Evolution of the data aggregation concepts for STS readout in the CBM experiment

To cite this article: Wojciech M. Zabołotny *et al* 2025 *JINST* **20** C03018

View the article online for updates and enhancements.

TOPICAL WORKSHOP ON ELECTRONICS FOR PARTICLE PHYSICS
UNIVERSITY OF GLASGOW, SCOTLAND, U.K.
30 SEPTEMBER–4 OCTOBER 2024

Evolution of the data aggregation concepts for STS readout in the CBM experiment

Wojciech M. Zabołotny [a,*] **David Emschermann,** [b] **Marek Gumiński,** [a]
Michał Kruszewski, [a] **Jörg Lehnert,** [b] **Piotr Miedzik,** [a] **Walter F.J. Müller,** [b]
Krzysztof Poźniak [a] **and Ryszard Romaniuk** [a]

[a] *Institute of Electronic Systems, Faculty of Electronics and Information Technology,*
Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warszawa, Poland

[b] *GSI — Helmholtzzentrum für Schwerionenforschung GmbH,*
Planckstraße 1, 64291 Darmstadt, Germany

E-mail: wojciech.zabolotny@pw.edu.pl

ABSTRACT: The STS detector in the CBM experiment delivers data via multiple e-links connected to GBTX ASICs. In the process of data aggregation, that data must be received, combined into a smaller number of streams, and packed into so-called microslices containing data from specific periods. The aggregation must consider data randomization due to amplitude-dependent processing time in the FEE ASICs and different occupancy of individual e-links. During the development of the STS readout, the continued progress in the available technology affected the requirements for data aggregation, its architecture, and algorithms. The contribution presents considered solutions and discusses their properties.

*Corresponding author.

Contents

1   Introduction

Silicon Tracking System (STS) [1] is one of the detectors in the CBM experiment prepared in FAIR/GSI in Darmstadt. The experiment uses triggerless free streaming data acquisition [2]. The STS detector Front-End (FE) ASICs (SMX) mounted on Front-End Boards (FEBs) produce 24-bit data containing, among others, timestamped hits and epoch markers,[1] which are 8b/10b encoded [3] and sent at 320 Mb/s via more than 20000 e-links connected to GBTX ASICs [4] in readout boards (ROBs). The data are further transmitted via more than 1700 GBT-links at 4.8 Gb/s to the data aggregation system.[2] The STS will be placed in a confined space in a dipole magnet [1]. That significantly limits the available uplink bandwidth. Because STS is a tracking detector with significant redundancy, it is possible to maximize the bandwidth at the cost of accepting a small amount of corrupted data. Therefore, no CRC is used for uplink detector data in e-links, nor is forward error correction (FEC) used in the GBT uplinks.[3] In the aggregation system, the data must be received, combined into a smaller number of streams, and packed into so-called microslices containing data from specific time intervals [2, chapter 4.2.1], which are later combined into time-slices used for event reconstruction [5]. The aggregation must consider that data are not ordered according to their timestamp due to readout delay caused by different occupancy of individual e-links and amplitude-dependent processing time in the FE ASICs [6, 7]. Finally, the concentrated data must be delivered via the PCIe interface to the First Level Event Selector (FLES) entry node, connected via the InfiniBand network to FLES computing nodes in the Computer Center. The general structure of the STS readout chain is shown
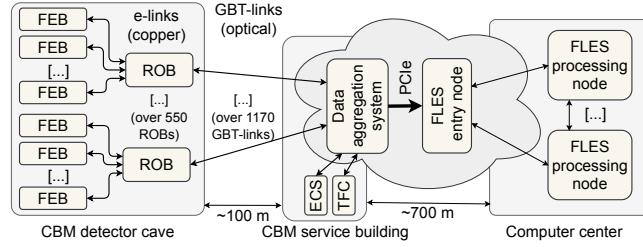
---

[1]The SMX produces the 14-bit timestamp. Its resolution is 3.125 ns for a 320 Mb/s uplink rate, so the wrap-around period is 51.2 μs. The hit messages transmit only 10 lower bits of the timestamp. The epoch markers (TS-MSB) contain the bits 13–8, delivering the remaining 13–10 bits. Bits 9 and 8 are transmitted both in hits and epoch markers [3]. The TS-MSB is protected with 4-bit CRC, and bits 13–8 are triplicated to enable recovery of a partially corrupted marker.

[2]Each e-link delivers maximally one 24-bit 8b/10b encoded data word every 30 b/320 Mb/s = 93.75 ns. A single GBT-link transmits data from up to 14 e-links delivering up to 149.333 Mwords/s.

[3]The main reason for data corruption is the non-zero bit error rate in e-links. The hit data corruption is similar (but significantly less frequent) to the effect of noise or limited efficiency of the STS sensors (additional wrong hits, lost hits). However, data aggregation must consider the possible corruption of epoch messages, which is more serious as it affects the interpretation of a higher number of hits.
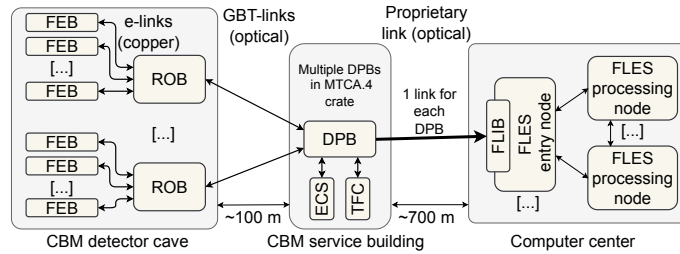
in figure 1. During the development of the STS readout, the continued progress in the available technology affected the requirements for data aggregation, its architecture, and algorithms. Therefore, multiple concepts have been implemented and verified.

Figure 1. The general structure of the STS readout chain in the CBM experiment. The GBT-links transceivers must be located in the CBM service building. The location of the FLES entry node depends on the solution. (ECS — Experiment control system, TFC — Timing and Fast Control).

2   The first concept of the readout

The first proposed STS readout version [8] assumed the use of the intermediate FPGA-based Data Processing Boards (DPB) in the MTCA.4 standard. The MTCA.4 crate provided the possibility to deliver high-speed TFC signals and an IPbus-based control interface. The available MTCA.4 interconnect infrastructure could be reused for non-local preprocessing based on data received by multiple boards. The DPB output was connected via a 10 Gb/s Aurora link to the PCIe FLES interface boards (FLIB) [9].
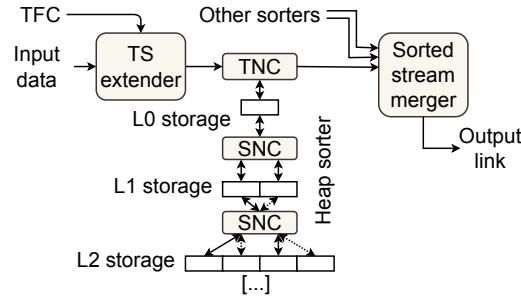

Figure 2. Block diagram of the first DPB-based prototype of the STS readout chain in the CBM experiment.

Each DPB board was expected to aggregate the data from 6 (possibly extensible to 8) GBT-links (see figure 2). For 24-bit data words, the expected maximum data bandwidth was 21.5 Gb/s for 6 GBT links (28.7 Gb/s for 8 GBT links) and was significantly higher than the available output bandwidth. Therefore, a significant data preprocessing was considered to enable data volume reduction. At least, introducing of a context-based data format was planned, where the number of bits needed to transfer a hit information could be reduced. Such processing required perfect sorting of the hit data according to their timestamp (TS). A dedicated TS extender[4] and a heap sorter (see figure 3) blocks have been implemented for that purpose. Unfortunately, in the beam tests, that solution appeared extremely sensitive to sorter overflow due to variations in the data rate caused by beam intensity fluctuations

---

[4]The TS extender restores additional bits of timestamps based on epoch markers received in the particular link.
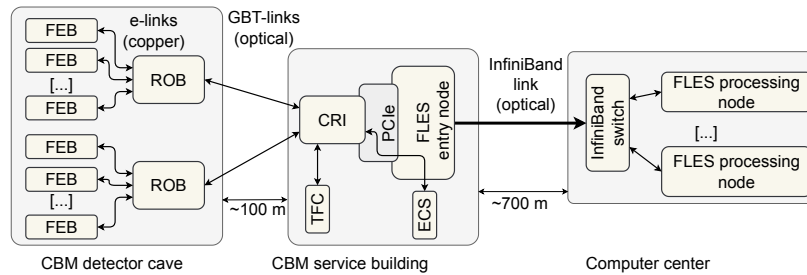
and the data timestamps corrupted by transmission errors in the e-links. Also, implementing any FPGA-based non-local data processing appeared difficult due to high resource consumption.

Figure 3. The data path with the heap sorter and stream merger. (TS — timestamp, TNC — top node controller, SNC — sorting node controller). Reproduced with permission from [10].

3 The second concept of the readout

Avoiding the early non-local data processing enabled the elimination of the MTCA.4 crate and the intermediate FPGA layer. The functionalities of the DPB and FLIB boards have been integrated into new Common Readout Interface (CRI) boards [2, chapter 5.2], which have been implemented as PCIe boards placed in the FLES entry nodes. That change required FLES entry nodes to be moved to the CBM service building, which resulted in significant advantages. The proprietary optical link with a rate limited by the transceivers in the FPGA in the DPB board could be replaced with a standard long-distance InfiniBand link to the FLES processing nodes in the Computer Center, which can run at higher speed and be easier maintained and upgraded (see figure 4). The PCIe interface used to connect the CRI boards offers much higher bandwidth than the proprietary optical link.[5]

Figure 4. Block diagram of the second prototype of the STS readout chain for the CBM experiment.
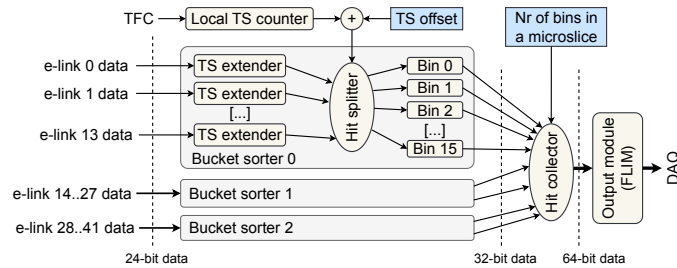
The CRI implements GBT links for ROB connectivity and the FLES Interface Module (FLIM) for PCIe. The hardware platform for the first CRI prototype was the FELIX BNL-712v2 board developed for the ATLAS experiment at CERN [11] but used with CBM-developed firmware. The board may support a higher number of GBT-links — up to 47. Currently 24 GBT-Links are used for STS. The measured PCIe bandwidth ($2 \cdot 6.65$ GB/s = 106.4 Gb/s) is higher than the expected maximum input

---

[5]The theoretical maximum bandwidth of the PCIe Gen3x16 interface is 128 Gb/s. The theoretical maximum bandwidth of the PCIe Gen4x16 interface is 256 Gb/s. Even if it can't be fully utilized, it is much higher than the 10 Gb/s offered by the proprietary optical link offered by the DPB.

bandwidth (86.02 Gb/s for 24 GBT-links), eliminating the need for context-based data aggregation, and thereby relaxing the requirement for perfect sorting.

4 Bucket sorter-based data aggregation

The relaxed requirements for data compression enabled the replacement of the heap sorter with the bucket sorter, providing partially sorted data [2, appendix B.1.4]. In that approach (see figure 5), the data received from a group of 14 e-links are concentrated, and their TS is extended. Then, four bits of TS are used to select the bin. The lower bits are ignored. The higher bits define the acceptable range of timestamps.[6]

Figure 5. Data aggregation with bucket sorters used in the second version of the CBM DAQ readout. Data from 3 GBT-links are delivered to 3 bucket sorters. The data are sorted into 16 bins based on 4 selected bits of their timestamp. The hit collector receives the data from the same bins from all 3 sorters and sends them to the output. The 24-bit e-link data are supplemented with the source ID, forming 32-bit data. Finally, data are packed into 64-bit words used by FLIM. Reproduced from [2]. CC BY 4.0.

This solution better handles intermittent peaks in the data rate. The data delivered to an already full bin are rejected, and the bin's "data lost" flag is set. The data with corrupted timestamps are rejected if their timestamp is outside the limit matching 16 available bins. Of course, the corruption of lower bits may lead to data being stored in an incorrect bin. There is still a small risk of disturbing the operation of the TS extender by the corrupted epoch markers in the data stream. The main problem with that solution is the static allocation of the same memory amount for each bin. In case of data rate fluctuations, it is not possible to compensate for the higher memory occupancy in one bin with the lower memory usage in another bin. In the beam tests, with the bucket sorter handling data from a single GBT-link for reasonable bin duration (3.2 µs) and memory size (1024 words), the data loss due to bin overflow occurred unacceptably often.

5 Aggregation of data with simple concentration

The previously described solutions heavily depend on the timestamps contained in received data, which makes them sensitive to data corruption. Additionally, they significantly modify the data stream. In case of problems, reconstructing original data and diagnosing the problem is impossible. The debugging requires adding a dedicated diagnostic mode (which increases the FPGA resource consumption) or providing a special diagnostic firmware (which requires reconfiguring the FPGAs).
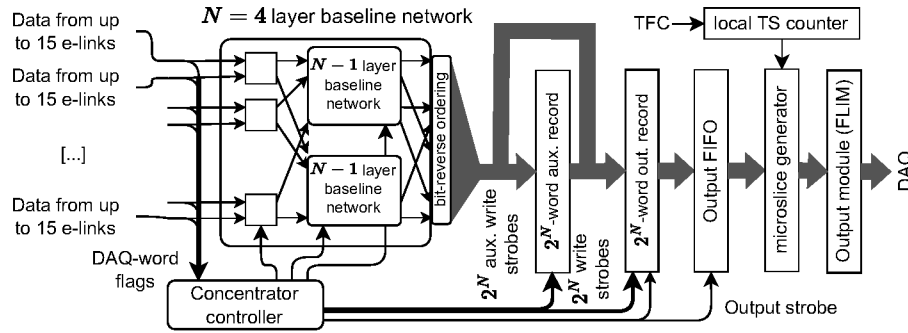
---

[6]The least significant bit of TS corresponds to 3.125 ns. Therefore, if, for example, bits 10 to 13 are chosen for bin selection, the ten lowest bits are ignored and the time period covered by each bin is $1024 \cdot 3.125$ ns = 3.2 µs.

Therefore, yet another aggregation concept was tested. The last progress in PCIe technology (Gen4 and Gen5) further increases the bandwidth available for data transmission. Based on that, a new concept utilizing the simple concentration of data has been created. Other experiments have also used the concepts of using FPGA-based boards as almost transparent data concentrators. For the LHCb experiment at CERN, the PCIe40 board [12] was developed. The transparent data concentration has also been proposed for the ATLAS experiment at CERN [13], and the boards developed for that project are used as the hardware platform for prototype CRI boards in the CBM readout.

In the simple concentration approach, the data words received from a group of e-links (up to 15) are serialized at 160 MHz, supplemented with the source ID (number of e-link and number of the GBT link), creating the 32-bit words. However, the PCIe output module uses a wider data word (256, 512, or 1024 bits). Therefore, the data from multiple groups may be stored in a single output word. A dedicated high-speed concentrator [14] has been created to pack such data into wider words without leaving empty places and wasting clock cycles.

In that approach (see figure 6), the boundaries of the microslices are determined by the arrival time of the data, hence completely eliminating the influence of corrupted data.[7] The original data stream may be fully reconstructed, so the detection of anomalies (including those related to data corrupted in readout) may be implemented in the software. One modification of the original data stream is necessary if an individual e-link does not deliver data for a prolonged time. In such a case, artificially created epoch markers must be inserted.

That solution has been successfully tested with a DMA engine developed for the GERI board [15] with a 256-bit output word, and in a simplified form (with a 64-bit output word) in the first prototype of the CRI board. The tests have confirmed that this aggregation scheme offers the best handling of high data rates among the described solutions.

Figure 6. System performing the simple concentration of the data in STS readout. Reproduced from [14]. CC BY 4.0. As proven in [14], the N-layer baseline network is capable of reordering the data from up to 2^N inputs so that the DAQ words from consecutive inputs are routed to consecutive (modulo 2^N, and in bit-reversed order) outputs. Non-DAQ words are skipped. That enables packing the DAQ words into the output record without leaving empty spaces, preserving their order according to their arrival time at the concentrator and the input number. In the particular clock period, the output record may be partially occupied. In that case, some data are preserved in the auxiliary record until the next clock period.

---

[7]Simplified generation of microslices may lead to a situation where data from the same period are stored in neighboring microslices. However, later analysis uses timeslices built as overlapping sequences of microslices [2, chapter 4.4]. Therefore, each timeslice contains all the data from a given time interval despite potential time disorder in microslices.

6 Conclusions

The preparation of the CBM experiment inspired the development and testing of various methods for the aggregation of detector data. The selection of the particular method depended on the currently available technology. The currently selected solution utilizes the progress in the FPGA and PCIe technology to ensure almost transparent transmission of the detector-produced data stream, eliminating the need for a separate diagnostic mode. All information contained in the original data is available for software processing in the FLES computing nodes. However, the effort invested in developing earlier solutions is not void. The elaborated solutions may be reused in other systems where perfect or partial data sorting in the FPGA layer is necessary. Most solutions described in the paper are open-source and available for the HEP community.

References

[1] J. Heuser et al. eds., *[GSI Report 2013-4] Technical Design Report for the CBM Silicon Tracking System (STS)*, GSI, Darmstadt (2013).

[2] CBM collaboration, *Technical Design Report for the CBM Online Systems — Part I, DAQ and FLES Entry Stage*, GSI Helmholtzzentrum fuer Schwerionenforschung, GSI, Darmstadt (2023), DOI:10.15120/GSI-2023-00739.

[3] K. Kasinski et al., *A protocol for hit and control synchronous transfer for the front-end electronics at the CBM experiment*, *Nucl. Instrum. Meth. A* **835** (2016) 66.

[4] P. Moreira et al., *GBTX manual; V0.18 draft*, (2021), https://cds.cern.ch/record/2809057.

[5] V. Friese, *Event Reconstruction in the Tracking System of the CBM Experiment*, *EPJ Web Conf.* **226** (2020) 01004.

[6] K. Kasinski, R. Szczygiel and W. Zabołotny, *Back-end and interface implementation of the STS-XYTER2 prototype ASIC for the CBM experiment*, 2016 *JINST* **11** C11018.

[7] K. Kasinski et al., *Characterization of the STS/MUCH-XYTER2, a 128-channel time and amplitude measurement IC for gas and silicon microstrip sensors*, *Nucl. Instrum. Meth. A* **908** (2018) 225.

[8] J. Lehnert et al., *GBT based readout in the CBM experiment*, 2017 *JINST* **12** C02061.

[9] D. Hutter, J. de Cuveland and V. Lindenstruth, *CBM First-level Event Selector Input Interface Demonstrator*, *J. Phys. Conf. Ser.* **898** (2017) 032047.

[10] W.M. Zabołotny, *Dual port memory based Heapsort implementation for FPGA*, *Proc. SPIE* **8008** (2011) 80080E.

[11] ATLAS TDAQ collaboration, *FELIX: the Detector Interface for the ATLAS Experiment at CERN*, *EPJ Web Conf.* **251** (2021) 04006.

[12] J.P. Cachemiche et al., *The PCIe-based readout system for the LHCb experiment*, 2016 *JINST* **11** P02013.

[13] J. Anderson et al., *A new approach to front-end electronics interfacing in the ATLAS experiment*, 2016 *JINST* **11** C01055.

[14] W.M. Zabołotny, *Scalable data concentrator with baseline interconnection network for triggerless data acquisition systems*, *Electronics* **13** (2024) 81 [arXiv:2309.13690].

[15] W.M. Zabołotny, *Versatile DMA Engine for High-Energy Physics Data Acquisition Implemented with High-Level Synthesis*, *Electronics* **12** (2023) 883.