Published by IOP Publishing for Sissa Medialab

Received: *November 1, 2024*
Accepted: *January 13, 2025*
Published: *February 13, 2025*

Topical Workshop on Electronics for Particle Physics
University of Glasgow, Scotland, U.K.
30 September–4 October 2024

First results of an evaluation of 100Gb/s Ethernet as a future HEP readout link

Valentin Stümpert,[a,b,*] **Adam Klekotko**,[a,c] **Alberto Perro**,[a,d] **Christophe Gabriel Sigaud**,[a] **Daniel Hernandez Montesinos**,[a] **Francesco Martina**,[a] **Paschalis Vichoudis**,[a] **Sophie Baron**[a] **and Stefan Biereigel**[a]

[a] *CERN,*
Esplanade des Particules 1, 1211 Geneva 23, Switzerland
[b] *Institute for Data Processing and Electronics, Karlsruhe Institute of Technology,*
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
[c] *KU Leuven,*
Oude Markt 13, 3000 Leuven, Belgium
[d] *Universite d'Aix-Marseille III,*
58, bd Charles Livon, 13284 Marseille Cedex 07, France

E-mail: valentin.stumpert@cern.ch

Abstract: This work presents first results from a feasibility study to qualify commercial 100 Gb/s Ethernet as a readout link for future high-energy physics detectors. The motivation for this work is to examine the possibility of reading out future detector front-ends directly via commercial, high bandwidth, cost-optimized data centers, without the need of the intermediate custom-link/hardware layer used for this purpose thus far. In an attempt to determine how well 100 Gb/s Ethernet is adapted to the readout of such systems, this paper examines the protocol on aspects such as unidirectional operation and radiation hardness. Initial work towards a complete demonstrator system is also presented.

Keywords: Front-end electronics for detector readout; Optical detector readout concepts; Radiation-hard electronics; Data Processing

[*] Corresponding author.

https://doi.org/10.1088/1748-0221/20/02/C02024

Contents

1 Introduction

Current detector data-acquisition (DAQ) chains are based on custom radiation-hard datalinks. These systems rely on the use of additional custom hardware in the back-end (BE) to receive and decode the custom-protocol datalinks, aggregate and often pre-process the data and subsequently forward the resulting information to data centers via commercial datalinks. For future systems we envisage to eliminate the need for such intermediate custom BE layers from the DAQ chain. This can be achieved using a high-capacity, commercial datalink such as 100 Gb/s Ethernet (100GbE), allowing the front-end (FE) to communicate directly with the data center. Processing could then be centralized, profiting from flexible, cost-optimized and quickly-evolving data center technologies.

This work shows the first results of a feasibility study, to determine how well commercial 100GbE is adapted to the challenges inherent to high-energy physics (HEP) readout. This complements efforts currently underway at CERN, developing radiation-hard transmitters for future FEs, based on four wavelength-multiplexed 25 Gb/s lanes over a single fibre [1, 2].

It is important to note, that this work currently focuses solely on the flow of physics data from the detector, often referred to as “uplink”. As such, considerations regarding the low-bandwidth control and timing distribution in the opposite direction towards the detector — the “downlink” — are outside of the scope of this paper. The operation of 100GbE as a unidirectional, uplink-only connection is discussed in section 2. The radiation-hardness of 100GbE is assessed in section 3. Here, the results of a statistical analysis of radiation test data are presented, giving a realistic estimate of error resilience under radiation. Section 4 discusses the ongoing development of a full 100GbE demonstrator system.

Complementing this development is an intermediate effort to allow existing custom datalinks to feed realistic detector data directly into Ethernet infrastructure. To do so, we are working on the development of standard-format pluggable transceiver modules that can perform protocol conversion from the current FE custom links to Ethernet in a fully transparent manner.

2 Unidirectionality

Upstream links are mostly dedicated to high-volume detector data extraction. The command and control information transmitted on the downlink requires only minor bandwidth (varying with application), leading to highly asymmetric link usage. As an example, in front-end link implementations of HL-LHC experiments based on lpGBT ASICs [3], the bandwidth is asymmetric by design: the uplink features a bandwidth of up to 10.24 Gb/s, whereas the downlink bandwidth is reduced to 2.56 Gb/s. While the uplink bandwidth requirement is bound to increase due to higher interaction rates in the HL-LHC and other factors, such as future detectors potentially leveraging triggerless readout, the downlink usage is not projected to follow this trend, furthering link asymmetry. Ethernet does provide standards for asymmetric connections, but not at 100 Gb/s speeds. The complexity of a four-lane 100GbE receiver implementation in the FE is prohibitive for such a low-usage downlink.

The 100GbE standard [4, section 81.1.7.4] only defines full-duplex operation. Regardless, unidirectional links can work by sending specific link control messages on the transmit-only side. These message pretend to receive signals from the RX-only side, therefore appearing as a bidirectional link. This fact was verified in testing, using an FPGA (Xilinx VMK180) as a traffic generator and commercial-off-the-shelf (COTS) transceivers of various standards. The target was a commodity network interface card (FS E810-CAM2) in a PC. Both point-to-point and multi-hop network configurations have been tested. The multi-hop intermediary was a standard 100GbE switch (Juniper QFX5200-32C-SAFO). In all tests, the link was correctly identified as active and full link saturation was achieved, with UDP packets being routed to their intended destinations. There have been no observations of adverse behavior of the connection, e.g. data corruption, packet loss or connection drops.

3 Radiation hardness

lpGBT and other custom links currently used in detector FEs are specifically adapted to protect the transmitted data against radiation-induced errors. In particular, data is organized into small frames. This reduces the size of buffers and thus the number and physical footprint of vulnerable memory cells on the ASIC. In contrast, Ethernet is not radiation-hardened by design and an Ethernet frame is orders of magnitude longer. It is therefore paramount to assess the compatibility of the Ethernet standard with the radiation environment of high-energy physics detectors, especially since FEs are generally not equipped with capabilities for data re-transmission in case of packet corruption.

Both lpGBT and 100GbE feature Forward Error Correction (FEC) in the form of Reed-Solomon (RS) codes. These codes cluster bits into symbols, which allows multiple bit errors to be corrected with only one operation. This increases the robustness against burst errors typically induced by single event effects due to radiation, as multiple corrupted bits in the same symbol do not consume additional resources for correction. The key properties of an RS code are the number of symbols that can be corrected per frame and the size of the symbol. These properties are summarized in table 1 for both links. The lpGBT FEC features smaller symbols with fewer corrections per frame, but its small frames enable frequent corrections to take place. It is thus particularly well adjusted to protect against short but frequent burst errors. The most common Ethernet FEC — so-called KR4 — has a low ratio of correctable symbols to frame size. Frequent errors quickly saturate the correction capability. Advantages of KR4 are the lower overhead and large symbols, allowing longer burst errors to be corrected. KR4 is suited to protect against long but infrequent burst errors.

Table 1. Summary of key properties of lpGBT and Ethernet KR4 FEC schemes. KR4 is more efficient, but cannot tolerate as high a bit error rate (BER). The larger symbols of KR4 are more robust against longer burst errors.

FEC scheme	Frame Size	Symbol Size	corr. Sym./Frame	Overhead	max. corr. BER
lpGBT	240b	4b	6	20%	10%
Ethernet KR4	5280b	10b	7	3%	1.3%

Heavy-ion test data taken in December of 2023 by the developers of the DART28 ASIC allows a real-world analysis of Ethernet FEC capabilities. Recorded error patterns were validated against the Ethernet KR4 correction capabilities. Data was recorded with the triple modular redundancy (TMR) of the ASIC both enabled and disabled. Figure 1 shows histogram data for recorded corrupted symbol counts. While enabling TMR does not prevent all errors, the data shows no errors that cannot be corrected by the KR4 FEC. Data with TMR disabled shows only a single non-correctable error.

Figure 2 shows the derived heavy-ion radiation cross section with the TMR disabled, dependent on the linear energy transfer (LET) of the incoming ion. Data is shown for both a single 25 Gb/s link and a configuration mirroring a 100GbE link with four parallel lanes. If no errors above the correction limit are observed for a given LET, one error is nevertheless assumed in calculation to obtain a safe estimate [5]. When integrated over an LHC-like LET distribution [6], for example for the CMS inner tracker close to the beam pipe, a total cross section of 2.4×10^{-13} cm^2 is derived. This represents a worst-case value and is assumed for all environments for simplicity.

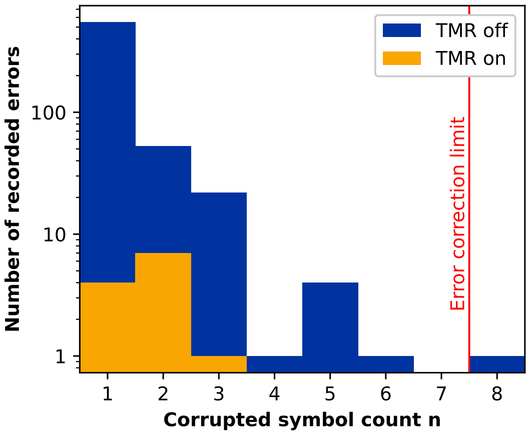

Figure 1. Logarithmic histogram of error lengths. Most errors are 1 to 3 symbols long. Only one recorded error cannot be corrected. TMR reduces the error count and length.

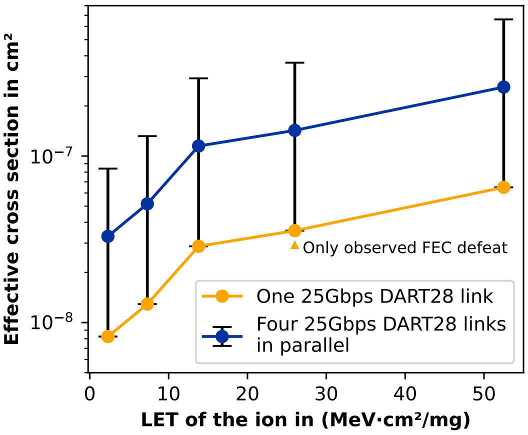

Figure 2. Calculated heavy-ion cross section of a 100GbE KR4 link. Data given for a single 25 Gb/s link and four parallel lanes in 100GbE configuration.

Figures for the mean-time between non-correctable errors (MTBE) and bit error rate (BER) are provided in table 2 for a wider range of environments. The BER is calculated under the worst-case assumption of 8 error symbols with the FEC decoder amplifying the error with a 7 symbol maximum-length mis-correction, leading to 15 total erroneous symbols per non-correctable frame.

Table 2. MTBE and BER for simulated 100GbE deployment (four parallel 25 Gb/s lanes) in different HL-LHC environments. Flux data is taken from public reports.

Simulated area	Assumed Flux	MTBE	BER
CMS Inner Tracker	1.5 GHz/cm^2 [7]	46.30 min	5.24×10^{-13}
CMS Muon Gas Electron Multiplier ME0	3.28 MHz/cm^2 [8]	352.87 h	1.15×10^{-15}
CMS Muon Resistive Plate Chamber RE4/1 iRPC	10 kHz/cm^2 [9]	13.21 a	3.49×10^{-18}

4 Protocol translators

To be able to test an Ethernet-based DAQ system with real-world traffic, protocol translators between 10.24 Gb/s lpGBT and 10GbE are being developed. These translators also serve as a way to connect lpGBT nodes into an Ethernet network. FPGAs are used to ensure flexibility and adaptability for arbitrary use cases.

As the translators are to be tightly integrated with Ethernet infrastructure, they are designed as pluggable modules. Two relevant standards exist for these modules: the single-lane small form-factor pluggable (SFP) and four-lane Quad-SFP (QSFP). These modules plug directly into the transceiver ports of commodity networking hardware such as switches and network cards, simplifying deployment.

The module interfaces are fixed by the standard, limiting power draw, number of transceiver lanes and speed. For both module standards lane speeds can be either 10 Gb/s or 25 Gb/s. No dedicated clock input is possible, preventing the synchronization of the lpGBT links to a central master clock.

The firmware implemented in the modules is based on firmware developed for the NetGBT project [10], developing custom hardware for a rack-mountable lpGBT-to-Ethernet converter.

4.1 SFPs

The requirements for the SFP modules are summarized in table 3. Commercial solutions satisfying these requirements were identified and a product based on a Microchip application note [11] was selected. As the design is based on a Microchip Polarfire FPGA, the NetGBT firmware was adapted and ported to this platform. The functionality was verified using an evaluation kit (Microchip MPF300-EVAL-KIT). The FPGA board was connected to a VLDB+ lpGBT test board [12] and an Ethernet switch (Juniper QFX5200-32C-SAFO). Full throughput of lpGBT frames into UDP traffic was verified at a PC connected to the same switch, using a COTS network interface (FS E810-CAM2). Manufacturing of the modules is currently ongoing.

Table 3. Summary of requirements for the SFP protocol translator modules.

Requirement	Value
Cost	< 1000 CHF per module
Throughput	10 Gb/s
Maximum power draw	2 W via SFP Edge card
Clocks	Internal generation. No external clocks
Transceiver lanes	2 in 2 different frequency domains
SFP standard compatibility	Must comply mechanically and electrically.

4.2 QSFPs

The requirements for the QSFP module are summarized in table 4. No suitable commercial solution exists.

Table 4. Summary of requirements for the QSFP protocol translator modules.

Requirement	Value
Throughput	$\geq$ 4x 10 Gb/s
Maximum power draw	10 W via QSFP Edge card
Clocks	Internal generation. No external clocks
Physical FPGA size	<18 mm wide
Transceiver lanes	8 in 2 different frequency domains
QSFP standard compatibility	Must comply mechanically and electrically.
Module management interface	Active management interface <1.5 W (start-up requirement)
Program memory	Flash capacity > 2 images
Configuration memory	Persistent memory for addresses and configuration
Device configuration	Possibility of reconfiguration without module disassembly

The Artix Ultrascale+ (US+) AU10P and AU15P FPGAs were identified as the only options with 8 transceiver lanes within a suitably small package. No 25 Gb/s transceivers exist at this form factor, limiting the design to 4x 10 Gb/s throughput. As the lpGBT is a bidirectional protocol that can require a master clock to be provided to the front-end via the downlink, a configurable clock source is included in the module.

To comply with the low-power startup requirement, an STM L071 microcontroller is used to communicate with the host device. The FPGA power supply is initially off, and only enabled once the host transmits high-power clearance. A TPS6521905 PMIC with efficiencies above 80 % is used to provide the FPGA supply in a minimal footprint. Power- and space-optimized parts are used to remain within constraints given by the QSFP standard. Final power draw is estimated at 8.07 W.

The microcontroller features built-in EEPROM. 32 kb of external EEPROM is provided for the FPGA. Up to 4 FPGA images can be stored in 256 Mb of flash memory. The microcontroller can be programmed via the I2C interface of the QSFP host-side connector. A dedicated ribbon-cable connector exposes the JTAG interface of the FPGA for initial programming. It is planned to provide a golden image to configure the FPGA remotely via the Ethernet interface after initial programming, enabling in-situ over-the-network reconfiguration once deployed.

The selected FPGAs are already compatible with the NetGBT firmware. The firmware was modified for operation with four independent links and verified using an evaluation kit (OpalKelly XEM8320) with the previously described network setup.

5 Conclusion and outlook

Commercial-standard 100GbE has been validated against HEP uplink requirements. Results show feasibility for its use as a front-end data readout link. Radiation hardness was shown to be sufficient with a simulated BER of 5.24×10^{-13} in HL-LHC conditions, close to the interaction point, when leveraging silicon-proven, radiation-tolerant serializer designs — even without additional hardening through TMR. Unidirectional operation of 100GbE has been demonstrated using commodity hardware. The high bandwidth per fibre would enable the use of triggerless, streaming readout systems. Suitability for the use as a timing and control link require further study.

Work towards a full demonstrator system including data center processing is ongoing. In parallel, integration of custom protocols into commodity Ethernet infrastructure using protocol translators has been successfully tested. Scale deployment will be tested with pluggable modules, the specification and design considerations of which have been presented.

Acknowledgments

We thank the CERN EP-LBC team for their insights and cooperation.

This work has been sponsored by the Wolfgang Gentner Programme of the German Federal Ministry of Education and Research (grant no. 13E18CHA).

2025 JINST 20 C02024

References

[1] T. Prousalidi et al., *System Development of Radiation-Tolerant Silicon Photonics Transceivers for High Energy Physics Applications*, *IEEE Trans. Nucl. Sci.* **70** (2023) 2373.

[2] M. Baszczyk et al., *Dual use driver for high speed links transmitters in the future high energy physics experiments*, 2024 *JINST* **19** C03013.

[3] The lpGBT Team, *The Low Power GigaBit Transceiver ASIC (lpGBT)*, manual https://lpgbt.web.cern.ch/lpgbt/v1/.

[4] *IEEE Standard for Ethernet*, IEEE Std 802.3-2022 (Revision of IEEE Std 802.3-2018) (2022) 1 [DOI:10.1109/IEEESTD.2022.9844436].

[5] J.R. Schwank, M.R. Shaneyfelt and P.E. Dodd, *Radiation Hardness Assurance Testing of Microelectronic Devices and Integrated Circuits: Test Guideline for Proton and Heavy Ion Single-Event Effects*, *IEEE Trans. Nucl. Sci.* **60** (2013) 2101.

[6] M. Huhtinen and F. Faccio, *Computational method to estimate single event upset rates in an accelerator environment*, *Nucl. Instrum. Meth. A* **450** (2000) 155.

[7] CMS collaboration, *The Phase-2 Upgrade of the CMS Tracker*, CERN-LHCC-2017-009, CERN, Geneva (2017) [DOI:10.17181/CERN.QZ28.FLHW].

[8] CMS collaboration, *The Phase-2 Upgrade of the CMS Muon Detectors*, CERN-LHCC-2017-012, CERN, Geneva (2017).

[9] CMS Muon collaboration, *CMS iRPC FEB development and validation*, *Nucl. Instrum. Meth. A* **1064** (2024) 169400.

[10] A. Perro, M. Vodnik and P. Durante, *A Low-Cost, Low-Power Media Converter Solution for Next-Generation Detector Readout Systems*, in the proceedings of the *Topical Workshop on Electronics for Particle Physics*, University of Glasgow, Scotland, U.K., 30 September–4 October 2024 [arXiv:2410.23173].

[11] M. Technologies, *AN4364: PolarFire FPGA SFP+ Module*, Application Note Microchip Technologies (2021).

[12] D.H. Montesinos, S. Baron, N. Guettouche and J. Mendez, *The Versatile Link+ Demo Board (VLDB+)*, 2022 *JINST* **17** C03032.