EPJ Web of Conferences **116**, 05008 (2016)
DOI: 10.1051/epjconf/201611605008

NaNet^3: The on-shore readout and slow-control board for the KM3NeT-Italia underwater neutrino telescope

R. Ammendola[2], A. Biagioni[1,a], O. Frezza[1], F. Lo Cicero[1], M. Martinelli[1], P.S. Paolucci[1], L. Pontisso[3], F. Simula[1], P. Vicini[1], F. Ameli[1], C.A. Nicolau[1], E. Pastorelli[1], F. Simeone[1], L. Tosoratto[1], and A. Lonardo[1]

[1] INFN Sezione di Roma, Sapienza, P.le Aldo Moro, 2, 00185 Roma, Italy
[2] INFN Sezione di Roma, Tor Vergata, via della Ricerca Scientifica, 1, 00133 Roma, Italy
[3] INFN Sezione di Pisa, via F. Buonarroti 2, 56127 Pisa, Italy

Abstract. The KM3NeT-Italia underwater neutrino detection unit, the tower, consists of 14 floors. Each floor supports 6 Optical Modules containing front-end electronics needed to digitize the PMT signal, format and transmit the data and 2 hydrophones that reconstruct in real-time the position of Optical Modules, for a maximum tower throughput of more than 600 MB/s. All floor data are collected by the Floor Control Module (FCM) board and transmitted by optical bidirectional virtual point-to-point connections to the on-shore laboratory, each FCM needing an on-shore counterpart as communication endpoint. In this contribution we present NaNet^3, an on-shore readout board based on Altera Stratix V GX FPGA able to manage multiple FCM data channels with a capability of 800 Mbps each. The design is a NaNet customization for the KM3NeT-Italia experiment, adding support in its I/O interface for a synchronous link protocol with deterministic latency at physical level and for a Time Division Multiplexing protocol at data level.

1. Introduction

The ever-increasing bandwidth requirements for High Energy Physics experiments are often so constrained as to require custom technological solutions to be satisfied. The KM3NeT-Italia experiment [1], being an evolution of the NEMO Phase-2 experiment [2], inherits the interface of the on-shore endpoints for the electronic boards that gather the data collected by the underwater detector. Compared to NEMO Phase-2, the desired ramp-up in number of channels that KM3NeT-Italia aims for would cause, if the same readout boards as NEMO Phase-2 were used as endpoints, an unmanageable encumbrance at the on-shore receiving installation. The need of more tightly aggregated channels calls for an improved endpoint but incurs the obvious costs of its redesign. This endpoint consolidation can instead be achieved with a modicum of implementation effort – mainly due to the enforcement of deterministic latency on the board transceivers for system-wide distribution of a common timing – by

[a] e-mail: andrea.biagioni@roma1.infn.it

a customization of the NaNet design [3], the NaNet^3 board, that allows for a cost-effective scaling in the receiving installation.

2. The KM3NeT-Italia DAQ and data transport architecture

KM3NeT-Italia [4] is an underwater experimental apparatus for the detection of high energy neutrinos in the TeV÷PeV range based on the Čerenkov technique. The detector consists of a tridimensional array of photomultiplier tubes (PMTs) exploiting the Čerenkov effect induced by superluminal charged particles in water. The hit is the response of a PMT at the passage of particles. The temporal and spatial distribution of the hits allows reconstructing the particles direction. Accurate knowledge of the position of the PMTs is mandatory to accomplish this task. The required spatial accuracy is 40 cm.

The final assessment of the experiment foresees the installation of 8 detection units. The detection unit is called *tower* and consists of 14 floors vertically spaced 20 meters apart. The floor consists of 6 optical modules (OMs) containing a 10 inch PMT each, the front-end electronics needed to digitize the signal, format and transmit the data and 2 hydrophones to reconstruct in real-time the OM position. A board called *Floor Control Module* (FCM) collects the data coming from the underwater devices and communicates with the on-shore laboratory through an optical fiber.

The main requirement for the endpoint board in the laboratory is the distribution of a common timing all over the system in order to correlate signals from different parts of the apparatus with the required ~ 1 ns resolution. Every data frame needs to be labelled with a “time stamp” in order to reconstruct the tracks of charged particles. The described constraints hinted to the choice of a synchronous link protocol which embeds clock and data with deterministic latency; due to the distance between the apparatus and shoreland, the transmission medium is forced to be an optical fiber.

The second requirement is keeping the infrastructure cost-effective: an on-shore readout board supporting multiple channel allows for more efficient scaling of the PC farm size.

3. NaNet^3

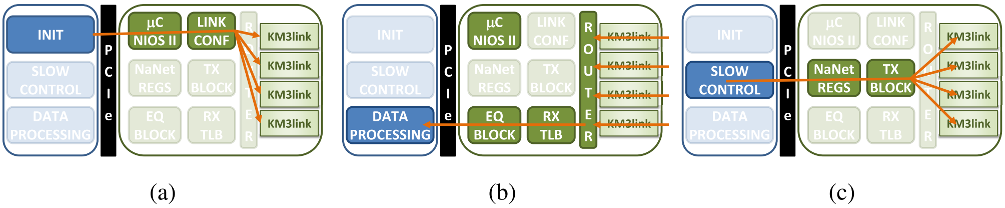
The NaNet board family aims at providing a bridge between the readout of the detectors and the computing nodes of a PC farm by means of a real-time data transport mechanism.

NaNet^3 is a PCIe Gen2 Network adapter implemented on the Terasic DE5-net board, which is equipped with an Altera Stratix V FPGA and features four I/O channels with maximum capability of 800 Mbps each. The hardware follows the RDMA paradigm to manage the I/O towards both CPU and GPU – for this latter the protocol is more precisely the GPUDirect RDMA [5], – in this way avoiding the use of bounce buffers and minimizing the latency jitter of data transfers to and from application memory. The virtual-to-physical address translation is entrusted to a proprietary Translation Look-aside Buffer based on Content Addressable Memory [6]. In order to sustain the ~ 320 MB/s aggregate bandwidth of the multi-channel system, incoming data are sent to the computing node by PCIe DMA write processes. The outbound flow consists instead of messages for slow control management of the underwater devices; since requested bandwidth does not exceed ~ 240 kB/s, the task can be accomplished with PCIe target read processes.

The NaNet^3 Router module can support up to 6 data streams with a capability of 2.8 GB/s, multiplexing the incoming messages towards the receiving port and applying a deterministic routing policy.

Finally, the KM3link data transmission system is based on Time Division Multiplexing (TDM) protocol and guarantees deterministic latency of the communication.

GPS signals – i.e. clock and IRIG data – are received from the two SMA connectors on board.



(a) (b) (c)

Figure 1. Basic functions paths within NaNet^3. Software tasks are on the left side of the PCIe bus, the hardware tasks are on the right.

3.1 NaNet^3 configuration and data flow

In this section, overall board operativity with a focus on major functionalities is provided.

During *Hardware/Software Initialization* (Fig. 1a), several hardware modules are enabled by the NaNet driver which initializes the board configuration registers. Incoming data are collected in circular lists of persistent buffers (CLOPs) on the target device memory. The driver registers the buffers and memory pages are pinned and locked. The Nios II microcontroller on board sets up the translation of their addresses from virtual memory space to the physical one the network adapter is able to handle. A custom hardware module takes care of setting up the CLOP parameters – i.e. number of buffers, buffer size and virtual address – for the KM3links. The result is a channel able to receive and transmit messages without further OS involvement – the network stack protocol is totally hardware-managed.

In the *Data Reception and Processing* stage (Fig. 1b), incoming TDM data is translated in a packet-based protocol optimizing PCIe communication. A custom module computes the destination virtual address to allocate the data into the CLOPs and patches it into the packet header. The four data stream, one for each KM3link, are dispatched to the receiving port through the *Router*. The receiving module manages the data flow and operates Virtual-to-Physical translations exploiting a TLB. Its entries are pre-registered in order to speed-up the hardware processing. A DMA write transaction moves data to the computing node memory and an event queue collects the completion messages of the writing process; an interrupt is raised to notify the system. The driver signals the reception of new frames to the application, kicking off the computing process.

Finally, *Slow Control Transmission* (Fig. 1c) is handled by the kernel driver. Data addressed to the underwater devices are stored in registers (one for each underwater device) through PCIe target mode, then a custom TX module dispatches the messages to the proper channel according to the register address. Data is allocated in a 10 kB sized frame according to TDM protocol. One data frame is transmitted every 125 μs, with a GPS tagging every frame with a 12.5 ns precision timestamp.

3.2 Deterministic latency

A channel featuring deterministic latency is the main requirement of the DAQ architecture in order to achieve system-wide common timing. The KM3link exploits the Altera Deterministic Latency PHY IP core enabling accurate delay measurements and known timing for the transmit (TX) and receive (RX) datapaths. The core Physical Coding Sublayer offers, among other things, 8B10B encoder/decoder, Word Aligner and TX bit slipper.

NaNet^3 is in charge of distributing the timing information and signal to the connected underwater devices enforcing the constraint of having a *ns*-precise offset between the wave-fronts of the system clocks. A GPS clock acts as reference and it is used for the optical link transmission from the on-shore board towards the underwater FCM. NaNet^3 acts as master of the communication managing the reference clock. The FCM recovers the clock from the incoming stream and employs it to send its payload from the apparatus back to NaNet^3. Finally, the receiving module of KM3link recovers

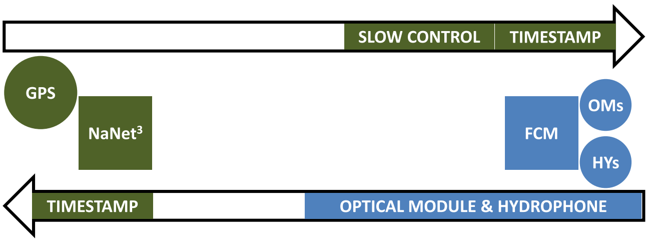


Figure 2. NaNet^3-FCM data flow; *timestamp* and *Slow control* segments are injected by NaNet^3 while *Optical module & Hydrophone* segment is injected by the FCM.

the clock by decoding this payload at loop end, therefore guaranteeing the establishment of a fully synchronous link. A scheme of the data flow is depicted in Fig. 2. Full control of the communication latency between NaNet^3 and FCM board, based on Xilinx FPGA, is mandatory to obtain the knowledge of timing of the overall system. The Altera Deterministic Latency PHY IP guarantees its functioning when sender and receiver are directly connected. In the present case, the off-shore device breaks the link making the TX bit slipper ineffective. The latency varies whenever alignment procedure of the communication channel is executed – i.e. the reboot of the host server or substitution of the cable. A custom module was developed to tackle the issue. The hardware logic allows for setting the desired bit slip value during the alignment procedure through a configuration register of NaNet^3. The procedure runs until the number of bit slip needed to acknowledge the alignment-word matches the required one.

4. Conclusion

In this paper we presented the NaNet^3 board, a customization of the NaNet design dedicated to the KM3NeT-Italia experiment. The NaNet -to-FCM slow control transmission was successfully tested for distances up to 100 Km. The multichannel reception system is being lab-tested and has reached 2 continuous weeks of error-free data collection.

References

[1] *KM3NeT ITALIA*, accessed: 26/Gen/2016, `https://web2.infn.it/KM3NeT-Italia/`

[2] S. Aiello, F. Ameli, M. Anghinolfi, G. Barbarino, E. Barbarito, F. Barbato, N. Beverini, S. Biagi, B. Bouhadef, C. Bozza et al., Astroparticle Physics **66**, 1 (2015), ISSN 0927-6505, `http://www.sciencedirect.com/science/article/pii/S0927650514001960`

[3] A. Lonardo, F. Ameli, R. Ammendola, A. Biagioni, A.C. Ramusino, M. Fiorini, O. Frezza, G. Lamanna, F. Lo Cicero, M. Martinelli et al., Journal of Instrumentation **10**(04), C04011 (2015), `http://stacks.iop.org/1748-0221/10/i=04/a=C04011`

[4] A. Margiotta, Journal of Instrumentation **9**(04), C04020 (2014), `http://stacks.iop.org/1748-0221/9/i=04/a=C04020`

[5] R. Ammendola et al., *GPU Peer-to-Peer Techniques Applied to a Cluster Interconnect*, in *Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW), 2013 IEEE 27th International* (2013), pp. 806–815

[6] R. Ammendola et al., *Virtual–to–Physical address translation for an FPGA–based interconnect with host and GPU remote DMA capabilities*, in *Field-Programmable Technology (FPT), 2013 International Conference* (2013), pp. 58–65