Journal of Physics: Conference Series

GPUs for real-time processing in HEP trigger systems

To cite this article: G Lamanna *et al* 2014 *J. Phys.: Conf. Ser.* **513** 012017

View the article online for updates and enhancements.

GPUs for real-time processing in HEP trigger systems

G.Lamanna
INFN, Pisa, Italy
E-mail: gianluca.lamanna@cern.ch

R.Ammendola,[0], M.Bauce[1,2], A.Biagioni[2], R.Fantechi[3,6], M.Fiorini[4], S.Giagu[1,2], E.Graverini[5,6], G.Lamanna[6], A.Lonardo[2], A. Messina[1,2,3], F.Pantaleo[5,6], P.S.Paolucci[2], R.Piandani[6] , M.Rescigno[2], F.Simula[2], M.Sozzi[5,6] and P.Vicini[2].

[0] INFN, Rome "Tor Vergata", Italy
[1] University, Rome "Sapienza", Italy
[2] INFN, Rome "Sapienza", Italy
[3] CERN, Geneve, Switzerland
[4] University & INFN, Ferrara, Italy
[5] University, Pisa, Italy
[6] INFN, Pisa, Italy

Abstract. We describe a pilot project for the use of Graphics Processing Units (GPUs) for online triggering applications in High Energy Physics (HEP) experiments. Two major trends can be identified in the development of trigger and DAQ systems for HEP experiments: the massive use of general-purpose commodity systems such as commercial multicore PC farms for data acquisition, and the reduction of trigger levels implemented in hardware, towards a pure software selection system (trigger-less). The very innovative approach presented here aims at exploiting the parallel computing power of commercial GPUs to perform fast computations in software both at low- and high-level trigger stages. 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 throughputs, the use of such devices for real-time applications in high-energy physics data acquisition and trigger systems is becoming very attractive. We discuss in details the use of online parallel computing on GPUs for synchronous low-level trigger with fixed latency. In particular we show preliminary results on a first test in the NA62 experiment at CERN. The use of GPUs in high-level triggers is also considered, the ATLAS experiment (and in particular the muon trigger) at CERN will be taken as a study case of possible applications.

1. Introduction

The trigger system of any HEP experiment has a crucial role deciding, based on limited and partial information, whether a particular event observed in a detector is interesting and must be recorded or not. Every experiment is characterised by a limited amount of DAQ bandwidth and disk space for storage, therefore the use of real-time selections to reduce data throughput selectively, rejecting uninteresting events only, is fundamental to make an experiment affordable and at the same time maintaining its discovery potential.

This paper describes the idea to use GPUs for triggering in HEP experiments. This work is included in a project wider in scope, named GAP (GPU Application Project) concerning the use of GPUs for advanced scientific computation in real-time application. The Project covers both applications in HEP for event selection and in medical imaging (CT, PET and NMR), however this paper focuses only on recent results on real-time triggering in HEP.

2. GPU in low-level trigger

Graphic processors represent a viable alternative to fill the gap between a multi-level trigger system with a hardware lowest level and a system which does not require any real-time hardware processing on reduced event information. Indeed GPUs do provide a large raw computing power on a single device, thus allowing to take complex decisions with a speed which can match significant event rates.

In a standard multi-level trigger architecture GPUs can be easily exploited in the higher software levels: being powerful computing devices, they can boost the capabilities of the processing system, thus allowing more complex computations to be performed without increasing the scale of the system itself. This is the case of the use of GPUs in the software trigger level (LVL2) of the CERN ATLAS experiment, as discussed later on.

The use of GPUs in lowest trigger levels, on the other hand, requires a careful assessment of their real-time performances. A low total processing latency and its stability in time (on the scales of "hard" real-time) are indeed requirements which are not of paramount importance in the applications for which GPUs have been originally developed. As mentioned above, the issue is related to the fact that in common usage of GPUs as graphics co-processors in computers, data is transferred to the GPU - and results are transferred back - through the PCI-express computer bus. Moreover, in order to better exploit the parallel GPU architecture, computing cores must be saturated, thus requiring the computation on a significant number of events after a buffering stage.

To address the problem of the latency we follow two ways: the first is based on the use of a special high performance driver to increase the speed of data trasmission from the NIC (Network Interface Card) buffers to the RAM, the second employs a custom "smart" NIC to allow the copy of data directly inside the GPU.

2.1. PFRING

Standard drivers do not guarantee the highest packet capture and transmission speed, especially with small-size packets and high rates. PFRING is a new kind of socket that works in connection with the DNA (Direct NIC Access) driver (both developed by NTOP [1]) in order to allow direct copy of packets from the NIC's FIFO to the memory through DMA (Direct Memory Access). Based on PFRING, the scheduler managing data transfer from the NIC to the GPU's memory has been implemented in a partially preemptive way in parallel streams. This allows to "hide" data transfer latency by exploiting concurrent copy-execution enabled in last generation GPUs. To measure the transmission latency we used a readout board (TEL62) to send input data and a PC equiped with a standard Intel I350T2 NIC and a NVDIA Tesla K20 GPU as receiver. The latency is measured with an oscilloscope, by using the "send of packet" in the TEL62 as "start" and the "end of computing" in the GPU (through the PC LTP port) as "stop". Figure 1 shows the comparison between data transfer based on standard intel driver (in red) and on PFRING (in blue) as a function of the size of the received packet (MTP - Multi Trigger Packet). The transfer time improves by more than a factor of 2 and, most importantly, the fluctuations are reduced to a negligible level. Figure 2 shows the processing time per event as a function of the number of buffered events (GMTP - GPU Multi Trigger Packet) including data transfer from

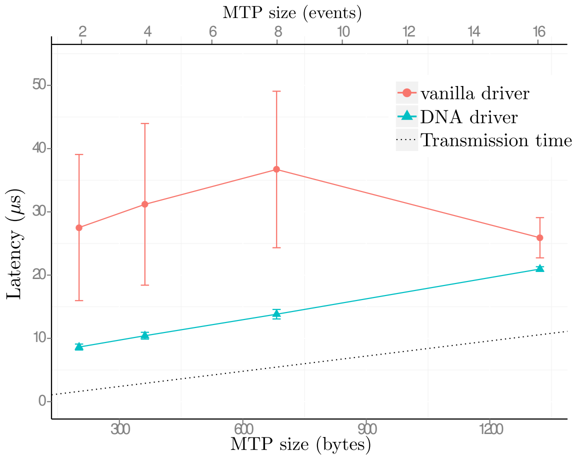

Figure 1. Comparison between standard data transmission (red) and by using PFRING socket (blue).

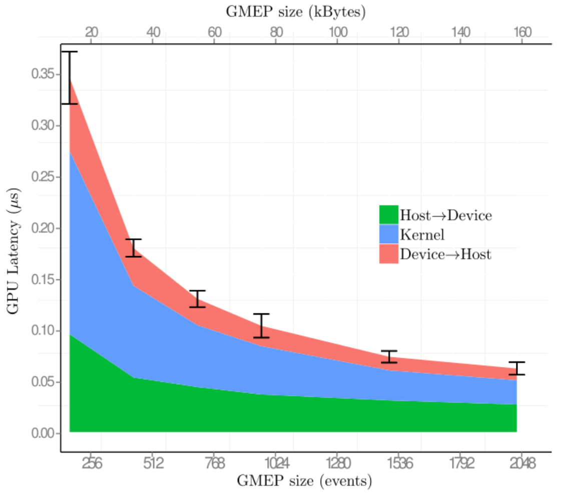

Figure 2. Total time (transfer+processing) per event.

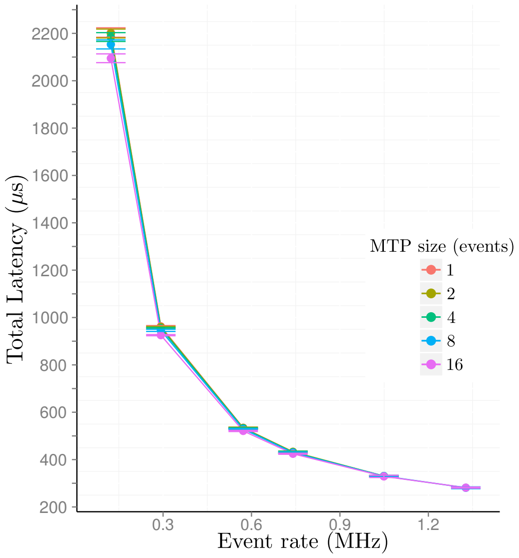

Figure 3. Total latency (including transfer times and computing) for a buffer of 256 events.

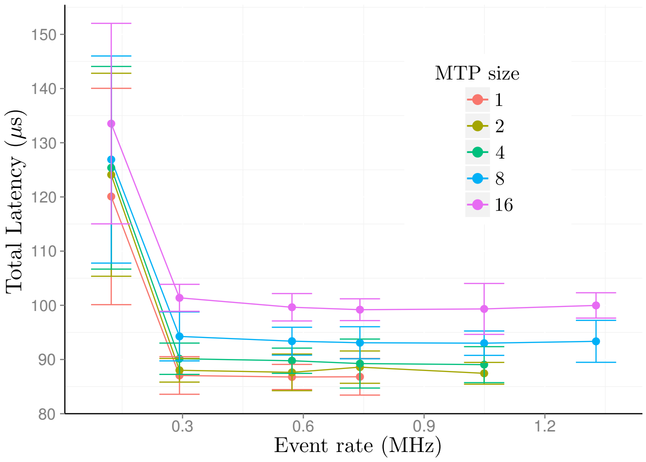

Figure 4. Total latency of the system. Gathering time excluded.

host to device, computing time for a simple kernel (described below), and copy of results from device to host. Since the processing time is relatively small, the latency depends on the gathering time needed to buffer events (shown in Fig. 3). For this plot the start signal corresponds to the first event of the GMTP buffer, while the end is produced when all the events in the buffer are processed. Figure 4 shows the component of the latency due to the GPU system only, by excluding the gathering time (considering as "start" the recording of the last packet of the GMTP buffer).

2.2. NaNet

The NaNet board [3] is a customized version of the APEnet+ [2] NIC designed to be integrated in the GPU-based low level trigger system of the NA62 RICH detector (Section 2.3). The communication tasks are entirely offloaded to a dedicated UDP protocol-handling block directly communicating with the P2P logic: this allows direct data transfer (no data coalescing or staging is performed on NIC and/or PC) with low and predictable latency on the GbE link → GPU data path. A wide discussion on NaNet implementation and tests can be found in [4].

2.3. NA62

As a first use case, we studied the possibility of reconstructing, in GPUs, the ring-shaped hit patterns in a RICH Cerenkov detector of the NA62 experiment [5]. Such detector, described in [6], can provide a measurement of the velocity and direction of charged particles (such as muons and pions) which traverse it, thus contributing to the computation of other physical quantities such as the decay vertex of the K^+ and the missing mass. The use of such information allows to implement highly selective trigger algorithms also for other interesting decay modes.

As described in [7] the “math” algorithm, based on a simple coordinate transformation of the hits which reduces the problem to a least square procedure, was found to be the best one in terms of computing throughput (for single rings). This algorithm was implemented and tested on different GPUs, such as the NVIDIA Tesla C1060, Tesla C2050 and GeForce GTX680 (in increasing order of processing core generation). The computing performance of the C2050 and GTX680 proved to be a factor 4 and 8 higher than that of the C1060. In figure 5 we show 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 in one go; the horizontal line shows the requirement related for an online trigger based on the RICH detector in NA62.

Figure 6 shows instead (for NVIDIA Tesla C2050 and GeForce GTX680 devices) the total latency, which includes data transfer times to and from the GPU and the kernel execution time. The significant reduction of the latency for the newer GTX680 GPU is due to the faster data transfer due to the presence of the gen.3 PCI express bus. Also in this case the maximum latency allowed by the NA62 application is seen to be attainable when a reasonable number of events is processed in one batch.

3. GPU in High-Level trigger

HLT systems, in particular those of LHC experiments, offer a complementary environment with respect to the one discussed so far for NA62 for rate, bandwidth, and latency. Typical values are latencies from ms up to seconds, input rates of few hundred KHz, and bandwidths of hundreds of Gb/s. HLT systems are nowadays implemented as customized software algorithms executed on farms of commodity PCs. The LHC upgrade with the consequent increase of luminosity and pile-up, poses new challenges for the HLT systems in terms of rates, bandwidth and signal selectivity. To exploit more complex algorithms aimed at better performances, higher computing capabilities and new strategies are required. Moreover, given the tendency of the computing industry to move away from the current CPU model towards architectures with high numbers of small cores well suited for vectorial computation, it is becoming urgent to investigate the possibility to implement higher level of parallelism in the HLT software.

The GAP project is studying the deployment of GPUs for the HLT in LHC experiments, using as a study case the ATLAS muon HLT. The ATLAS trigger system is organized in 3 levels [8]. The first-level trigger is built on custom electronics, while the second-level (LVL2) and the event-filter are implemented in software algorithms executed by a farm of about 1600 PCs with different Xeon processors each with 8 to 12 cores. Currently, a first upgrade is foreseen in 2018 [9], when real-time tacking capabilities will also be available, followed by a complete renovation

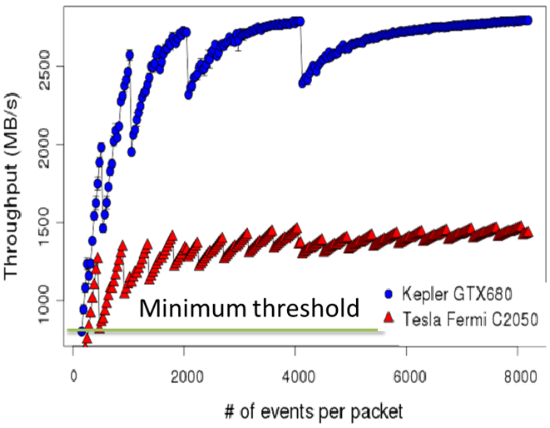

Figure 5. Throughput as a function of number of events for last generation GPUs.

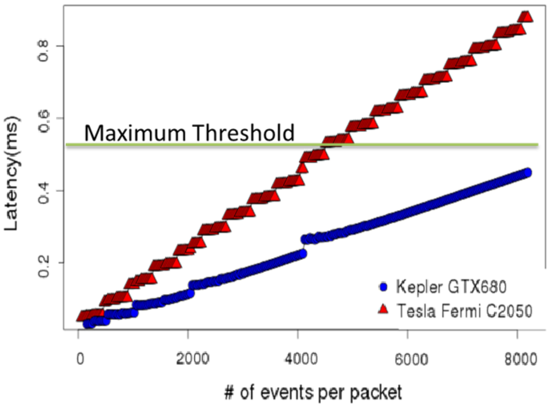

Figure 6. Total latency (including data transfer and computing).

of the trigger and detector systems in 2022. We intend to explore the potential improvements attainable in the near future by deploying GPUs in the ATLAS LVL2 muon trigger algorithms. Such algorithms are now implemented as approximated solutions of complex primitives; the high computing capabilities of GPUs would allow the use of refined algorithms with higher selection efficiency, and thus to maintain the sensitivity to interesting physics signals even at higher luminosity.

Acknowledgments

The GAP project is partially supported by MIUR under grant RBFR12JF2Z “Futuro in ricerca 2012”. This work is partially supported by the EU Framework Programme 7 project EURETILE under grant number 247846.

References

[1] *http://www.ntop.org*
[2] R. Ammendola, A. Biagioni, O. Frezza, F. Lo Cicero, A. Lonardo, P. S. Paolucci, D. Rossetti and F. Simula *et al.*, J. Phys. Conf. Ser. **396** (2012) 042059.
[3] A. Lonardo and others,”Building a Low-Latency Real-time GPU-based stream processing system”, *http://on-demand.gputechconf.com/gtc/2013/presentations/S3286-Low-La tency-RT-Stream-Processing-System.pdf*
[4] A. Lonardo, “NaNet: a low-latency NIC enabling GPU-based, real-time low level trigger systems.” at the same conference (CHEP2013).
[5] G. Lamanna, J. Phys. Conf. Ser. **335** (2011) 012071.
[6] B. Angelucci, G. Anzivino, C. Avanzini, C. Biino, A. Bizzeti, F. Bucci, A. Cassese and P. Cenci *et al.*, Nucl. Instrum. Meth. A **621** (2010) 205.
[7] G. Collazuol, G. Lamanna, J. Pinzino and M. S. Sozzi, Nucl. Instrum. Meth. A **662** (2012) 49.
[8] *The Atlas Collaboration*, JINST **3** (2008) P08003.
[9] *The Atlas Collaboration*, -CERN-LHCC-2011-012 (2012).