ALICE Physics Data Processing

in LHC Run 3+4

Andreas Morsch for the ALICE O^2-PDP Project

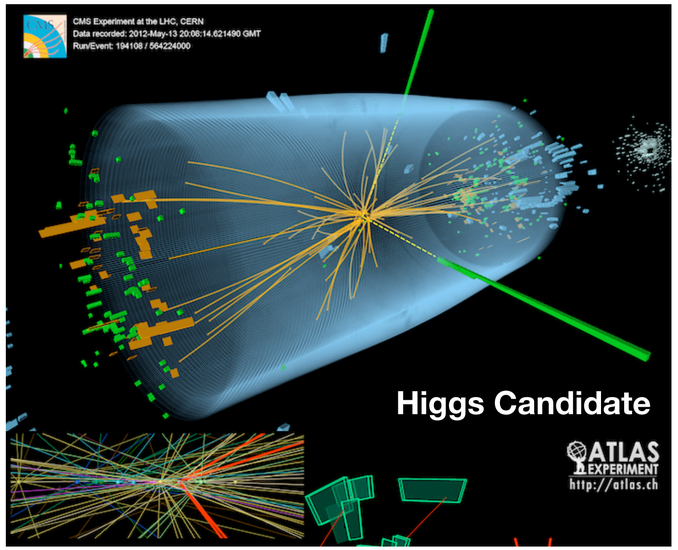
IT Seminar, 17/1/2024

Outline

- Some background concerning the ALICE physics programme
- Motivation for ALICE upgrade during LHC Long Shutdown 2
- New online processing and the O^2 facility at Point 2
- Asynchronous (offline) reconstruction
- New software frameworks for calibration, reconstruction, simulation and analysis
- First experience with high-rate Pb-Pb
- Future

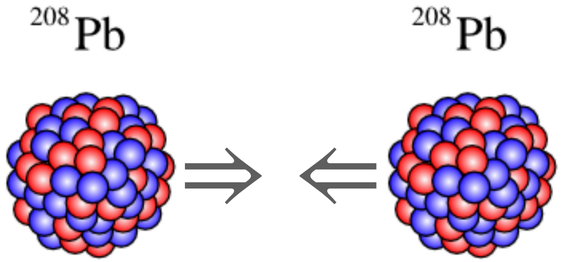
What is the nature of the Universe? What is it made of?

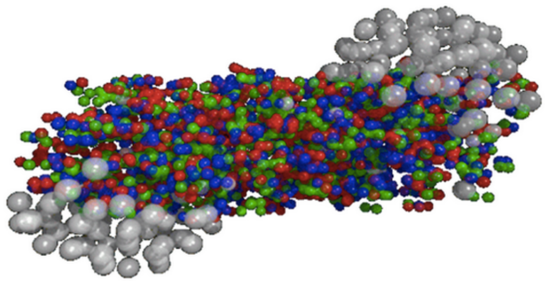
- Study **elementary particles** and their **interactions** at high energies.
  - High centre of mass energy of colliding particles, ex.
    - LEP $e^+e^- \rightarrow Z^0$ factory to study the Z-boson
    - Search for new particles in pp-collisions @ LHC (discovery of the Higgs-boson)



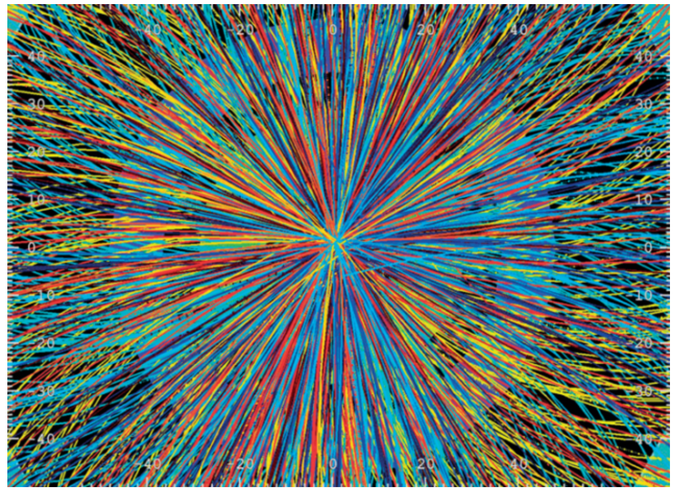
Our quest is not complete without …

- Study **matter** under extreme conditions prevailing in the early Universe / neutron stars
  - Ultra-relativistic heavy ion collisions
    - Distribute large amount of energy over sizeable volume
  - Phase transition from normal matter (neutrons and protons) with quarks and gluons as their confined constituents
  - …. to a **plasma of free quarks and gluons (Quark Gluon Plasma (QGP))**





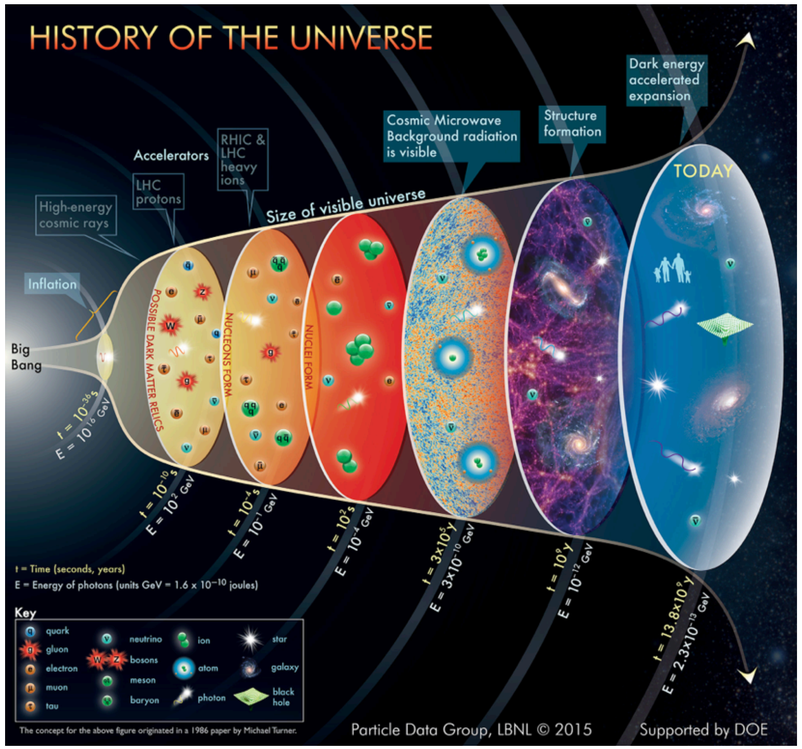
Quantum Chromo Dynamics (QCD) predicts phase transition to QGP



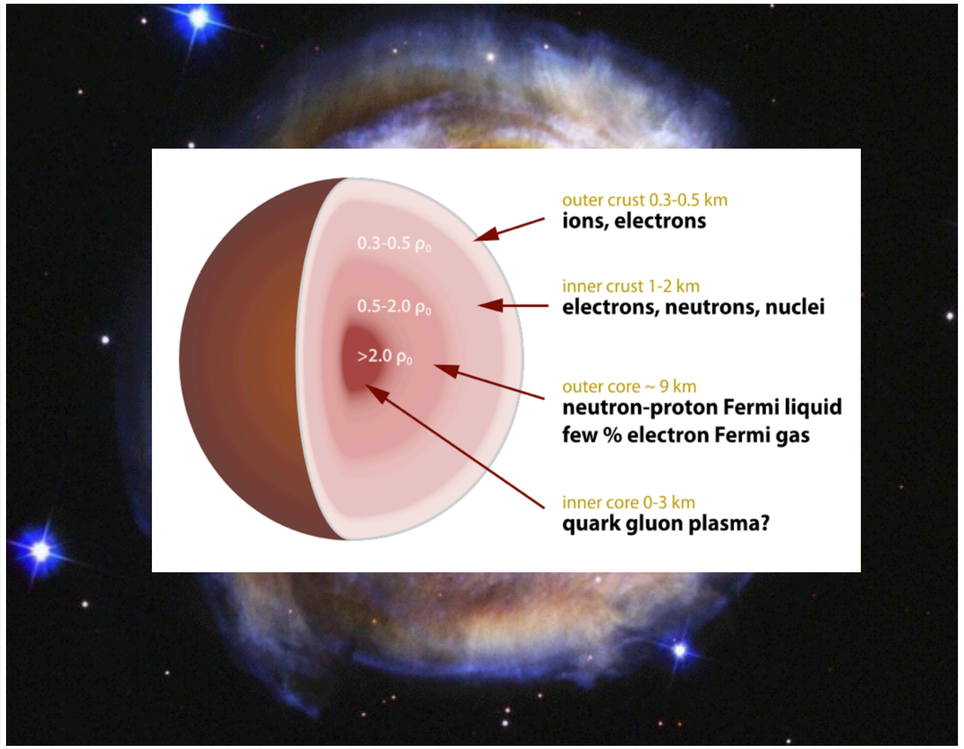
Matter under extreme conditions

- Transformation of normal matter confined in hadrons to the Quark Gluon Plasma needs:
  - high temperatures: $4 \cdot 10^{12}$ K (10^5 x temperature in interior of sun) or
  - high pressure / density: 10^{34} Pa (road roller placed on the area of the size of a proton)

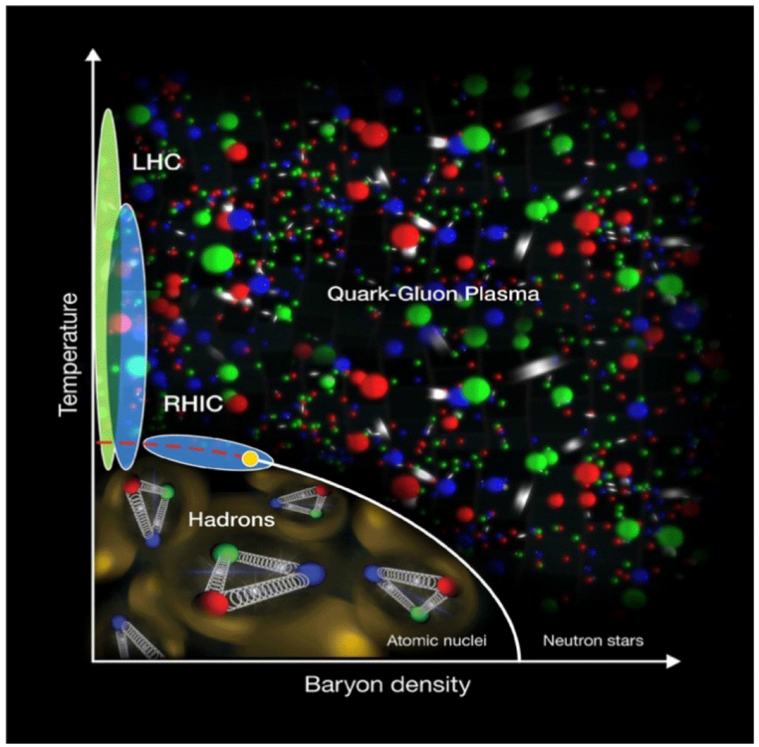
QGP phase ~10 μs after Big Bang



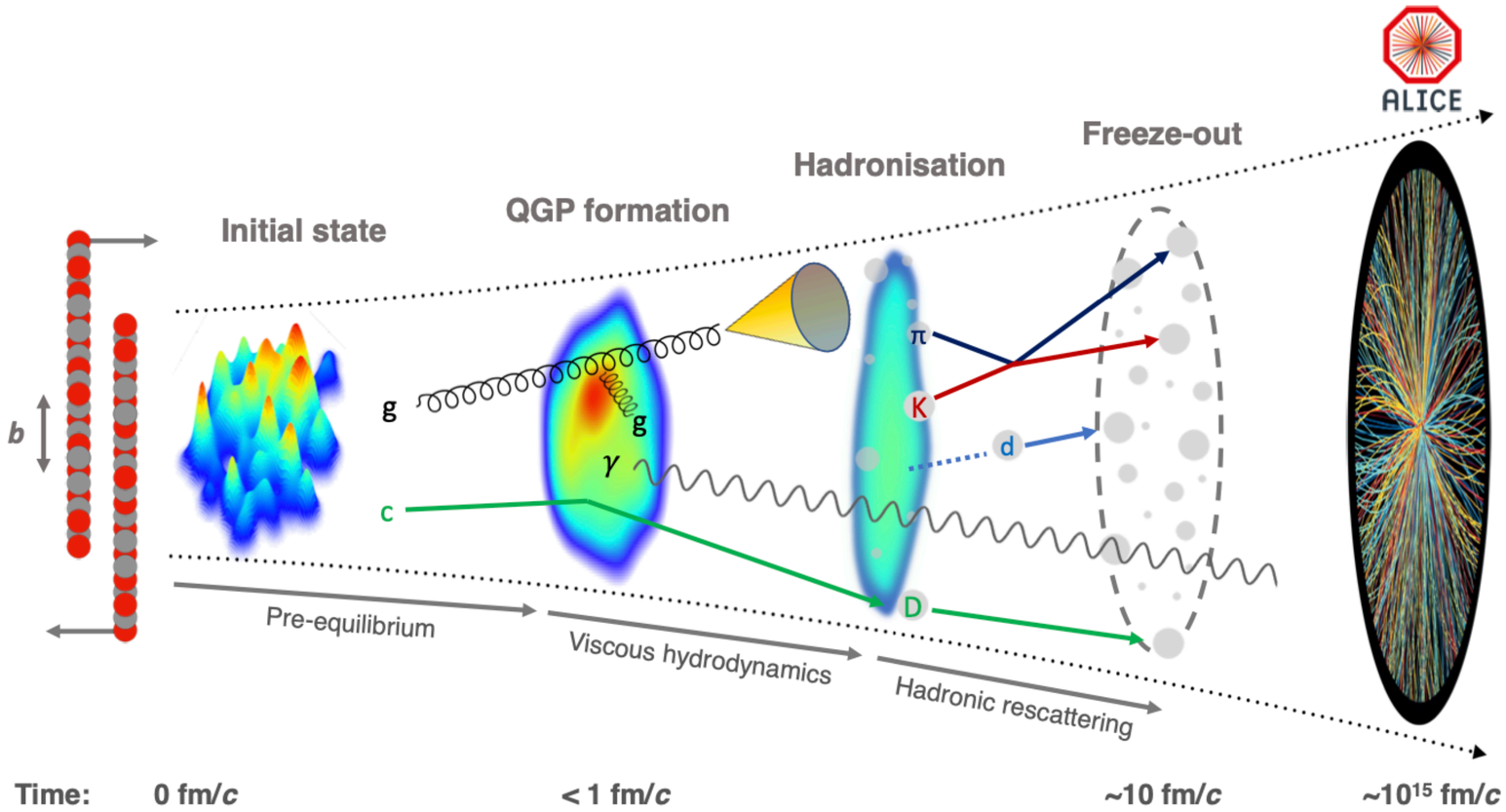
Neutron star



QGP phase transition at colliders



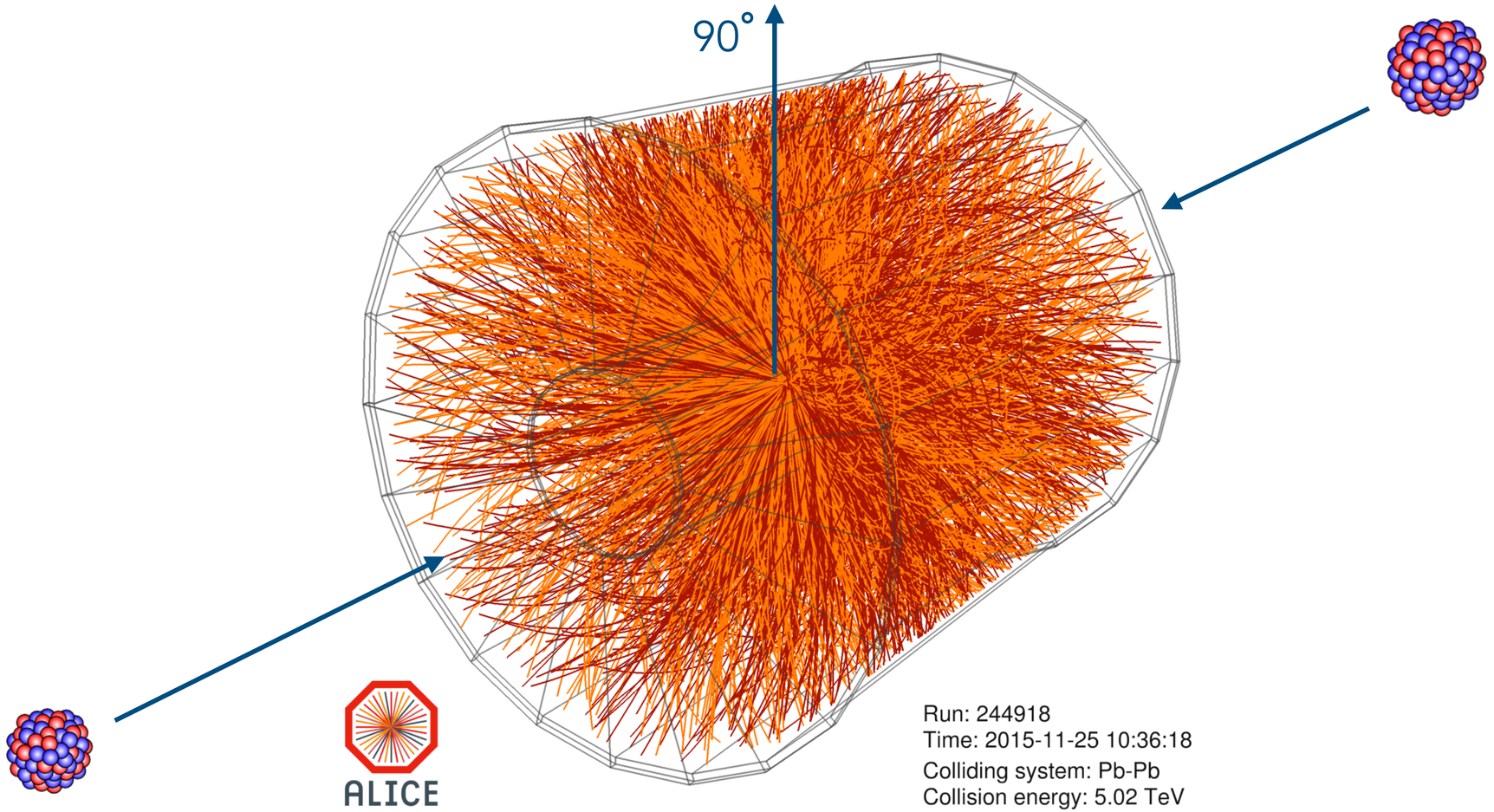
QGP created in HI collision is ephemeral



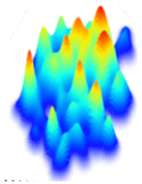
- In HI collision QGP produced under "explosive conditions"
  - Matter evolving through several phases
  - QGP phase lasts only $\mathcal{O}(10^{-23}$ s)
  - **Only final state particles are directly observable as messengers of the earlier phases … and there are many of them …**

Biggest experimental challenge for HI physics

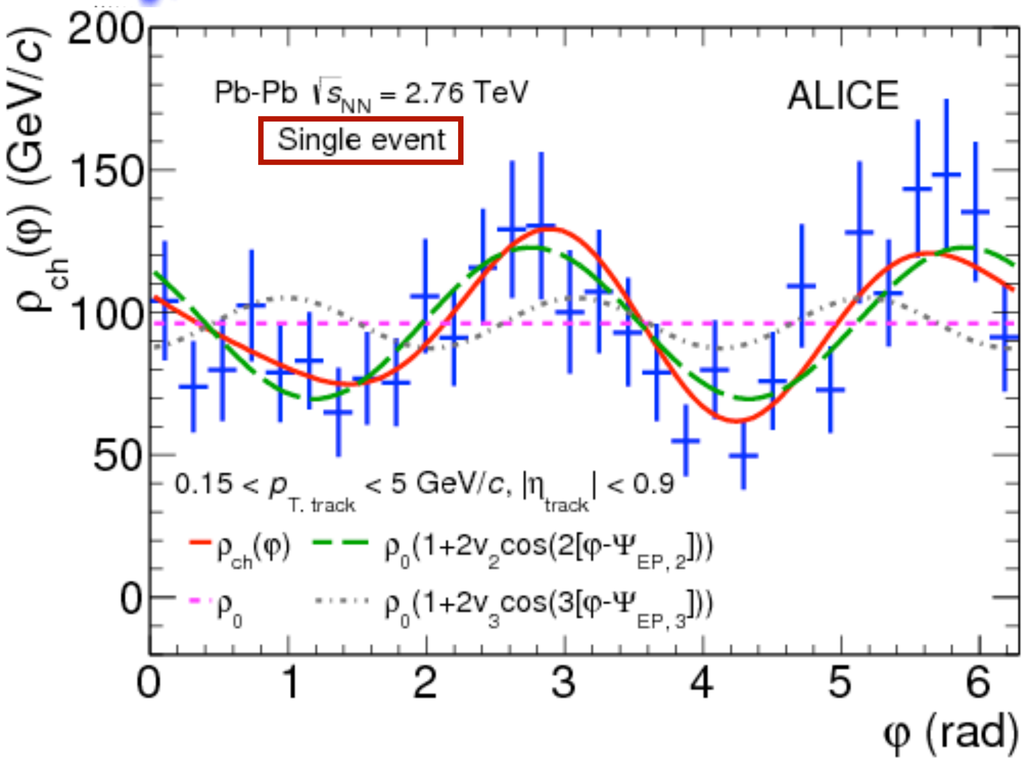
- High charged particle density
- Up to 4000 particle from a single collisions in central detector region: 45°-135° range (perpendicular to beam axis)



Strong signals

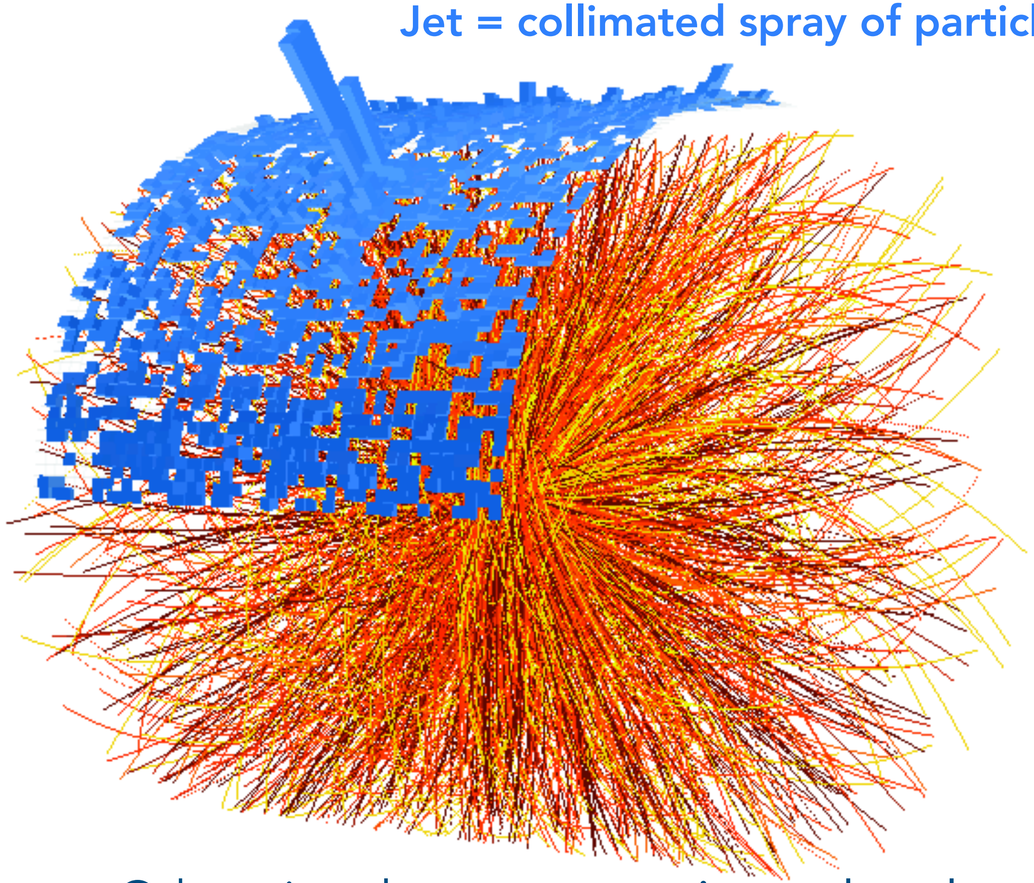


collision geometry $\mathcal{O}(10^{-15}\ \mathrm{m}) \Rightarrow \mathcal{O}(\mathrm{m})$



- Some effects seen in single collisions (energy flow)

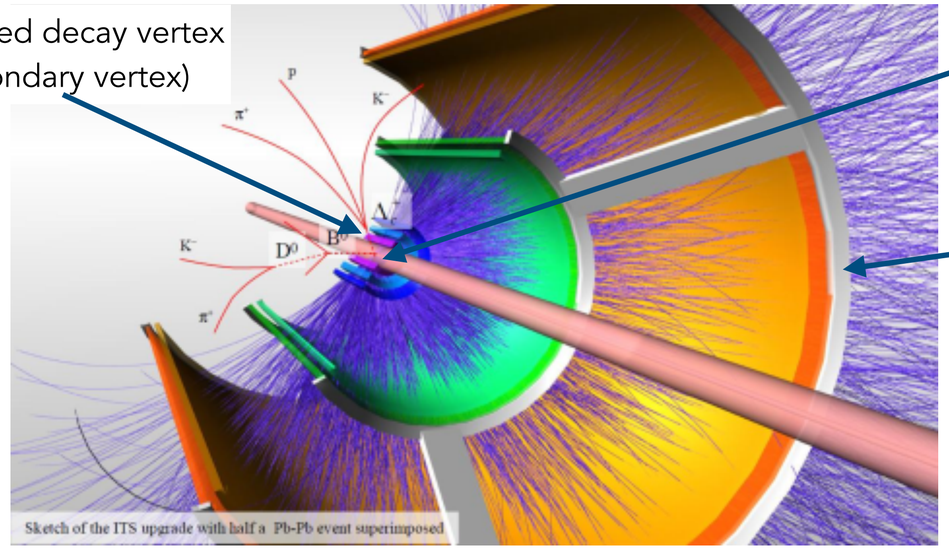
Jet = collimated spray of particles

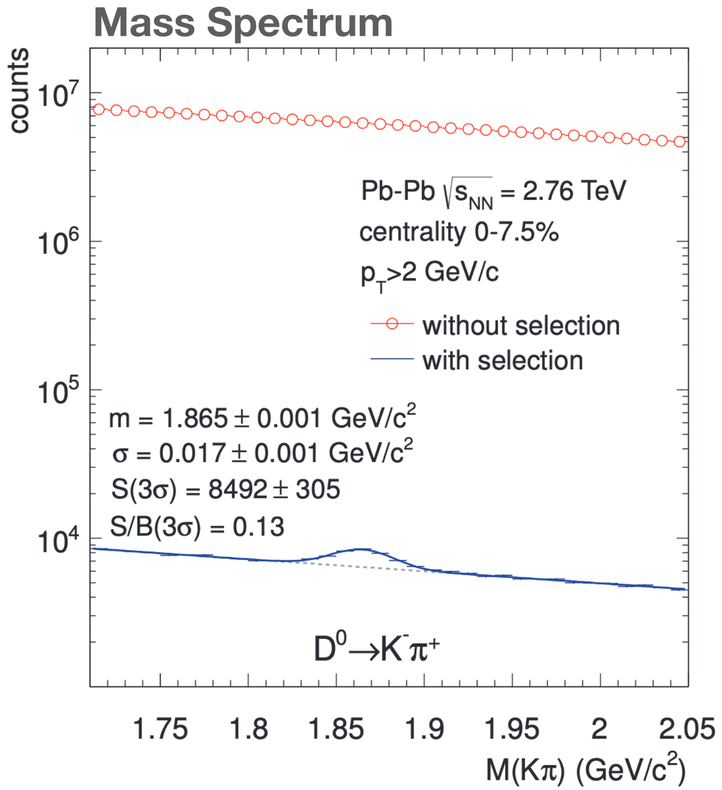


- Other signals are so prominent that they can be rel. easily distinguished from the background
- **Selectable by hardware / online triggers.**

Small signal / large noise

- **… and then there are signals that need to be identified over large backgrounds**
- In particular, heavy, slow particles (charm and beauty hadrons)
- Get "kicked around" by the QGP, and are therefore important "messengers"
  - No direct observation but identification via their decay products
  - Look for particle production points displaced from main collision point ($\mathcal{O}(100\ \mu\mathrm{m})$)

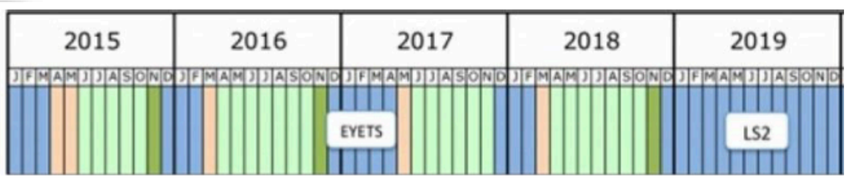


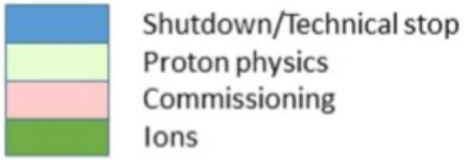


- Large number of random combinations below signal peak
- ≥ one candidate in each collision.
- **Cannot be selected by hardware / online triggers**
- **Need for large minimum bias event samples**

Heavy Ion HL-LHC has already started

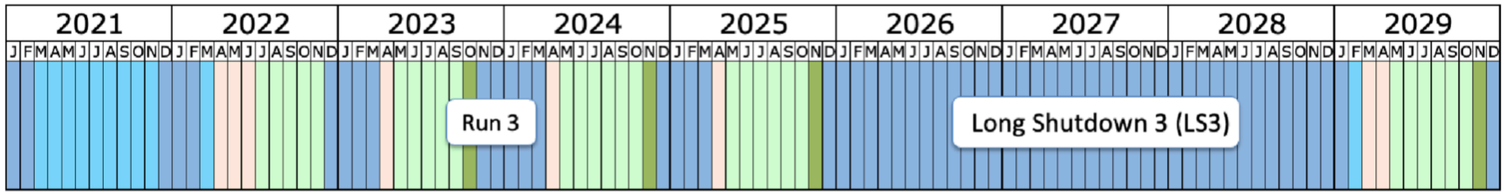
~1 month per year dedicated to HI physics

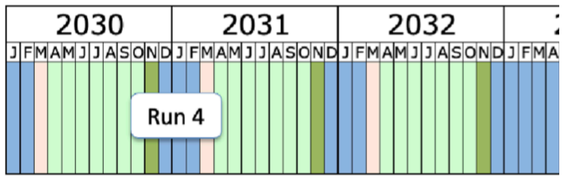




Run 2: $\int dt\mathscr{L}_{\mathrm{PbPb}} \approx 1\ \mathrm{nb}^{-1}$ (8·10⁹ inelastic collisions)

Run 3: $\int dt\mathscr{L}_{\mathrm{PbPb}} \approx 6\ \mathrm{nb}^{-1}$ (~5·10¹⁰ inelastic collisions)





Run 4: $\int dt\mathscr{L}_{\mathrm{PbPb}} \approx 7\ \mathrm{nb}^{-1}$ (~6·10¹⁰ inelastic collisions)

For ALICE: Expected increase of statistics in Run 3+4

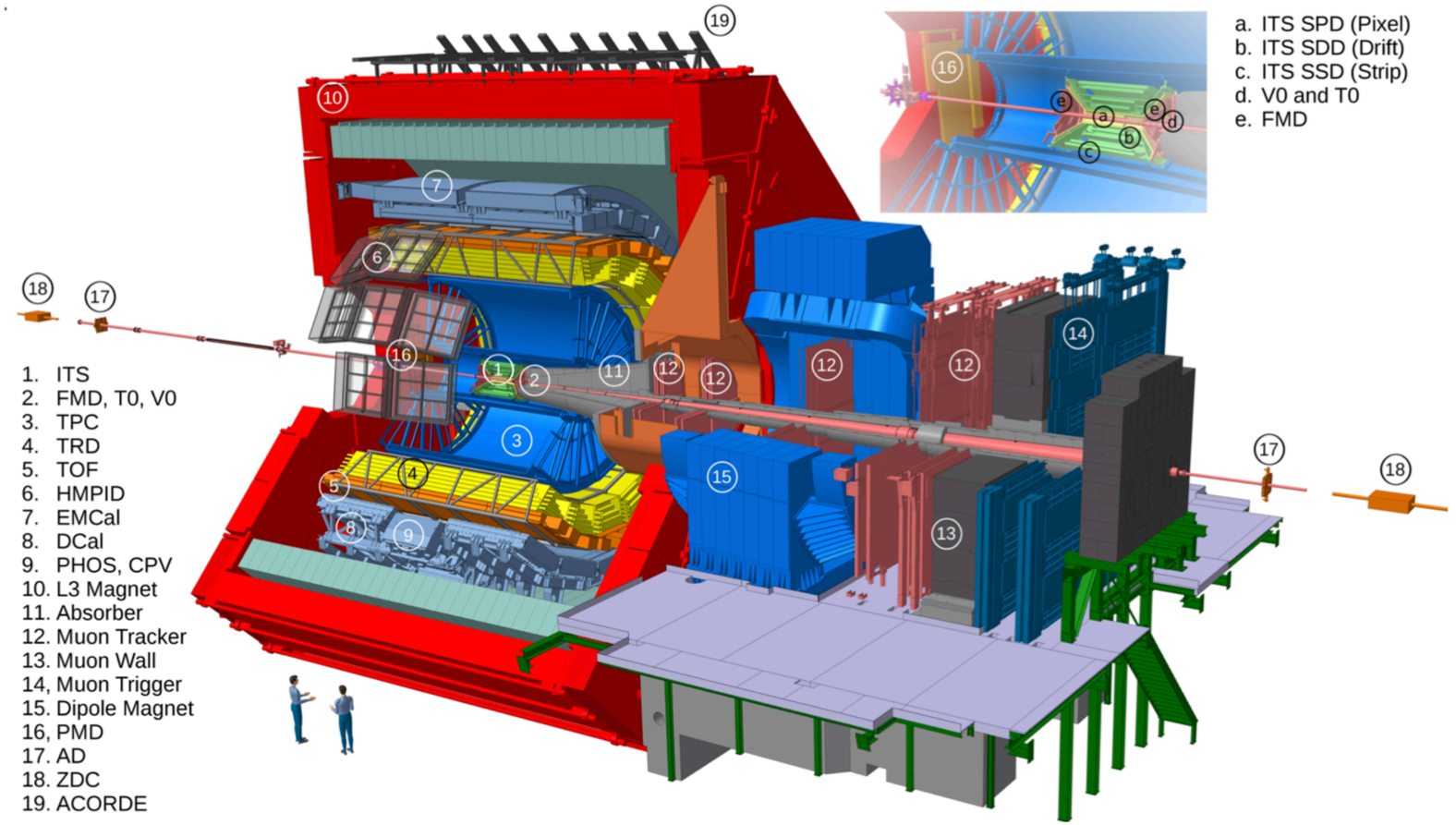
x ~10 for triggered signals
x ~100 for un-triggered collisions

Small collision systems

- Heavy Ion physics programme not complete without analysis of small collision systems: pp at Pb-Pb centre of mass energy and p-Pb
  - as reference data for Pb-Pb
- Observation of QGP-like effects in pp and p-Pb has created a novel field of research opening a new window into QCD.
- ALICE also collects large statistics data samples during LHC standard pp operation.

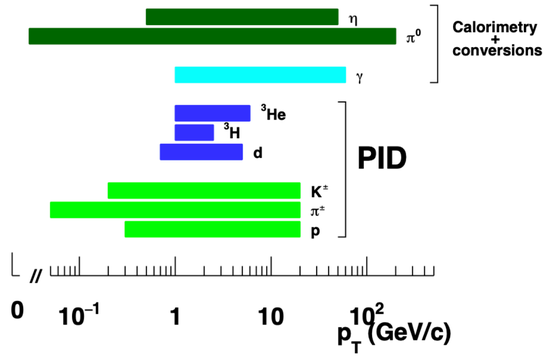
A Large Ion Collider Experiment

Optimised for QGP studies with Heavy Ions



Run 1/2 Configuration

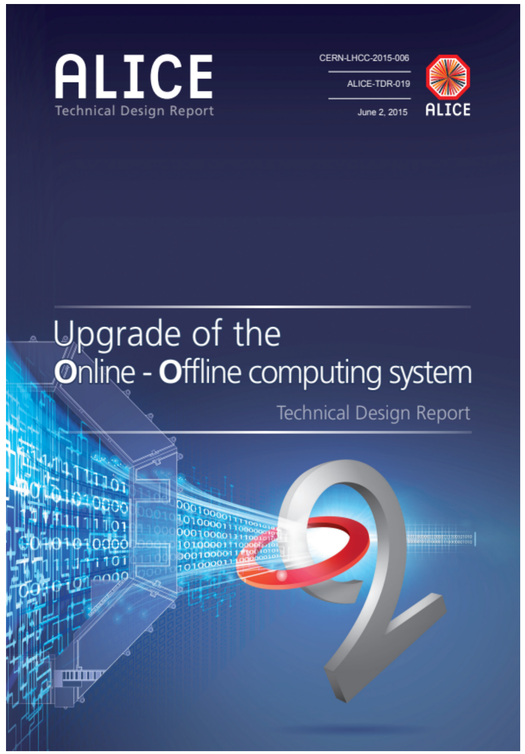
- Barrel tracking geometrical acceptance $|\eta| < 0.9$ and full azimuth
- Precision tracking in rel. low B-field (0.5 T) down to low momenta < 100 MeV/c optimised for large particle densities
- Particle identification via all known techniques (dE/dx, time of flight, transition radiation, Cherenkov radiation, …)
- Barrel electromagnetic calorimeters
- Forward Muon Spectrometer $2.5 < y < 4$



ALICE upgrade during LHC LS2

- Improve secondary vertex resolution
  - Improved Inner Tracking System
  - Add secondary vertexing capabilities for forward muons
- Make full use of the LHC Pb-Pb luminosity
  - $\mathscr{L}_{\mathrm{PbPb}} \approx 6$ Hz/mb; $\sigma_{\mathrm{PbPb}} \approx 8000$ mb ⇒50 kHz interaction rate
  - Triggering not possible for probes with very low S/B: inspect every event offline
    - compare to Run 1/2: max read out rate 1 kHz
    - mainly limited by TPC dead time (gating, < 3 kHz) and bandwidth (< 1 kHz)
      - Multi Wire Proportional Chamber use gating to limit ion back-flow

⇒Continuous detector readout, in particular upgraded TPC readout



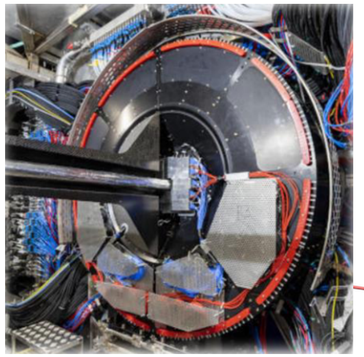
- >100x more data than in Run 1+2 while WLCG resources increase only by ~x4 over 10 years

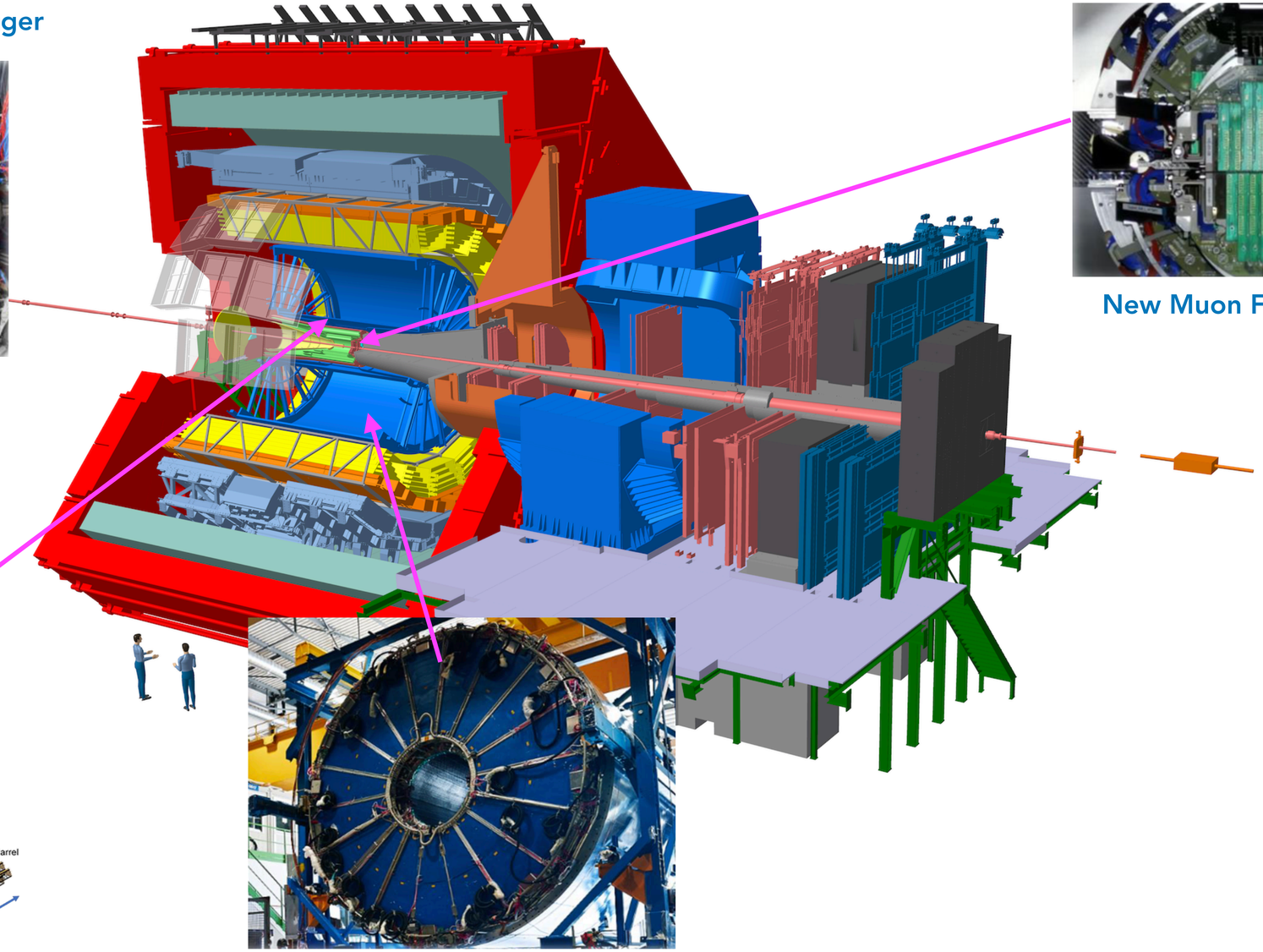
⇒New system (O^2) for **O**nline and **O**ffline data processing

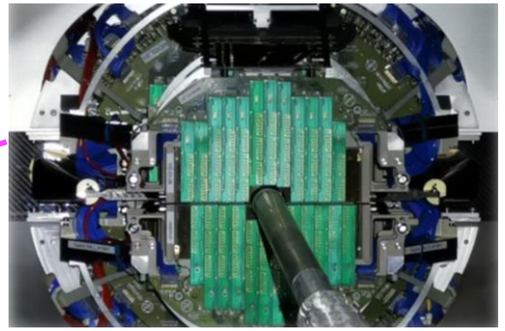
- Cope with high data rates: 3.5 TB/s from detectors; dominated by TPC
- Minimise costs and requirements for data processing and storage
  - Data compression
  - Leveraging parallel processing for reconstruction, simulation and analysis

ALICE hardware upgrades in LHC LS2

New Fast Interaction Trigger

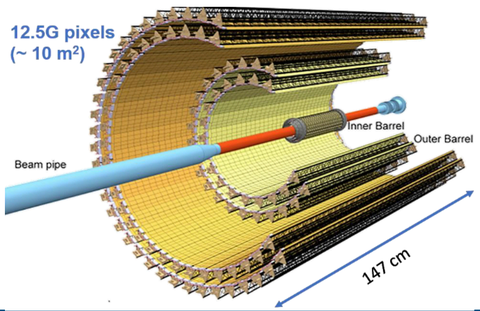






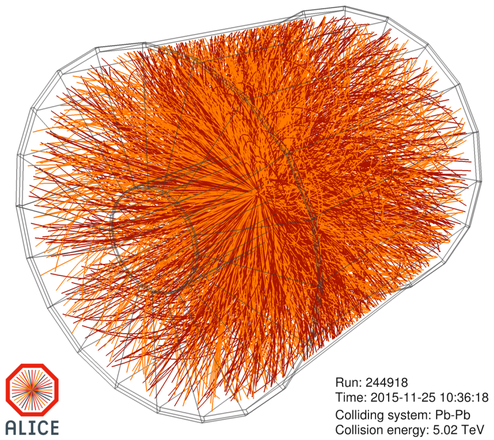
New Muon Forward Tracker

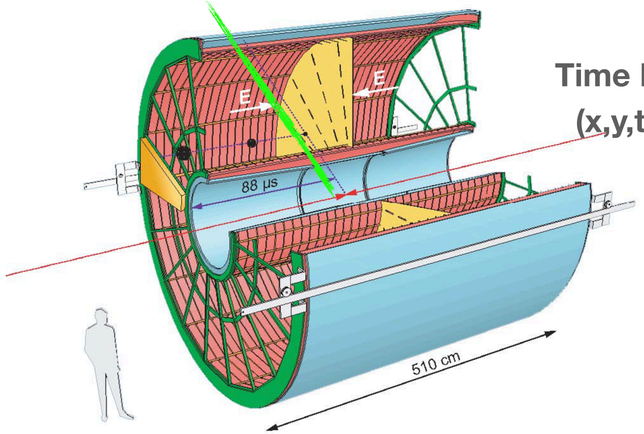
New Inner Tracking System

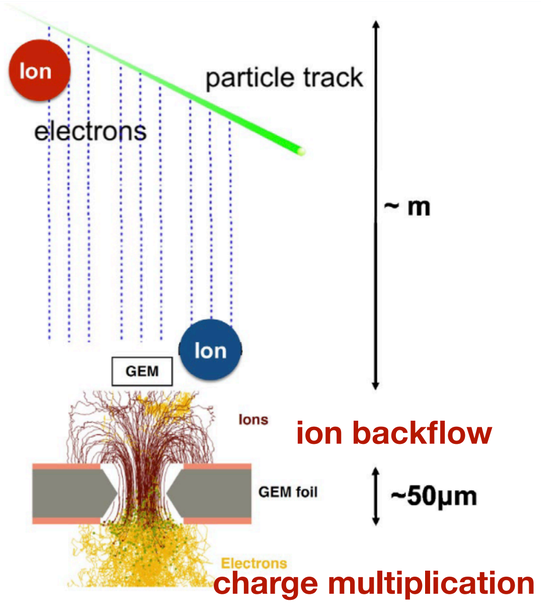


Gas Electron Multiplier (GEM) Based TPC continuous read-out

One more challenge: TPC space charge distortions



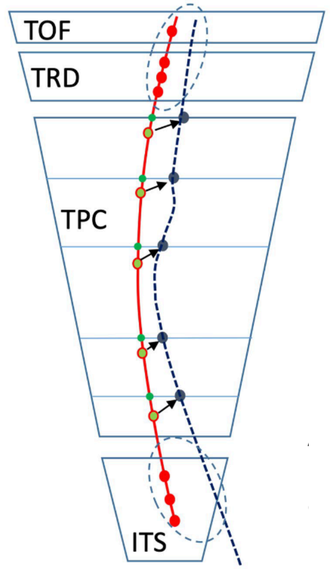




- TPC detection mechanism (charge avalanche) creates slowly moving disks of ionised atoms in drift volume.
  - At 50 kHz Pb-Pb, ions from $\mathcal{O}(10^4)$ collisions in the drift volume
  - Charge from these ions leads to time-dependent modifications of E-field.
  - Effect limited by optimised Gas Electron Multiplier technology (<1% ion back-flow)

⇒ distortions of the hit positions which need to be taken into account in track reconstruction.

- Correction from $\mathcal{O}$(10 cm) to intrinsic TPC resolution $\mathcal{O}(100\ \mu\mathrm{m})$
  - $\mathcal{O}$(1 mm) during synchronous reconstruction pass
- Corrections obtained as average difference between TPC hits and tracks obtained from surrounding detectors.



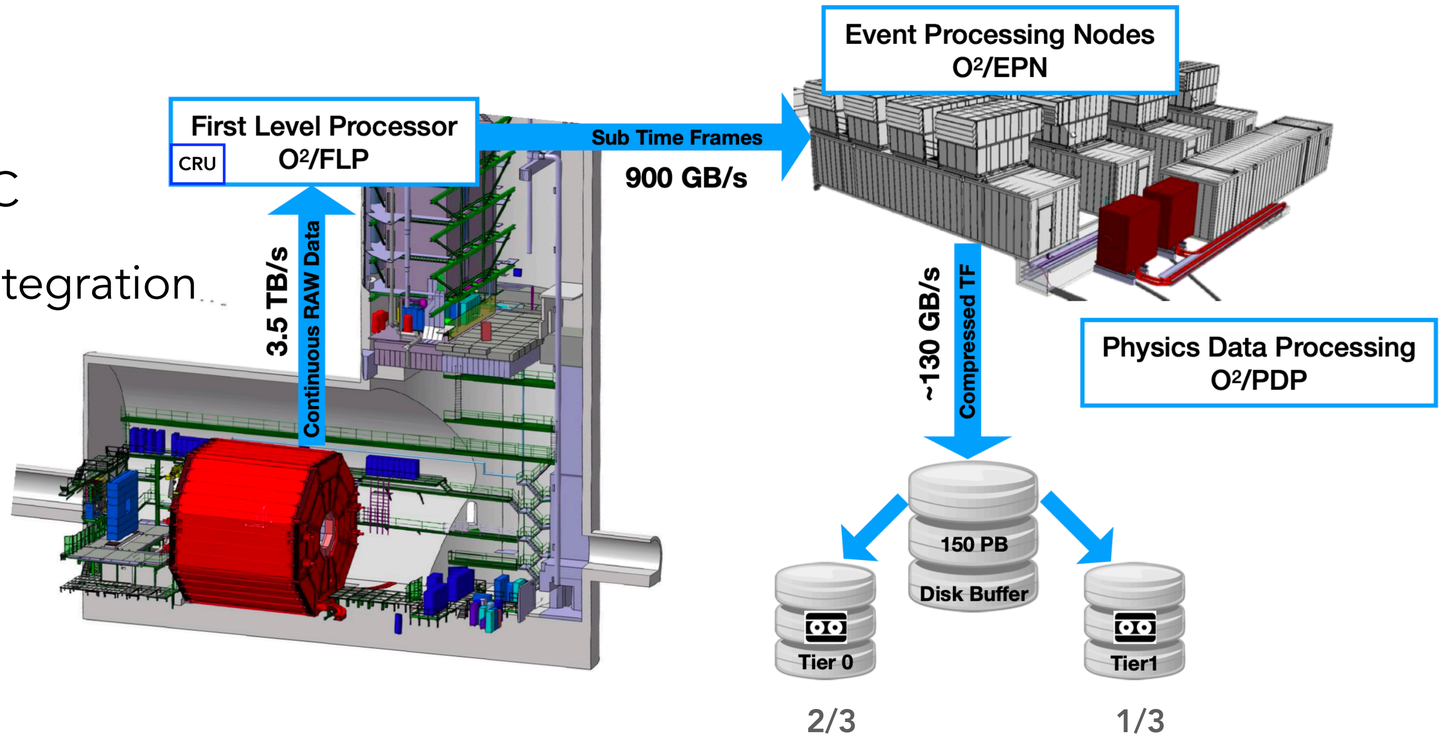
Physics data processing synchronous with data taking

Main design considerations

- Minimise costs and requirements for data processing and storage
  - Aim for maximum compression of the data volume read out from detectors synchronously with data taking
    - through online TPC track reconstruction
  - Extraction calibration data online to avoid offline calibration passes over full data sets
- Support for continuous and triggered read out
  - Subsystems that have been not upgraded to cont. readout need hardware trigger signal
  - Triggered read out also needed for commissioning and calibration runs

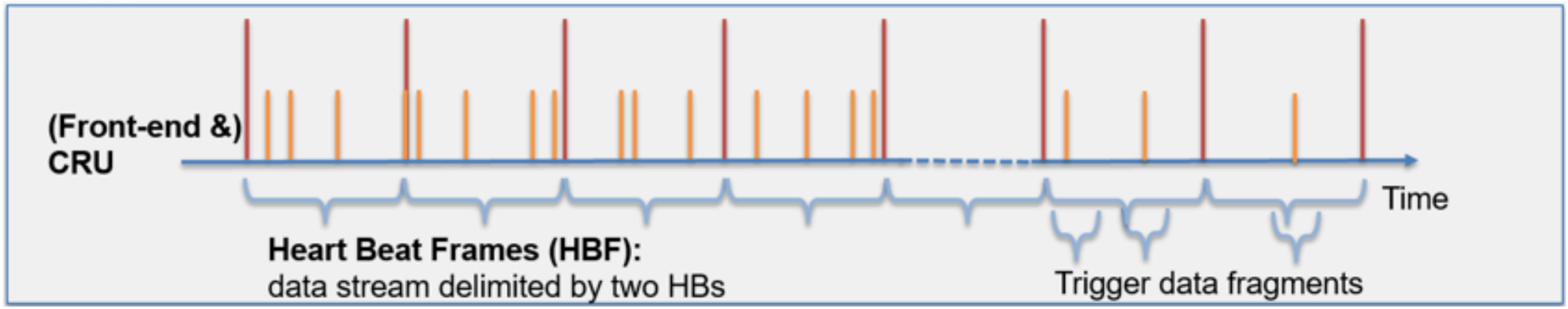
Readout data flow

- Online data processing performed in two steps on O² facility @ Point 2
  - Two types of computing nodes
    - First Level Processors (FLP) located in the experiment access shaft (CR1)
    - Event Processing Nodes (EPN) in dedicated computing containers (CR0)
  - Facility provides also
    - the network for data distribution
    - EOS based storage capacity @ Meyrin CC
    - connection to Tier 0 storage and GRID integration

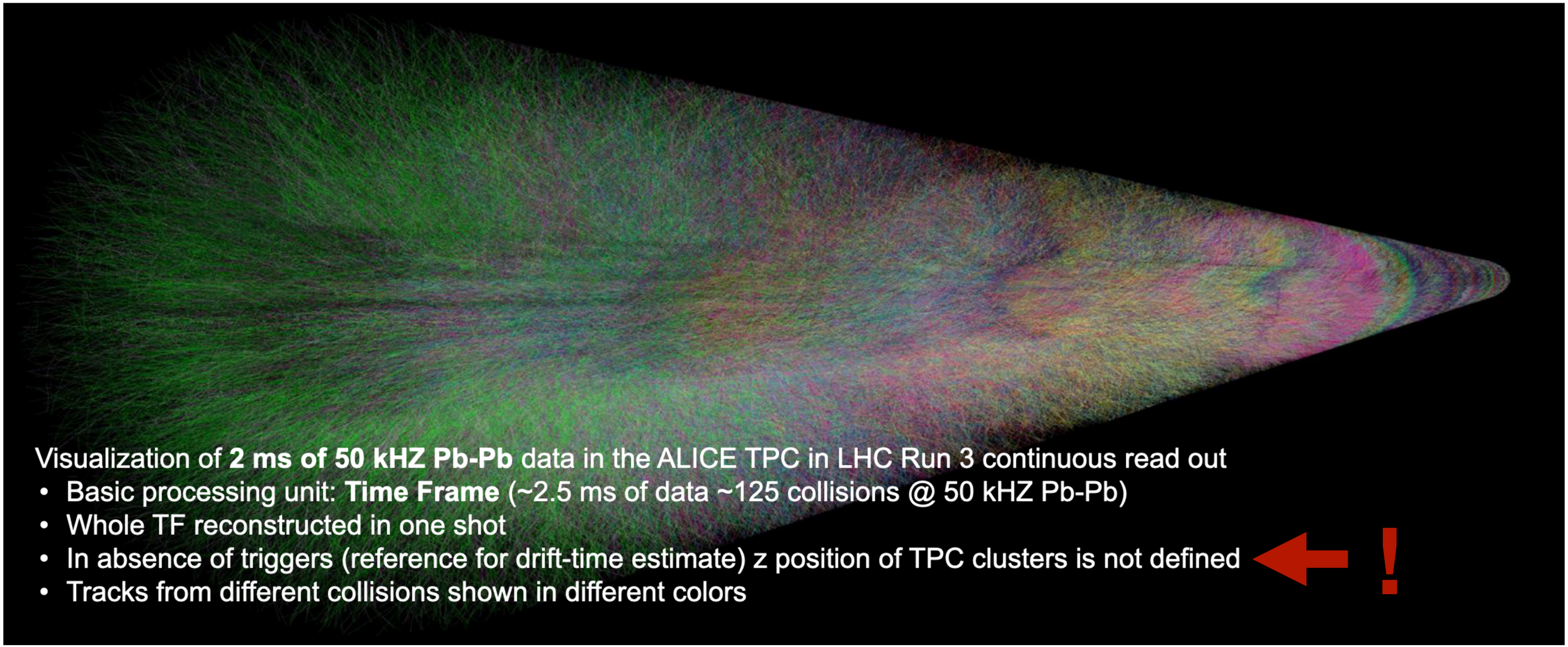


Event identification with continuous read out

- Timing information is needed to **match data streams** coming from different detectors
- Solution for continuous read-out: **Heart Beat Trigger**
  - Heart beat frequency given by the revolution frequency of the LHC = 1 orbit (11.25 kHz)
  - ALICE data stream divided into **heart beat frames (HBF)** with duration of 1 orbit synchronised with LHC clock.
  - Within one orbit: fixed number **bunch crossings (BC)** where collisions can occur (max 3564, depends on LHC fillings scheme)
- Configurable number of HBFs form a **Time Frame**
  - **data processing unit replacing traditional event entity**
  - nominal length is now 32 orbits (2.85 ms) and was 128 orbits in 2022
  - @ 50 kHz Pb-Pb, it contains on average 142 collisions.
  - Continuous and triggered data are tagged by BC and HBF identifiers

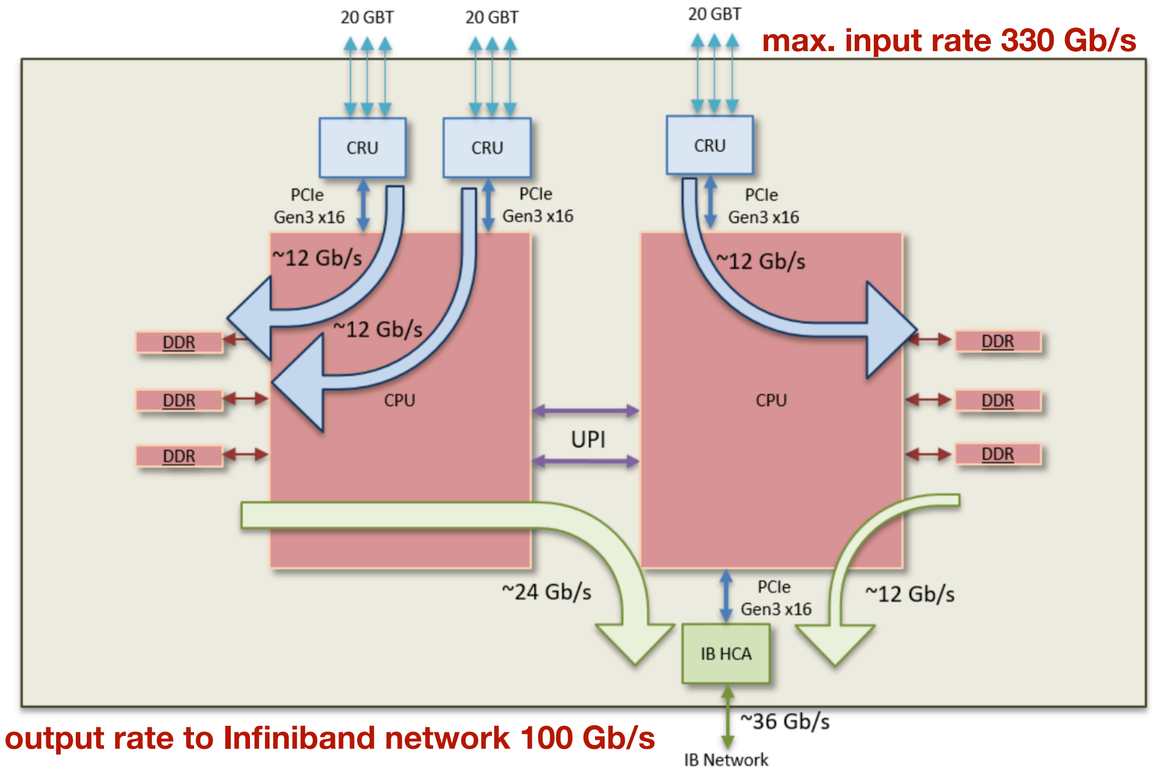


Continuous read out visualisation



First Level Processors (FLP) readout farm

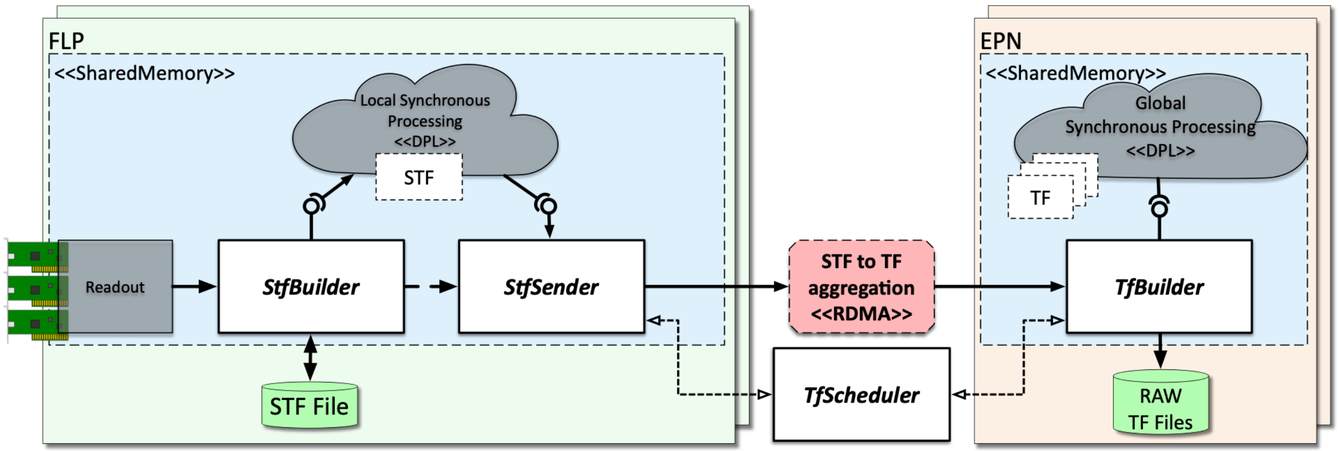
- 199 nodes (Dell Poweredge R740 servers)
  - 488 readout cards (472 new FPGA based Common Readout Units (CRUs) + 16 legacy cards)
  - Up to 3 CRUs per server connected to the detectors via optical links (8000 links in total, 20688 fibres)
  - 192 GB DDR memory
  - 2 CPUs/server (Intel Cascade Lake: 10 core Silver 4210 (mainly) and 20 core Gold 6230 depending on detector needs)



First Level Data Processing

General Tasks

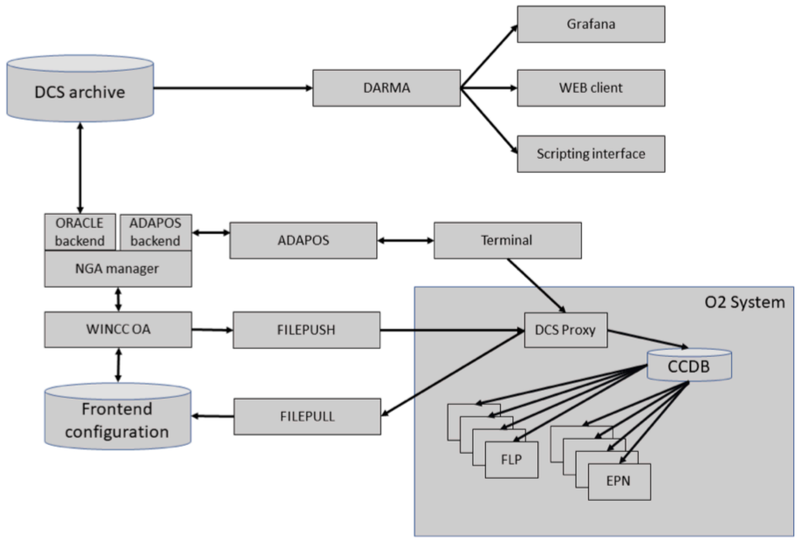
- First level of **data compression** by 0-suppression to <900 GB/s
- Possibility to perform **calibration** tasks based on local information from the part of the detector they serve.
  - Example TPC calibration on CRU (FPGA) important at high particle densities
    - Ion tail suppression
    - Common baseline restoration
- Build **Sub Time Frame** (STF)
  - Sub Time Frame comprises all HBFs belonging to a TF from one FLP node
  - After all FLPs have built their STFs of an individual TF, an available EPN is selected and all STFs are sent
  - **Full TF is built on that EPN node**



First Level Data Processing (FLP)

Dedicated Tasks: Calibration Objects and Trigger Signals

- Dedicated FLP node is used to collect and process data from the Detector Control System (DCS)
  - Detector conditions like voltages, temperature, and pressure
  - Processes configuration files sent by detectors as well as LHC information
- The calibration objects are stored in the condition and calibration database (CCDB) and from there they are read by the following processing stages.



- A 2nd dedicated FLP is used to collect all trigger signals sent by the Central Trigger Processor to the detectors

.

Synchronous reconstruction

- Main objectives
  - Reduction of the data rate from TPC (main contributor to raw data volume)
  - Extraction of data for calibration
- Strategy
  - Production of **Compressed Time Frames (CTF) for further asynchronous processing passes**
    1. Clustering and full track reconstruction in the TPC
    2. Removing background hits from TPC data
    3. Cluster space point data are stored in relative coordinates reducing the entropy
       - allows for efficient ANS entropy encoding (compression)
  - Small fraction of tracks fully reconstructed using in addition ITS, TOF and TRD information
    - used to extract data for TPC space charge distortion (SCD) calibration
- **TPC reconstruction is the most demanding step in terms of computing time.**
- **Reconstruction algorithm particularly suited for parallel processing (experience from Run 1+2)**

TPC reconstruction on GPU

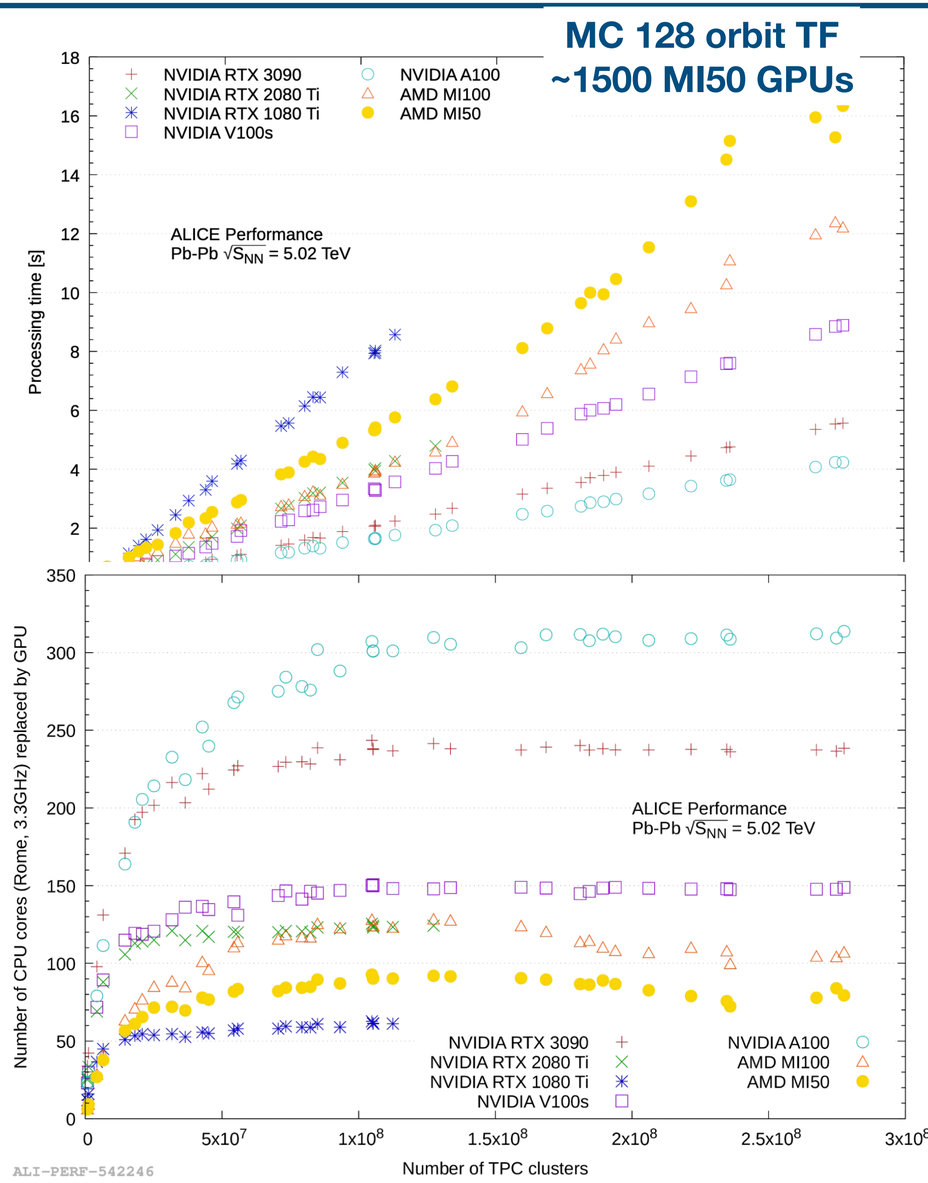
- Developed vendor-and architecture-independent software
  - All algorithms are written in generic C++
  - Can be dispatched to HIP, CUDA, OpenCL on GPUs or OpenMP on CPUs using small wrappers
    - developed in ALICE
  - GPU libraries are linked dynamically on demand
  - Same binaries can be distributed to CPU and GPU nodes

- Processing time increases almost linearly with number of TPC clusters.
- MI50 GPU ~55 faster than CPU (AMD Rome CPU core @ 2.35 GHz)

- CPU processing would need >2000 servers + corresponding networking

  ⇒ GPU processing most cost effective

  and only feasible solution within budget

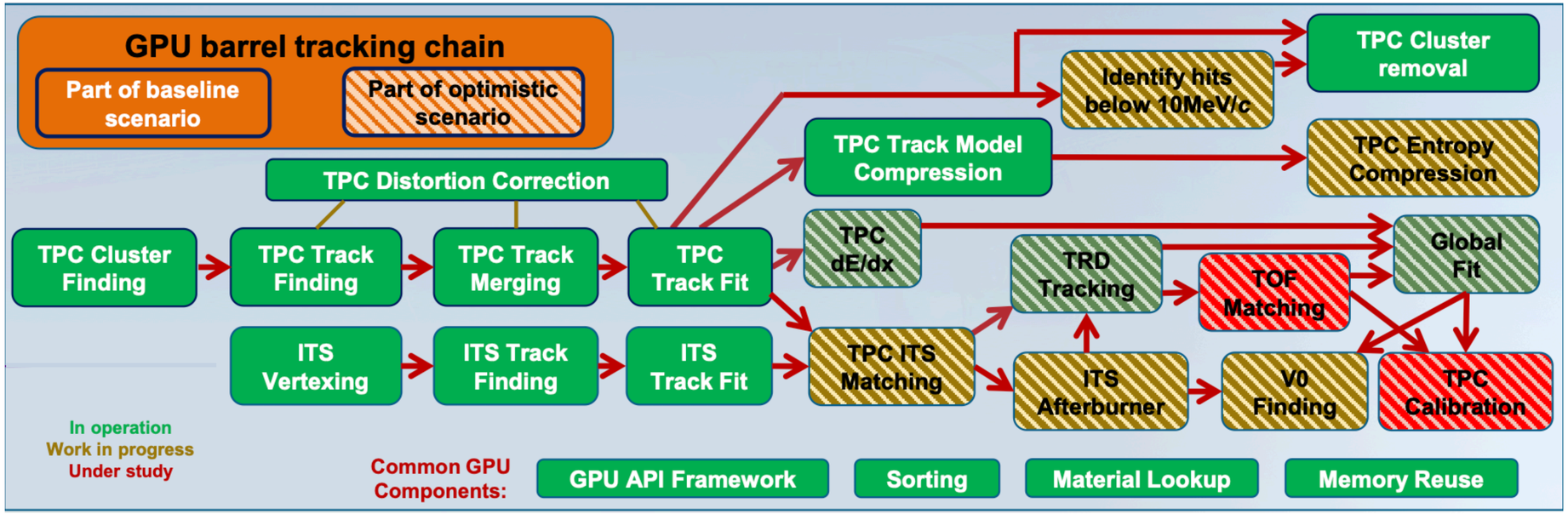


Event Processing Nodes (EPN)

- Original EPN configuration
  - 280 SuperMicro servers (AS-4124GS-TNR)
  - 8xMI50 32 GB GPUs per server
  - 2x32 core AMD Rome 7452 CPUs
  - 512 GB 3.2 GHz main memory
  - 100 Gb/s Infiniband Host Channel Adaptors (HCA)
- Consolidation after 2022 Pb-Pb test run
  - 30% more hits than expected exhausting our margin
  - 70 additional servers with 8 x MI100 GPUs added (35% faster)
    - 2x 48 core AMD Rome, 1 TB main memory

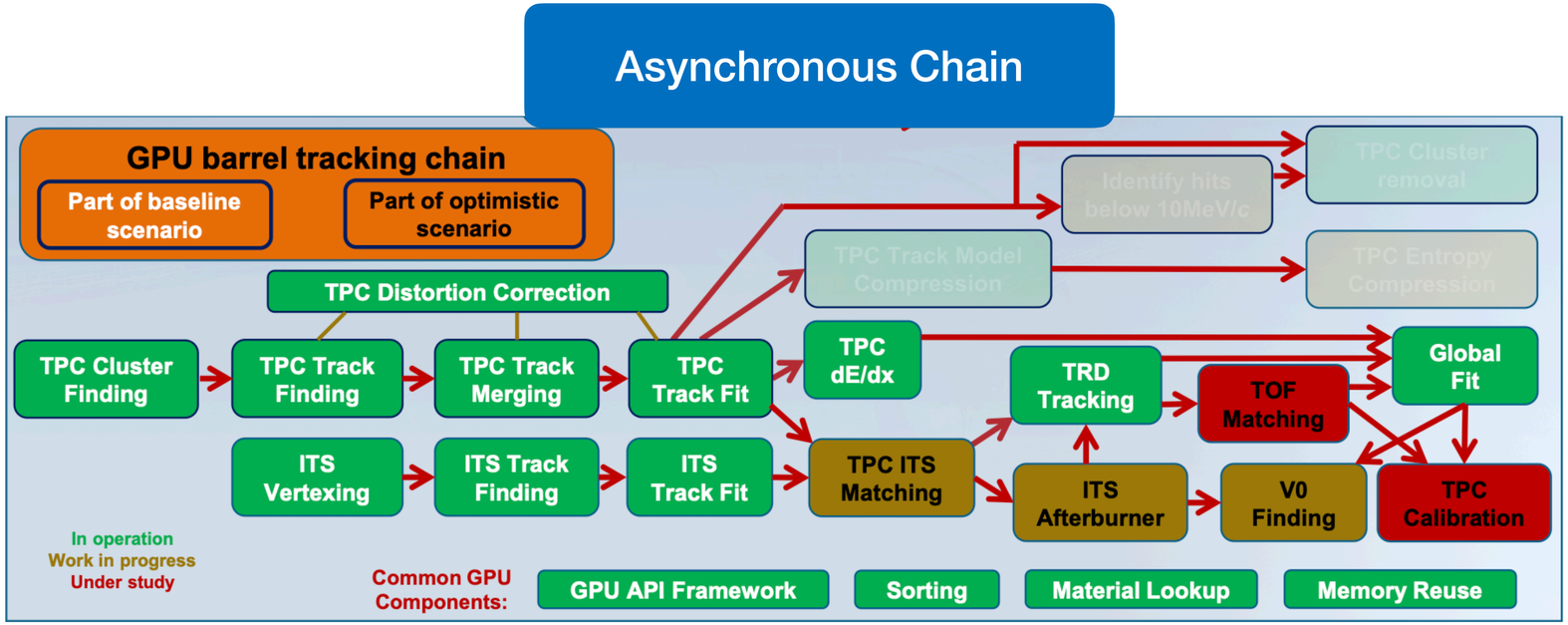
.

GPU barrel tracking

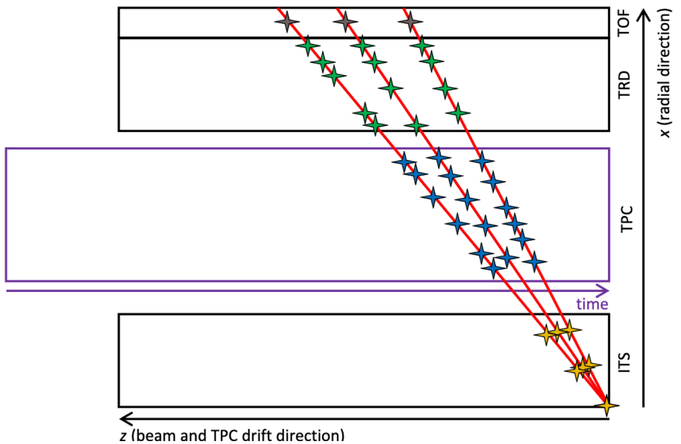


- Mandatory baseline scenario has been implemented
  - TPC tracking and data compression during synchronous reconstruction.
- Optimistic scenario
  - port full barrel tracking to GPU

GPU global tracking in asynchronous reconstruction

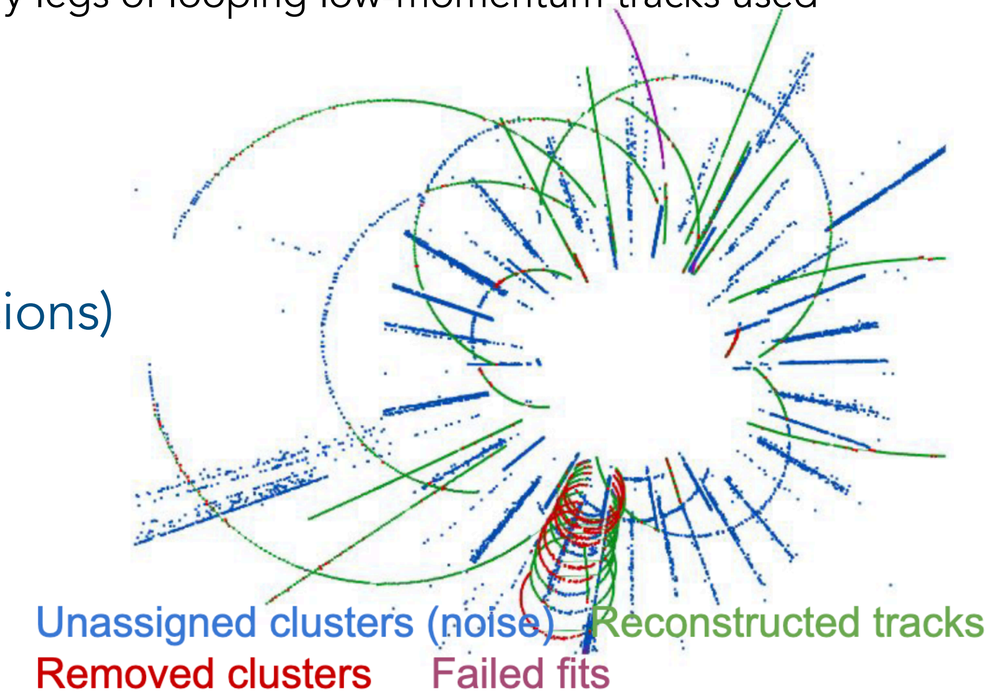


- Steps still to be implemented for asynchronous reconstruction on GPU (existing on CPU)
  - Matching to ITS
  - Matching to TOF
  - **Secondary Vertexing**
  - TPC interpolation for SCD calibration



TPC data reduction: Option A

- Reject
  - Clusters from background
    - ex. noisy pads or charge clouds related to low momentum protons
  - Clusters that are associated to or are in the proximity of background tracks.
    - ex. tracks from very low momentum particles spiralling around the magnetic field lines, track segments with large inclination with respect to the TPC pad rows, and clusters from secondary legs of looping low-momentum tracks used for physics.
- Protect
  - Clusters in a tube around good tracks are protected.
- Average size/Pb-Pb collision: 3.7 MB (74 PB for $2 \cdot 10^{10}$ collisions)



TPC data reduction: Option B

- Keep
  - Clusters that are attached to, or in the proximity of **identified good tracks** that may be used for physics analysis.
- Expected average data size per collision: < 2.4 MB

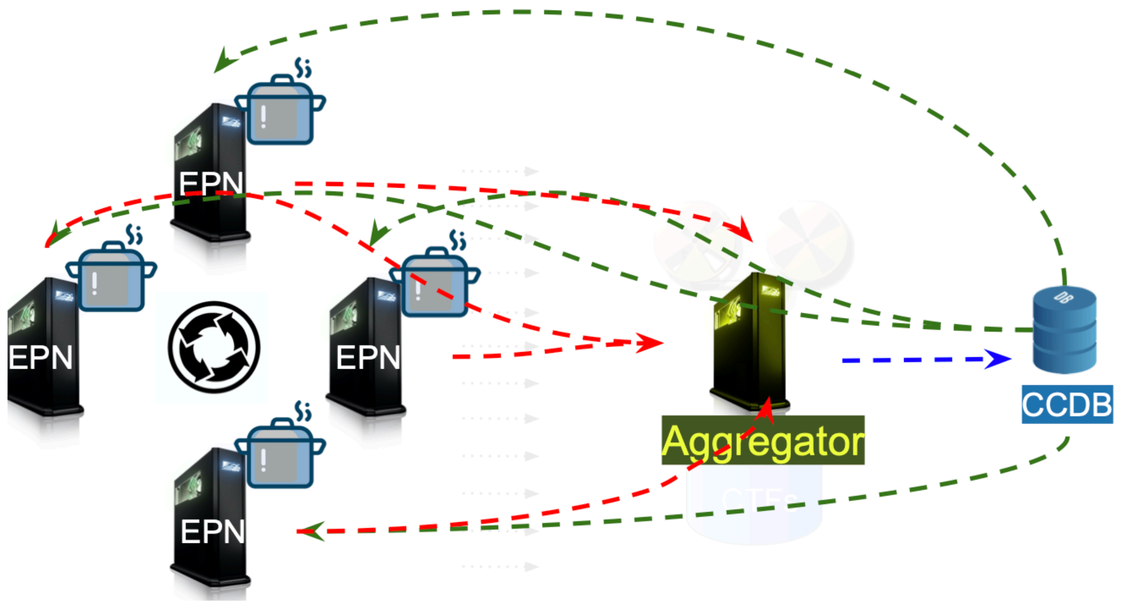
- Option B yields lower data size
- However, it bears the risk of losing tracks in case calibration is not good enough.
- Currently ALICE is still using Option A
  - Reduction and physics performance with real data under investigation

TPC data compression

- Goal: Efficient storage of TPC raw cluster coordinates
- Entropy reduction
  - Coordinates of hits that are **not assigned to tracks** are sorted by geometrical coordinates and the difference to the previous hit is stored.
  - Raw coordinates of hits **assigned to tracks** are stored relative to the extrapolated track
  - Encode cluster properties (maximum charge, total charge, and cluster size) together
    - profit from their correlation
- Data rounded to relevant precision and encoded with ANS entropy compression

Calibration in synchronous processing

General Scheme

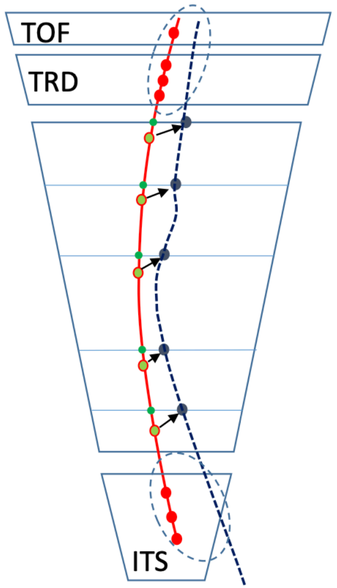


- Every EPN node produces compact data objects relevant for production of data objects for calibration
- Data is sent to dedicated **aggregator servers.**
- **Time slots** with granularity characteristic for each calibration type is attributed to this data.
- **CCDB object** is created once enough data has been accumulated.
- During synchronous processing, also input data are accumulated for those calibration constants that need a large amount of data or are too demanding to be determined synchronously.
  - The corresponding calibration information is extracted before the asynchronous reconstruction takes place, and the CCDB is updated.

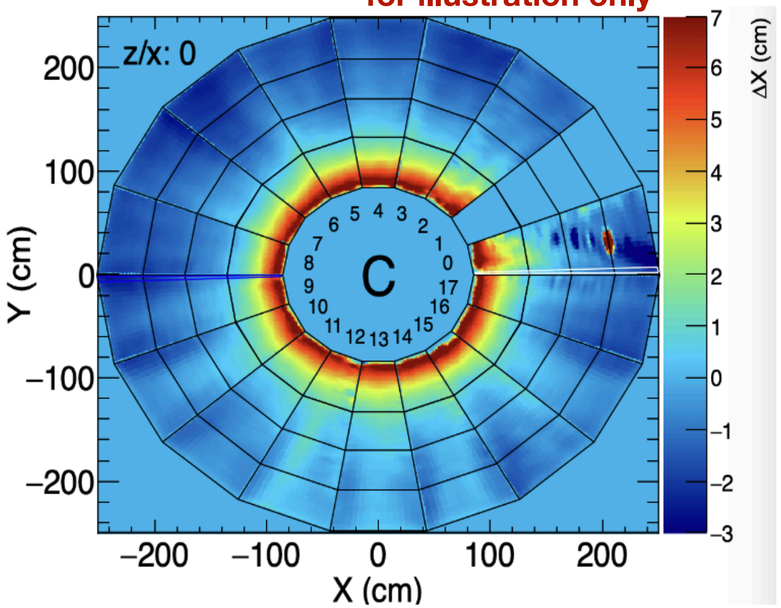
.

Global tracks for calibration data

SCD calibration (baseline scenario)



for illustration only



- Full tracking including all barrel detectors for subsample of tracks selected from peripheral collisions (overall < 1%).
- Residuals between refitted ITS-TRD-TOF track segments and TPC clusters are used to create 3-dim SCD maps
  - granularity of 1-2 minutes in Pb-Pb and 10 minutes in pp collisions.
- These maps together with the TPC integrated digital currents recorded during synchronous processing become part of the calibration used in asynchronous (offline) processing.

.

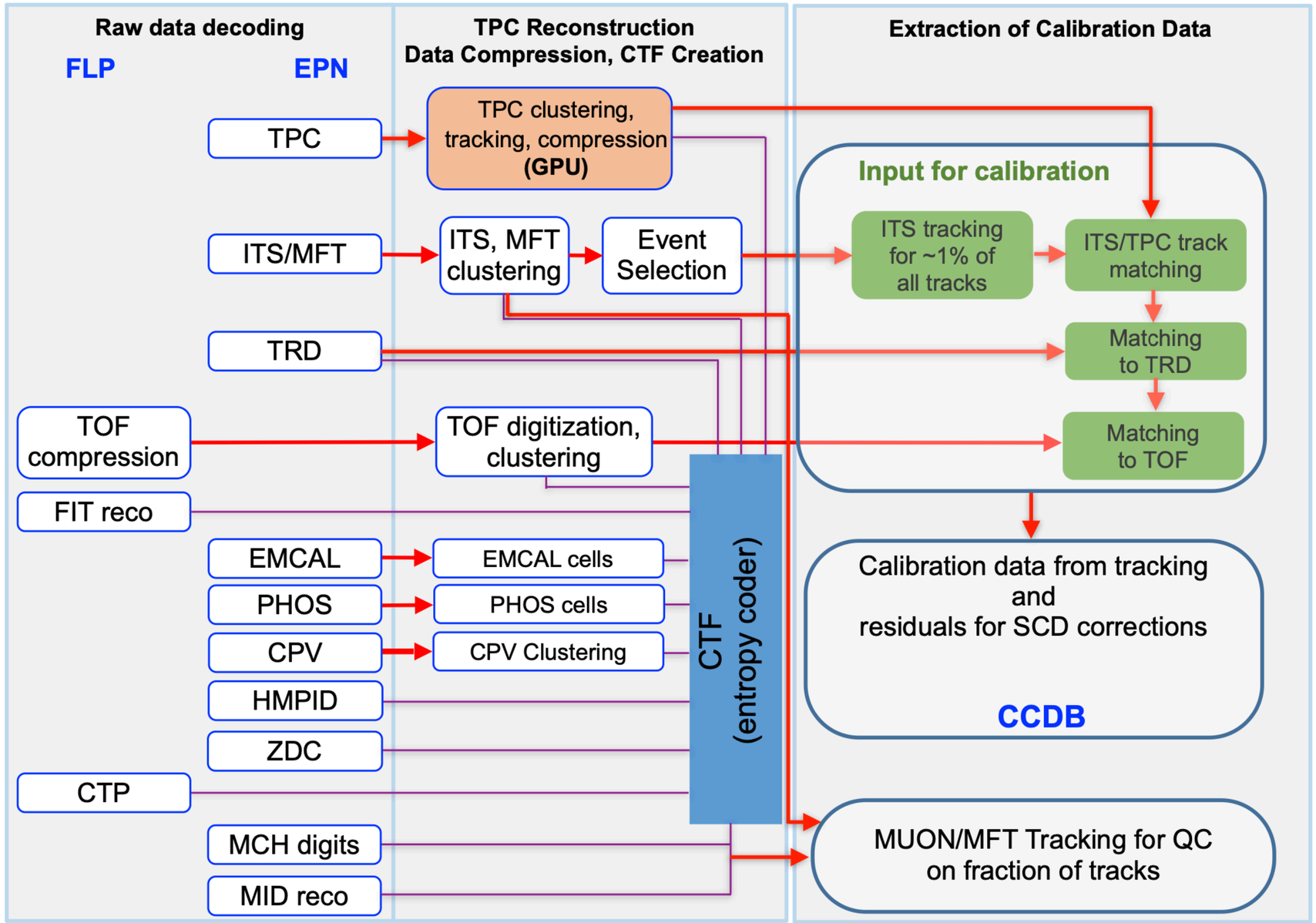
Global tracks for calibration data

Other tasks

- Global barrel tracks also used for
  - TPC drift time and TRD calibration (gain, t_0, ExB, drift velocity)
  - LHC clock drift from tracks matched to TOF
  - TOF channel level time offset
- Interaction region, calorimeter bad channels, and gain parameters …

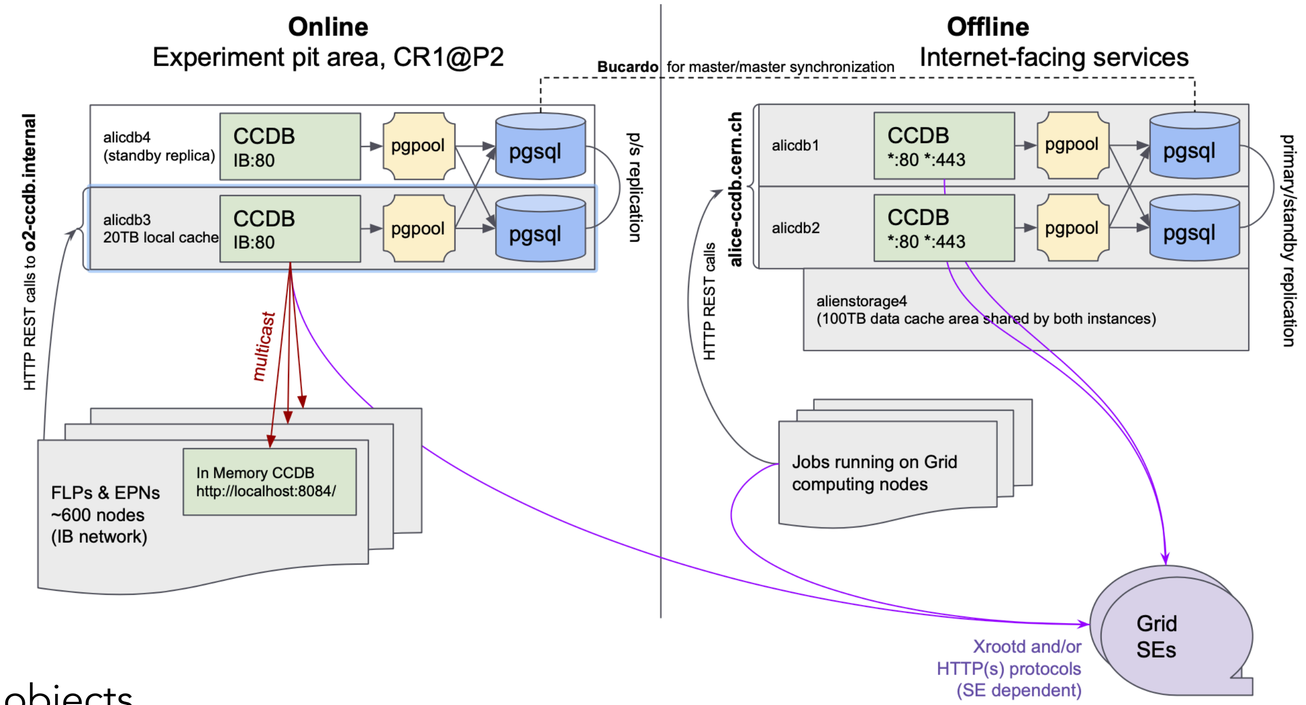
.

Synchronous reconstruction workflow



Condition and Calibration Data Base (CCDB)

- **PostgreSQL database with two instances (Online / Offline)**
- **Online service**
  - Synchronous processing **pushes** objects to the online instance
  - **Distributed** immediately to all IB-equipped nodes …
    - via multicast
- **Online-Offline independent operation via:**
  - loosely coupled master-master database
  - replication and large caches on either side
- **QCDB instance (Quality Control):**
  - Same CCDB engine with Grid replication for a curated list of objects

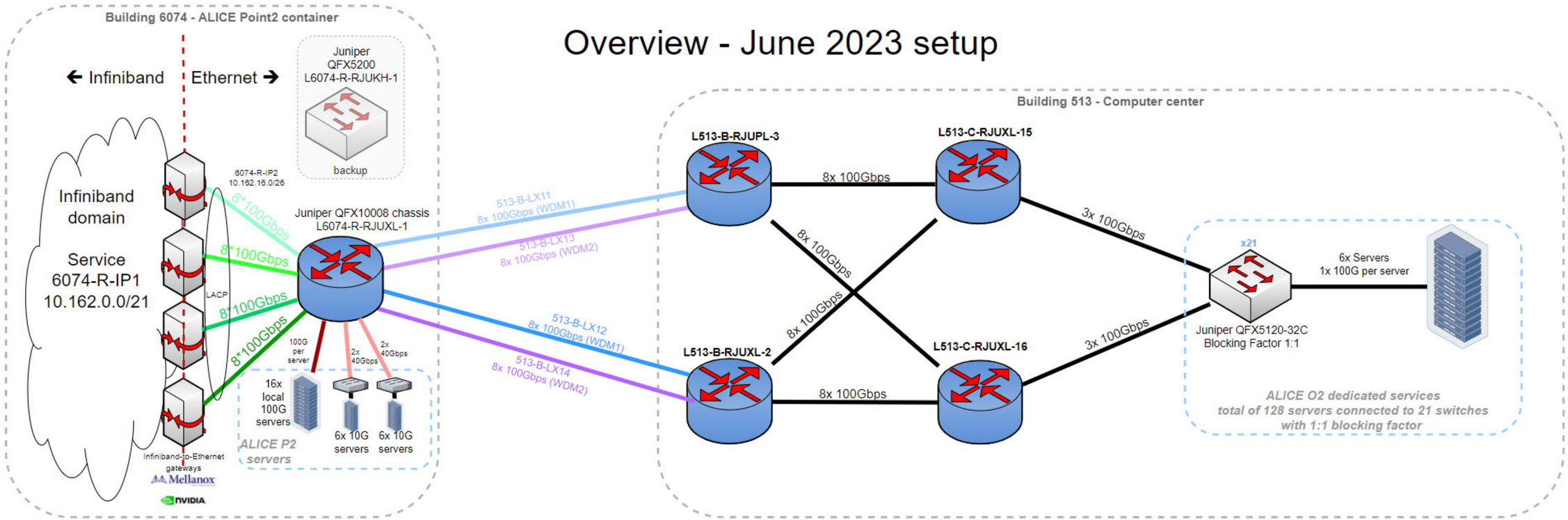


- **Some Figures:**
  - 1.8 TB of data for 8.2 M calibration objects
  - 1.3 kHz of requests to the Offline instances (1week average),
    - with long periods of 7 kHz of requests (depends on the running job mix)
  - Average response time: 6.7 ms
  - New object rate in PbPb was 0.35 Hz
  - Active object set in the Online nodes is 435 MB in 249 objects

A CERN IT - ALICE collaboration

- Need for a new robust and scalable data storage system for the ALICE upgrade
  - Unprecedented projected data rates from P2: ~130 GB/s in Pb-Pb and ~90 PB of data in 1 ½ months of data taking
  - Offline calibration and processing requires keeping the data on disk for extensive periods of time …
  - … serve it at speeds up to 250 GB/s
  - Conventional large local data buffer + transfer to disk + tape to Computing Center
    - too complex, too expensive and actually lacking the necessary performance
- Solution: a centralised storage instance with high I/O throughput and large capacity in Meyrin CC
  - Leverages the latest advancements in network fibre optics (**WDM** next slide) and in-house developed storage technology
  - Data flow and retention maintained seamlessly through **common ALICE-CERN IT software**

High throughput redundant network service



- Two independent Wavelength Division Multiplexing (WDM) systems using 2 pairs of fibres
  - Total throughput 3.2 Tbit/s (400 GB/s)
- Provides redundancy in case of one fibre pair or CC router failure - 1.6 Tbit/s (200 GB/s)
  - Spare router at P2
- Fully operational for p-p and Pb-Pb data taking periods

High throughput and redundant EOS-based storage

- EOS based instance in Meyrin CC with 150 PB capacity (EOSALICEO2)
  - Largest EOS instance @CERN
  - 12000 hard drives spread over 126 servers
- Spare instance at P2 with 14 PB capacity (EOSP2)
  - In case of total network or CC storage failure
  - Capable of storing up to 3 days of Pb-Pb data
  - Automatic switch-over and data transfer to main instance upon network recovery
    - using ALICE developed eps2eos suite

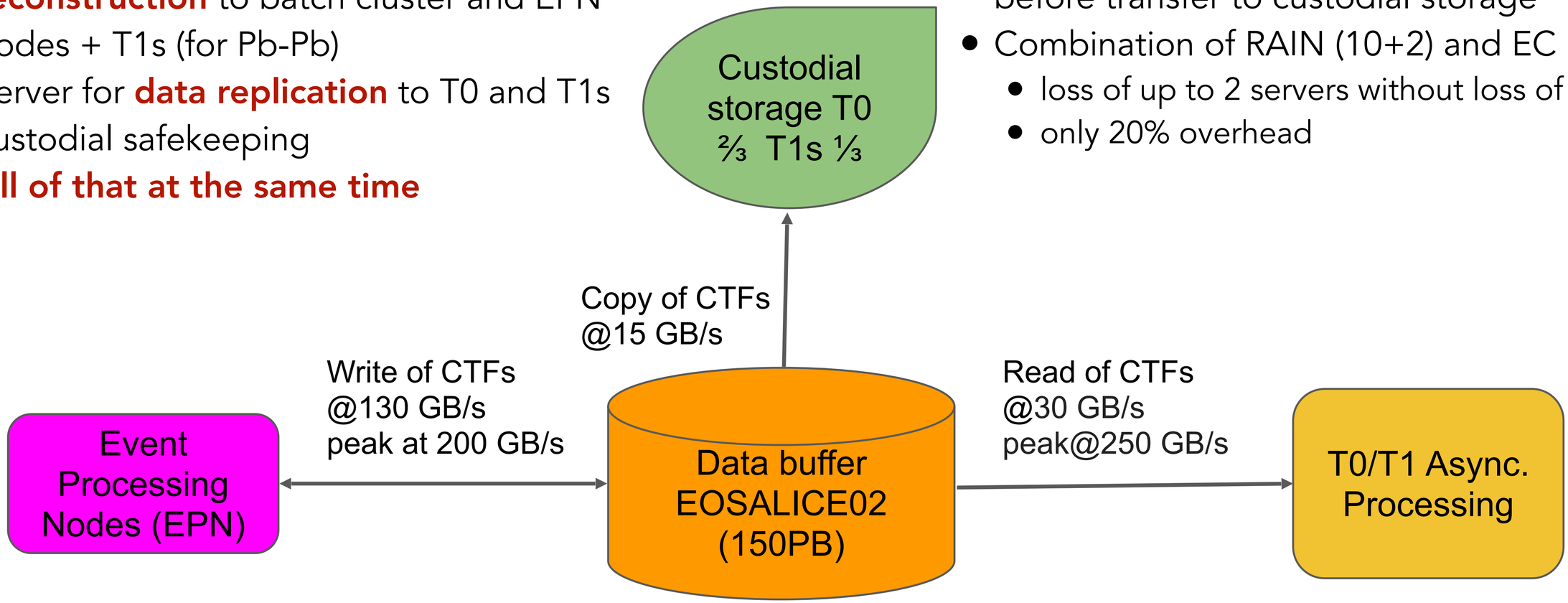
Main storage - the EOSALICEO2 instance

Complex function:

- High throughput storage for **data taking**
- Server for data for **asynchronous reconstruction** to batch cluster and EPN nodes + T1s (for Pb-Pb)
- Server for **data replication** to T0 and T1s custodial safekeeping
- **All of that at the same time**

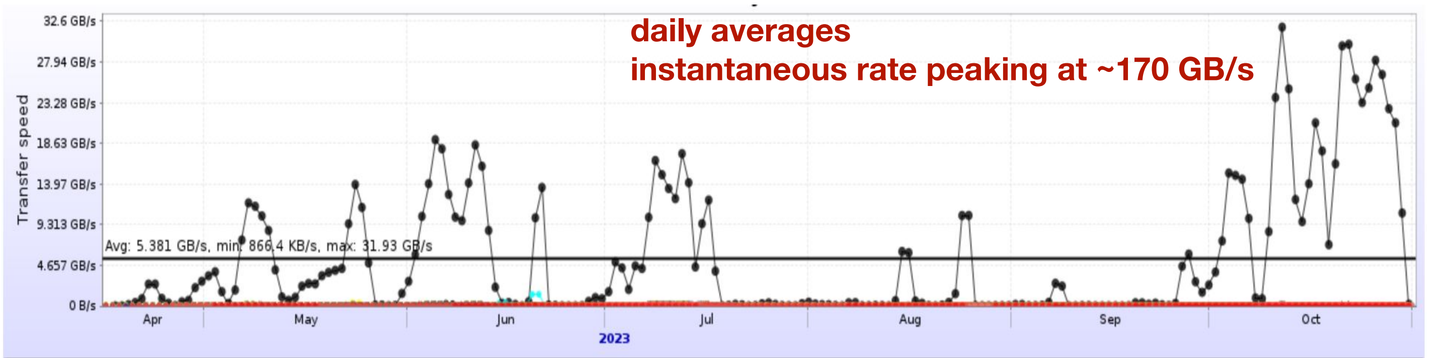
High redundancy:

- Precious physics and calibration data with lifetime on instance of up to one year before transfer to custodial storage
- Combination of RAIN (10+2) and EC
  - loss of up to 2 servers without loss of data
  - only 20% overhead



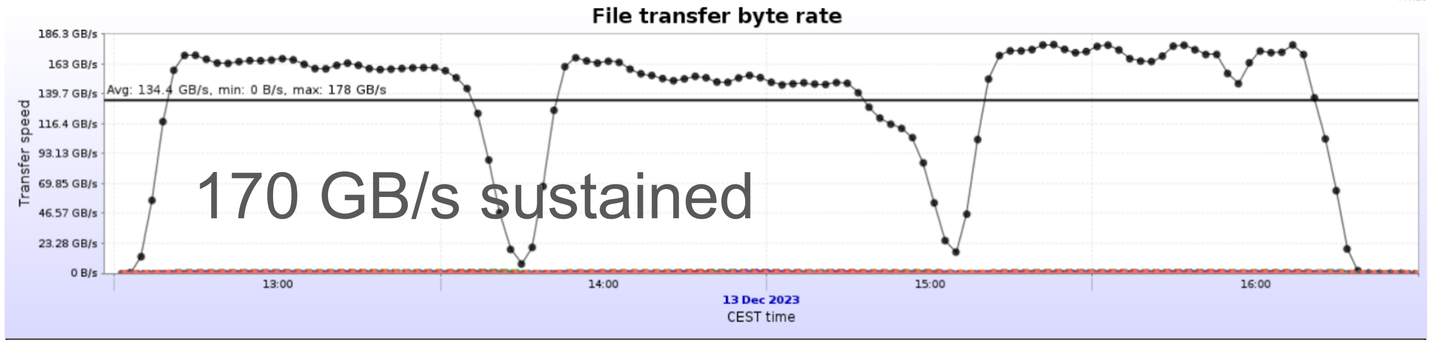
Performance

Data rates: 2023 pp and Pb-Pb



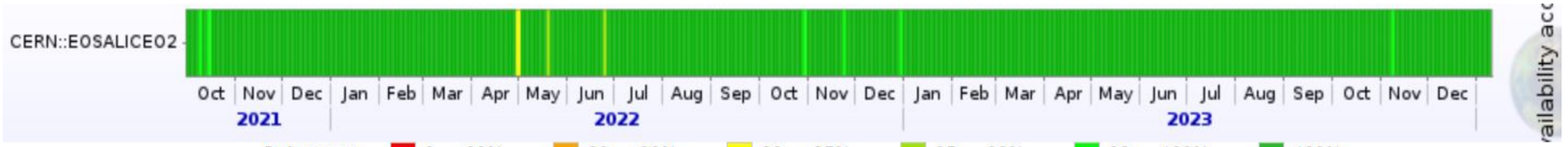
December '23 full chain P2-CC test

- Internal CC test: 380 GB/s



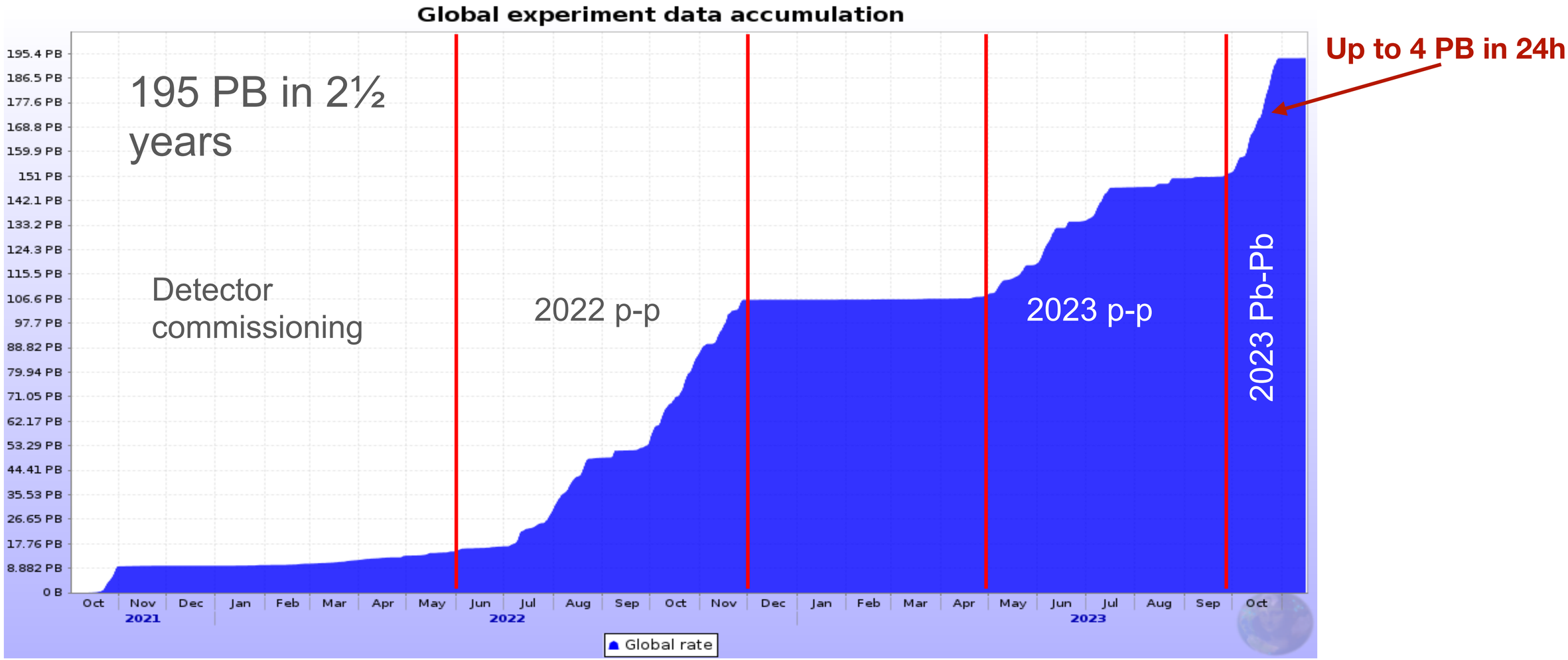
Reliability and data safety

- 99.93% over 2½ years!
- 100% during data taking
- No data loss



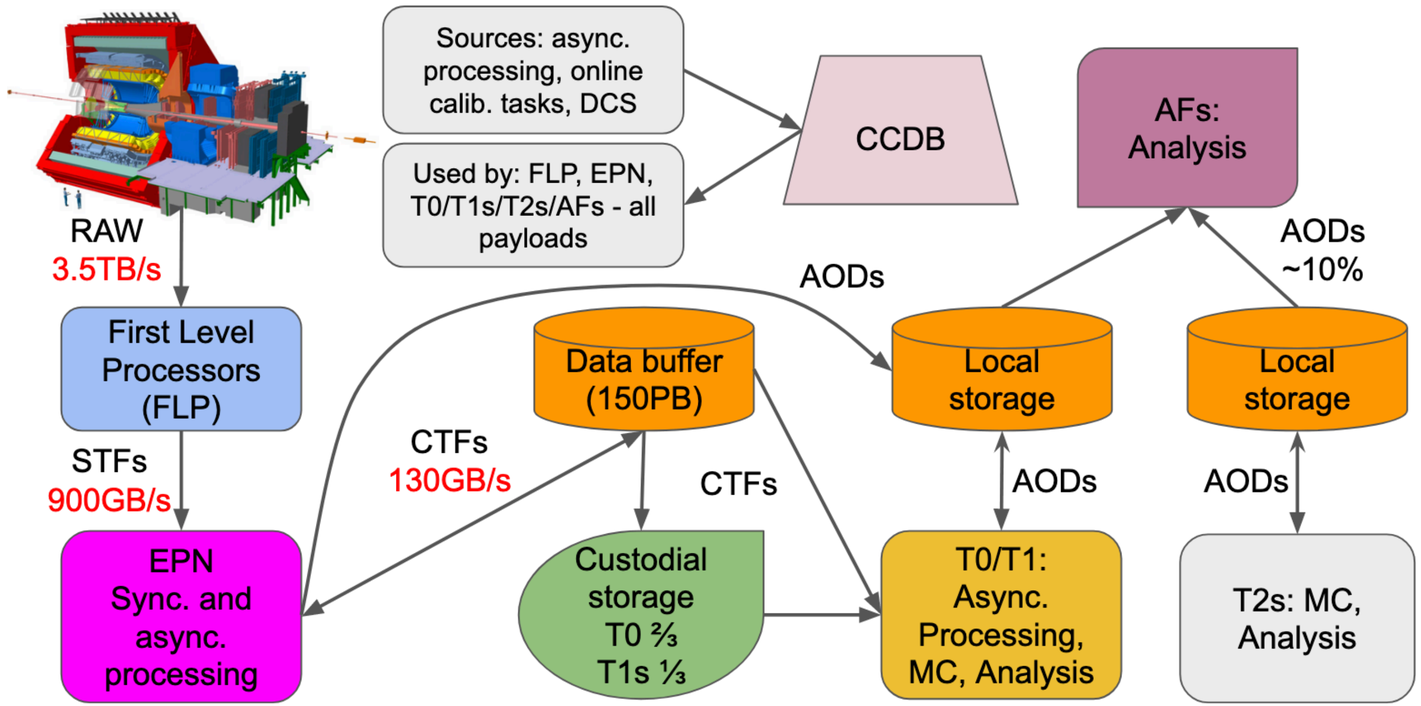
Link name	Data		Individual results of reading tests			Overall
	Starts	Ends	Successful	Failed	Success ratio	Availability
CERN::EOSALICEO2	01 Oct 2021 00:00	10 Jan 2024 23:46	19962	15	99.92%	99.93%

CTF and calibration data on EOSALICEO2



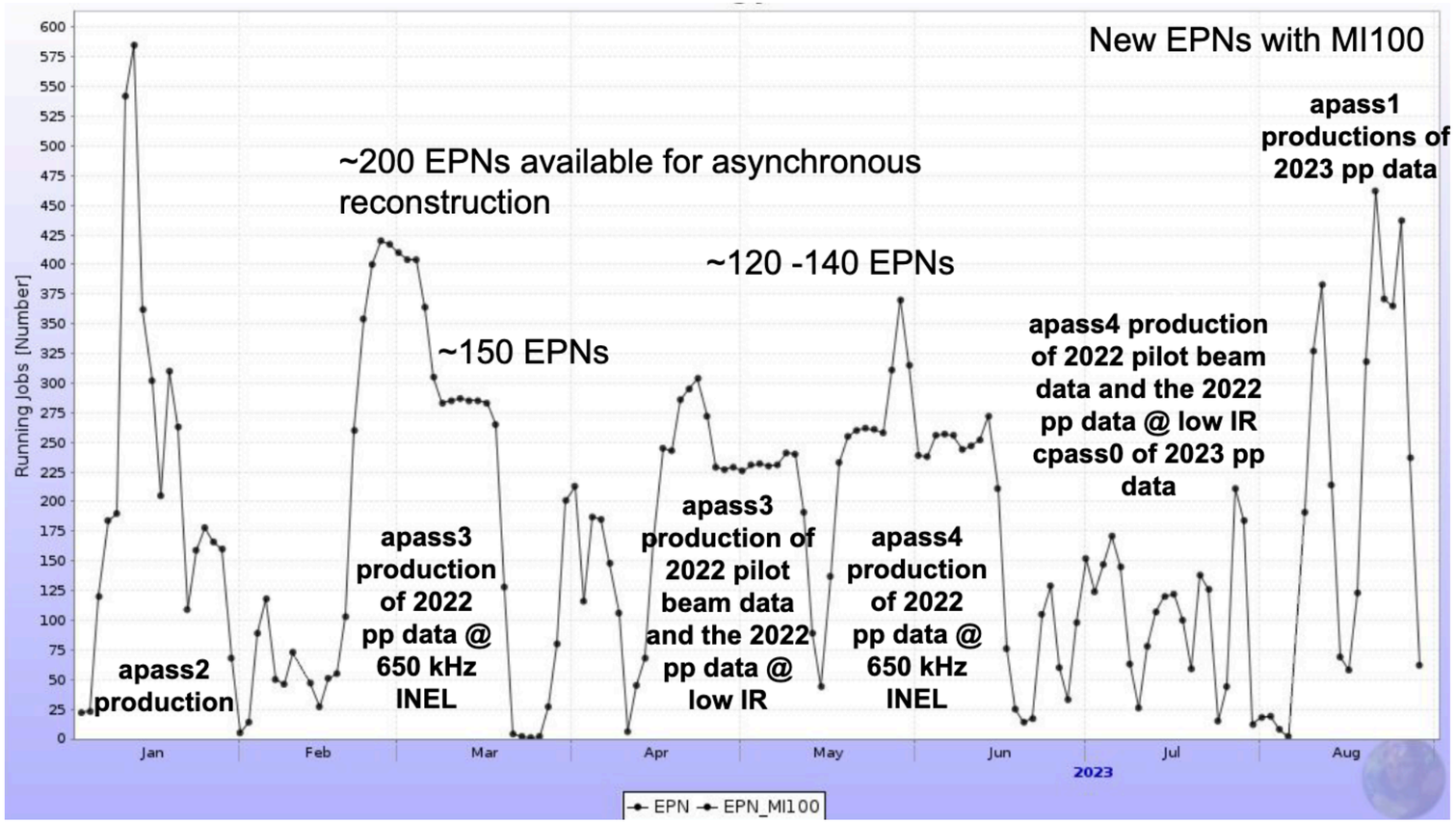
Asynchronous reconstruction

- Reconstruction passes after data taking on CTFs
- Full reconstruction and calibration for all detectors
  - Full TPC SCD calibration including fluctuations
  - TPC-ITS matching
  - track propagation to outer detectors
  - Global track fits
  - Primary and secondary vertex reconstruction
  - Particle Identification hypothesis
- Output: Analysis Object Data (AOD)
  - ~ factor of 7-10 smaller than CTFs



- EPN farm used for asynchronous reconstruction when not used / fully used for synchronous processing
  - insures 100% duty cycle
- Rel. compute time for TPC reconstruction smaller in asynchronous reconstruction
  - GPU processing less important, but still could profit from moving barrel processing to GPU
  - For Pb-Pb at present fully CPU bound since dominated by large combinatorics for secondary vertexing.

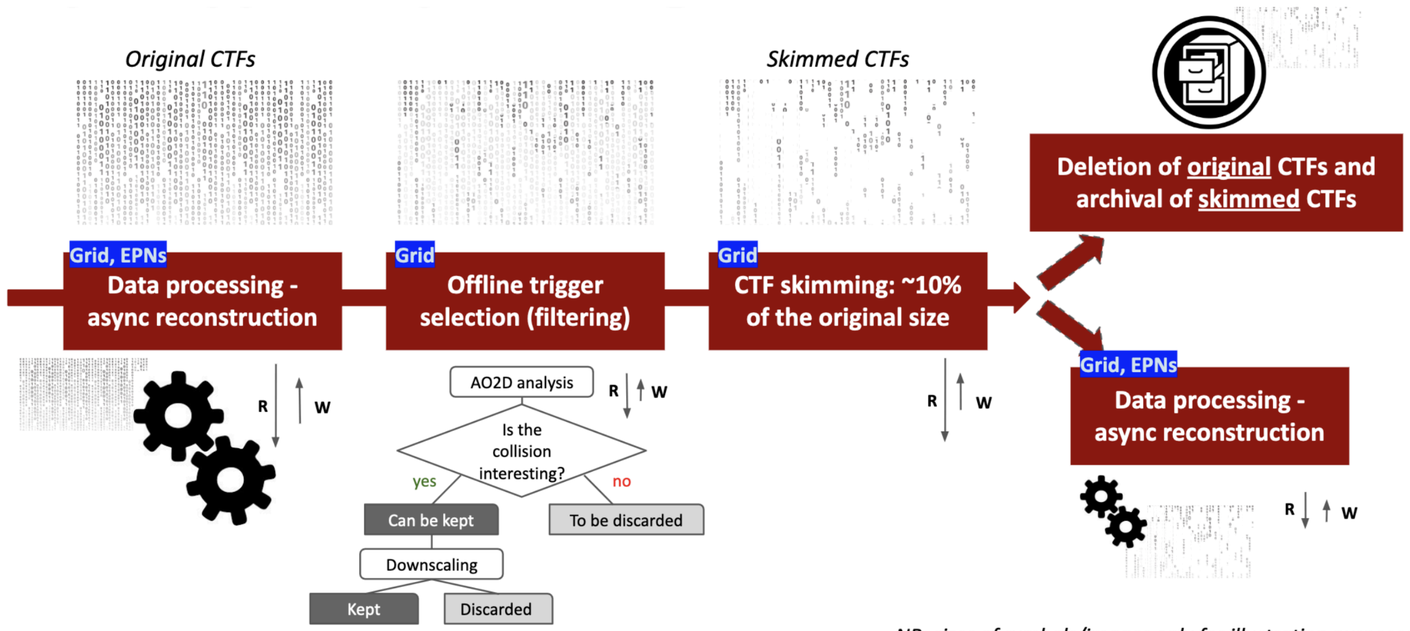
Asynchronous reconstruction on EPN



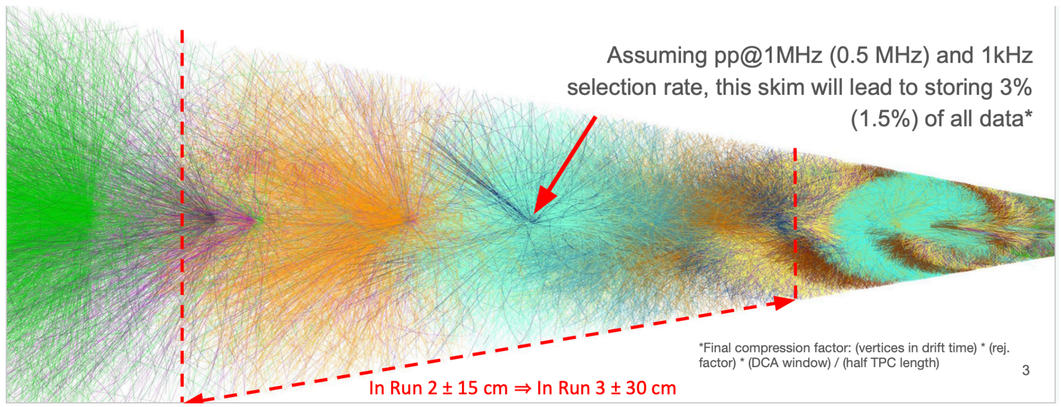
- Current configuration for pp: 1/2 EPN node (32 cores, 256 GB memory in same NUMA domain with 4 GPUs)
- 1/3 (2/3) of 2023 pp processing on EPN (Grid)
- GPU speed-up x ~2.5 rel. to GRID node (normalised to 8 core)
- Recent 2023 Pb-Pb asynchronous pass1: 56% on EPN (20% of data in ~ 1 week); EPN ~1.3x faster

An additional data reduction step for pp

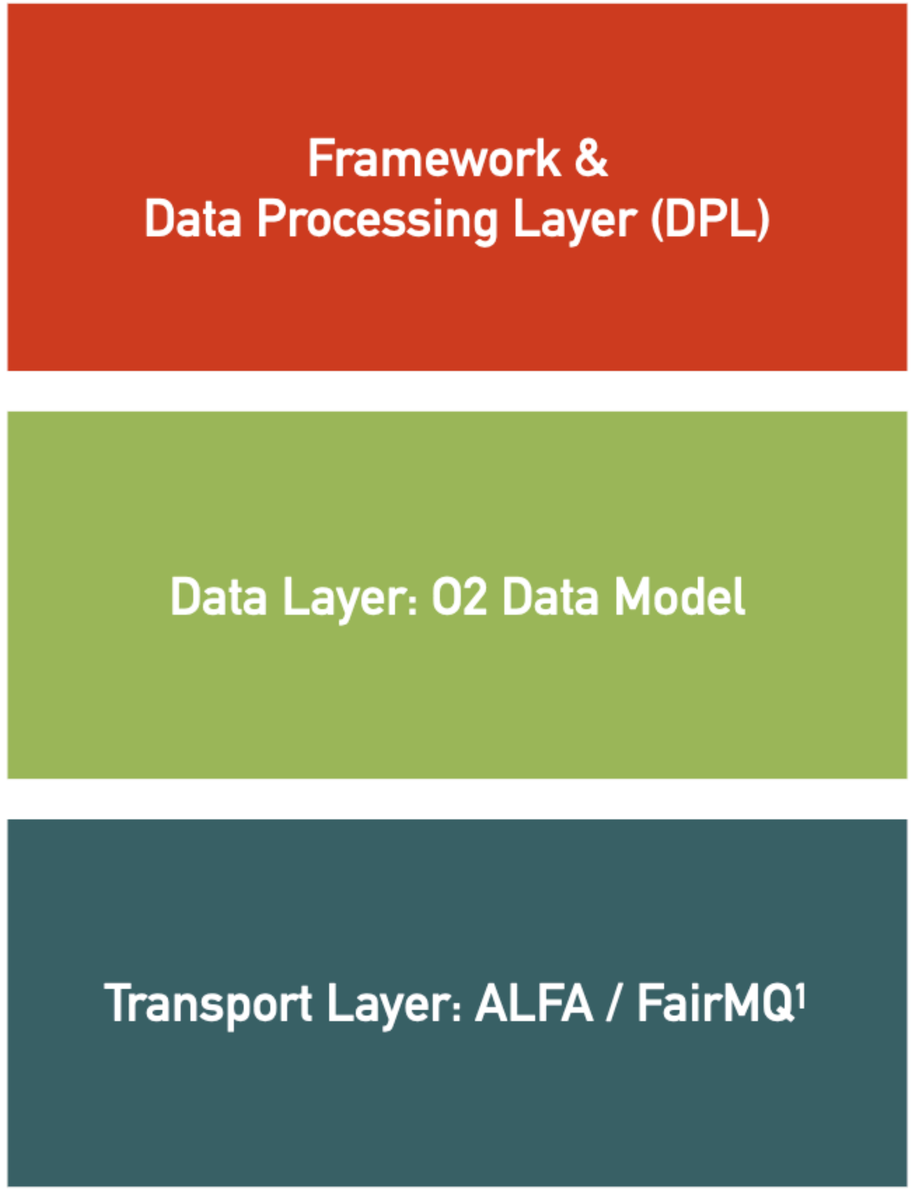
- pp data taking during standard LHC operation at max. rate (0.5 - 1 MHz) creates data volumes similar to Pb-Pb
  - for which ALICE does not have the storage resources.
- Strategy: reduce data volume by more than factor 10 through offline trigger selection and filtering out the interesting collisions
- Since disk buffer has to be freed before Pb-Pb data taking, this process has to run in parallel with data taking
- Caveat: need to store a fiducial volume to store also clusters adjacent to tracks belonging to the interesting collision and secondary vertices



NB: sizes of symbols/images only for illustration purposes



Software Framework

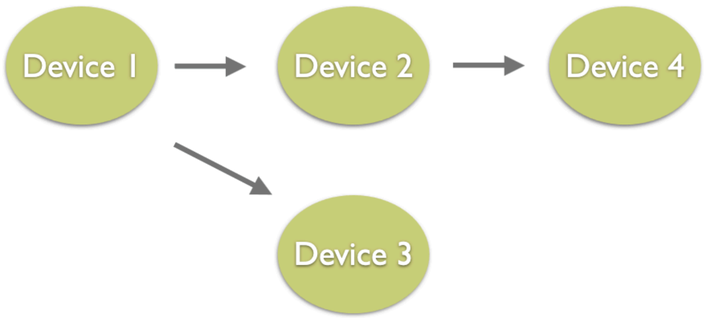


Software Framework

Transport Layer: ALFA / FairMQ[1]

Joined collaboration with GSI and FAIR

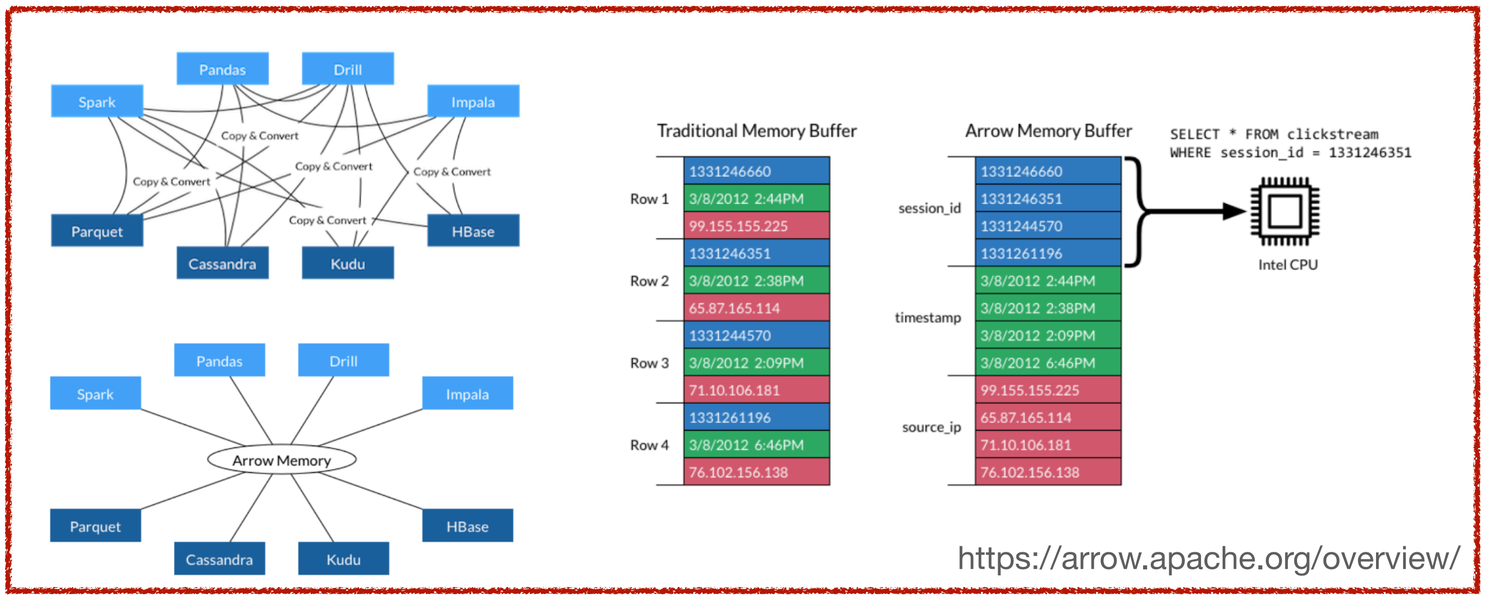
- Uses standalone processes (devices) communicating via messages.
  - On same device communicate via pointers to shared memory
  - Seamless remote communication



Software Framework

Data Layer: O2 Data Model

- Message passing aware data model with support for multiple backends:
  - Simplified, zero-copy format optimised for performance and direct GPU usage.
  - ROOT based serialisation. Useful for QA and final results.
  - Apache Arrow based.
    - Backend of the analysis data model using column-wise table
    - for integrating with other tools.



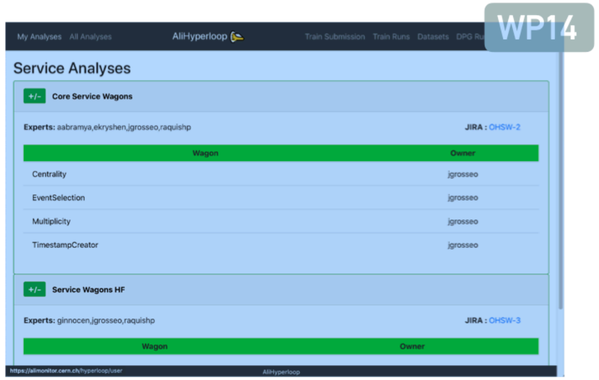
Software Framework

Framework &
Data Processing Layer (DPL)

- Hides the details of a distributed system, presenting a familiar "Data Flow" system.
  - **Has reactive-like design** (push data, don't pull)
  - **Uses implicit workflow definition** via modern C++ API.
  - **Provides Core common tasks:** topological sort of dependencies, deployment of generated topologies, data lifecycle handling, service management, common infrastructure services, plug-in manager.
  - **Integration** with the rest of the production system, e.g. Monitoring, Logging, …

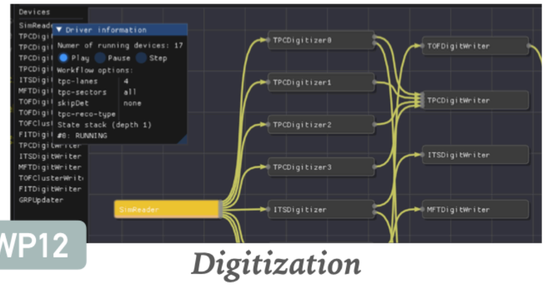
Data Processing Layer (DPL)

- Data Processing Layer (DPL) acts as an integration platform for many activities
- Plays a crucial role in leveraging parallel data processing

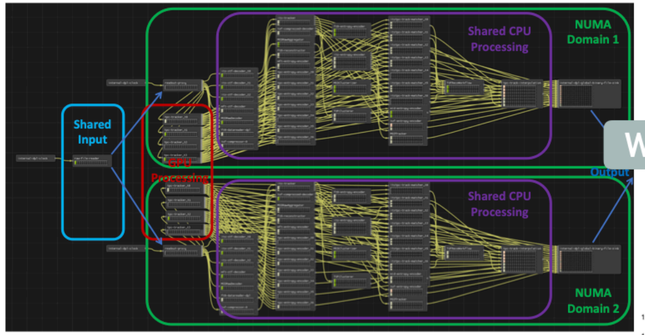


AliHyperloop Integration

DPL workflows in O2 are automatically integrated within the new "AliHyperloop" Analysis Train service

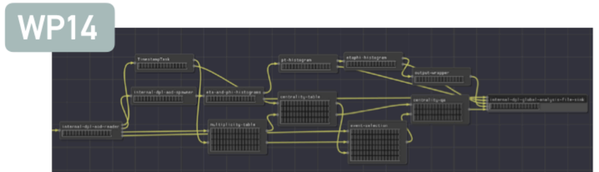


Digitization



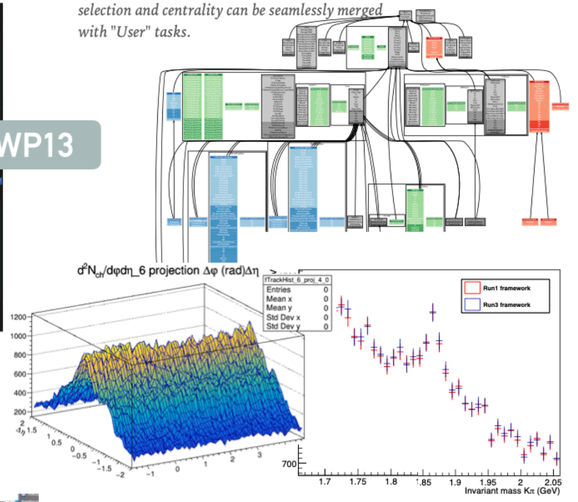
Full system test

DPL provides the backbone of the full system integration, in particular successfully allowing 8 GPU processes to share the CPU part via DPL "time pipelining" feature (courtesy of David).



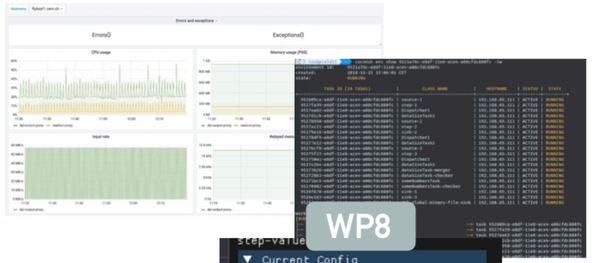
Analysis Framework

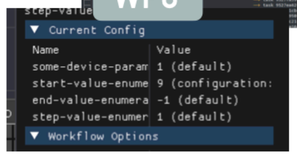
The AliceO2 Analysis Framework is based on DPL and uses its features to provide parallelism and describe deployments. Common tasks like event selection and centrality can be seamlessly merged with "User" tasks.



Real physicists!

Analysis Framework is enabling real physicists to produce actual plots! :-) See report on the ongoing analysis challenge.

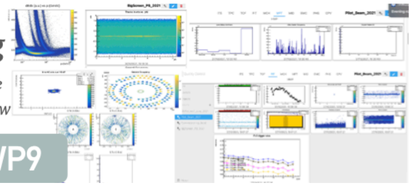




Integration with WP8 provided Monitoring, InfoLogger, Configuration & Control packages

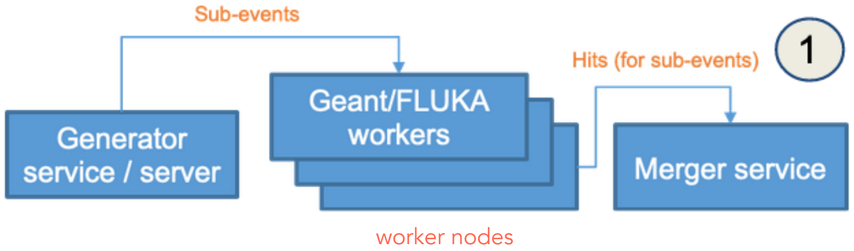
QC & DataSampling

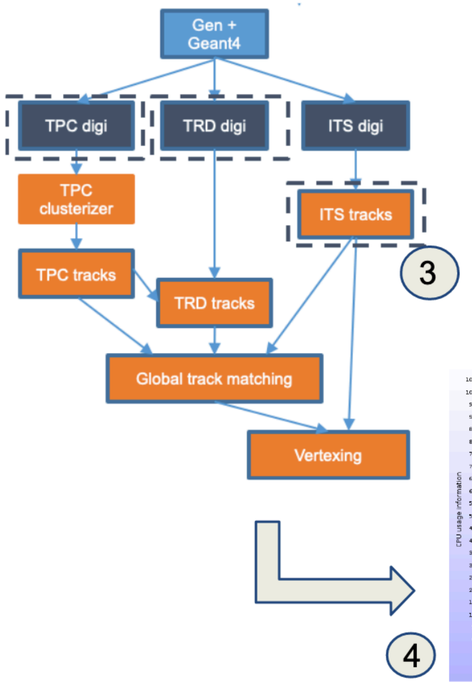
DataSampling & QC use DPL to integrate with the rest of the data processing workflow (courtesy of Daniele & Barth)

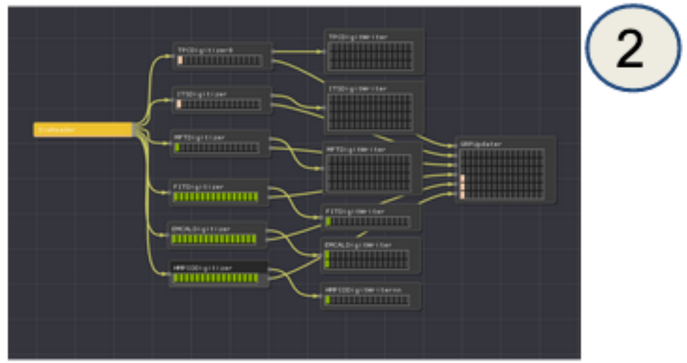


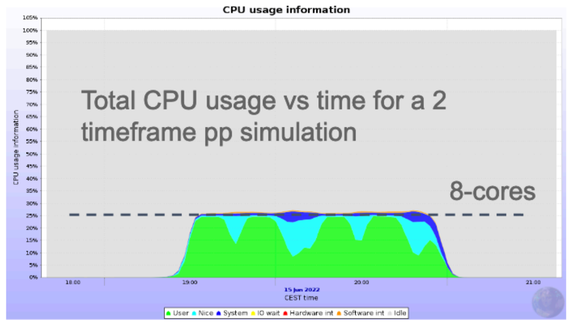
Simulation framework highlights

- **Detector simulation framework**
  - multi-core and sub-event parallel simulation
  - for any transport engine interfaced via Virtual Monte Carlo
    - GEANT4 (default), GEANT3, FLUKA
    - using exactly the same user code
- DPL-based **digitisation framework** handling Time Frames and performing signal to background embedding
  - compute time gain by using background several times
- Graph-based **MC workflow description**
  - integrate all processing steps from event generation to AOD creation in global workflow
- **Multi-core graph execution engine**
  - compute workflow in stages on the GRID
  - yields very good CPU efficiency



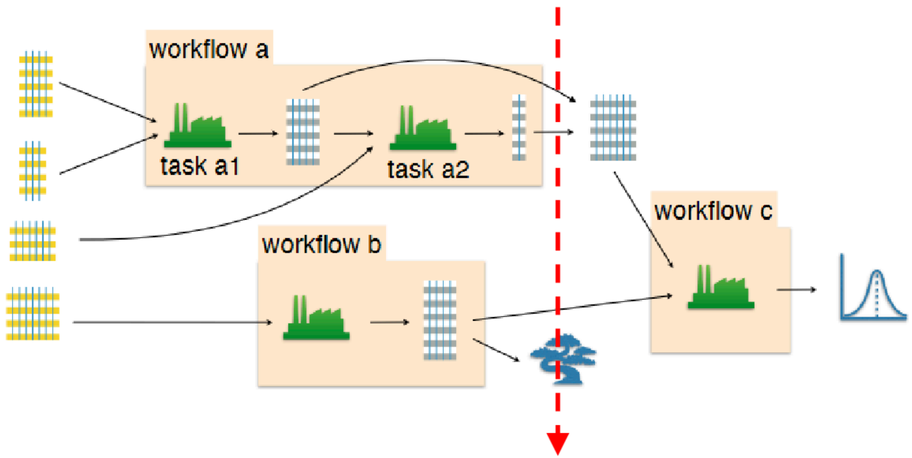


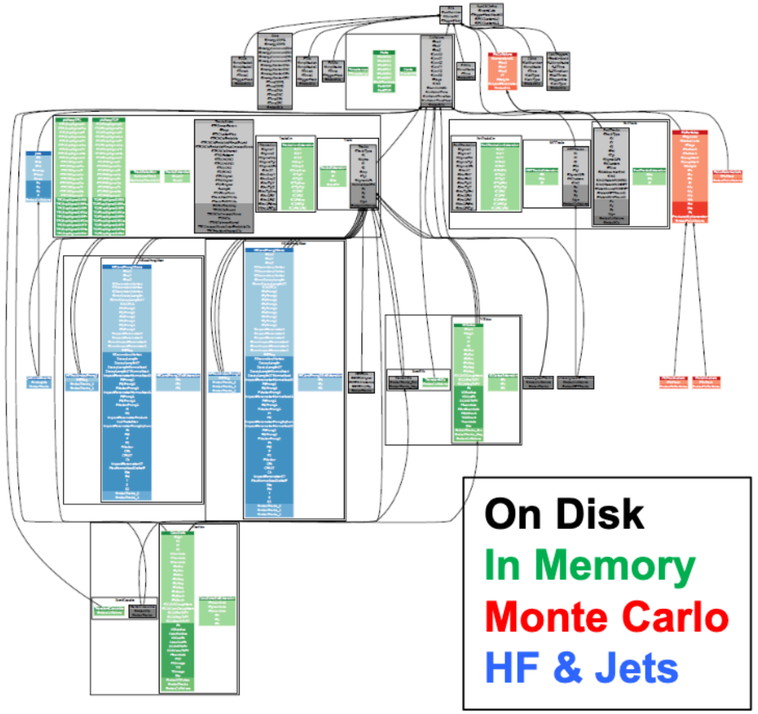


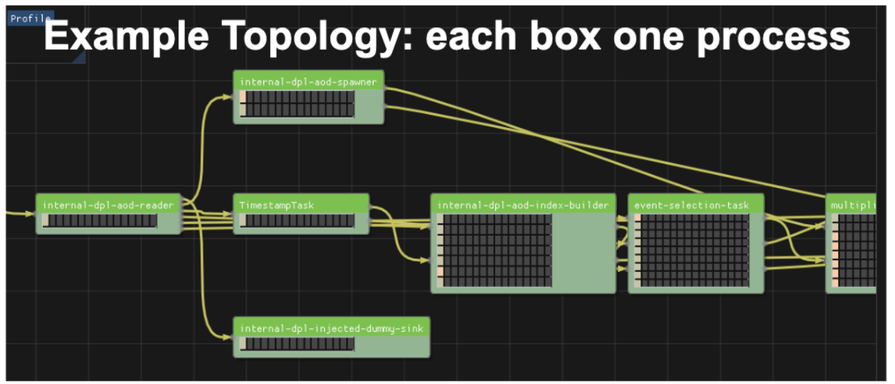


Analysis Framework: Modular and Multi-Core

- DPL based: analysis split into processing blocks and organized in workflows
  - Each block consumes trees/tables, produces histograms or trees/tables
  - Common service tasks (event selection, track selection, secondary vertex finder, …) reused
  - Information flow through tables between processes (zero copy, shared memory)
- Data model based on flat tables arranged in a relational database-like manner
  - minimises I/O costs
  - improves vectorisation / parallelism
  - High I/O throughput using ROOT bulk read
- Analysis code
  - Declarative (say what you want): allows for automatic optimisation
  - Imperative: specify the processing algorithm



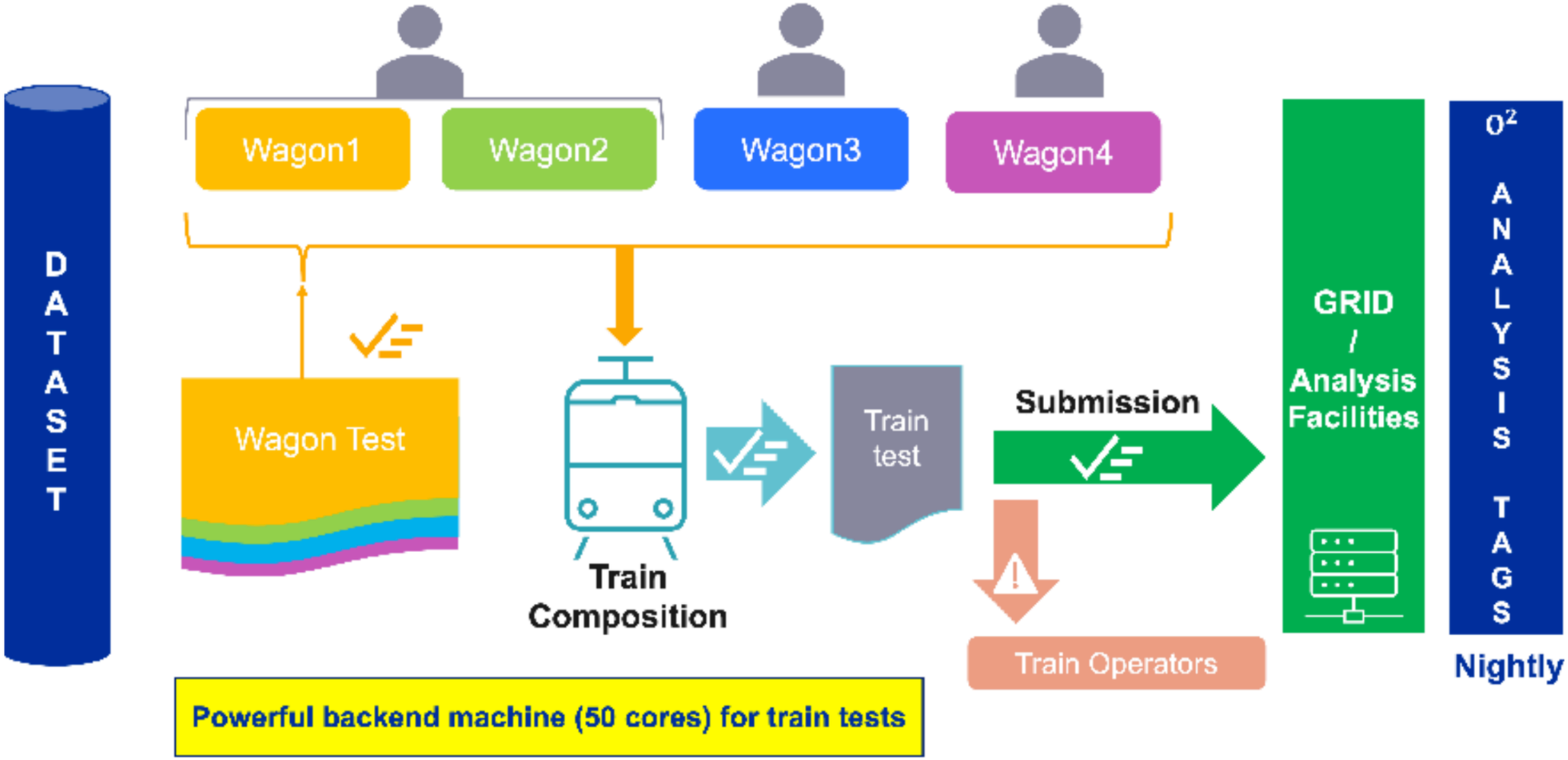




Analysis: Further Optimisation

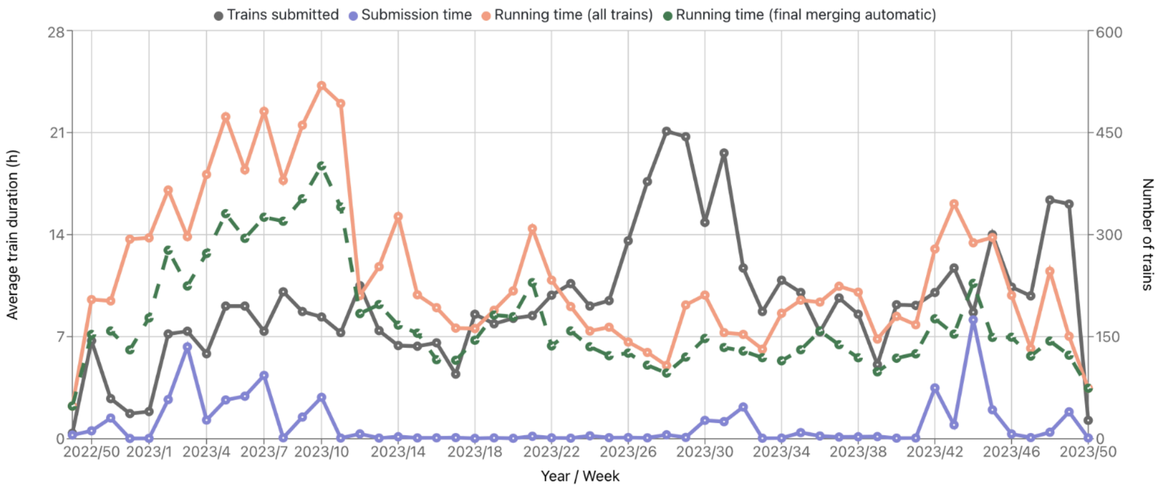
- Perform analysis on skimmed data sets
  - reduced information and / or collision selection
- 10% of AODs (from data and MC) copied to Analysis Facilities (AF)
  - for fast turnaround on validation and cut-optimisation
  - The AF needs to be able to digest more than 5 PB of AODs in a 12-hour period
  - 20000 cores and 5→10 PB of disk on very-performing file system
  - Currently: GSI, Wigner, Berkeley
- Full samples analysed on the GRID
  - Use organised analysis to minimise data access
  - Analysis Trains (Hyperloop)

Organised Analysis: Train Concept



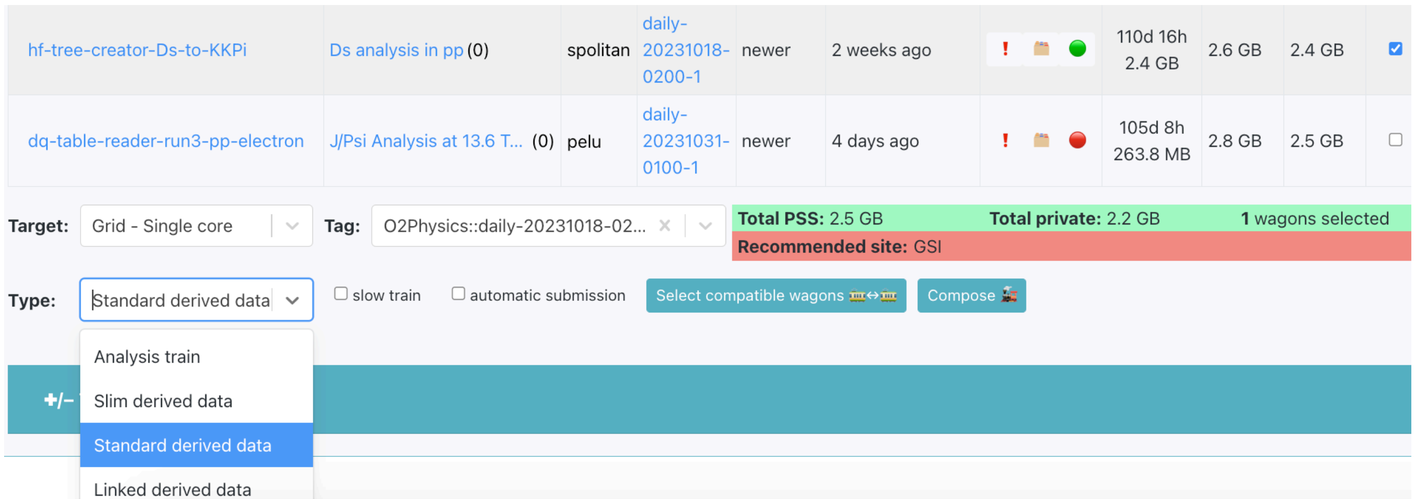
Organised Analysis: Hyperloop

- Organised analysis on GRID and Analysis Facilities
  - Run 3 data and converted Run 2 data
- Fully integrated with O2, allowing task configuration
- Individual workflows (wagons) are combined into trains
- Book-keeping
- 24/5 operator support (4 institutes in 3 different time-zones)
- Technologies:
  - Back-end: within MonALISA, Java-based model
    - Database: PostgreSQL
  - Front-end: JavaScript, React



Analysers: 228
Analyses: 315
Train runs: 10725
Jobs serviced: 32.4 M

Total CPU time: 4773y 262d 16h
CPU time / train: 162d 11h
Total input size: 367.6 PB
Input size / train: 45.1 TB



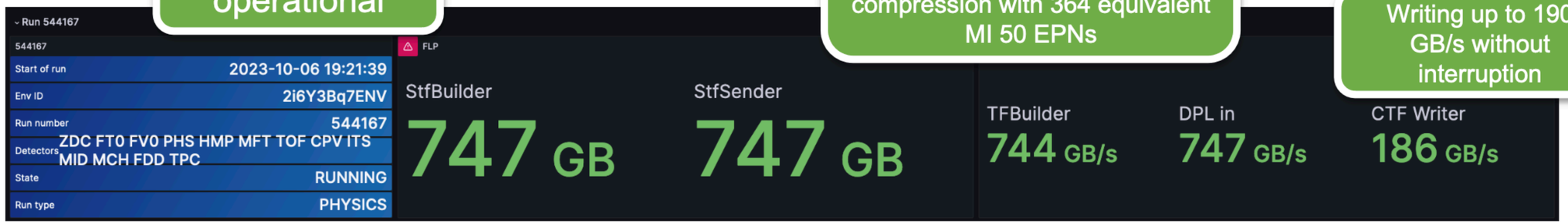
First Experience with high-rate Pb-Pb

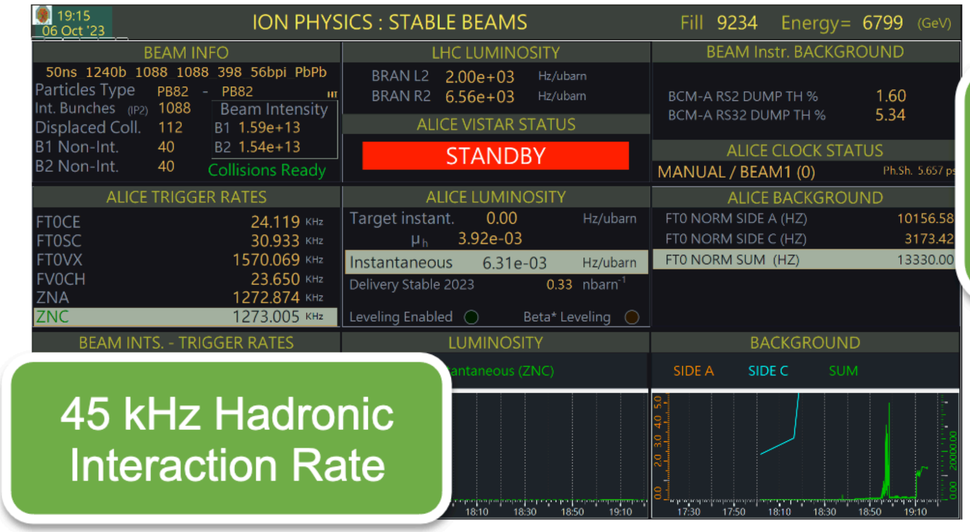
Heavy Ion : a success

Upgrade fully operational

15 detectors
Data Volume as predicted
Acquisition and online compression with 364 equivalent MI 50 EPNs

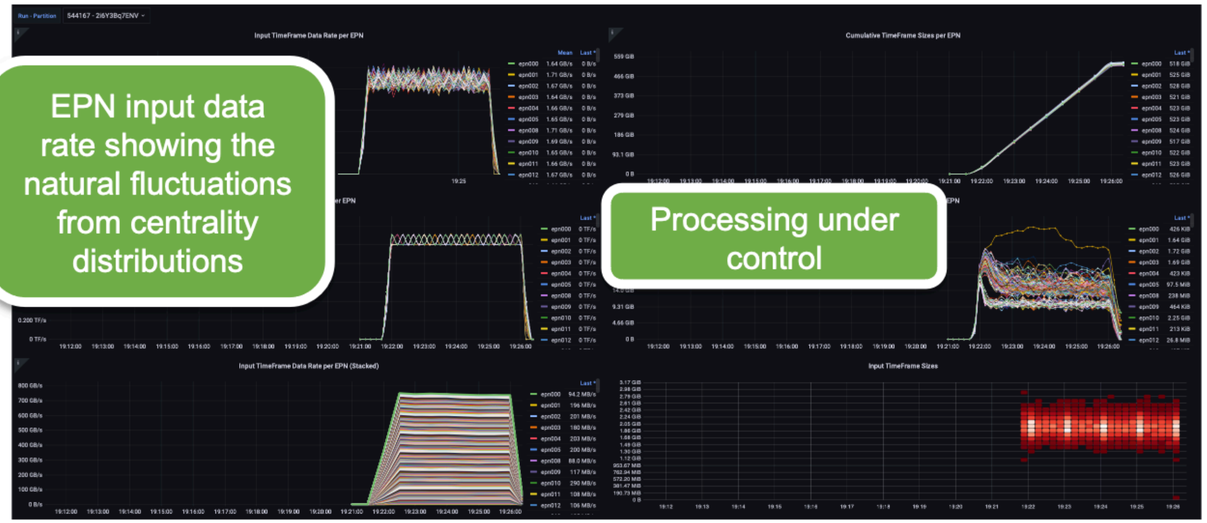
Writing up to 190 GB/s without interruption





45 kHz Hadronic Interaction Rate

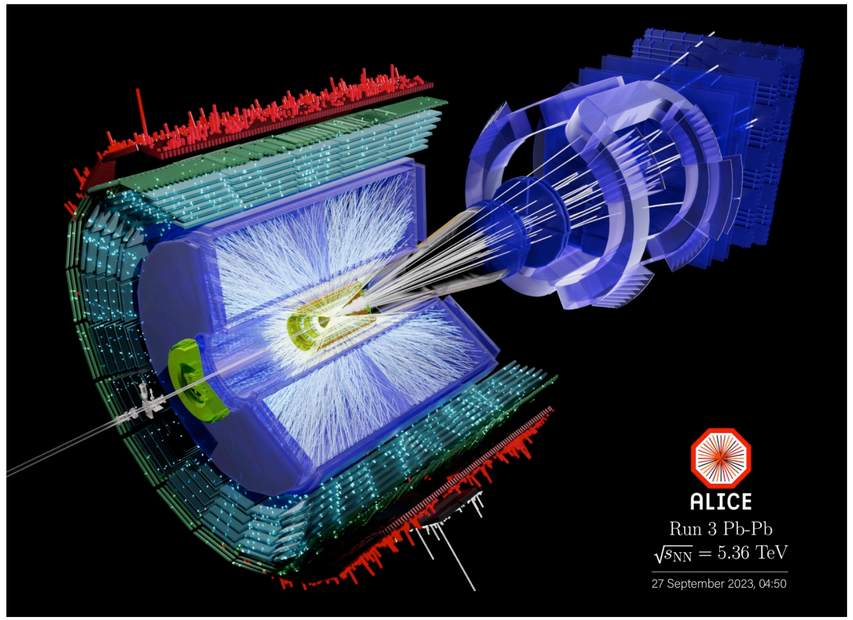
EPN input data rate showing the natural fluctuations from centrality distributions

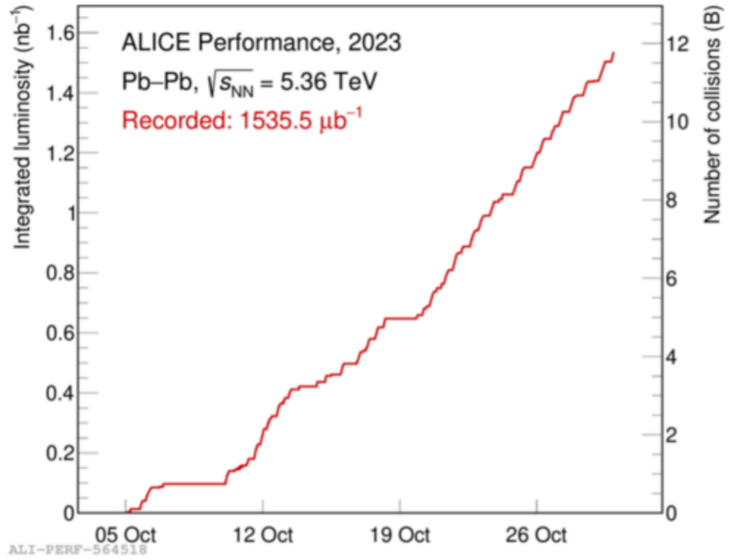


Processing under control

First Experience with high-rate Pb-Pb

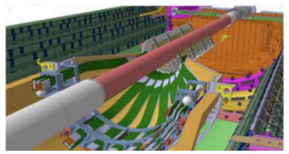
- Maximum interaction rate reached: 47 kHz
  - Raw data rate into EPNs 770 GB/s
  - CTF rate to storage 170 GB/s
  - up to 4 PB of CTFs written in 24 h
  - EPN compute margin 17.5%
  - 33 CPU cores used
- All detectors performed well
- Collected 1.4 10^{10} collisions
  - less than expected due to
    - machine background problems at the begin of Pb-Pb period
    - interaction rate reduced by machine
- First asynchronous reconstruction pass:
  - on 20% of data during 2 week end-of-year break



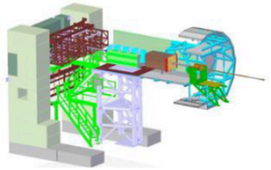


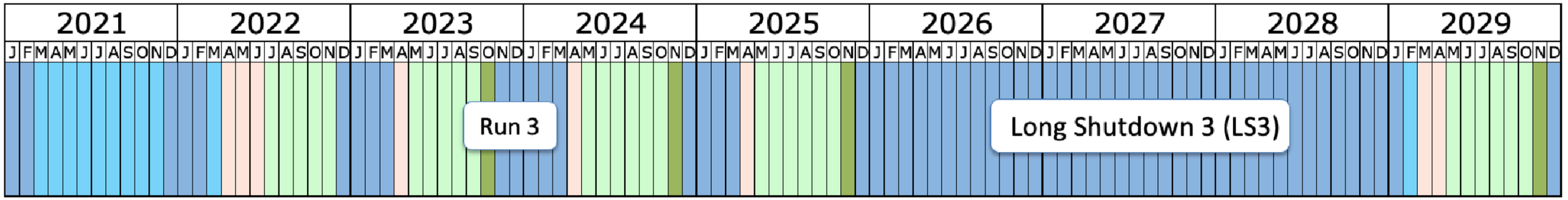
ALICE upgrade plans for run 4-6

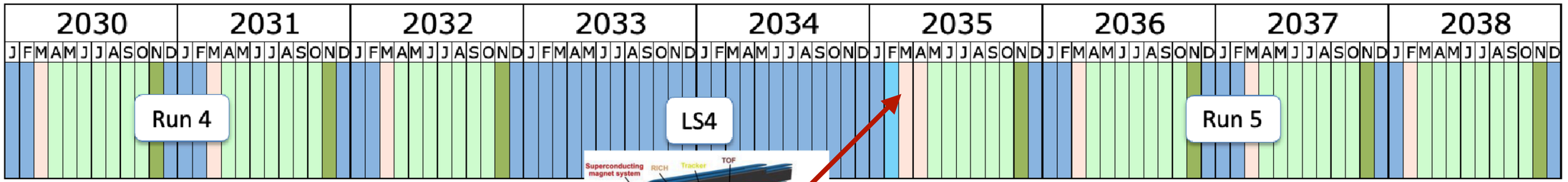
ITS3

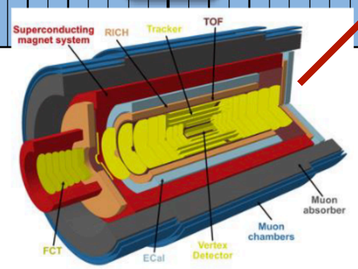


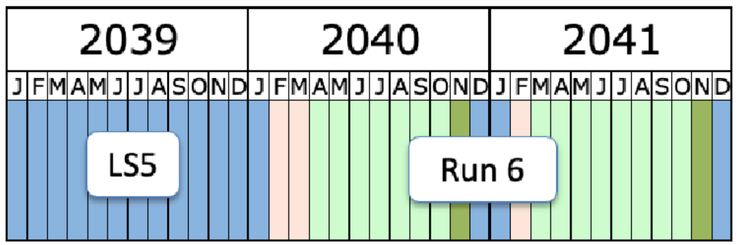
FOCAL





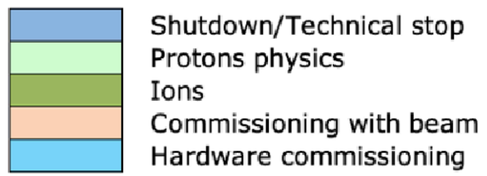




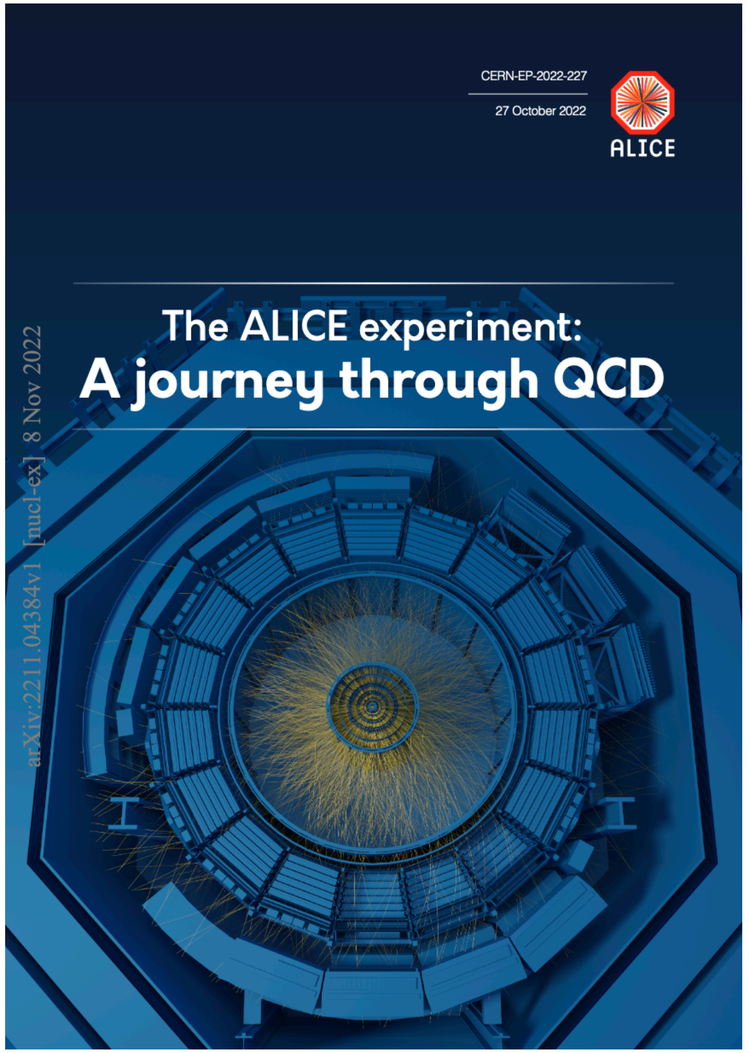


Last update: April 2023

More computing challenges ahead!



- Impressive physics output from Run 1+2 documented here ⇒
- Successful high-rate pp and Pb-Pb data taking
  - with upgraded detectors
  - and the novel O^2 online/offline data processing system
- **Our journey continues with more precision**
- **… and CERN-IT is an essential part of it!**



Acknowledgments and thanks to …

- Communication systems IT-CS-NE
- Storage & data management IT-SD-PDS, IT-SD-TAB
- Technical delivery IT-TD
- Fabric IT-FA
- And indirectly many other groups in the IT department involved in the ALICE data taking and processing activities