University
POLITEHNICA
of Bucharest

Faculty of
Automatic
Control and
Computers

EPN2EOS Data Transfer System

Computing in High Energy & Nuclear Physics - May 2023

Author

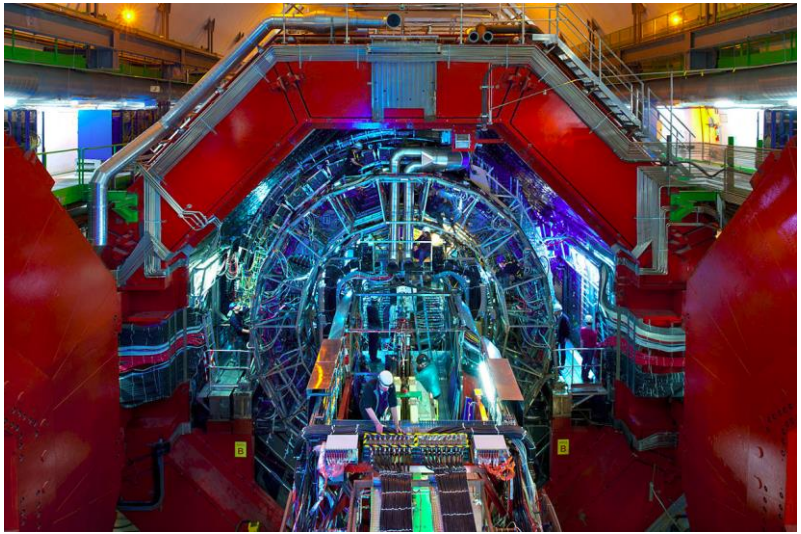
Alice-Florenţa Suiu
asuiu@cern.ch

Scientific Advisor(s)

Latchezar Betev
Costin Grigoras

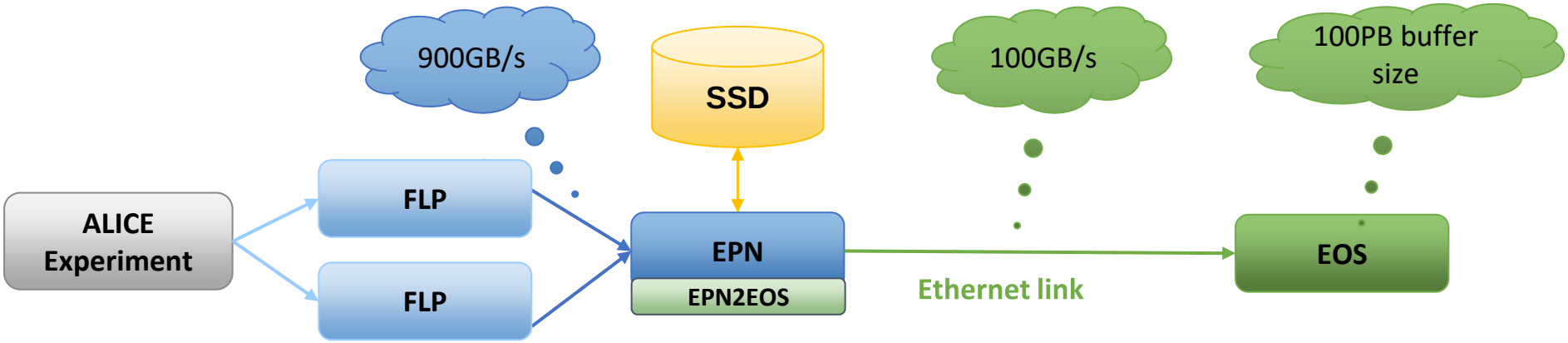
On behalf of the ALICE Collaboration

Context



- 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



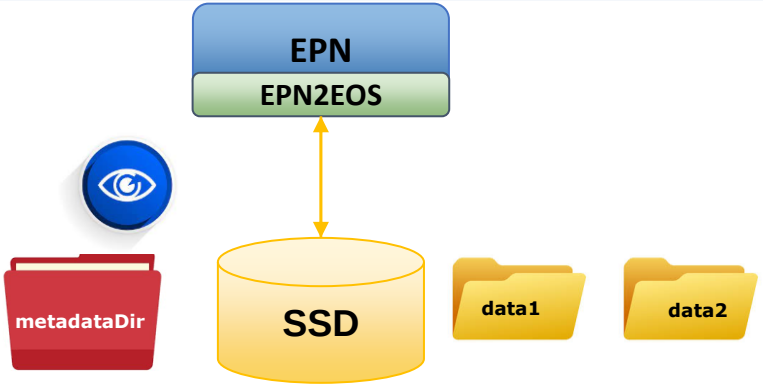
- 250 EPN nodes, each equipped with one 4TB SSD
- SSD buffer capacity sufficient for ~3h of data taking

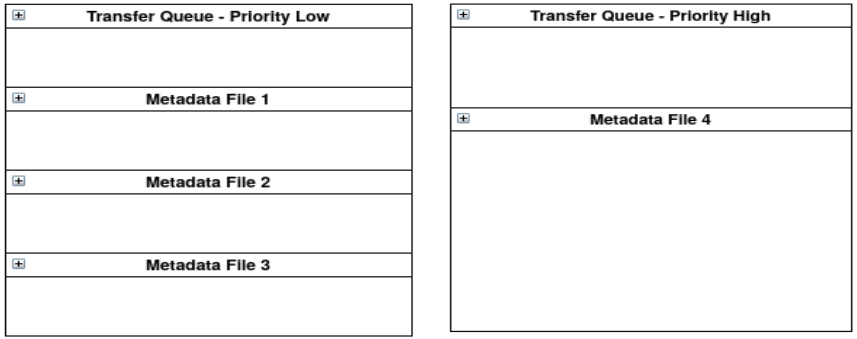
- EPNs produce ~2GB data files with frequency of 0.2Hz
- These must be transferred to EOS promptly and removed from the nodes

EPN2EOS 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





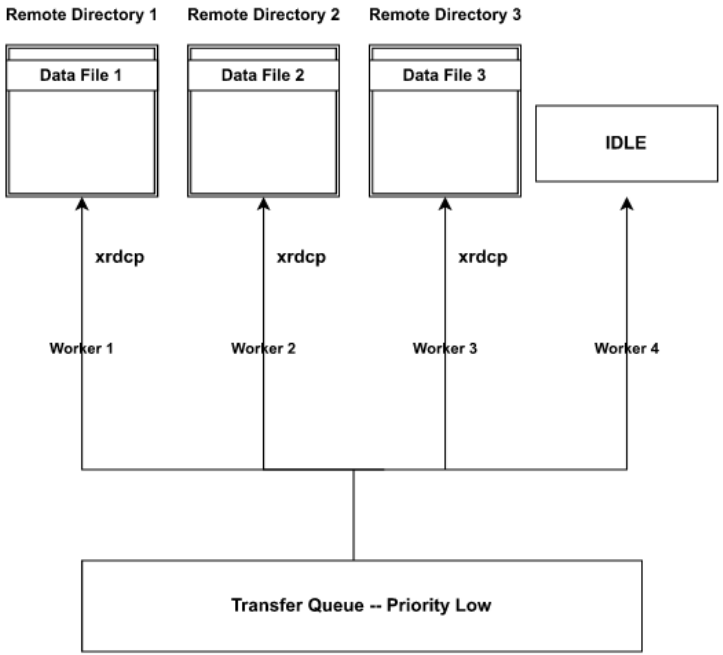
Metadata File 1
priority: low

Metadata File 2
priority: low

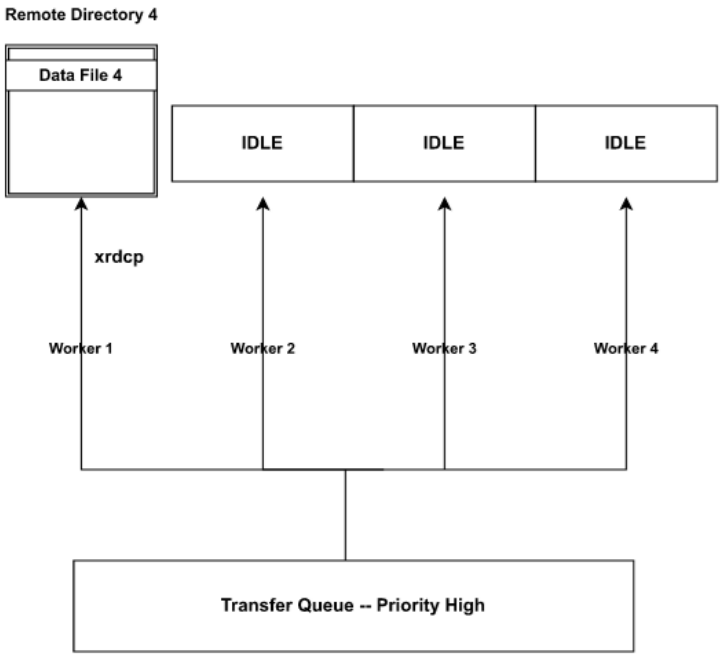
Metadata File 3
priority: low

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:

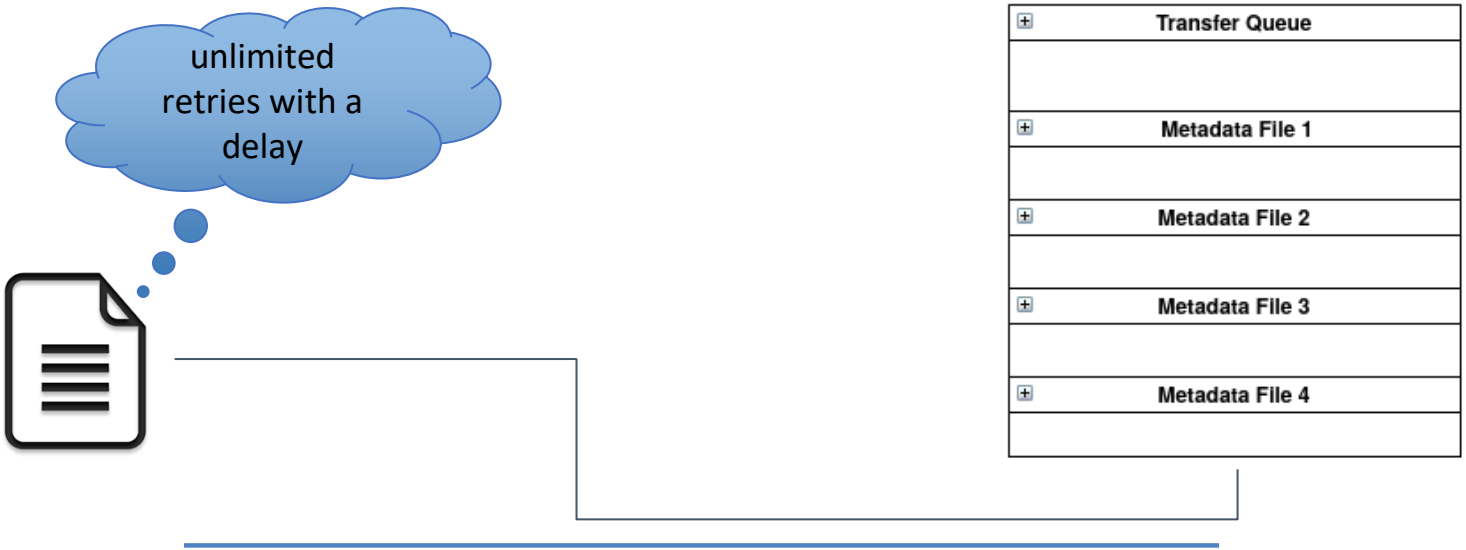
  - Is a data transfer protocol optimized for quick and efficient transfer over LAN and WAN
  - Implemented by all ALICE Grid storage endpoints, including EOS
- Verify that the data was correctly transferred

  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))



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

EPN2EOS file transfer service

What is this about?

Global view

	Services	Data file transfers						Catalogue registration				
Location	reporting	Ongoing	Slots	Queued	Copy rate	Success rate	Failure rate	Ongoing	Slots	Queued	Success rate	Failure rate
P2	282	2	2	0	0	0	0	0	0	0	0	0

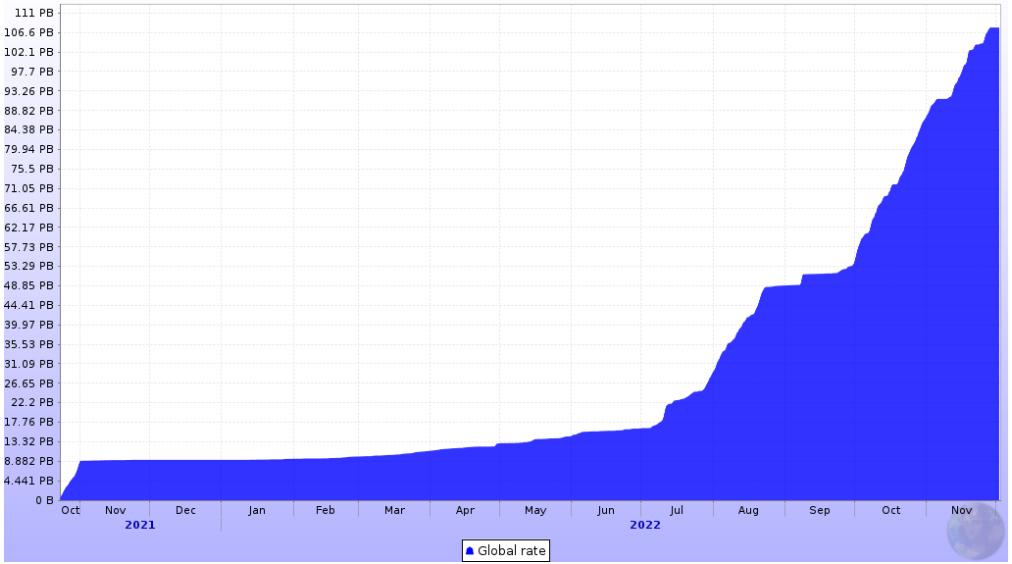
What is this about?

Detailed machine view

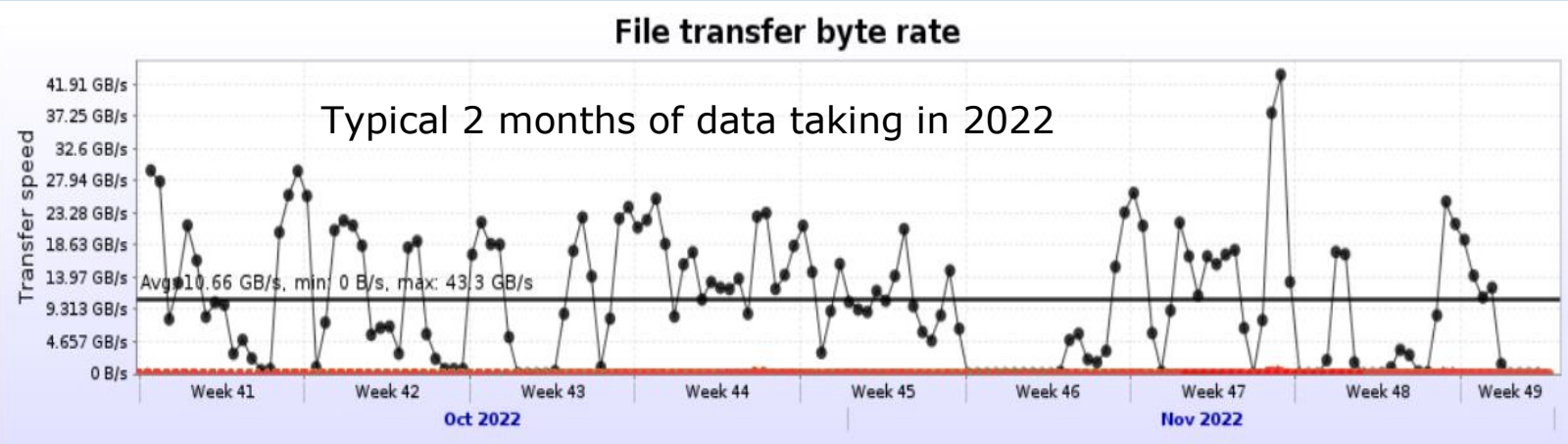
			Data file transfers							Catalogue registration					Accumulated error files			Target SE
Machine	Uptime	Version	Ongoing	Slots	Queued	Queued size	Copy rate	Success rate	Failure rate	Ongoing	Slots	Queued	Success rate	Failure rate	Rejected	Transfer	Missing source	Invalid meta	Write status
1. epn000	14d 0:55	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
2. epn001	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	1	0	0	0	
3. epn002	24d 2:14	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
4. epn003	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
5. epn004	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
6. epn005	24d 2:12	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
7. epn006	24d 2:12	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
8. epn007	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
9. epn008	24d 2:14	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
10. epn009	24d 2:14	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
11. epn010	24d 2:12	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
12. epn011	24d 2:12	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
13. epn012	24d 2:14	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
14. epn013	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
15. epn014	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
16. epn015	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
17. epn016	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
18. epn017	24d 2:13	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
19. epn018	24d 2:12	v.1.28	0	0	0	0	-	-	-	0	0	0	-	-	0	0	0	0	
20. epn019	20d 1:18	v.1.28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21. epn020	20d 1:16	v.1.28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22. epn021	20d 1:18	v.1.28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23. epn022	20d 1:16	v.1.28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24. epn023	20d 1:16	v.1.28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25. epn024	20d 1:16	v.1.28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	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

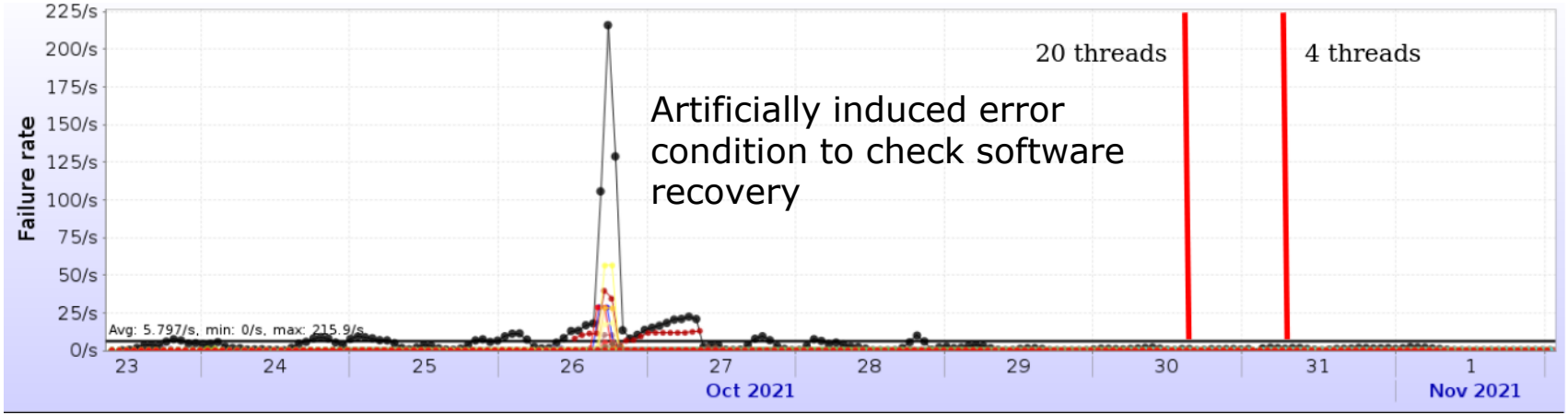


Transfer Speed



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

Transfer Error Handling



- 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