A FPGA-based Network Interface Card with GPUDirect enabling real-time GPU computing in HEP experiments.

Alessandro Lonardo[a]*, *Fabrizio Ameli*[a], *Roberto Ammendola*[b], *Andrea Biagioni*[a], *Angelo Cotta Ramusino*[c], *Massimiliano Fiorini*[c], *Ottorino Frezza*[a], *Gianluca Lamanna*[de], *Francesca Lo Cicero*[a], *Michele Martinelli*[a], *Ilaria Neri*[c], *Pier Stanislao Paolucci*[a], *Elena Pastorelli*[a], *Luca Pontisso*[f], *Davide Rossetti*[g], *Francesco Simeone*[a], *Francesco Simula*[a], *Marco Sozzi*[fe], *Laura Tosoratto*[a], *Piero 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

DOI: http://dx.doi.org/10.3204/DESY-PROC-2014-05/16

The capability of processing high bandwidth data streams in real-time is a computational requirement common to many High Energy Physics experiments. Keeping the latency of the data transport tasks under control is essential in order to meet this requirement. We present NaNet, a FPGA-based PCIe Network Interface Card design featuring Remote Direct Memory Access towards CPU and GPU memories plus a transport protocol offload module characterized by cycle-accurate upper-bound handling. The combination of these two features allows to relieve almost entirely the OS and the application from data tranfer management, minimizing the unavoidable jitter effects associated to OS process scheduling. The design currently supports one GbE (1000Base-T) and three custom 34 Gbps APElink I/O channels, but four-channels 10GbE (10Base-R) and 2.5 Gbps deterministic latency KM3link versions are being implemented. Two use cases of NaNet will be discussed: the GPU-based low level trigger for the RICH detector in the NA62 experiment and the on-/off-shore data acquisition for the KM3Net-IT underwater neutrino telescope.

1 Introduction

In many fields of Experimental Physics ranging from Radio Astronomy to High Energy Physics, the GPU adoption effort that in High Performance Computing brings such outstanding results

*Corresponding author. E-mail: alessandro.lonardo@roma1.infn.it

is hampered by the real-time constraints that the processing of data streams outflowing from experimental apparatuses is often subject to. GPUs processing latency is mostly stable once data has landed in their own internal memories; a data transport mechanism with deterministic or at least bound latency is then crucial in building a GPGPU system honouring these constraints.

Our NaNet design is one such mechanism, reusing several IPs developed for the APEnet+ 3D torus NIC [1] targeted at HPC hybrid clusters; its real-time features were achieved by adding dedicated modules and support of multiple link technologies, both standard and custom. The resulting FPGA-based, low-latency PCIe NIC is highly configurable and modular, with RDMA and GPUDirect capabilities. It has been employed in widely varying 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 low-latency PCIe NIC supporting standard — GbE (1000BASE-T) and 10GbE (10Base-R) — and custom — 34 Gbps APElink [4] and 2.5 Gbps deterministic latency optical KM3link [5] — network links. NaNet bridges the gap between real-time and GPGPU heterogeneous computing by inheriting the GPUDirect P2P/RDMA capability from its HPC-dedicated sibling — the APEnet+ 3D torus NIC — and enhancing it with real-time-enabling features, *e.g.* a network stack protocol offload engine for stable communication latency. 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: following the OSI Model, the Physical Link Coding stage implements 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 implements a parametric full crossbar switch responsible for data routing, sustaining multiple data flows @2.8 GB/s.

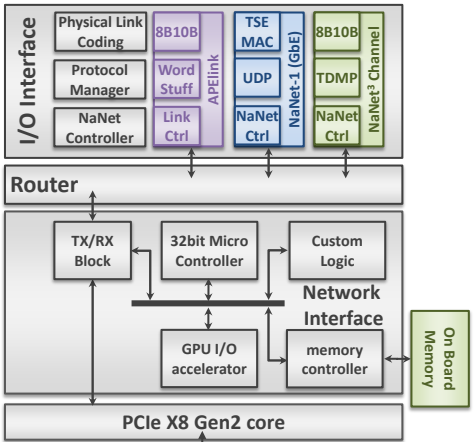

Figure 1: NaNet architecture schematic.

The *Network Interface* block acts on the trasmitting side by forwarding PCIe-originated data to the Router ports and on the receiving side by providing 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.

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** is implemented on the Altera Stratix IV FPGA Dev Kit; it sports a PCIe Gen2 x8 host interface, a GbE channel and three optional 34 Gbps APElink ones; **NaNet^3** is implemented on the Terasic DE5-net Stratix V FPGA Dev Kit; it supports a PCIe Gen2 x8 host interface while the four SFP+ cages of the board are used for the KM3link channels which are 2.5 Gbps optical links with deterministic latency;**NaNet-10** is another implementation on the Terasic DE5-net board; the PCIe Gen2 x8 host interface is the same as NaNet^3 while the SFP+ ports support four 10GbE channels.

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

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. 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 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 better I/O channels. NaNet-1 real-time characterization 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 performs fast virtual-to-physical address translations.

A bandwidth-downscaled GL0TP setup was realized in a loopback configuration by a host system simulating TEL62 UDP traffic through one of its GbE ports towards a NaNet-1 NIC streaming the incoming data into a GPU memory circular list of receive buffers; once landed, buffers are consumed by a CUDA Kernel implementing the pattern matching algorithm. Communication and kernel processing tasks were serialized in order to perform the measure; in Fig. 2 (left) the results for the K20Xm system are shown. During normal operation, this serialization constraint can be relaxed, and kernel processing task overlaps with data communication. Actually this is what was done to measure system throughput, with the results in Fig. 2 (right).

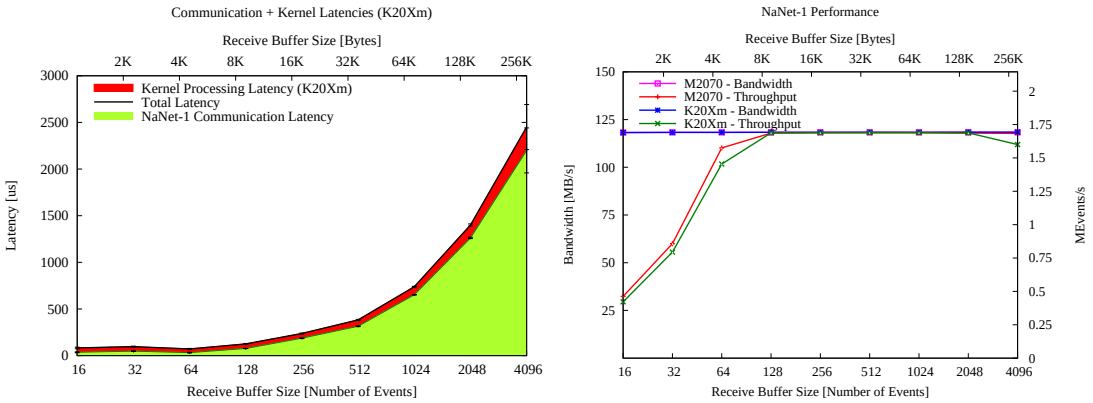

Figure 2: (left) NaNet-1 latency over GbE towards GPU memory for varying datagram payload sizes (nVIDIA Fermi K20Xm); (right) NaNet-1 GbE bandwidth towards GPU memory at varying datagram payload sizes.

We see that using GPU receive buffer sizes ranging from 128 to 1024 events allows the system to stay within the 1 ms time budget while keeping a ~ 1.7 MEvents/s throughput.

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 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 was implemented in the NaNet Controller module that triggers the DMA of received data towards CPU/GPU memory on a configurable deadline rather than on the filling of a buffer. The prototype to be deployed at the NA62 experiment site will integrate into NaNet-1 a TTC HSMC daughtercard and a nVIDIA Kepler K20 GPU.

4 NaNet^3: KM3NeT-IT neutrino telescope on-shore board

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 KM3NeT-IT detection unit consists of 14 floors vertically spaced 20 meters apart, with ∼8 m long floor arms bearing 6 glass spheres each called Optical Modules (OM); each OM contains one 10 inches photomultipliers (PMT) and front-end electronics that digitizes, formats and emits the PMT signal. All data produced by OMs and auxiliary floor instrumentation is collected by an off-shore electronic board called *Floor Control Module* (FCM) contained in a vessel at the floor centre; the FCM manages the communication between the on-shore laboratory and the underwater devices, also distributing the timing information and signals. Due to the distance between apparatus and shoreland the connecting medium is optical fiber.

The spatially distributed DAQ architecture requires a common clock distributed all over the system to correlate signals, with respect to a fixed reference, coming from different parts of the apparatus. For this purpose data acquisition and transport electronics labels each signal with a "time stamp", *i.e.* hit arrival time. Time resolution is also fundamental for the track reconstruction accuracy. These constraints pushed the choice of a synchronous link protocol

embedding clock and data with deterministic latency; Floors are independent from each other; each is connected to the on-shore laboratory by a bidirectional virtual point-to-point link.
A single floor data stream delivered to shore has a rate of ∼300 Mbps, while shore-to-underwater communication data rate is much lower, consisting only of slow-control data for the apparatus. The small data rate per FCM compared with state-of-the-art technologies led us to designing NaNet^3, an on-shore readout board able to manage multiple FCM data channels.
This is a NaNet customization for the KM3NeT-IT experiment, with added support for a synchronous link protocol with deterministic latency at physical level and for a Time Division Multiplexing protocol at data level (see Fig. 1).

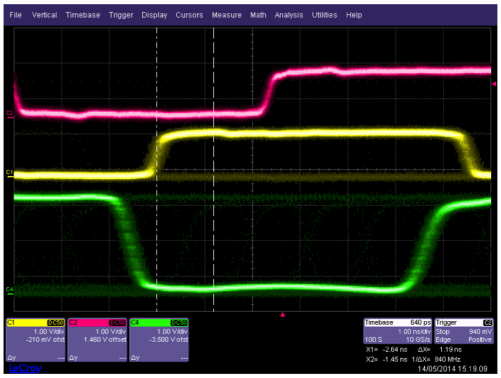

Figure 3: Deterministic latency of NaNet^3 SerDes: phase alignment of the transmitting (upper) and receiving (central) clocks.

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 with four SFP+ channels and a PCIe x8 edge connector. To match time resolution constraint, time delta between wavefronts of three clocks must have *ns* precision: the first clock is an on-shore reference one (coming from a GPS and redistributed), 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 decoding the payload at the end of the loop. The deterministic latency constraint must be enforced on the Stratix device as the FCM does on paths both ways to allow correct time stamping of events on the PMT. The established link is synchronous, *i.e.* clock rate is equal for both devices with fixed phase displacement. We developed a test setup to explore latency capabilities of a complete link chain leveraging on the fixed latency native mode of the Altera transceivers, on the NaNet^3 board, and on the hardware fixed latency implementation for a Xilinx device on the FCM board [9]. 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 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. 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 on every initialisation sequence, *e.g.* for a hardware reset or at SerDes hardware powerup, 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. Fig. 3 shows a scope acquisition in infinite persistence mode sampled over 12 h issuing every 10 s a new *reset and align*. The NaNet^3 transmitter parallel clock (the upper line) maintains exactly the same phase difference with the receiver parallel clock (the central line) and with the FCM recovered clock (the lower line).

5 Conclusions and future work

Our NaNet design proved to be an efficient real-time communication channel between the NA62 RICH readout system and GPU-based L0 trigger processor over a single GbE link. With four 10 GbE ports, the currently under development NaNet-10 board will exceed the bandwidth requirements for the NA62 RICH, enabling integration of other detectors in the GPU-based L0 trigger. Along the same lines, the deterministic latency link of the NaNet[3] customization makes it a viable solution for the data transport system of the KM3NeT-IT experiment while the GPUDirect RDMA features imported from NaNet will allow us later on to build a real-time GPU-based platform, to investigate improved trigger and data reconstruction algorithms.

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. *Journal of Physics: Conference Series*, 396(4):042059, 2012.

[2] Gianluca Lamanna. The NA62 experiment at CERN. *Journal of Physics: Conference Series*, 335(1):012071, 2011.

[3] A Margiotta. Status of the KM3NeT project. *Journal of Instrumentation*, 9(04):C04020, 2014.

[4] R Ammendola et al. APEnet+ 34 Gbps Data Transmission System and Custom Transmission Logic. *Journal of Instrumentation*, 8(12):C12022, 2013.

[5] A. Aloisio et al. The NEMO experiment data acquisition and timing distribution systems. In *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE*, pages 147–152, Oct 2011.

[6] B Angelucci et al. The FPGA based trigger and data acquisition system for the CERN NA62 experiment. *Journal of Instrumentation*, 9(01):C01055, 2014.

[7] R Ammendola et al. NaNet: a flexible and configurable low-latency NIC for real-time trigger systems based on GPUs. *Journal of Instrumentation*, 9(02):C02023, 2014.

[8] R. Ammendola et al. NaNet: a low–latency NIC enabling GPU–based, real–time low level trigger systems. *Journal of Physics: Conference Series*, 513(1):012018, 2014.

[9] R. Giordano and A. Aloisio. Fixed latency multi-gigabit serial links with Xilinx FPGA. *IEEE Transaction On Nuclear Science*, 58(1):194–201, 2011.