PROCEEDINGS OF SCIENCE

Preparation for ALICE data processing and analysis in LHC Run 3

Giulio Eulisse[a,*] **for the ALICE Collaboration**

[a] *CERN*

E-mail: giulio.eulisse@cern.ch

After the ALICE Long Shutdown 2 detector upgrades, including a new silicon tracker and a GEM-based readout for the TPC, the experiment will operate during LHC Run 3 at a peak Pb-Pb collision rate of 50 kHz, about 50 times higher than in previous running periods. To maximise the significance of physics signals with low S/B ratios for which triggering is not possible, all events will be read out and written to permanent storage without any selective trigger. In order to minimise the costs and computing time of the online and offline systems, data volume reduction is performed synchronous with data taking on the newly installed Online/Offline facility O^2. The facility consists of two types of compute nodes, the First Level Processors (FLP) and the Event Processing Nodes (EPN). Each FLP receives data from parts of individual detectors, performs a first level of data compression by zero suppression as well as calibration tasks, and sends its output to the EPNs over an InfiniBand network. Using the EPN's CPU cores and GPUs, data is reconstructed and further compressed. Moreover, data for detector calibration is created. Online data processing is followed by offline reconstruction passes using fully calibrated data producing the input for data analysis (AOD). In addition, large samples of simulated data as input for detector response and performance studies will be produced.

Here we describe the data processing chain and give an overview of the design choices and implementations for the newly developed software frameworks, which can cope with the unprecedented data rates and volumes. The status of the preparation for data processing and analysis in view of the first physics runs in 2022 is presented.

**** The European Physical Society Conference on High Energy Physics (EPS-HEP2021), ****
**** 26-30 July 2021 ****
**** Online conference, jointly organized by Universität Hamburg and the research center DESY ****

[*] Speaker

1. Introduction

The ALICE experiment [1] is undergoing a major update of its detector during the Long Shutdown 2. In particular this upgrade will consist of a completely new detector readout electronics, a number of completely revamped subdetectors including a completely new silicon tracker (ITS), a new Muon Forward Tracker (MFT) which extends the capabilities of the MUON arm and new readout chambers based on the Gas Electron Multiplier (GEM) technology [2] for its Time Projection Chamber (TPC). This will bring the peak data rate from less than 1kHz of triggered events to sustained 50kHz of continuous readout data during the Pb-Pb period. The net result of this hardware upgrade will be that around one hundred times more events will have to be reconstructed online and stored. All this will have to happen in a so called "flat budget" scenario where the computing resources are projected to grow only by a factor of four during the next ten years.

In order to cope with the computing challenges posed by the new detector a radical redesign of the experiment software and computing architecture was needed. Such a new design was based on the experience gained by the ALICE HLT during Run 1 and Run 2, expanded in scope to handle every aspect of the experiment data processing.

The main design decision is driven by the need to compress TPC data online to make sure it can be stored. For this reason only analysis objects, dubbed AODs, will be readily available. The computational cost needed for their production will be therefore traded to reduce the storage needs. Moreover, a simplified data model will be adopted to further improve the I/O performance. The use of general purpose Graphics Processing Units (GPUs) and algorithm appropriately chosen to exploit their parallel architecture will be key in order to match the needed computing performance.

However, the technical aspects alone will not be sufficient. In order to tackle the challenge, the software and computing experts will have to work in close collaboration with the physics community in order to organise the analysis efforts. Moreover a close collaboration with the GSI Helmholtzzentrum für Schwerionenforschung and the Facility for Antiproton and Ion Research (FAIR) has been established on a common software stack [3].

2. ALICE data processing in Run 3

The new software and computing architecture is depicted in Fig. 1.

The most notable feature of the new architecture is the fact that the traditional boundary between online and offline is blended in a unique system implemented with a coherent software solution. In such a design, there are two main processing phases, one called "Synchronous Reconstruction" which happens while data is being taken and second one which is called "Asynchronous Reconstruction" which follows the Pb-Pb data taking periods and it's interleaved with the less demanding p-p periods. In the first phase the goal is to compress TPC data as much as possible, to minimize the cost related to its storage, while in the second phase, improved calibration will be used and the totality of the events and detector will be reconstructed. In order to implement this, two layers of processing nodes will be required. A set of 192 First Level Processor nodes will be equipped with the readout cards and are responsible to extract the data from the various detectors. Depending on the detector, local (on the given FLP) processing of the data might happen and some initial Quality Control (QC) might be performed on a subset of the data. The data acquired by each FLP will be

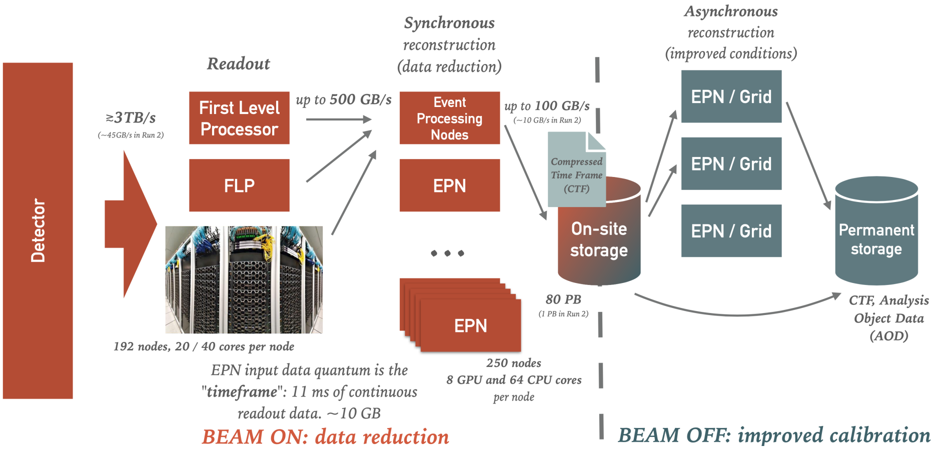

Figure 1: ALICE computing architecture for Run 3

subdivided in equal lenght time periods of 11ms and it will be sent in a synchronized manner to one of the nodes in the second cluster, composed of Event Processing Nodes (EPNs).

The EPN cluster is composed by 250 dual core AMD Rome nodes, for a total of 64 physical CPU cores. Each node will be equipped with 8 AMD MI 50 GPUs. The bulk of the synchronous processing will then happen on the EPNs, and in particular thanks to the massively parallel computing abilities of the GPUs.

3. Synchronous reconstruction

Instrumental for the Synchronous Reconstruction phase is the usage of GPUs to reconstruct the barrel tracks in the TPC and use the reconstructed quantities to reduce the precision needed to store the clusters associated to a track, hence improving the compression ratio achievable by the subsequent entropy compression step [4]. In our benchmarks one modern GPU core provides the computing capacity equivalent of 40 CPU cores, with a net benefit in terms of cost of a factor 4. In the Pb-Pb baseline strategy, reconstructing the ITS tracks of the 5% most peripheral collisions will also be done on the CPU to better calibrate the TPC. In pp, all the ITS tracks will be reconstructed on the CPU during the synchronous phase.

At the time of writing 99% of the computing time required to perform the synchronous phase will be on the GPU.

In order to evaluate different platforms and for better risk management, the code which runs on the GPU is implemented in platform agnostic C++ with special macros which allow targeting different architectures, in particular: AMD HIP, nVidia CUDA, OpenCL. To simplify development and debugging the same code can also target normal CPU based processing.

The result of the synchronous reconstruction will be a number of Compressed Time Frames (CTFs) which will be store on a 80 PB disk buffer.

4. Asynchronous processing

The asynchronous processing phase follows the Pb-Pb period and it will be interleaved with pp synchronous processing. Two processing cycles per data taking period will be performed, using each time more refined calibration and alignment. Processing is foreseen to happen on the EPN farm for 2/3 of the CTF volume and on the Worldwide LHC Computing Grid for the remaining 1/3.

Currently, over 80% of the CPU - equivalent computing time is running on the GPU. While not strictly necessary in terms of latency, making good use of the GPUs is crucial for effectively using the hardware resources at disposal while the EPN farm is not taking data.

After the 2nd cycle the CTF will remain on tape. Any subsequent reprocessing cycle will have to wait until the next LHC Long Shutdown period.

The final result of the Asynchronous reconstruction will be the persistent analysis object output, dubbed Analysis Object Data (AOD). All the analysis will have to be performed on such data and the derived objects.

5. Data Processing Layer

To coordinate the processing of the data and to simplify the task of writing algorithms by physicists, ALICE developed a software framework, named Data Processing Layer [5] (DPL), which hides the complexities of having to deal with a distributed system and presents a traditional, task based, interface to the user. In such a framework the user provides a high level description of the computation, specifying inputs and outputs and how to process the former to obtain the latter, and the DPL takes care of creating the actual distributed topology implemented using the message passing toolkit FairMQ [3]. All the data processing needs of the experiment, from data taking to analysis will be provided by such a framework, ensuring familiarity to the user and reducing the maintenance effort.

6. Analysis

Like the rest of the software stack, also the Analysis Framework has been revamped for Run 3 and it was redesigned to be built on top of the DPL. That said, the main ideas on how the computing will happen will remain the same. In particular ALICE has a long standing effort towards organized Analysis, nicknamed "trains", with the goal of amortizing the cost of the data access by different tasks, or so called wagons. This approach was particularly successful in Run 1 and Run 2 and we are confident that it will allow us to process 100 times more collisions as required in Run 3. In order to scale the system to match the requirements two different approaches have been followed. On the software side, we have streamlined the data model, trading generality for speed and flattening data structures to achieve higher I/O performance. On the computing side, a new entity, the Analysis Facility (AF) has been introduced in our computing model. The AF is a computing cluster which has been particularly tuned for highly performant data access and which will allow the train system to run on 10% of the data in an optimised way. AFs will be used to provide the needed rapid turnaround required by the development of an analysis. When a given analysis qualifies on the smaller subset of data present at the analysis facility, the final production will be launched on the whole Grid. To complete this approach, a set of highly targeted ntuples will be produced, to optimize the turnaround for some key analysis, like the heavy flavour ones which would profit from highly filtered dataset.

The goal, as set in the ALICE O^{2}TDR [6] is to have an Analysis Facility to go through the equivalent of 5PB of AODs every 12 hours.

6.1 Analysis Framework

The Analysis Framework of the experiment has been completely rewritten to run on the same software stack, the O^{2}DPL, as the rest of the data processing.

Each Analysis Task is a DPL device, taking advantage of the innate parallelism of a message passing framework.

The data model has been simplified and now consists of cross indexed tables, like in a relational database, rather than a hierarchy of objects. However, in order to keep an object oriented feel for the user, an Object Relational Mapping (ORM) API has been provided to hide the backing store and allow users to write familiar statements like `track.pt()` [7]. A simplified analysis example is provided in Listing 1

```
struct MyAnalysisTask {
  Filter vertexFilter = nabs(collision::posZ) < 7;
  Filter ptFilter = track::pt > 0.5f;
  OutputObj<TH1F> hist{TH1F("pt", "pt", 20, 0., 10.)};
  void process(Collision const& collision,
               Tracks const& tracks)
  {
    for (auto &track : tracks)
      hist.Fill(track.pt());
  }
};
```

Listing 1: a simplified example of an analysis, implemented in the new Analysis Framework. The declarative part of the filters are expressions which are constructed and stored as member variables, while the imperative part is implemented inside the `process` function

The actual backing store for the table is provided by the Open Source project Apache Arrow [8], which provides a backend for in memory storage of columnar data. The backend was selected as it is already widely used in the data analysis industry at large and it provides abstraction of the underlying memory model and seamless integration with programming languages like Python and machine learning toolkits like TensorFlow.

While designing the new framework, special care has been taken to allow both declarative statements, like filters, which can be bulk applied to the data upfront in an optimized way and a so called imperative part, where object oriented expressiveness of an object oriented framework is retained. For the declarative part, the user can provide an expression on the column of a given table. All the specified expressions are collected by the framework, compiled upfront and bulk applied on the data, using the Arrow provided expression engine Gandiva [9].

6.2 Analysis Operations

The infrastructure used for the analysis production operations, now dubbed Hyperloop, have also been revamped to integrate the new framework and take advantage on the new capabilities provided. The new version has been put in place and it's currently being used for all the Run 3 analysis tests. It shields the user from the mechanics of Grid jobs submission, handling of occasional failures and output merging. Moreover, it allows the users to locally test their tasks, providing top-level metrics about their performance as well as detailed profiling information using perf[1] and SpeedScope[2].

7. Conclusions

During the Long Shutdown 2 of the LHC, ALICE overhauled the computing and software infrastructure of the experiment to cope with the challenges posed by Run 3. The new system is currently being commissioned and data recorded from the first Run 3 collisions is being reconstructed.

References

[1] The ALICE Collaboration, *The alice experiment at the cern lhc*, *Journal of Instrumentation* **3** (2008) S08002.

[2] ALICE Collaboration collaboration, *(W)hole new field: the new GEM Time Projection Chamber of ALICE*, *J. Phys.: Conf. Ser.* **1561** (2019) 012017. 11 p [1912.08673].

[3] M. Al-Turany, D. Klein, A. Manafov, A. Rybalchenko and F. Uhlig, *Extending the fairroot framework to allow for simulation and reconstruction of free streaming data*, vol. 513, p. 022001, 2014, http://stacks.iop.org/1742-6596/513/i=2/a=022001.

[4] D. Rohr, *Gpu-based reconstruction and data compression at alice during lhc run 3*, *EPJ Web of Conferences* **245** (2020) 10005.

[5] G. Eulisse, P. Konopka, M. Krzewicki, M. Richter, D. Rohr and S. Wenzel, *Evolution of the ALICE software framework for Run 3*, *EPJ Web Conf.* **214** (2019) 05010.

[6] P. Buncic, M. Krzewicki and P. Vande Vyvre, *Technical Design Report for the Upgrade of the Online-Offline Computing System*, Tech. Rep. CERN-LHCC-2015-006, ALICE-TDR-019, The ALICE Collaboration (2015).

[7] Alkin, Anton, Eulisse, Giulio, Grosse-Oetringhaus, Jan Fiete, Hristov, Peter and Kabus, Maja, *Alice run 3 analysis framework*, *EPJ Web Conf.* **251** (2021) 03063.

[8] The Apache Arrow team, *From the apache arrow wiki: Physical memory layout*, 2016.

[9] Dremio, *Introducing the gandiva initiative for apache arrow*, 2018.

[1] https://perf.wiki.kernel.org

[2] https://www.speedscope.app