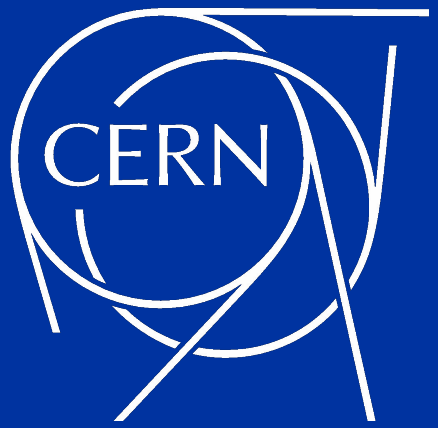


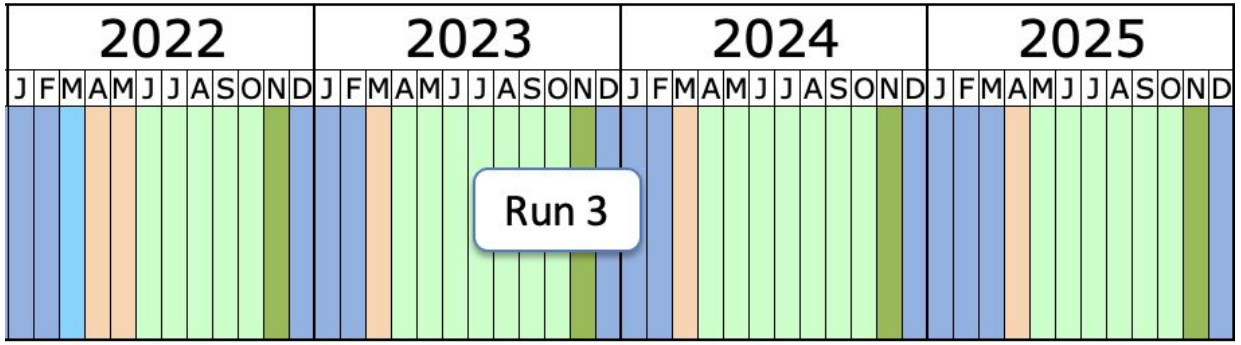
Storage for the LHC

Operations during Run 3

Presented by Cedric Caffy on behalf of the CERN IT-SD group

HEPiX Autumn 2022
03/11/2022

Introduction



- **CERN Run 3 started in July 2022**
  - Experiments generate data (a lot!)
- **Different workflows put in place to suit each experiment needs in term of storage**
  - Export data from the experiments' storage buffer to T0 data center
  - Long-term archival of the data
  - Data export from T0 to T1, T2…
- **Different CERN-made softwares used**

EOS

CERN Tape Archive

FTS File Transfer Service

Outline

I. **Run 3 experiments expectations**
   A. Storage
   B. Throughput

II. **The software stack and infrastructure**
   A. EOS
   B. CTA
   C. FTS

III. **Experiments workflows**
   A. General overview
   B. The ALICE(O2) setup

IV. **EOS File replication - Erasure coding**
   A. File replication
   B. Erasure coding
   C. Status at CERN

V. **From the beginning of RUN 3**
   A. EOS transferred data
   B. CTA transferred data
   C. FTS transferred data

Run 3 experiments expected storage

	Disk storage 2022	Disk storage 2023	Disk storage 2024	Tape storage per year
ALICE	50 PB	58.5 PB	67.5 PB	At least 150 PB
ATLAS EXPERIMENT	32 PB	40 PB	46 PB	
CMS	35 PB	45 PB	52 PB	
LHCb	26.5 PB	30.3 PB	46.8 PB	

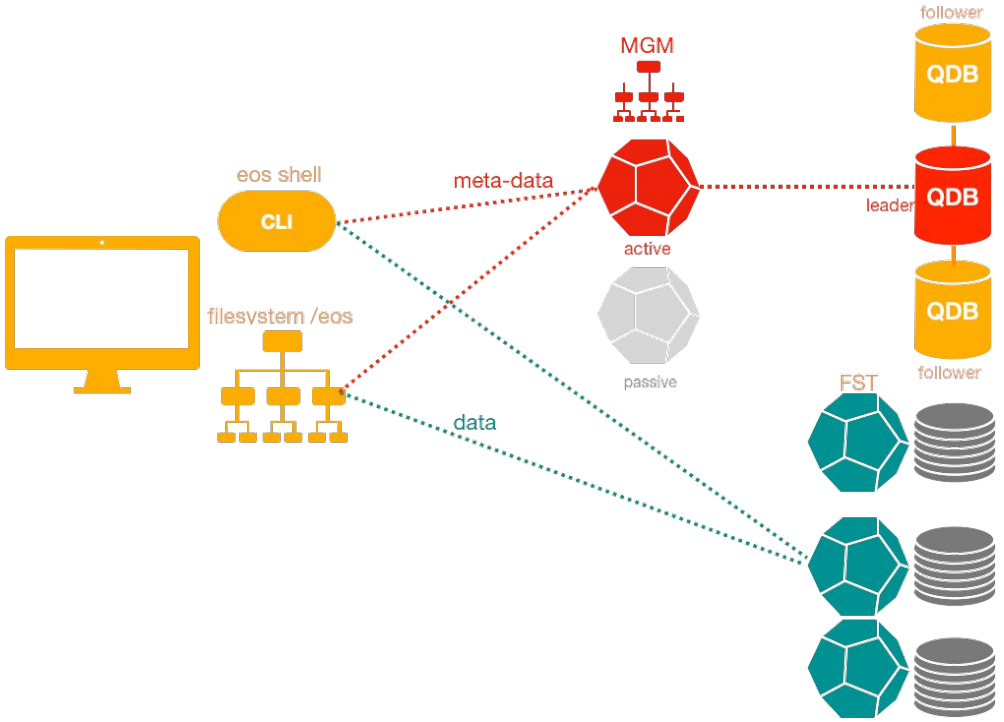
Run 3 experiments expected throughput

	Experiments pits to T0 disks	Disks to T0 tapes
ALICE	100 GB/s (ALICE) + 150 GB/s (ALICEO2)	10 GB/s
ATLAS EXPERIMENT	10 GB/s	10 GB/s
CMS	20 GB/s	10 GB/s
LHCb	10 GB/s	10 GB/s

The software stack and infrastructure

The software stack and infrastructure

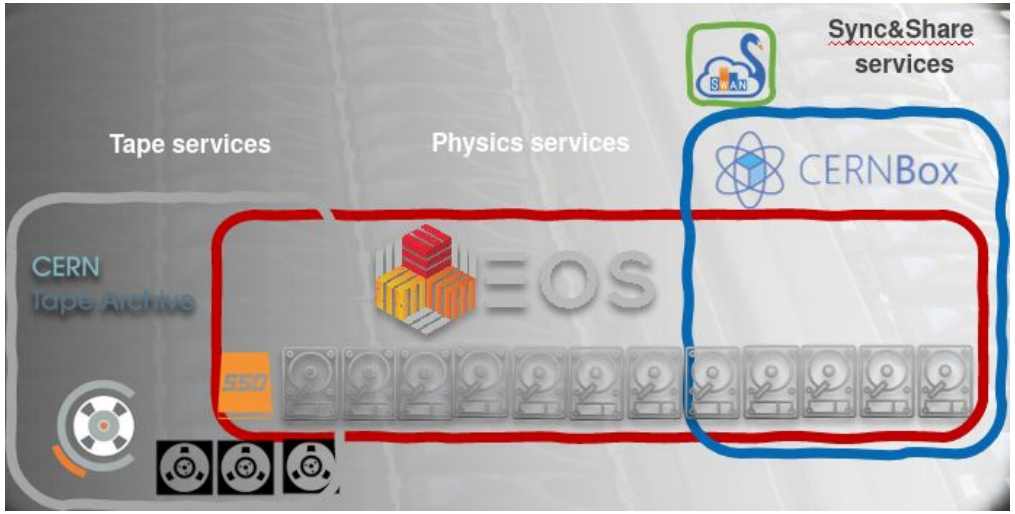
- **EOS (EOS Open Storage)**
  - Distributed disk-based storage system
  - Highly available and low latency namespace
    - Namespace persisted on a distributed key-value store
    - Working entries cached in-memory
  - Highly available and reliable file storage
    - Based on (cheap) JBODs
      - File replication across independent nodes and disks

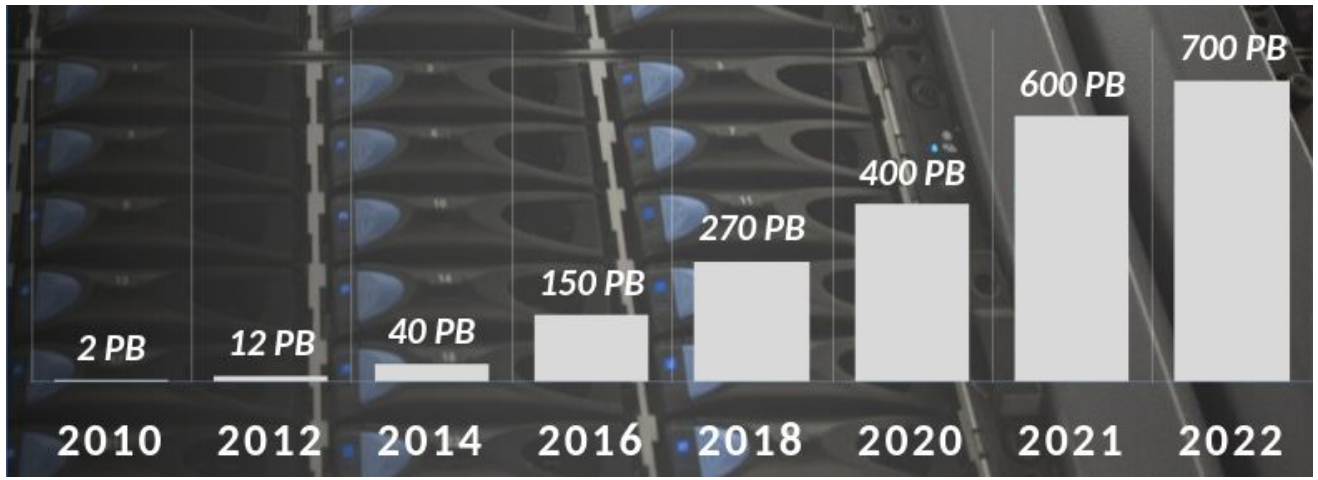


The software stack and infrastructure

- **EOS by numbers**

Total Space	700 PB
Files Stored	~7.4 Bil
# Storage Nodes	~1000
# Disks	~80000



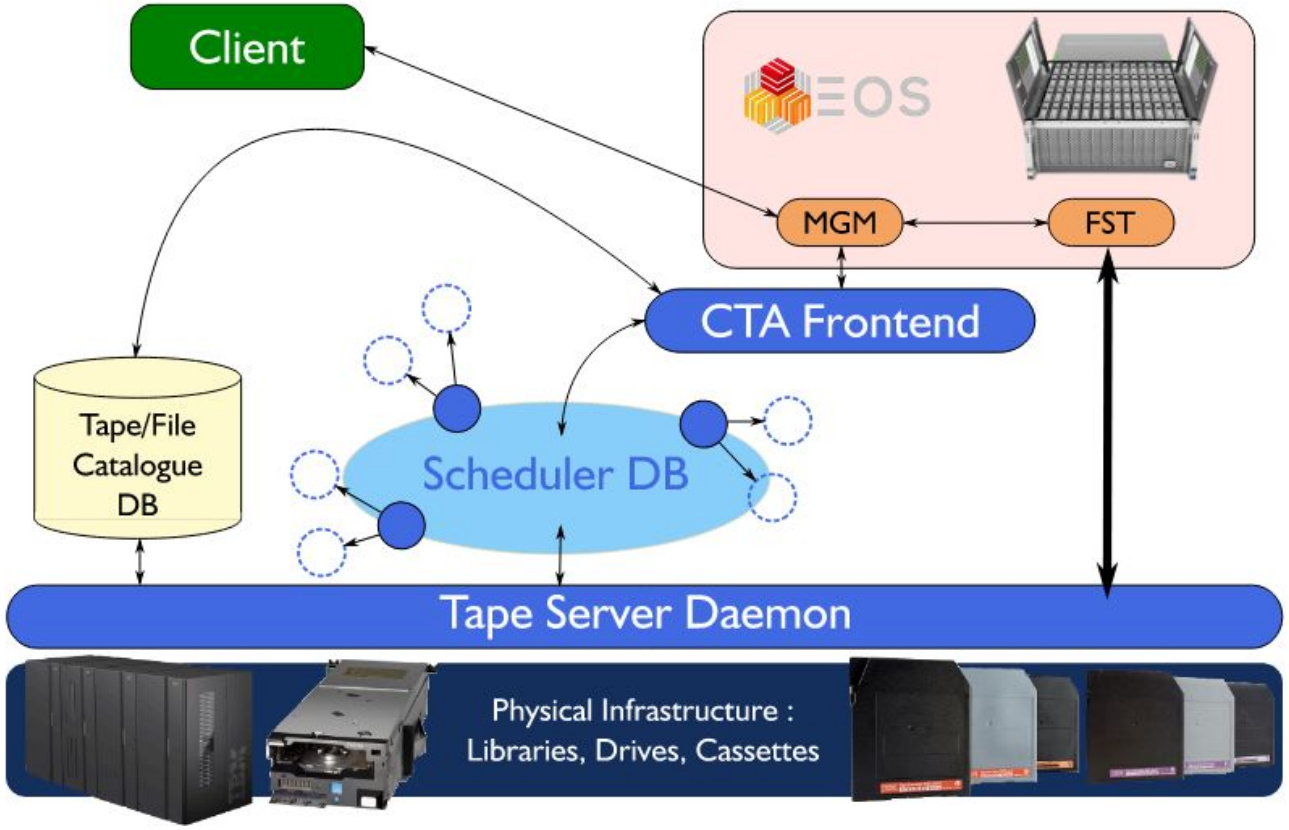


The software stack and infrastructure

CERN Tape Archive

- **CTA (CERN Tape Archive)**

  - High performance tape archival storage system
  - Tape backend to EOS
    - Provides file operation and disk pool
  - Mount scheduling logic and tape operations



The software stack and infrastructure

CERN Tape Archive

- **CTA by numbers**

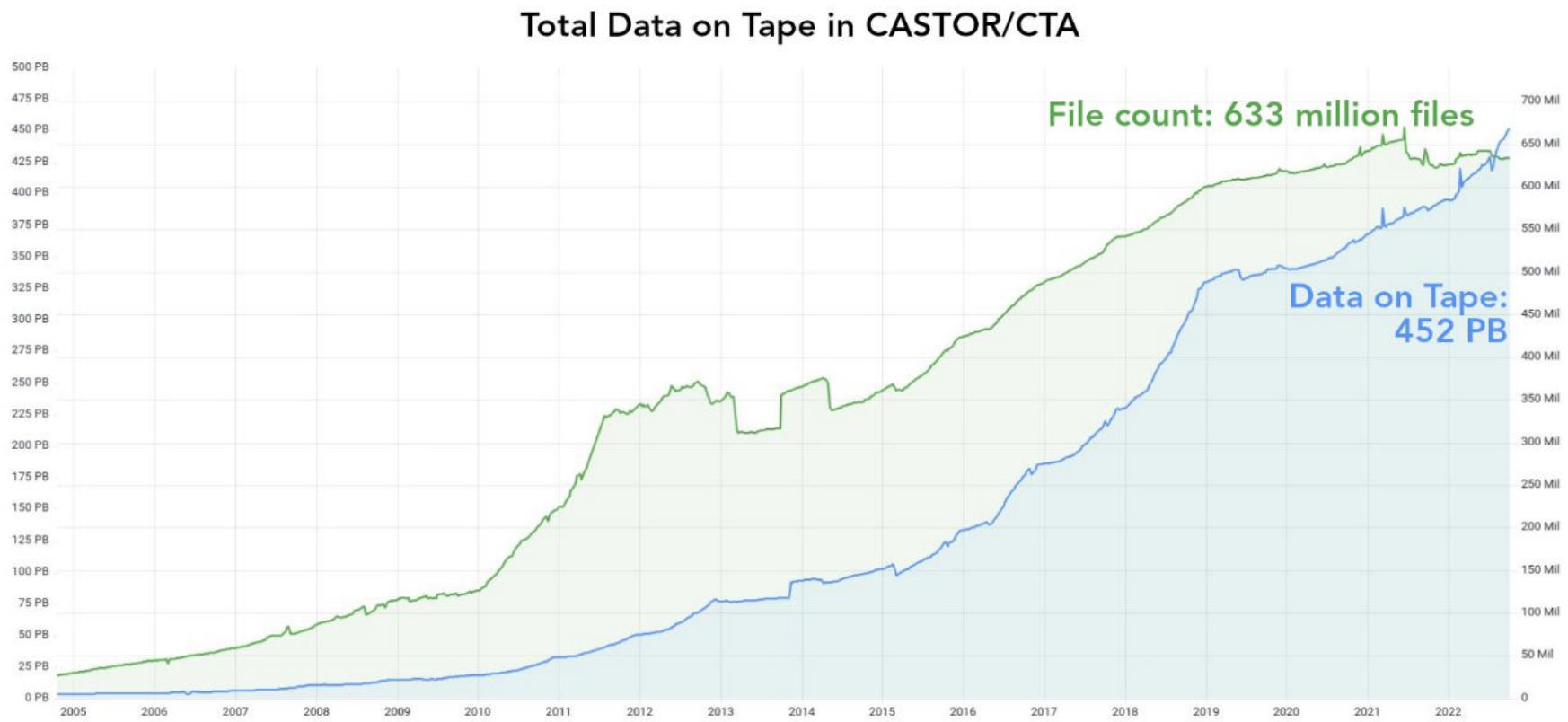
Shared infrastructure

	Enterprise	Non-enterprise	Total
Tape libraries	3x IBM TS4500	2x Spectra Logic TFinity	5
Drives	10x IBM TS1155 76x IBM TS1160	10x LTO-8 98x LTO-9	194
Media (tapes)	34 PB on 3592JC 227 PB on 3592JD 84 PB on 3592JE	62 PB on LTO–7M 29 PB on LTO–8 83 PB on LTO–9	519 PB

The software stack and infrastructure

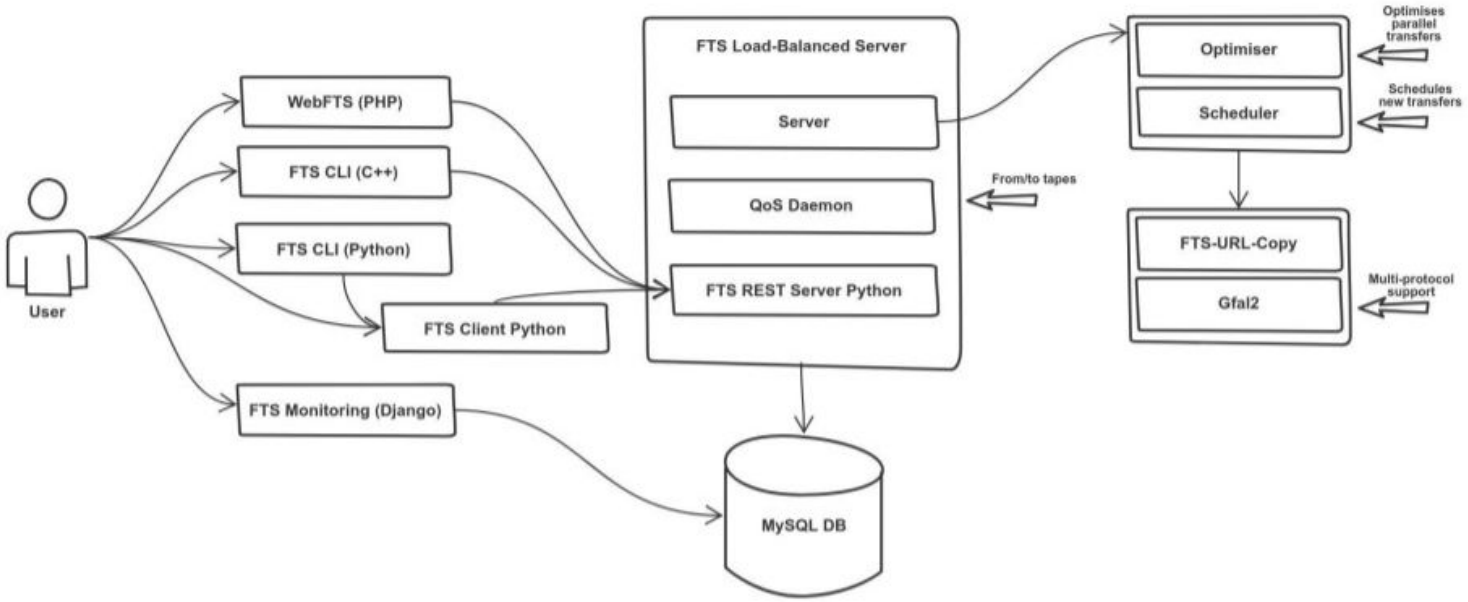
CERN Tape Archive

- **CTA by numbers**



The software stack

- **FTS (File Transfer Service)**
  - Offers reliable and large-scale data transfers functionalities
    - Orchestrates data transfers from different storage endpoints throughout the entire WLCG

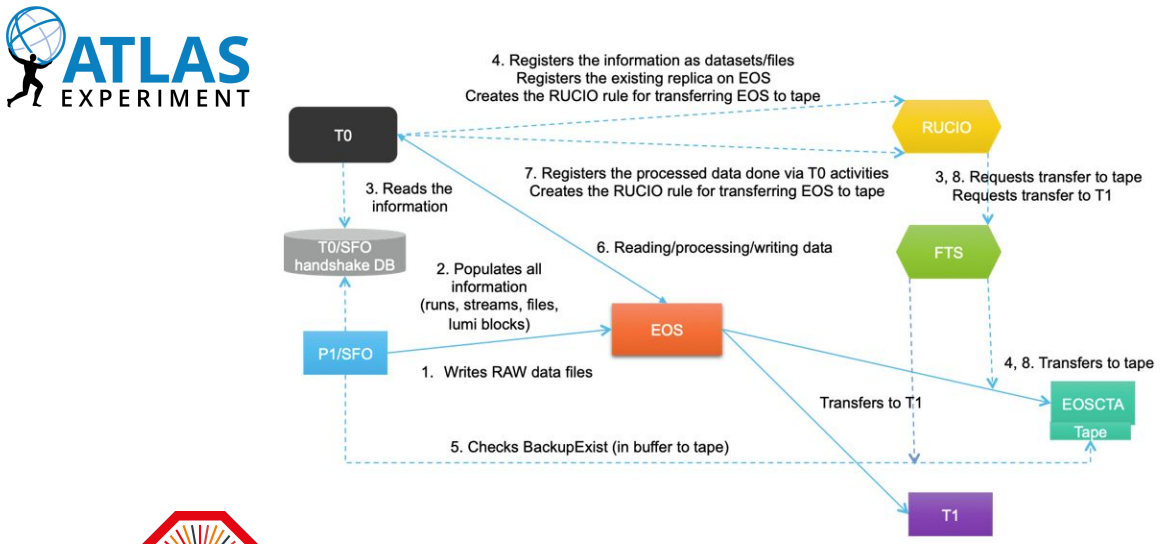


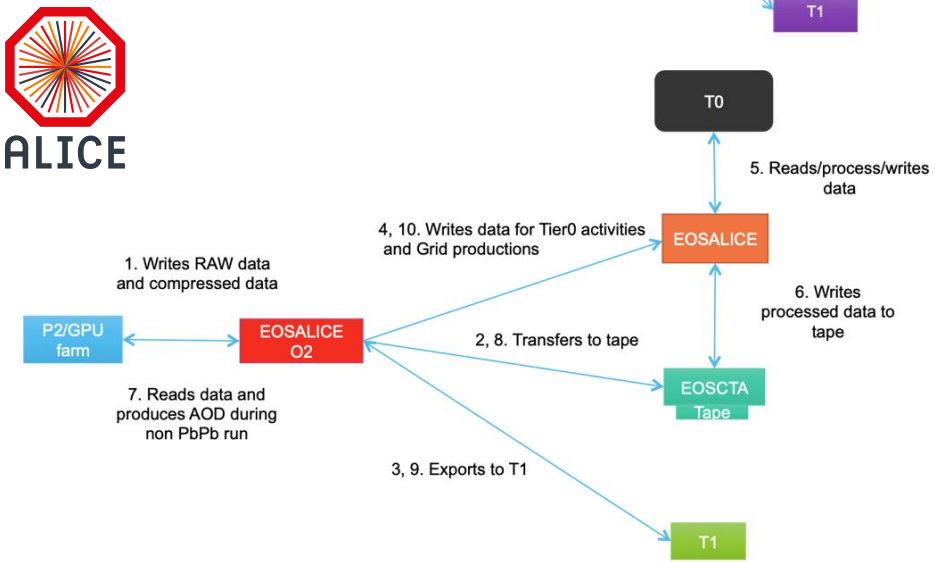
The software stack

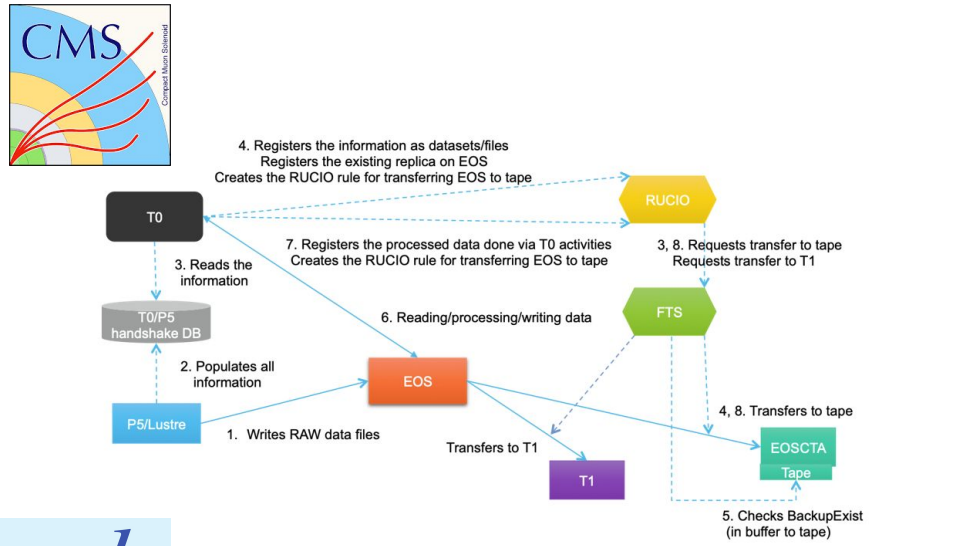
- **FTS by numbers**

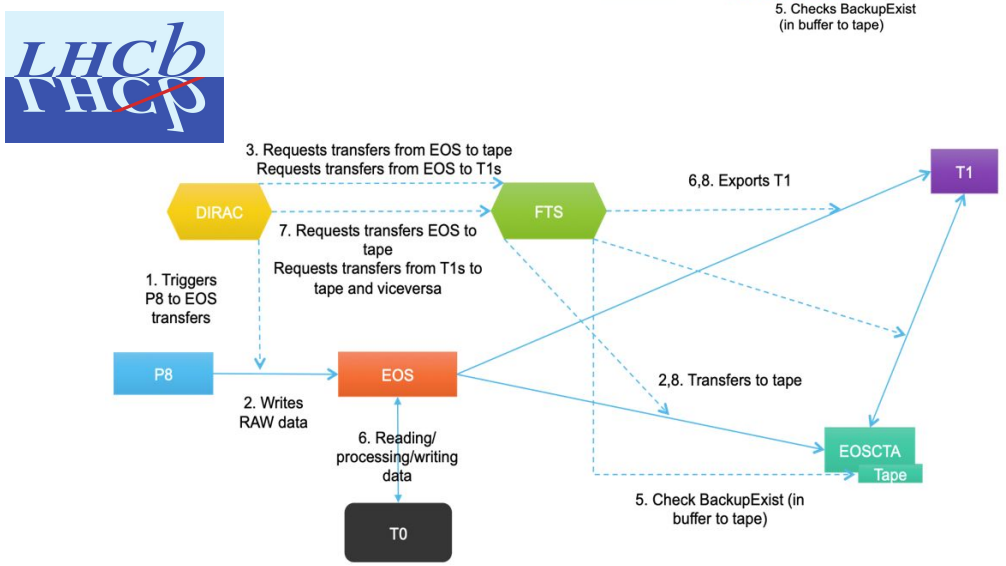
	# Servers
ATLAS EXPERIMENT	10
CMS	10
LHCb	5
Non-LHC experiments	4
Total	29

Experiments workflows - General overview

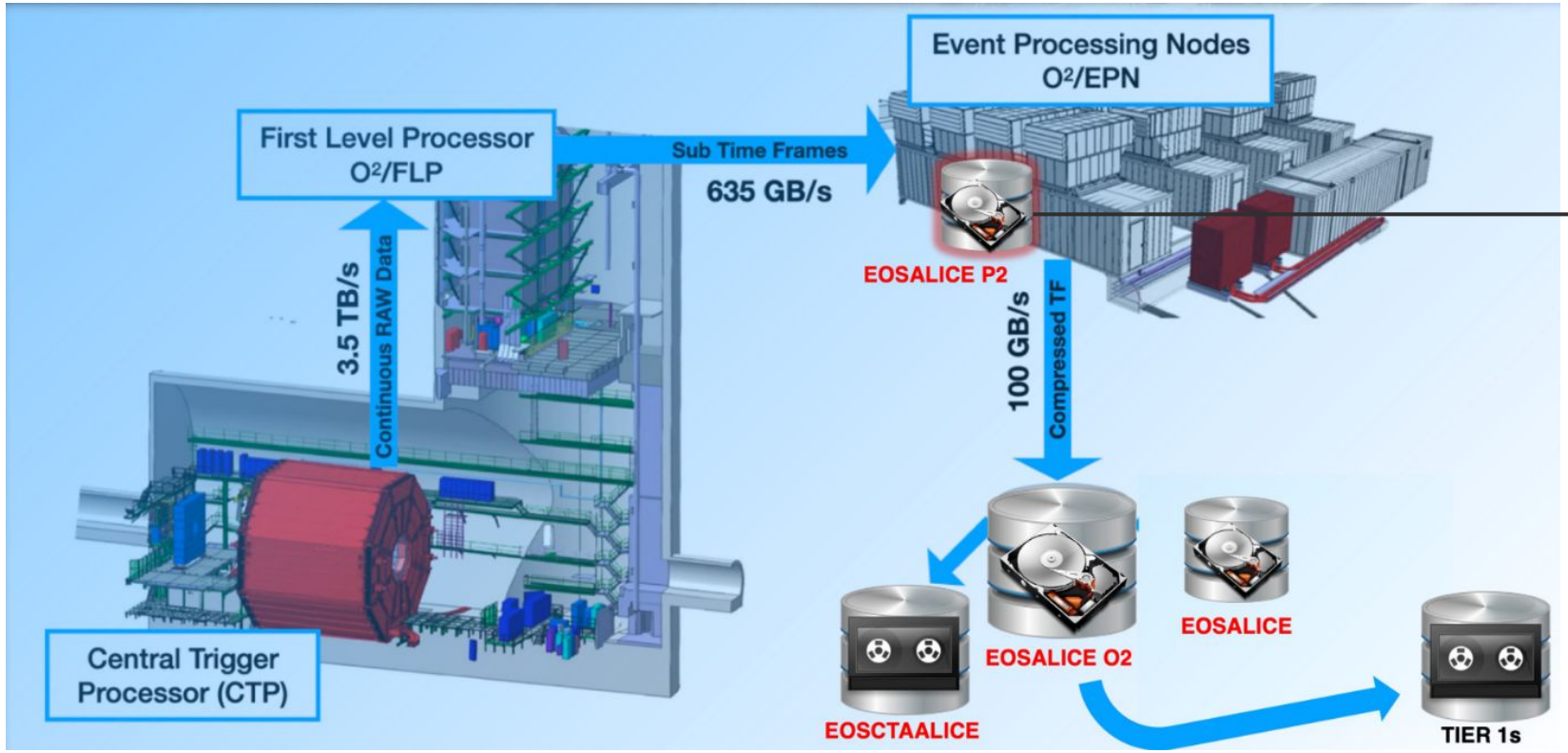






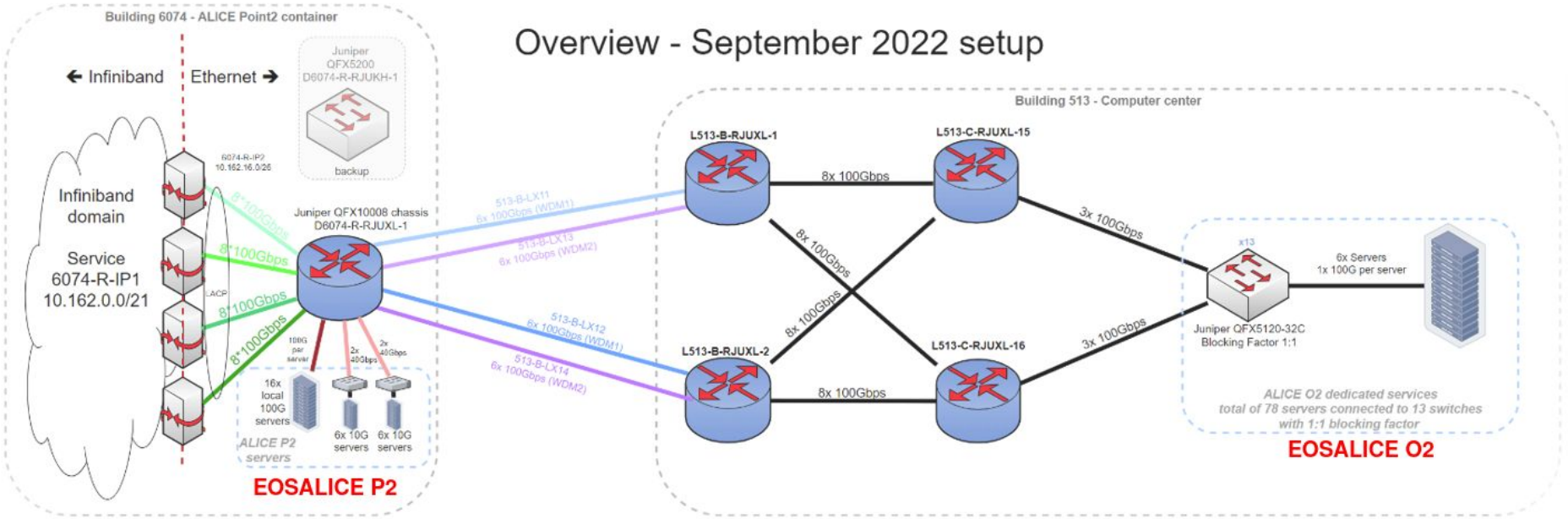


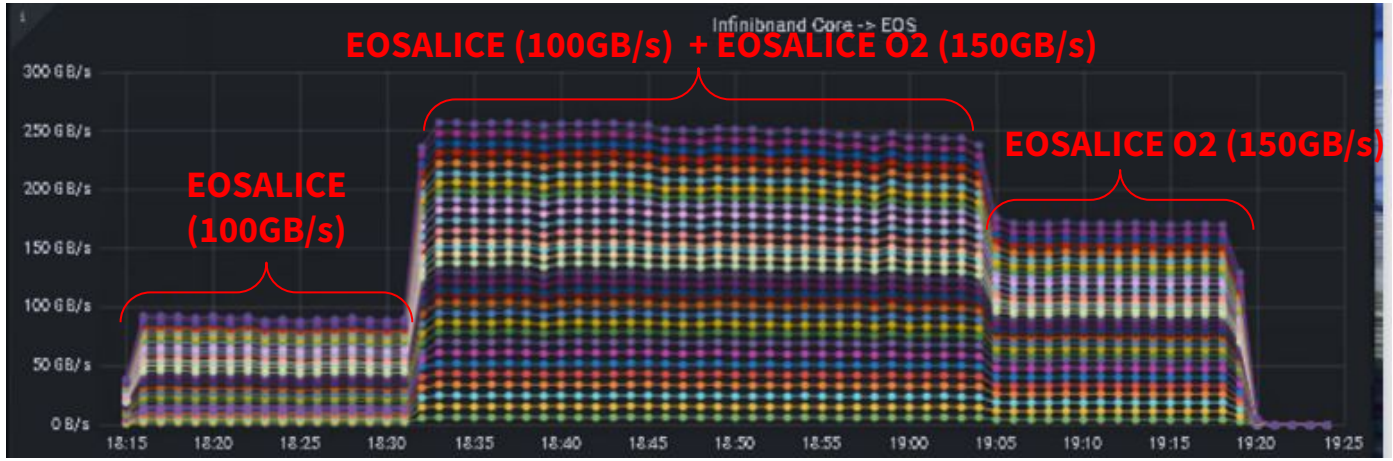
ALICE(O2) setup



- "Backup" in case of link failure between EPNs and CERN Data Center
- Sustains 100GB/s
- 13.5 PB → 18.5h buffer

ALICE(O2) setup



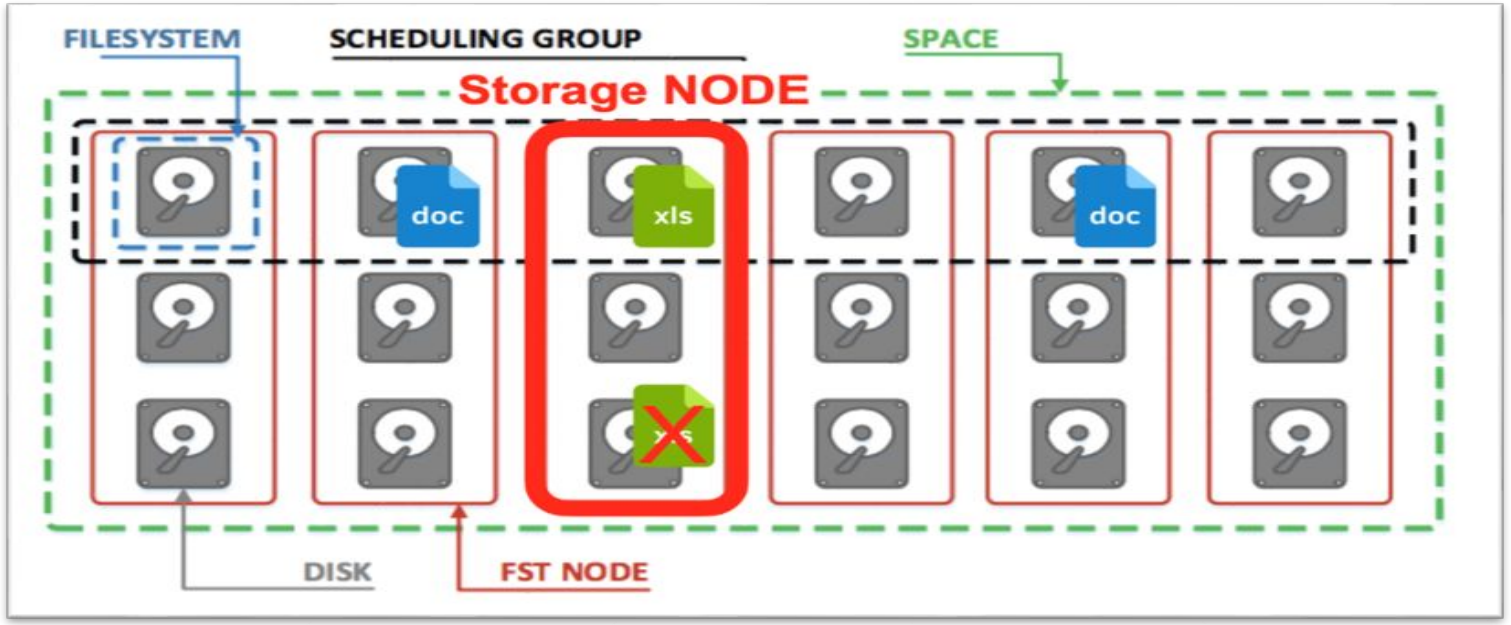


EOS File replication

- **File availability usually insured by replication**
  - RAID vs RAIN
  - Files are replicated n-times on different disks on different machines
  - Protect against disk failure and storage node failure

File stored as RAID

File stored as RAIN

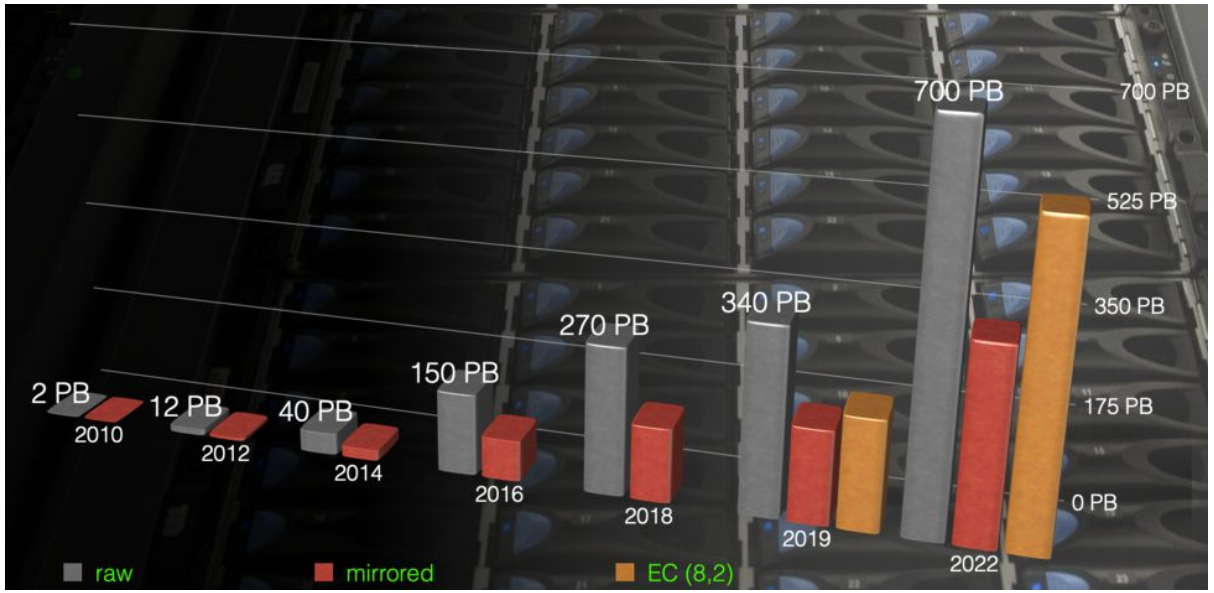


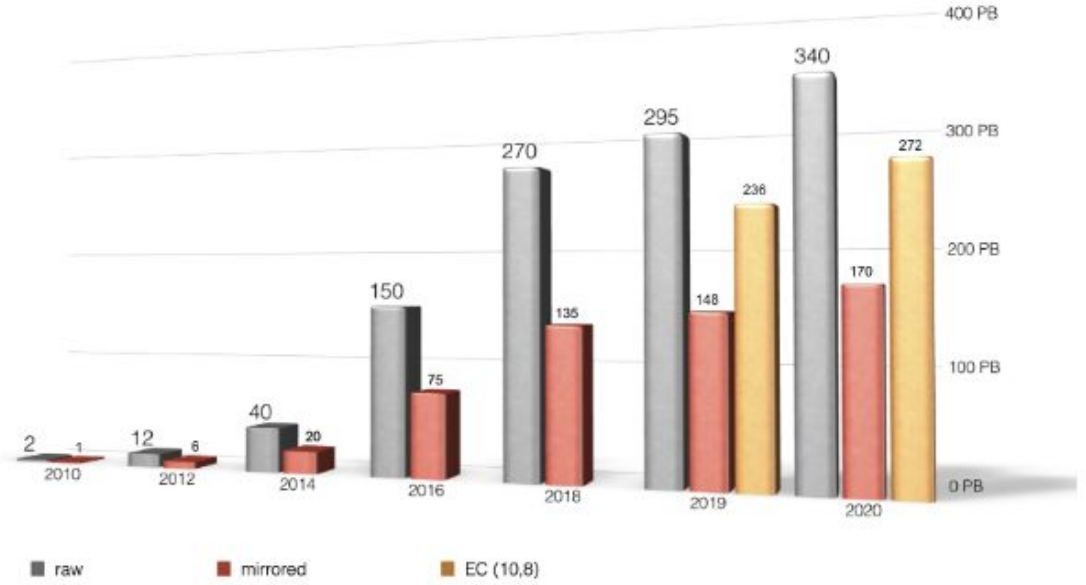
EOS File replication

Do you want to store more data?

Erasure coding

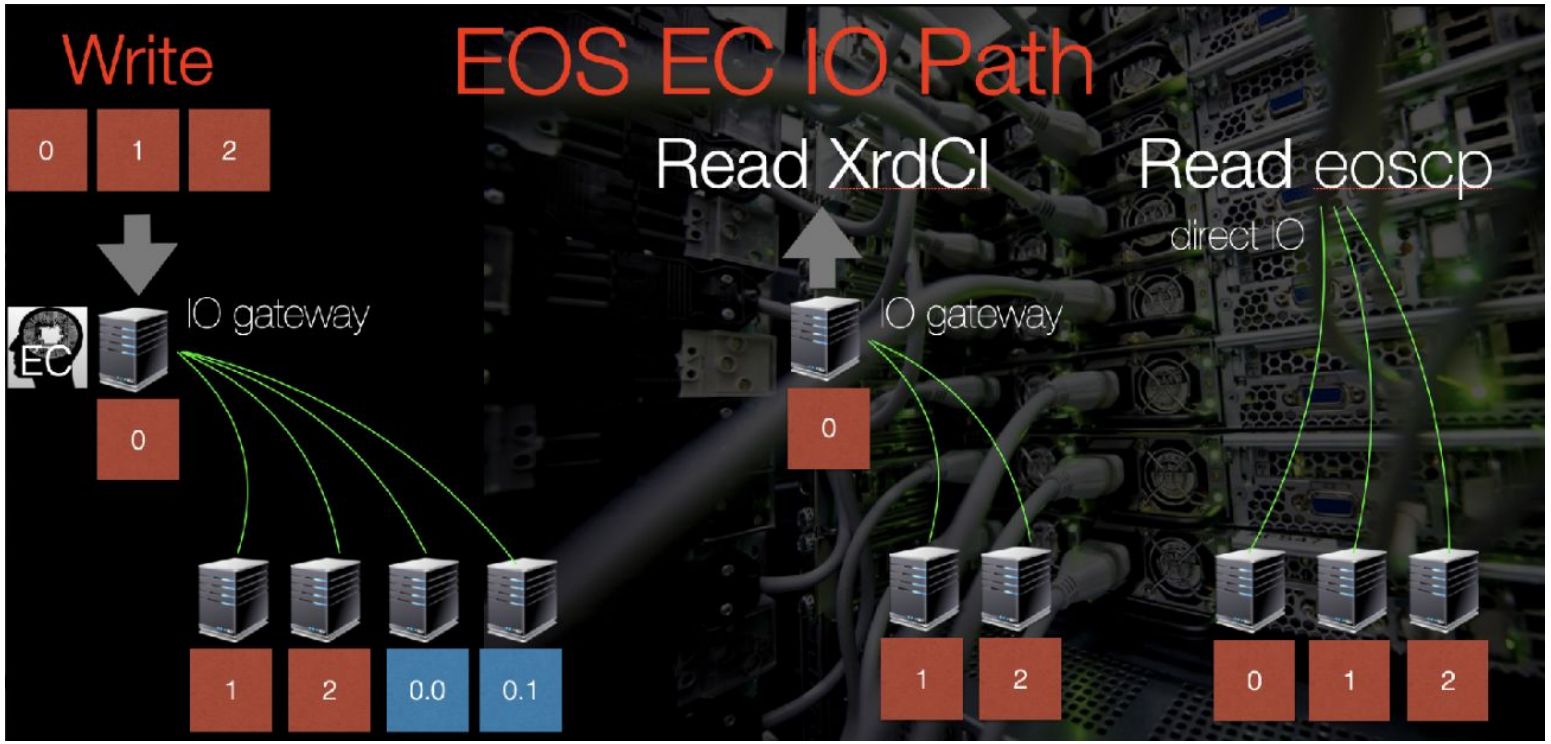
- **With erasure coding**
  - Instead of storing n-times the data…
  - …split it into different locations and add parity blocks to protect it





Erasure coding

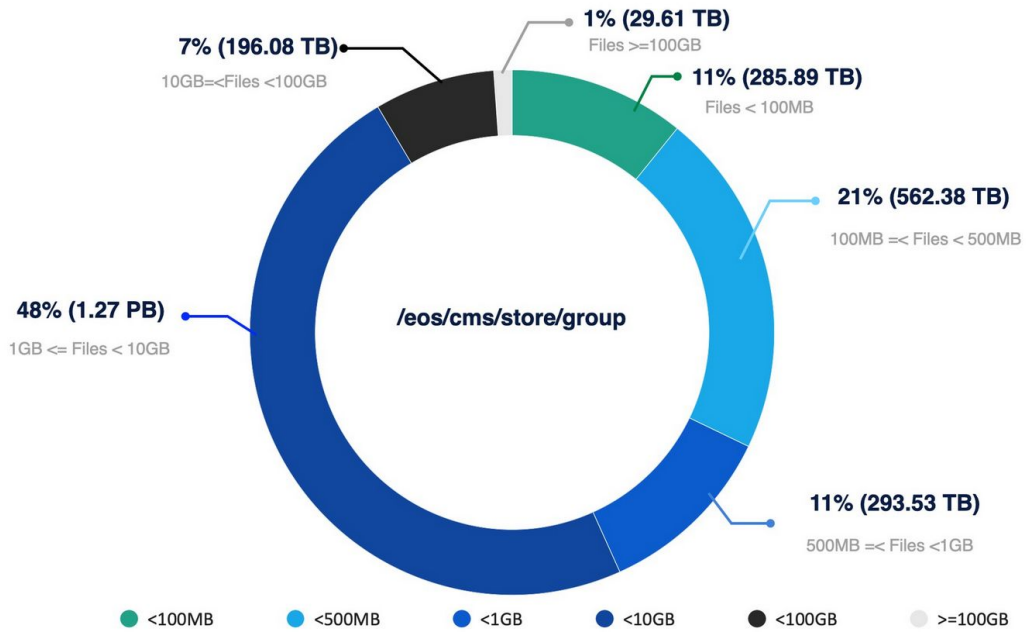
Ex: EC(3 + 2) with 1MB block size



- **Best case scenarios**
  - Writing 2x faster than replication
    - Almost same network usage
  - Reading more than 2x faster than replication
    - Twice as much network usage as replication (GW model)…
    - … Except if eoscp is used

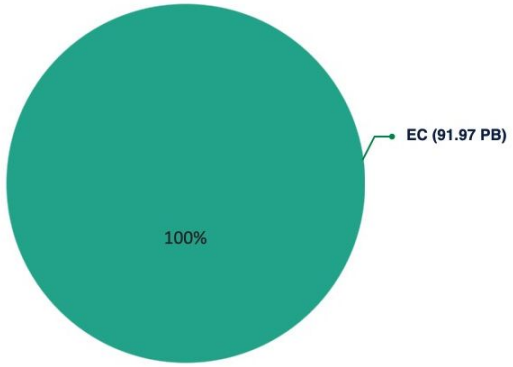
Erasure coding - Status at CERN

- Files data protection is usually insured by creating a second replica of every files
  - Need twice as much raw storage that is logically needed…
- Fully deployed on EOSALICE O2
- Some physics groups of EOSCMS are erasure coded
  - Files bigger than 100MB are converted

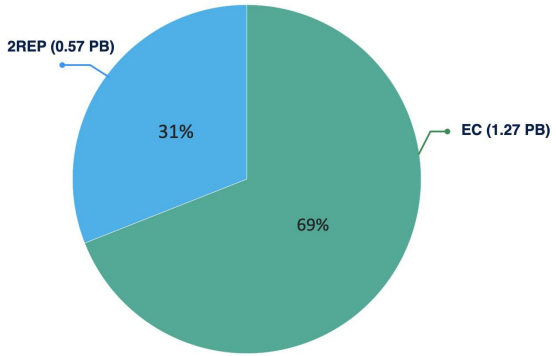


Erasure coding - Status at CERN

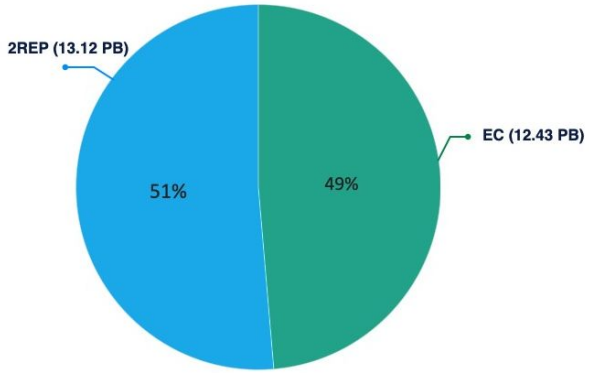
	EC layout	Logical bytes stored	Raw storage used	Raw storage gain (compared to 2-rep layout)
EOSALICE O2	EC(10+2)	75.82 PB	91.97 PB	59.67 PB
Some EOS CMS Physics groups	EC(10+2)	1.05 PB	1.27 PB	0.83 PB
EOSAMS (non-LHC experiment)	EC(8+2)	9.95 PB	12.43 PB	7.47 PB



EOSALICE O2



EOSCMS Physics groups



EOSAMS

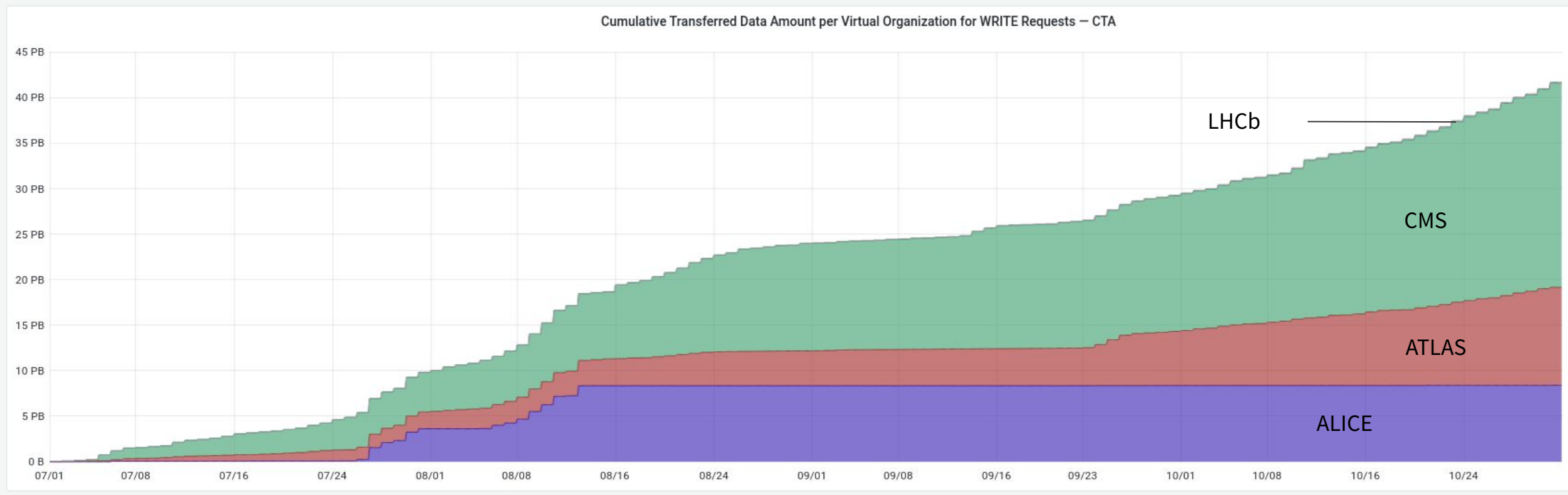
From the beginning of Run 3

LHC experiments transfers from and to EOS since July 2022

	READ		WRITE	
	Total amount of files read	Total amount of bytes read	Total amount of files written	Total amount of bytes written
ALICE	2.26 Bil	253 PB	110 Mil	73.6 PB
ATLAS EXPERIMENT	972 Mil	253 PB	105 Mil	40.8 PB
CMS	649 Mil	216 PB	251 Mil	50.9 PB
LHCb	123 Mil	40.8 PB	77.5 Mil	24.2 PB
TOTAL	4.00 Bil	762 PB	544 Mil	190 PB

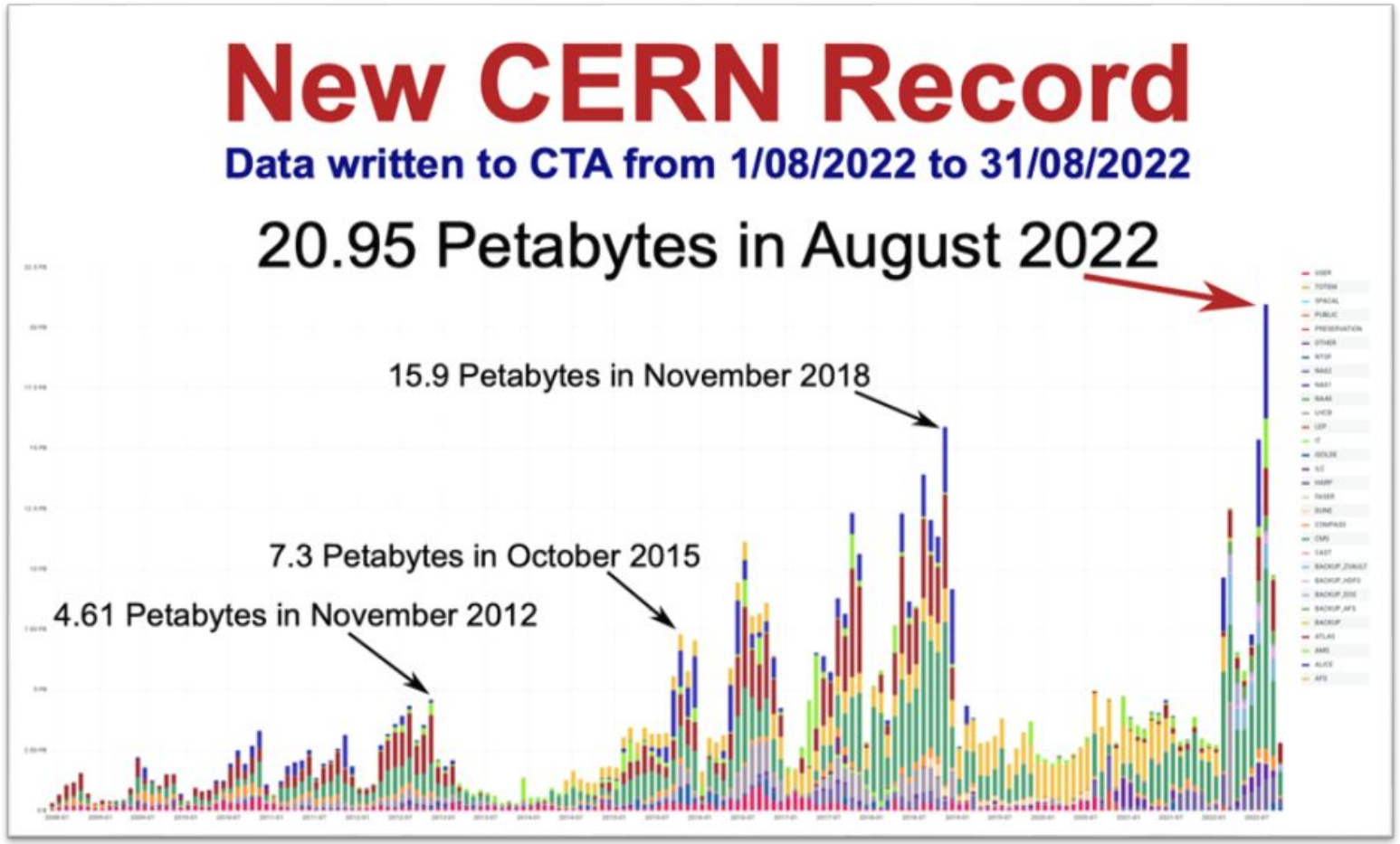
From the beginning of Run 3

CTA transfers - Archived data



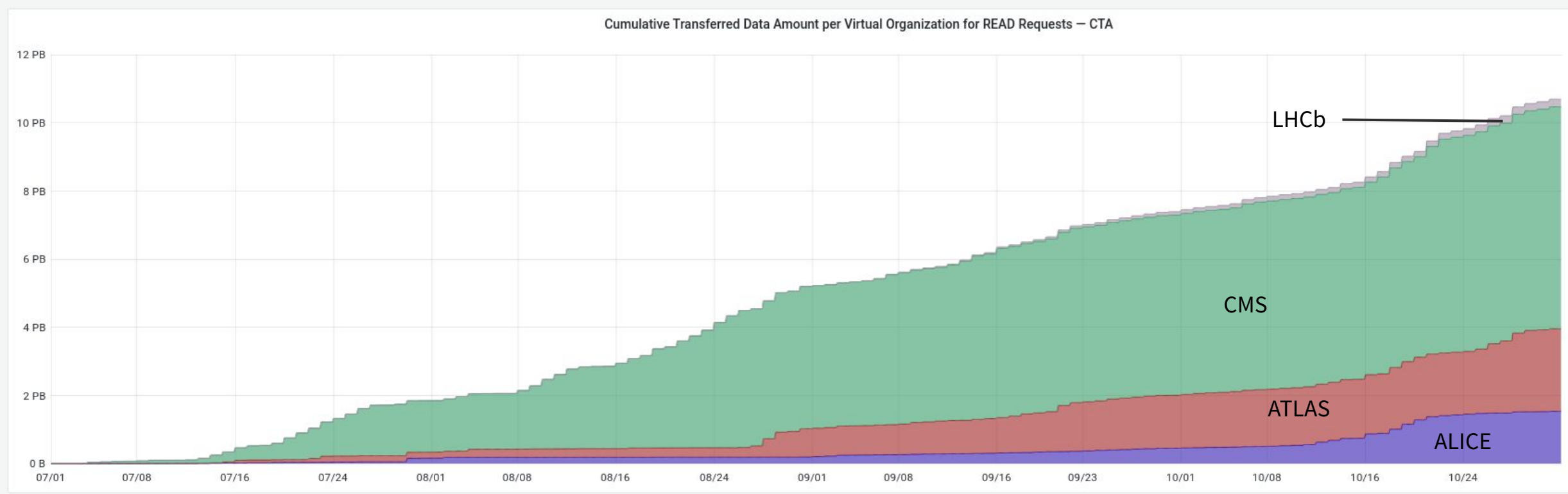
From the beginning of Run 3

CTA transfers - new record!



From the beginning of Run 3

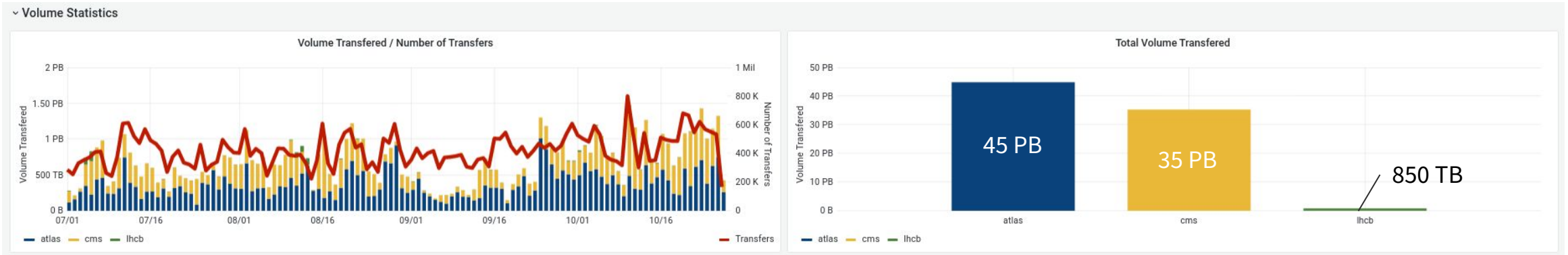
CTA transfers - Recalled data



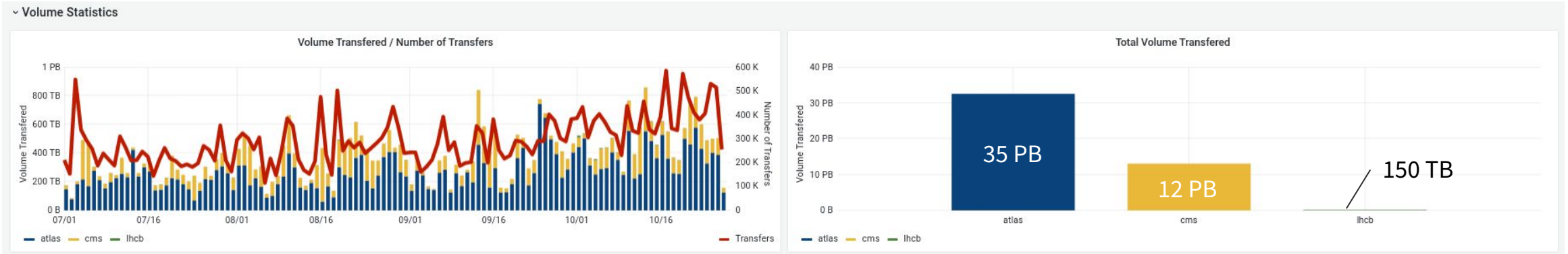
From the beginning of Run 3

FTS transfers

Out of CERN to anywhere in the world

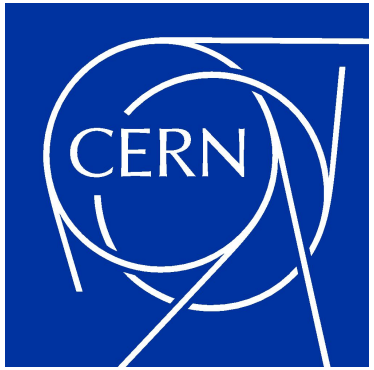


From anywhere in the world to CERN



Conclusion

- **LHC experiments data flows are managed by 3 CERN-made softwares**
  - EOS provides disk storage
    - Large disk infrastructure (700 PB, ~80k disks and 1k storage nodes)
  - CTA provides tape storage
    - ~520 PB, 194 tape drives and 5 tape libraries
  - FTS orchestrates data transfers between different WLCG sites
- **Slowly introducing erasure coding in some EOS instances**
  - Allows to save raw space
- **Since beginning of Run 3**
  - ~760 PB exported from EOS and ~190 PB written to it
  - 42 PB of data archived and 10.5PB recalled
  - Ready to take the challenge of next year ;-)



home.cern