PAPER

Progress report on the online processing upgrade at the NA62 experiment

Published by IOP Publishing for Sissa Medialab

Received: *October 24, 2021*
Accepted: *January 26, 2022*
Published: *April 4, 2022*

Topical Workshop on Electronics for Particle Physics 2021
20–24 September, 2021
Online

Progress report on the online processing upgrade at the NA62 experiment

R. Ammendola,[b] A. Biagioni,[a] A. Ciardiello,[a] P. Cretaro,[a] O. Frezza,[a] G. Lamanna,[e,1] F. Lo Cicero,[a] A. Lonardo,[a] M. Martinelli,[a] R. Piandani,[f] L. Pontisso,[a] M. Raggi,[d,2] F. Simula,[a] D. Soldi,[c] M. Turisini[a,*] and P. Vicini[a]

[a] *INFN, Sezione di Roma, Roma, Italy*
[b] *INFN, Sezione di Roma Tor Vergata, Roma, Italy*
[c] *INFN, Sezione di Torino, Torino, Italy*
[d] *Università Sapienza di Roma, Roma, Italy*
[e] *Università di Pisa, Pisa, Italy*
[f] *University of Chinese Academy of Science, Beijing, China*

E-mail: matteo.turisini@roma1.infn.it

Abstract: A new FPGA-based low-level trigger processor has been installed at the NA62 experiment. It is intended to extend the features of its predecessor due to a faster interconnection technology and additional logic resources available on the new platform. With the aim of improving trigger selectivity and exploring new architectures for complex trigger computation, a GPU system has been developed and a neural network on FPGA is in progress. They both process data streams from the ring imaging Cherenkov detector of the experiment to extract in real time high level features for the trigger logic. Description of the systems, latest developments and design flows are reported in this paper.

Keywords: Pattern recognition, cluster finding, calibration and fitting methods; Trigger concepts and systems (hardware and software); Data processing methods; Trigger algorithms

*Corresponding author.
[1] Also at INFN, Sezione di Pisa, Pisa, Italy.
[2] Also at INFN, Sezione di Roma, Roma, Italy.

https://doi.org/10.1088/1748-0221/17/04/C04002

Contents

1 Introduction

NA62 is a kaon physics experiment located in the north area of the SPS complex at CERN. It adopts a novel decay-in-flight technique to record ultra rare decays of K mesons from a 75 GeV/c positively charged hadron beam. The apparatus can detect kaons and related decay products using a wide range of detectors distributed along a 270 meters long beamline. A proximity focusing Ring Imaging Cherenkov (RICH) is used to increase the muon/pion separation in the final state in addition to calorimetry and kinematic detectors. The signals of interest can occur with a frequency 10 orders of magnitude less than the background, requiring a beam of high intensity coupled with good trigger selectivity. The nominal intensity is 750 MHz for the hadron beam with a 10 MHz particles rate on the RICH.

The online data selection is operated by a fully digital multilevel trigger together with a network-based data acquisition system (TDAQ). A common infrastructure distributes clock, backpressure and reference signals to all detectors. The lowest level trigger is called Level Zero (L0) and it is implemented on an Altera Stratix IV FPGA. The task of L0 is to produce trigger decisions based on minimal information (called trigger *primitives*) received from a subgroup of fast detectors with a latency less than one millisecond.

Our work has focused on the hardware upgrade of the L0 trigger processor (L0TP+) and on the development of two innovative modules that improve the information encoded in the RICH primitives. At present the RICH primitives encode only hit multiplicity, while the proposed systems can calculate ring related quantities relying on high throughput trackless seeded algorithms. The paper is organized as follows: section 2 describes a GPU-based ring reconstruction system currently under finalization at the experiment (GPURICH); section 3 presents L0TP+ hardware which has been recently installed at the experiment and is currently under comparative validation test; section 4 introduces a preliminary neural network classifier on FPGA for real time inference over the RICH data stream (RiNNgs).

2022 JINST 17 C04002

2 GPURICH

The RICH detector of NA62 was designed for pion/muon separation in the range 15 to 35 GeV/c. A typical performance plot and an event with three charged particles are shown in figure 1 as examples of its particle identification (PID) capability. The RICH electronics consists of an array of 1952 photomultiplier tubes (PMT), 64 frontend cards and 4 readout boards in charge of data digitization, timestamping, buffering and networking. In the baseline TDAQ configuration the RICH primitives are generated by means of a fifth readout board. The signal sent to L0 is derived from a simple OR of 8 adjacent PMTs, resulting in a trigger logic that is based on hit multiplicity at low granularity and no ring information. This solution is motivated by an intrinsically high time precision at the level of 70 picoseconds [1], and practical absence of random coincidences. However, it does not exploit any PID capability of the detector, which is fully used during offline analysis instead.

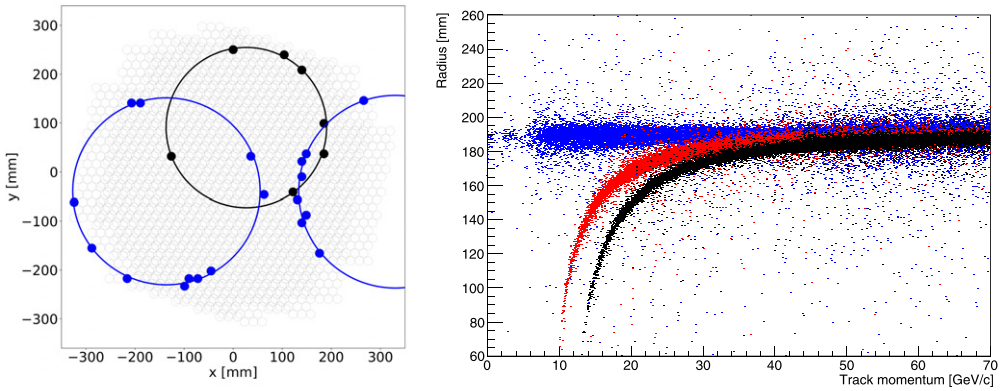

Figure 1. Left: a hit pattern on the RICH array with reconstructed rings geometry superimposed. Right: ring radius versus particle momentum for 130k events. Electrons in blue, muons in red and pions in black. Particles are identified offline combining RICH, tracking and calorimetry data.

A new GPU-based system for online ring fitting and physics-related primitive generation, called GPURICH, has therefore been proposed and installed. It operates on the full detector granularity and leverages the FPGA-based custom network interface called NaNet [2] to transfer data with low and controlled latency from the detector to the GPU memory. Primitives are then computed using a histogram based ring detection and a standard ring fitting algorithm. The adopted boards are a Terasic DE5 with an Altera Stratix-V FPGA for NaNet, and a NVIDIA Pascal P100 GPU for the ring geometry reconstruction. The two boards communicate directly through GPUdirect RDMA protocol on PCIe bus. Data from the detector to the trigger processor are sent using eight 1 GbE links aggregated through a switch to a 10 GbE link. Since GPURICH provides physics related quantities like the number of particles and the ring radius information already at L0, it can participates in the trigger decision as a smart additional detector.

During NA62 run1 (in 2018) GPURICH was installed in the experimental hall and the synchronization with other detectors generating primitives was validated. A total latency well below

2022 JINST 17 C04002

one millisecond has been measured, with an average processing time of 130 ns per event [3]. For run 2 (2021–2022) the efforts are devoted to specialize the algorithms to electron identification.

3 L0TP+

The Level Zero Trigger Processor (L0TP) of NA62 is entirely hardware based and implemented on an FPGA [4]. It receives trigger *primitives* from a subset of detectors, operates logical coincidences between them according to user defined masks and sends the trigger decision to the NA62 common infrastructure depending on a programmable downscaling factor for each mask. A proximity PC is used to interface with the run control of the experiment, debug and collect data for offline trigger analysis. Current implementation is a on a Terasic DE4 board housing a Stratix-IV FPGA device and eight 1 GbE links.

The hardware and logic resources available on the current L0TP cannot sustain a beam intensity increase nor additional computations to improve selectivity. To overcome these limitations a more recent platform has been adopted and the new system is called L0TP+ [5]. An illustration of the assembly based on a Xilinx Virtex UltraScale+ FPGA VCU118 board is shown in figure 2. In parallel with a test bench for characterization and debugging at the INFN laboratories, the L0TP+ has been installed in the experimental hall and validation is ongoing to replicate the performance obtained so far of the previous system before adding new features.

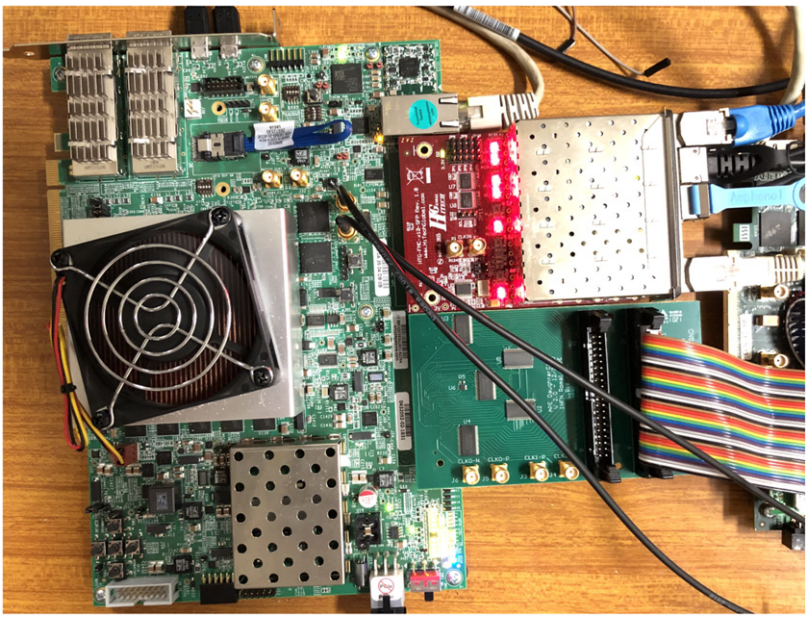

Figure 2. L0TP+ assembly: Xilinx Virtex UltraScale+ FPGA VCU118 board with 2 mezzanine cards to interface with detectors, PC farm and control signals from NA62 infrastructure.

A 10 GbE channel has been already added to support an output rate of up to 10 MHz and a microcontroller soft IP has been embedded in the new FPGA device for a smoother integration with the run control. Currently occupied resources are 23% of IO pins, 19% gigabit transceivers,

2022 JINST 17 C04002

19% BRAM, 8% of LUT, 6% of registers and less than 1% of DSP blocks. The additional resources allows also for new trigger algorithms to be implemented as described in the next section.

4 RiNNgs

An event classifier for the RICH detector compatible with L0 timing requirements has been developed using machine learning techniques and high level synthesis tools for FPGA design. This is motivated by a potential benefits for the trigger selectivity given by the full exploitation of the Cherenkov signals at the online level and is an example of how offline processing can be turned online due to the latest developments in hardware and software. As a first case study, the task of counting the number of rings was considered and even if ring features extraction can be a trivial task for modern deep neural networks, the 10 MHz particle flux on the RICH still represents a challenge. The classifier is called Rings Neural Network (RiNNgs) and is composed of 3 fully connected layers with respectively 64, 16 and 4 output neurons activated by ReLu functions. It takes the event hit list as input (normalized by the number of frontend channels) and produces the probability of having 0, 1, 2, 3 or more rings in output (labels). The dataset consists of 80k events per label and was obtained within the NA62 analysis framework using experimental data and track seeded ring reconstruction.

Figure 3 presents the inference performance after training in the standard TensorFlow framework. For the confusion matrix (figure 3, left) the average accuracy performance is about 80%, satisfactory for the L0 trigger. Moreover, some trigger conditions could tolerate an overestimation of the number of particles, rising the efficiency above 80% for all the cases. On the right, the receiver operating characteristics are shown for the four labels. The curves can be used to optimize the trigger for high efficiency or high purity.

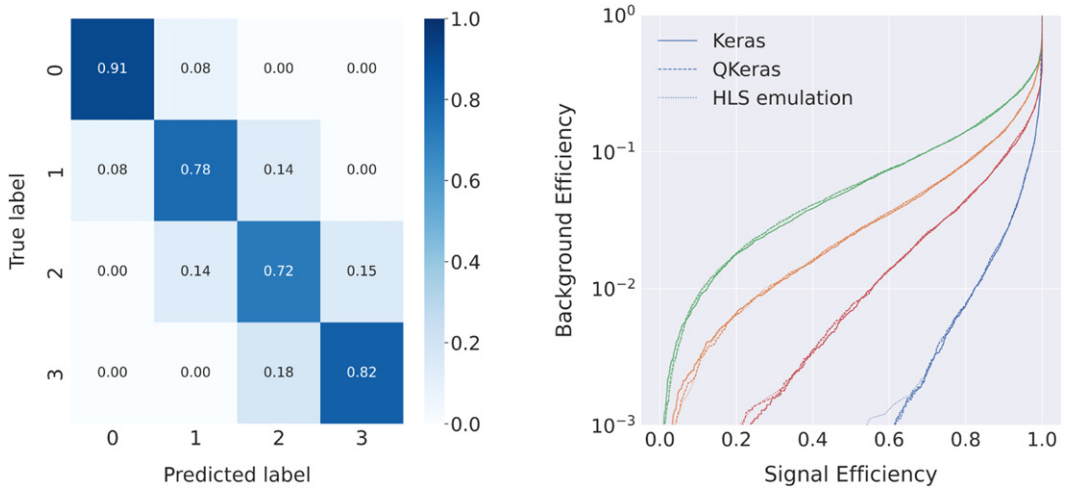

Figure 3. RiNNgs particle counting performance, the model is trained to predict the number of charged particles (label) in each event. Left: confusion matrix result for validation dataset (8k events). Right: simulation results for the 4 output labels i.e. rings: 0 (blue), 1 (orange), 2 (green), 3 or more (red). Reducing the numerical precision (QKeras) does not degrade the floating point (Keras) performance. This result is confirmed using an emulated version of the FPGA design (HLS emulation).

2022 JINST 17 C04002

The corresponding FPGA design is obtained using the HLS4ML Python package [6] in combination with the Xilinx Vivado High Level Synthesis (HLS) tool. The target board is the VCU118, since the L0TP+ will be the system were RiNNgs is to be deployed. As shown in table 1 the design occupies a modest fraction of the available resources, allowing the neural network and the trigger processor to be placed on the same FPGA. Table 1 also shows the HLS estimate for the timing characteristics which is compatible with a 10 MHz throughput. After an initial 18 bits fixed point data representation (*baseline*, 8 bits of integer part) a reduced precision (*quantized*) attempt was made, obtaining the same performance with less DSP resources and better timing (see table 1 for details). A necessary step to achieve this goal was to perform part of the training phase using a Python library called QKeras [7] that supports arbitrary quantization datatypes.

Table 1. HLS report summary using different arithmetic representations. The Instantiation Interval (II) is the minimum time distance between consecutive events to be processed. With a conservative estimation of 10 nanoseconds clock period the II results in 100 nanoseconds, thus compatible with the 10 MHz throughput requirement. The quantization is determined from the numerical range of the weights and biases after an initial training phase. The results showed are obtained with 7 and 9 bits fixed point representation with 1 bit dedicated to the integer part.

	BRAM [%]	DSP [%]	FF [%]	LUT [%]	II [Clock ticks]	LATENCY [Clock ticks]	FMAX [MHz]
Baseline	1	10	<1	3	10	40	100
Quantized	0	6	<1	15	10	20	150

The design flow proved to be very effective for the fast implementation of a neural network on FPGA. However the following two limitations prevent at the moment a better implementation: (a) a limit of 4096 cycles on the unrolling factor for the pipelined design in Vivado HLS, and (b) unsupported binary representation of the variables in QKeras (only weights and biases can be quantized). Attempts are ongoing to overcome these limitations, both by editing manually the C++ code generated by the HLS4ML library and by migrating to the more recent Vitis HLS platform.

5 Conclusions

The new low level trigger processor of NA62 is currently under test. With 10 GbE support and significantly more logic, it will support the future NA62 physics programme, including an increase in beam intensity. For a better trigger selectivity two systems are under development to extract high level features from the RICH online data stream at full granularity. A GPU-based system performing a geometrical ring reconstruction algorithm is under finalization in the experimental hall. A preliminary neural network model to be implemented on the L0TP+ FPGA is under development. They both satisfy the requirements to run online and are examples of distributed trigger computing paradigms with potential benefits for physics discovery.

2022 JINST 17 C04002

Acknowledgments

This work has been carried out within the EuroEXA project (G.A. EU H2020 FP No. 754337) and TEXTAROSSA project (G.A. H2020-JTI-EuroHPC-2019-1 No. 956831).

References

[1] G. Anzivino et al., *Light detection system and time resolution of the NA62 RICH,* 2020 *JINST* **15** P10025.

[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.

[3] P. Cretaro et al., *NaNet: a reconfigurable PCIe network interface card architecture for real-time distributed heterogeneous stream processing in the NA62 low level trigger, PoS* **343** (2019) 118.

[4] R. Ammendola et al., *The integrated low-level trigger and readout system of the CERN NA62 experiment, Nucl. Instrum. Meth. A* **929** (2019) 1.

[5] R. Ammendola et al., *L0TP+: the upgrade of the NA62 level-0 trigger processor, EPJ Web Conf.* **245** (2020) 01017.

[6] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics,* 2018 *JINST* **13** P07027.

[7] C.N. Coelho et al., *Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors*, arXiv:2006.10159 [physics.ins-det].

2022 JINST 17 C04002