Contribution ID: **73** Type: **15 minute talk**

Efficient Computing with the ALICE Event Processing Nodes GPU farm

Thursday 15 May 2025 11:45 (15 minutes)

\section{Introduction}

The Large Hadron Collider (LHC) at CERN resumed operation in 2022, achieving 13.6 TeV proton–proton collisions.
During the 2019–2021 shutdown, the \mbox{ALICE} detector was upgraded to handle a 50 kHz interaction rate for Pb–Pb collisions,
increasing data volume tenfold compared to previous data-taking periods.
The ALICE Run 3 and 4 computing model, called O2 (Online-Offline), enables continuous readout of sub-detectors,
allowing for synchronous processing of raw data during data taking.
Run˜3 and Run˜4 workflows involve continuous readout from detector front-end electronics, with FPGA boards performing Zero Suppression (ZS).
The TPC outputs 3.3˜TB/s of raw data, reduced by ZS for processing in the EPN farm via an InfiniBand network,
with compressed results transferred to CERN's central IT data center.

\subsection{Efficient data distribution, processing and compression with the EPN farm}

This approach unifies online and offline data processing relying on a single data structure, the Time Frame (TF), which contains all information from the sub-detectors for a given time interval.
TFs are built and compressed on the \mbox{ALICE} Event Processing Nodes (EPN) farm composed of 350 servers and 2800 GPUs achieving compressed data rates up to 4 PB/day,
while the average during Pb–Pb data taking in 2024 with the full LHC orbit was around 2.5 PB/day (comparable with the high rate Pb–Pb period of 2023).
The EPNs are also able to perform the first data calibration pass online. Calibration tasks are split between global calibrations,
which run on CPU-only nodes of the EPN farm, and detector-specific calibrations, which are executed directly on the readout nodes at each LHC fill.
The use of GPU hardware accelerators, with high intrinsic parallelism, reduces costs and energy usage, requiring eight times fewer servers than CPUs for equivalent performance.
The EPN farm is designed for high throughput and low latency, with a focus on energy efficiency and cost-effectiveness.
The data is distributed to the EPN farm by an EPN software module managing the generation of partial TFs containing data from only one detector, directly from the readout nodes.
This module also handles the scheduling and aggregation of the partial TFs into complete TFs at the EPN farm level.
The overall computing efficiency heavily relies on data compression using lossy methods, such as ZS, to reduce data size, while lossless techniques optimize storage.
The TFs from each sub-detector are processed using subsystem-specific algorithms: the resulting integer arrays are compressed using rANS entropy coding,
which efficiently encodes symbols based on their probability distribution, achieving compression ratios close to the entropy limit.
Compared to Huffman coding, rANS reduces the TF sizes by 3\% and outperforms standard compression libraries by up to 15\%.

ALICE's vectorized rANS implementation, using AVX2 instructions, achieves compression speeds of 3200 MB/s for 32-bit symbols,
doubling the performance of standard CPU implementations (Lettrich, M., Fast entropy coding for ALICE Run 3. Proceedings of Science, 2021, https://arxiv.org/abs/2102.09649).

\subsection{Energy-efficiency of the EPN IT infrastructure}

The EPN farm data center, located at the LHC Point 2, uses modular IT containers for scalability
with adiabatic cooling to ensures effective energy use. The containers are air-cooled, and each container has a dedicated Air Handling Unit (AHU) which provides the necessary air flow to cool the servers with a temperature set point of 27 ℃.
The cooling units are designed to operate in adiabatic mode when the temperature of the outside air is too high to be used directly
to maintain the set point temperature of the cold aisle inside the containers.
The adiabatic system uses purified water (produced on-site by the ancillary reverse osmosis plant) to irrigate the cooling units heat
exchanges whenever the air-to-air heat exchange alone is not sufficient to maintain the set point temperature of supply air to the racks.
The adiabatic cooling system is designed to operate in a closed loop, with the sprayed water being recuperated to the unit water tanks.
However, periodic water flush cycles are needed to keep the bacteria that accumulate in the water circuit to an acceptable level.
Additionally, annual shock treatment with biocides is performed to ensure a deeper cleanliness the water circuits.
The reverse osmosis plant does not use any chemical products to purify the water apart from the salt consumption needed to soften the incoming raw water.
The EPN IT infrastructure can operate at a Power Usage Effectiveness (PUE) lower then 1.10,
significantly reducing energy consumption compared to pure mechanical cooling techniques with PUE values around 1.5.
To limit the salt and water consumption, the adiabatic cooling usage is limited to the months of May to September, when the outside temperature exceeds the
capacity of the air-to-air heat exchangers to keep the chosen set point temperature.

\subsection{Topic of the contribution}

The talk will focus on the use of GPUs in the \mbox{ALICE} EPN farm, highlighting their role in data processing and calibration and the energy efficiency of its IT infrastructure.
Performance metrics from the 2023 and 2024 heavy ion running periods will be presented, including the achieved data rates and compression ratios.
Differences between synchronous and asynchronous processing will be discussed, including resource sharing between the two modes.
Comparisons with CPU-only processing will be made, emphasizing the advantages of GPU-based computing in terms of performance and energy efficiency.
Considerations on the GPU effectiveness, compactness, and favorable
cost-benefit ratio will be done, along with the evaluation of benefits of using GPUs in high-energy physics domain,
including the other major LHC experiments.
Details of the EPN farm's energy-efficient infrastructure will be provided, including the adiabatic cooling system and its impact on PUE.
Considerations on the EPN farm's modular design, scalability and adaptability in view of future upgrades, will be also discussed.
The talk will also address the challenges and lessons learned from the EPN farm's implementation,
and provide and overview of the unique expertise gained by the \mbox{ALICE} collaboration in the field of GPU-based computing since 2010,
when the first GPU-based High-Level Trigger farm was deployed and successfully used in the Run 1 and Run 2 periods.

Author: RONCHETTI, Federico (CERN)

Co-author: ERBA, Giada (CERN)

Presenter: RONCHETTI, Federico (CERN)

Session Classification: Submitted Talks