# Common Gigabit Ethernet interfaces for HLT and L1-trigger links of LHCb

H.Müller, F.Bal, A.Guirao

CERN, 1211 Geneva 23, Switzerland
Hans. Müller @cern.ch

## *Abstract*

The LHCb experiment has standardized on Gigabit Ethernet (GBE) and the Internet Protocol (IPv4) for the transport of trigger data streams between L1 buffers and a switch-based readout network. The High Level Trigger (HLT) operates at low rate and large event sizes, the Level-1 Trigger (L1T) at high rate and small event size. In order that both streams can be transmitted from a single FPGA associated to the L1 buffers, an LHCb-specific GBE mezzanine was specified and built. Interfaced via the 100 MHz industry SPI-3 bus, a Medium Access Controller (MAC) on the mezzanine transmits IPv4 packets which originate from the FPGA and directs these towards 2 or 4 physical GBE links according to the separate streams. Up to 4 GBE links are required per mezzanine for handling the large data volumes of some detectors. The GBE mezzanine connector definition of LHCb includes also a 16-bit @ 10 MHz control interface for the configuration of the MAC chip via a local embedded PC. The first, 2-channel GBE mezzanine reference design of LHCb is presented here.

## I. DATA STREAMS IN THE LHCB EXPERIMENT

In the LHCb experiment, both HLT and L1T data streams are transmitted from the same common hardware source, (Tell-1 [1]). L1 event fragments are promptly transmitted whilst HLT events are queued in their L1 latency buffers until a Level1-Yes trigger decision is received. An FPGA driver assembles HLT and L1T data fragments into Multiple Event Packets (MEP's [2]), adds LHCb-specific transport formats [3] under IPv4 framing conventions and transmits the MEP's as separate data streams via the common SPI-3 bus [4] to the multi-channel GBE mezzanine cards [5] described in this paper. Future implementations of such GBE mezzanine cards have to comply with the LHCb connector conventions for the SPI-3 and control bus. The first reference design is depicted in Fig.1). This 2-channel mezzanine contains two bi-directional Rx/Tx 1000Base-T copper channels as well as an option for 8b/10b coded 1000Base-LX fibre. As a follow-up project, a 4-channel Rx/Tx GBE mezzanine design is under way.

The stream controller in the L1 buffer modules is to be implemented in FPGA logic as IPv4 packet driver which generates two, quasi simultaneous data streams consisting of MEP packets, transmitted via the SPI-3 bus to the GBE mezzanine. The L1T stream is assembled from 1 MHz raw input event fragments of the L1 buffer modules. Up to 32 such event fragments are grouped into MEP's and formatted as IPv4 frames. Hence individual MEP's of the L1T data stream are transmitted at a modest rate down to 34 kHz. The MEP driver assigns the destination addresses for each MEP after receiving it from the LHCb Readout Supervisor [6] via long broadcasts of the TTC optical network [7]. Each MEP has to be dynamically addressed towards an subfarm node of the combined CPU farm for HLT and L1T. The HLT stream is generated in a similar way as L1T, only that IP packets are made from the buffered and accepted L1 event data, hence at a lower rate. The HLT and L1T links are connected to distinct router switches of the LHCb switching network. They are directed, via the IPv4 address mechanism, to free resources of the CPU farm which consists of O(100) sub-farms, each of which contains O(15) CPU's. Since the number of sub-farms corresponds to a relatively small range of less than 10 bit, the dynamic destination allocation by the FPGA driver merely requires updating of the 10bit LSB in the IPv4 destination address. Hence for each MEP, 10 bits are received from the Readout Supervisor via the TTC system.

The envisaged LHCb push architecture for the L1T and HLT data streams requires that the FPGA driver writes to the "Egress FIFO" of the MAC controller on the GBE card. In case that a pull-architecture is required, the full bandwidth of GBE in the opposite (Ingress or Rx) direction is available between all sub-farms and the L1 source modules.

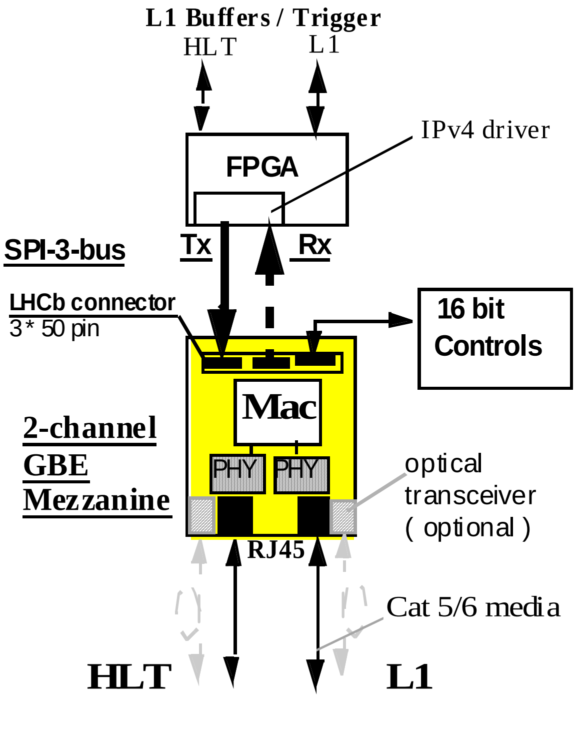


Figure 1: Definition of the 2-channel GBE mezzanine

On all Tell-1 boards equipped with GBE mezzanines of type A, the Rx direction is available for the debugging purposes with standard computing equipment.

## A. *SPI-3 bus*

The SPI-3 bus [4] is used in the networking industry as OC-48 level packet interface between link layer devices and physical link controllers. In the LHCb experiment, the link layer device is an FPGA and the physical link controller is a MAC chip on the GBE mezzanine (Fig. 2). Ethernet packets are made in the FPGA and sent via the SPI-3 bus to the MAC chip.

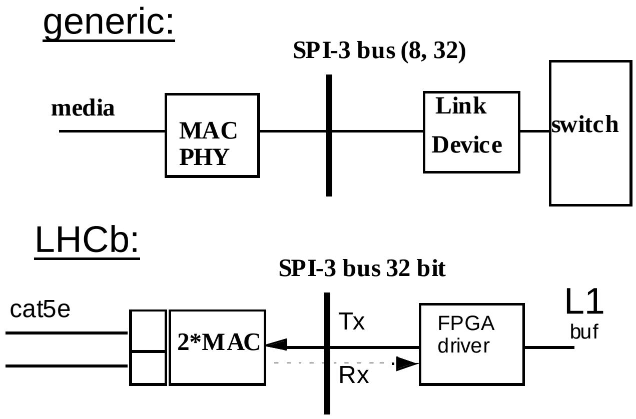


Figure 2: Industry SPI-3 protocol in LHCb

The implementation of an FPGA-based IPv4 driver for Gigabit Ethernet is much facilitated by using the SPI-3 protocol for OC48c networking equipment (OC48 = 2.488 Gbps). A single 32 bit SPI-3 bus, operating at 104 MHz allows quasi simultaneous use of two physical GBE links. The protocol is based on one-directional "master write" operations for each of the two GBE directions. This concept provides that latencies are low, bandwidth utilization is high and that only a single direction can be implemented if needed. The "in band address selection" of the physical link is inherent to SPI-3 and allows taking full advantage of the SPI-3 bandwidth for directing IPv4 packets to their different physical links.

## B. *LHCb connector for SPI-3*

SPI-3 is a FIFO-like protocol with a low pin count (~50) for each direction. The common connector for GBE mezzanines consists of 3 blocks (50 pins each) for Rx, Tx and Control (Fig. 3). The SAMTEC QSS connector with 50 OHM transfer impedance is very well suited for SPI-3 clock rates beyond 100 MHz. The central bus-bar contacts on the two SPI-3 blocks guarantee good signal ground and the bus-bar on the control block provides a low impedance 3.3 Volt supply to the mezzanine. Voltages different from 3.3 V have to be generated via onboard DC-DC converters.

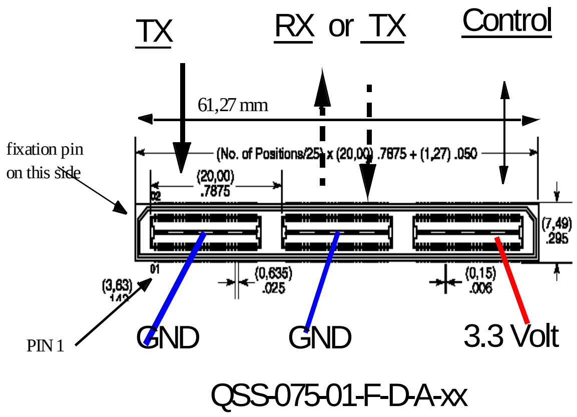


Figure 3: LHCb connector on mezzanine for Rx , Tx and Control

The control interface block is used for a generic 16-bit bus for configuring the control and status registers of the MAC or PHY chips. Three transfer types (Tx), (Tx + Rx), (Tx + Tx) can be implemented in two 2 types of mezzanine cards: type A = (Tx + Rx) and type B = (Tx + Tx). Type A is the bi-directional LHCb Reference Design for 2 GBE channels described here. Type B can be designed for very high bandwidth requirements over multiple GBE links in a push-only application.

## C. *Gigabit Ethernet*

Each PHY chip handles four PAM-encoded 250 MHz twisted pairs as part of 8-wire CAT5e or CAT 6 cables, resulting in a total bandwidth of 1 Gbit/s in both directions (Fig 4). The transmission standard over copper is called 1000 BASE-T symbolizing that the Rx direction uses the same line as the transmit one (T).

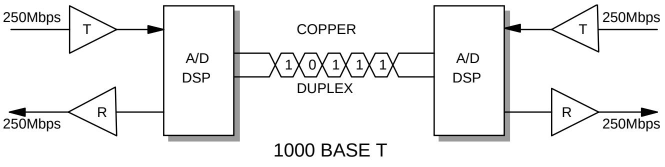


Figure 4: Physical 1000 Base –T over one of four 250 Gbps links

The guaranteed distance specified by the IEEE802.3ab standard for un-shielded CAT5 cable is 100 m for a BER figure of less than 10E-10. Using quality cables and advanced PHY controllers, distances of 180 m are possible [10]. Direct serialization out of the MAC chips via integrated SER-DES logic also enables the alternative use of optical fibre links (1000BASE-LX). In this case, the PHY chips are redundant and the SERDES ports have to be equipped with 850 nm optical fibre GBE transceivers. The 2-channel GBE reference design has both options.

## D. *Packing of MEP's into IPv4 frames*

The FPGA driver generates one IPv4 packet from one MEP packet, following the exact rules for data fields within IPv4 packets. Fig.5 illustrates details: An address/destination

Ethernet header is followed by a data payload of 46…1500 byte and a CRC check. For packets with more than 1500 byte, the Ethernet “Length” field becomes “Type of protocol” which means that the expanded IPv4 packet conventions apply and a higher level format header is contained in the first 20 bytes of the original Ethernet data field. The Length/Type field takes one of two meanings depending on its numerical value. The higher protocol applies if the numerical value is greater than 0x0600. After the 20 byte of the higher protocol header follows a 14 byte header, used as data descriptor of LHCb for one MEP. The following data block contains up to ca. 1500 byte of concatenated and event-numbered event fragments. The maximum amount of MEP data carried by the IPv4 protocol is 64 kbyte, conveniently large since an average MEP within the L1T stream contains ~2 kbyte. Special partitions may however contain up to 22 kbyte corresponding at 1.1 MHz input rate to 99% of a single GBE load. These partitions require that a second L1T link is used, hence calls for the quad channel card design.

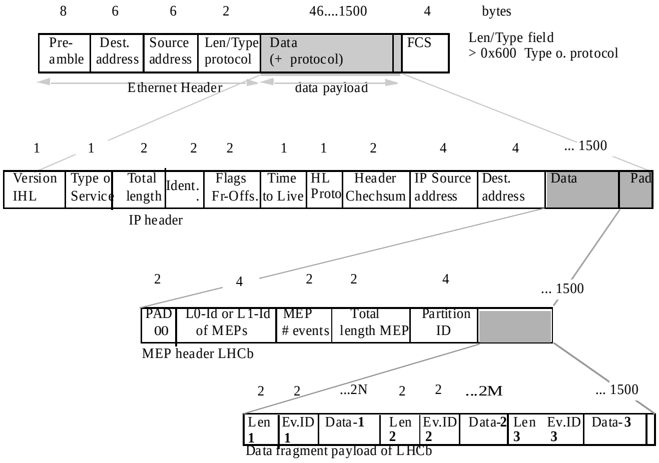


Figure 5: Ethernet / IPv4 packet with one MEP and LHCb data

### *E. GBE Medium Access Controller (MAC)*

The MAC chip[8], chosen for the reference design is a two-port full-duplex Gigabit Ethernet Controller with an industry standard SPI- 3 system interface. It directly connects either to fibre optic transceivers via two internal Serializer/Deserializer (SERDES) or to physical layer devices for copper via two GMII interfaces. The SPI-3 Tx and Rx buses connect to internal 16kbyte egress and 64kbyte ingress FIFOs per channel.

### *F. Physical link Controller (PHY)*

A dual PHY controller chip [9] is connected to the MAC chip via the Gigabit Media Independent Interface (GMII) bus. The PHY chip handles 1000BASE-T and also 100BASE-TX (fast Ethernet). An auto-negotiation protocol selects the speed according to the connected media. According to the manufacturer, the DSP-based internal error cancellation performance allows for a copper cable length of up to 180 m [10].

## II. 2-CHANNEL MEZZANINE PROTOTYPE

The layout of the 2-channel Gigabit Ethernet prototype is shown in Fig. 6.

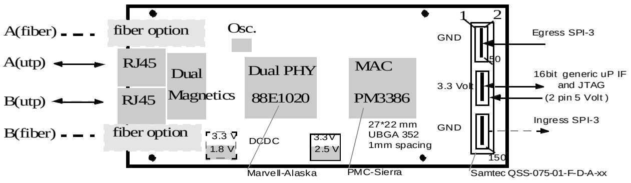


Figure 6: layout of the 2 channel reference mezzanine

This reference design uses dual MAC and dual PHY chips which greatly reduces power and chip-surface on a 74 * 149 mm mezzanine card. Fig.7 shows the photo of the first

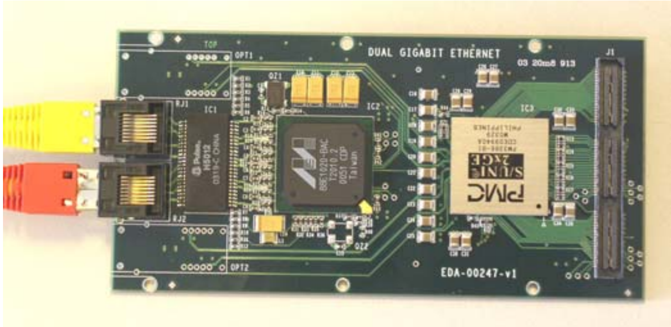


Figure 7: Photo of the 2-channel GBE reference mezzanine

2-channel GBE card prototype, equipped with RJ45 connectors for 1000Base-T operation. The emplacements for the optical transceivers (multimode LC fibre optics) are visible next to the RJ45 plugs.

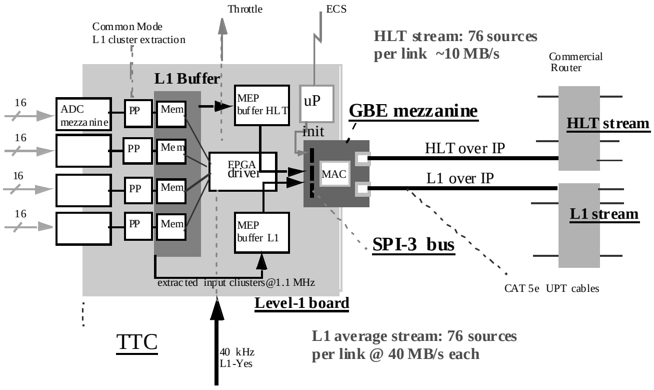


Figure 8: GBE mezzanine in LHCb VELO readout system

## III. EXAMPLE: VELO DETECTOR READOUT

An example application for the dual GBE card is the development of an IP packet driver for the HLT and L1T data streams for the Vertex Locator (VELO). Each VELO L1T link produces ~ 40 Mbyte/s data streams. Analogue detector data are received at 1 MHz by 76 Tell-1 [1] boards, each handling 64 input channels. The input fragments are event-wise processed and stored in a Level-1 latency buffer which can contain up to 58.000 input events. On reception of a L1 trigger-Yes via the TTC link, several kHz of selected HLT fragments are passed to the FPGA driver. Also under control of the TTC receiver, the FPGA driver generates two Gigabit streams with encapsulated MEP packets, formatted according to IPv4 standard of Fig.5.

## IV. TEST AND RESULTS

The Aroc PCI card [11] is used as an FPGA-based test platform since it is hardware-wise very similar to the output stage of the Tell-1 board. Fig.9 shows details of the test environment for which 2 PC's were used.

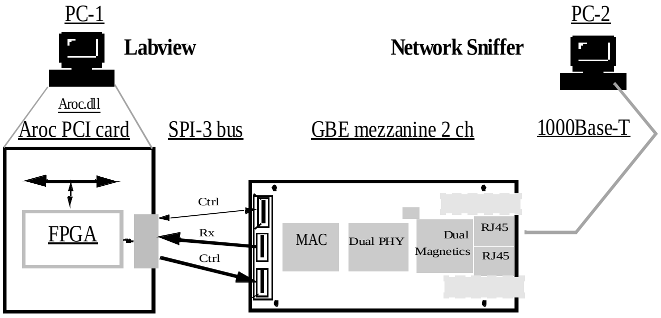
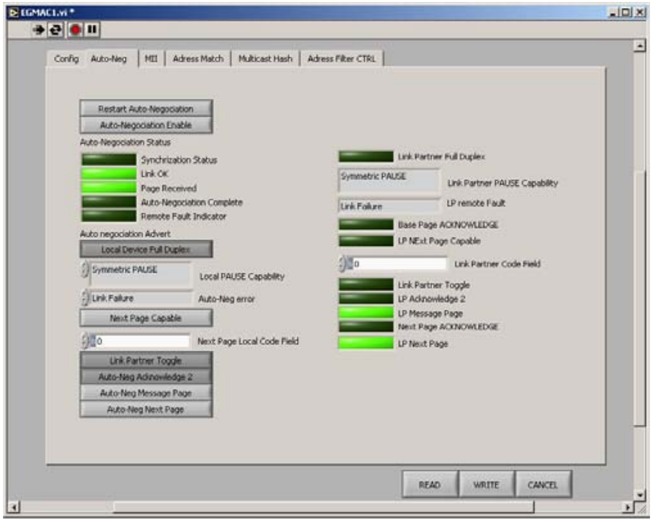


Figure 9: GBE card test environment on Aroc PCI card

A Labview test environment [12] in PC-1 communicates via a PCI driver (Aroc.dll) with the FPGA resources of the Aroc PCI card which carries a GBE mezzanine. In this way, Labview allows configuring the MAC and PHY registers and to write test-data to it's egress FiFos. In real time tests, the FPGA driver on the Aroc writes IPv4 packets via the SPI-3 bus to the GBE mezzanine. The FSC trailer is added by the MAC chip. On the receiving side, a PC equipped with a standard GBE card and Network Analyzer software like "Ethereal" [13] is used to detect, verify and analyse the received IPv4 packets.

### G. Register control interface

The Labview test screen of Fig.10 transfers user parameters and commands via a PCI driver (Aroc.dll) to FPGA-internal control logic which communicates with the GBE card. A generic, 16 bit control bus on the one LHCb connector block (Fig. 3) maps 16 bit R/W register addresses into a 32 bit PCI address space.

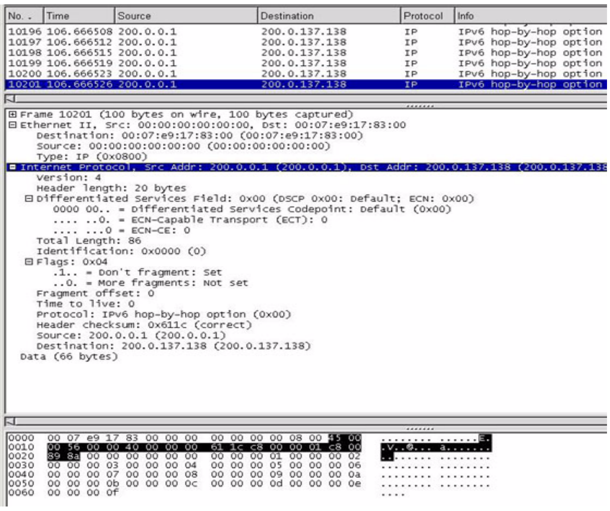


Symmetric PAUSE Local PAUSE Capability
Link Failure Auto-Neg error
0 Next Page Local Code Field
Symmetric PAUSE Link Partner PAUSE Capability
Link Failure LP remote Fault
0 Link Partner Code Field

Figure 10: Labview configuration of MAC registers

### H. First generated IP Packets

After configuration of registers via Labview, the FPGA-based IP driver can send IPv4 packets via the SPI-3 bus interface to the MAC chip's egress FiFos. The MAC verifies and completes the packets with a CRC-32 frame check sequence (FCS field) before transmitting them via the GMII interface to the PHY chip. The latter transmits them over magnetically coupled Cat5e cables to an external PC which is equipped with commercial GBE card and network analyser software. The Ethereal screen dump of the first IP packets which were recorded in an externally addressed PC is shown in Fig. 11.



```
No. .  Time       Source      Destination    Protocol  Info
10196 106.666508 200.0.0.1   200.0.137.138  IP        IPv6 hop-by-hop option
10197 106.666512 200.0.0.1   200.0.137.138  IP        IPv6 hop-by-hop option
10198 106.666515 200.0.0.1   200.0.137.138  IP        IPv6 hop-by-hop option
10199 106.666519 200.0.0.1   200.0.137.138  IP        IPv6 hop-by-hop option
10200 106.666523 200.0.0.1   200.0.137.138  IP        IPv6 hop-by-hop option
10201 106.666526 200.0.0.1   200.0.137.138  IP        IPv6 hop-by-hop option

⊞ Frame 10201 (100 bytes on wire, 100 bytes captured)
⊟ Ethernet II, Src: 00:00:00:00:00:00, Dst: 00:07:e9:17:83:00
    Destination: 00:07:e9:17:83:00 (00:07:e9:17:83:00)
    Source: 00:00:00:00:00:00 (00:00:00:00:00:00)
    Type: IP (0x0800)
⊟ Internet Protocol, Src Addr: 200.0.0.1 (200.0.0.1), Dst Addr: 200.0.137.138 (200.0.137.138
    Version: 4
    Header length: 20 bytes
  ⊟ Differentiated Services Field: 0x00 (DSCP 0x00: Default; ECN: 0x00)
      0000 00.. = Differentiated Services Codepoint: Default (0x00)
      .... ..0. = ECN-Capable Transport (ECT): 0
      .... ...0 = ECN-CE: 0
    Total Length: 86
    Identification: 0x0000 (0)
  ⊟ Flags: 0x04
      .1.. = Don't fragment: Set
      ..0. = More fragments: Not set
    Fragment offset: 0
    Time to live: 0
    Protocol: IPv6 hop-by-hop option (0x00)
    Header checksum: 0x611c (correct)
    Source: 200.0.0.1 (200.0.0.1)
    Destination: 200.0.137.138 (200.0.137.138)
  Data (66 bytes)

0000  00 07 e9 17 83 00 00 00  00 00 00 00 08 00 45 00   ........ ......E.
0010  00 56 00 00 40 00 00 00  61 1c c8 00 00 01 c8 00   .V..@... a.......
0020  89 8a 00 00 00 00 00 00  00 00 00 01 00 00 00 02   ........ ........
0030  00 00 00 03 00 00 00 04  00 00 00 05 00 00 00 06   ........ ........
0040  00 00 00 07 00 00 00 08  00 00 00 09 00 00 00 0a   ........ ........
0050  00 00 00 0b 00 00 00 0c  00 00 00 0d 00 00 00 0e   ........ ........
0060  00 00 00 0f                                        ....
```

Figure 11: Network Analyzer output of first generated IP packet ( contained data is a 32 bit up-counter )

## V. Summary

The reference design of a two channel, LHCb-standard GBE mezzanine has been successfully completed. Further GBE mezzanines with up to 4 GBE channels can be designed, building up upon the technology and test environment of the current reference card. All such GBE mezzanines use the same LHCb connector carrying the SPI-3 industry bus and a 16 bit MAC configuration bus. A test system, based on the hardware-programmable Aroc PCI card was used to initialise the MAC registers. A first version of an Aroc-based FPGA driver for generating IPv4 packets with the GBE mezzanine was successful. The FPGA driver for IP packets will be further developed into the full LHCb data stream driver which will transport a balance of HLT and L1T streams as IPv4-encoded MEPs between the Tell-1 buffers and the readout network of LHCb.

## VI. Acknowledgements

We wish to thank BeiBei Shao of Tsinghua University Beijing for the fast and high quality production of the printed circuits.

## VII. References

[1] *Tell1: Common L1 read out board for LHCb* (this conference ) G.Haefeli (University of Lausanne)

[2] *A common implementation of the Level 1 trigger and HLT Data Acquisition*, A.Barczyk, J.P.Dufey, B.Jost, N.Neufeld, LHCb note LHCb 2003-079, DAQ

[3] *Raw-data transport format*, B. Jost, N. Neufeld, LHCb note 2003-063, DAQ

[4] *SPI-3 specification* Optical Interface Forum http://www.oiforum.com/public/documents/OIF-SPI3-01.0.pdf

[5] *Gigabit Ethernet mezzanines for DAQ and Trigger links of LHCb* LHCb note 2003-021 DAQ H.Muller, F. Bal, A. Guirao

[6] *Implementing the L1 trigger path* R.Jacobsson, LHCb note 2003-080, DAQ

[7 ] *TTC: trigger and Timing Control systems for LHC* http://ttc.web.cern.ch/TTC/intro.html

[8] *PM3386 Dual Gigabit ethernet Controller*, Data Sheet, PMC Sierra, July 2001, http://www.pmc-sierra.com/index.html

[9] *Marvell Alaska 88E1020 dual port Transceiver* http://www.marvell.com/products/transceivers/dualport/88E1020.js

[10] Marvell White paper "*Gigabit Ethernet over Copper Performance*" http://www.marvell.com/products/transceivers/singleport/Gigabit_Performance_White_Paper_final.pdf

[11] *Advanced Readout Controller for PCI-based test systems of LHCb* (this conference) Angel Guirao, Ken Wyllie, Francois Bal, Hans Muller CERN

[12] *Développement de logiciel de configuration d'un chip MAC via logique FPGA sur port PCI*, Antonis Bonos, Stagiaire Français CERN April -Sept 03 (french only) Source :

http://cern.ch/ep-div-ed/Documents/Mini-these-Antonis.pdf

[13]*Ethereal*, freely available Network Analyzer software http://www.ethereal.com