Available on CMS information server **CMS IN -2011/023**

# CMS Internal Note

*The content of this note is intended for CMS internal use and distribution only*

**03 October 2011**

# A Combined Conventional and Time-multiplexed Calorimeter Trigger for CMS

A. Rose[1)]
*Blackett Laboratory, Imperial College, London SW7 2BW, UK*

**Abstract**

A calorimeter trigger design is demonstrated through which it would be possible to operate as either a Time-multiplexed trigger or as a Conventional trigger, with the choice being made at run-time.

The design requires a total of 36 cards of two designs and thus fits in three $\mu$TCA crates.

The design may operate using either 9.6Gbit/s or 6.4Gbit/s links, with the trigger tower resolution being adjusted accordingly.

[1)] awr01@imperial.ac.uk

# Contents



# 1. Introduction

For the upgrade of the CMS level-1 calorimeter trigger, two alternative schemes have so far been proposed; a Conventional trigger and a Time-multiplexed trigger[1].

The Conventional trigger design, like the current trigger, divides the detector into sections, with each section being handled by a separate board and the boundaries between the regions handled by sharing data between neighbouring boards. To perform global algorithms, the data must be reduced sufficiently to be received by a single board; this may either be achieved by summing data into coarser objects, or by sorting objects and only keeping those of the highest priority. In contrast to the existing system, the proposed conventional trigger design employs a μTCA infrastructure, instead of VME, uses optical fibres instead of SCSI cables and replaces ASICs with FPGAs.

A Time-multiplexed trigger is more like a cross between the existing Level-1 trigger and the existing Higher-level trigger, with a series of trigger nodes, typically around 10, operating in a round-robin fashion. Each node receives data corresponding to a single bunch-crossing and because it only handles every "Nth" bunch-crossing, may receive data spread over a period of up to "N" bunch-crossings. By this means, it has been shown that the full detector data may be received into a single trigger node at full granularity, allowing algorithms to be run at any resolution and allowing "global" algorithms to be performed before the rejection of any candidates. The proposed Time-multiplexed design is also based on a μTCA infrastructure, optical links and FPGAs.

Because a Time-multiplexed trigger need not throw any data away, unlike a Conventional trigger, it necessarily needs more links between the first and second stages and the optics are, therefore, more complex. On the other hand, the proposed Time-multiplexed trigger requires no sharing of data across the backplane, which the proposed Conventional trigger does require.

An alternative to sharing data between neighbouring boards in a Conventional trigger design is to duplicate the data at the output from the previous stage. This has the advantage of lowering the latency and simplifying the architecture within each processing stage at the cost of making the links between processing stages more complex, sometimes considerably more so.

A generalized Conventional trigger architecture featuring no sharing between neighbours is shown in Figure 1, whilst a generalized Time-multiplexed trigger is shown in Figure 2. Although these figures clearly show that there are similarities between the two systems, they are, in fact, overly simplified. To build any "real-world" system which can operate in both Conventional and Time-multiplexed modes requires some level of configurability of the link architecture. The configuration does not change whilst running, however, only when changing between modes of running.

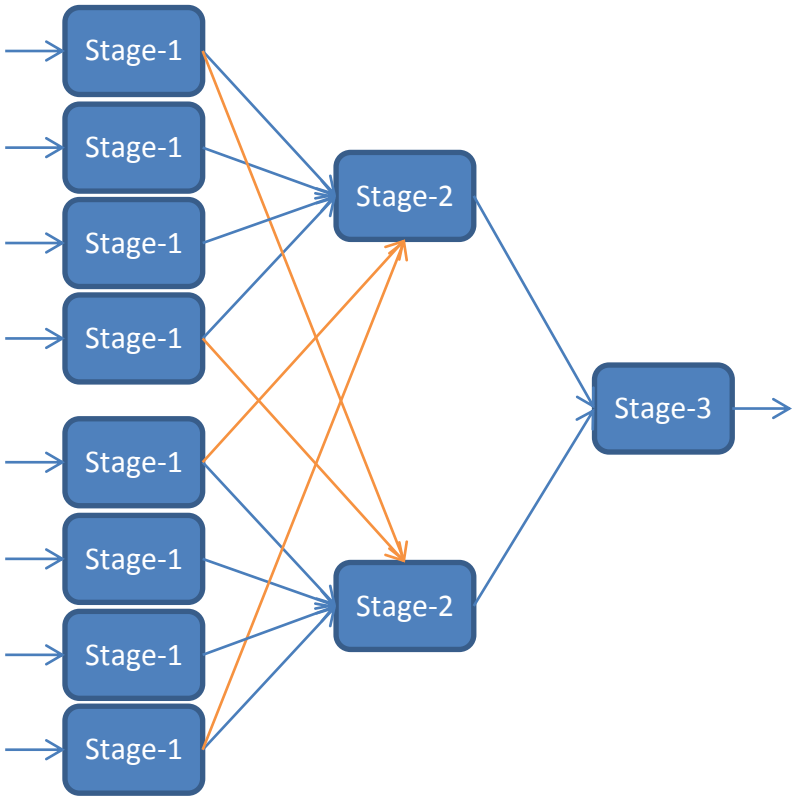


Figure 1 : A generalized Conventional trigger featuring no sharing between neighbours. Each stage-2 board handles a different region of the detector. To handle the boundaries between regions, a subset of the data must be duplicated at an earlier stage and sent to the boards on either side of the boundary, shown in orange.

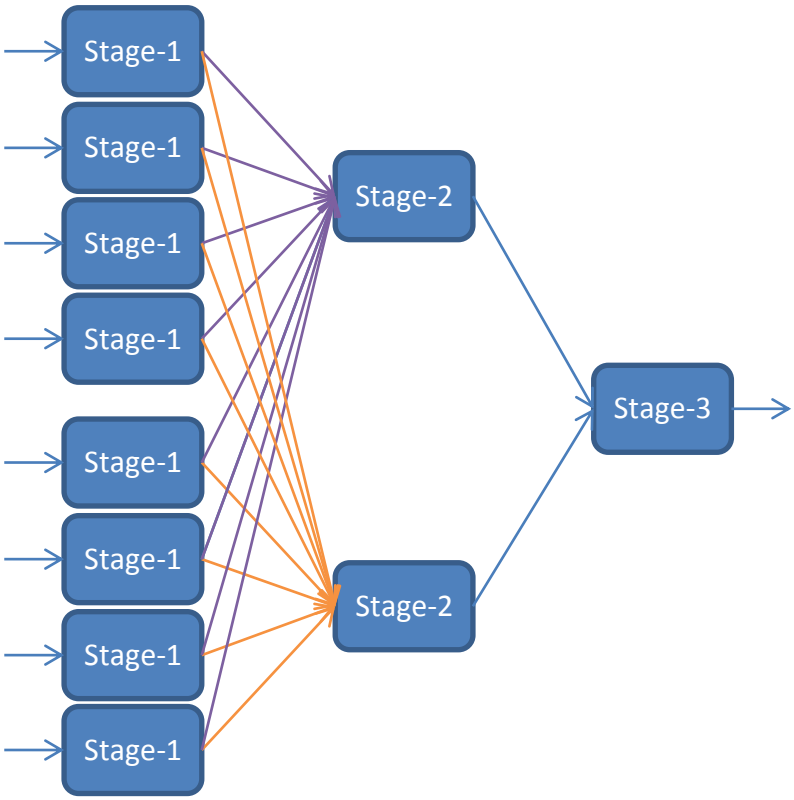


Figure 2 : A generalized Time-multiplexed trigger. Each stage-2 board handles a different bunch-crossing. Data from the first bunch-crossing is indicated in purple, whilst data corresponding to the second bunch-crossing is shown in orange.

# 2. Key technologies

The scheme presented here is based on three key technologies; FPGAs with integrated high-speed SerDes, high-density parallel optics and protocol agnostic cross-point switches.

To make use of prior experience, it is assumed here that Xilinx Virtex-series FPGAs will be used. The exact choice of FPGA, however, depends on which speed the internal links are to be run at. The maximum link speed which is both a multiple of the LHC machine clock and which may be handled by standard optics is 9.6Gbit/s, and to run at such a speed would require the use of Virtex-7[3] series FPGAs. If, however, the links are run at 6.4Gbit/s, it would be feasible to use either Virtex-6[4] or Virtex-7 series FPGAs.

Previous designs (for example [2], [8]) have made use of SNAP-12[5] or PPOD[6] optics which, although used very successfully, require front-panel space and have a large footprint. To make use of all the SerDes on the next generation of FPGA, there is insufficient space on a μTCA form-factor board to use PPODs and, instead, it is proposed that a new, higher density form-factor, such as Avago MicroPODs[7], are used. These devices are designed to be placed as close to the SerDes as possible, thereby avoiding long, high-speed traces and optimizing signal integrity. Should board space allow, another possibility is to use Avago MiniPOD[7] devices, which play a similar role to the MicroPODs but are easier to handle at the cost of having a significantly larger footprint.

Protocol agnostic cross-point switches allow serial signals to be programmably routed, duplicated and/or distributed, with very low latency since no deserialization step is used. A 72×72-way cross-point switch by Mindspeed was successfully included on the CMS GCT Matrix card[8]. For the scheme discussed here, a 144×144-way cross-point switch is required, with a higher bandwidth than is currently available using Mindspeed switches. Instead, we consider here the Vitesse VSC3144-12[9] (although alternatives may become available in the future); a 144×144 cross-point switch with a per-channel bandwidth of 10.709 Gbit/s.

# 3. Proposed hardware

Two hardware designs are required for the proposed scheme, a processing board and an optical cross-point board. The two board designs are ambitious technically, but it is the author's opinion that they are not unreasonably so[1]. This assertion is based on the success with which the required technologies have been used previously within the CMS GCT and on the widespread uptake of these technologies across the telecommunications industry and within the field of supercomputing[10].

## Master Processor board

The master processor board (abbreviated MP6 for a Virtex-6 design or MP7 for a Virtex-7 design) is an FPGA-based processing card with 72 high-speed serial data links in each direction. The optical links are implemented through 6 microPOD transmitters and 6 microPOD receivers. Such a design would require a single large Virtex-7 FPGA or, should this part be unavailable, two medium-sized Virtex-7s or two large Virtex-6s. If two FPGAs are used, as much bandwidth as possible should be made available between the chips, in order to maintain flexibility. It is also worth noting that a two FPGA design places an additional burden on the "Finalization" stage, since the global sorting must also be performed there. Two possible configurations for the master processor board can be seen in Figure 3.

## Optical cross-point board

The optical cross-point board (abbreviated Opti-X) consists of a single Vitesse VSC3144-12 cross-point switch (or similar), 12 microPOD transmitters and 12 microPOD receivers, Figure 4. It is expected that a very small "system's" FPGA would be included to provide the Gigabit-Ethernet interface to the μTCA crate.

[1] In the Appendix, a comparison is made of the key technologies proposed here with those used in three boards from the existing CMS trigger and an existing μTCA demonstrator board which has been developed for CMS

## Common features

Both boards would share many design features in common. For example, the “unit-cells” for the optical transmitters and the optical receivers would be identical on each; the physical layout of the microcontroller and associated sensors, as well as the software providing the IPMI interface, would similarly be identical on both board designs; as would firmware elements, such as the Gigabit-Ethernet interface. Furthermore, many of these features have already been implemented in existing designs[1] and could be directly ported to any new design.

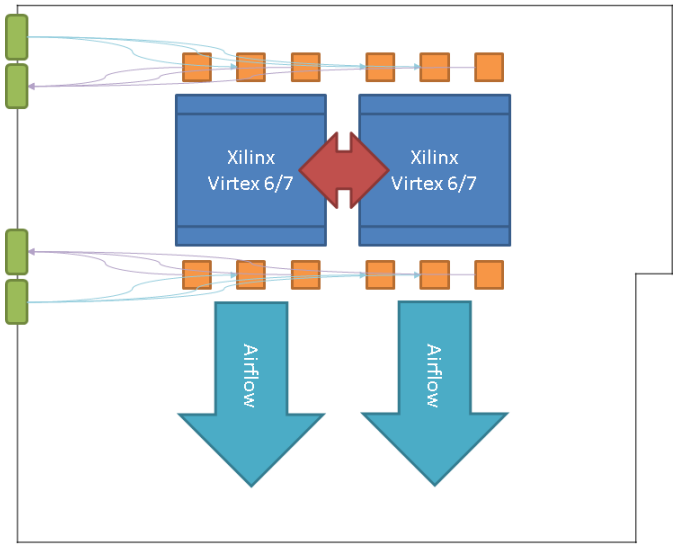


Figure 3a : The Master Processor board featuring two medium sized Virtex 7 FPGAs or two large Xilinx Virtex 6 FPGAs. Receiver fibres are indicated in blue whilst transmitter fibres are indicated in purple.

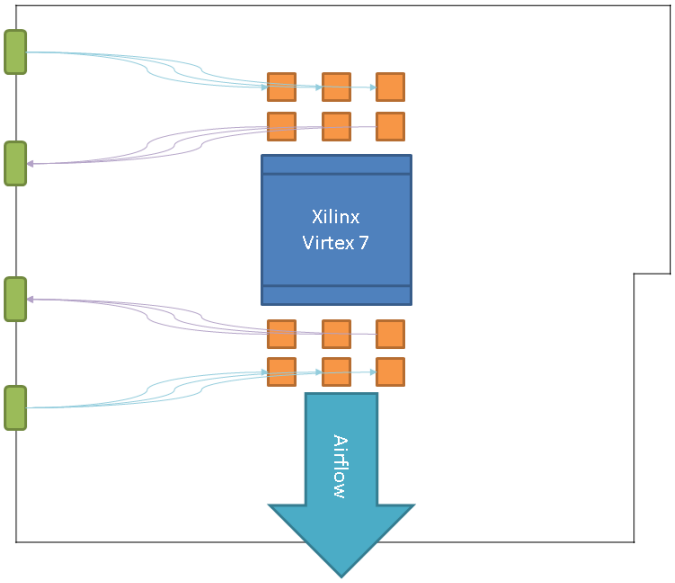


Figure 3b : The Master Processor board featuring a single large Xilinx Virtex 7 FPGA. Receiver fibres are indicated in blue whilst transmitter fibres are indicated in purple.

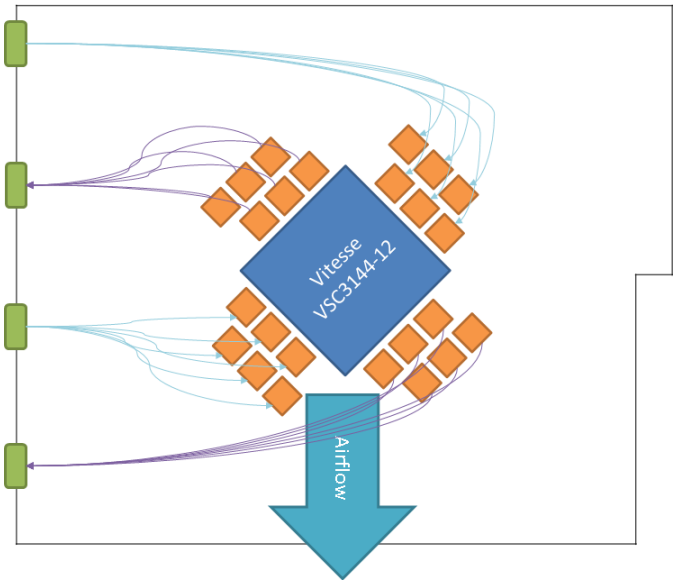


Figure 4 : The Optical Cross-point (Opti-X) board based on a Vitesse VSC3144-12 protocol agnostic cross-point switch. Mounting the cross-point switch at a 45° angle would improve fibre management and produce a more even airflow. Receiver fibres are indicated in blue whilst transmitter fibres are indicated in purple.

# 4. The stage-1 system

Based on the geometry of the detector, it is natural to divide the detector into 9 segments in φ and 2 in η, with each segment handled by one processor board, resulting in 18 stage-1 boards[2] in total. Such a region covers 28×8 (η×φ) trigger-towers in the barrel+endcap region and a further 8×4 (η×φ) trigger-towers in the forward region.

It is assumed that the trigger-tower data will arrive on optical fibres at 4.8Gbit/s, each fibre covering a region of 2×4 (η×φ) trigger towers, except for the boundary between the barrel and the endcap, where the links are only half occupied. As such, to receive all the Ecal barrel+endcap towers (EB+EE) requires 16×2 (η×φ) fibres, and similarly for the Hcal barrel+endcap (HB+HE). The forward Hcal (HF) requires a further 4 fibres, resulting in a total of 68 input fibres at 4.8Gbit/s.

# 5. Link architecture for a Time-multiplexed trigger

Since the link structure of a Time-multiplexed trigger is considered more complicated than a Conventional trigger, we shall first consider a link architecture suitable for a Time-multiplexed trigger. We shall initially consider the highest-bandwidth scenario, although other scenarios will be discussed later.

We consider the scenario where the raw data is simply re-serialized in a Time-multiplexed fashion on links running at 9.6Gbit/s. For each bunch-crossing of data, the EB+EE towers may be transmitted to a single processing node on 2 fibres over a period of 7 BX. Similarly the HB+HE, may be transmitted on a further 2 fibres over 7 BX. Rather than having a dedicated fibre for HF data, it is in fact optimal for the processing algorithms to have the HF data transmitted on the same fibres as the HB+HE, over an additional 2BX. As such, to transmit 1BX of raw data in a Time-multiplexed fashion requires 4 fibres transmitting over a period of 10BX. In all Time-multiplexed schemes proposed so far, the data is received by 12 trigger nodes[3] (including 2 spares), and so each stage-1 board must output 48 channels, 4 fibres to each of 12 trigger nodes.

Since each of the 18 stage-1 cards sends 4 ribbon cables and each Opti-X card may receive a total of 12 ribbon cables, a total of 6 Opti-X cards are required.

A technical detail which is worth noting here is that the optimal arrangement for the outputs is 4 12-channel ribbon cables, with each ribbon carrying data of one particular "type"; as opposed to placing all 4 fibres which handle the same bunch-crossing in the same ribbon. This is a trivial firmware detail for the stage-1 cards, but greatly improves the flexibility of the optical cross-point system.

For a Time-multiplexed trigger, the "Nth" channel of the "Mth" input ribbon must always map to the "Mth" channel of the "Nth" output ribbon, where N and M run from 1 to 12. One possible arrangement for distributing the 4 output ribbons from each of the 18 stage-1 cards to the 6 input ribbons of each of the 12 stage-2 cards is shown in Figure 5.

[2] The nomenclature "stage-1 board" refers merely to the board's purpose as defined by its position in the system and its firmware. Physically, the stage-1 board is a master processor card.

[3] In the scheme presented here, each trigger node is a single master processor card.

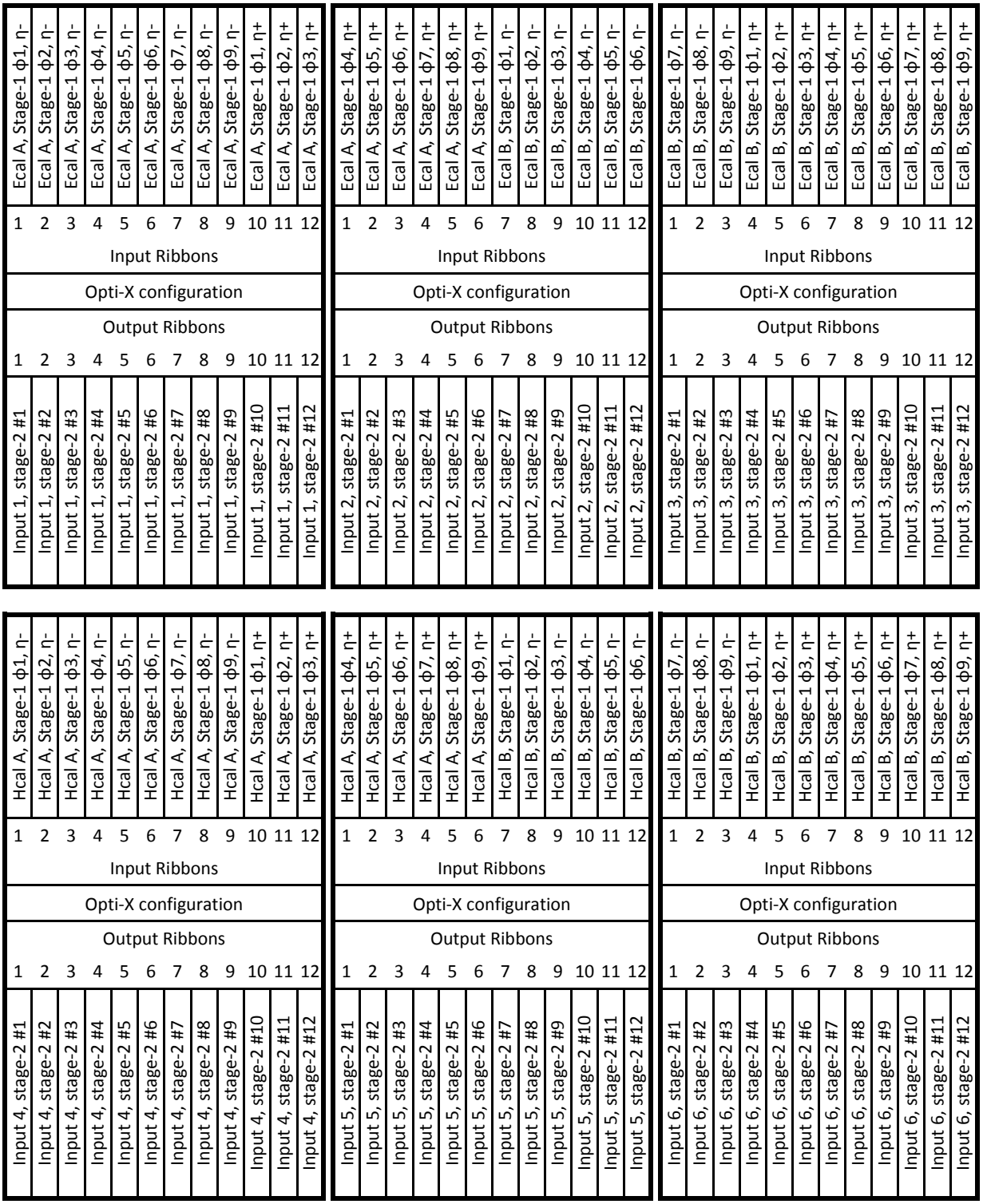


Figure 5 : One possible arrangement of ribbon cables from the 18 stage-1 cards to the 12 stage-2 cards for the 6 Optical cross-point cards. Despite the different input data types on each board, the configuration of the cross-point switch is identical on each: the cross-point switch must map the “Nth” channel of the “Mth” input ribbon to the “Mth” channel of the “Nth” output ribbon, where N and M run from 1 to 12.

# 6. Conventional trigger

Using the physical link structure suggested in the previous section, we may also consider the form of a conventional trigger implemented using the same hardware.

For the sake of simplicity, it is assumed that only 9 of the 12 stage-2 cards are used, each processing $1/9^{th}$ detector in φ and the entire detector in η. In such a scheme, only $9/12^{ths}$ of the total number of optical channels are available and so each stage-1 card may transmit data on a maximum of 36 of the 48 links.

The most trivial pre-processing which may be considered at stage-1 would be to convert the raw Ecal and Hcal towers into combined 18-bit "trigger" towers. For each stage-1 card, there are 224 barrel+endcap towers and 32 forward towers, resulting in an aggregate bandwidth of 4608bits per BX, or 230.4Gbit/s (including 8b10b encoding). To send this data to stage-2 at 9.6Gbit/s requires a total of 24 channels, leaving 12 channels free.

Each stage-2 card must, therefore, receive 48 fibres carrying this "core" data (from η+ and η-), leaving a further 24 channels available to receive data concerning neighbouring regions to handle the boundaries. Depending on how the data was organized within the links, it may be possible to perform the duplication within the cross-point switches, although even if this were not possible, the duplicated data could easily be transmitted through the remaining 12 available links on each stage-1 card, or by a combination of the two methods. If greater than 50% sharing is required, the tower resolution may be further lowered from 18bits/tower to 16bits/tower or even 12bits/tower. Although the bandwidth calculations show that the conventional trigger is fully implementable, it is not currently known if the link configuration required can be made as elegant as that of the Time-multiplexed design.

The scheme presented above represents the most trivial case; it is probable that more advanced processing, such as the formation of eγ-candidates, regions, mini-jets, etc., would be performed at stage-1 and that these objects would be sent to stage-2 for completion and final sorting, lowering the bandwidth requirements.

# 7. A combined Conventional and Time-multiplexed calorimeter trigger

Since the scheme presented features a single hardware architecture, we may consider the following scenario: Each stage-1 card, stage-2 card and Opti-X card has a flash memory featuring, one safe/boot image and two operational images, one for a Time-multiplexed configuration and one for a Conventional trigger architecture. Upon powering up the board, the safe/boot image is loaded and the system sits idle awaiting a command from the control-system as to which operational image to load. In this way, the choice of whether to use a Conventional or Time-multiplexed architecture is made at runtime, rather than being fixed at the time of construction.

Two further advantages of this system over others previously proposed are that, firstly, it requires no break-out cables or manual optical patch-panels, all the links are on multi-fibre ribbon cables and link boards directly. Secondly, this system is very compact; it requires a total of 36 boards and as such could fit in three uTCA crates, Figure 6.

Although the figures presented so far assumed a link-speed of 9.6Gbit/s, the scheme presented works equally well at 6.4Gbit/s, provided the tower resolution is lowered accordingly. For a Time-multiplexed design this corresponds to lowering the resolution from 24-bits/tower to 16-bits/tower and for the "minimal" Conventional trigger presented above, the resolution lowered from 18-bits/tower to 12-bits/tower.

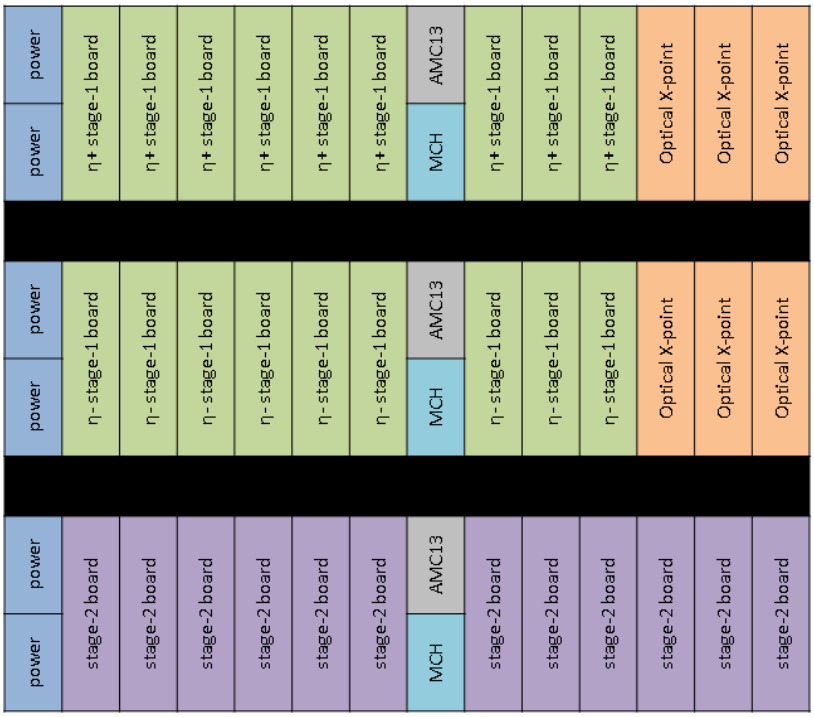


Figure 6 : A potential rack layout for a trigger system which may be operated as either a Time-multiplexed or as a Conventional trigger

# 8. Conclusions

A hardware configuration has been demonstrated through which it would be possible to choose between a Time-multiplexed and Conventional trigger architecture at run-time. The hardware required for such a scheme is ambitious technically, but is not considered to be prohibitively so, being a logical extension of hardware which has already been successfully implemented and tested.

The design can operate at either 9.6Gbit/s or 6.4Gbit/s, with the trigger tower resolution being adjusted accordingly.

The design requires a total of 36 cards of two designs. Both card designs have a lot of common features and could reuse many features from existing designs, facilitating development and improving maintainability.

# References


[1] G. Iles, A. Rose et al, A Time-multiplexed Calorimeter Trigger for CMS with Addendum, CMS IN 2011/008, http://giles.web.cern.ch/GILES/projects/slhc/TMT_TriggerUpgrade_IN2011_008.pdf

[2] Status of the Mini-T, G. Iles, CMS Trigger Upgrades Meeting, 25th March 2010, https://indico.cern.ch/materialDisplay.py?contribId=4&materialId=slides&confId=86618

[3] Virtex-7 FPGAs, Xilinx Inc., http://www.xilinx.com/products/silicon-devices/fpga/virtex-7/index.htm

[4] Virtex-6 Family Overview, DS150 v2.3, Xilinx Inc., 24th March 2011, http://www.xilinx.com/support/documentation/data_sheets/ds150.pdf

[5] InterBOARD Snap-12, Reflex Photonics Inc., SN-970-004-00 Rev 3.6, December 2010, http://reflexphotonics.com/PDFs/SN-970-004-00_Rev_3-6_InterBoard_Board-Edge_Data_Sheet.pdf

[6] AFBR-775BxxxZ / AFBR-785BxxxZ: Twelve-Channel Transmitter and Receiver Pluggable Parallel-Fiber-Optics Modules, AV02-2179EN, Avago Technologies, 7th October 2009, http://www.avagotech.com/docs/AV02-2179EN

[7] MicroPOD and MiniPOD 120G Transmitters/Receivers, Avago Technologies, http://www.avagotech.com/pages/minipod_micropod

[8] The GCT Matrix Card and its Applications, J. Jones, C. Foudas, G. Iles, M. Hansen, TWEPP-09: Topical Workshop on Electronics for Particle Physics, Paris, France, 21st-25th September 2009, pp.259-264, http://cdsweb.cern.ch/record/1234905/files/p259.pdf

[9] Vitesse VSC3144-12 10.709Gbps 144×144 Asynchronous Crosspoint Switch, VMDS-10336 rev. 4, Vitesse Semiconductor Corporation, January 2011, http://www.vitesse.com/products/product.php?number=VSC3144

[10] Avago Technologies Collaborates with IBM on High Bandwidth Optical Interconnect Breakthrough for Supercomputers, OFC Conference, San Diego, 22nd March 2010, http://www.avagotech.com/pages/en/press/avago_ibm_embedded_optics

# Appendix

| | | CMS GCT Leaf Card | CMS GCT Matrix Card | CMS GCT OptoGTI Card | Mini-T5 µTCA dev. board | MP7 | Opti-X |
|---|---|---|---|---|---|---|---|
| **Year** | | 2006 | 2008 | 2010 | 2010 | Est. 2012 | Est. 2012 |
| **Designer** | | M. Stettler | M. Stettler | G. Iles | G. Iles | - | - |
| **Form-factor** | | Dual PMC | Single width, full height µTCA | Dual CMC | Double width, full height µTCA | Double width, full height µTCA | Double width, full height µTCA |
| **Processing FPGA** | **Manufacturer** | Xilinx | Xilinx | Xilinx | Xilinx | Xilinx | - |
| | **Family** | Virtex-2 Pro | Virtex-5 | Virtex-5 | Virtex-5 | Virtex-7 | |
| | **Model** | P70 | LX110T | LX110T | TX150T TX240T | VX415T VX485T VX550T VX690T[4] | |
| | **Package** | FF1513 | FF1136 | FF1136 | FFG1759 | FFG1927 | |
| | **Dimensions** | 34.5mm × 34.5mm | 35mm × 35mm | 35mm × 35mm | 42.5mm × 42.5mm | 45mm × 45mm | |
| **Cross-point switch** | **Manufacturer** | - | Mindspeed | - | - | - | Vitesse |
| | **Model** | | M21141 | | | | VSC3144-12 |
| | **Configuration** | | 72×72 4.25Gb/s | | | | 144×144 10.709Gb/s |
| | **Package** | | CBGA 1156 | | | | FCBGA 1072 |
| | **Dimensions** | | 35mm × 35mm | | | | 45mm × 45mm |
| **Serial Links** | **Optical Channels ( In / Out )** | 32 / 0 | 32 / 32 | 16 / 16 | 32 / 20 | 72/72 | 144/144 |
| | **Max. Speed** | 2Gb/s | 3Gb/s | 3Gb/s | 5Gb/s | 10Gb/s | 10Gb/s |
| | **Optics form-factor** | SNAP-12 | SNAP-12 + POP-4 | POP-4 | PPOD + QSFP | miniPODs or microPODs | miniPODs or microPODs |
| **IPMI Microcontroller** | | - | Phillips NXP LPC2368 | - | Atmel AT32 UC3A | Atmel AT32 UC3A | Atmel AT32 UC3A |

Table A : A comparison of the key technology features of the two boards proposed in this note (MP7 and Opti-X boards) with three boards in the existing CMS trigger (Leaf, Matrix and OptoGTI cards) and an existing µTCA demonstrator board developed for CMS (Mini-T5 board).

[4] The four parts listed here are pin compatible and differ only in the number of serial links and the number of internal resources. For the single-FPGA scheme proposed in this document the VX690T is required: The smaller parts may, however, be used during development or to lower the cost of the board for use in projects where fewer links are required.