NaNet: a flexible and configurable low-latency NIC for real-time trigger systems based on GPU

R. Ammendola[a], A. Biagioni[b], O. Frezza[b], G. Lamanna[c], F. Lo Cicero[b], A. Lonardo[b], F.Pantaleo[c], P.S. Paolucci[b], D.Rossetti[b], A. F. Simula[b], L. Tosoratto[b], M. Sozzi[c], P. Vicini[b]

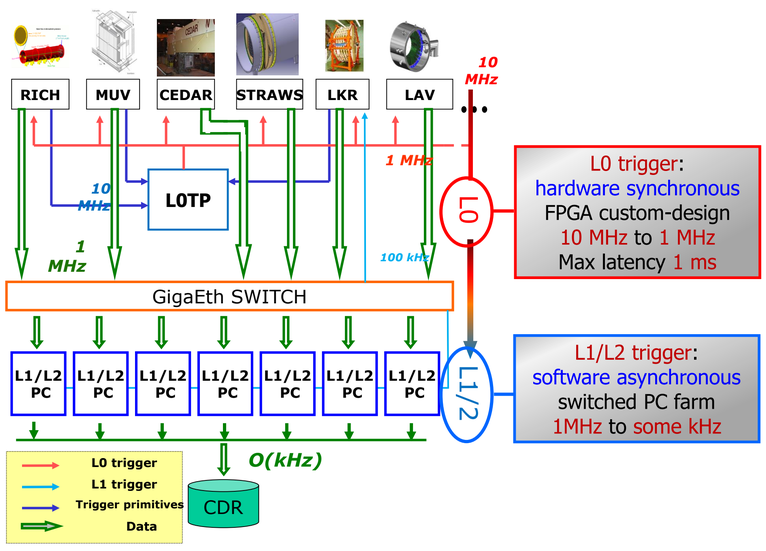
(a) INFN Sezione di Roma Tor Vergata (b) INFN Sezione di Roma (c) INFN Sezione di Pisa

TWEPP-13 Topical Workshop on Electronics for Particle Physics

Perugia, Italy, 23-27 September 2013

ABSTRACT –The adoption of GPUs in the low level trigger systems is currently being investigated in several HEP experiments. While GPUs show a deterministic behaviour in performing computational tasks, data communication is the main source of fluctuations in the response time of such systems. We designed NaNet, a FPGA-based NIC supporting 1/10GbE links and the custom 34 Gbps APElink channel. The design has GPUDirect RDMA capabilities, i.e. is able to inject the input data stream directly into the Fermi/Kepler class GPU(s) memory, and features a network stack protocol offloading engine. We will provide a detailed description of the NaNet hardware modular architecture and a comparative performance analysis on the NA62 RICH detector GPU-based L0 trigger case study using the NaNet board and a commodity GbE NIC. Figures of merit for the system when using the APElink and 10GbE links will also be provided.

The NA62 Trigger and Data Acquisition System



L0 trigger: hardware synchronous, FPGA custom-design, 10 MHz to 1 MHz, Max latency 1 ms

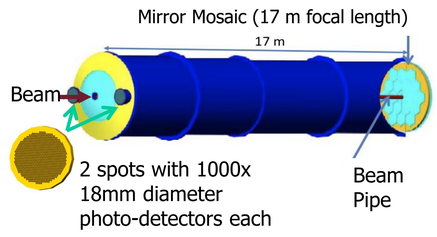
L1/L2 trigger: software asynchronous, switched PC farm, 1MHz to some kHz

GPUs in the NA62 L0 trigger

Replace custom hardware with a GPU-based system performing the same task but:

- ❑ Programmable
- ❑ Upgradable
- ❑ Scalable
- ❑ Cost effective
- ❑ Increasing selection efficiency of interesting events implementing more demanding algorithms.

The RICH Detector Case Study



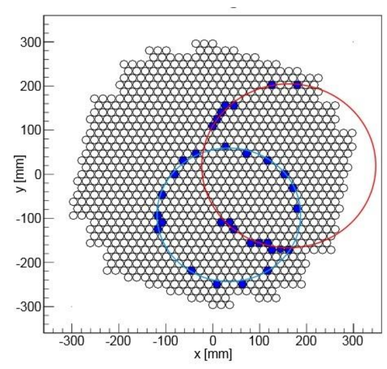
Ring-imaging Čerenkov detector: charged relativistic particles produce cone-shaped light in Neon-filled vessel.

- ❑ 100 ps time resolution
- ❑ **10 MHz** event rate
- ❑ 20 photons detected on average per single ring event (**hits** on photo-detectors)
- ❑ **70 byte** per event (max 32 hits per event)

Rings pattern recognition and fit performed on GPU:

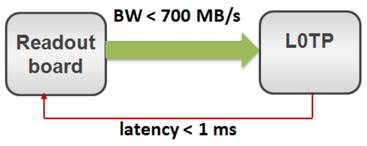
- 50 ns for single ring
- 1 us for multiple ring

New algorithm ("Almagest") developed for trackless, fast, and high resolution ring fitting.

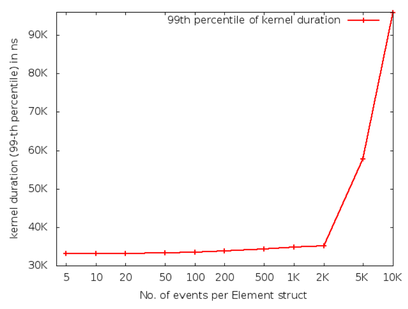


NA62 RICH L0 Trigger Proc Requirements

- ❑ Network Protocols/Topology: UDP over Point-to-Point (no switches) GbE.
- ❑ Throughput
  - Input event primitive data rate < 700MB/s (on 7 GbE links)
  - Output of trigger results < 50 MB/s (on 1 GbE link)
- ❑ System response latency < 1 ms
  - determined by the size of Readout Board memory buffer storing event data candidated to be passed to higher trigger levels.

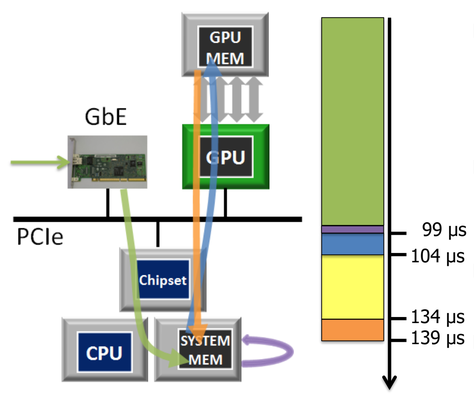


Processor - Processing Latency



- ❑ **lat$_{proc}$** : time needed to perform rings pattern-matching on the GPU with input and output data on device memory.
- ❑ 10K events = 70 kB
- ❑ **lat$_{proc}$** is stable
- ❑ max 1/10 of the time budget available

Processor – Communication Latency



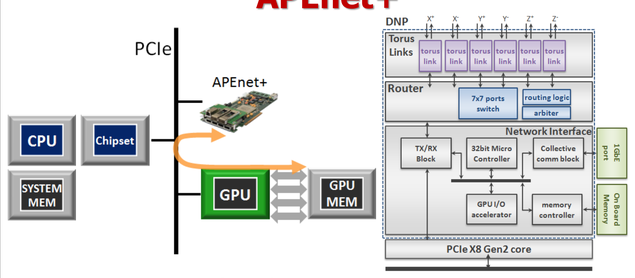
- ❑ **lat$_{comm}$** =110 µs avg (4 x **lat$_{proc}$**)
- ❑ Fluctuations on the GbE component of **lat$_{comm}$** may hinder the real-time requisite, even at low events count: **Min 60 µs, Max 650 µs!**

- ❑ **lat$_{comm}$** : time needed to receive input event data from GbE NIC to GPU memory and to send back results from GPU memory to Host memory.
- ❑ 20 events data (1404 byte) sent from Readout board to the GbE NIC are stored in a receiving host kernel buffer.
- ❑ Data are copied from kernel buffer to a user space buffer
- ❑ Data are copied from system memory to GPU memory
- ❑ Ring pattern-matching GPU Kernel is executed, results are stored in device memory.
- ❑ Results are copied from GPU memory to system memory (322 bytes – 20 results)

NaNet

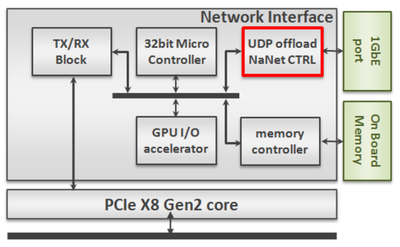
- ❑ Problem: lower communication latency and its fluctuations.
- ❑ Solution:
  - ❑ Injecting directly data from the NIC into the GPU memory with no intermediate buffering, re-using the APEnet+ GPUDirect RDMA implementation.
  - ❑ Adding a network stack protocol management offloading engine to the logic (UDP Offloading Engine) to avoid OS jitter effects.

APEnet+

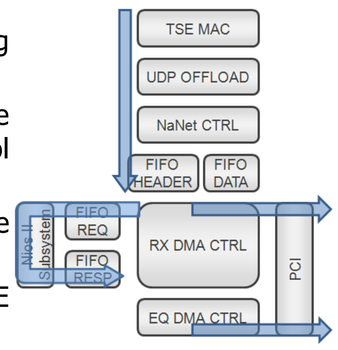


- ❑ First non-Nvidia device supporting GPUDirect RDMA (2012).
- ❑ No bounce buffers on host. APEnet+ can target GPU memory with no CPU involvement.
- ❑ GPUDirect allows direct data exchange on the PCIe bus between NIC and GPU, using P2P protocol.
- ❑ Latency reduction for small messages.

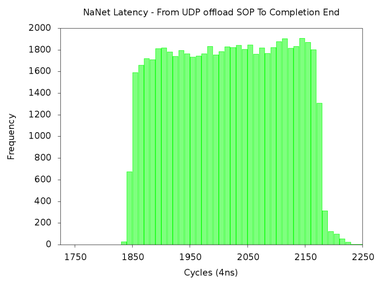
NaNet Architecture and Data Flow

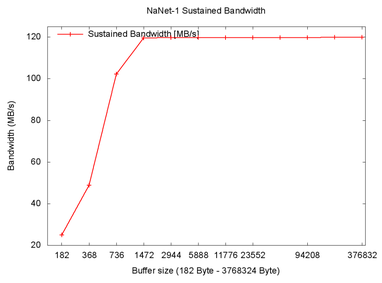


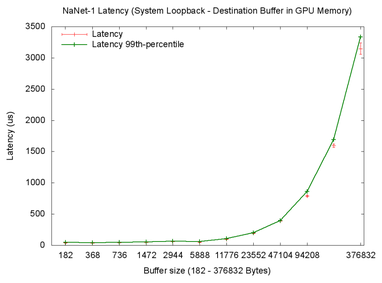
- ❑ APEnet+ Firmware Customization
- ❑ UDP offload collects data coming from the Altera Triple-Speed Ethernet Megacore (TSE MAC) and redirects UDP packets into an hardware processing data path.
- ❑ NaNet Controller encapsulates the UDP payload in a newly forged APEnet+ packet and send it to the RX Network Interface logic.
- ❑ RX DMA CTRL manages CPU/GPU memory write process, providing hw support for the Remote Direct Memory Access (RDMA) protocol.
- ❑ Nios II handles all the details pertaining to buffers registered by the application to implement a zero-copy approach of the RDMA protocol (OUT of the data stream).
- ❑ EQ DMA CTRL generates a DMA write transfer to communicate the completion of the CPU/GPU memory write process.
- ❑ A Performance Counter is used to analyze the latency of the GbE data flow inside the NIC.

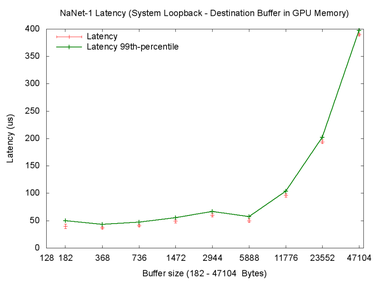


NaNet Benchmark



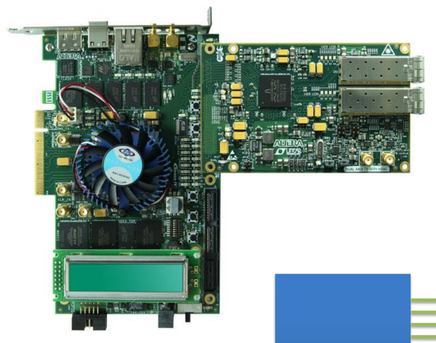






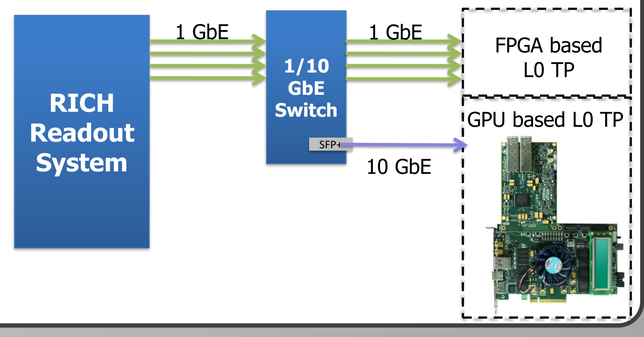
- ❑ Latency of a 1472 bytes payload UDP Packet through the NIC hardware path is quite stable (7.3 µs ÷ 8.6 µs).
- ❑ Sustained Bandwidth ~119.7 MB/s.
- ❑ In the 1 GbE link L0 GPU-based Trigger Processor prototype the sweet spot between latency and throughput is in the region of 70-100 Kb of event data buffer size, corresponding to 1000-1500 events.

Future Work



NaNet-10 (dual 10 GbE)

- ❑ Implemented on the Altera Stratix IV dev board + Terasic HSMC Dual XAUI to SFP+ daughtercard.
- ❑ BROADCOM BCM8727 a dual-channel 10-GbE SFI-to-XAUI transceiver.



References

http://on-demand.gputechconf.com/gtc/2013/presentations/S3286-Low-Latency-RT-Stream-Processing-System.pdf

http://apegate.roma1.infn.it/mediawiki/index.php/Main_Page

http://euretile.roma1.infn.it/mediawiki/index.php/Main_Page

http://na62.web.cern.ch/na62/

Nucl.Instrum.Meth.A662:49-54,2012

Contacts

alessandro.lonardo@roma1.infn.it

andrea.biagioni@roma1.infn.it

piero.vicini@roma1.infn.it

gianluca.lamanna@cern.ch

This project was partially funded by the Euretile european FP7 grant 247846.