EOS for ALICE O^2

Evolution & challenges (season 2022-2023)

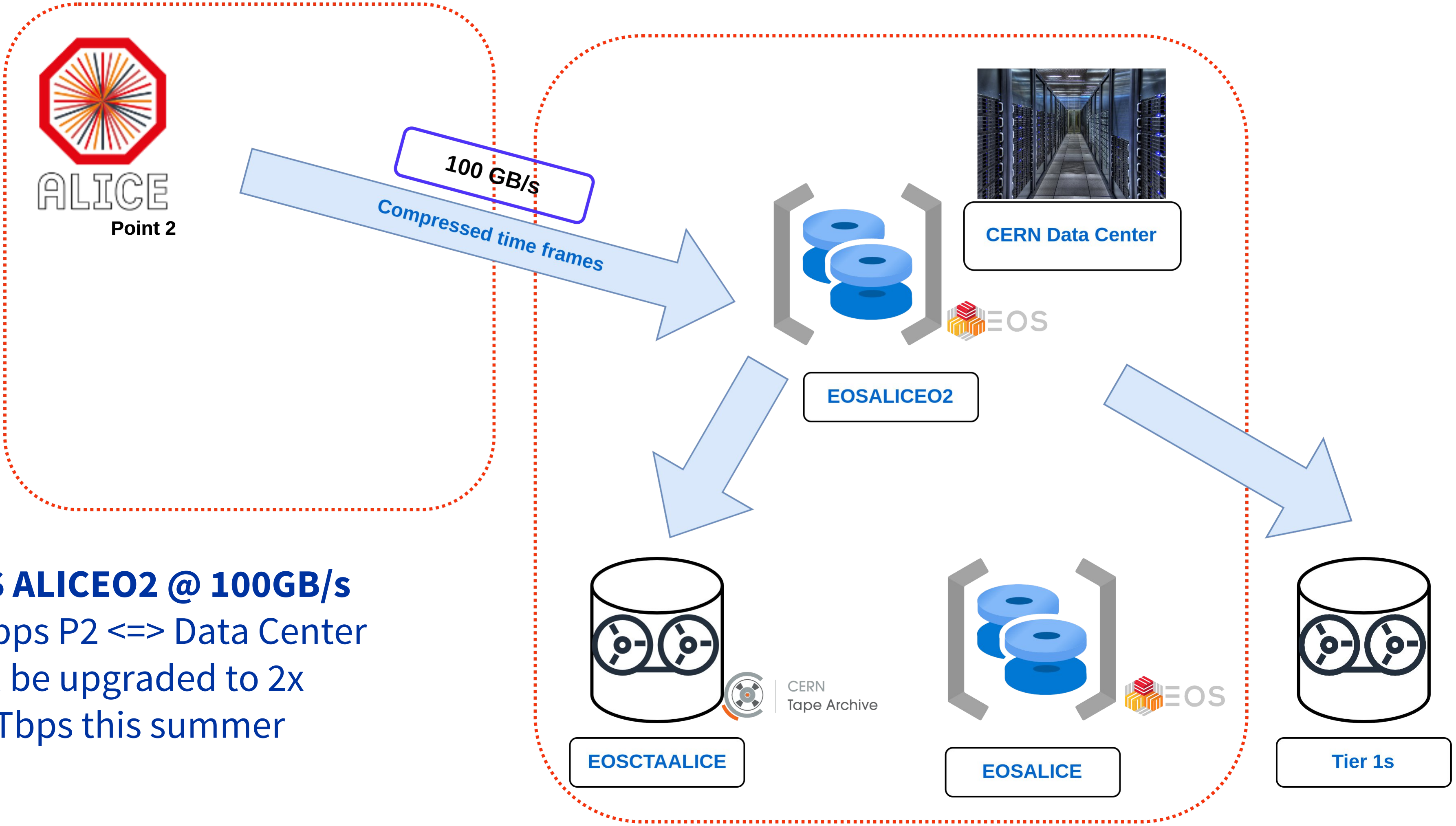
25.04.2023

Before summer 2022...

Before summer 2022...

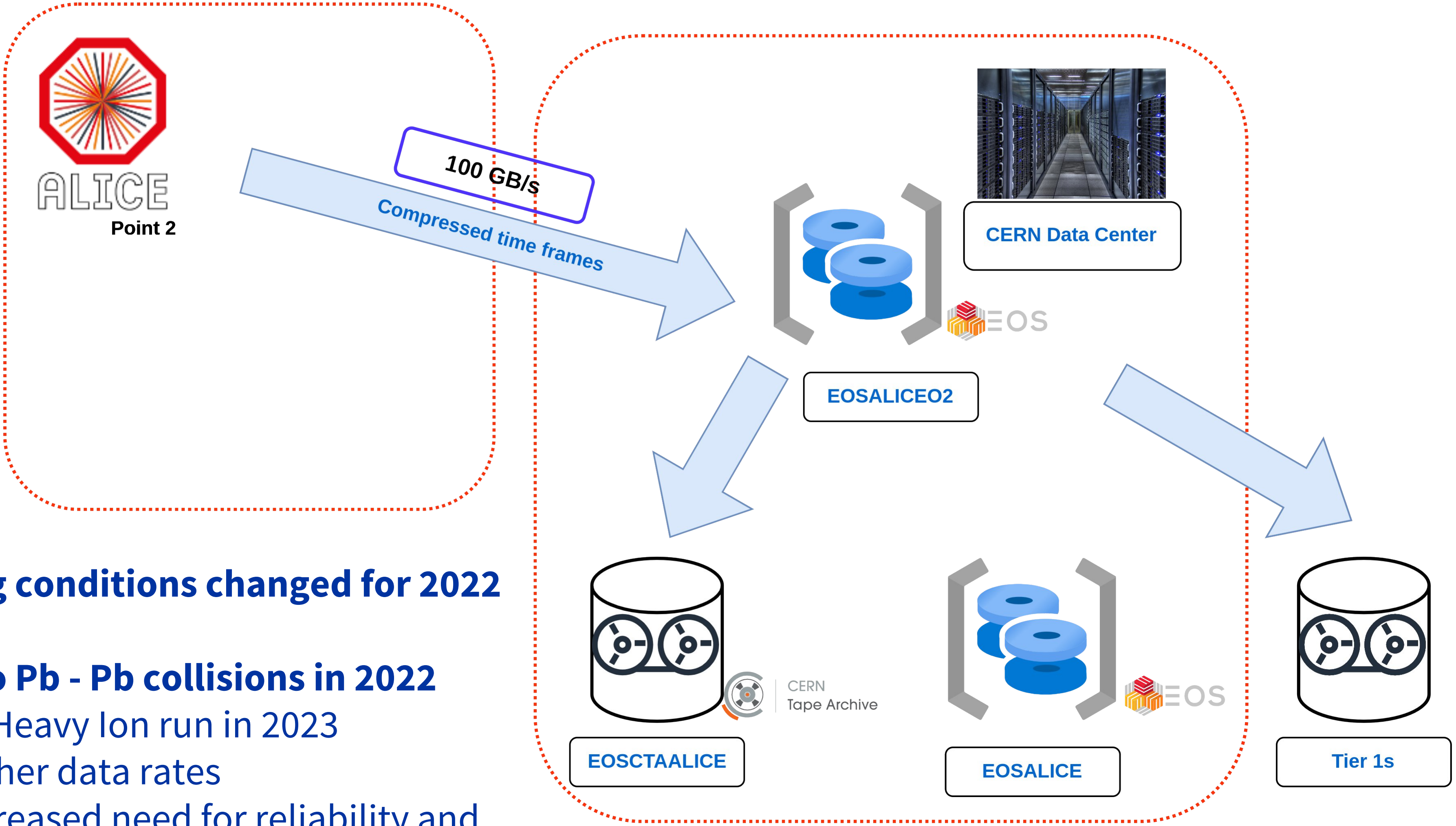
- EOS for ALICE O2 - HW setup and OS challenges (EOS Workshop 2021)
- Data flowing on the stream (EOS Workshop 2022)
- High-throughput EOS instance for ALICE O2 (HEPiX Fall 2021)

Run3 data taking for ALICE: Beginning of 2022 setup



- **EPN => EOS ALICEO2 @ 100GB/s**
  - 2x 1.2Tbps P2 <=> Data Center
    - will be upgraded to 2x 1.6Tbps this summer

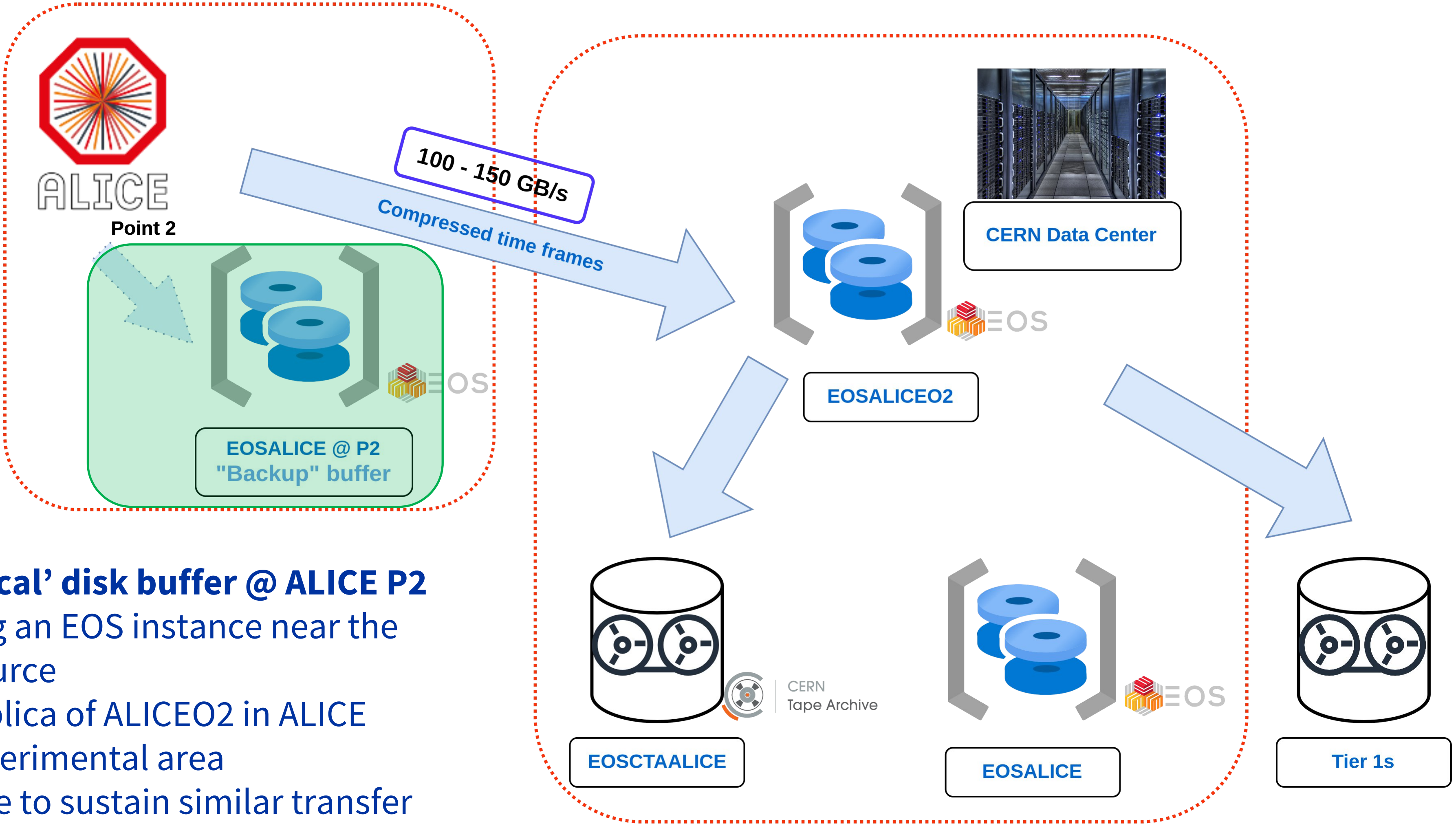
Run3 data taking for ALICE: beginning of 2022 setup



- **Data taking conditions changed for 2022 - 2023**
- **(Almost) no Pb - Pb collisions in 2022**
  - longer Heavy Ion run in 2023
    - higher data rates
    - increased need for reliability and availability

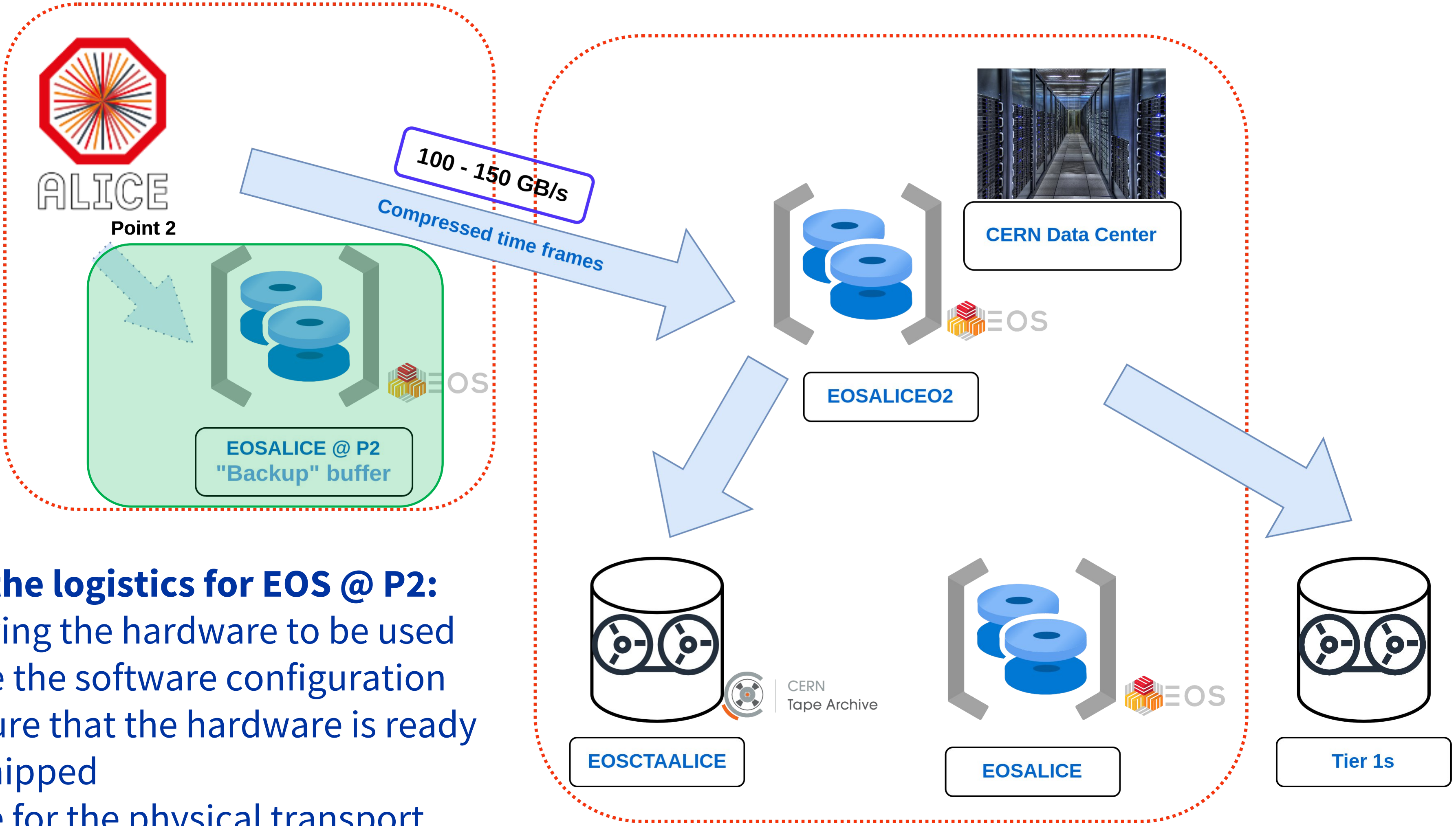
After summer 2022...

Run3 data taking for ALICE: current storage setup



- **Increase 'local' disk buffer @ ALICE P2**
  - Creating an EOS instance near the data source
    - Replica of ALICEO2 in ALICE experimental area
    - Able to sustain similar transfer rates

Run3 data taking for ALICE: current storage setup



- **Preparing the logistics for EOS @ P2:**
  - Identifying the hardware to be used
  - Prepare the software configuration
  - Make sure that the hardware is ready to be shipped
  - Arrange for the physical transport

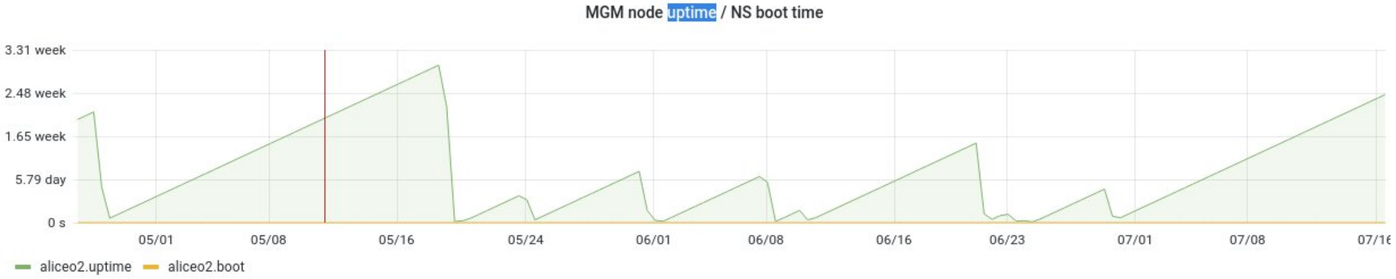
The journey to the new setup...

May 2022: choosing what to move

- **How did we pick the nodes to be moved? Easy…**
  - Two types of storage nodes in the EOSALICEO2 cluster (74 nodes in total):
    - 58x nodes with 96 disks
    - 16x high-density nodes with ‘only’ 60 disks per node — ‘breaking’ slightly the uniformity of the cluster
  - Compute nodes (for MGMs and QuarkDB cluster) from spare pool
    - 12 ‘CPU’ nodes (3x quad servers): 3x for MGM/QDB and 9x for tests/spare
- **Waiting for hardware deliveries…**
  - New hardware available only ~ 1 month before the transport was scheduled
  - ~10PB of data to be drained
- **Started the node draining campaign**
  - New EOS subsystem created for this (Group Drainer)
    - Allows draining data on specific groups
      - Destination is any other group that is not scheduled for draining
      - Needed by groups with not enough capacity to accommodate the drained data
  - Not without hurdles

May – June 2022: Draining hurdles

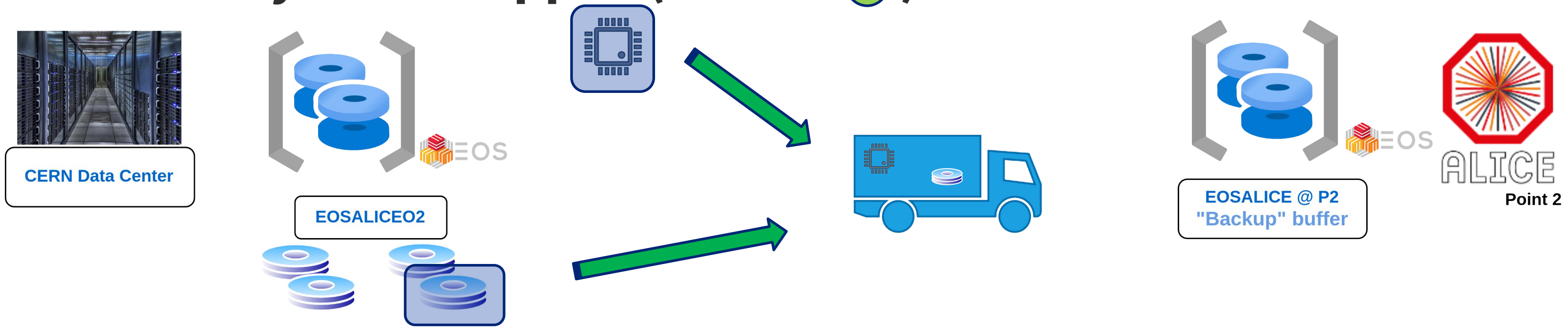
- **End of May:**
  - New storage nodes added to EOSALICEO2
  - Draining started
- **Off to a slow start**
  - code tweaks needed to increase performance of the GroupDrainer (many thanks to Abhishek and Elvin)
  - Quick release / deploy loop
- **The draining and its hurdles:**
  - Started the draining of the nodes to be moved
  - ALICE test runs in view of the LHC Run3 start



- **MGM started to become unstable:**
  - Random(?) segmentation faults triggered by the Conversion jobs generated by the GroupDrainer
  - Debugging the issues observed and getting a couple of patches for the XRootD client

End of June 2022: finishing up the draining

- **Address sanitizer EOS builds: spotting the underlying issues**
- **Workaround in place before fixes pushed into XRootD**
  - Avoid 0-sized files while converting for the time being
  - Create an external tool to pick the files to be converted => sent manually to the EOS converter
    - Rinse and repeat until only problematic files remained
      - These files were never properly transferred => old bug leaving Delete_On_Close files lingering in the namespace
- **‘Draining’ ramped up**
- **Hardware ready to be shipped (in time 🙂)**



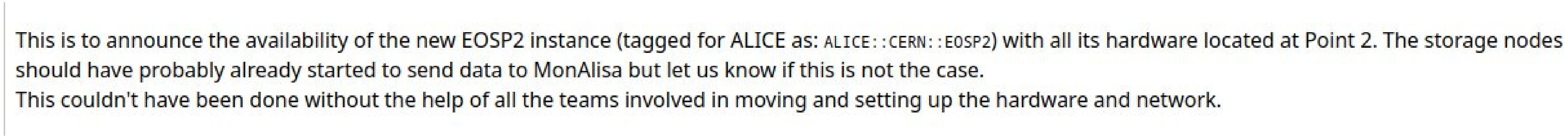
And it got there...

And it got there...

...I have proof

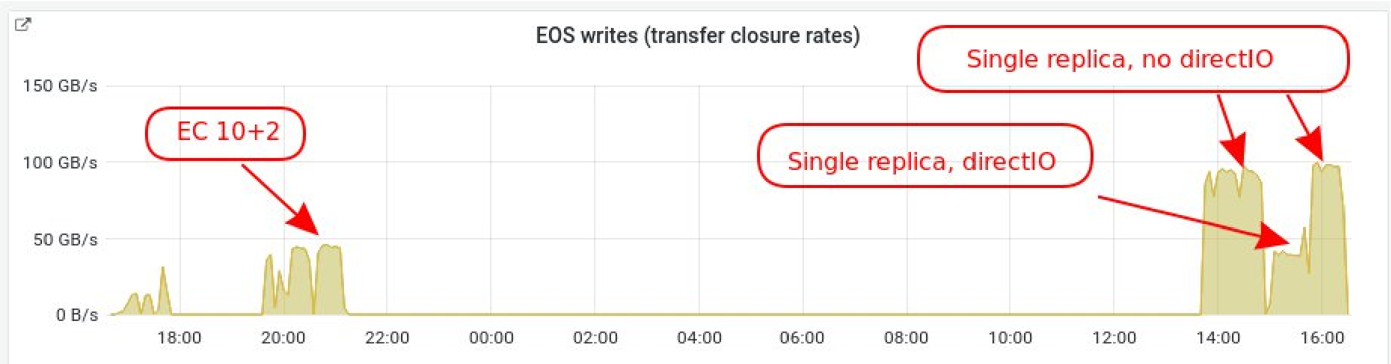
August 2022: A new EOS instance was born

- New hardware installed and wired in the ALICE barracks @ Point2 during July
- In parallel, we prepared the installation + configuration of the cluster
- On 4th of August, after a couple of days:



> This is to announce the availability of the new EOSP2 instance (tagged for ALICE as: `ALICE::CERN::EOSP2`) with all its hardware located at Point 2. The storage nodes should have probably already started to send data to MonAlisa but let us know if this is not the case.
> This couldn't have been done without the help of all the teams involved in moving and setting up the hardware and network.

- ...and it reached the expectations:

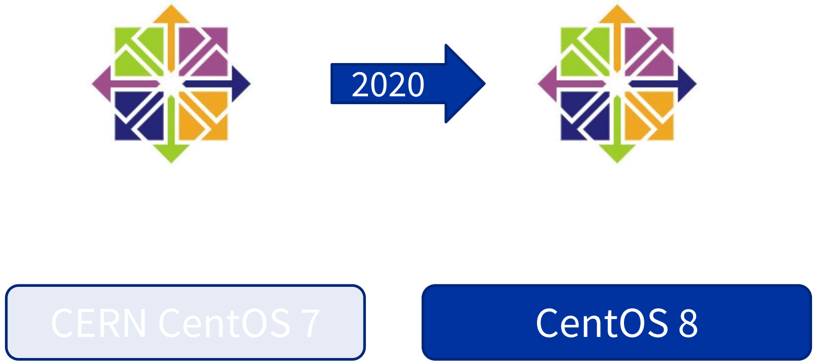


Some EOSALICE for O^2 news from 2023…

Change of Operating System (again)

CERN CentOS 7

Change of Operating System (again)



2020

CERN CentOS 7

CentOS 8

Change of Operating System (again)

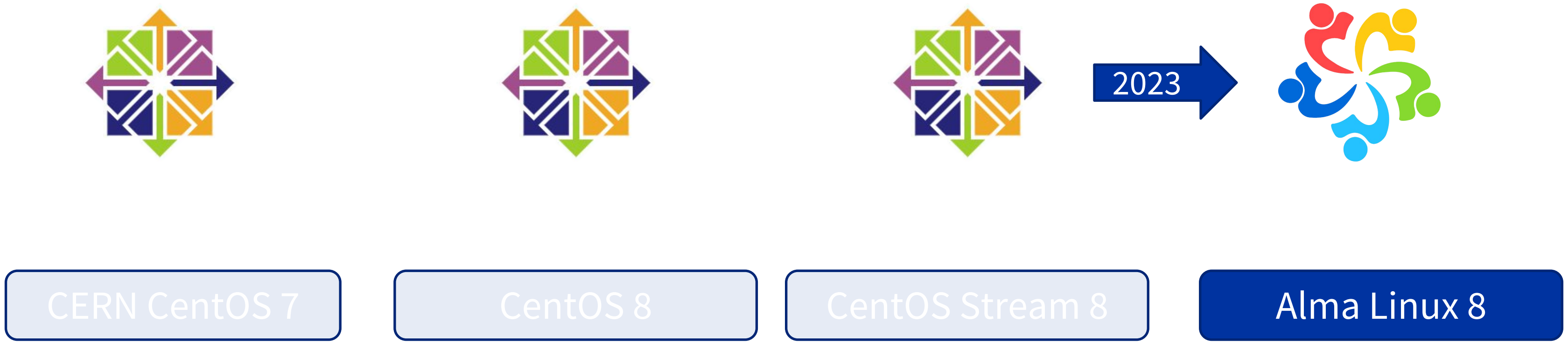
CERN CentOS 7

CentOS 8

2021

CentOS Stream 8

Change of Operating System (again)



CERN CentOS 7

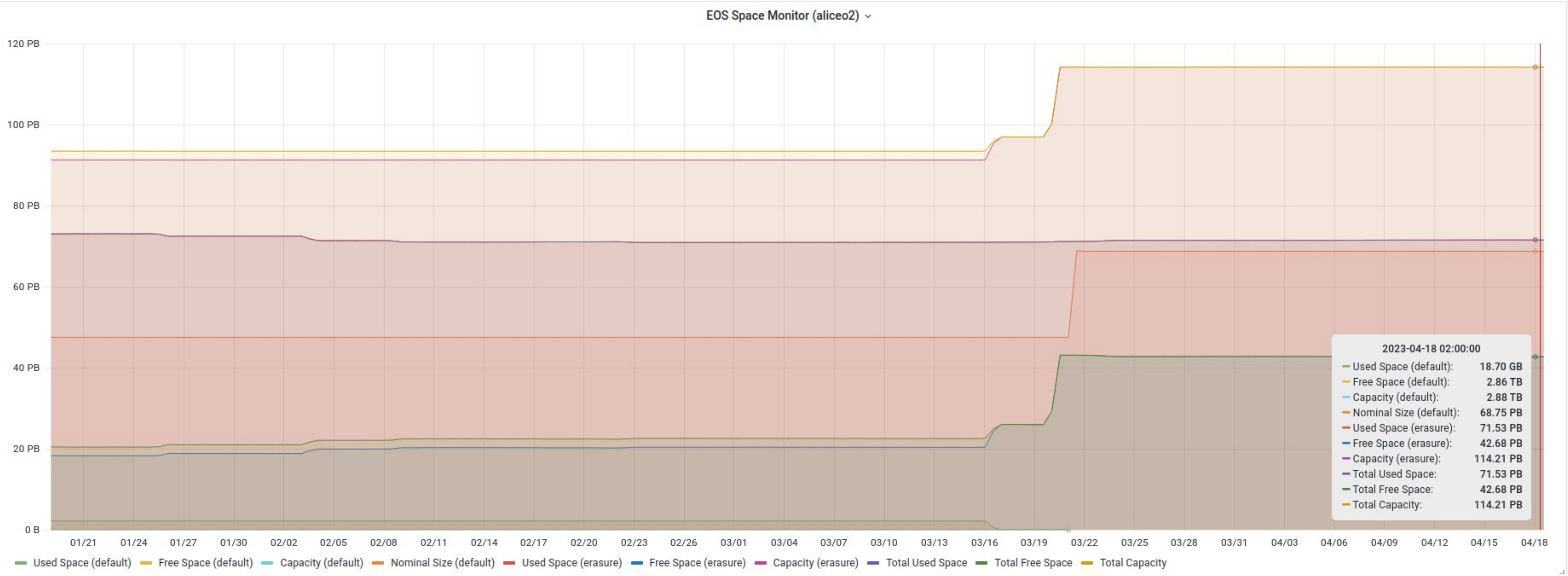
CentOS 8

CentOS Stream 8

2023

Alma Linux 8

ALICEO2 capacity increase (March 2023)



In place of conclusions

- **Bugs became more insidious … and that’s a good thing**
  - Less occurrences in ‘normal’ running conditions
  - Our developers have the right tools and knowledge to find and squash them
- **Draining might take more time than expected**
  - Do plan and start in advance
- **Operating system changes are not a big hurdle**
  - But again, things go better if planned in advance

Thank you!

home.cern