Published by IOP Publishing for Sissa Medialab

Received: *October 24, 2021*
Accepted: *February 25, 2022*
Published: *June 14, 2022*

Topical Workshop on Electronics for Particle Physics 2021
20–24 September, 2021
Online

2022 JINST 17 C06002

Hermes — a robust, low latency, optical link protocol for synchronous data transfer at commercial asynchronous line rates

K. Adamidis,[a,*] I. Bestintzanos,[a] C. Foudas,[a] A. Howard,[b] G.M. Iles,[b] S. Mallios[c] and T. Williams[d]

[a] *University of Ioannina, Ioannina, Greece*
[b] *Blackett Laboratory, Imperial College, Prince Consort Road, London, SW7 2BW, U.K.*
[c] *European Organization for Nuclear Research (CERN), Geneva, Switzerland*
[d] *STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot, OX11 0QX, U.K.*

E-mail: kosmas.adamidis@cern.ch

Abstract: The CMS Level-1 Trigger at HL-LHC and associated upstream systems employ more than 10000 25 Gb/s optical links, transferring almost a Pb/s of data synchronously between the different back-end processing nodes. Stable operation of these links is essential to avoid the injection of erroneous signals into the trigger path, potentially leading to a flood of false triggers or data loss. The Hermes protocol, implemented on Xilinx UltraScale+ FPGAs, provides this stability while operating at asynchronous, industry standard line rates. The protocol design as well as results on the performance from extensive testing are presented here.

Keywords: Digital electronic circuits; VLSI circuits

*Corresponding author.

© 2022 The Author(s). Published by IOP Publishing Ltd on behalf of Sissa Medialab. Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

https://doi.org/10.1088/1748-0221/17/06/C06002

Contents

2022 JINST 17 C06002

1 Introduction

The CMS Level-1 Trigger at HL-LHC [1] and associated upstream systems must synchronously transfer data between many different processing nodes with optical links running at 25 Gb/s. Four independent data processing paths deliver trigger objects to the Global Trigger where the Level-1 Accept decision is made. This complex system will comprise hundreds of ATCA boards interconnected with thousands of optical fibers. The pipeline design of the Level-1 Trigger system requires that the input data are aligned to the accelerator clock (LHC clock). This requirement, however, cannot be achieved when commercial transceivers are used since their line rate differs from the algorithm data production rate, which is synchronous to the LHC clock. Furthermore, the Forward Error Correction (FEC) methods used in industry (e.g. RS-FEC(528, 514) for 25G Ethernet [2]) to protect against bit errors are not suitable here due to their large latency. The above are two of the main reasons that led to the development of a custom protocol, called Hermes.

Hermes is an asynchronous optical link protocol, proposed to be used at CMS for the communication between back-end trigger processor boards during Phase-2. Its key objectives are to keep the total latency as short as possible while maintaining the utilization footprint considerably low. It is built around the GTH and GTY [3] Multi-Gigabit Transceiver (MGT) generations of Xilinx's Ultrascale and Ultrascale Plus FPGA families, reaching line rates up to 25.78125 Gb/s.

2 Protocol description

2.1 Encoding layer

The protocol is designed to handle two kinds of 64-bit words, the Data and Control words, similar to the Aurora 64b/66b protocol [4]. It utilizes the 64b/67b encoding scheme in order to transmit an additional 3-bit Header attached on every 64-bit word, the usage of which is described in section 2.7. The overhead added by the chosen scheme is 4.68%.

To achieve DC balance over the optical medium and provide enough transitions for the recovery of the receiving clock, the scrambling method defined by IEEE 802.3 and introduced by 10 Gb/s Ethernet [2] is used for every 64-bit word. The 3-bit Header is balanced using a custom method, called Toggling Header. This method assumes two polarities for each Header transmitted. Polarity A transmits two 1s and one 0, while polarity B transmits two 0s and one 1. To achieve balancing, the Header polarity toggles with every word transmitted. When the tag of a word needs to change from Data to Control, and vice versa, instead of toggling the Header the same polarity is being transmitted. That way the receiver compares the previous value of the Header with the one received and is aware of the nature of the received word.

2.2 Asynchronous architecture

In this design the link and the algorithms are operating on different clock domains, commonly referred to as algorithm domain, synchronous with the LHC clock, and link domain. Hence, Hermes is designed to operate asynchronously with respect to the algorithm domain. These two domains are chosen to be independent so that the line rate and MGT's reference clock frequency can be chosen freely, following the standards set by industry. At the same time, synchronization of physics data with LHC is guaranteed as well.

The implementation of the asynchronous architecture, which was introduced in CMS already in 2008 and also used during Phase-1 [5], instantiates a FIFO and a dual port Block RAM (BRAM) in the transmitter and the receiver data paths respectively. As illustrated in figure 1, both of them are placed between the algorithm block and the transceiver. The architecture operates successfully only when the algorithm domain clock rate is lower than that of the link domain clock rate. Since the latter's rate is fixed by the link line rate, the algorithm can run with maximum frequency which is multiple of the 40 MHz LHC clock and yet remains lower than the link clock. This technique guarantees that no data loss would ever occur due to FIFO overflow.

The bandwidth difference occurring in this case, called filler bandwidth, is compensated by transmitting padding words, generated when the Tx FIFO is empty. Its size is the subtraction of the algorithm bandwidth from the total link bandwidth. For example, in 25.78125 Gb/s links the maximum payload clock can reach 360 MHz resulting in 23.04 Gb/s of algorithm bandwidth. The remaining 2.74125 Gb/s is the filler bandwidth. Since the latter is generated in the link domain of the transmitter, it is imperative that it will be detected and removed in the link domain of the receiver. Otherwise, data coming out of the Rx BRAM will go out of sync. To avoid any such occurrences, the usage of two FEC schemes has been introduced and is described in section 2.7.

2022 JINST 17 C06002

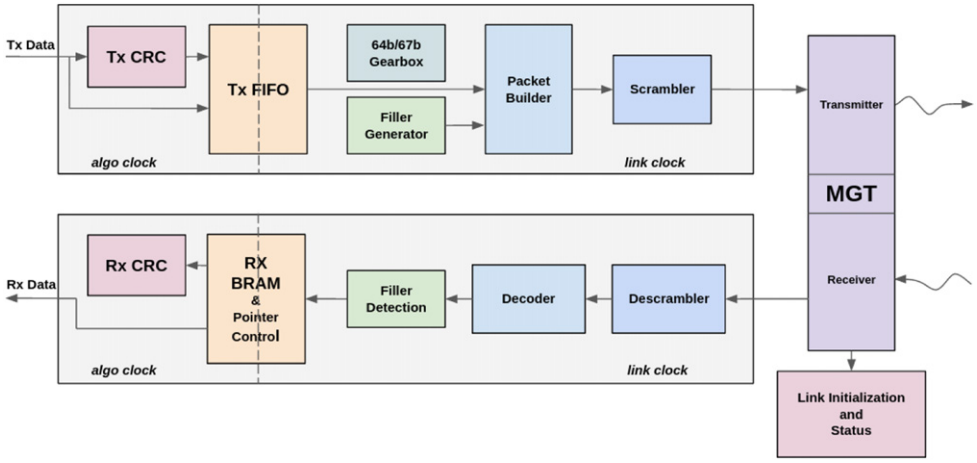

Figure 1. Block diagram of Hermes protocol. Arrows illustrate data propagation through the transmitter and receiver data paths, flowing to and from the MGT block.

2.3 Framing layer

Hermes protocol specifies two kinds of words, Data and Control words, with the latter consisting of two categories: Idle and Filler Control words, as shown in table 1. The difference between Idle and Filler words is that Idles belong to the algorithm bandwidth, same way as Data do. Idle words are characterized as such by the user and by controlling the so called "Valid Bit". When no valid data exist to be transmitted by the algorithm block, the value of this bit can be set to low, forcing the transmission of Idles. Hence, they are written to the Rx BRAM and act as a method to separate data packets. A Control Word Type (CWT) defines the identity of control words.

Table 1. Hermes frames specification. The Header specifies the transmission of either Data or Control words, the type of which is defined in the Control Word Type field.

Header	Byte 7	Byte 6 <> Byte 0
Data	Data	
Control	CWT	Idle Payload
Control	CWT	Filler Payload

In order to devote the whole algorithm bandwidth solely to physics data, the filler bandwidth is utilized to transmit information such as link metadata, Cyclic Redundancy Check (CRC) checksums and alignment markers. This is achieved by artificially generating two kinds of filler words that propagate the above information, the CRC and the Align Marker Filler words.

The format of Hermes's filler words can be seen in table 2. The least significant byte is used to carry link id information, while the next 3 bytes are user defined and could contain such information as board id, crate id, subsystem kind, etc.

2022 JINST 17 C06002

Table 2. Format of the Padding, CRC and Align Marker Filler words.

Byte 7	Byte 6	Byte 5	Byte 4	Byte 3	Byte 2	Byte 1	Byte 0
CWT	Reserved	CRC	CRC	User Info	User Info	User Info	Link Id

2.4 Cyclic redundancy check

CRC codes 16-bits long are used by Hermes to determine the integrity of data transmitted. Two checksum codes are calculated both at the transmitter and receiver sides independently, enabled by the Valid Bit. Tx and Rx CRC blocks perform the calculation and check in the algorithm clock domain. This way domain crossing of data is included in the checksum propagation and potential flaws in their operation can be detected. For this reason, the CRC value must cross domains and does so using four out of the seventy two bits of FIFO's width. The checksum code breaks into four 4-bit chunks, each of which crosses separately.

2.5 Data transmission modes

The main objective of Hermes protocol is to support all physics data packet types that meet the requirements of CMS Level-1 Trigger, while at the same time maintaining ease of use. These requirements can be addressed by implementing two transmission modes, called Packet and Streaming mode.

Packet mode is used when data transmission by the algorithm block includes Data and Idles. The packets can be of any length and have at least one Idle word between them. The user defines a packet by asserting the Valid Bit to declare the start and deasserting it at the end. A Data Start marker is generated on the receiver side at the rising edge and is used as the alignment marker. The falling edge triggers the transmission of the CRC checksum.

Streaming mode is introduced to facilitate the transmission of back to back packets, i.e. packets that are not separated by Idle words. In this scenario, the algorithm is constantly streaming valid data frames through the link, hence Valid Bit is always asserted. The boundaries of a packet are defined by the "End of Packet" bit. When asserted it declares the last word to be used in the CRC calculation and also triggers its transmission. An alignment marker can be transmitted by asserting the "Align Marker" bit.

2.6 Link alignment

Link alignment, or channel bonding, in Hermes is implemented by using the alignment marker bits of all receiving channels and by controlling the read pointer of each Rx BRAM separately. Both Packet and Streaming modes transmit this marker through all links on the same clock cycle. Due to differences in propagation delay, these markers may not arrive simultaneously but in different clock cycles, placed between the first and the last received markers. The method of aligning, or syncing, all channels together counts for every channel the number of clock cycles required after the reception of its own marker and until the reception of the last. It then subtracts this number from the read pointer of the BRAM of the individual channel, so that all channels are aligned to the link with the largest latency.

2022 JINST 17 C06002

2.7 Protection mechanism

Assuming a channel is aligned to a specific bunch crossing and frame, the receiving algorithm expects a strict sequence of incoming physics data. If the alignment of an individual link is lost, it results in misinterpretation of all data frames until the link is re-aligned. The above scenario takes place in cases where fillers are not properly recognized and get mixed with words of the algorithm bandwidth, or more specifically when either the Header or the Control Word Type is received incorrectly. In order to protect the link from such occurrences, Hermes encodes both of them into Forward Error Correction codes, each of which is capable of correcting up to one bit flip of the original code. The 3rd Header bit is used as a secondary check to determine the original value and the 4-bit CWT is encoded in Hamming (7, 4) codes.

3 Performance

The firmware of Hermes protocol has undergone detailed testing by transmitting data between several ATCA processors. All 192 links used had been running error free while transferring more than 675 Petabytes at 25 Gb/s. Link alignment loss has been tested with forced errors using attenuators to introduce errors over a link. With the implementation of the two FEC mechanisms, Hermes link alignment has shown to be immune to single bit flips to error ratio up to 10^{-9}. Furthermore, link robustness has been tested by repeatedly forcing reset and re-configuration of links in different packet mode scenarios.

4 Conclusion

A novel optical link protocol has been developed, which fully satisfies the requirements of the CMS back-end electronics. This protocol is exceptionally ruggedized against erroneous data transmission, maintaining the ability to operate at commercial asynchronous line rates. While the performance has been validated, further stress testing on hardware is ongoing to detect and improve potential operational flaws.

Acknowledgments

The work of this publication is financed from the project DeTAnet — MIS 5029538 — Ministry of development & Investments.

References

[1] CMS collaboration, The Phase-2 Upgrade of the CMS Level-1 Trigger, CERN-LHCC-2020-004, CMS-TDR-021 (2020), https://cds.cern.ch/record/2714892?ln=en.

[2] Ethernet Working Group, *IEEE 802.3-2018 — IEEE Standard for Ethernet* (2018), https://standards.ieee.org/ieee/802.3/7071/.

[3] Xilinx Incorporated, *UltraScale Architecture GTY Transceivers UG578 Rev v1.3.1* (2021), https://docs.xilinx.com/v/u/en-US/ug578-ultrascale-gty-transceivers.

2022 JINST 17 C06002

[4] Xilinx Incorporated, *Aurora 64B/66B Protocol Specification SP011 (v1.3)* (2014), https://docs.xilinx.com/v/u/en-US/aurora_64b66b_protocol_spec_sp011.

[5] CMS collaboration, CMS Technical Design Report for the Level-1 Trigger Upgrade, CERN-LHCC-2013-011, CMS-TDR-012 (2013), https://cds.cern.ch/record/1556311?ln=en.

2022 JINST 17 C06002