Journal of Physics: Conference Series **513** (2014) 012037 doi:10.1088/1742-6596/513/1/012037

O^2: A novel combined online and offline computing system for the ALICE Experiment after 2018

Ananya[17], A. Alarcon Do Passo Suaide[22], C. Alves Garcia Prado[22], T. Alt[5], L. Aphecetche[26], N. Agrawal[17], A. Avasthi[17], M. Bach[5], R. Bala[21], G. Barnafoldi[2], A. Bhasin[21], J. Belikov[15], F. Bellini[10], L. Betev[1], T. Breitner[5,9], P. Buncic[1], F. Carena[1], W. Carena[1], S. Chapeland[1], V. Chibante Barroso[1], F. Cliff[1], F. Costa[1], L. Cunqueiro Mendez[1], S. Dash[17], C. Delort[1,7], E. Dénes[2], R. Divià[1], B. Doenigus[8], H. Engel[9], D. Eschweiler[5], U. Fuchs[1], A. Gheata[1], M. Gheata[11], A. Gomez Ramirez[9], S. Gorbunov[5], L. Graczykowski[16], A. Grigoras[1], C. Grigoras[1], A. Grigore[1,3], R. Grosso[14], R. Guernane[28], A. Gupta[21], I. H˘rivna´˘cov´a [19], P. Hristov[1], C. Ionita[1], M. Ivanov[8], M. Janik[16], S. Kalcher[5], N. Kassalias[22], U. Kebschull[9], R. Khandelwal[17], S. Kushpil[24], I. Kisel[5], T. Kiss[2,4], T. Kollegger[5], M. Kowalski[27], M. Kretz[5], I. Kulakov[9], V. Lafage[19], C. Lara[9], I. Legrand[13], V. Lindenstruth[5], A. Maevskaya[25], P. Malzacher[8], A. Morsch[1], B. Nandi[17], M. Niculescu[11], P. Pillot[26], M. Planinic[20], J. Pluta[16], N. Poljak[18], S. Rajput[21], K. Read[23], A. Ribon[1], D. Rohr[5], G. Rubin[2], R. Shahoyan[1], A. Sharma[21], G. Simonetti[1,6], C. So´os[1], M. Szymanski[16], A. Telesca[1], J. Thaeder[8], A. Udupa[17], P. Vande Vyvre[1], F. Vennedey[5,9], B. von Haller[1], S. Wenzel[1], C. Zampolli[12], and M. Zyzak[9] for the ALICE collaboration

[1] European Organization for nuclear Research (CERN), Geneva, Switzerland
[2] Wigner RCP Hungarian Academy of Sciences, Budapest, Hungary
[3] Politehnica University of Bucharest, Bucharest, Romania
[4] Cerntech Ltd., Budapest, Hungary
[5] Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe-Universität, Frankfurt, Germany
[6] Dipartimento Interateneo di Fisica 'M. Merlin', Bari, Italy
[7] Ministère des Affaires Etrangères et Européennes, Paris, France
[8] GSI - Helmholtzzentrum fur Schwerionenforschung GmbH, Darmstadt, Germany
[9] Institut für Informatik, Johann Wolfgang Goethe-Universität Frankfurt, Frankfurt, Germany
[10] Dipartimento di Fisica e Astronomia dellUniversit'a and Sezione INFN, Bologna, Italy
[11] ISS - Institute of Space Science, Bucharest, Romania
[12] INFN and University, Italy
[13] California Institute of Technology, Pasadena, California, United States
[14] University of Houston, Houston, Texas, United States
[15] Institut Pluridisciplinaire Hubert Curien, Strasbourg, France
[16] Warsaw University of Technology, Warsaw, Poland
[17] IIT- Indian Institute of Technology, Mumbai, India
[18] Institute Rudjer Boskovic, Zagreb, Croatia
[19] Institut de Physique Nucléaire (IPNO), Université Paris-Sud, CNRS-IN2P3, Orsay, France
[20] University of Zagreb, Zagreb, Croatia


Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

[21] University of Jammu, Jammu, India
[22] University of São Paulo, Brasil
[23] Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States
[24] Nuclear Physics Institute, Academy of Sciences of the Czech Republic, Řež u Prahy, Czech Republic
[25] Institute for Nuclear Research, Academy of Sciences, Moscow, Russia
[26] Laboratoire de Physique Subatomique et des Technologies Associées - Subatech, Nantes, France
[27] The Henryk Niewodniczanski Institute of Nuclear Physics, Polish Academy of Sciences, Cracow, Poland
[28] Laboratoire de Physique Subatomique et de Cosmologie (LPSC), Université Joseph Fourier, CNRS-IN2P3, Institut Polytechnique de Grenoble, Grenoble, France
E-mail: pierre.vande.vyvre@cern.ch

Abstract. ALICE (A Large Ion Collider Experiment) is a detector dedicated to the studies with heavy ion collisions exploring the physics of strongly interacting nuclear matter and the quark-gluon plasma at the CERN LHC (Large Hadron Collider). After the second long shutdown of the LHC, the ALICE Experiment will be upgraded to make high precision measurements of rare probes at low p_T, which cannot be selected with a trigger, and therefore require a very large sample of events recorded on tape. The online computing system will be completely redesigned to address the major challenge of sampling the full 50 kHz Pb-Pb interaction rate increasing the present limit by a factor of 100. This upgrade will also include the continuous un-triggered read-out of two detectors: ITS (Inner Tracking System) and TPC (Time Projection Chamber)) producing a sustained throughput of 1 TB/s. This unprecedented data rate will be reduced by adopting an entirely new strategy where calibration and reconstruction are performed online, and only the reconstruction results are stored while the raw data are discarded. This system, already demonstrated in production on the TPC data since 2011, will be optimized for the online usage of reconstruction algorithms. This implies much tighter coupling between online and offline computing systems. An R&D program has been set up to meet this huge challenge. The object of this paper is to present this program and its first results.

1. Introduction

The ALICE Experiment [1] has been taking data since the LHC first beam in 2008. During Run1 - till February 2013, the beginning of the first Long Shutdown (LS1) - the online and offline systems [2, 3] have performed particularly well, allowing the experiment to collect and process data with a delivered performance surpassing the design specifications.

After the second Long Shutdown (LS2) currently scheduled for 2018, the LHC will progressively increase its luminosity, eventually reaching for Run3 an interaction rate of about 50 kHz with Pb beams. To be able to make high precision measurements of rare probes at low p_T, ALICE proposed an upgrade [4] which includes a modification of the detectors and the computing systems so that all particle interactions can be examined:

- The present Inner Tracking System (ITS) will be replaced with a new, high-resolution, low-material detector [5].
- The Time Projection Chamber (TPC) will be upgraded with replacement of the chambers by Gas Electron Multipliers (GEMs) and a new pipelined readout electronics based on a continuous read-out scheme[6].
- A new 5-plane silicon telescope will be placed in front of the hadron absorber covering the acceptance of the Muon Spectrometer [7].
- The forward trigger detectors and the electronics of the Transition Radiation Detector (TRD), the Time Of Flight (TOF), and several other detectors will be upgraded [8].

Our approach is to read out all Pb-Pb events at the anticipated interaction rate of 50 kHz, representing an increase by a factor of 100 beyond the present limit. ALICE will then be in a position to accumulate 10 nb^{-1} of PbPb collisions, inspecting about 10^{11} interactions.

The detector electronics and the computing system are designed to achieve nominal performance, even in case of noise or background larger than anticipated with capability to scale up to twice this performance for higher interaction rates. The main physics topics for this upgrade require measurements characterized by very small signal-to-background ratios and large statistics. With large backgrounds, using traditional triggering or filtering techniques is very inefficient for most physics channels. Moreover, the continuous read-out of some detectors and the online calibration and reconstruction will impose a major shift of paradigm for the offline and online computing with a complete re-design of the latter.

2. Requirements

The general strategy is to read out the ITS, TPC, TRD and TOF detectors at the planned interaction rate of 50 kHz and ship the data to the computing system. The Electromagnetic Calorimeter (EMC) and Muon system will be read out from their own triggers at a lower rate. The architecture of the computing system should allow an event rate of 50 kHz, corresponding to the highest instantaneous luminosity. It should also be able, if needed, to scale up by a factor of two to deal with an increased event rate.

It will be necessary to reduce the massive data volume resulting from the combination of the high interaction rate and large event size produced at the expected luninosity. The optimal way to achieve this is to reconstruct the data, at least partially, before recording them in permanent data storage together with online calibration and quality assurance.

Table 1 lists the detector event size sent to the computing system (after zero suppression) and recorded data (after compression by online reconstruction). The table summarizes the estimated overall data throughput to the input of the computing system and the peak and average rates to the mass storage. The peak rate must be sustained through the whole computing system up to the local data storage, whereas the average rate must be sustained all the way to data recording at the CERN Computing Center. For the latter case, data storage at the experimental area provides the necessary buffer.

Detector	Event Size (MByte)		Input to computing system (GByte/s)	Compressed output to data storage (GByte/s)	
	After zero suppression	After data compression		Peak	Average
TPC	20.0	1.0	1000	50.0	8.0
TRD	1.6	0.2	81.5	10.0	1.6
ITS	0.8	0.2	40	10.0	1.6
Others	0.5	0.25	25	12.5	2.0
Total	22.9	1.65	1146.5	82.5	13.2

Table 1. Expected event sizes and data rates

3. The O^2 Project

The present ALICE organization for online and offline computing is comprised of the three groups: the DAQ, HLT and Offline. Run2 will proceed with the same organization as Run1, whereas Run3 will require a far more integrated structure with a single common system. The O^2 project has been launched to design and build the system for Run3 with the new common computing structure.

The work of the O^2 Project was started by establishing a software panel composed of representatives of the DAQ, HLT and Offline groups. As recommended in the panel's report [9], several Computing Working Groups (CWGs) were created which are currently working on different topics for the future system. Table 2 summarizes the topics and goals of the CWGs. The main deliverables of the O^2 CWGs will be used for the Technical Design Report scheduled to be submitted to the LHC Committee in September 2014 which include demonstrators of the key (alternative) technologies for the O^2 system and the software framework.

CWG	Name	Topics	Deliverables
1	Architecture	Global design	Requirements and design documents
2	Tools and procedures	Organization and code development	Procedures, policies and tools
3	Dataflow	Data transport, simulation	Dataflow simulation, design, demonstrator
4	Data model	Data structures	Data model design and demonstrator
5	Computing platforms	Parallel platforms hardware and software	List of platforms and benchmarks, guidelines for using parallel platforms
6	Calibration	Detector calibration	Design and demonstrator of the online detector calibration
7	Reconstruction	Cluster finding, tracking	Design and demonstrator of the online reconstruction
8	Physics simulation	Detector simulation	Design and optimization of the detector simulation
9	Quality Assurance	Data Quality Monitoring, Visualization	Data Quality Monitoring framework, Event Display software
10	Control and configuration	Control, computing farm monitoring	Methods and tools survey, demonstrator
11	Software process	Code quality	Software life-cycle model
12	Hardware	Network, data storage	Design of custom hardware, recommendation of commercial hardware
13	Software framework	Large software framework	Design document, framework prototype

Table 2. The O^2 Computing Working Groups with their main topics and deliverables.

4. The O^2 system

4.1. Dataflow

The global architecture of the O^2 computing system is shown in figure 1. The total read-out capacity will be on the order of 25 Tbit/s, corresponding to 2500 detector links at 10 Gbit/s. This takes into account the construction constraints for detector read-out and provides some headroom for the link protocol.

The detector read-out is performed by Detector Data Links of the 3rd generation (DDL3). The present ALICE data collection is based on a common interface between all the detectors' read-out electronics and the DAQ and HLT systems: the Detector Data Link (DDL1). A common hardware platform [10] and an implementation for the second version of the link (DDL2) has been developed and will be used during the Run2. The bandwidth of the DDL1, the two variations of the DDL2, and a first DDL3 prototype based on Ethernet are shown in figure 2.

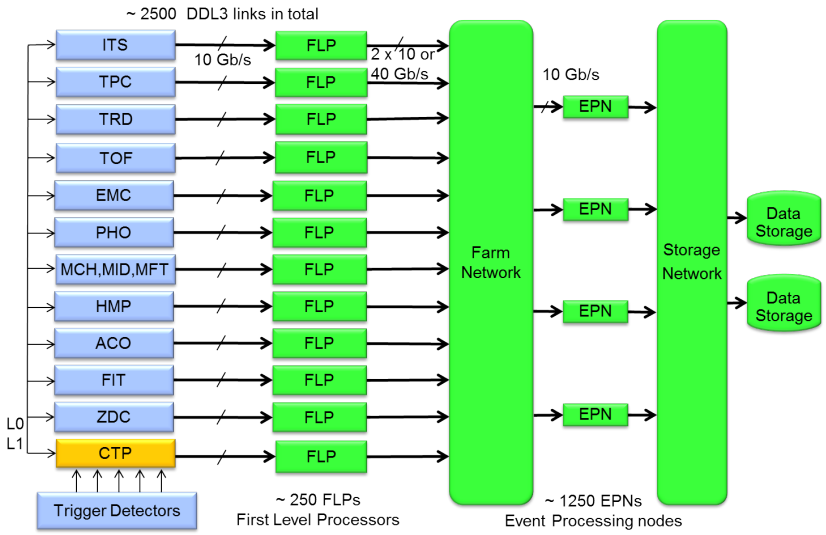


Figure 1. The global hardware architecture of the ALICE O^2 computing system.

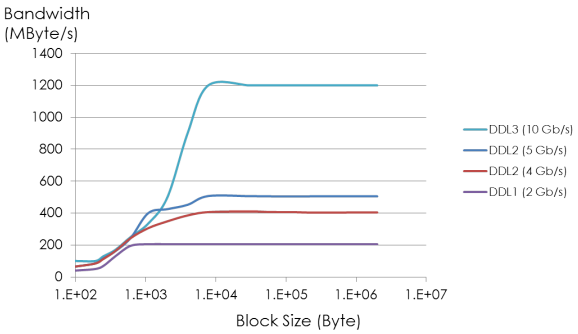


Figure 2. The bandwidth of: DDL1 clocked at 2.125 Gb/s; two variations of DDL2 clocked at 4.25 and 5.3125 Gb/s; prototype of DDL3 clocked at 10.3125 Gb/s.

The DDL3 will transfer raw physics data and detector control data from the read-out electronics to PC adapters which will also be the 3rd generation of Read-Out Receiver Card (RORC3). The RORC3 will ship the data from the detectors to the PC memory. The present HLT system makes use of the RORC FPGAs (Field Programmable Gate Arrays) to compress data on-the-fly by finding clusters. The same concept will be used with the FPGA of the RORC3 if it is a custom design, or of an additional FPGA card if the RORC3 is one of the currently available commercial multi-port network interface cards (already available today for 10 and 40 Gb/s Ethernet or 56 Gb/s Infiniband).

The data acquisition and processing will then be carried out by a large processor farm, based on CPUs and GPUs (or similar multicore) devices. The compressed detector data are transferred by the RORC3 to the memory of a First Level Processor (FLP), which will then carry out localized calibration and reconstruction steps. The full event will then be assembled and reconstructed in one Event-building and Processing Node (EPN) which will perform the final data compression and data recording. The local data storage will be centralized or distributed amongst all EPNs. The FLPs and EPNs are connected by the farm network. The EPNs access the local data storage through the storage network if it is centralized. These two networks are functionally separate but could be implemented by the same physical devices.

4.2. Triggered and continuous read-out

During the Run3, the ALICE detectors will run either in triggered or continuous mode. This will imply a different approach for the read-out based on time frames. Each time frame contains multiple events collected within common time intervals defined by heartbeat triggers.

Each detector read-out unit used for continuous read-out will autonomously tag the data using the local copy of the LHC Orbit and the BCID (Bunch Crossing Identification). The data will be sent as a continuous flow of successive time frames each preceded with a header containing the time based tagging. A trailer will indicate the error cases such as data truncation due to the early arrival of a physics or heartbeat trigger. The triggered read-out will function in the same way as now by sending a data block preceded by a header for every trigger, physics or heartbeat (see figure 3).

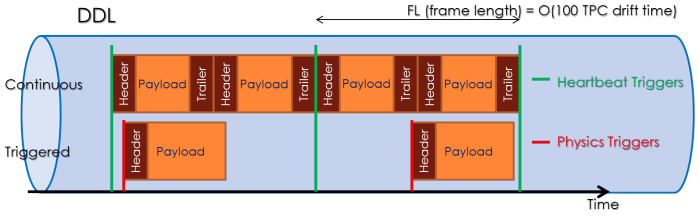


Figure 3. Using physics and heartbeat triggers for the continuous and triggered read-out.

4.3. Data processing

The data processing will be performed by the O^2 system in an iterative manner according to requirements and the underlying hardware as illustrated in figure 4. The software framework should be flexible enough to allow modification of the overall distribution and scheduling based on experience with the data. Three options are available for the online calibration and reconstruction: in the FLPs, or in the EPNs in a synchronous (on-the-fly) or asynchronous manner (from data temporarily stored to the data storage).

The online data processing will have two main goals: to reduce the data volume and to limit overall computational needs by performing part of the reconstruction before recording the data. This is not entirely new for ALICE. It has been successfully used since the Pb-Pb run of 2011. The original raw data of the TPC are discarded after the online cluster reconstruction, the transformation and quantization of cluster parameters, and a lossless data deflation and compression process. An average compression factor of 4.4 has been achieved with Pb-Pb data. Further reduction is possible by optimizing cluster parameters and removing irrelevant clusters by an appropriate online cluster analysis.

4.4. Impact of commercial hardware and software solutions

The High Energy Physics (HEP) community has traditionally been an insatiable consumer of electronics and computing technologies. This will, of course, continue with the continuous development of these technologies. The O^2 project anticipates a massive use of heterogenous commodity devices for the transfer, processing, and storage of data. This influence might as well extend to the software environment used for the upgrade. Large distributed data processing structures, the data grids, have been put in place to satisfy the processing needs of the present generation of HEP experiments. The processing of such huge amounts of data has only been possible by selecting events with a fast custom electronic trigger system. This has all been changed by the relentless evolution of technology. The ALICE upgrade will reduce the role of the usual trigger scheme and entrust the processing of complete data sets to standardized computing algorithms. HEP is evolving and might use some Big Data concepts.

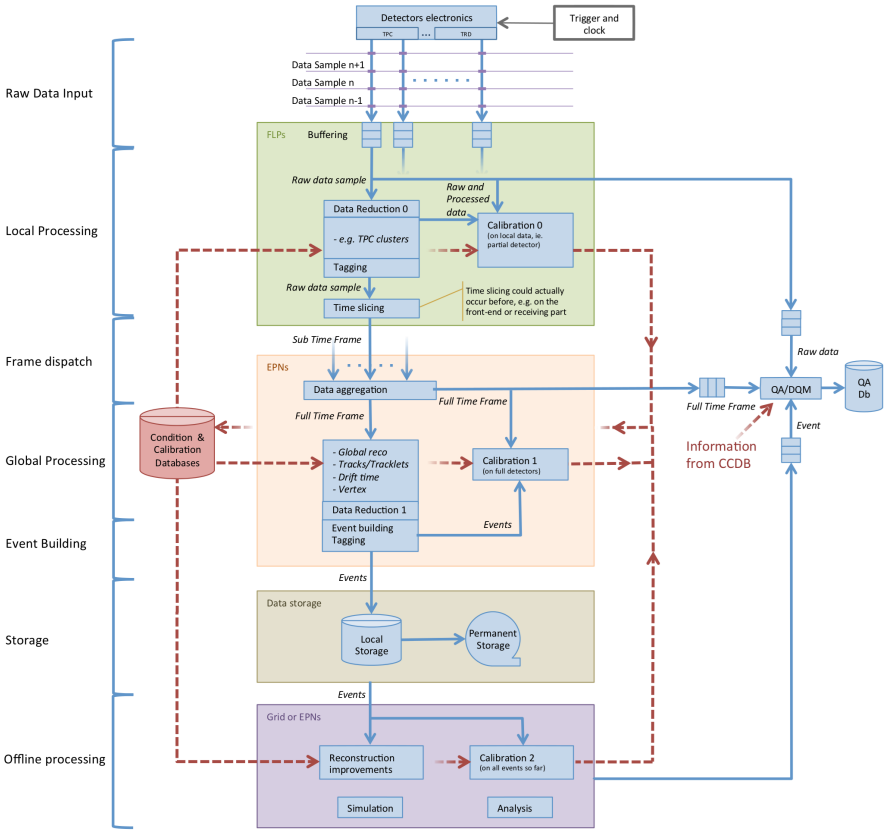


Figure 4. Outline of the data flow and processing in the ALICE O^2 computing system.

5. Conclusion

After LS2, the ALICE apparatus will be upgraded in order to make high precision measurements of rare probes at low p_T. The full 50 kHz Pb-Pb collision events will be entirely analyzed. The ALICE Collaboration has therefore initiated a large research and design effort: the O^2 project. The goal of this project is to design and build a completely new computing system capable of acquiring data for online processing at a rate increased by a factor of a 100 beyond the present limit.

References

[1] ALICE Collaboration, The ALICE experiment at the CERN LHC, 2008 JINST 3 S08002, 2008.
[2] ALICE Collaboration, ALICE Technical Design Report of the trigger, data-Acquisition, high level trigger, and control system, CERNLHCC- 2003-062, 2004.
[3] ALICE Collaboration, The ALICE Offline Bible, `http://aliweb.cern.ch/secure/Offline/sites/aliweb.cern.ch.Offline/files/uploads/-OfflineBible.pdf`.
[4] ALICE Collaboration, Letter of Intent of the ALICE Upgrade, CERN-LHCC-2012-012, 2012.
[5] ALICE Collaboration. Technical Design Report for the Upgrade of the Inner Tracking System, ALICE-UG-TDR1, Dec. 2013.
[6] ALICE Collaboration. Technical Design Report for the Upgrade of the Time Projection Chamber, ALICE-UG-TDR2, Dec. 2013.
[7] ALICE Collaboration. A Muon Forward Tracker for the ALICE Experiment - Letter of Intent, Dec. 2013.
[8] ALICE Collaboration, Technical Design Report for the High Rate Electronics upgrade of the ALICE detector, ALICE-UG-TDR3, Dec. 2012.
[9] L. Betev et al., ALICE Computing software framework for LS2 Upgrade, 2012.
[10] H. Engel, Common Read-Out Receiver Card for the ALICE Run2 Upgrade, Topical Worshop on Electronics for Particle Physics 2013, Sep. 2013, Perugia, Italy, 2013.