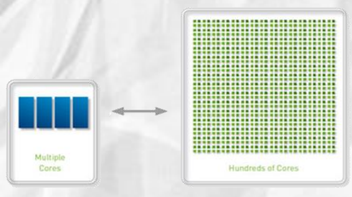
GPU for Real Time processing in HEP trigger systems

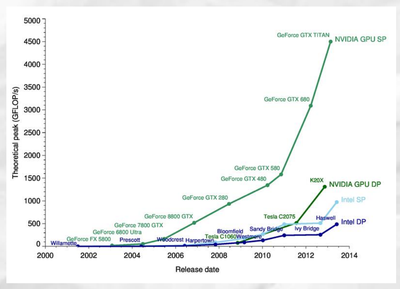
R.Ammendola [1], M.Bauce [2], A.Biagioni [3], S.Chiozzi [9], R.Fantechi [4], M.Fiorini [5], S.Giagu [2], A.Gianoli [9], G.Lamanna [8], A.Lonardo [3], A. Messina [6], F.Pantaleo [7], R.Piandani [8], M.Rescigno [3], F.Simula [3], M.Sozzi [7], P.Vicini [3]

(1) INFN sez.Roma-TorVergata, (2) University of Rome and INFN, (3) INFN sez. Roma-Sapienza, (4) INFN sez.Pisa and CERN, (5) University of Ferrara and INFN, (6) University of Rome and CERN, (7) University of Pisa and INFN, (8) INFN sez.Pisa, (9) INFN sez. Ferrara

The GPU

- The use of **GPU** (Graphics processing unit) for High Performance Computing is rising in the last years.
- The main differences between **GPU** and **CPU** are due to the different resources dedicated to computing and to the parallel architecture.
- Nowadays a single GPU can deliver more than **3 TFLOPS**.
- Vectorizable algorithms could benefit from the GPU computing power.
- GPUs are to be intended as a co-processor: the data must be brought on the video card using the PCI Express bus.





Video Card	N. Cores	Processing Power (GFLOPS)	Bandwidth (GB/s)
NVIDIA TESLA C1060 (2009)	240	933	102.4
NVIDIA TESLA C2050 (2011)	448	1288	144
NVIDIA GTX680 (2012)	1536	3090	192
NVIDIA TESLA K20 (2013)	2496	3520	208
AMD RADEON 5870 (2010)	1600	2720	153.6
AMD RADEON 7970 (2012)	2048	4096	288

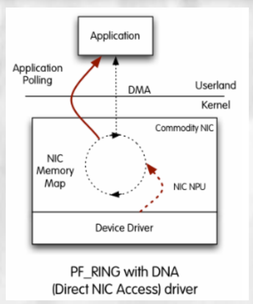
GAP

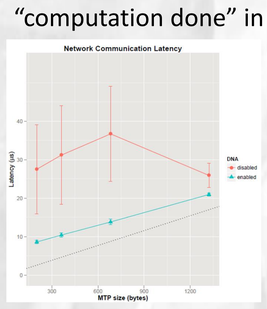
- The **GAP project** (GPU application for physics) aims at studying the use of **GPU** in real-time application.
- The main fields of study are trigger in High Energy Physics experiments and image reconstruction for medical purposes (PET, TAC and NMR).
- Three research units: INFN (Pisa and Rome Apenet+ Group), University of Ferrara and University of Rome.

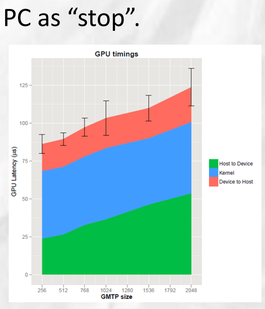
- More info: **http://web2.infn.it/gap**

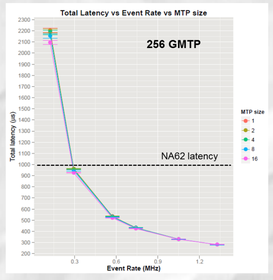
PFRING

- **PFRING-DNA** is a special socket–driver, developed by **NTOP** *(http://www.ntop.org)* that allow to directly copy data from **NIC's FIFO** to the user memory.
- A prototype system with a readout board (**TEL62**) and a PC equipped with an **NVIDIA TESLA K20** has been used to measure latency and computing time.
- The latency has been measured with an oscilloscope by using the start of the packet in the **TEL62** as "start" and the "computation done" in the PC as "stop".





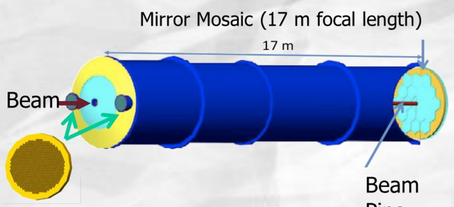




- **PFRING-DNA** allows to reduce the latency by a factor 3 and the fluctuations to a negligible level .
- The total latency is given as a function of the number of events to buffer before the start of the **GPU** computation.
- For real application the *"working point"* depends on the events rate and event dimension: for real applications the total latency (transfer time through ethernet+transfer to GPU+computing) is in the order of 100/200 us.

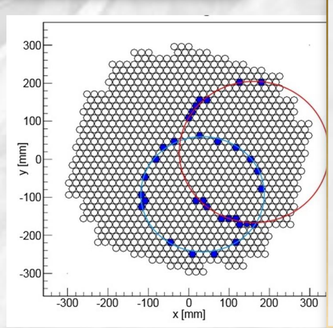
NA62 physics case

[see M.Fiorini – 12/9/2014 12:15]



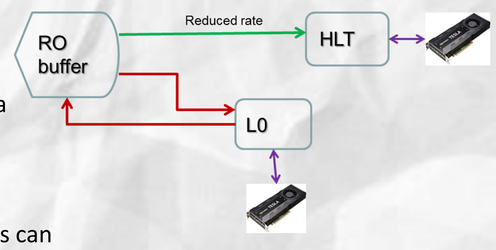
- The **NA62 RICH** is a Neon (1 atm) Cherenkov detector to distinguish between pions and muons in the **15-35 GeV** range.
- The identification of rings at low trigger level is useful to build very selective trigger conditions.

- Requirements:
  - **Fast**
  - **Offline quality**
  - **Trackless**
  - **Multi-rings**
  - **Noise tolerance**
- New parallel algorithm (called "Almagest") in two steps:
  - Pattern recognition based on parallel application of "Ptolemy's Theorem"
  - Fitting with Tobin algorithm
- Preliminary results very encouraging:
  - ~50 ns in single ring events.
  - ~1 us in multi rings events.

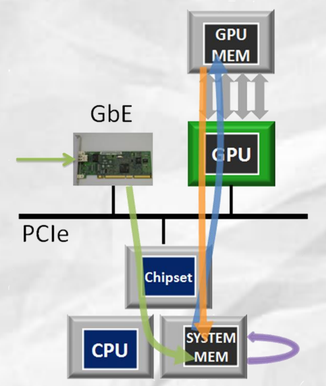


GPU in the trigger

- The **GPU** can be employed to build high selective triggers when the limit of the standard approach is due to the computing power.
- Usually the triggers are subdivided in "levels" in order to online select the interesting events:
  - The **lower level** is realized with custom electronics to apply a fast and rough selection, with low latency.
  - The **High Level Trigger (HLT)** is realized in Software allowing more precise decisions in a longer time.
- In **HLT** the use of **GPU** is quite straightforward: the video processors can increase the computing power of the online farm reducing the total number of PCs where the trigger algorithms run.
- The use of GPU in low level trigger is more difficult: the total time to have an answer from a system based on **PC+GPU** and standard links (i.e. Ethernet) is given both from the computing time and the time to bring the data in the GPU.
- For a system based on Ethernet most of the time is spent on data transfer instead of computing on **GPU** (example: packet of 1404 B with an algorithm to reconstruct ring in a Cherenkov counter, see below).



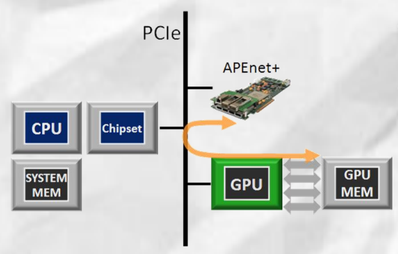
- In order to reduce the data transport latency we are investigating two ways:
  - **PFRING-DNA** Driver: a driver for fast packet capture.
  - **NANET**: direct transport of data from the **NIC** to the **GPU** without **CPU** involvement.
- In both cases a part of the latency, since we are interested in real-time systems, the fluctuations of the latency must be studied and reduced.



NaNet

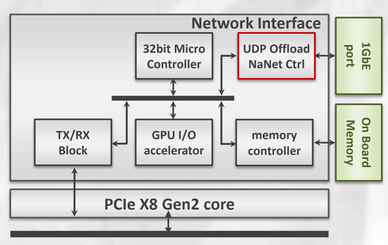
[see A.Lonardo – 12/9/2014 12:45]

An **FPGA** based PCIe 8x gen 2 board derived from the **APEnet+ 3D NIC** design, able to directly inject an UDP data stream from an external **GbE** source into the memory of a Fermi- or Kepler-class NVIDIA GPU exploiting GPUDirect RDMA capabilities.

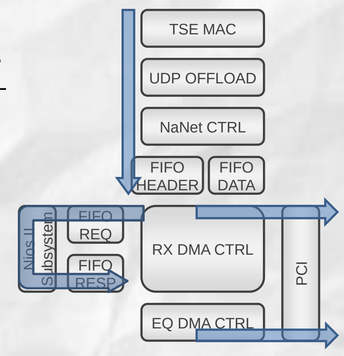


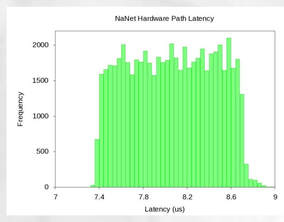
- PCIe P2P protocol between Nvidia Fermi/Kepler devices and APEnet+
- First non-NVIDIA device (2012) supporting **GPUDirect RDMA**:
  - No bounce buffers on host. APEnet+ can target GPU memory with no CPU involvement
  - **Latency reduction for small messages**

Project	Board	Comb. ALUT	Register	Memory [MB]
NaNet-1	EP4SGX230KF40C2	54635 (30%)	54415 (30%)	1.00 (55%)

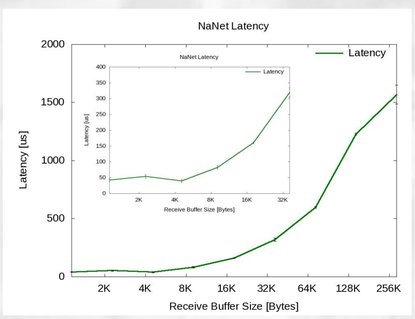


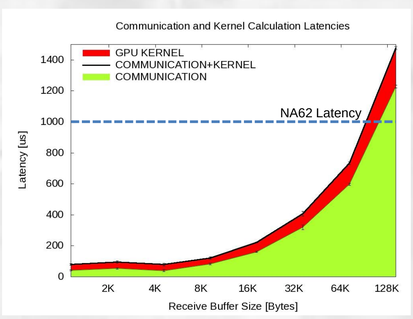
- **NaNet**: reduce comm. latency and its fluctuations to meet real-time constraints
- UDP offload collects data coming from the Altera Triple-Speed Ethernet Megacore (TSE MAC) and redirects UDP packets into a hardware processing data path.
- NaNet Controller encapsulates the UDP payload in a newly forged APEnet+ packet and sends it to the RX Network Interface logic
- RX DMA CTRL manages CPU/GPU memory write process
- 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.

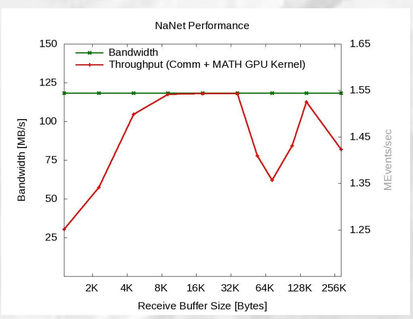




Latency of a 1472 bytes payload UDP Packet through the NIC hardware path is quite stable (7.3 µs ÷ 8.6 µs).







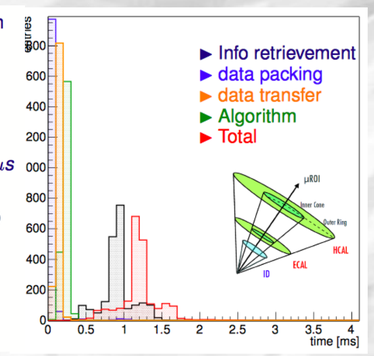
The "natural" place: GPU in HLT

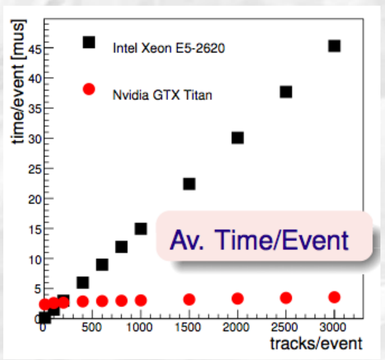
[see M.Bauce – 10/9/2014 9:45]

1. Retrieve calorimeter cell information
2. Pack data for transfer
3. Send them to APE server
4. Perform algorithm
5. Send back the output to the main trigger framework
6. Total time: ~1.2 ms

$\Delta t \sim 250\ \mu s$

Overhead well within time budget!





- GPU in High Level triggers is the most "natural" place
- Algorithm parallelization is mandatory in next high luminosity hep experiments.
- L2 muon isolation algorithm in ATLAS could benefit from sped-up in computing time.

Conclusions

- The **GPU** is a specialized processor designed for fast images handling.
- The parallel **GPU's** architecture can be exploited for general purpose computation.
- The use of the **GPUs** for online applications, such as the trigger in **HEP** experiments, is very challenging.
- We are investigating two different ways to allow the use of **GPUs** in real-time: **PFRING and NANET**.
- The preliminary results on prototypes designed for the **NA62** experiment are very encouraging.
- We are investigating the use of **GPUs** in the ATLAS experiment.

Acknowledgements

The GAP project is supported by MIUR under grant RBFR12JF2Z ``Futuro in ricerca 2012''.
The NaNet project is partially supported by the EU Framework Programme 7 project EURETILE under grant number 247846.

Other GAP talks

[see G.Di Domenico – 12/9/2014 15:00]

[see M.Palombo – 12/9/2014 15:45]