

ECMWF Site Report – Conf. Stockage de Masse – Lyon, Juin 2026

Pierre-Emmanuel BRINETTE and Enrico FATTIBENE, Bologna/Italia



Outline

- ECMWF
- Data handling @ ECMWF
- Storage infrastructure and HPSS
- Future challenges

ECMWF: three sites, one unique role

ECMWF is the **E**uropean **C**entre for **M**edium-Range **W**eather **F**orecasts.

Established in 1975

- Intergovernmental Organisation
- 23 Member States
- 12 Cooperating States
- over 400 staff in 3 sites

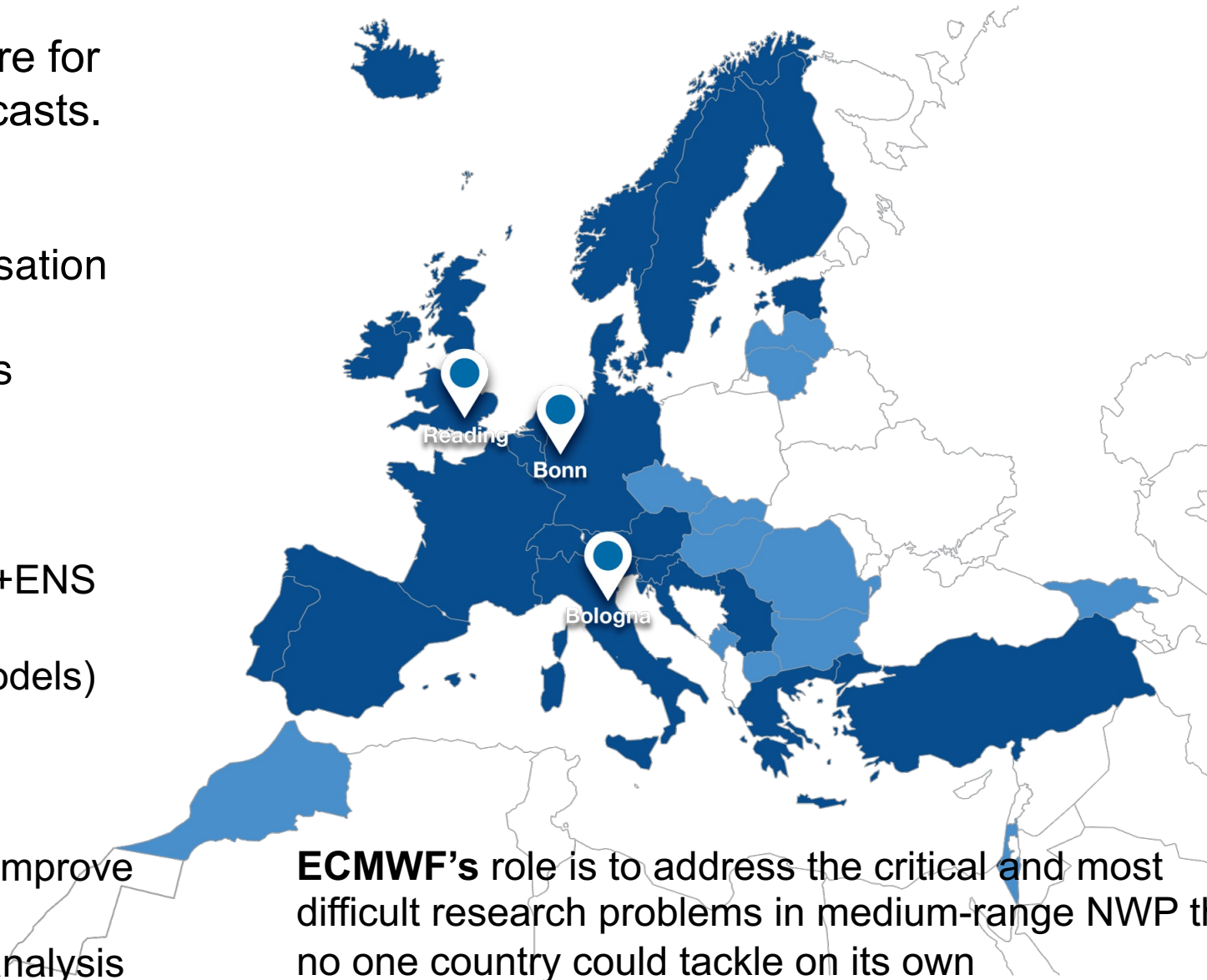
24/7 operational service

- Operational NWP - 4x HRES+ENS forecasts / day
- Supporting NWS (coupled models) and businesses

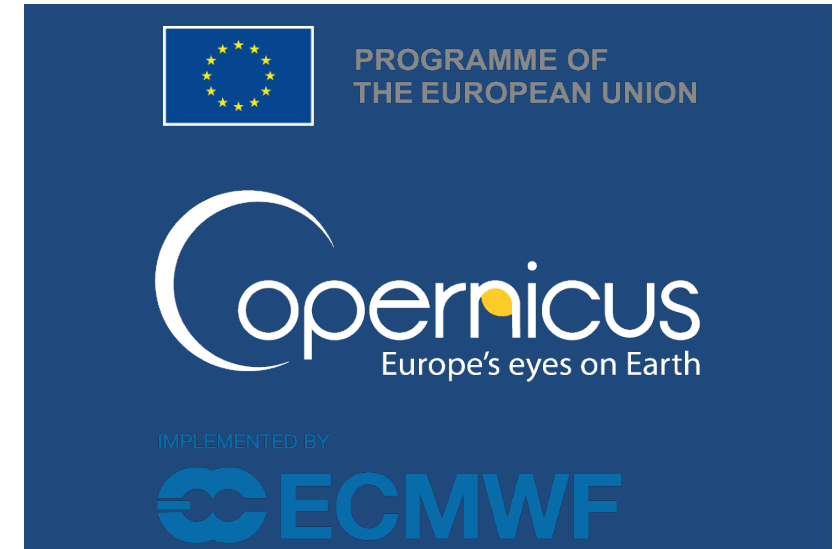
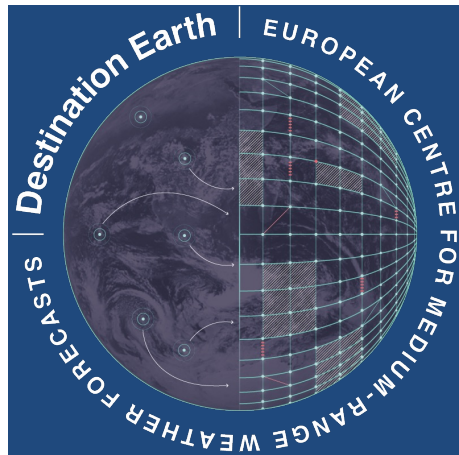
Research institution

- Experiments to continuously improve our models
- Reforecasts and Climate Reanalysis

ECMWF's role is to address the critical and most difficult research problems in medium-range NWP that no one country could tackle on its own



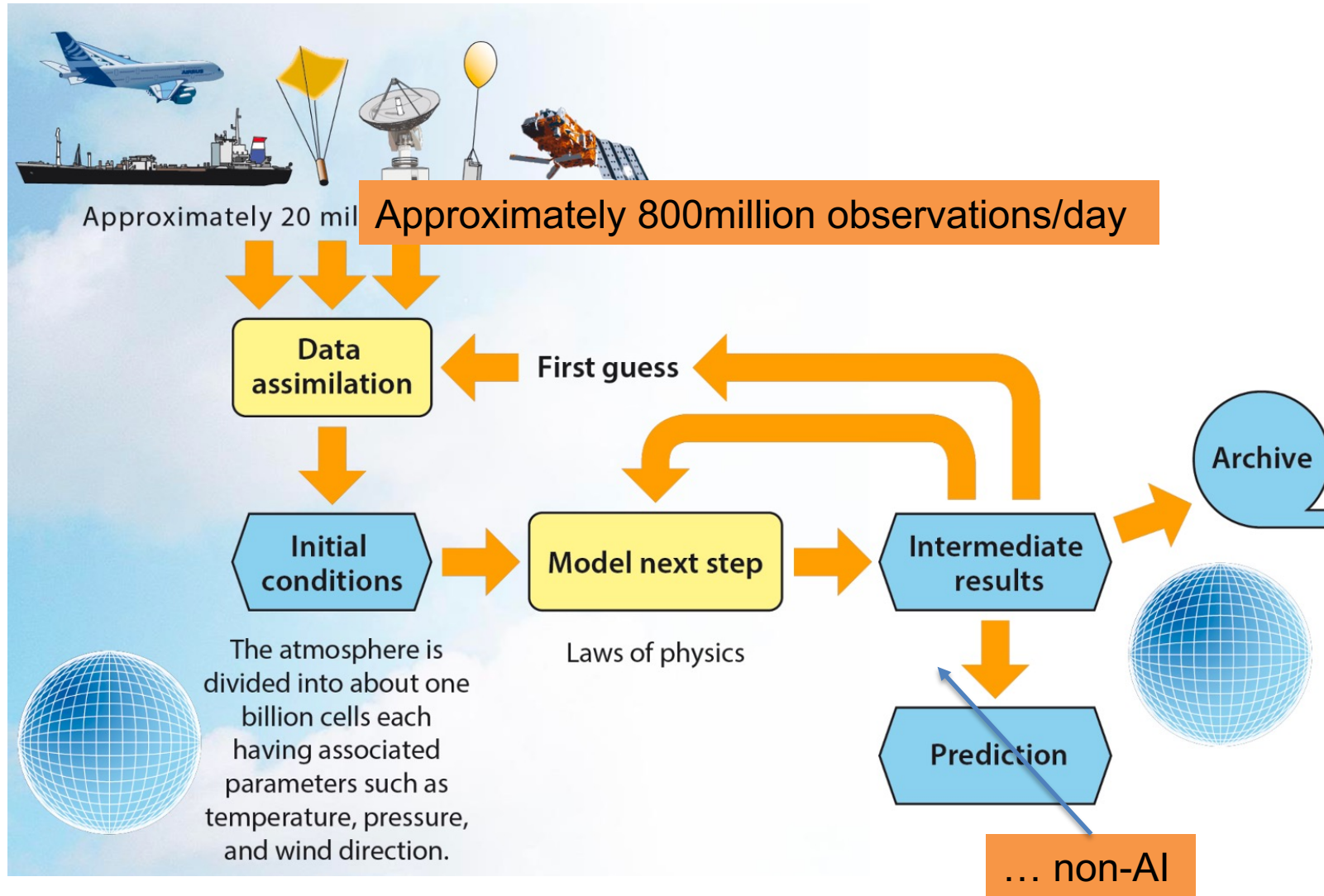
Collaboration with the EU



- *Entrusted entities of the EU initiative DestinE*
- *Develop and operate*
 - *DestinE Digital Twin Engine (DTE)*
 - *Extreme Digital Twin*
 - *Climate Adaptation Digital Twin*

- *Operating the Copernicus Climate Change (C3S) and Atmosphere Monitoring (CAMS) Services contributing EFAS and FIRE to the Copernicus Emergency Management Service,*
- *Building CAMS emission services (co2+...)*

Traditional NWP Numerical Weather Prediction Workflow

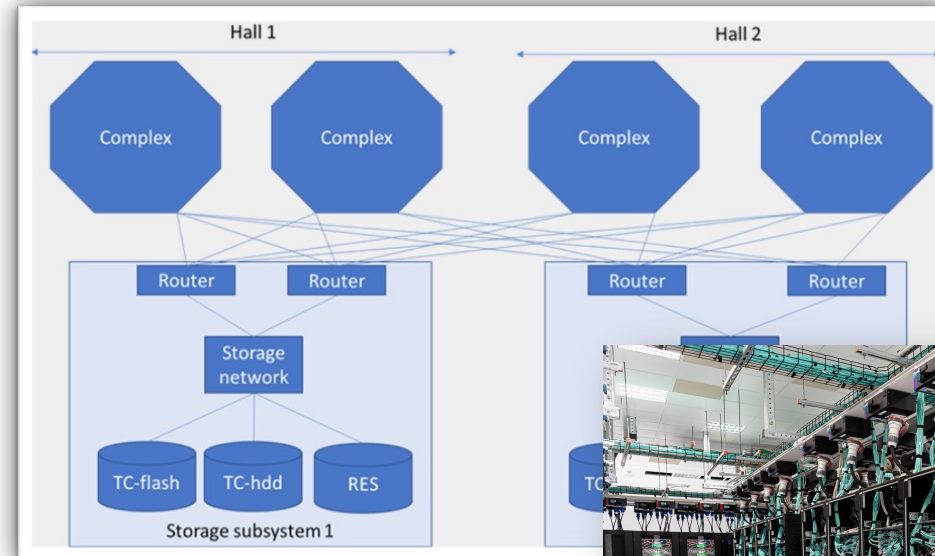


Archiving global weather data on tape for 50 years

Current HPC facility: Atos

- 4 self-sufficient compute complexes
 - each complex consists of two partitions:
 - Parallel:
 - 20 ATOS XH2000 Water cooled racks, >42t
 - 1920 nodes for parallel compute, diskless
 - Dual AMD Rome 64 core processors, 256GB
 - General Purpose/Post-Proc (“GPIL”)
 - 112 nodes for general purpose use
- Separate SLURM scheduler in each complex
- 1 GPU rack (32 blades, each with 4 nvidia A100)
- Production since 2022 service through 2027-Oct
- 2024-June top500 listings: 138 → 141
- HPC resources
 - 25% ECMWF time-critical suites (4 runs/day)
 - 25% allocated to ECMWF Member States
 - Remainder to ECMWF research department
- >>500,000 jobs/day

- >8000 2x64c AMD Rome nodes, 256GB each
- >1M cores, 2.1 PB RAM, >40PF/s peak aggreg.
- 90PB lustre, 2.8TB/s IOR aggregate (in 10 filesystems)
- Dragonfly+ HDR Infiniband w/o oversubscription,
 - 200Tb/s bisection bandwidth



Data Handling

Handled data

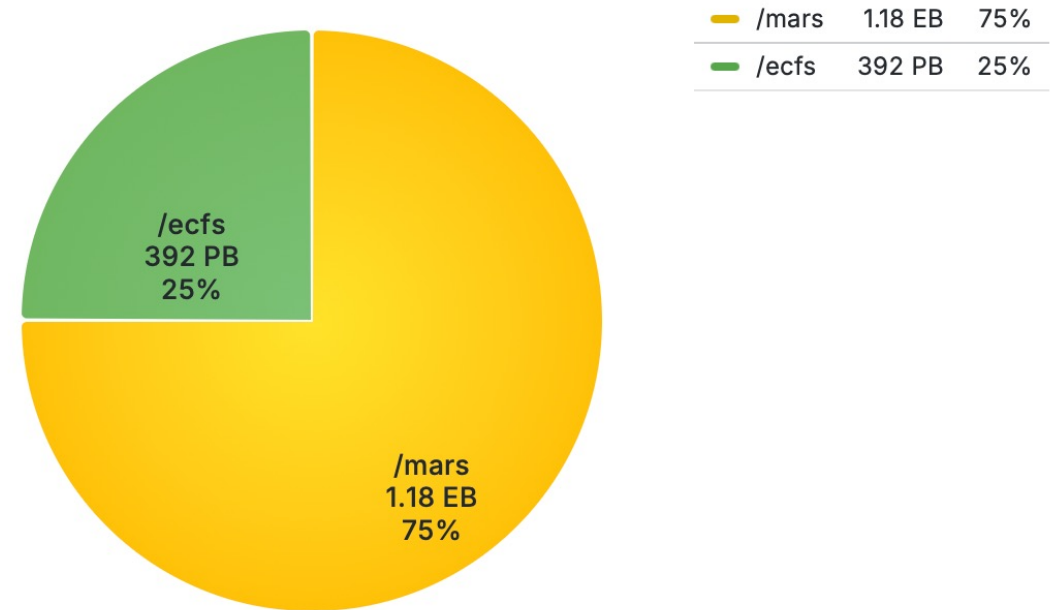
- **MARS** is ECMWF's **M**eteorological **A**rchival and **R**etrieval **S**ystem

- 75% of the volume of data
- 4% of the number of files
- **TAPE ONLY**
- No disk in HPSS, two tape layers
- Migrations are tape to tape
- Ability to handle partial recalls from tape

- ECFS is **EC**MWF's **F**ile **S**torage system

- 25% of the volume of data
- 96% of the number of files
- Disk and two tape layers
- More **traditional HSM**

MARS vs ECFS data distribution - all_data



Current data ingestion

- On an average day
 - 800 TB of data written
 - 600 TB primary copy
 - 200 TB backup copy
 - 200 TB of data read

Space used for all copies

1.46 EB

Space used for primary copy

1.087 EB

Space used for backup copy

371.14 PB

Avg data added per day - all copies

815.08 TB

Avg data added per day - primary copy

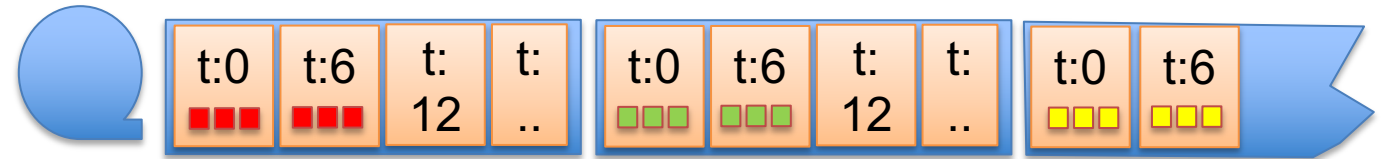
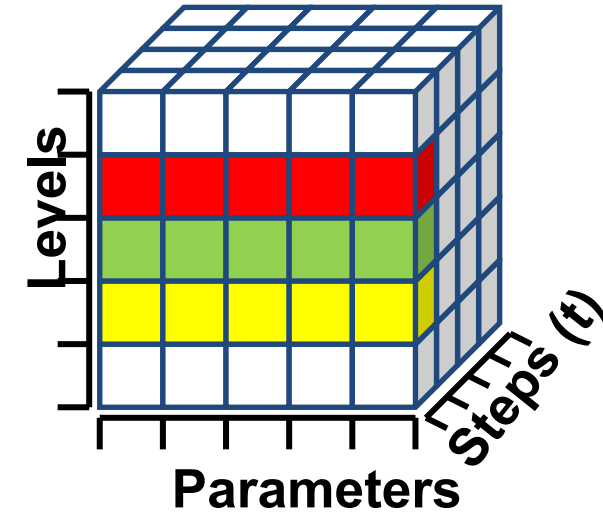
615.40 TB

Avg data added per day - backup copy

199.69 TB

MARS (1/2)

- MARS store for GRIB objects
 - Store geographic grid **fields** (long/lat)
 - Level (ie: 800 hPa, 500 hPa)
 - Parameters (physical variable, ie: T° / %RH / wave level)
 - Step (time step of the simulation)
- Grouping:
 - Logical grouping over time (step)
 - Multi TB files
 - Tape segregation per dataset (families)
- Recall:
 - **Only partial** recalls (just the requested **fields**)
 - Never the full file
 - Sequence of seek/read/seek on tape.



MARS is a global meteorological data archive and retrieval system, which relies largely on **TAPES**

MARS (2/2)

- A Tape Only storage ???
 - Original Idea : Use disk cache capacity on MARS rather for the HSM.
 - Avoid cache bottleneck on HSM
 - 1 Write Stream == 1 tape drives
 - Write large objects to expect maximum drive speed.
 - Grow horizontal (more drives mean more bandwidth)
- Operational drawback
 - Request need to be optimized and scheduled in order to minimize tape mount.
 - MARS has a straight control on HPSS
 - 2nd copies (backup) need 2 tape drives for migration
 - Complex HPSS configuration
 - 800 of tape family
 - Drives and Tapes are **heavily** used !

Storage Infrastructure & HPSS

Storage infrastructure figures

Disk Cache Storage	Capacity/Tape
IBM Storage Array Systems	~50 PB
CEPH clusters	~39 PB

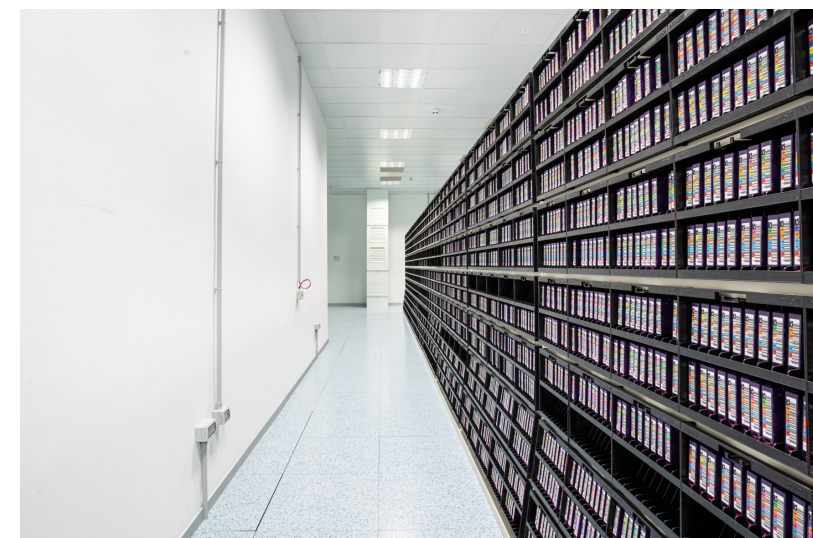
Libraries Primary Copy	#Drives 1160	#Drives 1170
11 x IBM TS4500 Enterprise	336	352
1x Spectra – Tinfinity Enterprise	60	30

Tapes Primary Copy	Data stored	#Tapes
JE1160	1.03 EB	44200
JF1170	54.5 PB	1220
Totals	1.09 EB	45420

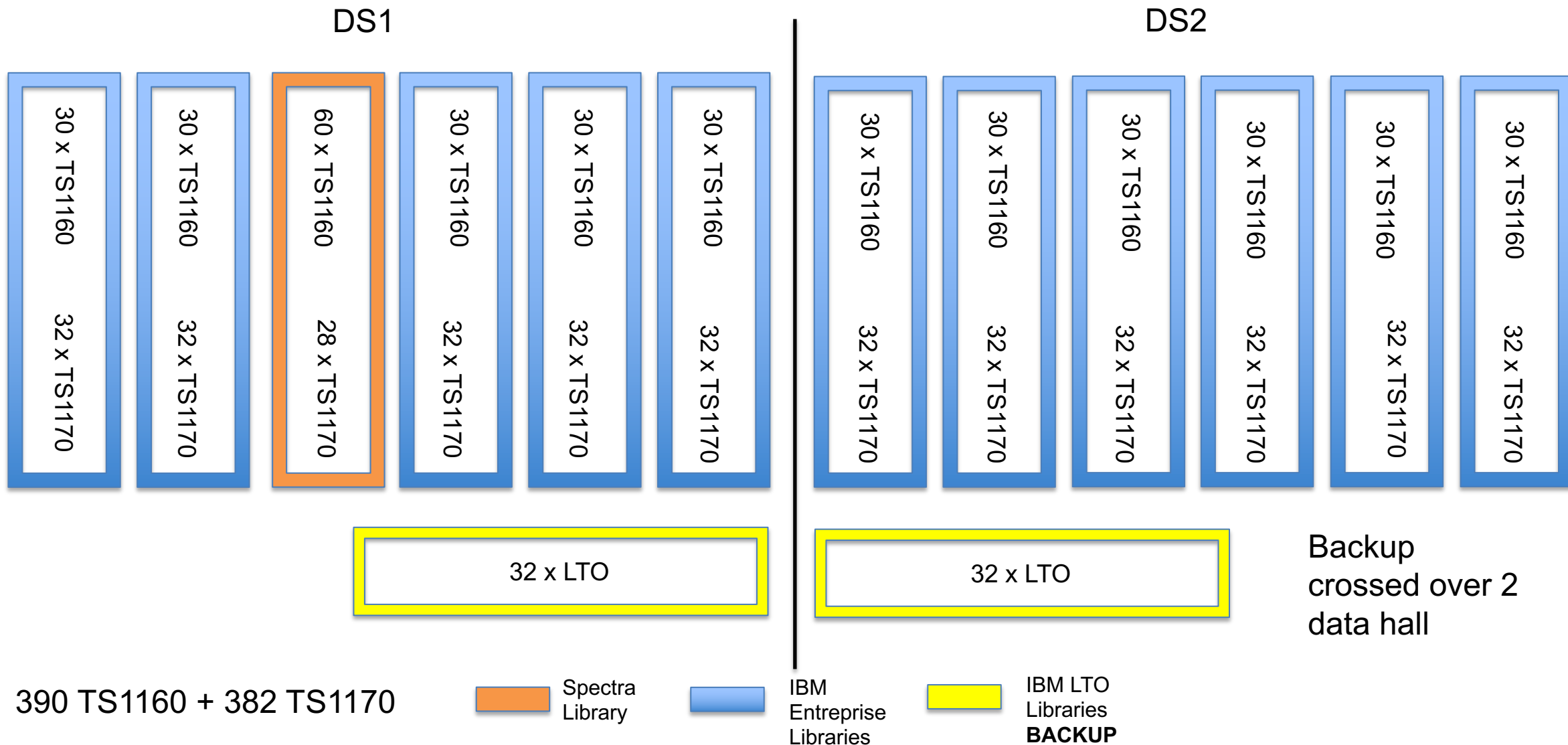
Libraries Secondary Copy	#LTO
2 x IBM TS4500 LTO	64

Tapes Secondary Copy	Data stored	#Tapes
LTO-7(M8)	61 PB	6420
LTO-8	199 PB	13000
LTO-9	115 PB	5270
Totals	371 PB	24690

Tapes Secondary Copy shelved	#Tapes
LTO-(M8/8)	15280



Logical view



HPSS environment

- 3 instances
 - Prod, preprod, test
 - 1 failover instance for prod
 - RHEL 7.9
- Version 10.3 in production
 - Core server with 1 TB RAM, partitioned database
 - Upgrade core to RHEL 9.6+ (for DB2 12)
- Production instance managing:
 - 14 PVR
 - 136 tape mover
 - 15 disk movers
 - 14 active Storage Classes / 54 COS
 - Hundreds of families (> 800)



Video Monitoring

ZoneMinder

Console Options Log Groups Filters Cycle Montage Montage Review Audit Events Report

High

Load: 22.30126953125 DB: 89/151 Storage /tmp: 57%

administrator RUNNING v1.36.12

cam15f10 2026-06-10 08:56:19 D

cam15f9 2026/06/10 08:56:20

State: - fps

cam15f8 2026/06/10 08:56:19

cam15f7 2026/06/10 08:56:19

Video monitoring

- Each library is equipped with 5 webcam
 - Inhouse solution
 - POE Powered
 - Poe Switch on top of rack
- Zoneminder
 - Installed in BSD Jail
 - Run on a TrueNas server
- Use full for troubleshooting

RS PRO Network Indoor Power Over Ethernet CCTV Camera 1920 x 1080 pixel

RS Stock No.: 146-4649 | Brand: [RS PRO](#)



JF Tapes go Early End Of Life

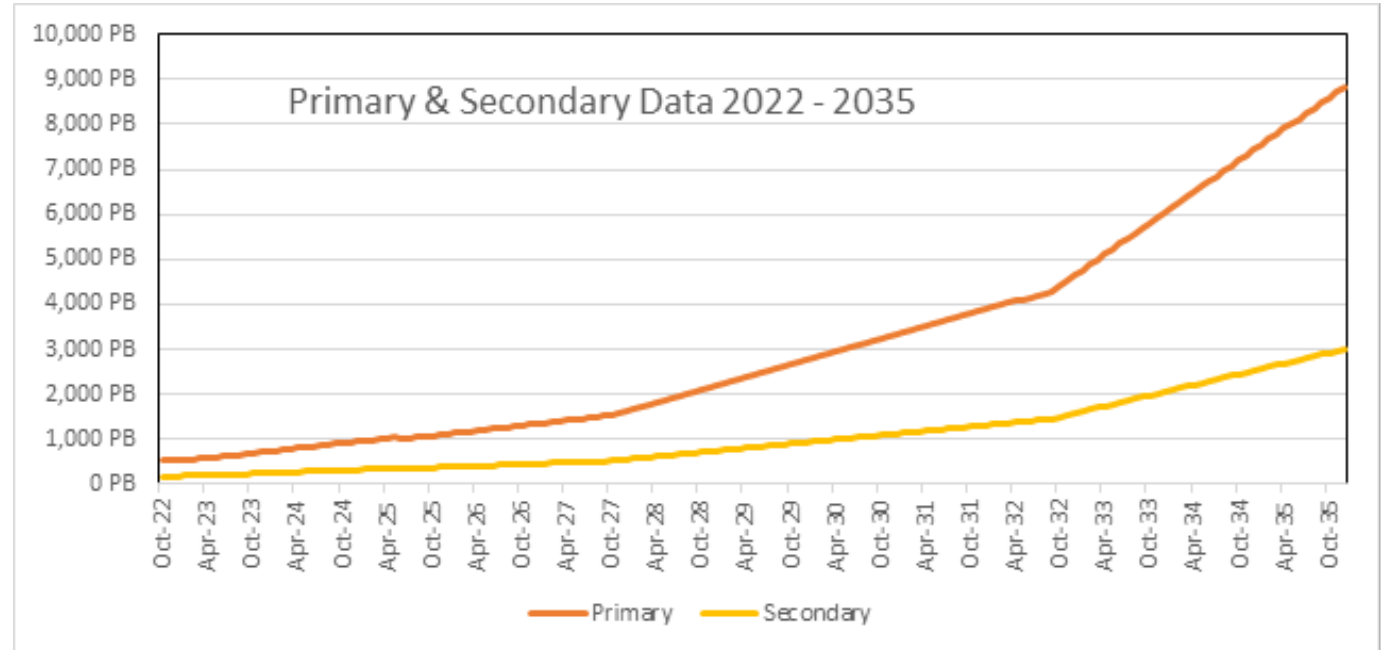
- Using TS1170 technology since June 2025
 - Installed in TS4500
 - 47 PB written so far
- Many I/O errors detected early 2026 :
 - Libraries reported remaining lifetime to 0
 - Only after few hundreds mounts
 - **After less than 3 months in used in HPSS**
 - **Motion meters Lifetime JF << JE !**

Media	# Broken tape	Motion mètres	Remainig life time	AVG mount count	Days in HPSS
Jaguar E	9	28 M	0 %	2,900	342 d
Jaguar F	3	16 M	0 %	101	73 d

Future Challenge

Future challenges

- Data demanding weather forecasts
 - Avg daily growth of ~1.8x starting from 2027
- Technology evolution
 - Enterprise vs LTO
- New libraries technology
 - Linear vs compact modular silos
- Secondary copy colocation
 - To enforce the DR strategy



Future challenges

- Jaguar decommissioning start in 2030-32
 - What about our Enterprise libraries ?
- We need REALIABLE tapes
 - Enterprise LTO would be the only option
- Great expectation for the LTO11
 - Capacity !
 - Reliability !

Thank you!

Questions?



ECMWF site report
HPSS-FR 2026 – CC-IN2P3

Enrico Fattibene
Pierre-Emmanuel Brinette
Storage and Servers Team