KM3NeT – INFRADEV – H2020 – 739560

Report on the improvement of the WR Switch with increased transmission rate aimed at 10 Gbps

KM3NeT – INFRADEV GA DELIVERABLE: D9.4

Document identifier:	**KM3NeT-INFRADEV-WP9-D9.4**
Date:	**13/01/2020**
Work package:	**WP9**
Lead partner:	**UVEG**
Document status:	**Final**
Dissemination level:	**Public**

Abstract

This document reports on the status and the activities connected to the study of the feasibility of a White Rabbit Switch working at 10 Gbps, which is interesting for KM3NeT as well as for most of the High Energy Physics experiments requiring high level of synchronization. In KM3NeT, White Rabbit is used for the distribution of time and frequency within a sub-nanosecond accuracy to KM3NeT nodes. KM3NeT has performed preliminary studies on the feasibility of a White Rabbit Switch with an increased data rate transmission. In particular, the FPGAs available in the market that could be used with this type of switch have been investigated, focusing in the stability of the Serializer- Deserializer latency, which determines the time synchronization that can be obtained. The oscillator system has also been studied, implementing, for evaluation, some of the possible solutions in one of the KM3NeT White Rabbit nodes.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/WP	Date
Author(s)	D. Real	UVEG / WP9	11/11/2019
Reviewed by	T. Chiarusi	INFN Bologna	12/12/2019
Reviewed by	A. D'Amico	Nikhef	15/12/2019
Reviewed by	E. Tzamariudaki	NCSR Demokritos	22/12/2019
Approved by	PMB		20/01/2020

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	11/11/2019	1st version circulated to reviewer	D. Real UVEG
2	20/12/2019	Version implementing reviewers' comments	D. Real UVEG

IV. APPLICATON AREA

This document is a formal deliverable for the GA of the project, applicable to all members of the KM3NeT INFRADEV project, beneficiaries and third parties, as well as its collaborating projects.

V. TERMINOLOGY

A complete project glossary is provided:

ARCA: Astroparticle Research with Cosmics in the Abyss
CDR: Clock Data Recovery
CLB: Central Logic Board
DDMTD: Digital Dual Mixer Time Domain

DOM:	Digital Optical Module
DU:	Detection Unit
FIFO:	First In First Out
FPGA:	Field Programmable Gate Array
GPS:	Global Positioning System
MGT:	MultiGigabit Transceivers
ORCA:	Oscillation Research with Cosmics in the Abyss
PCS:	Physical Coding Sublayer
PLL:	Phase-Locked Loop
PMA:	Physical Media Attachment
PMT:	PhotoMultiplier Tubes
PTP:	Precision Time Protocol
SC:	Slow Control
Sync-E:	Synchronous Ethernet
WR:	White Rabbit
WRS:	White Rabbit Switch

VI. LIST OF FIGURES

VII. PROJECT SUMMARY

KM3NeT is a large Research Infrastructure that will consist of a network of deep-sea neutrino telescopes in the Mediterranean Sea with user ports for Earth and Sea sciences. Following the appearance of KM3NeT 2.0 on the ESFRI roadmap 2016 and in line with the recommendations of the Assessment Expert Group in 2013, the KM3NeT-INFRADEV project addresses the Coordination and Support Actions (CSA) to prepare a legal entity and appropriate services for KM3NeT, thereby providing a sustainable solution for the operation of the research infrastructure during ten (or more) years. The KM3NeT-INFRADEV is funded by the European Commission's Horizon 2020 framework, and its objectives comprise, amongst others, activities on technology transfer and innovation in the KM3NeT Collaboration (work package 9).

VIII. EXECUTIVE SUMMARY

The main goal of WP9 is to establish methodologies both for exposing to interested parties in the industrial sector technological choices and innovative solutions that have been developed or adapted by KM3NeT and for following the technological advances in key areas of interest to KM3NeT. In addition, technology transfer in the form of services can be provided from KM3NeT to the industry or institutions with potential interest.

The main goal of this task is to study the feasibility of an improved version of the White Rabbit Switch working at 10 Gbps, needed due to the constant increase in the data produced by the particle physics experiments. This kind of device will simplify the architecture that is being prepared for next phases of KM3NeT based on White Rabbit standard. Using a 10 Gbps White Rabbit Switch any data transfer bottleneck at the base of the Detection Units will be avoided, thereby enabling the possibility to use one single White Rabbit Switch instead of the two needed at present.

KM3NeT has performed preliminary studies on the feasibility of a White Rabbit Switch with an increased data rate transmission. In particular, FPGAs available in the market that could be used with this type of switch have been investigated, focusing on the stability of the Serializer-Deserialiser SERDES latency, which determines the time synchronization that can be obtained. An evaluation has been carried out both for a long operation term and after a reset or power cycle. The oscillator system of White Rabbit has also been studied, implementing some of the possible solutions in one of the KM3NeT White Rabbit nodes for evaluation.

Table of Contents

1. Introduction

The KM3NeT Collaboration is currently installing two neutrino telescopes at the bottom of the Mediterranean Sea, one, called ORCA (Oscillation Research with Cosmics in the Abyss), 40 km away from Toulon and the second one, called ARCA (Astroparticle Research with Cosmics in the Abyss), 100 km away from the southern tip of Sicily. The ARCA telescope is designed to detect high energy neutrinos from distant astrophysical sources, while ORCA is dedicated to the study of the oscillations of atmospheric neutrinos with energy of a few GeV in order to measure the neutrino mass ordering. The neutrino detectors comprise a three-dimensional grid of photomultiplier tubes (PMTs) for measuring the Cherenkov photons induced. The ARCA and ORCA detectors measure the Cherenkov photons induced by the relativistic charged particles produced in the neutrino interactions with the matter around the detector. A signal recorded by a PMT housed in a Digital Optical Module (DOM) is called a hit. Synchronization of the different DOMs with a resolution better than one nanosecond is mandatory in order to fully explore the advantage provided by the different directionality of the PMTs inside the KM3NeT DOMs to achieve a precise reconstruction of the neutrino direction. The synchronization of the DOM is performed in the Field Programmable Gate Array (FPGA) of the Central Logic Board (CLB) by means of the White Rabbit (WR) protocol, which allows the data communication and the synchronizations using the available optical links.

WR is a data transmission and synchronization protocol. It is based on Ethernet, ensuring sub-nanosecond synchronization and deterministic data transfer using Precision Time Protocol (PTP) and Synchronous Ethernet (Sync-E). WR devices are connected in a master-slave way, providing sub nanosecond accuracy and picosecond precision between WR nodes. White Rabbit is a master-slave network, controlling hierarchically several endpoints and spreading the synchronization down the network. The WR switch at the vertex in the network hierarchy receives the time from Global Positioning System (GPS) and distributes it down to the lower layer of the WR hierarchy.

The main peculiarity of the KM3NeT network distribution is its physical topology connected to the necessity of a synchronized system. Every single DOM has a unidirectional 1Gbps uplink to reach the on-shore station. On the other side, the on-shore station has a unidirectional 1 Gb/s downlink to reach all the DOMs; this is called the broadcast or slow control (SC) link. The SC is shared by every single DOM, so when a DOM receives a packet, all the other DOMs also receive it. This topology highly reduces the communication resources, but it has required to customize the communication elements of the switches and nodes as the KM3NeT architecture is outside the standard White Rabbit.

In order to simplify the optical network, reducing the amount of electro-optical cables needed and easing the maintenance of the White Rabbit firmware and software, the KM3NeT Collaboration is evaluating the use of a completely standard White Rabbit architecture. The use of a White Rabbit Switch at the bottom of the Detection Units (DUs) will be needed (A development that in KM3NeT is called Wet White Rabbit Switch (WWRS). The increase of the transmission rate will challenge the current capabilities of the WR switch, currently limited to 1 Gbps. For this reason, and taking into account the limit of 18 ports in the 1Gb switch, it is necessary to use two switches working in parallel. The increase of the data rate, ideally to 10 Gbps, and the increase of the number of ports would simplify the current design, composed of two WR switches working in parallel.

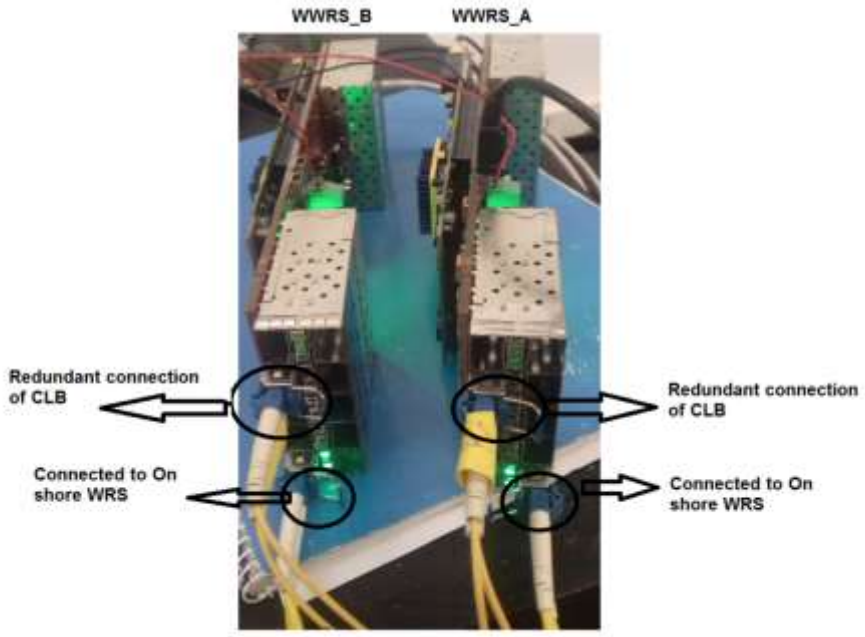

Figure 1: A Figure of the current development for the DU base showing the two White Rabbit Switches (WRS) that will be implemented in one DU.

Figure 1 shows two of the WWRS prototypes and Figure 2 and Figure 3 present the architecture proposed for Phase II, both for ARCA and ORCA. It can be observed that in both cases two WRS switches are needed.

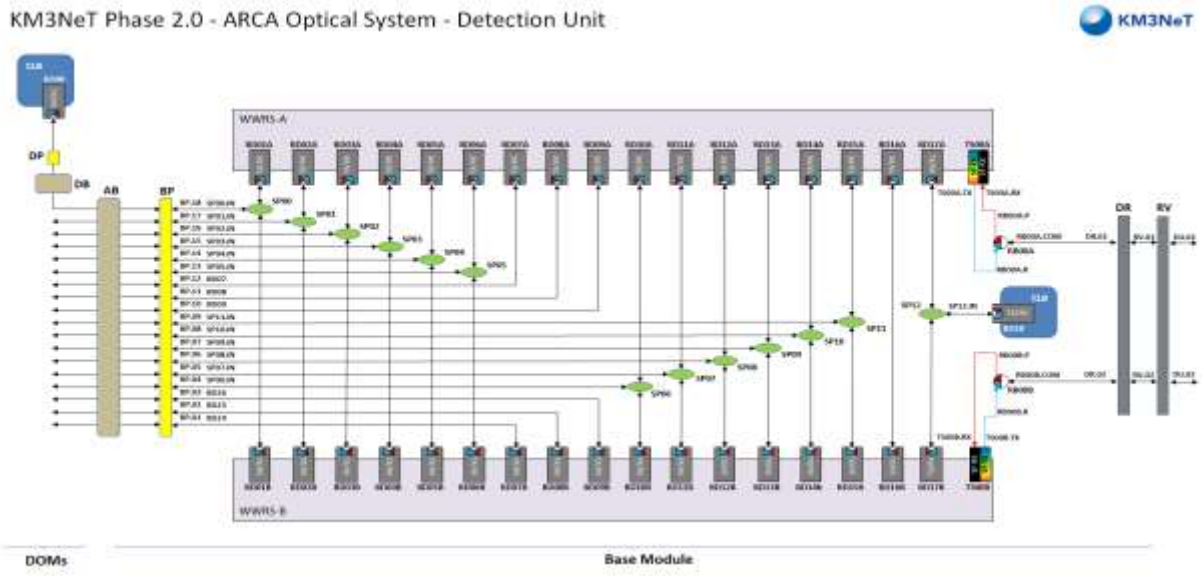

Figure 2: Schematics of the WWRS interconnections within themselves, the DOMs, the Base-module CLB and the shore station for ARCA DU.

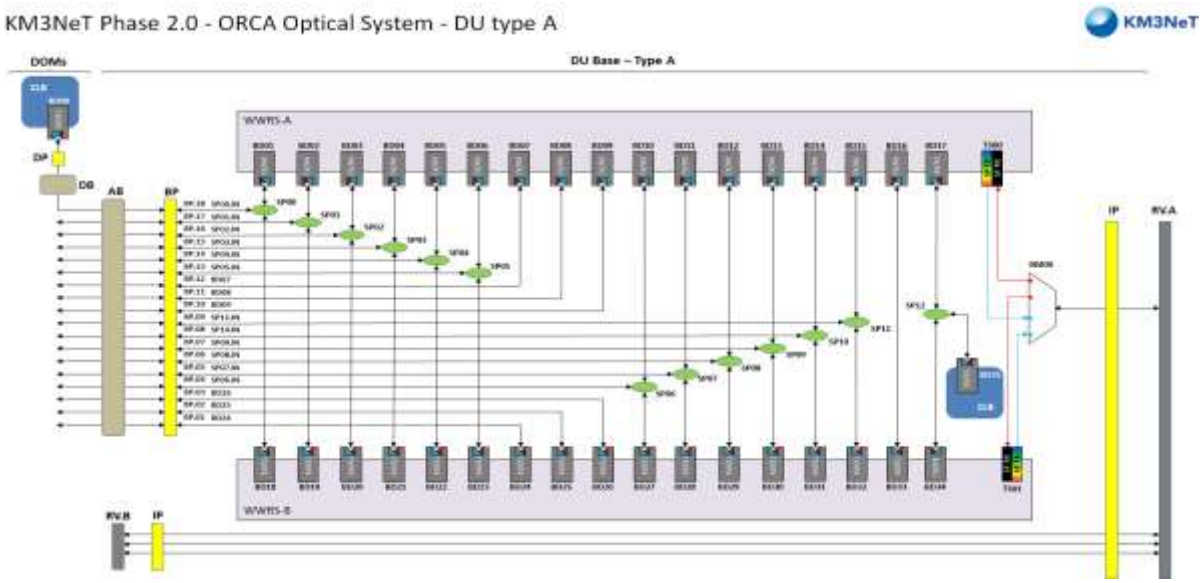

Figure 3:Schematics of the WWRS interconnections for ORCA (bottom) DU.

A White Rabbit switch of 10 Gbps will therefore greatly simplify the current design as it will allow for the possibility to use only one WR switch. This working package presents the efforts carried out to investigate the feasibility of a WRS working at 10 Gbps including the study of the oscillator system to be used as well as the FPGA latency of the devices that could be used in KM3NeT, which represents one of the limiting factors in a White Rabbit at 10 Gbps.

2. Improvements of the White Rabbit Oscillator System

Presently, enough CLBs have been produced to equip all DOMs of 31 DUs (each DU has 18 DOMs), some of which have already been deployed and are currently taking data. Taking into account the return of the experience and in order to increase the performances and the reliability of the boards, a new version of the CLB has been developed. Taking advantage of this new development, the clock system has been improved in order to reduce the phase noise of the CLB, of great importance for the improvement of the synchronization of the telescope. Other improvements are the addition of new instrumentation, such as a pressure sensor or an upgraded version of the compass sensor. In addition the SFP transceiver has been replaced by a high reliability transceiver from Glenair. A crucial implementation is the addition of another type of clock generation, which has allowed the evaluation of the performance of the boards and an investigation of the feasibility of the use of a new version of the White Rabbit switch.

In order to operate, White Rabbit needs, at the FPGA input, two different clocks, one with a frequency of 125 MHz, and another one oscillating at 124.992 MHz. This difference in phase allows for the precise delay measurements that WR is able to obtain. The two clocks are used by a Dual Mixer Time

Difference (DMTD) phase detector. The DMTD is able to decrease the fast frequencies to a lower frequency where it is possible to measure the phase shifts. For instance, a reference clock of 125 MHz and an offset clock of 124.99 MHz, as the ones implemented in the CLB, will produce an output signal of 10 kHz. Then, the phase shift can be very accurately measured using a simple counter.

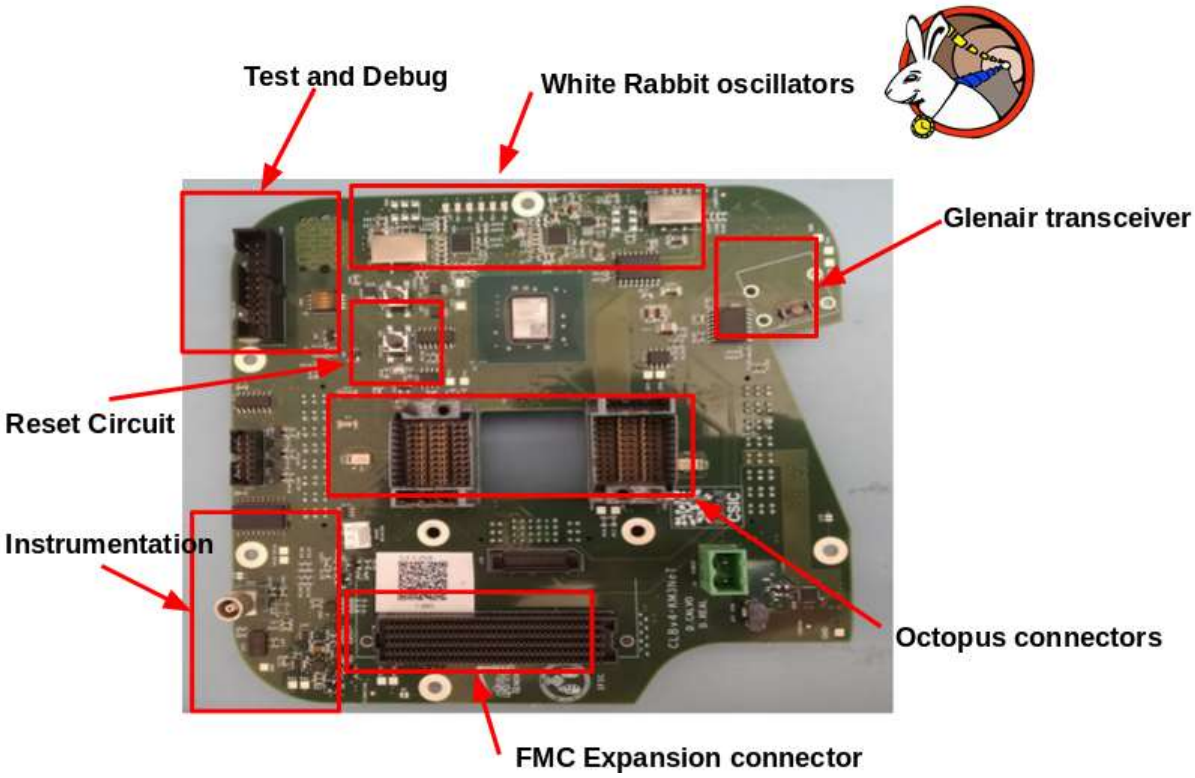

Figure 4: Prototype of CLBv4, improved version of the CLB.

The clock scheme of the CLBv4 is presented in Figure 5. In the first system implemented, a clock of 25 MHz is generated by a quartz oscillator, and, in order to create the needed 125 MHz, a CDCM (Digital Clock Manager) is used. The same applies for the generation of the 124.992 MHz. Instead, in the second system, the 125MHz and the 124.99 MHz are generated directly by a quartz oscillator manufactured to oscillate at that frequency. As the quartz creates the frequency with higher accuracy it improves the quality of the clock signal. The signals generated by both systems are available at the FPGA and it is possible to use either of them.

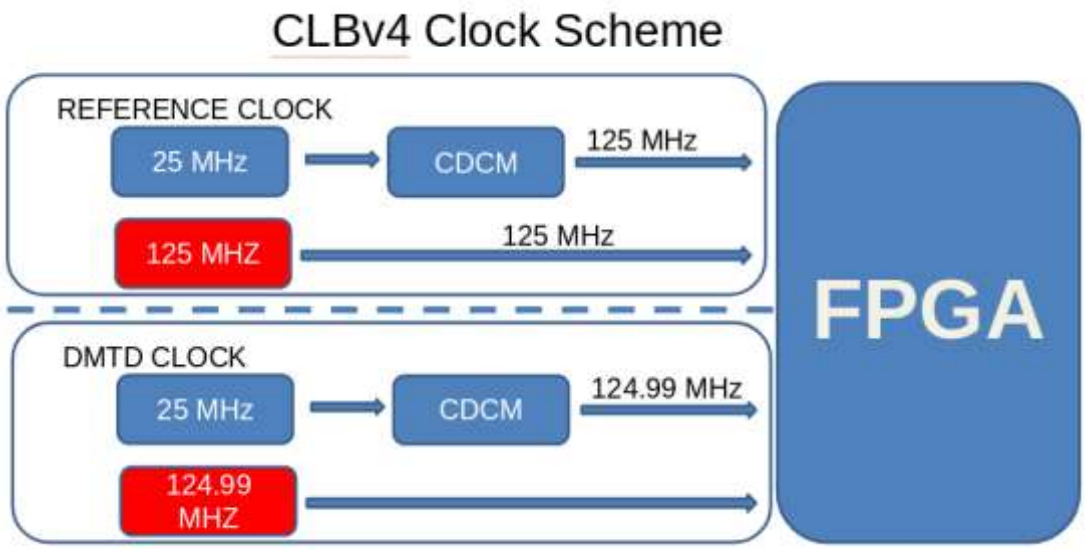

Figure 5: White Rabbit Scheme of the CLBv4.

Four prototypes have been produced, where the new clock generation system has been tested and positive results have been obtained for the stability of the clock. Tests were performed (see Figure 6) to evaluate the stability of the clock generated when used by White Rabbit. The skew of the Pulse Per Second (PPS) presents very low jitter, with values below +- 25 ps for the standard deviation. These values include also the effect of the reset and power down of the board as it is in this operation when usually higher differences are obtained. The overall results of the prototype boards are very promising for their future use in a White Rabbit switch with higher bandwidth.

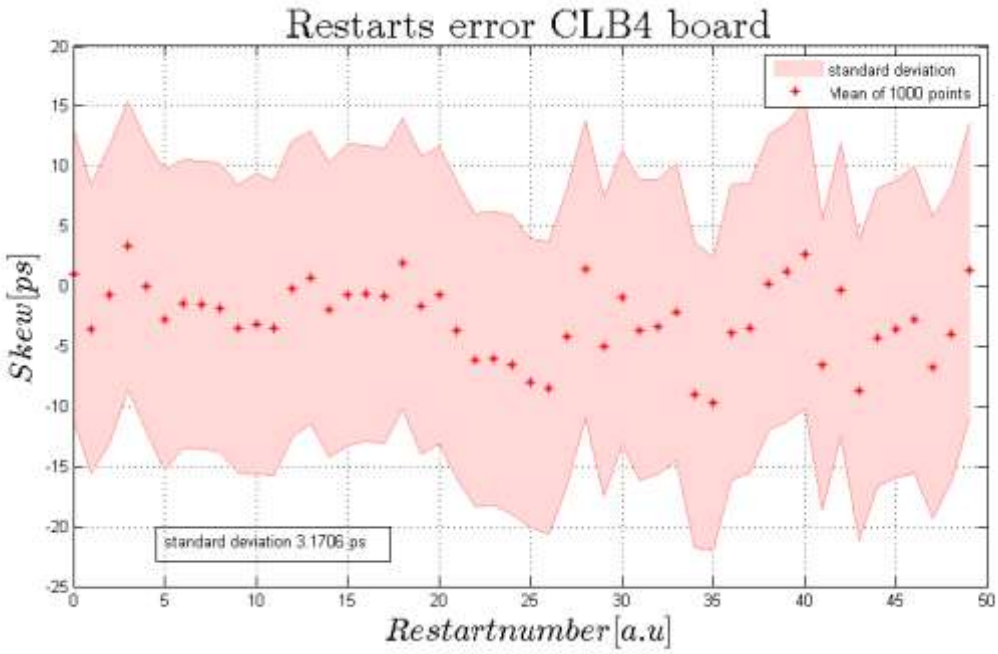

Figure 6: Tests performed on the CLBv4. The skew of the Pulse Per Second signal has been measured after the restart of the board.

3. FPGA study and latency measurement at 10 Gbps

Most of the high-speed transceivers embedded in FPGAs do not include the capability of keeping a fixed and deterministic latency, through the data-path after a reset, a resynchronization process or a power cycle. However, the use of links with fixed and deterministic latency is a requirement for the implementation of timing synchronization systems for astroparticle and high-energy physics experiments as KM3NeT, CMS or ATLAS experiments.

Figure 7: Board where the latency of the transceivers has been tested. It contains a Xilinx Virtex 7 FPGA.

In this section, the implementation of high-speed links with fixed and deterministic latency using the Tile PreProcessor Demonstrator board design at IFIC is presented. Xilinx MultiGigabit Transceivers (MGT) are composed of different hardware blocks than can be enabled or bypassed according to the requirements of the design. Figure 7 shows the board used for the present tests using the Xilinx FPGA with the MGT. The architecture of the MGT can be divided in two parts referred to as the Physical Coding Sublayer (PCS) and the Physical Media Attachment (PMA).

The PMA block contains all the analog circuitry of the MGT needed to acquire the incoming signal and transform it into a digital signal, and opposite. The Clock Data Recovery (CDR) unit is one of the most important units in the transceiver. Its role is to extract the clock from the received stream with good quality in terms of jitter and re-time the serial data. On the other hand, the PCS operate at digital level handling the encoding/decoding of the data stream, word alignment or comma detection. This layer also includes First In First Out (FIFO) memories for phase adjustment between PMA and PCS clocks, which solve any phase difference between the PMA and PCS clock domains.

The CDR and the elastic buffers are two potential sources of uncontrolled latency. The CDR provides a word clock which is generated from a higher frequency clock in the GTH (Gigabit Transceiver) Phase-Locked Loop (PLL). In order to avoid any change in the phase in the word clock after a reset or power cycle, the phase of the clock has to be monitored and properly aligned. In addition, the elastic buffers could also introduce latency variations since the number of words written into the FIFO memories is not known. For this reason, the elastic buffers have to be bypassed in order to keep fixed and deterministic latency.

Experiment testbench

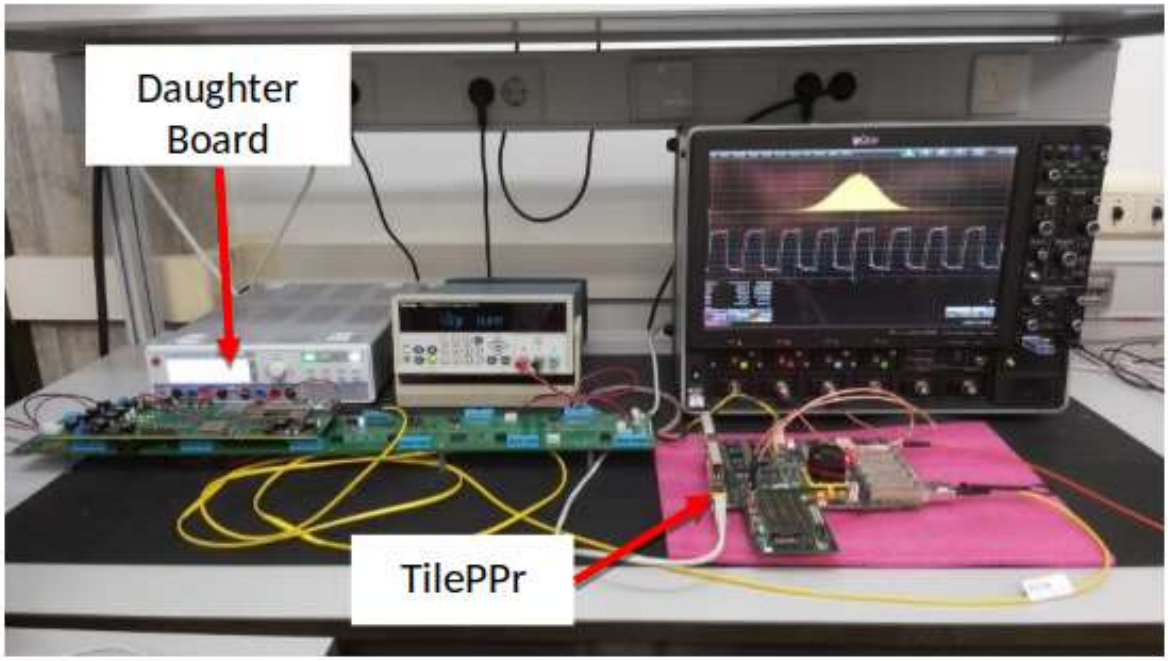

Figure 8: Experimental tests bench used at IFIC for testing the latency of the 10 Gbps transceivers of Xilinx FPGAs of family 7.

The test-bench (See Figure 8) is composed of one Tile PreProcessor Demonstrator board and one Tile DaughterBoard. Four high-speed links have been implemented in both Tile PPr Demonstrator and DaughterBoard boards using the GigaBit Transceiver (GBT) protocol with an asymmetric bandwidth (4.8 Gbps / 9.6 Gbps). In addition, a phase measurement circuit based on the Digital Dual Mixer Time Domain (DDMTD) circuit has been implemented in firmware, in order to measure the phase difference between the transmitted and recovered clock in the Tile PPr Demonstrator.

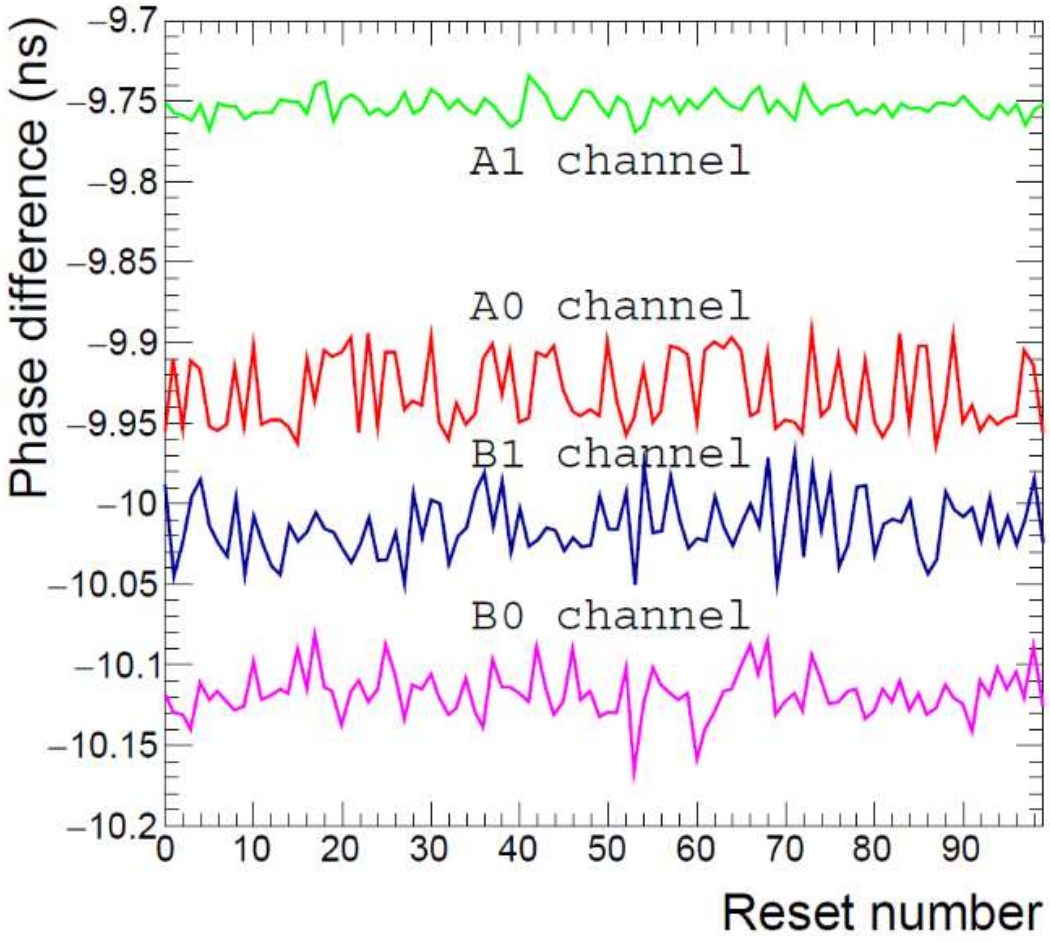

Figure 9: Experimental results of the latency of the family seven transceivers. The observed latency is below 100 ps and it meets the requirements for their use in a White Rabbit switch working at 10 Gbps. A0,A1,B0 and B1 represent each of the four tested MGTs at 9.6 Gbps.

Figure 9 presents the average of 1000 phase difference measurements between recovered and transmitted clocks of four MGTs at 9.6 Gbps path after resetting the Daughter Board 100 times. As can be observed in this Figure, the variations between the phase of the distributed and recovered clock are below 100 ps peak-to-peak. Thus, the operation of the Xilinx MGT links with fixed and deterministic latency is validated.

4. Next steps and conclusions

In order to reconstruct the neutrino direction with high precision, time synchronization of the KM3NeT Digital Optical Modules with an accuracy better than 1 ns is required. At the same time the transfer of a high quantity of data is required. This document describes the implementation of a new oscillator system and the high-speed links with fixed and deterministic latency which fulfil the requirements for the implementation of the White Rabbit protocol above 10 Gbps.

The latency of the links (in both of the studied cases) showed small phase deviations in the distributed clock produced after resetting the links. The results obtained using these transceivers and oscillator system permit the extrapolation of the results to other similar transceiver architectures as the Xilinx ZYNQ UltraScale+ MGTs. The advantage of this board is that it includes a control processing unit integrated in the same chip. This would allow to simplify the current hardware of a WRS as in the current version the microprocessor is implemented as a separate device.

Further steps involve the development of a board which include the oscillator system tested in the KM3NeT node using a Xilinx ZYNQ ultrascale FPGA, a hardware that can be used in a WRS working at 10 Gbps.

IX. REFERENCES

1. *Letter of Intent for KM3NeT 2.0.* Adrián-Martínez, S., et al. 2016, Journal of Physics G: Nuclear and Particle Physics, Vol. 43 (8), p. 084001. arXiv:1601.07459 [astro-ph.IM]. DOI: 10.1088/0954-3899/43/8/084001.

2.*"White Rabbit RFC 2889 Benchmarking Methodology for LAN Switching Devices"*, Cesar Prados and Jiaoni Bai, Timing Group, CSCO, GSI, August 2016.

3. "The Trigger and Data Acquisition System for the KM3NeT neutrino telescope", T. Chiarusi, C. Pellegrino, EPJ Web of Conferences 116, 05005 (2016) DOI: 10.1051/epjconf/201611605005

4. . Moreira, P. Alvarez, J. Serrano, I. Darwezeh, and T. Wlostowski,"Digital Dual Mixer Time Difference for Sub-Nanosecond Time Syn-chronization in Ethernet,"FCS 2010 IEEE International, 2010.