NaNet-10: a 10GbE network interface card for the GPU-based low-level trigger of the NA62 RICH detector.

2016 JINST 11 C03030

(http://iopscience.iop.org/1748-0221/11/03/C03030)

Published by IOP Publishing for Sissa Medialab

Received: *November 15, 2015*
Accepted: *January 25, 2016*
Published: *March 15, 2016*

Topical Workshop on Electronics for Particle Physics 2015,
September 28th – October 2nd, 2015
Lisbon, Portugal

NaNet-10: a 10GbE network interface card for the GPU-based low-level trigger of the NA62 RICH detector.

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

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

E-mail: andrea.biagioni@roma1.infn.it

Abstract: A GPU-based low level (L0) trigger is currently integrated in the experimental setup of the RICH detector of the NA62 experiment to assess the feasibility of building more refined physics-related trigger primitives and thus improve the trigger discriminating power. To ensure the real-time operation of the system, a dedicated data transport mechanism has been implemented: an FPGA-based Network Interface Card (NaNet-10) receives data from detectors and forwards them with low, predictable latency to the memory of the GPU performing the trigger algorithms. Results of the ring-shaped hit patterns reconstruction will be reported and discussed.

Keywords: Data processing methods; Trigger concepts and systems (hardware and software); Online farms and online filtering

[1] Corresponding author.

doi:10.1088/1748-0221/11/03/C03030

2016 JINST 11 C03030

Contents

2016 JINST 11 C03030

1 Introduction

The NA62 particle physics experiment at CERN SPS aims at measuring the ultra rare kaon decay $K^+ \to \pi^+ \nu\bar{\nu}$ as a highly sensitive test of the Standard Model and in search for hints of New Physics. A multi-level trigger is designed to manage the high rate required by the experiment. The lowest level (L0) trigger represents an essential element because it is able to handle an input event rate of the order of 10 MHz and apply a rejection factor of 10, with a maximum latency of 1 ms. In the standard implementation of the L0 trigger, data contributing to the final trigger decision are processed on FPGA devices and are mostly based on event multiplicity and topology.

Work is underway in testing a real-time software-based approach for this decision system, one that exploits the parallel computing power of a commercial GPU (Graphics Processing Unit) within the L0 trigger for the NA62 experiment [1]. The use of a GPU in this level would allow for building of more refined physics-related trigger primitives, such as energy or direction of the final state particles in the detectors, therefore leading to a net improvement of trigger conditions and data handling. GPU architectures have been designed to optimize computing throughput with no particular attention to their usage in real-time contexts, such as the one we are considering here. While execution times are rather stable on these architectures, also data transfer tasks are to be taken into account: assessment of the real-time features of the whole system needs a careful characterization of all subsystems along data stream path, from detectors to GPU memories.

We identified the standard network subsystem as the main source of fluctuations for the total system latency. To address this problem, we designed and implemented two generations of FPGA-based Network Interface Cards (NICs), NaNet-1 [2] and NaNet-10 [3], supporting respectively GbE and 10GbE I/O channels. To achieve a low and stable communication latency, NaNet design combines support for GPUDirect [4], i.e. the direct data transport between the GPU memory

and the external I/O channels, with a network protocol offloading module implemented in the FPGA logic. A GPU-based L0 trigger using NaNet is currently integrated in the experimental setup of the RICH Čerenkov detector of the NA62 experiment in order to reconstruct the ring-shaped hit patterns; results obtained with this system will be reported and discussed. A strictly related NIC design implementing four deterministic latency links named NaNet^3 is currently being deployed in the data transport system for the KM3NeT-IT underwater neutrino telescope. NaNet^3 will not be discussed it in this paper, a description of this design can be found in [5].

2 Related works

Data acquisition and high-throughput network mechanisms interfacing the detectors readout are currently under development in several CERN experiments in order to face the increase of luminosity planned for the next years. In [6] the FELIX (Front End LInk eXchange) is presented, a PC-based device to route data from and to multiple GBT links via a high-performance general purpose network capable of a total throughput up to O(20 Tbps). The new data acquisition system under development for the next upgrade of the LHCb experiment at CERN is presented in [7] focusing on the PCIe board PCIe40 aiming to achieve a data throughput of 100 Gbps. The firmware design and implementation of Common Readout Unit (CRU) for data concentration, multiplexing and trigger distribution at ALICE experiment is described in [8].

The adoption of GPU to implement trigger algorithm to achieve the computing power to cope with the LHC luminosity increase is under investigation in several CERN experiments [9, 10].

3 A GPU-based low-level trigger of the NA62 RICH detector

We focus on ring reconstruction in the RICH detector to study the feasibility of a GPU-based L0 trigger system for NA62 (GPU_L0TP). The RICH identifies pions and muons with momentum in the range between 15 GeV/c and 35 GeV/c. Čerenkov light is reflected by a composite mirror with a focal length of 17 m focused onto two separated spots equipped with $\sim$ 1000 photomultipliers (PM) each. L0 is a low latency synchronous level and GPU usage must be verified. Furthermore, data communication between the readout boards (TEL62) and the L0 trigger process happens over multiple GbE links using UDP streams. The final system consists of 4 GbE links to move primitives data from the readout boards to the GPU_L0TP (see figure 1). Each link contains the data coming from $\sim$ 500 PMs. The main requirement for the communication is the deterministic response latency of GPU_L0TP. The entire response latency comprising both communication and computation tasks must be lower than 1 ms.

3.1 MultiRing Čerenkov ring reconstruction on GPUs

In order to build stringent conditions for data selection at trigger level, it could be useful to take the parameters of Čerenkov rings into account when making trigger decisions. This implies that circles have to be reconstructed using the coordinates of activated PMs.

We focus on two pattern recognition algorithms based only on geometrical considerations and particularly suitable for exploiting the intrinsic parallel architecture of GPUs.

2016 JINST 11 C03030

Ptolemy's Theorem states that when four vertices of a quadrilateral (ABCD) lie on a common circle, it is possible to relate four sides and two diagonals: $|AC|\times|BD| = |AB|\times|CD|+|BC|\times|AD|$. This formula can be implemented in a paralled way allowing for a fast multi-ring selection. This is crucial either to directly reconstruct the rings or to choose different algorithms according to the number of circles. The large number of possible combinations of four vertices, given a maximum of 64 points for physics event, can be a limitation to this approach. To greatly reduce the number of tests, a possibility is to select few triplets — i.e. a set of three hits — trying to maximize the probability that all their points belong to the same ring and iterating through all the remaining hits to search for the ones satisfying the aforementioned formula [11].

We also make use of an histogramming method in which the XY plane is divided into a grid and an histogram is created with distances from these points and hits of the physics event. Rings are identified looking at distance bins whose contents exceed a threshold value. In order to limit the use of resources, it is possible to proceed in two steps, starting the histogram procedure with a 8×8 grid and calculating now distances from such squares. Afterwards, to refine their positions, the calculation is repeated with a grid 2×2 only for the square selected according to the threshold in the previous step.

Once the number of rings and points belonging to them have been found, it is possible to apply e.g. Crawford's method [12] to obtain centre coordinates and radii with better spatial resolution.

4 NaNet: a PCIe NIC family for HEP

The NaNet project goal is the design and implementation of a family of FPGA-based PCIe Network Interface Cards for High Energy Physics to bridge the front-end electronics and the software trigger computing nodes [5]. The design of a low-latency and high-throughput data transport mechanism for real-time systems is a mandatory requirement in order to accomplish this task.

NaNet features multiple link technologies to increase the scalability of the entire system allowing for lowering the numerosity of PC farm clusters. The key characteristic is the management of custom and standard network protocols in hardware, in order to avoid OS jitter effects and guarantee a deterministic behaviour of communication latency while achieving maximum capability of the adopted channel. Furthermore, it integrates a processing stage which is able to reorganize data coming from detectors on the fly, in order to improve the efficiency of applications running on computing nodes. Ad hoc solutions can be implemented according to the needs of the experiment (data decompression, reformatting, merge of event fragments).

Finally, data transfers to or from application memory are directly managed avoiding bounce buffers. NaNet accomplishes this zero-copy networking by means of an hardware implemented memory copy engine that follows the RDMA paradigm for both CPU and GPU — this latter supporting the GPUDirect V2/RDMA by nVIDIA to minimize the I/O latency in communicating with GPU accelerators. The quirks in the interactions of this engine with the bulky virtual memory management of the GNU/Linux host are smoothed out by adopting a proprietary Translation Look-aside Buffer based on Content Addressable Memory [13].

The block diagram of the NaNet architecture is in figure 2. Exploiting the modularity of the hardware design, the Network Interface and Router modules are shared among all boards of the family, ensuring fast and reliable implementation of different NaNet design configurations. The PCIe interface can be customized — i.e. Gen2 or Gen3 — and the number or switch ports chosen,

2016 JINST 11 C03030

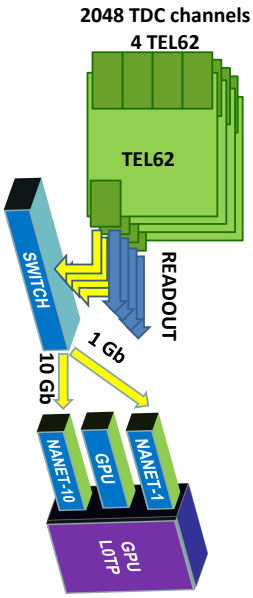

Figure 1. Pictorial view of GPU-based Trigger.

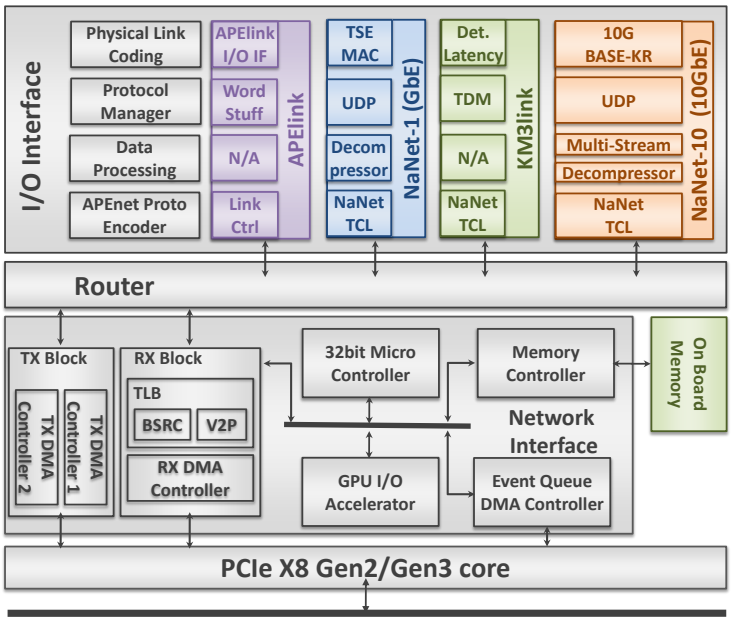

Figure 2. NaNet PCIe Network Interface Card family architecture

to best match experimental requirements. The I/O interface allows for implementation of custom or standard communication protocols, as described in section 4.1.

4.1 NaNet-1 and NaNet-10

One of the key features of the NaNet design is the I/O interface; it consists of 4 elements: Physical Link Coding, Protocol Manager, Data Processing, APEnet Protocol Encoder. The Physical Link Coding covers the Physical and DataLink Layers of the OSI model, managing transmission of data frames over a common local media and performing error detection. The Protocol Manager covers the Network and Transport Layers, managing the reliability of communication data flow control. A processing stage, Data Processing, that applies some function to the data stream in order to ease the work for the applications running on the computing node, can be enabled on the fly. Finally, the APEnet Protocol Encoder performs a protocol translation to a format more suited for PCIe DMA memory transaction. NaNet-1 was developed in order to verify the feasibility of the project; it is a PCIe Gen2 x8 network interface card featuring GPUDirect RDMA over GbE. A Timing Trigger and Control (TTC) HSMC daughtercard captures either trigger and 40 MHz clock streams distributed from the experiment TTC system via optical cable, along with Start/End of Burst information.

The Altera Triple Speed Ethernet Megacore (TSE MAC) provides a 10/100/1000 Mbps Ethernet IP modules with SGMII standard interface to connect the MAC to the PHY. The MAC is a single module in FIFO mode for both receive and transmit sides. The UDP offloader deals with UDP packets payload extraction. The decompressor performs application-dependent modifications to reformat events data in a GPU-friendly fashion on the fly. The NaNet Transmission Control Logic (NaNet TCL) encapsulates the received streams into the APEnet Protocol allowing for reuse of

many IPs of the APEnet+ design, a 3D-Torus NIC developed by the INFN APE lab [14]. Several parameters can configure the NaNet TCL (i.e. packet size, port id, target device) and whatever is needed to fulfill the key task of virtual address generation for the APEnet packets.

NaNet-10 is a PCIe Gen2 x8 network adapter implemented on Terasic DE5-net board equipped with an Altera Stratix V FPGA featuring 10GbE SFP+. The TTC signals are received through 2 onboard SMA connectors. The Physical Link Coding is implemented by the Altera 10GBASE-KR PHY and the 10Gbps MAC: the former delivers serialized data to a module driving optical fiber at a line rate of 10.3125 Gbps, the latter supports operating modes starting from 10 Mbps up to 10 Gbps with an Avalon-Streaming interface of 64-bit wide interface running at 156.25 MHz and MII/GMII/SDR XGMII on the network side. A 10 Gbps UDP/IP Core providing full UDP, IPv4 and ARP protocols is in charge of Protocol Manager task. The module offers an AXI-based 64-bit data interface. UDP header settings — e.g. source/destination port and destination IP address — are exposed in both transmit and receive sides. Zero-latency between the Protocol Manager and the Physical Link Coding is guaranteed avoiding internal buffer but packet segmentation and reassembly could not be supported. A multi-stream hardware module inspects the received stream and separates the packets according to the destination port to redirect data to different GPU memory buffers to satisfy the application requirements. The decompressor and NaNet TCL units are mostly shared with those in NaNet-1.

5 Results: NaNet-1 at NA62

Current setup of a GPU-based trigger at CERN comprises 2 TEL62 readout boards connected to a HP2920 switch and a NaNet-1 board with a TTC HSMC daughtercard plugged into a SuperMicro server consisting of a X9DRG-QF dual socket motherboard — Intel C602 Patsburg chipset — populated with Intel Xeon E5-2620 @2.00 GHz CPUs (i.e. Ivy Bridge micro-architecture), 32 GB of DDR3 memory and a Kepler-class nVIDIA K20c GPU.

Such a system allows for testing of the whole chain: the data events move towards the GPU-based trigger through NaNet-1 by means of the GPUDirect RDMA interface. Data arriving within a given time frame — which is configurable — are gathered and then organized in a Circular List Of Persistent buffers (CLOP) in the GPU memory. Buffer number and size are tunable in order to optimize computing and communication. Clearly, this time frame must necessarily be shorter or equal on average to how long the GPU takes for multi-ring reconstruction, to be sure that buffers are not overwritten by incoming events before they are consumed by the GPU. Events coming from different TEL62 need to be merged in GPU memory before the launch of the ring reconstruction kernel. Each event is timestamped and the ones coming from different readout boards that are in the same time-window are fused in a single event describing the status the PMs of RICH detector. The current GPU implementation of multi-ring reconstruction is based on the histogram algorithm and is executed on an entire CLOP buffer as soon as the NIC signals to the host application that the receiving process is complete for that buffer. We report the time length distribution for different phases of the Multi-ring reconstruction performed on K20c nVIDIA GPU for a gathering time of 450 μs under a beam intensity of 34×10^{11} protons per spill. Data events come from one TEL62 readout board only. The phases the task is split into are data merger and histogram fitter. In case of single event source, the merger process consists of a simple reformat of

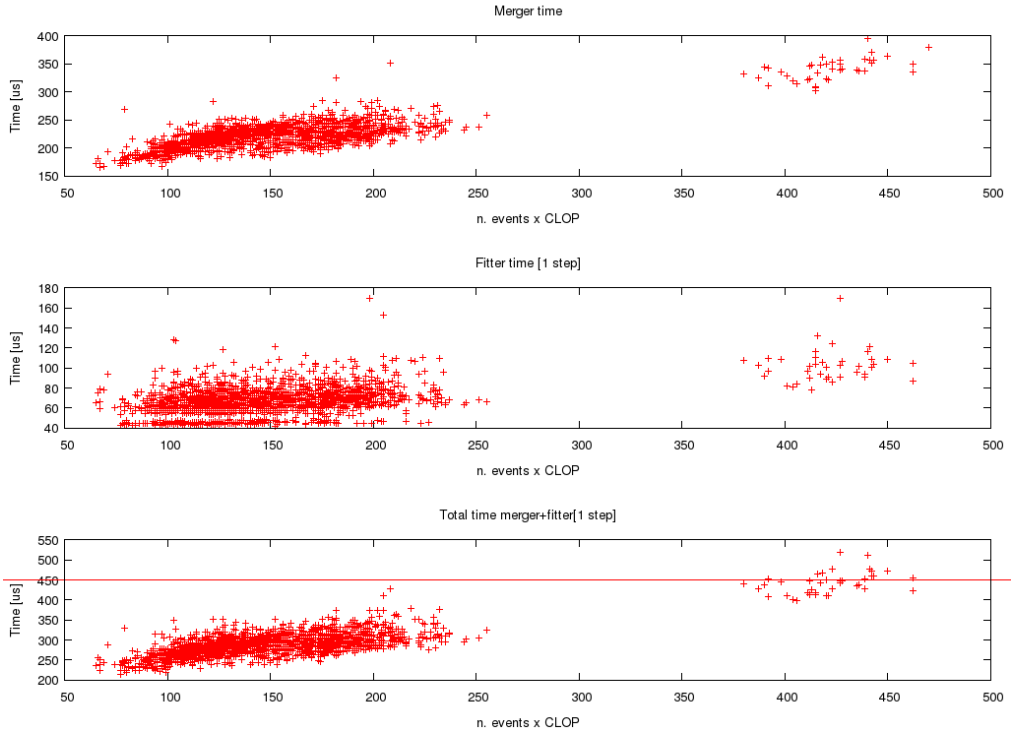

Figure 3. Multi-ring reconstruction of events coming from single source performed on K20c nVIDIA GPU.

the data. The results of the histogram fitter with a single step are reported — i.e. the procedure applies an 8x8 grid only; the runtime of $\sim 100\mu$s is well within the allotted time. The distributions are in figure 3; the CLOP size measured as number of received events is on the X-axis being and the different durations of the phases are on the Y-axis. Globally, reconstruction takes almost always less than the gathering time — most points lie below the horizontal threshold in the total time plot.

The results reported in figure 4 are for events coming from 2 readout boards, for a gathering time of 206 μs under a beam intensity of 2.4×10^{11} protons per spill; the GPU execution time is 4 times the gathering time so that a continuous data stream overwrites the CLOP buffers causing data loss. No major differences are in the fitting time of $\sim 150\mu$s. On the other hand, the reorder of the merge operation is non-trivially parallelizable, i.e. it is an ill-suited problem to both the GPU architecture and mode of execution. We acknowledge the double clustering apparent in the plot; to optimize the communication bandwidth, the current event gathering implemented within the NaNet board tries maximizing the packet size from the FPGA into the GPU memory which causes a latency hit. To prevent it, a different implementation of the gathering mechanism is foreseen.

5.1 NaNet-10 synthetic benchmark

NaNet-10 benchmarking setup was hosted on a X9DRG-HF dual socket motherboard by SuperMicro with an Intel C602 Patsburg chipset equipped with Intel Xeon E5-2630 @2.60 GHz CPUs (i.e. Ivy Bridge micro-architecture), 64 GB of DDR3 memory and a Kepler-class nVIDIA K40m GPU.

In this setup, 2 ports out of the 4 available ones of the Terasic DE5-net board are closed in "loop-back" while a custom hardware module feeds the UDP TX with outgoing packet payload;

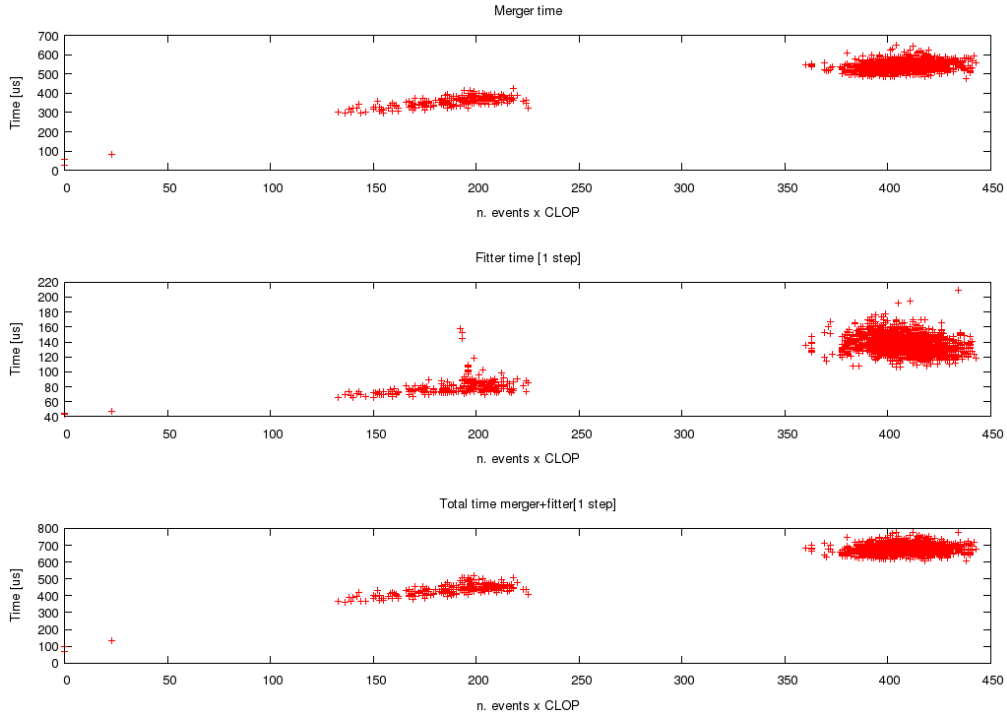

Figure 4. Multi-ring reconstruction of events coming from double source performed on K20c nVIDIA GPU.

NaNet-10 registers are then programmed to set UDP protocol destination port and variable packet size, number of transmitted packets and delay between packets.

Time-of-flight from UDP TX to UDP RX as measured by the SignalTap II Logic Analyzer tool of the Altera Quartus II suite is 64 clock cycles @156.25 MHz (409.6 ns).

At different stages of the packet processing pipeline, cycle counters were added whose values were then patched into the completion payload and stored into the event queue, in order to profile the receiving hardware path traversal latency. The adoption of a data transmission custom hardware module ensures that results are not affected by external latency sources (i.e. DMA memory reading process). The custom module is always ready to send data exploiting the entire link capability mimicking the detector readout system "worst-case".

In figure 5, NaNet-10 and NaNet-1 latencies in the range of interest are compared; NaNet-10 guarantees sub-μs hardware latency for buffers up to ~1kByte in GPU/CPU memory. The data transmission system reaches its 10 Gbps bandwidth peak already for a $\sim$ 1 kByte buffer size (figure 6). NaNet-10 satisfies the read-out system capability requirement of 4 GbE channels from the TEL62, starting from a $\sim$ 256 bytes buffer size. We notice that performance of moving data from the I/O interface of the NIC to target device memory is the same for both the CPU and the GPU.

6 Conclusion

In this paper we described a data transport mechanism bridging the readout system of HEP experiments with the software trigger computing nodes. The results obtained with NaNet-1 at NA62 show that moving data from the RICH detector to a GPU-based low-level trigger characterized by a time

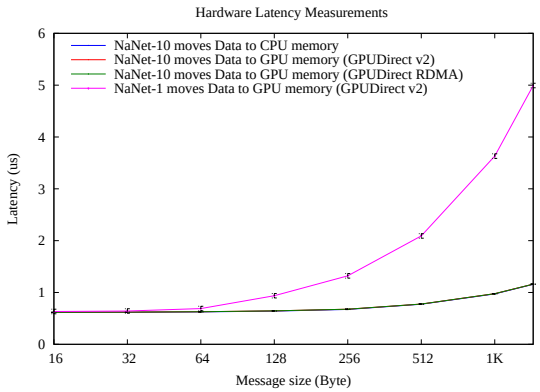

Figure 5. NaNet-10 vs. NaNet-1 hardware latency.

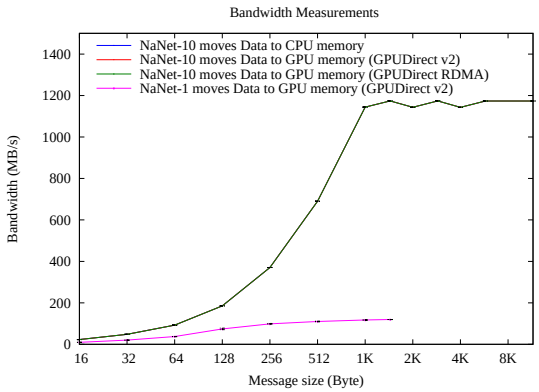

Figure 6. NaNet-10 vs. NaNet-1 bandwidth.

budget of 1 ms does not represent a real issue. Results obtained for a single stream of data event are encouraging. The synthetic benchmarks of NaNet-10 clarify that the Network Interface Card is capable to sustain more than 4 streams coming from the readout of the experiment, fully satisfying the requirements. The unacceptably high latency of the merger task when performed on a GPU strongly suggests to offload such duties to a hardware implementation in the Data Processing stage of the I/O interface of the network adapter.

Acknowledgments

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

[2] 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].

[3] R. Ammendola et al., *A multi-port 10GbE PCIe NIC featuring UDP offload and GPUDirect capabilities.*, *J. Phys. Conf. Ser.* **664** (2015) 092002.

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

[5] A. Lonardo et al., *NaNet: a configurable NIC bridging the gap between HPC and real-time HEP GPU computing*, 2015 *JINST* **10** C04011.

[6] ATLAS collaboration, *FELIX: a High-Throughput Network Approach for Interfacing to Front End Electronics for ATLAS Upgrades*, ATL-DAQ-PROC-2015-014 (2015).

[7] P. Durante, N. Neufeld, R. Schwemmer, G. Balbi and U. Marconi, *100 GBPS PCI-Express readout for the LHCb upgrade*, 2015 *JINST* **10** C04018.

[8] J. Mitra, S.A. Khan, S. Mukherjee and R. Paul, *Common Readout Unit (CRU) — a new Readout Architecture for ALICE Experiment*, 2015 *JINST* **11** C03021.

2016 JINST 11 C03030

[9] V. Halyo, A. Hunt, P. Jindal, P. LeGresley and P. Lujan, *GPU Enhancement of the Trigger to Extend Physics Reach at the LHC*, 2013 *JINST* **8** P10005 [arXiv:1305.4855].

[10] ATLAS collaboration, D. Emeliyanov and J. Howard, *GPU-based tracking algorithms for the ATLAS high-level trigger*, *J. Phys. Conf. Ser.* **396** (2012) 012018.

[11] G. Lamanna, *Almagest, a new trackless ring finding algorithm*, *Nucl. Instrum. Meth.* **A 766** (2014) 241, in *Proceedings of the Eighth International Workshop on Ring Imaging Cherenkov Detectors*, Shonan, Kanagawa, Japan, December 2-6, 2013.

[12] J.F. Crawford, *A NONITERATIVE METHOD FOR FITTING CIRCULAR ARCS TO MEASURED POINTS*, *Nucl. Instrum. Meth.* **211** (1983) 223.

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

[14] R. Ammendola, A. Biagioni, O. Frezza, F. Lo Cicero, A. Lonardo, P.S. Paolucci et al., *APEnet+: A 3D Torus network optimized for GPU-based HPC systems*, *J. Phys. Conf. Ser.* **396** (2012) 042059.

2016 JINST 11 C03030