NaNet: a configurable NIC bridging the gap between HPC and real-time HEP GPU computing

2015 JINST 10 C04011

(http://iopscience.iop.org/1748-0221/10/04/C04011)

PUBLISHED BY IOP PUBLISHING FOR SISSA MEDIALAB

RECEIVED: *November 15, 2014*
REVISED: *January 15, 2015*
ACCEPTED: *January 26, 2015*
PUBLISHED: *April 10, 2015*

TOPICAL WORKSHOP ON ELECTRONICS FOR PARTICLE PHYSICS 2014,
22–26 SEPTEMBER 2014,
AIX EN PROVENCE, FRANCE

NaNet: a configurable NIC bridging the gap between HPC and real-time HEP GPU computing

A. Lonardo,[a,1] **F. Ameli,**[a] **R. Ammendola,**[b] **A. Biagioni,**[a] **A. Cotta Ramusino,**[c] **M. Fiorini,**[c] **O. Frezza,**[a] **G. Lamanna,**[d,e] **F. Lo Cicero,**[a] **M. Martinelli,**[a] **I. Neri,**[c] **P.S. Paolucci,**[a] **E. Pastorelli,**[a] **L. Pontisso,**[f] **D. Rossetti,**[g] **F. Simeone,**[a] **F. Simula,**[a] **M. Sozzi,**[f,e] **L. Tosoratto**[a] **and P. Vicini**[a]

[a] *INFN Sezione di Roma - Sapienza,*
P.le Aldo Moro, 2 - 00185 Roma, Italy
[b] *INFN Sezione di Roma - Tor Vergata,*
Via della Ricerca Scientifica, 1 - 00133 Roma, Italy
[c] *Università degli Studi di Ferrara and INFN Sezione di Ferrara,*
Polo Scientifico e Tecnologico,
Via Saragat 1 - 44122 Ferrara, Italy
[d] *INFN Laboratori Nazionali di Frascati,*
Via E. Fermi,40 - 00044 Frascati (Roma), Italy
[e] *CERN,*
CH-1211 Geneva 23, Switzerland
[f] *INFN Sezione di Pisa,*
Via F. Buonarroti 2 - 56127 Pisa, Italy
[g] *NVIDIA Corp,*
2701 San Tomas expressway, Santa Clara, CA 95050

E-mail: alessandro.lonardo@roma1.infn.it

ABSTRACT: NaNet is a FPGA-based PCIe Network Interface Card (NIC) design with GPUDirect and Remote Direct Memory Access (RDMA) capabilities featuring a configurable and extensible set of network channels. The design currently supports both standard — Gbe (1000BASE-T) and 10GbE (10Base-R) — and custom — 34 Gbps APElink and 2.5 Gbps deterministic latency KM3link — channels, but its modularity allows for straightforward inclusion of other link technologies. The GPUDirect feature combined with a transport layer offload module and a data stream processing stage makes NaNet a low-latency NIC suitable for real-time GPU processing. In this

[1] Corresponding author.

doi:10.1088/1748-0221/10/04/C04011

paper we describe the NaNet architecture and its performances, exhibiting two of its use cases: the GPU-based low-level trigger for the RICH detector in the NA62 experiment at CERN and the on-/off-shore data transport system for the KM3NeT-IT underwater neutrino telescope.

KEYWORDS: Trigger concepts and systems (hardware and software); Computing (architecture, farms, GRID for recording, storage, archiving, and distribution of data)

Contents

1 **Introduction** 1

2 **NaNet design overview** 1

3 **NaNet-1: a NIC for the NA62 GPU-based low-level trigger prototyping** 3

4 **NaNet^3: the on-shore readout and slow-control board for the KM3NeT-IT underwater neutrino telescope** 5

5 **Conclusions and future work** 8

2015 JINST 10 C04011

1 Introduction

The GPGPU paradigm, i.e. general purpose computing performed on graphics processing units, has strongly established itself in the High Performance Computing (HPC) arena: GPU-accelerated clusters have been taking the highest ranks of Top500 list over the last few years. Applications in virtually all fields of Computational Physics — such as Lattice Quantum Chromo-Dynamics or Fluid Dynamics — have now been reimplemented exploiting the fine-grained parallelism of GPUs so as to take advantage of their large processing power; the gains in execution times are significant, often outstanding. Similar reimplementation efforts are ongoing in several fields of Experimental Physics, ranging from Radio Astronomy to High Energy Physics.

In these contexts, processing data streams incoming from experimental apparatuses is most often characterized by real-time constraints. GPUs show a stable processing latency once data are available in their own internal memories, therefore a data transport mechanism with deterministic or at least bound latency is crucial in building a GPGPU system honouring these constraints. Such a mechanism of transport towards processor or GPU memories of experimental data is implemented by our NaNet design. The design reuses several intellectual properties developed for the APEnet+ 3D torus NIC [1] which is dedicated to HPC hybrid clusters, adding support for several standard and custom link technologies and modules for the system real-time characterization. The result is a highly configurable, modular design of a FPGA-based low-latency PCIe NIC with RDMA and GPUDirect capabilities that has been employed, with different configurations and physical device implementations, in two different High Energy Physics experiments: the NA62 experiment at CERN [2] and the KM3NeT-IT underwater neutrino telescope [3].

2 NaNet design overview

NaNet is a modular design of a low-latency PCIe RDMA NIC supporting different network links, namely standard GbE (1000BASE-T) and 10GbE (10Base-R), besides custom 34 Gbps APElink [4]

and 2.5 Gbps deterministic latency optical KM3link [5]. NaNet includes a network stack protocol offload engine yielding a very stable communication latency, a feature making it suitable for use in real-time contexts. Moreover, NaNet acts as a bridge between the worlds of real-time and GPGPU heterogeneous computing thanks to its GPUDirect P2P/RDMA capability, inherited from its HPC-dedicated sibling, the APEnet+ 3D torus NIC.

NaNet design is partitioned into 4 main modules: *I/O Interface*, *Router*, *Network Interface* and *PCIe Core* (see figure 1).

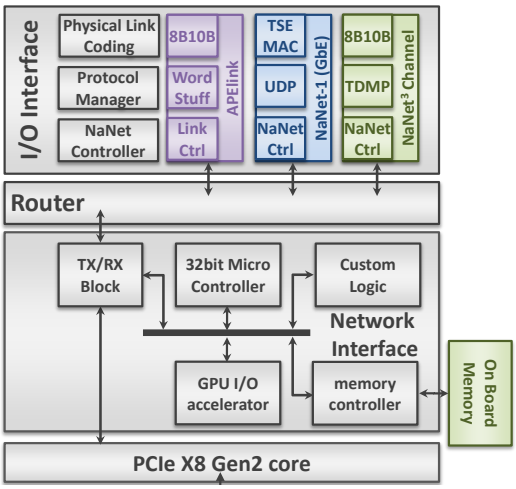

Figure 1. NaNet architecture schematic.

The I/O Interface module performs a 4-stages processing on the data stream: following the OSI Model, the Physical Link Coding stage implements, as the name suggests, the channel physical layer (e.g. 1000BASE-T) while the Protocol Manager stage handles, depending on the kind of channel, data/network/transport layers (e.g. Time Division Multiplexing or UDP); the Data Processing stage implements application-dependent reshuffling on data streams (e.g. performing de/compression) while the APEnet Protocol Encoder performs protocol adaptation, encapsulating inbound payload data in APElink packet protocol — used in the inner NaNet logic — and decapsulating outbound APElink packets before re-encapsulating their payload into output channel transport protocol (e.g. UDP).

The Router module supports a configurable number of ports implementing a full crossbar switch responsible for data routing and dispatch. Number and bit-width of the switch ports and the routing algorithm can all be defined by the user to automatically achieve a desired configuration. The Router block dynamically interconnects the ports and comprises a fully connected switch, plus routing and arbitration blocks managing multiple data flows @2.8 GB/s.

The *Network Interface* block acts on the trasmitting side by gathering data incoming from the PCIe port and forwarding them to the Router destination ports; on the receiving side it provides support for RDMA in communications involving both the host and the GPU (via a dedicated *GPU I/O Accelerator* module). A Nios II μcontroller handles configuration and runtime operations.

Finally, the PCIe Core module is built upon a powerful commercial core from PLDA that sports a simplified but efficient backend interface and multiple DMA engines.

2015 JINST 10 C04011

This general architecture has been specialized up to now into three configurations, namely NaNet-1, NaNet^3 and NaNet-10, to match the requirements of different experimental setups.

NaNet-1 features a PCIe Gen2 x8 host interface plus a GbE one, three optional 34 Gbps APElink channels and is implemented on the Altera Stratix IV FPGA Development Kit (see figure 2).

NaNet^3 is implemented on the Terasic DE5-net Stratix V FPGA development board sporting four SFP+ cages and supports four 2.5 Gbps deterministic latency optical KM3link channels and a PCIe Gen2 x8 host interface.

NaNet-10 features four 10GbE SFP+ ports along with a PCIe Gen2 x8 host interface and is also being implemented on the Terasic DE5-net board.

3 NaNet-1: a NIC for the NA62 GPU-based low-level trigger prototyping

The NA62 experiment at CERN aims at measuring the Branching Ratio of the ultra-rare decay of the charged Kaon into a pion and a neutrino-antineutrino pair. The NA62 goal is to collect $\sim$ 100 events with a 10:1 signal to background ratio, using a novel technique with a high-energy (75 GeV) unseparated hadron beam decaying in flight. In order to manage the 25 GB/s raw data stream due to a $\sim$ 10 MHz rate of particle decays illuminating the detectors, the trigger system is designed as a set of three cascaded levels that decrease this rate by three orders of magnitude [6]. The low-level trigger (L0) is a synchronous real-time system implemented in hardware by means of FPGAs on the readout boards and reduces the stream bandwidth tenfold: whether the data on the readout board buffers is to be passed on to the higher levels has to be decided within 1 ms to avoid data loss. The upper trigger levels (L1 and L2) are implemented in software on a commodity PC farm for further reconstruction and event building. In the baseline implementation, the FPGAs on the readout boards compute simple trigger primitives on the fly, such as hit multiplicities and rough hit patterns, which are then timestamped and sent to a central trigger processor for matching and trigger decision. A pilot project within NA62 is investigating the possibility of using a GPGPU system as L0 trigger processor (GL0TP), exploiting the GPU computing power to process unfiltered

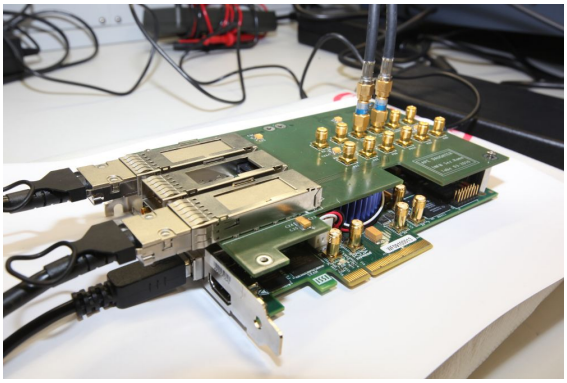

Figure 2. NaNet-1 on Altera Stratix IV dev. board EP4SGX230KF40C2 with custom mezzanine card + 3 APElink channels.

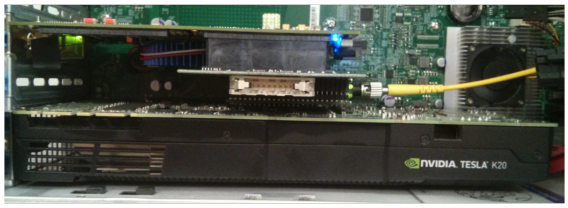

Figure 3. GL0TP prototype: NaNet-1 with a plugged TTC HSMC card is coupled with a nVIDIA Kepler K20 GPU for low-latency operation.

2015 JINST 10 C04011

data from the readout in order to implement more selective trigger algorithms. In order to satisfy the real-time requirements of the system, a fast, reliable and time-deterministic dedicated link must be employed. NaNet-1 has been designed and developed with the motivation of building a fully functional and network-integrated GL0TP prototype, albeit with a limited bandwidth with respect to the experiment requirements, in order to demonstrate as soon as possible the suitability of the approach and eventually evolve the design to incorporate more capable I/O channels.

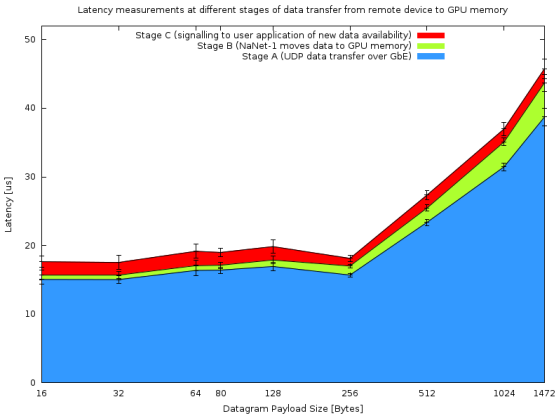

Figure 4. NaNet-1 latency at different stages of UDP transfer over GbE towards GPU memory for varying datagram payload sizes. (nVIDIA Fermi C2050).

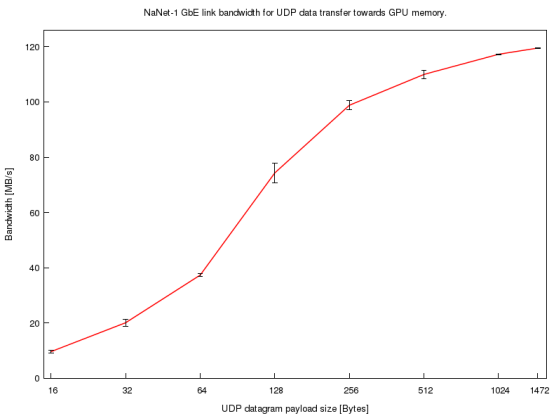

Figure 5. NaNet-1 GbE bandwidth of UDP data transfer towards GPU memory at different datagram payload sizes. (nVIDIA Fermi M2070).

The real-time characterization of NaNet-1 has been carefully assessed on dedicated testbeds [7, 8]. The results of this activity have driven the latest developments towards a lower latency design. Allocation of time-consuming RDMA related tasks has been moved from the Nios II μcontroller to dedicated logic blocks. The Virtual Address Generator (VAG) included in the NaNet Controller module is in charge of generating memory addresses of the receiving buffers for incoming data while a Translation Lookaside Buffer (TLB) module, implemented as an associative cache, performs fast virtual-to-physical address translations: a single mapping operation takes only $\sim$ 200 ns [9]. When compared with μcontroller based operations, datagram handling latency is lower and, more importantly, its jitter is greatly reduced: for 1 KB payload, datagrams latency is almost halved while its standard deviation value drops by a factor of about ten. In figure 4, the latency of a UDP GbE data transfer from a remote device to GPU memory is split in three consecutive stages: stage A accounts for time elapsed since start of send operation on the remote device until the first word of the incoming datagram exits from NaNet-1 GbE MAC; stage B ends at the completion of the PCIe DMA of datagram payload towards GPU memory and yields datagram traversal time for NaNet-1; stage C accounts for the delay of the hw/sw stack signaling to the userspace application that a new datagram payload was received. For payload sizes up 128 Bytes, NaNet-1 logic handles datagrams in less than 1 μs. Figure 5 shows NaNet-1 GbE link to GPU memory bandwidth figures at varying datagram payload sizes.

Besides optimizing performances, we undertook several developments to cope with the NaNet-1 integration within the NA62 experiment.

A Timing Trigger and Control (TTC) HSMC daughtercard was designed and developed to provide NaNet-1 with the capability of receiving either trigger and 40 MHz clock streams distributed from the experiment TTC system via optical cable: this common reference clock was used to perform latency measurements discussed above.

A decompressor stage was added in the I/O interface to reformat events data in a GPU-friendly fashion on the fly.

Finally, a timeout mechanism was implemented in the NaNet Controller module, triggering the DMA of received data towards CPU/GPU memory on a configurable deadline rather than on the filling of a receive buffer.

The first prototype of the GL0TP has recently been deployed at the NA62 experiment site. The prototype integrates NaNet-1 with a TTC HSMC daughtercard and a nVIDIA Kepler K20 GPU as shown in figure 3. The GL0TP will operate in parasitic mode with respect to the main L0 trigger processor and, at least in the initial phases of the study, will process data from only one detector, the Ring Imaging Čerenkov (RICH) detector.

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

KM3NeT-IT is an underwater experimental apparatus for the detection of high energy neutrinos in the TeV÷PeV range by means of the Čerenkov technique. The detector consists of an array of photomultipliers (PMT) that measure the visible Čerenkov photons induced by charged particles propagating in sea water at speed larger than that of light in the medium. The charged particle track can be reconstructed measuring the time of arrival of the Čerenkov photons on the PMTs, whose positions must be known. The KM3NeT-IT detection unit is called *tower* and consists of 14 floors vertically spaced 20 meters apart. The floor arms are about 8 m long and support 6 glass spheres called Optical Modules (OM): 2 OMs are located at each floor end and 2 OMs in the middle of the floor; each OM contains one 10 inches PMT and the front-end electronics needed to digitize the PMT signal, format and transmit the data. Each floor hosts also two hydrophones, used to reconstruct in real-time the OM position, and, where needed, oceanographic instrumentation to monitor site conditions relevant for the detector. All data produced by OMs, hydrophones and instruments is collected by an electronic board contained in a vessel at the floor centre; this board called *Floor Control Module* (FCM) manages the communication between the on-shore laboratory and the underwater devices, also distributing the timing information and signals. Timing resolution is fundamental in track reconstruction, i.e. pointing accuracy in reconstructing the source position in the sky. An overall time resolution of about 3 ns yields an angular resolution of 0.1 degrees for neutrino energies greater than 1 TeV. Such resolution depends on electronics but also on position measurement of the OMs, which is, in fact, continuously tracked. The spatial accuracy required should be better than 40 cm.

The DAQ architecture is heavily influenced by the need of a common clock distributed all over the system in order to correlate signals from different parts of the apparatus with the required nanosecond resolution. The aim of the data acquisition and transport electronics is to label each signal with a "time stamp", i.e. the hit arrival time, in order to reconstruct tracks. This implies that the spatially distributed parts of the readout electronics require a common timing and a known delay

2015 JINST 10 C04011

with respect to a fixed reference. The described constraints hinted to the choice of a synchronous link protocol which embeds clock and data with a deterministic latency; due to the distance between the apparatus and shoreland, the transmission medium is forced to be an optical fiber.

All floor data produced by the OMs, the hydrophones and other devices that monitor the apparatus status and environmental conditions is collected by the Floor Control Module (FCM) board, packed together and transmitted along the optical link. Each floor is independent from the others and is connected by an optical bidirectional virtual point-to-point connection to the on-shore laboratory.

A single floor data stream delivered to shore has a rate of $\sim$300 Mbps, while the shore-to-underwater communication data rate is much lower, consisting only of slow-control data for the apparatus. To preserve optical power budget, the link speed is operated at 800 Mbps, which, using an 8B10B encoding, accounts for a 640 Mbps of user payload, well beyond experimental requirement.

Each FCM needs an on-shore communication endpoint counterpart. The limited data rate per FCM compared with state-of-the-art link technologies led us to designing NaNet^3, an on-shore readout board able to manage multiple FCM data channels.

This design represents a NaNet customization for the KM3NeT-IT 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 (see figure 1).

The first design stage for NaNet^3 was implemented on the Terasic DE5-net board, which is based on an Altera Stratix-V GX FPGA and supports up to 4 SFP+ channels and a PCIe x8 edge connector. The first constraint to be satisfied by the design is having a time delta with nanosecond precision between the wavefronts of three clocks:

- the first clock is an on-shore reference one (typically coming from a GPS and redistributed by custom fanout boards) and is used for the optical link transmission from NaNet^3 towards the underwater FCM;
- the second clock is recovered from the incoming data stream by a Clock and Data Recovery (CDR) module at the receiving end of the FCM which uses it for sending its data payload from the apparatus back on-shore;
- a third clock is again recovered by NaNet^3 while decoding this payload at the end of the loop.

The second fundamental constraint is the deterministic latency that the Altera Stratix device must enforce — as the FCM does — on both forward and backward paths to allow correct time stamping of events on the PMT. The link established in this way is synchronous, i.e. clock rate is the same for both devices with a fixed phase displacement.

In this way, the NaNet^3 board plays the role of a bridge between the 4 FCMs and the FCM-Server — i.e. the hosting PC — through the PCIe bus. Control data en-route to the underwater apparatus are correctly sent over the PCIe bus to the NaNet^3 board, which then routes the data to the required optical link. On the opposite direction, both control and hydrophones data plus signals from the front-end boards are extracted from the optical link and re-routed on the PCIe bus towards an application managing all the data. The GPUDirect RDMA features of NaNet, fully imported into NaNet^3 design, will allow us, at a later stage, to build an effective, real-time, GPU-based platform, in order to investigate improved trigger and data reconstruction algorithms.

2015 JINST 10 C04011

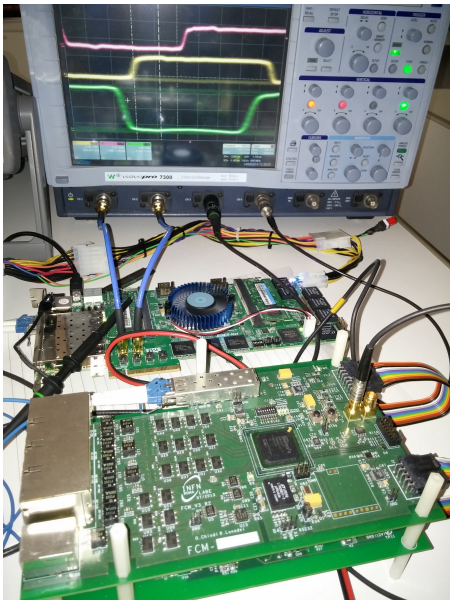

Figure 6. NaNet[3] testbed for link deterministic latency: in foreground the FCM board, in background NaNet[3] on a Terasic DE5-net.

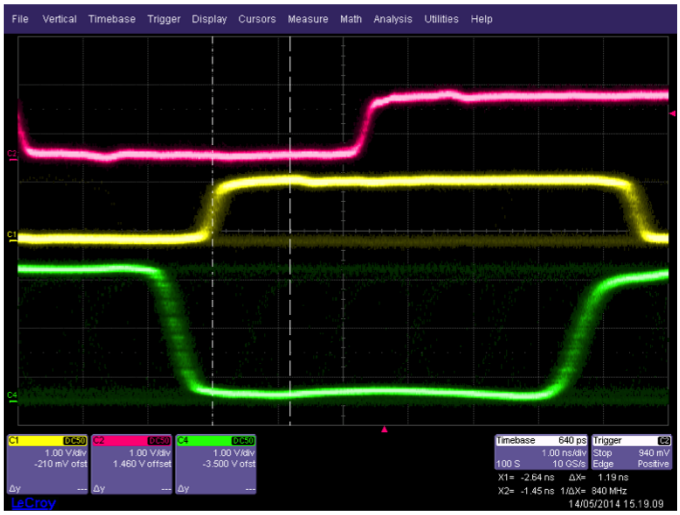

Figure 7. Deterministic latency feature of NaNet[3] SerDes: the plot scope shows the phase alignment of the transmitting (purple) and receiving (yellow) parallel clocks after 12 h test of periodic reset and initialisation sequence.

At a higher level, two systems handle the data that come from and go to the off-shore devices: the Trigger System, which is in charge of analysing the data from PMTs extracting meaningful data from noise, and the so-called Data Manager, which controls the apparatus. The FCMServer communicates with these two systems using standard 10GbE network links.

Preliminary results show that the interoperability between different vendors FPGA devices — Altera for the on-shore device and Xilinx for the off-shore one — can be achieved and the timing resolution complies with the requirements of the physics experiments.

We developed a test setup to explore the fixed latency capabilities of a complete links chain leveraging on the fixed latency native mode of the Altera transceivers and on the hardware fixed latency implementation for a Xilinx device [10]. The testbed is composed by the NaNet[3] board and the FCM Xilinx-based board respectively emulating the on-shore and off-shore boards connected by optical fibers (see figure 6). The external GPS-equivalent clock has been input to the NaNet[3] to clock the transmitting side of the device. A sequence of dummy parallel data are serialised, 8b/10b-encoded and transmitted, together with the embedded serial clock, at a data rate of 800 Mbps along the fiber towards the receiver side of the FCM system. The FCM system recovers the received clock and transmits the received data and recovered clock back to the NaNet[3] board. Lastly, the receive side of NaNet[3] deserializes data and produces the received clock.

Testing the fixed latency features of the SerDes hardware implementation is straightforward when taking into account that every time a new initialisation sequence is done, e.g. for a hardware reset or at powerup of the SerDes hardware, we should be able to measure the same phase shift between transmitted and received clock, equal to the fixed number of serial clock cycles shift used to correctly align the deserialised data stream. Figure 7 is a picture taken from scope acquisition in

2015 JINST 10 C04011

infinite persistence mode displaying sampled data points for 12 h period of time with a new *reset and align* procedure issued every 10 s. The NaNet^3 transmitter parallel clock (the purple signal) maintains exactly the same phase difference with the receiver parallel clock (the yellow signal) and with the FCM recovered clock (the green signal).

5 Conclusions and future work

Our NaNet design proved to be efficient in performing real-time data communication between the NA62 RICH readout system and GPU-based L0 trigger processor over a single GbE link. We demonstrated that the NaNet^3 design customization is a viable solution for the data transport system of the KM3NeT-IT experiment, implementing the fundamental requirement of a deterministic latency link. With its four 10 GbE ports, the currently under development NaNet-10 board will broadly exceed the bandwidth requirements for the NA62 RICH and will enable the integration of other detectors in the GPU-based L0 trigger.

Acknowledgments

This work was partially supported by the EU Framework Programme 7 EURETILE project, grant number 247846; R. Ammendola and M. Martinelli were supported by MIUR (Italy) through the INFN SUMA project. G. Lamanna, I. Neri, L. Pontisso and M. Sozzi thank the GAP project, partially supported by MIUR under grant RBFR12JF2Z "Futuro in ricerca 2012".

References

[1] R. Ammendola et al., *APEnet+: a 3D Torus network optimized for GPU-based HPC systems*, *J. Phys. Conf. Ser.* **396** (2012) 042059.

[2] G. Lamanna, *The NA62 experiment at CERN*, *J. Phys. Conf. Ser.* **335** (2011) 012071.

[3] KM3NeT collaboration, A. Margiotta, *Status of the KM3NeT project*, 2014 *JINST* **9** C04020 [arXiv:1408.1132].

[4] R. Ammendola et al., *APEnet+ 34 GBPS data transmission system and custom transmission logic*, 2013 *JINST* **8** C12022.

[5] A. Aloisio, F. Ameli, A. D'Amico, R. Giordano, V. Izzo and F. Simeone, *The NEMO experiment data acquisition and timing distribution systems*, in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE*, Valencia, Spain, October 2011, pp. 147–152.

[6] B. Angelucci et al., *The FPGA based Trigger and Data Acquisition system for the CERN NA62 experiment*, 2014 *JINST* **9** C01055.

[7] R. Ammendola et al., *NaNet: a flexible and configurable low-latency NIC for real-time trigger systems based on GPUs*, 2014 *JINST* **9** C02023 [arXiv:1311.4007].

[8] R. Ammendola et al., *NaNet: a low-latency NIC enabling GPU-based, real-time low level trigger systems*, *J. Phys. Conf. Ser.* **513** (2014) 012018 [arXiv:1311.1010].

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

[10] R. Giordano and A. Aloisio, *Fixed latency multi-gigabit serial links with Xilinx FPGA*, *IEEE Trans. Nucl. Sci.* **58** (2011) 194.

2015 JINST 10 C04011