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

Kosmas Adamidis[a]†, Ioannis Bestintzanos[a], Costas Foudas[a], Alexander Howard[b], Gregory Michiel Iles[b], Stavros Mallios[c], Tom Williams[d]

a. University of Ioannina, b, Imperial College London, c. CERN, d. STFC Rutherford Appleton Laboratory

Phase-2 CMS Level-1 Trigger

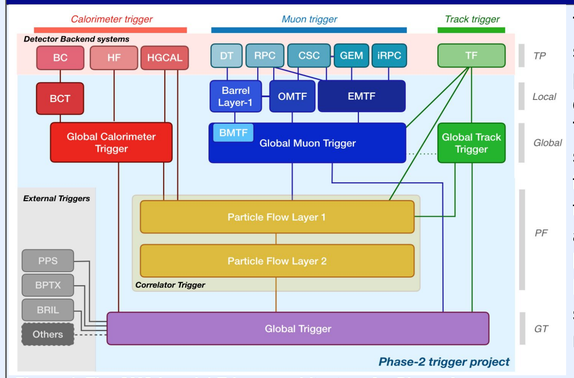

Figure 1. The CMS Level-1 Trigger architecture for Phase-2

The CMS Level-1 Trigger[1] for HL-LHC and associated upstream systems must synchronously transfer data between many different processing nodes with optical links running at 25Gb/s. Four independent data processing paths (*Figure 1*) deliver trigger objects to the Global Trigger where the Level-1 Accept decision is made. This complex system will comprise of hundreds of ATCA boards interconnected with thousands of optical fibers. The data transmitted must remain aligned to the LHC beam to avoid data loss and accidentally swamping CMS with a large number of false triggers due to misinterpretation of link data. The Forward Error Correction (FEC) methods used in industry (e.g. RS-FEC(528,514) for 25G Ethernet) to protect against bit errors are not suitable here due to their large latency. Instead, a new approach is needed, which has led to the development of the Hermes link protocol.

Hermes Protocol

- **Encoding Scheme:** Hermes utilizes the 64b/67b encoding to transmit 64-bit words over optical medium. An additional 3-bit Header is sent for every single word to identify it as either a **Data** or **Control Word** (*Table 1*).
- **Asynchronous Architecture:** Given that the line rate of high speed optical modules is determined by Industry and the desire not to drive the FPGA transceivers with the LHC clock, Hermes is based on the asynchronous architecture introduced at CMS during Phase-1. This implementation uses FIFO blocks in both the transmitter and the receiver, whilst the write and read sides, respectively, are driven by the LHC clock. The TX input of the link and the RX output of the link are are driven by the link clock (*Figure 2*).
- **Filler Bandwidth:** For the above architecture to operate, the link clock frequency must be higher than that of the LHC clock. This method leads to splitting the link bandwidth into a synchronous part with a fixed number of data words per LHC orbit and an asynchronous, called Filler Bandwidth. The Filler Bandwidth is mainly filled with padding control words.
- **Non-deterministic Filler words:** In order to dedicate the whole LHC bandwidth to physics data, non-deterministic words are artificially introduced, carrying information such as link metadata, CRC checksums and alignment markers, used to synchronize to the LHC orbit.
- **Protection Mechanism:** For the asynchronous model to operate successfully, it is imperative that Filler words are never written in the RX FIFO. To achieve this, two independent FEC techniques are used. One for the 3-bit Header and one for the Control Word Type to ensure that Fillers will always and reliably be identified in cases of single bit errors.

Header	Byte 7	Byte 6 <> Byte 0
Data	Data	
Control	CWT	Control Word Payload

Table 1. Hermes uses a 3-bit Header to separate Data from Control words. A Control Word Type (CWT) field is used to identify the different Control words the protocol transmits

- ***Control Words supported by Hermes***
  - **Idle Control Word:** Transmitted when the TX user has no valid data to send. It is visible to the RX user and used to define packet structures
  - **Padding Filler Word:** Pads the link with filler words when TX FIFO is empty.
  - **CRC Filler Word:** CRC checksums are calculated in the LHC clock domain, cross to/from the link domain inside the FIFOs and are transmitted as a non-deterministic filler word.
  - **Align Marker:** Non-deterministic filler word that is transmitted simultaneously in order to align receiver links on the same bunch crossing number.

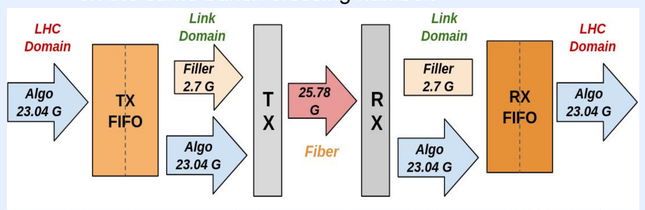

Figure 2. Asynchronous architecture block diagram. In the case of 25 G links, the actual line rate 25.78125 Gbps. The maximum Algorithm clock can run at 360 MHz, resulting in bandwidth of 23.04 Gbps. The remaining 2.74125 Gbps occupy the so called Filler Bandwidth which on the must be identified on the receiving end and not be written on the RX FIFO block.

Data Transmission Modes

Packet

Packet	Frame	VB	64 bit user words
N - 1	53	0	IDLE
N	0	1	DATA
	1	1	DATA
	2	1	DATA
	3	1	DATA
	4	1	DATA
	5	1	DATA
	....		.....
	50	1	DATA
	51	0	IDLE
	52	0	IDLE
	53	0	IDLE
N + 1	0	1	DATA
	1	1	DATA

Streaming

Packet	Frame	VB	64 bit user words
N - 1	0	1	DATA
	1	1	DATA
	2	1	DATA
	3	1	DATA
	4	1	DATA
	....		.....
	52	1	DATA
	53	1	DATA
N	0	1	DATA
	1	1	DATA
	2	1	DATA
	3	1	DATA
	4	1	DATA
	....		.....
	52	1	DATA
	53	1	DATA
N + 1	0	10	DATA
	1	1	DATA

Figure 3. Left: Example of a packet with 51 out of the 54 frames being valid data and the rest of them Idle words. Right: Only valid data words are constantly streamed through the link

Hermes Protocol is capable of supporting two data transmission modes, the Packet mode and Streaming mode (*Figure 3*), that are most commonly used in CMS Level-1 Trigger. In both of them, link metadata, CRC checksums and alignment markers are propagated through the Filler Bandwidth. The only words visible to the user are Data and Idle control words.

Packet Mode: In Packet mode the user transmits either Data or Idle words by controlling the so called Valid Bit. Packets can be defined with at least one Idle word in between them. Idle words are the only Control Words that are not Filler words, hence, they are written to the RX FIFO and read by the algorithm block.

Streaming Mode: Streaming Mode targets optimization to the bandwidth. In this mode physics data are constantly streamed through the link. Packet structures can be defined either by synchronous counters on both sides or by using the CRC checksum flag as an end of the packet indicator.

Performance - Results

So far, the Hermes Protocol Firmware have undergone detailed testing by transmitting data between several Phase-2 ATCA processors, all of which have been successful. No bit errors have even been observed during run time of more than 60 kilo link hours. Link alignment loss has been tested with forced errors using attenuators to introduce jitter over a link. With the implementation of the FEC mechanisms Hermes has shown to be immune to single bit flips at BERs exceeding 1 in 10^9. The adoption of the 64b67b encoding has raised concerns about the lack of DC balance, the effect of which in normal conditions is small (*Figure 4*). This imbalance is introduced by the 3-bit Header, since, apart from the scrambled 64-bit words, the Header will always have disparity of +1 or -1 (polarity A and polarity B). One proposal to counter this effect is to toggle the Header between two polarity modes.

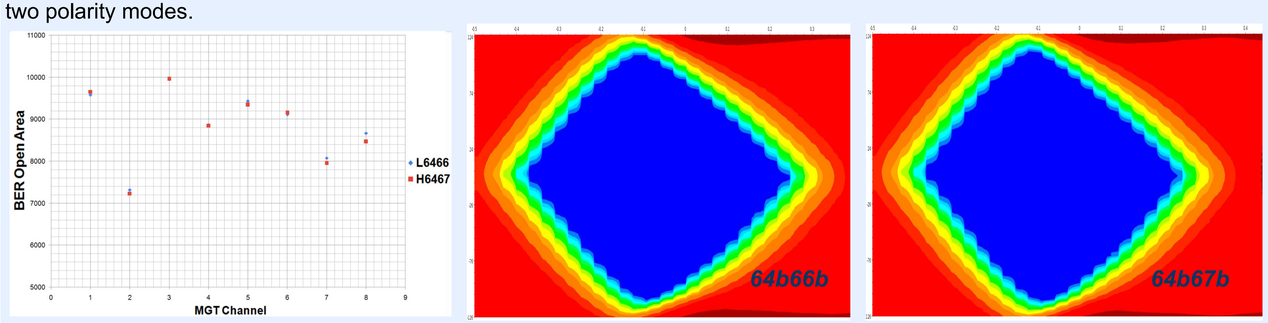

Figure 4. Left: BER Open Area comparison of 64b/66b (blue) and 64b/67b (red). Right:Eyescan of a link running the 64b66b version of the protocol (left) and one of the same link running the 64b67b version of the protocol (right).

References

1. CMS Collaboration The Phase-2 Upgrade of the CMS Level-1 Trigger, CERN-LHCC-2020-004 ; CMS-TDR-021

† contact : kosmas.adamidis@cern.ch