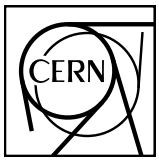
EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH



DRAFT ALICE-TECH-2016-001

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

A. Kluge, P. Vande Vyvre

CERN, EP Department

Version 1.7.2, 2 March 2017

Abstract

This note presents the detector read-out of the ALICE experiment after the upgrade scheduled for the Long Shutdown 2. It reviews the requirements and presents a detailed concept for the detector read-out electronics design, the trigger and computing systems.

Version	Date	Change	Author
1.5	3 June 2016	first release	AK
1.5.1	14 June 2016	updated ITS/MFT CRU&link number	AK
		updated MID PbPb data rate	
		added J. Schambach to technical responsible	
		added DCS/A. Augustinus,	
		and P. Chocula to technical responsible/PL	
1.5.2	15 July 2016	added graphical table for data formats	AK
		added more trigger types to table 7	
1.6	23 August 2016	added CPV project leader	AK
		added trigger functionality description to 2.2	AK
		added clarification on CRU status bits	AK
		Section 4.2 updated	AK
1.7	13 Jan 2017	modified signal message flow diagram and	
		text to reflect that nominal max. delay of HBF ack is 8 orbits.	AK
1.7.1	2 Feb 2017	suppressed reference to SOC/EOC/SOC/SOT	
		as special triggers, page 11, section 3.1, page 22, section 4.1	AK
1.7.2	2 March 2017	changed LHC orbit period to 88.9 us	AK

hb

Table 1: Change log.

Sub-system	released to technical responsible	released to project leader
ACO	A. V. Tello	A. F. Tellez
CPV	Y. Kharlov	S.Sadovsky
CTP	M. Krivda, R. Lietava	D. Evans
CRU	E. David, J. Imrek	T. Kiss, T. Nayak
DCS	P. Chochula	A. Augustinus
EMC	M. Poghosyan	T. Cormier
FIT	T. Karavicheva, D. Serebryakov	W. Trzaska
HMP	G. De Cataldo	G. De Cataldo
ITS	G. Aglieri, P. Giubilato, J. Schambach	L. Musa
MCH	H. Borel, C. Flouzat	A. Baldisseri
MFT	S. Panebianco, C. Guerin	G. Martinez
MID	P. Dupieux, C. Renard	A. Baldisseri
O^2	F. Costa	P. Vande Vyvre
PHS	Y. Kharlov	
TOF	D. Falchieri	P. Antonioli
TPC	T. Alt, C. Lippmann	H. Appelshäuser
TRD	J. Mercado-Perez	J. Stachel
ZDC	P. Cortese	N. Di Marco
TC		W. Riegler, A. Tauro

Table 2: Distribution list.

Contents

1 Introduction

The requirements for the Run 3 and 4 have evolved since the submission of the Readout and Trigger TDR [1] and of the O^2 TDR [2]. The design of the detector read-out has been refined and prototypes have been developed. This note 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. Finally it elaborates on the system behavior in continuous and triggered read-out and defines ways to throttle the data read-out.

2 System architecture

2.1 O^2 system

The overall data-flow from the detector into the O^2 system is shown in Fig 1.

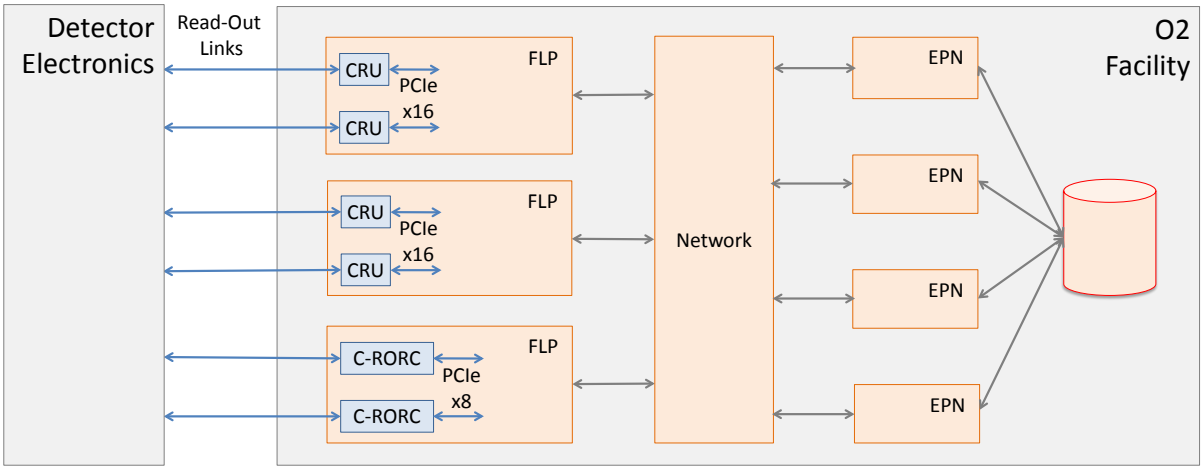


Figure 1: ALICE data-flow from the detector electronics up to the O^2 system.

Data produced by the detectors are transferred to the common read-out units (CRU) in a continuous or triggered read-out mode over the GBT or DDL based read-out links. The triggered data will be tagged with the LHC clock information as it is the case for Run 1 and 2. The continuous streams of data samples are split into heartbeat frames (HBF) using periodically occurring heartbeat triggers (HB) distributed by the trigger system synchronously with the data transfer over the timing and trigger links (TTS) and read-out links. The HBFs are tagged with the corresponding HB ID. The data are compressed and multiplexed in the CRUs and transferred to the memory of the FLPs. Several streams may be aggregated on each FLP and buffered in memory. Via the TTS system 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 has 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 STFs are then dispatched to the Event Processing Nodes (EPNs) for aggregation. The STFs related to the same time period and from all FLPs are received by the 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.

The O^2 facility, located at the experimental area at Point2, will include all the FLPs, EPNs, networking and data storage. It will also provide the interfaces with the Grid and the permanent data store at the Tier 0.

2.2 Trigger

The architecture of the upgraded ALICE read-out and trigger system as described in [4] is reproduced in Figure 2. For the triggered ALICE operation the upgraded CTP system will produce trigger signals (LM, L0, L1) with several latencies derived from the input trigger signals provided by the trigger detectors. For both the continuous and triggered operation the CTP will produce the periodic heart beat (HB) trigger and specific software triggers. All trigger types provided by the CTP are sent via the LTUs and the TTS to the detector read-out systems.

The CTP system will assemble and evaluate the HB acknowledge messages sent from the CRUs to the CTP which will 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 timing and trigger distribution (TTS) link from the CRU to the CTP. The detectors with the largest number of CRUs and HB map entries are: TPC (360), MCH (24), ITS (24), MFT (10), TRD (29), TOF (3), MID (2). All other detectors using the CRU as read-out use only one CRU. It needs to be defined whether for some detectors the granularity of the HB acknowledge message needs to be increased to more than one bit per CRU. In case the HB map evaluation gives a too high number of incomplete HBF the CTP has the possibility to act automatically or manually as described in Section 3.

2.3 Common Read-Out Unit

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 Trigger, Timing, and clock distribution System (TTS). Figure 2 shows the general ALICE detector read-out scheme with its three variations.

For all detectors, but the TRD, the read-out links to the CRU from the on-detector electronics are GBT links. 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 maximum number of read-out link connections per CRU given by the hardware is 48. The number of actually used input link connections depends on the application. The TPC system will for example use a maximum number of read-out links per CRU of 20 which corresponds to a maximum aggregate throughput of 20 * 4.48 Gb/s = 89.6 Gb/s. The TRD detector will not use the GBT protocol but a 2 Gb/s 8b10b protocol and plans to use up to 36 input links. Table 3 shows the number of read-out links connected to one CRU and the corresponding input data bandwidth for each detector.

The performance of the PCIe bus interface (PCIe Gen 3 16 lanes) towards the FLPs in the O^2 system has been characterised and has a practical bandwidth of up to 90 Gb/s. Depending on the detector application the instantaneous CRU input bandwidth might be higher than the sustainable PCIe bandwidth. Depending on the selected data processing mode in the CRU, but also on operation conditions the CRU firmware needs to buffer input data bandwidth fluctuations accordingly to adapt to the PCIe bandwidth. In case of a buffer overflow, the buffer control integrity needs to be maintained and the surplus of data needs to be deleted.

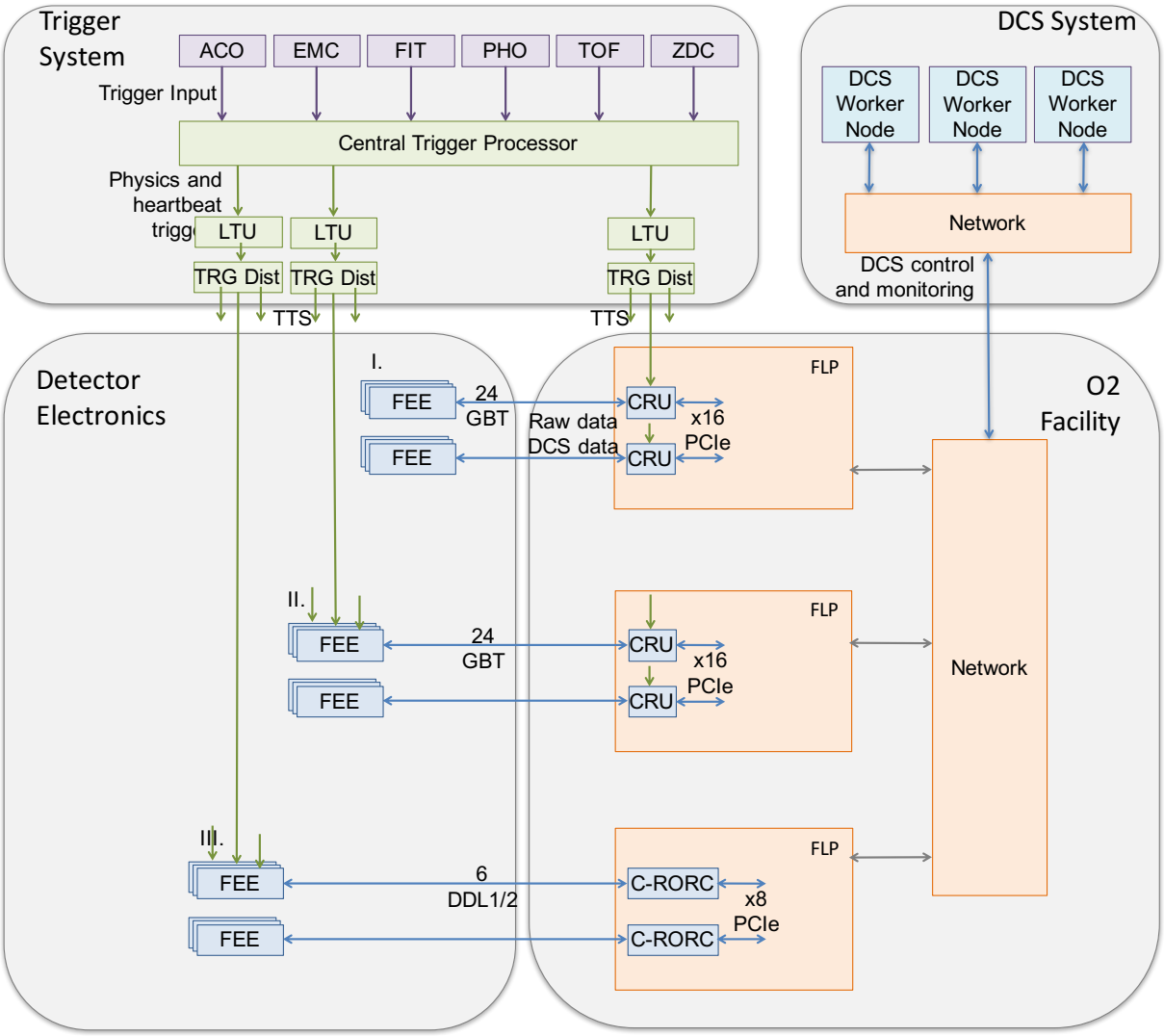


Figure 2: ALICE detector read-out system block diagram. Three configurations exist: I. Detectors using the CRU which receive the TTS information via the CRU only. II. Detectors using the CRU which receive the TTS information via the CRU and via the on-detector electronics (ITS/MFT via GBT, TRD via TTC). III. Detectors which use the C-RORC and receive the TTS information on the on-detector electronics via the TTC protocol.

The interface of the CRU to the CTP/LTU is based on bi-directional optical timing and trigger system (TTS) links. The links from the CTP to the CRUs carry the LHC clock with a jitter performance below 20 ps rms and transmit trigger information.

The 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, zero-suppression and cluster finding). Figure 3 shows the block diagram of the CRU-TPC with its interface to the detector, CTP and FLP. 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. However, the acquisition duration must span at least one TPC drift time.

Table 3: Number of read-out links using the CRU, CRUs, read-out links per CRU, maximum possible data rate per read-out link, maximum possible input data rate per CRU.

Detector	# read-out links	# CRU	# CRU read-out links	max. link rate [Gb/s]	max. CRU rate [Gb/s]
CTP	19	1	19	3.2	60.8
FIT	26	1	26	3.2	83.2
ITS	576	24	24	3.2	76.8
MCH	575	24	24	3.2	76.8
MFT	160	10	24	3.2	76.8
MID	32	2	24	3.2	76.8
TOF	72	3	24	3.2	76.8
TPC	6552	360	20	4.48	89.6
TRD	1044	29	36	2	72
ZDC	1	1	1	3.2	3.2
Total	9057	455			

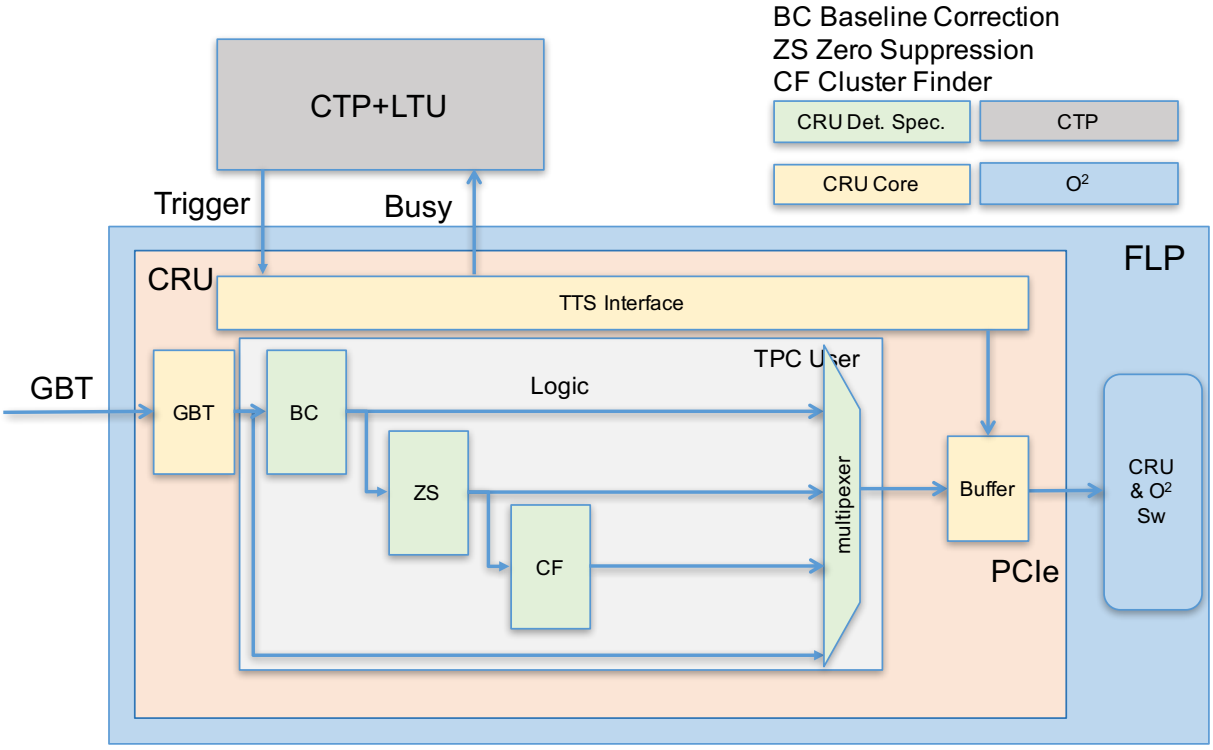


Figure 3: Block diagram of the CRU-TPC implementation.

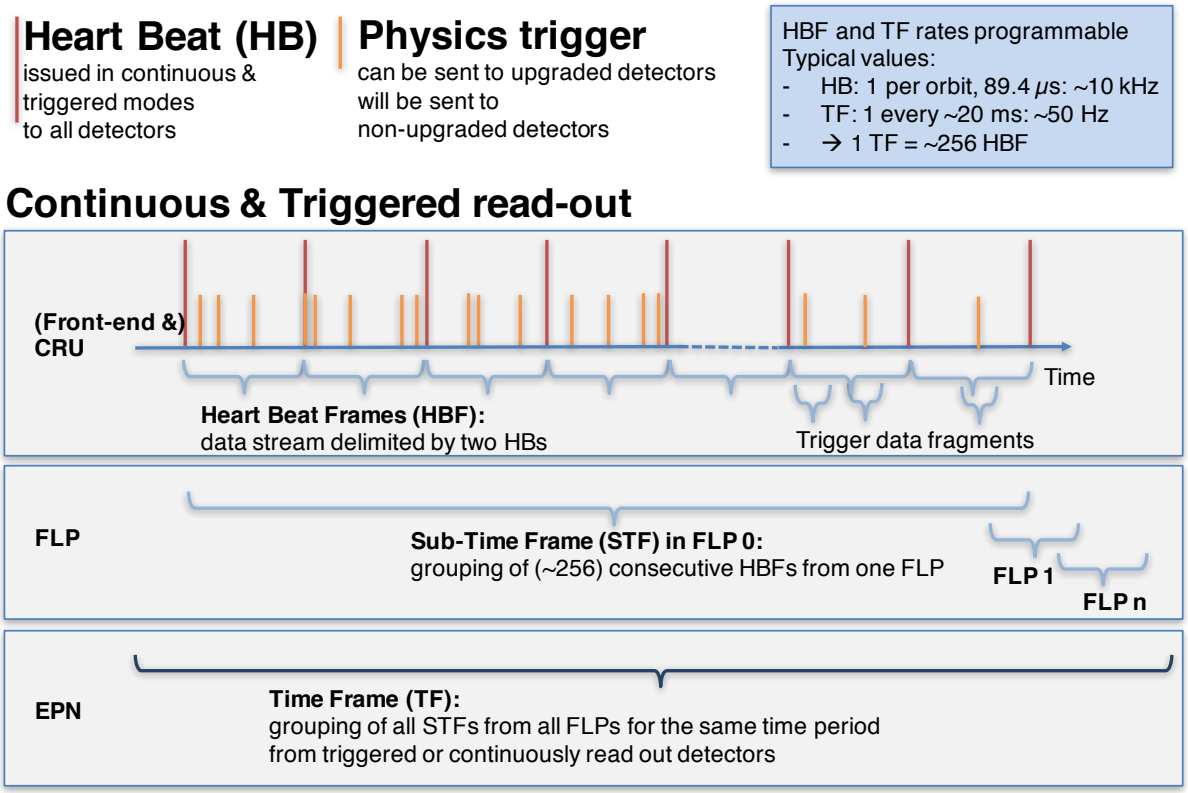


Figure 4: Heartbeat triggers, physics triggers and the corresponding data blocks.

2.4 Functional requirements

After the upgrade of the Long Shutdown 2 (LS2), the ALICE detectors will use two modes of operation: triggered and continuous read-out.

2.4.1 Continuous and triggered 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 data streams, the heart beat frames (HBF) delimited by periodic HB triggers (HB) that will be transferred from the detectors to the computing system. In triggered operation, the data transmitted in a HBF are filtered by the physics trigger. As shown in Fig. 4, dedicated time markers, the HBs, will be used to chop the data flow to the First Level Processor (FLP) into manageable pieces the HBFs. The HBFs are assembled in Sub-Time Frames (STF) by the FLPs and then in Time Frames (TF) by the event processing nodes (EPN). The LHC clock will be used as a reference to synchronise, aggregate and buffer the data.

Nominally, the time between two HBs is 88.9 μs (one LHC orbit period) and a TF consists of 256 HBF or approximately 22 ms. However, HBF lengths and number of HBF in a TF are programmable parameters. 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 CTP transmits a HB trigger to the CRUs. The HB trigger contains an additional flag when a TF starts.

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 start-up of the detector, 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 throttling schemes (see section 3.3) 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.

Each CRU sends the HB acknowledge message to the Central Trigger Processor (CTP) stating whether data has been discarded, where geographically in the detector and which HBF is affected. In triggered mode, in addition to the HB acknowledge message for each individual trigger a trigger acknowledge message is sent. As the trigger acknowledge message has the same format as the HB acknowledge message it is seen as a sub-form of the HB acknowledge message. Thus, in the following description throughout the document it is only referred to HB acknowledge message. In continuous and in triggered mode the CTP takes the HB acknowledge message to assemble a HB map covering the whole detector and forwards it to the O^2 system via the CTP read-out. The CTP applies pre-defined algorithms on the HB map and if needed takes action to reduce the data throughput.

In order to throttle the system and to allow for dedicated physics trigger conditions, all detectors must be compatible with triggered read-out. The CTP will provide one trigger signal with several latencies (LM, L0, L1). Upon reception of a trigger signal, each detector needs to respond and send the data corresponding to the triggered time period. For all detectors, with the exception of the TPC and ITS/MFT, the trigger refers to the read-out data of a given bunch crossing. The TPC for example will provide the data of the 102.4 μs following this bunch crossing. Special trigger sequences, such as calibration triggers or wake-up signals can be requested and need to be defined. Exceptions exist for detectors not being upgraded to receive more than one trigger signal.

The CTP in Run3 has no dead time anymore, which means that it can assert a subsequent trigger in the following bunch crossing (after 25 ns). The control loop (detector gets full, sends HB acknowledge message to CTP, CTP sends trigger to detector) is in any case longer than 25 ns and might be longer than 10 μs. As a result, a new trigger might have been released from the CTP to the detector before the corresponding message indicating that the detector cannot read out any more events arrives at the CTP. Therefore, the detectors need to be able to process a trigger signal, even though they are full and the CRUs need to send an acknowledge message to the CTP and a data header/trailer pair without payload acknowledging the trigger reception to the FLPs.

Not upgraded detectors with a minimum read-out time longer than the control loop cycle can be protected by the CTP from triggers once they are full.

2.4.2 Detector read-out operation modes

The TPC detector read-out modes are described in this section as an example. Similar modes may exist for the other detectors and need to be defined.

The functional requirements of the TPC have been described in [3] and include the need for different data types:

- Unmodified raw Data (UR): raw SAMPA ADC values from the TPC FECs;

Table 4: TPC requirements for triggering and read-out options depending on the operating mode (physics, calibration, debugging and commissioning) and the run type (Physics, Pedestal, Pulser, Laser, Krypton and Technical). The periodicity can be on demand or on a regular basis such as once per fill or once per year. The data types collected in each case are indicated by symbols with the following meaning: "•" corresponds to the data type selected, "o" corresponds to a choice between data types and "&" corresponds to a combination of data types read-out concurrently for the same time period.

Mode	Run type	Period.	Trigger		Acqu. wind.	Data types			
			Mode	Rate (Hz)		UR	BC	ZS	CF
Physics	Phy.		Cont.	N.A.	Cont.				•
Calib.	Ped.	1/Fill	Trig.	10	TDT	•			
	Pul.	1/Fill	Trig.	10	TDT	o	o	o	o
	Las.	Demand	Trig.	10	TDT	o	o	o	o
	Las.	Physics	Trig.	10	TDT				•
	Kry.	1/year	Cont.	N.A.	Cont.			o	o
Debug.	Tech.	Demand	Trig.	10	TDT	o	o	o	o
	Tech.	Demand	Trig.	10	TDT			&	&

– Base-line Corrected raw data (BC): base-line corrected raw data with the pedestal offset in each individual read-out channel subtracted;

– Base-line corrected and zero suppressed raw data (ZS): base-line corrected raw data with zero suppression and run-length encoding;

– Clusterised data (CF): detector data from which the cluster finder has extracted the particle cluster's information.

During normal operation with physics beam, the TPC detector will be continuously read out without any trigger. During calibration or for debugging and commissioning, the TPC will be slowly triggered (typically 10 Hz) and the data will be acquired with an acquisition window of 102.4 μs corresponding to the TPC Drift Time (TDT).

The TPC also requires different triggering and read-out modes. Depending on the data taking mode (physics, calibration, debugging and commissioning) different run types (Physics, Pedestal, Pulser, Laser, Krypton and Technical) are used with variable periodicity: on demand or on a regular basis such as once per fill or once per year. Different triggering, read-out options and data types from the list above are selected according to the TPC requirements.

A summary of all the TPC triggering and read-out operation modes is presented in Table 4.

2.5 Data rates, sizes and throughput

Table 5 shows the maximum detector read-out rate, as well as the average data throughput and data size per interaction (or per trigger, for those being triggered at lower rate) at the detector electronics output for PbPb collisions at 50 kHz. The data rate indicated corresponds to the data volume transferred by the read-out links up to the Common Read-out Unit boards (CRU).

A significant change to the TPC requirements as known at the time of submission of the O^2 TDR calls for a revision of the detector read-out architecture and of the system throttling. This change consists of transferring untouched raw data from the front-end electronics to the CRUs. The CRU will implement the baseline correction and the zero suppression in addition to the cluster finder. The total data throughput from the TPC to the CRU amounts now to 3.6 TB/s with the selected sampling frequency of 5 MHz.

Table 5: Detector parameters at the CRU input: maximum read-out rate, data rate and data size. The data rate and data size are estimated for PbPb interactions at a rate of 50 kHz. The numbers have been extracted from published sources of information where references available and from draft documents otherwise.

Detector Name	Maximum read-out rate (kHz)	Data rate for PbPb collisions at 50 kHz (GB/s)	Average data size per interaction or trigger at the detector (MB)
ACO	100	0.014	0.00028
CPV	50	0.9	0.018
CTP	200	0.02	0.0004
EMC	42	4.0	0.08
FIT	100	0.115	0.023
HMP	7.5	0.06	0.024
ITS	100	45	0.8
MCH	100	2.2	0.04
MFT	100	10.0	0.2
MID	100	0.7	0.006
PHS	42	2.0	0.04
TOF	200	2.5	0.05
TPC	50	3670	73.4
TRD	38	20	0.5
ZDC	100	0.06	0.0012
TOTAL		3753	75.2

3 System operation in continuous mode

3.1 Global control and synchronisation

The overall control of the experiment during Run 3/4 will be performed by the Control, Configuration and Monitoring (CCM) of the O^2 system. It will be the responsibility of the CCM to issue the appropriate commands to the detector electronics, the trigger system and the O^2 itself to configure the system and initiate the data taking periods.

The data streams also need to be synchronised. The data taking in continuous mode will be initiated and terminated by commands issued by the CCM: *StartOfContinuoustoCRUs* (*SOC*) and *EndOfContinuous-toCRUs* (*EOC*). These commands will be used to synchronise all data streams, like the *StartOfData* and *EndOfdata* in the present online systems used during Run 1 and 2. The *SOC* and *EOC* will be sent as attributes to the HB trigger message (see Table 7).

3.2 Heartbeat acknowledge message and heartbeat map

3.2.1 Heartbeat map assembly

Each CRU responds to a HB and physics trigger with a HB acknowledge message containing the data transmission status of the corresponding HBF/trigger and the CRU buffer status word. The concept of HB maps has been introduced in order to allow evaluation and throttling of the data throughput in case the data throughput exceeds the instantaneous system bandwidth. The HB map is representing the read-out buffer states of the entire ALICE detector with a granularity of one HBF and one CRU.

It needs to be foreseen that one CRU is divided into several geographical HB acknowledge regions in order to increase the HB map granularity. In that case one CRU will deliver one acknowledge bit for each of its geographical regions, stating whether the corresponding HBF has been read out successfully

or whether data is missing and has been deleted. It is considered that one buffer status word per CRU is sufficient.

The HBF data is considered collected successfully once all data corresponding to one HBF have left the PCIe buffer in the CRU. In that case, the CRU sends a positive HB acknowledge message containing the HB ID to the CTP via the optical TTS links. In case data of this HBF is fully or partially missing the CRU will autonomously delete data of that HBF and send a negative HB acknowledge message. In addition the CRU will send one word indicating the CRU buffer status at the time of transmission of the HB message. The CTP will assemble all HB messages of a given HBF, after all CRUs have replied with a positive or negative message or after a programmable time-out, in case the HB map message is missing or delayed for too long. Depending on the throttling mode, selected by the CTP and shown in section 3.3, CRU buffer status is taken into account for the HB map evaluation. The CRU buffer status field in the HB acknowledge message will report whether the CRU buffers are full/almost full/almost empty/empty. Depending on the selected throttling mode these status bits are taken into account when assembling the HB map.

3.2.2 Heartbeat map elaboration

Once the HB map has been fully assembled, the CTP applies one out of a pre-defined set of algorithms to evaluate whether the corresponding HBF is considered useful or should be deleted. This decision is called the HB map decision. Depending on the throttling mode, the subsequent HB triggers will be set to HBa or HBr.

3.2.3 Heartbeat map decision

Via TTS the HB map decision is transmitted to all CRUs which will forward it to the FLPs between the transmission of a HBF trailer and the HBF header of the next HBF. In case of negative decision, those HBFs still in the buffers of CRU and FLP, which have not yet been forwarded to the EPN, will be deleted. The CTP also will transmit the full HB map and decision via the CTP read-out to the CTP-FLP and thus is made available to the EPNs where again HBFs can be deleted. Depending on the throttling operation mode, described in section 3.3, the CRUs, FLPs and EPNs will act differently on the data corresponding to the HB map decision.

3.3 Throttling modes

The succession of triggered runs and continuous operated runs will allow to have 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 (see Table 4). The CRU used in continuous mode will receive the HB triggers from the LTU and the corresponding data from the GBT read-out links. 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 modes of throttling envisaged 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.

These modes are described in more detail below.

3.3.1 Autonomous

The autonomous throttling mode consists of giving the right to each CRU to delete data autonomously when it cannot buffer them. The CRU will use the TTS links to send negative HB acknowledge messages when it partially or fully deletes a HBF. The bookkeeping of which HBF has been deleted will be performed by the CTP which will provide to the O^2 system a HB map for each HBF. This mode requires that the reconstruction software is able to deal with TFs with HBFs missing.

Fig. 5 shows an illustration of the autonomous read-out mode. Each green rectangle corresponds to one HBF containing a HBF header, the payload and the HBF trailer. The payload can be empty if no hits have been collected during that HBF. Each red rectangle corresponds to a HBF for which the CRU data has not been collected completely due to buffer overflow and the HBF data has been autonomously deleted. In autonomous mode no further action is taken. Note that HBFs which have been deleted due to a HBr trigger are considered collected correctly and thus the HB acknowledge message will be positive. Each FLP collects all HBFs from its CRUs and assembles the nominal number of 256 HBFs to one STF. The FLPs forward the STFs to the EPN which builds the TF containing the data of all HBFs of all detectors.

The Finite-State Machine (FSM) diagrams of the LTU and CRU for this mode is shown in Fig 6.

3.3.2 Scaling

Fixed scaling: The fixed scaling mode will establish a repetitive policy of acquiring and deleting HBFs according to the estimated capacity of the system. The CTP transmits a fixed pre-defined pattern of accepted and rejected HB triggers (HBa/HBr) selected by the operator. The CRUs send only the HBFs of those HBF selected by an accepted HB trigger and delete those corresponding to rejected HB triggers. This systematic allows tuning of the effective on-time of the continuous read-out with a fine granularity. The sequences can be adapted to the amount of data loss due to buffer overflow. The mode relies on the tracking software being able to work with incomplete or completely missing HBFs in a TF. For rejected HB triggers (HBr), instead of sending no message, a dedicated HBr messaged is transmitted. This allows the detector electronics to continue the counter integrity checks and to send HB acknowledges messages and to send a header/trailer pair with empty data packets to the FLPs. In this mode, for the HB trigger accepted/rejected pattern, it needs to be considered that nominally the HB period is 88.9 μs and the TPC drift time is 102.4 μs. In order to ensure the acquisition of a full TPC drift time, one accepted HB trigger must be followed by a second one. The CRU and LTU state machines are shown in Fig 7.

Autonomous scaling: Alternatively in the scaling mode, the sequence of accepted/rejected HB triggers

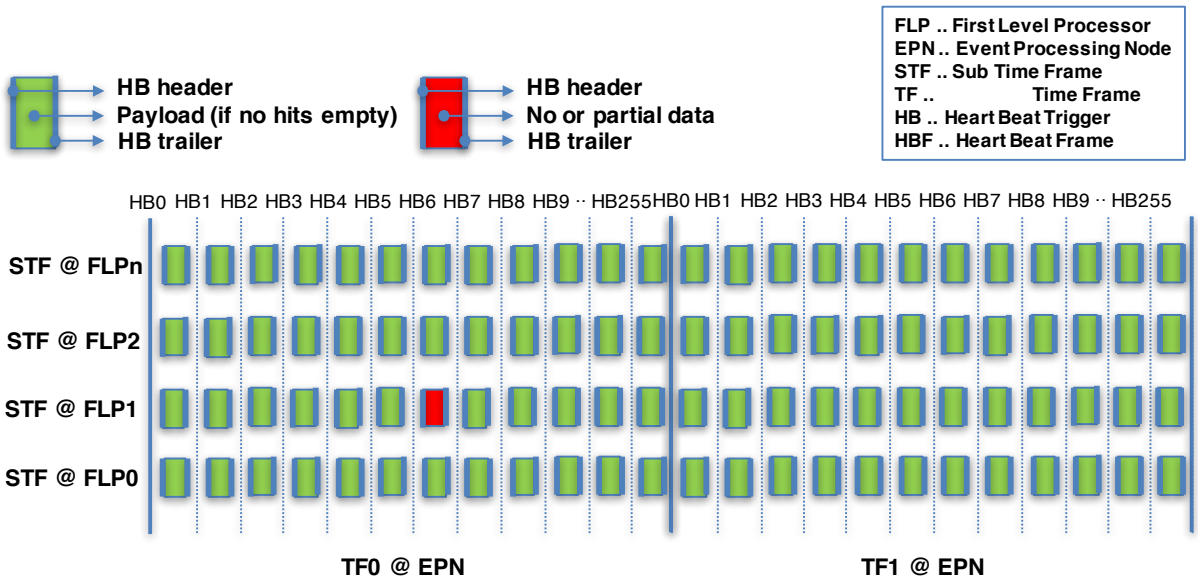


Figure 5: Illustration of autonomous read-out mode.

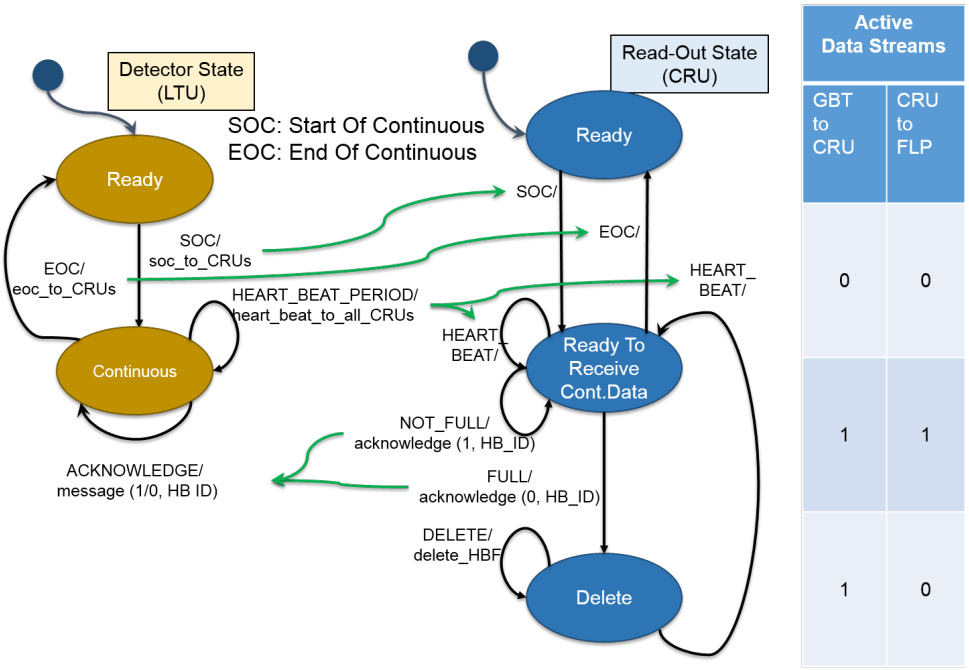


Figure 6: FSM diagrams of the LTU and CRU in continuous reading with CRU autonomous mode.

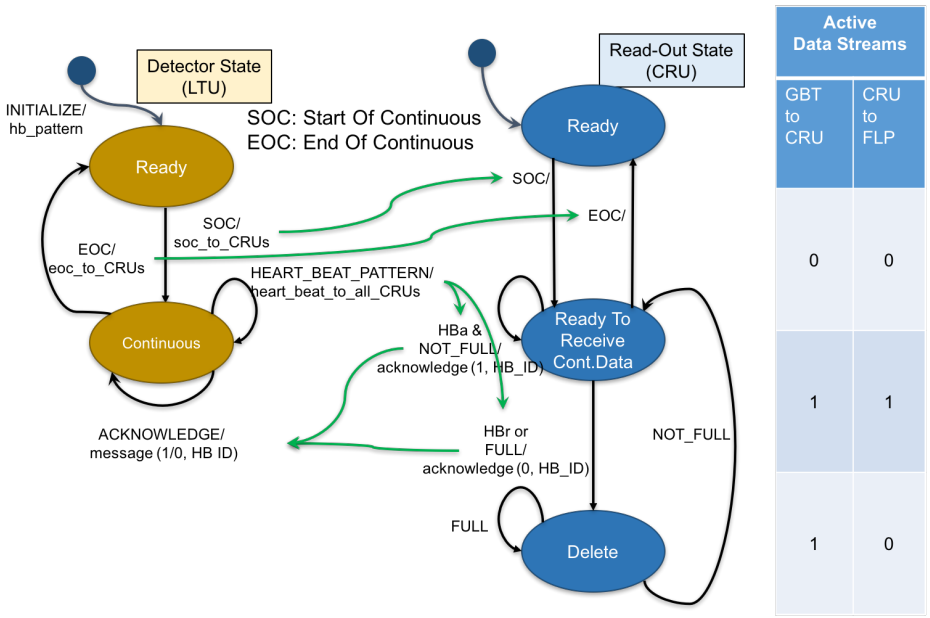


Figure 7: FSM diagrams of the LTU and CRU in continuous reading with CRU fixed scaling mode.

can be created in an autonomous way by evaluation of the HB map in the CTP or a PC using predefined algorithms. The algorithms take ineffective detector read-out regions in geographic proximity but also within successive HBFs into account and adapt the sequence of accepted and rejected HB triggers in an autonomous fashion. As in fixed scaling mode, the CRUs send only the HBFs of those frames selected by an accepted HB trigger and delete those corresponding to rejected HB triggers. The CRU and LTU state machines are shown in Fig 8.

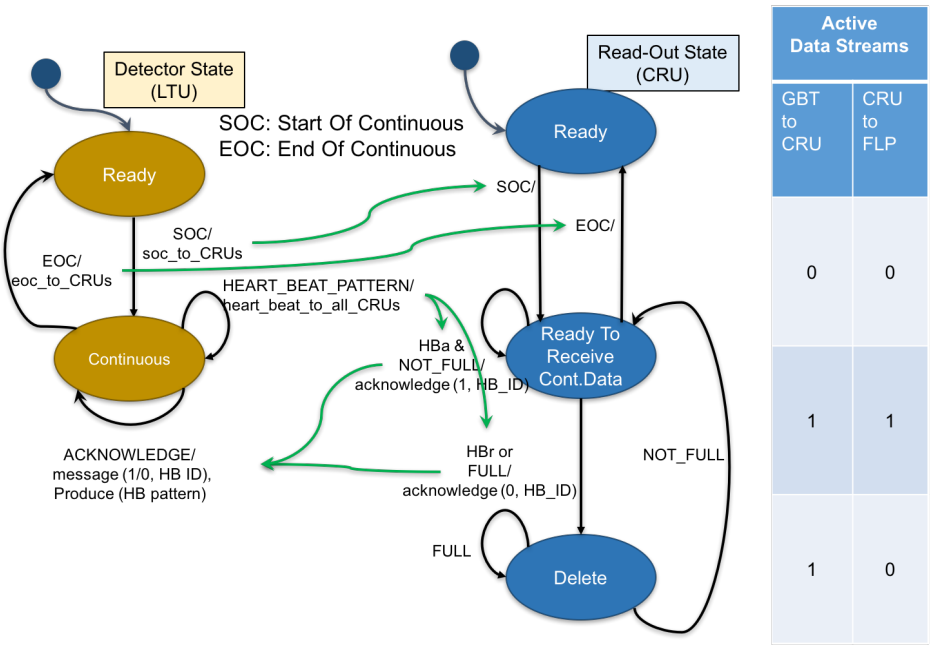


Figure 8: FSM diagrams of the LTU and CRU in continuous reading with CRU autonomous scaling mode.

In both scaling modes it is important that the HB trigger arrives before the corresponding data in the CRU, as otherwise CRU buffer space would be sacrificed to wait for the HB trigger.

Fig. 9 shows an illustration of the scaling read-out mode. Depending on the operation settings a fixed

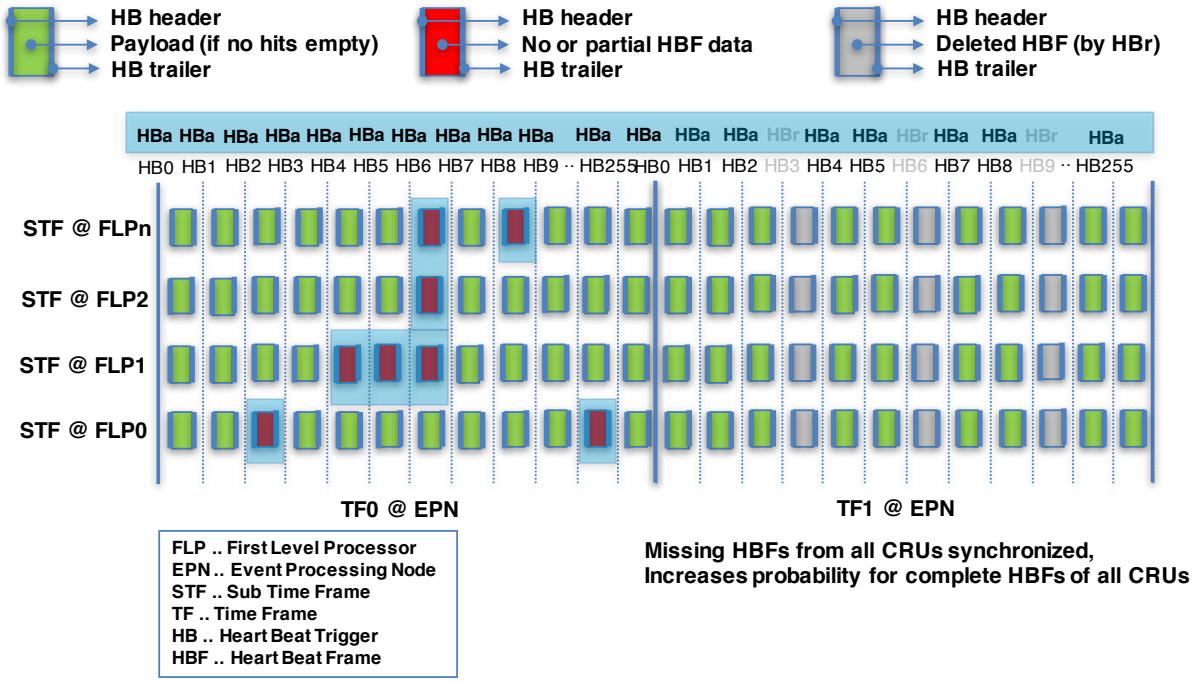


Figure 9: Illustration of scaling read-out mode.

HBa/HBr sequence or a sequence derived from the number of incomplete HBFs (in the figure the red rectangles) by elaboration of the HB map is transmitted. The transmission of HBr triggers initiates a synchronised deletion of HBFs (in the figure the grey rectangles) in the CRU and increases the probability to collect complete HBFs in all CRUs.

3.3.3 Collective

In collective mode, once at least one single CRU or a programmable minimum number of CRUs are reporting one or a programmable number of sequential HBFs missing or incomplete, all subsequent HB triggers are set to HBr until the next TF starts and the buffers of all or a programmable minimum of CRUs have been cleared completely. In collective mode the HB map is built using the CRU buffer status information instead of the HBF acknowledge information. From the moment on the HBr sequence is started all HBFs of an incomplete TF (containing typically 256 HBFs) are deleted from the CRU buffers and are not sent to the FLPs. The collective mode allows commissioning the online tracking efficiently.

Fig. 10 shows an illustration of the collective read-out mode. After the occurrence of a programmable number of incomplete HBF transfers (red rectangles) the CTP issues HBr triggers (grey rectancles) until the end of the TF and all CRUs cleared their buffers completely. This mode increases the probability to collect a complete TF or at least one compete HBF in all CRUs.

Fig. 11 shows an illustration of the collective read-out mode with the use of the HB decision to reduce the data flow in the system. In this example the collective mode has stopped sending HBFs from HB5 on to clear the CRU buffers. At the start of the next TF all CRUs have reported their buffers to be empty and the first HBF is collected in all CRUs. From HB2 on CRUs are starting to deliver incomplete HBFs. If the aim is to collect HBFs with all CRUs actively contributing only then the HBFs marked with blue colour are collected but will not be used later and thus occupy bandwidth inefficiently. The HB decision arriving in the CRUs and FLPs allows the deletion of those HBFs still in their buffers and before the data are shipped to the EPNs.

This mode is foreseen for debugging and commissioning, as it assumes that data loss due to buffer overflow is high. This mode makes sure that as many as possible complete HBFs and possibly at least

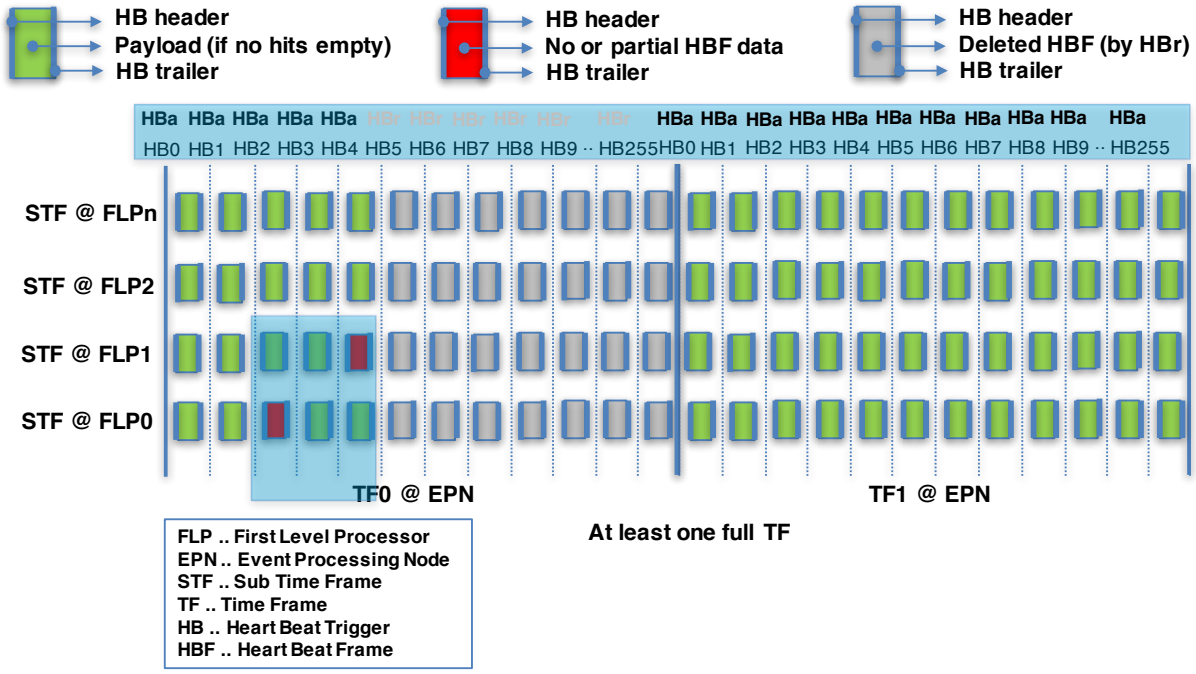


Figure 10: Illustration of collective read-out mode.

one full TF without missing HBFs are acquired.

The corresponding state machines of the LTU and CRU are identical to those of the autonomous scaling mode in Fig 8. The difference in operation is the composition of the HB pattern once the threshold of CRUs reporting incomplete HBFs has been exceeded. In the collective mode the HB triggers are set to HBr at least until the next TF and when a sufficient number of CRUs report again empty buffers.

It should be noted that the initial states (Ready) of the FSMs of the LTU shown in Figs 6, 7, 8, 12 and 14 refer to the same state. The same applies to the FSMs of the CRU in Figs 6, 7, 8, 12 and 14. From its initial state, the system can start a continuous or a triggered period of data taking.

3.4 Data format

The data generation in the data flow chain is depicted in Fig. 13. In continuous mode the detectors only receive the HB triggers as timing reference. As a result, HBFs sent from the CRU to the FLP are referenced to HB triggers. A data transmission must not span over two HBFs. Scheduled data transmissions spanning over two HBFs must be broken in two packets. The reception of the HB trigger must be followed by the transmission of a HBF header followed by the detector data and a HBF trailer or if there were no hits in the corresponding HBF by transmission of a HBF header and trailer only. Additional detector status information can be sent in the HBF header or trailer.

Some read-out detectors require a sequence of auxiliary triggers to initiate a constant stream of read-out transfers. The CRU needs to reformat the data packets to be compatible with continuous HB based read-out detector format. Throttling of these detectors by changing the auxiliary trigger pattern is not foreseen. These detectors are throttled switching to triggered mode or using the throttling techniques in the CRU described in the section 3.3.

The HBF data transmission can be split into several sub HBFs and each of the sub HBF data headers indicate the corresponding HB ID. The HBF transmission is terminated with the HBF trailer, which contains the number of words in a HBF and the information whether this HBF transmission has been completed successfully or whether data has been deleted or only partially transmitted due to buffer overflow. This

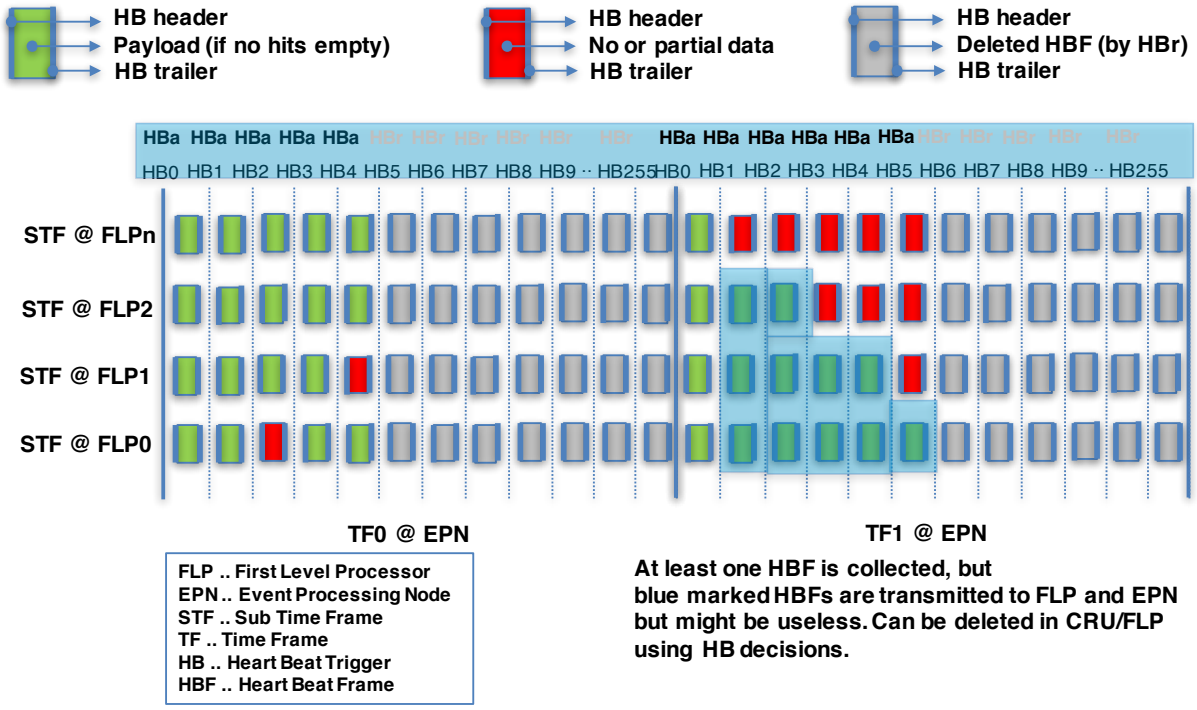


Figure 11: Illustration of collective read-out mode using the HB decision to reduce data flow in the system.

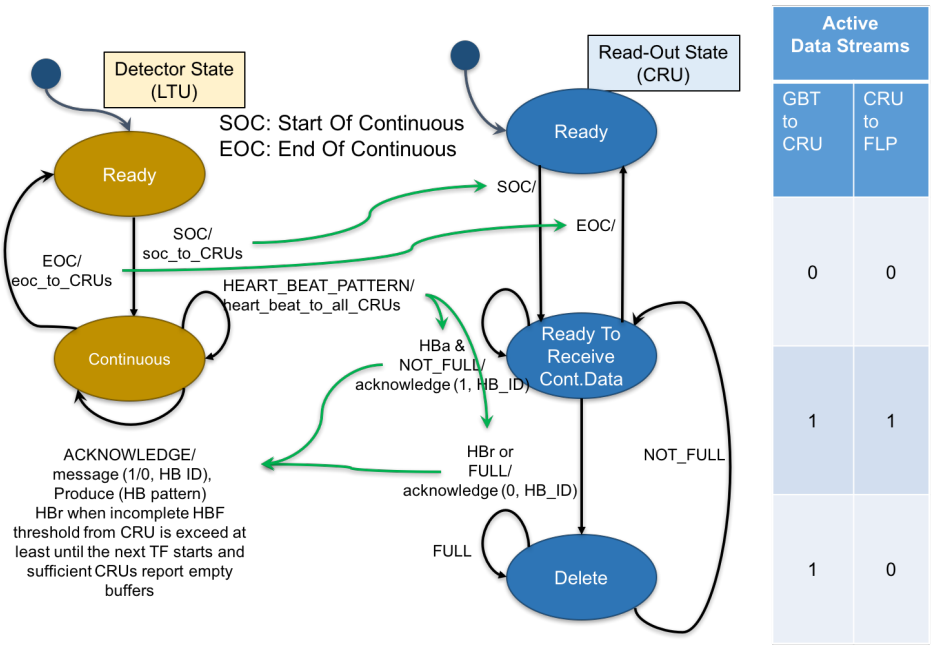


Figure 12: FSM diagrams of the LTU and CRU in continuous reading with CRU collective mode.

information is sent in the trailer as the CRU might not know the number of words or whether the buffer will overflow at the begin of the data transmission. It can be considered that the FLP extracts the HBF trailer information and moves it to the HBF header to facilitate further data processing. In case no hits were recorded during a HBF, the HBF header and trailer indicating zero hits will be transmitted.

Some detectors send their data already divided in sub HBFs from the on-detector electronics to the CRU. An example is shown in fig. 13 (Detector B, link B1/B2). The CRU will multiplex the sub HBF data streams of all links connected into HBFs. In order to allow this multiplexing, the data must be labeled by

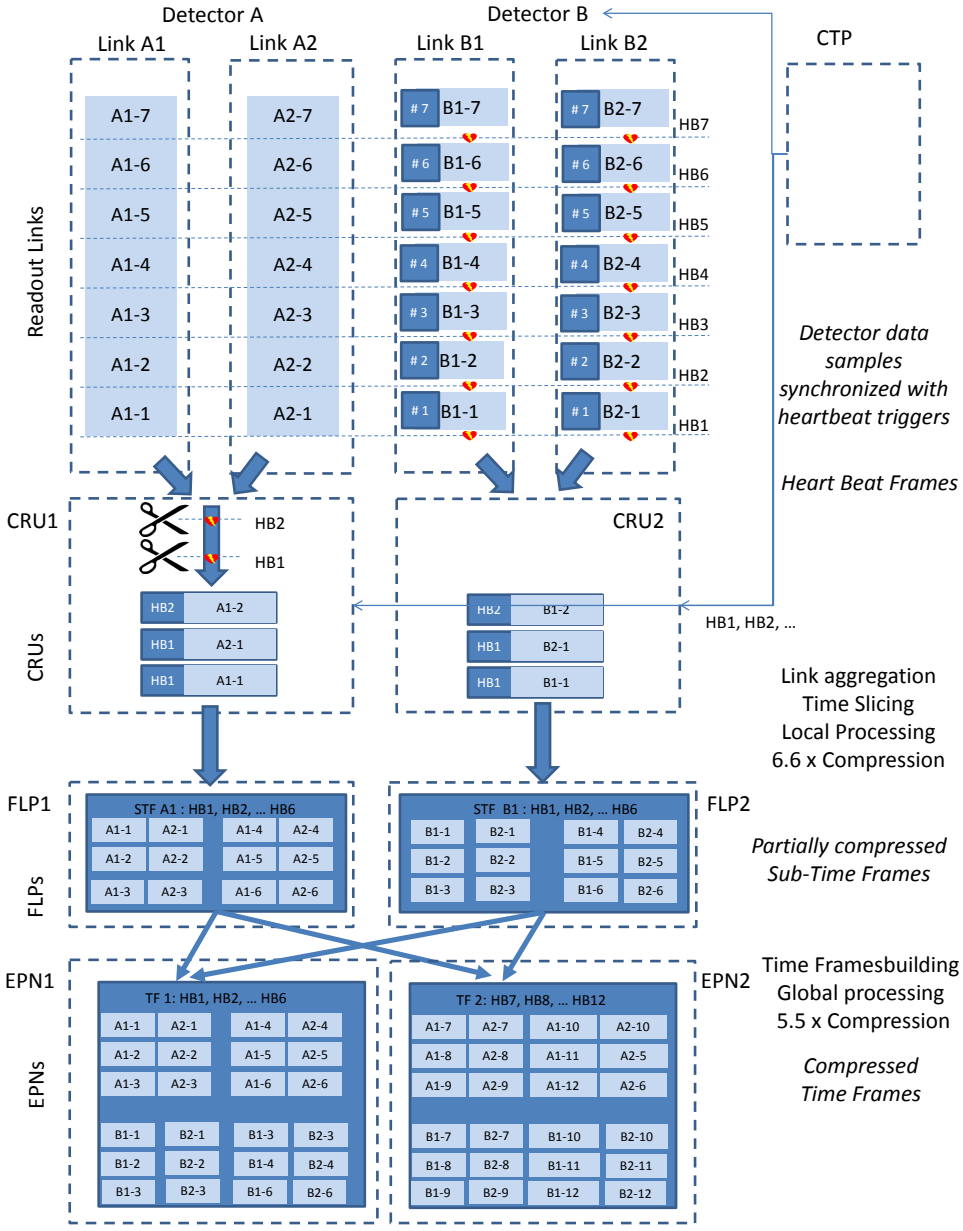


Figure 13: Data aggregation along the data flow from the continuous data streams from the detectors up to the full TFs in the EPNs.

the on-detector electronics with the HB ID or the content of a local HB counter which will be replaced by the HB ID in the CRU. In the most ideal case the CRU will order the data from different sub HBFs in time in order to create a time ordered HBF. In some detector implementations (TPC) the HB trigger is not taken into account in the on-detector electronics. However, the data stream arrives synchronously with respect to the HBF. An example is shown in fig. 13 (Detector A, link A1/A2). The CRU extracts data belonging to one HBF from each link and multiplexes the sub HBF data streams of all links connected into HBFs. The FLPs assemble STFs and forward them to the EPNs.

The CTP transmits the HB map decision to the CRUs for each HBF whenever they are available. After each HBF trailer and before the next HBF header the CRU transmits the HB map decisions of all HBFs received in the meanwhile to their FLPs.

The content of the HBF header, trailer and decision data are shown in Table 6. The content of the HB trigger, acknowledge and decision are described in [5] and shown in Table 7.

Table 6: Content of the HBF/trigger header and trailer.

	Field name	Field length (bits)	position from bit to	Comment
HBF Header (Label 2a in Fig. 15)		64		
Word 1	Block type	4	63..60	=1 HBF/trigger Header
	Header length	4	59..56	
	Reserve	4	55..52	
	Trigger type	4	51..48	HB, TF, physics or software
	Reserve	4	47..44	
	BC ID	12	43..32	Bunch crossing ID
	Orbit	32	31..0	Orbit ID
HBF Header		64		
Word 2-16	Detector specific status		63..0	Additional words according to header length
HBF Trailer (Label 2a in Fig. 15)				
Word 1	Block type	4	63..60	=5 HBF Trailer
	Trailer length	4	59..56	
	HBaccept/reject	1	55	=1 HBa/0 HBr received
	HBF status	1	54	=0 HBF correctly transmitted
	HBF truncated	1	53	=1 in case a new physics trigger arrived within one read-out period
	Status	21	52..32	detector specific status words
	Data Length	32	31..0	Length in words
HBF Trailer		64		
Word 2-16	Detector specific status		63..0	Additional words according to trailer length
HB Decision Data(Label 5 in Fig. 15)				
Word 1	Block type	4	63..60	=3 HBF Decision
	Decision length	4	59..56	by default 1
	Reserve	8	55..48	
	Decision	1	47	=1 positive, =0 negative
	Reserve	3	46..44	
	BC ID	12	43..32	Bunch crossing ID
	Orbit	32	31..0	Orbit ID

Table 7: Content of the HB trigger, acknowledge and decision.

	Field name	Field length (bits)	position from bit to	Comment
HB Trigger (Label 1 in Fig. 15) 64 bit				
Word 1	Trigger type	16	63..48	Orbit, HBa, HBr, TFstart, Physics, Prepulse, Start/End Triggered/Continuous Data Calibration, Detector special tirggers
	Reserve	4	47..44	
	BC ID	12	43..32	Bunch crossing ID
	Orbit	32	31..0	Orbit ID
HB Acknowledge (Label 2b in Fig. 15) 48 bit				
Word 1	Message type	2	47..46	= 1 HB Acknowledge
	Reserve	12	45..34	
	Buffer status	2	33..32	= 3 Buffer full, = 2 buffer almost full = 1 buffer not empty, = 0 buffer empty
	Acknowledge	1	31	= 1 positive, = 0 negative
	Reserve	3	30..28	
	BC ID	12	27..16	Bunch crossing ID
	Orbit	16	15..0	Orbit ID
HB Decision Message (Label 4 in Fig. 15)				
Word 1	Trigger type	2	63..62	= 3 HB Decision
	Reserve	14	61..48	
	Decision	1	47	= 1 positive, = 0 negative
	Reserve	3	46..44	
	BC ID	12	43..32	Bunch crossing ID
	Orbit	32	31..0	Orbit ID

Table 8 shows a graphical representation of the data formats.

4 System operation in triggered mode

4.1 Global control and synchronisation

The ALICE detector will also operate in triggered mode. The trigger is based on the FIT detector and can be complemented by trigger decisions from additional detectors (ACO, EMC, PHS, TOF, ZDC). The trigger architecture has only one level but provides trigger signals with different latencies (LM, L0, L1) and each detector selects one signal according to the detector specification. The system will operate in a very similar way as in the continuous mode.

The data taking in triggered mode will be initiated and terminated by commands issued by the CCM: StartOfTriggeredtoCRUs (*SOT*) and *EndOfTriggeredtoCRUs* (*EOT*). These commands will be used to synchronise all data streams, like the *StartOfData* and *EndOfdata* in the present online systems used during Run 1 and 2. The *SOT* and *EOT* command will be sent as attributes to the HB trigger message (see Table 7). The FSM diagrams of the LTU and CRU for this mode is shown in Fig 14.

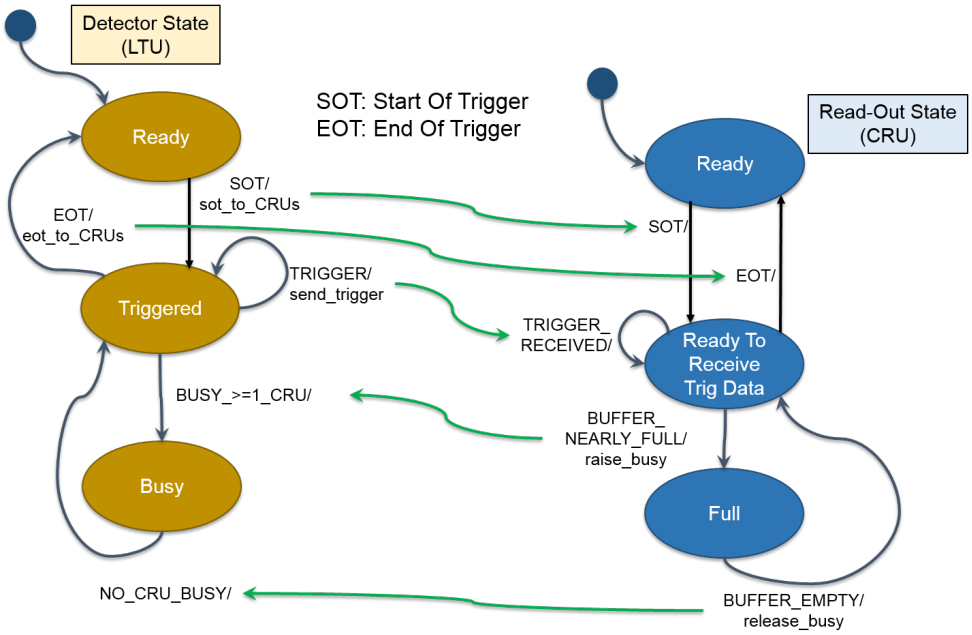
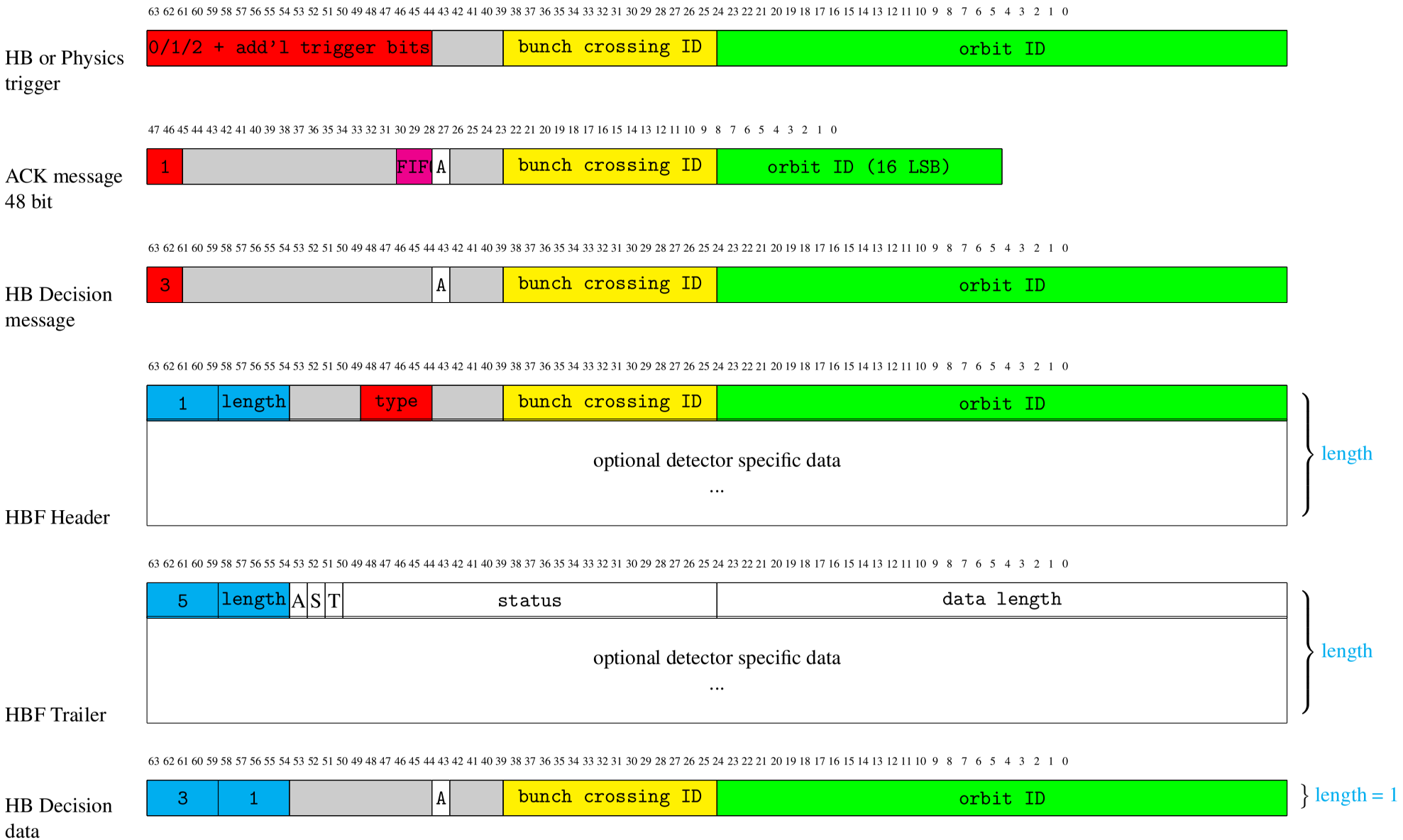


Figure 14: Finite-State Machine (FSM) diagrams of the LTU and CRU in triggered read out with CRU collective mode.

The CTP will continue to send also HB triggers, which need to be responded to with the transmission of a HBF header/trailer pair and a trigger acknowledge message. For non-upgraded detectors in case the overhead of responding to each HB trigger is too high, it is considered to send a dedicated HB trigger at the time when the TF starts. Alternatively for every physics trigger the HB counter could be transmitted.

The CTP will release triggers to each detector and the detector systems will reply for each detector with a dedicated data packet, containing the trigger ID in the header and the detector data corresponding to the trigger in the payload. In all cases with the exception of the TPC the detector data corresponds to the bunch crossing the trigger is referring to. The TPC and ITS/MFT will transmit the data of the corresponding bunch crossing and the data of the TPC drift time afterwards. For some detectors, for example the TPC, in the triggered operation the trigger manager in the CRU will extract the data the trigger signal is referring to and deletes all other data. In this case, the triggered operation can be seen as a special case of the continuous read-out.

Table 8: Data format



Due to the fact that the TPC, ITS/MFT will read out a time period of programmable length corresponding to the drift length, it is possible that another trigger arrives referring to a time span already partially covered by the preceding trigger. In that case, the CRU will truncate the data packet belonging to the previous trigger, inform the FLPs in the data trailer and immediately schedule the transmission of a trigger packet (header, payload, trailer) with the full lengths for the new trigger signal (see HBF truncated field in Table 6).

The CRUs in triggered mode will transmit the full word in the CRU buffer status once the detector cannot process anymore triggers. However, due to the long control loop (CRU-CTP-CRU/front-end), additional triggers might have been issued before the busy signal has arrived at the CTP. Thus, all detectors need to be able to process triggers and maintain the counter and state machine integrity even when they receive a trigger once they are busy. The reaction when receiving a trigger, although being full, is the transmission of an empty data packet (header/trailer), confirming the reception of the trigger and the busy state (HBF or trigger status = 1 in Table 6).

As during Run1/2, the CTP in Run 3 provides modes where combinations of detectors can be grouped to be triggered and read out together when all detectors in this group are not busy. Thus, in triggered mode the CTP assembles a busy map completely identical to the HB map in continuous read-out and evaluates the state of all detectors by analyzing the CRU buffer status for the upgraded detectors and for the non-upgraded detectors the busy information. In nominal operation conditions this mode will loose its importance compared to the Run1/2 scenario, as the majority of the detectors has a read-out bandwidth higher than the interaction rate. However, this mode allows commissioning and operation in non-nominal conditions.

4.2 Data format

The data format is identical to the continuous mode. Data frames are delimited by HB triggers to from HBFs. The payload of these frames contain only triggered events. Non-triggered events have been filtered out by the detector electronics or the CRU. The definition and formatting of the trigger, acknowledge, decision signals, header and trailer as shown in Table 6 and Table 7 are identical in the continuous and triggered operation.

5 Summary of signal and message flow

Fig. 15 summarises the HB trigger and message flow. In the figure only the HB signals/messages are shown. The physics/software triggers and corresponding acknowledge messages have been omitted for simplicity of the graphic illustration. However, the same physical channels and data format are used for the HB signal/messages and the physics/software signal/messages. The read-out procedure starts with a HB trigger and message being sent to all CRUs (nominally at a fixed bunch crossing ID in the orbit gap, but the position is programmable). This trigger can either be set to 'accepted' (HBa, signal bit '1') or 'rejected' (HBr, signal bit '0') but contains always the message with the HB ID. In fig. 15 this action is marked with the digit '1' in a rectangle.

If the CRUs receive a HBa they assemble the data belonging to the HBF and transmit the HBF header, HB payload (detector data) and the HBF trailer to the FLPs. This action is marked in the figure with '2a'. If no hits occurred in this HBF the CRU will send only the HBF header and trailer. In any case the CRUs will send a positive HB acknowledge signal and the corresponding message (HB ID) to the CTP (in the figure marked with '2b').

In case the CRU cannot send the HBF detector hit data due to buffer overflow or any other malfunction the HBF header/trailer pair is sent to the FLP, stating in the trailer status field the absence of hit data (2a). If data from a given HBF is missing partially, the data of the this HBF should not be transmitted at all to the FLP. However, it can occur that for a given HBF the data transmission started already before the

buffers overflow. In that case, the transmission will be terminated as soon as possible with a HBF trailer stating that the HBF is incomplete. Furthermore, a negative HB acknowledge message is sent to the CTP (2b).

In summary the receipt of the HB trigger is acknowledged by the CRUs to the CTP with the HB acknowledge signal/message. It is active (signal bit '1') if the corresponding HBF has been fully transmitted and inactive (signal bit '0') if the HBF has been collected incomplete and has been deleted. The HB acknowledge message contains the HB ID the signal refers to. One CRU can send more than one HB acknowledge bit in case the geographic granularity needs to be increased. In addition the acknowledge message contains the CRU buffer status (full/almost full/not empty/empty) at the time of message transmission.

In case the HB signal from the CTP was inactive the CRU will send a HBF header/trailer pair to the FLP (2a) and a positive HB acknowledge signal/message to the CTP (2b).

For each HBF the CTP builds one HBF map containing all HB acknowledge signals (marked with '3' in the figure). The CTP evaluates the HB map and produces the HB map decision indicating whether the data of this HBF is considered useful in the selected operation mode.

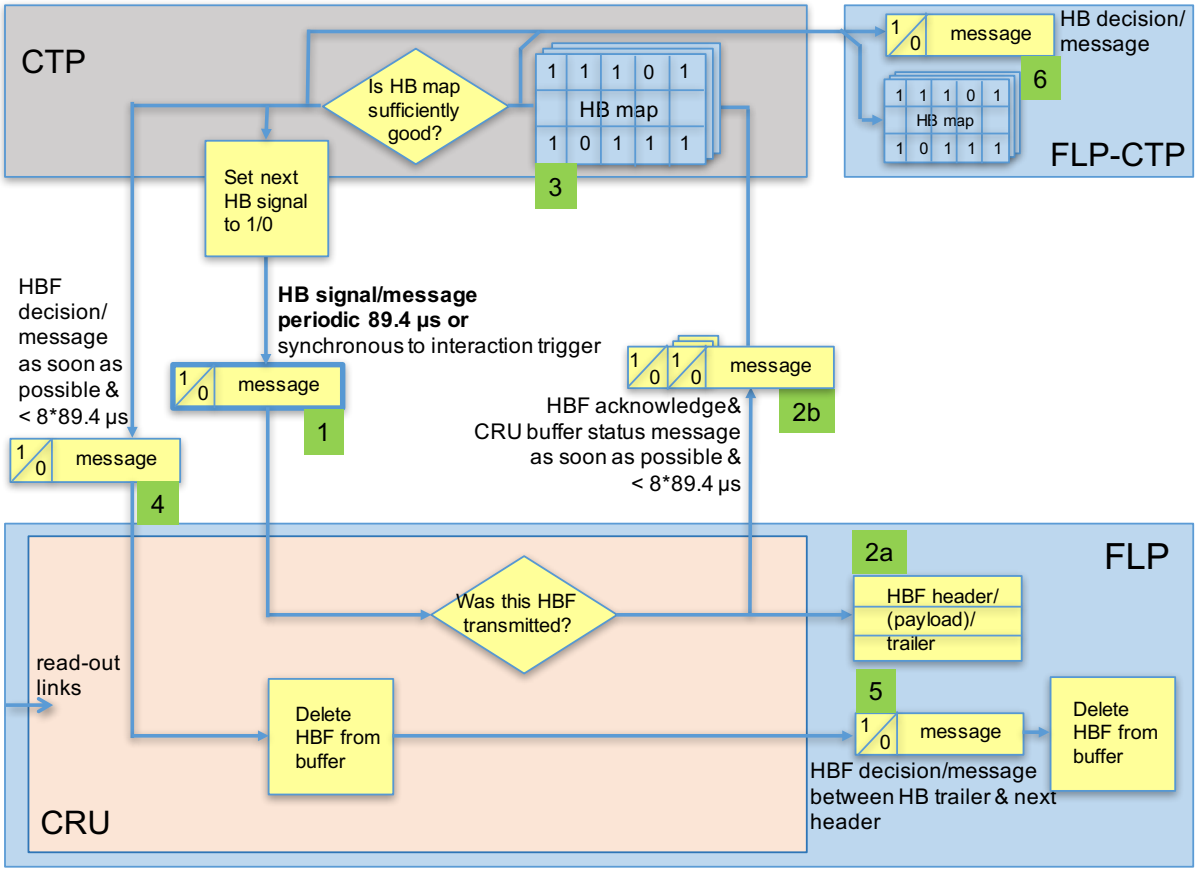


Figure 15: Signal and message flow. The reader is guided by the digits in the rectangles which are described in the text.

According to the HB map decision and depending on the operation mode the following HB triggers can be set to accept/reject (marked with '1' in the figure) in order to adapt the data flow. In addition to the throttling by adapting the HBa/HBr pattern, the HB map decision and message (HB ID) is transmitted to the CRUs. In case of a negative decision the corresponding HBFs are deleted provided the HBFs are still in the CRU buffer (marked with '4' in the figure). The HB map decision/message is forwarded to the FLPs where the HBFs or in collective mode full TFs can be deleted if they are still in the FLP buffer

(marked with '5' in the figure).

The CTP read-out to the CRU-CTP contains the HB map and HB decision which allows the EPNs to delete inactive HBFs or in collective mode full TFs (marked with '6' in the figure).

The transmission time of the HB trigger/message is given by the periodic occurrence of the bunch crossing ID reserved for the transmission. The HB acknowledge signal/message is sent from the CRU to the CTP as soon as the corresponding HBF has left the PCIe buffer or data loss occurred. A maximum time of 8 times the orbit duration (8*88.9 μs) is aimed at, but needs to be programmable. The estimate of 8 * 88.9 μs is derived from the maximum CRU buffer depth of 56 Mb and the TPC raw data rate of 89.6 Mb/s (20 GBT links * 4.48 Gb/s). In case the PCIe interface gets unavailable it takes (56 Mb/89.4 Gb/s =) 630 μs (8 orbit durations) until the buffer is full and the CRU can send the information via the HB acknowledge message to the CTP that data loss occurred. The HB map decision signal/messages is sent from the CTP to the CRU immediately after the HB map has been assembled and evaluated. In case a negative HB decision signal/message arrives too late in the CRU and the CRU already started the transmission to the FLP, the CRU attempts to truncate the HBF transmission and sends a HBF trailer. The HB map decision signal/message is sent from the CRU to the FLP after the HBF trailer and before the next HBF header. This allows the FLP to discard HBFs with negative HB map decision signals.

References

[1] ALICE Collaboration. Upgrade of the Readout and Trigger System, Technical Design Report. CERN-LHCC-2013-019 / ALICE-TDR-015.

[2] ALICE Collaboration. Upgrade of the Online-Offline computing system, Technical Design Report. CERN-LHCC-2015-006 / ALICE-TDR-19. URL: https://cds.cern.ch/record/2011297/files/ALICE-TDR-019.pdf .

[3] A. Kluge, C. Lippmann, J. Wiechula et al. Run 3 calibration and raw data requirements for the ALICE-TPC. Feb. 10th, 2016.

[4] M. Krivda and J. Pospisil for the ALICE Collaboration. The ALICE Central Trigger Processor (CTP) Upgrade. Proc. Of the TWEPP 2015 conference.

[5] D. Evans et al., Trigger system design review, April 2016.