AUTOMATIC DATA SELECTION WORKFLOW FOR THE RECONSTRUCTION PROCESS IN ALICE EXPERIMENT

Alice-Florenţa Şuiu[1], Costin Grigoraş[2], Nicolae Țăpuș[3], Latchezar Betev[4]

The LHC, run by CERN, is the world's biggest and most potent particle accelerator. Four large experiments that study high-energy proton-proton and heavy-ion collisions are located at the LHC. The main goal of ALICE, one of the four experiments, is to investigate the physics of heavy-ion collisions. A thorough update of the detector, readout and data processing software was followed by the start of the third data-taking phase (Run 3) in July 2022. Since then, about 355 PB had been collected as of September 30, 2024. The collected physics data is reconstructed, then analyzed. Specialized reconstruction software extracts from the data as much information as possible about the kinematic properties and identity of the particles produced in the collisions. The reconstruction processing is applied only to datasets that meet specific quality criteria.

This article details the implementation of an automatic data selection workflow for the reconstruction process within the ALICE experiment. This tool has replaced the previously used manual workflow managed by ALICE operators, representing a significant step in optimizing data processing and management in the ALICE experiment.

Keywords: data acquisition system, distributed computing infrastructure, big data, reconstruction process

1. **Introduction**

Located along the French-Swiss border, the European Organization for Nuclear Research (CERN) [1] offers sophisticated accelerators and technological infrastructure for studies of the basic structure of matter. With the aim of testing the Standard Model and explore fundamental questions in particle physics, four large experiments are operated at the Large Hadron Collider

[1] PhD student, Faculty of Automatic Control and Computer Science, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: `alice_-florenta.suiu@upb.ro, asuiu@cern.ch`

[2] Software Engineer, CERN, ALICE experiment, Geneva, Switzerland, e-mail: `costin.grigoras@cern.ch`

[3] Professor, Faculty of Automatic Control and Computer Science, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: `nicolae.tapus@upb.ro`

[4] Software Engineer, CERN, ALICE experiment, Geneva, Switzerland, e-mail: `latchezar.betev@cern.ch`

(LHC) [2], a particle accelerator built specifically for research of high energy proton-proton and heavy ion collisions.

Specifically designed to study heavy-ion collisions and provide accurate measurements of the properties of the quark-gluon plasma, A Large Ion Collider Experiment (ALICE) [3], is one of the four experiments at the LHC. So far, this experiment has had three major periods of operation. During Run 1 (2009-2013) [4], approximately 7.4 PB of data were collected and during Run 2 (2015-2018) [5], approximately 28 PB, totaling about 35 PB. After that, there was a long shutdown period in which the entire Data Acquisition System (DAQ) was decommissioned and replaced with a different structure whose architecture is described in section 2. In July 2022, the third data-taking phase (Run 3) [6] commenced following the substantial update of the detector, readout and data processing software. Since then, up until the time of writing (September 30, 2024), around 355 PB of data had been collected. The Run 3 phase of ALICE aims to record 100x more minimum bias collisions than Run 1 and 2 combined. This large volume of data needs to be transferred from the experiment location to the persistent storage located at CERN main computing centre at a rate up to 160 GB/s. The data volume is composed of three categories of data files: *raw*, *calibration* and *other*.

Time Frames (TFs) [7] and Compressed Time Frames (CTFs) [7] are the two file formats classified under the *raw* category. A TF file contains a contiguous 2.85 ms time range of the data collected from all ALICE subdetectors. Multiple TFs are compressed using entropy compression to reduce the size and some data that is not needed for physics analysis is discarded in this process. This is known as CTF compression and achieves a compression factor 5, corresponding to a reduction in size relative to the uncompressed size of 80%. Due to storage constraints, mostly CTFs but also sample TFs are stored on the persistent storage at CERN and after a series of reconstruction cycles are made available for the physicists to analyze the experiment's data [8].

The second, the *calibration* category, contains information about the accelerator and the ALICE sub-detectors status and conditions that is periodically collected and stored as a parallel stream of data files.

The data, which belongs to the *other* category is collected during commissioning tests or sub-detector specific tests that have a limited intended group of users.

Long and contiguous periods of time in which the experiment's and LHC's conditions do not change significantly are organized in so called *runs*. A run can last from several minutes to the time it takes for the LHC beam fill to be dumped, which can extend up to 12 hours. The files in the three previously mentioned categories are internally partitioned based on the run number and further split into 10 minute periods to simplify data organization and workflow management. More details about the attributes of a run can be found in Ref. [8].

The reconstruction process involves reproducing the events that occurred during the beam collisions and uses a purpose written software which operates on the CTFs [9]. This process is applied to a list of runs that meet a set of criteria detailed in section 3. The results are saved in Analysis Object Data, second generation (AO2D) [10] files. These files are then used in the physics analysis.

This article describes an automatic workflow for initiating the reconstruction process for the ALICE experiment. It assists ALICE operators by enabling them to more quickly and efficiently select the dataset they wish to reconstruct and subsequently analyze for the physical events observed in the experiment. Each dataset must meet specific criteria to be considered for reconstruction. The automatic tool thus operates only on those datasets that satisfy all the required criteria set by the operators. This workflow has been in production since May 2024 and has until now (Sep 2024) automatically triggered the reconstruction workflow on 385 run numbers corresponding to 148 PB of data collected by the experiment. Prior to its implementation, ALICE operators performed manual selections, identifying new datasets and applying necessary filters to ensure they met the requirements for valid reconstruction. The introduction of the automatic workflow eliminated the tedious manual operations, allowing the operators to focus more effectively on the core objectives of the experiment.

The structure of this article is as follows: section 2 describes the path that data takes from the moment it is produced by the ALICE detector until it is selected for the reconstruction process, section 3 provides an overview of the implementation details underlying the automatic data selection workflow, and section 4 presents the tool's performance since its deployment in May 2024.

2. Data flow from collection to reconstruction

The ALICE experiment generates a large volume of data that is transmitted with a total throughput of 3.5 TB/s to the first layer of computing nodes, called First Level Processors (FLPs) [11]. This layer of nodes performs the initial compression of the data (noise corrections and zero suppression) and forwards a throughput of 900 GB/s to the next layer of computing nodes, known as Event Processing Nodes (EPNs) [11]. The 350 EPNs assemble the individual detector streams into TFs containing the full detector information and perform the second level of compressions, which results in the CTFs. The CTFs are then stored on local disks, at a rate of up to 500 MB/s per disk. From here, a tool called EPN2EOS [12] handles the transfer of the CTFs to the persistent storage in quasi-synchronous operation with the data-taking. An instance of EPN2EOS runs as a daemon on each EPN node and is exclusively responsible for managing the collected experiment data.

When the EPN farm is not fully employed for the online compression, it is used for processing the data that was collected so far. This step is called asynchronous (with regard to data-taking by the experiment) reconstruction, which takes CTF files as input and produce the AO2D files used for physics analysis.

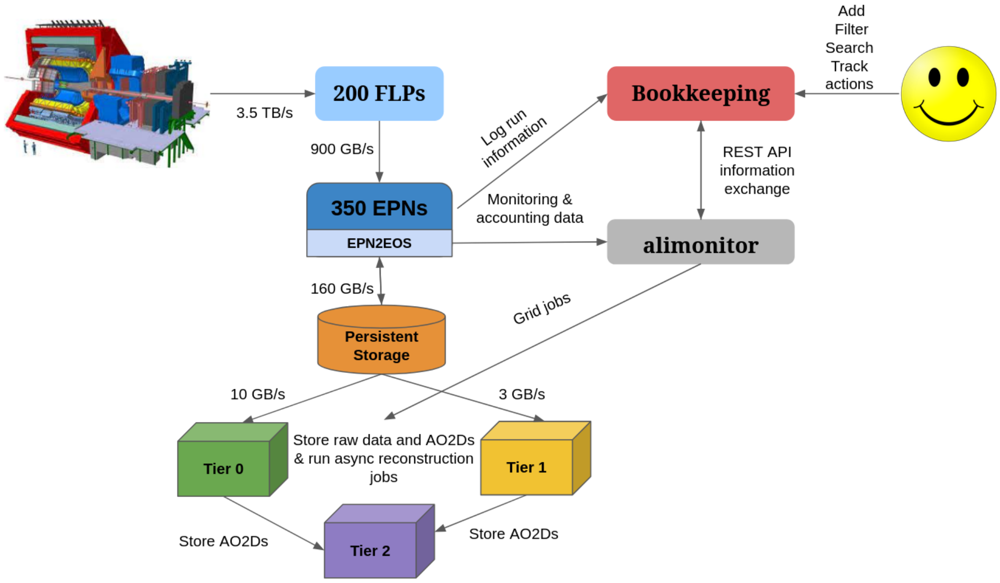


Fig. 1. Data flow diagram

Once on the persistent storage, the data is available for the asynchronous reconstruction step. It can be copied or moved to external storage or can be deleted if a Quality Assurance (QA) process [13], which runs over parts of the CTFs and determines if the data is below the quality threshold required for asynchronous processing [9].

The persistent storage is the data entry point in the ALICE distributed computing infrastructure and is part of the CERN storage infrastructure located in Meyrin site. The ALICE computing model organizes the constituent computing centers in tiers [14].

Tier 0, located at CERN, serves as the central hub for data processing and storage for the four LHC experiments. It is equipped with 201 PB of disk storage, 825 PB of tape storage, and substantial computational power, providing 2690 kHS23 [15]. For the ALICE experiment specifically, Tier 0 is used to store raw data and the results obtained from the reconstruction process, with current pledged capacity of 67 PB of disk storage, 181 PB of tape storage, and 600 kHS23 [16].

Tier 1 is the next level of computing centers with a total capacity of 418 PB of disk storage, 1072 PB of tape storage, and computational power of 3650 kHS23 for all LHC experiments [16]. These centers are connected to

Tier 0 through dedicated optical networks (LHCOPN [17]). Among the Tier 1 centers, 7 are dedicated to the ALICE experiment, providing a combined capacity of 70 PB of disk storage, 110 PB of tape storage, and 630 kHS23 of computational power.

These two categories of computing centers ensure redundancy of the data collected from the experiment, which is initially saved on the persistent storage. Thus, by moving data from the persistent storage to the Tier 0 or Tier 1 elements, enough free space is ensured on the persistent storage for future data collection.

The third category of computing centers, Tier 2, consists of 37 centers for the ALICE experiment that offer a smaller storage capacity, approximately 70 PB, and a computational power of about 650 kHS23 [15]. These elements mainly store the results obtained from physics analyses for which two copies of data are kept.

Bookkeeping [18] is a real-time logging platform that provides a graphical interface, allowing ALICE operators to add new actions, filter, search, or track actions that have already been logged. Bookkeeping is a web application that keeps track of all the actions performed by ALICE operators during the experiment, providing an overview of the detector operation and general data-taking conditions. Through this platform, information is collected, processed and redistributed to a wide range of clients, including both people (ALICE operators) and software applications.

Alimonitor [19] is a computing node that collects monitoring data from the distributed computing infrastructure and takes automatic actions to steer the activity on it. One of the applications that is running on this node is the automatic data selection tool. It is written in Java and regularly retrieves data from Bookkeeping using its REST interface (HTTP GET requests), specifically focusing on runs from the past week that have not yet been flagged for reconstruction. After pulling the data from Bookkeeping, the process applies certain filtering criteria, which are detailed in section 3. The actual reconstruction process is then initiated for the runs that fulfill these conditions.

In summary, the system architecture and data flow discussed throughout this section are illustrated in Fig. 1, offering a comprehensive view of the processes from data collection to reconstruction.

3. Automatic data selection workflow: implementation details

This section provides a comprehensive description of the implementation details of the automatic data selection workflow for the reconstruction process.

A run is represented by contiguous periods of time in which the sub-detector compositions do not change significantly. Since the beginning of Run 3 and up until the time of writing, the average run duration is about 4 hours.

Each run is characterized by a set of attributes, detailed in the paper [8]. The following attributes are relevant for the automatic data selection tool used in the reconstruction process:

- Run number: a unique integer designating a continuous period of time during which data is collected;
- Quality: a global indicator of a run's quality, initially set to N/A and later updated to Good, Test, or Bad;
  - Good: denotes that the run's data is of a high enough quality to be utilized for physics study in the future;
  - Test: denotes that the run's data is ephemeral and may be removed from the primary storage;
  - Bad: denotes that the run's data is of insufficient quality. After the data is analyzed to identify the causes of their invalidation, it can be removed from the primary storage.
- Length: duration of a run;
- Beam type: type of particles involved in the collision; some examples are presented below
  - Pb-Pb signifies collisions between lead nuclei;
  - pp denotes proton collisions;
  - and combinations of the two above.
- Run type: category of a run, which can be:
  - PHYSICS: contains physics-related events resulting from particle collisions;
  - COSMICS: data collected in the absence of an LHC beam and without a magnetic field in the experiment, measuring the cosmic background radiation and used for calibration purposes.
- Detectors: a list of (sub)detectors involved in a specific run.

The values of these attributes for each newly created run are retrieved from Bookkeeping via its REST interface, providing the run number as argument and expecting a JSON formatted response. This type of request allows to retrieve any field associated with a run from Bookkeeping. Figure 2 provides an example of the response received following an HTTP GET request for a specific run.

```
{
  "runNumber": 529038,
  "runType":      "PHYSICS",
  "runQuality":   "good",
  "beamType":     "PROTON - PROTON",
  "runDuration":  22171000,
  "lhcBeamMode":  "STABLE BEAMS",
  "lhcBetaStar":  19.2,
  "aliceL3Current":       29999.9,
  "aliceDipoleCurrent":   5999.95,
  "aliceL3Polarity":      "POSITIVE",
  "aliceDipolePolarity":  "POSITIVE",
  "lhcPeriod":      "LHC22q",
  "pdpBeamType":    "pp",
  "detectors": [
    "CPV","EMC","FDD","FT0","FV0","HMP",
    "ITS","MCH","MFT","MID","PHS","TOF",
    "TPC","TRD","ZDC"
  ]
}
```

Fig. 2. GET response from Bookkeeping for run 529038

The automatic data selection tool, reading from Bookkeeping, is triggered automatically after the data files have been successfully transferred to the persistent storage element. Upon receiving the response from Bookkeeping, the tool parses it, extracts the relevant attributes for selection and verifies if the values meet the following criteria:

- Run quality is GOOD;
- Run type is PHYSICS;
- Beam type is either Pb-Pb or pp;
- Run duration exceeds 300 seconds for pp collisions or 120 seconds for Pb-Pb collisions;
- Detectors list includes the critical detectors selected by ALICE operators: TPC, ITS, FT0, +ZDC [6] (for Pb-Pb).

When a run's attributes meet the specified criteria, it is considered suitable for the reconstruction process and is then queued. The experiment's workload manager schedules jobs corresponding to the requested processing on the Grid computing nodes as computing resources become available. Consequently, this process operates entirely asynchronously from the data acquisition itself.

For each run whose data files have been successfully transferred, the data selection workflow outlined in Algorithm 3 is applied. A request is made to the Bookkeeping to retrieve the attributes of the processed run. Subsequently, it is checked whether the run meets the criteria for reconstruction processing.

If the run's quality has not yet been set, it will temporarily be ignored and revisited later, allowing time for its quality status to be determined. Only runs marked as GOOD are eligible for reconstruction. Consequently, if the run's quality is set but classified as BAD or TEST, it will fail the selection process and will not be considered for reconstruction.

If the run's quality is marked as GOOD, the next step is to verify its type. Runs that are not of the type PHYSICS are excluded, as they do not contain physics-related events and are therefore irrelevant to the reconstruction process. If the run type is correct, the beam type used during data collection is then checked. Runs with a beam type other than Pb-Pb or pp will also be excluded from the selection process.

If all the above-mentioned criteria are met, the next check is performed on the run's duration. This duration depends on the beam type: for Pb-Pb beams, the run must last longer than 120 seconds, while for pp beams, the minimum duration is 300 seconds. Only runs meeting these duration requirements are considered sufficiently long to qualify for reconstruction.

Finally, the run must pass a check on the detector list. This list must not be empty and must include the critical detectors selected by ALICE operators, namely TPC, ITS, and FT0. If the beam type is Pb-Pb, the ZDC detector is also required. Only runs satisfying this final detector check advance further in the reconstruction process.

For runs that successfully pass all selection criteria, the reconstruction process is initiated. The experiment's workload manager assigns and executes the required processing jobs on the Grid computing nodes as resources become available.

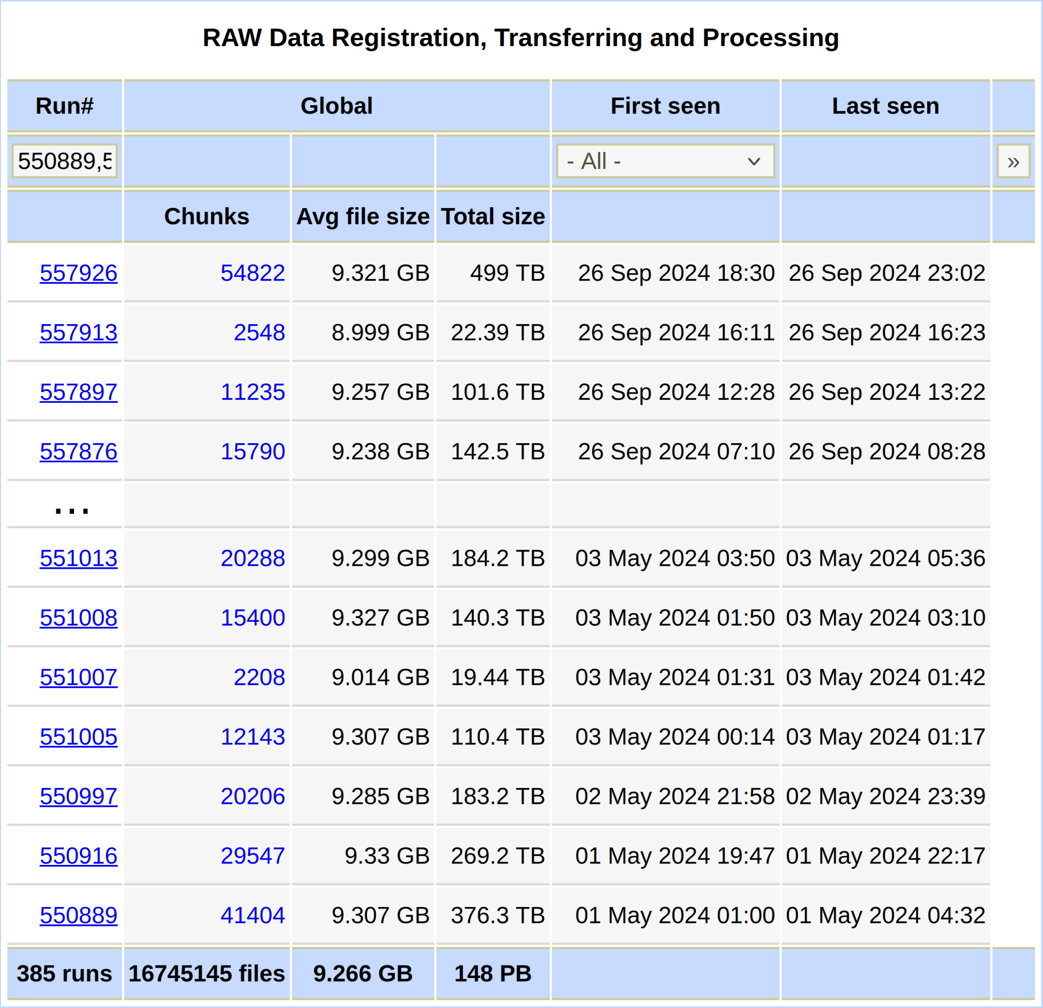
Algorithm 3.1 Run Selection for Reconstruction

```
Require: $runNumber \in \mathbb{N}$
Ensure: Only good physics runs of sufficient length with all required detectors
    are sent to reconstruction.
                                                        ▷ Definitions
$runInfo \leftarrow$ FetchRunInfoFromBK($runNumber$)
$Quality \leftarrow runInfo$.quality          ▷ one of {N/A, BAD, TEST, GOOD}
$RunType \leftarrow runInfo$.type             ▷ PHYSICS or other
$BeamType \leftarrow runInfo$.beamType        ▷ Pb-Pb, pp, ...
$D \leftarrow runInfo$.duration               ▷ total duration in seconds
$Detectors \leftarrow runInfo$.detectors      ▷ set of detector names
                                                        ▷ Constants
$D_{PbPb} \leftarrow 120$                     ▷ min. duration for Pb–Pb runs (s)
$D_{pp} \leftarrow 300$                       ▷ min. duration for pp runs (s)
$CoreDet \leftarrow$ {TPC, ITS, FT0}
$ValidBeams \leftarrow$ {Pb-Pb, pp}
                                                        ▷ Checks
```

```
if $Quality$ = N/A then
    Log(”Quality flag pending; skipping run.”)
    return
end if
if $Quality$ = BAD or $Quality$ = TEST then
    Log(”Run marked as BAD/TEST; skipping run.”)
    return
end if
if $RunType \neq$ PHYSICS then
    Log(”Non-physics run; skipping run.”)
    return
end if
if $BeamType \notin ValidBeams$ then
    Log(”Unsupported beam type; skipping run.”)
    return
end if
if $BeamType$ = Pb-Pb and $D < D_{PbPb}$ then
    Log(”Pb–Pb run too short; skipping run.”)
    return
end if
if $BeamType$ = pp and $D < D_{pp}$ then
    Log(”pp run too short; skipping run.”)
    return
end if
if $Detectors = \emptyset$ then
    Log(”No detectors recorded; skipping run.”)
    return
end if
if $CoreDet \not\subseteq Detectors$ then
    Log(”Missing core detectors; skipping run.”)
    return
end if
if $BeamType$ = Pb-Pb and ZDC $\notin Detectors$ then
    Log(”Missing ZDC detector for Pb–Pb; skipping run.”)
    return
end if
Log(”Run accepted; starting reconstruction.”)
StartReconstruction($runNumber$)
```

4. **Automatic data selection workflow: results**

Since its deployment in May 2024, the automatic data selection tool has been operational. As of September 30, 2024, the tool has effectively selected 385 out of 1794 runs — representing a data volume of approximately 148 PB — for which the reconstruction procedure was initiated and completed. This outcome is displayed in Fig. 3. Out of the total runs, 1409 runs were TEST or BAD and occupied a data volume of approximately 2 PB. Therefore, although we reconstructed only 22% of the collected runs, we processed 99% of the total data volume during reconstruction. This high exclusion rate is primarily because the majority of runs are designated as TEST or BAD to ensure only high quality physics data enters the reconstruction workflow.



RAW Data Registration, Transferring and Processing

Run#	Global			First seen	Last seen	
550889,5				- All -		»
	Chunks	Avg file size	Total size			
557926	54822	9.321 GB	499 TB	26 Sep 2024 18:30	26 Sep 2024 23:02	
557913	2548	8.999 GB	22.39 TB	26 Sep 2024 16:11	26 Sep 2024 16:23	
557897	11235	9.257 GB	101.6 TB	26 Sep 2024 12:28	26 Sep 2024 13:22	
557876	15790	9.238 GB	142.5 TB	26 Sep 2024 07:10	26 Sep 2024 08:28	
...						
551013	20288	9.299 GB	184.2 TB	03 May 2024 03:50	03 May 2024 05:36	
551008	15400	9.327 GB	140.3 TB	03 May 2024 01:50	03 May 2024 03:10	
551007	2208	9.014 GB	19.44 TB	03 May 2024 01:31	03 May 2024 01:42	
551005	12143	9.307 GB	110.4 TB	03 May 2024 00:14	03 May 2024 01:17	
550997	20206	9.285 GB	183.2 TB	02 May 2024 21:58	02 May 2024 23:39	
550916	29547	9.33 GB	269.2 TB	01 May 2024 19:47	01 May 2024 22:17	
550889	41404	9.307 GB	376.3 TB	01 May 2024 01:00	01 May 2024 04:32	
385 runs	**16745145 files**	**9.266 GB**	**148 PB**			

Fig. 3. Displaying runs selected for reconstruction

5. Conclusion

This article describes the implementation of an automatic data selection workflow for the reconstruction process in the ALICE experiment. The development of this tool has significantly simplified the work of ALICE operators and is in production since May 2024. From that time up to the writing of this paper (September 30, 2024), the tool has successfully managed the selection of 385 runs, representing a data volume of approximately 148 PB, for which the reconstruction process was initiated and executed. This tool has replaced the previously manual process carried out by ALICE operators, thereby allowing them to focus more directly on the core objectives of the experiment. Consequently, this tool marks an important step towards optimizing data processing and management within the ALICE experiment.

REFERENCES

[1] CERN Organization. CERN Accelerating science. Last accessed: 30 September 2024.

[2] CERN Organization. The Large Hadron Collider. `https://home.cern/science/accelerators/large-hadron-collider`. Last accessed: 30 September 2024.

[3] CERN Organization. ALICE detects quark-gluon plasma, a state of matter thought to have formed just after the big bang. Last accessed: 30 September 2024.

[4] R. Divià. Run 2 DAQ systems. `https://indico.cern.ch/event/471309/contributions/1981069/attachments/1256304/1854691/2016.04.12.ALICE.Run2.DAQ.pdf`. Last accessed: 30 September 2024.

[5] Latchezar Betev. ALICE Experiment Status and Run2 Plans. `https://indico.cern.ch/event/366989/contributions/1784720/attachments/729520/1001018/LBNL_run2_programme.pdf`. Last accessed: 30 September 2024.

[6] Acharya, Shreyasi and others. ALICE upgrades during the LHC Long Shutdown 2. *JINST*, 19(05):P05062, 2024.

[7] Rohr, David. Usage of gpus in alice online and offline processing during lhc run 3. *EPJ Web Conf.*, 251:04026, 2021.

[8] Şuiu, Alice Florenţa and Grigoraş, Costin and Ţăpuş, Nicolae and Betev, Latchezar. Automatic Data Workflow and Disk Management Tool for the ALICE Experiment. 10 2023.

[9] David Rohr, Sergey Gorbunov, Marten Ole Schmidt, and Ruben Shahoyan. Track reconstruction in the alice tpc using gpus for lhc run 3, 2018.

[10] Sergiu Weisz, Costin Grigoras, Alice-Florența Șuiu, Latchezar Betev, Mihai Carabaș, and Nicolae Țăpuș. Optimizing large data transfers for the ALICE experiment in Run 3. In *22nd RoEduNet Conference: Networking in Education and Research*, 9 2023.

[11] Richter, Matthias for the ALICE Collaboration. A design study for the upgraded ALICE O^2 computing facility. *Journal of Physics: Conference Series*, 664:082046, 2015.

[12] Şuiu, Alice Florenţa and Grigoraş, Costin and Weisz, Sergiu and Betev, Latchezar. EPN2EOS Data Transfer System. *EPJ Web Conf.*, 295:01023, 2024.

[13] Barthélémy von Haller and Piotr Konopka. The alice data quality control. *EPJ Web of Conferences*, 295, 05 2024.

[14] F. Carminati and Y. Schutz and for the ALICE Collaboration. ALICE Computing Model. `https://alice-collaboration.web.cern.ch/sites/default/files/static/Documents/TDR/Computing/Computing\_Model/alice\_computing\_model.pdf`. Last accessed: 30 September 2024.

[15] HEPiX Organization. Benchmarking Working Group. https://w3.hepix.org/benchmarking.html. Last accessed: 30 September 2024.

[16] CERN Organization. Computing Resource Information Catalog. http://wlcg-cric.cern.ch/. Last accessed: 30 September 2024.

[17] CERN Organization. LHCOPN - Large Hadron Collider Optical Private Network. https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome. Last accessed: 30 September 2024.

[18] Boulais, Martin, Raduta, George C., and Huijberts, Jik. Bookkeeping, a new logbook system for alice. *EPJ Web of Conf.*, 295:01010, 2024.

[19] CERN Organization. MonALISA Repository for ALICE. https://alimonitor.cern.ch/. Last accessed: 30 September 2024.