A demonstrator for a level-1 trigger system based on MicroTCA technology and 5Gb/s optical links

PUBLISHED BY IOP PUBLISHING FOR SISSA

RECEIVED: *October 13, 2010*
ACCEPTED: *November 4, 2010*
PUBLISHED: *November 25, 2010*

TOPICAL WORKSHOP ON ELECTRONICS FOR PARTICLE PHYSICS 2010,
20–24 SEPTEMBER 2010,
AACHEN, GERMANY

A demonstrator for a level-1 trigger system based on MicroTCA technology and 5Gb/s optical links

C. Foudas,[a,1] R. Frazier,[b] G. Hall,[a] G. Iles,[a,2] J. Jones,[c] J. Marrouche,[a] D. Newbold[b] and A. Rose[a]

[a] *Blackett Laboratory, Imperial College,*
London SW7 2BW, U.K.
[b] *H.H. Wills Physics Laboratory,*
Tyndall Avenue, Bristol BS8 1TL, U.K.
[c] *Weathertop,*
Claverton Down Road, Bath BA2 7AL, U.K.

E-mail: g.iles@imperial.ac.uk

ABSTRACT: A demonstrator for the CMS Level-1 calorimeter trigger system has been designed, manufactured, tested and a time-multiplexed trigger implemented. The prototype card uses the AMC double width form factor, 5Gb/s links and a Xilinx XC5VTX150T or XC5VTX240T FPGA. A possible implementation of such a trigger architecture in CMS is described.

KEYWORDS: Trigger concepts and systems (hardware and software); Image filtering; Data processing methods; Digital signal processing (DSP)

[1] Since moved to University of Ioannina, Greece
[2] Corresponding author.

doi:10.1088/1748-0221/5/11/C11015

Contents

1 Introduction

The CMS experiment Level-1 trigger system selects interesting physics events at a rate of 100 kHz from an input rate of 40 MHz. It is designed to operate up to a luminosity of 10^{34} cm^{-2} s^{-1}. The luminosity will increase to 2×10^{34} cm^{-2} s^{-1} with the planned LHC phase I upgrade in 2016 and reach 5×10^{34} cm^{-2} s^{-1} with the LHC phase II upgrade in 2020. The Level-1 trigger system will operate well up to the nominal luminosity, but beyond that it will degrade due to pile up events that will make distinguishing physics objects from background more challenging.

To counter this it is planned to upgrade the calorimeter clusterisation algorithms and improve the resolution at which these operate so that they take full advantage of the 0.087η x 0.087ϕ granularity of the trigger primitives generated by ECAL (Electromagnetic Calorimeter) and HCAL (Hadronic Calorimeter). The upgrade will also leave open the potential to include trigger information from the Tracker at Phase II.

The extra algorithm complexity is only now feasible because of the continuing advance in digital signal processing performance in reconfigurable programmable logic (FPGAs). This should allow CMS to build a much more powerful, yet simpler and easier to maintain trigger than we currently have. This technology has characteristics that are significantly different from the technologies used in the past.

Firstly, high-speed serial links have emerged as an excellent way to bring large volumes of data into FPGAs, but they have a high latency, typically 100–200ns. It is therefore essential that the number of serialisation stages be kept to a minimum. Consequently, all our new designs are based on just 3 or 4 serialisation stages, which must include both the serialisation stage from ECAL and HCAL to the calorimeter trigger and those to the GT (Global Trigger).

Secondly, the time per processing step within the algorithms has gone down substantially so that FPGAs now operate with a few 100 MHz clock. This lends itself to a pipelined architecture rather than one that is simply clocked. For example, imagine data being clocked in from a serial link at 240MHz (i.e. LHC bunch crossing clock x6). A conventional architecture will wait until all the data are present and then processes it in parallel, first doing task A, then B, then C, etc until the next bunch crossing (i.e. it consists of 6 stages that are only active for 1/6 the time). In a true pipelined design all tasks are running concurrently with new data fed into task A on each 240MHz clock cycle. This approach is only really applicable when you start to have many clock cycles of data to process.

These observations led to a re-evaluation of the current calorimeter trigger architecture, which follows a conventional design of nodes that process small parts of the detector for every bunch crossing to one based on a time-multiplexed architecture that processes large areas of the detector over many bunch crossings. It was first proposed by John Jones [1].

The calorimeter trigger is one of the most challenging aspects of CMS because the large data volume of several Tb/s has not just to be processed, but also (a) the data must be shared/duplicated between processing nodes to satisfy boundary constraints (b) the resulting physics objects need to be sorted in order of significance (c) this must be achieved within a latency budget of $\sim 1\mu$s.

The data sharing is a particularly significant constraint, which has in the past required complex systems/backplanes to share/duplicate data between processing nodes. In a time-multiplexed trigger, data from a single bunch crossing (bx) are concatenated and delivered to processing system over several bx. This approach requires several processing systems operating in a round-robin fashion (i.e. processing system 1 takes bx = n, processing system 2 takes bx = n+1). We currently envisage approximately 10 processing systems. The major advantage with this approach is that the whole system becomes much more efficient because the ratio of the area processed to the boundary area is substantially increased. This results in fewer cards, which also makes the subsequent sort simpler.

The obvious drawback with this approach is that there is an immediate latency increase to time multiplex the data; however we expect this to be offset by the ability to build a much more compact trigger, requiring fewer serialisation stages. The system also has other advantages. For example, it is possible to prototype the entire trigger system with just 10% of the hardware. It also offers redundancy because if one of the processing systems were to fail the data could be redirected to a backup processing node. The system does not requite complex active or passive backplanes and can be built with a single card design.

2 Technology demonstrator: MINI-T5

In order to evaluate the feasibility of different trigger architectures, gain experience in the latest technologies (e.g. MicroTCA) and develop the core firmware and software blocks that are common to many designs we have developed a double-width, full-height AMC card, MINI-T5, to prototype new trigger designs.

The card is compatible with either a Xilinx XC5VTX150T or XC5VTX240T FPGA. The card has 32 input and 20 output 5Gb/s optical links that provide 160Gb/s (input) and 100Gb/s (output) of optical IO capability. The optical modules are two QSFPs [2] that each provide 4 bi-directional links and 2 input, 1 output SNAP12 (Rev0) / POD (Rev1) that are uni-directional and transmit or receive up to 12 optical links each. The switch between SNAP12s to PODs between revision 0 and 1 of the board was driven simply by a lack of availability of SNAP12 devices, however these are once more available. The 32 optical links (SNAP12/QSFP) were tested in two stages, each with 20 in external fibre loopback and the rest in internal transceiver loopback because of the single SNAP12/POD output. The links were operated for 12 hours without error, corresponding to $\sim 7 \times 10^{15}$ bits. The POD optics on Rev1 will be tested shortly.

The majority of the remaining high-speed serial transceivers are connected so that the card is compatible with the services available from a standard MicroTCA telecom crate [3] with the MCH in the primary slot (i.e. Ports: 0 (Ethernet); 2 (SATA/SAS); 4–7 (Fat-Pipe — e.g. SRIO, 10GbE, PCIe)). The absence of a dedicated PCIe clock in a telecom crate does require that any PCIe devices support the PCIe independent clock option. The last two high-speed serial transceivers are connected to AMC ports 1 and 8 so that they can utilise the DAQ functionality provided by the CMS service card [4]. The remaining AMC ports are connected to LVDS.

In addition to high-speed serial connectivity MINI-T5 also has dual 40×800 Mb/s LVDS IO via a 40 way differential Samtec connector on either side of the card. These can be joined together via an off-the-shelf Kapton cable from Samtec. It provides a 2×32Gb/s low-latency connection. An Atmel AVR32 microprocessor provides IPMI control and a USB2 interface.

3 Firmware infrastructure

The firmware core architecture is relatively simple. It comprises 5Gb/s GTX transceiver elements configured to have the minimum latency possible without bypassing elements such as the transmit FIFO or receive elastic buffer. These can be bypassed to reduce latency, but with extra complexity. A low latency design in the GCT project suffered from data corruption depending on the firmware build [5]. The problem was eventually traced to subtle clock routing issues which were eventually solved, but it shows that considerable care must be taken when using the serdes blocks in a non-standard configuration.

The GTX transceiver transmit data path or the algorithm input can be driven by a pattern derived from a bunch crossing counter or from a pattern injection RAM. The RAM can also capture incoming data. The firmware validation involves driving data out of the GTX transceivers, onto optical fibres (different lengths), and then receiving it with different GTX transceivers. This ensures that the low-latency synchronisation blocks that align the incoming data are fully tested.

The algorithm input is simply an array of 32bit wide data generated from the GTX transceivers which makes swapping in/out different algorithms very simple. A DAQ capture unit is placed both before and after the algorithm enabling algorithm verification in software. The DAQ units capture an array of 32bit wide data of arbitrary length in a pipeline which is transferred to a DAQ buffer upon receipt of a Level-1 trigger.

4 Trigger geometry of CMS

The region of CMS in which both ECAL and HCAL trigger input data are present spans a range of +/- 3 η and all of ϕ. It is segmented into 56 towers in η and 72 towers in ϕ with a granularity of 0.087 η x 0.087 ϕ up to +/- 1.74 η. The HF (Forward Hadron Calorimeter) extends η coverage up to +/- 5 η, albeit at a coarser resolution. On each side it is segmented into 8 units in η and 36 units in ϕ, which is mapped onto 16 towers in η and 72 in ϕ. Plans exist to double the resolution of HF in both η and ϕ so that it matches that of the rest of the calorimeter.

5 Firmware algorithms and laboratory system

The laboratory demonstrator system is simplified to focus on the resource-hungry components of the trigger algorithms. It assigns 8bits per tower for both calorimeters, which is used solely for the energy deposition. In the existing system there are 9bits per tower, with the extra bit used for calorimeter-specific information. The demonstrator system also ignores the HF because the lower resolution (double tower) and missing ECAL in this region reduces the data rate by a factor of 8.

The demonstrator system is therefore a good approximation to the challenges posed by a real system. In the full lab system there would be 28 links running at 4.8Gb/s, each loading 2 ECAL and 2 HCAL energy depositions per 120MHz clock. A single clock cycle therefore loads an entire row of constant ϕ (i.e. 56 towers in η). At present we have limited the system to 12 input links because with a single card we cannot drive all 28 input links.

It therefore takes 72 clocks to loop over the full ϕ span of 72 towers. The entire calorimeter is therefore loaded in 24bx. This is far too long for the final system, but in the laboratory this is perfect for testing algorithms. The test system required 22% of registers and 29% of LUTs and almost all BRAMS within the smaller FPGA (XC5VTX150T). External RAM will be available in a final design.

The current system only has the electron-finding algorithm implemented, which has been verified using a C++ testbench in conjunction with ModelSim's Foreign Language Interface.

Work on the jet finding and subsequent sort has been delayed so that a robust software structure for hardware access can be put in place (i.e. similar CMS HAL for VME access). The electron algorithm used is the 2×2 clustering algorithm [6], albeit implemented for a time-multiplexed trigger.

6 Time-multiplexed system for CMS

There are many ways to time multiplex the incoming calorimeter trigger information, of which one is shown below (figure 1). It assumes the worst case (i.e. that the HF calorimeter is upgraded to operate at tower resolution rather than the current 2×2 tower resolution and that the system is upgraded to transmit 12bits per tower rather than 9bits.

The Main-Processor (MP) nodes are split across two cards (MP+ and MP-) so that we have some margin (i.e. not operating at extreme limit of FPGA link technology). The cards each process either positive or negative η. There are 10 of these MP nodes operating in a round robin scheduling manner, each only receiving data for every tenth bunch crossing. The two cards receive a single 9.6Gb/s link from each Pre-Processor (PP) card in their respective η half. They also receive 4 links from the 4 adjacent towers in the opposite η half so that they have sufficient boundary information to build physics objects at the boundary between the two processing nodes.

The Pre-Processor cards spanning the barrel and endcap each receive ECAL and HCAL data in a ring that is 1 tower wide in η and spans the full ϕ circumference. The lack of ECAL data in

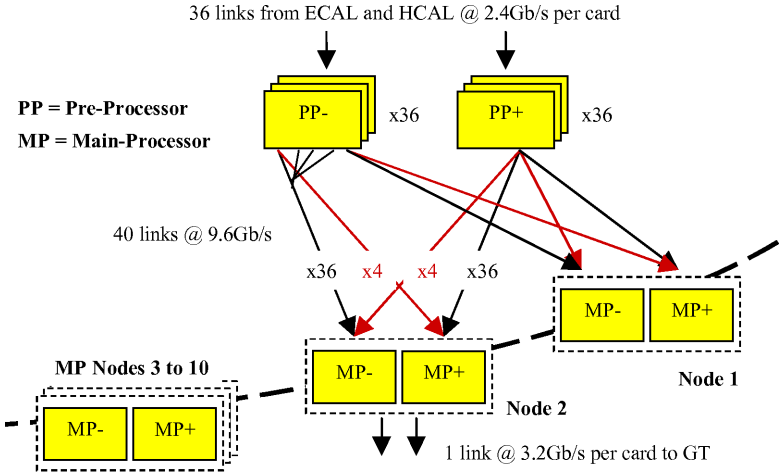

Figure 1. An example of a time multiplexed trigger within the CMS experiment. A full explanatiion is given in the text.

the HF region enables these rings to be 2 towers wide in η. This requires 2 × 28 cards for the the barrel and endcap and a further 2 × 8 for the HF and thus 72 PP cards in total.

7 Latency

The current latency of the calorimeter trigger path from the input of the serdes blocks on the Synchronization & Link Boards that are mounted on the ECAL and HCAL trigger cards through to the output of the serdes blocks within the GT Pipelined Synchronising Buffer is approximately 47bx (bunch crossings).

In the time-multiplexed example outlined here the latency is expected to be 44bx with 3bx contingency. It assumes that the final jet clustering and sort (not yet fully implmented) will take an additional 4bx beyond the 4bx used for the electron clusterisation and that the link will take 6bx. This is quite conservative but may be wise given that there is no guarantee that the final FPGA transceiver will operate in precisely the same way.

The transceivers on the MINI-T5 are operated at 5Gb/s with a 32bit wide fabric interface running at 125MHz and use the receiver side elastic buffer with minimum latency. The latency of a GTX transceiver on the MINI-T5 has been measured to be 5.3bx in internal loopback mode through the PMA (Physical Media Attachment) section. In theory the minimum latency should have been closer to 4.0bx. The discrepancy has yet to be investigated. Note that in a conventional trigger system the algorithm waits until all the bunch crossing data are present and thus it is necessary to add another bunch crossing of latency. The slowest data path through the custom synchronization block will only pass through a single LUT and thus the additional latency is very small.

8 Conclusions

The core part of a prototype time-multiplexed calorimeter trigger for CMS has been built and tested in a relatively short time span. It demonstrates that such a scheme is feasible and provides a useful test bench on which to develop firmware and software for the final system.

References

[1] J. Jones et al., *The GCT Matrix Card and its Applications*, in *TWEPP-09: Topical Workshop on Electronics for Particle Physics*, Paris, France, 21–25 Sep 2009, pp.259–264.

[2] *Micro Telecommunications Computing Architecture PICMG® Base Specification MTCA.0 R1.0*, http://www.picmg.org, July 6 (2006).

[3] *QSFP Transceiver*, http://www.sffcommittee.com/ie/index.html, Rev 1.0 November 2006.

[4] E. Hazen et al., *Development of a MicroTCA Carrier Hub for CMS at SLHC*, in *Topical Workshop on Electronics for Particle Physics 2010*, 20–24 September 2010, Aachen, Germany.

[5] G. Iles et al., *Trigger R&D for CMS at SLHC*, in *TWEPP-09: Topical Workshop on Electronics for Particle Physics*, Paris, France, 21–25 Sep 2009, pp.249–253.

[6] P. Klabbers et al., *CMS Regional Calorimeter Trigger Upgrade: Hardware and Firmware Proposals and Development*, 2010 CMS Internal Note (awaiting reference number).