EUROPEAN
SPALLATION
SOURCE

EPICS Archiver Appliance Infrastructure at ESS

ESS deployment

PRESENTED BY STEPHANE ARMANET

2024-04-16

Agenda

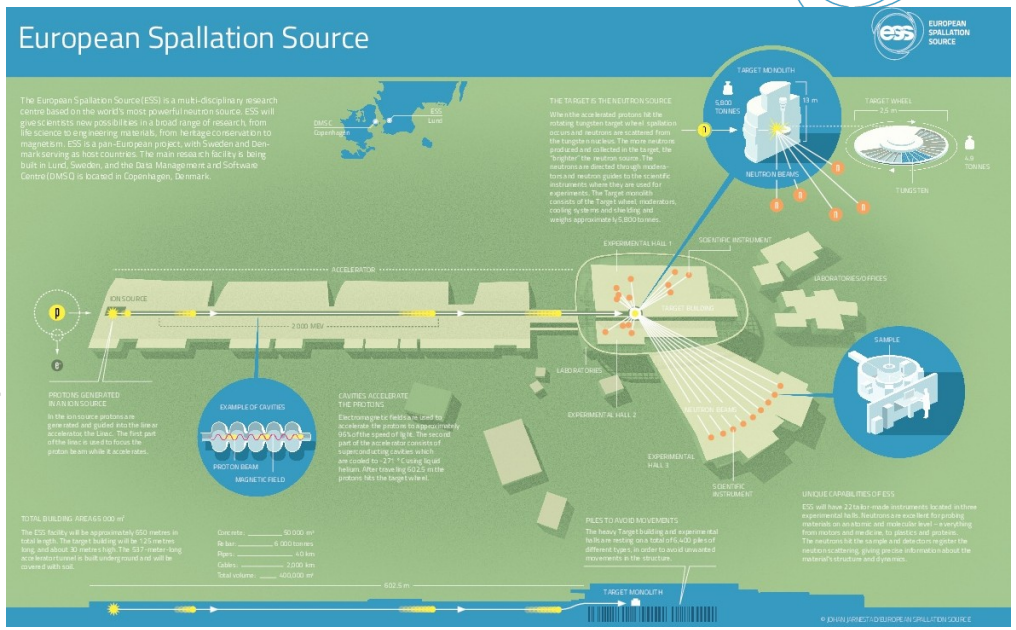
1	ESS introduction
2	Controls Infrastructure
3	EPICS Archiver deployment
4	Monitoring
5	Questions

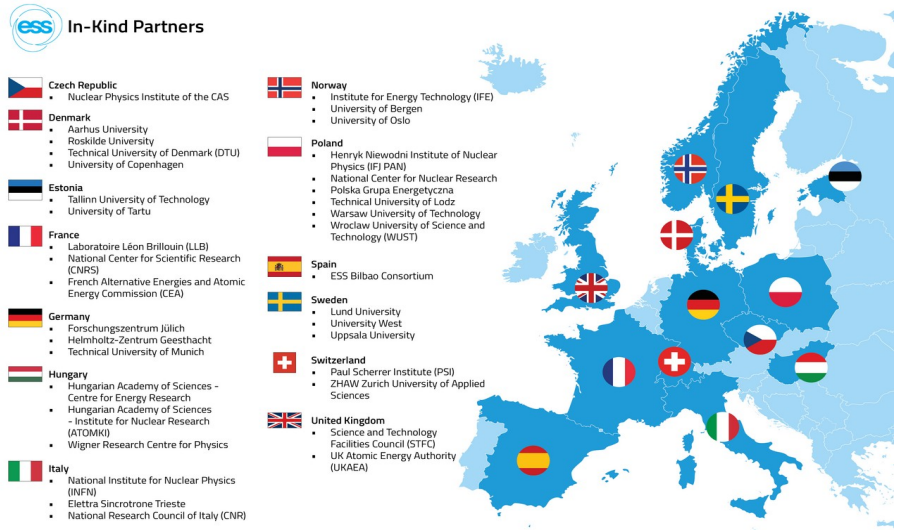
European Spallation Source

In a nutshell

ESS is :

- accelerator based neutron source in construction in Lund – Sweden
- A collaboration of 13 European Countries
- To become the most powerful neutron source in the world
  - 5MW beam power (initial target set at 2MW), 2.5GeV proton energy, pulsed at 14 Hz
  - Initial set of 15 neutron instruments (as part of the construction budget)
- Will cover a large spectrum of research (energy, Health and life science, Information Technology, Nanoscience , Environment Sciences, Heritage …)
- Project deliveries are a collaboration between ESS internal solutions and in-kind deliveries





Infrastructure for Control systems

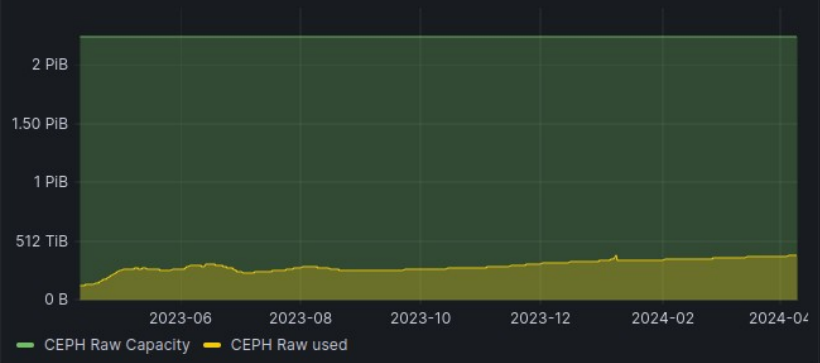
IT Infrastructure for Controls

Virtualisation and storage overview

- IT infrastructure for Controls is built with openSource solutions:
  - Virtualization :
    - Proxmox VE – 35 hosts, 5 clusters , 1300 VMs
  - Storage:
    - CEPH :
      - RBD Block storage for VM backend
      - CephFS for shared filesystem (native or NFS gateways)
      - Combination of HDD and NVMe
      - 2.2 PB raw, 500 OSDs, 60 servers → will increase by 50% this summer
    - ZFS NAS
      - NFS servers
      - Long term storage (3 PB net)
      - Backups (4 PB net)

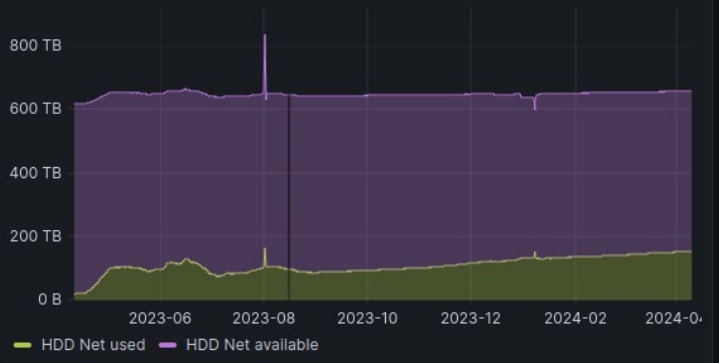
Storage

CEPH Global Raw Capacity
Last 365 days

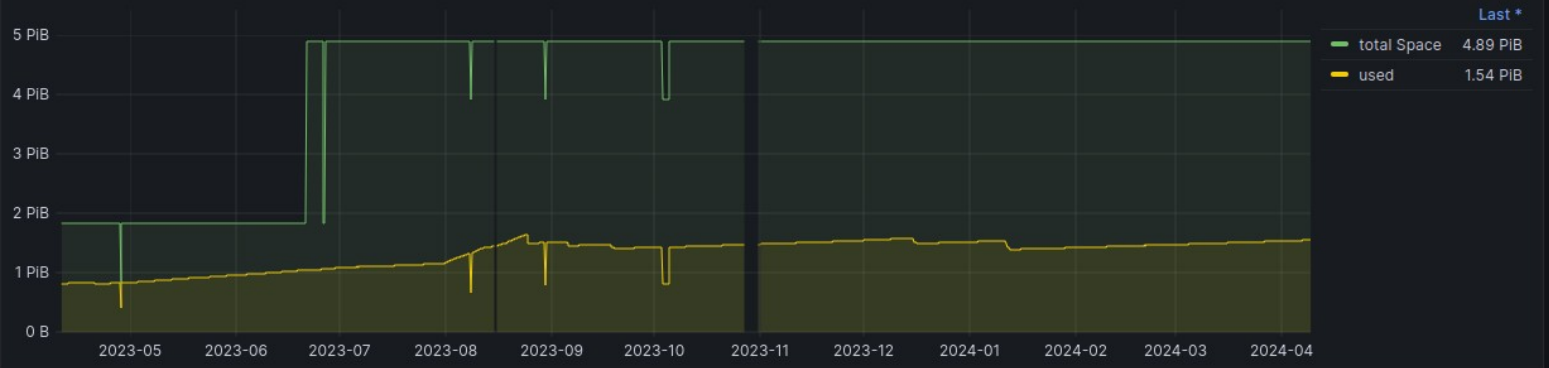


CEPH NVMe Net capacity (fast layer)
Last 365 days

CEPH HDD Net capacity (capacitive layer)
Last 365 days



ZFS storage - Global (LTS and backup)
Last 365 days



Number of Storage server ...
59

Number of ZFS Storage se...
9

Number of CEPH Storage s...
50

Virtualisation

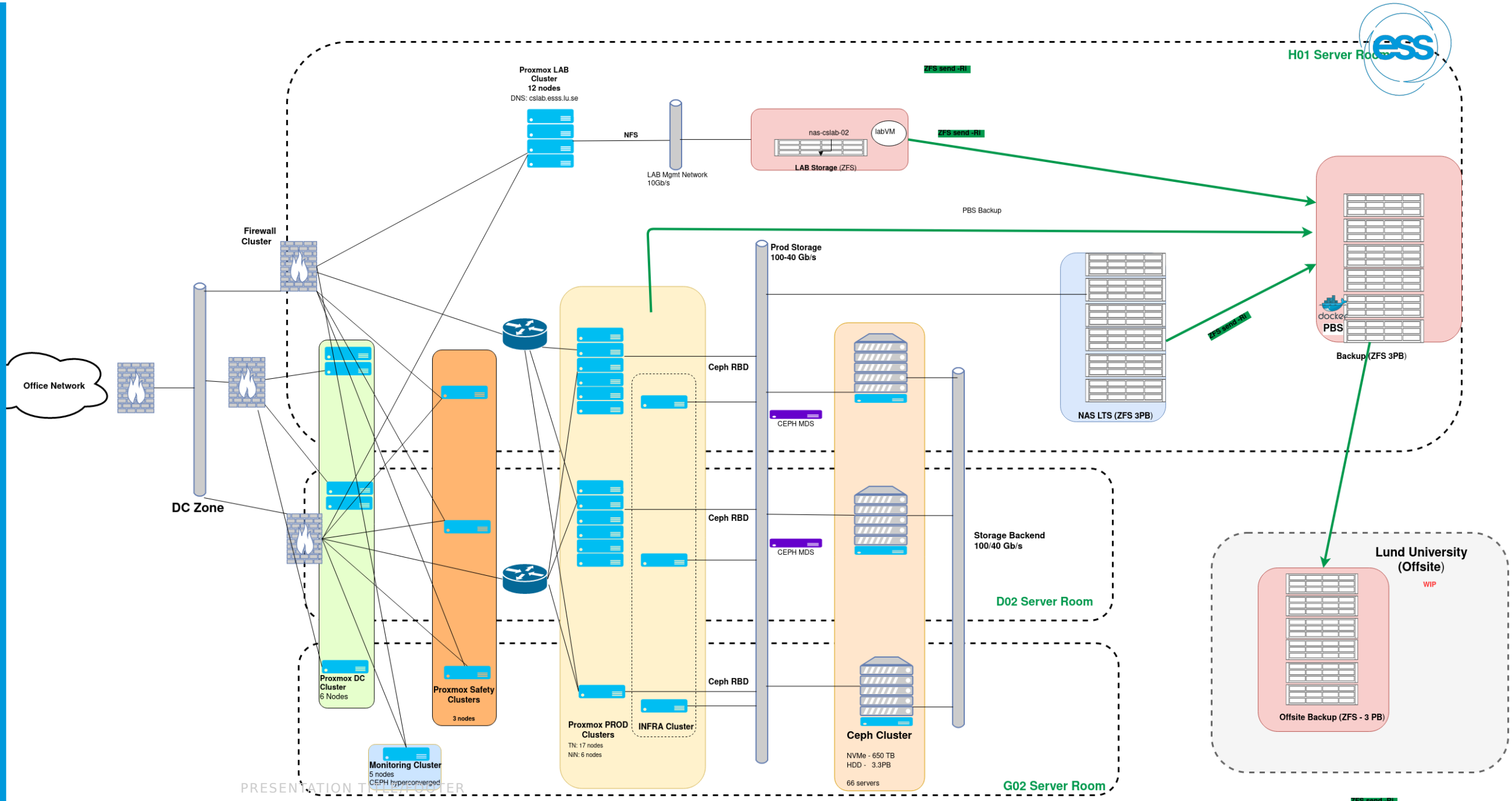
Number of virtualisation servers (total)
36

Proxmox Nodes per cluster

cluster/CSLAB-2022	cluster/DC-2022	cluster/NIN2021	cluster/TN-CLUSTER-2022
10	6	3	17

Number of VMs per cluster

CSI Infrastructure 2024

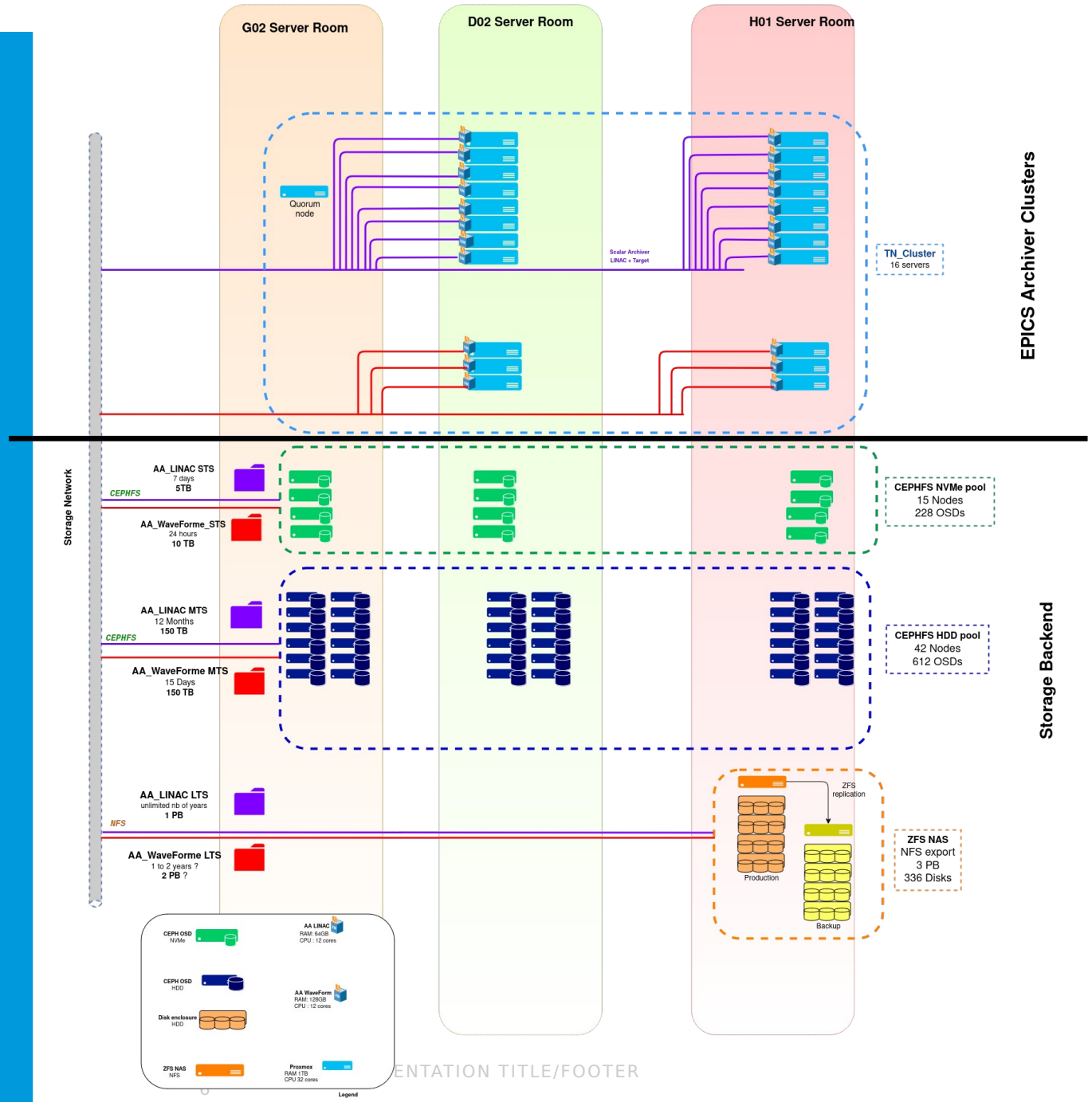


Infrastructure for EPICS Archivers

Control System overview

ESS is an EPICS based Facility

- Technical Network → Linac and Target
  - EPICS 7
  - Includes CRYO, Vaccum, RF, SRF, Target, safety systems (PSS and MPS), BI …
  - ~ 1500 IOCs
  - IOCs running on µTCA, IPCs (Lenovo basic servers) and Vms → mostly centos7 (~ 500 hosts)
  - 2 Control room : LCR (CRYO and Test Stands), MCR for 24/7 operation
  - 2 archiver appliance clusters (next slide)
- CSLAB environment
  - Dev and Test
  - Same size as TN but no real archiving (test clusters)
- Neutron instrument Controls
  - Should be similar size than TN
  - 1 network per instrument but with central archiving (does not cover instrument detector data)
  - Only for technical systems (motion Control, neutron choppers, sample env., EPICS control of the instruments)



- TN Archiving : 2 clusters → 2 policies
  - 1 for scalars, 1 for large wave forms Pvs (75k - 400k points)
  - STS on CephFS NVMe pool
    - Retention: WF 24 hours – Scalars 7 days
  - MTS on CephFS HDD pool
    - Retention: WF 14 days – Scalars 12 months
  - LTS on NFS/ZFS
    - No data deletion
    - Inline Compression (ZFS)
    - Easy to backup/replicate (AA append to protoBuf files → ZFS send/recv)

- Running on VM
  - Not too big (easier to fail-over)
  - Scalars Cluster : Max 100k Pvs per instance (soft limit)
  - Scale-out architecture (max 2 VMs per nodes)

Issues and challenges

What we learned … so far

- Governance → ESS Machine data management maturity
  - Hard to identify which Pvs are strictly required (~15% are actual signals, the rest are parameters)
  - High level policy (per cluster) instead of system or per type of signals
  - Hard to gather future requirements
    - we jumped from 25k Pvs (2022) to 300K for NCL commissioning (2023) to 700k+ … so far!
  - No “quality of service” → same solution cover all types of Pvs (Safety, operation, instrumentation …)
  - Policy we has been keeping from the beginning:
    - No applied decimation
    - 14Hz, forever by default → applied to everything
    - balance between request from system owners and operators vs integrators and Infra

Issues and challenges

What we learned … so far

- Wave Form archiving is challenging
  - Try to archive large WF (up to 400k points) → dedicated cluster (best effort)
  - Upcoming SDS solution :
    - triggered/event based data acquisition to HDF5 files
    - will write to CephFS
    - will allow to inject meta-data into data collections (user tags, post-mortem events, pulseID…)
    - Should provide a good alternative for large waveform
    - Will not answer all use cases
- Performance and retrieval limitations
  - Data retrieval has to go through archiver API (default JSON → custom python pkg to read raw PB)
  - No easy way to discard data after the archiving has wrote to protoBuf (1 file per PV per partition)
  - No way to tag datasets after acquisition (only timestamp) → no way to query the data via a high level language
  - Hard to have a clean environment during installation/maintenance phases
    → lots of disconnected PVS = high broadcast
    → new : ChannelFinder add PVs to the archiver

System Monitoring

System Dashboard

Prometheus and Grafana

- System monitoring
  - Prometheus node_exporter
- Storage monitoring
  - Ceph MGR to Prometheus
  - ZFS extend node_exporter metrics
- Archiver Appliance
  - Custom Prometheus exporter (gather metrics from archiver API)
- Dashboard (Grafana)
  - high level information (PV count and status, storage and network status)
  - Per instance in-depth monitoring (JVM stats, event/s, system load …)

  → has been a great tool to help understanding some internals of the archiver appliance

  → help to plan and monitor cluster expansion (also maintenance progress)

Grafana Dashboard

Grafana Dashboard

Grafana Dashboard

Thanks !!
Questions ?

Title

Sub_title

Text