The Compact Muon Solenoid Experiment
Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

26 October 2015 (v4, 25 January 2016)

Algorithm and implementation of muon trigger and data transmission system for barrel-endcap overlap region of the CMS detector

Wojciech Zabolotny^{1,2)}, Adrian Byszuk^{1,2)}, on behalf of the CMS collaboration

Abstract

The CMS experiment is currently undergoing upgrade of its trigger, including the Level-1 muon trigger. In the barrel-endcap transition region it is necessary to merge data from 3 types of detectors - RPC, DT and CSC. The Overlap Muon Track Finder (OMTF) uses the novelty approach to concentrate and process those data in an uniform manner. The paper presents the algorithm and FPGA firmware implementation of the OMTF and its data transmission system in CMS. The OMTF is subject to significant changes during optimizations based on physical simulations. Therefore a special, high level, parametrized HDL implementation is necessary.

Presented at *TWEPP 2015 TWEPP 2015 - Topical Workshop on Electronics for Particle Physics*

¹⁾ Institute of Electronic Systems, Warsaw University of Technology, Warszawa, Poland

²⁾ University of Warsaw, Faculty of Physics, Warszawa, Poland

Algorithm and implementation of muon trigger and data transmission system for barrel-endcap overlap region of the CMS detector

Wojciech M. Zabolotny^{a,b*} and Adrian Byszuk^{a,b}, on behalf of the CMS collaboration

^aInstitute of Electronic Systems, Warsaw University of Technology

ul. Nowowiejska 15/19, 00-665 Warszawa, Poland

E-mail: wzab@ise.pw.edu.pl

^bUniversity of Warsaw, Faculty of Physics

ul. Pasteura 5, 02-093 Warszawa, Poland

ABSTRACT: The CMS experiment Level-1 trigger system is undergoing an upgrade. In the barrel-endcap transition region, it is necessary to merge data from 3 types of muon detectors - RPC, DT and CSC. The Overlap Muon Track Finder (OMTF) uses the novel approach to concentrate and process those data in a uniform manner to identify muons and their transversal momentum. The paper presents the algorithm and FPGA firmware implementation of the OMTF and its data transmission system in CMS. It is foreseen that the OMTF will be subject to significant changes resulting from optimization which will be done with the aid of physics simulations. Therefore, a special, high-level, parameterized HDL implementation is necessary.

KEYWORDS: Digital electronic circuits, Trigger concepts and systems, Trigger algorithms.

*Corresponding author.

Contents

1. Introduction	1
2. Algorithm of the trigger	2
3. Hardware Platform	4
4. Data transmission and preprocessing	5
5. Implementation of the algorithm	5
6. Results & Conclusions	8

1. Introduction

Compact Muon Solenoid [1] (CMS) is one of the experiments at the Large Hadron Collider (LHC) at CERN. The CMS was designed to study proton-proton (and lead-lead or proton-lead) collisions at a centre-of-mass energy of 14 TeV (5.5 TeV nucleon-nucleon) and at luminosities up to $10^{34} \text{cm}^2 \text{s}^{-1}$ ($10^{27} \text{cm}^2 \text{s}^{-1}$ in the case of lead-lead collisions). During the Run-1, the LHC delivered about 30fb^{-1} of proton-proton data at a center-of-mass energy $\sqrt{s} = 7 \text{TeV}$ in 2010 and 2011, and at $\sqrt{s} = 8 \text{TeV}$ until the end of 2012 (see for example [2]). The collected data allowed physicists to perform a variety of measurements and searches and led to the discovery of a Higgs particle [3].

After the end of the LHC Run-1, the CMS detector is undergoing the upgrade of its trigger [4], including the Level-1 muon trigger [5]. In the barrel-endcap transition region (see Figure 1, based on a figure published in [6]) it is possible to combine signals from 3 types of muon detectors - RPC, DT and CSC. The Overlap Muon Track Finder (OMTF) [7] is a dedicated electronic system, that:

- Analyzes the data received from those detectors from every crossing of the LHC proton (heavy ion) bunches. Collisions occur at every bunch crossing of the LHC which is every 25 ns. However not all LHC bunches are always filled.
- Identifies tracks of muons, and estimates their transversal momentum p_T . Bunch crossings in which there is more than one muon are possible. Therefore, the system must be able to identify and deliver up to 3 muon candidates from each bunch crossing.
- Transfers the muon candidates to the CMS Level-1 Global Muon Trigger.

Figure 1. Quarter longitudinal schematic view of the CMS detector (based on a figure published in [6]). The muon detectors (RPC - denoted as RB and RE, CSC - denoted as ME and DT - denoted as MB), located outside the solenoidal magnet are shown. Additionally the approximate partitioning of CMS into Barrel, Overlap and Endcap areas, serviced by different subsystems of the new Moun Track Finder is shown.

The layout of the CMS detector is shown in Figure 1.

Results produced by the OMTF are used to derive the Level 1 trigger decision. To simplify synchronization of the trigger system, the data processing latency should be constant. The latency budget of OMTF is about 29 BX, but from those 9 BX are devoted for the data deserialization and serialization. In each overlap area¹ the OMTF receives data from eight layers of the muon detectors from the barrel partition of the detector and seven layers from the endcap partition². Each overlap area is serviced by six OMTF processors, each covering a 70° range of the azimuthal angle Φ . There are 10° overlaps between neighbouring OMTF processors. Each OMTF processor in each layer handles inputs from three (from the barrel partition) or from seven (from the endcap partition) neighbouring chambers. Each chamber may provide up to two hits in each BX, which finally gives up to 14 input hits per layer in the OMTF processor.

2. Algorithm of the trigger

The trigger design used during the LHC Run-1 utilized detector-specific muon identification algorithms: PACT Pattern Comparator for RPC [8], Drift Tube Track Finder for DT [9] and Cathode Strip Chamber Track Finder [10] for CSC. Unfortunately, none of these solutions may be used to

¹There are two overlap areas, one on each end of the detector

²In the barrel, there are 5 RPC layers (RB1in, RB1out, RB2in and RB2out, RB3), and three DT layers (MB1, MB2 and MB3) connected. In the endcap, there are 3 RPC layers (RE1/3, RE2/3 and RE3/3), and 4 CSC layers (ME1/2, ME1/3, ME2/2 and ME3/2) connected.

process combined signals from different types of detectors (RPC, CSC and DT) in the Overlap area. Therefore, it was necessary to find a common representation of the signals delivered by those detectors. When the charged particle crosses a chamber in one of the detectors, the chamber generates a signal that must be converted to the uniform format containing the azimuthal angular position (Φ) at which the particle crossed the chamber. This signal will be called a hit³. The chamber in which the hit occurred is univocally defined by the link through which the hits data are delivered to the OMTF trigger.

The complex detector geometry in the overlap region and potentially high number of hits in the event motivated us to develop an algorithm based on the comparison of reconstructed signals from detectors (hits) with a set of precomputed patterns, called Golden Patterns (GPs). Each GP is an object intended to represent a muon tracks with defined transverse momentum range and sign. It contains information about average track bending in the CMS magnetic field⁴ between consecutive detector layers and possible track deviations (due to stochastic effects), represented in associated probability density functions (PDF).

From 15 available detector layers (barrel: 5 RPC, 3 DT layers; endcap: 4 CSC, 3 RPC), eight have been selected⁵ to be so called “reference layers”.

To provide analysis of multi-muon events, up to 4 hits in these layers (reference hits) are selected to start the muon reconstruction within OMTF. To detect reference hits, up to 128 reference hit ranges are defined, each covering a certain range of azimuthal angular position (Φ) in one of reference layers. The reference hit ranges are ordered (assigned a priority) in such a way, that processing of up to 4 reference hits with the highest priorities maximizes the probability of detection of all muons (up to 4) in a multi-muon event.

For each hit in the event, the difference in the angular azimuthal positions ($\Delta\Phi$) between this hit and the particular reference hit is computed. Further processing is performed for each defined GP independently, consecutively in all layers. Taking into account bending of the average track ($\Delta\Phi_{mean}$) kept in the GP, the distance of each hit in the particular layer from the average position in that layer is calculated (Φ_{dist}) (see Figure 3).

The hit with the smallest Φ_{dist} is found, and the PDF value corresponding to its position is extracted from the GP. The PDF values from all layers are summed. The layers with hits providing non-zero PDF value for particular GP are called “active layers”. Finally, the best matched GP is found basing first on the highest number of active layers and then on the sum of PDF values from all layers⁶.

³Such representation may be used for all three detectors with one exception: the DT detector produces information not only about the azimuthal angular position of the hit, but also about the particle track azimuthal direction (bending angle)

⁴The curvature of the muon track depends on the transversal momentum of the muon.

⁵The selection was based on the high resolution in Φ , low noise, and good coverage in Φ and η . Currently, the MB1, ME2/3, MB2, ME1/3, RE2/3, MB3, RB1in, and RB1out layers are selected, but this selection results from simulations and may be modified in the further optimization of the algorithm.

⁶The data received from DT are handled in a special way, as they contain the hit Φ and the estimated track azimuthal direction. Therefore, each layer consisting of the DT chambers is treated as two layers: one standard layer delivering Φ , and the second one delivering the track direction. Of course, the Φ_{dist} is calculated only for the first layer. The direction data are processed directly. The PDF values are generated for the both layers independently. The DT layer is treated as “active” only if both Φ and direction layers deliver non-zero PDF values.

Figure 2. Schematic view of the analysis of a multi-muon event based on reference hits. There are up to 128 reference hit ranges defined. In the figure the reference hit ranges from 0 to 13 are schematically shown. The index of the range defines its priority. (0 is the highest priority). In the example shown there are two muons detected. Each of them generates three hits in different layers. The first muon generates one reference hit in the reference hit range 2. The second muon generates two reference hits in the reference hit ranges 6 and 13. Taking the reference hits in the order of their priority (2, 6, 13) our algorithm should first detect the 1st muon, and then twice the 2nd muon. The Ghostbuster algorithm at the end of the OMTF processor should filter out the superfluous detection of the 2nd muon.

Thus up to 4 muon candidates (given by GP), associated with the reference hits, can be reconstructed. Within these candidates duplicates (one physical muon may result in several candidates) are removed and best candidates in terms of reconstruction quality are selected as a result of algorithm reconstruction.

3. Hardware Platform

The chosen hardware platform for upgraded CMS trigger is μ TCA. The OMTF processor requires an μ TCA board equipped with FPGA chips with sufficient logic resources and link capabilities. To minimize the design and maintenance costs, a versatile MTF7 board [11] designed for the endcap trigger was selected. The MTF7 board is an μ TCA double width AMC board equipped with two FPGA chips: XC7VX690T and XC7K70, and with an optional memory daughterboard. The OMTF processor occupies the bigger XC7VX690T chip [12].

4. Data transmission and preprocessing

In the pre-upgrade trigger, the RPC, CSC and DT detectors are connected to different trigger systems (see section 2. Therefore, they transmit data via optical links at different rates (1.6, 3.2 and 10 Gbps respectively) using different transmission modes (synchronous or asynchronous). Before these data are processed by the OMTF trigger, it is necessary to deserialize them and combine

Figure 3. The principle of operation of the Golden Pattern algorithm. The figure shows, how contribution from a certain layer is calculated basing on the azimuthal angles of the hit in the reference layer and hit in the contributing layer Φ_i . Φ_{mean} describes the average bending of the tracks in the i^{th} layer (for the particular reference layer). It also corresponds to the maximum of the PDF function. Φ_{dist} describes the distance between the analyzed hit and the maximum of the PDF function.

them in the data sets corresponding to the particular event. To ensure that each data source is independently synchronised with the LHC clocking system, synchronization circuits are controlled via IPbus [13] software and specific delay coefficients are determined (and tuned) at runtime. To allow uniform further processing, those data must be converted into a unified format, containing the azimuthal angle (Φ) of each hit. Conversion parameters (specific to the data source) can also be controlled with IPbus. Finally, the data from various detectors must be aligned in time before passing to the OMTF algorithm. The whole process can be divided into 3 stages:

1. Data reception and alignment to local BC0
2. Preprocessing
3. Data realignment to provide common head-of-line

Mechanisms of data reception and alignment differ greatly between all 3 sources. The simplest one is CSC receiver: synchronous data from each of 35 links is packed in four-word packets, running at 160 MHz. This means that simple HDL logic is sufficient for extracting required half-strip number N_{hs} into generic 40 MHz domain. At present, a simple linear formula is used to convert each hit into the OMTF scale. For the RPC, the conversion is done with look-up tables. The DT conversion algorithm is still under development now. In all cases, the conversion parameters are programmed via IPbus.

5. Implementation of the algorithm

The OMTF processor has been implemented in a pipelined architecture. To allow checking of the

Figure 4. Data transmission. CDC - Clock domain crossing block.

Figure 5. Block diagram of the OMTF trigger processor firmware.

multiple reference hits in a single event, the internal clock frequency of the OMTF processor is a multiple of the LHC clock frequency (ca. 40 MHz). The multiplication factor is parametrized and may be changed before compilation. The current implementation of the OMTF is able to operate with the multiplication factor of 4, i.e. with the internal clock frequency of ca. 160 MHz, allowing to check up to 4 reference hits in each event.

The OMTF processor must be implemented in a flexible and easy to maintain way in order to fulfill High Energy Particle Physics requirements to allow further development, optimization and adaptation to running operating conditions of LHC.

Therefore, a special high level parametrized VHDL description was used. Various parameters, like: the number of reference hits, the length of the word on different stages and the number of patterns may be adjusted before synthesis. Most data structures are described using the hierarchy of VHDL record types, which allows to minimize code changes during optimization of the algorithm [7].

The block diagram of the OMTF trigger processor firmware is shown in Figure 5. The particular blocks of the processor implement consecutive steps of the algorithm described in Section 2.

Figure 6. Block diagram of the Golden Pattern Unit (GPU) calculating the PDF values for the group of patterns in a single layer.

The high-speed priority encoder is responsible for delivery of up to 4 reference hits in each BX, in decreasing priority order. The azimuthal angle of the selected reference hit Φ_{ref} is passed to the shift register so that it can be delivered to the final processing block together with the information about the best matched GP. Basing on the reference hit range, only a subset of OMTF inputs (currently up to 6 inputs, delivering data with Φ reasonably near to the Φ_{ref}) is used for further processing, this operation is performed by the “Connection area builder”. The next parts of the system are designed to take full advantage of the parallel data processing in FPGA. The $\Delta\Phi = \Phi - \Phi_{ref}$ is calculated simultaneously for each hit delivered in selected inputs. The calculated $\Delta\Phi$ values are then processed in parallel by 20 Golden Pattern Processors (GPP). Each GPP checks matching of hits to two or four GPs. In total, all GPPs handle 52 Golden Patterns. The GPP consists of multiple (currently 18) Golden Pattern Units (GPU). Each GPU calculates the PDF values for a single detector layer⁷. The block diagram of a single GPU is shown in Figure 6.

The GPU first calculates the $\Phi_{dist} = \Delta\Phi - \Delta\Phi_{mean}$ for each of six inputs, and the input with the lowest absolute value of Φ_{dist} is found. To ensure optimal operation both for “wide” GPs corresponding to low p_T muons and for “narrow” GPs corresponding to high p_T muons, it is possible to select which bits of the resulting Φ_{dist} are used to address the look-up table (LUT) with PDF values. The PDF LUTs are implemented with Virtex 7 BRAM blocks storing 1024 18-bit words. They offer 10 address bits. Seven of them are used for Φ_{dist} . If the calculated Φ_{dist} is too big to fit in the seven bits, the overflow is detected, and the $PDF = 0$ is returned. The remaining three bits of the LUT address are produced from the reference hit – namely they are the number of the reference layer in which the currently used reference hit was detected. Each BRAM stores two PDF

⁷The number of layers handled by a GPU (18) is higher than the number of layers in the detector (15) because the three DT layers are split into two layers each - the first one producing the Φ data, and the second one producing the bending angle.

values. Therefore, GPUs in GPPs servicing two GPs contain a single BRAM block while GPUs in GPPs servicing four GPs contain two BRAM blocks. If non-zero PDF value for the particular GP is produced, it is added to the sum of PDFs (SPDF) for that GP, and the number of fired layers for that GP is increased⁸.

From the results delivered by all GPPs, the best matched (i.e. with the highest number of active layers and the highest SPDF value) GP is found. The number of best GP, together with the corresponding Φ_{ref} is delivered to the final processing block (see Figure 5). After collecting all (up to 4) results for the currently processed BX, this block sorts muon candidates, removes possible duplicates and transmits data to the L1 GMT.

The OMTF processor is configured with the XML data prepared basing on physical simulations. Those data describe the reference hit ranges, the configuration of the “Connection area builder”, and the GP definitions. A special tool written in Python is used to automatically generate the VHDL files and memory initialization files from the XML data.

6. Results & Conclusions

A functional implementation of the algorithm is already available although further upgrades are expected. Given that the maintenance of firmware for complex pipelined systems is a time consuming operation, a special methodology for automatic equalization of latency in parallel paths has been developed [14] and should be used in future versions of the OMTF. The interesting finding was that utilization of DSP blocks available in the Virtex 7 chip didn’t improve performance and didn’t reduce the logic resource utilization. The DSP blocks introduced their own latency, and in the pipelined architecture it was necessary to use slice registers to buffer other data. With short word fixed-point arithmetics used in the GPU blocks, automatically synthesized adders offered similar performance and resource utilization, but their VHDL description was much simpler and easier to maintain⁹.

The whole OMTF Processor (full project with surrounding infrastructure) was successfully synthesized for the XC7VX690T chip available in the MTF7 board. The chip occupancy is as follows:

- Slices used: 94.48% (102319 from 108300 available)
- Slice LUTs: 64.88% (281040 from 433200 available)
- Slice Registers: 32.87% (284752 from 866400 available)
- Block RAM tiles: 51.22% (753 from 1470 available)

Due to the fact, that synthesis of the OMTF firmware is a time-consuming process (typically ca. 8 hours on 8 cores 64GB Xeon machine), it is important, that correctness of the algorithm after any modification may be verified in simulation. Therefore, a dedicated environment for generation

⁸As it was described in section 6, the DT layers are treated in a special way. However, the detailed description of the implementation is too complex to be included in this paper.

⁹In the next version of OMTF, it may be possible to use the DSP blocks if Xilinx adds support for SIMD mode in inferred DSP48 blocks [15].

of input data and expected results from simulation XML files has been created. Full verification of design, including timing, has been also performed in real hardware. The tests - both in simulations and in hardware have confirmed correct operation of the OMTF trigger.

References

- [1] CMS Collaboration, *The CMS experiment at the CERN LHC*, 2008 JINST **3** S08004. Also published by CERN Geneva in 2010.
- [2] M. Konecki, *CMS: Performance, physics, perspectives*, Acta Physica Polonica B **45** (2014), no. 07 1427–1445.
- [3] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys.Lett. **B716** (2012) 30–61, [arXiv:1207.7235].
- [4] CMS Collaboration, *CMS Technical Design Report for the Level-1 Trigger Upgrade*. No. CERN-LHCC-2013-011. CMS-TDR-12 in Technical Design Report CMS. Geneva, Jun, 2013.
- [5] CMS Collaboration, *CMS. The TriDAS project. Technical Design Report, Volume 1: The Trigger Systems*. No. CERN/LHCC 2000-038, CMS TDR 6.1 in Technical Design Report CMS. CERN, Geneva, 2000. <http://cds.cern.ch/record/706847>.
- [6] CMS Collaboration, *CSC local efficiency measurements from run II data*, . <http://cds.cern.ch/record/2064410>.
- [7] W. M. Zabolotny et al., *FPGA implementation of overlap MTF trigger: preliminary study*, Proc. SPIE **9290** (2014) 929025–929025–11.
- [8] M. Andlinger et al., *Pattern comparator trigger PACT for the muon system of the CMS experiment*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **370** (1996), no. 23 389 – 395.
- [9] J. Erö et al., *The CMS drift tube trigger track finder*, 2008 JINST **3** P08006.
- [10] V. Barashko, A. Drozdetskiy, A. Korytov, G. Mitselmakher, and Y. Pakhotin, *Fast algorithm for track segment and hit reconstruction in the CMS cathode strip chambers*, Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **589** (2008), no. 3 383–397.
- [11] D. Acosta et al., *The CMS modular track finder boards, MTF6 and MTF7*, 2013 JINST **8** C12034.
- [12] A. Byszuk et al., *OMTF firmware overview*, Proc. SPIE **9662** (2015) 966241–966241–6.
- [13] R. Frazier, G. Iles, D. Newbold, and A. Rose, *Software and firmware for controlling CMS trigger and readout hardware via gigabit ethernet*, Physics Procedia **37** (2012) 1892 – 1899. Proceedings of the 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP 2011).
- [14] W. M. Zabolotny, *Automatic latency balancing in VHDL-implemented complex pipelined systems*, ArXiv e-prints (Sept., 2015) [arXiv:1509.08111].
- [15] Xilinx, *Vivado synthesis - what features are not supported for dsp48 inference?*, 2015. <http://www.xilinx.com/support/answers/55196.html> .