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

Wojciech M. Zabołotny[1], David Emschermann[2], Marek Gumiński[1], Michał Kruszewski[1], Jörg Lehnert[2], Piotr Miedzik[1], Walter F.J. Müller[2], Krzysztof Poźniak[1], Ryszard Romaniuk[1]

[1]Warsaw University of Technology, Faculty of Electronics and Information Technology, Institute of Electronic Systems,
[2]GSI - Helmholtzzentrum für Schwerionenforschung GmbH
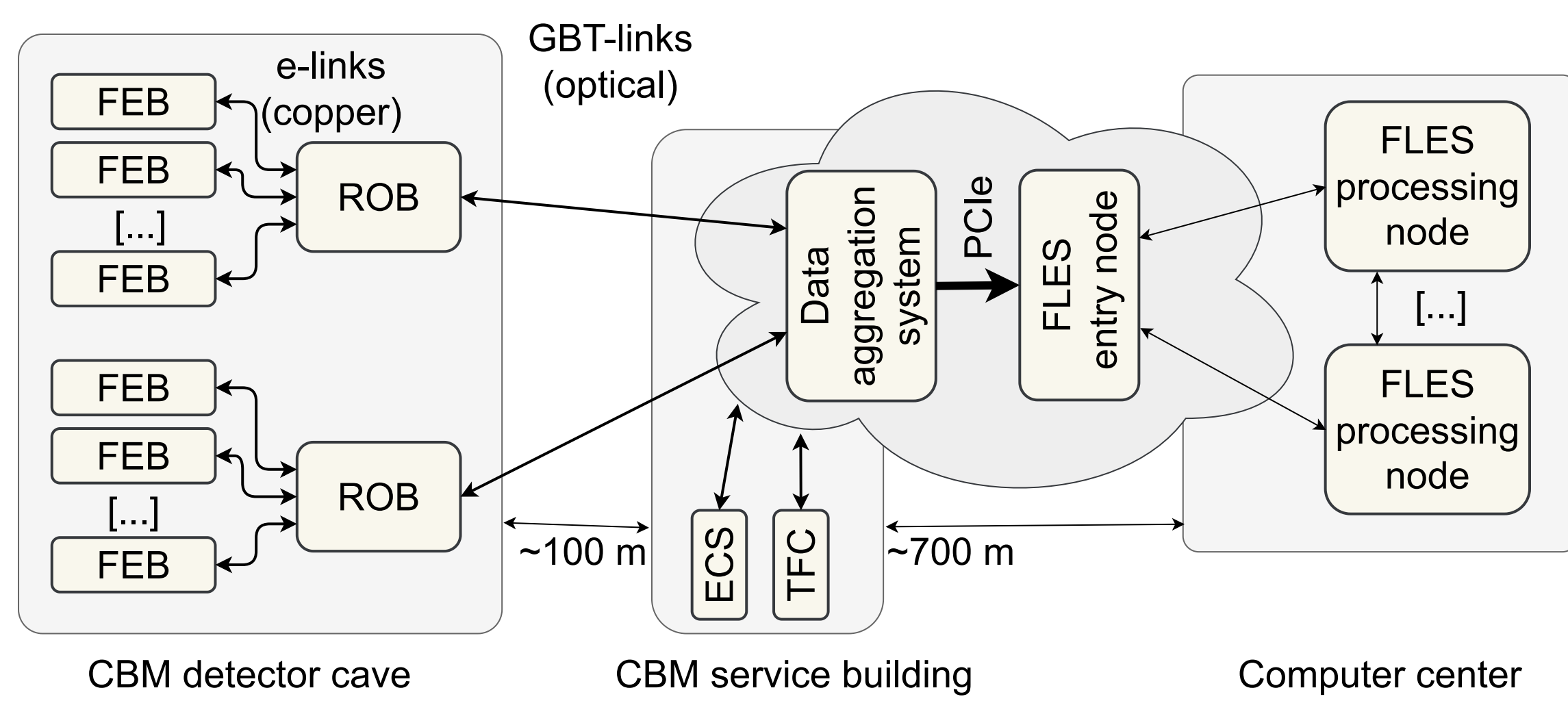
wojciech.zabolotny@pw.edu.pl

Introduction

STS is one of the detectors in the CBM experiment prepared in FAIR/GSI in Darmstadt. The experiment uses triggerless free streaming data acquisition [1]. The STS detector FEE ASICs [2] deliver timestamped data via more than 20000 e-links (8b/10b encoded, 320 Mb/s) connected to GBTX ASICs [3] in readout boards (ROBs), which transmits them further via more than 1700 GBT-links (4.8 Gb/s) to the data aggregation system.
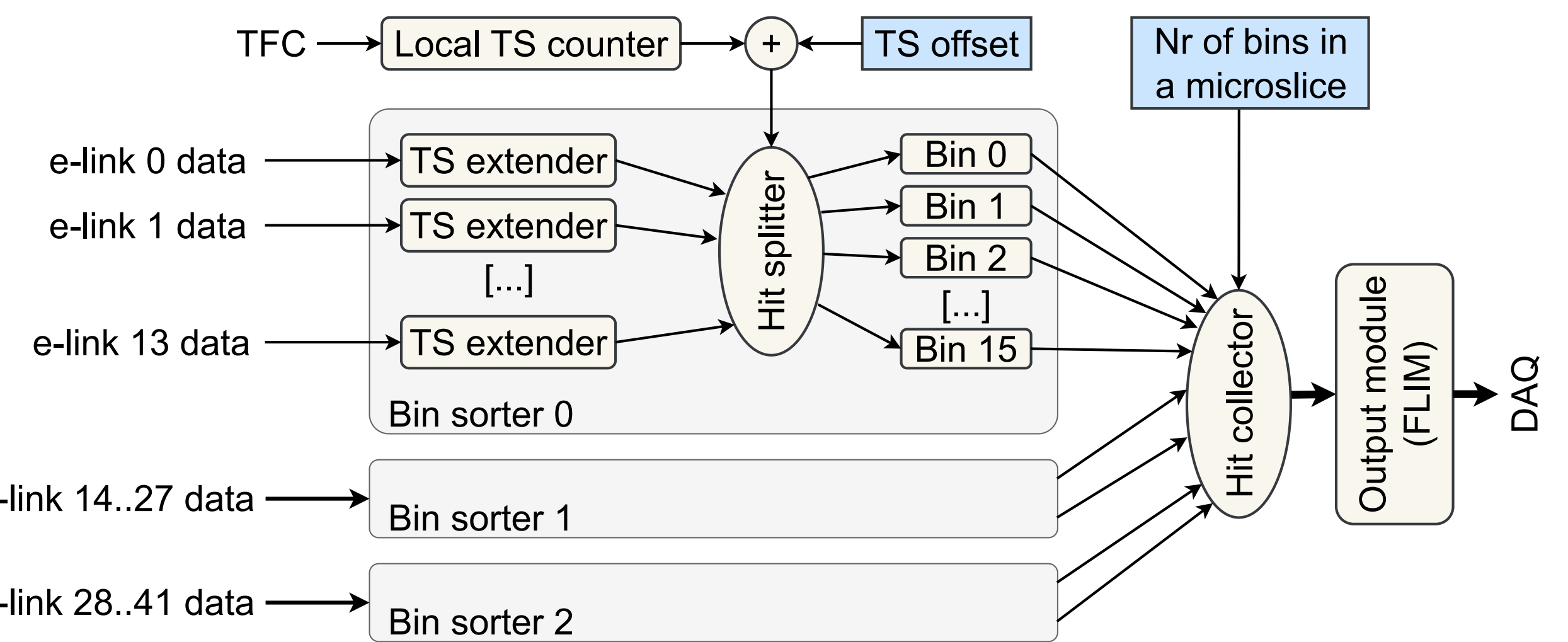
Here, that data must be received, combined into a smaller number of streams, and packed into so-called microslices containing data from specific time intervals. The aggregation must consider that data are not ordered according to their timestamp due to readout delay caused by different occupancy of individual e-links and amplitude-dependent processing time in the FEE ASICs. 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 located in the Computer Center. During the development of the STS readout, the continued progress in the available technology affected the requirements for data aggregation, its architecture, and algorithms. Here, we present the solutions that have been considered and their properties.

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 input node depends on the solution. (ECS - Experiment control system, TFC - Timing and Fast Control)

The first concept of the readout

The first proposed STS readout version [4] 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 data preprocessing based on data received from multiple channels. The DPB output was connected via a 10 Gb/s Aurora link to the PCIe FLES interface boards (FLIB) [5].
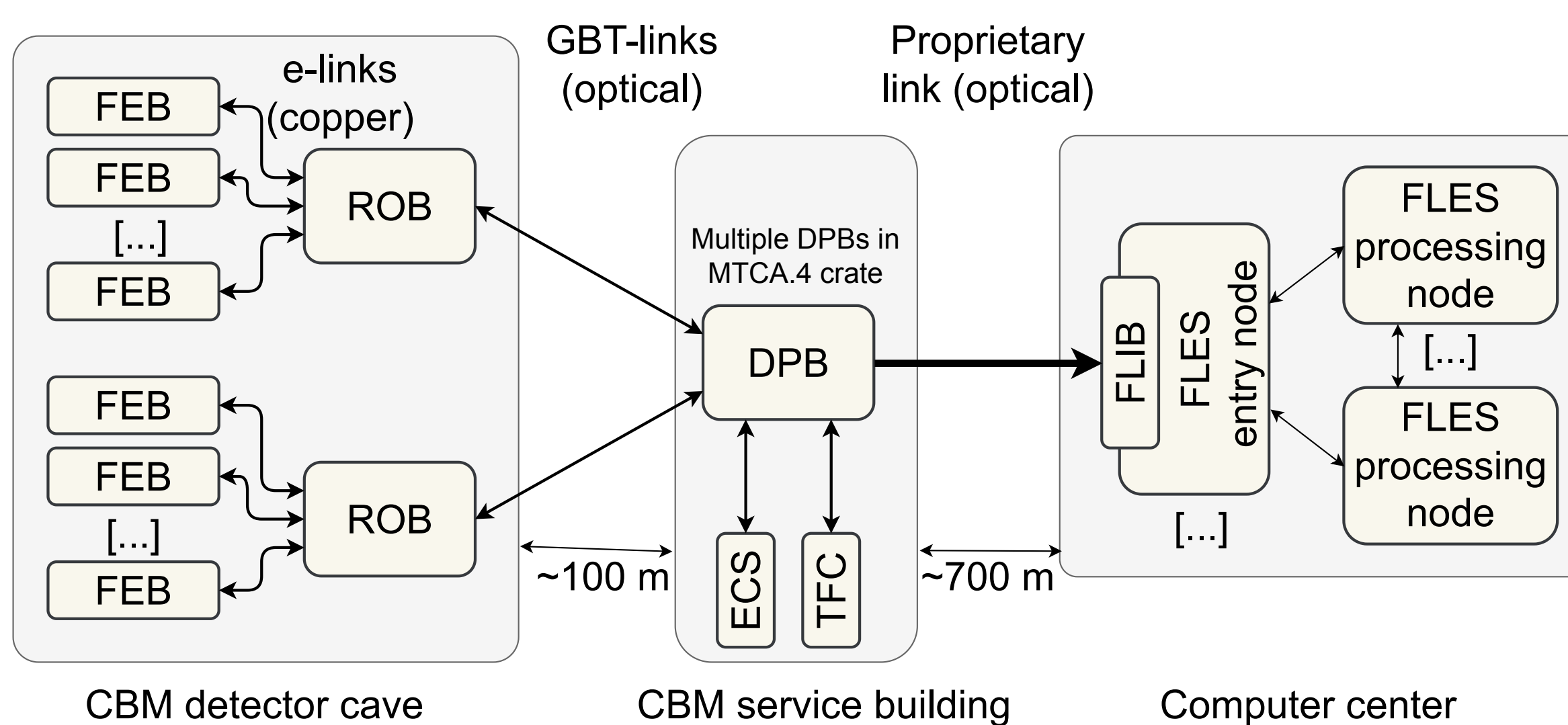
Block diagram of the first prototype of the STS readout chain in the CBM experiment (description in text).

The throughput of the output link (10 Gb/s) was lower than the expected maximum data bandwidth (21.5 Gb/s for 6 GBT links up to 28.7 Gb/s for 8 GBT links). Therefore, a heap sorter was introduced to perfectly sort incoming data according to their timestamp (reconstructed in the TS extender). It enabled context-based data format, leading to a reduction in data volume. Unfortunately, in the beam tests, such a sorter appeared extremely sensitive to overflow caused by fluctuations in the FEE data rate and the data with timestamps corrupted by transmission errors. Implementing the non-local data processing appeared difficult due to too much FPGA resource consumption and, in fact, unnecessary.

The data path with the heap sorter and stream merger. (TS - timestamp, TNC - top node controller, SNC - sorting node controller).

The second concept of the readout

The lack of need for 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, which have been implemented as PCIe boards placed in the FLES entry nodes. That change eliminated the proprietary optical link but required the FLES entry nodes to be moved to the CBM service building and connected via a standard long-distance InfiniBand link to the FLES processing nodes in the Computer Center. The CRI implements GBT links for ROB connectivity and the FLES Interface Module (FLIM) for PCIe. In the first CRI prototype, the measured PCIe bandwidth is higher than the expected maximum input bandwidth, relaxing the requirement for perfect sorting and context-based data aggregation.

Block diagram of the second prototype of the STS readout chain for the CBM experiment.

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. In that approach, the data received from a group of 14 e-links are concentrated, and their TS is extended. Then, the data are divided into bins based on four user-selected bits of the TS (the lower bits are ignored in bin selection). The higher bits define the acceptable range of timestamps ([1] appendix B.1).
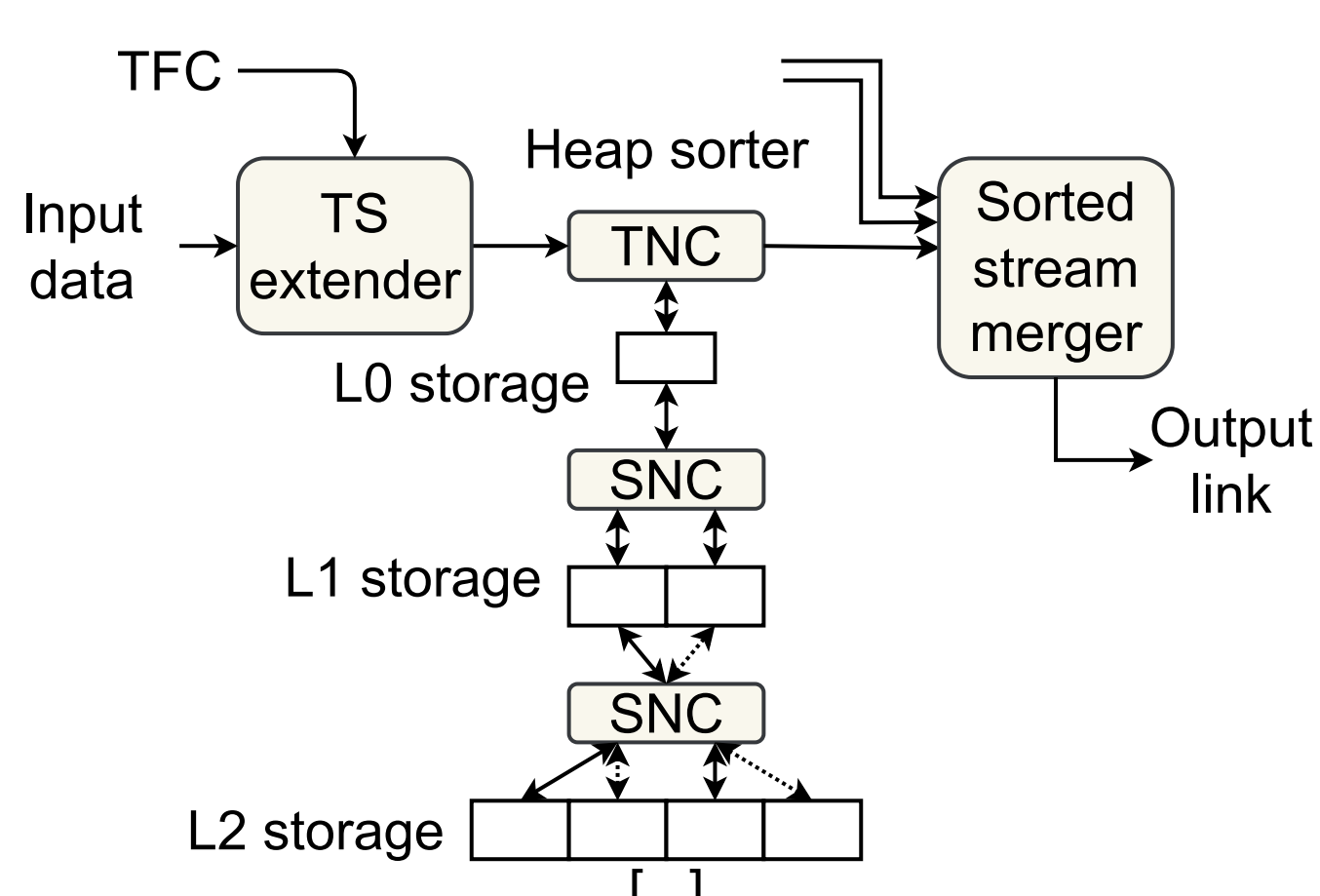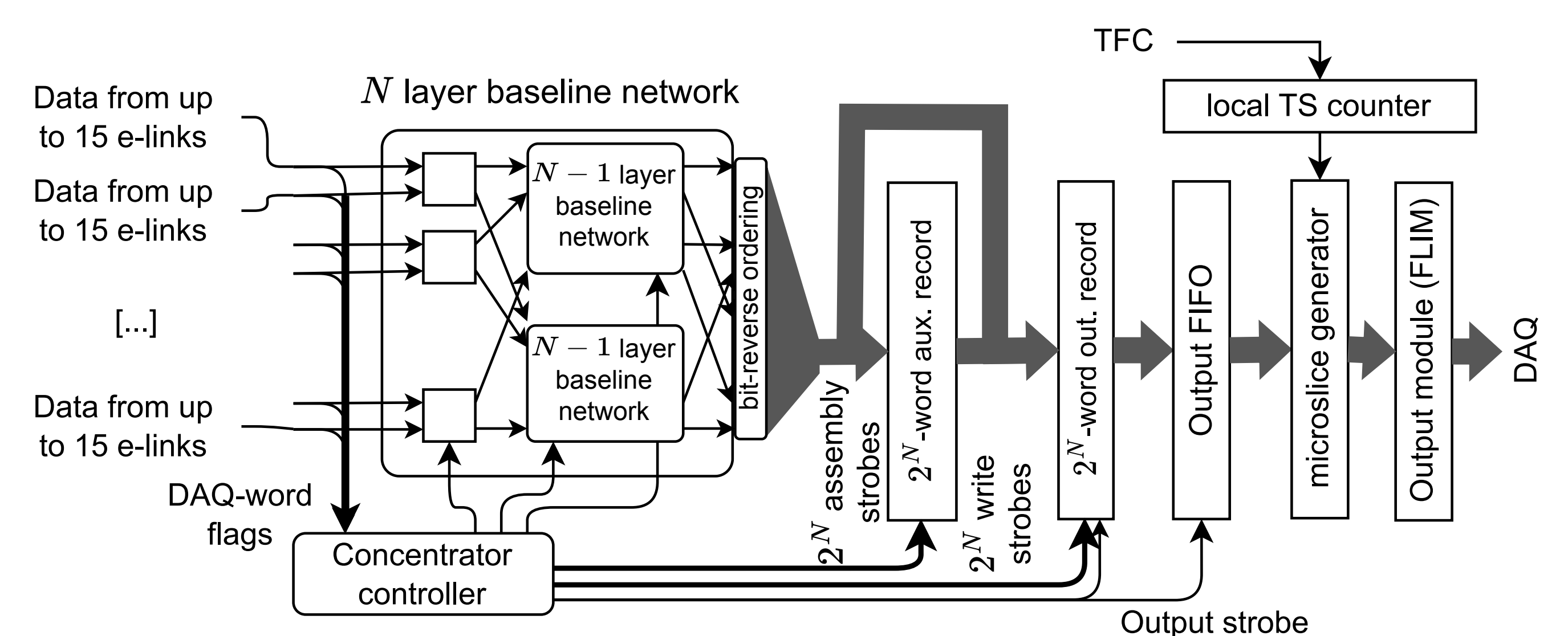

Bucket sorter used in the second version of the CBM DAQ readout. The data in each group of e-links are sorted into 16 bins based on 4 selected bits of their timestamp. The more significant bits decide whether the data are accepted. The hit collector combines the data from the same bins from all groups.

The amount of memory allocated for any single bin is the same. Therefore, this solution better handles intermittent peaks in the hit intensity (the superfluous data are rejected, and the data lost flag is set). However, in the beam tests, for reasonable bin duration (3.2μs) and memory size (1024 words), such data loss occurred too frequently. The data with corrupted timestamps are also rejected with the flag set or stored in an inappropriate bin. There is still a small risk of disturbing the operation of the TS extender.

Aggregation of data with simple concentration

Both described solutions heavily depend on the timestamps contained in received data, which makes them sensitive to data corruption. Both solutions also significantly modify the data stream. In case of problems, reconstructing original data and diagnosing the problem is impossible. Therefore, a special diagnostic version of the FPGA firmware or additional resources for debugging are required.
The last progress in PCIe technology (Gen4 and Gen5) further removes the limits on the transmitted data volume. Based on that, a new concept utilizing the simple concentration of data has been created.
In that solution, 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. The PCIe output module uses a wide 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 [6] has been created to pack such data into wider words without leaving empty places and wasting clock cycles. The boundaries of the microslices are determined by the arrival time of the data, hence eliminating the influence of corrupted data. The original data stream may be fully reconstructed. However, inserting additional words with artificially created timestamps may be needed in case no data is transmitted in a particular e-link for a longer time.
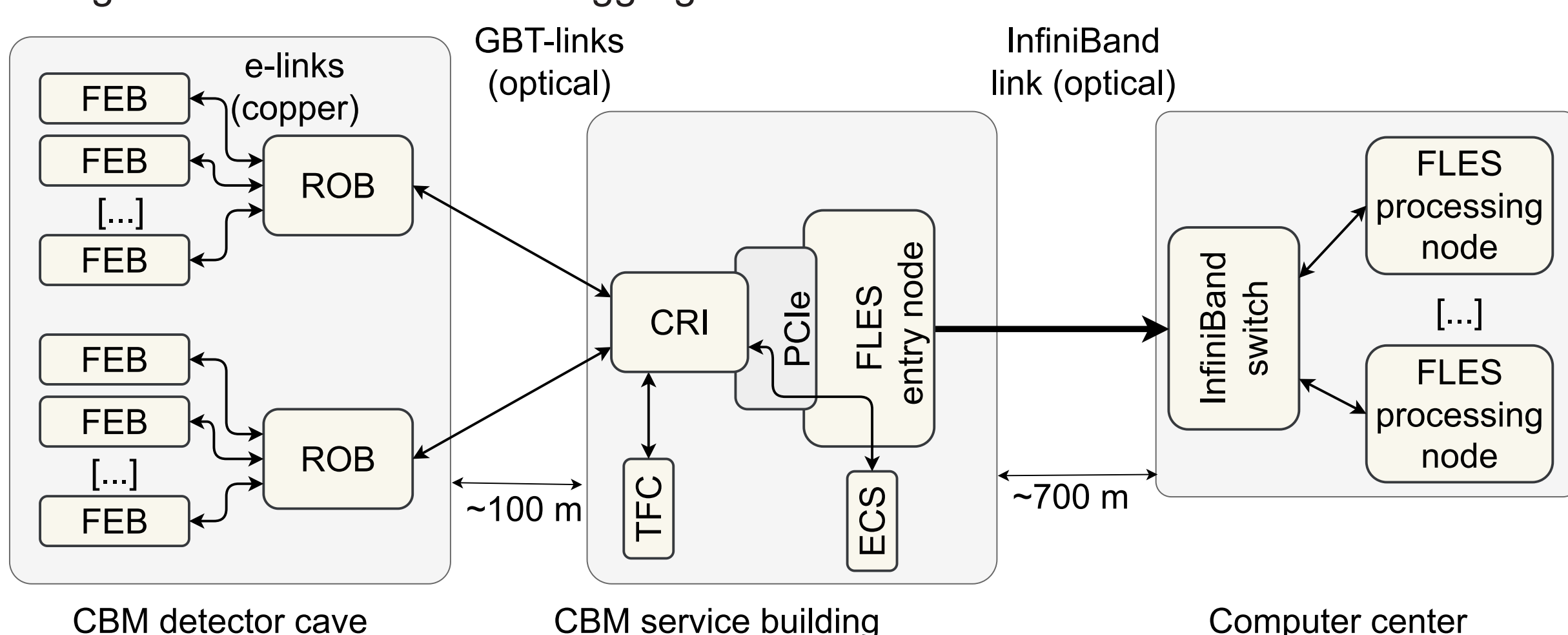That solution has been successfully tested in a GERI board using the 256-bit PCIe word and in a simplified form (with 64-bit output word) in the first prototype of the CRI board.
The tests have confirmed that this aggregation scheme offers the best handling of high FEE data rate.

Block diagram of the system performing the simple concentration of the data in CBM readout

Conclusions

The preparation of the CBM experiment inspired the development and testing of various methods for aggregation 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. It enables the 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. They may be reused in other systems where perfect or partial data sorting at the FPGA level is necessary.

Acknowledgments

References

[1] "Technical Design Report for the CBM Online Systems – Part I, DAQ and FLES Entry Stage", doi:10.15120/GSI-2023-00739
[2] K. Kasiński, R. Szczygiel, et al. "A protocol for hit and control synchronous transfer for the front-end electronics at the CBM experiment", NIMA 835 66, 2016, doi:10.1016/j.nima.2016.08.005
[3] P.Moreira, J. Christiansen et al. "GBTX Manual", https://cds.cern.ch/record/2809057/
[4] J.Lehnert,A.P.Byszuk et al. "GBT based readout in the CBM experiment", JINST 12(02)C02061, 2017, doi:10.1088/1748-0221/12/02/C02061
[5] Dirk Hutter, J. de Cuveland, et al. "CBM First-level Event Selector Input Interface Demonstrator", J. Phys.: Conf. Ser. 898 032047, 2017, doi:10.1088/1742-6596/898/3/032047
[6] Zabołotny, W.M. "Scalable Data Concentrator with Baseline Interconnection Network for Triggerless Data Acquisition Systems", Electronics 2024, 13, 81. doi:10.3390/electronics13010081