PROCEEDINGS OF SCIENCE

GPUs for Online Processing in Low-Level Trigger Systems

Stefano Chiozzi, Angelo Cotta Ramusino, Massimiliano Fiorini*†, Alberto Gianoli, Ilaria Neri
Università degli Studi di Ferrara and INFN Sezione di Ferrara, Italy

Riccardo Fantechi, Gianluca Lamanna, Roberto Piandani, Marco Sozzi
INFN Sezione di Pisa, Italy

Matteo Bauce, Andrea Biagioni, Ottorino Frezza, Alessandro Lonardo, Andrea Messina, Pier Stanislao Paolucci, Francesco Simula, Piero Vicini
INFN Sezione di Roma and Università di Roma "La Sapienza", Italy

Roberto Ammendola
INFN Sezione di Roma "Tor Vergata", Italy

Mauro Piccini, Cristiano Santoni
INFN Sezione di Perugia, Italy

We describe a pilot project for the use of GPUs (Graphics Processing Units) in online triggering applications for high energy physics experiments. General-purpose computing on GPUs is emerging as a new paradigm in several fields of science, although so far applications have been tailored to the specific strengths of such devices as accelerator in offline computation. With the steady reduction of GPU latencies, and the increase in link and memory throughput, the use of such devices for real-time applications in high-energy physics data acquisition and trigger systems is becoming ripe. We will discuss in details the use of online parallel computing on GPU for synchronous low level trigger systems. We will show the results of two solutions to reduce the data transmission latency: the first based on fast capture special driver and the second based on direct GPU communication using NaNet, a multi-standard, FPGA-based, low-latency, PCIe network interface card with GPUDirect capabilities. We will present preliminary results on a first field test in the CERN NA62 experiment. This study is done in the framework of GAP (GPU Application Project), a wider project intended to study the use of GPUs in real-time applications.

Technology and Instrumentation in Particle Physics 2014,
2-6 June, 2014
Amsterdam, the Netherlands

*Speaker.
†E-mail: fiorini@fe.infn.it

PoS(TIPP2014)208

1. Introduction

Nowadays, in order to overcome some shortcomings of the present microprocessor technology, the use of GPUs (Graphics processing units) for scientific computation is gaining ground in several fields of scientific research, e.g. hydrodynamics, molecular simulation and medical imaging among others. GPUs offer a parallel architecture with most of the chip resources devoted to computation, in contrast to a traditional processor (CPU) where a large fraction of the resources are used for other functions such as caching and handling of peripherals. For these reasons, GPUs can be used to achieve a large computing power using a limited amount of space and power [1].

Another important topic that may benefit from the usage of GPUs for scientific computation is real-time triggering in High Energy Physics (HEP). In fact, in HEP experiments the trigger system plays a central role because it must decide, generally based on limited information, whether a physical event observed in a detector should be recorded or not. Due to the fact that every experiment features a limited amount of data acquisition bandwidth and disk space for data storage, the use of real-time selection becomes fundamental to make the experiment affordable and at the same time preserve its discovery potential. Moreover, only uninteresting events would be rejected, thus selectively reducing the throughput of data. GPUs provide a huge computing power on a single device, allowing complex decisions to be made with a significantly high speed capable of matching valid event rates.

Online selection of significant events can be performed by arranging the trigger system in a cascaded set of computation levels. The first level (hereinafter referred as to low-level) is usually realised in hardware, often based on custom electronics devoted to the experiment and normally developed using programmable logic (FPGA) that offers flexibility and possibility of reconfiguration. The following (higher) trigger levels are commonly implemented in software, by using dedicated farms of commodity PCs. In this standard multi-level trigger architecture, GPUs can be easily exploited in the higher software level, where thanks to their computing power 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 (LVL2) of Atlas experiment at Cern [2] and are implemented with encouraging results in tracking of Pb-Pb Events in Alice experiment [3].

Low-level triggers can also benefit from the use of GPUs but a careful assessment of their online performances is required. In fact, low level trigger systems are designed to perform very rough selections based on a sub-set of the available information, 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 read-out electronics, such systems typically require very low latency. While in the applications for which GPUs have been originally developed a low total processing latency is not of paramount importance, in the case of low level triggers data transfer latency to the GPU and its stability in time become a very important issue.

Within the framework of the GAP (GPU Application Project) project [4], this paper presents recent results on the use of GPUs in real-time low-level triggering for HEP experiments.

2. The use of GPUs in low-level triggers: the NA62 physics case

The NA62 particle physics experiment at the CERN SPS aims at measuring the ultra rare decay $K^+ \to \pi^+ \nu\overline{\nu}$ as a highly sensitive test of the Standard Model (SM) and a search for New Physics [5]. This measurement is extremely challenging from the experimental point of view due to the large backgrounds coming from the other decay channels. Because of the large rate reduction required before recording events on tape, efficient online selection of candidate events represents a very important issue for this experiment.

The NA62 trigger consists of three levels: the first hardware level (L0) is based on FPGA boards which perform detector data readout [6], while the next two levels are implemented in software. L0 must handle an input event rate in 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 second trigger level (L1). L1 together with the following trigger level (L2) must reduce the rate to about 10 kHz in order to permit permanent data storage for later offline analysis. The maximum allowed L0 latency in NA62 is 1 ms. In the standard implementation of L0, trigger primitives contributing to the realisation of the final trigger decision are computed on the readout board FPGAs, and are mostly based on the event hit pattern. 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 RICH Cherenkov detector represents the first case of study on the use of GPUs at Level 0 trigger in the GAP project. Such detector, described in [7], provides a measurement of the velocity and direction of the charged particles crossing its volume. It therefore contributes 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 [1], several ring reconstruction algorithms have been studied in order to assess the best for a 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 first GPU in the list has a Fermi architecture and features 240 CUDA cores, 4 GB RAM and a PCI Express 2.0 connection; the second one has the same architecture but almost twice the number of cores (448), 3 GB RAM and PCIe 2.0; the third one has a Kepler architecture, 1536 cores, 2 GB memory and a PCIe 3.0 connection. The computing performance of the C2050 and GTX680 proved to be a factor 4 and 8 higher, respectively, than that of the C1060.

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. For both plots, NVIDIA Tesla C2050 and GeForce GTX680 devices have been used for measurements. The significant increase in data throughput and latency reduction for the

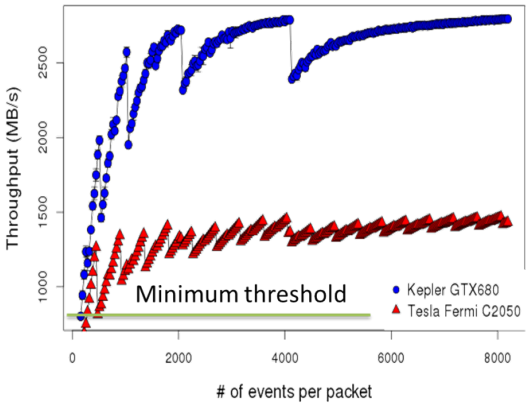

Figure 1: Throughput as a function of number of events for last generation GPUs.

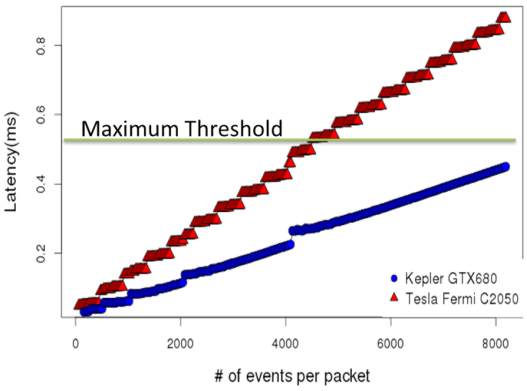

Figure 2: Total latency (including data transfer and computing). Here the maximum threshold does not take into account the data transfer from readout, but it only estimates the total latency within the GPU.

newer GTX680 GPU is due to the faster data transfer allowed by the 3^{rd} generation PCI Express 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. The discontinuous structure in both plots is due to the modularity of the buffer dimension used for data transfer.

In a standard approach for GPU computing, considering a system in which data transfer occurs through a standard ethernet network, data from the detector reaches 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 reading of the complete data. 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. Furthermore, another caveat of GPUs architecture is the need for saturation of computing cores, that requires a significant number of events and a buffering stage, this fact further weighing on trigger answer latency.

Two approaches will be described in detail in order to reduce the maximum total latency, i.e. the use of a dedicated NIC device driver with very low latency (PFRING) and a direct data transfer protocol from a custom FPGA-based NIC to the GPU (NaNet).

2.1 Driver DNA-PFRING

In order to allow the copying of packets directly from the FIFO of the NIC to the memory through Direct Memory Access (DMA), PFRING has been employed as a special socket that works in connection with the Direct NIC Access (DNA) driver. PFRING and DNA are both developed by

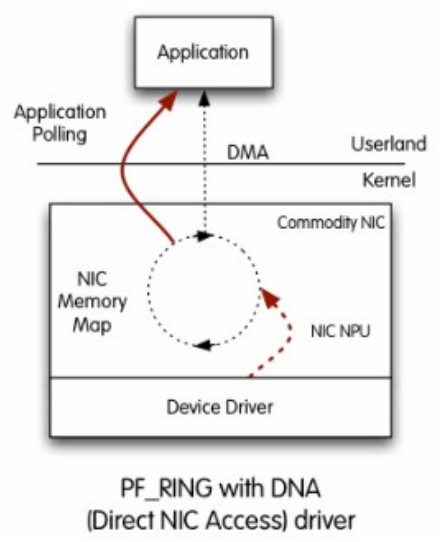

Figure 3: Scheme of PFRING data processing (courtesy of NTOP [8]).

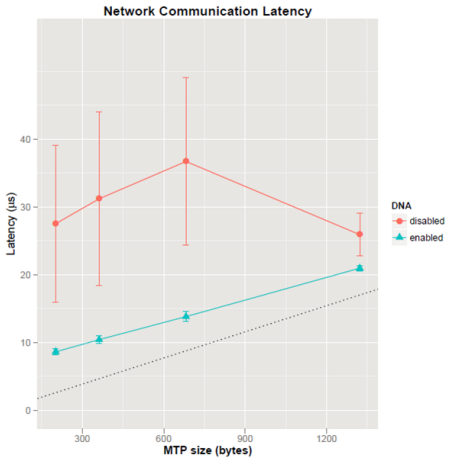

Figure 4: Comparison between the use of standard data transmission and the PFRING socket.

NTOP [8] and they represent a way to map NIC memory and registers to the userland so that there is no additional packet copy besides the DMA transfer done by the Network Process Unit (NPU) of the NIC. This results in better performance because CPU cycles are only used for consuming packets and not for moving them off the adapter. A schematic diagram of data processing through PFRING is shown in Figure 3: by using the DNA driver, data transfer is managed by the NPU processor, through circular buffers that are directly accessible for the applications, and therefore for the copy to the GPU's memory.

To test the performance of PFRING, a comparison between the use of standard data transmission and the PFRING socket has been done and results are shown in Figure 4, where the error bars represent the r.m.s. error for a large number of repeated measurements. The transfer time due to PFRING improves by more than a factor of 2 with respect to standard transmission, and fluctuations are reduced to a negligible level. For packet size close to the maximum MTP size, the prefetching in the standard vanilla driver allows to have a smaller latency.

2.2 NaNet

NaNet is a modular design of a low-latency NIC with GPUdirect capability developed at the INFN Sezione di Roma, 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 5). 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, equipping a hosting PC to directly inject into their 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 avoid

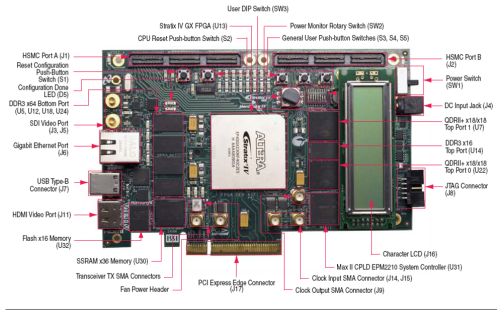

Figure 5: NaNet as an FPGA-based NIC implemented using an Altera development board (Stratix IV GX230 FPGA).

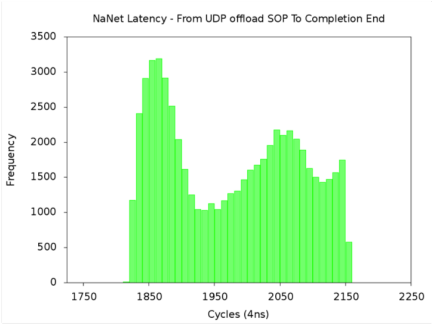

Figure 6: Distribution plot over 60000 samples of a NaNet packet traversal time.

jitter effects that usually affect system response time stability, NaNet has been partitioned so that the hosting CPU can be offloaded from any data communication or computing task, leaving to it only system confguration and GPU kernel launch tasks. This means that data communication tasks are entirely offloaded to a dedicated UDP protocol-handling block directly communicating with the P2P logic: this allows a direct data transfer with low and predictable latency on the data path among GbE link and GPU.

The UDP offload block comes from an open core module (NIOS II UDP) built for a Stratix II 2SGX90 development board. Focus of that design is the unburdening of the Nios II soft-core microprocessor onboard the Stratix II from UDP packet management duties by a module that collects data coming from the Avalon Streaming Interface (Avalon-ST) of the Altera Triple-Speed Ethernet Megacore (TSE MAC) and redirects UDP packets along a hardware processing data path. The Nios II subsystem executes the InterNiche TCP/IP stack to setup and tear down UDP packet streams which are processed in hardware at the maximum data rate achievable over the GbE network. Bringing the open core into the NaNet design required some modifications, first of all the hardware code was upgraded to work on the Stratix IV FPGA family; this upgrade made available the improved performances of an FPGA which is two technology steps ahead in respect to the Stratix II. The synthesis performed on a Stratix IV achieves the target frequency of 200 MHz (in the current APEnet+ implementation, the Nios II subsystem operates at the same frequency). Current NaNet implementation provides a single 32-bits wide channel; it achieves 6.4 Gbps at the present operating frequency, 6 times greater than what is required for a GbE channel. Data coming from the single channel of the modified UDP offload are collected by the NaNet CTRL. NaNet CTRL is a hardware module in charge of managing the GbE flow by encapsulating packets in the typical APEnet+ protocol (Header, Payload, Footer).

Latency inside the NIC was measured adding four cycles counters at different stages of packet processing; their values are stored in a profiling packet footer with a resolution of 4 ns; for a standard 1472 bytes UDP packet, traversal time ranges between 7.3 μs and 8.6 μs from input of NaNet CTRL to the completion signal of the DMA transaction on the PCIe bus (see Figure 6).

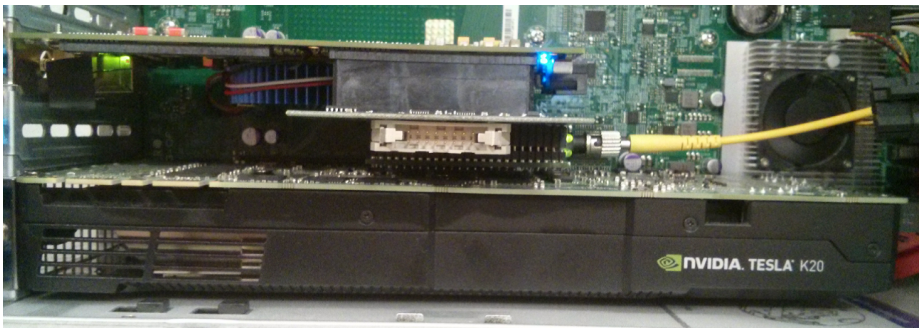

Figure 7: The TTCrq-interface-NaNet system together with a NVIDIA Tesla K20 GPU.

For the same packet size, saturation of the GbE channel is achieved, with 119.7 MB of sustained bandwidth.

2.2.1 Synchronisation 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 synchronisation commands, which are all distributed on a single optical fibre. In particular, a mezzanine card (TTCrq) [13], developed by the CERN Micro-Electronics 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 synchronisation information to the front-end electronics controllers in the detector.

Synchronisation of NaNet with the TTC system through the TTCrq has been achieved by realising an interface board between NaNet and the TTCrq, connected to the High Speed Mezzanine Card (HSMC) port B of the FPGA board. This interface card has been recently produced at Physics and Earth Science Department at the University of Ferrara and the whole system TTCrq-interface-NaNet (see Figure 7) proved to correctly receive all the signals necessary to synchronise the detectors.

3. Conclusions

The GAP Project has been recently funded to study the application of GPUs in real-time HEP trigger systems, for both low and high level trigger systems, and in medical imaging. In this paper, the low-level trigger case has been presented. Two approaches are being pursued: the first employs a special driver that allows direct copy of the data from the NIC buffers avoiding redundant copies; the second one foresees to use a FPGA-based board to establish a peer-to-peer connection with the GPU. The NA62 experiment at CERN represents a possible application of this technique, in particular in the reconstruction of photon rings in the RICH detector. Preliminary results show that Cherenkov rings pattern recognition within the total L0 latency of 1 ms seems possible with current GPUs.

Acknowledgment

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

References

[1] 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.

[2] 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.

[3] 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.

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

[5] M. Fiorini [on behalf of the NA62 Collaboration], "The NA62 experiment at CERN", PoS HQL **2012** (2012) 016.

[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] *http://www.ntop.org*

[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*