FPGA-based Network Interface Cards Implementing Real-time Data Transport for HEP Experiments

R. Ammendola[†], A. Biagioni[*], P. Cretaro[*], O. Frezza[*], G. Lamanna[§], F. Lo Cicero[*], A. Lonardo[*], M. Martinelli[*], P. S. Paolucci[*], E. Pastorelli[*], L. Pontisso[‡], D. Rossetti[¶], F. Simula[*], M. Sozzi[‡] and P. Vicini[*]

***Abstract*—NaNet is a modular design of a family of FPGA-based PCIe Network Interface Cards implementing a low-latency, real-time data transport between its network channels and the host CPU and GPU accelerators memories.**

The design features a network stack protocol offloading module that, operating in conjunction with a high performance PCIe Gen2/3 X8 core, yields a low and predictable communication latency, making NaNet suitable for real-time applications. A reconfigurable processing module is also available to implement application-specific processing on inbound/outbound data streams with highly reproducible latency.

As of now, NaNet design has been specialized in the NaNet-1 (single 1GbE port) and NaNet-10 (four 10GbE ports) configurations, which are employed in the GPU-based real-time trigger of the CERN NA62 experiment, and in the NaNet^3 (four 2.5 Gbit optical channels) configuration adopted in the data acquisition system of the KM3NeT-Italia underwater neutrino telescope. Assessment of the real-time characteristics and performances of the resulting systems will be provided and analyzed.

I. NaNet design overview

NANET is a modular design of a low-latency PCIe RDMA NIC supporting different network link technologies: standard GbE (1000BASE-T) and 10-GbE (10GBase-KR), besides custom 34 Gbps APElink [1] and 2.5 Gbps deterministic latency optical KM3link [2]. The design includes a network stack protocol offload engine yielding a very stable communication latency, an enabling feature for use in real-time contexts; NaNet GPUDirect RDMA capability, inherited from the APEnet+ 3D torus NIC dedicated to HPC systems [3], extends its range of application into the world of GPGPU heterogeneous computing.

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

The I/O Interface module performs a 4-stages processing on the data stream: according to the OSI Model, the Physical Link Coding stage implements, as the name suggests, the channel physical layer (*e.g.* 10GBASE-R) while the Protocol Manager stage handles data/network/transport layers (*e.g.* UDP); the Data Processing stage implements application dependent manipulations on data streams (*e.g.* performing compression/decompression) 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 the output channel transport protocol (*e.g.* UDP).

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 Number and bit-width of the switch ports and the routing algorithm can each be defined by the user to automatically achieve a desired configuration. The *Network Interface* block acts on the trasmitting side by gathering data coming in from the PCIe port and forwarding them to the Router destination ports, while on the receiving side it provides support for RDMA in communications with the CPU or the GPU (via the dedicated *GPU I/O Accelerator* module). A Nios II microcontroller is included to support configuration and runtime operations. Finally, the PCIe Core module sports a simplified but efficient backend interface and multiple DMA engines and provides PCIe endpoint functionality.

As it will be described in the following sections, this general architecture has been specialized in the NaNet-1 (single 1GbE port) and NaNet-10 (four 10GbE ports) configurations employed in the GPU-based real-time trigger of the CERN NA62 experiment, and in the NaNet^3 (four 2.5 Gbit optical channels) configuration adopted in the data acquisition system of the KM3NeT-Italia underwater neutrino telescope.

II. GPU real-time processing in the NA62 trigger

The NA62 experiment at CERN [4] 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 high-rate data stream due to a $\sim$ 10 MHz rate of particle decays illuminating the detectors, a set of trigger levels will have to reduce this rate by three orders of magnitude. The entire trigger chain works on the main digitized data stream [5].

The low-level trigger (L0), implemented in hardware by means of FPGAs on the readout boards, reduces the data stream by a factor 10 to meet the maximum design rate for event readout of 1 MHz. The upper trigger levels (L1 and L2) are software-implemented on a commodity PC farm for further reconstruction and event building.

[*]INFN Sezione di Roma, Italy.
[†]INFN Sezione di Tor Vergata, Italy.
[‡]INFN Sezione di Pisa, Italy.
[§]INFN Laboratori Nazionali di Frascati, Italy.
[¶]NVIDIA Corporation, U.S.A.
M. Martinelli is the corresponding author (michele.martinelli@roma1.infn.it.)

Manuscript received May 30, 2016.

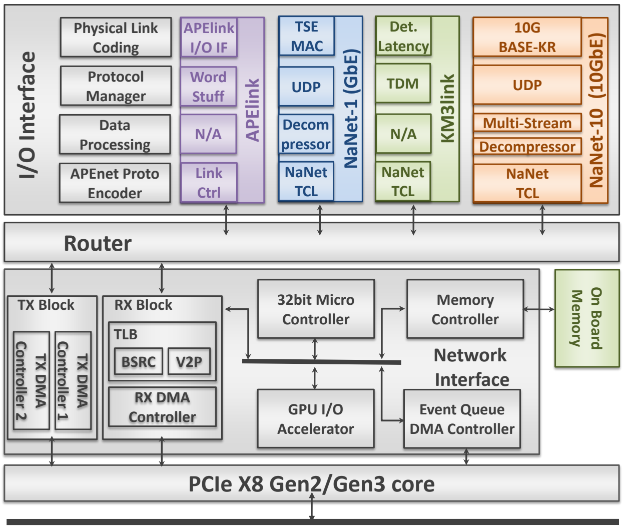


Fig. 1. NaNet architecture schematic.

In the standard 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 time-stamped and sent to a central processor for matching and trigger decision. The maximum latency allowed for the synchronous L0 trigger is thus related to the maximum data storage time available on the data acquisition boards and corresponds to a 1 ms time budget. In this case study we use a GPU-based system (GPU_L0TP) to generate in real-time refined trigger primitives for the RICH detector, (*i.e.* centres and radii of Čerenkov ring patterns on the photomultipliers (PMs) arrays), as this knowledge is strictly related to physics parameters and allows building stringent conditions for data selection at trigger level. Data from the PMs are gathered by four readout boards (TEL62), each sending primitives to the GPU_L0TP (see fig. 2) by UDP streams through a 1 GbE dedicated link.

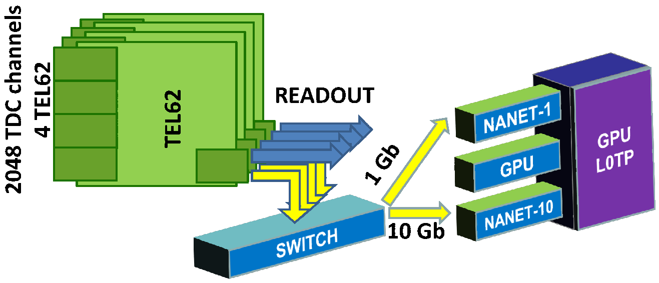


Fig. 2. RICH read-out boards (TEL62) send primitives to the GPU-based processing system (GPU_L0TP) through the NaNet NIC.

In the 2015 experimental setup, the GPU_L0TP at CERN included 2 TEL62 boards connected to a HP2920 switch and a NaNet-1 [6] board with a TTC HSMC daughtercard plugged into a server made of a X9DRG-QF dual socket motherboard populated with Intel Xeon E5-2620 @2.00 GHz CPUs (*i.e.* Ivy Bridge architecture), 32 GB of DDR3 RAM and a Kepler-class NVIDIA K20c GPU. For operation at nominal 10 Mhz event rate with the complete set of mboxread-out links, we have implemented NaNet-10 [7], a 10 GbE version of the board currently under integration in the experimental setup.

This test setup worked at $\sim 1/3$ of the nominal event rate with half of the read-out channels, collecting primitives with the single 1 GbE channel of the NaNet-1 card and allowing us to test the whole chain with the data events moving towards the GPU-based trigger through NaNet-1 by means of the GPUDirect RDMA interface. Data received within a configurable time window are gathered in a Circular List Of Persistent buffers (CLOP) in GPU memory. This time window must always be shorter or equal to how long multi-ring reconstruction takes on the GPU, to be sure that buffers are not overwritten before they are consumed, working in zero-copy. Events are timestamped; those sharing a time-window but coming from different boards are fused into one event describing the PMs status in the RICH detector in GPU memory by a software kernel (Merger). The multi-ring parameters are reconstructed by a GPU kernel (Fitter) consuming coalesced event data [8].

Results are reported in Fig. 3, with the CLOP size measured as number of received events on the X-axis and the latencies of different stages on the Y-axis. Events were received from 2 readout boards with a beam intensity of 4×10^{11} protons per spill, with a gathering time of 400 μs and a receive buffer (CLOP) size of 8kB.

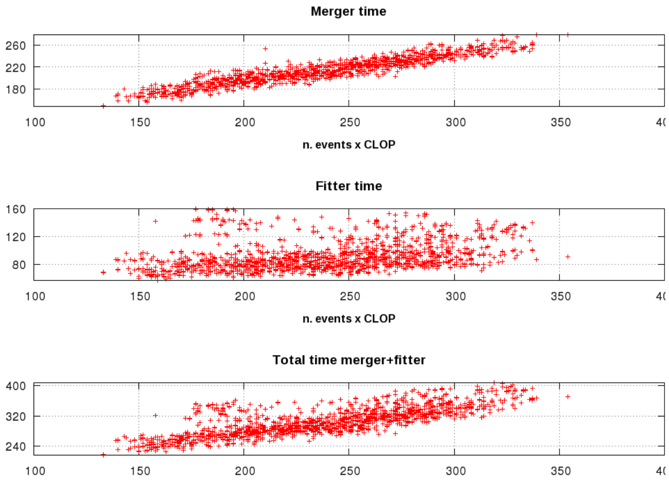


Fig. 3. Multi-ring reconstruction of RICH events performed on a K20c NVIDIA GPU during the 2015 run of the NA62 experiment.

Since the merge operation requires synchronization and serialization and does not exhibit much parallelism, as such it is an ill-suited problem to the GPU architecture. The high latency of this task suggests to offload it to a dedicated implementation in the FPGA Data Processing stage [8].

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

KM3NeT-Italia 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 seawater. 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 final assessment of the experiment foresees the installation of 8 detection units. The KM3NeT-Italia 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, are collected by an electronic board contained in a vessel at the centre of the floor; 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.

A. The KM3NeT-Italia DAQ and data transport architecture

The DAQ architecture is heavily influenced by the need of a common timing distributed all over the system in order to correlate signals from different parts of the apparatus with the required ~ 1 ns 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 need implies that the readout electronics, which is spatially distributed, require common timing and a known delay 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 used to monitor the apparatus status and the environmental conditions, are collected by the Floor Control Module (FCM) board, packed together and transmitted through 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.

The data stream that a single floor delivers to shore has a rate of ~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 requires an on-shore communication endpoint counterpart. The limited data rate per FCM compared with state-of-the-art link technologies led us to consider the design of NaNet^3, an on-shore readout board able to manage multiple FCM data channels.

This is a NaNet customization implementing the data transport requirements of the KM3NeT-Italia experiment. The I/O interface has support for a synchronous link protocol with deterministic latency at physical level and for a Time Division Multiplexing protocol at data level. Moreover, an on-shore readout board supporting multiple channels allows for more efficient scaling of the PC farm size, thus keeping the infrastructure more cost-effective (see Fig. 4).

B. NaNet^3 implementation

The first stage design for NaNet^3 was implemented on an evaluation board from Terasic, the DE5-net board, which is based on Altera Stratix-V GX FPGA, supports up to 4 SFP+ channels and a PCIe x8 edge connector.

The first constraint to be satisfied requires 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 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 link established in this way is fully synchronous.

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.

In this way, the NaNet^3 board plays the role of a bridge between the 4 FCMs and the FCMServer — *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. At a higher level, two systems handle the data that come from and go to off-shore: 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 10-GbE network links.

C. NaNet^3 performances

We verified that the interoperability between different FPGA devices vendors can be achieved and the timing resolution complies with physics requirements.

This needed the development of a test setup to explore the fixed latency capabilities of the complete links chain.

We leveraged on the fixed latency native mode of the Altera transceivers and on the hardware fixed latency implementation for Xilinx devices [9]. The testbed was composed by the NaNet^3 board and the FCM Xilinx-based board, emulating respectively the on-shore and off-shore boards connected by optical fibers.

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 from the received clock and transmits the received data and recovered clock back to the NaNet^3 boards. Lastly, the received side of NaNet^3 deserializes data and produces the received clock.

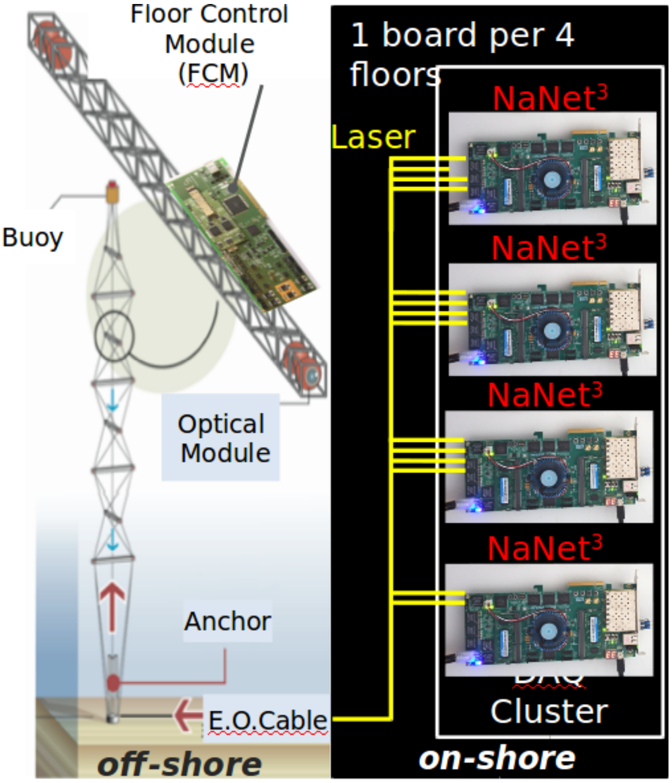
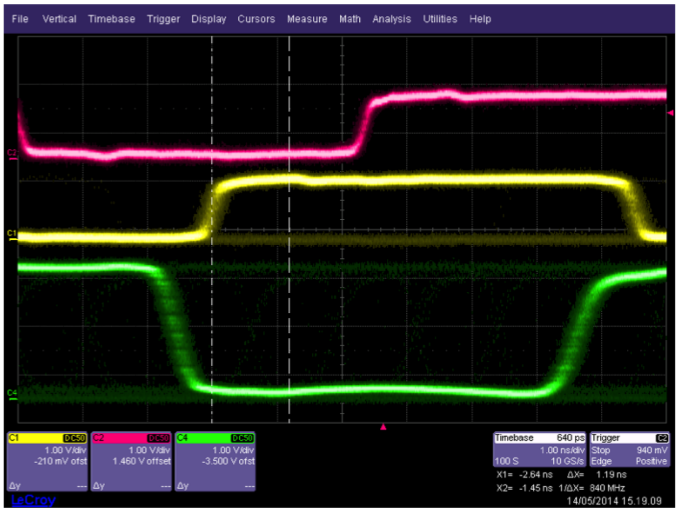


Fig. 4. NaNet^3 in the DAQ of the KM3NeT-Italia experiment.

The way to test the fixed latency features of the SerDes hardware implementation is quite straightforward considering that every time a new initialisation sequence, following an hardware reset or a powerup of the SerDes hardware, has been issued, we are 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. Fig. 5 is a picture taken from scope acquisition in Infinity Persistence mode showing the results of a 12 h test where every 10 s a new *reset and align* procedure has been issued. 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).


File Vertical Timebase Trigger Display Cursors Measure Math Analysis Utilities Help
C2
C1
C4
LeCroy

C1 DC50 1.00 V/div -210 mV ofst Δy ---
C2 DC50 1.00 V/div 1.460 V offset Δy ---
C4 DC50 1.00 V/div -3.500 V ofst Δy ---
Timebase 640 ps 1.00 ns/div 100 S 10 GS/s
Trigger C2 Stop 940 mV Edge Positive
X1= -2.64 ns ΔX= 1.19 ns
X2= -1.45 ns 1/ΔX= 840 MHz
14/05/2014 15.19.09

Fig. 5. 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.

Further tests have been conducted, verifying zero frame loss in a week of non-stop operation. The deterministic latency capability has been verified on the on/off-shore ~100 km long optical link, during the setup and configuration phases of the underwater tower base devices.

IV. Conclusions and Future Work

Our NaNet-1 design proved to be efficient in performing real-time data communication between the NA62 RICH readout system and the GPU-based L0 trigger processor during the 2015 run of the experiment. We are currently integrating the NaNet-10 10 GbE card in the experiment in order to collect data at the nominal 10 MHz event rate over the full set of read-out channels. We successfully finalized and tested the NaNet^3 implementation for the KM3NeT-Italia experiment readout system, demonstrating the feasibility of deterministic latency channels over very long distances (~100 km). We plan to develop further the NaNet architecture, incorporating higher bandwidth I/O channels (e.g. 40 GbE) and host interfaces.

References

[1] R. Ammendola, A. Biagioni, O. Frezza, A. Lonardo, F. L. Cicero, P. S. Paolucci, D. Rossetti, F. Simula, L. Tosoratto, and P. Vicini, *APEnet+ 34*

Gbps data transmission system and custom transmission logic, *Journal of Instrumentation* **8** (2013), no. 12 C12022.

[2] 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*, Oct 2011, pp. 147–152.

[3] R. Ammendola, A. Biagioni, O. Frezza, F. L. Cicero, A. Lonardo, P. S. Paolucci, D. Rossetti, F. Simula, L. Tosoratto, and P. Vicini, "APEnet+: a 3D Torus network optimized for GPU-based HPC systems," *Journal of Physics: Conference Series*, vol. 396, no. 4, p. 042059, 2012. [Online]. Available: http://stacks.iop.org/1742-6596/396/i=4/a=042059

[4] G. Lamanna, "The NA62 experiment at CERN," *Journal of Physics: Conference Series*, vol. 335, no. 1, p. 012071, 2011. [Online]. Available: http://stacks.iop.org/1742-6596/335/i=1/a=012071

[5] C. Avanzini *et al.*, "The trigger and DAQ system for the NA62 experiment," *Nucl. Instrum. Methods Phys. Res., A*, vol. 623, pp. 543–545, 2010.

[6] R. Ammendola, A. Biagioni, O. Frezza, G. Lamanna, A. Lonardo, F. L. Cicero, P. S. Paolucci, F. Pantaleo, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "Nanet: a flexible and configurable low-latency nic for real-time trigger systems based on gpus," *Journal of Instrumentation*, vol. 9, no. 02, p. C02023, 2014. [Online]. Available: http://stacks.iop.org/1748-0221/9/i=02/a=C02023

[7] R. Ammendola, A. Biagioni, M. Fiorini, O. Frezza, A. Lonardo, G. Lamanna, F. Lo Cicero, M. Martinelli, I. Neri, P.S. Paolucci, E. Pastorelli, L. Pontisso, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "NaNet-10: a 10GbE network interface card for the GPU-based low-level trigger of the NA62 RICH detector," *Journal of Instrumentation*, vol. 11, no. 03, p. C03030, 2016. [Online]. Available: url=http://stacks.iop.org/1748-0221/11/i=03/a=C03030

[8] R. Ammendola, A. Biagioni, P. Cretaro, S. Di Lorenzo, R. Fantechi, M. Fiorini, O. Frezza, G. Lamanna, F. Lo Cicero, A. Lonardo, M. Martinelli, I. Neri, P. S. Paolucci, E. Pastorelli, R. Piandani, L. Pontisso, D. Rossetti, F. Simula, M. Sozzi and P. Vicini, "GPU-based Real-time Triggering in the NA62 Experiment," *arXiv* 2016. [Online]. Available: url=https://arxiv.org/abs/x.y

[9] R. Giordano and A. Aloisio, "Fixed latency multi-gigabit serial links with Xilinx FPGA," *IEEE Transaction On Nuclear Science*, vol. 58, no. 1, pp. 194–201, 2011.