Performance and advantages of a soft-core based parallel architecture for energy peak detection in the calorimeter Level 0 trigger for the NA62 experiment at CERN

R. Ammendola, M. Barbanera, M. Bizzarri, V. Bonaiuto, A. Ceccucci, B. Checcucci, N. De Simone, R. Fantechi, L. Federici, A. Fucci, M. Lupi, G. Paoluzzi, A. Papi, M. Piccini, V. Ryjov, A. Salamon, G. Salina, F. Sargeni and S. Venditti

Abstract: The NA62 experiment at CERN SPS has started its data-taking. Its aim is to measure the branching ratio of the ultra-rare decay $K^+ \rightarrow \pi^+ \nu \bar{\nu}$. In this context, rejecting the background is a crucial topic. One of the main background to the measurement is represented by the $K^+ \rightarrow \pi^+ \pi^0$ decay. In the 1-8.5 mrad decay region this background is rejected by the calorimetric trigger processor (Cal-L0). In this work we present the performance of a soft-core based parallel architecture built on FPGAs for the energy peak reconstruction as an alternative to an implementation completely founded on VHDL language.

Keywords: Trigger detectors; Calorimeters

Corresponding author.
1 Introduction

The goal of the NA62 experiment [1, 2], located in the CERN North Area, is to improve the precision in measuring the branching ratio of the ultra-rare decay $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ with 100 expected events and less than 10% background [3–5]. The 400 GeV/c high-intensity SPS (Super Proton Synchrotron) proton beam impinges on a beryllium target, producing a 750 MHz secondary hadron beam of which 6% are kaons. They are selected with a momentum of 75 GeV/c and decay in flight along a 65 m fiducial decay region. Because of the small branching ratio of the processes under study, a complex trigger system has been designed.

In particular, as depicted in figure 1, various detectors are distributed along a 270 m long region and each of them is devoted to particle identification and to the high precision measurement of the various physical quantities.

- A Differential Cerenkov Counter (CEDAR) that is placed on incoming beam to positively identify the kaons in a high rate environment before their decay.
- A spectrometer (GigaTracker — GTK) provides the precise measurements of momentum, time and angle of incoming kaon beam.
- A detector to veto charged particles to reduce the background (CHarged ANTI-Counter — CHANTI).
- Twelve photon veto system stations (Large Angle Photon Veto — LAV) to reject the dominant background from channels with final state photons.
- The Straw Chamber Magnetic Spectrometer (STRAW), to measure with good accuracy the direction and the momentum of secondary charged particles originating from the decay region.
Figure 1. Schematic view of the NA62 detectors.

- The Ring Imaging Cherenkov detector (RICH), to identify and reduce muons contaminating pion samples and measure the arrival time of charged tracks.
- The Charged Particle Hodoscope (CHOD) that is used to detect possible photo-nuclear reaction in the RICH mirror plane.
- Two small-angle vetoes: Small Angle Calorimeter (SAC) and Intermediate Ring Calorimeter (IRC). They complete the coverage of the NA62 photon veto system for particles that would otherwise escape down the beam pipe.
- The Liquid Krypton (LKr) calorimeter [9] used for rejecting photons from the various K\(^+\) decay modes, in particular \(K^+ \rightarrow \pi^+\pi^0\) in the 1-8.5 mrad forward region.
- The Muon-Veto detectors (MUV1, MUV2 and MUV3): a set of three hadronic calorimeters and muon detectors used for the suppression of the background with muons in the final state.

The experiment data-taking phase started in 2016 and it has been scheduled until the end of 2018.

2 The TDAQ system

The CERN SPS 400 GeV/c primary beam will provide \(3 \times 10^{12}\) protons per spill (4.8 s burst duration with a period of 16.8 s) impinging on a beryllium target. The selected 75 GeV/c secondary hadron beam will result in an instantaneous kaon rate of about 50 MHz. Since this high-intensity kaon beam is required to collect enough statistics, the L0 trigger plays a fundamental role in both the background rejection and in the particle identification. For this reason, a complex Trigger and Data Acquisition (TDAQ) system has been designed [6].

The TDAQ System handles up to 25 GB/s raw data from twelve sub-detectors for a total of about 80000 channels. It has been structured on three different trigger levels [7, 8].

In particular, the first level (L0) is a hardware synchronous level that is able to reduce the data rate from a value of 10 MHz to 1 MHz with a maximum latency of 1 ms.

The second level handles data from a single detector and is carried out in software. It is able to reduce the data rate from 1 MHz to 100 kHz with a maximum latency of O(1s).
The last level (L2) is again a software level where the algorithms are executed on the complete event.

In particular, as depicted in figure 2, the RICH, the MUV1 and MUV2, the CEDAR, the LKr[9], the IRC, the SAC, one of the LAVs (LAV12) and the CHOD are involved in L0. While the GTK, the CHANTI, the STRAW, the other LAV stations (from 1 to 11) and the last MUV detector (MUV3) are involved from L1 only.

### 3 The Liquid Krypton Level 0 trigger

The LKr is a high-performance electromagnetic calorimeter, about 27 radiation lengths, with 13248 channels consisting of 2×2 cm² cells of thin copper-beryllium ribbons, kept at high voltage, and immersed in a 10 m³ liquid krypton bath at 120 K acting as active medium. This paper focuses on the calorimeters level 0 trigger (Cal-L0) that receives the data from the three electromagnetic calorimeters readout boards (Calorimeter REAdout Module — CREAM) [9]: the LKr, the SAC, the IRC, MUV1 and MUV2. In particular, the CREAM digitizes the analogue channels of the calorimeter, buffer them for the trigger L1 decision and provides a Trigger Sum Links (TSL) of 16 (4×4) calorimeter cells to the calorimetric Level-0 Trigger (figure 3) (864 TSL for the LKr, 1 for IRC, 1 for SAC, 12 for MUV1 and 6 for MUV2).

The Cal-L0 is a sequence of three connected layers.
The first layer (Front End in figure 4) is made of 29 electronic boards (figure 5), and collects the data coming from the calorimeters and detects the energy peak underlying parametrized constraints. The majority of the first layer, 28 boards each one corresponding to a vertical slice of the LKr calorimeter, sends the detected peaks to the second layer. In particular, this layer identifies the electromagnetic clusters in the calorimeter and prepares a time-ordered list of reconstructed clusters together with the arrival time, position, and energy measurements of each cluster.

The second layer is composed of seven boards (Merger), each of which is able to merge information coming from four LKr slices and send the aggregated information to the last level.

Finally, the third (Concentrator) layer aggregates all the information received from the LKr second layer as well as the other calorimeter boards (IRC, SAC, MUV1, MUV2) and sends it to the central level 0 trigger processor [10–14].

3.1 Trigger algorithm

The identification process consists of the energy peak reconstruction and the estimation of some shape properties. It is composed of four different channel by channel tasks performed in parallel on each sub-set elements time sample. This is performed by the analysis of the 896 data streams (i.e. the sum of the values of the energy deposit over 16 cells) coming directly from the CREAMs (16 bit samples per channel every 25ns).
The first task is a threshold check on incoming samples to avoid counting the activity of noisy channels as valid events (figure 6a).

The second task is a data analysis on the over-threshold samples for the identification of an in-time trend compatible with an energy deposit (peak in time) defined as $E_{i[n-2]} < E_{i[n-1]} \text{ AND } E_{i[n-1]} > E_{i[n]}$. Where $E$ is the sample data of the energy deposit, $i$ is $i$-th channel (i.e. the tile number) and $n$ the sample number (figure 6b).

The third task consists of a peak search in space, where a peak in space is defined by the following condition: $E_{i-1}[n] < E_{i}[n] \text{ AND } E_{i}[n] > E_{i+1}[n]$.

The fourth task is the estimation of some energy deposit shape properties: its maximum value (i.e. the real maximum energy deposit) and fine time (i.e. the computation of a more precise value of the energy deposit starting time). In particular, the estimation of the value of the maximum real deposit shape is obtained by mean of a parabolic in time interpolation with the least square method applied on constant time references, while the event precise time estimation can be computed by the bisection or the linear fit method (figure 6c and figure 6d).

3.2 Trigger processor implementation

The trigger algorithm previously introduced is physically implemented on FPGAs located on the Front-End layer of the Liquid Krypton Level 0 system. Attention will be placed on the two alternatives implementations that can be used to obtain the fine time estimation of the energy peak determined by the particles decay products in the calorimeter.
The first possible implementation is to design the required hardware by using a code totally written with combinatorial logic by means of VHDL language while the second implementation deals with an architecture based on soft-core processors. In particular, we present the performances of the peak detection procedure implemented by using both the approaches.

The first solution takes advantage of the speed and less occupied resources compared to the second one, which, by using of the C language to program the soft-core processors, simplifies the algorithm implementation and adds higher flexibility to the system [15]. Moreover, it takes advantage of the digital signal processing (DSP) blocks available on a last generation FPGAs.

3.2.1 VHDL implementation

The logic exploited to compute the fine time estimation is based on the eight steps bisection method. The parabolic fit around the detected maximum value was carried out as a pipeline module with: a multiplier (a 17-bit word as input and a 34-bits word as output), a divider entity (a 17-bit word as denominator and a 34-bit word as numerator for a 34-bit word as quotient). Moreover, a proper number of delay lines has been added in order to take into account the relative latency of each block. The bisection method receives as input the data samples that have been identified as part of an energy peak and it applies on them, in sequence, the bisection operation by comparing the outcome with the mean value of the range under investigation. The number of the bisection stages corresponds to the required resolution of the fine time (i.e. each step computes a bit of the output word).

3.2.2 Embedded processors for trigger logic

The main component of this solution is the Altera® Nios® II 32-bit soft-core processor [16] that can be surrounded by other logic blocks able to interface it with the rest of the project. In particular, in this design, these interfaces consist of two data memories suited to minimize the resources’ occupancy per core as well as to achieve the storage of the required number of data words to work on.

The powerfulness of a design like this is to reduce the requested entities to those that are strictly needed by the project. In particular, a couple of dual port 1024 bits RAM sized to host four words 256 bits wide each have been reserved for input data and the other for output data respectively (figure 7). Each of the input words consists of a conveniently arranged deposited energy information coming from the eight channels spanned by one PP (Pre-Processing) FPGA and that already passed the peak in time and threshold selections. Moreover, these memories serve as data buffers for data waiting to process. When the next available location in the Nios II interface memory is free, the new data word to process move to that location. The new word written in the soft-core input data memory, always hosting the event time reference, will be the trigger for the computational core of new data available to process.

The soft-core goes out its reading loop on the location when it finds a new data word and starts processing again. All the memories in the Nios II project are configured as tightly coupled memories; with this choice, the memories out of the computational core are used as processor’s caches with the same latency of this kind of temporary memory avoiding the needs to implement them inside the processor. Therefore, the data do not need to be copied in temporary variables but can be accessed directly for computation.

Furthermore, the input data memory location can be made available when the access to the original package is no longer required. Since the output memory may not be available when the fine-time
Figure 7. Block diagram of embedded processors and memories.

computation ends, the process checks if the location is ready to be overwitten before moving results to the next location. After the green light, the data are copied in a precise sequence inside the 256 bits output word and a next location to point is computed. The process is now ready to restart on new data when available. The Nios II core also has a 2 kB instruction memory at its disposal to host the compiled C code needed for computations and all of the temporary structures needed to run the software. This specific memory size is the results of the compromise between precision, resources’ occupancy and speed. In particular, a short version of the C libraries was chosen to minimize the memory resources required, for the same reason all the operations are made on integers instead of using the floating point operations that are also available. This last choice also has implications in performance because floating point operations are time consuming and their dedicated libraries are bigger.

Thanks to the great flexibility given by the possibility to implement new functionality simply writing C code, leaving all the hardware structure unchanged, the algorithm was initially atomized to take advantage of the pipeline properties of the Nios II and to analyse which section of the computation requires more time to be concluded.

A single Nios II core working with a 320 MHz clock with the optimized configuration explained above presents a latency of about 1.4 $\mu$s that means a supportable rate of about 700 kHz. In order to hold up the expected 5 MHz worst peak rate per PP-FPGA, a multi soft-core architecture was realized with the number of requested cores given by:

$$\frac{1.4 \mu s}{N_c} \leq \frac{1}{5MHz} \rightarrow N_c \geq \frac{1.4}{0.2}$$

where $N_c$ is the number of soft-cores processors. In particular, in the described system 8 Nios II cores are used. The data word to process are distributed to the cores with a round robin scheduling.
The data distribution logic explores the input memory’s location of each computational core in a cyclic path. While waiting for new data this distribution entity is pointing to a free input memory location. When the new word arrives, it is written in the pointed location and the distribution entity points to the next scheduled location in the following computational module. Once all of the first input memories’ locations have been visited for the first time, the scheduling proceeds with the second location of the first computational core, and so on. If the next location that is pointed to is not available, the round robin scheduler waits until this location is free. A similar output logic exists.

4 Performance tests

By means of Altera SignalTap II tool, the block of logic described with VHDL and delegated to the computation of the fine time of energy deposit has a latency of 37 clock cycles that, working at 40 MHz corresponds to a latency of 925 ns. The resolution of this logic is related to the 8-bits word used to represent the fine time. Because an eight bisection stage is used, the resolution can be computed as $|s_n - s| \leq \frac{x_n - x_{n-1}}{2^n}$ with $n$ equal to 8 and $x_n - x_1 = \Delta t = 25\, ns$.

$$\frac{25}{2^8}\, ns \approx 0.0976\, ns \approx 98\, ps$$

To provide this section measures the Eclipse framework [17] for the Altera Nios II tool was used. With this tool, it is possible to control the running processor, check the memory content, set up parameter for the best suite of the processor to available memory space, etc. Another Nios available feature that is possible to exploit to measure the soft-core performance is the Performance Counter [18]. This is a logic block that must be included in the Nios project in order to execute performance measures in section of C code that implements the processor functionalities. In particular, in the software written to achieve the precise time estimation 5 section were monitored. The section “1st division” include that C code rows that include the division used to find the peak real maximum via the parabolic fit while the “2nd division” rows are included in the block of code where the fine time is estimated. All these sections contribute to the “Total Time”. The spread is due mainly to the latency variability associated to the hardware divide logic, included in project, implemented with multicycle divide circuit inside the Arithmetic Logic Unit (ALU). The variability reported by Altera is from 4 to 66 cycles per instruction. The maximum value associated to the “Total Time” section is 456. Since the Nios II runs at a work frequency of 320 MHz, the latency associated to a single Nios II core built as explained in the section 3.3 is:

$$456 \cdot \frac{1}{320} \cdot 10^{-6} s \approx 1.425 \cdot 10^{-6} \approx 1.4\, \mu s$$

So, an architecture made of 8 Nios II with an associated latency of about 178 ns can support a maximum data rate of about 5.6 MHz.

5 Discussion and conclusions

The results shown in the previous section prove that a soft-core based implementation can achieve the same goal of the VHDL one. In particular, while the VHDL design is faster than the Nios II one (~0.9 µs per logic block with respect to ~1.4 µs per soft-core) the soft core design presents a
better resolution (32-bits instead of 8-bits). In particular, the 8 bits resolution is a design constrain imposed by the TDAQ system that gives a time resolution around 98 ps better than that one of 260 ps estimated for the calorimeter. This result highlights that the comparison between the two implementations cannot be limited to the execution speed only. The main result of this analysis is that a conveniently implemented parallel architecture that has a soft-core processor as computational core can hold up the same input data rate than a pure VHDL implementation with the surplus value of the greater flexibility and a faster development stage. The Nios II, indeed, is a general-purpose RISC processor that take advantage of the DSP block inside the FPGA to accomplish complex computations and it is programmed by means of the standard C language. This aspect was crucial also from the development stage point of view because C code is easier and faster to write, test and maintain than VHDL code. The Nios II is also able to run a Linux based operative system taking advantage of libraries implemented in this environment. Compared to a microcontroller, instead, a soft-core processor like the Nios II can be arranged inside an FPGA and the resources that it uses can be tuned to save logic elements without giving up in performance.

The idea behind this alternative implementation is to exploit a more flexible and easy to implement solution for those portions of the firmware that are more prone to changes and improvements. The alternative between more flexibility with respect to an improvement in the computational speed surely represents a very critical issue. Nevertheless, it is worth to note that the use of embedded processors could surely lead advantages in the readability of the code and, as a consequence, an improvement of the reliability a well as the maintainability of the whole system.

References

[1] NA62 collaboration, Proposal to measure the rare decay $K^+ \to \pi^+\nu\bar{\nu}$ at the CERN SPS, CERN-SPSC-2005-013, CERN, Geneva Switzerland, (2005).

[2] E949 collaboration, A.V. Artamonov et al., New measurement of the $K^+ \to \pi^+\nu\bar{\nu}$ branching ratio, Phys. Rev. Lett. 101 (2008) 191802 [arXiv:0808.2459].

[3] J. Brod, M. Gorbahn and E. Stamou, Two-loop electroweak corrections for the $K \to \pi\nu\bar{\nu}$ decays, Phys. Rev. D 83 (2011) 034030 [arXiv:1009.0947].

[4] NA48/2 collaboration, K. Massri, Precision tests of the Standard Model with kaon decays at CERN, J. Phys. Conf. Ser. 631 (2015) 012040.

[5] NA62 collaboration, V. Duk, LFV and exotics at the NA62 experiment, J. Phys. Conf. Ser. 556 (2014) 012067.

[6] NA62 collaboration, NA62 technical design, NA62-10-07, CERN, Geneva Switzerland, (2010).

[7] M. Sozzi, A concept for the NA62 trigger and data acquisition, NA62-07-03, CERN, Geneva Switzerland, (2007).

[8] NA48 collaboration, V. Fanti et al., The beam and detector for the NA48 neutral kaon CP-violations experiment at CERN, Nucl. Instrum. Meth. A 574 (2007) 433.

[9] NA48 collaboration, A. Gianoli et al., The NA48 LKr calorimeter readout electronics, IEEE Trans. Nucl. Sci. 47 (2000) 136.

[10] A. Ceccucci et al., The NA62 liquid krypton calorimeter’s new readout system, 2014 JINST 9 C01047.
[11] E. Pedreschi et al., *Firmware approach for TEL62 trigger and data acquisition board*, in *Topical Workshop on Electronics for Particle Physics (TWEPP 2012)*, Oxford U.K., 17–21 September 2012.

[12] B. Angelucci et al., *The FPGA based trigger and data acquisition system for the CERN NA62 experiment*, *2014 JINST 9 C01055*.

[13] A. Fucci et al., *The NA62 liquid krypton electromagnetic calorimeter Level 0 trigger*, in *Proc. 13th ICATPP Conf.*, (2011).

[14] V. Bonaiuto et al., *Status of the NA62 liquid krypton electromagnetic calorimeter Level 0 trigger processor*, *2013 JINST 8 C02054*.

[15] M. Barbanera et al., *First operation of the Level 0 trigger of the NA62 liquid krypton calorimeter*, *2015 JINST 10 C03050*.

[16] D. Badoni et al., *Use of FPGA embedded processors for fast cluster reconstruction in the NA62 liquid krypton electromagnetic calorimeter*, *2014 JINST 9 C01010*.

[17] *Altera webpage*, http://www.altera.com.

[18] *Eclipse webpage*, http://www.eclipse.org.

[19] *Profiling Nios II systems*, application notes AN-391-3.0 on http://www.altera.com.