The GCT Matrix Card and its Applications

J. Jones [a], C. Foudas [b], G. Iles [b], M. Hansen [c]

[a] Princeton University, Princeton, NJ, USA
[b] Imperial College, London, UK,
[c] CERN, Switzerland

neutrinodeathray@gmail.com

Abstract

The Matrix card is the first in what is expected to be a series of xTCA cards produced for a variety of projects at CMS. It was developed as a joint collaboration between colleagues at Princeton, Imperial College, LANL and CERN. The device comprises the latest generation of readily-available Xilinx FPGAs, cross-point switch technology and high-density optical links in a 3U form factor. In this paper we will discuss the development and test results of the Matrix card, followed by some of the tasks to which it is being applied.

I. Introduction

The Matrix card was originally designed as part of the CMS GCT Muon and Quiet Bit System [1]. As such it was developed to provide a combination of reconfigurable optical links and firmware that can be adapted to different tasks without the redesign of the hardware itself. In this paper we will discuss the board's design, and the testing of the prototypes. This includes the infrastructure required to control the board (based on Ethernet). The I/O and computing performance of the card have been studied in detail and these results are also discussed. Since the production of two prototypes the board has been included in the design of a number of projects, including the LLRF control system for the FERMI free electron laser at Trieste and the calorimeter trigger upgrade project at CMS. In the FERMI project, the Matrix card provides a central timing and control point for the RF system. For the calorimeter trigger, its flexibility allows for changes in the algorithms without modification of the basic hardware and a reduction in latency by utilising wire-speed data duplication.

A. Card Specifications

The Matrix card design has been specified previously in [1][2][3]. In summary, it is a 3U (standard width), full height Advanced Mezzanine Card (AMC). The key components are MTP optics (SNAP12 and POP4), a large Xilinx Virtex-5 FPGA (XC5VLX110T-3) and a Mindspeed 21141 72x72 4Gb/s protocol-agnostic cross-point switch. The latter of these components is the key feature of this design, allowing the reconfiguration of a system to handle different processing topologies. It also provides the possibility of wire-speed data duplication and dynamic redundancy management.

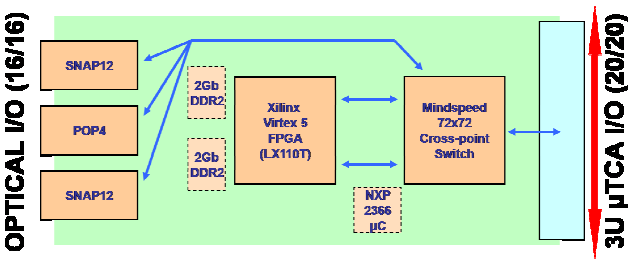

Figure 1: Schematic of the matrix card. 16 input and output channels are provided by the MTP optics on the front of the card while there are 20 channels on the edge connector that plugs into the backplane.

A variety of host functionality is required for an AMC, and this is provided by an NXP LPC2366 micro-controller. The controller is also responsible for programming the FPGA and its corresponding FLASH PROM, and has its own dedicated Ethernet interface, shown in figure 2. This interface is only used for testing. Reprogramming of the board in a crate can also be achieved using I^2C over the backplane.

B. Prototype Testing

A prototype board was received from manufacture in December 2008. Since then the design has been extensively tested. Several minor flaws were discovered in the original design. However none of these were critical, as most of them involved design oversight resulting in missing bias resistors on non-BGA components or configuration lines to the micro-controller. All of these faults were corrected for by board re-work.

The clock system has been tested, including driving it to and from a standard micro-TCA backplane and MCH. No issues have been observed.

The DDR2 memory has been tested at 300MHz (600Mb/pin DDR), with no errors during a 24 hour test period on one of the prototypes. However, the tests so-far carried out are not believed to be thorough enough to guarantee long-term reliability and further study is required.

For the micro-controller, a UDP/IP firmware has been implemented and tested allowing 4MB/s communication with the board (performance is limited by the micro-controller clock frequency). A packetized FIFO interface has been built that connects the micro-controller and the FPGA. Further to this a programming interface has been developed that allows the Xilinx Impact tools to view the FPGA and PROM as devices attached to a parallel cable, whereas in fact the JTAG control is forwarded over a UDP interface to the LPC2366. This allows for seamless reprogramming of the board.

The serial links and cross-point switch have been tested extensively on all channels at a line rate of more than 3Gb/s. Results show a BER of less that one part in 10^{12} at 95% C.L. in most cases. However six of the transmitter links have shown data instability which has been traced to a correlation with noise from the switching regulators on the board. This issue is currently under investigation.

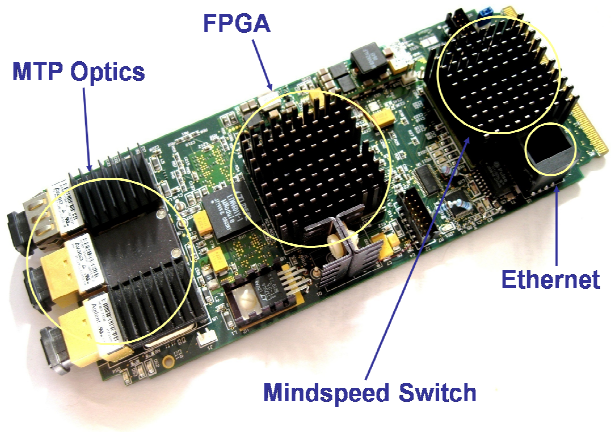

Figure 2: Top view of a Matrix card. MTP optics, the FPGA, cross-point switch and Ethernet interface can be seen as well as various power regulators.

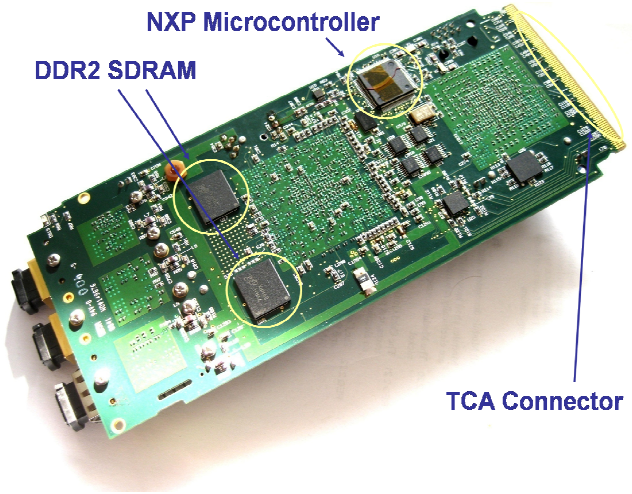

Figure 3: Bottom view of a Matrix card. The DDR2 memory, edge connector and micro-controller are visible. Also note the large number of capacitors.

II. The CMS Trigger Upgrade

It is envisaged that from 2011 onwards the CMS Level 1 trigger will be progressively upgraded to adapt to the physics requirements of the experiment. The Matrix card is expected to play a key role in this process as a development tool for new algorithms. Based on this a new implementation of the trigger system has been studied. This new algorithm will be described in the context of the calorimeter trigger.

A. The Current Calorimeter Trigger

The calorimeter trigger of CMS can be divided into four distinct components: the first of these is the front end of the detector and its corresponding readout system off-detector, which produces trigger primitives (energy clusters) and is therefore called a Trigger Primitive Generator (TPG). There are two kinds of calorimeter TPGs in CMS, those that come from the hadronic calorimeter and those that come from the electromagnetic calorimeter. The second link in the chain is the Regional Calorimeter Trigger (RCT), which performs electron finding and coarse-graining of data. The third component is the Global Calorimeter Trigger, which sorts the electrons by energy and searches for jets using the coarse-grained information from the RCT. Finally the results of this process are passed to the Global Trigger (GT) which makes a decision on whether the data collected about the proton collision is worth saving based on the summary information provided by the GCT and Global Muon Trigger (GMT – not described here). In CMS, the top four candidates ranked by energy of every type of trigger object (jet, electron, etc.) are used to make this decision.

A multi-layered system like this creates several complications when considering changes to the trigger system. Most improvements in processing algorithms will require a corresponding increase in the data density of the processing system, and so ideally one would wish to merge the RCT, GCT and GT into a combined processing unit. The front-end is an exception to this because it its output bandwidth is determined by the capabilities of the on-detector digitisation and readout electronics, which in turn is determined by many factors (e.g. power consumption) that are not critical for the off-detector components of the trigger system. In CMS each stage apart from the front end results in approximately a 20x reduction in data rate. An obvious target for an improved reconstruction path in CMS would be the use of full-resolution information in the reconstruction of jets, as is currently used in the Higher Level Trigger (HLT) [5]. Even with the advent of modern FPGAs with fast serial links, a brute-force attempt at this often runs into several issues [4]. Ultimately, increasing the input bandwidth of a processing system does not resolve scaling issues until the bandwidth of the link technology significantly exceeds the bandwidth requirements of data sharing imposed by the size of a trigger object. As a result of the typical size of a jet in CMS, the data sharing fraction required to contain it is significant. In fact this is a key reason why the RCT and GCT were separated in CMS in the first place, combined with the fact that serial link technology was far slower ten years ago than it is today.

B. A Future Calorimeter Trigger

It is often stated that a serious issue pertaining to the use of serial links in a trigger system is their latency. In the context of the latest generation of modern hardware, this statement can be seen to be incorrect for two reasons:

Firstly, the latest generation of serial links (as found in a Xilinx Virtex-5) are capable of operating in an extremely low-latency mode, using fewer than four bunch-crossings of latency to serialise and de-serialise a parallel data stream. At a 6.5Gb/s line rate, this decreases to a latency similar to that of a standard I/O.

Secondly, serial links were not designed to be used for simplistic processing in geometric fashion where the data remains in a given device (FPGA) for a very brief time

(~50ns). One should attempt to pipeline a processing algorithm and process data within a single device for as long as possible. The name itself implies the correct serial link usage model: *serialise*.

Based on the second of these points, we have considered a radically different topology for a future trigger system. The topology lends itself to comparison with the HLT in CMS, which is a time-multiplexed system where a many PCs are used. A single PC is responsible for processing all the data in any given event over many bunch crossings.

In the current CMS trigger the TPG system receives approximately the same number of fibres as it transmits. Of these each input fibre transmits the data representing a specific detector region for each bunch crossing; in other words, the dimension of time flows through the fibre whereas the dimensions of eta and phi flow across the fibres themselves.

One can imagine a system where this is not the case, but instead the dimension of phi flows through the fibre for a given bunch crossing, and the dimensions of eta and time flow across the output fibres up to a user-defined granularity (see figure 4).

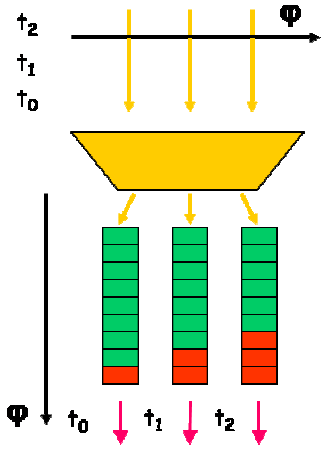

Figure 4: Time-multiplexed serialisation. The input data to the TPGs arrives in phi, eta segments per bunch crossing. A dynamic multiplexer is implemented in the TPG that converts this so that the output fibres have an entire detector segment in phi in each outgoing fibre.

This data ordering cannot be achieved on-detector given the fact its fundamental mode of operation is to capture data in a time-ordered fashion. However the TPG is capable of re-ordering this data in such a way.

This creates a latency penalty equal to the number of bunch crossings delay caused by the multiplexer (which itself is equal to the number of fibres entering the TPG). In typical implementations that have been studied a 16:16 multiplexer was implemented at a resolution of 32 bits per channel. Such an implementation has a synthesised resource utilisation of 2% of an FPGA similar to the one on the Matrix card. The estimated maximum clock speed is so high as to have no effect on any algorithm implemented in the device (Xilinx tools estimate the performance at approximately 1GHz). 32 bit resolution corresponds to a 6.5Gb/s link at a quarter of the byte clock frequency (~160MHz). One of the important advantages of this implementation which will be discussed later is the fact that redundancy can be easily incorporated into such a system by expanding the output bandwidth of the TPGs.

At first it might seem strange to deliberately delay the data processing chain at the start for no obvious gain. However the benefits further along the processing chain more than out-weigh the additional TPG latency.

In the context of CMS, such a system can absorb an entire phi-ring of data from the calorimeter in a single fibre when using a serial link operating at the peak line rate of a Matrix card (3.75Gb/s). Corresponding to 72 towers in phi, this implementation eliminates *all* boundary data sharing in that dimension and therefore also allows the serialisation of the processing algorithms, something that has never been previously achievable. Hence one observes a dramatic improvement in clock speed from the pipelined processing architecture, a task that FPGAs are well suited to.

When considered in equivalent terms, a traditional brute-force approach would result in a data sharing link to input link ratio of approximately 32:1 at 6.5Gb/s line rate for a full granularity processing system in phi and a quarter resolution in eta. The slowest line rate at which the links are even usable in a traditional scheme is 6.5Gb/s due to the data sharing constraints. By contrast the new system requires a sharing ratio of approximately 2:1 at a line rate of 3.75Gb/s. Such a system can be achieved using 16 copies of a 10 Matrix card system, or 160 cards in total. If one desires a full-granularity system, adding a further six cards per partition makes this achievable. For the brute-force approach this would result in approximately a 129:1 data sharing ratio and several thousand processing cards, which is completely infeasible. Table 1 shows the relationship between line rate and data sharing for each architecture.

Table 1: Data sharing ratios for different link speeds and architectures at CMS. These calculations assume a processing card with sixteen inputs and sixteen outputs, like the Matrix card. The numbers in brackets are the number of processing cards required for the implementation of a full trigger system.

	3.75Gb/s	6.5Gb/s
Serialised, partial granularity	1.82 (160)	1.27 (64)
Non-serialised, partial granularity	N/A	32 (1440)
Serialised, full granularity	2.91 (256)	1.39 (64)
Non-serialised, full granularity	N/A	129 (5544)

A unique feature of this new approach is that the trigger system after the TPGs is effectively split into N identical modules (most likely individual processing crates), one of which might look like the one shown in figure 5.

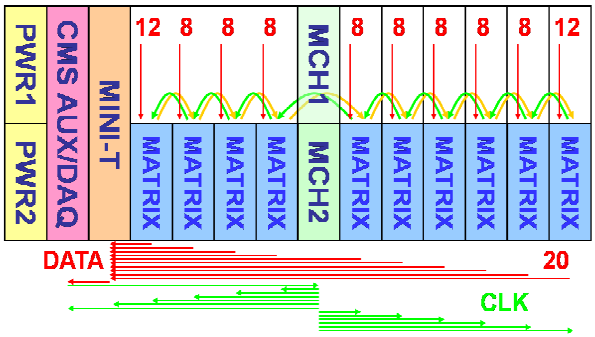

Figure 5: A possible configuration for a trigger partition in a new trigger system. The input data from the TPGs for a given time slice would be received through 88 fibres into 10 matrix cards. Each card would share 4 fibres with each nearest neighbour, corresponding to the overlap region of a coarse-grained jet finder. The results would be sent to a final decision card for sorting. A matrix card would be inappropriate for the final decision card as it has too few input links, so a new board called the Mini-T is under construction that has a higher link capacity. In addition, an auxiliary card is required to provide CMS interfaces.

This system conveys several advantages:

- **System redundancy** – by providing additional spare output channels at the TPG, backup crates can be included that take over from a failed partition at run-time. Furthermore if one does fail, it results in increased trigger dead time rather than a blind spot in the detector.
- **System reliability** – reduced data sharing requirements lower the demands on system connectivity. Ultra high-speed links are harder to manufacture and use, and should be avoided if possible.
- **Capacity for future expansion** – corresponding to the previous point, the lower line rate / fibre usage provides room for the addition of muon and tracker information in the future.
- **Separate testing partitions** – during periods when the LHC beam is not available, the trigger can be split into its partitions. Rather than requiring an individual sub-detector to only use their system component, a full trigger chain can be made available to each one for a 'slice' of the time. Furthermore the full trigger chain can be easily tested in a small setup at an individual institution.
- **Ease of understanding** – the system is only partitioned in one dimension, making the individual processing elements far easier to understand.
- **Processing speed** – while the initial multiplexer loses most likely eight or sixteen bunch crossings by serialisation of the data, the final sorting algorithm gains a similar performance benefit, negating the effect. Furthermore the serialisation of the processing algorithm results in significantly higher clock speed. In the GCT the processing system runs at 40MHz. Studies of the new system show that it will operate at over 200MHz.

III. Time Synchronisation at FERMI

FERMI is a 4th generation Free Electron Laser (FEL) light source, currently under construction at the Synchrotrone Trieste site in Italy [6]. It operates as an approximately 3GHz RF system with a few components operating at approximately 12GHz using Travelling Wave Tubes (TWTs) for electron acceleration. As with all FELs, the quality of the light source is directly dependent on the accuracy of the phase and amplitude of the power driving each TWT in the system. In Trieste these constraints are very difficult to achieve, with a specification of less than 0.1 degrees error in phase relative to a master time reference and less than 0.1% in amplitude per cavity. 0.1 degrees of a 3GHz system corresponds to approximately 300fs, and so a timing precision greater than this must be achieved.

When converted to its master reference frequency, the system clock is approximately 2.4GHz. This is in an ideal operating frequency for a Xilinx gigabit transceiver, and so is used directly to provide a star-topology control system with the Matrix card at its centre. It is envisaged that this central processing system will be able to calibrate itself by measuring the loop propagation delay through a bi-directional optical link to each RF station. Knowing this it is theoretically possible to re-phase all the RF stations such that the control system is aligned to within 50ps at all stations. This has been achieved with an accuracy of 300ps but so far there is an error of 1UI which appears to be caused by the internal operation of the Xilinx GTPs. However, this already greatly exceeds the requirements for the operation of the control system (4ns resolution). Furthermore it is believed that using the Matrix card, the GTPs can be substituted with LVDS I/O operating at up to 1.25Gb/s, which have a completely deterministic behaviour.

The advantage of this approach is that any variation in the propagation through a fibre in one direction will likely correspond to the change in propagation time in the other direction (for example due to temperature variation). While the absolute limits of this approach are not yet known, for many applications the current results are already more accurate than necessary. One important detail of this approach though is that the reference clock at each end of the system must have a constant phase relationship. Therefore one must either have a reliable global clock network or a local VCXO at the slave end that can be locked to the recovered clock from the serial link. In figure 6 the first of these approaches is shown; the LLRF stations in Fermi also have a high-performance OCXO on each station that can be used instead of a global clock network.

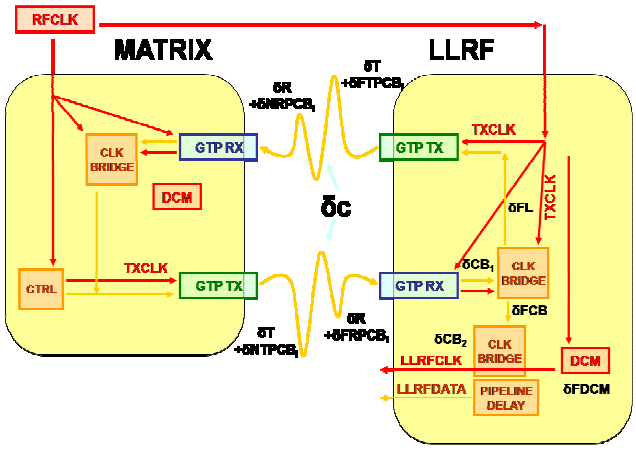

Figure 6: The timing synchronisation system for the Matrix – Low Level RF (LLRF) system at FERMI. The cable delay is expected to be equal in each direction so by measuring the loop time through the system one can accurately calibrate and correct for it.

IV. Conclusions

The Matrix card has been tested thoroughly and shown to operate at its anticipated maximum performance with few problems. Most of these issues have been resolved in a revision of the board that was recently received from manufacture. It remains to be seen whether the board revision also resolves the issues seen with six serial link transmitters.

A L1 trigger architecture has been developed that uses the Matrix card as its template and that results in a significant performance improvement over previous generations of hardware. It is expected that within the next year a development platform will be built based on this architecture that uses the Matrix card or one of its successors, and processing algorithms will be demonstrated.

The FERMI light source is currently undergoing installation and commissioning and it is expected to begin operation by the end of 2010.

V. Acknowledgements

The authors acknowledge the prior work of Matt Stettler and John Power at LANL in the development of the Matrix card hardware and the involvement of Tony Rohlev at Synchrotrone Trieste in the conception of the LLRF control system.

VI. References

[1] M. Stettler et al., “The GCT Muon and Quiet Bit System, Design Production and Status”, TWEPP, September 2008, Naxos, Greece.

[2] M. Stettler et al., “Modular Trigger Processing, the LHC GCT Muon and Quiet Bit System”, IEEE NSS, October 2008, Dresden, Germany.

[3] J. Jones et al., “DAQ and Control Interfaces for the CMS Global Calorimeter Trigger Matrix Processor”, IEEE NSS, October 2008, Dresden, Germany.

[4] G. Iles et al., “Trigger R&D for CMS at SLHC”, TWEPP, September 2009, Paris, France.

[5] G. Bagliesi, “CMS High-Level Trigger Selection”, Eur. Phys. J. C 33 (2004) s1035-s1037.

[6] T. Rohlev et al., “Sub-Nanosecond Machine Timing and Frequency Distribution via Serial Data Links”, TWEPP, September 2008, Naxos, Greece.