“Il progetto GAP: stato, prospettive & company”

Preventivi CSN1, Pisa, 26.6.2014

Gianluca Lamanna (INFN)

GAP at a glance

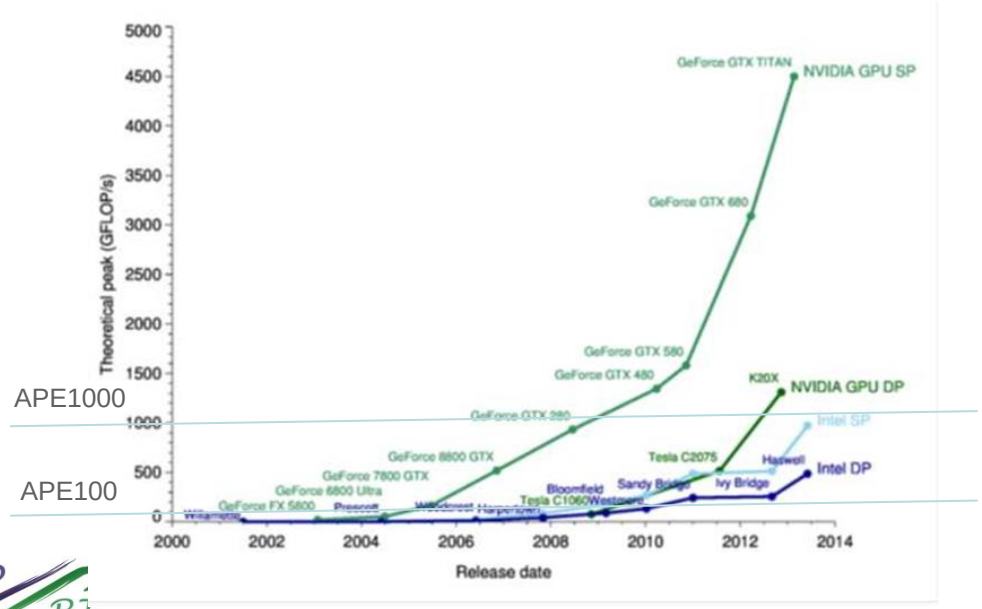
PI – responsabile unità INFN (Pisa):
gianluca.lamanna@pi.infn.it
Responsabile unità di Ferrara:
massimiliano.fiorini@fe.infn.it
Responsabile unità di Roma:
andrea.messina@roma1.infn.it

Web site:
http://web2.infn.it/gap

"The aim of the GAP project is the deployment of Graphic Processing Units (GPU) in real-time applications, ranging from high-energy physics online event selection (trigger) to medical imaging reconstruction. The final goal of the project is to demonstrate that GPUs can have a positive impact in sectors different for rate, bandwidth, and computational intensity.
Most crucial aspects currently under study are the analysis of the total latency of the system, computational algorithms optimisation, and integration with the data acquisition systems."

- 3 anni: termina Q1 2016
- Costo totale: 836620
- Annualità di persone prese per il progetto ad oggi (100% GAP): 12
- Giovani: 2 tesisti (laurea) + 1 summer student

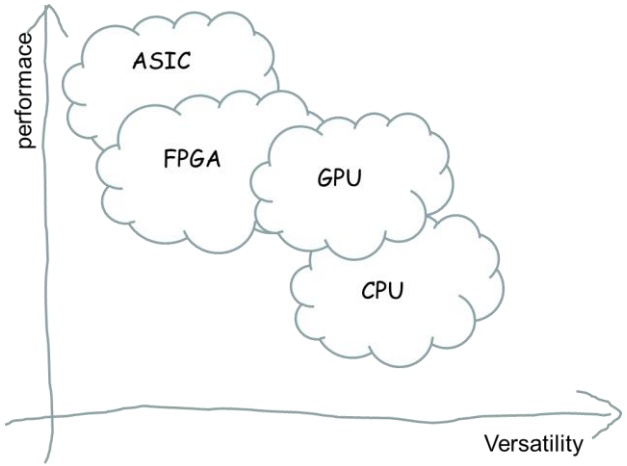
Perchè le GPU?



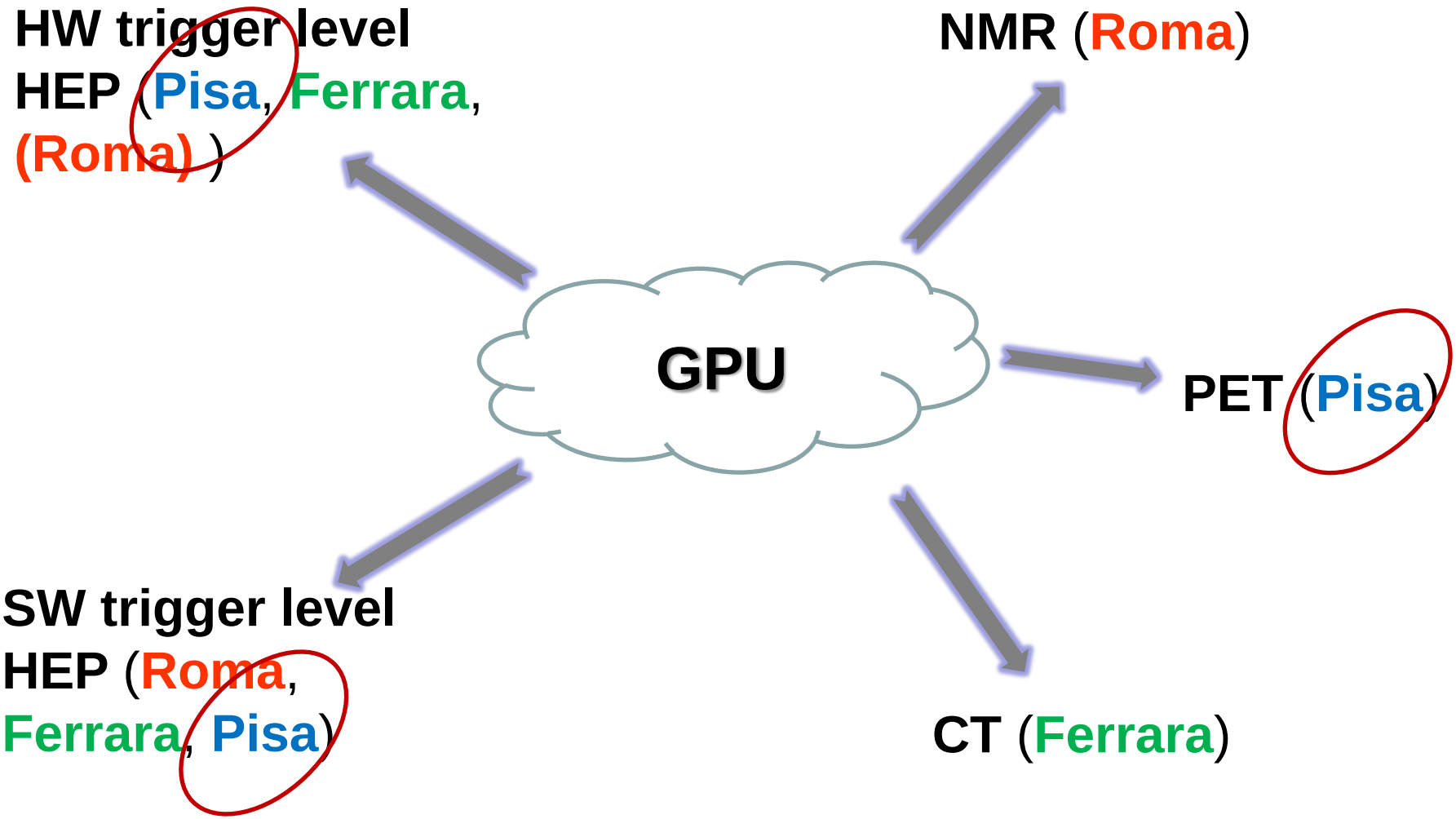
- Dal 1999 (anno di introduzione della GeForce 256) le GPU hanno rivoluzionato il mondo della computer grafica.
- La struttura parallela dei processori grafici può essere impiegata per calcolo parallelo general purpose (GPGPU).
- Le GPU sono impiegate nei principali supercomputer per calcolo scientifico.

Perchè le GPU in HEP?

- L'High Performance Computing è fondamentale in diverse applicazioni HEP (Montecarlo, Analisi, Offline, Online, …).
- In diverse applicazioni l'uso di acceleratori dà un vantaggio in termini di costo e tempo.
- La possibilità di applicare algoritmi dell'offline direttamente online potrebbe dare vantaggi in termini di potenziale di scoperta.
- "Sì, vabbè: ma perchè le GPU e non qualcosa di dedicato?": GAP vuole capire se con HW general purpose sviluppato per altri scopi si può fare quello che *sicuramente* si potrebbe fare costruendo HW dedicato.



GAP goals

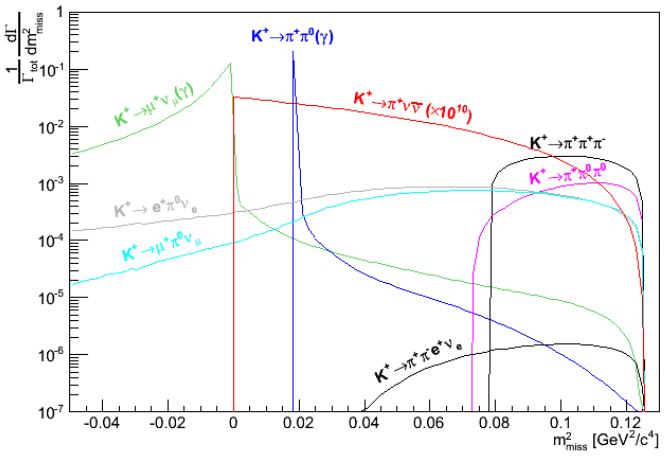

HW trigger level
HEP (Pisa, Ferrara, (Roma))
NMR (Roma)
GPU
PET (Pisa)
SW trigger level
HEP (Roma, Ferrara, Pisa)
CT (Ferrara)

Due problemi nell'usare le GPU a L0

- Potenza di calcolo: Le GPU sono abbastanza veloci per prendere decisioni di trigger su un rate di eventi di decine di MHz ?

- Latenza: La latenza per evento delle GPU è piccola abbastanza per un basso livello di trigger? La latenza è sufficientemente stabile per un sistema sincrono ?

Physics case: NA62



- Kaon decays in flight
  - High intensity unseparated hadron beam (6% kaons).
  - Event by event K momentum measurement.

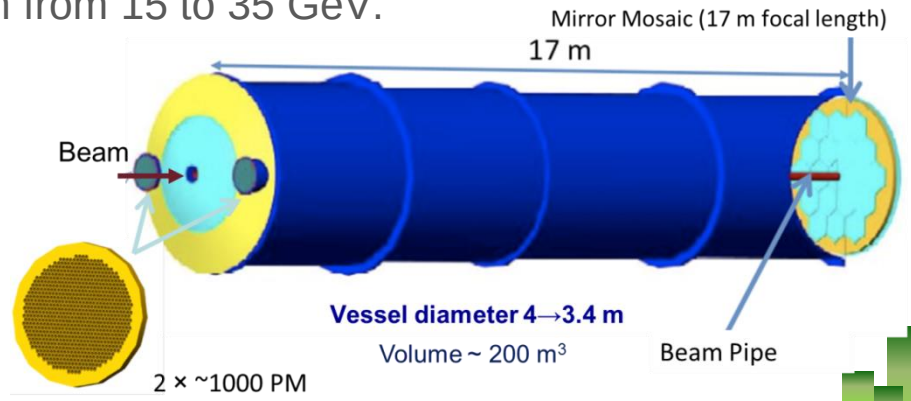
Huge background from kaon decays

- ~10^8 background wrt signal
- Good kinematics reconstruction.
- Efficient veto and PID system for not kinematically constrained background.

Trigger/DAQ:

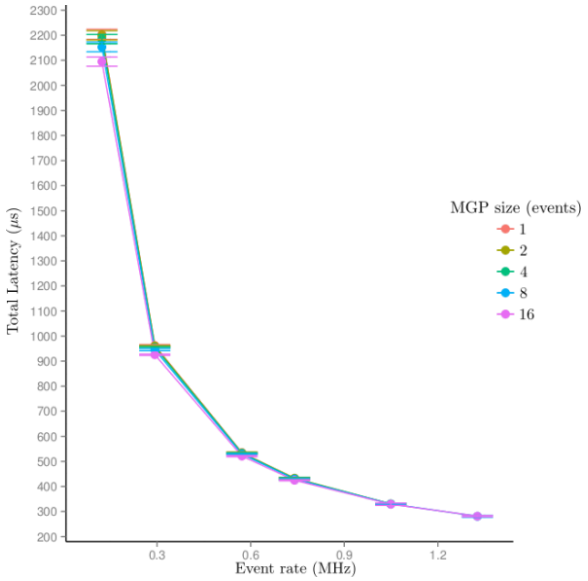
- Three levels
- L0 max latency: 1ms
- 10 MHz events on main detectors

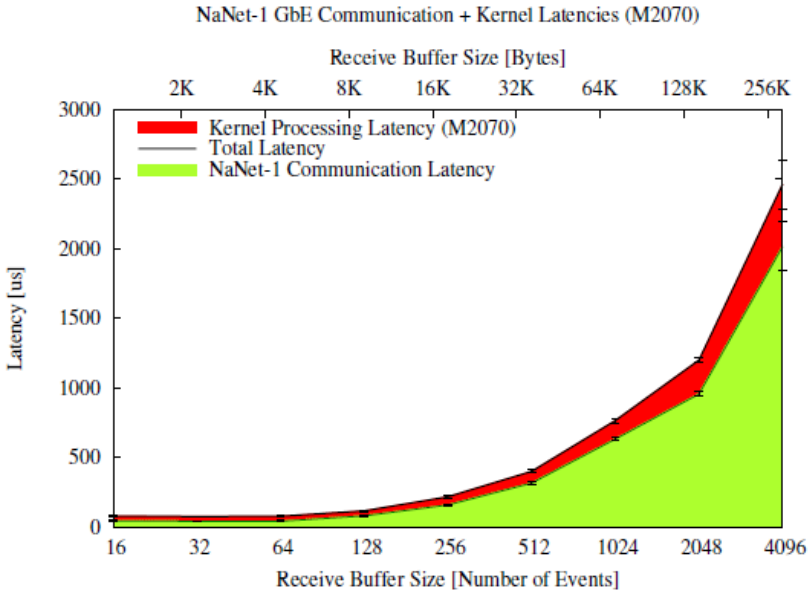
- RICH:
  - 17 m long, 3 m in diameter, filled with Ne at 1 atm
  - Distinguish between pions and muon from 15 to 35 GeV.
  - 2 spots of 1000 PMs each
  - Time resolution: 70 ps
  - MisID: 5×10^{-3}
  - 10 MHz events: about 20 hits per particle



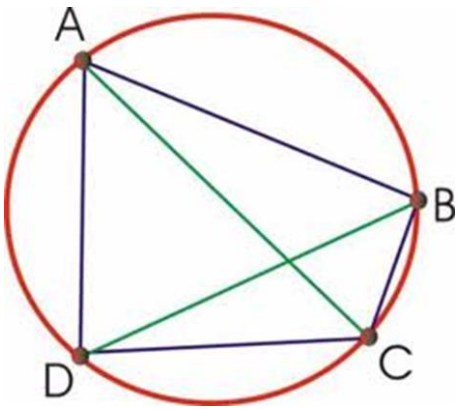
Caratterizzazione della latenza

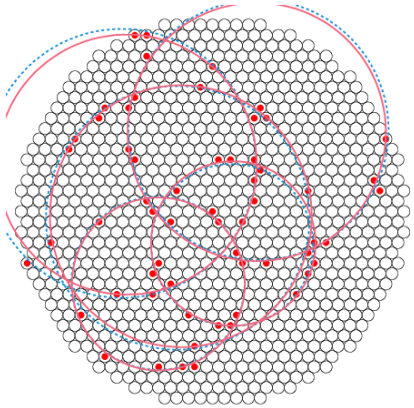
- 2 soluzioni:
  - PFRING
  - NANET
- Ambedue le soluzioni sono competitive
  - Latenza paragonabile (O(200 us) per 256 eventi)
  - Pro e contro



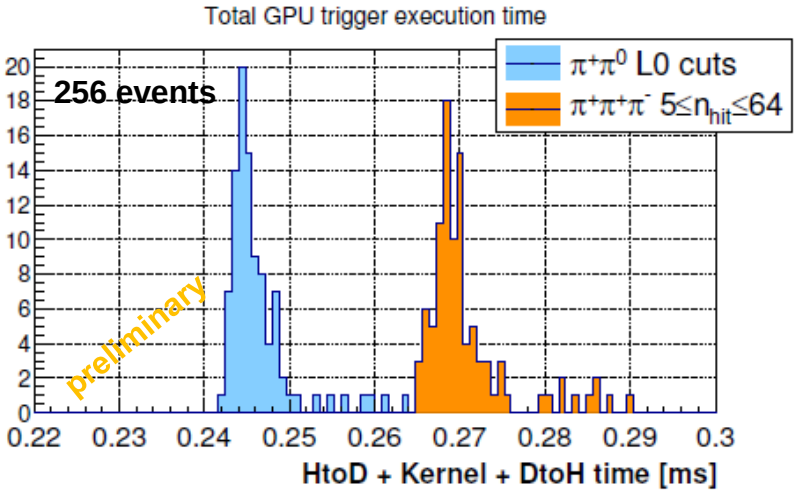


Algoritmi





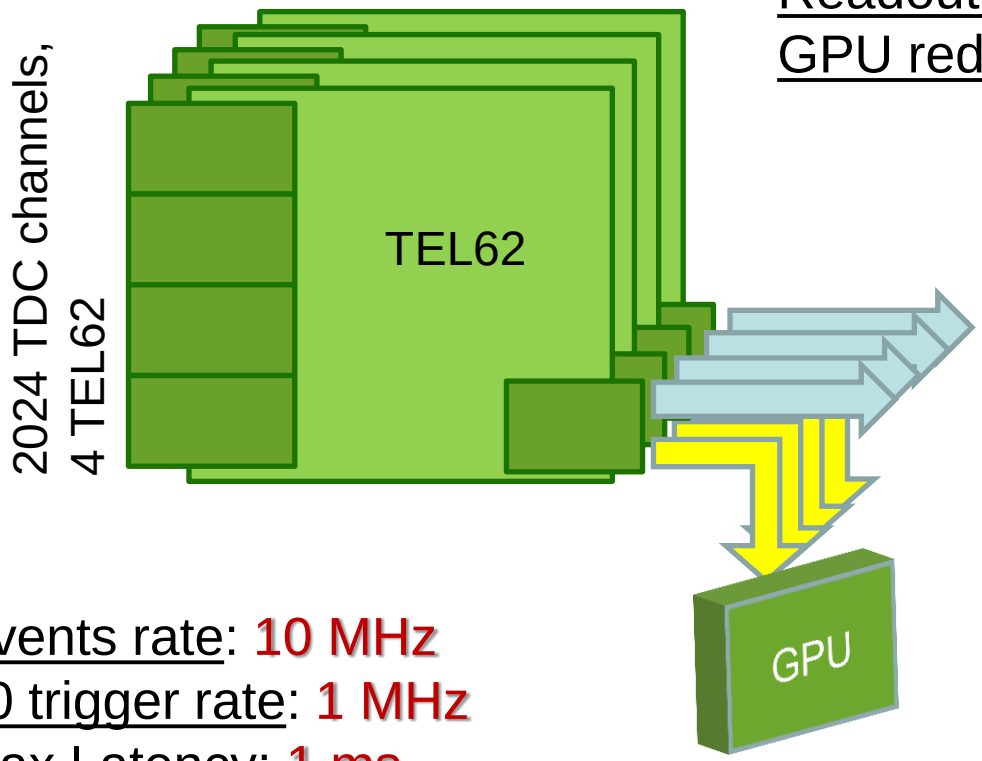
- Fit geometrico basato su geometria elementare
- Idoneo alla parallelizzazione
- Versione da ottimizzare ha già buone prestazioni



GAP-PET

- Studi preliminair per identificare gli algoritmi idonei:
  - MLEM (Maximum Likelihood expectation maximization)
  - Punti in comune con la ricostruzione CT(retro-projectors)
  - Indicazione indiretta di miglioramenti rilevanti, essenziale in nuovi scanner con alta risoluzione (aumento di LOR)

NA62 GPU trigger system



Readout event: 1.5 kb (1.5 Gb/s)
GPU reduced event: 300 b (3 Gb/s)

8x1Gb/s links for data readout
4x1Gb/s Standard trigger primitives
4x1Gb/s GPU trigger

Events rate: 10 MHz
L0 trigger rate: 1 MHz
Max Latency: 1 ms
Total buffering (per board): 8 GB
Max output bandwidth (per board): 4 Gb/s

GPU NVIDIA TITAN:

- 2688 cores
- 4.5 Teraflops
- 6GB VRAM
- PCI ex.gen3
- Bandwidth: 288 GB/s

Prospettive

- Test di integrazione con il readout di NA62: 18 agosto
- Parassiticamente nella presa dati di ottobre

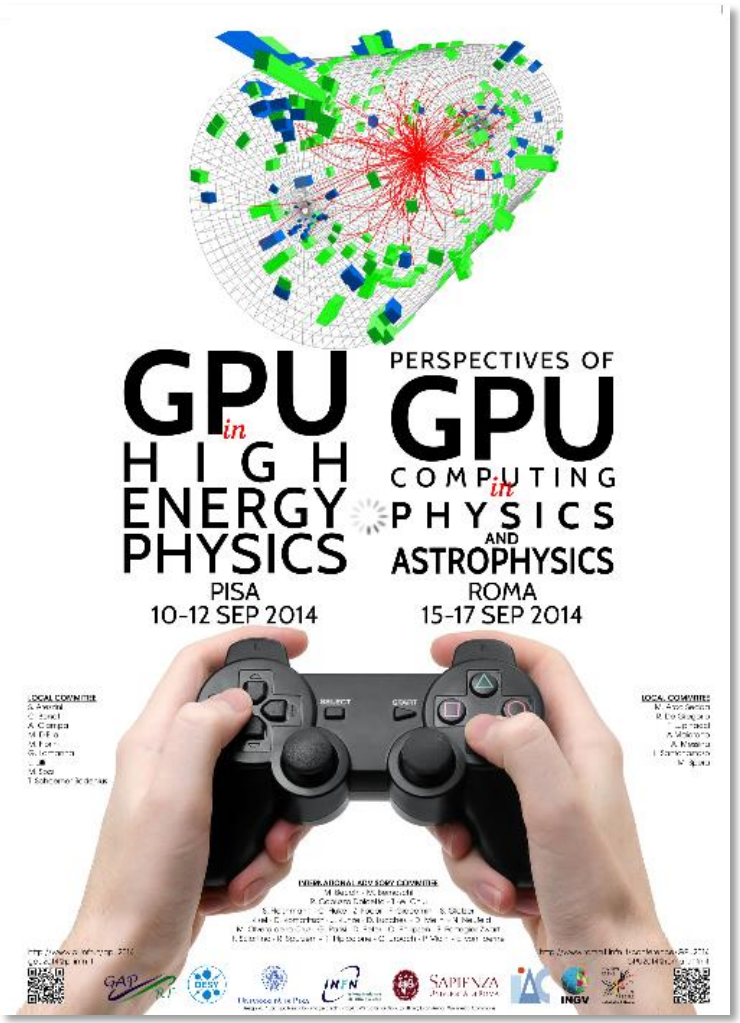
- Realizzato hardware per l'interfacciamento
- Firmware speciale per la trasmissione dei dati in formato opportuno
- Dati da 2 TEL62 (uno spot = 1000 PMs)
- Misura della latenza e tempo di processamento in condizioni reali

Conferences and Schools

- Amburgo – Workshop – 15/16 apr 2013 (Lamanna G., Messina A., Fiorini M.)
- Pechino – ACAT2013 – 16/21 apr 2013 (Vicini P.)
- Stoccolma – EPS2013 – 18/24 lug 2013 (Lamanna G., Vicini P., Fantechi R.)
- Amsterdam CHEP2013 – 14/18 apr 2013 (Lonardo A., Ammendola R.)

- Perugia – TWEPP2013 – 23/27 set 2013 (Biagioni A.)
- Seoul – IEEE-NSS2013 – 27ott/3nov 2013 (Fiorini M.)
- Hakayama – RICH2013 – 2/6 dic 2013 (Lamanna G., Collazuol G., Piccini M., Fiorini M.)
- Scuola di Bertinoro (Pinzino J., Santoni C., Bauce M.)
- Nara - RT-IEEE – (Bauce M., Vicini P.)
- Amsterdam –TIPP2014 – (M.Fiorini)

Conferenza GPU2014



- Accettati gli abstract: molti più rispetto alle aspettative!
- 3 giorni: circa 35 talk e 10 poster
- Attese almeno 80 persone.
- Diversi sponsor
- Conferenza gemellata a Roma.

Preventivi 2014

Persona	Qualifica	Sezione	%	Mesi in rend. MIUR 2014
Gianluca Lamanna	Ricercatore ex.art.23	Pisa	100	12
Riccardo Fantechi	Primo Ricercatore	Pisa	11	1
Mauro Piccini	Ricercatore	Perugia	20	2

“L'attività di ricerca di Gianluca Lamanna, Fantechi Riccardo, Mauro Piccini finanziata nel FIRB-LAMANNA è sinergica con l'esperimento NA62 della CSNI. “

Preventivi 2015

- Stesso schema dell'anno scorso?
- Un assegnista che ha già vinto il concorso ma ancora non ha preso servizio va messo?
- Per gli associati INFN come ci comportiamo?
- Cosa succede se dovessi cambiare sezione di afferenza? Posso mettere una piccola percentuale in un'attività di gruppo 5?

Difficoltà future

- Ovviamente il mio (eventuale) trasferimento non avrà impatto nullo sul progetto:
  - Ancora non mi sono chiare le limitazioni burocratiche e neppure cosa è meglio fare per il futuro (fare un gruppo (eventualmente) a Frascati? Potenziare il gruppo a Pisa?)
- Preoccupazione per la parte di ricostruzione PET
  - Situazione in stallo per molti mesi
  - Qualche sviluppo: chiediamo aiuto e supporto alla sezione

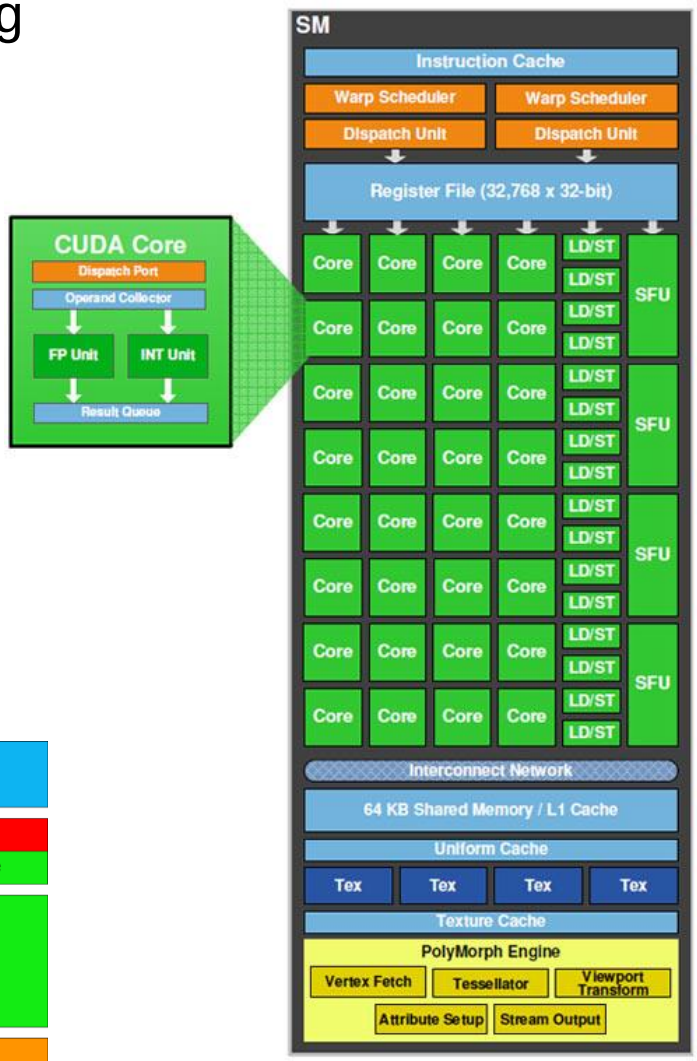
Spares

GPU vs CPU

- A GPU is tailored for highly parallel operation while a CPU executes programs serially
- For this reason, GPUs have many parallel execution units and higher transistor counts, while CPUs have few execution units and higher clock speeds
- A GPU is for the most part deterministic in its operation
- GPUs have much deeper pipelines (several thousand stages vs 10-20 for CPUs)
- GPUs have significantly faster and more advanced memory interfaces as they need to shift around a lot more data than CPUs

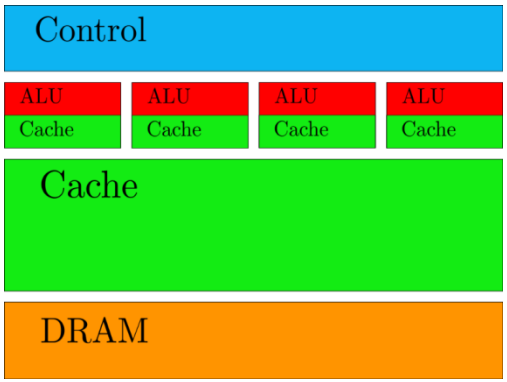
Different architecture

- SMX executes kernels (aka functions) using hundreds of threads concurrently.
- SIMT (Single-Instruction, Multiple-Thread)
- Instructions pipelined
- Thread-level parallelism
- Instructions issued in order
- No Branch prediction but Branch predication



Streaming Multiprocessor (SM)

- Large caches
- ILP
- Brach prediction
- Few powerful ALU
- Short pipeline



CPU

GAP Team

- 3 Research units, 3 years (started April 2013)

- **Pisa (INFN): Gianluca Lamanna**
  - Marco Sozzi
  - Riccardo Fantechi
  - Mauro Piccini (PG)
  - Gianmaria Collazuol (PD)
  - Flavio Costantini
  - Niccolò Camarlinghi

- **Roma: Andrea Messina**
  - Stefano Giagu
  - Marco Rescigno
  - Silvia Capuani
  - Marco Palombo
  - Andrea Laghi
  - Matteo Bauce

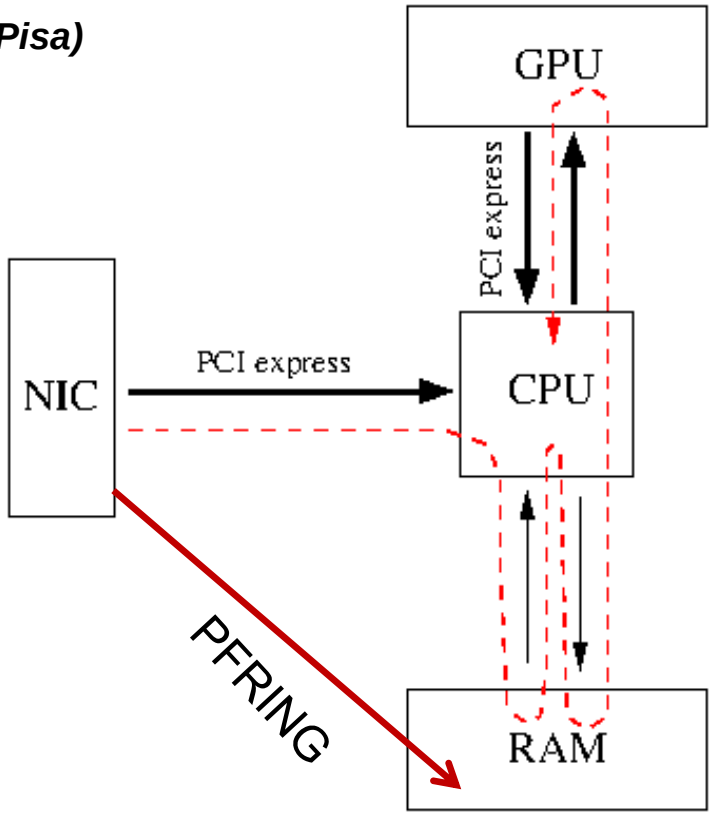
- **Ferrara: Massimiliano Fiorini**
  - Luciano Milano
  - Guido Zavattini
  - Angelo Cotta Ramusino
  - Stefano Chiozzi
  - Alberto Gianoli
  - Giovanni Di Domenico
  - Marco Corvo

GAP "extended" collaboration

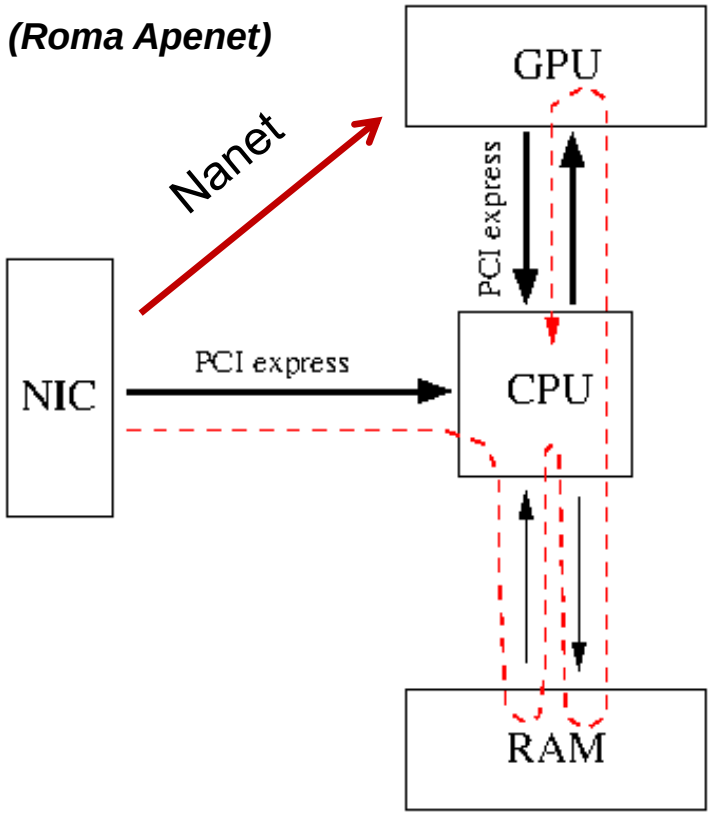
- Strong collaboration with the Apenet group (INFN-Roma1):
  - Piero Vicini, Alessandro Lonardo, Andrea Biagioni, etc.
  - HW development (con unità Pisa e Unità Ferrara)
- Possibility to collaborate with LHCb group (Padova):
  - Mainly Ferrara's unit
  - Preliminary collaboration for the tracking
  - Possible work in the RICH reconstruction
- Students & Ph.D. Students
  - Graverini Elena (PI), Pantaleo Felice (PI), Jacopo Pinzino (PI), Roberto Piandani (PI) and Santoni Cristiano (PG)

Two solutions: PFRING & NANET

(Pisa)

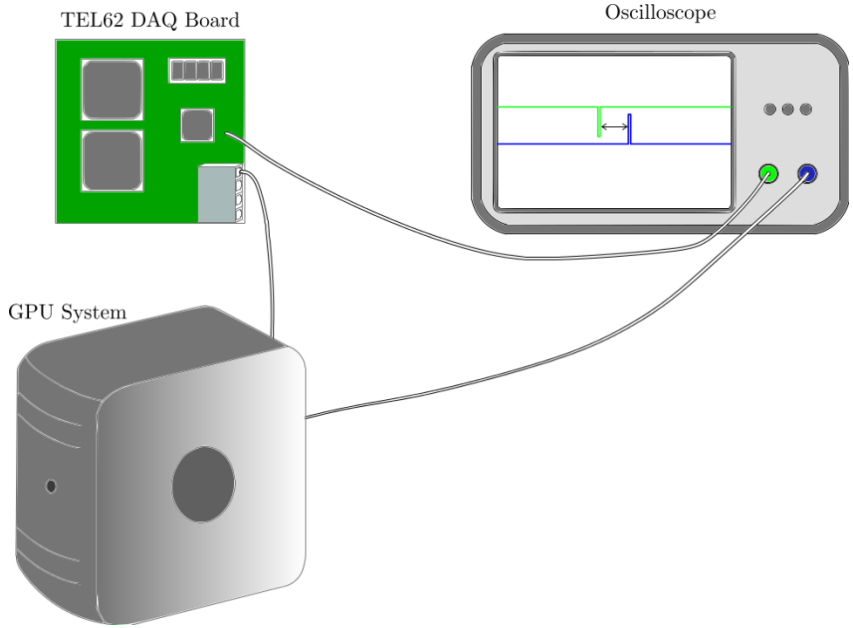


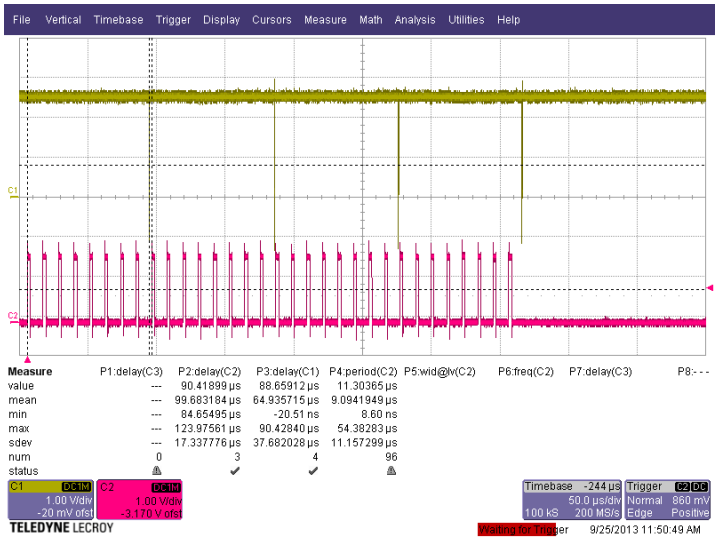
(Roma Apenet)



Latency measurement

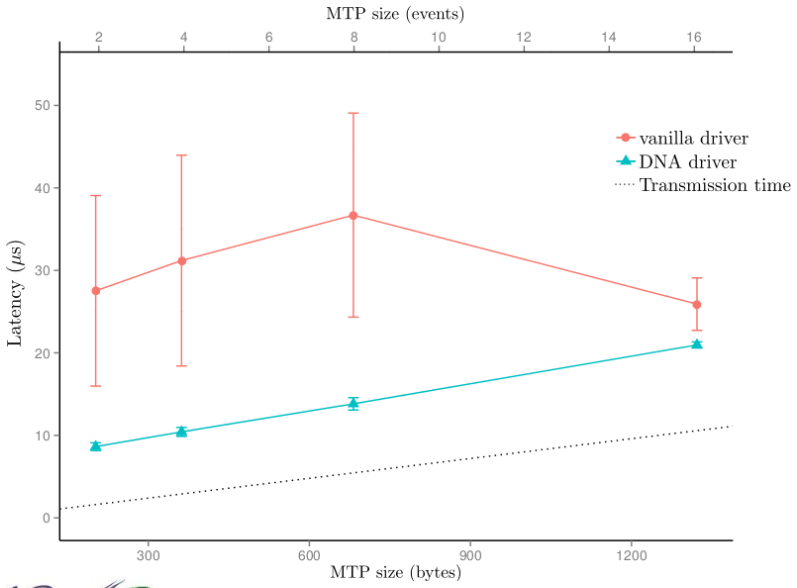
- It is not easy to measure (with high precision) the latency between two systems running with uncorrelated clocks
- Use an oscilloscope to have a common time reference

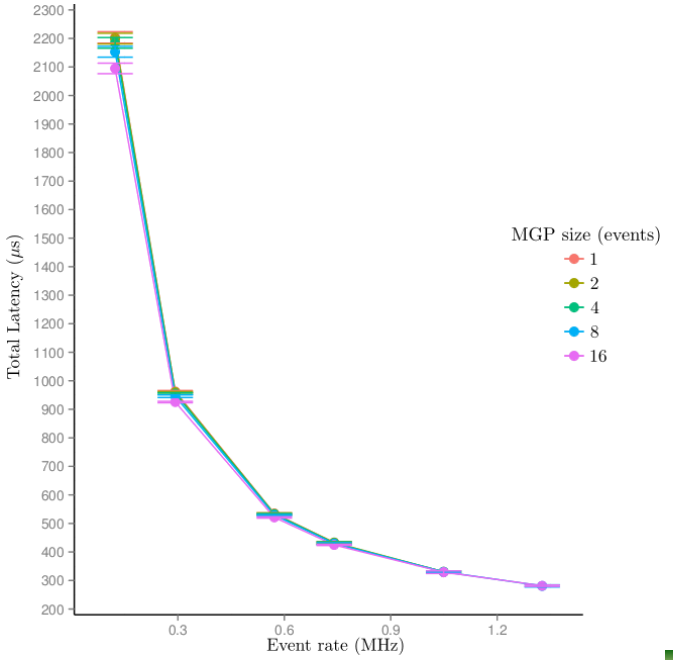




Results: PFRING

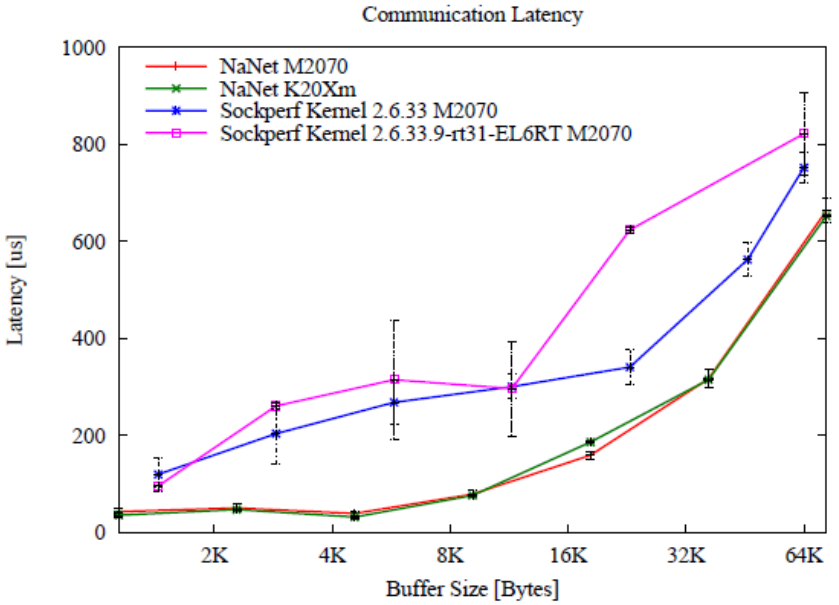
- Big improvement thanks to the DNA driver (both in absolute value and in fluctuations)
- Concurrent data copy – kernel execution
- Simple single ring kernel
- Latency strongly dominated by gathering
- Intrinsic latency due to the GPU computing (all components): ~200 us
  Data throughput: above 10 MHz (on single board)

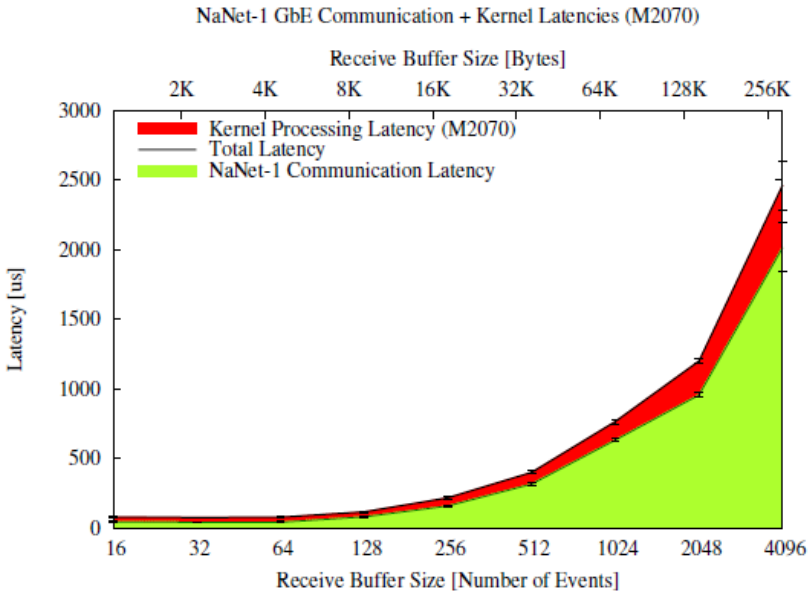




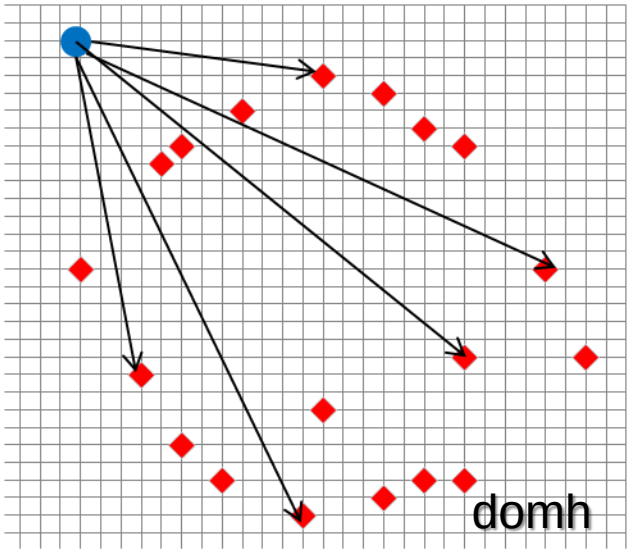
Results: NANET

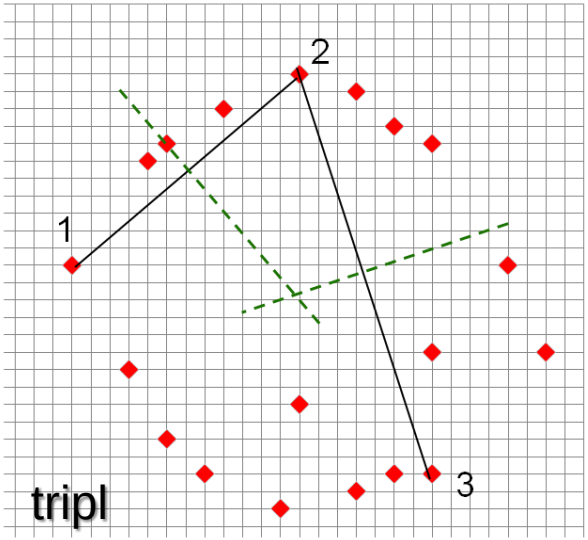
- Low latency-deterministic data transmission in GPU memory
- Latency and throughput: same order of magnitude with respect of PFRING (smaller fluctuations)
- Fine comparison ongoing
- Possibility of preprocessing in the FPGA not yet exploited
- For the moment “base line” solution (but PFRING is competitive and “easier”)

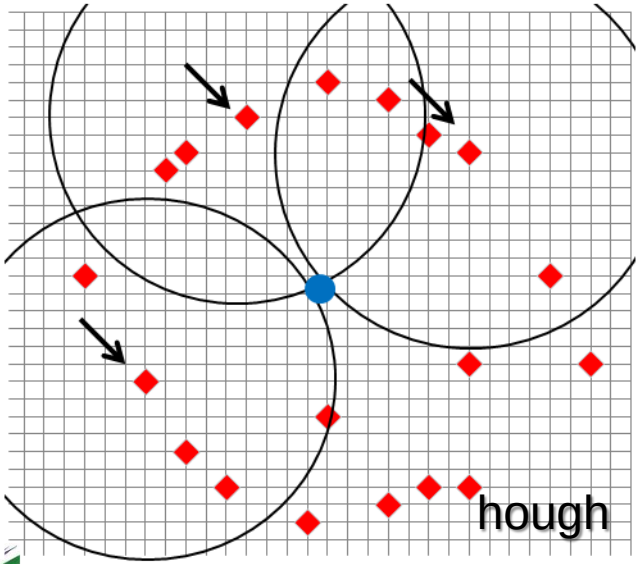


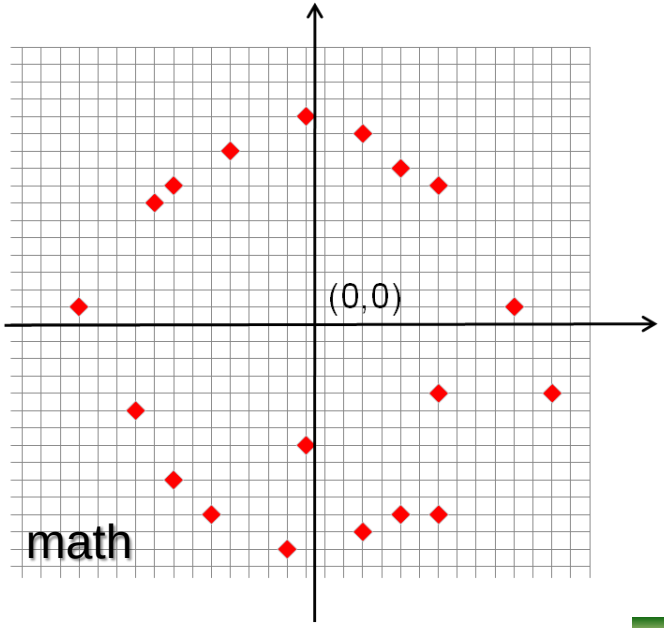


Single ring

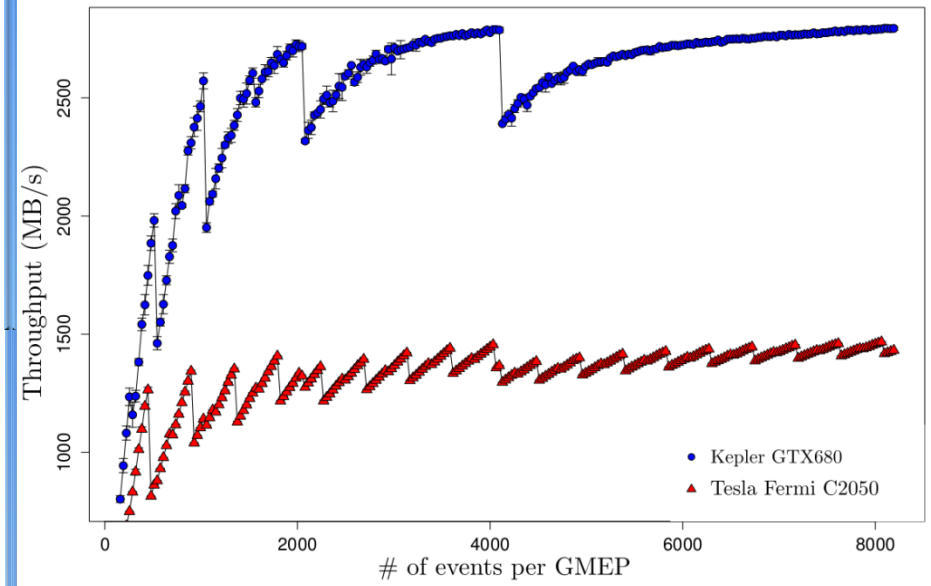


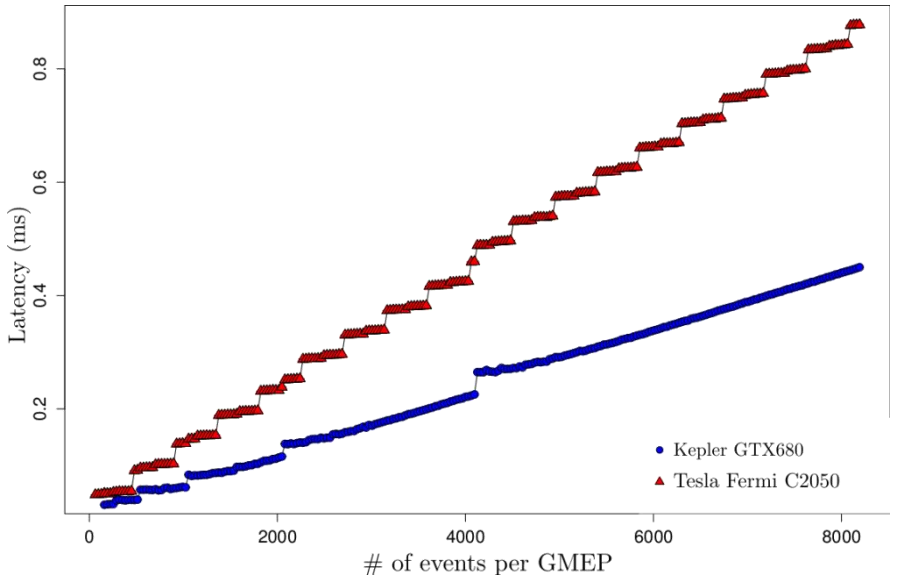






Single Ring Fit – Throughput & Latency



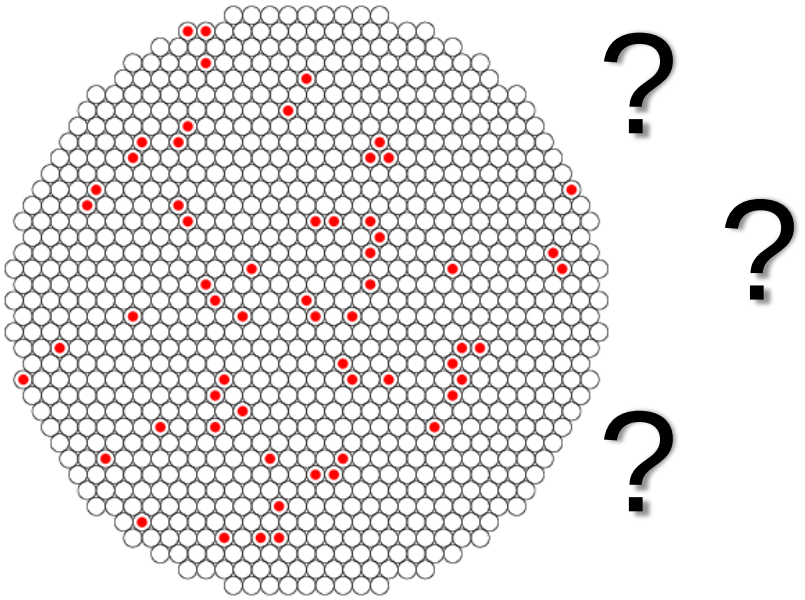


"Fast online triggering in high-energy physics experiments using GPUs" Nucl.Instrum.Meth.A662:49-54,2012

- discrete oscillations due to the discrete nature of the GPU
- saturation plateau (1.4 GB/s and 2.7 GB/s)
- A lower number of events inside the buffer is better to achieve a low latency
- A larger number of events guarantees a better performance and a lower overhead
- The choice of the buffer size is a compromise

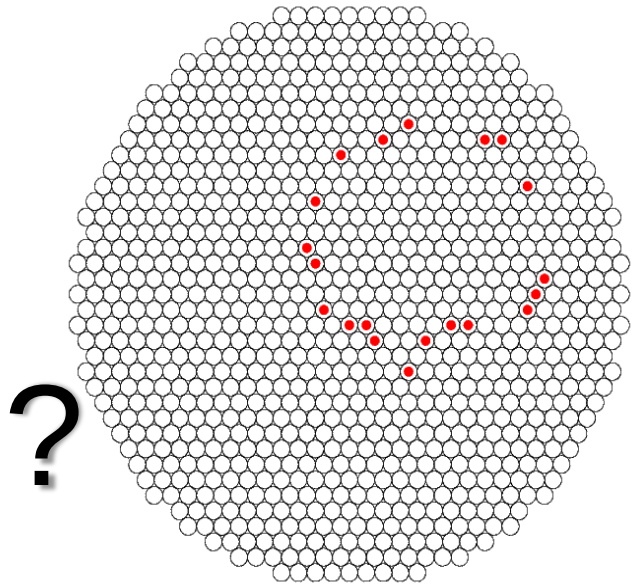
A look at the market: double rings

- Multi rings:
  - With seeds: Likelihood, Constrained Hough, …
  - Trackless: fiTQun, APFit, possibilistic clustering, Metropolis-Hastings, Hough transform, …



Requirements for an online multi-rings algorithm

- Trackless
  - no information from the tracker
  - Difficult to merge information from many detectors at L0
- Multi rings
  - Multi body decay in the RICH acceptance
- Fast
  - Not iterative procedure
  - Events rate at levels of tens of MHz
- Low latency
  - Online (synchronous) trigger
- Accurate
  - Offline resolution required



Almagest

i) Select a *triplet* (3 starting points)

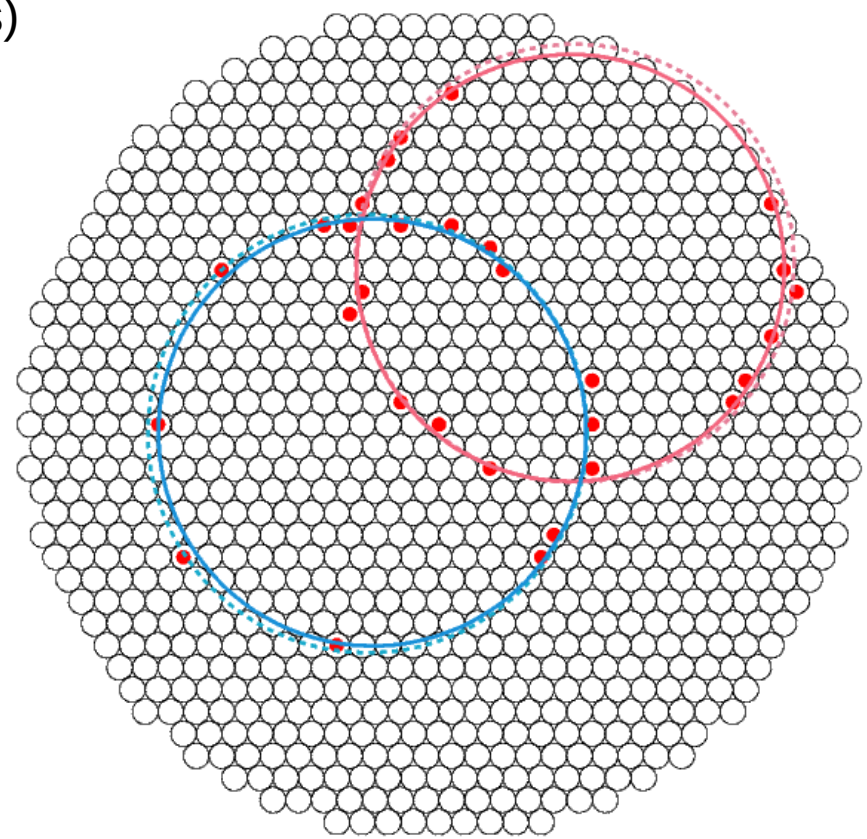
ii) Loop on the remaining points: if the next point does not satisfy the Ptolemy's condition then reject it

iii) If the point satisfy the Ptolemy's condition then consider it for the fit

iv) …again…

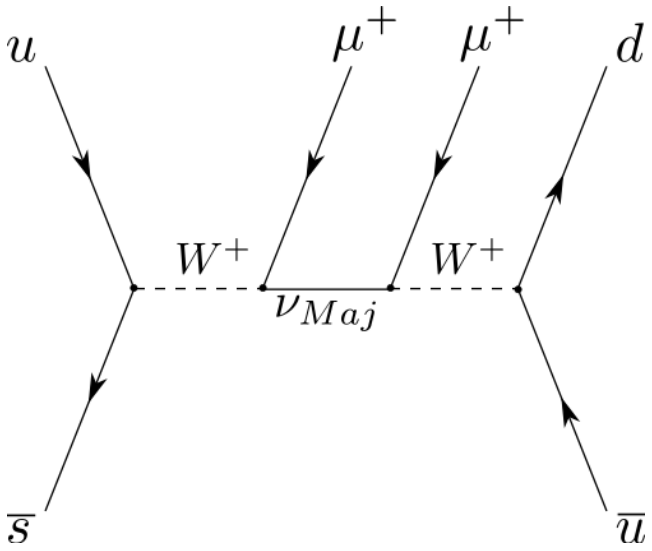
v) Perform a single ring fit

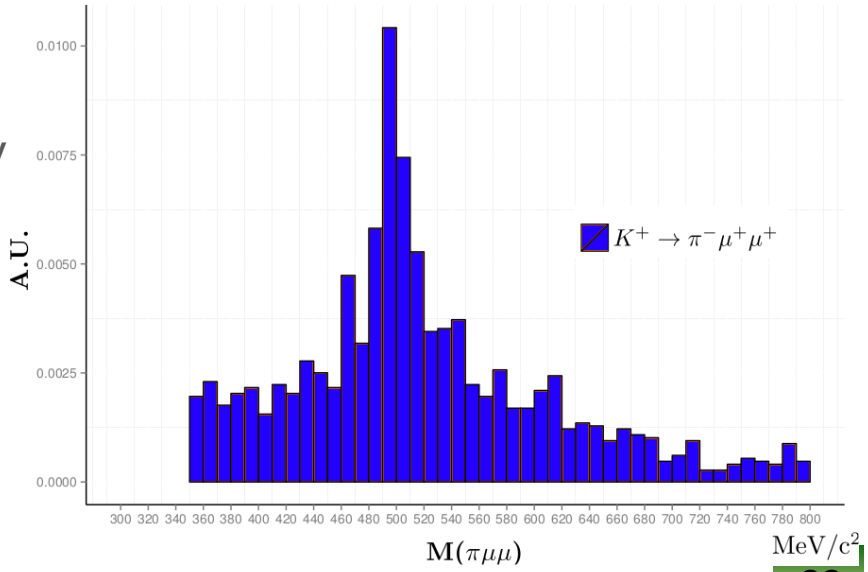
vi) Repeat by excluding the already used points



Benefit for NA62 (example)

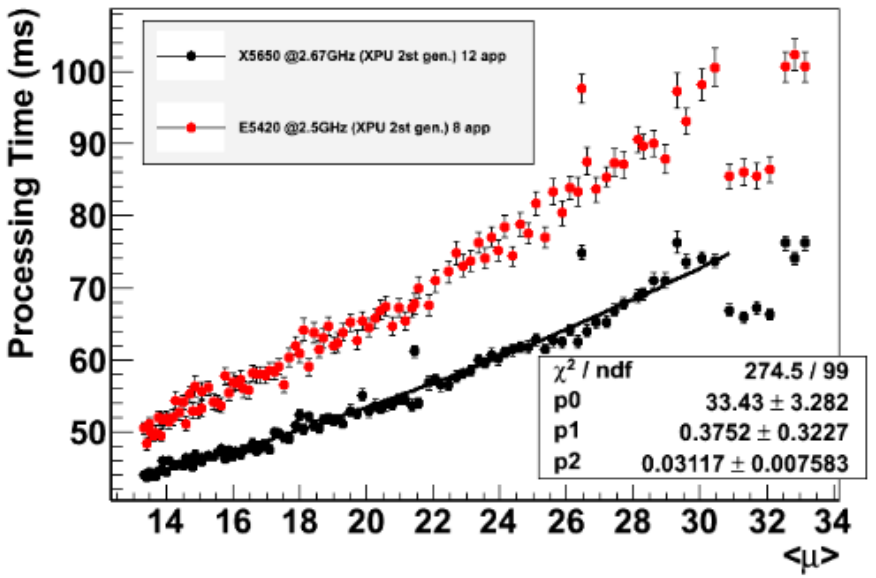
- $K^+ \to \pi^- \mu^+ \mu^+$:
  - LFV transitions
  - Rates extremely suppressed in SM
  - Any observable rate leads to evidence of NP
  - BR($K^+ \to \pi^- \mu^+ \mu^+$) < 1.1×10^{-9} @ 90% CL
- The presence of muons in the final stage isn't compatible with the NA62 main trigger:
  - Selective dedicated L0 trigger
  - Total mass can be reconstructed by using the rings information only (assuming the masses of the particles)
  - The center is proportional to the track's slope and the radius to the speed.





GAP-ATLAS

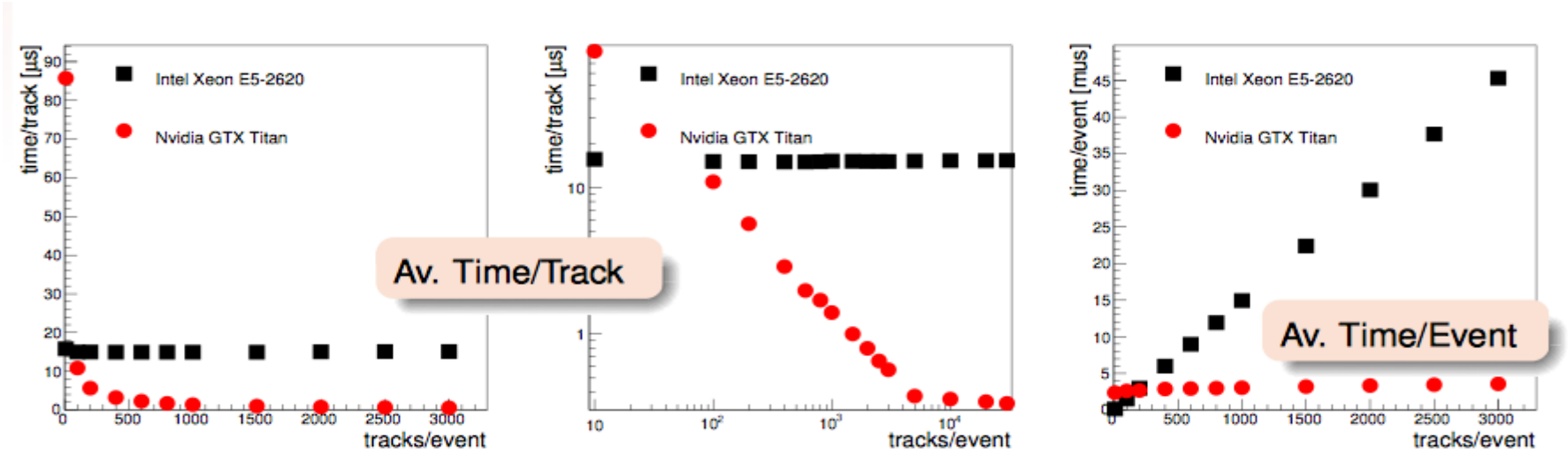
- The increase in LHC luminosity and in the number of overlapping events poses new challenges to the trigger system, and new solutions have to be developed for the fore coming upgrades (2018--2022)
  - A simple increase of the threshold can reduce signal efficiency drastically
  - More resolution and more complex reconstruction in HLT
- Reconstruction complexity and computing time scales with number of tracks
  - Higher throughput means increase network and CPU capabilities
  - Parallel computing is the solution



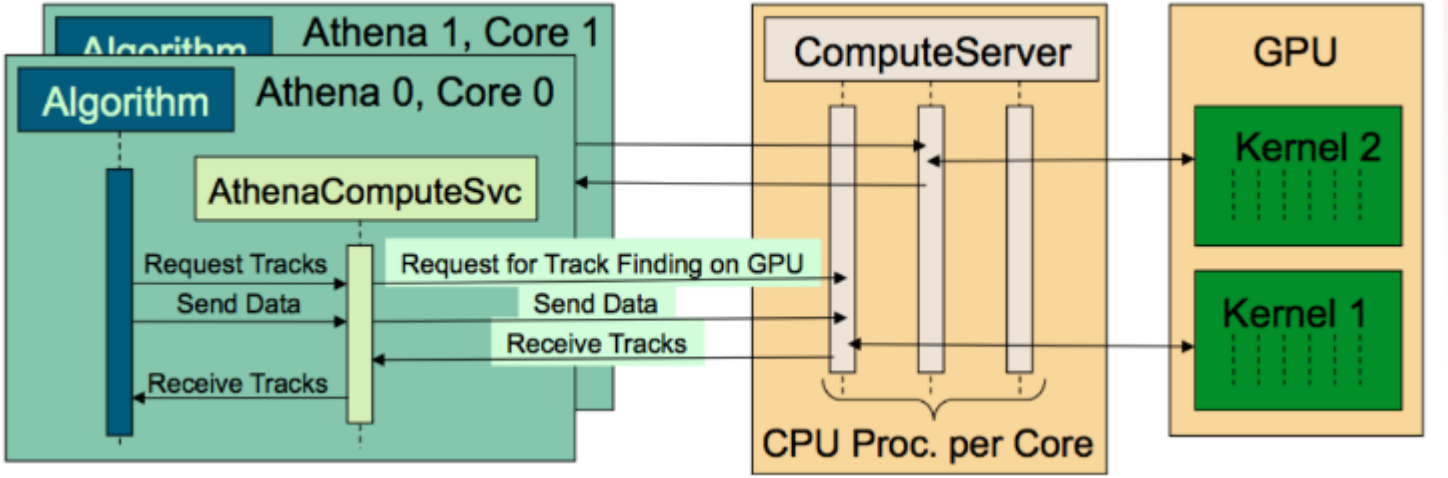
- The need to increase the computer performances (Flops/watt) in HLT will be relevant for all the LHC experiments.
- GPUs are an appealing solution to be investigate for this problem

GAP-ATLAS

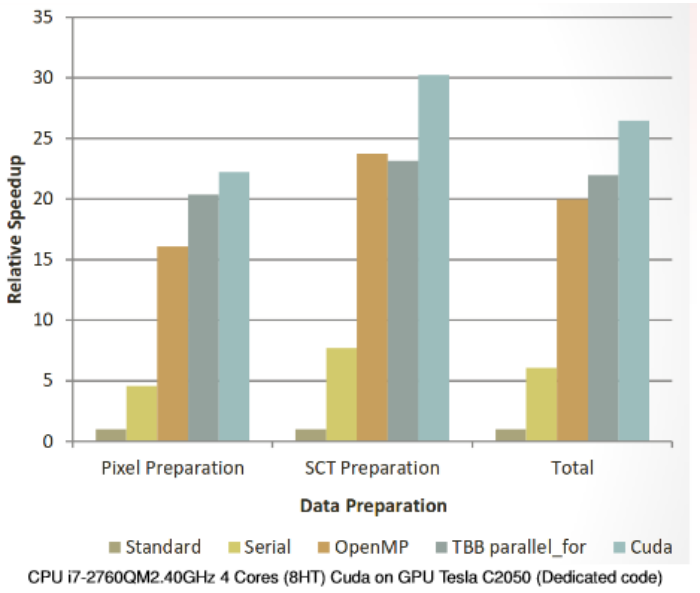
- Porting of Track reconstruction (same serial algorithm)
- Using standard CPUs the processing time is proportional to the pileup
- The GPUs give a processing time almost independent
- Start the porting of the muon trigger algorithm



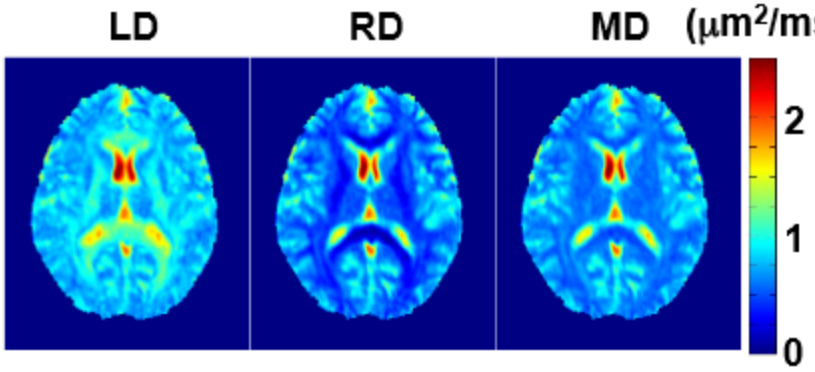
GAP-ATLAS



- Several groups interested in using GPUs in HLT and EB (Wuppertal, Frankfurt, Bonn, Edimburg, Oxford,…)
- Client-server layer to allow heterogeneous computing solution
- Encouraging preliminary tests



GAP-NMR



Conventional DTI

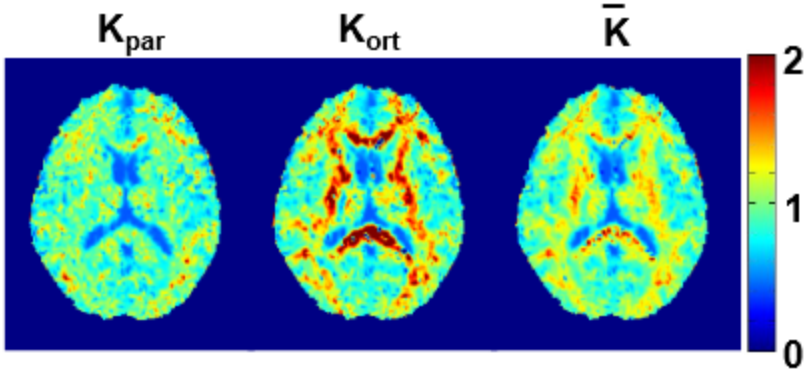
128x128x32x15

=

7.864320 10^6

linear systems to solve

~ 15 s on CPU*



Kurtosis tensor imaging

128x128x32x15

=

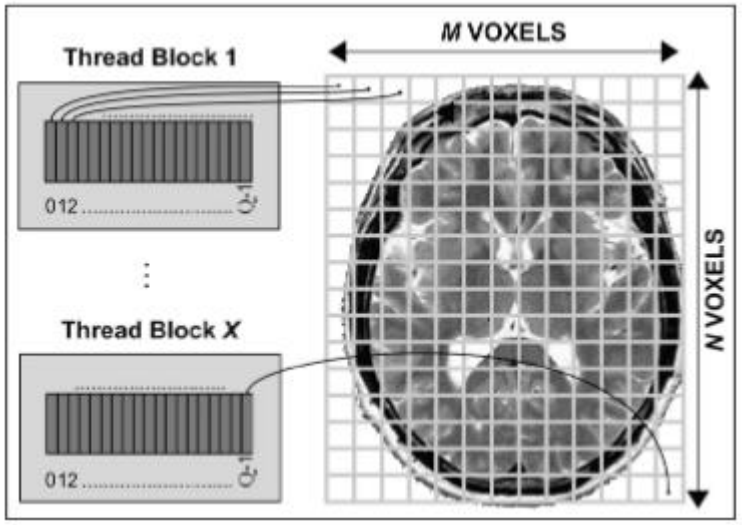
7.864320 10^6

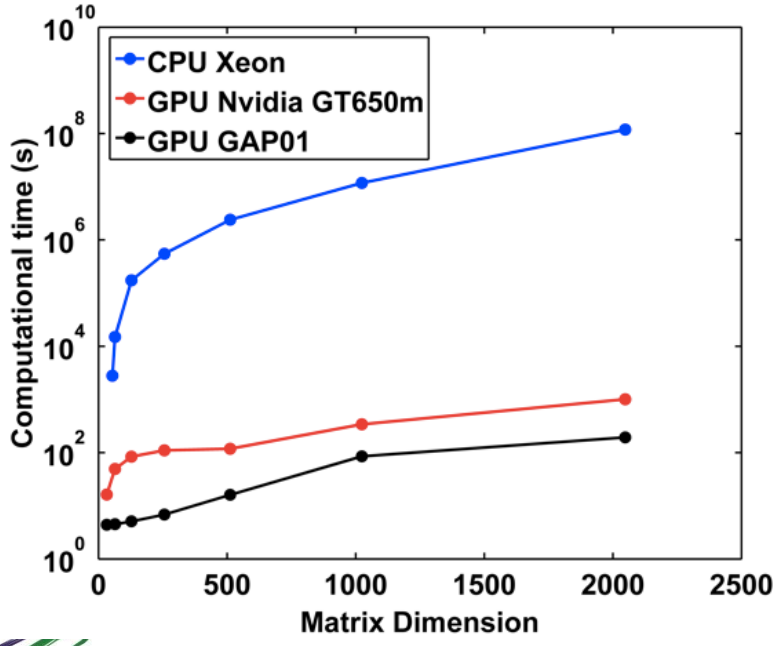
non-linear fit to perform

~7200 s on CPU*

* 8 threads on an Intel Xeon E5-2609 CPU at 2.4 GHz

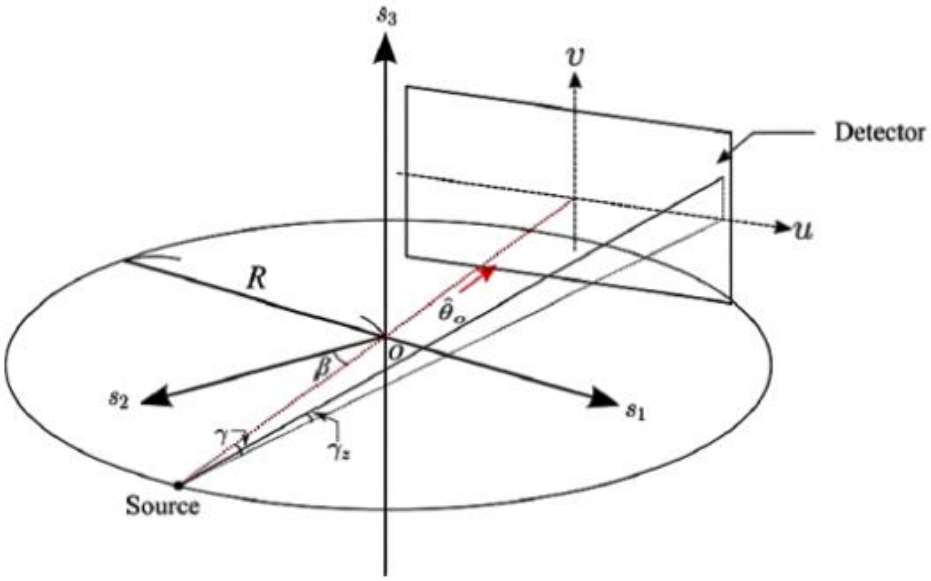
GAP-NMR





- From preliminary estimation:
  - Assuming 512x512 pixels
  - Factor 20-100 including optimization part
- With this gain in image reconstruction time, both Kurtosis and stretched exponential techniques would be used in diagnostic.

GAP-CT



- Porting of FDK algorithm on GPU
- Big improvement with respect to already parallel version
- Big benefit from next generation (already on the market) GPU (K40)

$$\hat{f}(s_1, s_2, s_3) = \frac{1}{2}\int_0^{2\pi} d\beta \frac{1}{U^2}\int_{-u_m}^{u_m} du \frac{D}{\sqrt{D^2+u^2+v^2}} \cdot g_\beta(u,v) h(u-u')$$

step	GTS 450	GTX 680	Titan
weighting	0.32	1.5	0.39
filtering	49.17	15.01	13.57
backlproj	17.45	4.37	3.08
Total	66.94	20.88	17.04

step	OpenMP	Titan	Perf. ratio
weighting	0.215	0.39	**0.55**
filtering	1.26	13.57	**0.092**
backlproj	80.85	3.08	**26.3**
Total	82.325	17.04	**4.83**

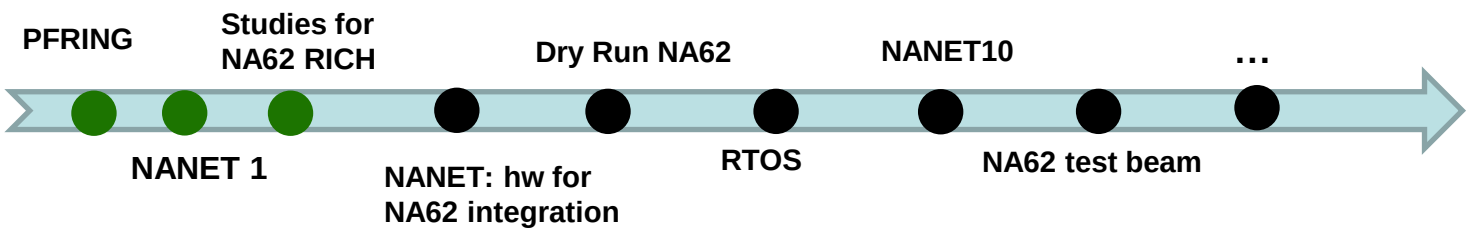
GAP-PET

- Preliminary studies to identify suitable algorithms:
  - MLEM (Maximum Likelihood expectation maximization)
  - Common points with CT development (retro-projectors)
  - Indication of relevant improvement using GPUs, essential for new scanners with higher resolution (more LOR)

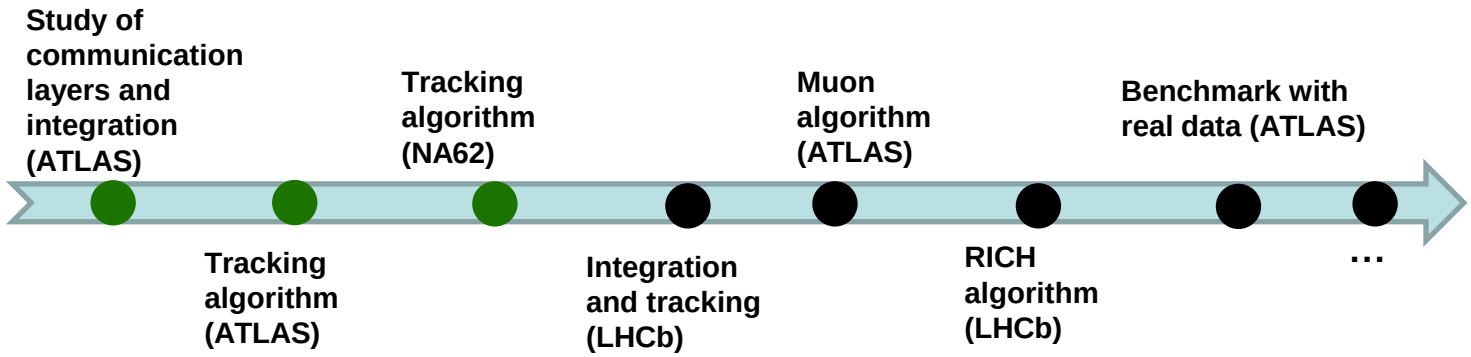
GAP Plan

	I ANNO Realizzazione	II ANNO Caratterizzazione	III ANNO Integrazione
PISA	• Gestione della latenza (10, 40 Gb/s) • Algoritmi per Ring • Algoritmi list-mode per PET	• Misura della latenza di trasferimento in varie modalità • Studio dell'integrazione in NA62 • Studio dell'integrazione con PET	• Integrazione in NA62 • Misure di efficienza in NA62 • Integrazione in PET e studio di performance e vantaggi
FERRARA	• Sviluppo UNICA/NANET • Algoritmi per GTK •Algoritmi per CT (FDK iterativo)	• Caratt. UNICA/NANET su sistemi sincroni • Algoritmi per GTK • Algoritmi SIRT e EM e studio per integrazione con CT	• Integrazione in NA62 • Misure di efficienza in NA62 • Integrazione in CT e studio di performance e vantaggi
ROMA	• Sistemi RTOS • Algoritmi per mu • Algoritmi per NMR (kurtosis diffusionale)	• Integrazione dei sistemi di riduzione della latenza in sistemi RTOS •Studio integrazione con NMR	• Integrazione in ATLAS • Analisi di dati veri e/o simulati • Integrazione in NMR e studio di performance e vantaggi

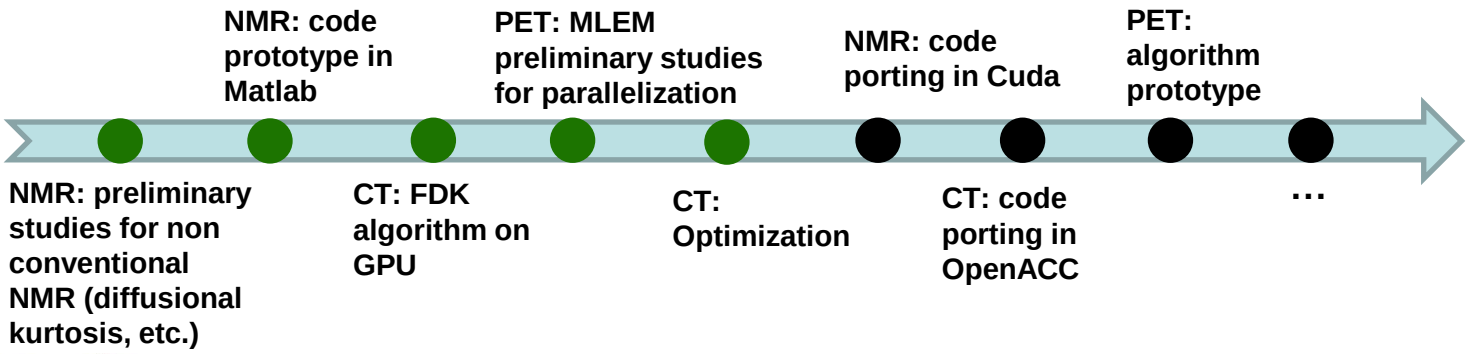
Where we are?



Demonstration of the use of GPUs in low latency trigger level.



Study of the advantages given by GPU in HLT.



Proof of the GPU improvement in real time image reconstruction in NMR, PET and CT.

Conferences and Schools

- Amburgo – Workshop – 15/16 apr 2013 (Lamanna G., Messina A., Fiorini M.)
- Pechino – ACAT2013 – 16/21 apr 2013 (Vicini P.)
- Stoccolma – EPS2013 – 18/24 lug 2013 (Lamanna G., Vicini P., Fantechi R.)
- Amsterdam CHEP2013 – 14/18 apr 2013 (Lonardo A., Ammendola R.)
- Perugia – TWEPP2013 – 23/27 set 2013 (Biagioni A.)
- Seoul – IEEE-NSS2013 – 27ott/3nov 2013 (Fiorini M.)
- Hakayama – RICH2013 – 2/6 dic 2013 (Lamanna G., Collazuol G., Piccini M., Fiorini M.)
- Scuola di Bertinoro (Pinzino J., Santoni C., Bauce M.)

Computing vs LUT

Where is this limit?
It depends …
In any case the GPUs
aim to shrink this space

Supercomputers

Titan (supercomputer)

Active	Became operational October 29, 2012
Sponsors	US DOE and NOAA (<10%)
Operators	Cray Inc.
Location	Oak Ridge National Laboratory
Architecture	18,688 AMD Opteron 6274 16-core CPUs 18,688 Nvidia Tesla K20X GPUs
Power	8.2 MW
Operating system	Cray Linux Environment
Space	404 m^2 (4352 ft^2)
Memory	693.5 TiB (584 TiB CPU and 109.5 TiB GPU)
Storage	40 PB, 1.4 TB/s IO Lustre filesystem
Speed	17.59 petaFLOPS (LINPACK) 27 petaFLOPS theoretical peak
Cost	$97 million
Ranking	TOP500: #2, June 2013[1]
Purpose	Scientific research
Legacy	Ranked 1 on TOP500 when built. First GPU based supercomputer to perform over 10 petaFLOPS
Web site	www.olcf.ornl.gov/titan/

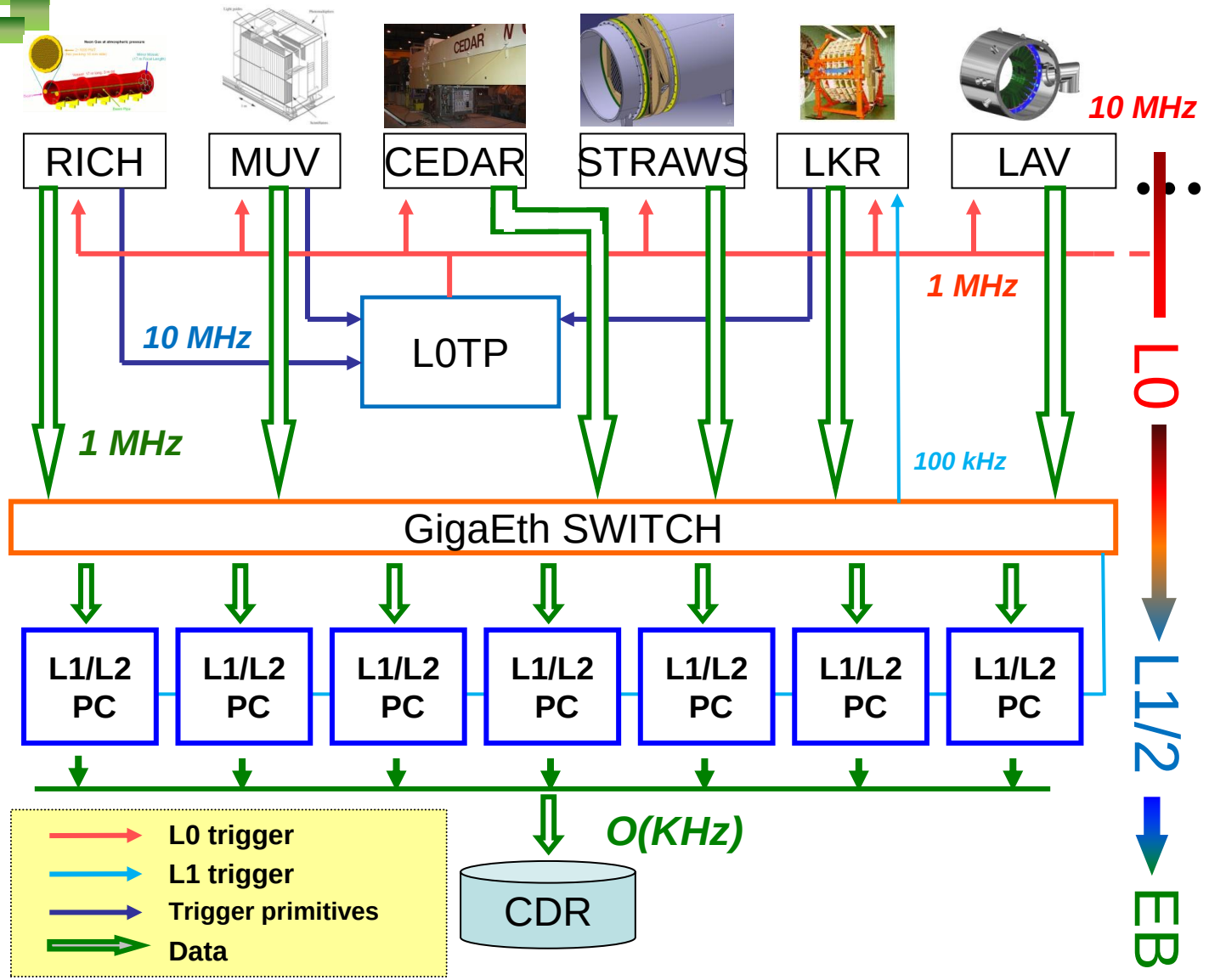
System	**Eurora supercomputer: 64 nodes, 128 CPUs, 128 GPUs**
Node Card	Intel Xeon E5-2687W (150W)
n.2 nVIDIA K20s, n.1 Infiniband QDR	NVIDIA® Tesla® K20
Ambient Temperature	20°C+/-1°C
Coolant Temperature	19°C+/-1°C
Coolant	water
Flowrate	120lph +/-7lph each EuroraBoard

CINECA

NVIDIA. TESLA™

EUROTECH

The NA62 TDAQ system

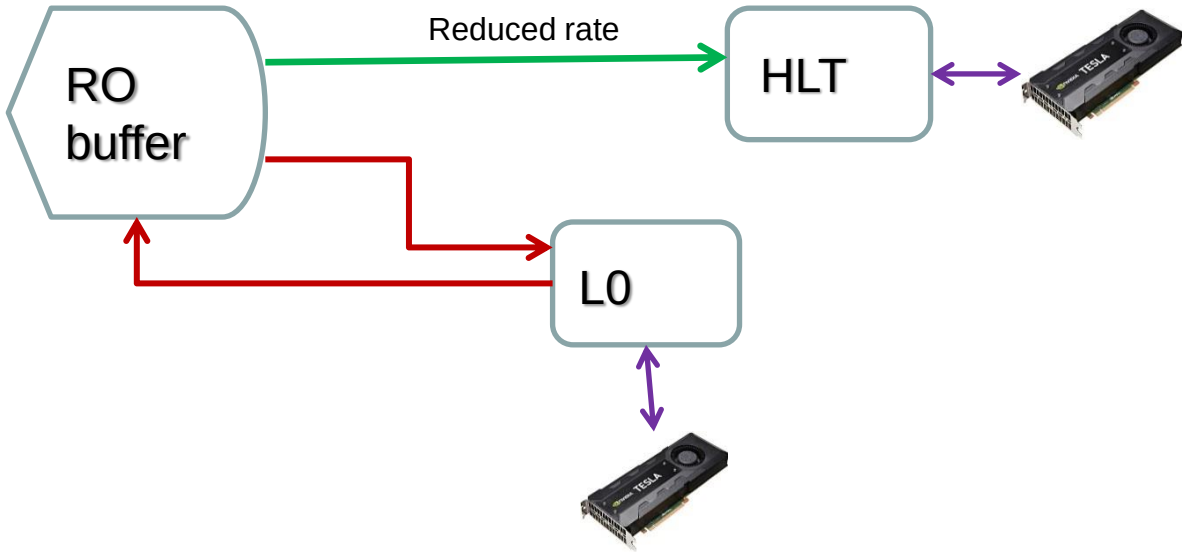


L0: Hardware synchronous level. 10 MHz to 1 MHz. Max latency 1 ms.

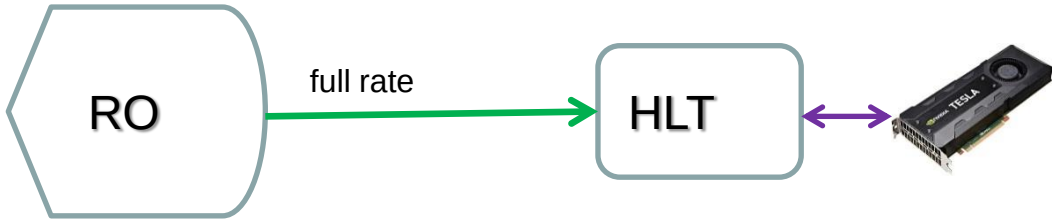
L1: Software level. "Single detector". 1 MHz to 100 kHz

L2: Software level. "Complete information level". 100 kHz to few kHz.

GPU: where?



«classical trigger»



«triggerless»