root      7 Oct   2   2024 lib -> usr/lib
lrwxrwxrwx.      1 root root      9 Oct   2   2024 lib64 -> usr/lib64
drwxr-xr-x.      2 root root      6 Oct   2   2024 media
drwxr-xr-x.      2 root root      6 Oct   2   2024 mnt
drwxr-xr-x.      3 root root     18 Feb 13 12:17 opt
drwxrwxr-x.      3 root root     47 Jul 17   2024 orfeo
dr-xr-xr-x    1174 root root      0 Apr   4 10:25 proc
dr-xr-x---.     16 root root 4.0K Apr   8 16:21 root
drwxr-xr-x      40 root root 1.3K Apr   8 15:21 run
lrwxrwxrwx.      1 root root      8 Oct   2   2024 sbin -> usr/sbin
drwxr-xr-x.     35 root root 4.0K Jan 20 16:24 srv
dr-xr-xr-x      13 root root      0 Apr   4 10:25 sys
drwxrwxrwt.     13 root root 4.0K Apr   9 20:26 tmp
lrwxrwxrwx.      1 root root     19 Jul 17   2024 u -> /orfeo/cephfs/home/
drwxr-xr-x.     12 root root    144 Nov 23 13:05 usr
drwxr-xr-x.     19 root root 4.0K Nov 23 13:30 var
```

Note symbolic link!

[fbazzocchi@login02 ~]$ ssh 10.128.2.231

nfs01

```
[fbazzocchi@nfs01 ~]$ ls /srv
STRAS                          illumina_decode          nep
read_the_docs
analisi_da_consegnare          illumina_run             onp_run_1
soundsafe_long_term_storage
borg_repos                     lade_long_term_storage   opt                        storage
burlo_long_term_storage        lage_archive             orfeo_home_backup          tmp_stuff
burlo_long_term_storage2       lage_long_term_storage   orfeo_replicated_share
zfs_collaborations
cdslab_long_term_storage       lala_storage             orfeo_repo                 zfs_lage
cro_long_term_storage          lame_long_term_storage   plus
decolibus                      laptop_backup            post_run
iga_long_term_storage          long_term_storage        proxmox
```



"Ghost Mount" Pattern

Balancing Path Consistency and Access Control in HPC

- "Ghost" Mount Points
  - **Definition:** Local directories (e.g., `/srv/`) created on all nodes but left unmounted on restricted servers.
  - The structure exists physically on the local disk but the remote data is not connected.
- Why Maintain Empty Directories?
  - Using automation tools (Ansible/Terraform) to ensure identical file paths across Login, Compute, and Storage nodes (path consistency)
  - Simplifies pipeline development; users don't need to change paths based on the node they are using.

Network Zoning & Security Implementation

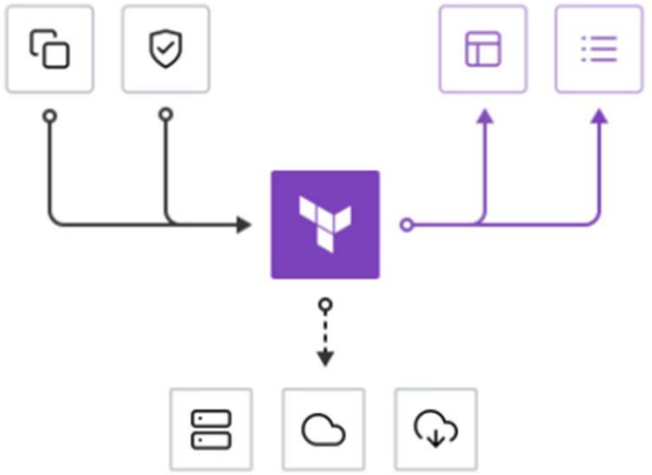
Securing Data via Segmented NFS Exports

- Network Segmentation
  - Implemented to isolate the "Login Zone" from the "Data Transfer Zone."
- The Three Pillars of This Architecture:
  - **Security:** Creates a DMZ; if a public-facing login node is compromised, the core NFS storage remains unreachable.
  - **I/O Performance:** Prevents massive data transfers (e.g., 10TB+ syncs) from saturating the management/SSH bandwidth.
  - **Granular Export Control:** Restricted access via

What is Terraform?

Terraform is an infrastructure as code tool that lets you build, change, and version infrastructure safely and efficiently. This includes low-level components like compute instances, storage, and networking; and high-level components like DNS entries and SaaS features.

Terraform | HashiCorp Developer

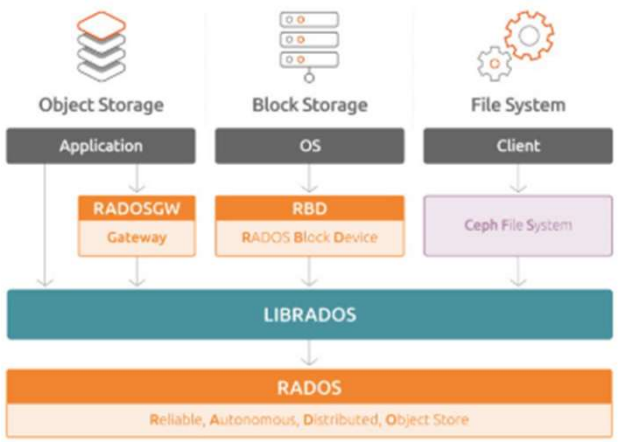


Object Storage

Flat Object Spaces

- **Flat Namespace:** Unlike NFS/CephFS, Object Storage uses a flat hierarchy (Buckets and Objects). No nested directory performance penalties.
- **RESTful Access:** Communication happens over **HTTP/HTTPS** via APIs (S3 or Swift) rather than kernel-level mounting.

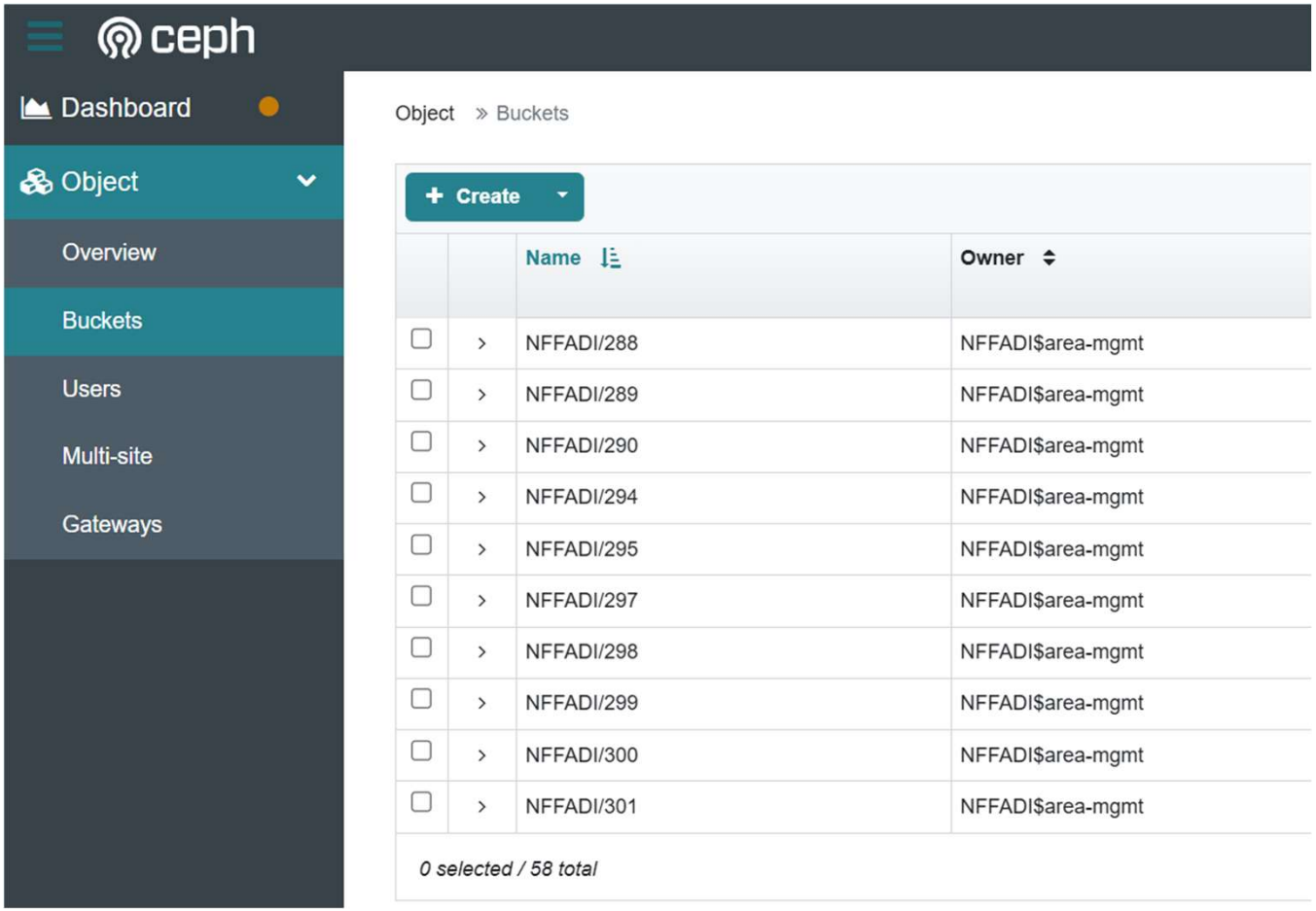
- **Infinite Scalability:** Designed to handle billions of files (images, audio, logs) where traditional filesystems would struggle
- Data is accessible from any node in the cluster (or even outside) via a URL/Endpoint, bypassing the need for complex network mounts.
- **RGW (RADOS Gateway) a**cts as the translator between S3/Swift API calls and the underlying Ceph RADOS cluster.



RGW_ENDPOINT = https://buckets.areasciencepark.it

USER_ID= "***"

ACCESS_KEY='***'
SECRET_KEY='***'



The Importance of Documentation

Parenthesis

What you could call the Cluster (Data) Management Plan

- File systems
- Virtual File Systems
- Links
- Network configurations
- Nodes configurations
- ……..

A jungle

```
[fbazzocchi@login02 /]$ stat -f /orfeo/cephfs/home/area/
  File: "/orfeo/cephfs/home/area/"
    ID: b92f1c82ffffffff Namelen: 255     Type: ceph
Block size: 4194304    Fundamental block size: 4194304
Blocks: Total: 1290859070 Free: 1028595412 Available: 1028595412
Inodes: Total: 469906844  Free: -1


[fbazzocchi@login02 /]$ stat -f /u/area/
  File: "/u/area/"
    ID: b92f1c82ffffffff Namelen: 255     Type: ceph
Block size: 4194304    Fundamental block size: 4194304
Blocks: Total: 1290859070 Free: 1028594500 Available: 1028594500
Inodes: Total: 469907422  Free: -1


[fbazzocchi@nfs01 /]$ stat u
  File: u -> /orfeo/cephfs/home/
  Size: 19              Blocks: 0          IO Block: 4096   symbolic link
Device: 803h/2051d      Inode: 6283        Links: 1
Access: (0777/lrwxrwxrwx)  Uid: (    0/    root)   Gid: (    0/    root)
Access: 2025-04-09 10:36:45.962915056 +0200
Modify: 2024-07-17 17:17:30.081319370 +0200
Change: 2024-07-17 17:17:30.081319370 +0200
 Birth: 2024-07-17 17:17:30.081319370 +0200
```

Note symbolic link!

Which FS?

```
  File: "proc"
    ID: 0          Namelen: 255     Type: proc
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 0          Free: 0          Available: 0
Inodes: Total: 0          Free: 0
  File: "root"
    ID: 80300000000 Namelen: 255     Type: xfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 39302400   Free: 15792587   Available: 15792587
Inodes: Total: 78643200   Free: 78511636
  File: "run"
    ID: db9da7e0d7518f56 Namelen: 255     Type: tmpfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 4842911    Free: 4842253    Available: 4842253
Inodes: Total: 819200     Free: 816098
File: "srv"
    ID: 80300000000 Namelen: 255     Type: xfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 39302400   Free: 15792587   Available: 15792587
Inodes: Total: 78643200   Free: 78511636
```

Which FS?

```
 File: "sys"
    ID: 0        Namelen: 255     Type: sysfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 0          Free: 0          Available: 0
Inodes: Total: 0          Free: 0
  File: "tmp"
    ID: 80300000000 Namelen: 255     Type: xfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 39302400   Free: 15792587   Available: 15792587
Inodes: Total: 78643200   Free: 78511636
  File: "u"
    ID: b92f1c82ffffffff Namelen: 255     Type: ceph
Block size: 4194304    Fundamental block size: 4194304
Blocks: Total: 1290859070 Free: 1028258499 Available: 1028258499
Inodes: Total: 470799222  Free: -1


[fbazzocchi@nfs01 /]$ stat -f srv
  File: "srv"
    ID: 80300000000 Namelen: 255     Type: xfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 39302400   Free: 15792587   Available: 15792587
Inodes: Total: 78643200   Free: 78511636
```

Long-term storage

```
fbazzocchi@nfs01 srv]$ ls -l
total 41
drwxrwx---   4 root       cnriom          30 Sep 12  2024 STRAS
drwxrwx---  12         33 tape           227 Mar 31 09:30 analisi_da_consegnare
drwxr-xr-x   4 root       root            47 Jan 19  2022 borg_repos
drwxrwx---  13 root       burlo          222 Nov  8 11:00 burlo_long_term_storage
drwxr-x---  17   10070001 burlo         4096 Jul  3  2024 burlo_long_term_storage2
drwxrwx---  26 root       cdslab         4096 Sep 25  2024 cdslab_long_term_storage
drwxrwx---+  5 root       cro_aviano       84 Apr  1 15:26 cro_long_term_storage
drwxr-xr-x   3 root       root            24 Jan 21 16:44 decolibus
drwxrwx---  13 root       iga            4096 Mar 17 14:31 iga_long_term_storage
drwxrwxr-x  15 iscan      lage_new        170 Mar 20 16:46 illumina_decode
drwxr-xr-x   4 root       lageinstruments  41 Jul 15  2024 illumina_run
drwxrwx---  31 root       lade           4096 Mar  6 14:24 lade_long_term_storage
drwxrwx---   8 root       lage            170 Jul  5  2024 lage_archive
drwxrwx---   4 root       area             37 Jul 24  2023 lage_long_term_storage
drwxrwx---  12       1000 lage           4096 Apr 24  2024 lala_storage
drwxrwx---   2 root       lame             6 Mar 20 11:13 lame_long_term_storage
drwxrwxr-x   6 root       area           118 Jul  8  2024 laptop_backup
drwxr-xr-x  11 root       root           4096 Apr 22  2022 long_term_storage
drwxr-xr-x  11 root       root            211 Jun  6  2022 nep
drwxrwx---  16       1000 lage           4096 Aug  2  2023 onp_run_1
drwxr-xr-x.  3 root       root            18 Jul  8  2024 opt
drwxr-xr-x  16 root       root           259 Jun  5  2023 orfeo_home_backup
```

Machine write

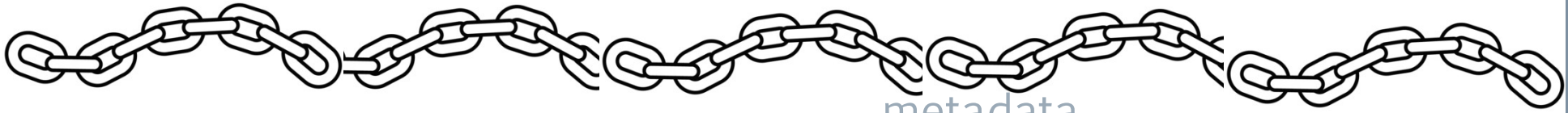
```
drwxr-xr-x  15 root       root            4096 Mar  7 11:37 orfeo_replicated_share
drwxr-xr-x   4 root       root            4096 Jun 28  2024 orfeo_repo
drwxrwx---   3 root       plus              20 May 10  2023 plus
drwxrwx---  82       1000 lage            4096 Jan 16 11:33 post_run
drwxr-xr-x   5 root       root              47 Feb  4 16:08 proxmox
drwxr-xr-x   7 1000030001 orfeohelpers     132 Feb  7  2023 read_the_docs
drwxr-xr-x   5       1000 lage             111 Jul  3  2024 storage
drwxrwx---   4       1000 lage              36 Jun 21  2024 tmp_stuff
drwxr-xr-x   3 root       root               3 Oct 21 17:18 zfs_collaborations
drwxr-xr-x   5 root       root               5 Aug 27  2024 zfs_lage
drwxr-xr-x.  2 root       root               6 Oct 10 10:44 zfs_lage_backup
```

```
[fbazzocchi@nfs01 zfs_lage]$ ls
delivery  post_run  run
[fbazzocchi@nfs01 zfs_lage]$ cd run/
[fbazzocchi@nfs01 run]$ ls
AREA  IGA
[fbazzocchi@nfs01 run]$
```

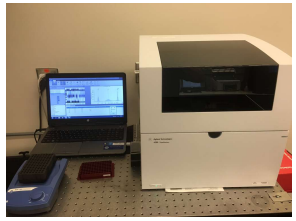
The Importance of Data Management Plan

Parenthesis

FAIR-by-design in a genomic lab



ELectronic Notebook (ELN)



metadata &data

metadata csv

metadata Pdf; csv a

Bcl+ xml

fastq

Online presequencing machines

Nextcloud

Data handling

```
[fbazzocchi@nfs01 srv]$ stat zfs_lage
  File: zfs_lage
  Size: 5               Blocks: 1          IO Block: 4096   directory
Device: 25h/37d Inode: 34          Links: 5
Access: (0755/drwxr-xr-x)  Uid: (    0/    root)   Gid: (    0/    root)
Access: 2025-04-08 15:44:45.367192404 +0200
Modify: 2024-08-27 17:51:20.211330256 +0200
Change: 2024-08-27 17:51:20.211330256 +0200
 Birth: 2024-08-26 14:08:09.052420247 +0200
[fbazzocchi@nfs01 srv]$ stat -f zfs_lage
  File: "zfs_lage"
    ID: b2b5a14c0013261f Namelen: 255     Type: zfs
Block size: 131072     Fundamental block size: 131072
Blocks: Total: 425429788  Free: 425429787  Available: 425429787
Inodes: Total: 108910025671 Free: 108910025656
```

zfs

Remember zfs ideal for archive!

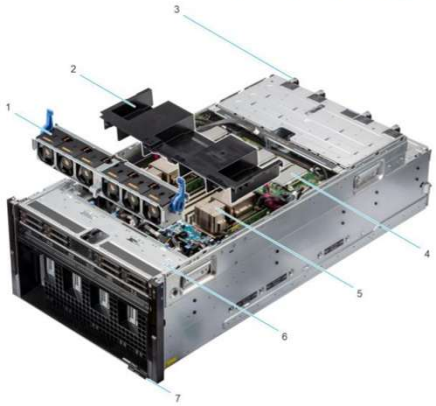
```
[fbazzocchi@nfs01 srv]$ stat -f zfs*
  File: "zfs_collaborations"
    ID: b2ab198d00a57f18 Namelen: 255     Type: zfs
Block size: 131072     Fundamental block size: 131072
Blocks: Total: 970458437  Free: 970458436  Available: 970458436
Inodes: Total: 248437359729 Free: 248437359720
  File: "zfs_lage"
    ID: b2b5a14c0013261f Namelen: 255     Type: zfs
Block size: 131072     Fundamental block size: 131072
Blocks: Total: 425429788  Free: 425429787  Available: 425429787
Inodes: Total: 108910025567 Free: 108910025552
  File: "zfs_lage_backup"
    ID: 80300000000 Namelen: 255     Type: xfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 39302400   Free: 15792721   Available: 15792721
Inodes: Total: 78643200   Free: 78511636
```

xfs

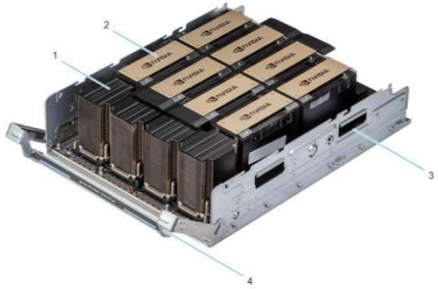
```
[fbazzocchi@nfs01 srv]$ stat analisi_da_consegnare/
  File: analisi_da_consegnare/
  Size: 227             Blocks: 0          IO Block: 4096   directory
Device: fd0ah/64778d    Inode: 96          Links: 12
Access: (0770/drwxrwx---)  Uid: (   33/ UNKNOWN)   Gid: (   33/    tape)
Access: 2025-04-08 15:19:08.883116263 +0200
Modify: 2025-03-31 09:30:49.307931039 +0200
Change: 2025-03-31 09:30:49.307931039 +0200
 Birth: 2022-12-21 11:49:45.894167000 +0100

[fbazzocchi@nfs01 srv]$ stat -f analisi_da_consegnare/
  File: "analisi_da_consegnare/"
    ID: fd0a00000000 Namelen: 255     Type: xfs
Block size: 4096       Fundamental block size: 4096
Blocks: Total: 2440883712 Free: 299333480  Available: 299333480
Inodes: Total: 976562176  Free: 976539790
```

ORFEO: THE COMPUTING FACILITY

Inside view of the system and System board connectors



1. Cooling fans
2. Air shroud
3. Power supply unit
4. Power Distribution Board
5. Heat sink for processor
6. Drive cage
7. Express Service Tag



1. NVSwitch Heatsink
2. GPU Heatsink
3. Chassis
4. Handle

ORFEO HPC nodes after PNRR

TYPE OF NODE	RAM per node	COREs per node	GPUs per node
10 x THIN intel nodes	768 GB	24	-
2 x FAT intel nodes	1536 GB	36	-
4 x GPU intel nodes	256 GB	24	2 x Nvidia V100 (32GB)
8 x EPYC Amd nodes	512 GB	128	-
2 x DGX Nvidia Station	1024 GB	128	8 x Nvidia A100 (40GB)
13 x GENOA Amd nodes	512 GB	64	-
3 x Dell XE9680	1024 GB	112	8 x Nvidia H100 (80GB)
42 Compute Nodes	**~ 25 TB of RAM**	**2856 cores**	**48 GPUs**

New CPU nodes: GENOA

13 x DELL PowerEdge R6625 server 1U

Each one configured with:

- CPU: 2 x AMD EPYC 9374F 3.85GHz, 32C/64T, 256M Cache, 320W codename GENOA
- Network:
  - 1 Broadcom 5720 Dual Port 1GbE (OutBand management)
  - 1Broadcom 57414 Dual Port 25GbE SFP28 (InBand management)
  - 1 Nvidia ConnectX-7 Single Port NDR200 OSFP (Infiniband)
- Memory:
  512GB DDR5 @ 4.800 MT/s

New GPU nodes: H100

3 x DELL PowerEdge XE9680, server 6U

Each one configured with:

- CPU: 2 x Intel Xeon Platinum 8480+ 2.0G, 56C, 350W
- Network:
  - 1 Broadcom 5720 Dual Port 1GbE (OutBand management)
  - 1 Broadcom 57414 Dual Port 25GbE SFP28 (InBand management)
  - 8 x Nvidia ConnectX-7 Single Port NDR200 OSFP (Infiniband)
- Memory:
  1024GB DDR5 @ 4.800 MT/s
- GPU:
  8 x GPU NVIDIA HGX H100 SXM 80GB 700W

ORFEO: THE NETWORK

NETWORK CLUSTER CLASSIFICATION

- High Speed Network
  - Parallel computation
    - Low latency / High bandwidth
    - Usual choices: **Infiniband !**
- I/O Network
  - I/O requests (NFS and/or parallel FS)
  - latency not fundamental/ good bandwidth
  - We use Infiniband with **IPoIB**, so the TCP stack is implemented on top of the Infiniband layer.
- In band Management network
  - management traffic of all services (LRMS/NFS/software etc..)
- Out of band Management network:
  - Remote control of nodes and administration tasks.

NETWORK

- Ethernet network is an efficient spine-leaf configuration comprising 10 switches.
- Infiniband network contains switches that are now interconnected via a high-speed link with a bandwidth of 1.6 Tbit/s.

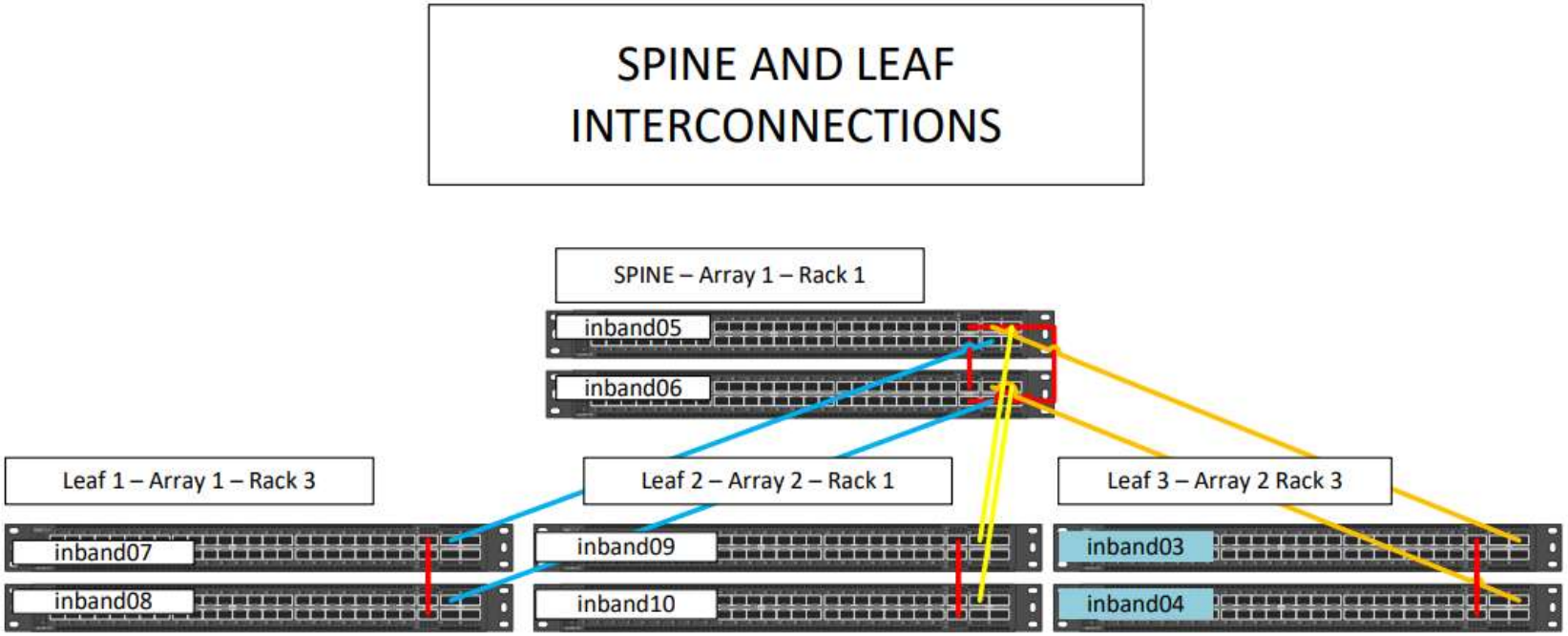
ORFEO network after PNRR

• HPC Network • I/O Network	100 Gbit & 200 Gbit Infiniband QM8700 + QM9700 Mellanox - 16 + 51.2 Tb/s aggregate throughput
In band Management network	2 x 25Gbit Ethernet link 10x dell S5148F switches: - 3.6 Tbps aggregate throughput - Fully redundant network
Out of band Management network	1Gbit Ethernet

Spine-Leaf Network Architecture

- Modern data center network topology for scalability and performance
- Spine switches: high-speed backbone interconnecting all leaf switches
- Leaf switches: connect directly to servers and edge devices
- All leaf switches connect to every spine switch
- Reduces bottlenecks and supports east-west traffic efficiently
- Ideal for highly virtualized environments and cloud-scale infrastructure

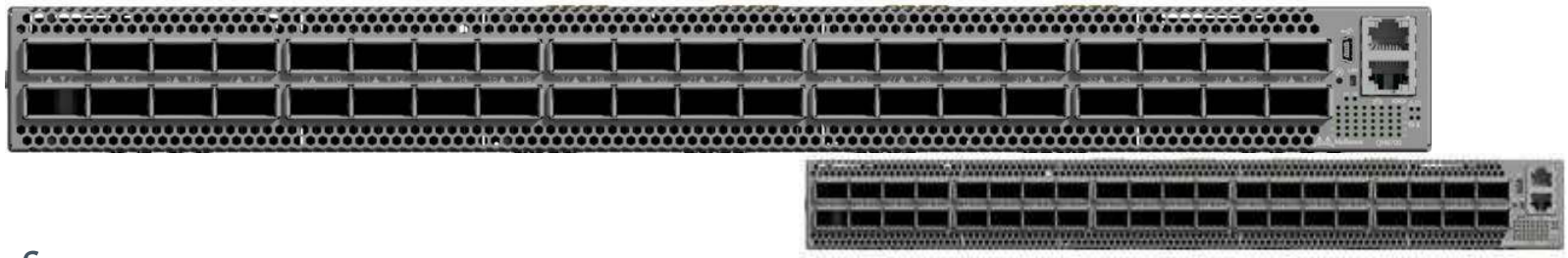
ORFEO 25 Gbit network topology



Networks

```
[fbazzocchi@nfs01 /]$ ls -lh /sys/class/net/
total 0
lrwxrwxrwx 1 root root    0 Apr  9 02:01 bond0 -> ../../devices/virtual/net/bond0
-rw-r--r-- 1 root root 4.0K Apr  9 21:00 bonding_masters
lrwxrwxrwx 1 root root    0 Apr  4 10:25 eno1 ->
../../devices/pci0000:17/0000:17:02.0/0000:19:00.0/net/eno1
lrwxrwxrwx 1 root root    0 Apr  4 10:25 eno2 ->
../../devices/pci0000:17/0000:17:02.0/0000:19:00.1/net/eno2
lrwxrwxrwx 1 root root    0 Apr  8 01:00 hpc.711 -> ../../devices/virtual/net/hpc.711
lrwxrwxrwx 1 root root    0 Apr  9 02:02 hpc.713 -> ../../devices/virtual/net/hpc.713
lrwxrwxrwx 1 root root    0 Apr  4 10:25 ibp94s0 ->
../../devices/pci0000:5d/0000:5d:00.0/0000:5e:00.0/net/ibp94s0
lrwxrwxrwx 1 root root    0 Apr  9 02:02 lage.714 -> ../../devices/virtual/net/lage.714
lrwxrwxrwx 1 root root    0 Apr  4 10:25 lo -> ../../devices/virtual/net/lo
lrwxrwxrwx 1 root root    0 Apr  9 00:44 orfeo1.717 ->
../../devices/virtual/net/orfeo1.717
lrwxrwxrwx 1 root root    0 Apr  9 13:11 podman1 -> ../../devices/virtual/net/podman1
lrwxrwxrwx 1 root root    0 Apr  9 13:11 veth0 -> ../../devices/virtual/net/veth0
```

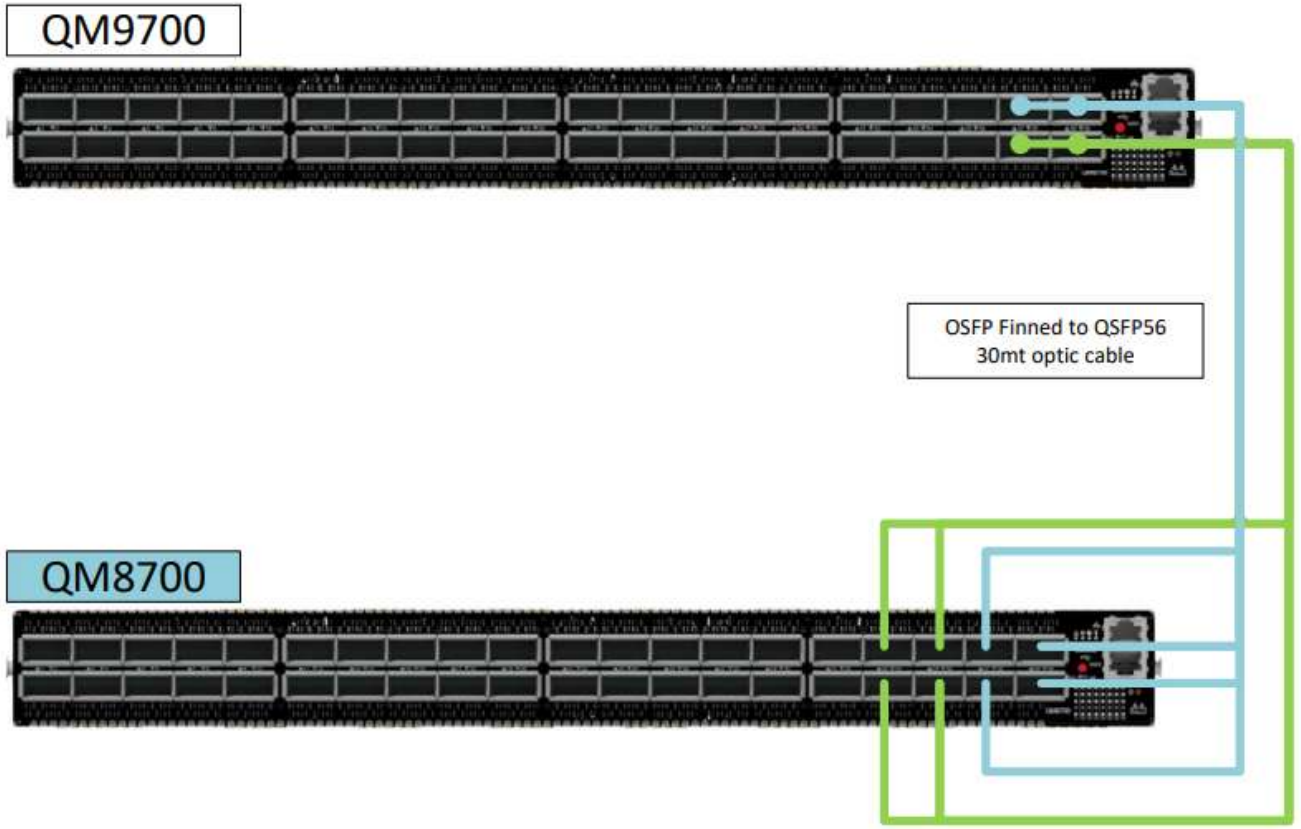
ORFEO IB network



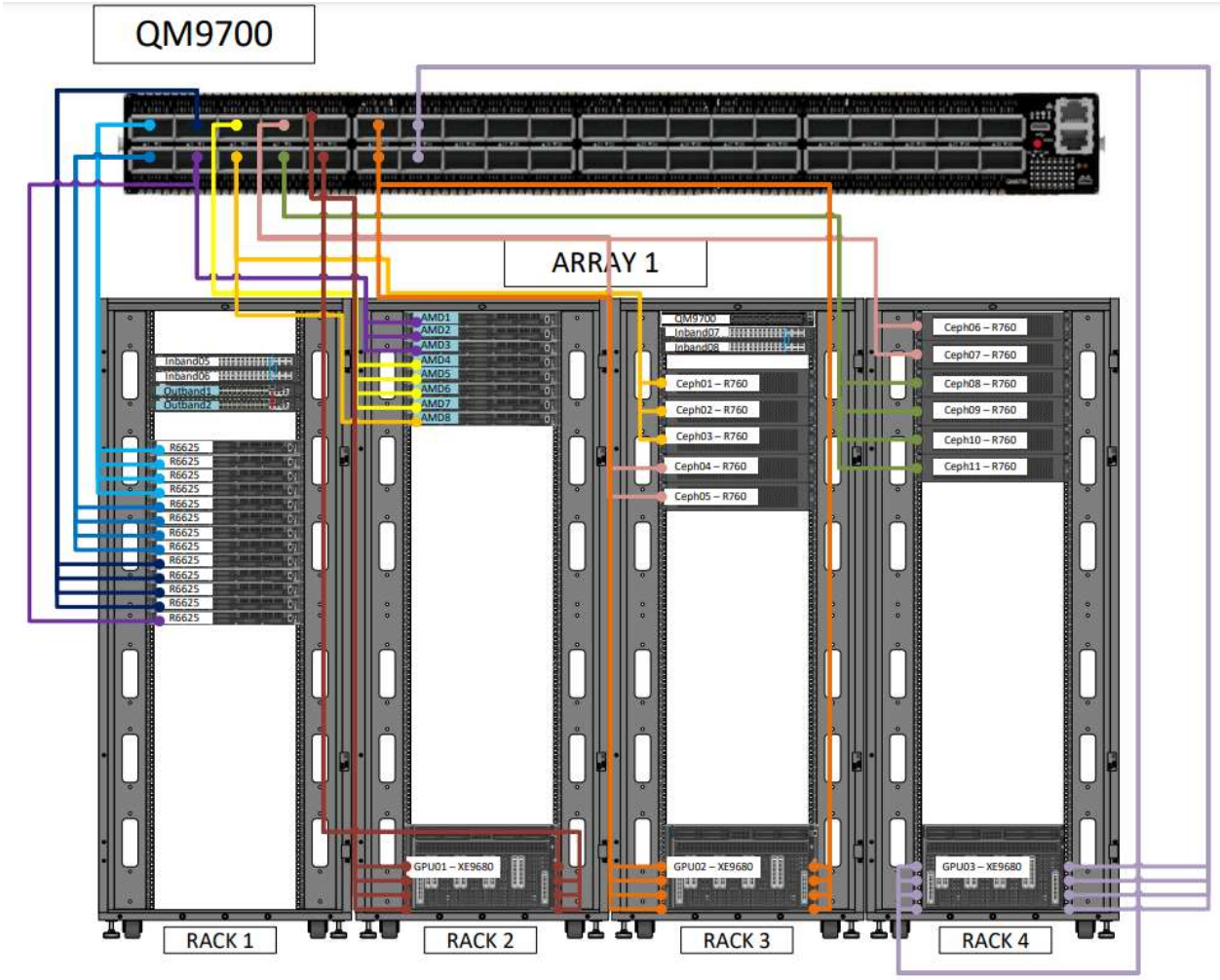
Performance

- 40 x HDR 200Gb/s ports in a 1U switch
- 80 x HDR100 100Gb/s ports (using splitter cables)
- 16Tb/s aggregate switch throughput
- Sub-130ns switch latency

Infininband network topology



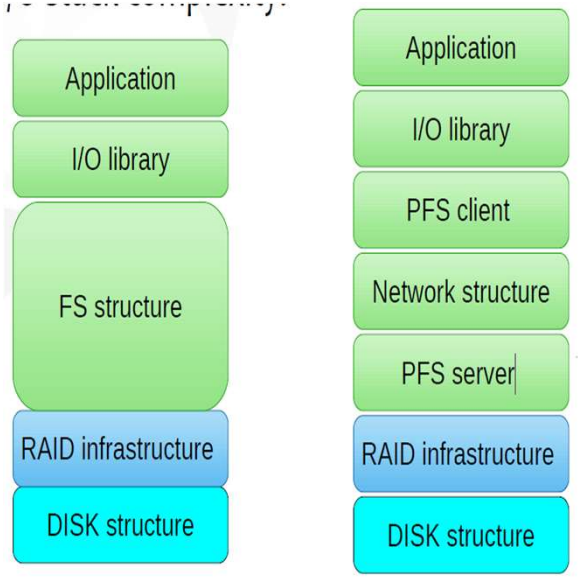
ORFEO Infiniband cabling

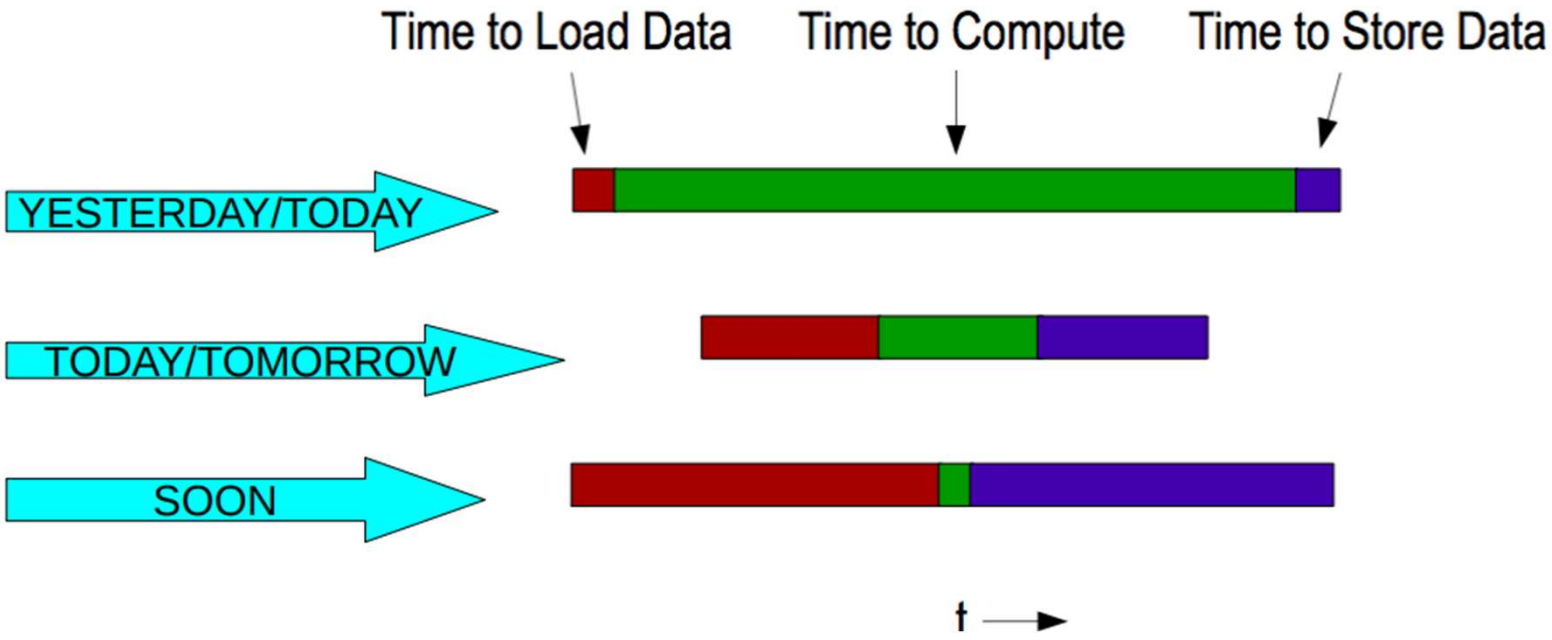


ORFEO: I/O BENCHMARKING

BENCHMARKING

- It is becoming more and more important
- I/O performance tends to be trickier than CPU/memory ones





Imagine the burden of packing and unpacking when the transit time itself is near-zero.

How to test a complex I/O infrastructure ?

- Benchmark all the single component of the infrastructure.
- Compare simple component Peak performance with measured numbers.
- Combine all numbers together to get a performance model and some expected value.
- Perform the high level benchmark and compare against what you evaluated.

What we can measure?

- Bandwidth:
  How fast we can [write | read], sequentially or randomly.
  Tools: iozone, ior, fio.
- IOPS:
  How many I/O operations per second can be performed, IOPS has no standard definitions, since the size of the payload could be variable. A common choice is 4KB.
  Tools: iozone, ior, fio.
- Metadata:
  How fast we can perform action on metadata: how many [directories | files] I can [create | delete | stat].
  Tool: mdtest

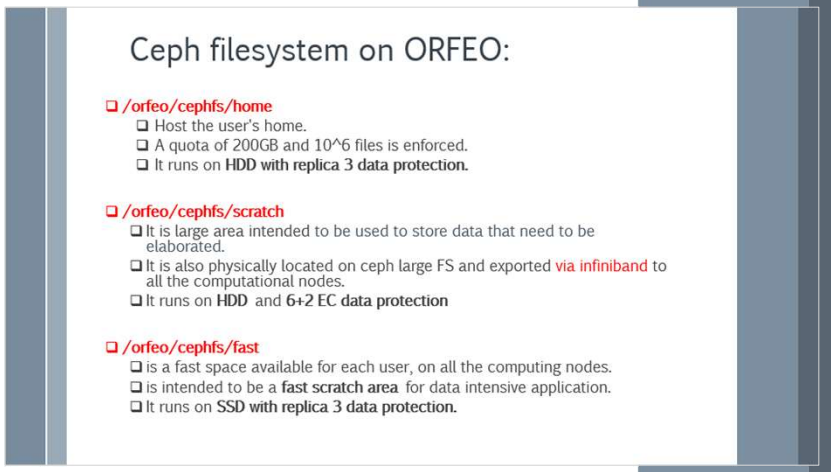
IOR

Is the de-facto standard tool to benchmark large parallel file system. It supports natively MPIIO and POSIX standard. It is distributed and could scale across several nodes exploiting all FS available bandwith

- Official docs and repository

IOR - results

	Scratch	Fast	Home
Write[MiB/s]	4878.94	5818.53	1564.87
Read [MiB/s]	4964.75	12713.33	4620.71



- This results give us an idea about the infrastructure's performance and may not accurately represent an application's overall performance.
- Factors like caching and IO patterns can significantly influence actual performance.
- The fast pool excels in reading tanks to its SSD and the X3 replication rule.
- Optimizing the number of PGs could enhance the performance of the home filesystem.

IOR - IOPS

To measure the max IOPS of our FS it has been used the following definitions of IOPS:

"*move smallest unit of storage from/to arbitrary location, when with "smallest" usually we take 4 KiB of data."*

Experiment:

dd if=/dev/zero of=here bs=1G count=1

VS

dd if=/dev/zero of=here bs=1G count=1 oflag=direct

IOR - IOPS results

- Fast result:
- $ mpirun -npernode 128 ./ior -F -C -e -g -b 1m -t 4k -s6 -D45 -w -z --posix.odirect -l random -i 10 -o /fast/area/ntosato/test
- Summary of all tests:

Operation	Max(MiB)	Min(MiB)	Mean(MiB)	Max(OPs)	Min(OPs)	Mean(OPs)
write	314.67	188.81	264.35	80555.34	48335.38	67673.61

- Scratch result:
- $ mpirun -npernode 128 ./ior -F -C -e -g -b 1m -t 4k -s6 -D45 -w -z --posix.odirect -l random -i 10
- Summary of all tests:

Operation	Max(MiB)	Min(MiB)	Mean(MiB)	Max(OPs)	Min(OPs)	Mean(OPs)
write	20.87	17.64	18.97	5343.75	4515.07	4856.69

MDTEST

› Mdtest accompanies IOR for assessing metadata operation performance. In ORFEO configuration, it applies pressure on the metadata server.

FILE RESULTS

Operation	Home	Scratch	Fast
File creation	11630.367	11196.186	10201.188
File stat	1067745.405	1578904.628	1134152.213
File removal	9221.775	8886.967	7355.955

Results of file operations per second, averaged over 10 iterations.

File creation could simulate checkpoint saving, file stat could resemble filesystem traversal, and file removal could represent purging actions.

it measures latency on small file operations!

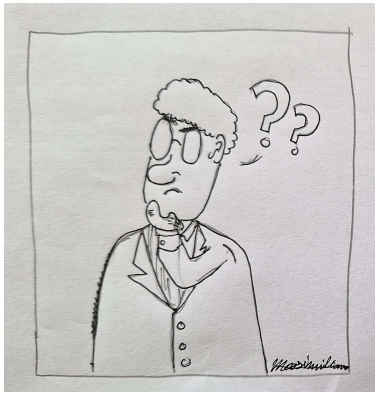
MDTEST

DIRECTORY RESULTS

Operation	Home	Scratch	Fast
Directory creation	10090.489	10176.229	11385.986
Directory stat	824791.487	1348808.262	1131034.540
Directory removal	7797.126	7634.967	7113.125

Results of directory operations per second, averaged over 10 iterations.

How much does it cost?



Utility

Total Power Facility Level

Electrical power:

- UPS
- Energy storage
- Switchgear
- Cables

Cooling:

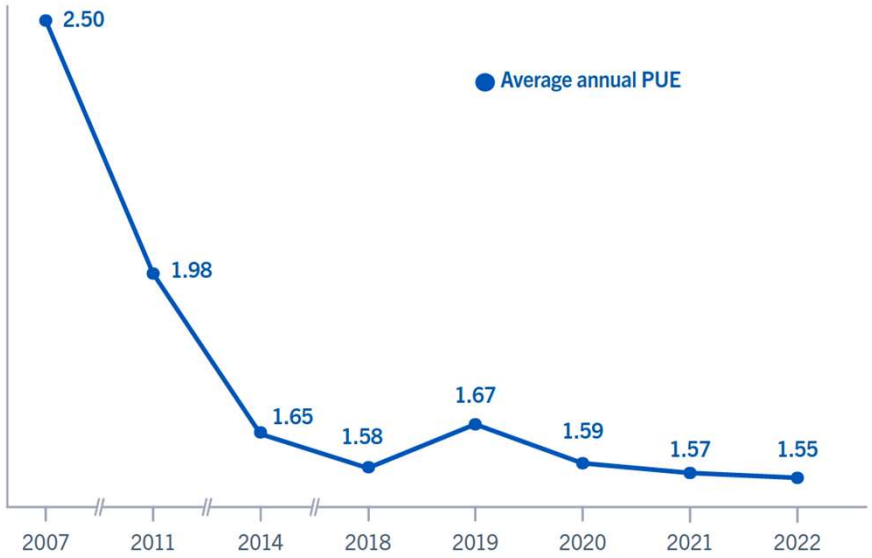
- Outdoor units
- Indoor units

IT Active Equipment Power

- Servers
- Switches
- Storage
- Telecom equipment

$$\text{PUE} = \frac{\text{Total Power Facility Level}}{\text{IT Active Equipment Power}}$$

Energy consumption efficiency metrics



Uptime Institute, Average annual PUE, 2021 Data Center Industry Survey Results

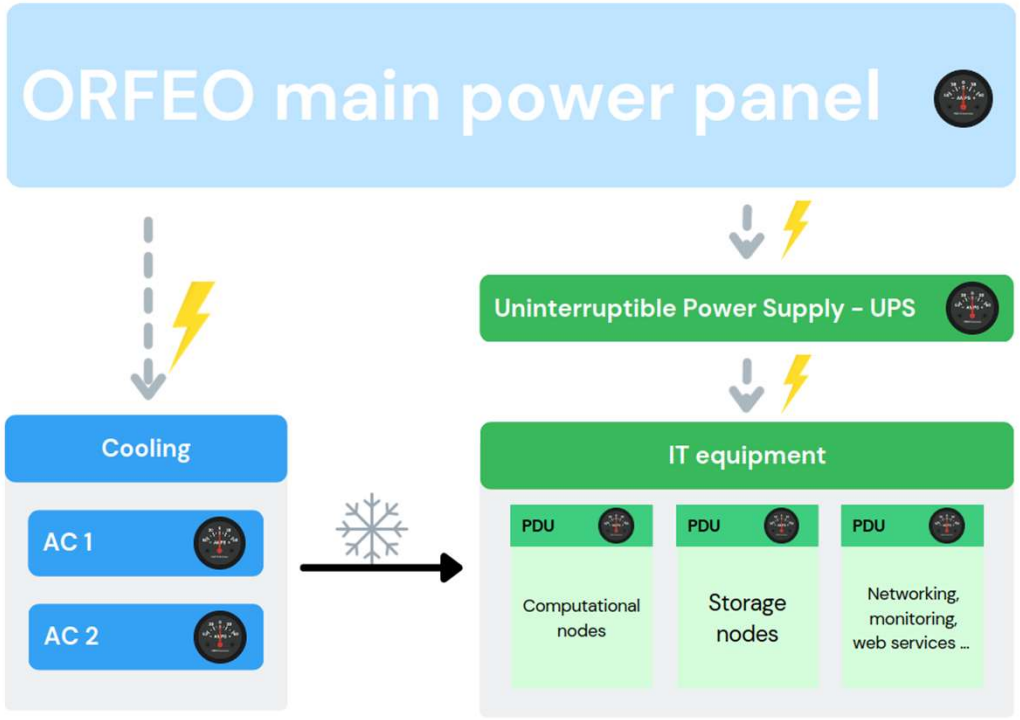
Efficiency and performance metrics:

- **PUE –** Power usage efficiency

$$\mathbf{PUE} = \frac{P_{total}}{P_{IT}}$$

The ratio between the amount of power given to the cluster and the effective amount of power used for IT tasks

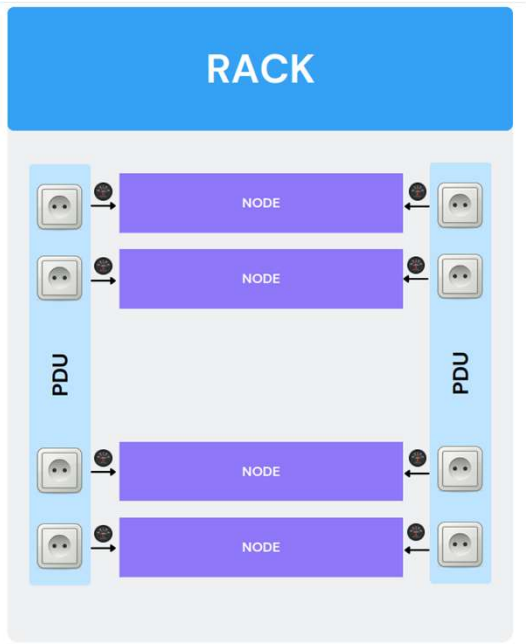
"If You Can't Measure It, You Can't Improve It"



Data sources:

- Main power panel
- Air conditioners
- UPS
- PDU
- PSU

ORFEO datacenter – Rack details



Coarse grain metrics:

- Time scale : [1-60] **s**
- Power measure: 1 **W**

Improvements:

- Better cooling strategy
- Job scheduling
- DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

GRAFANA

open-source platform for monitoring and observability, allowing visualizing, analizing and alerting on metrics and logs from various data sources

Key Features

•**Data Visualization**: Grafana provides customizable dashboards to visualize data in real-time using graphs, charts, heatmaps, and more1

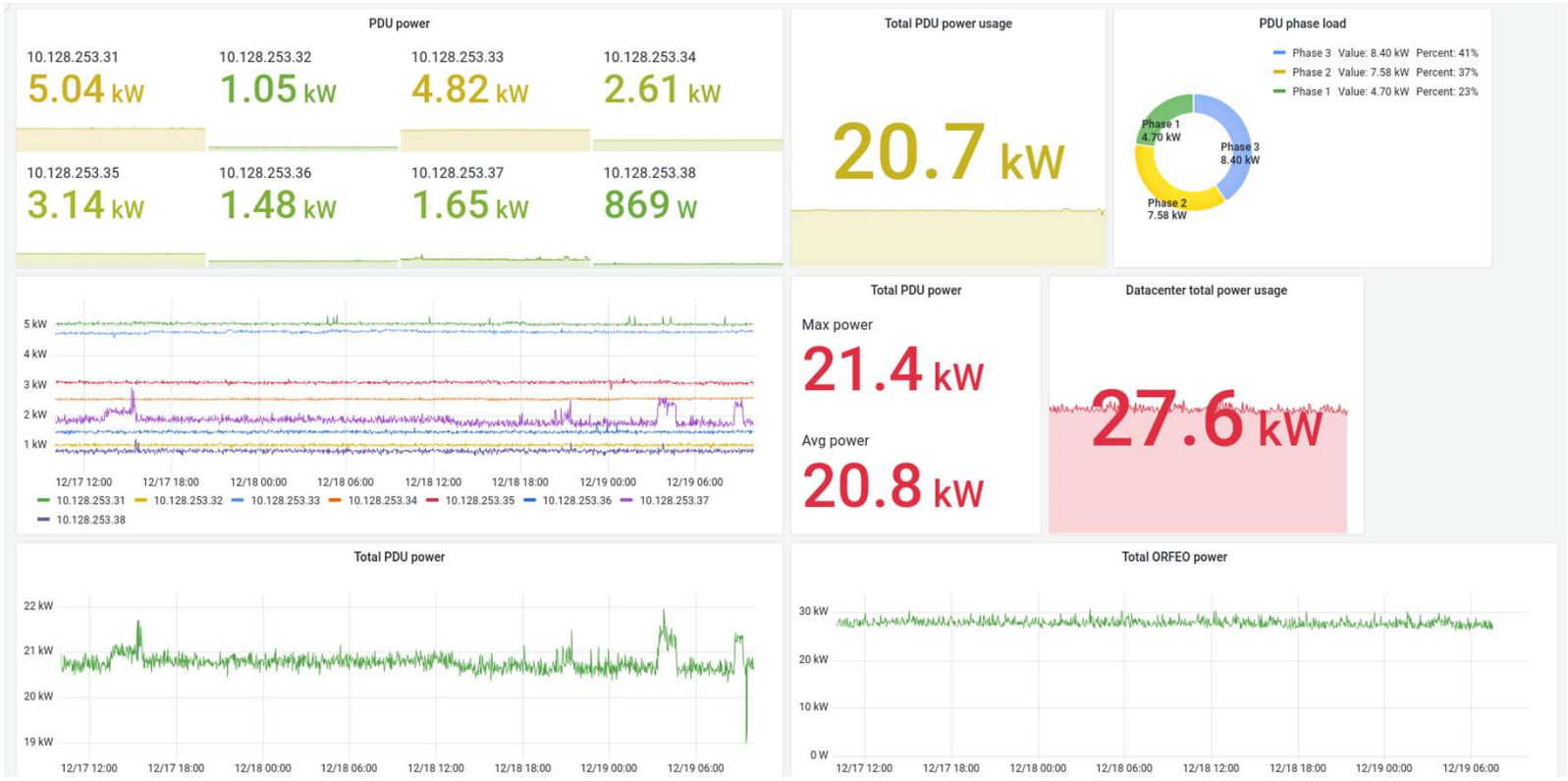
•**Alerting**: alerts may be setup to notify when certain conditions are met

•**Data Source Integration**: Grafana supports a wide range of data sources, including Prometheus, Elasticsearch, MySQL, and many others2

Grafana connects to data sources, queries the data, and displays it in a user-friendly interface. It is possible creating dashboards to monitor systems, applications and infrastructure

Grafana dashboards – coarse grain metrics

Grafana dashboards – coarse grain metrics



Grafana dashboards – coarse grain metrics