EPJ Web of Conferences **245**, 01036 (2020) https://doi.org/10.1051/epjconf/202024501036
CHEP 2019

Low Latency, Online Processing of the High-Bandwidth Bunch-By-Bunch Observation Data From the Transverse Feedback System in the LHC

Martin Söderén[1,*] and *Daniel* Valuch[1]

[1]CERN, Geneva, Switzerland

Abstract. During long shutdown 2 (2019-2020) the transverse observation system (ADTObsBox) in the LHC will undergo a substantial upgrade. The purpose of this upgrade is to allow for true low latency, online processing of the 16 data-streams of transverse bunch-by-bunch, turn-by-turn positional data provided by the transverse feedback system in the LHC (ADT). This system makes both offline and online analysis of the data possible, where the emphasis will lie on online analysis, something that the older generation was not designed to provide. The result of the analysis is made available for accelerator physicists, machine operators, and engineers working with LHC. The new system allows users to capture buffers of various lengths for later analysis just like the older generation and it provides a platform for real-time analysis applications to directly capture the data with minimal latency while also providing a heterogeneous computing platform where the applications can utilize CPUs, GPUs and dedicated FPGAs. The analysis applications include bunch-by-bunch instability analysis and passive bunch-by-bunch tune extraction to name a few. The ADTObsBox system uses commodity server technology in combination with FPGA-based PCIe I/O cards. This paper will cover the design and status of the I/O cards, server, firmware, driver, analysis applications and results of early performance testing.

1 Introduction

In 2015, a system called ObsBox (Observation Box) [1] was introduced by the Radio-Frequency group at CERN which allowed for buffering of multiple high-bandwidth data-streams from the Low-Level RF systems. The system was designed to be general purpose with the first application in the LHC for both the transverse [1, 2] and longitudinal plane [3]. Later it was deployed in the SPS for the transverse plane. There are plans to deploy it in the AD [4] and ELENA machines for the longitudinal Schottky measurements. The main purpose of the system for the transverse plane in the LHC is to make buffers with beam data (e.g. a bunch-by-bunch transverse position) of different lengths available for users. The buffers are ranging from 2^{12} turns to analyze injection oscillation transients, to 2^{17} turns to analyze transverse beam oscillations caused by civil engineering works close to the LHC beam tunnel [5]. Over time, the system has evolved into an important tool providing live beam parameter and transverse stability data to the accelerator operation, and it is an absolutely vital tool

*e-mail: martin.soderen@cern.ch

for the machine development sessions, where new ideas or methods are being tested in the machines. The ObsBox machines, which analyze the data in the transverse plane from the ADT are specifically called "ADTObsBox". The ADTObsBoxes receive the data from the 16 ADT Beam Position Monitor (BPM) VME modules, each with a link speed of 1 Gbit/s and a data transfer rate of 0.8 Gbit/s. After decoding, this results in a combined rate of 1.28 GB/s.

They have been a proving-ground to test the limits of its computing system, since the transverse plane is where the analysis has gradually moved towards an online (<1 second, or 11k turns latency), or after the upgrade a real time analysis (few turns, or a milli-second latency). In 2016, an online transverse instability detection system [6] was introduced that analyzed the beam positional data for exponential oscillation amplitude growths to detect an onset of transverse instability (all analysis performed bunch-by-bunch). In 2017, a system for low-frequency spectrum analysis on large data sets was introduced with the purpose of analyzing the correlation between the ground motion and transverse bunch oscillation [5]. In the same year, a system for online compression of the data was introduced with aim to compress and save all triggered buffers and store them on an NFS server for further offline analysis. One new functionality, that is proposed for the LHC restart in 2021 is a passive bunch-by-bunch tune extraction application which will combine data from multiple pickups and perform spectrum analysis on each bunch.

2 Comparison to the data generated by the experiments

The amount of data analyzed by the ADTObsBox is on the same scale as the data from the high level trigger (HLT) in the four major experiments in the LHC, a comparison can be seen in Table 1. It should be noted that an event in the experiments is a collision and an event for the ADTObsBox is one revolution. Each LHC experiment has ≈ 2000 processing nodes while the ADTObsBox has only three powerful and well-configured processing nodes. In terms of storage, the ADTObsBox has 144 TB of temporary local storage and an additional 144 TB on a dedicated storage server which stores relevant data to be used for further offline analysis. The data stored on the dedicated storage server will be archived long term with backups done regularly to CASTOR [7]. The combined storage capacity is roughly one thousandth of the storage capacity of the CERN computing center [8].

Table 1. Comparison of data generated by the ADTObsBox and the experiments at CERN [9]

	Level-1 Rate (Hz)	Event size (Bytes)	Readout Bandw. (GB/s)	HLT Out MB/s (Events/s)
ALICE (Pb-Pb)	500	5×10^7	25	1250 (10^2)
ALICE (p-p)	10^3	2×10^6	25	200 (10^2)
ATLAS	10^5	1.5×10^6	50	≈ 1000(10^2)
CMS	10^5	10^6	100	≈ 1000(10^2)
LHCb	10^6	5×10^4	50	700 (1.2×10^4)
ADTObsBox	10^4	10^5	2	1280 (10^4)

3 Requirements

- Latency from the time when a complete frame has been received until it is available for analysis on the host should be a few µs
- The acquisition card must aggregate multiple data-streams received from fibres into a single data-stream that can be transferred to the host

- The acquisition card must perform all necessary pre-processing steps so the readers in the user-space can read the data directly without any pre-processing required on the host
- The driver must support multiple concurrent readers
- The driver must be able to acquire and maintain a large circular buffer 32 GB (and more)
- The system must not miss any received data
- The system must be able to receive data-streams with a bandwidth of up to 2.5 Gbit/s
- The data rate should be configurable by the driver

4 Hardware

The I/O card selected for the upgrade is a PCIe based TEC0330-4 [10] from Trenz Electronics which features a Xilinx Virtex-7 XC7VX330T-2FFG1157C FPGA and a VITA 57.1 compliance HPC FMC connector with 10 MGT pairs. It also features a 8-lane PCIe Gen3 interface which allows for 7.88 GB/s to the host system, this would allow for up to ten channels at 5 Gbit/s, more than specified. This was paired with a HTG-FMC-X10SFP+ [11] mezzanine module from HiTech Global housing 10 SFP+ ports and a TI LMK61E2 high-performance low-jitter programmable oscillator to generate the reference clock for the GTH transceivers. Both cards can be seen assembled in Fig. 1.

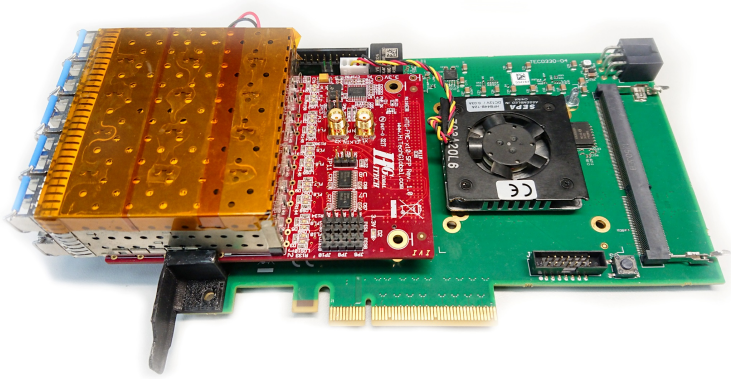

Figure 1. The Trenz Electronics TEC0330-4 coupled with the HiTech Global HTG-FMC-X10SFP+

The server model chosen for the upgrade is the Gigabyte G481-HA0 [12], depicted in Fig. 2. The reasoning is, that these servers can fit ten 16-lane PCIe gen3 full-length, full-height and double width cards. This is enough for dual acquisition cards, timing card, white-rabbit interface, GPUs and potential upgrades. It is planned to install four NVIDIA RTX 2080 TI in one of the new servers for performing the bunch-by-bunch tune extraction. Another reason was that it can accommodate 12 3.5" SATA or SAS disks which will be used for a long circular buffer.

5 Firmware

The duties of the I/O card are to receive up to 10 1 Gbit/s 8B/10B encoded fiber-optic serial data-streams sent from the BPM VMEs in form of data frames containing the transverse

Figure 2. GIGABYTE G481-HA0 with I/O card and connected fibers

bunch-by-bunch position data of all bunches in the LHC for one revolution of the beam. These frames are then pre-processed to prepare them for the host to analyze them. The low level pre-processing includes:

- Finding the start of the frame
- Alignment of the data
- Removal of control characters
- Extraction of triggers and signalling from the frame
- Calculate and control the checksum
- Swapping endianness
- Adding metadata about any faults that are associated with each frame

Each receiver has a pipeline that performs these tasks and an associated FIFO. These FIFOs are read by an IP block in charge of a triple buffer to which it writes the data from the pipelines.

When a pipeline signals the start of a new frame, the IP block starts to write the data from that pipeline into the next buffer. When all currently active pipelines have signaled the start of a new frame, the BufferController swaps to the next buffer and initializes a transfer of the filled buffer to the host server. The transfer is done by the Xilinx XDMA engine in streaming mode with the bypass descriptor control enabled. The bypass descriptor control allows the logic on the FPGA to read a DMA address from a buffer which is continuously filled by the driver and setup a transfer by itself. This means that the I/O card can work completely asynchronous from the driver (in polling mode) and still transfer data even if the driver is temporarily busy.

When a transfer is completed, the driver is notified either by an interrupt or by polling the write-back address which is incremented every time a transfer is completed.

6 Driver

The purpose of the driver is to initialize the card, make sure that the I/O card has DMA addresses to which it can transfer the data and to notify the readers when new data is available. To do this, the driver needs to allocate a large buffer of continuous physical memory in the range of 32 GB and upwards. The safest way to do this is to reserve the memory at boot so it is never a part of the memory pool. This involves passing the memmap argument to the kernel and then setup the appropriate kernel page-table mapping using the ioremap function. The driver supports either using pre-allocated memory or dynamic allocation of memory but the dynamic allocation uses dma_alloc_coherent and it is only guaranteed to work for a few MBs.

Notification about a completed transfer can be done either by an interrupt or from the driver polling the writeback from the card. A comparison between the latency of the different methods was performed on a real-time system running Centos7 3.10.0-1062.rt56.1022.el7.x86_64. The latency was tested by having the driver toggle a register in the I/O card every time a new transfer was completed and then using the Xilinx integrated logic analyzer (ILA) to measure the time between receiving the start of a frame and the time until the register was toggled. The polling has an average latency of 1.4 µs less than when using interrupts but at the cost of using one core at 100% compared to 7% when using interrupts. Receiving all frames takes 89 µs (as long as they are synchronous) and sending them to the host takes ≈ 30 µs.

7 Testing and Development

The most complex parts of the system are the firmware and the driver, so the majority of tests were performed on these parts. Since the development mostly took place during LS2, there was no real machine data to test the system with. Instead a firmware of the new generation Beam Position Module was modified to generate simulated data from four pickups (including configurable bunch filling patterns, or uncorrelated measurement noise background). The development strategy for the firmware was incremental development and testing. This was well suited for this project since the processing chain is more or less one long pipeline. By starting with the implementation of the first element in the pipeline and then testing that it always outputs the correct results by using internal observation memories, then adding one more element while applying the same procedure can significantly reduce the development and debugging time.

After verifying that the firmware and the driver can receive the data correctly and that it recovers correctly from corrupt headers or from disconnection and re-connection of the fibre links, a reliability test was performed. This was done by having one analysis application per link subscribing to the data and checking so the turn counter in the header is incremented by one every time the new data is available. To simulate a server under load, the rest of the available cores were running stress-tests so all cores were running (at 100%). This configuration was running for one week and the result was on average one missed turn per 20 minutes $\approx 7.4 \times 10^{-8}$ of all turns. This test was performed once again but instead of triggering the user applications at every turn, they were triggered every other turn and received data for two turns, this lead to zero missed turns over the time of one week. The conclusion from this is that the data is received correctly and it is in the circular buffer but notifying the user application at a rate of 11 245 Hz is on the limit of what is acceptable. There is potential for improvements by more tuning of the system with IRQ and core affinity and also testing different notifications schemes instead of the current one which relies on spinlocks and busy-waiting.

8 Example Application

The new system will reduce the processing latency from 364 ms to 120 µs which will make the ADTObsBox an even more powerful tool for transverse instability detection in the LHC. With the long processing delay, it can be too late for some fast rise time instabilities to be detected. If the beam is dumped due to losses on the collimators then it would be captured by the post-mortem buffers but if not then the data could be lost. The current transverse instability detection implementation applies the Hilbert transform [13] to the bunch-by-bunch data. An instantaneous oscillation amplitude for each bunch is calculated. A growth can be detected by comparing the running average of the instantaneous amplitude for different time windows. A time window of 2^{10} turns can be seen in Fig. 3. When an instability is detected, dedicated observation buffers are triggered and saved to a storage server for further analysis. At the same time all data from the instability analysis, for example the bunch-by-bunch transverse activity is logged by the accelerator logging service (NXCALS) [14].

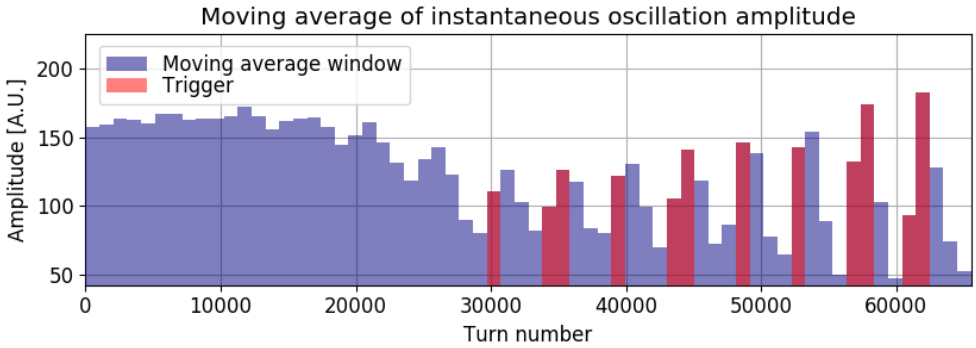

Figure 3. Moving average of instantaneous oscillation amplitude

9 Conclusion

ADTObsBox is a powerful computer system that interfaces the data-links transferring the bunch position data within the LHC transverse feedback with the users in forms of buffers. At the same time, the system provides a platform for online, low-latency beam data analysis. The transverse motion of all bunches in the LHC is analyzed and crucial beam parameters are extracted and distributed to the operators, or other equipment around the LHC. It is designed to be a general purpose system that can be coupled to any RF devices to provide a platform for analysis. The result of the integration and performance tests with the ADTObsBox to date indicate that it is on track to be ready for deployment in 2021.

References

[1] M. Ojeda Sandonís, P. Baudrenghien, A.C. Butterworth, J. Galindo, W. Hofle, T.E Levens, J.C. Molendijk, and D. Valuch, *Processing High-Bandwidth Bunch-by-Bunch Observation Data from the RF and Transverse Damper Systems of the LHC*, in *Proc. 15th Int. Conf. on Accelerator and Large Experimental Physics Control Systems (ICALEPCS'15)*, Melbourne, Australia, Oct. 2015, pp. 841–844. doi:10.18429/JACoW-ICALEPCS2015-WEPGF062

[2] L.R. Carver, X. Buffat, A. Butterworth, W. Höfle, G. Iadarola, G. Kotzian, K. Li, E. Métral, M. Ojeda-Sandonís, M.E Söderén, and D. Valuch, *Usage of the Transverse Damper Observation Box for High Sampling Rate Transverse Position Data in the LHC*, in *Proc. 8th Int. Particle Accelerator Conf. (IPAC'17)*, Copenhagen, Denmark, May 2017, pp. 389–392. doi:10.18429/JACoW-IPAC2017-MOPAB113

[3] J.F. Esteban Müller, *Longitudinal intensity effects in the CERN Large Hadron Collider*, Ph.D. Thesis, Ecole Polytechnique, Lausanne, 2016

[4] M.E. Angoletta, S.C.P. Albright, M. Jaussi, A. Findlay, V.R. Myklebust, and J.C. Molendijk, *A New Digital Low-Level RF and Longitudinal Diagnostic System for CERN's AD*, in *Proc. 10th Int. Particle Accelerator Conf. (IPAC'19)*, Melbourne, Australia, May 2019, pp. 3966–3969. doi:10.18429/JACoW-IPAC2019-THPRB070

[5] M. Schaumann, D. Gamba, M. Guinchard, L. Scislo, and J. Wenninger, *Effect of Ground Motion Introduced by HL-LHC CE Work on LHC Beam Operation*, in *Proc. 10th Int. Particle Accelerator Conf. (IPAC'19)*, Melbourne, Australia, May 2019, pp. 4092–4095. doi:10.18429/JACoW-IPAC2019-THPRB116

[6] M.E Söderén, *Online Transverse Beam Instability Detection in the LHC. High Throughput Real-Time Parallel Data Analysis*, Master's thesis, Linköpings Universitet, Linköping, 2017, CERN-THESIS-2017-401;LIU-IDA/LITH-EX-A--17/053--SE

[7] G.L. Presti, O. Barring, A. Earl, R.M. Garcia Rioja, S. Ponce, G. Taurelli, D. Waldron, and M.C. Dos Santos, *CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN*, in *24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007)*, San Diego, USA, Sept. 2007, pp. 275–280.

[8] "Key Facts and Figures – CERN Data Centre", http://information-technology.web.cern.ch/sites/information-technology.web.cern.ch/files/CERNDataCentre_KeyInformation_02March2018V1.pdf

[9] T. Colombo, *DAQ –Filtering Data from 1 PB/s to 600 MB/s* in *presented at CERN openlab Summer Student programme 2019, CERN, Geneva, Switzerland, July 2019*, 2019

[10] "Virtex-7 PCIe FMC Carrier, 8 Lane PCIe GEN2", https://shop.trenz-electronic.de/en/TEC0330-04-Virtex-7-PCIe-FMC-Carrier-8 -Lane-PCIe-GEN2

[11] "10-Port SFP+ (10G) FMC Module (Vita57.1)", http://www.hitechglobal.com/FMCModules/x10SFP+_FMC_Module.htm

[12] "G481-HA0 (rev. 200)", https://www.gigabyte.com/High-Performance-Computing-System/G481-HA0-rev-200

[13] Y. Liu, *Hilbert Transform and Applications*, in Fourier Transform Applications, S. Salih, Rijeka, Croatia: InTech, 2012, pp. 291–300.

[14] J. P. Wozniak and C. Roderick, *NXCALS - Architecture and Challenges of the Next CERN Accelerator Logging Service*, presented at the 17th Int. Conf. on Accelerator and Large Experimental Physics Control Systems (ICALEPCS'19), New York, NY, USA, Oct. 2019, paper WEPHA163.