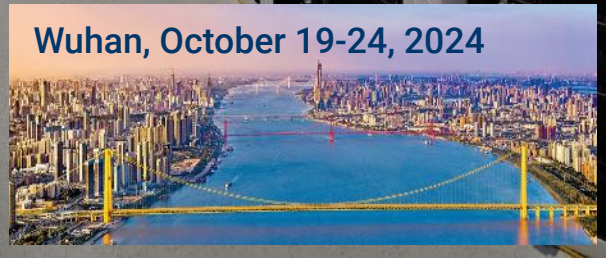
Efficient Computing with the ALICE Event Processing Nodes GPU farm

Federico Ronchetti
federico.ronchetti@cern.ch

On behalf of the ALICE Collaboration and the EPN team



Workshop on Advances, Innovations,
and Prospects in High-Energy Nuclear Physics

ALICE IN RUN 3

The ALICE detector underwent a major upgrade for Run 3

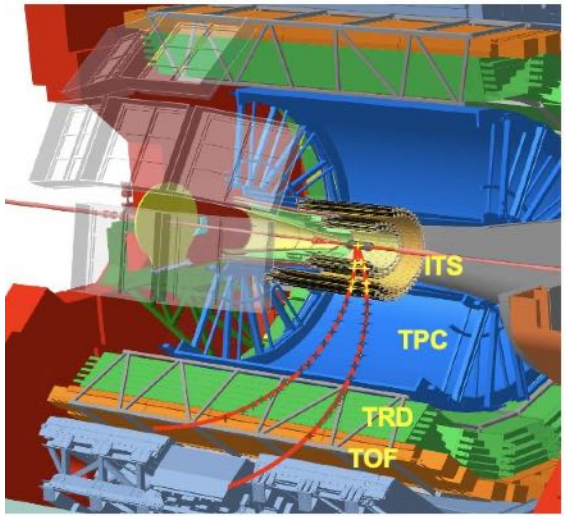
- In Run 3 (and Run 4) the LHC delivers an instantaneous luminosity of $6x10^{27}$ Hz/cm^2 in Pb-Pb
  - **The corresponding hadronic interaction rate is now 50 kHz (was 8 kHz in Run 1 and 2)**
  - Thanks to a now 50 ns filling schema (slip-stacking) such high luminosities can be sustained for some hours

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**

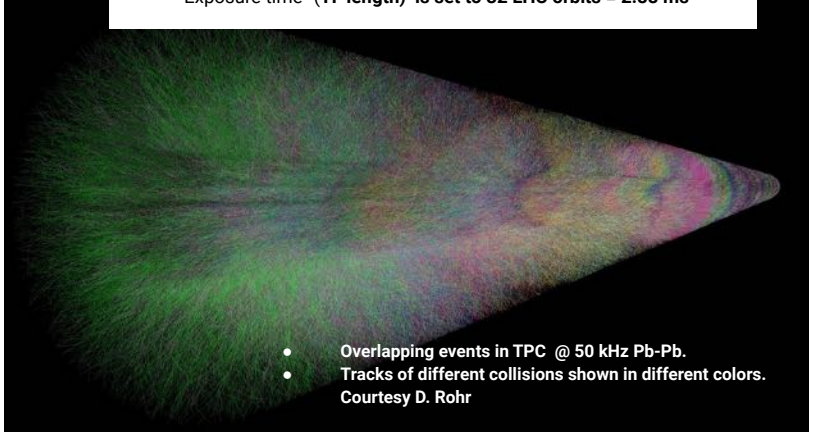
- As a consequence, **the ALICE computing facilities and software model were also upgraded to match the increase in detector performance and the much higher data rates.**

THE RUN 3 ALICE COMPUTING

The Time Frame concept

- All collisions stored for main detectors → **no trigger**
  - **no "collision event" but a collection of tracks in a given time windows**
  - **Similar to a "long exposure photograph"**
- **Continuous readout** → data in drift detectors overlap
- The "exposure time" or **TF length is selectable** and chosen to be a multiple of the LHC orbit time (90 us)
  - **32 orbits = 2.88 ms**
- **100x more collisions → in 2023 collected 1.5 nb^{-1} of Pb-Pb MinBias (as much as Run 1 + Run 2)**
- **The raw TF stream from readout is the input of the EPN GPUs** where tracking and reconstruction and compressions are performed in sync with the data taking (online calibration runs on dedicated CPUs).
- Cannot store all raw data → Storing COMPRESSED TF (CTF)
  - **No EPN online reconstruction and compression, no data daking.**
- **90% of synchronous reconstruction runs on GPUs**
- Online (**sync**) and offline (**async**) reconstruction **use exactly the same code**
  - For best efficiency async components not needed by sync are deactivated in the data taking phase

"Exposure time" (**TF length**) **is set to 32 LHC orbits = 2.88 ms**

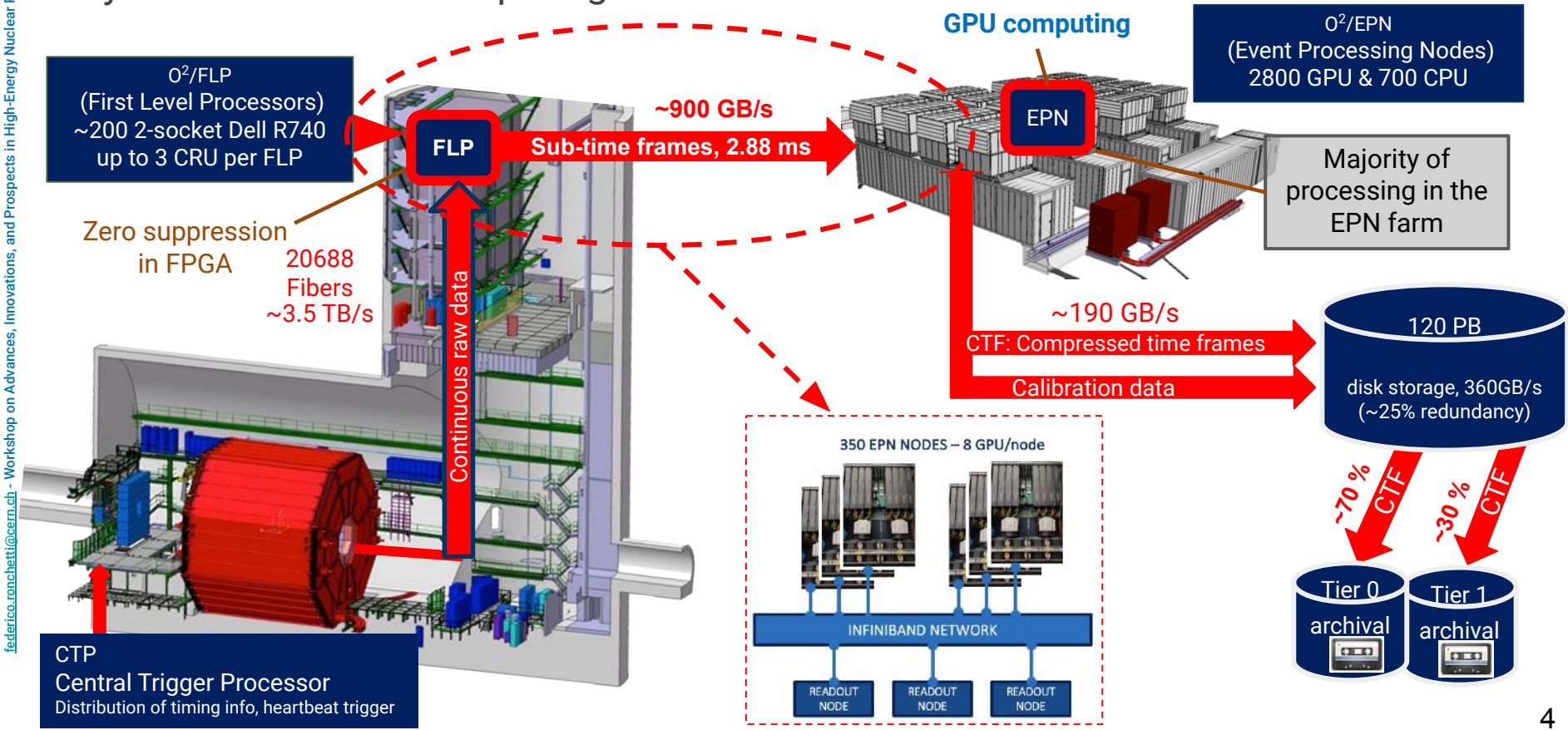


- **Overlapping events in TPC @ 50 kHz Pb-Pb.**
- **Tracks of different collisions shown in different colors.**

Courtesy D. Rohr

THE ALICE RUN 3 DATA FLOW

Layout of the ALICE computing at the LHC P2



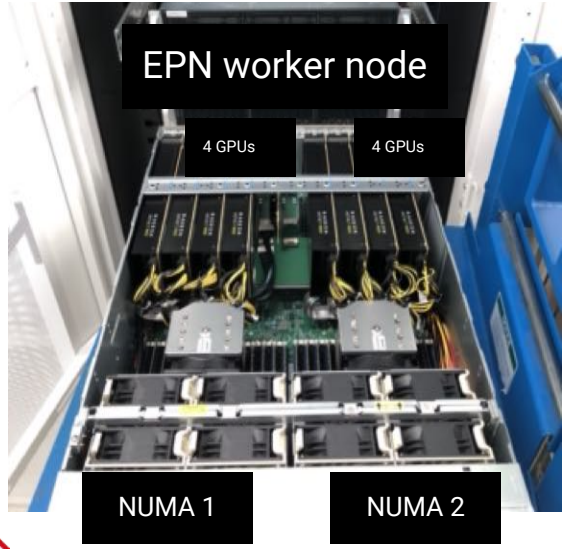
federico.ronchetti@cern.ch - Workshop on Advances, Innovations, and Prospects in High-Energy Nuclear Physics

THE ALICE EPN FARM

Hardware specifications

- **280 nodes** equipped with 8 AMD MI50 32GB GPUs
- Additional 70 nodes (installed in 2023) equipped MI100 32GB
- Grand total of 350 (280 MI50 + 70 MI100) nodes and 2800 GPUs (equivalent to ~373 MI50 nodes at MI100 = 4/3 MI50)

	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
Networks	IB 100 Gb/s, ETH 1 Gb/s		



PERFORMANCE	MI-100
Compute Units	120
Stream Processors	7,680
Peak BFLOAT16	Up to 92.3 TFLOPS
Peak INT4 | INT8	Up to 184.6 TOPS
Peak FP16	Up to 184.6 TFLOPS
Peak FP32 Matrix	Up to 46.1 TFLOPS
Peak FP32	Up to 23.1 TFLOPS
Peak FP64	Up to 11.5 TFLOPS
Bus Interface	PCIe® Gen 3 and Gen 4 Support³

PERFORMANCE	MI-50
Compute Units	60
Stream Processors	3,840
Peak INT8	Up to 53.6 TOPS
Peak FP16	Up to 26.5 TFLOPS
Peak FP32	Up to 13.3 TFLOPS
Peak FP64	Up to 6.6 TFLOPS
Bus Interface	PCIe®Gen 3 and Gen 4 Supported²

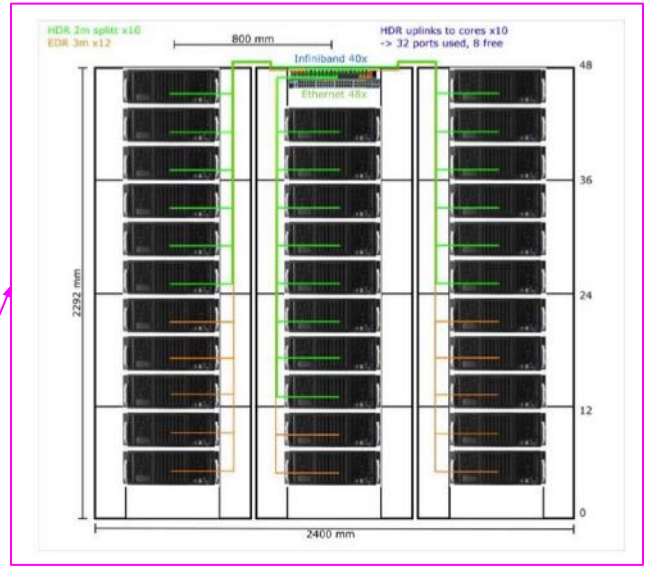
NODES	GPU FP32	TFLOPS (FP32)
280	13.3	3724
70	23.1	1617
350		5341

THE ALICE EPN FARM

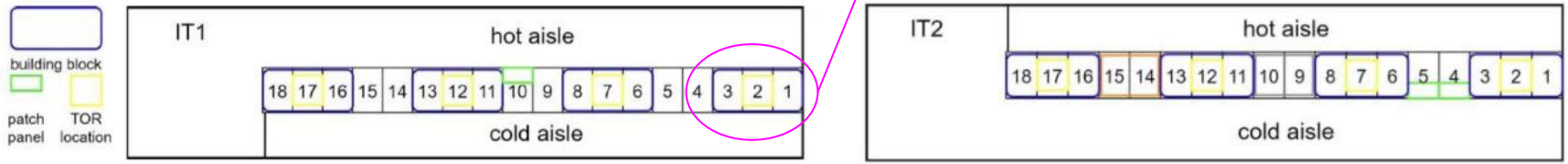
Farm layout

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:

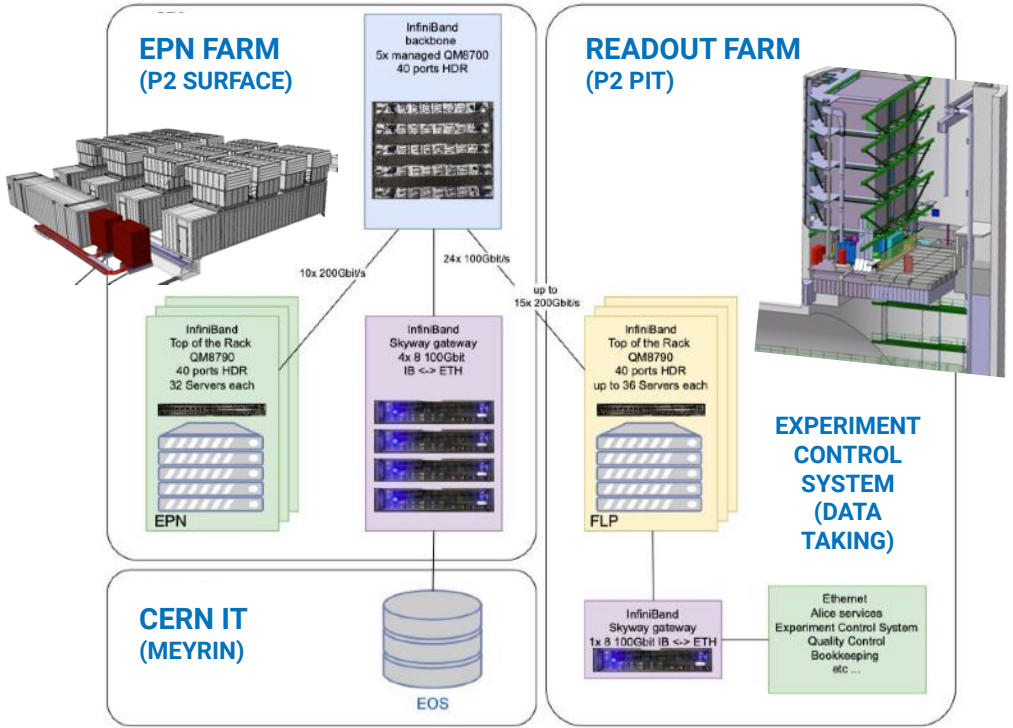


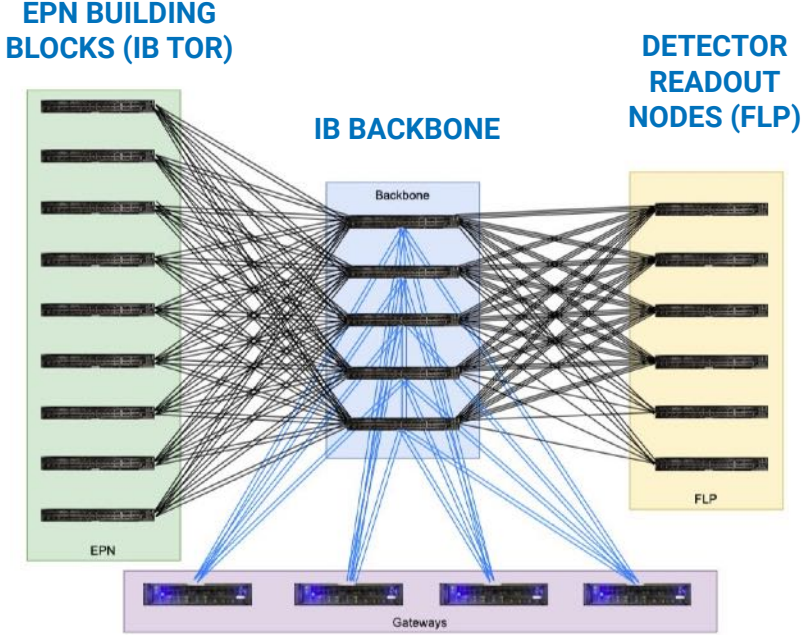
Credit: J. Lehrbach

THE EPN NETWORK TOPOLOGY

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

Credit: J. Lehrbach





THE PUMPING HEART OF THE EPN FARM

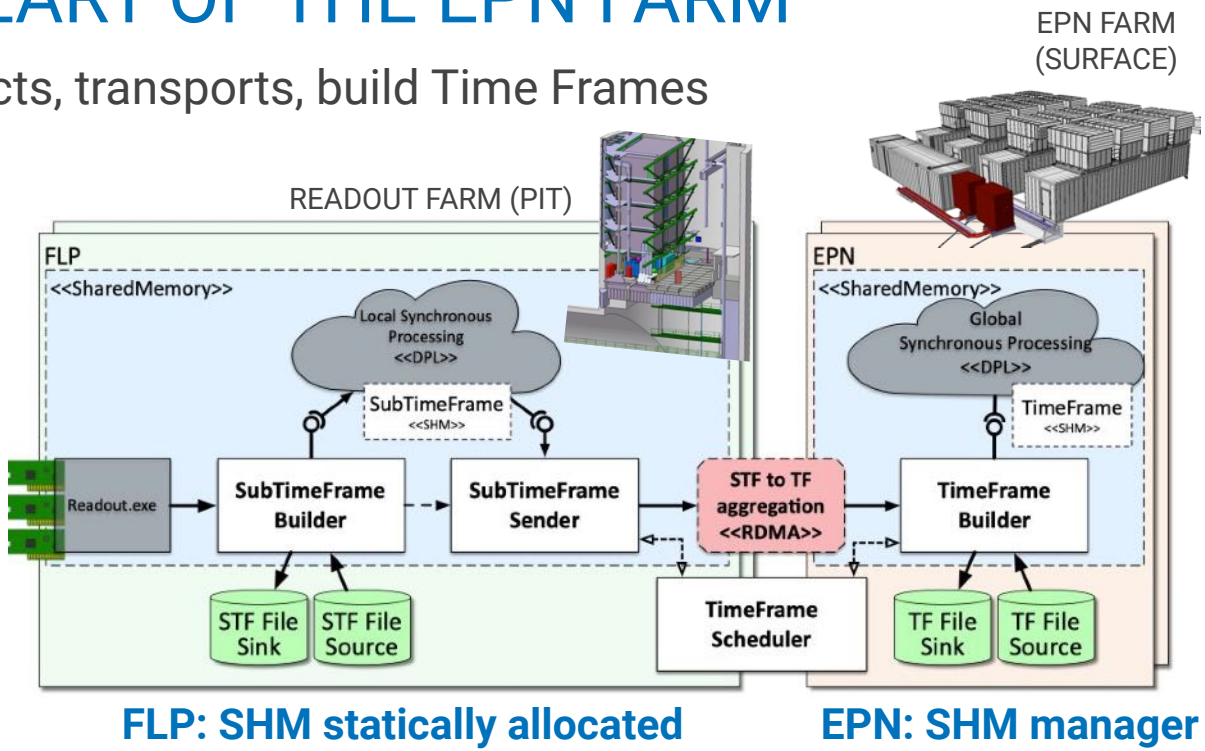
Data Distribution (DD) collects, transports, build Time Frames

Efficient transport

- Shared Memory (SHM) and RDMA transports

Balanced resource utilization

- EPNs nodes process multiple TFs in parallel
- DD selects a least utilized node to receive the new TF
- Support adding and removing EPN worker nodes

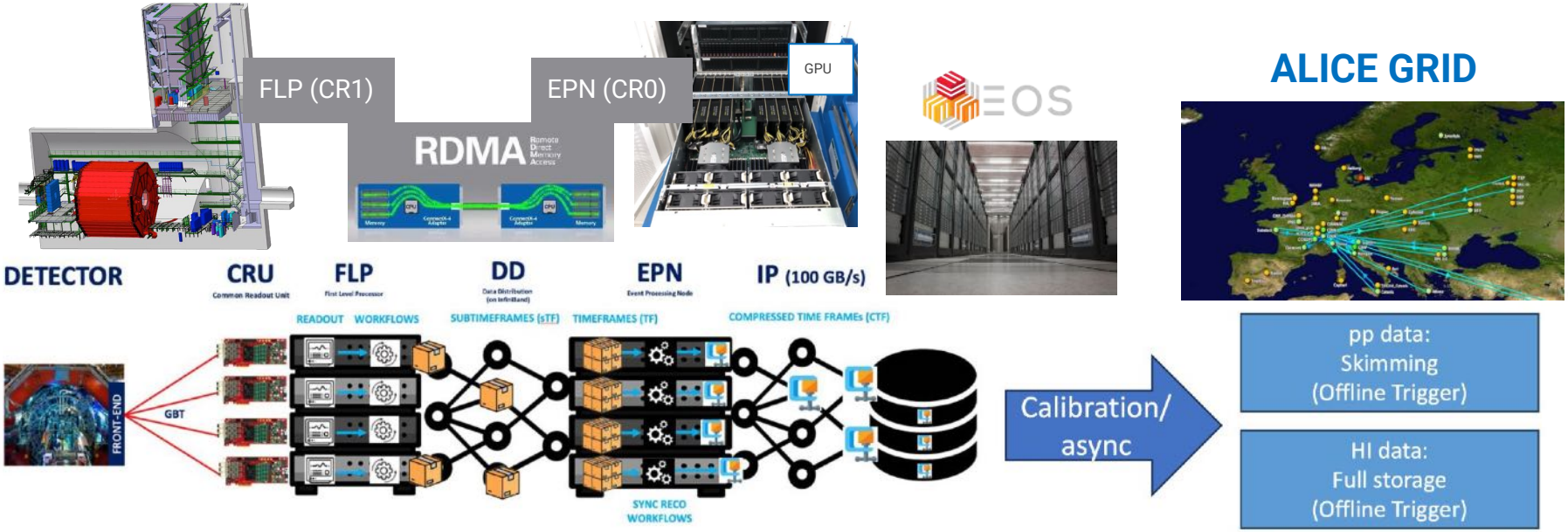


Application-level traffic shaping

- Spread TF transfers evenly over EPN TOR switches
  - Avoid IB-core to EPN-TOR congestion
- EPNs pull data from FLPs on different TORs
  - Avoid FLP-TOR to IB-core congestion

DATA DISTRIBUTION

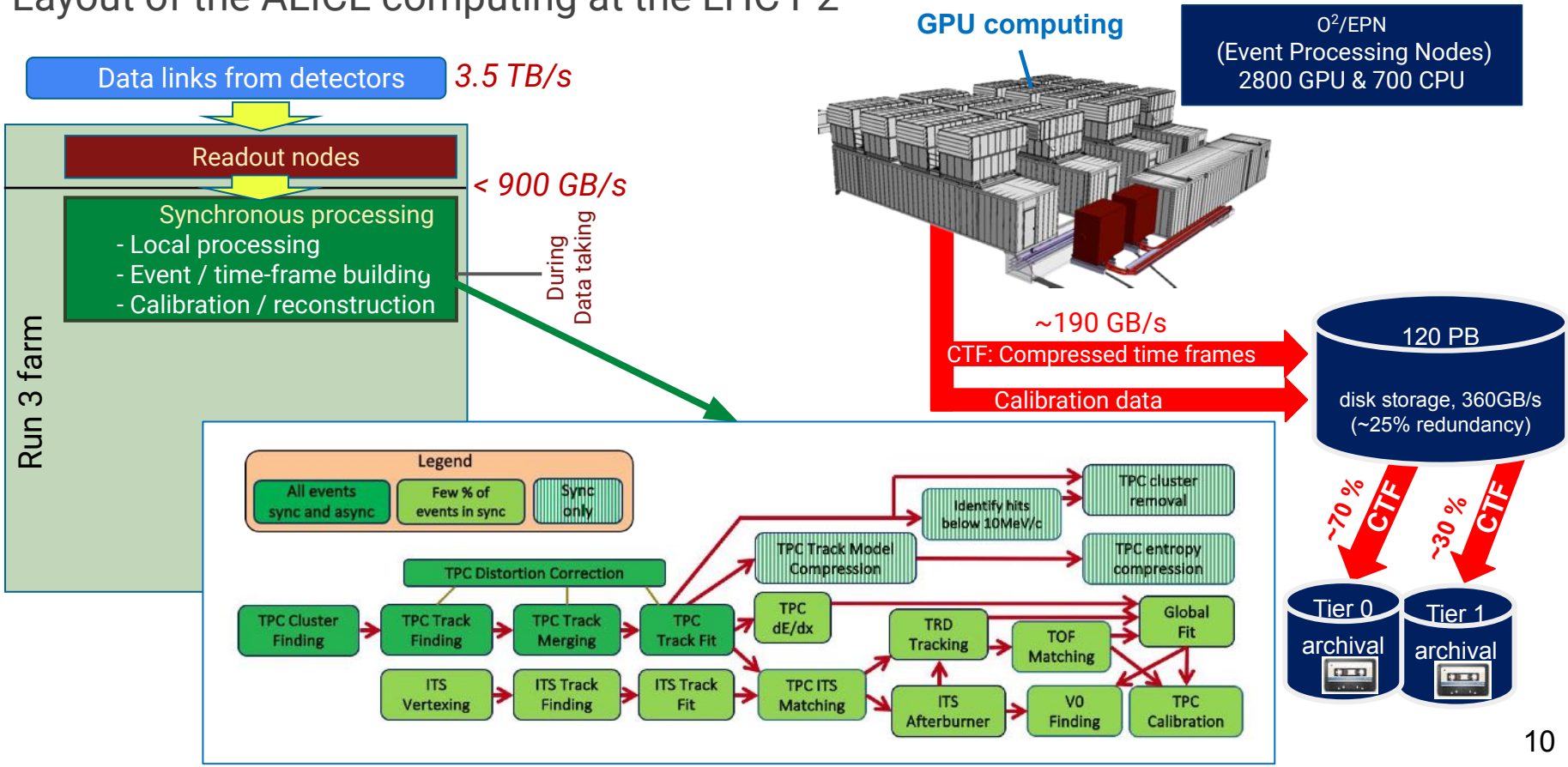
The EPN Data Distribution (DD) collects, transports, and build Time Frames



Data rate	3.51 TB/s	800-900 GB/s (PbPb) 100-200 GB/s (pp)	100-200 GB/s (PbPb) 30 GB/s (pp)	~2 PB per day
	Non Zero Suppressed	Raw Time Frames	Compressed TF	

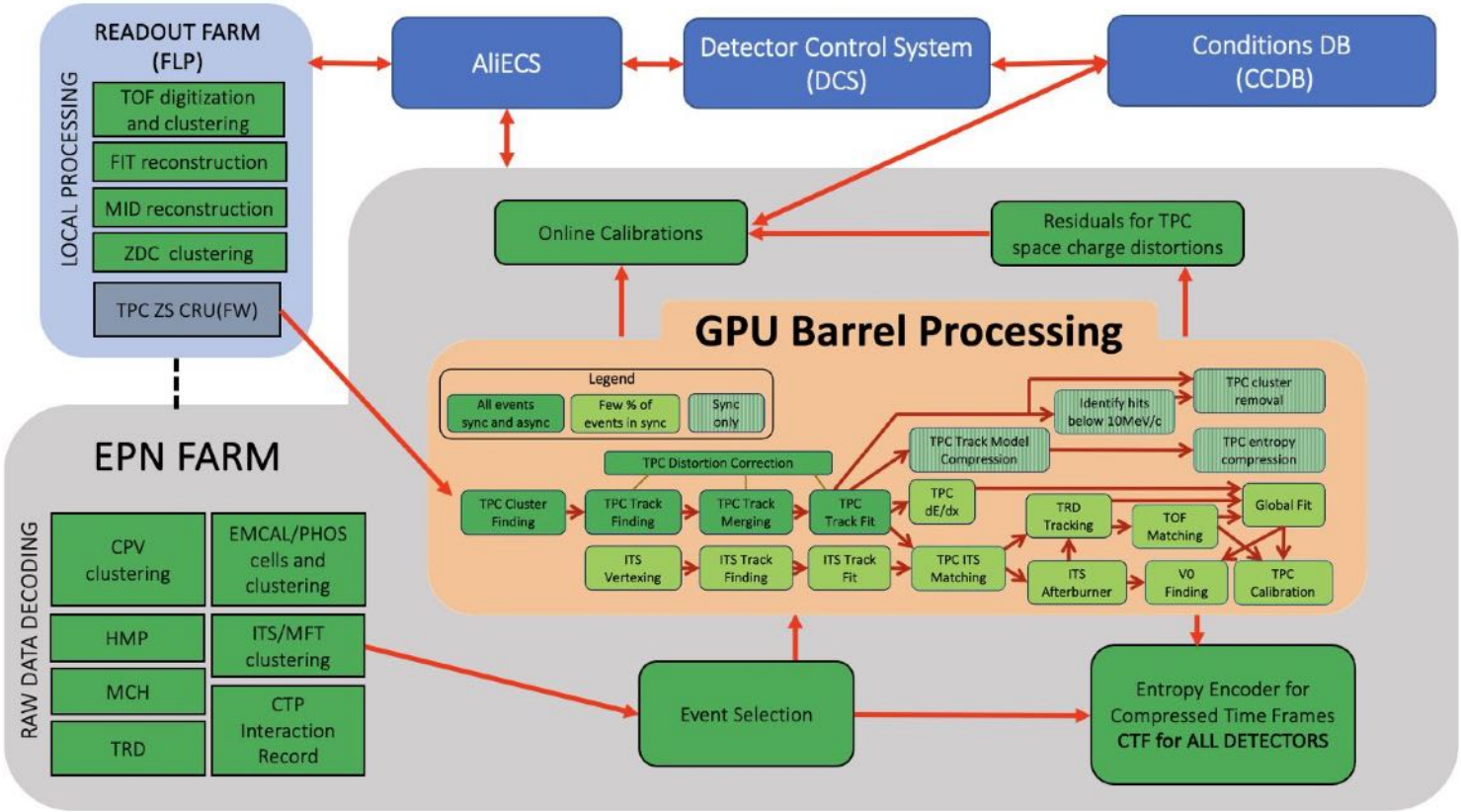
SYNCHRONOUS PROCESSING

Layout of the ALICE computing at the LHC P2



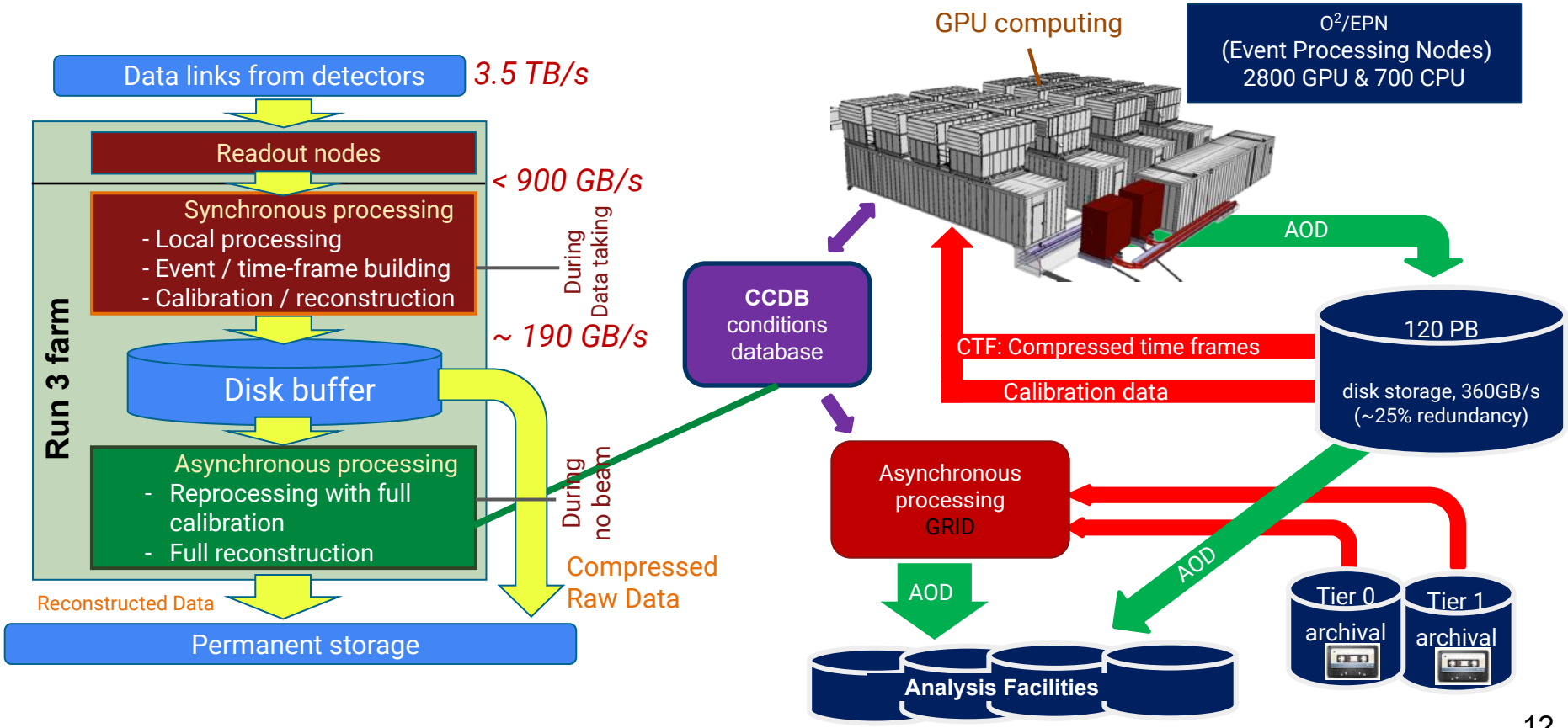
SYNCHRONOUS PROCESSING

Full view of the ALICE synchronous data taking



ASYNCHRONOUS PROCESSING

Layout of the ALICE computing at the LHC P2



federico.ronchetti@cern.ch - Workshop on Advances, Innovations, and Prospects in High-Energy Nuclear Physics

A/SYNCHRONOUS PROCESSING

The EPN farm processes, calibrate and reconstruct while the data taking is ongoing

Synchronous processing

- **99%** of compute time spent for TPC.
- EPN farm build for synchronous processing!

Synchronous processing
(50 kHz Pb-Pb, MC data)

Processing step	% of time
TPC Processing (Tracking, Clustering, Compression)	99.37 %
EMCAL Processing	0.20 %
ITS Processing (Clustering + Tracking)	0.10 %
TPC Entropy Encoder	0.10 %
ITS-TPC Matching	0.09 %
MFT Processing	0.02 %
TOF Processing	0.01 %
TOF Global Matching	0.01 %
PHOS / CPV Entropy Coder	0.01 %
ITS Entropy Coder	0.01 %
Rest	0.08 %

Running on GPU in baseline scenario

Only data processing steps
Quality control, calibration, event building excluded!

Asynchronous reprocessing :

- More detectors with significant computing contribution.
- **GPUs** provide the **required compute power.**
- Time frame concepts yields large enough GPU data chunks.

Asynchronous processing
(650 kHz pp, real data, calorimeters not in run)

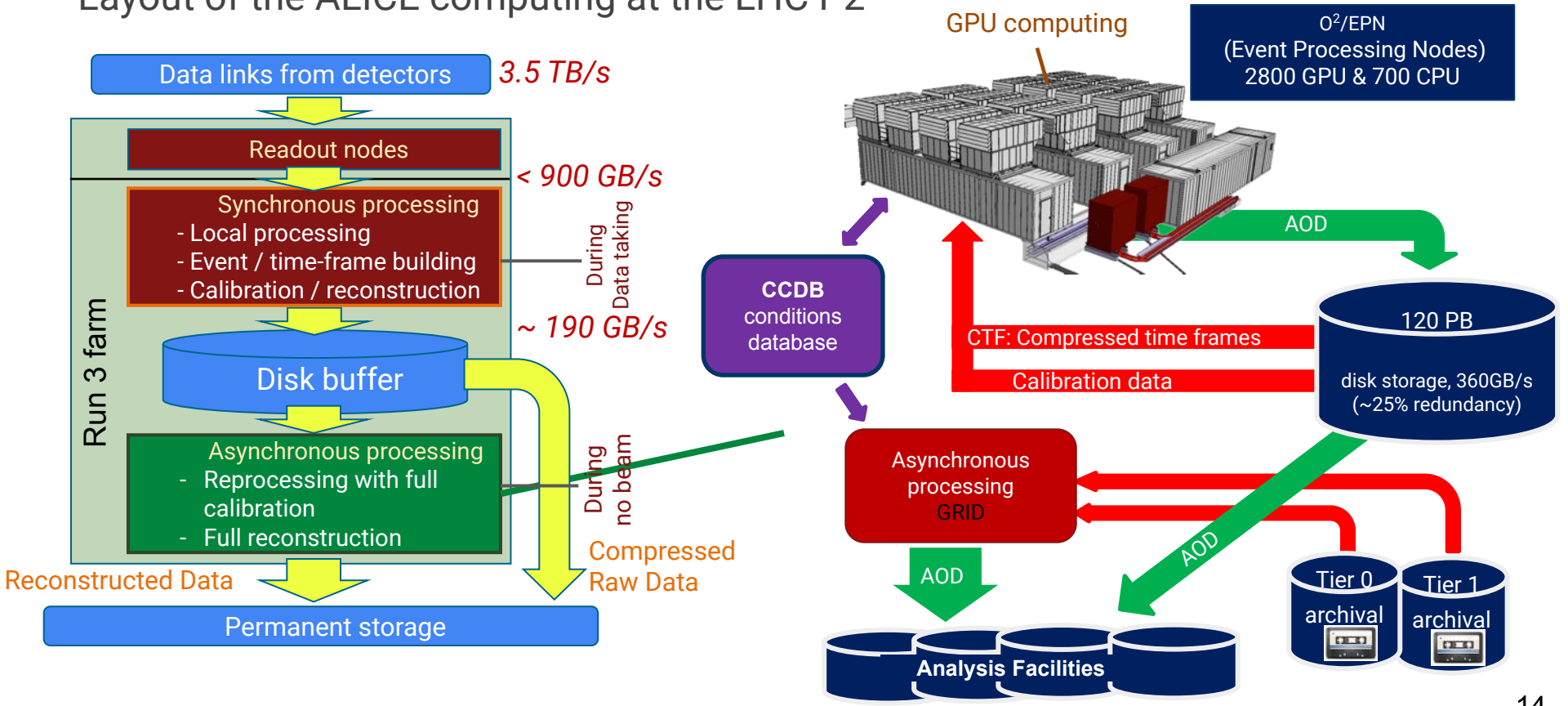
Processing step	% of time
TPC Processing (Tracking)	61.41 %
ITS TPC Matching	6.13 %
MCH Clusterization	6.13 %
TPC Entropy Decoder	4.65 %
ITS Tracking	4.16 %
TOF Matching	4.12 %
TRD Tracking	3.95 %
MCH Tracking	2.02 %
AOD Production	0.88 %
Quality Control	4.00 %
Rest	2.32 %

Running on GPU in optimizeed scenario

Under development:
- Achieve best GPU usage in async phase
- Run rest of central barrel tracking on GPU

ASYNCHRONOUS PROCESSING

Layout of the ALICE computing at the LHC P2



ENERGY EFFICIENCY: COMPUTING

Efficient usage of computing via GPUs

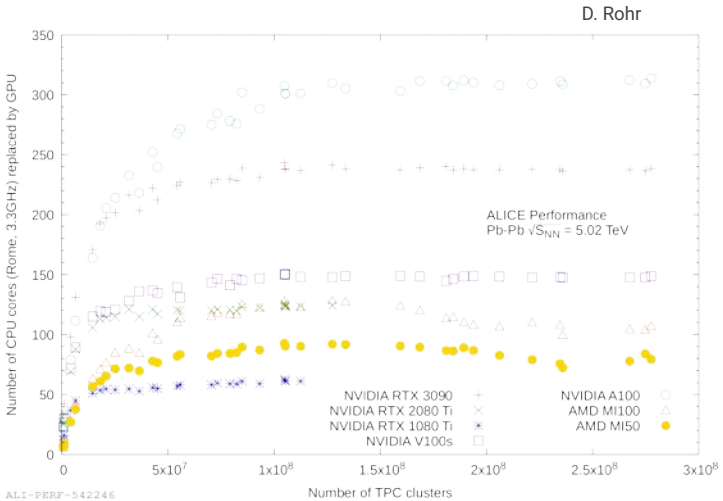
The ALICE EPN farm employs GPUs heavily to speed up online and offline processing

Synchronous processing:

- 99% of reconstruction compute load is handled by GPUs
- **1 MI50 GPU replaces ~80 AMD Rome CPU cores**
  - 1 MI100 GPU, is ca 35% faster then the MI50

Asynchronous processing:

- ~60% of full processing (for 650 kHz pp) on GPU
  - Could ideally increase to 80% with full barrel tracking
- Currently, 1 MI50 GPU replaces ~55 CPU cores in asynchronous reconstruction



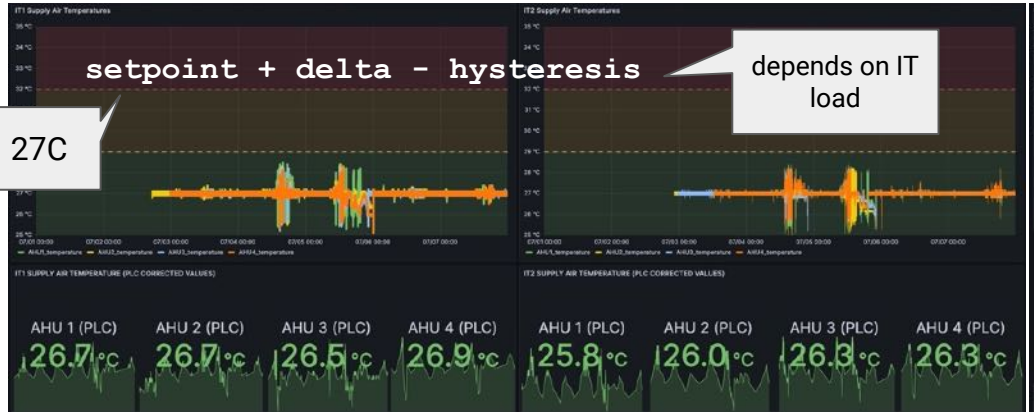
Without GPUs, more than 2000 64-core servers would be needed for online processing

ENERGY EFFICIENCY: INFRASTRUCTURE

Efficient usage of resources

Key features

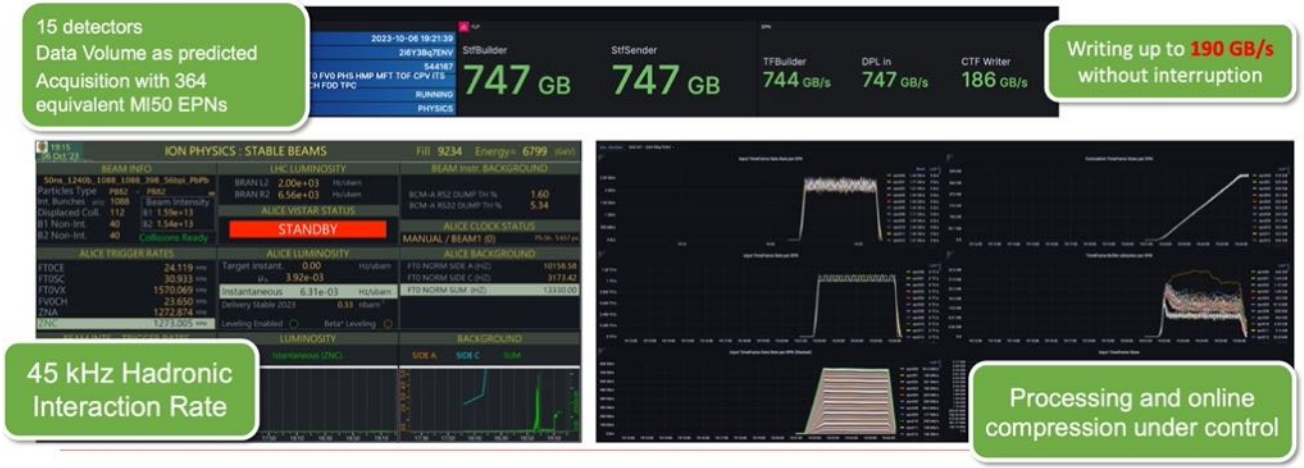
- **Modular containers** → allow easy extensions of the farm (also in view of possible upgrades)
- **Adiabatic cooling** → vaporize purified (Reverse Osmosis) water on heat exchangers when the container temperature is higher than the setpoint (otherwise use air-air cooling)
- **Commissioning and operation by the EPN team**
  - Preventive and second level maintenance from CERN (with the support of the ALICE Technical Coordination)



THE 2024 HEAVY ION RUN

Synchronous processing was successful in 2023 and ready for 2024

- High rate HI data was successfully taken in October 2023 with peak instantaneous luminosity of 47 kHz



47 kHz Pb-Pb

- **For 2024 the LHC performance is expected to be more stable: i.e. longer fills levelled at 50 kHz**
- Lower compute margin with respect to 2023 (more processing tasks, worker node mortality)
  - Still sufficient to cope with the expected rate of 50 kHz

Asynchronous reconstruction of 2023 ongoing in parallel (only during pp data taking)

- **EPN farm acts also as a GRID site (EPN nodes are 2.5x faster** than the CPU equivalent)
  - Other ALICE GRID sites use only CPUs

OUTLOOK AND CONCLUSIONS

Since 2012 ALICE has pioneered online data reconstruction/compression with GPUs

2010 → 2015 → Today → Long Shutdown 3 2027-2029

64 * NVIDIA GTX 480 in Run 1
Online TPC tracking

180 * AMD S9000 in Run 2
Online TPC tracking

2800 * AMD MI50/MI100 in Run 3
Online and Offline barrel tracking

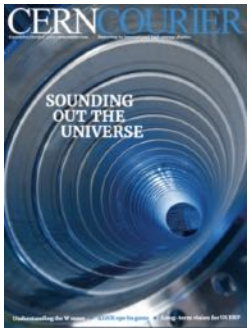
New detectors
All silicon ITS3

During the next Long Shutdown (LS3: 2027-29) ALICE will be receive further upgrades

- **The new all-silicon inner layers of the ITS and a new forward calorimeter: 20-30% more data**
- **The redout nodes will be consolidated and pushing more data on the IB network → 400+ Gb/s (estimation)**
- The current EPN farm nodes will not be replaced due to obsolescence of servers and GPUs
- An upgraded and consolidated EPN farm could be hosted in the current IT infrastructure and coexist with the old farm
  - Evaluation of the possible EPN upgrade scenarios considering the HPC/AI marker evolution are ongoing within the EPN project

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

THANK YOU…



…and if you want to learn more about the ALICE EPN farm please check out the following article on

- CERN Courier, September/October 2023

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 O² 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.

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 O² project, which merges online (synchronous) and offline (asynchronous) data processing into a single software framework. In addition to an upgrade of the

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

More EPN papers ahead…

- ***Efficient scientific computing with the ALICE Event Processing Nodes GPU-based farm***, **Frontiers in physics, to be submitted in December 2024**
- ***Online data processing with the EPN farm,*** **full technical paper in preparation**