ALICE Data Processing in LHC Run 3

David Rohr on behalf of the ALICE Collaboration

7.12.2022

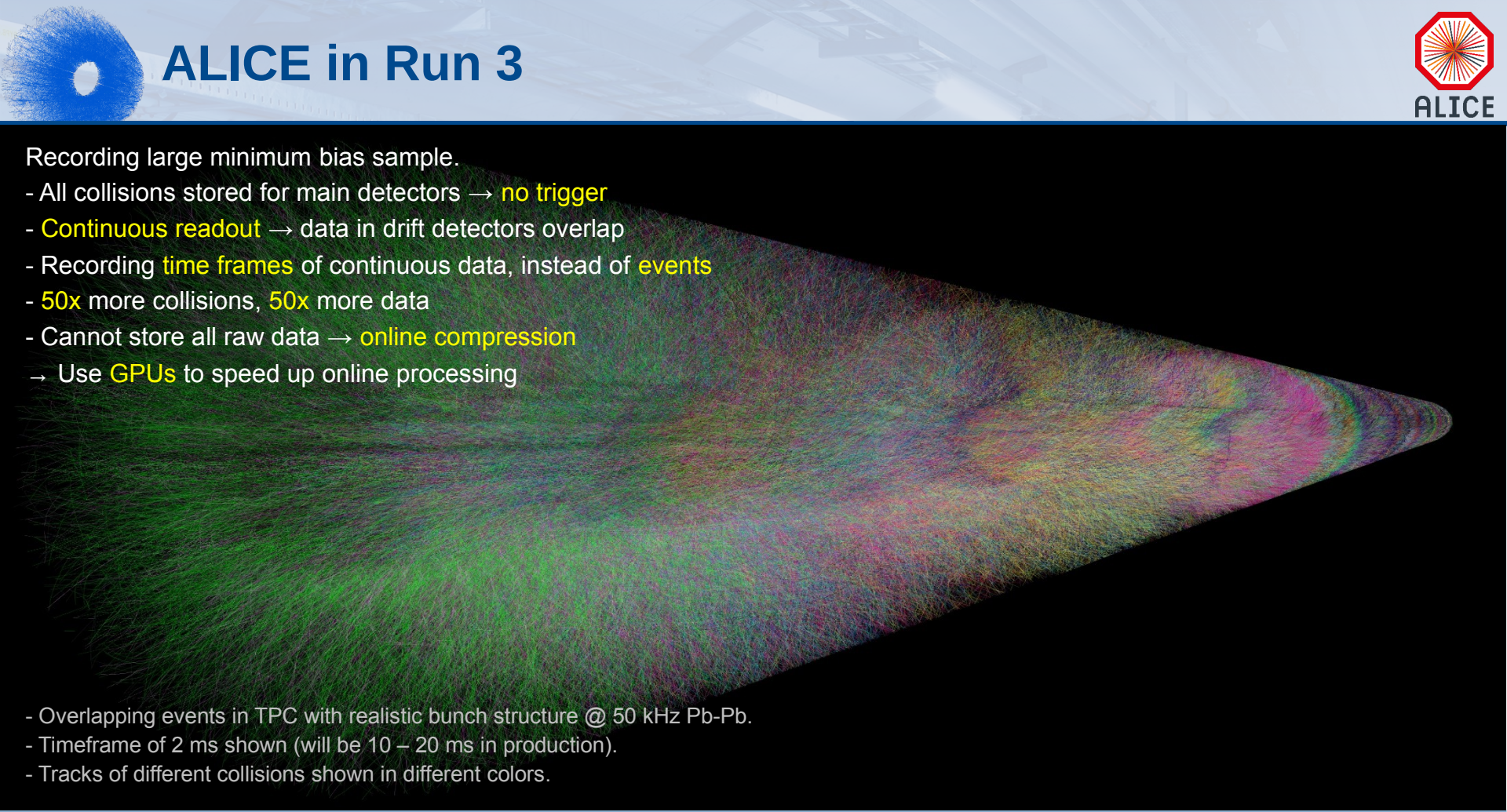
Kruger 2022: Discovery Physics at the LHC

drohr@cern.ch

ALICE in Run 3

Recording large minimum bias sample.
- All collisions stored for main detectors → no trigger
- Continuous readout → data in drift detectors overlap
- Recording time frames of continuous data, instead of events
- 50x more collisions, 50x more data
- Cannot store all raw data → online compression

→ Use GPUs to speed up online processing

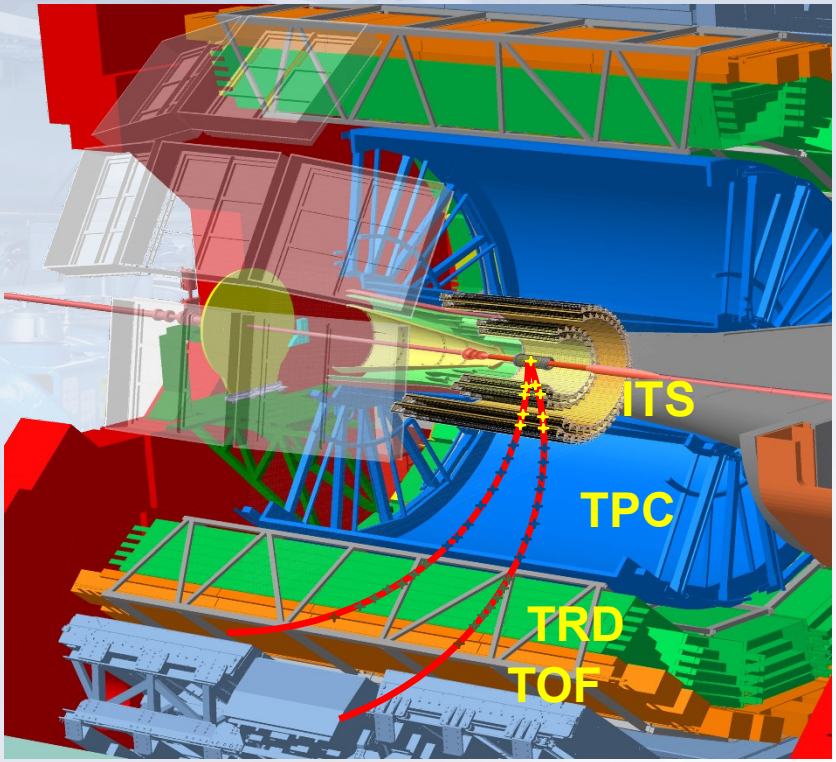


- Overlapping events in TPC with realistic bunch structure @ 50 kHz Pb-Pb.
- Timeframe of 2 ms shown (will be 10 – 20 ms in production).
- Tracks of different collisions shown in different colors.

The ALICE Detector in Run 3

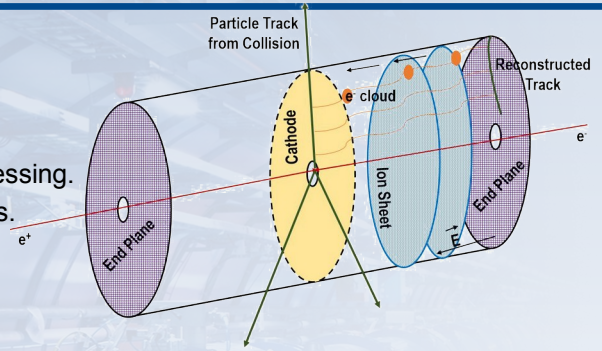
See talk of Jochen Klein!

- ALICE uses mainly 3 detectors for barrel tracking: ITS, TPC, TRD + (TOF)
  - **7 layers ITS** (Inner Tracking System – silicon tracker)
  - **152 pad rows** TPC (Time Projection Chamber)
  - **6 layers TRD** (Transition Radiation Detector)
  - **1 layer TOF** (Time Of Flight Detector)
- Several major upgrades before Run 3:
  - The TPC is equipped with a GEM readout
  - The ITS is completely replaced by 7 layers of silicon pixels
  - **Major computing upgrade in the O² project**
    - **Merges online and offline processing in the same software framework. Same code (with different cuts / parameters) running online and offline**
- Drivers behind design decisions:
  - Search for rare signals imposes large increase in statistics wrt. Run 1+2
  - Triggered TPC readout insufficient
    - Huge out-of-bunch pile up during one TPC drift time
    - → Need continuous readout



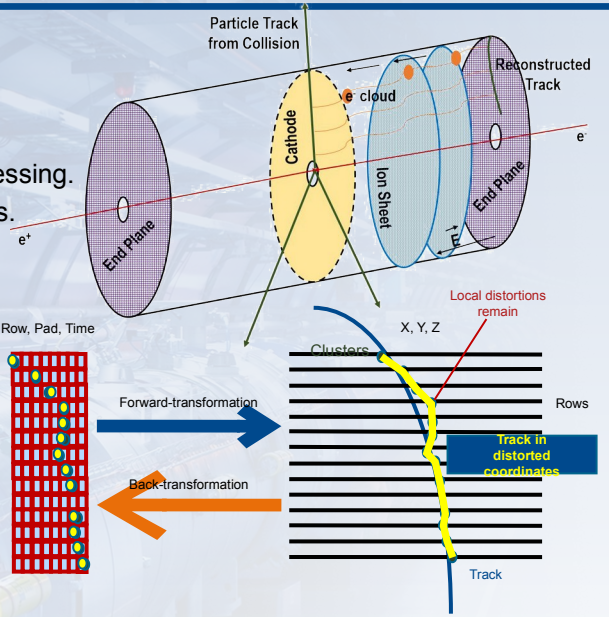
O2 Processing Steps

- **Synchronous processing (what we called online before):**
  - Extract information for **detector calibration**:
    - Previously performed in 2 offline passes over the data after the data taking.
    - Run 3 avoids / reduces extra passes over the data but extracts all information in the sync. processing.
    - An intermediate step between sync. and async. processing produces the final calibration objects.
    - The most complicated calibration is the correction for the TPC space charge distortions.



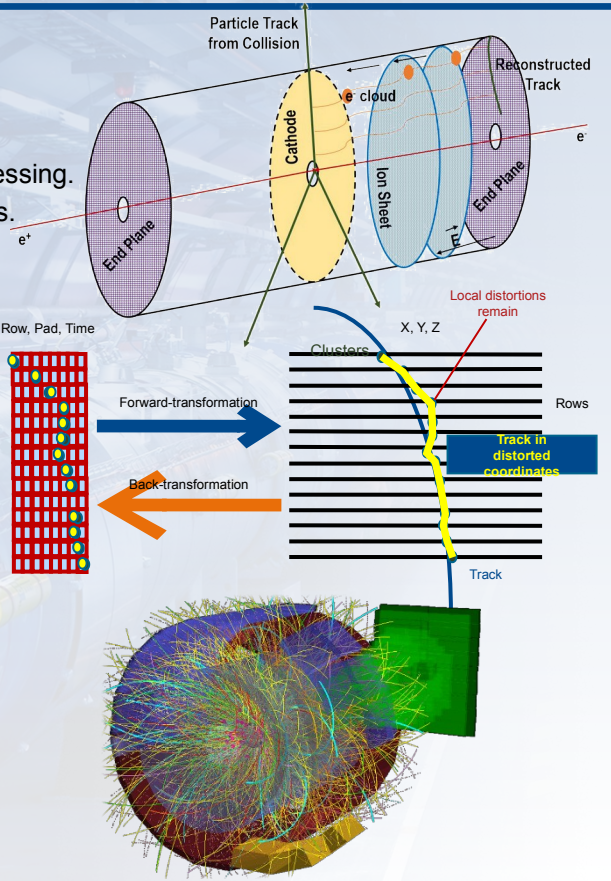
O2 Processing Steps

- **Synchronous processing (what we called online before):**
  - Extract information for **detector calibration**:
    - Previously performed in 2 offline passes over the data after the data taking.
    - Run 3 avoids / reduces extra passes over the data but extracts all information in the sync. processing.
    - An intermediate step between sync. and async. processing produces the final calibration objects.
    - The most complicated calibration is the correction for the TPC space charge distortions.
  - **Data compression**:
    - TPC is the largest contributor of raw data, and we employ sophisticated algorithms like storing space point coordinates as residuals to tracks to reduce the entropy and remove hits not attached to physics tracks.
    - We use ANS entropy encoding for all detectors.



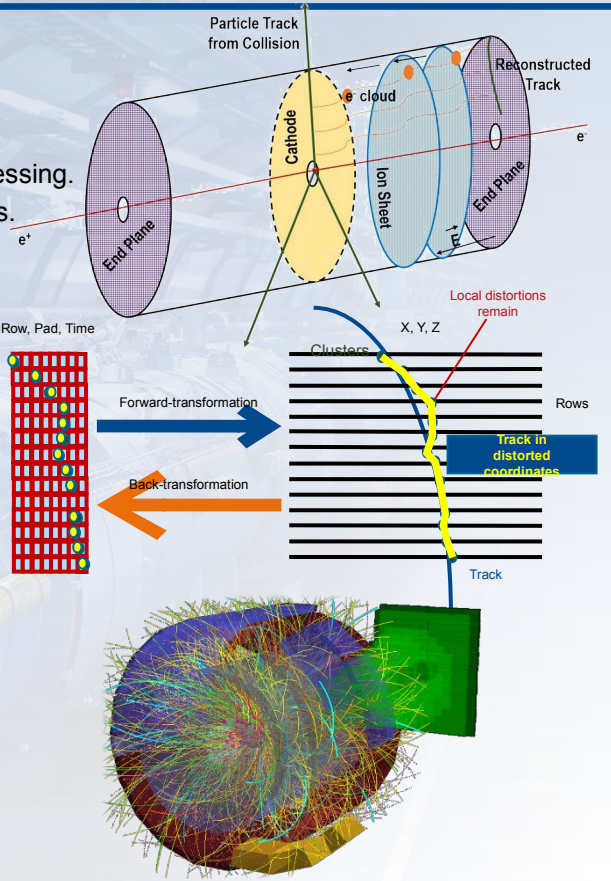
O2 Processing Steps

- **Synchronous processing (what we called online before):**
  - Extract information for **detector calibration**:
    - Previously performed in 2 offline passes over the data after the data taking.
    - Run 3 avoids / reduces extra passes over the data but extracts all information in the sync. processing.
    - An intermediate step between sync. and async. processing produces the final calibration objects.
    - The most complicated calibration is the correction for the TPC space charge distortions.
  - **Data compression**:
    - TPC is the largest contributor of raw data, and we employ sophisticated algorithms like storing space point coordinates as residuals to tracks to reduce the entropy and remove hits not attached to physics tracks.
    - We use ANS entropy encoding for all detectors.
  - **Event reconstruction** (tracking, etc.):
    - Required for calibration, compression, and online quality control.
    - Need **full TPC tracking** for data compression.
    - Need tracking in all detectors for ~1% of the tracks for calibration.
    - → **TPC tracking dominant part**, **rest** almost negligible (**< 5%**).



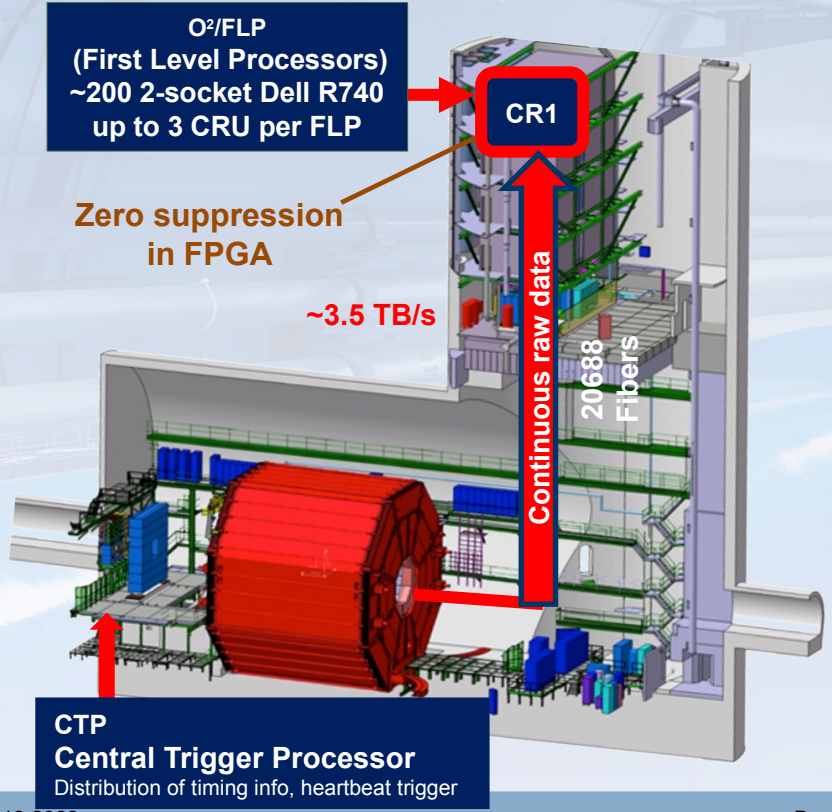
O2 Processing Steps

- **Synchronous processing (what we called online before):**
  - Extract information for **detector calibration**:
    - Previously performed in 2 offline passes over the data after the data taking.
    - Run 3 avoids / reduces extra passes over the data but extracts all information in the sync. processing.
    - An intermediate step between sync. and async. processing produces the final calibration objects.
    - The most complicated calibration is the correction for the TPC space charge distortions.
  - **Data compression**:
    - TPC is the largest contributor of raw data, and we employ sophisticated algorithms like storing space point coordinates as residuals to tracks to reduce the entropy and remove hits not attached to physics tracks.
    - We use ANS entropy encoding for all detectors.
  - **Event reconstruction** (tracking, etc.):
    - Required for calibration, compression, and online quality control.
    - Need **full TPC tracking** for data compression.
    - Need tracking in all detectors for ~1% of the tracks for calibration.
    - → **TPC tracking dominant part**, **rest** almost negligible (**< 5%**).

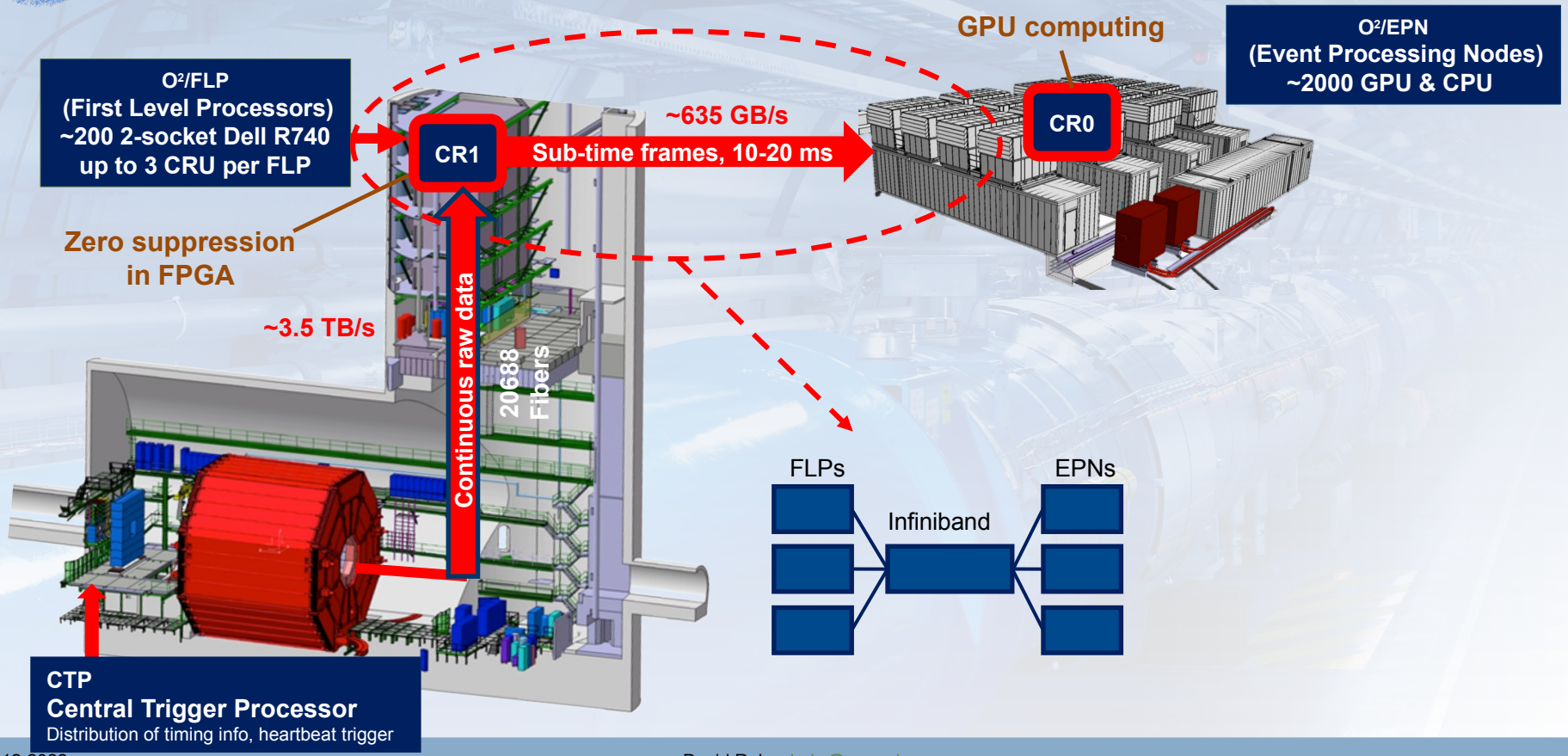


- **Asynchronous processing (what we called offline before):**
  - **Full reconstruction**, **full calibration**, **all detectors**.
  - TPC part faster than in synchronous processing (less hits, no clustering, no compression).
  - → **Different relative importance of GPU / CPU** algorithms compared to synchronous processing.

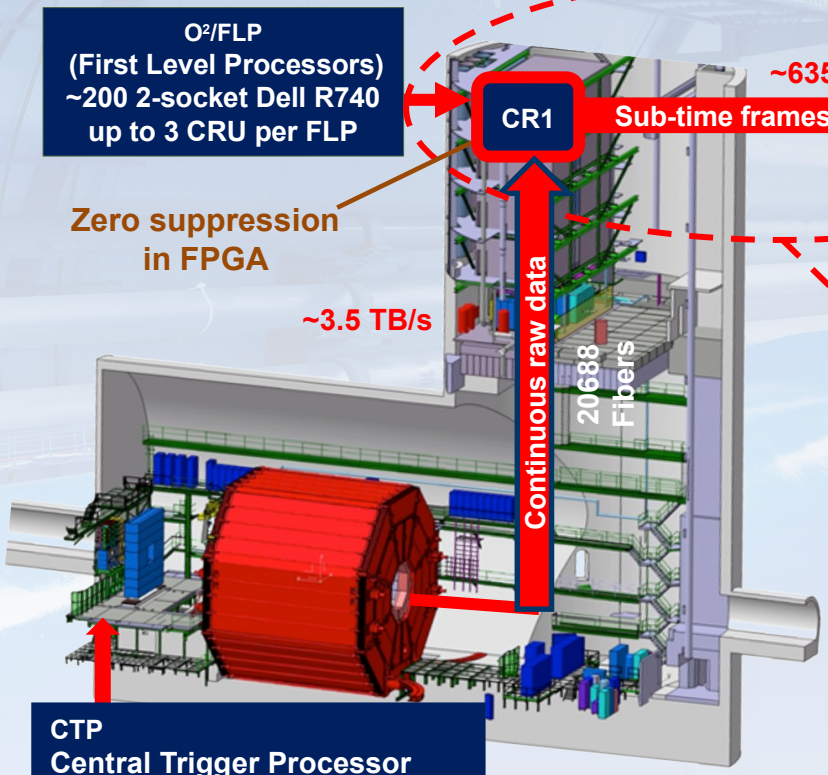
ALICE Data Flow in Run 3



ALICE Data Flow in Run 3



ALICE Data Flow in Run 3



O²/FLP
(First Level Processors)
~200 2-socket Dell R740
up to 3 CRU per FLP

CR1

~635 GB/s
Sub-time frames, 10-20 ms

GPU computing

CR0

O²/EPN
(Event Processing Nodes)
~2000 GPU & CPU

Zero suppression
in FPGA

~3.5 TB/s

Continuous raw data

20688
Fibers

~100 GB/s
CTF: Compressed time frames

Calibration data

80 PB
disk storage, 360GB/s
(~25% redundancy)

FLPs

Infiniband

EPNs

~70 % CTF

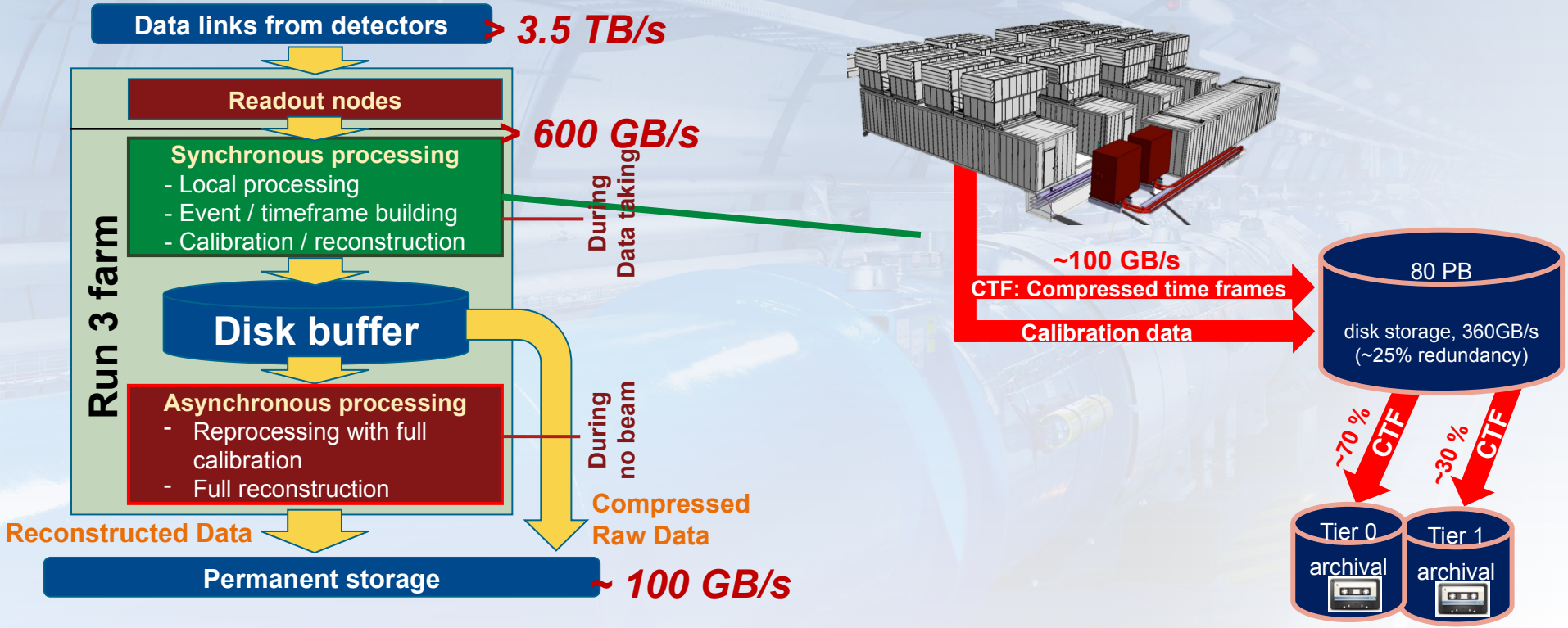
~30 % CTF

Tier 0
archival

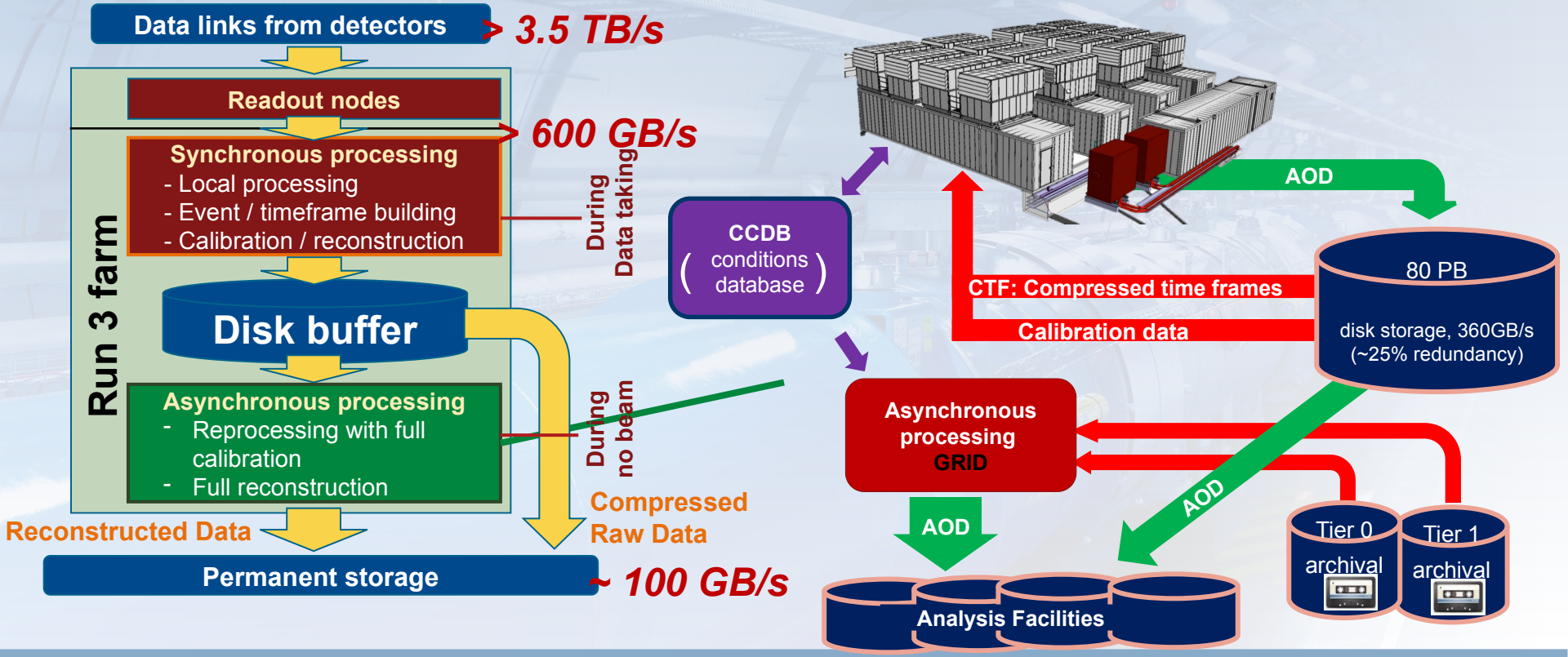
Tier 1
archival

CTP
Central Trigger Processor
Distribution of timing info, heartbeat trigger

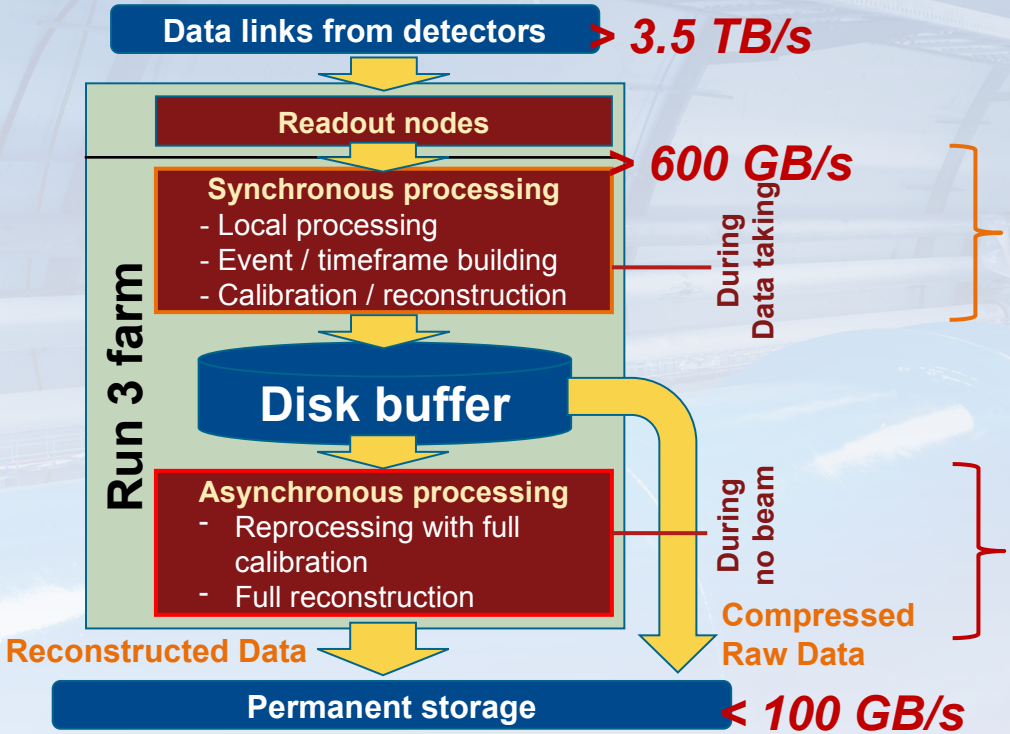
Synchronous and Asynchronous Reconstruction



Synchronous and Asynchronous Reconstruction



Synchronous and Asynchronous Reconstruction



- Calibration: Tracking for ITS / TPC / TRD / TOF for ~1% of tracks.
- Data compression: track-model compression requires full TPC tracking for all collisions.
  - → TPC tracking dominant workload during synchronous reconstruction.
  - → Well suited to run on **GPUs**, EPN farm designed for best TPC clusterization / tracking / compression performance.

- No clear single computational hot-spot.
- TPC reconstruction important but not dominant.
  - Actually faster than in the synchronous phase: no clusterization / no compression / less hits after hit removal in synchronous phase overcompensates the slowdown of more elaborate fits.
- Full reconstruction for all other detectors.

→ More heterogeneous workload.

Baseline / optimistic scenario for GPU processing

- **EPNs make massive use of GPUs to speed up the real time TPC processing (bulk of synchronous reconstruction).**
- **Aiming to use the GPUs as well as possible also in the asynchronous reconstruction.**
- **GPU processing developed for 2 scenarios:**

Baseline GPU solution
(fully available, used 2022):
TPC + part of ITS tracking on GPU

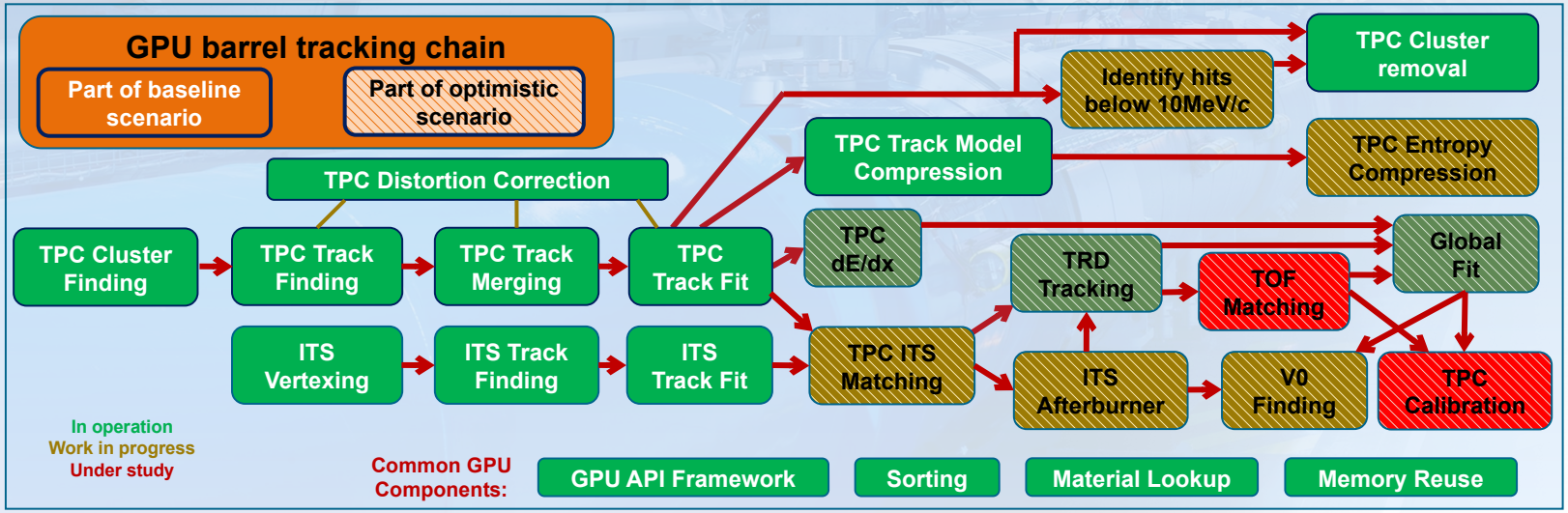
- This is the mandatory part to keep step with the data taking during the synchronous reconstruction.
  - Aiming for ~20% margin.
- Caching the raw data is impossible, i.e. if we are not fast enough here, we need to reduce the interaction rate.

Optimistic GPU solution
(what we are aiming for eventually):
Run full barrel tracking on GPU

- This aims to make the best use of the GPUs also in the asynchronous phase.
- Does not affect the synchronous processing much, though could offload slightly more steps to the GPU as well..
- If we cannot use the GPUs for a large part of the asynchronous reconstruction on the EPN, the processing would be CPU bound while the GPUs would be idle.

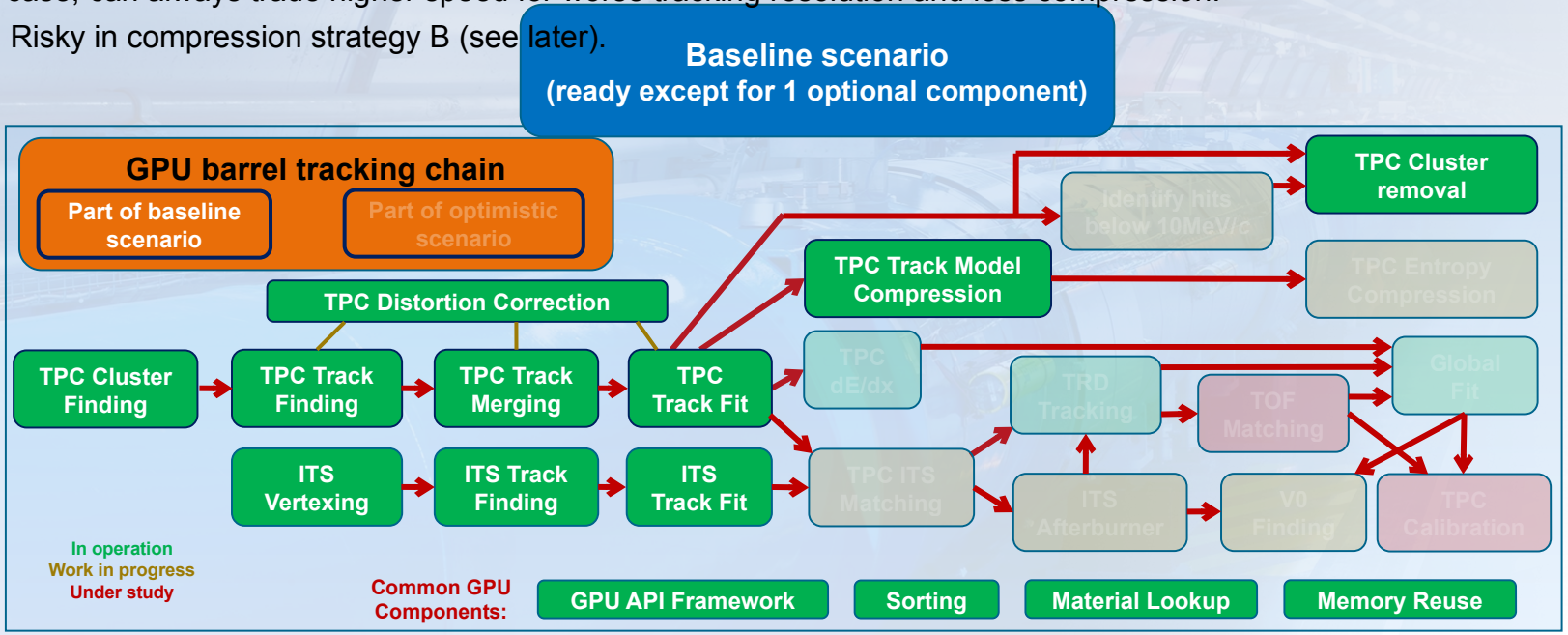
Central Barrel Global Tracking Chain

- Overview of reconstruction steps considered for GPU-offload:
  - Mandatory **baseline scenario** includes everything that must run on the GPU during synchronous reconstruction.
  - **Optimistic scenario** includes everything related to the barrel tracking.



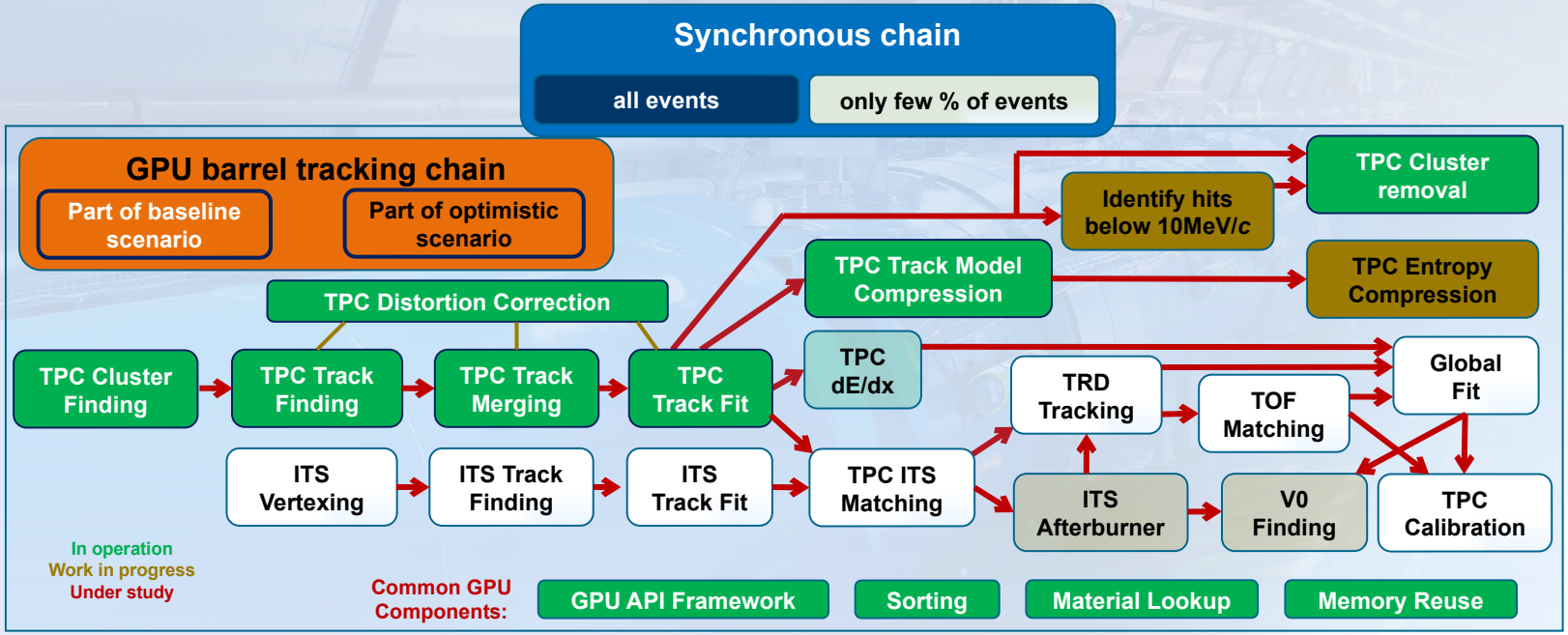
Central Barrel Global Tracking Chain

- Baseline scenario fully implemented (module some improvements e.g. distrotion correction).
  - Not mandatory to speed up the synchronous GPU code further, but we should try nonetheless.
  - If we add / improve reconstruction steps, we have to speed it up accordingly to remain in the 2000 GPU budget.
  - Worst case, can always trade higher speed for worse tracking resolution and less compression.
    - Risky in compression strategy B (see later).



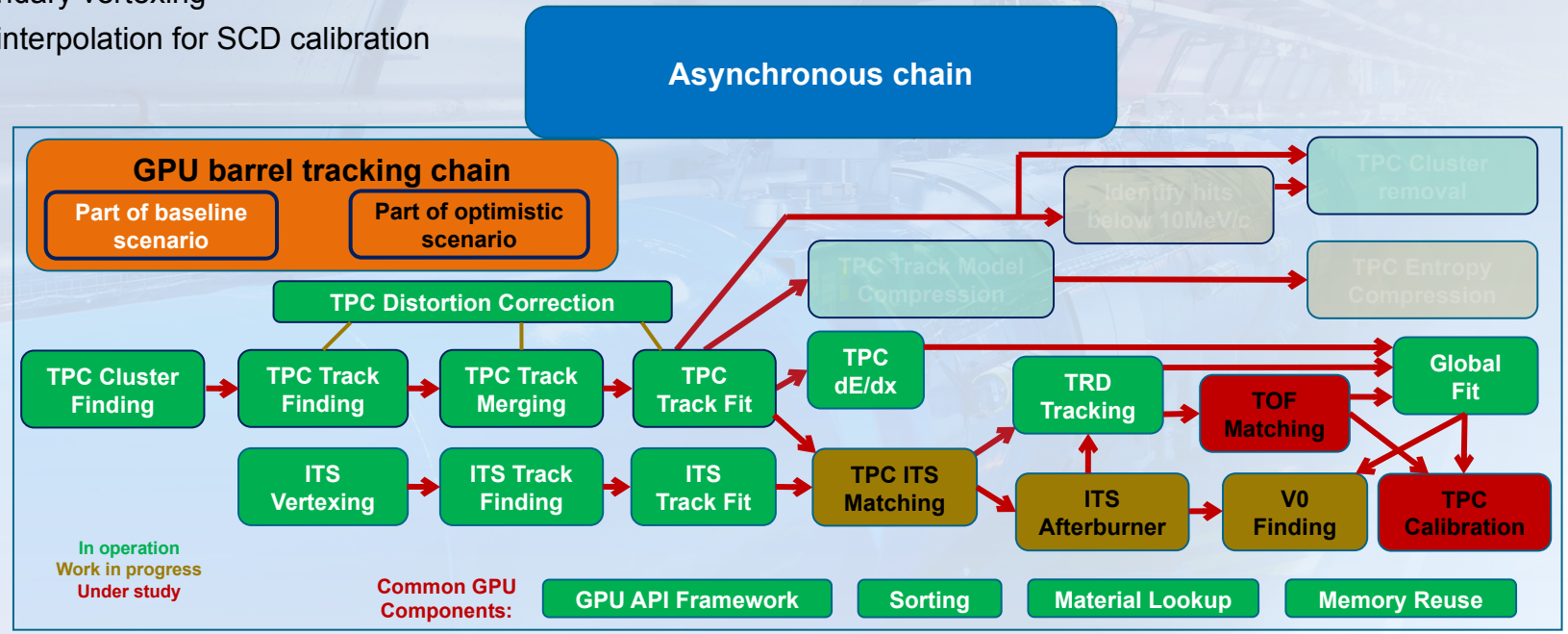
Central Barrel Global Tracking Chain

- Baseline scenario fully implemented (module some improvements e.g. distrotion correction).
  - 2 optional parts still being investigated for sync. reco on GPU: TPC entropy encoding / Looper identification < 10 MeV.



Central Barrel Global Tracking Chain

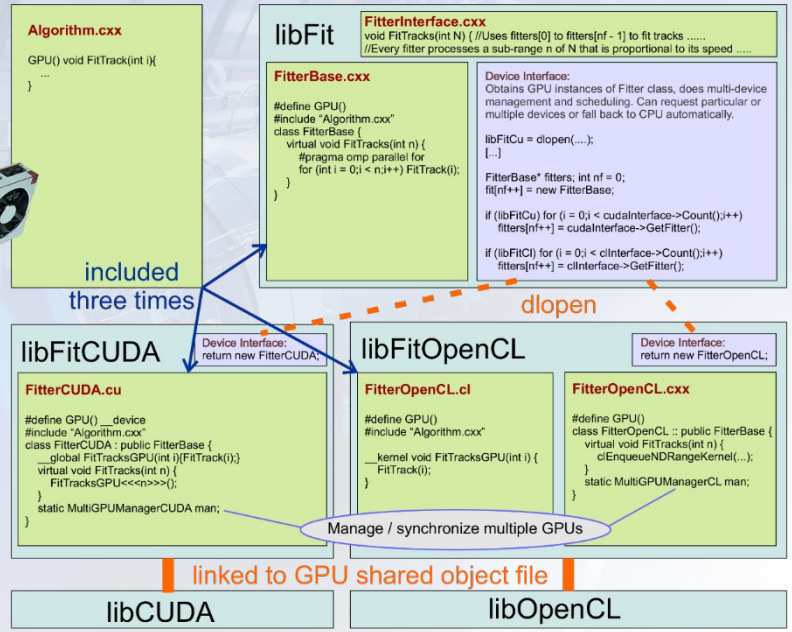
- Several steps missing in asynchronous reconstruction:
  - Matching to ITS
  - Matching to TOF
  - Secondary vertexing
  - TPC interpolation for SCD calibration



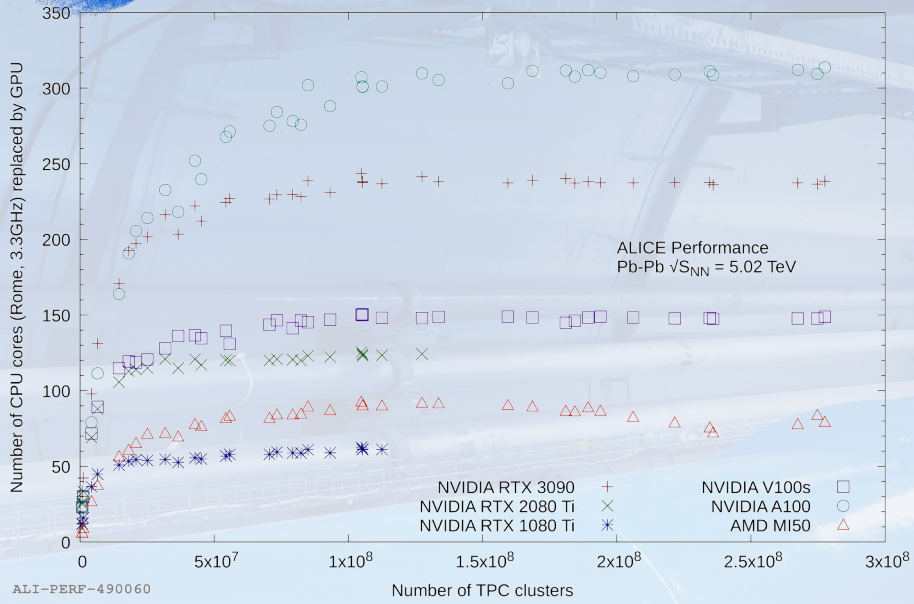
Compatibility with different GPU Frameworks

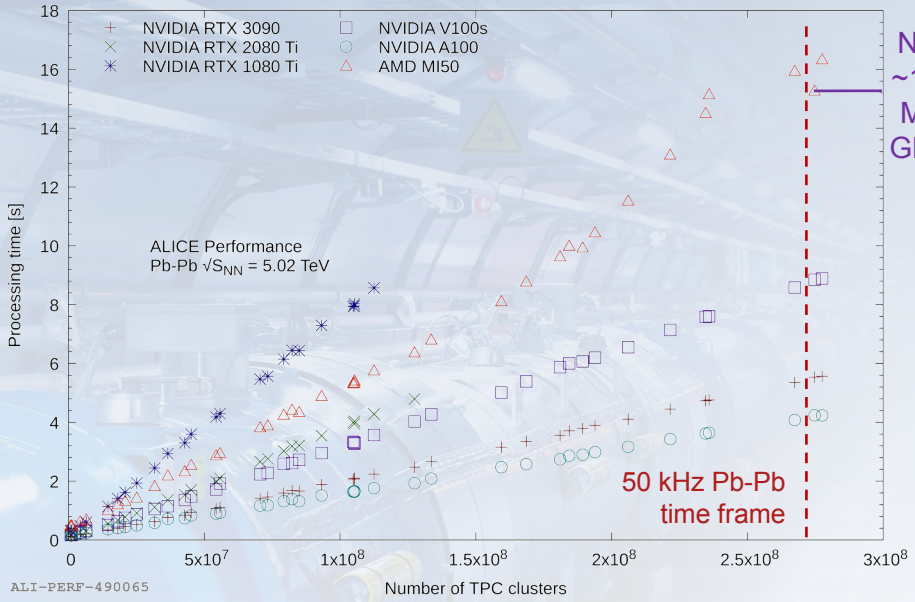
- Generic common C++ Code compatible to CUDA, OpenCL, HIP, and CPU (with pure C++, OpenMP, or OpenCL).
  - OpenCL needs clang compiler (ARM or AMD ROCm) or AMD extensions (TPC track finding only on Run 2 GPUs and CPU for testing)
  - Certain worthwhile algorithms have a vectorized code branch for CPU using the Vc library
  - All GPU code swapped out in dedicated libraries, same software binaries run on GPU-enabled and CPU servers

- Screening different platforms for best price / performance.
  (including some non-competitive platforms for c… …atio…
  - **CPUs** (**AMD Zen**, **Intel Skylake**)
    C++ backend with **OpenMP**, AMD **OCL**
  - **AMD GPUs**
    (**S9000** with **OpenCL 1.2**, **MI50** / **Radeon 7** / **Navi** with **HIP** / **OCL 2.x**)
  - **NVIDIA GPUs**
    (**RTX 2080** / **RTX 2080 Ti** / **Tesla T4** with **CUDA**)
  - **ARM Mali GPU** with **OCL 2.x**
    (Tested on dev-board with Mali G52)



GPU Performance

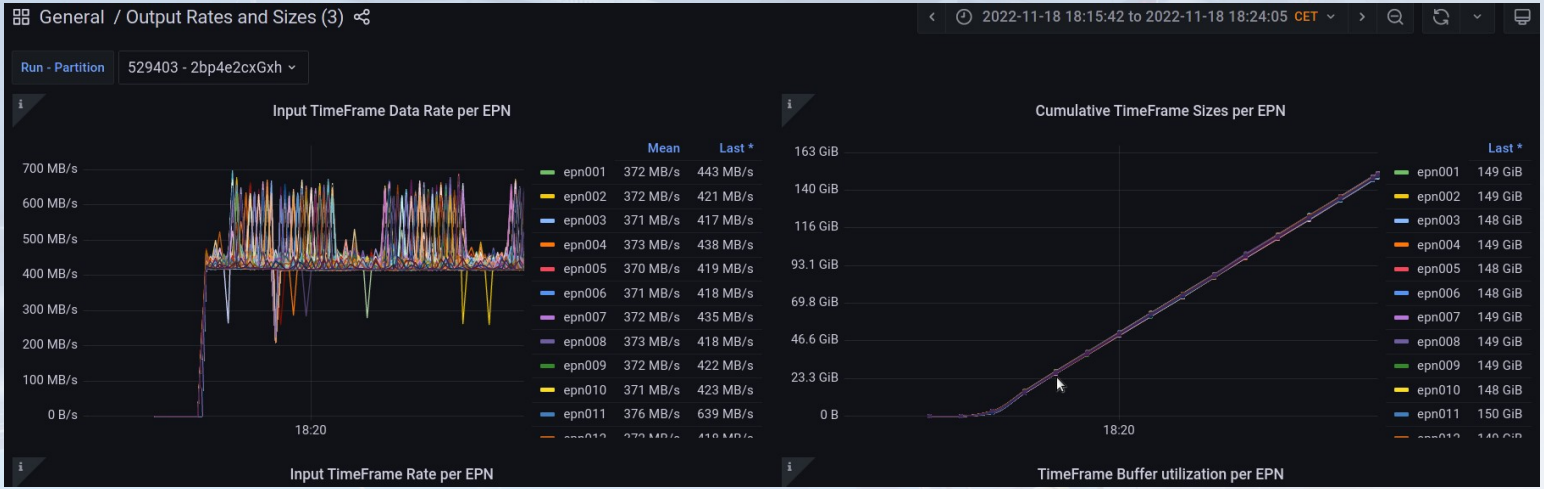




- **MI50 GPU** replaces **~80 Rome cores** in **synchronous** reconstruction.
  - Includes **TPC clusterization**, which is **not optimized** for the CPU!
- **~55 CPU cores** in **asynchronous** reconstruction (more realistic comparison).

Online Pb-Pb processing in 2022

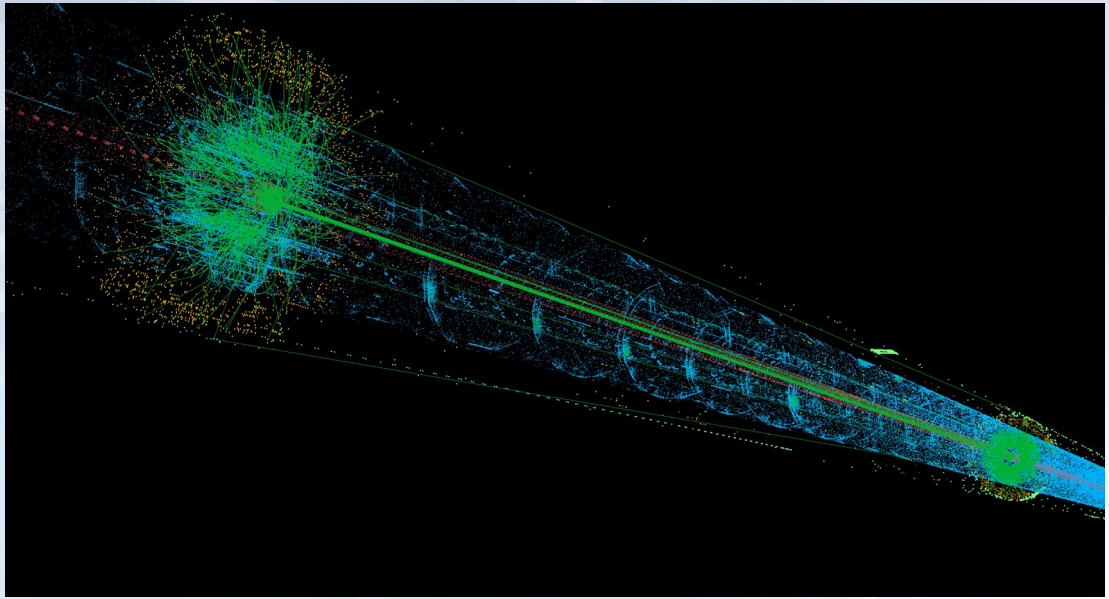
- **First stable Pb-Pb beam 2022** was processed onling on November 18th



- Not challenging from processing perspective, due to low IR (~100 Hz vs. 50 kHz design IR).
- Data taking very stable, no backpressure (as expected), recorded data size increasing linearly.

Online Pb-Pb processing in 2022

- **First stable Pb-Pb beam 2022** was processed onling on November 18th
  - Low interaction rate in beam test
  - <1 collision per time frame in average
  - One “dense” TF with 2 collisions is shown on the right



- Note: **baseline** for Pb-Pb is:
  - Calibration, Compression, QC
  - **Full TPC tracking**
  - **Few percent** of data from **other detectors** only

- Due to free compute resources, could run **additional processing** online:
  - Full ITS reconstruction
  - Forward muon reconstruction
  - TPC dEdx

- Validated **50 kHz Pb-Pb online** processing with **MC** data, **> 20 % margin for CPU** / **GPU load** and **host** / **GPU memory**.

Online pp processing in 2022

- Distinguish 3 cases for pp data taking:
  - Pb-Pb reference data taking: low IR rate, computationally not challenging.
  - Nominal pp data taking with up to 1 Mhz:
    - Much less challenging than 50 kHz Pb-Pb, i.e. EPN compute farm has plenty of margin.
    - Can optimize in 2 directions:
      1. Minimize the EPN load to run with fewer EPNs, i.e. more EPNs available for asynchronous reconstruction.
      2. Enable additional reconstruction steps (MUON, dEdx) using the available compute capacity.
      So far doing a mixture of the two.
  - High rate pp data taking for high rate tests (Pb-Pb equivalent rate is 4.5 Mhz inelastic IR).
    - Computationally challenging.
    - Full TPC processing is already problematic, since TPC occupancy does not scale linearly with mid-rapidity primaries.
    - Also other detectors don't necessarily scale linearly with primary particles.
    - In Pb-Pb doing only online processing of few percent for other detectors foreseen, don't have resources for "normal" pp processing at Pb-Pb equivalent rates.
    - Highest inelastic IR with "full" processing was 2.6 Mhz (limited by EPN CPU and memory resources).
    - Note: this is no real problem, since the high rate tests are special runs mostly not for physics. For many detector tests, it is enough to process only a subset of the time frames, and if we disable the TPC track model compression, we can also run full compression for all detectors.

Tuning of asynchronous pp reco on EPNs with GPU

- **Benchmarks of asynchronous reconstruction, pp collisions (real data), 650 kHz inellastic rate.**
- **Looked at 3 scenarios:**
  1. 8-core GRID node, CPU only: Allocated 4 physical cores (8 virtual cores) and 32 GB of RAM (4 per virtual core) on an EPN - closest to grid setup.
     - Note: GRID guarantees only 16 GB, but currently no workflow that can efficiently process TFs with 16 GB of memory.
  2. 1/8th of an EPN, or Single-GPU: 8 physical cores (16 virtual cores), 64 GB of memory, 1 MI50 GPU – **running today on EPNs**.
  3. ½ of an EPN / 1 NUMA domain: 32 physical cores (64 virtual cores), 256 GB of memory, 4 MI50 GPUs – closest to synchronous processing.

Setup	Seconds per TF
8 core, no GPU, unoptimized, 1 TF in flight, 32 GB	84.75
8 core, no GPU, optimized, 3 TF in flight, 64 GB	50.21
16 core + GPU, unoptimized, 1 TF in flight, 64 GB	71.25
16 core + GPU, optimized, 8 TF in flight, 64 GB	11.23
4 * (16 core + GPU, optimized, 8 TF in flight, 64 GB)	3.55 = (14.20 / 4)
64 core + 4 GPU, 24 TF in flight, 256 GB, partially optimized	3.45

CPU only

1/8th EPN

1/2 only

- Slowdown when running 4 independent 1-GPU workflows in one NUMA domain due to competition for resources: 14.20s vs. 11.23s
- Using the 1/2 EPN setup, we currently gain ~3%: 3.45s vs. 3.55s, but not yet fully optimized, and still causes some framework trouble.

Tuning of asynchronous Pb-Pb reco on EPNs

- Pb-Pb processing needs more memory than pp due to larger time frames.
- 1 GPU workflow not efficient, since we would need more memory. The 1NUMA domain workflow has synergy effects and is thus not limited by memory.
- Tuned only 2 cases:
  - the CPU-only workflow without memory constraint (for reference)
  - the 1 NUMA domain workflow.

- Tested on MC Pb-Pb data (since 2022 Pb-Pb real data is too low IR for realistic measurements).

Setup	Seconds per TF
8 core, no GPU, optimized, 3 TF in flight, 96 GB	200
64 core + 4 GPU, 24 TF in flight, 256 GB, partially optimized	25

  - Note: Running 8 * 8-core workflow in one NUMA domain does not speed up the throughput 8-fold, since they compete for resources such as memory.
  - GPU Pb-Pb async workflow currently heavily impacted by still poor performance of some CPU-bound algorithms. Average GPU load is only ~20%. Thus little improvement from GPU usage.

- Note: the 200s is very close to the conservative estimate we have been using for the resource estimates for Pb-Pb processing, so asynchronous processing of 2022 Pb-Pb data at 50 kHz should have worked nicely.

GPU / CPU fraction of workload (Pb-Pb 50 kHz MC)

- The table below shows the relative compute time (linux cpu time) of the processing steps running on the processor.
  - The synchronous reconstruction is fully dominated by the TPC (99%) which already fully runs on the GPU, some more processes might follow.
  - Basically no margin to offload mory synchronous reconstruction step to the GPU – or if we did, it wouldn't change anything.

Synchronous processing

Processing step	% of time
TPC Processing	99.37 %
EMCAL Processing	0.20 %
ITS Processing	0.10 %
TPC Entropy Coder	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 %
FIT Entropy Coder	0.01 %
TOF Entropy Coder	0.01 %
MFT Entropy Coder	0.01 %
TPC Calibration residual extraction	0.01 %
TOF Processing	0.01 %

Running on GPU in baseline scenario **Running on GPU in optimistic scenario**

GPU / CPU fraction of workload (650 kHz pp)

- Same table for asynchronous reconstruction.
- Compute time much more wide-spread: TPC is only ~60%.
- Imperative to offload more steps onto the GPU for good EPN usage

Asynchronous processing

Process	Compute Time [%]
TPC Tracking	61.64
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
QC	4.00
Rest	2.32

- Currently (**baseline scenario**) **61.6%** of the asynchronous workflow on the GPU.
- When the GPU offloading effort for TPC + ITS + TRD + TOF is finished (**optimistic scenario**), we will have **85%** on the GPU in this setup.
  - Already on GPU
  - To be offloaded
- For reference: **90%** of EPN compute power in the GPU (assuming GPU speedup for other detectors is similar as for TPC).

Conclusions

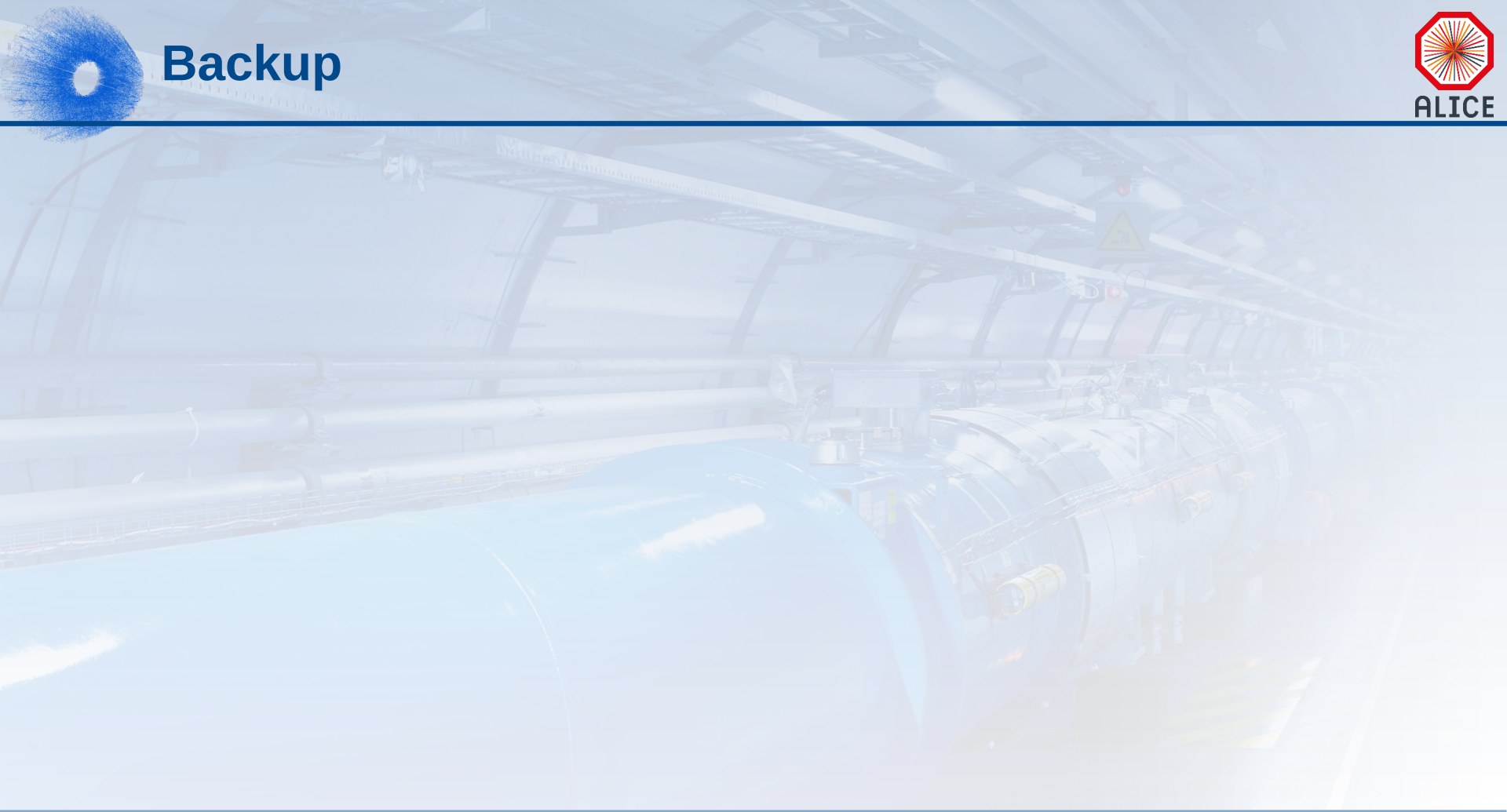
- ALICE employs **GPUs** heavily to speed up **online** and **offline** processing.
  - **99%** of **online processing** on the **GPU** (no reason at all to port the rest).
  - Currently **~60%** of **offline processing** (for 650 kHz pp) on **GPU** (offline jobs on the EPN farm).
  - Will increase to **>85%** with full barrel tracking (**optimistic scenario**).
  - Eventually aiming for **90%** of offline processing on GPU, since **90%** of **EPN compute power** is in **GPU**s.

- Synchronous processing **successful** in **2022**.
  - Pb-Pb 2022 not really stressfull for processing, since no 50 kHz Pb-Pb.
    - **50 kHz Pb-Pb** processing **validated** with data replay of **MC** data (> 20 % margin).
  - Performed **online pp** processing with **additional processing steps** up to **2.6 MHz** inelastic interaction rate.
    - Limitation is more on the CPU side / host memory side, not on the GPU.

- Work on **optimistic scenario** for GPU processing progressing (full barrel tracking **ITS/TPC/TRD/TOF** on **GPU**).

- Asynchronous reconstruction has started, processing the TPC reconstruction on the GPUs in the EPN farm, and in CPU-only style on the CERN GRID site.
  - Work ongoing to switch to the more efficient 1NUMA domain GPU workflow with less overhead.

Backup



Synchronous (CPU + GPU) reconstruction steps

Synchronous Reconstruction