GPU-Based Low-Level Trigger System for Real-Time Cherenkov Ring Fitting

Ilaria Neri
University of Ferrara and INFN, Italy
On behalf of the GAP Collaboration

San Diego November 2, 2015
Outline

- GAP (GPUs Application) Project
- Why GPUs?
- Low Level Trigger system
- Case study: NA62 RICH detector at CERN
- Status of GPU trigger
GAP Project

- GAP (GPU Application Project) for real-time in HEP and medical imaging is a 3 year project funded by the Italian Ministry of research, started in 2013.

- Collaboration between INFN Sezione di Pisa, University of Ferrara and University of Roma “La Sapienza”.

- Demonstrate the feasibility of using off-the-shelf computer commodities to accelerate real-time scientific computation.

- Application in different fields:
  - High Energy Physics (low and high level triggers).
  - Medical Imaging (NMR, CT and PET).
Trigger in High Energy Physics Experiments

- Limited amount of data acquisition bandwidth and disk space for data storage
  
  **Real-time selection** becomes fundamental to make the experiment affordable.

- **Online selection** of significant events is performed by arranging the trigger system in a cascaded set of computation level.

- A standard **multi-level trigger system** identifies the events to be saved exploiting dedicated hardware and flexible software in standard computers.

- The **lowest level (L0)** trigger is an essential component to perform very rough selections based on a sub-set of the available information. **Low latency** is required.
GPUs in Low Level Trigger System

- GPUs provide a huge computing power on a single device, thus allowing to take complex decisions with a significantly high speed capable to match valid event rates.

Two main issues

- **Latency**: Is the GPU “latency” per event small enough to cope with the requirements of a low level trigger system? Is the latency stable enough for synchronous trigger systems?
- **Computing power**: Is the GPU fast enough to take trigger decision at $10^7$ event/s?
NA62 case study

- **Kaon decays in flight**
  - High intensity unseparated hadron beam (6% kaons).
- **Huge background** from kaon decays
  - \( \sim 10^8 \) background wrt signal.
  - Good kinematics reconstruction.

**750 MHz beam rate**

**10 MHz rate from Kaon decays**

\[ K^+ \rightarrow \pi^+ \nu \bar{\nu} \text{ decay (BR } \sim 8 \times 10^{-11} \) ]
Trigger and DAQ

- Multi-level trigger system.
- L0: Hardware synchronous level. 10 MHz to 1 MHz. Max latency: 1 ms.
Primitives from RICH, MUV, LAV and LKR.

See poster N2AP-19 by E. Gamberini
RICH detector

Ring Imaging CHERenkov detector:
- Distinguish between pions and muons from 15 to 35 GeV.
- 2 spots of 1000 photomultipliers each.
- 70 ps time resolution.
- 10 MHz event rate.
- About 20 photons detected per single ring event (hits per particle).
- 40 Bytes per event.

Talk M. Pepe “Performances of the NA62 RICH Detector”
The aim: GPU in L0 RICH

- 4 TEL62 ReadOut boards for RICH detector
  - 8x1Gb/s links for data
  - 4x1Gb/s trigger primitives
  - 4x1Gb/s GPU trigger
  - L0 trigger rate (from L0TP): 1 MHz
  - Max Latency: 1 ms
Latency: main problem of GPU computing

- Decrease the data transfer time:
  - DMA (Direct Memory Access) to the memory of the GPU

- Avoid the increase of latency dominated by double copy in Host RAM
NaNet board

• NaNet: board based on the ApeNet+ (HPC project) card logic
  • PCIe interface with GPU Direct P2P/RDMA capability
  • Offload of network protocol
  • Multiple link support
  • Use FPGA resources to perform on-the-fly data preparation

A. Lonardo et al. 2015, JINST, 10 C04011

See also poster N2AP27 by M. Martinelli
The Chain

**RICH TEL62 board**

**NaNet**
- **GPURICH FW**: prepare “events” clustering hits. 2 sources: RICHJ1 & RICHJ2
- **NaNet**: receive data and drive direct DMA to GPU

**GPU**
- **GPU**: Apply the online algorithm and prepare trigger decision

**L0TP**
- **L0TP**: receive GPU results and merge it with other primitives
Test Setup

NaNet @NA62 experiment
Results

• About 1us per event for multi-ring.
• Test on single C2070 board.

OLD: 30 us trigger execution time

NEW: 1 us trigger execution time

Data transfer: TEL62 - NaNet - GPU
Almagest: multi-ring identification

• New algorithm (Almagest) based on Ptolemy’s theorem: “A quadrilateral is cyclic (the vertex lie on a circle) if and only if is valid the relation: \( AD \cdot BC + AB \cdot DC = AC \cdot BD \).”

• Select a triplet and check if all the other points lie on the same ring by checking the Ptolemy’s theorem.

• Design a procedure for parallel implementation.
Ring fitting

- Good quality of online ring fitting.
- Preliminary: efficiency greater than 80% (depends on the number of rings).
- Resolution, efficiency, comparison between algorithms (on one and multi-rings) is ongoing.
Conclusions and future plans

- The use of GPUs in HEP low level trigger systems could give several advantages provided that processing performances and latencies are carefully studied.
- Realization of a L0 trigger processor for the NA62 RICH is under way.
  - Cherenkov rings pattern recognition within the total L0 latency of 1 ms seems possible.
- Integration with the NA62 Trigger and DAQ system has been done.
  - First test was done during NA62 experimental run in summer 2015.
- NaNet10 will be implemented on the Altera StratixV dev. board which features 10 GbE channel to collect data from 4 TEL62 boards.
- Improve the kernel part by choosing the best computing algorithm.