ALICE experience with GPUs

David Rohr (CERN)
Stefano Piano (INFN TS)

Introduction

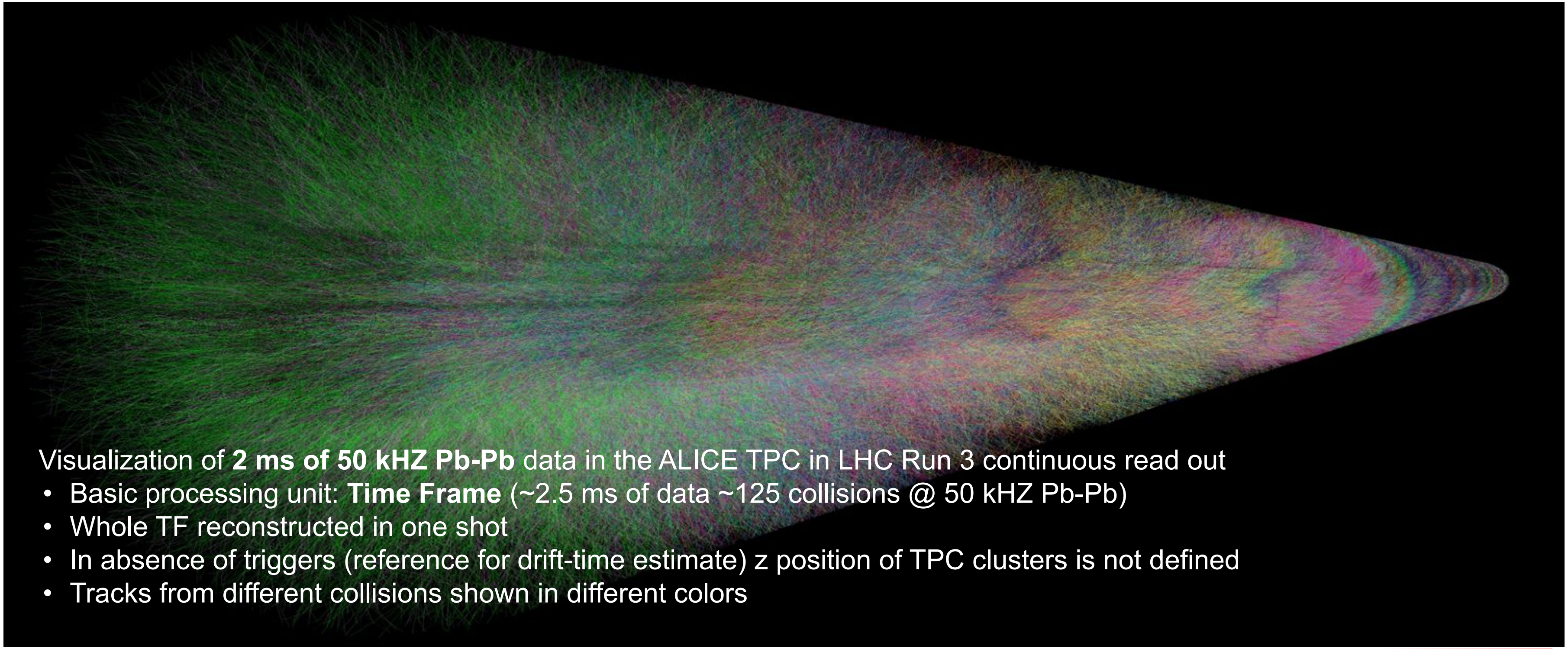
- ALICE long-standing and extensive experience in exploiting GPU in Run 1 / 2 HLT:
  - The available CPU supplemented by utilizing FPGAs and GPUs
  - In Run 1 with 64 Nvidia GeForce GTX 480, while in Run 2 with 180 AMD S9000
  - Significant cost savings, size of the Run 1 and Run 2 farms was nearly reduced by a half
- The HLT code could use both the CUDA framework for NVIDIA GPUs or the OpenCL framework with C++ extensions for AMD GPUs, added support for AMD ROCm as backend later
  - Small wrappers allowed the execution of the code on different GPU models and also on standard CPUs, parallelized via OpenMP
  - Much of the implemented functionality in Run 2 served as a testing bed for concepts related to the ALICE O2 system for LHC Run 3
- Investment in (outstanding) programming and computing skills crucial to get performant software and computing infrastructure

ALICE

From triggered Run 2 to un-triggered Run 3 and 4

- In Run 2 ALICE operated at Pb-Pb interaction rates ~7-10 kHz (inspected ~1 nb^{-1}) with trigger rate < 1 kHz for Central Barrel
- **LHC will deliver 50 kHz Pb-Pb interaction rate during Run 3**
- **ALICE in Run 3 & 4:**
  - **collect 13 nb^{-1} of Pb-Pb collisions at 5 TeV**
  - (of which 3 nb^{-1} with reduced magnetic field)
  - At high multiplicities (dN/dη ~ 2000 + pile-up) very low S/B for rare probes:
    - Dedicated (HLT) trigger is not realistic
  - Use **continuous readout** for most of the detectors
  - Read out all events and inspect all events offline **(100x MB events wrt Run 2)**
  - But ... cannot store all raw data
    - Online compression => **use GPUs to speedup online processing**

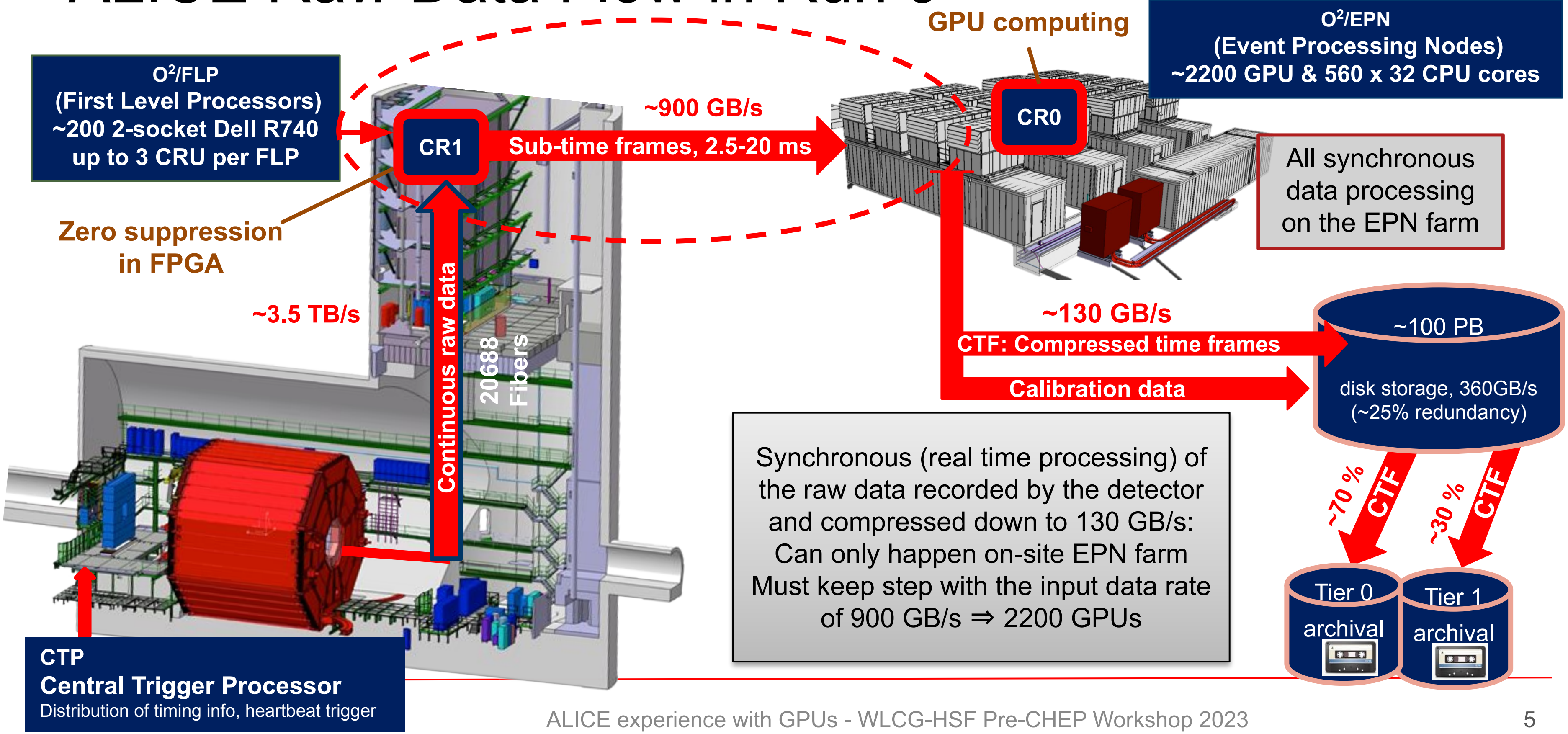
Continuous readout



Visualization of **2 ms of 50 kHZ Pb-Pb** data in the ALICE TPC in LHC Run 3 continuous read out

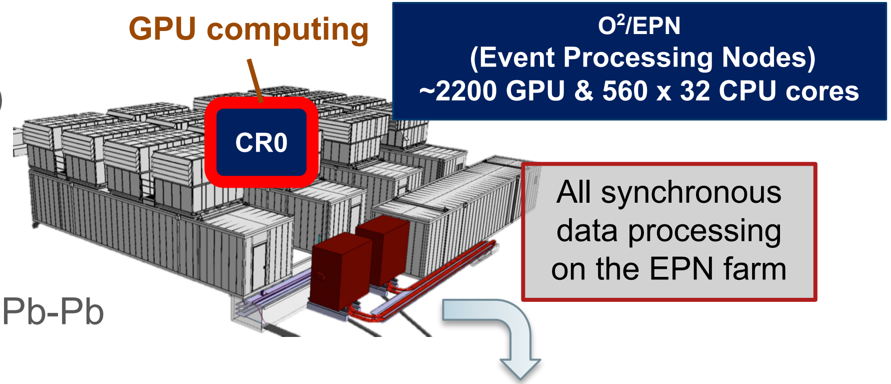
- Basic processing unit: **Time Frame** (~2.5 ms of data ~125 collisions @ 50 kHZ Pb-Pb)
- Whole TF reconstructed in one shot
- In absence of triggers (reference for drift-time estimate) z position of TPC clusters is not defined
- Tracks from different collisions shown in different colors

ALICE Raw Data Flow in Run 3



EPN Farm @ P2

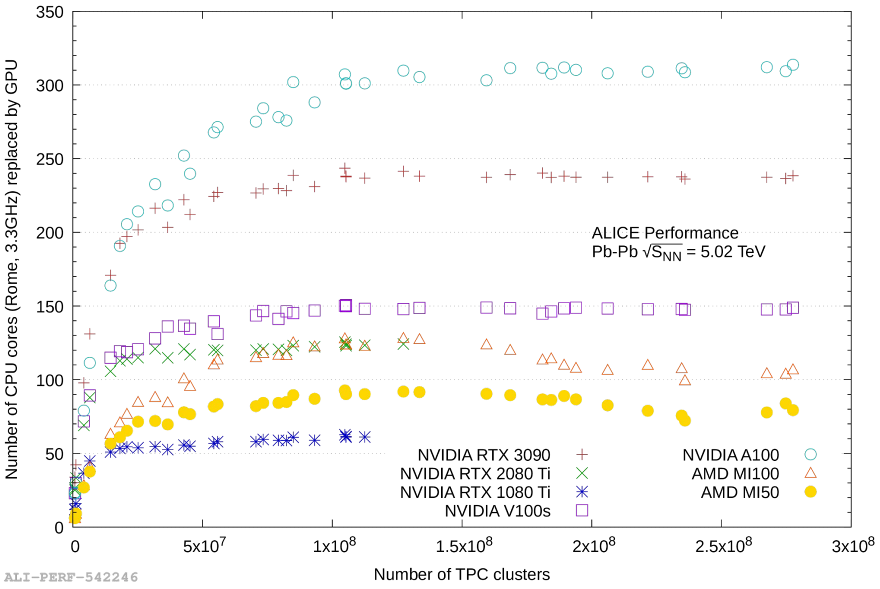
- EPN farm fast enough for synchronous (online) processing:
  - Event and timeframe building
  - Calibration and compression
- Tailored to absorb the maximum data rate with Pb-Pb central collisions at 50 kHz IR
- **280 EPNs, specs defined in 2020 with MC data:**
  - SuperMicro AS-4124GS-TNR
  - **Eight GPUs AMD MI50 – 32 GB**
  - Two CPUs 32 core AMD Rome 7452
  - 512 GB 3200 MHz Main Memory
  - 100 Gb/s InfiniBand HCA
- **GPUs provide the required compute power**
  - Multiple GPUs in a server minimized EPNs cost



ALICE synchronous processing

- Synchronous reconstruction fully dominated by the TPC (99%):
  - TPC online tracking, clustering, compression fully run on the GPUs
  - no margin and no reason to offload anything else to the GPU
- With GPU 280 Event Processing Nodes
  - 2 CPUs x 32 cores + 8 GPUs per EPN
- Without GPU 2800 Event Processing Nodes:
  - 2 CPUs x 32 cores per EPN (~180 kcores)
- **GPU based solution strong impact on hardware and operating cost savings**

Modern GPU replaces 50 - 300 CPU cores @ 3.3 GHz in the ALICE synchronous (online) processing (MI50 GPU replaces ~80 AMD Rome CPU cores)



See David's CHEP2023 plenary talk on Wednesday morning

ALICE

Some implementation details (*)

(*) For a comprehensive description see David's [CHEP2023 plenary talk] on Wednesday morning

- **Vendor- / architecture-independent software:**
  - All algorithms are written in **generic C++**, and can be dispatched to HIP, CUDA, OpenCL on GPUs or OpenMP on CPUs using small **wrappers** ⇒ good code maintainability, tested:
    - AMD GPUs (S9000 with OpenCL 1.2 , MI50 / MI100 / Radeon 7 / Navi with HIP / OCL 2.x)
    - NVIDIA GPUs (RTX 3090 and A100 with CUDA)
  - All GPU code swapped out in dedicated libraries, **GPU libraries linked dynamically on demand** ⇒ can distribute same binary software to CPU and GPU nodes (via CVMFS)
- **Multiple GPUs in a bi proc server:**
  - Synergies of using the same CPU and memory components for multiple GPUs
  - Split the node into 2 NUMA domains to minimize inter socket communication
  - GPUs are processing individual time frames: no inter GPU communication
    - Host processes can drive 1 GPU each, or run CPU only tasks
  - GPUs can be shared between algorithms (with memory reuse if within the same process)

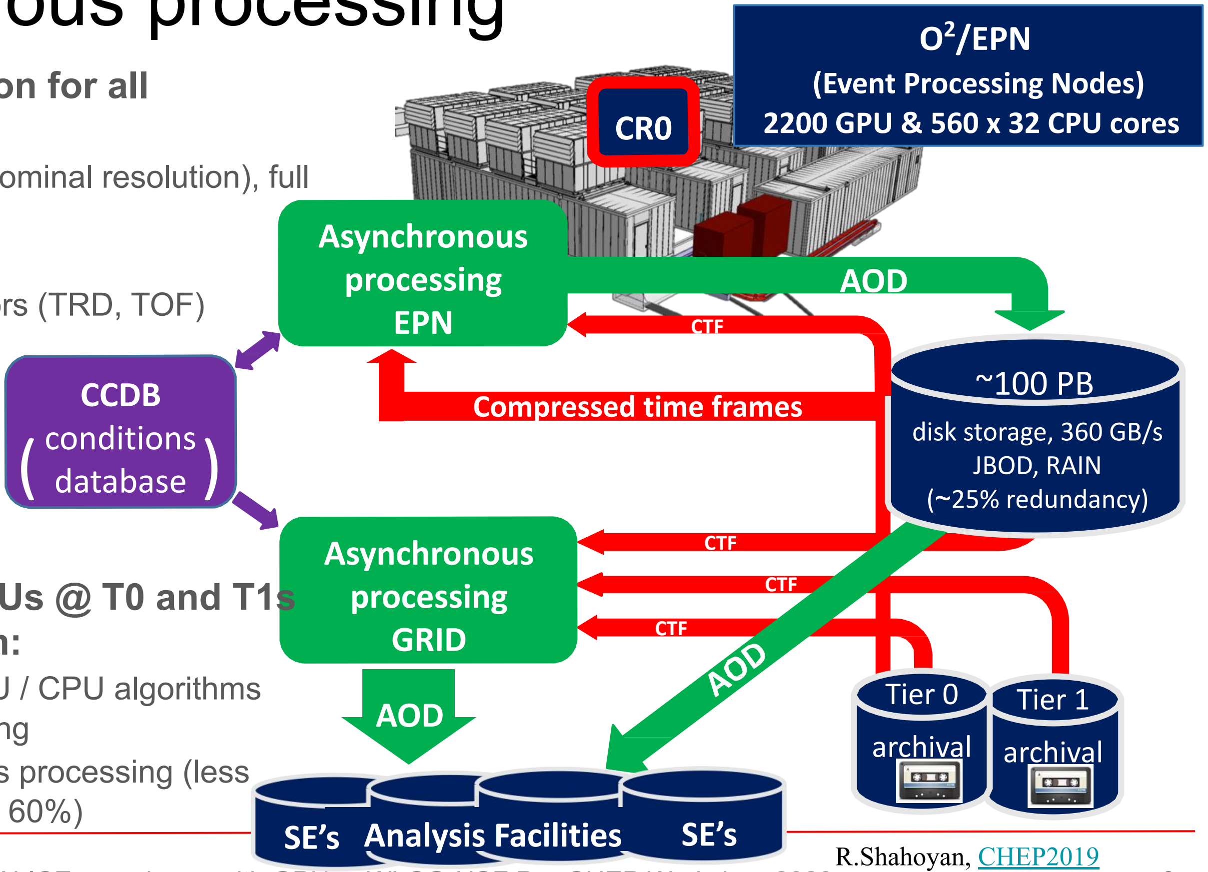
ALICE asynchronous processing

Full reconstruction, full calibration for all detectors:

- Full correction of TPC distortions (nominal resolution), full calibration for all detectors
- TPC – ITS tracks matching
- Tracks propagation to outer detectors (TRD, TOF)
- Global track fits
- Primary and secondary vertices
- PID hypothesis
- AOD as output

Asynchronous code runs on CPUs @ T0 and T1s and on CPUs+GPUs @ EPN farm:

- Different relative importance of GPU / CPU algorithms compared to synchronous processing
- TPC part faster than in synchronous processing (less hits, no clustering, no compression, 60%)

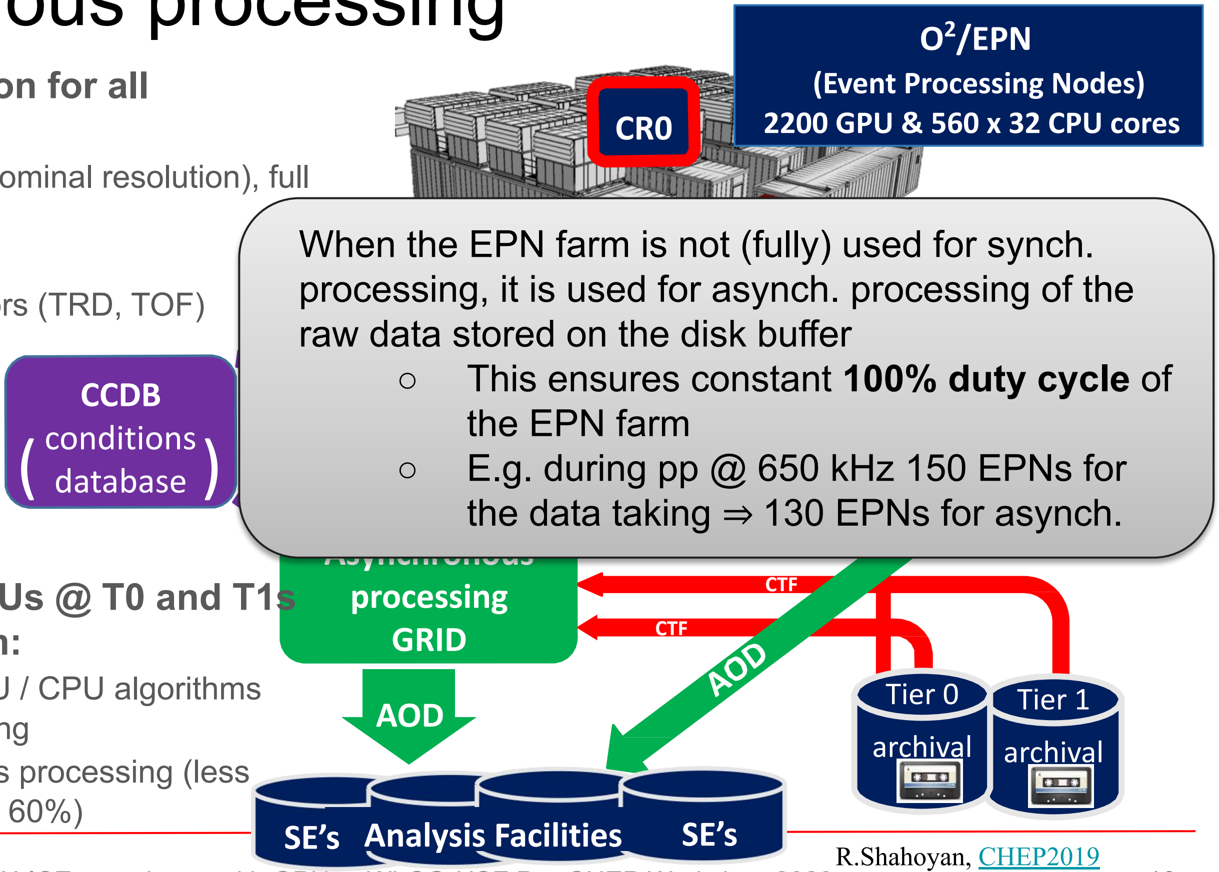


R.Shahoyan, CHEP2019

ALICE asynchronous processing

Full reconstruction, full calibration for all detectors:

- Full correction of TPC distortions (nominal resolution), full calibration for all detectors
- TPC – ITS tracks matching
- Tracks propagation to outer detectors (TRD, TOF)
- Global track fits
- Primary and secondary vertices
- PID hypothesis
- AOD as output



When the EPN farm is not (fully) used for synch. processing, it is used for asynch. processing of the raw data stored on the disk buffer

- This ensures constant **100% duty cycle** of the EPN farm
- E.g. during pp @ 650 kHz 150 EPNs for the data taking ⇒ 130 EPNs for asynch.

Asynchronous code runs on CPUs @ T0 and T1s and on CPUs+GPUs @ EPN farm:

- Different relative importance of GPU / CPU algorithms compared to synchronous processing
- TPC part faster than in synchronous processing (less hits, no clustering, no compression, 60%)

R.Shahoyan, CHEP2019

ALICE

Tuning of asynchronous reco on EPN with GPUs

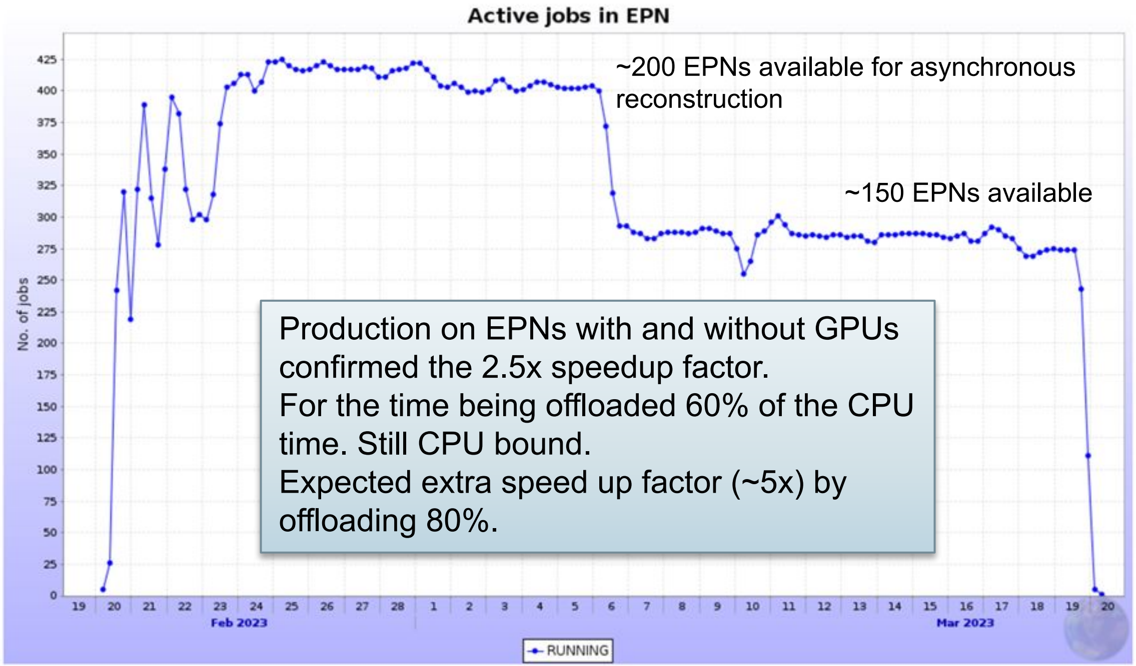
- **Benchmarks of asynchronous reconstruction, pp collisions (real data), 650 kHz inelastic rate, tested different scenarios:**
  - 8-core and 16-core, CPU only: Allocated 4-8 physical cores (8-16 virtual cores) and 16-32 GB of RAM (2 per virtual core) on an EPN
  - 1/8th of an EPN, or Single-GPU: 8 physical cores (16 virtual cores), 64 GB, 1 MI50 GPU
  - 1/2 of an EPN in 1 NUMA domain: 32 physical cores (64 virtual cores), 256 GB, 4 MI50 GPUs – closest to synchronous processing

Configuration (2022 pp, 650 kHz)	Time per TF (1 instance)	Time per TF (full server)
CPU 8 core	76.91s	4.81s
CPU 16 core	34.18s	**4.27s**
1 GPU + 16 CPU cores	14.60s	1.83s
1 NUMA domain (4 GPUs + 64 cores)	3.5s	**1.70s**

Factor 2.51

David's CHEP2023 plenary talk

Asynchronous reconstruction on EPN (CPU+GPU)

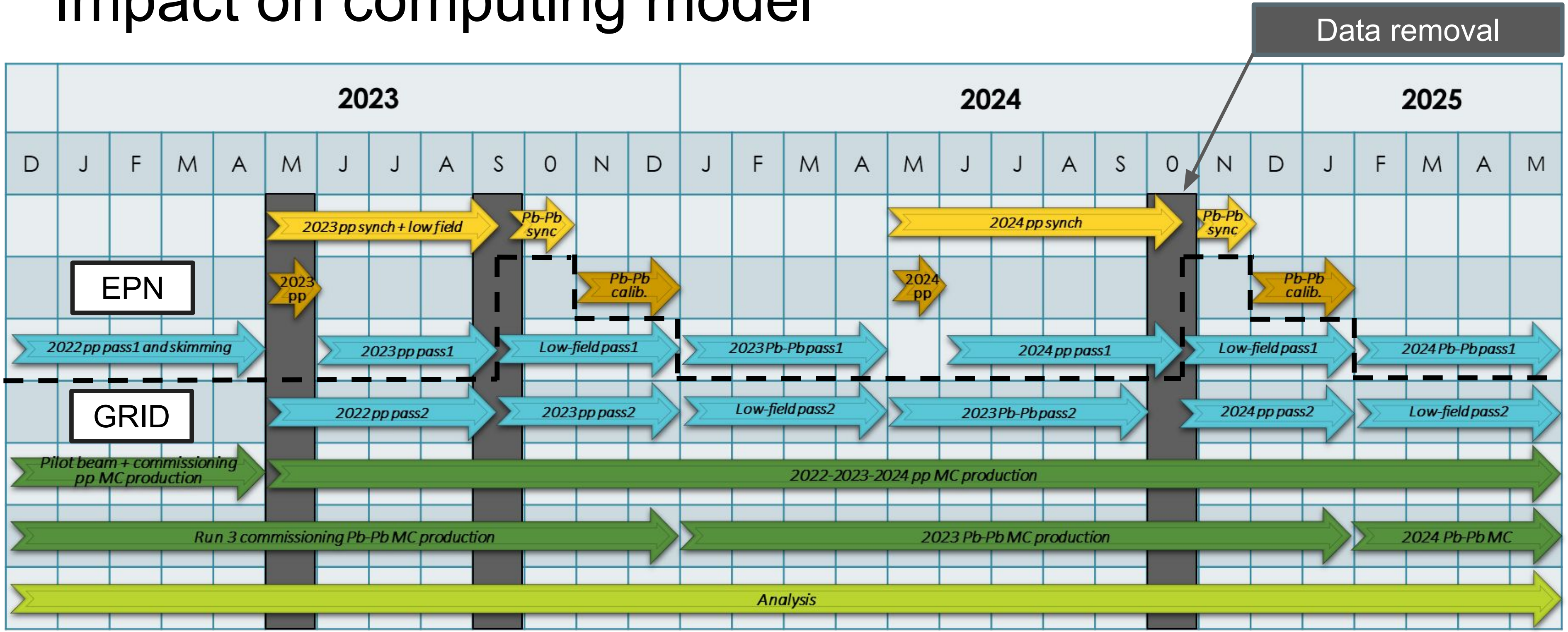


Production on EPNs with and without GPUs confirmed the 2.5x speedup factor.
For the time being offloaded 60% of the CPU time. Still CPU bound.
Expected extra speed up factor (~5x) by offloading 80%.

apass3 production of 2022 pp data @ 650 kHz INEL

- Dynamic allocation of the node partitions (sync and async) managed by Run Coordinator
- The concurrent async. reconstruction productions of **2022 pp data** on GRID (CPU only) and on EPN (CPU+GPU) allowed for the physics validation of the reconstruction with GPU and to speed up processing: 1/3 of data processed on EPN during recent pp apass3
- Tested different configurations in production as well:
  - 16 virtual CPU cores + 1 GPU per job
  - Half EPN (32 physical cores) in NUMA domain with 1 and 4 GPU per job

Impact on computing model



The achieved performances of the asynch. reconstruction on EPN with GPUs allowed to lower our 2024 CPU request

Requirements to run ALICE O2 code on GPUs

- ALICE O2 code is agnostic wrt. GPU vendors, though we don’t test all models on a regular basis.
  - **ALICE current builds on CVMFS support only AMD MI50 and NVIDIA RTX3090**
  - The same software from CVMFS can run on the GPUs, or on the CPUs only
    - Controlled only via variable from the JDL
- ALICE GPU code needs **RHEL8** or compatible
  - **On the EPN farm, we use the EL8 apptainer/singularity container from CVMFS**
- Supporting different architectures and different GPU models is feasible, it requires additional effort:
  - create additional builds for them or update the current builds such that they can support multiple GPU architectures per backend
  - dedicates tuning: relation between GPU performance and CPU performance will be different
- For the time being the processing 50 kHz Pb-Pb data requires:
  - AMD MI*** and NVIDIA GPU with 16 GB
  - From the performance perspective 1 GPU per 8 physical CPU cores is ideal
- For processing 650 kHz pp less resources are needed:
  - AMD MI*** and NVIDIA GPU with 8 GB (although not ideal)
- **It might be possible to run asynch. reconstruction on some grid sites with GPU**

ALICE

Summary

- Large effort in developing and realizing the Run 3 TPC synchronous reconstruction tasks for GPU platform
  - Vendor and architecture-independent software
  - 2200 AMD MI50 GPUs available for continuous readout of Pb-Pb at 50 kHz
- Porting of asynchronous reconstruction code to GPUs:
  - Offloaded 60% of the processing to the GPU with a speedup ~2.5x
  - By offloading 80% we might get close to ~5x
- First asynch. reconstruction productions on EPN farm in 2023:
  - 1/3 of data processed on EPN (CPU+GPU) during recent pp apass3, 2/3 on GRID (CPU)
- ALICE current builds on CVMFS support AMD MI50 and NVIDIA RTX3090:
  - Able to support different architectures and different GPU models
  - **Efficient reconstruction with GPU required tailored job tuning**, taking into account the GPU and CPU models and memory

Acknowledgements

*“This work is **partially** supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU”.*

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Backup

Tuning of asynchronous reco on EPN with GPUs

- **Benchmarks of asynchronous reconstruction, pp collisions (real data), 650 kHz inelastic rate, tested different scenarios:**
  - 8-core and 16-core, CPU only: Allocated 4-8 physical cores (8-16 virtual cores) and 16-32 GB of RAM (2 per virtual core) on an EPN
  - 1/8th of an EPN, or Single-GPU: 8 physical cores (16 virtual cores), 64 GB, 1 MI50 GPU
  - 1/2 of an EPN in 1 NUMA domain: 32 physical cores (64 virtual cores), 256 GB, 4 MI50 GPUs – closest to synchronous processing
- Benchmarks of Pb-Pb **asynchronous reconstruction** on MC Pb-Pb data (2022 data too low IR), tuned only 2 cases, the processing needs more memory than pp due to larger time frames:
  - the CPU-only workflow for reference and the 1 NUMA domain workflow
  - Excluded 1 GPU workflow, not efficient, since we would need more memory. The 1 NUMA domain workflow has synergy effects and is thus not limited by memory
- **Efficient reconstruction with GPU required tailored job tuning**, taking into account the GPU and CPU models and memory