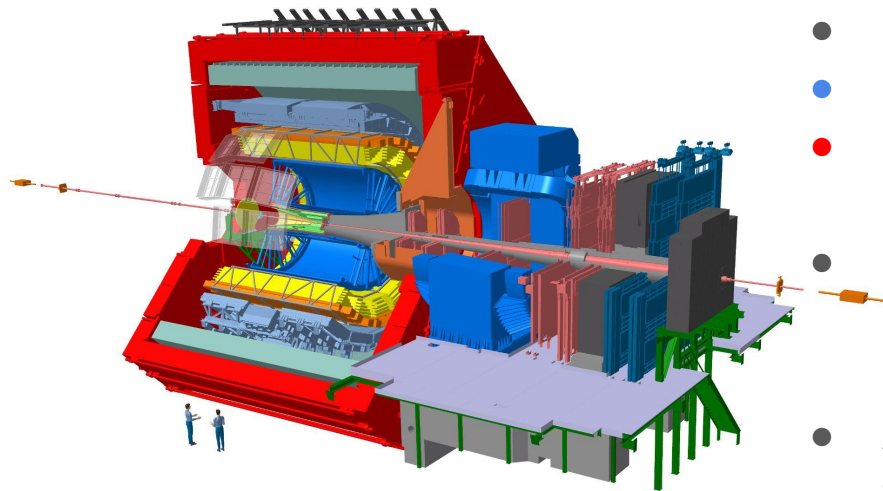




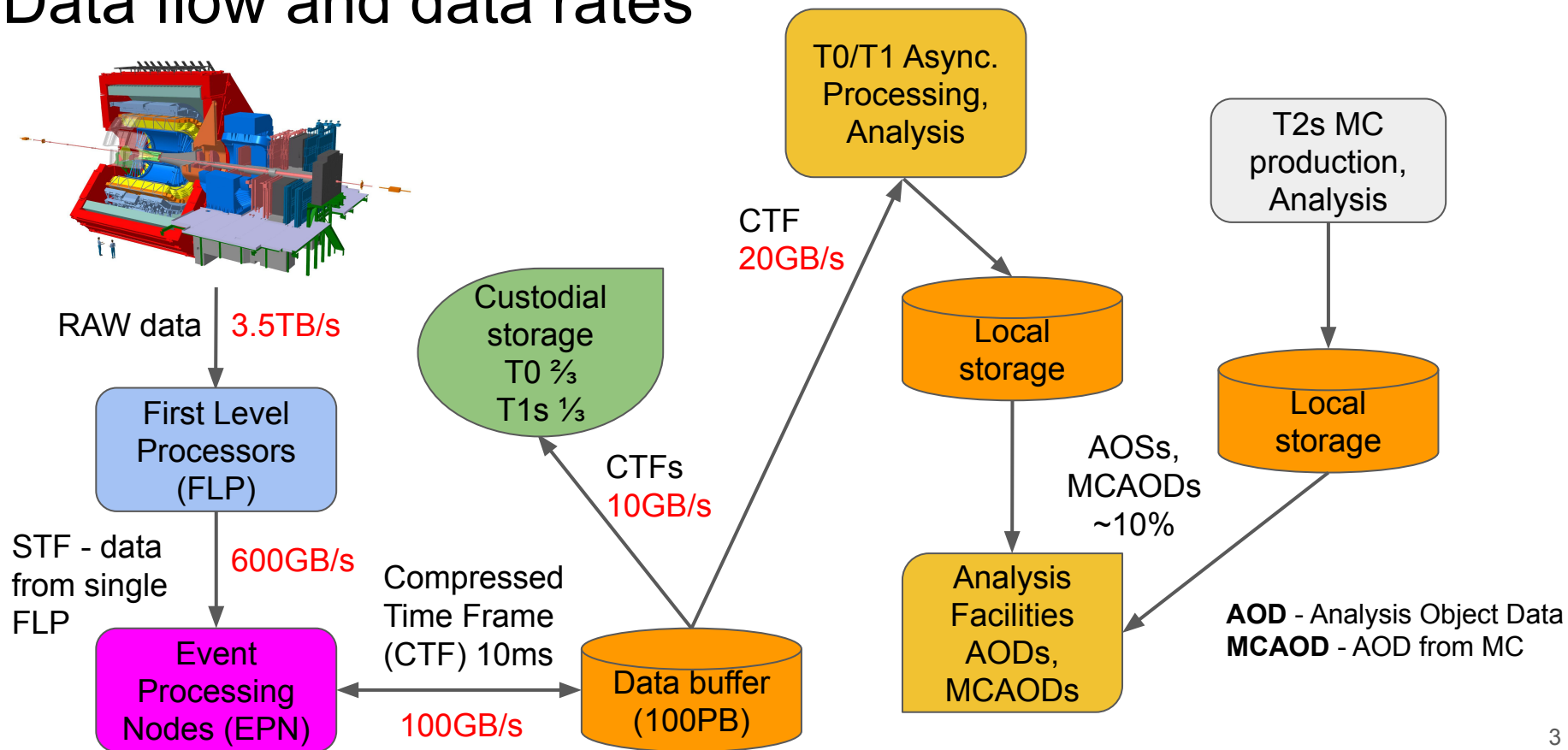
High-capacity, high-throughput EOS storage for ALICE data taking

ALICE upgrade general



- p-p and HI physics
- 10x integrated luminosity
- 100x event rate of Run 1/2, 10x more data
- Continuous readout
- Focus on data compression and real time (synchronous) data reconstruction
- => Reasonable rates and data volumes after compression to storage and secondary data formats
- Adherence to 'flat budget' resources funding for data processing and analysis

Data flow and data rates



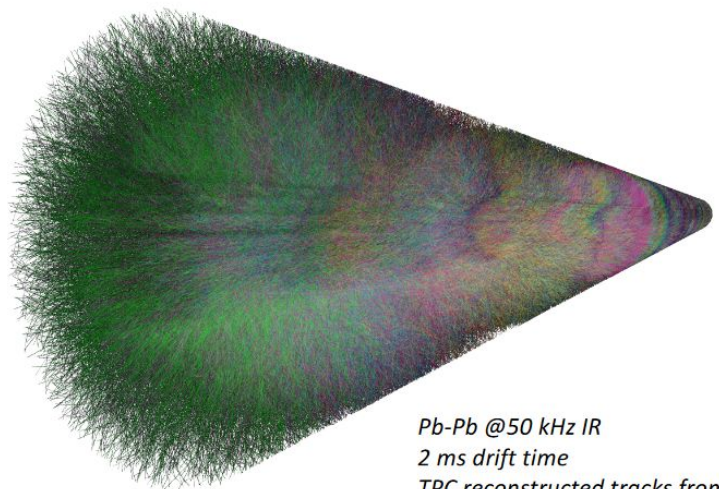
The O2 facility (EPNs)



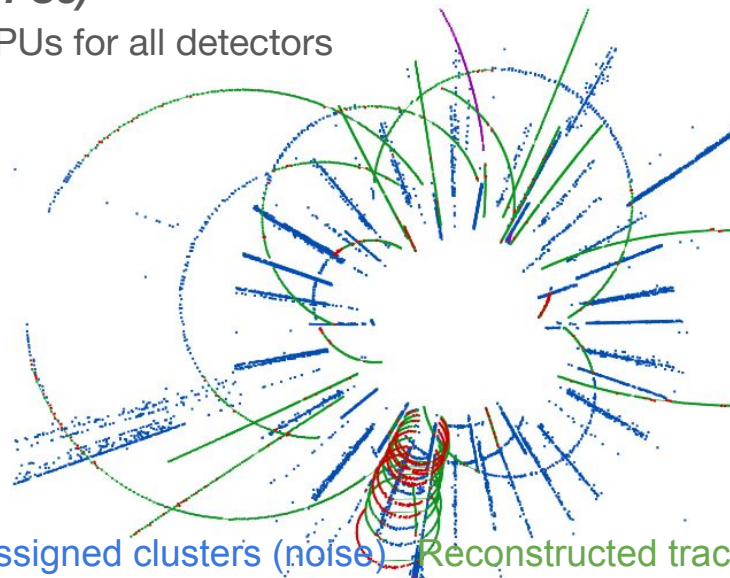
- Container-hosted computing facility located at the ALICE site, PUE<1.07
- High-throughput system, heterogeneous computing platform (CPU+GPU)
- 250 dual CPU nodes (ROME, 64 cores, 512GB RAM) with 8 AMD (MI50, 32GB) GPUs/node
- Functions
 - Data aggregation (Detector STFs to global CTF)
 - Synchronous global reconstruction
 - Calibration and data volume reduction
 - Quality control
 - Asynchronous data processing

Synchronous data processing

- Goal - to compress the RAW data by about factor 35 (3.5TB/s \rightarrow 100GB/s)
- Through zero suppression, clusterization, tracking, optimized data format
 - **Mandatory use of GPUs (~40x faster than CPUs)**
 - All synchronous level software is written for GPUs for all detectors



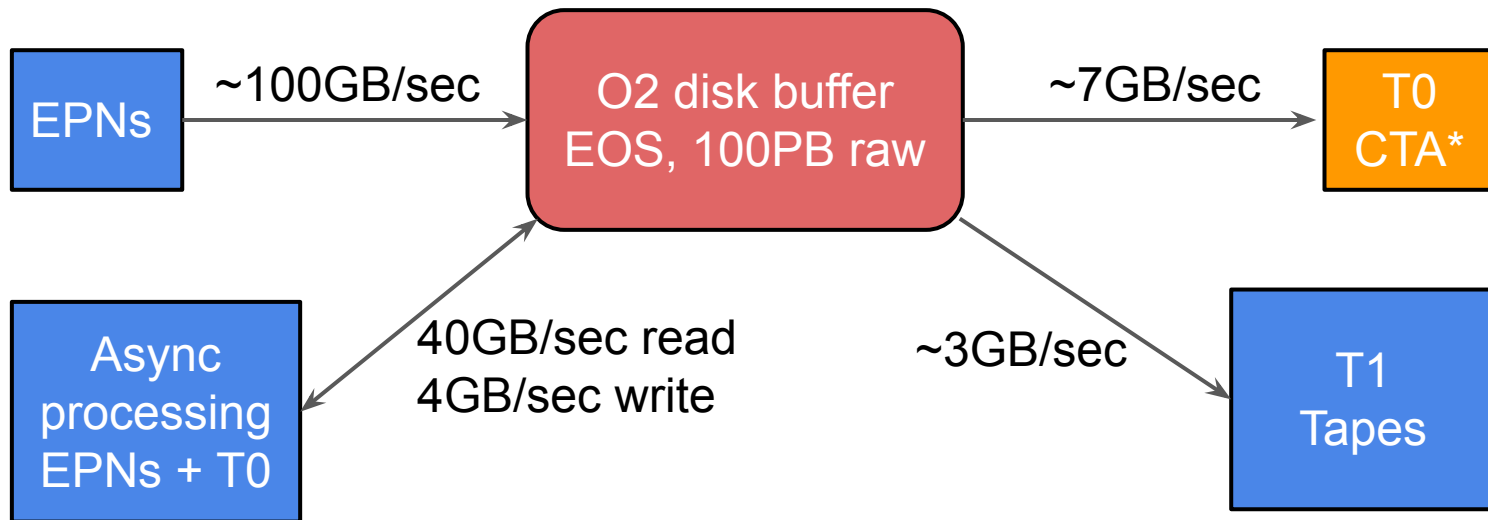
Pb-Pb @50 kHz IR
2 ms drift time
TPC reconstructed tracks from
different colour-coded events



Unassigned clusters (noise) Reconstructed tracks
Removed clusters Failed fits

EOS data buffer for O2 facility

- 100PB raw capacity, RS erasure coded (high level of data security)
- Based on cheap JBODs, SATA drives, EOS managed

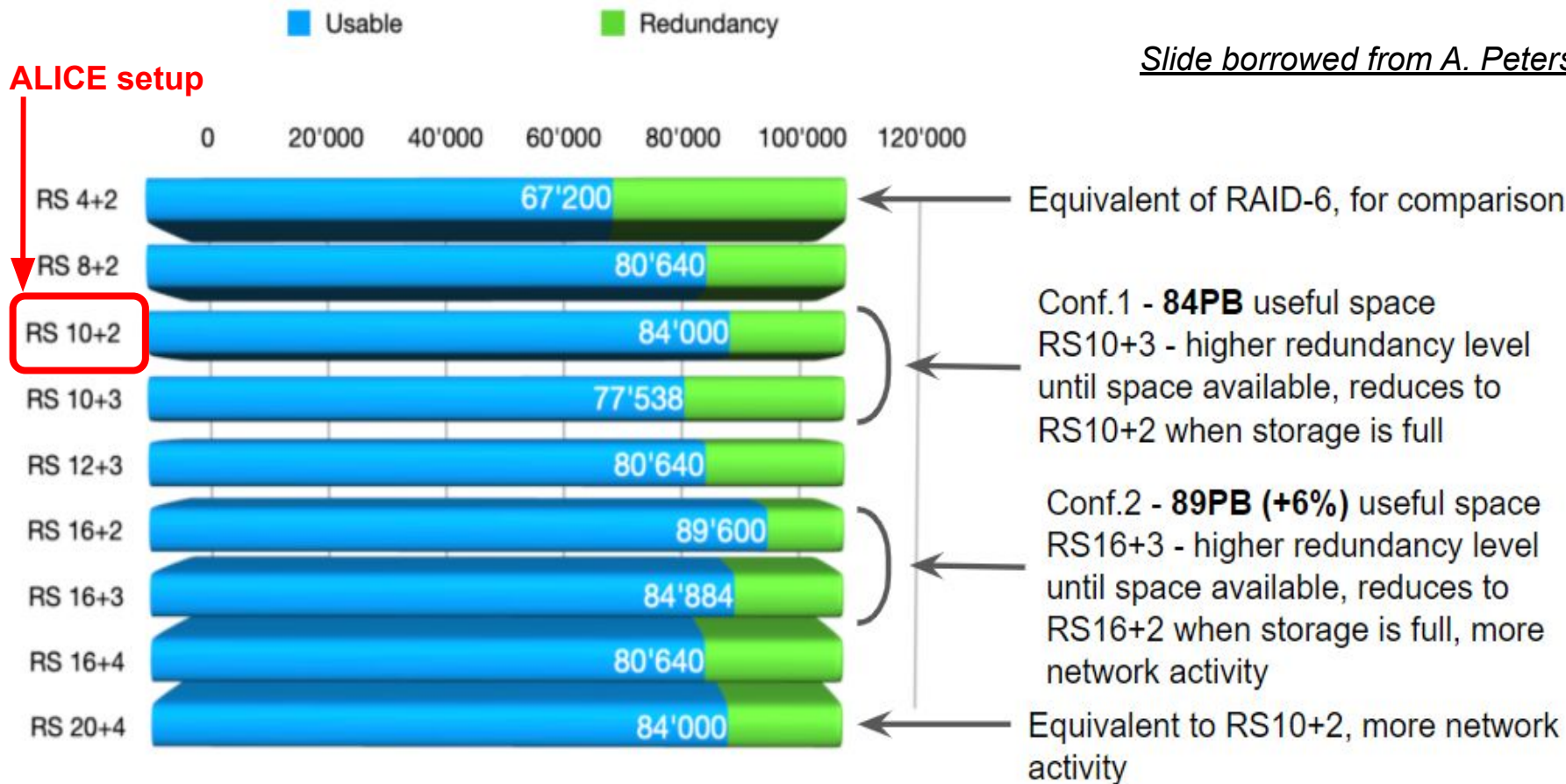


*CTA = CERN Tape Archive

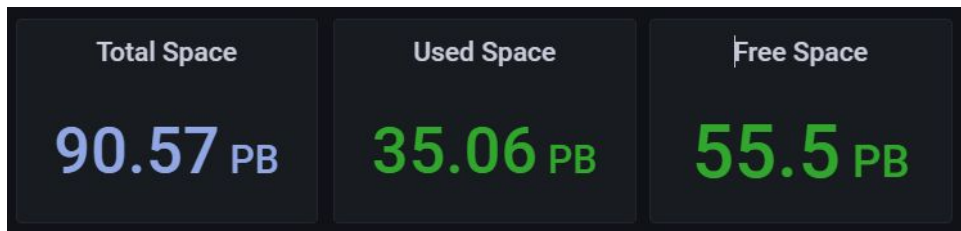
RAIN configuration effect on capacity

Slide borrowed from A. Peters

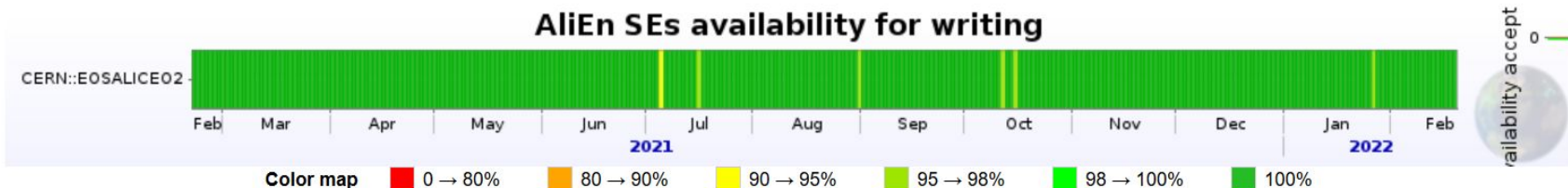
ALICE setup



EOSALICEO2 instance



Instance managed by CERN IT

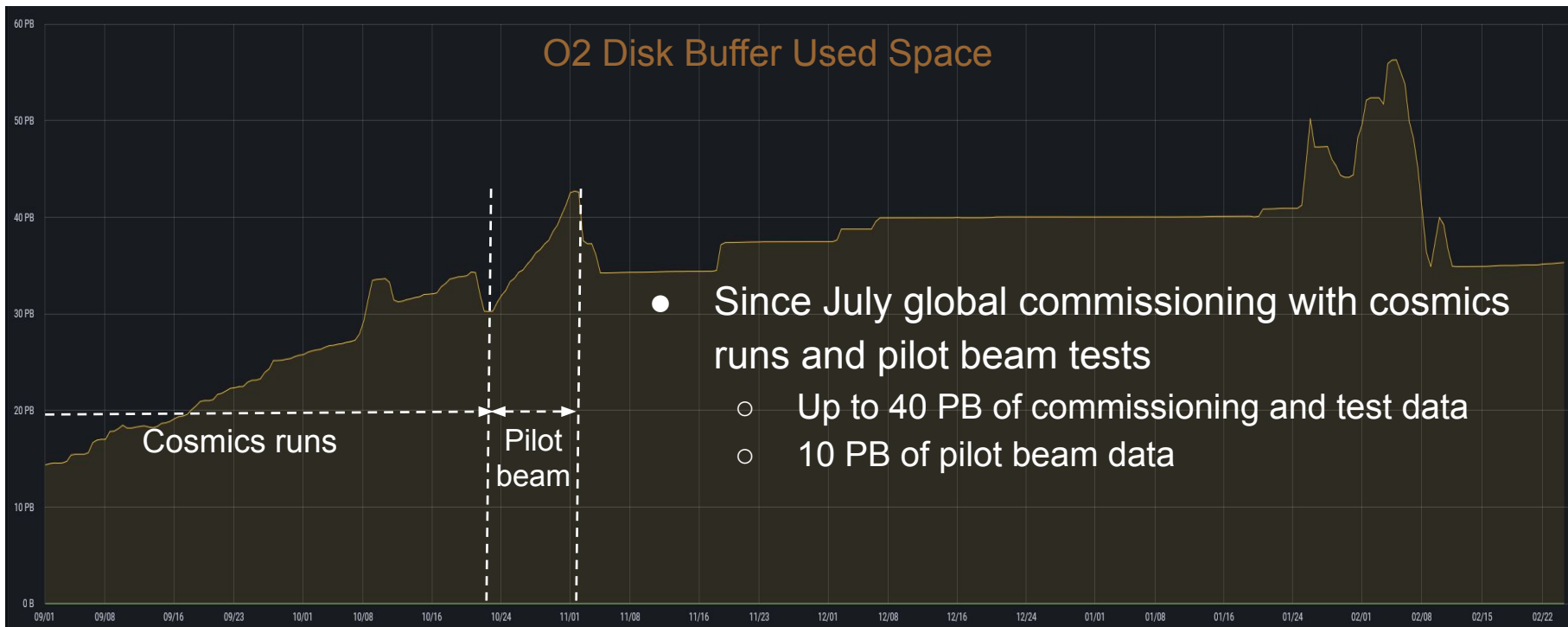


Not corrected for planned interventions

Statistics						
Link name	Data		Individual results of writing tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
CERN::EOSALICEO2	20 Feb 2021 01:07	20 Feb 2022 00:27	8744	7	99.92%	99.91%

ONLY 8 hours of down time for entire year

Usage of the disk buffer in the past year



Usage of EOSALICEO2 disk buffer - cont.

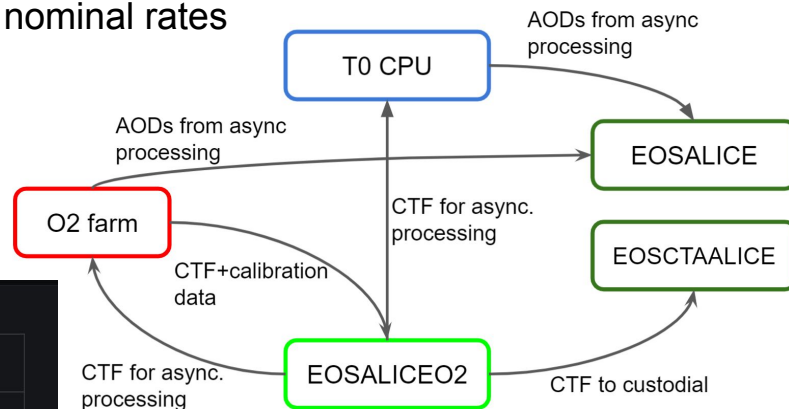
- The buffer is essentially in production for ALICE from mid-May
 - Used in all stages of ALICE commissioning data process
- Since then several updates
 - Space was increased to 90PB
 - Software updates were done as necessary
 - All was done transparently by IT experts and ***did not affect*** the availability of the storage
- In addition - several rate tests and challenges

CERN CTA data challenge

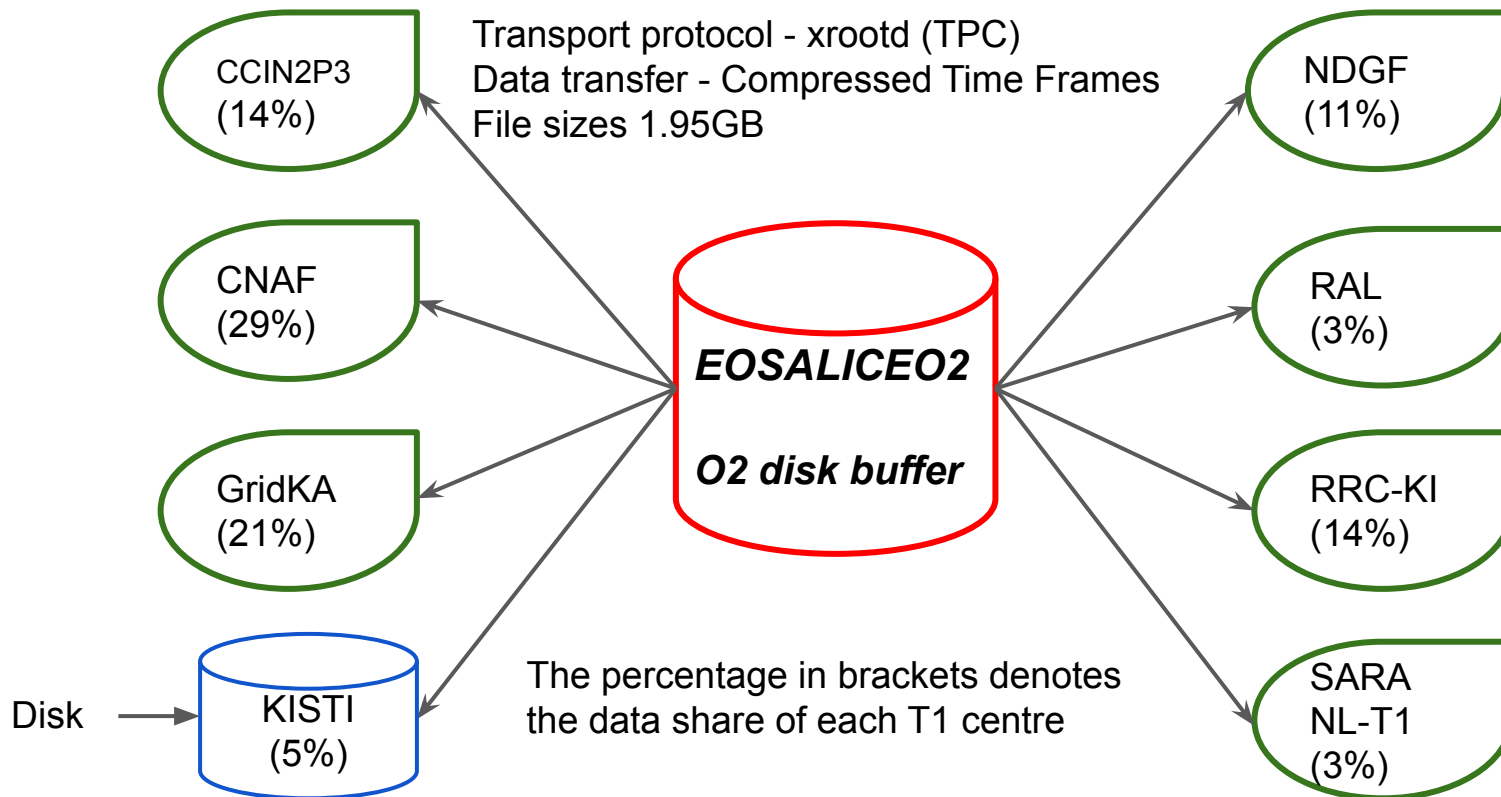
All together at (scaled) nominal rates

Transfer to CERN CTA @10GB/s

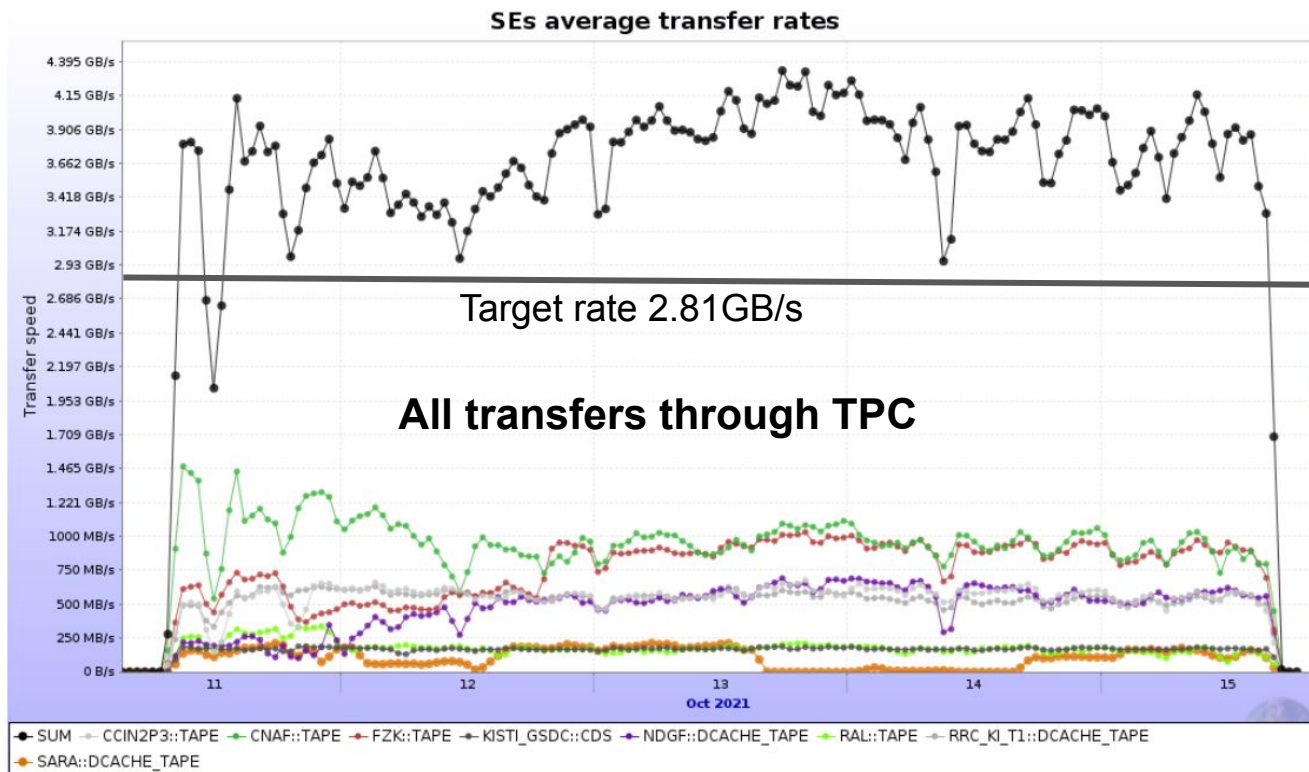
Write to EOSALICE02



Custodial storage challenge strategy



Data transfer rates over 5 days

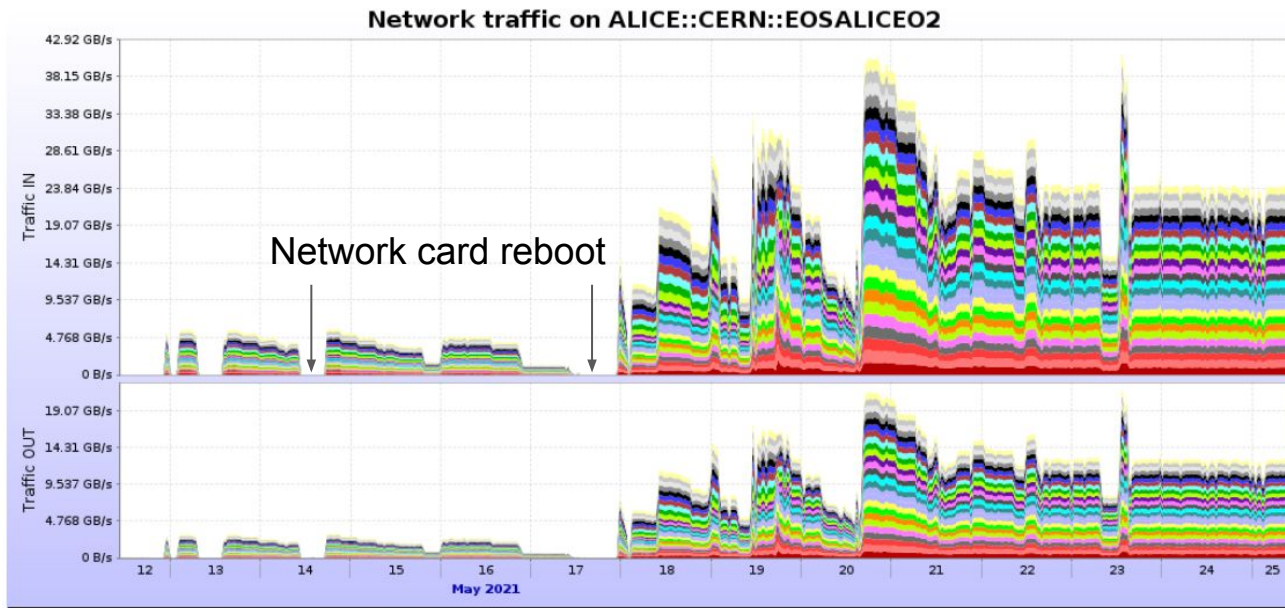


Client-server tests

- Test of data transport with xroot from EPN and comparison of rate and behaviour of two methods
 - C++ application streaming data from memory using the xroot API
 - **xrdcp from disk**
- Stable unattended operations for 72 hours with no losses (EPN to EOS)
 - At **nominal data taking** rate, IB to ENET gateway in place
 - With **nominal file size**, same as above
 - **Weekend-long unattended test** as soon as EPN machine room certified
 - Same workflow injecting/provoking common failures (EPN and EOS)

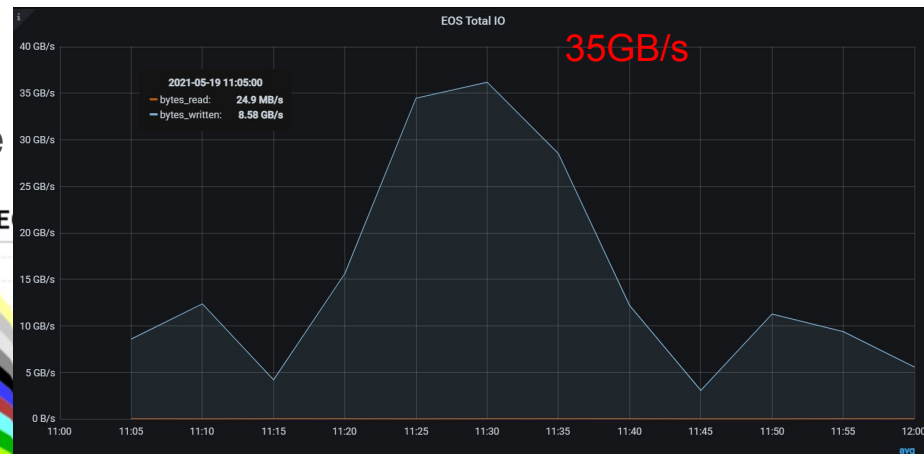
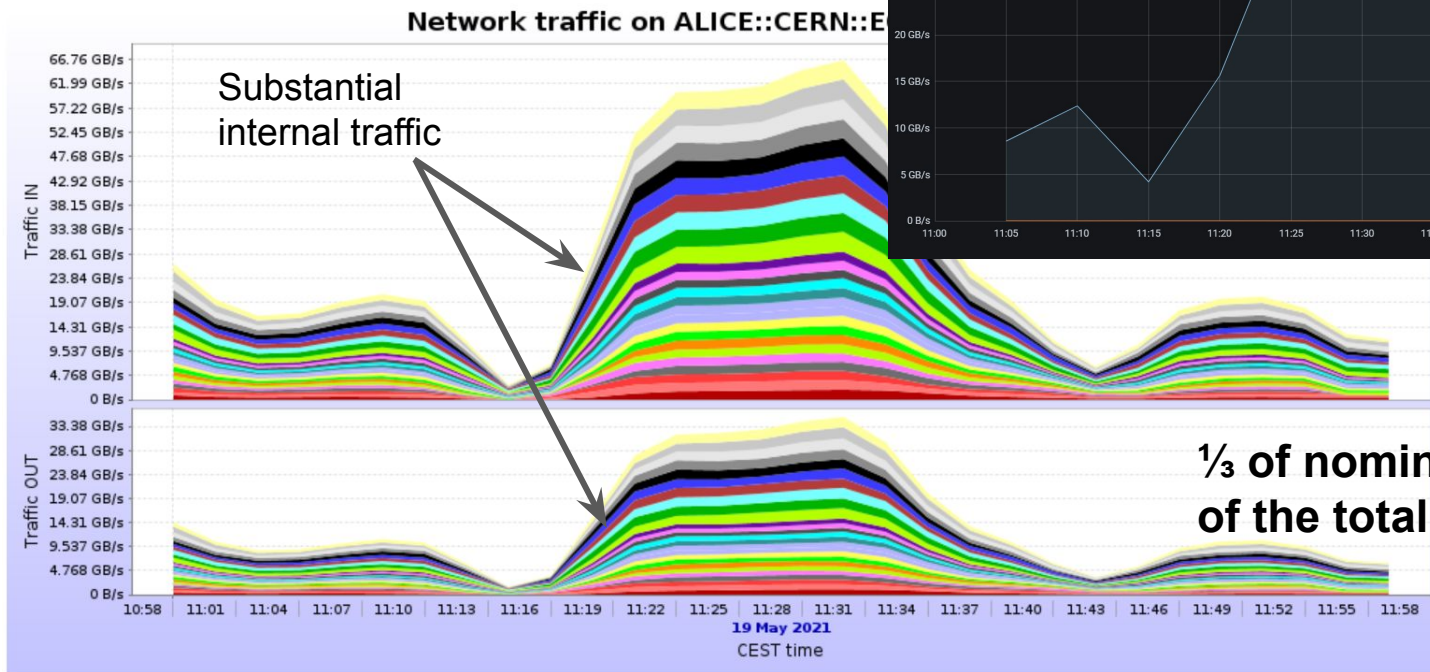
Test use cases - unattended operation

- Running since 13 May (12 days)
 - Minor and partially understood issues with network



Test use cases - rate tests

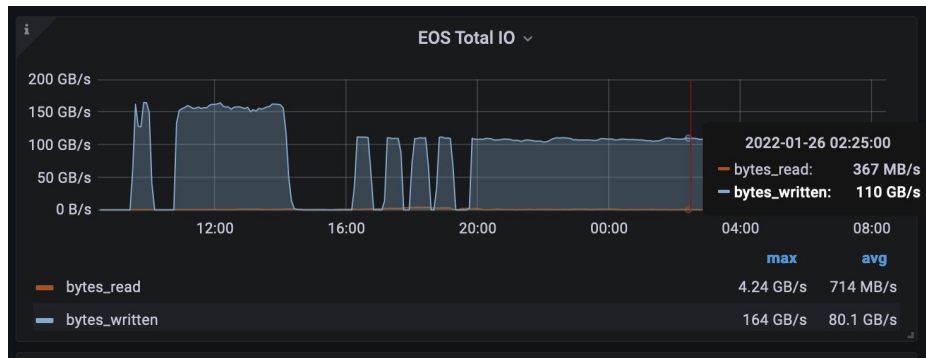
- 10 nodes, 20 streams/node, 2GB file



$\frac{1}{3}$ of nominal rate with $\frac{1}{3}$ of the total storage

Latest (25 January) rate tests from EOS

- Injecting 110 GB/s with 480 streams for 8 hours



- EOS performance with RS(10+2)
 - Avg bandwidth write-only > **140 GB/s**
 - Bandwidth reading+writing (1:2 ratio) **100 GB/s reading** + **110 GB/s writing** concurrently
 - Bandwidth read-only > **225 GB/s** (peaks at 248 GB/s)

Summary

- ALICE will use an IT developed and deployed EOS storage for the critical function of real-time raw data recording and offline processing
 - It has been installed and is called EOSALICEO2
- The size is 100PB (raw) 84PB useful
 - Sufficient to hold all collected data for a period of 1 year
 - It also implements adequate data protection through erasure coding
- The buffer functions
 - Receive data from the EPN compression facility
 - Serve it to the processing farms and to the custodial backup at CERN and T1s
 - With high r/w rates - up to 120GB/sec during Pb-Pb data taking (write) and up to 40GB/s sustained (read)
- Various test throughout the past year have confirmed the EOSALICEO2 capabilities for all foreseen use cases
 - Ready for Run3 data taking