University POLITEHNICA of Bucharest

Faculty of Automatic Control and Computers

ALICE

CERN

EPN2EOS Data Transfer System

Computing in High Energy & Nuclear Physics - May 2023

Author

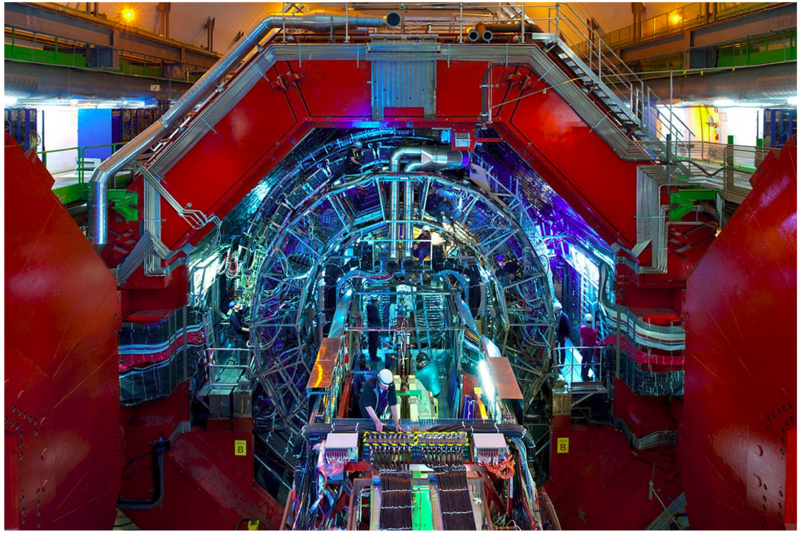
Alice-Florenţa Suiu
asuiu@cern.ch

Scientific Advisor(s)

Latchezar Betev
Costin Grigoras

On behalf of the ALICE Collaboration

Context

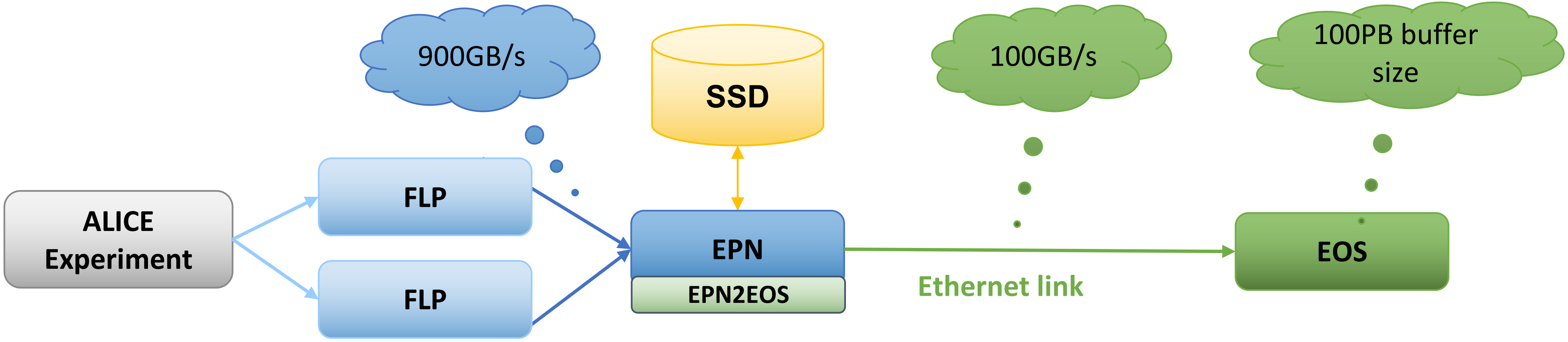


- EPN2EOS manages the transfer of files from volatile storage to persistent storage

- A Large Ion Collider Experiment - **ALICE** - a heavy-ion detector at the CERN LHC
  - Data rate to secondary storage: ~120GB/s

- Dedicated farm for online calibration and compression
  - Requires fast and secure system for transfer from experimental area to CERN IT storage

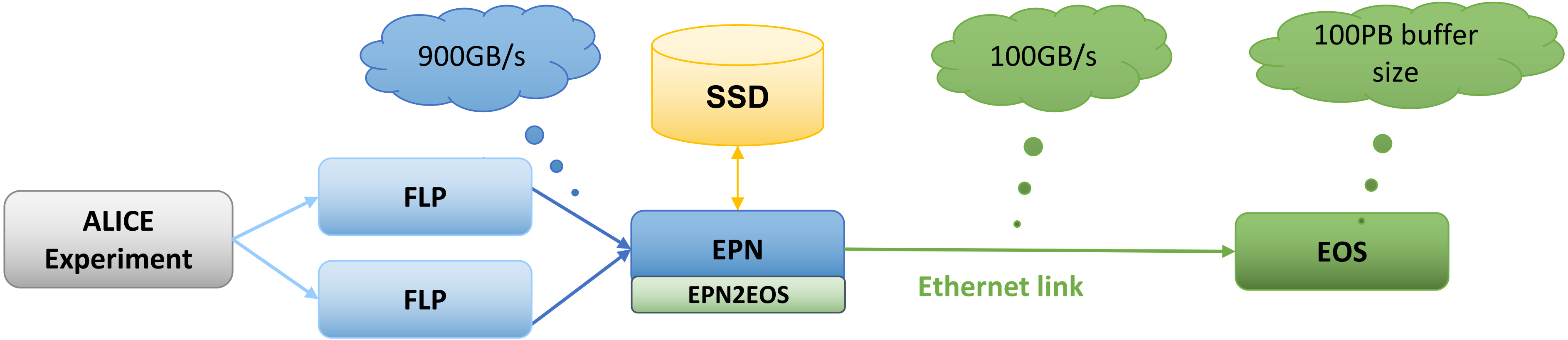
EPN2EOS in the Data Transfer Path



- **FLP** - First Level Processor
- **EPN** - Event Processing Node
- **EPN2EOS** runs inside each EPN node and shares resources with it

- ***Optical links*** between ALICE and FLPs
- ***Infiniband*** between FLPs and EPNs
- **SATA link** between EPN and SSD
- ALICE O2 software framework and GPU usage

Objectives



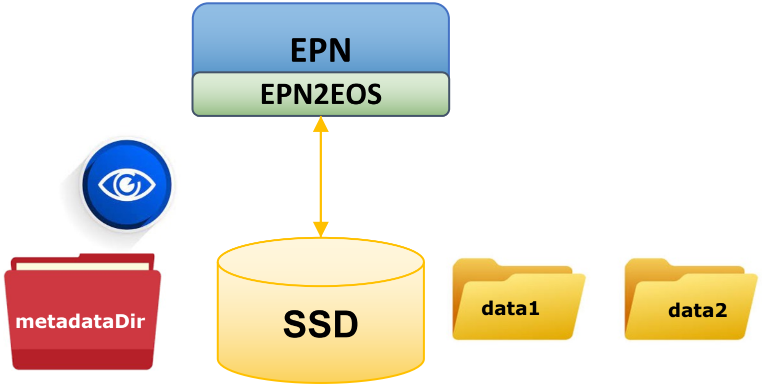
- 250 EPN nodes, each equipped with one 4TB SSD
- SSD buffer capacity sufficient for ~3h of data taking
- EPNs produce ~2GB data files with frequency of 0.2Hz

- These must be transferred to EOS promptly and removed from the nodes
- EPN2EOS has to use as few resources as possible
- EPN2EOS has to ensure that the data was not corrupted during the transfer

EPN2EOS Basic Functionality

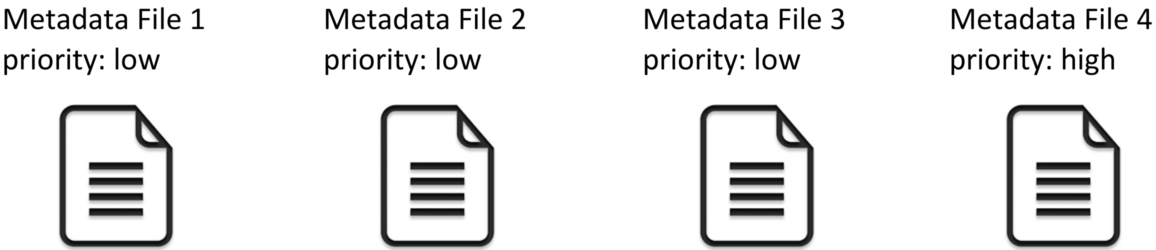
- On each EPN node maintains a queue of files to be transferred
- A metadata file is associated with every data file
- The metadata is used to steer the transfers and includes the following fields
    - **Size** — size of the data file
    - **Type** — raw, calib or other
    - ***Priority*** — low or high
    - **Spath** — local path to the data file
    - **Dpath** — path to a directory in EOS
    - **Persistent** — number of days that the data is available on storage, default is forever

Data Transfer Structure

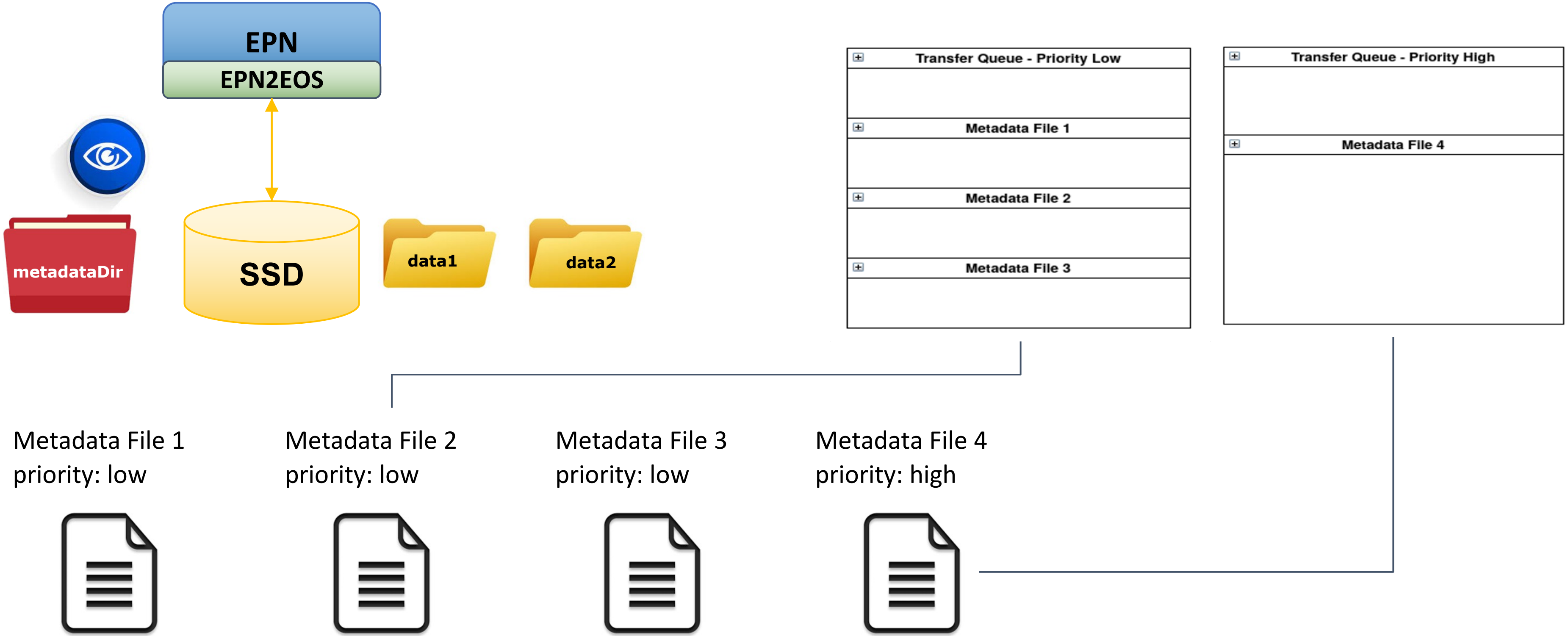


- All the metadata files are placed in the **metadataDir**

- A **watcher** is created on the **metadataDir**, a process that follows the activity of the directory and notifies the **EPN2EOS** when a new metadata file is added to it



Data Transfer Structure



EPN

EPN2EOS

metadataDir

SSD

data1

data2

Transfer Queue - Priority Low

Metadata File 1

Metadata File 2

Metadata File 3

Transfer Queue - Priority High

Metadata File 4

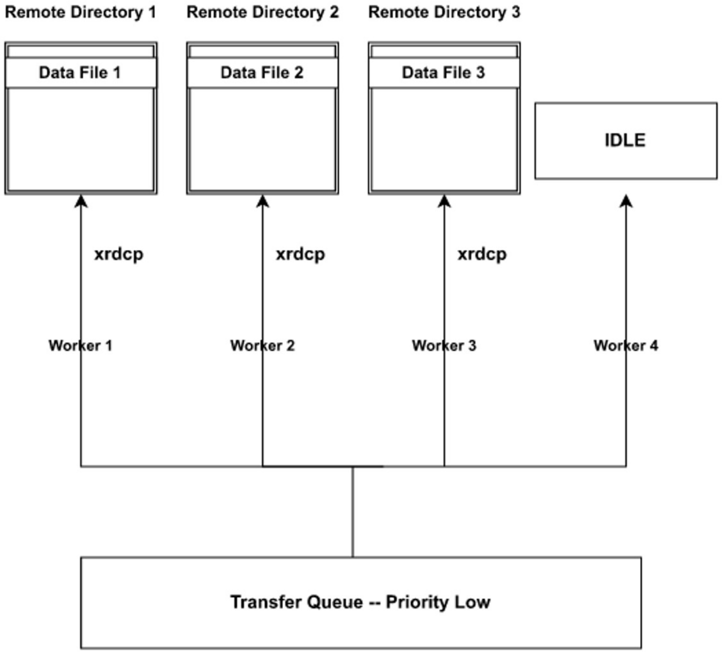
Metadata File 1
priority: low

Metadata File 2
priority: low

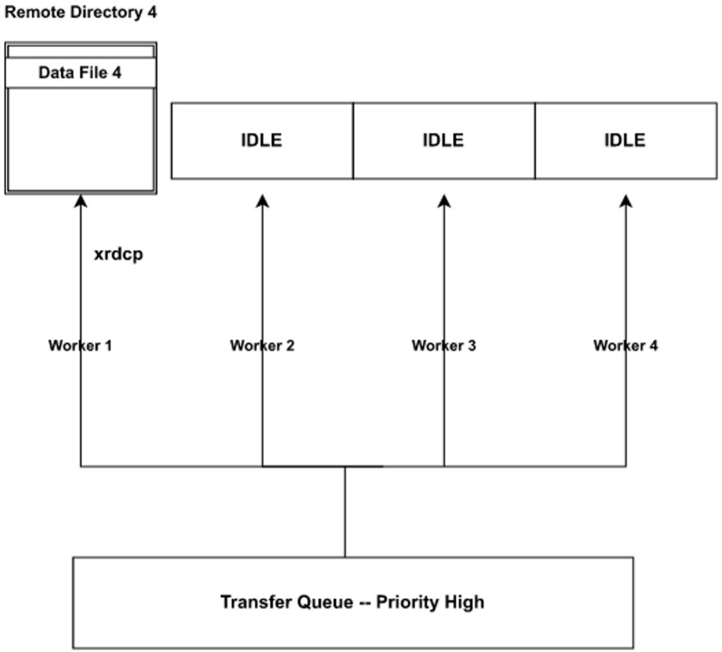
Metadata File 3
priority: low

Metadata File 4
priority: high

Parallel Transfer System and Priorities



4 threads for each priority



EPN2EOS Tasks and Tools

- Ensure that the data has been transferred quickly and successfully → xRootD

  xRootD:

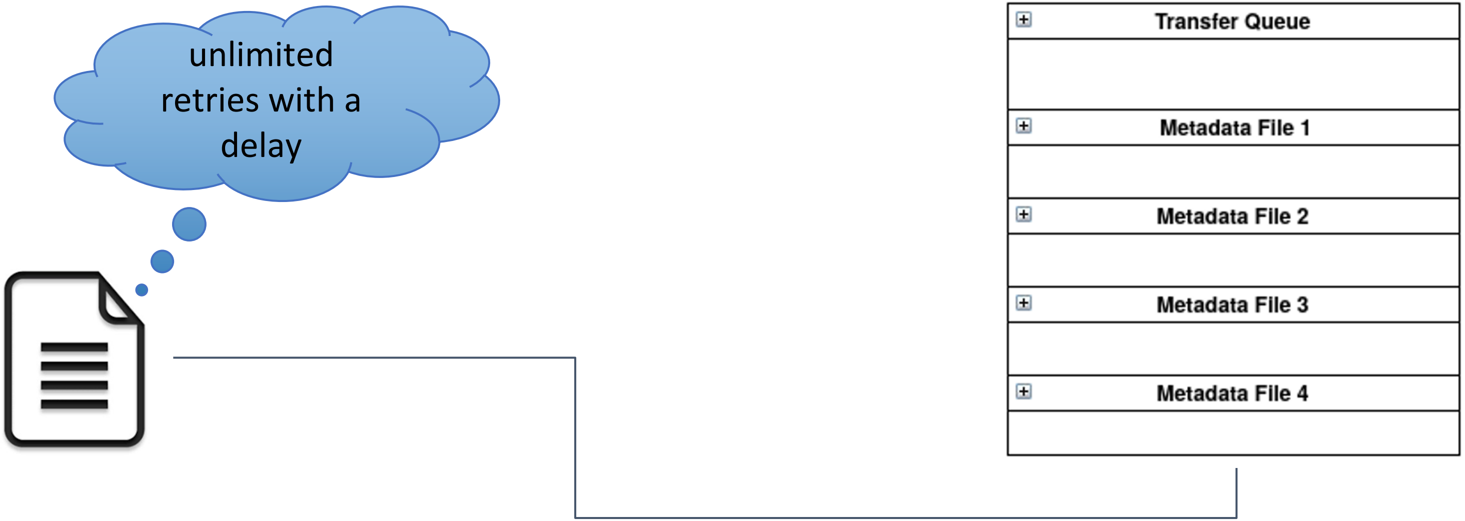
  - Is a data transfer protocol optimized for quick and efficient transfer over LAN and WAN
  - Implemented by all ALICE Grid storage endpoints, including EOS
- Verify that the data was correctly transferred → xxHash64

  xxHash64:

  - Fast and allows parallel processing of data blocks
  - Is implemented in EOS

Retry Mechanism

- On failure, compute a retry delay time for each file — exponential backoff
  - The delay time increases exponentially with the number of attempts to transmit the file (2^1 (second attempt), 2^2 . . . maxBackoff (60 seconds))



unlimited retries with a delay

Transfer Queue

Metadata File 1

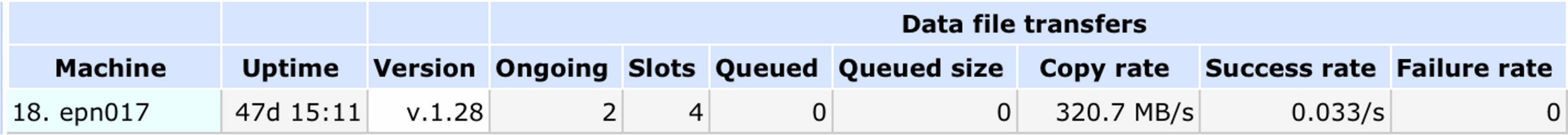
Metadata File 2

Metadata File 3

Metadata File 4

Monitoring System

- Log messages and monitor the system
  - Number of files in the queue for transmission
  - Number of successfully copied files
  - Number of failed transfers
  - Transferred bytes and transmission rate
  - Error rate
- Send alerts to list of recipients with details about the error condition and a message when recovered

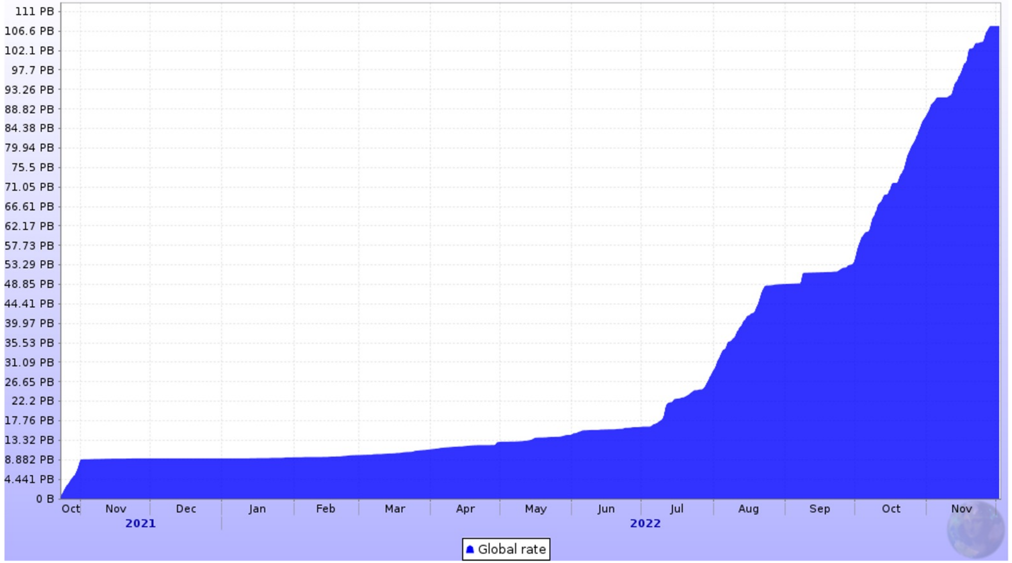


			Data file transfers						
Machine	Uptime	Version	Ongoing	Slots	Queued	Queued size	Copy rate	Success rate	Failure rate
18. epn017	47d 15:11	v.1.28	2	4	0	0	320.7 MB/s	0.033/s	0

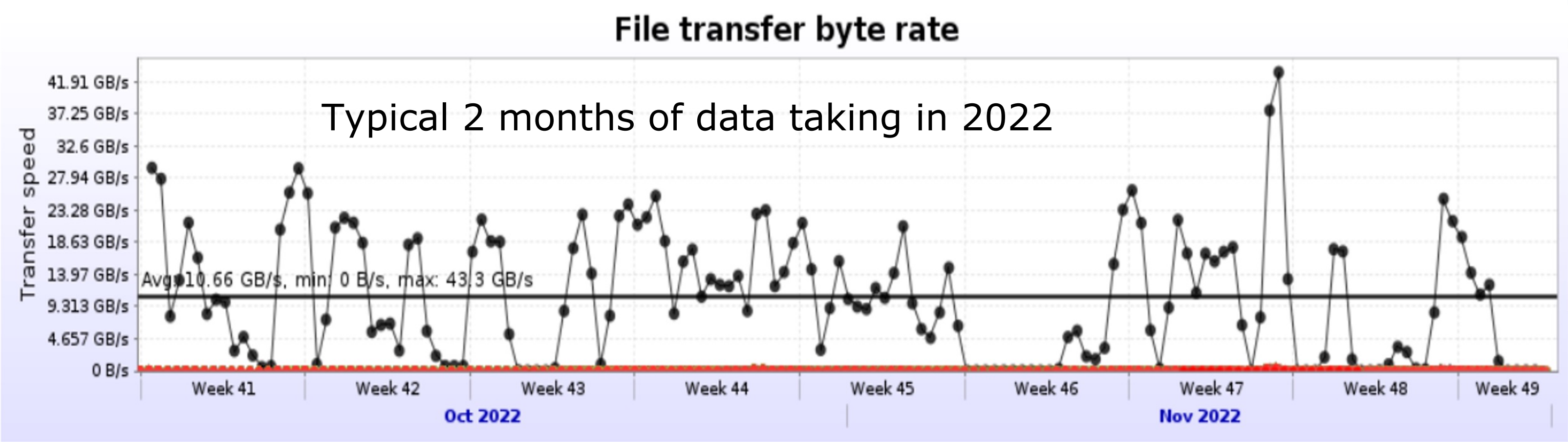
Cumulative Data Transfer

- System used in production
  - During the ALICE commissioning after upgrade in 2021
  - For the entire 2022 data taking year

- Total volume of transferred data - **107PB**
  - 75 M files, 1.4GB average file size



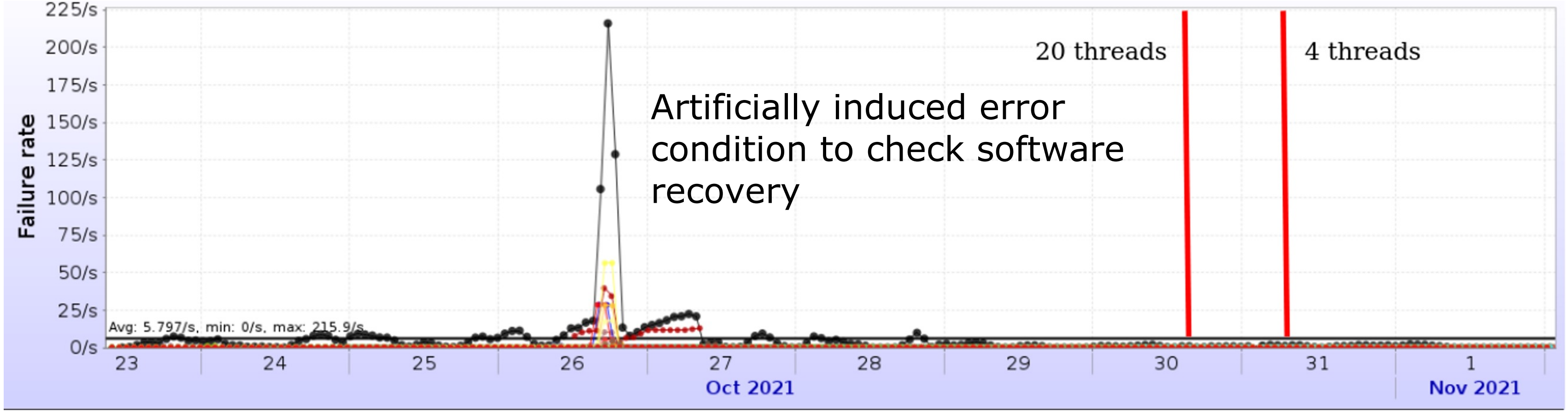
Transfer Speed



Typical 2 months of data taking in 2022

- Cyclical transfer structure due to standard LHC operation
- Optimization of parallel transfers
  - 20 transfer threads - aggregated transfer speed: 27GB/s
  - 4 transfer threads - aggregated transfer speed: 34GB/s
- 4 transfer threads adopted as standard - better use of available bandwidth from EPN to CERN IT

Transfer Error Handling



20 threads

4 threads

Artificially induced error condition to check software recovery

- Typical error: empty file on remote storage due to failed transfer
  - Since the files are write-once (for safety), the filename cannot be reused
- Solution: on retry, append the transfer attempt to filename on storage
  - The filename in the catalogue does not contain the retry number

Conclusions

- EPN2EOS is a **fully functional standalone system** for data transfer between the ALICE online processing cluster EPN and the IT-managed EOS storage
  - It works in the challenging condition of real time data taking
  - Uses xRootD for data transfer
  - Has transfer priority scheduling, robust error handling system, monitoring and messaging
- It is the **only** system used by ALICE to transfer **all data** from the experiment (including calibration) to storage and its registration for further processing