Journal of Instrumentation

Time error accumulation in a hierarchical time and clock distribution network with deterministic optical links

To cite this article: V. Sidorenko *et al* 2024 *JINST* **19** C03014

PRIME
PACIFIC RIM MEETING
ON ELECTROCHEMICAL AND SOLID STATE SCIENCE
HONOLULU, HI
Oct 6–11, 2024
Abstract submission deadline:
April 12, 2024
Learn more and submit!
Joint Meeting of
The Electrochemical Society
•
The Electrochemical Society of Japan
•
Korea Electrochemical Society

Published by IOP Publishing for Sissa Medialab

Received: *November 3, 2023*
Accepted: *January 30, 2024*
Published: *March 6, 2024*

Topical Workshop on Electronics for Particle Physics
Geremeas, Sardinia, Italy
1–6 October 2023

Time error accumulation in a hierarchical time and clock distribution network with deterministic optical links

V. Sidorenko,[a,*] W.F.J. Müller,[b] D. Emschermann,[b] W. Zabolotny,[c] I. Fröhlich[b] and J. Becker[a] on behalf of the CBM collaboration

[a] *Karlsruhe Institute of Technology,*
Engesserstraße 5, 76131 Karlsruhe, Germany
[b] *GSI Helmholtz Centre for Heavy Ion Research,*
Planckstraße 1, 64291 Darmstadt, Germany
[c] *Warsaw University of Technology,*
Pl. Politechniki 1, 00-661 Warsaw, Poland

E-mail: vladimir.sidorenko@kit.edu

Abstract: Accurate clock and time distribution is a key requirement for self-triggered streaming data acquisition in the CBM experiment. This distribution is handled by the Timing and Fast Control (TFC) system by clock forwarding and broadcasting the common time over latency-deterministic optical links in a hierarchical FPGA network. The point-to-point optical connections are served by the latency-optimized GBT-FPGA core, which has been developed at CERN. In the presented work, the performance of GBT-FPGA links for time and clock distribution in a scaled TFC system with multiple hops and endpoints has been investigated.

Keywords: Control and monitor systems online; Detector control systems (detector and experiment monitoring and slow-control systems, architecture, hardware, algorithms, databases)

[*] Corresponding author.

https://doi.org/10.1088/1748-0221/19/03/C03014

Contents

1 Introduction

As the requirements for timing systems have been growing together with the scale and measurement rates of the modern physics experiments, some experiments started to adopt the concept of data readout without using hardware triggers. One of such experiments, Compressed Baryonic Matter (CBM), is currently under development and is going to operate at the future Facility for Antiproton and Ion Research (FAIR) in Darmstadt, Germany. The goal of the experiment is to study strongly interacting matter at high baryonic densities, which requires identification of rare and complex decay topologies in an environment with high track density. Given the complex trigger signatures needed to identify the rare probes, and the high interaction rates of 10 MHz foreseen in the experiment, it is impossible to efficiently implement a hardware trigger [1]. For that reason, up to 1 TB/s of the timestamped data, already filtered by the predefined thresholds directly at readout channels, is generated by the self-triggered front-end electronics (FEE) and streamed to the data acquisition (DAQ) system. The DAQ system aggregates the data using the GBTx ASICs and the Common Readout Interface (CRI) PCIe cards equipped with Xilinx FPGAs, and forwards it to a high-performance computing farm, where the First-Level Event Selector (FLES) system performs event reconstruction and selection [2, 3].

Since the timestamps that are assigned to the data in the FEE are the only source of timing information for event reconstruction, it becomes critical to establish a stable common notion of time in the form of both clock and absolute time. In the CBM experiment, the global clock and time are distributed by the Timing and Fast Control (TFC) system to the CRI boards. The CRI boards, in turn, propagate clock and time to the FEE. Based on the configuration of the experiment, it is required that the time between any two nodes in the time distribution network is stable within 200 ps, including between partial and full system restarts [4].

However, given the target scale of the system of 200 CRI endpoints, validation of timing stability does not only require accurate clock skew measurements between nodes, but also appropriate analytical tools to estimate error accumulation in a scaled system in two dimensions: vertically, with added layers in the time distribution tree, and horizontally, with added nodes at the endpoint layer. The aim of the presented work is to establish this analytical framework.

2 TFC overview

The TFC system has a two-fold function in the CBM experiment: it ensures time stability across the experimental setup and provides the means for control message exchange between all system endpoints and the central TFC master with low latency. The latter function is referred to as fast control and is required to protect the DAQ system from congestion caused by the expected beam intensity fluctuations. Based on the prior investigation of potential throttling strategies, the throttling command must arrive to CRI boards within 6 µs from generation of the congestion status message [5]. This latency requirement includes the round-trip message transport and the delay caused by the throttling decision-making logic.

In addition to serving fast control commands, the TFC system provides phase stability of time counters in CRI boards with sub-clock accuracy. The source of the time in the TFC system is the master node. As no time alignment or propagation delay correction is required in the system, the system-wide time is distributed from the central master to CRI endpoints in a one-way fashion without compensating for the propagation delay, although the architecture is compatible with full two-way synchronization as well.

In order to successfully combine both functions and fulfill the scalability requirements at the same time, the TFC system is designed as a hierarchical optical network with syntonization of all nodes at the physical layer. It runs a reference time counter and sets a global clock frequency that can optionally be syntonized to an external frequency. From the master node, the time is continuously transmitted over downstream links to all endpoints at the same time. The purpose of the intermediate layers of so-called submaster node is to ensure scalability by retransmitting the time from one link to multiple downstream connections. The major prerequisite of accurate time distribution with this scheme is low-latency data transport in both upstream and downstream directions, with an additional requirement for fixed latency in the downstream direction, as latency variance in this case is directly connected with the accuracy of time distribution.

To ensure low link latency, using certain components and techniques must be avoided while designing a data transport link. Such components are, for example, FIFOs, queues, etc. As long data frames or individual node addressing are not required in the TFC design, general-purpose protocols, such as the Ethernet stack, would introduce unnecessary functional overhead, contributing to link latency. The GBT-FPGA core has been initially developed targeting the latency-critical application of handling the FPGA communication with the GBTx ASICs, and it features optional latency-deterministic datapaths. As the core is an HDL implementation of the symmetric GBT protocol, it can also be used for communication between two FPGA devices [6]. Combined with a clocking scheme optimization, GBT-FPGA demonstrates sufficiently low and deterministic link latency [7]. This core is used to serve the links between TFC nodes, with the resulting architecture of the time distribution chain shown in figure 1.

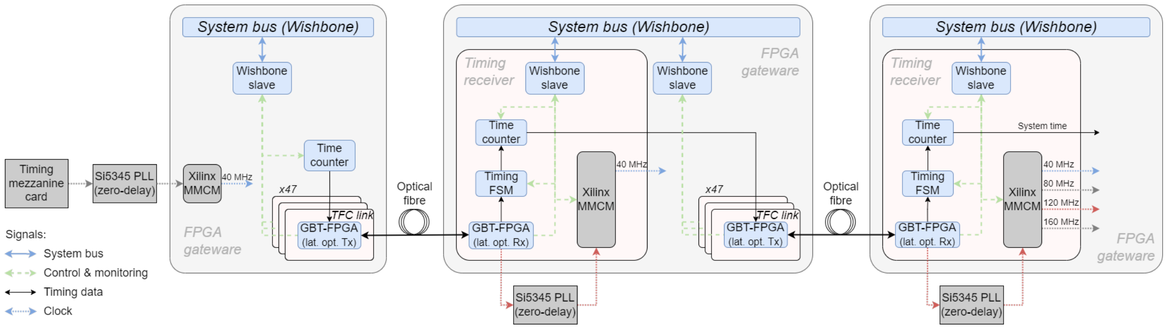

Figure 1. Architecture of the clock distribution chain in CBM.

3 Time error accumulation

In order to evaluate scalability of a timing system and estimate the accuracy of timing distribution in the final setup, it is crucial to have a clear unterstanding of how time error accumulates with added system complexity. As a hierarchical network, the TFC system can grow in two dimensions. On one hand, layers can be added to the hierarchy, thus extending a synchronization chain between the master node and each individual endpoint by an additional hop. In the context of this paper, this is called vertical time error accumulation. On the other hand, more nodes can be added in the same layer, making the system grow horizontally. As all clocks local to the TFC nodes are syntonized during normal operation, the current analysis is based on the model developed by the International Telecommunication Union for accumulation of time error in a chain of clocks [8]. The model of the vertical time error accumulation over N hops is expressed with equation (3.1) in terms of maximum absolute time error.

$$\max|TE_N(i)| \leq \sum_{i=1}^{N}|cTE_i| + \sum_{j=1}^{N-1}|\mathrm{link}TE_j| + \sqrt{\left\{\sum_{i=1}^{N}\left[\max|dTE_{L,i}(t)|\right]^2\right\} + \left[\max|dTE_{H,N}(t)|\right]^2} \tag{3.1}$$

In equation (3.1), cTE_i is the so-called constant time error. This error is immune to any filtering by the clock distribution chain and is accumulated linearly. It consists in the TFC system mainly of reset-to-reset phase jumps and the link propagation delay. In the current implementation, the reset-to-reset stability is affected by the known effects that have yet to be addressed [9]. For this reason, the focus of the current analysis is placed on the dynamic time error, $dTE(t)$. This error represents jitter and wander in the system and is composed of two factors: $dTE(t) = dTE_L(t) + dTE_H(t)$. The former, low-band, component is propagated through the chain with contribution from each intermediate node and is defined by the low-frequency variance of the clock skew between subsequent nodes. The latter, high-band, component $dTE_H(t)$ is filtered out by the PLLs in every node of the clock propagation chain, and is defined mainly by the local clock jitter in the last node in the chain. $linkTE_j$ represents a constant time synchronization error due to link latency asymmetry. As link latency asymmetry plays no role in one-way time distribution, and given the other considerations above, time error estimation can be simplified and expressed as follows:

$$\max|TE_N(i)| \leq \sqrt{\left\{\sum_{i=1}^{N}\left[\max|dTE_{L,i}(t)|\right]^2\right\} + \left[\max|dTE_{H,N}(t)|\right]^2} \tag{3.2}$$

As the TFC system is being designed with latency-deterministic downstream data transport, dynamic time errors can be considered consistent in this direction, with $\max|dTE(t)| = dTE(t)$. Besides, peak-to-peak value is commonly approximated in Gaussian variables with relation to standard deviation with $X_{p-p} = n\sigma$, with n being an arbitrarily selected approximation factor and σ being the standard deviation of the random variable. Assuming that all dynamic time errors in the system are normally distributed, peak-to-peak values are proportional to standard deviation, equation (3.1) can also be expressed in terms of standard deviations.

Horizontally, time error accumulates in the system with relative time error between any nodes in the same network layer. This error is connected with the absolute time error of these nodes and can be described with equation (3.3), given that the nodes obtain their frequency and phase synchronization from the same source. In case of the TFC system, this common source is the master node. Similarly to (3.1), horizontal error accumulation can be expressed in terms of peak-to-peak values and standard deviation.

$$TE_{xy}(i) = TE_x(i) - TE_y(i) \tag{3.3}$$

4 Model validation

To validate the model for estimating time error accumulation in the full-scale TFC system, clock distribution networks have been implemented in two minimal network configurations shown in figure 2. Serial configuration features a single clock distribution chain with two hops, allowing to study vertical time error accumulation. The second, parallel, configuration is meant to provide insight into how the time error accumulates horizontally. The parallel configuration features one time source, the master node, with two connected endpoints. In both configurations, all nodes are based on the BNL-712 platform, the same FPGA cards as will be used in the final setup, with the full-featured TFC gateware [10].

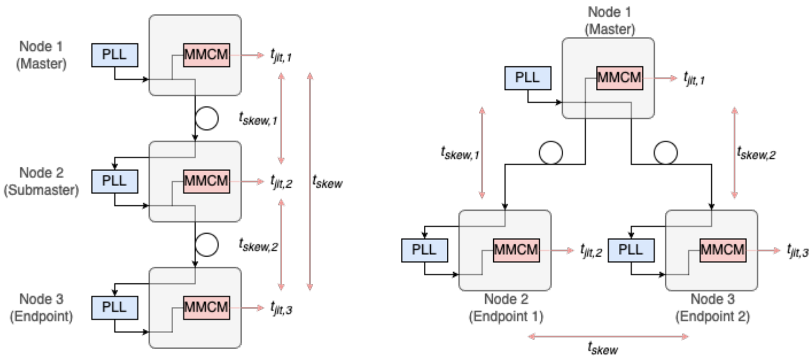

Figure 2. Clock distribution configurations for model validation: serial (left) and parallel (right).

In both cases, two major parameters have been measured in the tests: local clock jitter on each node, as it represents the high-band dynamic error $dTE_H(t)$, and relative clock skew between all nodes, which is meant to provide a value for the low-band dynamic error $dTE_L(t)$. Automated tests with extensive statistics require adjustments in the gateware and are planned in the near future. In the current work, focusing on measurement precision, three manual measurements of the standard deviation of the above parameters have been performed on each configuration using a Tektronix TDS6154C oscilloscope with the TDSJIT3 advanced jitter measurement application. Each parameter was measured with a sample size of at least 1 million. Before measuring the parameters in each test, a full system restart has been performed, including power cycling the FPGA boards. The clocks under test are the 40 MHz clocks derived with a Xilinx Mixed-Mode Clock Manager, the same way as the common system clock is produced in all subsystems of the experiment. The results of the measurements for the relevant parameters are presented in tables 1 and 2.

For comparison with the measurement results, $TE_N(i)$ in table 1 and $TE_{xy}(i)$ in table 2 have been computed using equations (3.2) and (3.3), respectively. To compute these values, $t_{jit,i}$ measurements have been used as $dTE_{H,i}$ and $t_{\text{skew},i}$ as $dTE_{L,i}$. As can be seen in table 1, there is a visible correlation between estimated time error accumulated over 2 hops and the measured clock skew, although a consistent bias can also be observed. This demonstrates general applicability of the model, with the bias being composed of minor contributors, that have yet to be investigated. The measured results in

Table 1. Measurement results for the serial configuration.

Test no.	$t_{jit,3}$, ps	$t_{\text{skew},1}$, ps	$t_{\text{skew},2}$, ps	$\boldsymbol{t_{\text{skew}}}$, ps	$\boldsymbol{TE_N(i)}$, ps
1	12.684	24.377	27.840	38.029	39.118
2	10.182	21.311	25.955	33.191	35.093
3	10.527	21.527	30.736	37.407	38.973

Table 2. Measurement results for the parallel configuration.

Test no.	$t_{jit,2}$, ps	$t_{jit,3}$, ps	$t_{\text{skew},1}$, ps	$t_{\text{skew},2}$, ps	$\boldsymbol{t_{\text{skew}}}$, ps	$\boldsymbol{TE_{xy}(i)}$, ps
1	12.997	8.538	24.593	31.396	30.869	4.720
2	14.497	10.400	24.102	31.948	31.482	5.472
3	14.411	7.836	24.945	34.212	34.050	6.289

the parallel configuration, however are not consistent with the estimated values by far. This indicates that there are major contributors to the horizontal time error accumulation that are not considered in the model and that the relative time error model cannot be used as-is and requires further research.

5 Conclusion

With the TFC system being the central source of time information in the streaming readout concept of the CBM experiment, it is critical to ensure that the sufficient time stability can be achieved in the final setup with at least 200 CRI boards. While direct measurements of time stability can be performed on minimal lab setups, a reliable model is necessary to estimate how time error will change as the time distribution system grows. For this purpose, a time error accumulation model developed by the ITU has been analysed and validated in application to the time error accumulation in the TFC system. While the model for vertical error scaling is generally applicable, although there is potential for adjustments, the model of horizontal time error accumulation requires further major analysis.

Acknowledgments

The project on which this report is based was funded by the German Federal Ministry of Education and Research under the funding code 05P21VKFC1. The responsibility for the content of this publication lies with the author. This work was also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 824093 (STRONG-2020).

References

[1] V. Friese, *The high-rate data challenge: computing for the CBM experiment*, *J. Phys. Conf. Ser.* **898** (2017) 112003.

[2] P. Moreira et al., *The GBT Project*, in the proceedings of the *Topical Workshop on Electronics for Particle Physics*, Paris, France (2009) [DOI:10.5170/CERN-2009-006.342].

2024 JINST 19 C03014

[3] J. de Cuveland and V. Lindenstruth, *A first-level event selector for the CBM experiment at FAIR*, *J. Phys. Conf. Ser.* **331** (2011) 022006.

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

[5] X. Gao, D. Emschermann, J. Lehnert and W.F.J. Müller, *Throttling strategies and optimization of the trigger-less streaming DAQ system in the CBM experiment*, *Nucl. Instrum. Meth. A* **978** (2020) 164442.

[6] M. Barros Marin et al., *The GBT-FPGA core: features and challenges*, 2015 *JINST* **10** C03021.

[7] V. Sidorenko et al., *Evaluation of GBT-FPGA for timing and fast control in CBM experiment*, 2023 *JINST* **18** C02052.

[8] ITU-T Recommendation G.8271.1/Y.1366.1 *Network limits for time synchronization in packet networks with full timing support from the network*, (2022), https://handle.itu.int/11.1002/1000/15130.

[9] E. Mendes et al., *Achieving Picosecond-Level Phase Stability in Timing Distribution Systems With Xilinx Ultrascale Transceivers*, *IEEE Trans. Nucl. Sci.* **67** (2020) 473.

[10] K. Chen et al., *A Generic High Bandwidth Data Acquisition Card for Physics Experiments*, *IEEE Trans. Instrum. Measur.* **69** (2019) 4569.