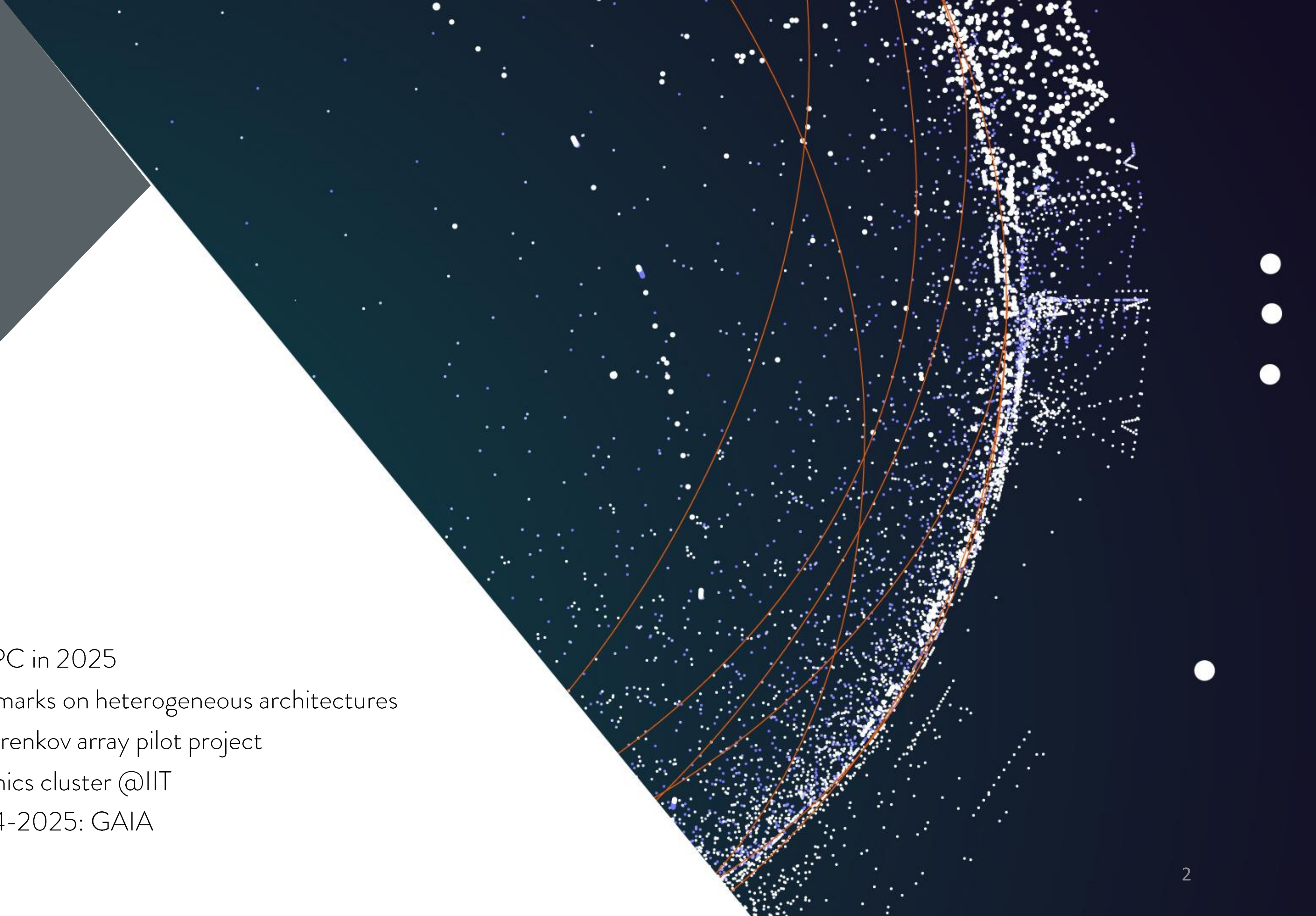
E4 COMPUTER ENGINEERING

WHEN PERFORMANCE MATTERS

HETEROGENOUS COMPUTING: FROM BENCHMARK TO RECONFIGURABLE INFRASTRUCURE

Giordano Mancini, CTO

Workshop sul calcolo INFN, La Biodola, 26 – 30 maggio 2025



AGENDA

- Who we are
- Limiting factors of HPC in 2025
- Case study A: Benchmarks on heterogeneous architectures
- Case study B: The Čerenkov array pilot project
- Case study C: Genomics cluster @IIT
- Flagship project 2024-2025: GAIA
- Conclusions

E4
COMPUTER ENGINEERING

E4 TIMELINE

2002
E4 COMPUTER ENGINEERING

2005
FIRST
INFINIBAND
CLUSTER

2009
FIRST
HYBRID CPU-GPU
NVIDIA TESLA
CLUSTER

2012
SGI ITALIA
ACQUISITION
sgi

2013
FIRST
ARM+GPU SERVER
FIRST
ARM CLUSTER

PEDRAFORCA
CLUSTER
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

2016-2017
D.A.V.I.D.E.
FIRST
OPENPOWER SERVER
& PETAFLOPS CLUSTER
TOP500 · The GREEN 500 · CINECA · PRACE
299° IN TOP500 (JUNE 2017)
14° IN GREEN500 (JUNE 2017)

2018
E4 ANALYTICS
E4 ANALYTICS
FOUNDATION

EPI SGA1
PARTNER OF
epi European Processor Initiative

2019
EUROHPC
JU CALLS

SW INTEGRATION
& CI/CD FACTORY

2021
E4 COMPUTER ENGINEERING DEUTSCHLAND
E4 DEUTSCHLAND
FOUNDATION

EPI SGA2
LIGATE
exaFOAM
ADMIRE
textarossa
ATTRACT
MAELSTROM

2022
DYNAMOS
epi European Processor Initiative
EUPEX

BOLOGNA TECNOPOLO
(ECMWF | CINECA)
+
FIRST
RISC-V BASED CLUSTER
(Monte Cimone)
RISC-V

2023
CHIRON LAB
DATACENTER
CHIRON LAB
NVIDIA GRACE &
NVIDIA GRACE
HOPPER

EUMaster4HPC
ISOLDE
TRISTAN
MAX DRIVING THE EXASCALE TRANSITION
DECICE
SPACE
EDOCC

2024
SmartHPC.QC
EoCoE
FUTURE
SEED
RISC-V MASTRO
Quack
MoseGad
SIREN

GenAI
ON PREMISE

GAIA
@ BOLOGNA
TECNOPOLO
(CINECA)

E4 IN NUMBERS

+287%

Revenues growth from 2021

92

Employees

>85%

Employees in Tech/R&D

>280

Active customers (last 3 years)

>10.000

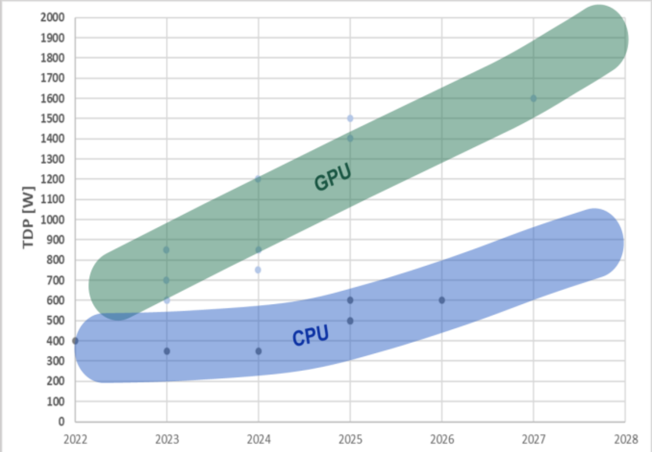
Systems produced in the last year

26

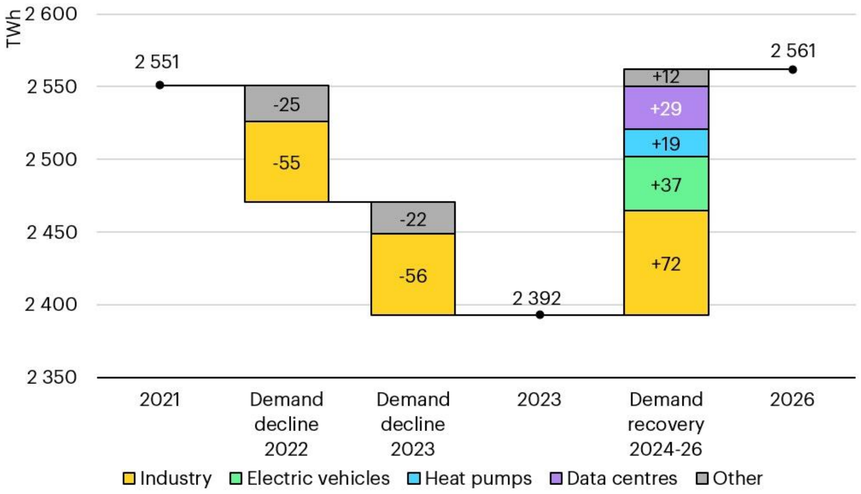
E4 contribution in European Projects

ENERGY CAP

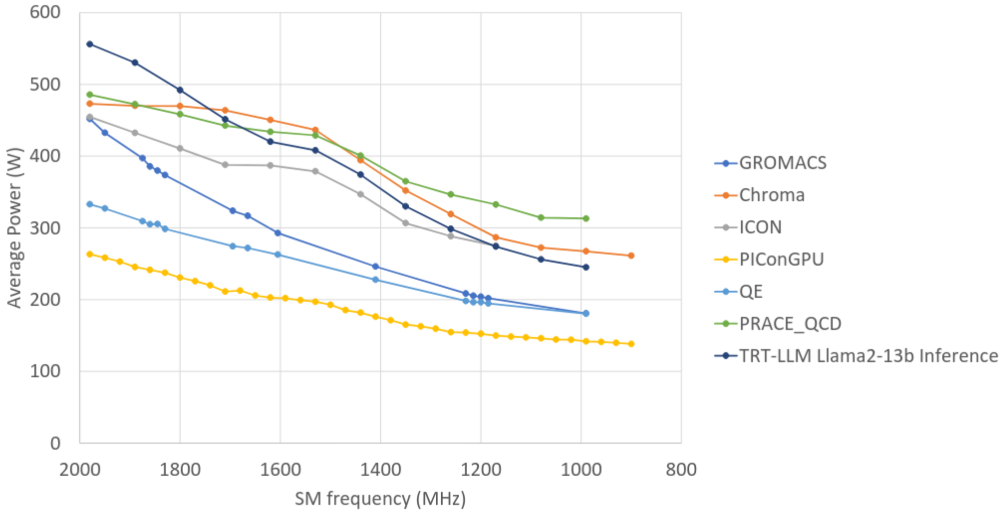
The thermal powers of GPUs will tend to 1,600 W by 2027

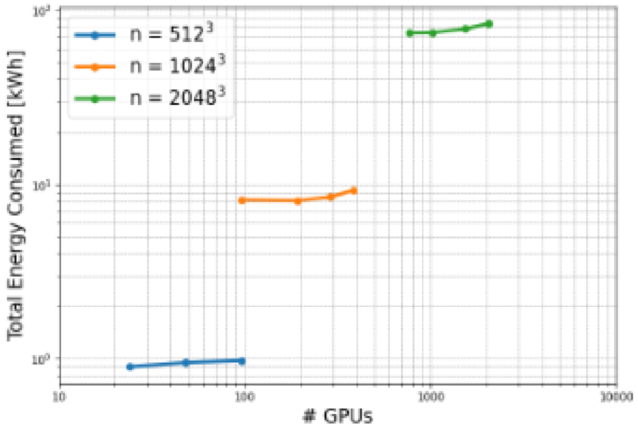


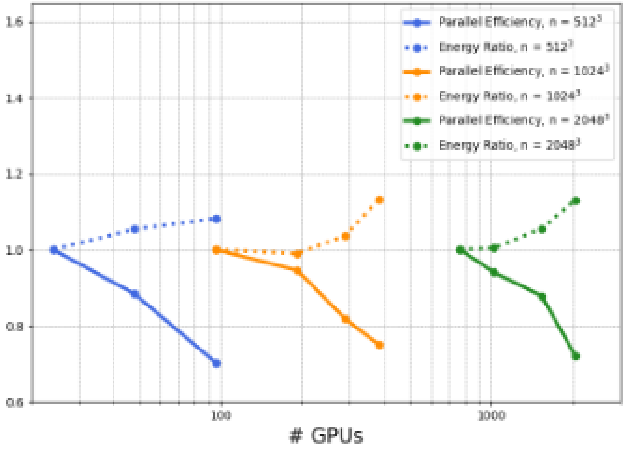
IEA report 2024 for the EU



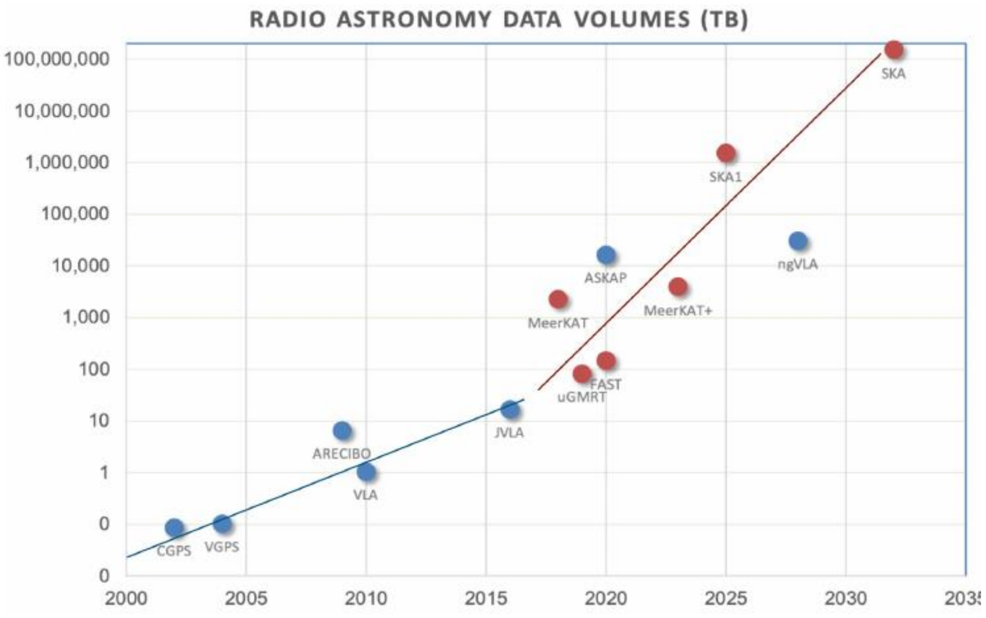
From "time to solution" to "energy to solution"

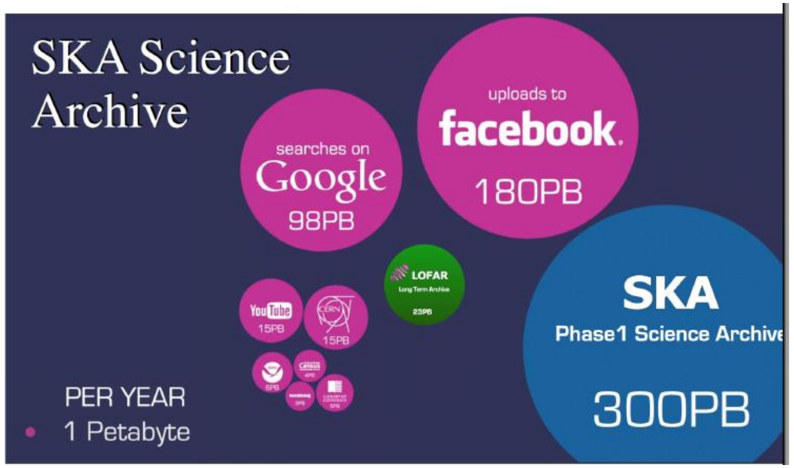




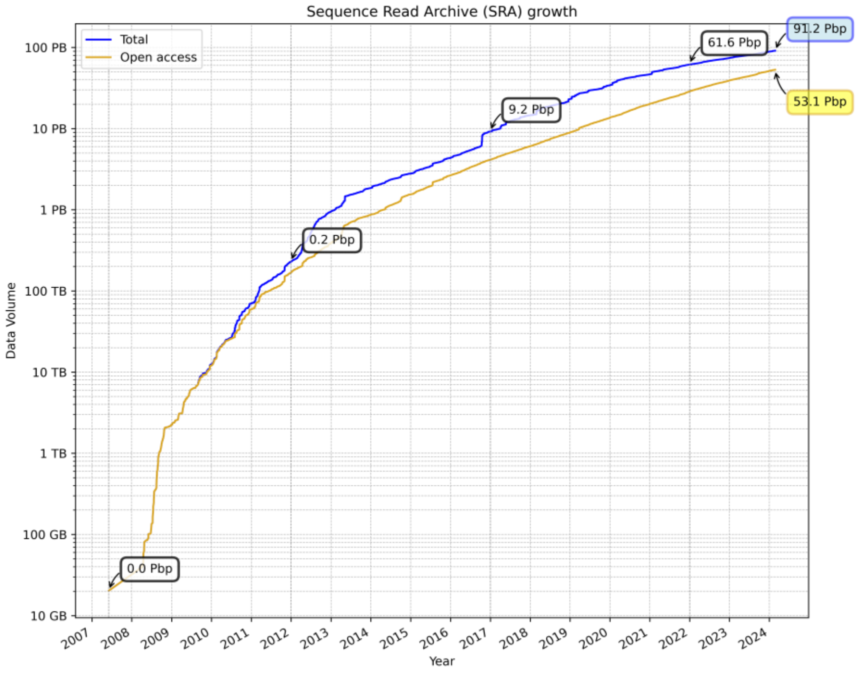


DATA PRODUCTION





Different technologies can produce short or long reads of ~ 10 Tb in 24- 72 hours



NIH short read archive database growth

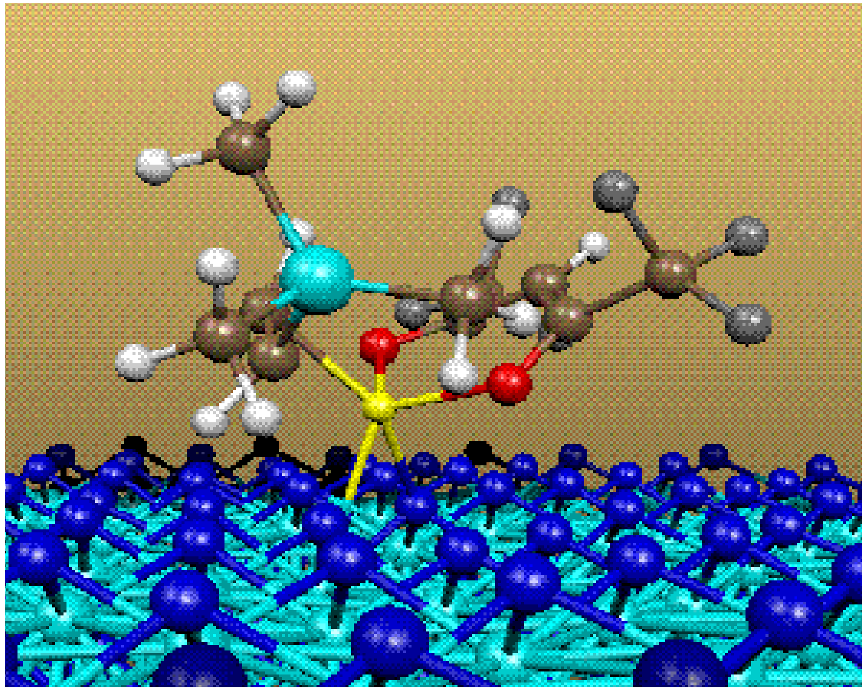
SIESTA:
A CODE FOR AB INITIO MOLECULAR DYNAMICS

Based on density-functional theory (DFT)

Written in Fortran

Parallelised with MPI and OpenMP

Uses BLAS, LaPACK, ScaLAPACK, and ELSI

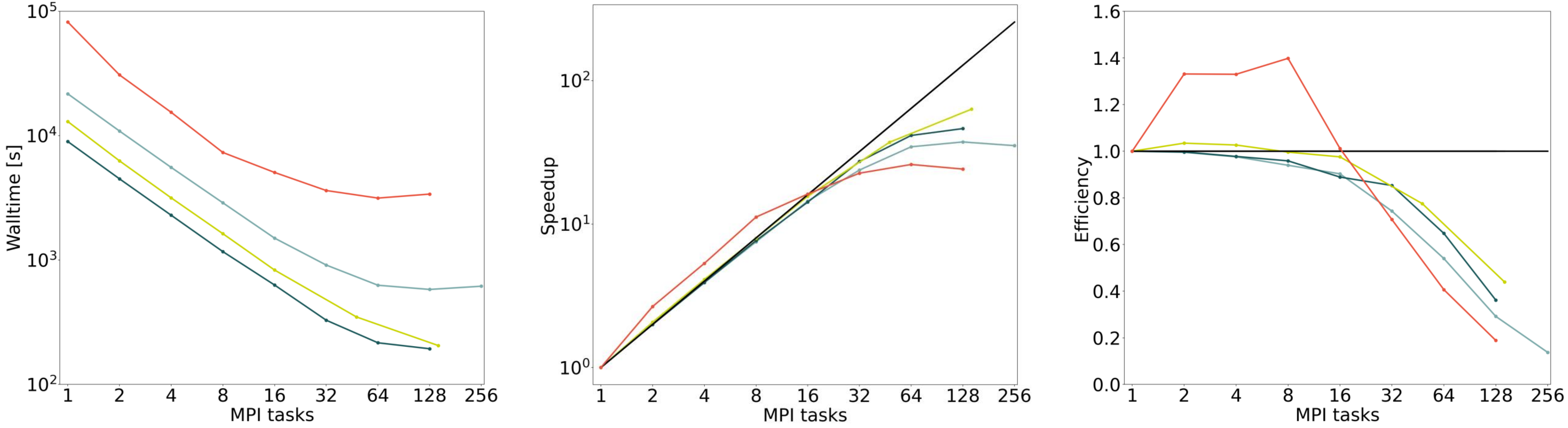


P. Ordejon, SIESTA (nano) TUTORIAL (2010).

IMPROVED PARALLEL EFFICIENCY ON GRACE WITH REDUCED ENERGY CONSUMPTION

Quantum dot

gcc 8.5.0 – 13.2.1 + OpenMPI 4.1.4

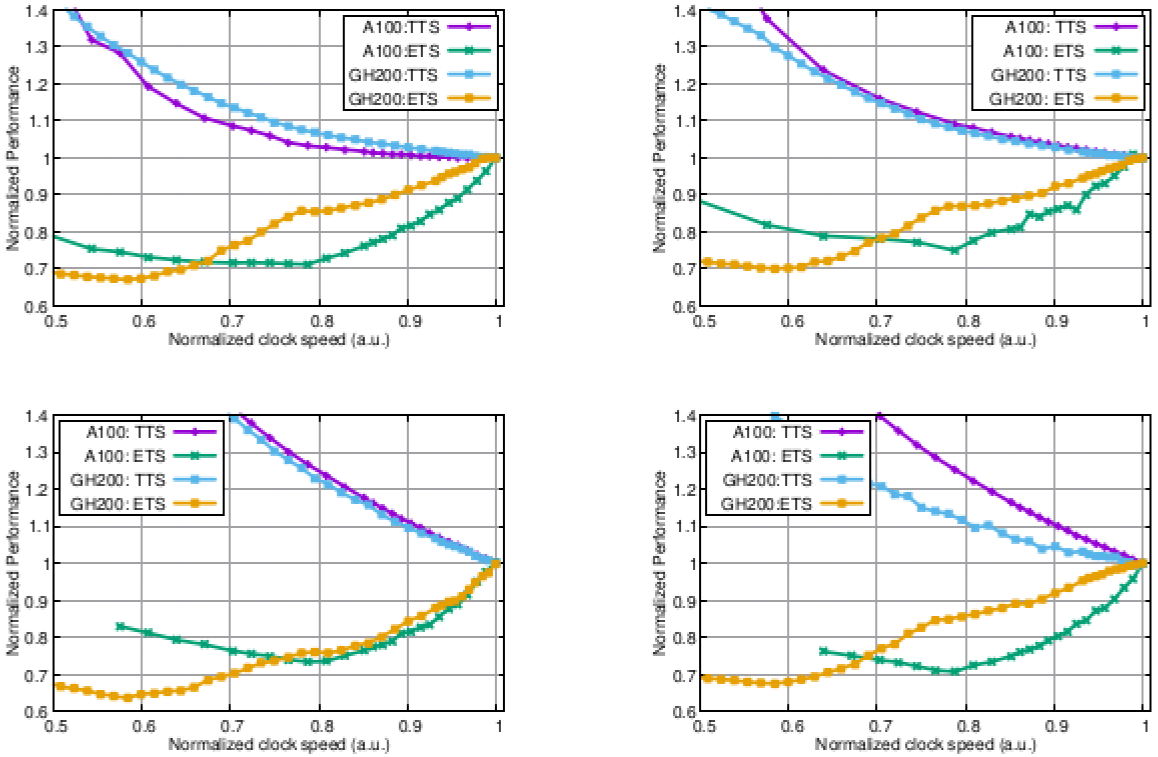


-- SOPHON SG2042, -- AMD GENOA, -- AMPERE ALTRA MAX, -- NVIDIA GRACE

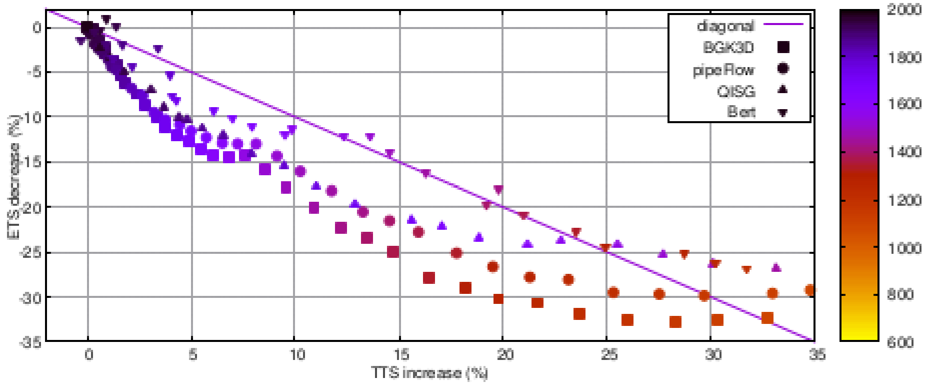
Energy-to-Solution full node estimate: AMD GENOA: 139.5 kJ, NVIDIA GRACE: 102.7 kJ

GH200 FREQUENCY SCALING FOR DIFFERENT WORKLOADS

- BGK3D: CFD / LB; F90 / OpenACC / MPI; memory bandwidth-limited code. Test: 3D lid-driven cavity flow, 512 x 512 x 512 grid, *Re* = 256
- PipeFlow: CFD code; CUDA Fortran / cuTensor / MPI; memory bandwidth-limited. Test: incompressible fluid flow in a circular pipe, *Re* = 2.85×10^5 .
- QISG: MC of 2D ising quantum spin glass; C / CUDA; memory bandwidth-limited
- BERT: transfomer; Python / TensorFlow; compute bound. Test: 25 epochs bert-large-un- cased variant (~340M parameters), batch size of 256.



ETS and TTS versus clock speed



GH200 ETS decrease vs TTS increase. Colors indicates the clock speed in MHz. Markers refers to the four codes under test.

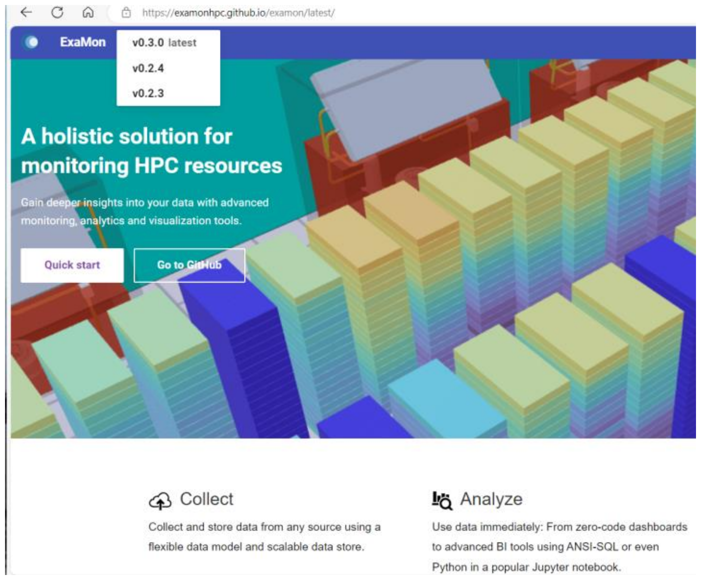
"Experience on clock rate adjustment for energy-efficient GPU-accelerated real-world codes" – Giorgio Amati, Matteo Turisini, Andrea Monterubbiano, Mattia Paladino, Elisabetta Boella, Daniele Gregori, and Danilo Croce

EXAMON DATA ANALYTICS AND VISUALIZATION

Community Initiative: https://examonhpc.github.io/examon

Current Installations:

- ENI: HPC4 & HPC5 Clusters, DaVinci 1



A holistic solution for monitoring HPC resources

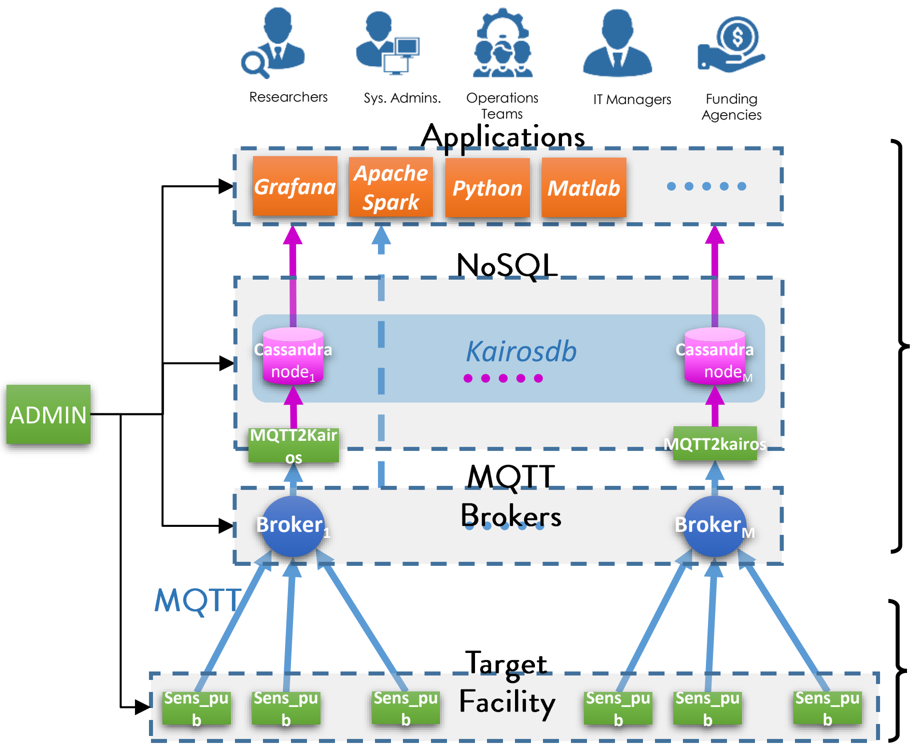
Gain deeper insights into your data with advanced monitoring, analytics and visualization tools.

Collect

Collect and store data from any source using a flexible data model and scalable data store.

Analyze

Use data immediately: From zero-code dashboards to advanced BI tools using ANSI-SQL or even Python in a popular Jupyter notebook.

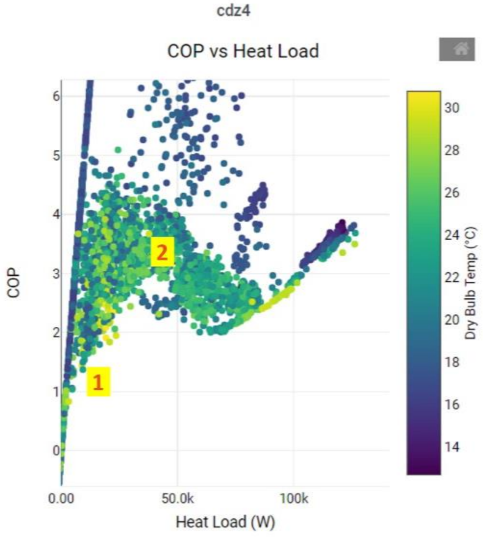


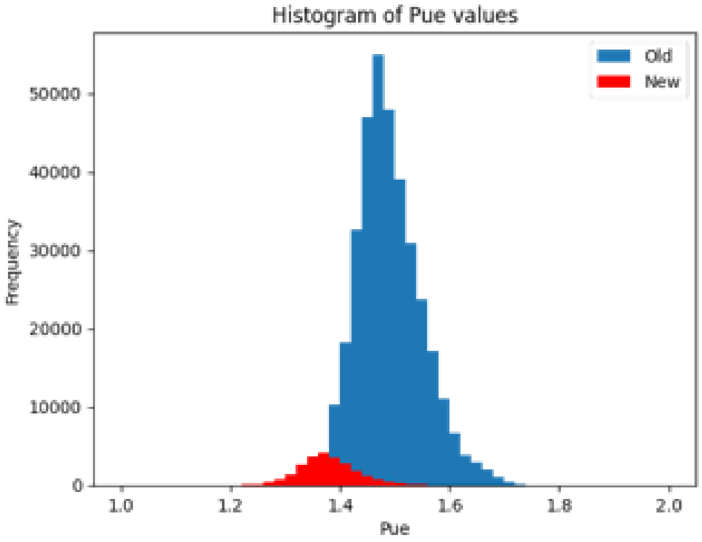
Front-end

- **MQTT** Brokers
- Data **Visualization**
- **NoSQL** Storage
- **Big Data** Analytics

Back-end

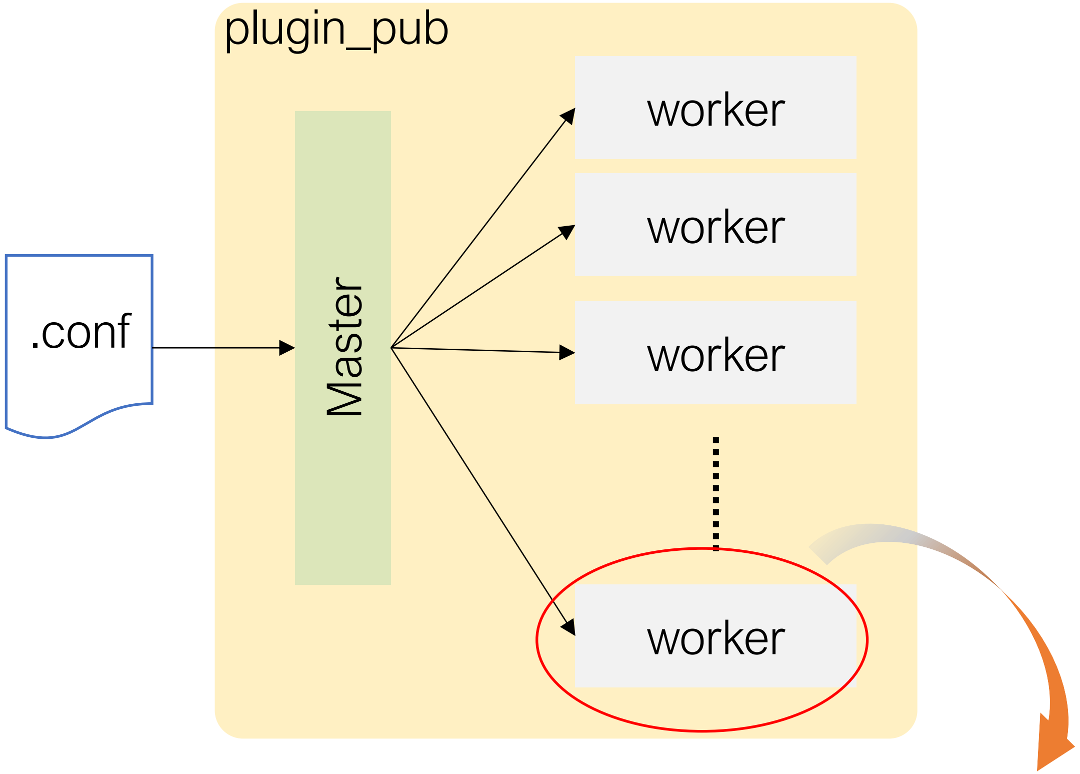
- **MQTT**–enabled sensor collectors



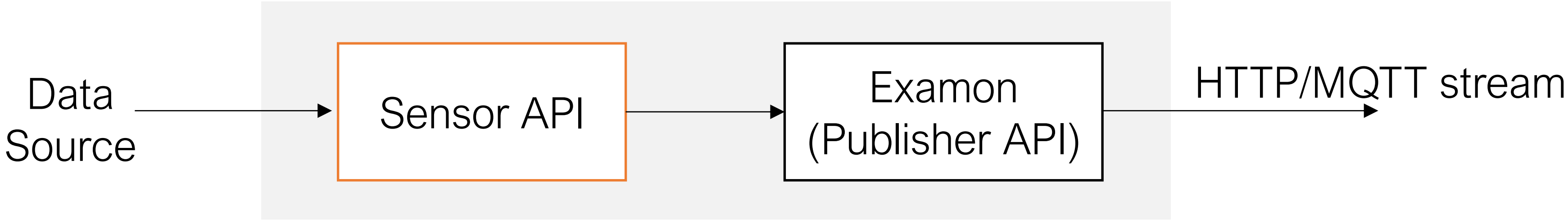


ExaMon Plugins: Internals

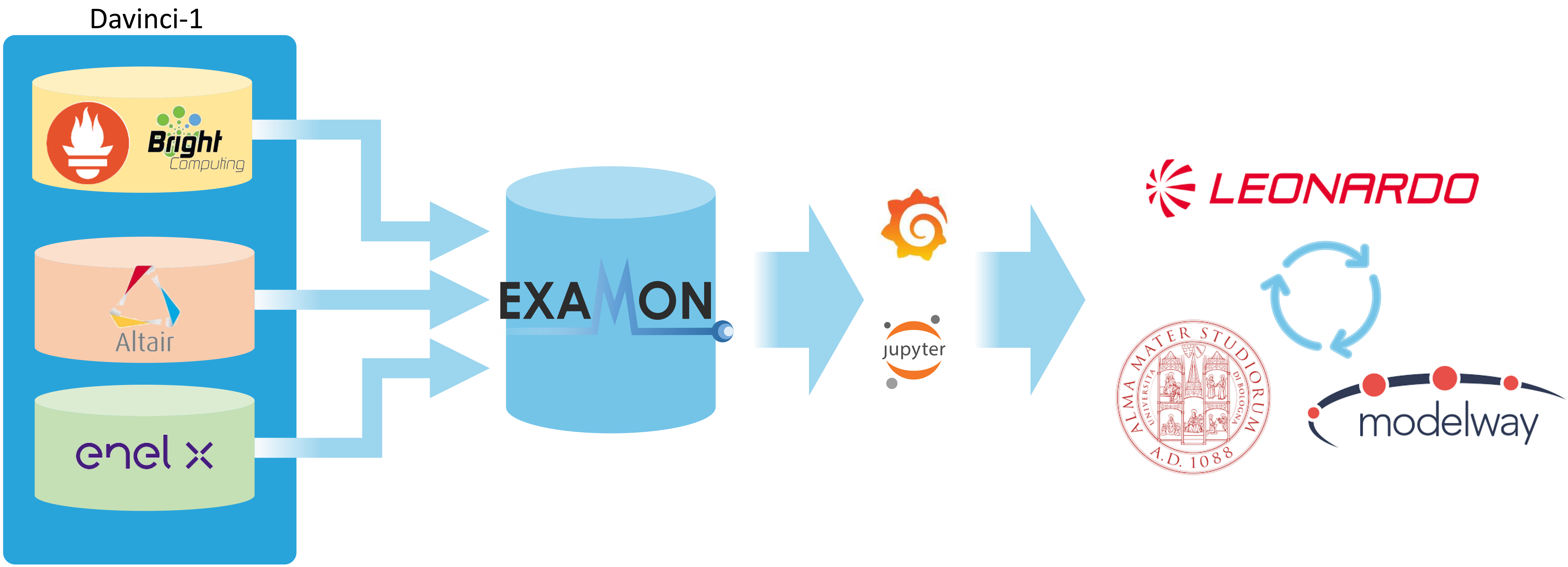
- Features:
    - Multithread/ multiprocessing
    - Internal scheduler / retry logic
    - logging



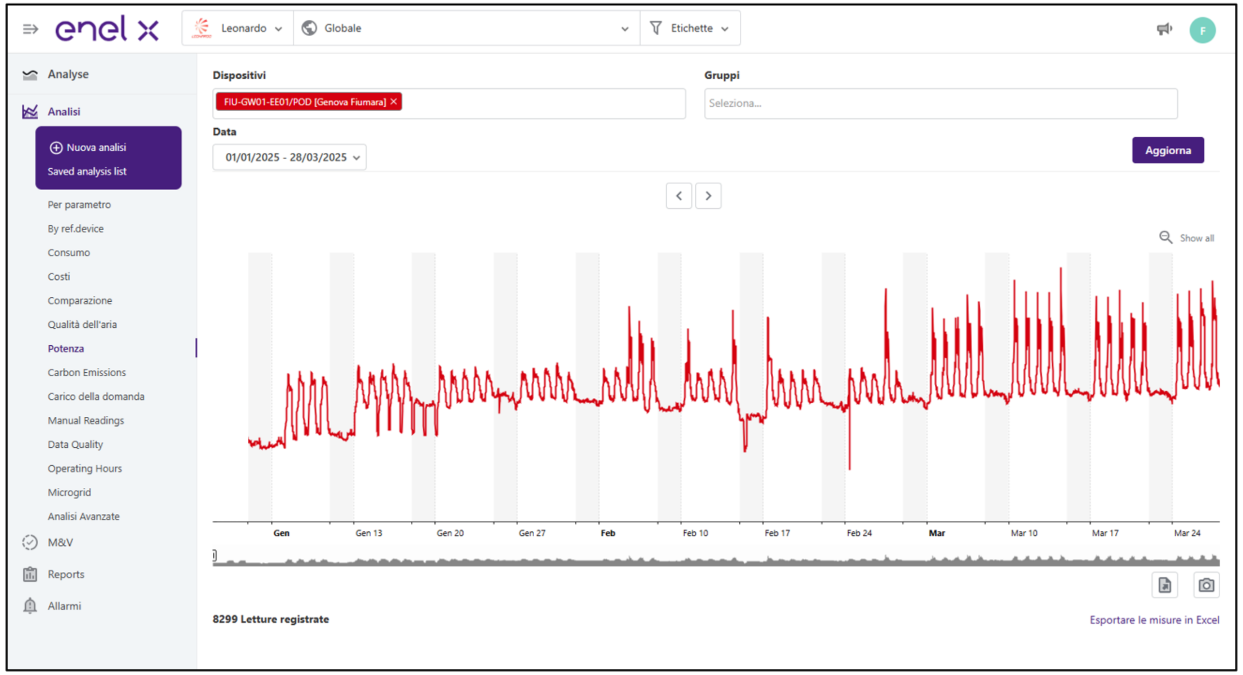
- Sensor API
    - Slurm, IPMI,…
- Publisher API
    - MQTT
    - HTTP
    - ….



Da Vinci POC: overview



EnelX – Power/Energy



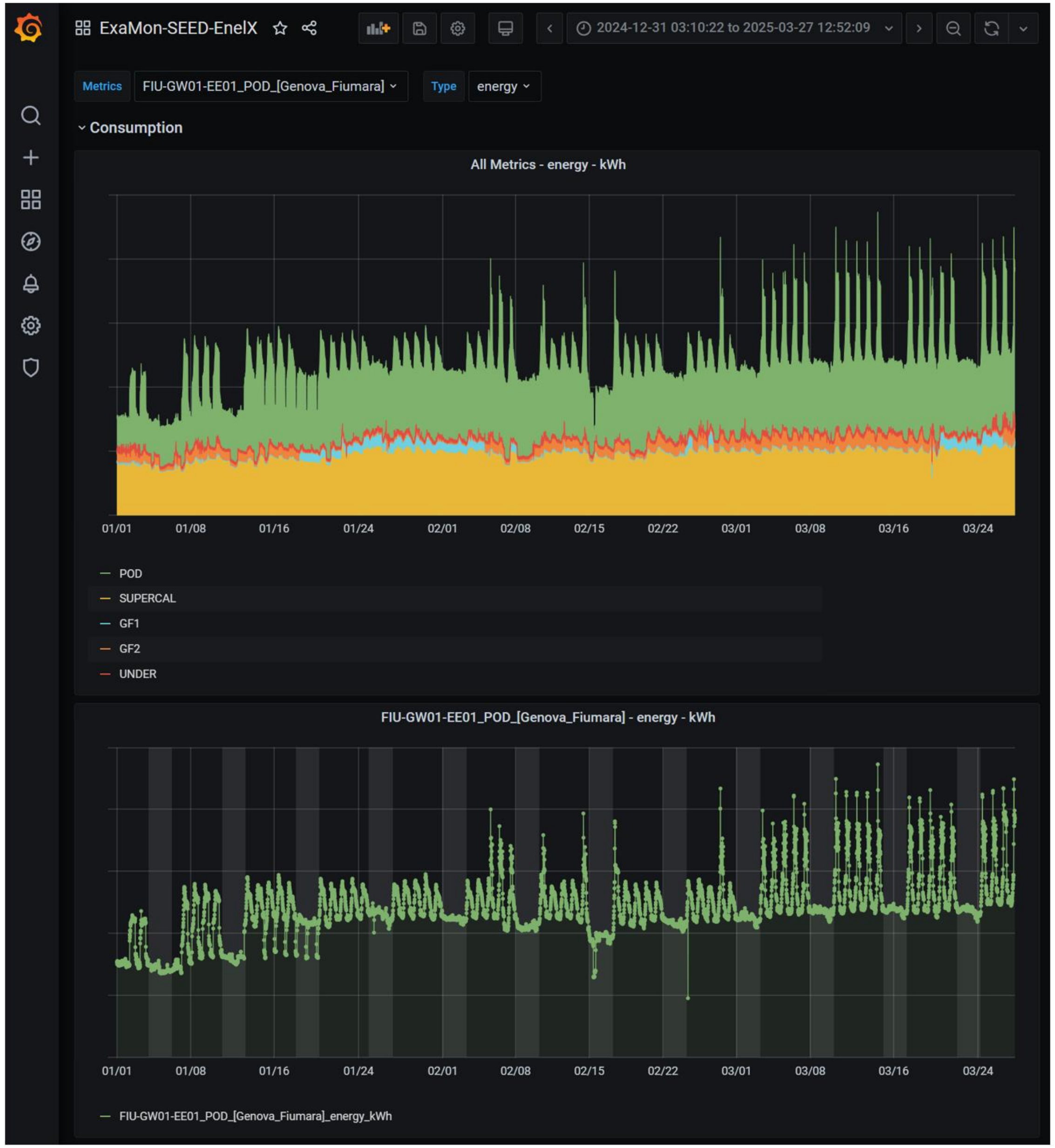
Dispositivi: FIU-GW01-EE01/POD [Genova Fiumara]

Gruppi: Seleziona...

Data: 01/01/2025 - 28/03/2025

8299 Letture registrate

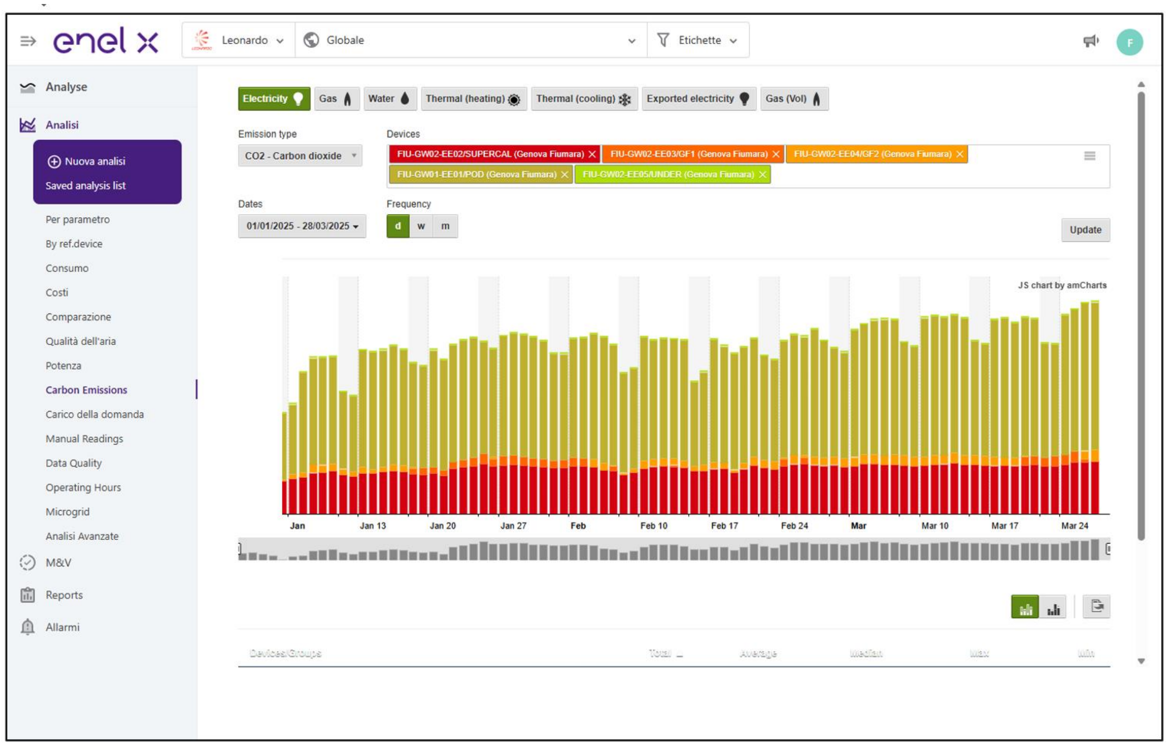
Ts=15min


ExaMon-SEED-EnelX
2024-12-31 03:10:22 to 2025-03-27 12:52:09
Consumption
All Metrics - energy - kWh
01/01
01/08
01/16
01/24
02/01
02/08
02/15
02/22
03/01
03/08
03/16
03/24
POD
SUPERCAL
GF1
GF2
UNDER
FIU-GW01-EE01_POD_[Genova_Fiumara] - energy - kWh
FIU-GW01-EE01_POD_[Genova_Fiumara]_energy_kWh

Metrics: FIU-GW01-EE01_POD_[Genova_Fiumara]

Type: energy

EnelX – Carbon Emissions

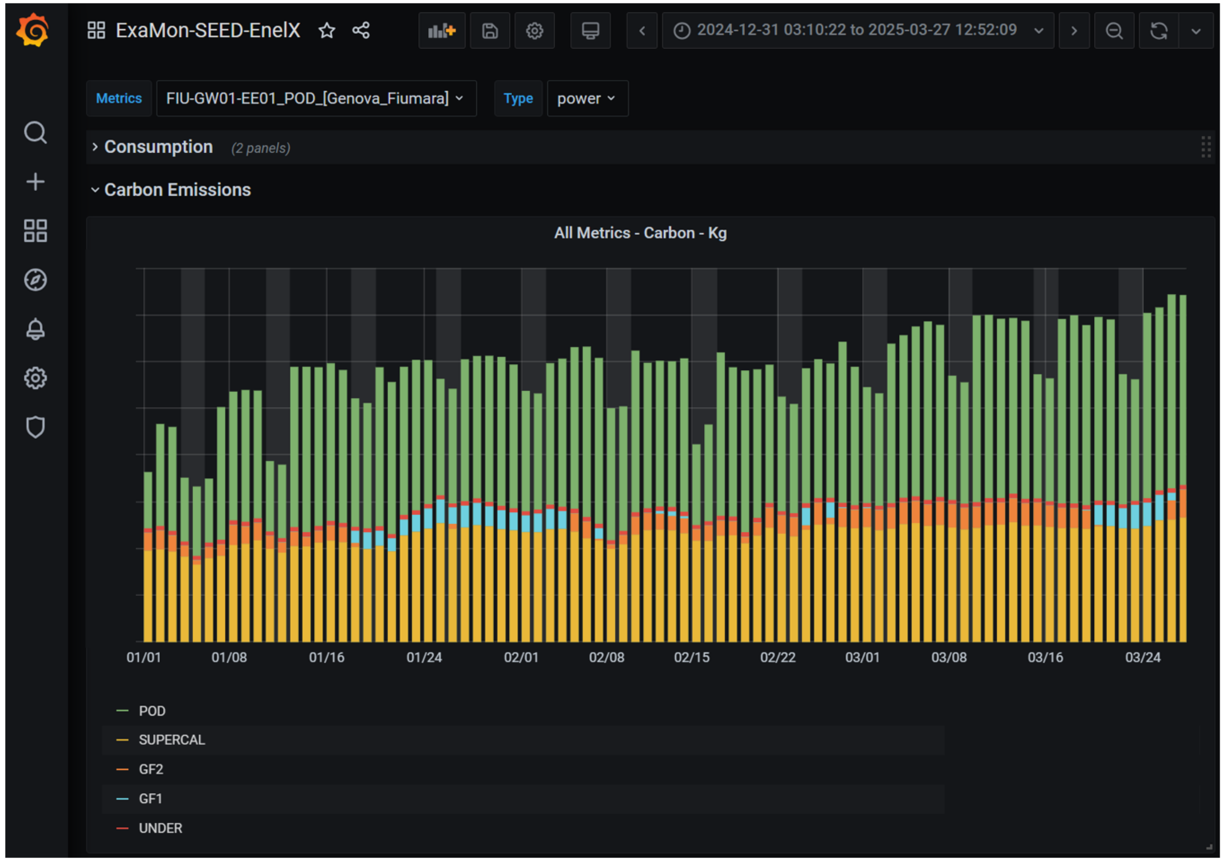


Emission type: CO2 - Carbon dioxide

Devices: FIU-GW02-EE02/SUPERCAL (Genova Fiumara), FIU-GW02-EE03/GF1 (Genova Fiumara), FIU-GW02-EE04/GF2 (Genova Fiumara), FIU-GW01-EE01/POD (Genova Fiumara), FIU-GW02-EE05/UNDER (Genova Fiumara)

Dates: 01/01/2025 - 28/03/2025

Frequency: d



2024-12-31 03:10:22 to 2025-03-27 12:52:09

Metrics: FIU-GW01-EE01_POD_[Genova_Fiumara]

Type: power

Ts=1day

R&D: DISAGGREGATED INFRASTRUCTURE

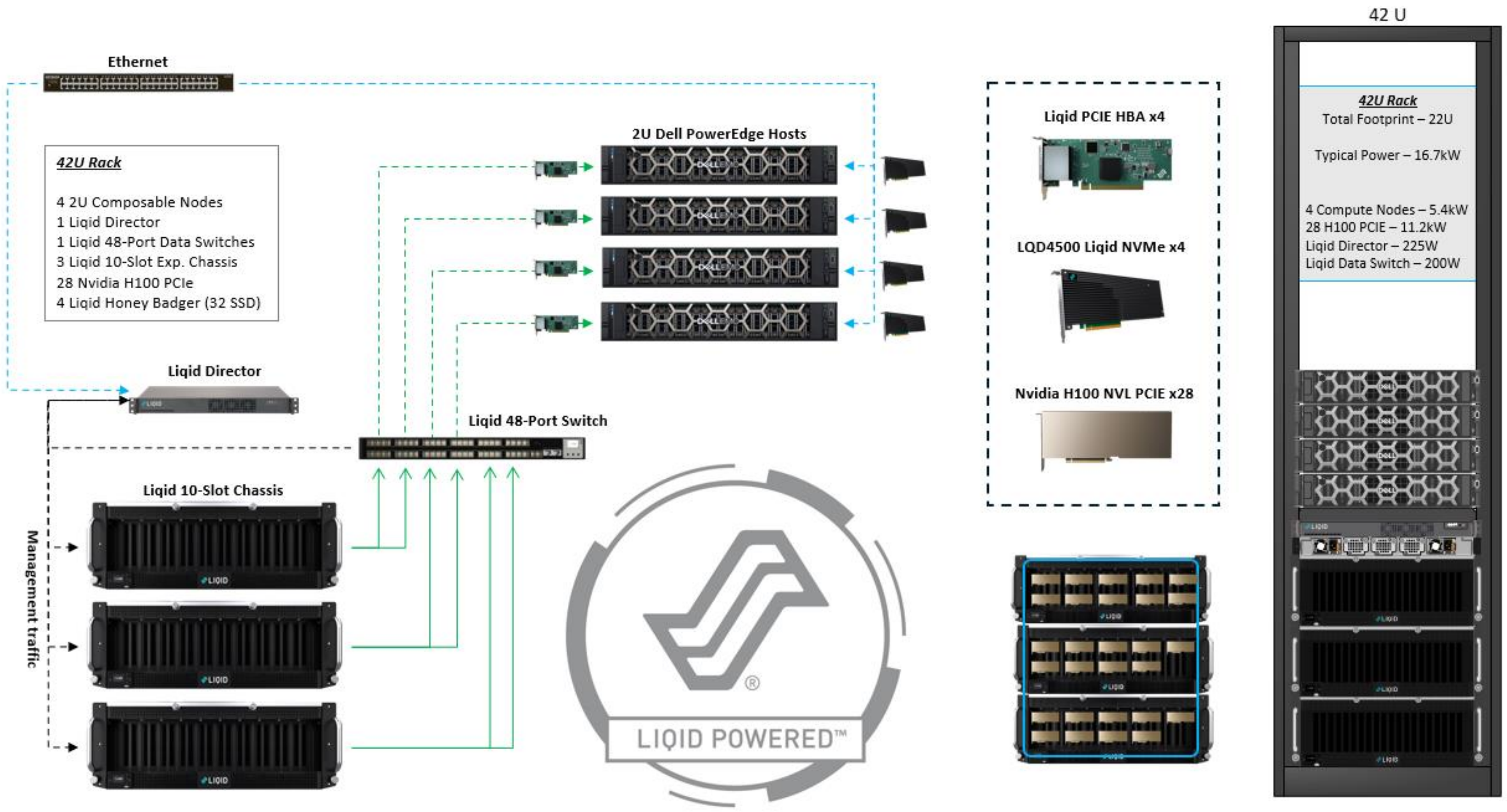
Liqid system installed in E4 Data Center and on customer's infrastructure

Different GPUs mounted (various models of NVIDIA and AMD

Connetected to ExaMon

Planned activity:

- Standard benchmark (e. g. HPL) and dedicated workloads (e. g. GROMACS)
- Check flop per watt tradeoff



Power Consumption estimates based on typical power usage

CHERENKOV TELESCOPE MINI ARRAY

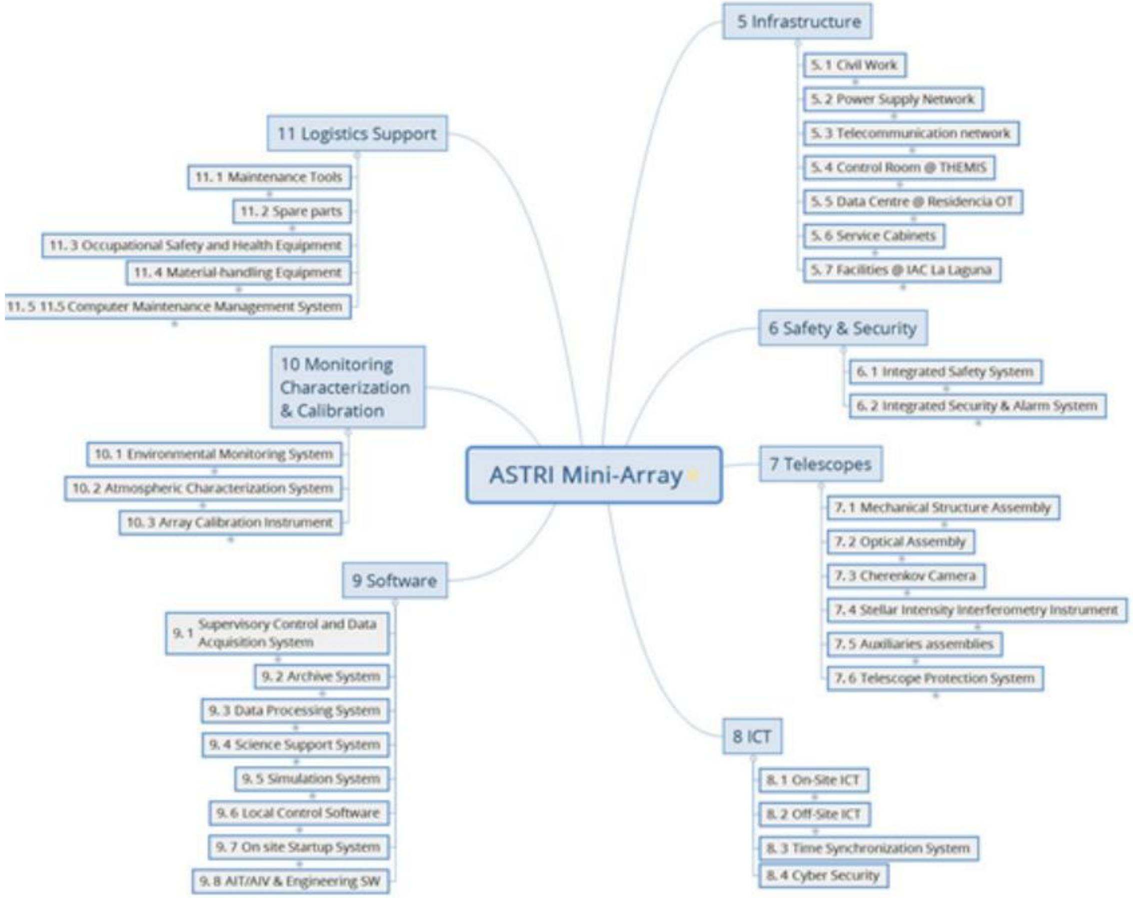
INAF
ISTITUTO NAZIONALE
DI ASTROFISICA

Year	2024-2025 (active)
Customer	INAF @Teide observatory
Solution	Design, Installation
	The **ASTRI Mini-Array** is an **INAF**-led project to build nine double-mirror Čerenkov telescopes to study astrophysical sources emitting high-energy gamma photons with the goal of **measuring the energy, direction, and arrival time of gamma photons arriving on Earth from astrophysical sources**, by exploiting the Čerenkov atmospheric imaging technique
	E4 has designed and installed an infrastructure to manage and monitor an array of nine Čerenkov telescopes and process the data collected in real time

ASTRI – MA: REQUIREMENTS

Hardware and operational requirements:

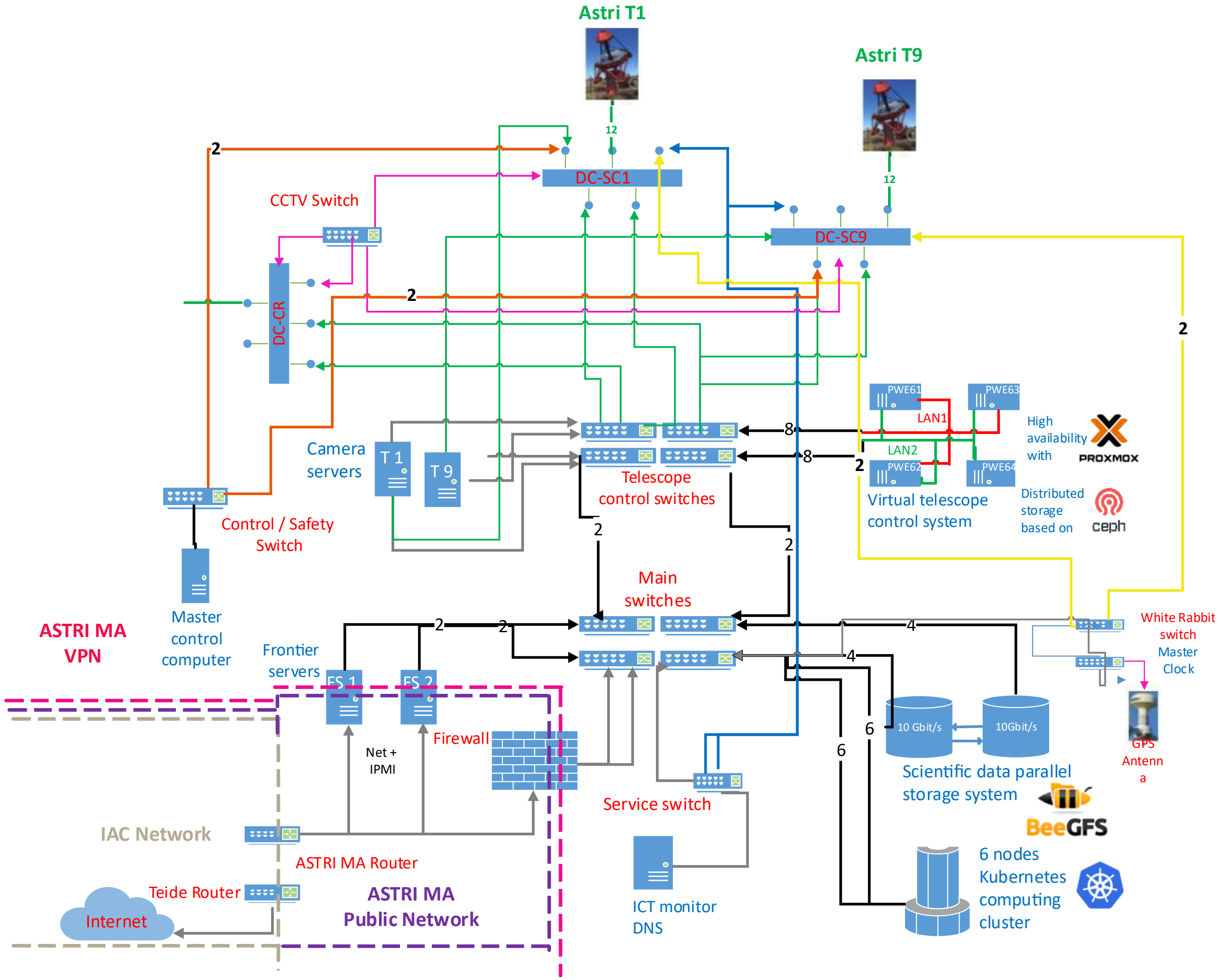
- Install a cluster to control the telescopes managed by a central control plane
- Install a data acquisition and preliminary analysis platform with the purpose of running tests on acquired data and transmit only relevant data
- Within the budget provide both computational power and resiliency to operate in a "semi unsupervised mode"
- Roll out the existing preliminary infrastructure while installing the new one
- Maximise ease of transistion by using as much possible software tools known to INAF staff
- *Assist the customer through the whole process up to the final run of the test bed in Tenerife*


ASTRI Mini-Array
5 Infrastructure
5. 1 Civil Work
5. 2 Power Supply Network
5. 3 Telecommunication network
5. 4 Control Room @ THEMIS
5. 5 Data Centre @ Residencia OT
5. 6 Service Cabinets
5. 7 Facilities @ IAC La Laguna
6 Safety & Security
6. 1 Integrated Safety System
6. 2 Integrated Security & Alarm System
7 Telescopes
7. 1 Mechanical Structure Assembly
7. 2 Optical Assembly
7. 3 Cherenkov Camera
7. 4 Stellar Intensity Interferometry Instrument
7. 5 Auxiliaries assemblies
7. 6 Telescope Protection System
8 ICT
8. 1 On-Site ICT
8. 2 Off-Site ICT
8. 3 Time Synchronization System
8. 4 Cyber Security
9 Software
9. 1 Supervisory Control and Data Acquisition System
9. 2 Archive System
9. 3 Data Processing System
9. 4 Science Support System
9. 5 Simulation System
9. 6 Local Control Software
9. 7 On site Startup System
9. 8 AIT/AIV & Engineering SW
10 Monitoring Characterization & Calibration
10. 1 Environmental Monitoring System
10. 2 Atmospheric Characterization System
10. 3 Array Calibration Instrument
11 Logistics Support
11. 1 Maintenance Tools
11. 2 Spare parts
11. 3 Occupational Safety and Health Equipment
11. 4 Material-handling Equipment
11. 5 11.5 Computer Maintenance Management System

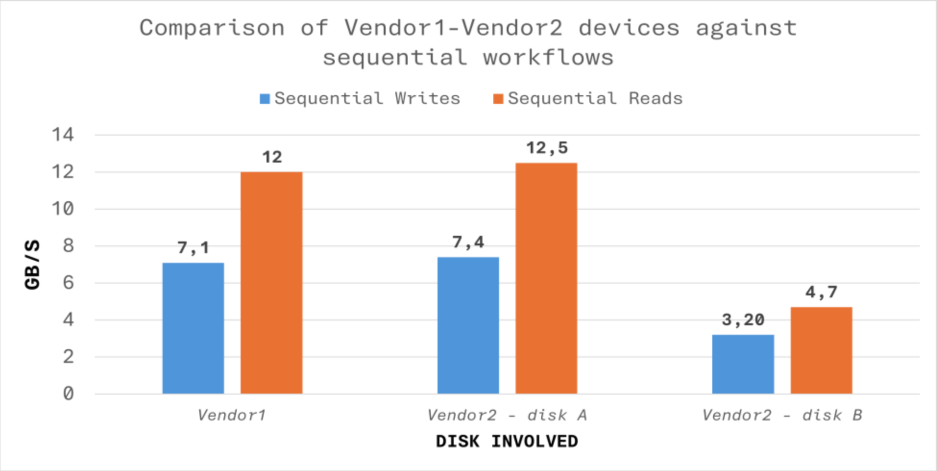
ASTRI – MA: IMPLEMENTATION

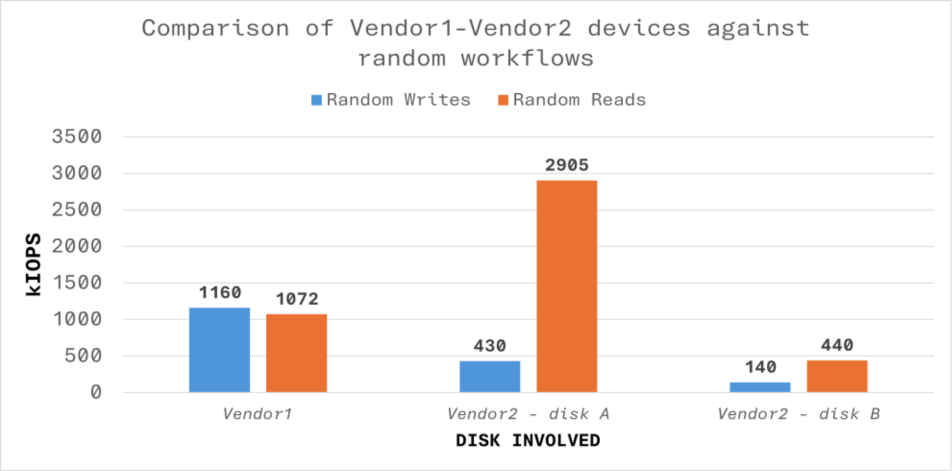
Key points

- Create a control plane based on ProxMox for simplicity but using OpenStack components (Ironic, Bifrost) for the lifecycle management of the infrastructure
- Use Medooza components for setup
- Create a Kubernets / Rancher cluster to analyze data from telescopes
- Use BeeGFS to provide the storage; to contain costs and complication consolidates Storage and Metadata Services into two storage systems, while Management is virtualized. Redundancy is ensured through Buddy Mirroring for both data and metadata servers.



Evaluating NVME and NICs performance under different conditions

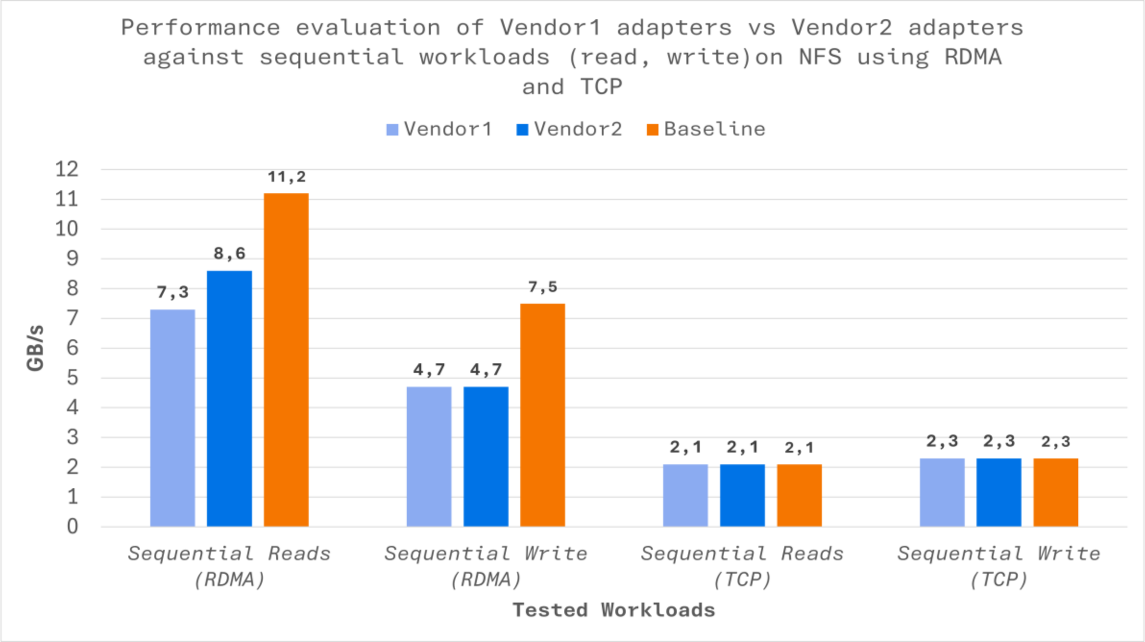

Comparison of Vendor1-Vendor2 devices against sequential workflows


Comparison of Vendor1-Vendor2 devices against random workflows

Test platform:

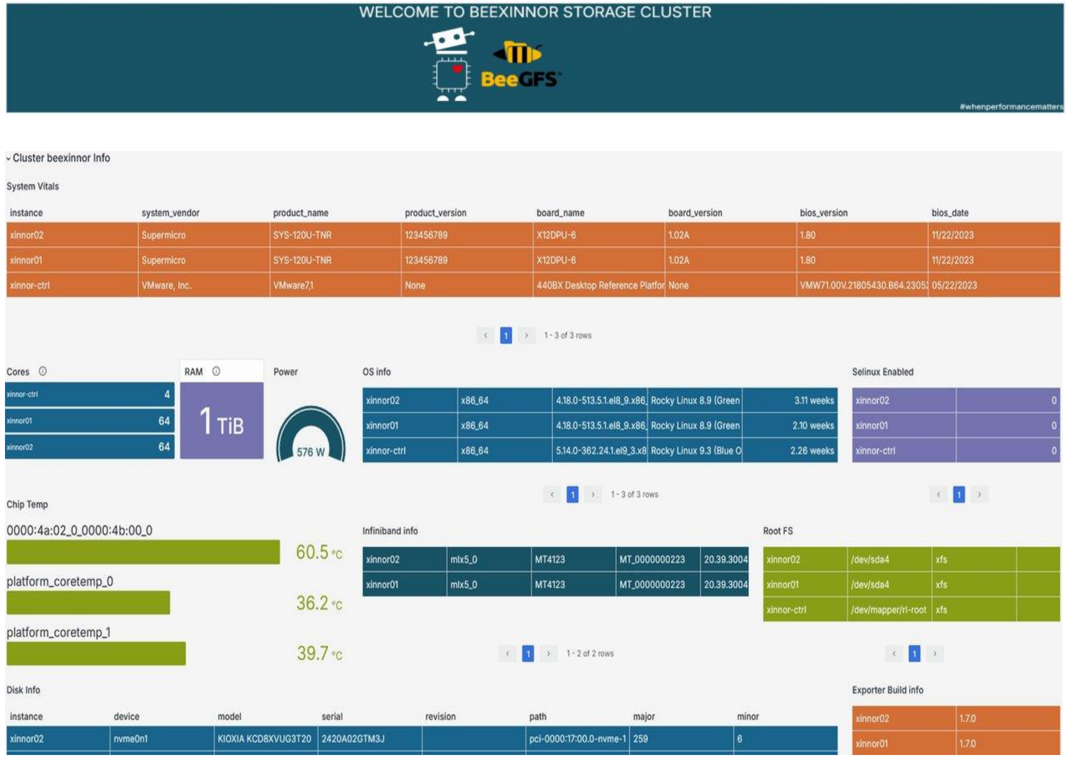
OS	AlmaLinux 9.5 (Teal Serval)
Kernel_version	5.14.0-503.31.1.el9_5.x86_64
FIO_version	fio-3.35
mdRAID_version	v4.3
NVME-CLI_version	2.9.1

NFS preliminary results carried out from VMs and baremetal nodes


Performance evaluation of Vendor1 adapters vs Vendor2 adapters against sequential workloads (read, write)on NFS using RDMA and TCP

STORAGE SOLUTIONS

Among the first to Implement NVMe RAID Solutions for the Beegfs Parallel Filesystem!



WELCOME TO BEEXINNOR STORAGE CLUSTER

#whenperformancematters

Cluster beexinnor Info

System Vitals

instance	system_vendor	product_name	product_version	board_name	board_version	bios_version	bios_date
xinnor02	Supermicro	SYS-120U-TNR	123456789	X12DPU-6	1.02A	1.80	11/22/2023
xinnor01	Supermicro	SYS-120U-TNR	123456789	X12DPU-6	1.02A	1.80	11/22/2023
xinnor-ctrl	VMware, Inc.	VMware7,1	None	440BX Desktop Reference Platfor	None	VMW71.00V.21805430.B64.2305	05/22/2023

1 - 3 of 3 rows

Cores

xinnor-ctrl	4
xinnor01	64
xinnor02	64

RAM

1 TiB

Power

576 W

OS info

xinnor02	x86_64	4.18.0-513.5.1.el8_9.x86_	Rocky Linux 8.9 (Green	3.11 weeks
xinnor01	x86_64	4.18.0-513.5.1.el8_9.x86_	Rocky Linux 8.9 (Green	2.10 weeks
xinnor-ctrl	x86_64	5.14.0-362.24.1.el9_3.x8	Rocky Linux 9.3 (Blue O	2.26 weeks

1 - 3 of 3 rows

Selinux Enabled

xinnor02	0
xinnor01	0
xinnor-ctrl	0

Chip Temp

0000:4a:02_0_0000:4b:00_0 60.5 °C

platform_coretemp_0 36.2 °C

platform_coretemp_1 39.7 °C

Infiniband info

xinnor02	mlx5_0	MT4123	MT_0000000223	20.39.3004
xinnor01	mlx5_0	MT4123	MT_0000000223	20.39.3004

1 - 2 of 2 rows

Root FS

xinnor02	/dev/sda4	xfs
xinnor01	/dev/sda4	xfs
xinnor-ctrl	/dev/mapper/rl-root	xfs

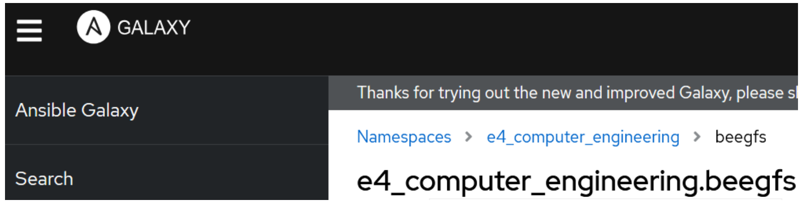
Disk Info

instance	device	model	serial	revision	path	major	minor
xinnor02	nvme0n1	KIOXIA KCD8XVUG3T20	2420A02GTM3J		pci-0000:17:00.0-nvme-1	259	6

Exporter Build info

xinnor02	1.7.0
xinnor01	1.7.0

Automating: Public Release of the Ansible Collection for Beegfs!

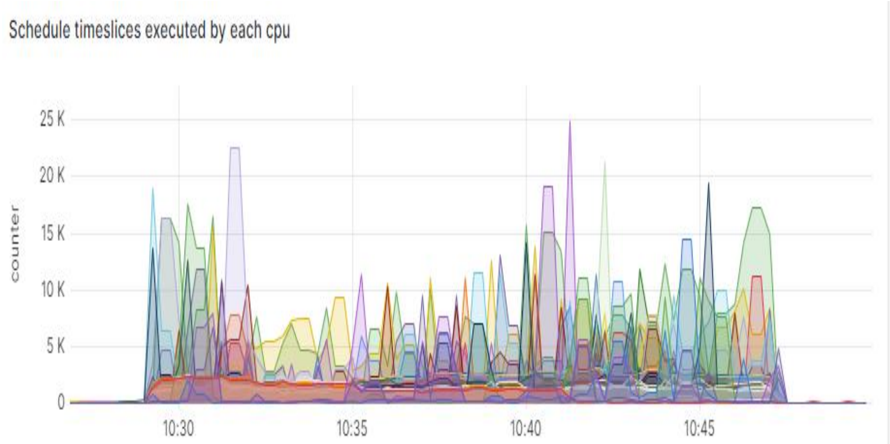


Thanks for trying out the new and improved Galaxy, please s

Namespaces › e4_computer_engineering › beegfs

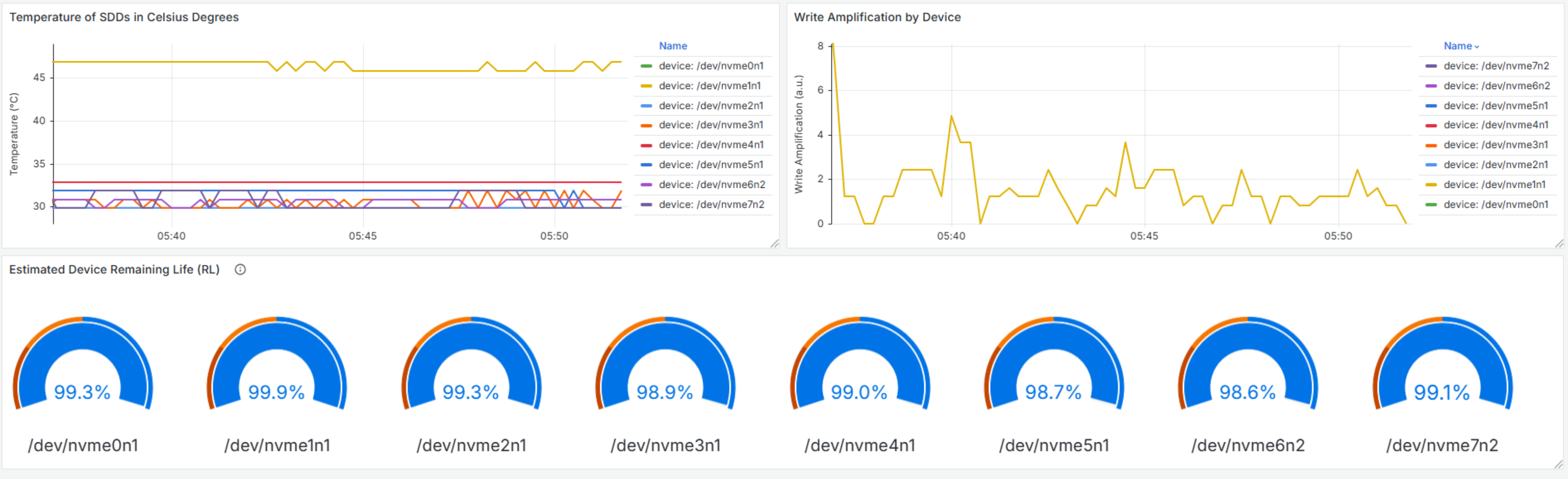
e4_computer_engineering.beegfs

Fair CPU usage with DRBD CPU masking



ARGUS
STORAGE MONITORING & ALERTING SOLUTION

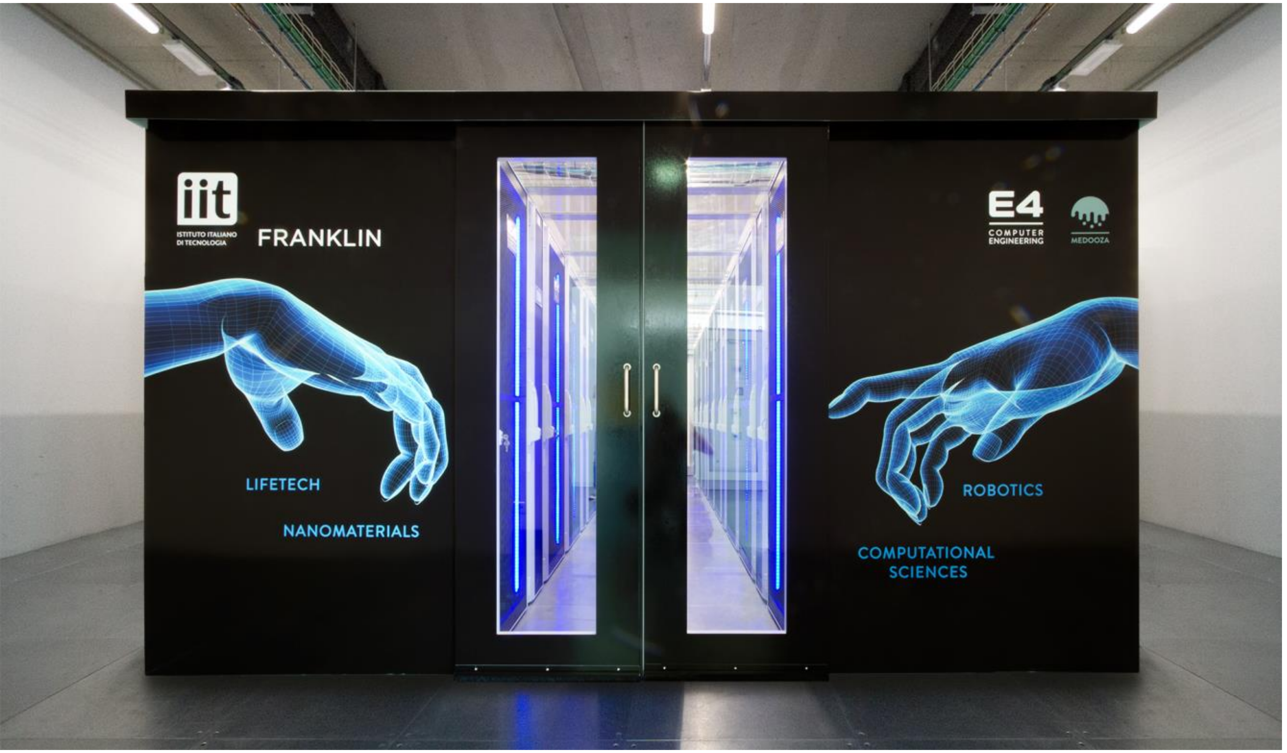
The upcoming release of Argus will allow observing the relative consumption of flash cells present in NVMe drives and thus predicting the lifespan of the drive and related carbon emissions: “Operational cost” and “Use phase”, in line with the **OCP sustainability initiative**.



Key features:

It will extend to monitoring the entire E4’s Medooza HPC solution & will integrate with Examon for metrics ingestion.

IIT (GENOA), FRANKLIN CLUSTER (1)

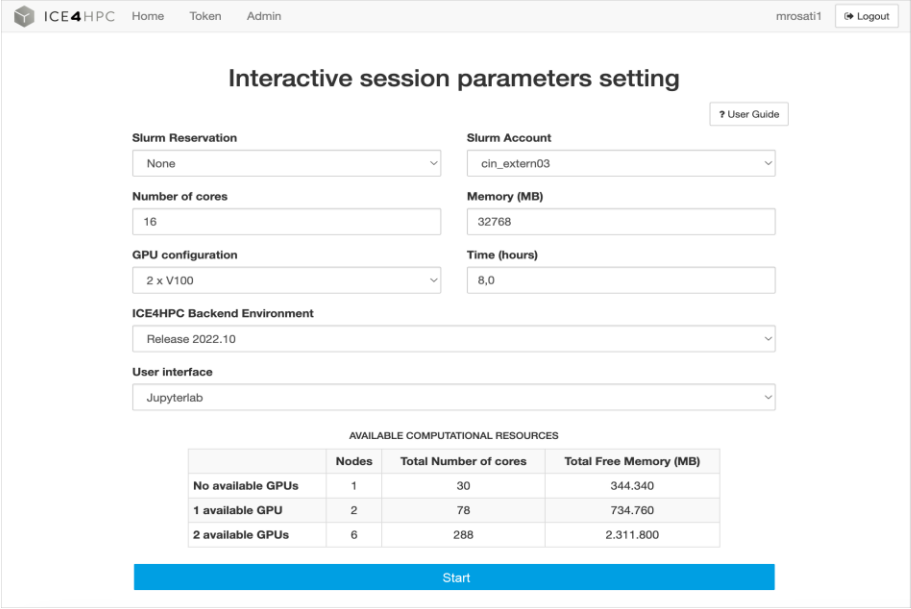


Installed 2021; starting situation - 2023

- CentOS 7.9
- 60 GPU nodes
- Intel based nodes
- 2 CPU + visualization nodes
- Two BeeGFS nodes
- xCAT based configuration control plane
- OpenPBS in HA
- Infiniband and 10 Gbit/s network
- Zabbix monitor
- Several software and software versions installed over the years

FRANKLIN CLUSTER @IIT (2)

RESOURCES SELECTION



Interactive session parameters setting

Slurm Reservation: None

Slurm Account: cin_extern03

Number of cores: 16

Memory (MB): 32768

GPU configuration: 2 x V100

Time (hours): 8,0

ICE4HPC Backend Environment: Release 2022.10

User interface: Jupyterlab

AVAILABLE COMPUTATIONAL RESOURCES

	Nodes	Total Number of cores	Total Free Memory (MB)
No available GPUs	1	30	344.340
1 available GPU	2	78	734.760
2 available GPUs	6	288	2.311.800

After upgrade

- Expansion of networking
- OpenStack Hyperconverged control plane on 3 nodes
- RockyLinux 8.9
- ~ 100 CPU and GPU nodes
- Intel and AMD mixed
- Interactive computing environment
- All software modules reinstalled with new stack and on different architectures

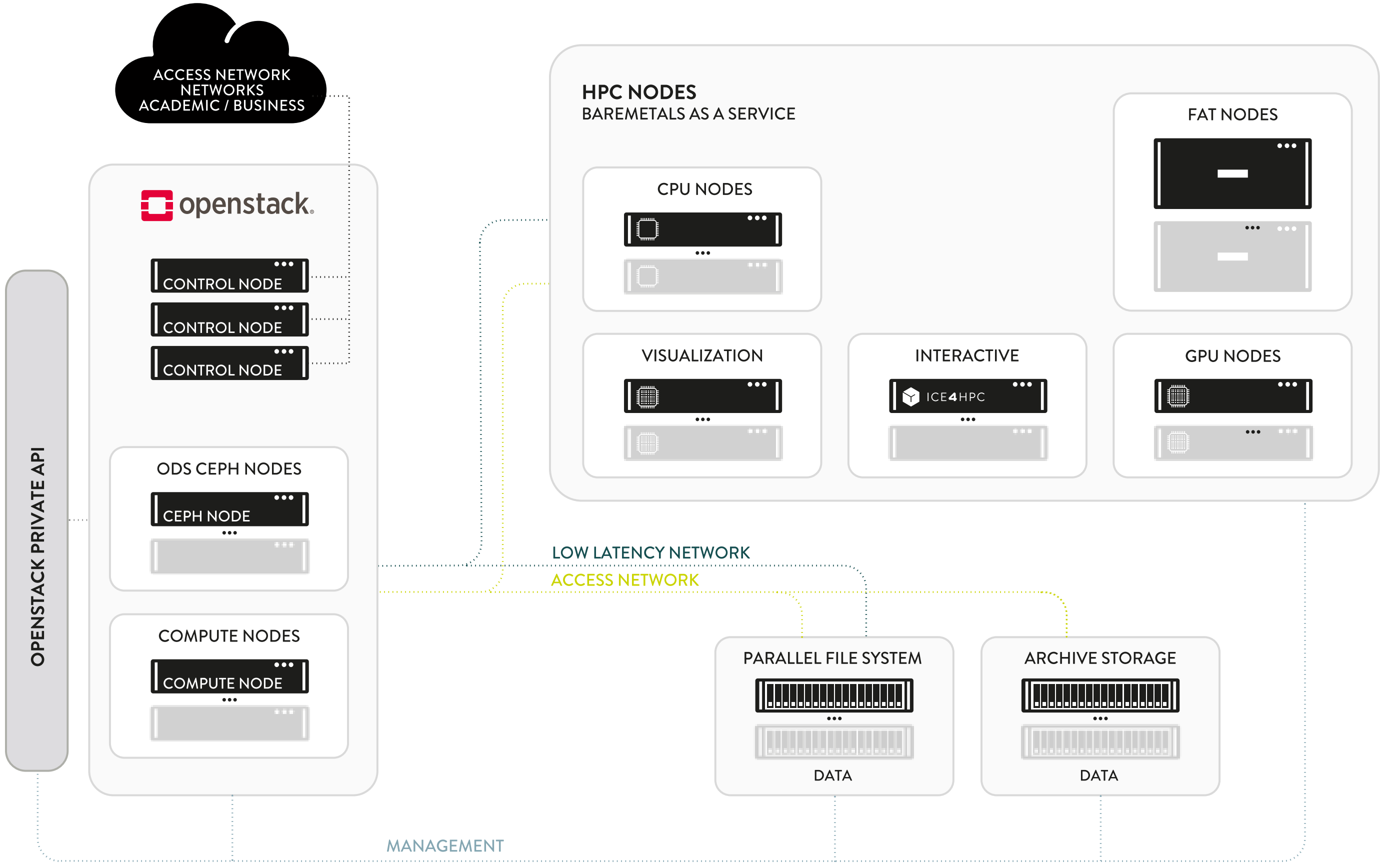
- Latest nodes with A100 GPUs needed by Genomics and Molecular Dynamics groups
- Baremetal & scheduled jobs vs container based workloads
- Strict isolation policies asked by DPO for Genomics data

→

- 100 nodes of mixed generation
- BeeGFS Storage
- Three node Open Stack control plane
- Cloud infrastructure allows to **reinstall with a different setup** nodes on demand
- Assign them to different networks
- Enforce firewall or other security measures

MEDOOZA
CURRENT REFERENCE ARCHITECTURE

MEDOOZA = + OpenTofu OpenBao + TALOS + + ARGUS + EXAMON + + ICE4HPC



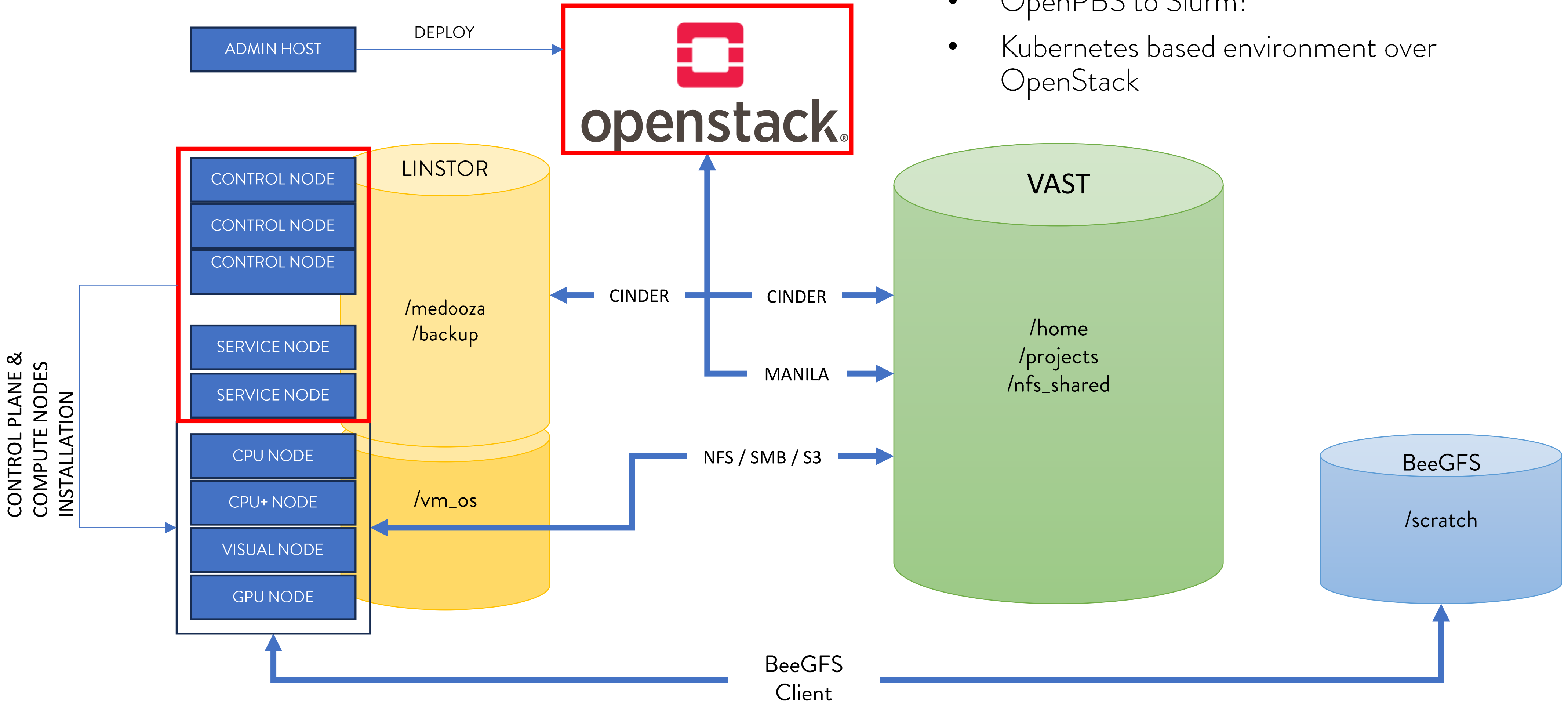
DEPLOYMENT AND MANAGEMENT

Medooza Infra & **Medooza Ansible** deployment solutions

- Create management VMs on the OpenStack control plane (e. g. Scheduler, BeeGFS)
- User's secrets management during deployment
- Bootstrap Ansible installation
- Additional OS packages, basic settings
- Drivers
- Configuration of services on VMs
- Base stack for including compilers, MPI and mathematical libraries, container stack and Python/R environments. Each release is tested against a set of versions and options. It’s basic environment the admins find on login.
- Compilers and mathematical libraries from scratch
- Release tests for the HPC Cluster from services to job submission and benchmarks
- Rocky 8.x, 9.x, Ubuntu Jammy and Noble
- Update plan: 1 release per year, ships a validated set of components and versions, support up to 3/5 years

NEW CLUSTER 2025

- Completely new cluster
- Standard Open Stack deployment with Kayobe
- Upgrade of Franklin after installation
- OpenPBS to Slurm?
- Kubernetes based environment over OpenStack



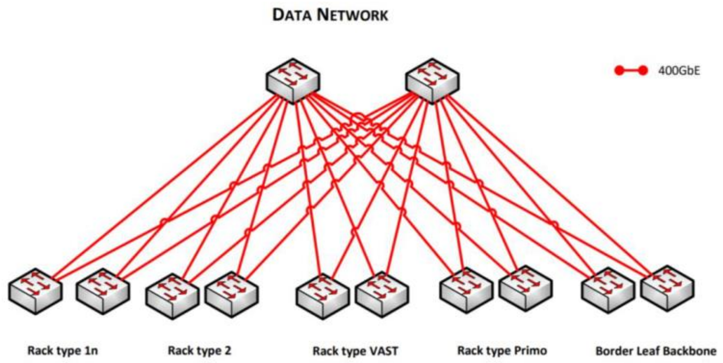
GALILEO100+ @ CINECA TECHNOPOLE

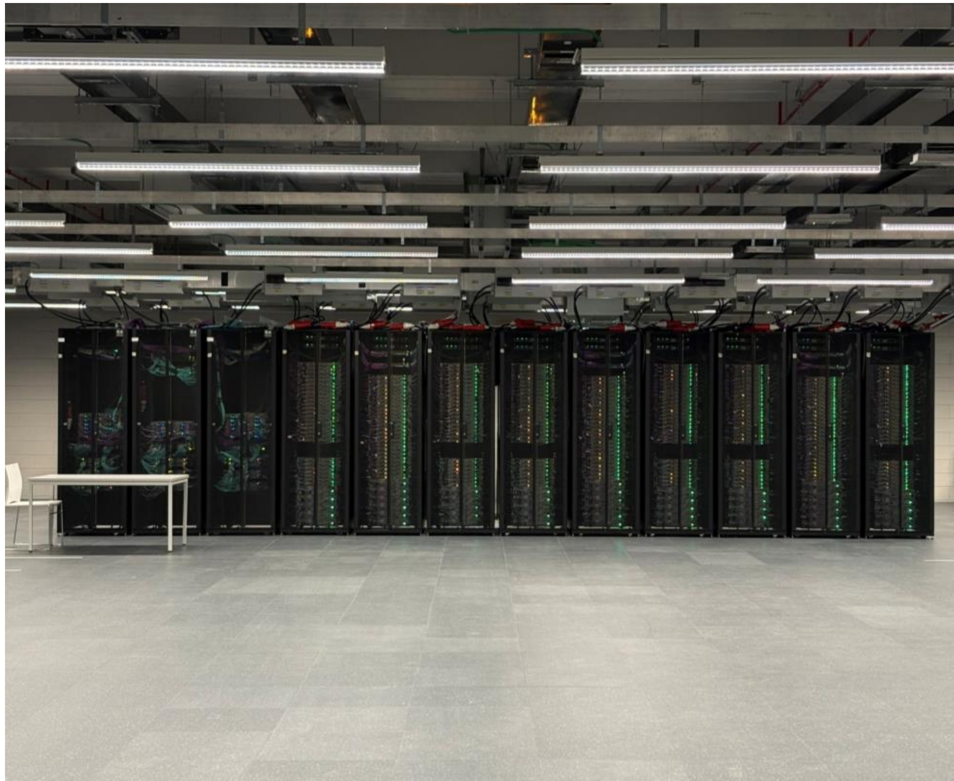
OpenStack Cluster:

- 420 CPU node DELL PowerEdge R670 Server CSP Edition with 2x CPU Intel Xeon Sierra Forest 144 core
- 80 GPU node with NVIDIA GPU 112x L40S + 80x A30 + 8x H100
- 4 GPU Server with 8x NVIDIA Hopper H100
- 62 node Ceph / Management / Service
- 400 GbE spine/leaf network with DELL Powerswitch Z9432F-ON + Z9664
- 50 PB full flash storage Datalake
- 4.35PB Disk-to-Disk Long Term Storage

- Examon HPC Monitoring (Exascale Monitoring) for Data Collection and Analysis.
- E4 Medooza Cluster Management Suite on the c. p.
- Complete system installed with OpenStack Kayobe
- E4 on-site permanent support with 60 minutes intervention for critical incident

TOTAL: 131.200 physical cores



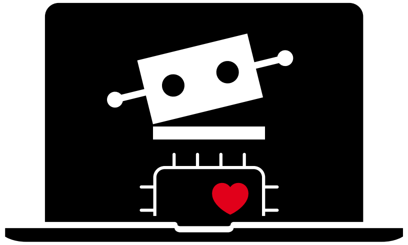


SUMMARY & PERSPECTIVES

- Cloud based HPC scales from small to big infrasctures
- «One size fits all» is impossible but is important to focus on a few but flexible tools
- Small penalties in performance lead up to 20% energy savings

- Support for mixed architectures in Medooza (x86_64, ARM, RiscV)
- Deployment of MLOps/AIOps solutions in production in ExaMon

THANK YOU FOR YOU ATTENTION!



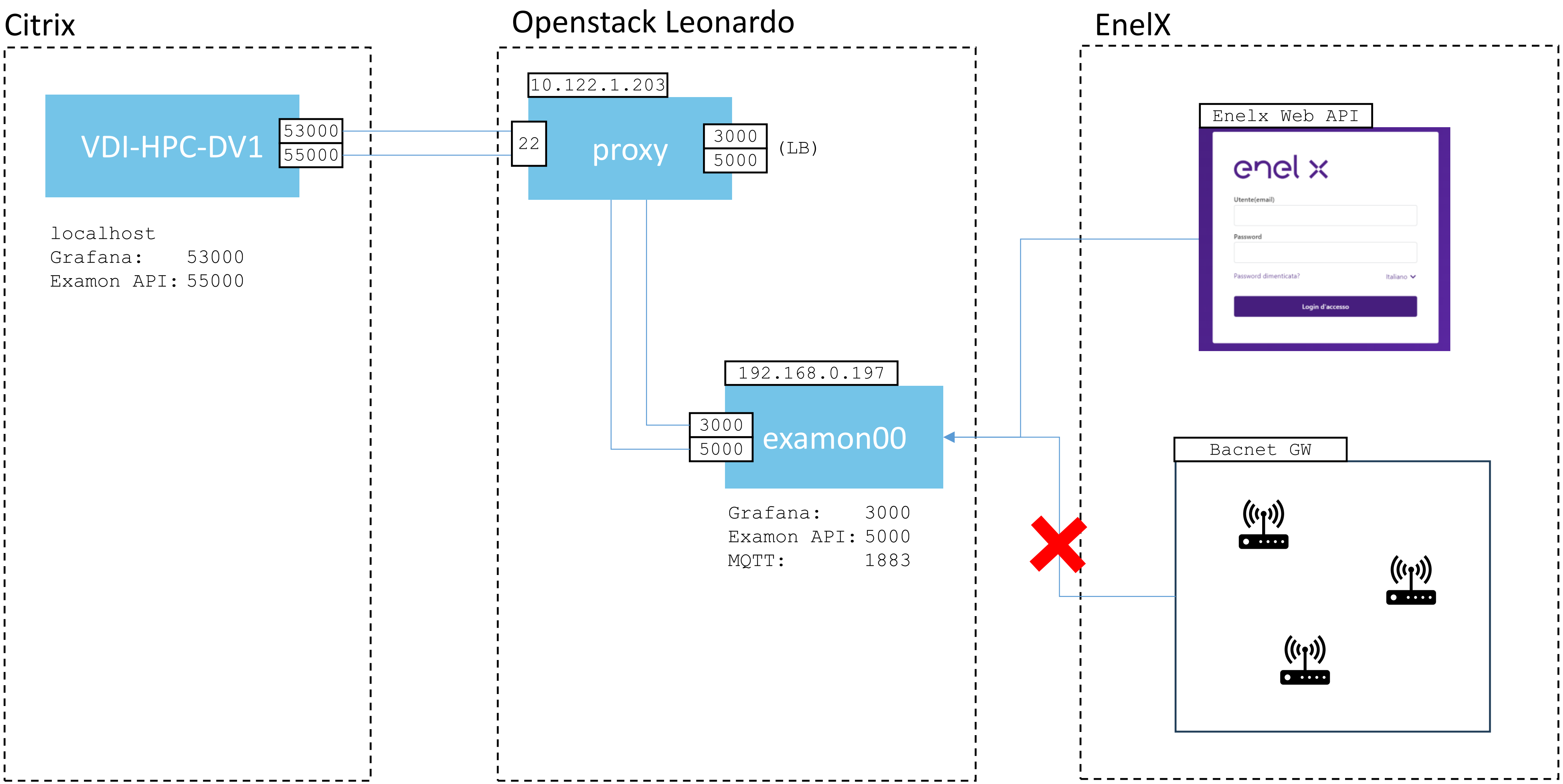
www.e4company.com

This slide left blank

E4
COMPUTER
ENGINEERING

EnelX – Room Meters

Confidential



Citrix

VDI-HPC-DV1

53000
55000

```
localhost
Grafana:     53000
Examon API: 55000
```

Openstack Leonardo

10.122.1.203

22

proxy

3000
5000
(LB)

192.168.0.197

3000
5000

examon00

```
Grafana:     3000
Examon API: 5000
MQTT:        1883
```

EnelX

Enelx Web API

Bacnet GW

Evolution of ExaMon: perspectives

Goals

- Find anomalies, solve problems (the current project)
- Improve efficency, investigate job properties from metadata (possible?)
- Extended to new HW in a standard way

Data flow (now / tomorrow):

- metrics ingested into Cassandra / IoTDB
- Streamed to Kafka for real-time processing
- written to S3 with iceberg
- Queried with Trino & Spark
- Visualized with Grafana
- Basic ML models trained on historical data for forecasting
- Integrated with foundation models & used with a conversational interface

Tool

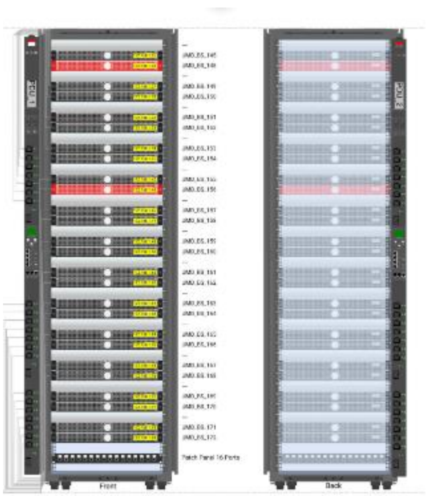
- Reference architecture for stand alone analytics platform: «ExaMon Box»

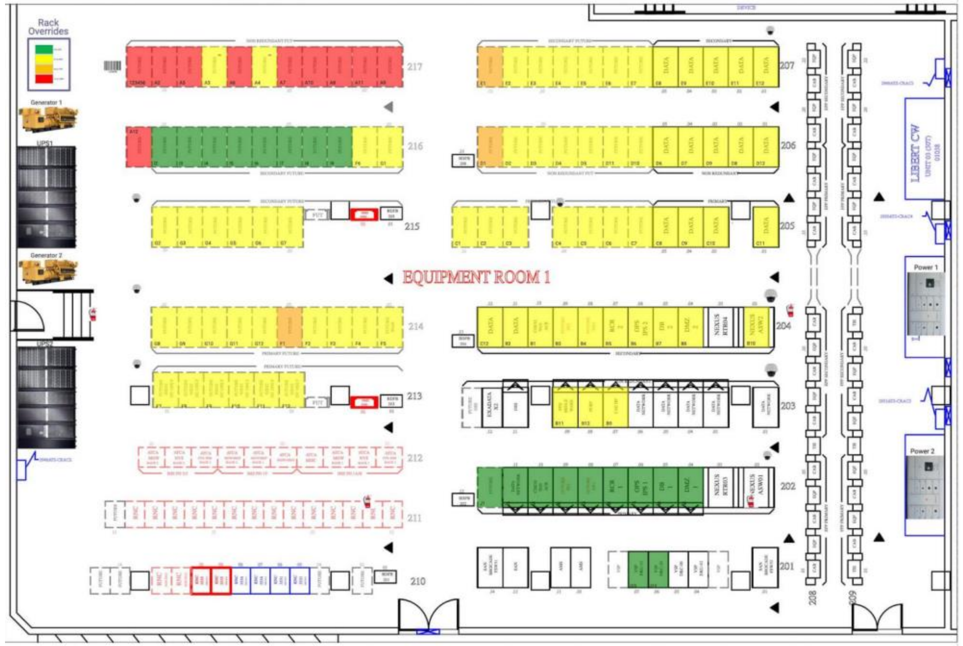
URANIA + EXAMON

E4'S SERVICES

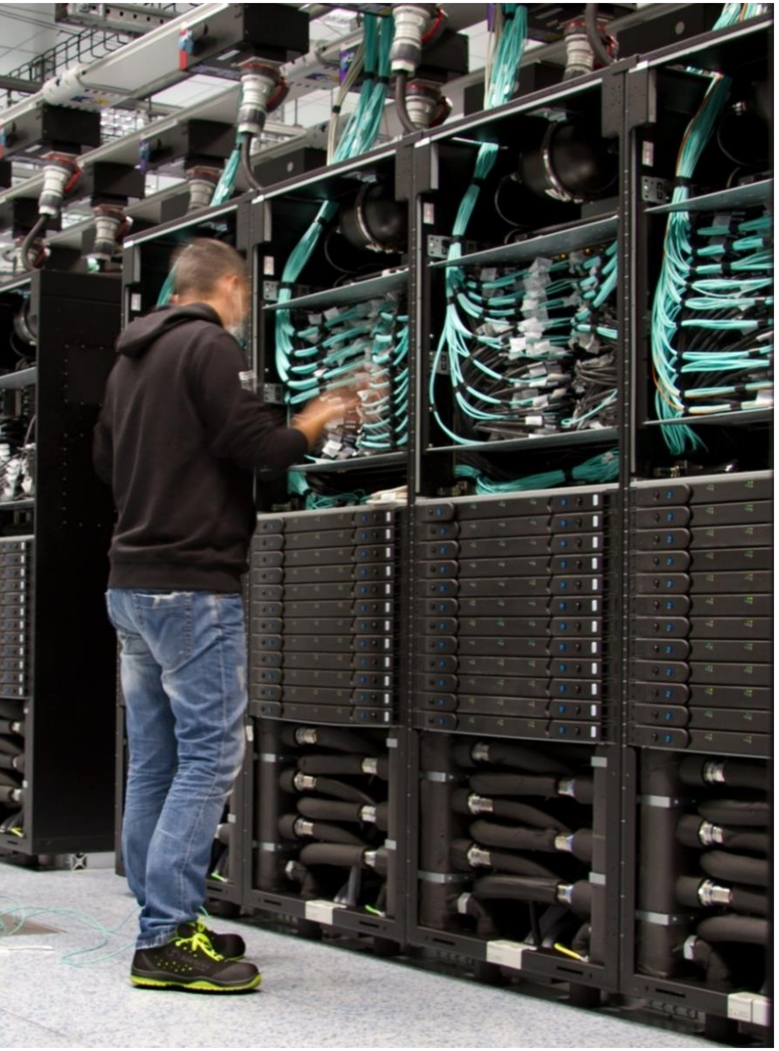
GUIDELINES FOR THE LAYOUT FOR RACKS AND CABINETS

- Track and visualize rack-mounted equipment and available power, networks, and physical space
- Decommission equipment and connections fully with user-friendly DCIM software
- Quickly reconfigure devices and reserve equipment space via drag-and-drop
- Monitor capacity and changes at-a-glance with easy-to-understand color coding
- View floor plans by rack units or weight, and oversubscribe racks based on actual power





E4'S SERVICES
REPORTING



- Comprehensive reporting, including data center capacity, asset usage, and connectivity reports
- Manage work orders and tasks with details on due dates, owners, and statuses
- Change management
- Easily track rack utilization, weight, power, device types, etc.
- Stay organized with DCIM's endpoint connectors, link types, and cable usage information