PROCEEDINGS OF SCIENCE

NaNet: a Reconfigurable PCIe Network Interface Card Architecture for Real-time Distributed Heterogeneous Stream Processing in the NA62 Low Level Trigger

Paolo Cretaro*, **Andrea Biagioni, Ottorino Frezza, Francesca Lo Cicero, Alessandro Lonardo, Michele Martinelli, Pier Stanislao Paolucci, Luca Pontisso, Francesco Simula, Piero Vicini, Cristiano Capone, Fabrizio Capuani, Giulia De Bonis, Elena Pastorelli**
INFN Sezione di Roma - Sapienza
Piazzale Aldo Moro, 2 - 00185 Roma, Italy
E-mail: name.surname@roma1.infn.it

Roberto Ammendola
INFN Sezione di Roma - Tor Vergata
Via della Ricerca Scientifica, 1 - 00133 Roma, Italy
E-mail: roberto.ammendola@roma2.infn.it

Gianluca Lamanna, Marco Sozzi
INFN Sezione di Pisa
Via F. Buonarroti, 2 - 56127 Pisa, Italy
E-mail: name.surname@pi.infn.it

Roberto Piandani
INFN Sezione di Perugia
Via Alessandro Pascoli, 23c - 06123 Perugia, Italy
E-mail: roberto.piandani@pg.infn.it

Dario Soldi
INFN Sezione di Torino
Via Pietro Giuria, 1 - 10125 Torino, Italy
E-mail: dario.soldi@to.infn.it

The NA62 experiment at CERN SPS is aimed at measuring the branching ratio of the the very rare kaon decay $K^+ \to \pi^+ \nu\bar{\nu}$. NaNet is the reconfigurable design of a FPGA-based PCI Express Network Interface Card with processing, RDMA and GPUDirect capabilities, supporting multiple link technologies. NaNet has been employed to implement a real-time distributed processing pipeline in the the low level trigger of the experiment, operating on the data streams produced by the RICH detector with an orchestrated combination of heterogeneous computing devices (CPUs, FPGAs and GPUs). Recent results collected during NA62 runs are presented and discussed.

Topical Workshop on Electronics for Particle Physics (TWEPP2018)
17-21 September 2018
Antwerp, Belgium

*Speaker.

1. Heterogeneous computing in the NA62 experiment

Heterogeneous computing with CPUs, GPUs and FPGAs is a paradigm that is getting widespread in the High Performance Computing field, thanks to the increasing computing power of GPUs and the huge quantity of resources available on state-of-the-art FPGAs. The NaNet project [1] was started to investigate the feasibility of the integration of this paradigm in the low-level trigger of particle physics experiments, which usually consists of purely dedicated hardware due to the strict latency requirements.

The NA62 experiment has the goal of measuring the branching ratio for the ultra-rare decay of the charged kaon into a pion and a neutrino anti-neutrino pair [2]. The rate of particle decays is $\sim$10 MHz and a trigger system consisting of three levels on cascade has to decrease this rate by three order of magnitude. Of these levels, only the low-level trigger (L0) is implemented in hardware, and it is a synchronous real-time FPGA-based system (L0TP) receiving primitives from the TEL62 readout boards [3] with a 1 ms time budget for trigger decision.

The Ring Imaging Čerenkov (RICH) detector is a key element for particle identification within the experiment and one of the subdetectors responsible for the L0 trigger [4]. Its main purpose is to separate pions from muons and to measure the particle arrival time with a resolution better than 100 ps. Thanks to its fast response, it is used in the L0 trigger by providing a multiplicity count, however much more information could be extracted from the detector.

As a first target for a heterogeneous pipeline application in the NA62 trigger system, we studied the possibility of reconstructing Čerenkov rings directly in the RICH (see figure 1). The information of center and radius of the rings is related to particle angle and velocity and can be used to produce refined primitives for the L0TP to increase the purity and the rejection power of the trigger.

2. NaNet architecture

NaNet features a modular design based on a low-latency PCIe RDMA (Remote Direct Memory Access) NIC, supporting different network link technologies [5]. This layout is functional for a straightforward deployment in multiple scenarios: standard GbE (1000BASE-T), 10 GbE (10GBASE-KR) and 40 GbE, plus a custom 34 Gbps APElink and 2.5 Gbps deterministic latency KM3link (see figure 2).

The NaNet-10 variant of the design, deployed in the NA62 experiment, is implemented on the Terasic DE5-net board[1], featuring an Altera Stratix V FPGA with 4 10GbE SFP+ links and a PCIe Gen2 x8 connector. On top of the 10GbE channel a fully compliant UDP/IP offloading hardware block has been realized, which also supports the Address Resolution Protocol (ARP) for MAC address translation.

3. Rings reconstruction on the GPU

Exploiting the GPUDirect RDMA capabilities of NVIDIA Tesla GPUs, NaNet can inject the input data stream, arriving from the detector front-end, directly into the GPU memory. To profit from the available resources, events are first collected in a buffer (called CLOP, being part of a

[1] https://www.terasic.com.tw/cgi-bin/page/archive.pl?Language=English&CategoryNo=164&No=526&PartNo=1

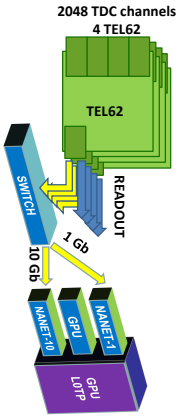

Figure 1: GPU-RICH architecture scheme.

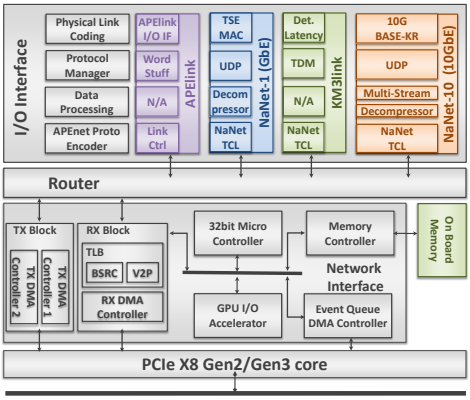

Figure 2: The NaNet PCIe Network Interface Card family architecture.

circular list of persistent buffers) and then a CUDA kernel is run, allowing the parallel processing of several events concurrently and exploiting pattern recognition parallel algorithms.

To cope with the data high throughput and the maximum latency requirement, and not being able to use informations from other detectors, the usage of a fast, pure geometrical, multi-ring fitting algorithm is crucial. The one adopted, named Histogram, divides the XY plane into a grid and computes the histograms of distances from the grid points and the hits measured in the RICH. Grid points having a histogram that has a bin over threshold is a ring center candidate, and the more frequent distance is the candidate radius. The Crawford's method [6] is then applied on the set of hits belonging to an annular region around the candidate ring.

Tests have been performed during NA62 2018 experimental runs using 4 TEL62 readout boards connected, through a HP2920 Ethernet switch, to a NaNet-10 board. NaNet-10 is plugged into a SuperMicro server consisting of a X9DRG-QF dual socket motherboard populated with Intel Xeon E5-2620 CPUs, 32 GB of DDR3 memory and a NVIDIA P100 GPU[2]. Figure 3 shows the total latency of operations during a single burst, with the 99th-percentile being at 260 µs and an average throughput of 0.13 µs per event.

4. Rings reconstruction on the FPGA

The increasing availability of resources inside state-of-the-art FPGAs is encouraging the usage of High-Level Synthesis (HLS) tools, which allow to describe a hardware design with a high-level programming language (e.g. C/C++). For its devices Intel provides two workflows:

- single IP generation from C/C++ code through the *hls* compiler;
- full design generation through the Intel SDK for OpenCL framework.

[2]https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-p100/pdf/nvidia-tesla-p100-PCIe-datasheet.pdf

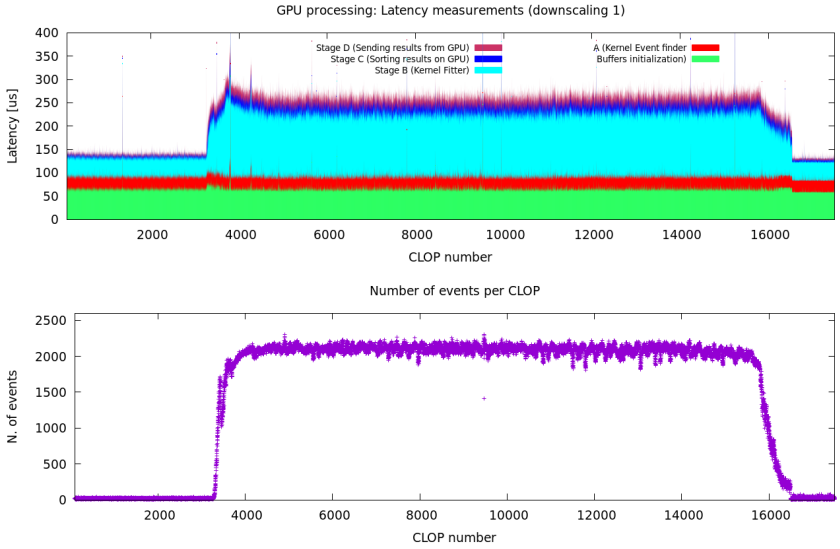

Figure 3: The upper plot shows the latency of the ring reconstruction, highlighting different stages of the process. The lower plot shows the number of events for each buffer, related to the burst shape.

The former approach is more suited for generating IPs which have to be integrated in a custom design, while the latter offers a more standard host-accelerator paradigm, which allows an easier prototyping and deployment of the kernel code, at the expense of a harder customization. It must be emphasized that for the second approach a Board Support Package (BSP) from the vendor is needed.

The Histogram algorithm has been rewritten in OpenCL and deployed on the Terasic DE5-net board, over a custom BSP which enhanced that of the vendor with the UDP and stream processing capabilities (see figure 4). The synthesized kernel runs at 184 MHz using 72% of the logic, 86% of the DSPs and 95% of the memory available on the board. This design has been tested offline sending events as UDP packets through the 10GbE interface and collecting ring fitting results over the same interface. A plot of the event processing latencies is depicted in figure 5, showing that after an initial transient, which seems to be correlated with the sender data rate, latency becomes extremely stable. The overall average throughput is about 5 µs per event, which is an order of magnitude worse than the measured GPU one, but two points must be taken into account:

- there is no dedicated block for floating point arithmetic on the targeted device;
- some of the optimizations for the synthesized kernel could not be enabled because of the lack of enough resources on the board.

State-of-the-art Stratix 10 FPGAs provide ∼3x more resources and ∼20x more DSPs with native variable precision arithmetic support, promising a big performance improvement.

5. Conclusions

The results achieved for the rings reconstruction on the GPU show the effectiveness of the heterogeneous pipeline approach of the NaNet framework, more tests are planned before the LHC

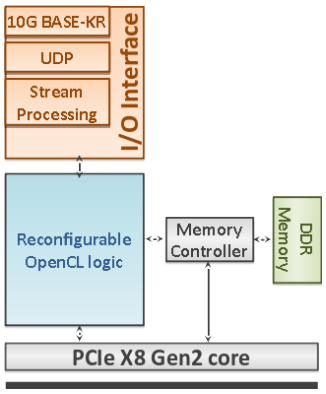

Figure 4: Custom Board Support Package used for kernels generation with the Intel SDK for OpenCL framework.

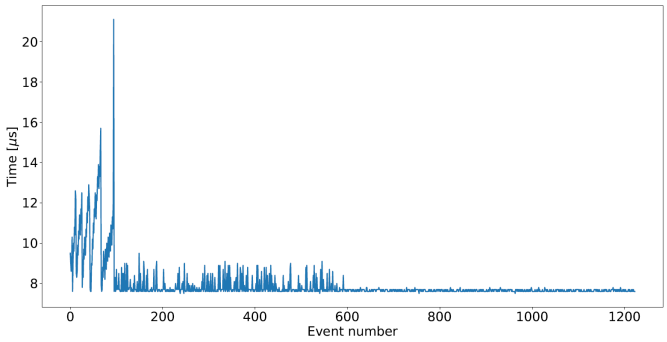

Figure 5: Latency for the OpenCL Histogram kernel performing rings reconstruction.

Long Shutdown 2 (LS2) to assess the quality of the reconstruction and finalize the integration with the other detectors. Moreover the promising results of the reconstruction on the FPGA point out that these devices must be taken into account for intensive computation tasks as well, especially when stable latencies are required.

Acknowledgments

This work was partially funded by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Specific Grant Agreements No. 671553 (ExaNeSt), No. 785907 (Human Brain Project SGA2) and No. 720270 (Human Brain Project SGA1).

References

[1] A. Lonardo, F. Ameli, R. Ammendola, A. Biagioni, A. C. Ramusino, M. Fiorini et al., *Nanet: a configurable nic bridging the gap between hpc and real-time hep gpu computing*, *Journal of Instrumentation* **10** (2015) C04011.

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

[3] F. Spinella, B. Angelucci, G. Lamanna, M. Minuti, E. Pedreschi, J. Pinzino et al., *The tel62: A real-time board for the na62 trigger and data acquisition. data flow and firmware design*, in *2014 19th IEEE-NPSS Real Time Conference*, pp. 1–2, May, 2014. DOI.

[4] B. Angelucci, R. Fantechi, G. Lamanna, E. Pedreschi, R. Piandani, J. Pinzino et al., *The fpga based trigger and data acquisition system for the cern na62 experiment*, *Journal of Instrumentation* **9** (2014) C01055.

[5] R. Ammendola, A. Biagioni, M. Fiorini, O. Frezza, A. Lonardo, G. Lamanna et al., *Nanet-10: a 10GbE network interface card for the GPU-based low-level trigger of the NA62 RICH detector.*, *Journal of Instrumentation* **11** (2016) C03030.

[6] J. Crawford, *A non-iterative method for fitting circular arcs to measured points*, *Nuclear Instruments and Methods in Physics Research* **211** (1983) 223 – 225.