GPUs for the realtime low-level trigger of the NA62 experiment at CERN

R. Ammendola[4], *M. Bauce*[3,7], *A. Biagioni*[3], *S. Chiozzi*[1,5], *A. Cotta Ramusino*[1,5], *R. Fantechi*[2], *M. Fiorini*[1,5,*], *A. Gianoli*[1,5], *E. Graverini*[2,6], *G. Lamanna*[2,8], *A. Lonardo*[3], *A. Messina*[3,7], *I. Neri*[1,5], *F. Pantaleo*[2,6], *P. S. Paolucci*[3], *R. Piandani*[2,6], *L. Pontisso*[2], *F. Simula*[3], *M. Sozzi*[2,6], *P. Vicini*[3]

[1]INFN Sezione di Ferrara, Via Saragat 1, 44122 Ferrara, Italy
[2]INFN Sezione di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy
[3]INFN Sezione di Roma"La Sapienza", P.le A. Moro 2, 00185 Roma, Italy
[4]INFN Sezione di Roma "Tor Vergata", Via della Ricerca Scientifica 1, 00133 Roma, Italy
[5]University of Ferrara, Via Saragat 1, 44122 Ferrara, Italy
[6]University of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy
[7]University of Rome "La Sapienza", P.le A.Moro 2, 00185 Roma, Italy
[8]INFN Sezione di Frascati, Via E.Fermi 40, 00044 Frascati (Roma), Italy
[*] Corresponding author. E-mail: fiorini@fe.infn.it

DOI: http://dx.doi.org/10.3204/DESY-PROC-2014-05/15

A pilot project for the use of GPUs (Graphics processing units) in online triggering applications for high energy physics experiments (HEP) is presented. GPUs offer a highly parallel architecture and the fact that most of the chip resources are devoted to computation. Moreover, they allow to achieve a large computing power using a limited amount of space and power. The application of online parallel computing on GPUs is shown for the synchronous low level trigger of NA62 experiment at CERN. Direct GPU communication using a FPGA-based board has been exploited to reduce the data transmission latency and results on a first field test at CERN will be highlighted. This work is part of a wider project named GAP (GPU application project), intended to study the use of GPUs in real-time applications in both HEP and medical imaging.

1 Introduction

The trigger system plays a fundamental role in any HEP experiment since it must decide whether a particular event observed in a detector should be recorded or not, based on limited information. Every experiment features a limited amount of DAQ bandwidth and disk space for data storage, the use of real-time selections is crucial to make the experiment affordable maintaining at the same time its discovery potential. Online selection of important events can be performed by arranging the trigger system in a cascaded set of computational levels. The low-level (hereinafter referred to as L0) is usually realized in hardware, often based on custom electronics devoted to the experiment and normally developed using programmable logic (FPGA) that offer flexibility and possibility of reconfiguration. The upper trigger levels (L1 and L2) are implemented in software on a commodity PC farm for further reconstruction

and event building. In the baseline implementation, the FPGAs on the readout boards compute simple trigger primitives on the fly, such as hit multiplicities and rough hit patterns, which are then timestamped and sent to a central trigger processor for matching and trigger decision.

This paper presents the idea of using GPUs for low-level triggering in HEP experiments. In fact, GPUs provide a huge computing power on a single device, thus allowing to take complex decisions with a significantly high speed. In particular, in the standard multi-level trigger architecture, GPUs can be easily exploited in the higher software levels, where the number of computing farm nodes can be reduced and the capability of the processing system can be improved without increasing the scale of the system itself. As an example, GPUs are currently under study in the software trigger level of ATLAS experiment at CERN [1] and are implemented with encouranging results in tracking of Pb-Pb Events in the ALICE experiment [2].

Low-level triggers can also take advantage from the usage of GPUs, but a careful assessment of their online performance is required especially in terms of computing power and latency. In fact, low level trigger systems are designed to perform very rough selection based on a sub-set of the available information, typically in a pipelined structure housed in custom electronics, in order to bring to a manageable level the high data rate that would otherwise reach the software stages behind them. Due to small buffers size in the read-out electronics, such systems usually require very low latency. A low total processing latency is usually not crucial in the applications for which GPUs have been originally developed. On the contrary, for a GPU-base trigger system, data transfer latency to the GPU and its stability in time become a very important issue.

In this paper we present recent results of the ongoing R&D on the use of GPUs in the lowest trigger level of the NA62 Experiment at CERN, within the framework of the GAP project [3]. Next section describes the idea of the adoption of GPUs in the NA62 L0 trigger while in the other subsection we address more specifically the latency control problem, and we show the results on the solution that we are currently developing. This is based on a custom "smart" NIC that allows copying data directly into the GPU.

2 GPUs in the low-level trigger of the NA62 experiment

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 goal is to collect ~ 100 events with a 10:1 signal to background ratio, using a high-energy (75 GeV/c) unseparated hadron beam decaying in flight [4, 5]. In order to manage the 25 GB/s raw data stream due to a ~ 10 MHz rate of particle decays illuminating the detectors, the NA62 trigger consists of three levels as illustrated in the previous section: the L0 is based on FPGA boards which perform detector data readout [6], while the next two levels are developed on PCs, thus implemented in software. L0 must handle an input event rate of the order of 10 MHz and apply a rejection factor of around 10, in order to allow a maximum input rate of 1 MHz to the L1 trigger. The latter, together with the L2 trigger, must reduce the rate to about 10 kHz in order to permit permanent data storage for later offline analysis. The maximum total latency allowed by the NA62 experiment for the L0 trigger is 1 ms.

A pilot project within NA62 is investigating the possibility of using a GPU 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 the standard L0 implementation, trigger primitives contributing to the final trigger decision are computed on

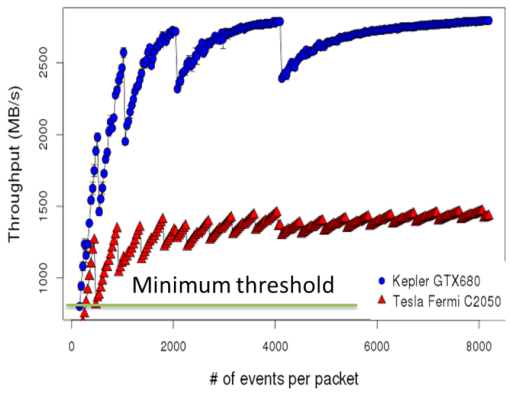

Figure 1: Throughput as a function of the number of events for last generation GPUs.

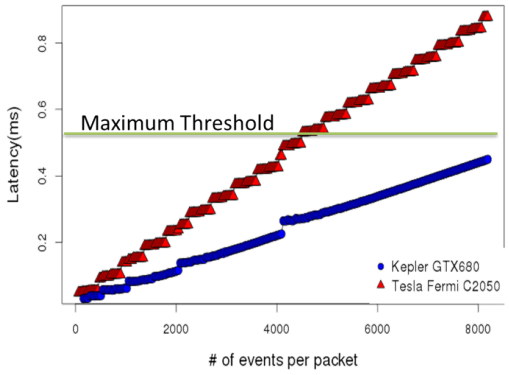

Figure 2: Total latency (including data transfer and computing). Here the maximum threshold does not take into account the data transfer from readout, but it only estimates the total latency contribution within the GPU.

the readout board FPGAs, and are mostly based on event hit patterns. The use of GPUs in this level would allow building more complex 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.

In particular, the reconstruction through GPUs of the ring-shaped hit patterns within the NA62 Ring Imaging Cherenkov (RICH) detector represents the first study case on the use of GPUs at low-level trigger in the GAP project. The proposed GL0TP will operate in parasitic mode with respect to the main L0 trigger processor by processing data coming only from the RICH detector. Such detector, described in [7], provides a measurement of the velocity and direction of the charged particles crossing its volume above the Cherenkov threshold. It can therefore contribute to the computation of other physical quantities, such as the decay vertex of the K^+ and the missing mass. On the basis of such information, highly selective trigger algorithms can be implemented for several interesting K^+ decay modes.

As highlighted in [8], several ring reconstruction algorithms have been studied in order to assess the best GPU-based implementation. In particular, the one based on a simple coordinate transformation of the hits which reduces the problem to a least square procedure was found to be the best ring-fitting algorithm in terms of computing throughput. This algorithm was developed and tested on different GPUs, such as the NVIDIA Tesla C1060, Tesla C2050 and GeForce GTX680. The computing performance of the C2050 and GTX680 proved to be a factor 4 and 8 higher than that of the C1060, respectively.

Figure 1 shows the computing throughput for these devices as a function of the number of events processed in one batch. The effective computing power is seen to increase with the number of events to be processed concurrently. The horizontal line shows the minimum throughput

requirement for an online trigger based on the RICH detector of the NA62 experiment.

Figure 2 points out the total computing latency that includes data transfer times to and from the GPU and the kernel execution time. Here NVIDIA Tesla C2050 and GeForce GTX680 devices have been used for the measurement. The significant reduction of the latency for the newer GTX680 GPU is due to the faster data transfer allowed by the 3rd generation PCIExpress bus. As can be seen, the maximum threshold (green horizontal line) seems to be attainable when a reasonable number of events is processed in one batch.

In a standard GPU computing approach, data from the detector reach the Network Interface Card (NIC) which copies them periodically on a dedicated area in the PC RAM, from where they are then copied to the user space memory where applications can process them. Here a sufficient data load of buffered events is usually prepared for the following stages, and they are copied to GPU memory through the PCI express bus. The host (the PC on which the GPU card is plugged) has the role of starting the GPU kernel, which operates on the data. Computation results can then be sent back to the host for further processing or distribution to the detectors, to ultimately trigger the read-out of the complete data event. In this system, the most important contribution to the total latency is due to the data transfer latency from the NIC to the GPU memory. Thus, in order to reduce the maximum total latency, an approach will be described in detail in the following, i.e., the use of a direct data transfer protocol from a custom FPGA-based NIC to the GPU.

2.1 NaNet

NaNet is a modular design of a low-latency NIC with GPUdirect capability developed at INFN Rome division, that is being integrated in the GPU-based low level trigger of the NA62 RICH detector [9, 10]. Its design comes from the APEnet+ PCIe Gen 2 x8 3D NIC [11] and the board supports a configurable number of different physical I/O links (see Figure 3). The Distributed Network Processor (DNP) is the APEnet+ core logic, behaving as an off-loading engine for the computing node in performing inter-node communications [12]. NaNet is able to exploit the GPUDirect peer-to-peer (P2P) capabilities of NVIDIA Fermi/Kepler GPUs enabling a host PC to directly inject into its memory an UDP input data stream from the detector front-end, with rates compatible with the low latency real-time requirements of the trigger system.

To measure the data transmission latency a readout board (TEL62) has been employed to send input data. In particular, data communication between the TEL62 readout boards and the L0 trigger processor (L0TP) happens over multiple GbE links using UDP streams. The main requisite for the communication system comes from the request for <1 ms and deterministic response latency of the L0TP, thus communication latency and its fluctuations are to be kept under control. The requisite on bandwidth is $400 \div 700$ MB/s, depending on the final choice of the primitives data protocol which in turn depends on the amount of preprocessing to be actually implemented in the TEL62 FPGA. This means that, to extract primitives data from the readout board towards the L0TP in the final system, $4 \div 6$ GbE links will be used.

NaNet latency was measured sending UDP packets from one of the host GbE ports to the NaNet GbE interface: using the x86 TSC register as a common reference time, a single-process test could measure the time difference between the moment before the first UDP packet of a bunch (needed to fill the receive buffer) is piped out through the host GbE port and when the signal of a filled receive buffer reaches the application. Within this measurement setup ("system loopback"), the latency of the `send` process is also taken into account.

Measurements in Figure 5 were thus taken; UDP packets with a payload size of 1168 Bytes

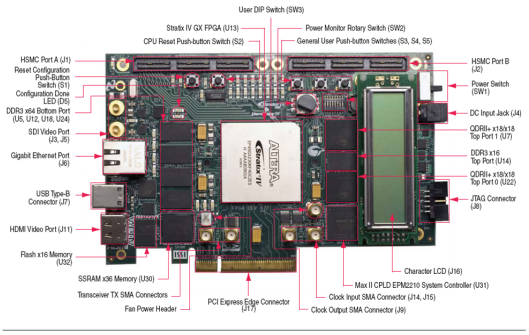

Figure 3: NaNet as a FPGA-based NIC implemented using an Altera development board (Stratix IVGX230 FPGA).

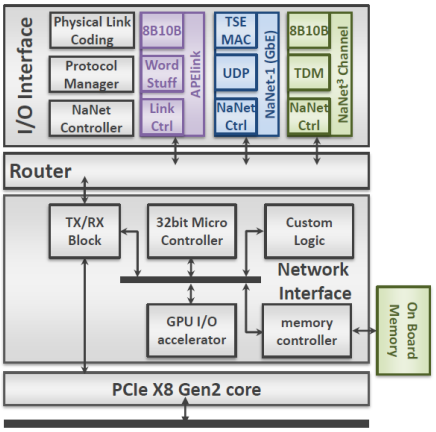

Figure 4: NaNet architecture schematic.

(16 events) were sent to a GPU memory receiving buffer of size variable between 1 and 64 times the UDP packet payload.

2.1.1 Synchronization of NaNet with the Timing Trigger and Control system at CERN

The Timing Trigger and Control (TTC) system distributes the system clock for the NA62 experiment as well as the first level triggers and synchronization commands, which are all distributed on a single optical fibre. In particular, a mezzanine card (TTCrq) [13], developed by the CERN microelectronics group, acts as an interface between the TTC system for NA62 detectors and its receiving end users. The card delivers the clock together with control and synchronization information to the front-end electronics controllers in the detector.

Synchronization of NaNet with the TTC system through the TTCrq, has been tested by firstly realizing an interface board between NaNet and the TTCrq. This has been produced at Physics Department at University of Ferrara (Italy) and connected to the High Speed Mezzanine Card (HSMC) port B of the FPGA board. The whole TTCrq-interface-NaNet system (see Figure 6) proved to correctly receive all the signals necessary to synchronise the detectors, i.e. clock, event counter reset and bunch counter reset (the last ones encoding start-of-burst and end-of-burst commands).

3 Conclusions

The GAP Project aims at studying the application of GPUs in real-time HEP trigger systems and in medical imaging. In this paper the application to a low-level trigger system for the NA62 experiment at CERN has been described: the critical point to be pursued is the reduction of

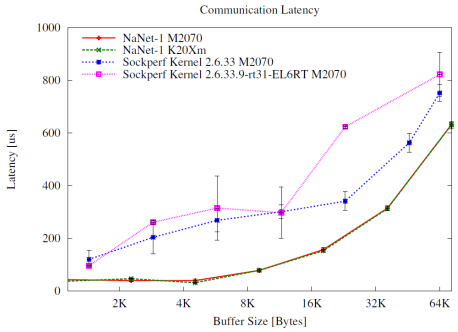

Figure 5: A comparison between NaNet communication latency and standard GbE interfaces results.

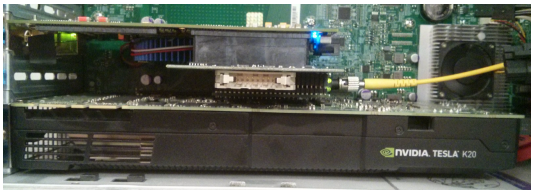

Figure 6: The TTCrq-interface-NaNet system together with a NVIDIA Tesla K20 GPU.

the contributions to the total latency due to data transfer from the detectors to the GPU. An approach based on a FPGA board to establish a peer-to-peer connection with the GPU is being developed. The reconstruction of photon rings in the RICH detector has been considered as first case of study on the use of GPUs at L0 in the NA62 experiment. Preliminary results show that current GPUs are suitable for sustaining the event rates, while at the same time minimizing the latency to an acceptable level.

Acknowledgment

The GAP project is partially supported by MIUR under grant RBFR12JF2Z “Futuro in ricerca 2012”.

References

[1] P.J. Clark, C. Jones. D Emeilyanov, M. Rovatsou, A. Washbrook, and the ATLAS collaboration, “Algorithm Acceleration from GPGPUs for the ATLAS Upgrade” Journal of Physics: Conference Series 331 (2011) 022031.

[2] D. Rohr, S. Gorbunov, A. Szostak, M. Kretz, T. Kollegger, T. Breitner and Torsten Alt, “ALICE HLT TPC Tracking of Pb-Pb Events on GPUs”, Journal of Physics: Conference Series 396 (2012) 012044.

[3] *http://web2.infn.it/gap*

[4] M. Fiorini [NA62 Collaboration], “The NA62 experiment at CERN”, PoS HQL **2012** (2012) 016.

[5] *http://cern.ch/NA62/*

[6] B. Angelucci, E. Pedreschi, M. Sozzi and F. Spinella, "TEL62: an integrated trigger and data acquisition board" Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE.

[7] B. Angelucci, G. Anzivino, C. Avanzini, C. Biino, A. Bizzeti, F. Bucci, A. Cassese and P. Cenci *et al.*, "Pion-muon separation with a RICH prototype for the NA62 experiment," Nucl. Instrum. Meth. A **621** (2010) 205.

[8] R. Ammendola, A. Biagioni, L. Deri, M. Fiorini, O. Frezza, G. Lamanna, F. Lo Cicero, A. Lonardo, A. Messina, M. Sozzi, F. Pantaleo, P.S. Paolucci, D. Rossetti, F. Simula, L. Tosoratto and P. Vicini, "GPUs for Real Time processing in HEP trigger systems" Journal of Physics: Conference Series 523 (2014) 012007 doi: 10.1088/1742-6596/523/1/012007 and references therein.

[9] R. Ammendola, A. Biagioni, O. Frezza, G. Lamanna, A. Lonardo, F. Lo Cicero, P. S. Paolucci, F. Pantaleo, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto and P. Vicini, "NaNet: a flexible and configurable low-latency NIC for real-time trigger systems based on GPUs" JINST **9** (2014) C02023, doi:10.1088/1748-0221/9/02/C02023.

[10] R. Ammendola, A. Biagioni, R. Fantechi, O. Frezza, G. Lamanna, F. L. Cicero, A. Lonardo, P. S. Paolucci, F. Pantaleo, R. Piandani, L. Pontisso, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto and P. Vicini, "NaNet:a low-latency NIC enabling GPU-based, real-time low level trigger systems", Journal of Physics: Conference Series, 513, (2014) 012018.

[11] R. Ammendola, A. Biagioni, O. Frezza, F. Lo Cicero, A. Lonardo, P. S. Paolucci, D. Rossetti and F. Simula, L. Tosoratto and P. Vicini, "APEnet+: A 3D Torus network optimized for GPU-based HPC systems", J. Phys. Conf. Ser. **396** (2012) 042059.

[12] A. Biagioni, F. L. Cicero, A. Lonardo, P. S. Paolucci, M. Perra, D. Rossetti, C. Sidore, F. Simula, L. Tosoratto and P. Vicini, "The Distributed Network Processor: a novel off-chip and on-chip interconnection network architecture", arXiv:1203.1536.

[13] *http://proj-qpll.web.cern.ch/proj-qpll/ttcrq.htm*