Operational challenges of the Event Processing Nodes GPU farm at ALICE Experiment

Federico Ronchetti – Giada Erba

federico.ronchetti@cern.ch
giada.erba@cern.ch

On behalf of the EPN team and the ALICE Collaboration

FIAS Frankfurt Institute for Advanced Studies

Sustainable HEP 2026 — 5th Edition

OVERVIEW

The ALICE detector was upgraded for continuous redout for the LHC Run 3 and 4

- **LHC instantaneous Pb–Pb luminosity** in Run 3 with 50ns spacing: **$6x10^{27}$ Hz/cm^2**
  - **Corresponding hadronic interaction rate: 50 kHz (was 8 kHz in Run 1 and 2)**
  - Detector readout upgraded (TPC: MWPC / Run1-2 → **GEM**, ITS: hybrid pixel sensor / Run1-2 → **MAPS**)
  - **Validation of alternating 25/50 ns filling scheme performed in 2025 and 2026 (1h stable beams)**

- **A new ALICE computing model (SW and HW) deployed successfully**
  - **Online pass-0 calibrations**
  - **GPUs computing for online reconstruction and compression** (inspired by the Run 1– 2 High Level Trigger)

- **Computing infrastructure: the ALICE Event Processing Nodes:**
  - **350 nodes – 2800 GPUs farm: operational core of ALICE data taking**
  - **Run 3 integrated lumi: 6.91 nb^{-1} of Pb-Pb collisions. Target for Run 3 + Run 4: 13 nb^{-1}**

- **Run 4 ALICE upgrades and challenges**
  - **New detectors: FoCal** (forward EM+H calorimeter) **and ITS 3** (all silicon inner tracking layers)
  - Higher peak Pb-Pb luminosities possible with 25 ns filling schemes from the LHC injectors
  - **New / refurbished EPN GPU farm**

DATA STRUCTURE

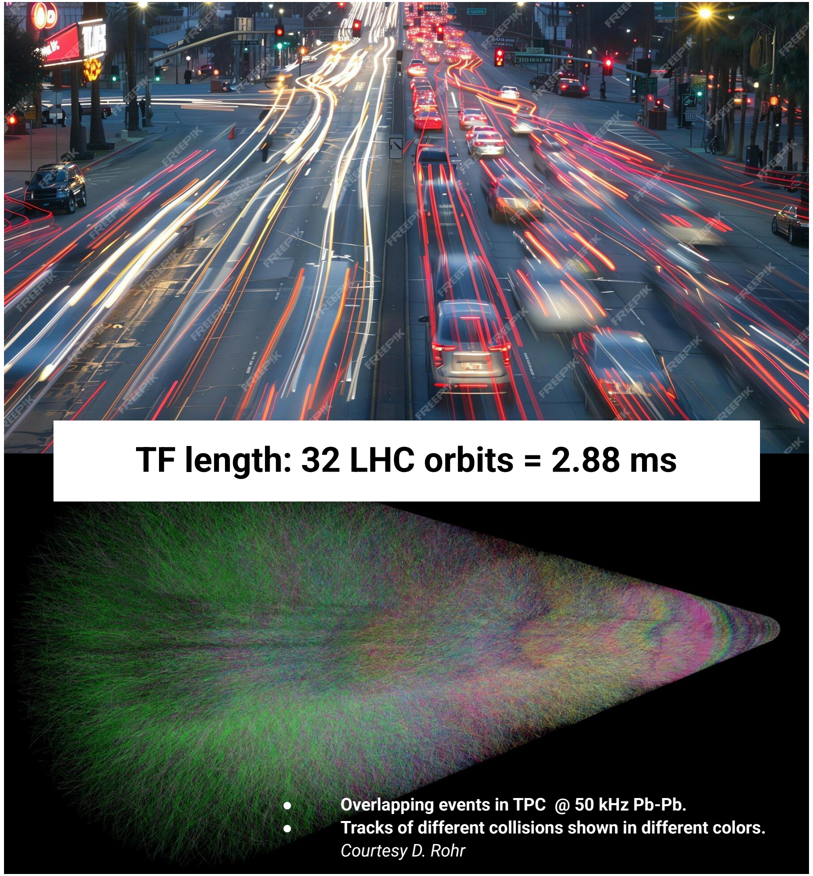
The Time Frame and Data Flow concepts

Time Frame → snapshot of raw data from the readout system
Data Flow → stream of raw TFs sent into the EPN farm

- **The EPN farm performs tracking, calibrations, reconstruction, and compressions synchronously with data taking**
    - TPC tracking is the main consumer of online GPU computing (99%)
    - Calibrations use a few dedicated CPU-only nodes
    - TF compression exploits CPU vector instructions (AVX2) for highly efficient entropy encoding
    - NOTE: **only compressed TFs (CTFs) can be stored, raw TFs are discarded.**

- **Online (sync) and offline (async) reconstruction run exactly the same code.**

- In Run 3 – 4, due to the data rates produced by the continuous readout, **data taking is impossible without the full operational status of the EPN synchronous reconstruction and compression. No "event tagging" mode possible.**

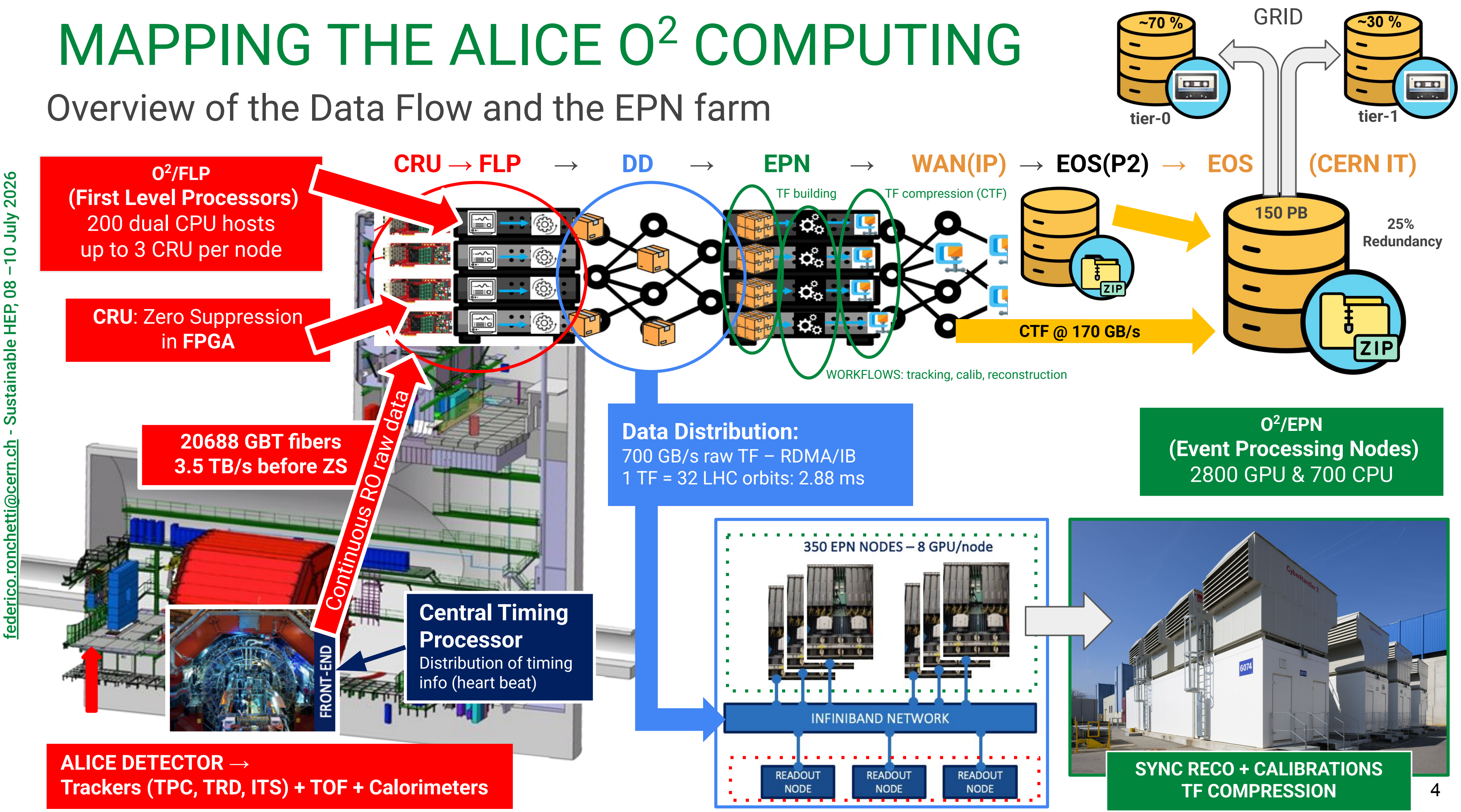


- **Overlapping events in TPC @ 50 kHz Pb-Pb.**
- **Tracks of different collisions shown in different colors.**

Courtesy D. Rohr

MAPPING THE ALICE O^2 COMPUTING

Overview of the Data Flow and the EPN farm

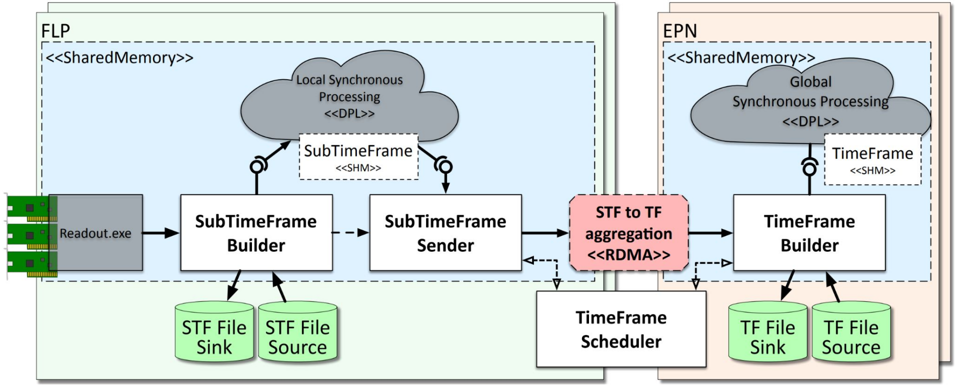


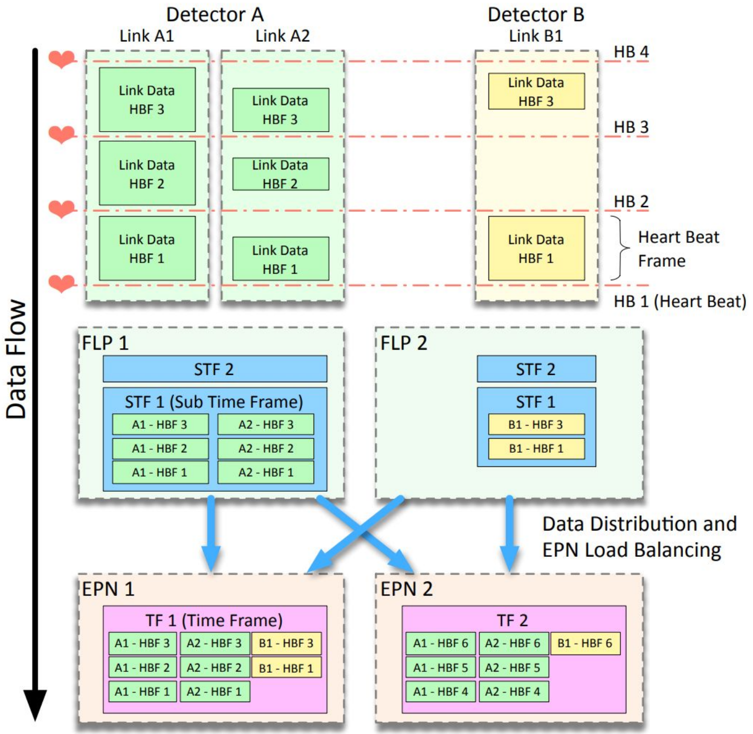
federico.ronchetti@cern.ch - Sustainable HEP, 08 –10 July 2026

DATA FLOW AND EVENT BUILDING

The EPN Data Distribution (DD) system collects, transports, and builds Time Frames

- **Detector data arriving at FLPs is aggregated** into sub Time Frames (sTFs) and **sent to EPNs via RDMA over IB**
- **All sTFs with the same ID go to the same EPN** for full TF assembly (round robin)
- Requires **all-to-all connectivity**
  - FLPs (~200 servers) - EPNs (≤350 servers)



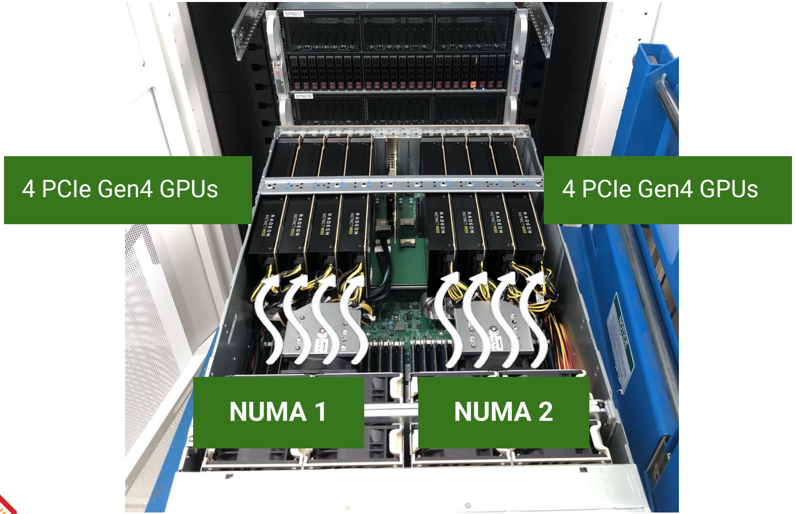


HARDWARE ACCELERATORS

The ALICE EPN farm heavily employs GPUs to speed up processing

1 MI50 GPU replaces:

- **~80 AMD Rome CPU cores** in synchronous processing (99% load on GPU → TPC tracking)
- **~55 AMD Rome CPU cores** in asynchronous processing (80% load on GPU)



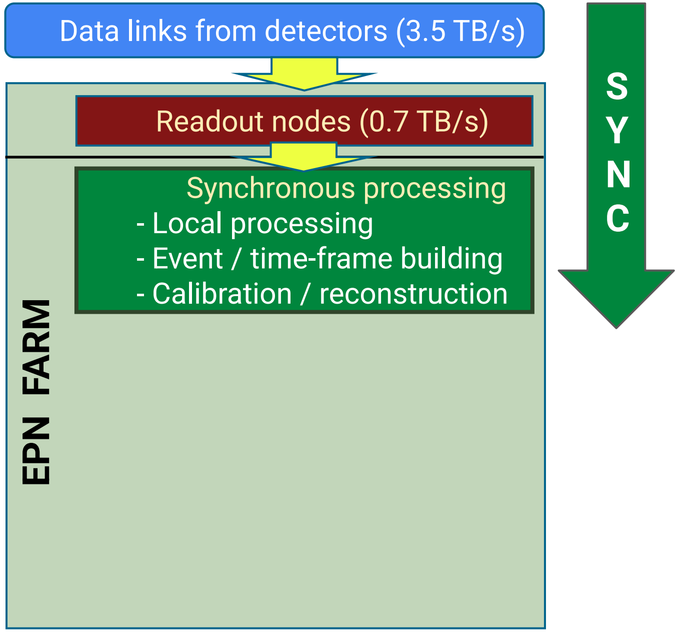
NODES	MI50 gpu/node	MI100 gpu/node	V-FP32 (peak) PFLOP	V-FP32(sust) PFLOP
280	8		30	18
70		8	13	8
350	2800		43	26

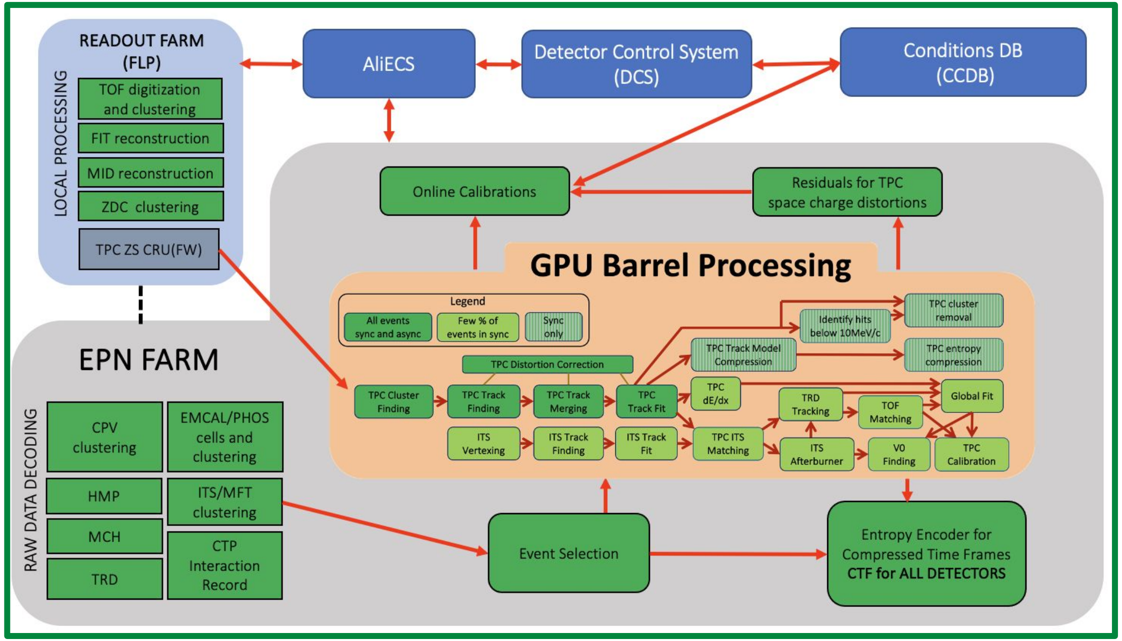
- **More than 2000 64-core servers would be needed for non-GPU online processing**
- Using a **CPU-only farm** would have resulted in **x2 energy consumption**

350+4	70 MI100 EPNs	280 MI50 EPNs	4 Calib Nodes
GPU	8 AMD Instinct™ MI100 32 GB	8 AMD Instinct™ MI50 32 GB	
CPU	2 AMD EPYC™ 7552 48 cores	2 AMD EPYC™ 7452 32 cores	2 AMD EPYC™ 7452 32 cores
MEMORY	1TB DDR4 3200 MHz	512GB DDR4 3200 MHz	512GB DDR4 3200 MHz
Networking	IB 100 Gb/s, ETH 1 Gb/s		

SYNCHRONOUS PROCESSING

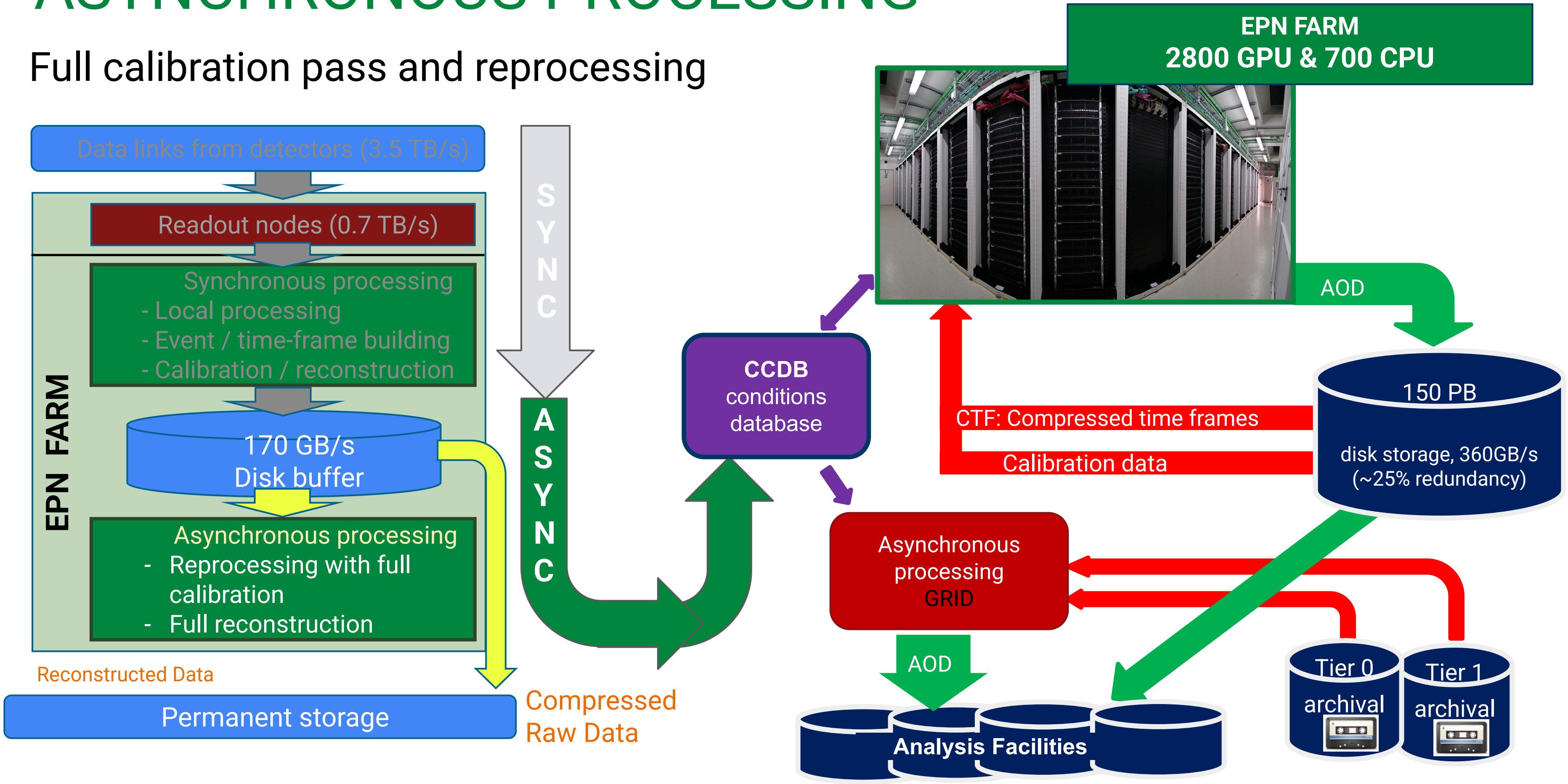
Tracking, calibration, and compression during data taking





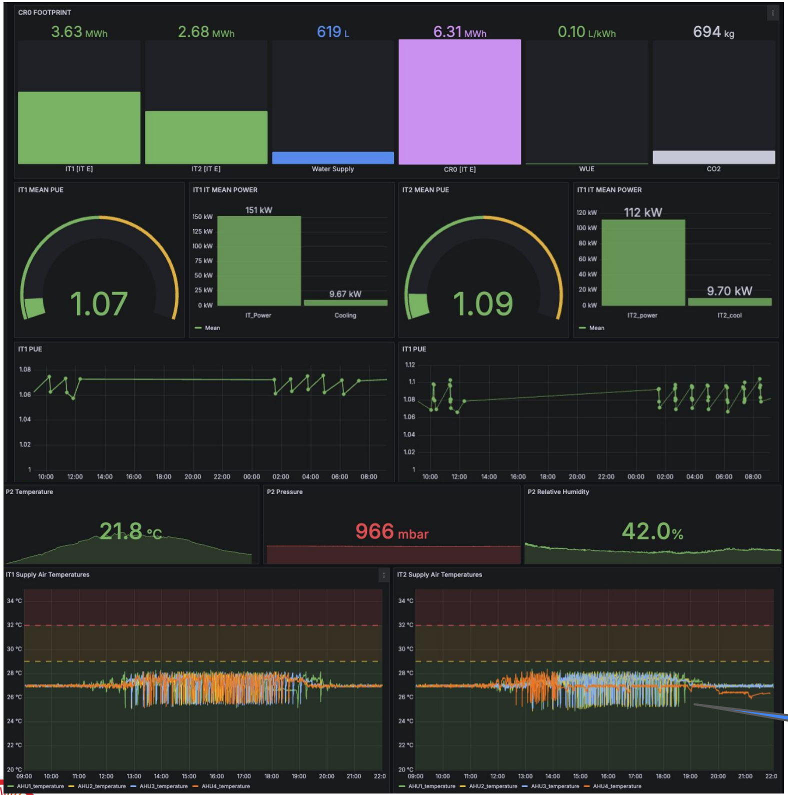
ASYNCHRONOUS PROCESSING

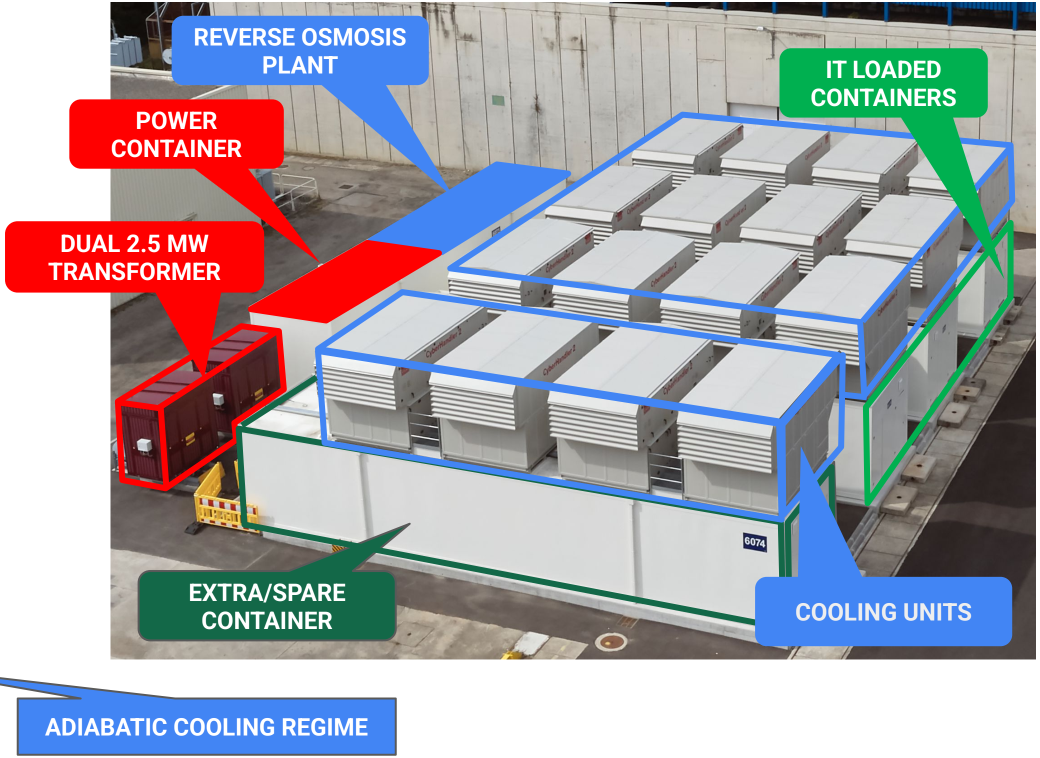
Full calibration pass and reprocessing



EPN IT INFRASTRUCTURE

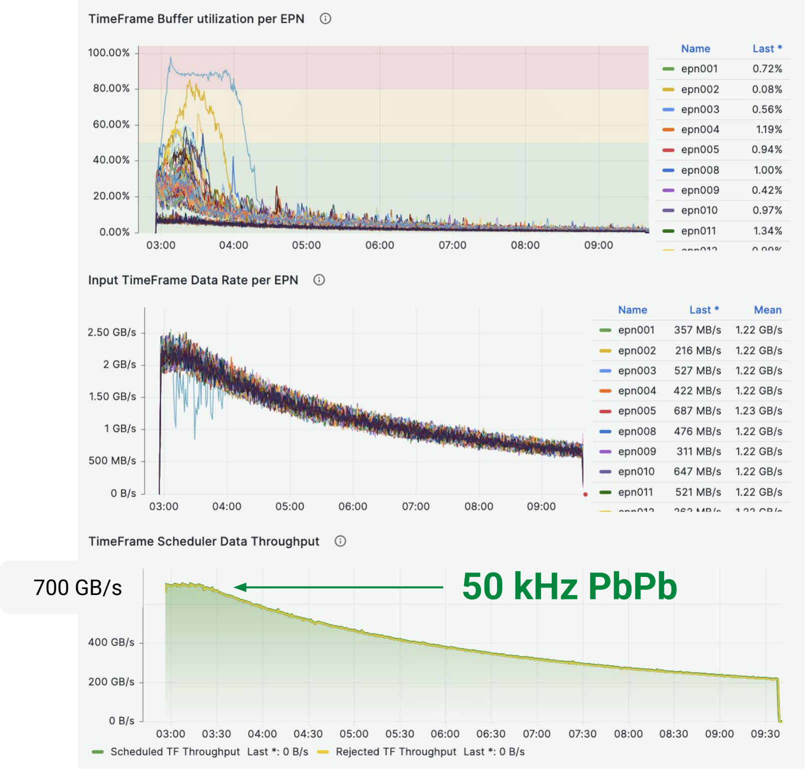
The EPN farm is hosted in an energy-efficient IT infrastructure





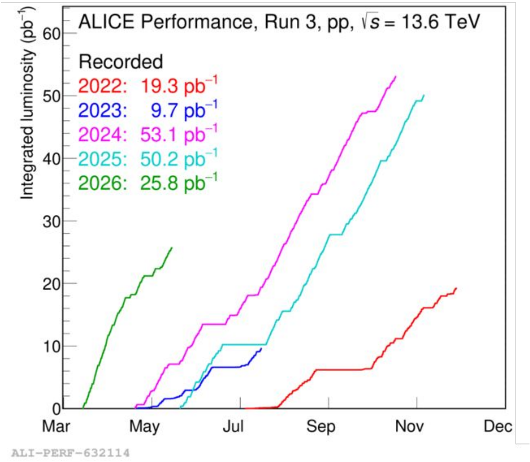
RUN 3 EPN PERFORMANCE

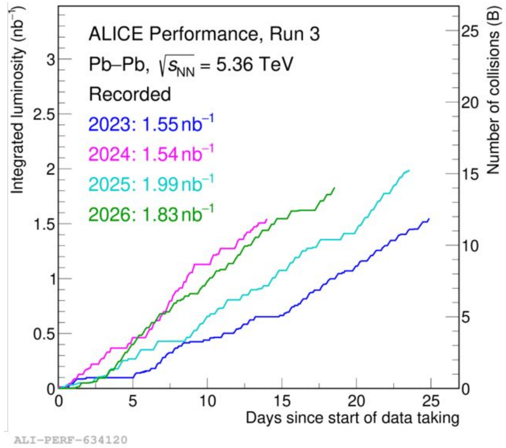
Synchronous processing successful in Run 3 (2022 – 2026)



In 2024–2025 the ALICE instantaneous luminosity was levelled at 50 kHz in Pb–Pb

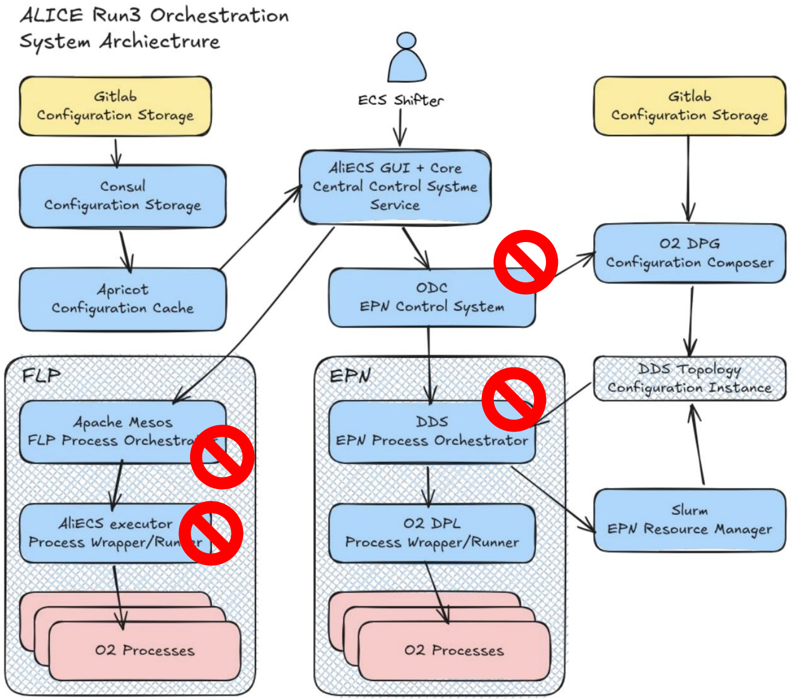
- Compute margin sufficient to cope with the expected rates, >90% of the EPN farm nodes used for online processing
- Average farm power consumption of **400 kW** during Pb-Pb at ~50 kHz

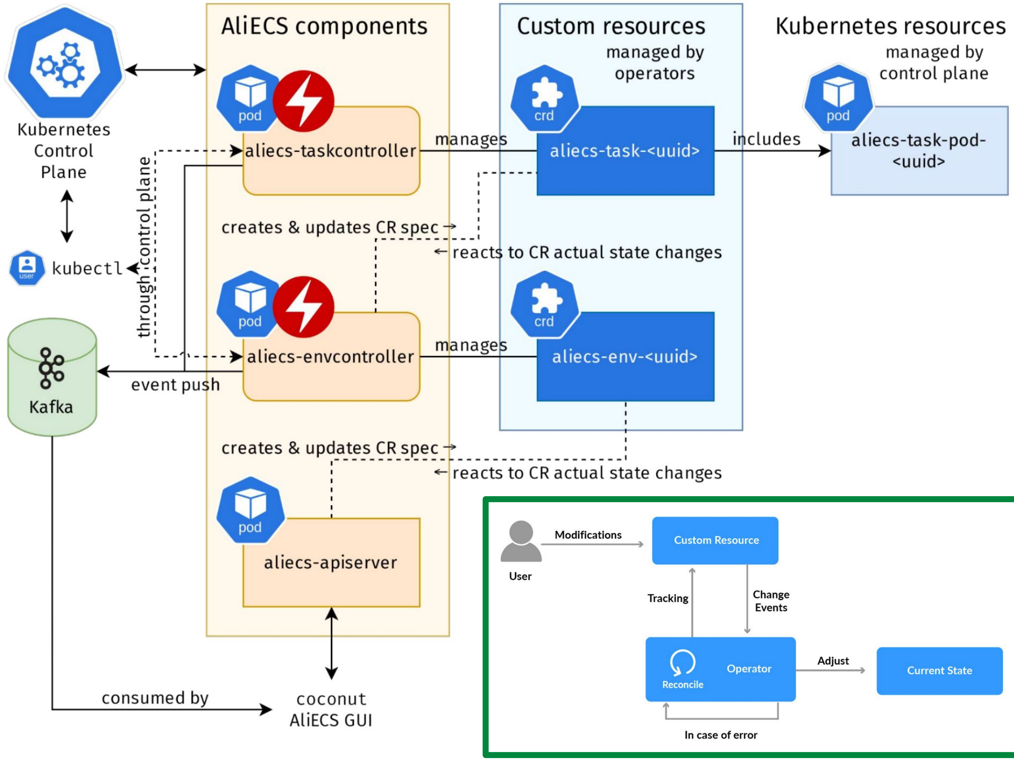




LS3 → RUN 4: ORCHESTRATION SW UPGRADE

Migration of online frameworks to industry standards for better resources exploitation





LS3 → RUN 4: HARDWARE UPGRADE

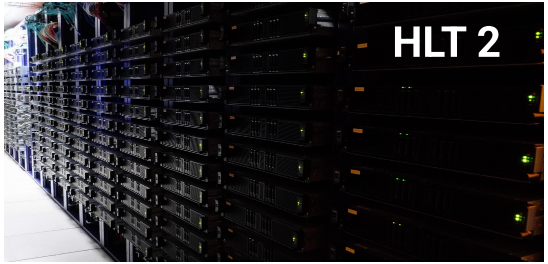
Since 2010 ALICE uses GPUs to optimize capital and operational costs

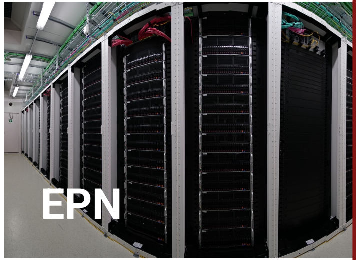
Run 1 – 2010
64 NVIDIA GTX 480
Online TPC tracking

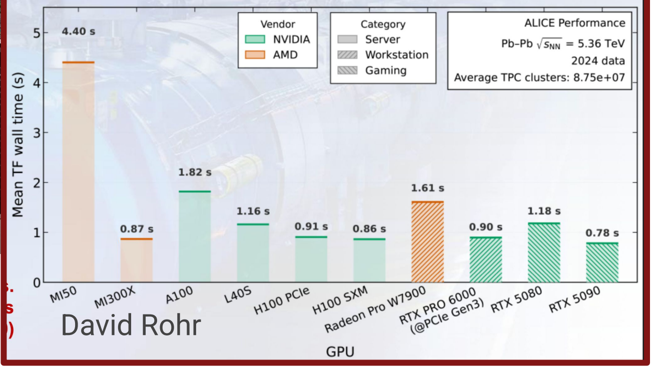
Run 2 – 2015
180 AMD S9000
Online tracking + **compression**

Run 3 – 2022/26
2800 AMD MI50/MI100
Calibration + tracking + compression

Run 4 – 2030/32
NVIDIA or AMD ?
PCIe or converged ?



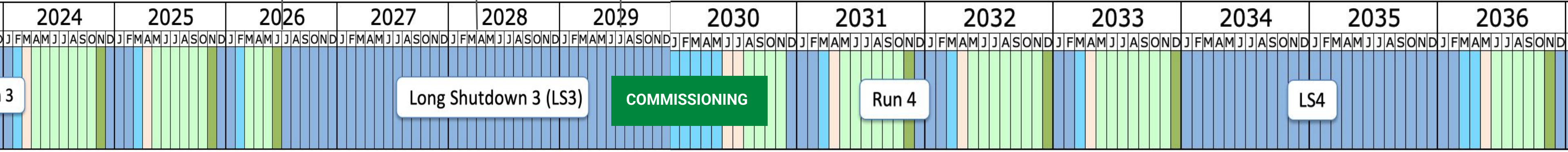




Vertical slice (predecessor)

Vertical slice (reference)

New farm available



SUMMARY & OUTLOOK

ALICE synchronous data reconstruction using GPUs was successful. We'll keep it up…

Summary

- Run 3 was a **major challenge** for ALICE
  - Continuous readout detectors
  - Online calibrations, reconstruction and compression
- **EPN GPU computing is crucial** for the experiment's success in **handling otherwise unmanageable rates,** keeping the overall costs (capital and environmental) under control
- **The EPN farm handled successfully the nominal 50 kHz of hardonic rate of Pb–Pb** (700 GB/s after zero suppression at input)
  - Farm availability was >98%
  - Infrastructure availability was 100% (will have to remain as such during LS3 for async re-processing)
- **Both the EPN farm compute and the hosting IT infrastructure feature high energy efficiency**

Outlook

- **Run 4 will see further upgrades of the ALICE detector** and **LHC accelerator performance**
- The current EPN farm will need upgrade/refurbishment to run until the end of Run 4 (2033+)
- **GPU benchmarks** and **market surveys are** underway to identify the optimal GPU and server candidate for the Run 4 farm

27-02-2025

frontiers | Frontiers in Physics

TYPE Brief Research Report
PUBLISHED 27 February 2025
DOI 10.3389/fphy.2025.1541854

Check for updates

OPEN ACCESS

EDITED BY
J. W. F. Valle,
Spanish National Research Council (CSIC), Spain

REVIEWED BY
Concezio Bozzi,
INFN Sezione di Ferrara, Italy
Jose Francisco Salt Cairols,
Spanish National Research Council (CSIC), Spain

*CORRESPONDENCE
Federico Ronchetti,
federico.ronchetti@cern.ch

RECEIVED 08 December 2024
ACCEPTED 23 January 2025
PUBLISHED 27 February 2025

CITATION
Ronchetti F, Akishina V, Andreassen E, Bluhme N, Dange G, de Cuveland J, Erba G, Gaur H, Hutter D, Kozlov G, Krčál L, La Pointe S, Lehrbach J, Lindenstruth V, Nešković G, Redelbach A, Rohr D, Weiglhofer F and Wilhelmi A (2025) Efficient high performance computing with the ALICE event processing nodes GPU-based farm. *Front. Phys.* 13:1541854. doi: 10.3389/fphy.2025.1541854

Efficient high performance computing with the ALICE event processing nodes GPU-based farm

Federico Ronchetti[1,2]*, Valentina Akishina[3,4], Edvard Andreassen[1], Nora Bluhme[3,4], Gautam Dange[3,4], Jan de Cuveland[3,4], Giada Erba[1], Hari Gaur[3,4], Dirk Hutter[3,4], Grigory Kozlov[3,4], Luboš Krčál[1], Sarah La Pointe[3,4], Johannes Lehrbach[3,4], Volker Lindenstruth[3,4,5], Gvozden Nešković[3,4], Andreas Redelbach[3,4], David Rohr[1], Felix Weiglhofer[3,4] and Alexander Wilhelmi[3,4]

[1]European Organization for Nuclear Research (CERN), Geneva, Switzerland, [2]Istituto Nazionale Fisica Nucleare (INFN), Laboratori Nazionali di Frascati, Frascati, Italy, [3]Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany, [4]Johann Wolfgang Goethe-Universität Frankfurt, Frankfurt am Main, Germany, [5]GSI Helmholtz Centre, Darmstadt, Germany

Due to the increase of data volumes expected for the LHC Run 3 and Run 4, the ALICE Collaboration designed and deployed a new, energy efficient, computing model to run Online and Offline O^2 data processing within a single software framework. The ALICE O^2 Event Processing Nodes (EPN) project performs online data reconstruction using GPUs (Graphic Processing Units) instead of CPUs and applies an efficient, entropy-based, online data compression to cope with Pb–Pb collision data at a 50 kHz hadronic interaction rate. Also, the O^2 EPN farm infrastructure features an energy efficient, environmentally friendly, adiabatic cooling system which allows for operational and capital cost savings.

KEYWORDS
scientific computing, sustainable computing, HTC, HPC, gpu, online data reconstruction and calibration, online data compression, synchronous data processing

1 Introduction

The Large Hadron Collider (LHC) accelerator at CERN returned to full operation on July 5^{th}, 2022 when proton–proton (pp) collisions occurred at a record center-of-mass energy of 13.6 TeV and data taking activities resumed. During the LHC shutdown (2019–2021), the ALICE detector [1] underwent a substantial upgrade [2] providing improved track reconstruction, and an increased interaction rate of up to 50 kHz for lead-lead (Pb-Pb) collisions in continuous readout mode. These advancements facilitated the collection of a pp data sample during the first year of Run 3 (2022), which is already ten times larger than the combined data samples from Run 1 (2010–2013) and Run 2 (2015–2018).

THANK YOU

09-10/2023

ALICE collab.

New nodes *The event processing node racks in the ALICE computing farm, part of a completely new computing model for Run 3 and beyond.*

ALICE UPS ITS GAME FOR SUSTAINABLE COMPUTING

The design and deployment of a completely new computing model – the O^2 project – allows the ALICE collaboration to merge online and offline data processing into a single software framework to cope with the demands of Run 3 and beyond. Volker Lindenstruth goes behind the scenes.

The Large Hadron Collider (LHC) roared back to life on 5 July 2022, when proton–proton collisions at a record centre-of-mass energy of 13.6 TeV resumed for Run 3. To enable the ALICE collaboration to benefit from the increased instantaneous luminosity of this and future LHC runs, the ALICE experiment underwent a major upgrade during Long Shutdown 2 (2019–2022) that will substantially improve track reconstruction in terms of spatial precision and tracking efficiency, in particular for low-momentum particles. The upgrade will also enable an increased interaction rate of up to 50 kHz for lead–lead (PbPb) collisions in continuous readout mode, which will allow ALICE to collect a data sample more than 10 times larger than the combined Run 1 and Run 2 samples.

ALICE is a unique experiment at the LHC devoted to the study of extreme nuclear matter. It comprises a central barrel (the largest data producer) and a forward muon "arm". The central barrel relies mainly on four subdetectors for particle tracking: the new inner tracking system (ITS), which is a seven-layer, 12.5 gigapixel monolithic silicon tracker (*CERN Courier* July/August 2021 p29); an upgraded time projection chamber (TPC) with GEM-based readout for continuous operation; a transition radiation detector; and a time-of-flight detector. The muon arm is composed of three tracking devices: a newly installed muon forward tracker (a silicon tracker based on monolithic active pixel sensors), revamped muon chambers and a muon identifier.

THE AUTHOR
Volker Lindenstruth *Goethe University Frankfurt, GSI Helmholtz Centre and Frankfurt Institute for Advanced Studies, on behalf of the ALICE Collaboration.*

Due to the increased data volume in the upgraded ALICE detector, storing all the raw data produced during Run 3 is impossible. One of the major ALICE upgrades in preparation for the latest run was therefore the design and deployment of a completely new computing model: the O^2 project, which merges online (synchronous) and offline (asynchronous) data processing into a single software framework. In addition to an upgrade of the

2026: Online data processing with the EPN farm

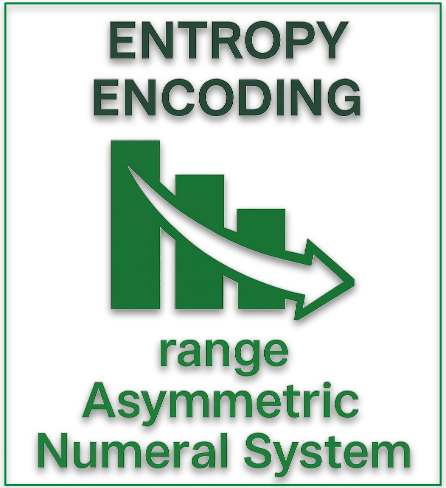
Full EPN technical paper in preparation

https://cerncourier.com/wp-content/uploads/2023/09/CERNCourier2023SepOct-digitaledition.pdf

Backup

OPTIMIZED REAL-TIME DATA REDUCTION

GPU TF compression is crucial for ALICE data taking



Data Structure and Compression

- Detector raw data reduced independently to a flat structure of integer arrays using custom lossy/lossless methods
- Output is further processed via **rANS-based entropy coding**

rANS efficiency and performance

- rANS outperforms Huffman (3% smaller) and zlib/gzip (up to 15% smaller)
- Approaches entropy limit (2–3× compression ratio) by modeling skewed 32-bit symbol distributions
- AVX2 vectorized implementation with up to 16 parallel encoders
- Achieves 3200 MB/s compression throughput, 2× faster than state-of-the-art CPU methods
- Exploits dynamic symbol distribution computation per TF for optimal compression

Other strategies

- event skimming: retains only ~3% to 4.5% of the original CTFs for offline processing

ALICE IN RUN 3

The ALICE detector underwent a major upgrade for Run 3

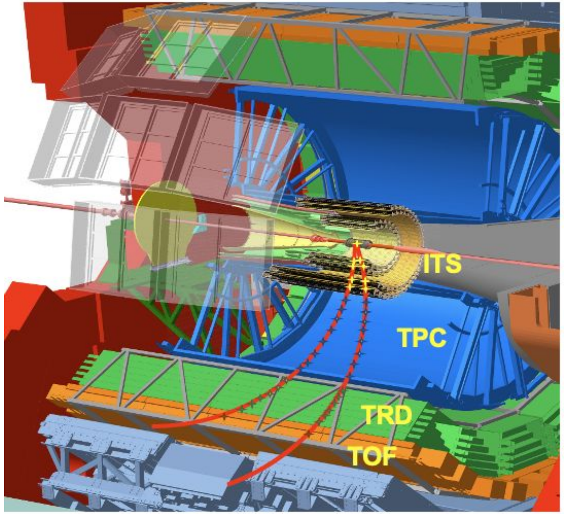
- In Run 3 (and Run 4) the LHC delivers an instantaneous luminosity of $6x10^{27}$ Hz/cm^2 in Pb-Pb
  - **The corresponding hadronic interaction rate is now 50 kHz (was 8 kHz in Run 1 and 2)**
  - Thanks to a now 50 ns filling schema (slip-stacking) such high luminosities can be sustained for some hours

To cope with such rates the **ALICE detector underwent a major upgrade** during LS2 (2019-21)

Detectors used for barrel tracking:

- **7 layers ITS**
  Inner Tracking System
  MAPS silicon tracker
- **152 pad rows TPC**
  Time Projection Chamber
- **6 layers TRD**
  Transition Radiation Detector
- **1 layer TOF**
  Time Of Flight Detector

UPGRADED ALICE DETECTOR



- The TPC **MWPC** chambers where **replaced by GEMs**
- The ITS was completely replaced by **7 layers of MAPS sensor for a total resolution of 12.5 Gigapixel**
- The triggered readout was abandoned in favour of a **continuous readout system**

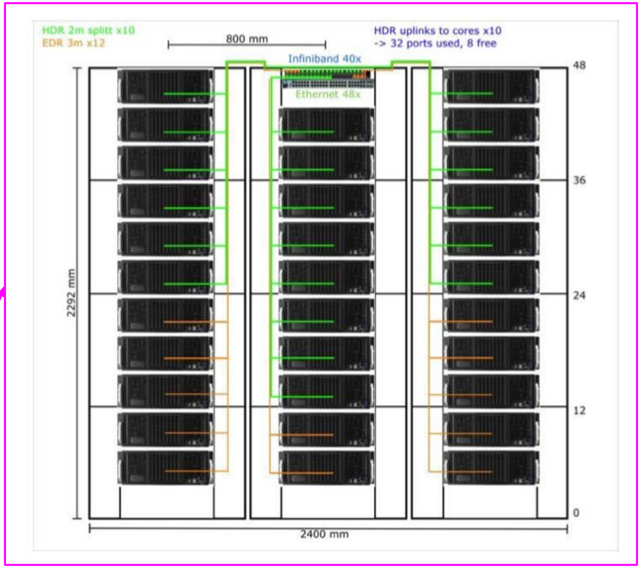
- As a consequence, **the ALICE computing facilities and software model were also upgraded to match the increase in detector performance and the much higher data rates.**

THE ALICE EPN FARM

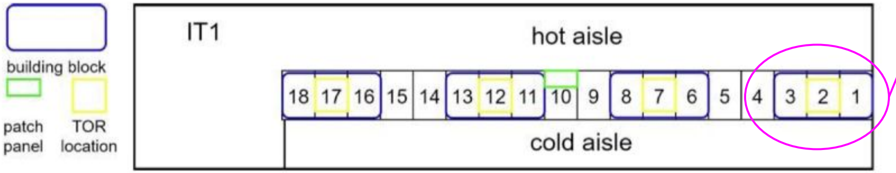
Farm layout

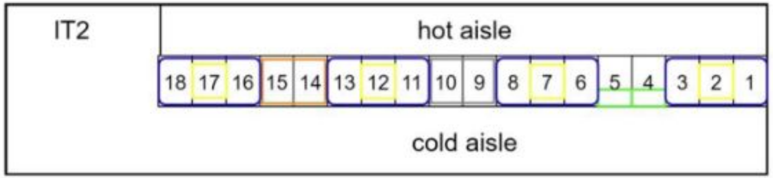
EPN Building Block: 35 worker nodes into 3 adiacent racks connected to a single IB switch (TOR, Top Of the Rack)

- Total of **10** building blocks across 3 IT containers
- All building blocks saturated with 35 worker nodes:
  - 28 “MI50” and 7 “MI100”



Example of Building Block layouts in the most dense containers:



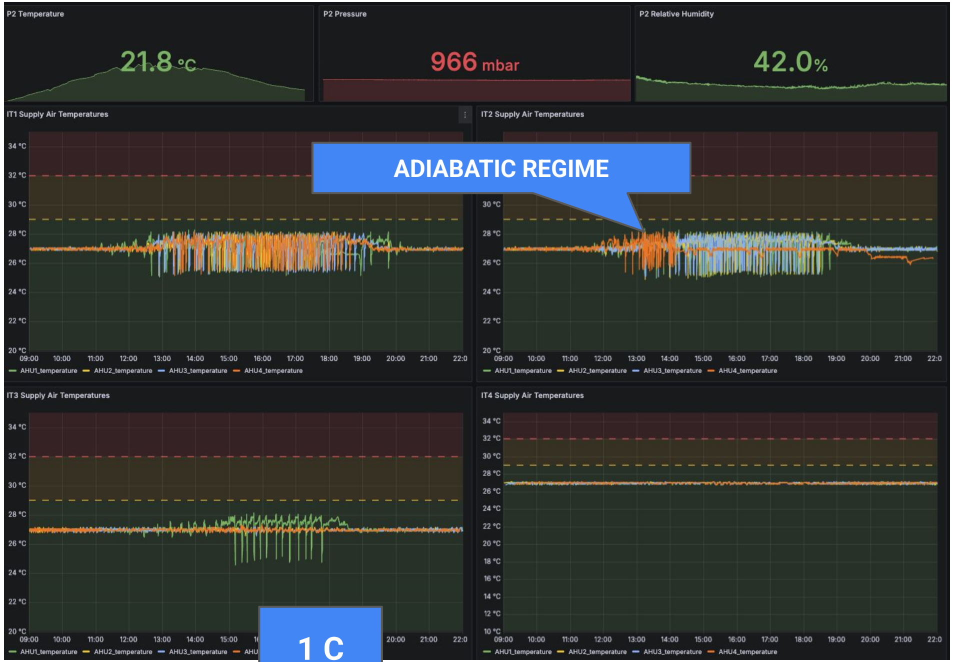


Credit: J. Lehrbach

ALICE EPN LOW-IMPACT IT INFRASTRUCTURE

The EPN farm is hosted within a sustainable, energy-efficient IT infrastructure

- **Air-to-air adiabatically-enhanced cooling →** vaporizes purified water on heat exchangers
- Exploits **free-cooling with favorable conditions**
- **Self-produced purified water** by reverse osmosis (no chemicals)
- **Modularized low-cost containers →** allow for easy extensions of the farm



Adiabatic cooling is more efficient than pure mechanical ventilation and has a lower energy, carbon, and water consumption footprint

27 C

1 C

`setpoint + delta - hysteresis`

pump feedback (depends on IT load)

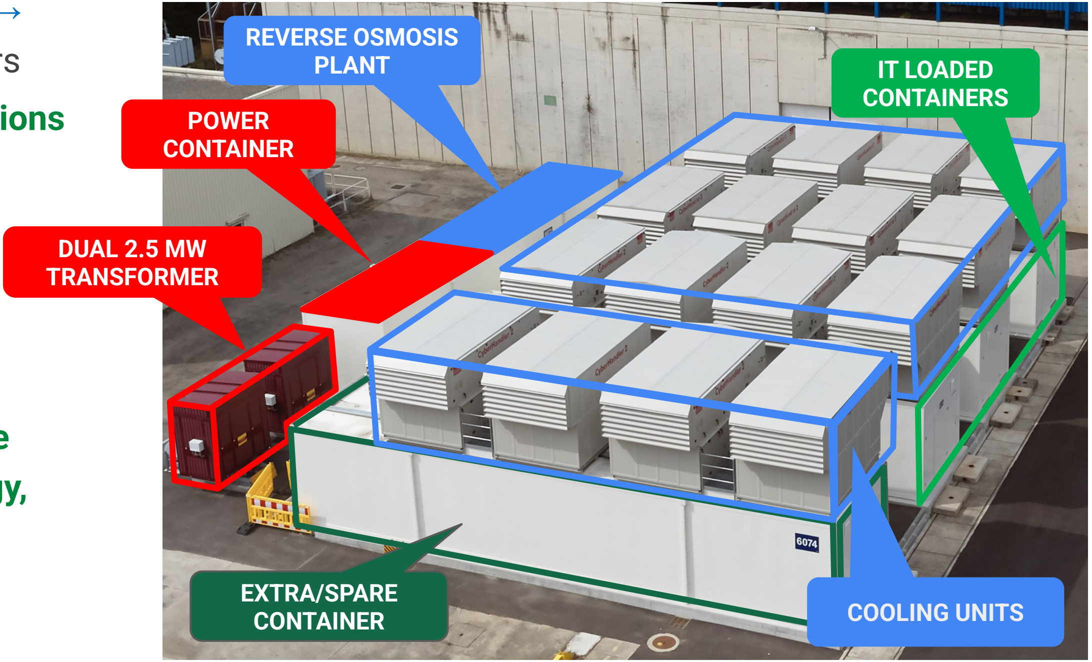
IT infrastructure is operated by the EPN team

- support from the ALICE Technical Coordination
- Preventive and second level maintenance from CERN and external contractors

ALICE EPN LOW-IMPACT IT INFRASTRUCTURE

The EPN farm is hosted within a sustainable, energy-efficient IT infrastructure

- **Air-to-air adiabatically-enhanced cooling →** vaporizes purified water on heat exchangers
- Exploits **free-cooling with favorable conditions**
- **Self-produced purified water** by reverse osmosis (no chemicals)
- **Modularized low-cost containers →** allow for easy extensions of the farm



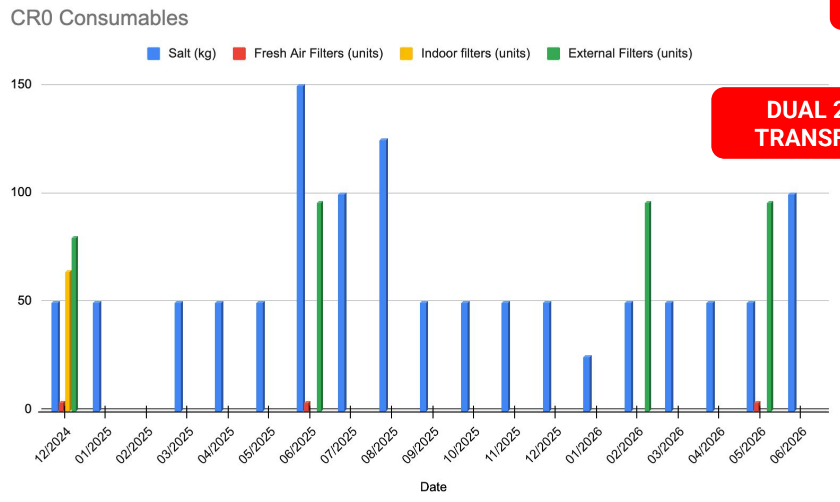
Adiabatic cooling is more efficient than pure mechanical ventilation and has a lower energy, carbon, and water consumption footprint

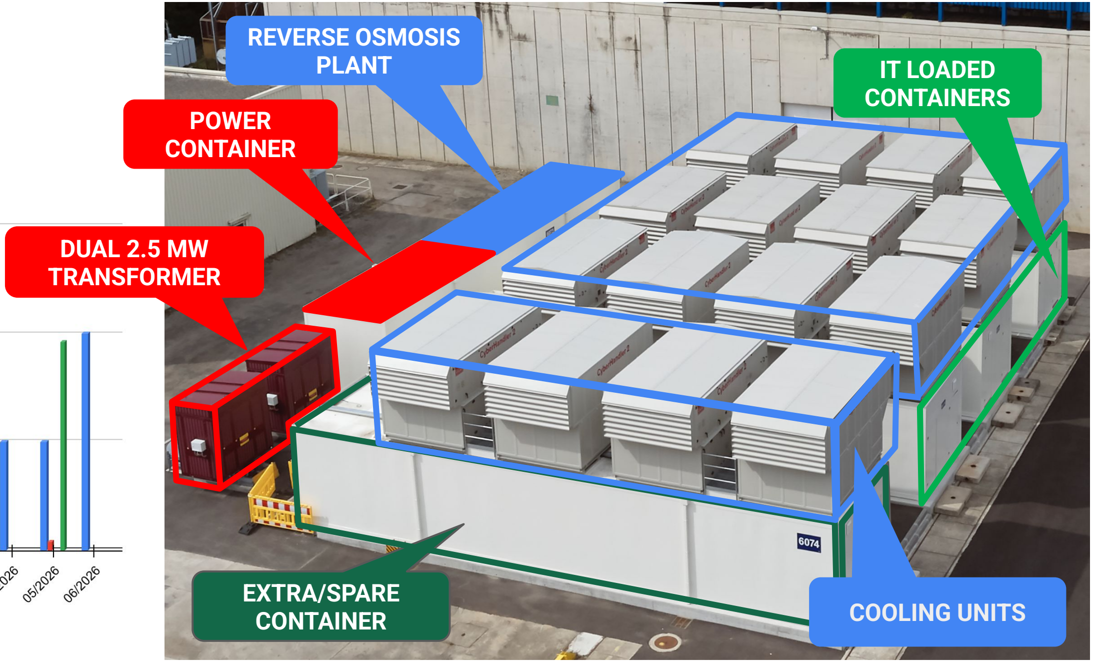
IT infrastructure is operated by the EPN team

- support from the ALICE Technical Coordination
- Preventive and second level maintenance from CERN and external contractors

ALICE EPN LOW-IMPACT IT INFRASTRUCTURE

The EPN farm is hosted within a sustainable, energy-efficient IT infrastructure

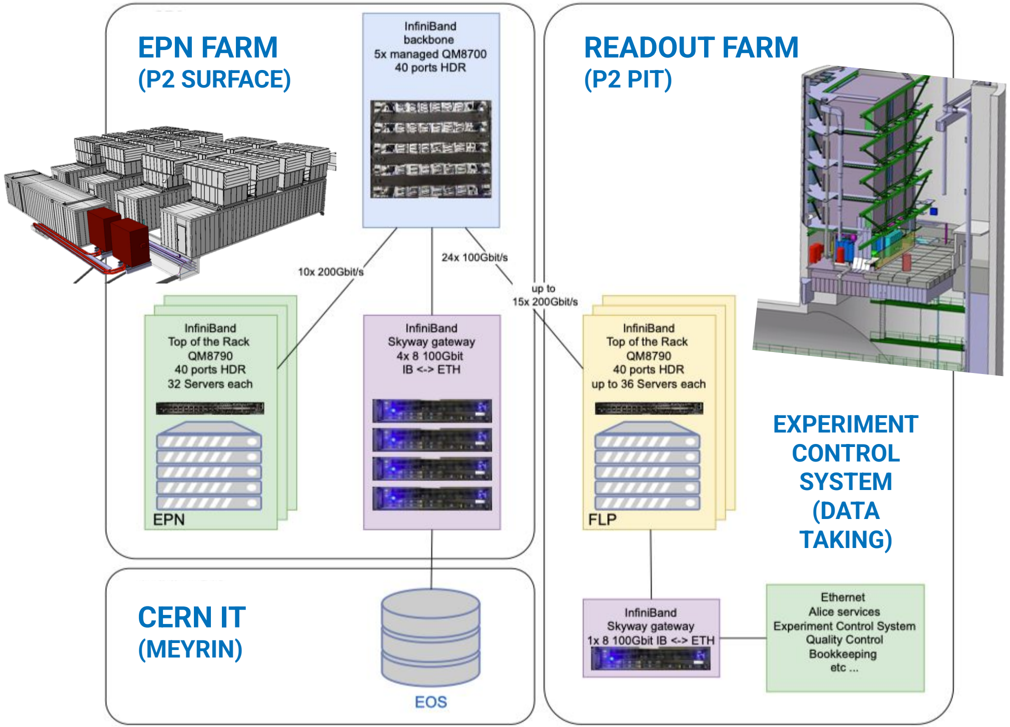




THE EPN NETWORK TOPOLOGY

The backbone of the EPN farm is based on EDR/HDR Infiniband

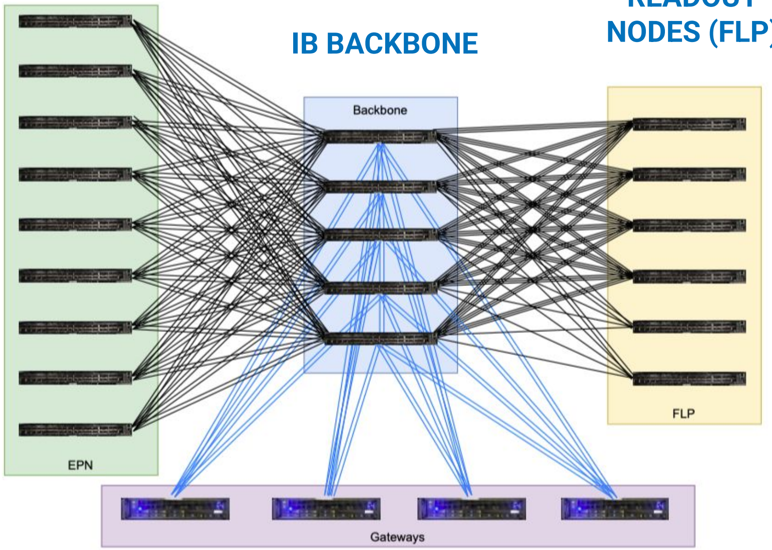
Credit: J. Lehrbach



EPN BUILDING BLOCKS (IB TOR)

IB BACKBONE

DETECTOR READOUT NODES (FLP)



IB TO ETHERNET GATEWAYS (to CERN IT mass storage)