The detector read-out in ALICE during Run 3 and 4

F Costa, A Kluge and P Vande Vyvre for the ALICE collaboration.
CERN, Switzerland
E-mail: Filippo.Costa@cern.ch

Abstract. ALICE, the general purpose, heavy ion collision detector at the CERN LHC is designed to study the physics of strongly interacting matter using proton-proton, nucleus-nucleus and proton-nucleus collisions at high energies. The ALICE experiment will be upgraded during the Long Shutdown 2 in order to exploit the full scientific potential of the future LHC. The requirements will then be significantly different from the original design of the experiment and will require major changes to the detector read-out. The main physics topics addressed by the ALICE upgrade are characterised by rare processes with a very small signal-to-background ratio, requiring very large statistics of fully reconstructed events. In order to keep up with the 50 kHz interaction rate, the upgraded detectors will be read out continuously. However, triggered read-out will be used by some detectors and for commissioning and some calibration runs. The total data volume collected from the detectors will increase significantly reaching a sustained data throughput of up to 3 TB/s with the zero-suppression of the TPC data performed after the data transfer to the detector read-out system. A flexible mechanism of bandwidth throttling will allow the system to gracefully degrade the effective rate of recorded interactions in case of saturation of the computing system. This paper includes a summary of these updated requirements and presents a refined design of the detector read-out and of the interface with the detectors and the online systems. It also elaborates on the system behaviour in continuous and triggered readout and defines ways to throttle the data read-out in both cases.

1. Introduction
The ALICE experiment [1] will be upgraded during the Long Shutdown 2 (LS2) in 2019-20 in order to exploit the full scientific potential of the future LHC. The requirements will then be significantly different from the original design of the experiment and will require major changes to the detector read-out as presented in the Readout and Trigger TDR [2] and the $O^2$ TDR [3].

The requirements for the Run 3 and 4 have evolved since the submission of these two TDRs. The main change concerns the read-out of the Time Projection Chamber (TPC) detector. All the data, unmodified and uncompressed, will be transferred by the Front-End Cards (FECs) to the read-out system in any circumstance. Neither a selection of data nor a detector data throttling via a busy signal are possible with the TPC FEC cards. The total data volume collected will increase significantly reaching a sustained data throughput of up to 3 TB/s. The reduction of the TPC data volume including the zero-suppression will be performed after the data transfer to the detector read-out system.

A flexible mechanism of bandwidth throttling has therefore been introduced to gracefully degrade the effective rate of recorded interactions in case of saturation of the computing system. This mechanism measures the saturation of the read-out system and discards a fraction of the data according to a predefined policy. It emulates the busy of conventional trigger systems by discarding data and therefore creating dead-time.
2. Read-out and data merging in continuous and triggered modes

After the upgrade during the LS2, the ALICE detectors will use two modes of operation: triggered and continuous read-out. The continuous read-out of the majority of the detectors is a substantial change from current practice. The data are not delimited by physics triggers but are composed of several constant data streams that will be transferred from the detectors to the computing system. As shown in figure 1, the data produced by the Front-End Cards (FEC) are transferred to the Common Read-out Units (CRU) which are the interfaces to a first farm of computers: the First-Level Processors (FLP) where an initial data volume reduction is performed. The data merging and the final data volume reduction is performed by a second farm of computers: the Event Processing Nodes (EPN).

Dedicated time markers, the heartbeat triggers (HB), will be used to chop the data flow to the FLPs into manageable pieces called HB Frames (HBF). The HBFs are assembled in Sub-Time Frames (STF) by the First-Level Processors (FLPs) and then in Time Frames (TF) by the Event Processing Nodes (EPN).

The data from triggered read-out are assembled into trigger data fragments and transferred to the FLPs together with the HBF ID the trigger belongs to.

At the time of each HB the Central Trigger Processor (CTP) transmits a HB trigger which contains an additional flag when a TF starts.

3. System architecture and read-out protocol

The overall data-flow from the detector into the O2 system is shown in figure 2.

The trigger system receives input from several detectors: the ALICE Cosmic Ray Detector (ACO), the Electro-Magnetic Calorimeter (EMC), the Fast Interaction Trigger detector (FIT), the PHOTon Spectrometer (PHOS), the Time-Of-Flight detector (TOF) and the Zero Degree Calorimeter (ZDC).

Data produced by the detectors FECs are transferred to the read-out cards (CRU or C-RORC) in a continuous or triggered read-out mode over the GBT [4] or DDL [5] based read-out links. The triggered data will be tagged with the LHC clock information as it is the case during
the LHC Run 1 and 2. The continuous streams of data samples are split into HBF tagged with the corresponding HB ID. The HBs are distributed by the trigger system synchronously with the data transfer over the Timing and Trigger Links (TTS) and read-out links. The data are compressed and multiplexed in the CRUs and transferred to the memory of a farm of servers, the FLPs. Several streams of HBFs may be aggregated on each FLP and buffered in memory.

For each HBF or physics trigger, the CRUs will send a HB acknowledge message to the CTP containing information whether the HBF data or trigger data have been sent successfully to the FLP and containing the occupancy status of the CRU data buffer.

The HBFs are accumulated into sub time frames (STF) during a time period of the order of 22 ms. All FLPs produce STFs, which could be empty for those FLPs receiving data from triggered detectors inactive during the corresponding time period. The nominal time between two HBs is 89.4 $\mu$s (one LHC orbit period) and a TF consists of 256 HBF or approximately 22 ms.

The STFs are then dispatched to the Event Processing Nodes (EPNs) for aggregation and processing. The STFs related to the same time period and from all FLPs are received by the

---

**Figure 2.** ALICE detector read-out system block diagram. Three configurations exist: I. Detectors using the CRU which receive the TTS information via the CRU only. II. Detectors using the CRU which receive the TTS information via the CRU and via the on-detector electronics. III. Detectors which use the C-RORC and receive the TTS information on the on-detector electronics via the TTC protocol.
same EPN and aggregated into a complete time frame (TF). The STF and TF duration of 22 ms is chosen to minimise incomplete data at the TF boundaries for the collisions producing tracks spanning across the TF boundaries.

The data volume will be reduced by processing the data on the fly in the EPNs synchronously with data taking and not by rejecting complete events. The \( O^2 \) system will perform a partial calibration and reconstruction and replace the original raw data with the compressed data. Data produced during this stage will be stored temporarily in the \( O^2 \) system. A second reconstruction stage will be performed asynchronously using the final calibration in order to reach the required data quality.

### 3.1. Trigger

The architecture of the upgraded ALICE read-out and trigger system as described in [6] is reproduced in figure 2. The CTP system will assemble and evaluate the HB acknowledge messages sent from the CRUs to the CTP. These acknowledge messages include the HB ID so that the CTP can assemble a complete HB map. This HB map represents the HBF data transmission and CRU buffer occupancy status of all the CRUs for each HBF. This HB map will be part of the CTP read-out to the \( O^2 \) system. The implementation does not require additional hardware, as the acknowledge message is sent via the bi-directional high bandwidth TTS link from the CRU to the CTP. In case the HB map evaluation gives too a high number of incomplete HBF, the CTP has the possibility to discard data as described in Section 4.

### 3.2. Detector read-out

Upgraded detectors will use the CRU as the interface between the front-end electronics, the \( O^2 \) facility, the Detector Control System (DCS via the \( O^2 \) facility), as well as the CTP via the Local Trigger Unit (LTU) and the TTS. Figure 2 shows the general ALICE detector read-out scheme with its three variations.

Most detectors use GBT links as read-out links from the on-detector electronics to the CRU. These links can be operated in wide bus mode, with a payload data bandwidth of 4.48 Gb/s or in forward error correction mode, with a payload data bandwidth of 3.2 Gb/s. Only the TPC foresees to use the wide bus mode. The CRU hardware can accommodate up to 48 read-out link connections. The number of input link connections actually used depends on the application. The TPC has the highest data throughput. Its read-out system will use up to 20 read-out links per CRU which corresponds to a maximum aggregate throughput of 20 · 4.48 = 89.6 Gb/s.

The TTS links from the CRUs to the CTP/LTU carry the HB acknowledge message and indicates whether a given HBF has been transmitted correctly and whether the CRU buffer is full.

The most demanding CRU implementation will be for the TPC due to the high data rate and signal processing: base-line correction (BC), zero-suppression (ZS) and cluster finding (CF). Depending on the operation mode, the TPC user logic will fill the PCIe read-out buffer with data from different processing stages (compressed data after the cluster finder or only zero suppressed or raw ADC data samples). For commissioning the TPC user logic can also fill the read-out buffer with two data types (for instance zero suppressed and after the cluster finder) for the same HBF. In this case the acquisition duration and rate will be degraded.

### 4. Throttling modes

The ALICE continuous read-out system is designed to operate in nominal conditions without data loss. In this context data loss is considered a rare exception and thus missing HBFs in the reconstruction are either considered negligible and are ignored or the full TF with missing data is discarded at the time of reconstruction. Under the above condition no coordinated approach between read-out modules to discard data fragments would be required.
However, the assumption of low data loss will not be true in all realistic scenarios. During the commissioning of the upgraded experiment, the $O^2$ system might not be fully available, the detectors might not have yet their data compression schemes optimised or the beam background might have been underestimated. The detectors also need to take samples without data compression during commissioning or for calibration runs. In those cases, the continuous read out detectors will continue to work with individual CRUs which will discard any surplus of data and inform the $O^2$ system.

In order to allow for efficient and coordinated throttling schemes even in continuous read-out, each HB trigger contains a single bit, the HBaccept (HBa)/HBreject (HBr) bit, stating whether the corresponding HBF should be transmitted to the FLP or deleted from the common read-out unit (CRU) buffer.

The succession of triggered runs and continuous operated runs will allow continuous and triggered periods of data taking following each other. This will be used for example for the TPC calibration runs or Laser runs during physics data taking. The CRU and the $O^2$ system are dimensioned to be able to handle the expected data flow during normal physics data taking with a fully deployed system and all the processing steps enabled (BC, ZS, CF) in the CRU. However, there will be circumstances when operation outside nominal conditions is required. For example the commissioning of the system in continuous mode with some of the processing steps disabled and an $O^2$ system partly deployed.

It is therefore important to have a way to throttle the system and acquire a fraction of the data corresponding as closely as possible to the capacity of the system. The role of throttling will be to keep the system working smoothly although with a lower data throughput and at lower performance. The three envisaged modes of throttling are:

- The autonomous mode relies on autonomous decisions of each CRU to delete data when it cannot buffer them. This mode will be used when the capacity of the system will be close to the needs and the probability of deleting data will therefore be low.
- In scaling mode, the CTP modulates the proportion of accepted HB triggers (HBa) and rejected HB triggers (HBr). For each HBa trigger the CRU transmits the corresponding HBF to the FLP. For each HBr trigger the CRU deletes the corresponding HBF from the CRU buffer and only transmits a HBF header/trailer pair. The proportion of accepted/rejected HB triggers sent by the CTP can be set by a predefined sequence initiated by the operator or a sequence autonomously applied and created by automatic evaluation of the HB map. Depending on the proportion of accepted/rejected HB triggers this mode can be tailored to operation conditions with low and high data loss due to buffer overflow.
- In collective mode, once at least one single CRU or a programmable minimum number of CRUs is reporting one or a programmable number of sequential HBFs missing, the CTP sets all subsequent HB triggers of the current TF to HBr, thus initiating the deletion of all subsequent HBFs belonging to this TF. This mode is foreseen for debugging and commissioning. It assumes that data loss due to buffer overflow is considered high.

Figure 3 summarizes the flow of messages needed to implement the throttling described above. The sequence of operations corresponds to the following steps indicated in figure 3:

1. A trigger message is sent to the CRUs for each HB with the HBa/HBr flag indicating whether the data has to be read out or discarded.
2. Each CRU transfers the data to the FLP (2a) and sends back an acknowledge message to the CTP (2b) indicating whether the HBF has been properly transmitted to the FLP.
3. All the acknowledge messages are used by the CTP to assemble the HB map which is used to send HBF decisions to the CRUs.
4 The HBF decisions are used by the CRUs to decide the deletion of HBFs in their own buffer.
5 The HBF decision is also transmitted to the FLP in order to be able to delete HBFs that have already been transmitted to the FLP.
6 The HB map is transferred to the CTP FLP so that all decisions can be checked in the EPN processing the corresponding TF.

5. Conclusion
The upgrade of the ALICE experiment during the LS2 will induce a major increase of the data volume transferred from the detector electronics to the read-out system. This increase can result in a saturation of the read-out system. The throttling system will allow the data traffic to be regulated adapting it to the available capacity of the read-out system. This throttling system does not require any additional hardware and is based on the existing trigger and read-out hardware. It emulates the dead time produced by the busy signal in conventional trigger systems.

References
[1] ALICE Collaboration. The ALICE Experiment at the CERN LHC, in: JINST 3.08 (2008), S08002; [doi 10.1088/1748-0221/3/08/S08002].
[2] ALICE Collaboration. Upgrade of the Readout and Trigger System, Technical Design Report. CERN-LHCC-2013-019 / ALICE-TDR-015.
[3] ALICE Collaboration. Upgrade of the Online-Offline computing system, Technical Design Report. CERN-LHCC-2015-006 / ALICE-TDR-019.

[4] Moreira P et al. The GBT: a Proposed Architecture for Multi-Gb/s Data Transmission in High Energy Physics, in proc. of TWEPP 2007, Prague, Czech Republic, September 37 2007, pp. 332-336; [doi:10.5170/CERN-2007-007.332]

[5] Costa F et al. for the ALICE collaboration. DDL, the ALICE data transmission protocol and its evolution from 2 to 6 Gb/s. in: *JINST* 10 (2015), C04008; [doi: 10.1088/1748-0221/10/04/C04008].

[6] Krivda M and Pospisil J for the ALICE Collaboration. The ALICE Central Trigger Processor (CTP) Upgrade, in : *JINST* 11 (2016), C03051; [doi:10.1088/1748-0221/11/03/C03051].