PCI Express Over Optical Links for Data Acquisition and Control

M. Bellato, R. Isocrate, G. Meng, M. Passaseo, G. Rampazzo, A. Triossi, S.Ventura

INFN Sezione di Padova, Italy

marco.bellato@pd.infn.it

Abstract

PCI Express is a new I/O technology for desktop, mobile, server and communications platforms designed to allow increasing levels of computer system performance. The serial nature of its links and the packet based protocols allows an easy geographical decoupling of a peripheral device. We have investigated the possibility of using an optical physical layer for the PCI Express, and we have built a bus adapter which can bridge remote busses (> 100m) to a single host computer without even the need of a specialized driver, given the legacy PCI compatibility of the PCI Express hardware. This adapter has been made tolerant to harsh environmental conditions, like strong magnetic fields or radiation fluxes, as the data acquisition needs of high energy physics experiments often require.

I. Introduction

The electronics developed for LHC detectors is expected to operate in a intense radiation field. This aspect has to be taken into account also when designing a board that will be located in the periphery of a detector. Concerning the CMS experiment, our practical need is to link the counting room with the VME creates placed on the balcony of the five barrel wheels. The radiation dose absorbed after ten years of operation in this environment is less than 0.2 Gy and the expected neutron fluence is not high enough to generate a relevant bulk damage (less than $2.5 \cdot 10^{10}$ n/cm^2) [1]. Nevertheless the detector electronics could still be fooled or even damaged because of Single Event Effects (SEE).

Figure 1: Block diagram of the hardware setup.

To cover the long distance (about 100 m) between the control room and the detector electronics we chose to manufacture an adapter that translates PCI express signals to/from the optical physical layer (Fig. 1) and which, using commercial bridges, could be fitted into legacy bus standards (PCI, CompactPci, VME). PCI Express is a new local bus generation using existing PCI programming concepts: it represents a radical move from traditional I/O architectures in that it replaces parallel multi-drop busses with serial switched point-to-point links. The link bandwidth is committed to the number of lane (individual serial pairs). Every lane utilizes two LVDS signals at 2.5 Gbit/s per direction so the capacity of the channel after 8b/10b encoding is fixed at 250 MByte/s times the number of lanes [2].

PCI Express adopts a "communication centric" approach: the load-store operations between two nodes are performed exchanging framed packets in accordance to a suite of stacked protocol layers taking care of the physical, link and transaction issues of the channel. In case of PCI Express all these activities are carried out at the hardware level, with no software intervention. Clearly this load-store model logically matches the model of field bus control in which a host and a networked peer node exchange software arranged packets to access memory and registers of the field bus for I/O operation. As the PCI Express base specifications version 1.0a make no assumption about the nature of the interconnection medium, to investigate the PCI Express I/O bus used like a distance communication protocol running on an optical fiber link we addressed a number of research activities in the context of the project LINCO (INFN Gr. V).

II. The New Physical Medium

The first goal of LINCO project is a feasibility study of an optical layer for the PCI Express protocol. The problems raised by the adoption of fiber as physical medium are mainly two: the jitter that, being specified both at the transmitter and the receiver implicitly constrains the in-between medium, and the reference clock distribution.

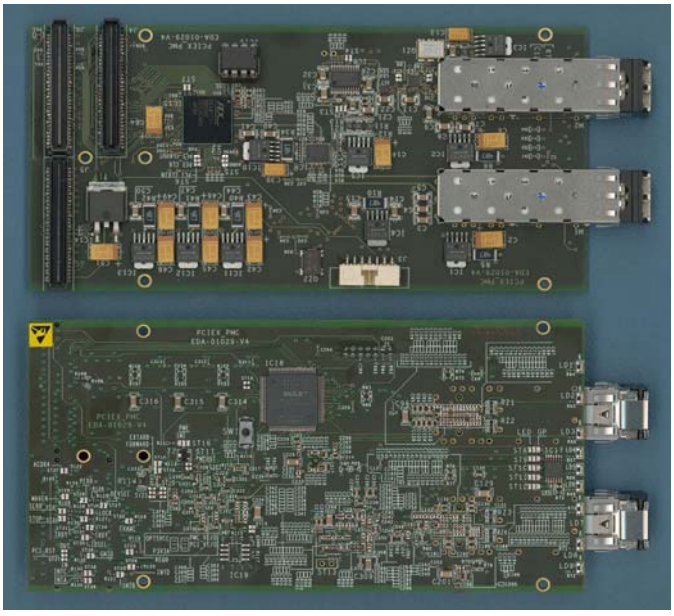

Figure 2: Picture of the board.

The figure of total jitter reserved for the interconnection, as fixed by the specifications, is a maximum of 0.3 Unit

Intervals (UI), e.g. 120 ps of total (random plus deterministic) jitter at a Bit Error Ratio (BER) level of 10^{-12} [3]. This figure refers to the jitter introduced by the chain of transmission side connector, electrical-to-optical conversion, fiber, optical-to-electrical conversion and receiver side connector. To this regard, important design parameters are related to the choice of the transceivers because their contribution to the overall jitter is in the critical path. For this reason we qualified for our application two types of transceivers: an Intel transceiver Infiniband compliant (2.5 Gbit/s) in a multimode fiber and a JDS Uniphase multirate (up to 2.7 Gbit/s) transceiver for OC-48 operation on a 1310 nm single mode fiber.

The other issue to face was designing the reference clock distribution. In fact the PCI Express reference clock (100 MHz) is usually supplied by the host board to the add-in cards but the specifications do not mandate its presence [4]: in principle a local supplied 100 MHZ ± 300 ppm clock should suffice to properly reconstruct the incoming stream. In practice, our tests revealed the need of bridging also the reference clock signal. A frequency offset in the clock generator of the PC host board we used showed that it may be impossible to lock the data signal. For these reasons we deserved a separate optical transmitter and fiber to broadcast the host reference clock.

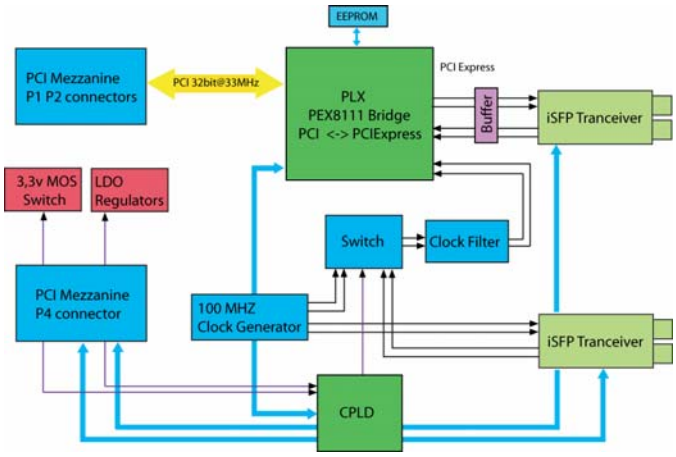

Figure 3: Block diagram of the mezzanine board.

III. Board Description

We chose the Pci Mezzanine Card (PMC) form factor (Fig. 2) in order to meet the two different bus standards the board will be inserted (CompactPCI and VME crate). A key component of the board is a PLX Technologies bridge (PEX 8111) that allows transparent bridging between a PCI bus and a PCI Express x1 lane (Fig 3). This kind of bridge can be configured as a device that can tunnel configuration cycles from the PCI side to the PCI Express channel (in our application when it sits in the control PC) or vice versa (when it is used in the remote VME field bus). Also the clock path is dependent on configuration indeed a Micrel SY58023U high speed switch is used to broadcast either the local generated clean 100MHz clock or the received one, both to the PEX8111 bridge and to the optical transmitter. A PLL with a small loop bandwidth is used to filter the jitter of the reference clock and to provide the proper differential and common mode voltages as per the PCI Express specifications. The presence of the optical transceiver in the data transmission path force us to use a Pericom PI2EQX4401 buffer to comply with the PCI Express receiver detect mechanism. An Altera programmable logic device is connected to the general purpose I/O bus of the PEX8111 and to the gate of a high power MOS switch; the use of this device will be illustrated in the CMS application section.

IV. Board Test

A direct measure of the voltage and timing margins at the entrance and exit of the electrical-optical-electrical chain, taken with an Agilent 54855A 6GHz real-time scope, shows that the contribution of the medium to total jitter is respectively of 4.88 ± 0.21 ps (Fig. 4). In order to measure the eye opening at a 10^{-12} we used the tail fit method [5]. We extracted the jitter's Probability Density Function (PDF) from the histogram of the data Time Interval Error (TIE) i.e. the time difference between the recovered clock (from the data stream) and the data signal. Since the deterministic jitter PDF is bounded below a certain jitter range, all the PDF comprises just random jitter processes. Through a Gaussian fit on these tail regions we extrapolated the PDF analytic function concerning rare events. Fig. 5 gives an illustration on the relationship between jitter PDF and BER that can be considered as the cumulative distribution function, essentially the integral of the PDF function.

From the so-called "bathtub" curve of the BER (showed in Fig. 5), we deduced the real eye opening as the time difference between its branches corresponding to a 10^{-12} BER level. From this extrapolation it turns out that both transceivers are compliant with the minimum eye opening fixed by specifications (183 ps), with the Intel transceiver outperforming the competitor with a measure of 243.69 ± 0.14 ps.

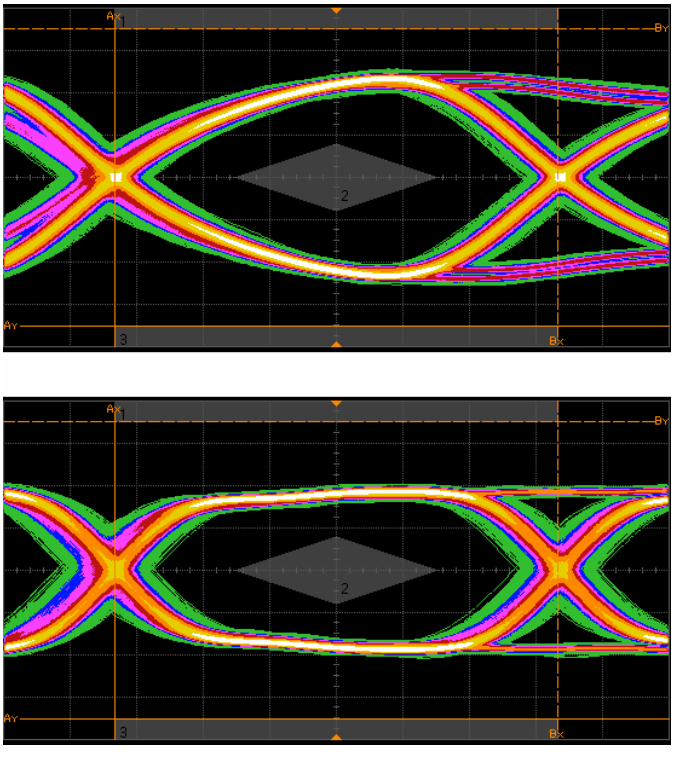

Figure 4: Contribution of the optical medium to signal jitter.

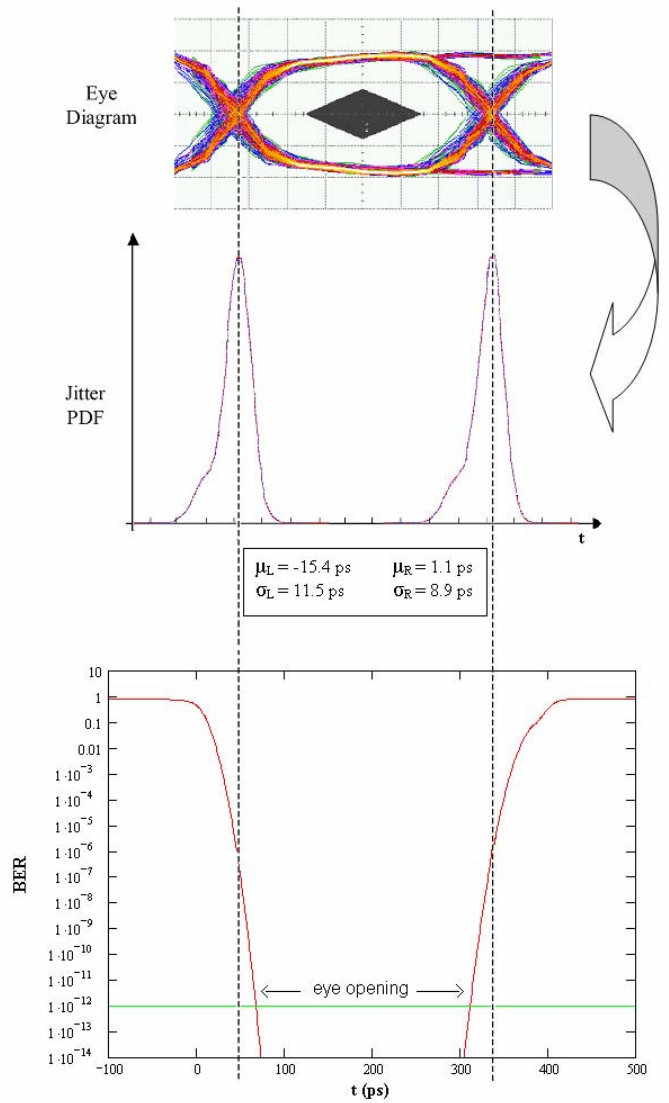

Figure 5: Modeling of Probability Density Function of jitter from measured data.

With the aim of understanding this difference, further tests were done on the data signals [6]. We used a real time individual jitter component separation technique, based on the examination of the TIE trend and its Fourier transform. For the estimation of the two jitter components probability density functions we assumed a dual Dirac delta model [7] that relies on two basic assumptions:

- Random jitter PDF follows a Gaussian distribution (hence fully described by its width σ),
- Deterministic jitter PDF follows a finite distribution formed by two Dirac delta functions separated by the entire amount of deterministic jitter.

The dual Dirac model is a simplistic model for jitter distribution. In fact it fits a real world jitter distribution only in presence of pure duty cycle distortion. In our case we obtained a better representation of the deterministic jitter PDF by deconvolving the PDF of the random jitter from the measured probability density function of total jitter. This method yields a more realistic evaluation of the deterministic jitter with respect to the dual Dirac. The only disparity between transceivers turned out to be the amount of deterministic jitter: the JDS device adds more than 30 ps p-p to this parameter with respect to the Intel one. The reason of this addition has to be charged to increased crosstalk levels due to the faster rise time of the JDS device.

We configured two prototypes: one for local and the other for remote operation. The first has been fit into a host PC, with passive adapter, while the second has been tested in a CompactPCI crate. The two boards were linked by a 100m multimode fiber with Intel transceivers. The BIOS could take control of the remote PCI bus without any driver and we could transfer data at the full legacy PCI throughput. Besides the jitter in the reference clock path has been measured holding a safe 85 ps (Fig. 6). Another test was done using a remote VME crate with an active adapter hosting a Tundra UniverseII PCI to VME bridge. In this case transfers were achieved on a VME target with a 3 to 4μs single access latency.

The only disparity between the PCI Express data signal reshaped at the end of the optical chain and the local one is the loss of the de-emphasis (e.g. a feature that affects a transmitter driving a trace with more then one bit of the same polarity; by lowering the voltage level of 3.5 dB with respect to the first bit it avoids an excessive charge of the parasitic capacitances of the transmission line). The limiting amplifier, which is in the optical chain, nullifies the de-emphasis making the optical interconnection not fully PCI Express compliant.

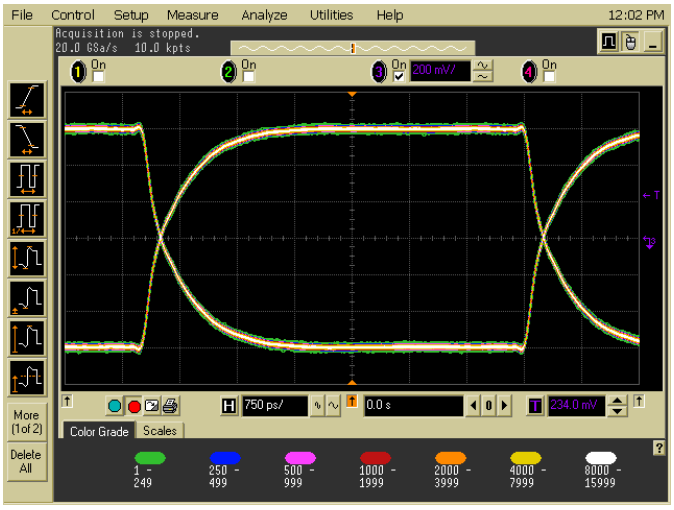

Figure 6: Remote reference clock.

V. CMS Environment

As reported above the operational environment of our PMC board presents harmful conditions. Due to the strong magnetic field (about 1 Tesla at the periphery of the detector [1]) we have been forced to avoid the use of core inductors. Usual LC filters in the power supplies have been replaced with lossy RC filters and the voltage drop associated with them has forced a careful design of the power distribution to the different devices of the board. For these reasons a proper decoupling strategy has been extensively analyzed with a full-wave electromagnetic field simulation software (Ansoft SiWave). The S-parameters of the high speed traces have been extracted from simulations and used in a Hspice program

to predict timing and voltage margins of the PCI Express signal at the input of the optical transceiver.

To qualify our design to Single Event Effects (SEE) we have setup a radiation test at the Paul Scherrer Institute in Zurich with protons of 60 MeV energy. The fluence measured on the board was $5 \cdot 10^{10}$ p/cm^2 and the board has been exposed for 8 hours, corresponding to a total ionizing dose of more than 6 Krad (ten years of LHC operation). In the picture (Fig. 7) you can see our board hosted by a commercial Tundra VME bridge. During the test, these boards were active and a certain number of VME registers were continuously written with random patterns, read back and compared for Single Event Upset checking and logging. Furthermore we checked the absorbed total current in order to verify the occurrence of Single Event Latch-up (SEL) events.

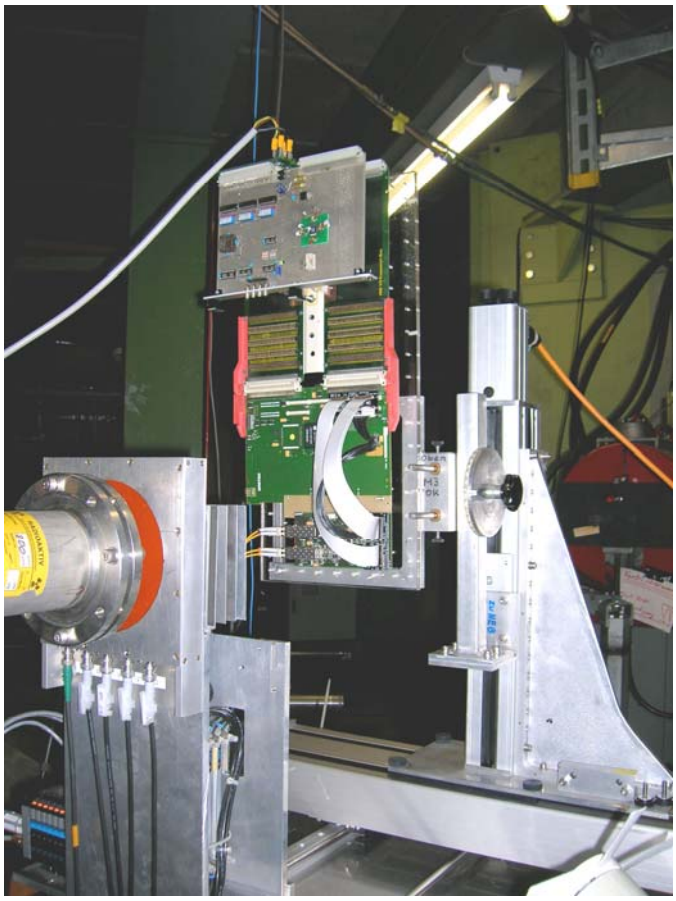

Figure 7: Radiation test setup at PSI.

We didn't observe any SEU, the only relevant effect being the maximum radiation dose absorbed by the Intel transceivers before latching-up. To ease operation at LHC, a watchdog-type mechanism for automatic reset has been put in place using the Altera programmable logic device on the board. The watchdog drives the gate of the high current MOS switch in series with the main power supply of the board and it is continuously reset through the GPIO bus of the PLX bridge. The GPIO bus is in turn periodically toggled by the remote host through the optical link; when a single event functional interrupt (SEFI) occurs on the remote board that stops the communication with the host, the watchdog doesn't get toggled any more and this results in a power off/on cycle, e.g. a cold reset that restores the PCI Express link functionality. Furthermore we continuously monitor the transceivers life parameters, like laser bias current, temperature and optical transmitted and received power, through I^{2}C protocol on GPIO bus.

VI. Conclusions

The implementation of a PMC board has been used both as a proof of concept and a test-bed for measurements of PCI Express compliance. In particular our tests show that the optical transport medium is very effective from the point of view of functionality and a modest jitter difference was found between the signals before and after the optoelectronic chain. Finally the application in the LHC harsh environmental condition doesn't seem to be an issue for our board.

VII. References

[1] CMS Collaborations, The Tracker Project Technical Design Report, CMS TDR 5, CERN/LHCC/98-6.

[2] MindShare, Inc, Ravi Budruk, Don Anderson, Tom Shanley, PCI Express System Architecture, Addison Wesley, 2003.

[3] PCI-SIG, Pci Express Base Specification 1.0a, 2003, www.pcisig.com.

[4] PCI-SIG PCI Express Card Electromechanical Specification Revision 1.0a, 2003, www.pcisig.com.

[5] M. Li, J. Wilstrup, Paradigm Shift For Jitter and Noise in Design and Test > 1Gb/s Comunication System, Proceedings of the 21st International Conference on Computer Design, 2003.

[6] A. Triossi, Link PCI Express su fibra ottica come interconnessione tra bus, Università degli Studi di Padova, 2005.

[7] Agilent Technologies, R. Stephens, Jitter Analysis: The Dual-Dirac Model, RJ/DJ, and Q-Scale, 2004, www.agilent.com.