7 Readout and Data Flow

This chapter discusses all aspects of the TRD data flow. Generally, there are two main data streams in three areas to be handled in real time. Both data streams require the simultaneous readout of all 64224 MCMs.

The first data stream is the shipping of the *tracklet* candidates produced on the individual MCMs to the global tracking unit (GTU). This *tracklet* shipping has to be performed during the critical trigger decision time and is limited to 400 ns (refer to Fig. 5.6, note the 200 ns setup time for the first *tracklet* to percolate through the readout tree). During that time, a total of up to 40×32 Bit *tracklets* per chamber have to be shipped to their appropriate φ sector of the GTU, resulting in an aggregate data stream of 216 GByte/s.

The second data stream is the raw data readout, which is performed upon a Level-2 accept (L2A). At that time, the event buffers on the MCMs are being read out. This readout is performed during the TRD dead time as the TRD front-end is not pipelined. For a detailed discussion of the TRD states and timing, refer to Chapter 5.

The data path begins at the MCMs, and ends, in case of a L2A, with the data shipment via the ALICE optical detector links (DDL). Therefore, there are three general regions of data shipping involved : the data flow on the chambers; the cabling between the chambers and the GTU; and, the data shipping off the GTU itself. All three regions are detailed below in the appropriate sections.

7.1 Data types and format

A major cost factor is the required connectivity between the MCMs as this drives the number of pins and connections required. Further, a large number of I/O signals increases the complexity of the readout plane. On the other hand, the tight latency requirement drives up the data transfer rates and bus widths. Therefore, the readout trees are designed to meet the requirements of the *tracklet* shipping. The raw data readout upon L2A uses the defined *tracklet* readout tree, which at that time is idle.

7.1.1 Tracklets

Tracklet candidates, which pass the defined p_t and PID cuts within one plane (MCM), have to be shipped to the global tracking unit for track matching. In order to assist the track matching, each *tracklet* candidate is projected onto the GTU reference plane prior to the shipping.

Table 7.1: Data fields of *tracklet* and TRD tracks

Type	*tracklet* Bits	TRD Track Bits	Description
y-position	13	18	$8 \times 18 \times 7.2$ mm with a resolution of 400 μm
y-deflection	5	7	± 8 mm to pass cut incl. one sign bit
z-position	4	10	max. 16 pad rows per chamber
charge	6	8	normalized charge above MIP
TR	2	4	TR quality flags
variance	1	4	fit quality flags
spare	1	4	fit quality flags
Σ	32	64	w/o Hamming Code

The *tracklet* parameters include y-position, y-deflection, the z-position or pad row number in the reference plane, the normalized charge relative to MIP, the fit variance, and some TR quality flags. The

number of bits required for each of these parameters is determined by the resolution and dynamic range. Table 7.1 shows the appropriate encoding. Correspondingly, each *tracklet* requires 32 Bit for encoding.

7.1.2 Raw data

All digitized ADC values are stored in event buffers, which are being read out upon a L2A. This readout is performed by the CPUs of the *tracklet* processors and, therefore, is completely programmable. Any preprocessing or reprocessing of the data is conceivable using the 256896 processors available. However, in order to understand the data flow, the largest typical data format is the zero suppressed raw data. Zero suppression is implemented in the standard form, running on the freely programmable *tracklet* processors. The zero suppression algorithm implements a configurable threshold, plus some pre and post history, while always reading out the appropriate neighboring channels in order to guarantee the complete readout of a cluster and while maintaining relatively high thresholds. The resulting data is run-length encoded in order to suppress the baseline zeroes. The raw data values are presented in Table 7.2. Here 'black event' represents all available pixels, including the readout of redundant borderline ADC channels.

Table 7.2: Average raw data parameters for Pb–Pb collisions.

Type	Value	Notes
Number of ADC channels	1348704	each MCM supports 18 PADs, plus three ADCs at borders
ADC resolution	10 Bit	
Number of time bins (event buffer can handle 32)	20	active drift time with start and end time configured
Size of 'black event'	39.4 MByte	assuming the readout of redundant borderline ADC channels
Overall occupancy	14%	overall pixel occupancy
Raw event size	7.1 MByte	zero suppressed raw event including 20% coding overhead

There are various options to compress the raw event further in a binary lossless fashion. This can be implemented both at the front-end and the back-end. Given that the TRD raw event is the second largest in ALICE, some effort will be invested within the framework of the high-level trigger project to reduce this sub-event to its minimum size. However, given the experience of the TPC data compression R&D [1] [2], it is expected that only a factor of 50% might be feasible. Huffman encoding can easily be implemented in the front-end. Other compression techniques might be implemented in the back-end.

7.2 Hardware implementation

As described in the introduction, the total transfer time for the trigger is limited to 600 ns. This is the sum of two contributions, i.e., the latency and the duration of the data transfer phase. To allow the operation of the GTU parallel to the data transfer, the readout sequence has to be chosen carefully as described later in this document. To maximize the overlap between processing and data transport, the latency has to be kept to a minimum.

7.2.1 Readout scheme

The organization of the readout follows the structure given by the hardware layout. A readout tree covers a plane of a supermodule consisting of five chambers with up to 16 pad rows. From simulations, it is known that a chamber will provide a maximum of 40 *tracklets* (Chapter 6) with a size of two 16 Bit words. Since the readout for the data acquisition poses a much lower constraint on the system than the trigger, the design is driven by the requirements of the latter.

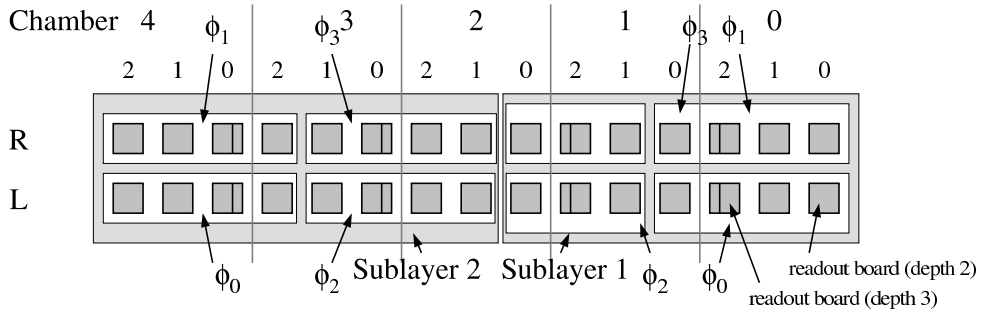

Figure 7.1: Layout of readout boards. The phases $\phi_0..\phi_3$ correspond to the time multiplexed inputs of the root of the tree feeding the optical detector links (refer to Fig. 7.2). Each chamber implements two rows of readout boards labelled as right and left. The readout boards within a chamber are numbered in ascending order in z direction. The same numbering scheme is applied to the chambers themselves. Each layer of a super module is being readout at both sides, therefore implementing two readout trees. The corresponding two logical areas are called sublayer 1 and 2.

To minimize the number of components in the trigger system, most of the readout tree is integrated into the digital part of the MCM, the LTU. To keep the system simple, the same frequency of 120 MHz is used for the *tracklet* processor and the transfer between MCMs. To achieve a small latency, the readout tree has to be as short as possible. However, the width is limited by the number of pins available. To achieve the necessary robustness, LVDS [4] and a 1 Bit error correction and 2 Bit error detection using Hamming coding [5] is foreseen. As a consequence, 46 pins for a 16 Bit data port are needed corresponding to the 16 Bit data, the 5 Bit for the Hamming code, 1 Bit for parity and 1 Bit for the strobe. A tree width of four requires five ports resulting in 230 I/O pins, which, together with pins for control signals and power, are a possible compromise. In the current design, five clock cycles are needed to ship data through a node of the tree. This is a result of the four cycles needed to register and synchronize the input and the one cycle to register the output. With the given clock rate, this corresponds to roughly 42 ns. For a tree with depth d, the latency is $(d \times 5 + 2)$ cycles.

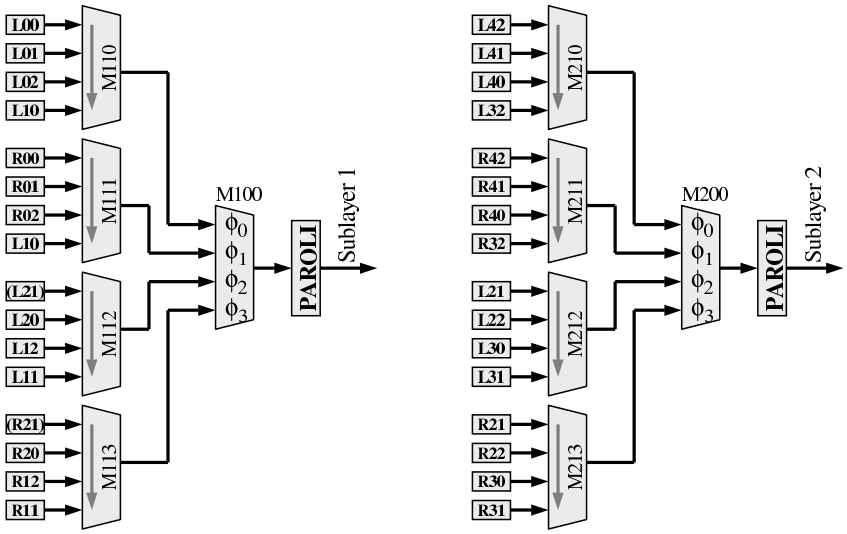

Figure 7.2: Tree structure for a layer. The readout boards are labeled in the following way : *ABC*. *A* distinguishes between the left and right row in Fig. 7.1. *B* is the chamber number and *C* indicates the pad row group inside a chamber. The gray arrows inside the mergers define the readout sequence.

The latency of the readout tree defines the worst case time between the shipment of the first *tracklet* and its receipt by the GTU. During this time no overlapped processing is possible and therefore it should

be kept as short as possible. On the other hand the depth of the readout tree is determined by the fixed number of data sources and the number of links one merger chip can process. The larger the link count of a merger chip is, the larger is its pin count and the more complex is the resulting routing on the readout board. Therefore the number of links and the readout trees latency are competing parameters to be optimized. A large number of scenarios were studied, where the given granularity effects were specifically taken into account. The result is a merger with four inputs and one output.

With the given five ports on a merger chip, a layer is partitioned as shown in Fig. 7.1 and 7.2. The chambers are read out by MCMs grouped together logically on two types of readout boards. These boards can house a maximum of 21 and 25-MCMs, as sketched in Fig. 7.3. One represents two levels and the larger one represents three levels of the tree. The larger sized boards are needed to cover the chambers with 16 pad rows. By ordering the readout sequence, the extra level of the second one can be hidden. As shown in Fig. 7.2, the output of four boards is merged together by the units named Mxxx. These mergers are the same MCMs as used for readout and *tracklet* processing. However the existing LTU functionality is disabled if the MCM is operated in merger mode. On the next tree level the outputs of the previous level are also grouped together by an MCM. The modules M100 and M200 send the data time multiplexed to a gigabit parallel optical link (PAROLI). The PAROLI device is described in greater detail in Section 7.2.2.1.

7.2.1.1 The readout logic on the MCM

The readout system is based on two different hardware components. The MCM, configured as readout tree module (Mxxx), and the parallel optical link (PAROLI). Note that LTU and readout tree functionality can be combined on one MCM by enabling both parts of the digital chip. Referring to the left scenario in Fig. 7.3, implementing 21 MCMs, the first four columns (MCM 4...1, 9...6, etc.) implement regular LTU functionality as discussed in chapter 6.3. The outputs of these LTUs are routed horizontally and are terminated at the rightmost column (MCM 0, 5, 10, 15). These MCMs operate both LTU and track merger functionality. Therefore they effectively merge the inputs of five LTUs into their output link. Finally the output links of the four MCM rows are combined by one MCM (sketched as trapezoid), which only operates track merger functionality, keeping its built-in LTU disabled. Therefore this MCM does not add internal data to the data stream like the ones discussed above. The output of the readout boards sketched in Fig. 7.3 form the inputs labelled Lxx, Rxx to the actual readout tree as sketched in Fig. 7.2. The numbering scheme is defined in Fig. 7.1. The subsequent layer of the readout tree implements MCMs in the same configuration. The last stage of the readout tree interfaces to the parallel optical output link (PAROLI) and is discussed in chapter 7.2.2.

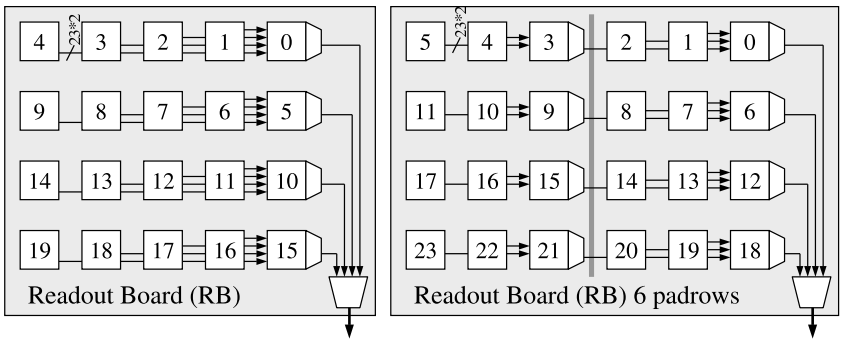

Figure 7.3: Structure within the readout boards. The left board represents a two-level tree, which is used for reading out five pad rows. The readout board on the right can handle six pad rows and adds an additional level to the tree.

Fig. 7.4 gives an overview of the data path inside the readout tree module. The ports and tree control are connected to the MIMD *tracklet* processor (TP) as a periphery mapped into the local and global

address space of the processors. Each CPU has a dedicated output interface, which it can serve independently and asynchronously. However it can also access any other port via the global I/O address space. To allow for maximum flexibility, each port is completely independent and the mapping of the ports is determined by configuration registers in the global I/O address space. In addition, the ports are designed for bi-directional use. These architectural measures simplify the layout and routing of the readout boards and their interconnects. Each physical port can be configured to be either input or output. Referring to Fig. 7.3, many MCMs do not fully utilize their available links, which can be used to implement alternate routes in order to implement some degree of fault tolerance.

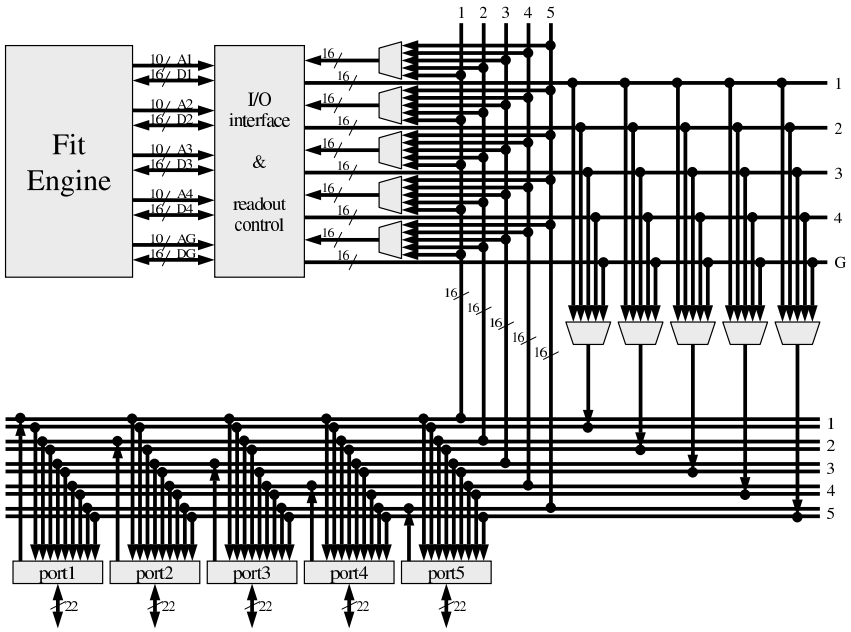

Figure 7.4: Data path of a readout module with the five bi-directional independent ports.

A more detailed view of the ports is given in Fig. 7.5. The bi-directional I/O ports synchronize the data of a previous MCM to its internal clock. This increases the latency, but makes a detector wide synchronous data transfer possible. The Hamming en-/decoder, increases the reliability of the system by implementing one-Bit error correction and two-Bit error detection. The Hamming status is evaluated and linked to the outgoing data stream. The physical signals adhere to the LVDS standard. Using differential signals improves the robustness of the system in a noisy environment.

7.2.2 GTU link

The link to the global tracking unit (see Chapter 5) is used for both *tracklets* and raw data. A layer, consisting of five chambers, is subdivided into two sublayers as shown in Fig. 7.1. To minimize the length for transmission each sublayer is read out to its closest side of the detector. The data are collected at the root of a readout tree and forwarded to a parallel optical link (PAROLI). This results 216 links off the detector.

7.2.2.1 Parallel optical link PAROLI

The Infineon[1] PAROLI links are parallel optical links for high speed data transmission. A complete system consists of a transmitter, a receiver and a fiber optic cable. A PAROLI link has the following main features:

[1] Infineon Technologies AG, *www.infineon.com*

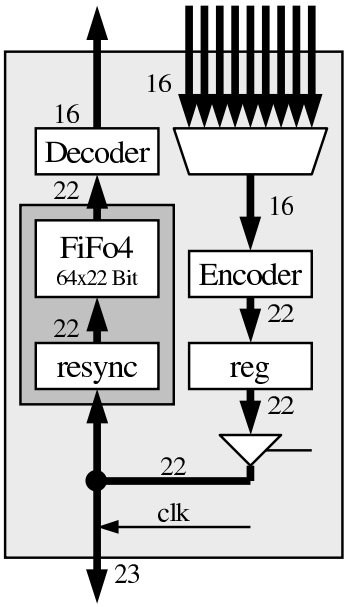

Figure 7.5: Bidirectional I/O port, including the Hamming en-/decoder, synchronization unit and input FiFo. The LVDS transceivers are not shown.

- 3.3 V power supply
- Low Voltage Differential Signal (LVDS) interface
- 22 data + 1 clock channel
- 12 optical data channels
- transmission rate of up to 500 MBit/s per channel
- transmission distance up to 75 m at max. data rate

The transmitter (V23814-K1306-M230) features multiplexing and encoding of 22 electrical data input channels to 11 optical data output channels via time multiplexing. It is closely coupled to the readout module, which builds up the root of a sublayer tree and quadruples the data bandwidth.

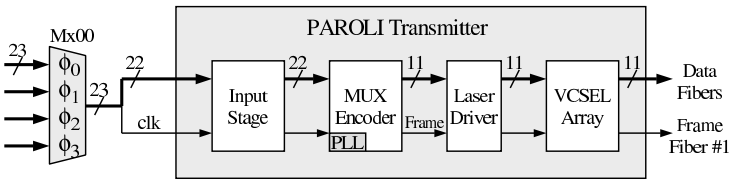

Figure 7.6: Interconnection of readout tree and PAROLI link.

The PAROLI link is operated at 4x the MCM readout link. Therefore the readout modules interfacing to the PAROLI will time multiplex their four inputs. Therefore the data of the four inputs correspond to four phases $\phi_0..\phi_3$ on the PAROLI link. The incoming data arrive at modules M100 and M200 and are being transmitted at the rising edge of the 120 MHz clock. These modules are running with an internal frequency of 480 MHz and forward the data of the four inputs with a 240 Mhz dual-edge clock, corresponding to 480 MWord/s, to the PAROLI link. The receiver (V23815-K1306-M230), as front-end of the global tracking unit, generates a synchronous data output with 480 MWord/s, which will be demultiplexed to 4×120 MWord/s using another instance of the readout tree MCM, operating in reverse mode.

7.2.3 Detector link

Readout of the compressed raw data, which are produced by and read from the local tracking units upon the Level-1 accept (L1A) condition, is triggered by a L2A. At that time the data reside at the TMUs within the GTU. A Level-2 reject (L2R) simply clears the appropriate buffers, which is implemented by advancing the appropriate readout pointers as both Level-1 (L1) and Level-2 (L2) triggers are executed in chronological order within their class and with respect to their associated interaction.

The readout off the detector from the GTU is performed in parallel to any other possible on-going TRD trigger activity, and thus does not contribute to the TRD dead-time except for the derandomizing readout buffer in the GTU threatening to overflow, which can be avoided by making this buffer reasonably large. For example, one 128 MByte DRAM SIMM per detector link will provide space for more than 300 compressed Pb–Pb raw events.

The individual TRD φ sectors can be operated independently even at the level of the global track matching (TMU). The only exception is the collection of the summary data for the L1 trigger decision. Therefore, there will be an event buffer as well as a detector link off the detector for each sector. The corresponding aggregate data bandwidth of 1.8 GByte/s allows for a maximum readout rate of 250 Hz, assuming the stated average event sizes. Should this prove inadequate, the number of links per sector can easily be increased by a factor of two, similar to the sublayer readout of the super modules (refer to Fig. 7.1).

One important processing scenario with respect to data analysis within the high-level trigger is the region of interest processing. Those regions are already distributed with a per sector granularity at L1 time. However, the high-p_t track candidates to be validated in the TPC are better defined in the GTU than in their appropriate sector number. Therefore, an appropriate summary event is planned to be compiled containing the track vectors of all identified high-p_t candidates. The data format is comparable to the one used for the *tracklet* candidates, however, implemented as 64 Bit word as sketched in Table 7.1. Only high-p_t tracks with configurable cuts would be shipped. However, even assuming the maximum shipment of all charged particle tracks, together with a TRD efficiency of 100%, it would result in 16k tracks being transmitted with an event size of 128 kB or a maximum L2 accept rate of 780 Hz.

The data required to be uploaded to the TRD are configuration and calibration parameters, such as the defined thresholds and TMU lookup tables. These data objects, however, are to be provided and maintained by the Detector Control System (DCS). It is an essential system requirement that the DCS be independent of any other system such as DAQ or trigger.

A second logical data stream is the uploading of test data for system integrity checking. An independent data path to the front-end will be implemented in order to enable uploading of test data, and to implement an alternate transparent monitoring data path allowing to monitor system integrity even during normal operation. This data path is an ideal method for redundant but slow readout of any TRD sector, allowing simple off line detector testing without the requirement of an operational DAQ or trigger system.

The TRD data link off the detector uses the ALICE DDL, which consists of three major components : the link feed (SIU), which resides within the GTU; the actual optical link itself; and, the optical receiver card (DIU). The interface to the SIU essentially implements a synchronous 32 Bit data bus running at 40 MHz [3].

In order to keep the TRD on-detector electronics as simple as possible, a data driven push architecture is planned. Upon a L2 accept, all available data is formatted and transmitted through the detector links at design speed. The back-end has to be designed to cope with this data stream of up to 100 MByte/s. Such requirement does not present any particular challenge and is also being used for the TPC readout. Given the availability of very large low-cost elasticity buffers at the back end, the latency requirement for the necessary throttling is relaxed. A canonical 1 GByte event buffer there corresponds to 2500 events or more than a minute of running time when running at full speed. Therefore, the necessary throttling can easily be implemented as a single dead-time signal off the detector, generated at one central place by

implementing an appropriate high water mark. No back pressure towards the detectors is being required or implemented as this would only move the throttling and creation of the dead-time signal onto multiple instances of the front-end, requiring their merging into one common TRD dead-time signal. However, the functionality of the DDL allows to implement a back pressure functionality, allowing to throttle the detector front-end. The potential use and implications of this feature will be revisited at a later stage. In any case, the minimal requirement for the DDL is sustained 100 MB/s half duplex throughput.