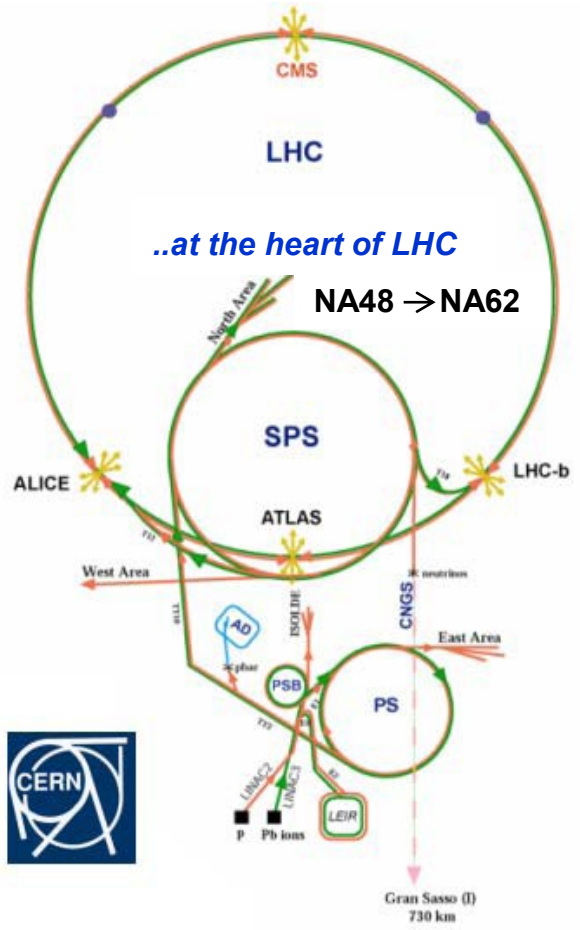
"GPUs for fast triggering in NA62 experiment"

TIPP 2011 Chicago, 11.6.2011

Gianluca Lamanna
(CERN)

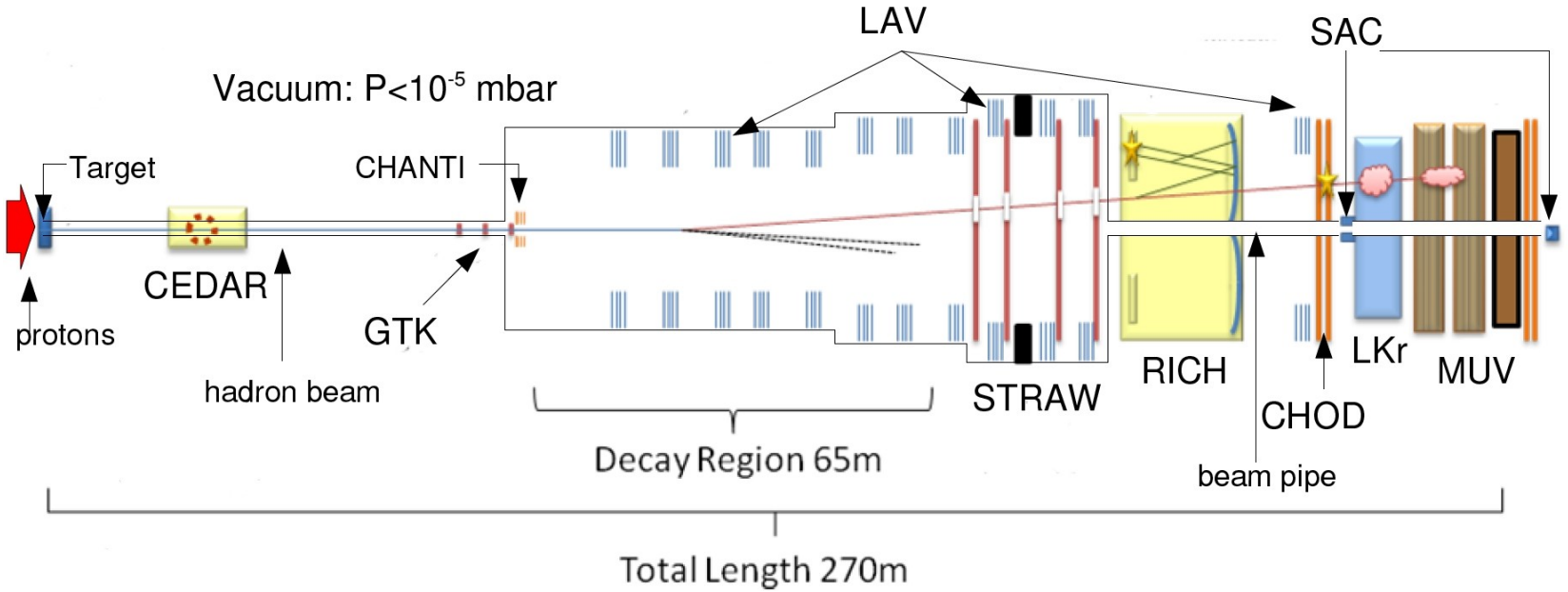
Outline

- Overview of the NA62 experiment @CERN
- Introduction to the Graphics processing unit (GPU)
- The GPU in the NA62 trigger
  - Parallel pattern recognition
  - Measurement of GPU latency
- Conclusions



NA62 Collaboration
Bern ITP, Birmingham, Bristol, CERN, Dubna, Ferrara, Fairfax, Florence, Frascati, Glasgow, IHEP, INR, Liverpool, Louvain, Mainz, Merced, Naples, Perugia, Pisa, Rome I, Rome II, San Luis Potosi, SLAC, Sofia, TRIUMF, Turin

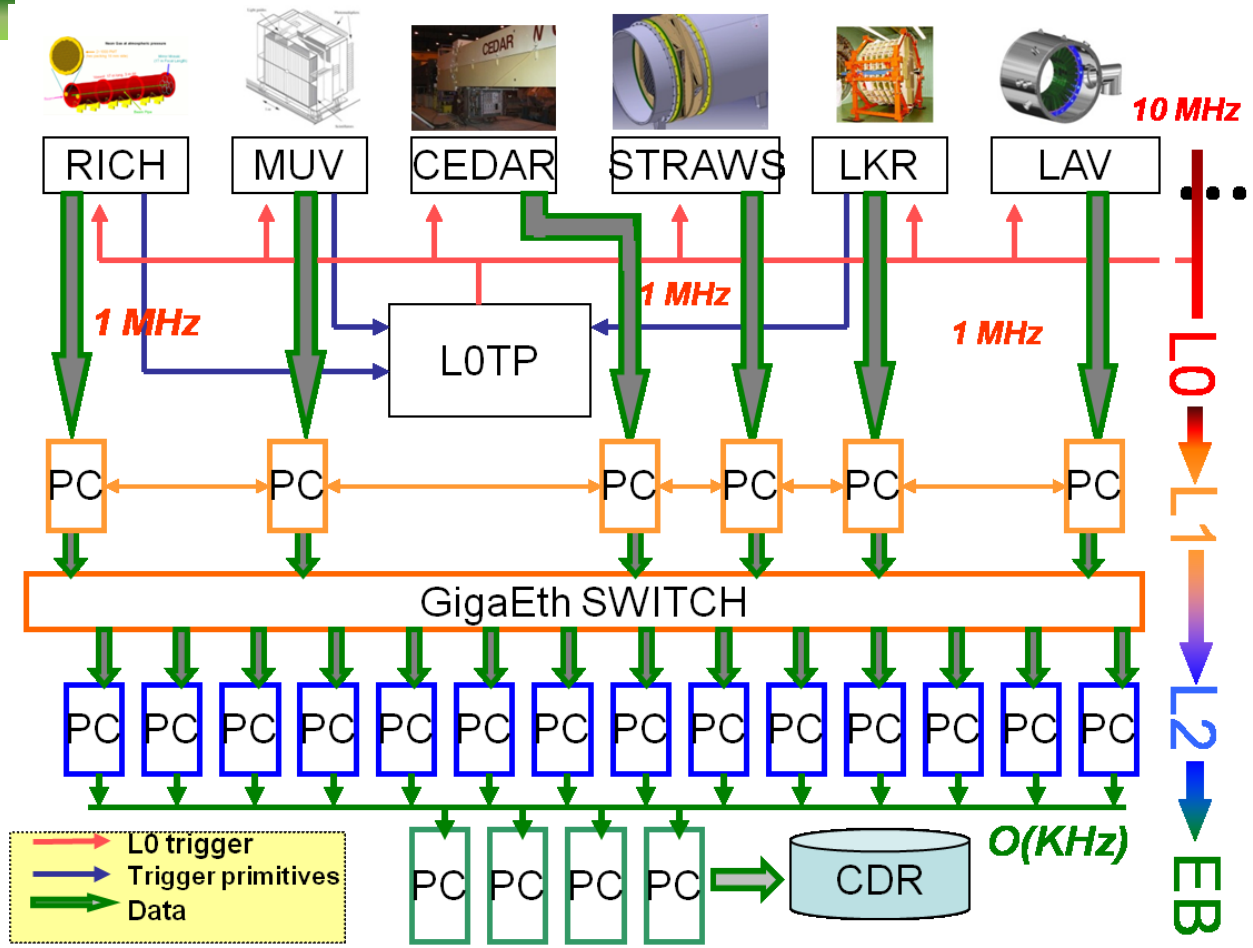
NA62: Overview



- Main goal: BR measurement of the ultrarare $K \to \pi\nu\nu$ ($BR_{SM}=(8.5\pm0.7)\cdot10^{-11}$)
- Stringent test of SM, golden mode for search and characterization of New Physics (complementary with respect to the direct search)
- Novel technique: kaon decay in flight, O(100) events in 2 years of data taking

- Huge background:
  - Hermetic veto system
  - Efficient PID
- Weak signal signature:
  - High resolution measurement of kaon and pion momentum
- Ultra rare decay:
  - High intensity beam
  - Efficient and selective trigger system

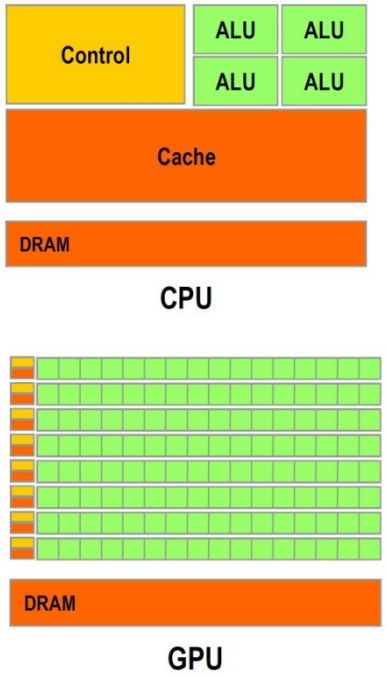
The NA62 TDAQ system



- L0: Hardware level. Decision based on primitives produced in the RO card of detectors partecipating to the trigger
- L1: Software level. "Single detector" PCs
- L2: Software level. The informations coming from different detectors are merged together

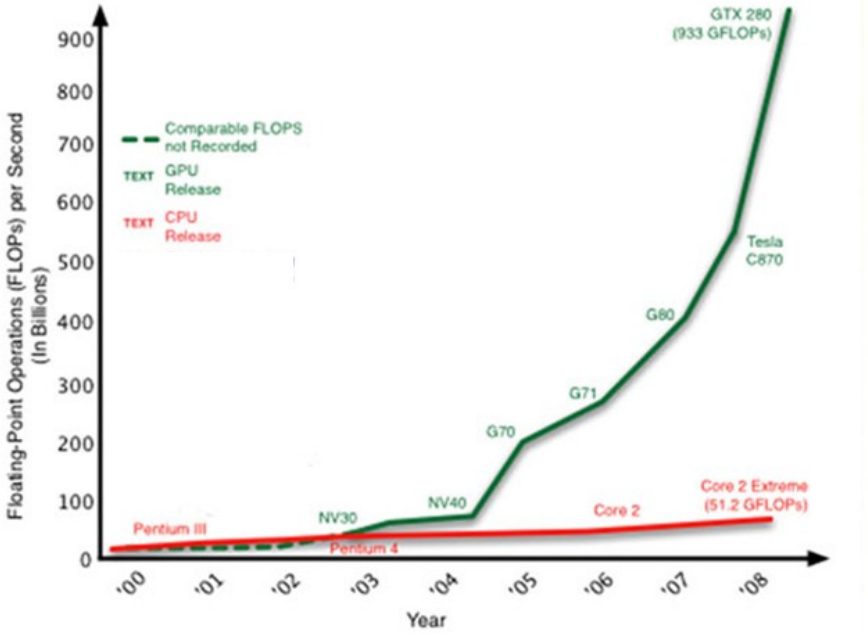
- The L0 is synchronous (through TTC). The input rate is ~10 MHz. The latency is order of 1ms. The trigger decision is based on information coming from RICH, LKr, LAV and MUV.
- The L1 is asynchronous (through Ethernet). The input rate is ~1 MHz. The maximum latency is few seconds (spill length). The output rate is ~100 kHz.

The Video Card processor: GPU



- The GPUs (Graphics processor units) are the standard processors used in the commercial Video Cards for PCs
- Main vendors: ATI (now AMD) and NVIDIA
- Two standards to program the GPU: OpenCL and CUDA
- GPU originally specialized for math-intensive, highly parallel computation
- More transistors can be devoted to data processing rather than data caching and flow control with respect to the standard CPUs

- SIMD (Single Instruction Multiple Data) architecture
- Very high computing power for "vectorizable" problems
- General purpose computing (GPGPU): several applications in lattice QCD, fluid dynamics, medical physics, etc.

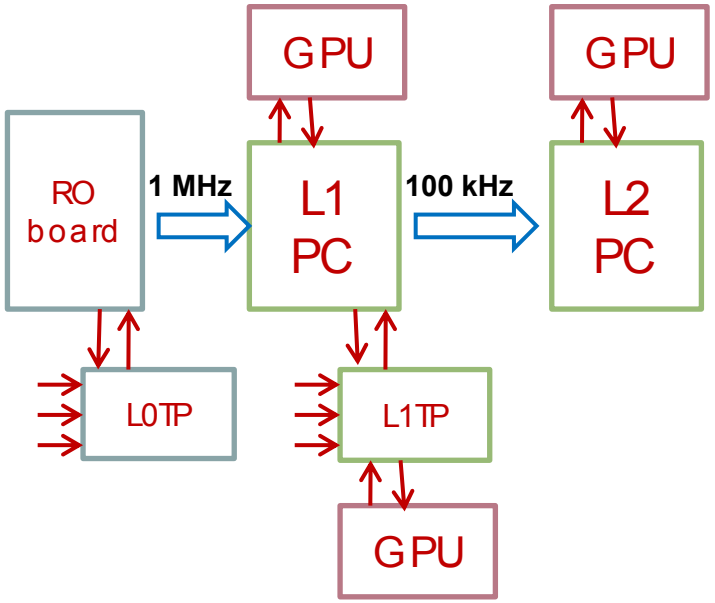


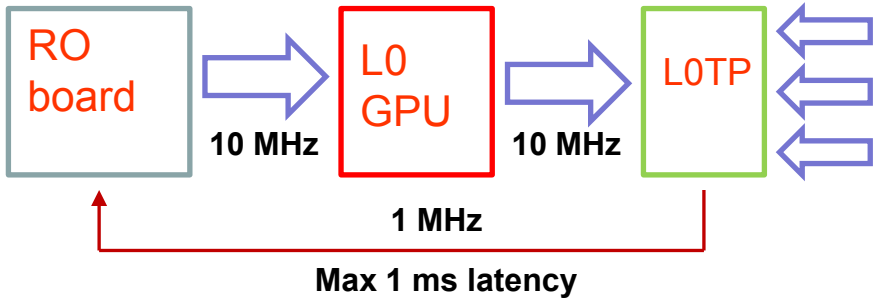
Video Cards used in the tests

	NVIDIA Quadro 600	NVIDIA Tesla C1060	NVIDIA Tesla C2050	AMD Radeon HD 5970
Number of multi-processors	2	30	14	20
Number of cores	96	240	448	320
Core Frequency (GHz)	0.64	1.3	1.15	0.725
Main memory (GB)	1	4	3	2
Main memory bandwidth (GB/s)	25.6	102	144	256
Computing power (TFLOPS)	0.246	0.93	1.03	4.6
Max power consumption	40	188	247	294

GPUs in the NA62 TDAQ system

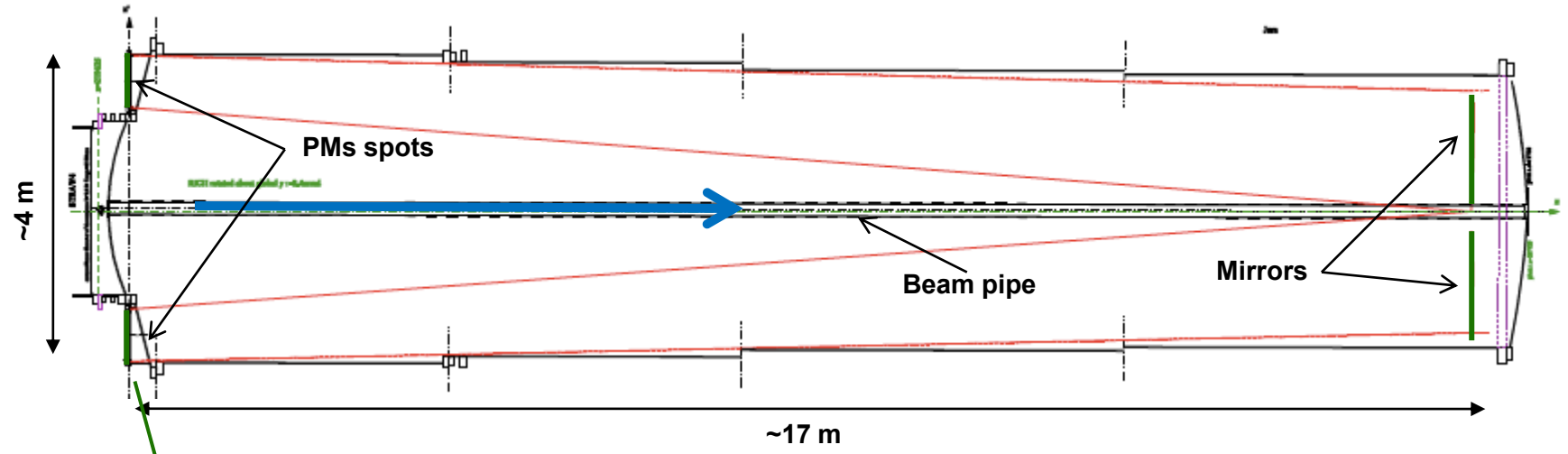
- The use of the GPU at the software levels (L1/2) is quite straightforward: just put the video card in the PC
- No particular changes to the hardware are needed
- The main advantages is to reduce the number of PCs in the L1 farms

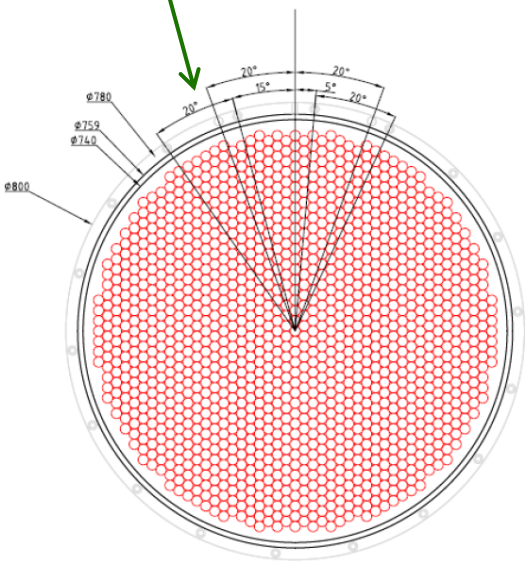




- The use of GPU at L0 is more challenging:
  - Fixed and small latency (dimension of the L0 buffers)
  - Deterministic behavior (synchronous trigger)
  - Very fast algorithms (high rate)

First application: RICH



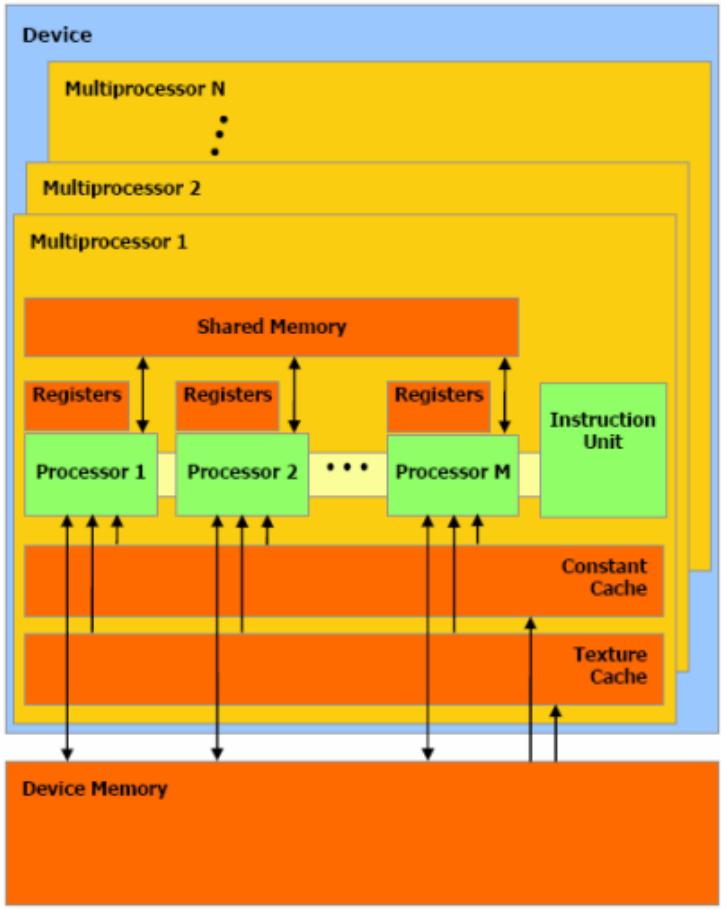


- ~17 m RICH
- 1 Atm Neon
- Light focused by two mirrors on two spots equipped with ~1000 PMs each (pixel 18 mm)
- 3σ π–μ separation in 15-35 GeV/c, ~18 hits per ring in average
- ~100 ps time resolution, ~10 MHz events rate
- Time reference for trigger

[See M.Pepe 9/6 at 16:00, PID Detectors]

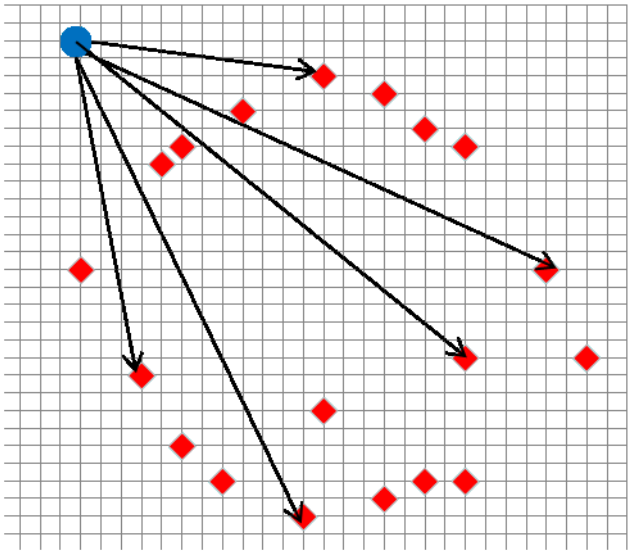
GPU @L0 RICH trigger

- 10 MHz input rate, ~95% single track
- 200 B event size → reduced to ~40 B with FPGA preprocessing
- Event buffering in order to hide the latency of the GPU and to optimize the data transfer on the Video card
- Two level of parallelism:
  - algorithm
  - data

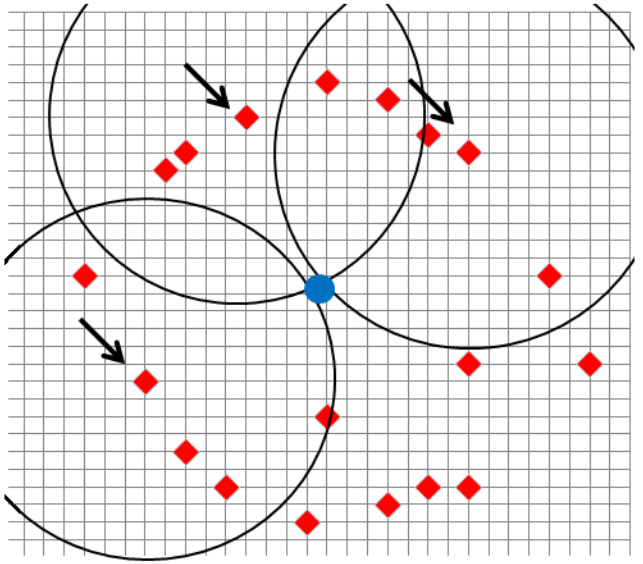


Algorithms for single ring search (1)

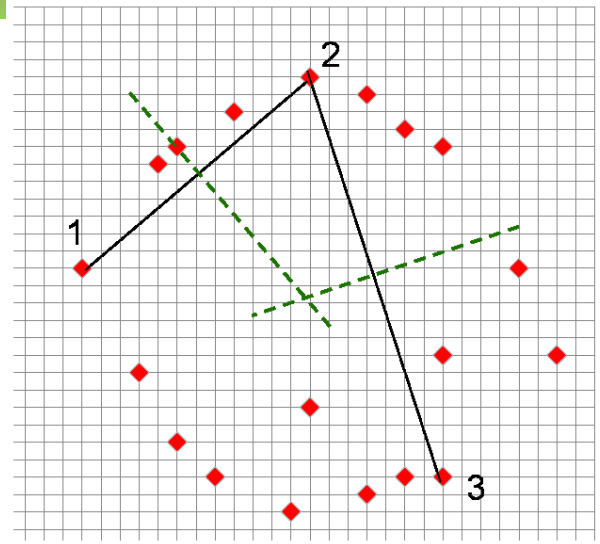
- DOMH/POMH: Each PM (1000) is considered as the center of a circle. For each center an histogram is constructed with the distances btw center and hits.

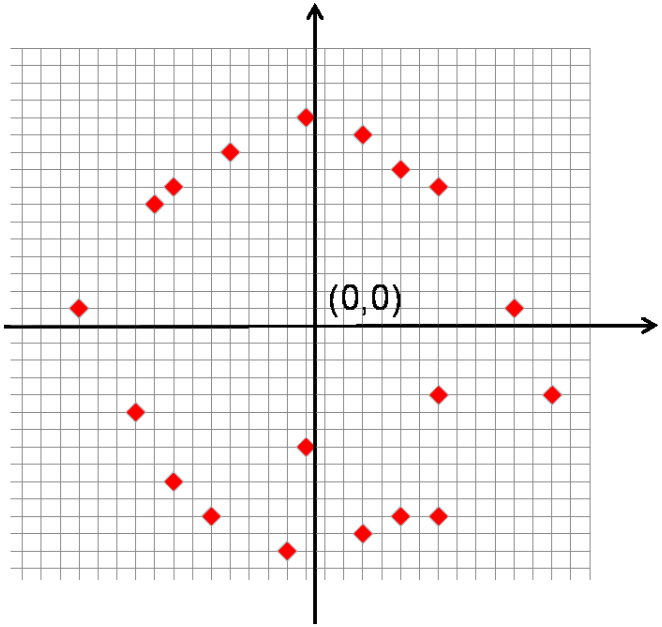


- HOUGH: Each hit is the center of a test circle with a given radius. The ring center is the best matching point of the test circles. Voting procedure in a 3D parameters space



Algorithms for single ring search (2)



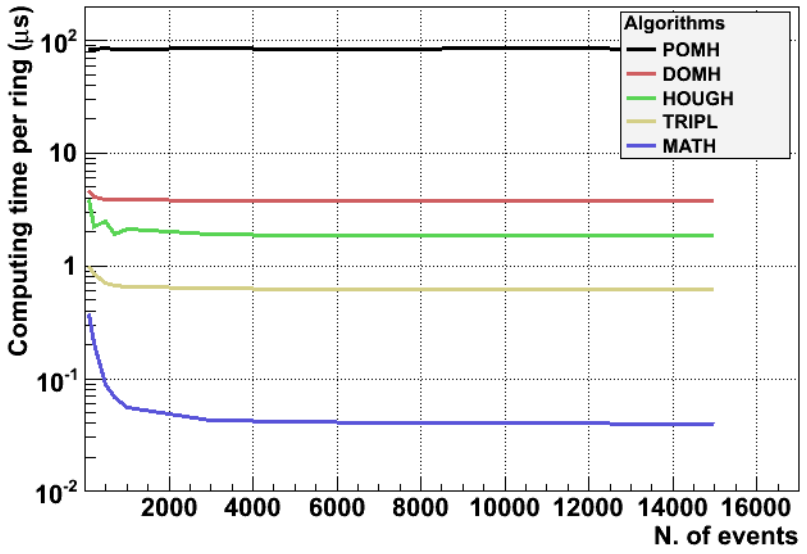


- TRIPL: In each thread the center of the ring is computed using three points (“triplets”) .For the same event, several triplets (but not all the possible) are examined at the same time. The final center is obtained by averaging the obtained center positions

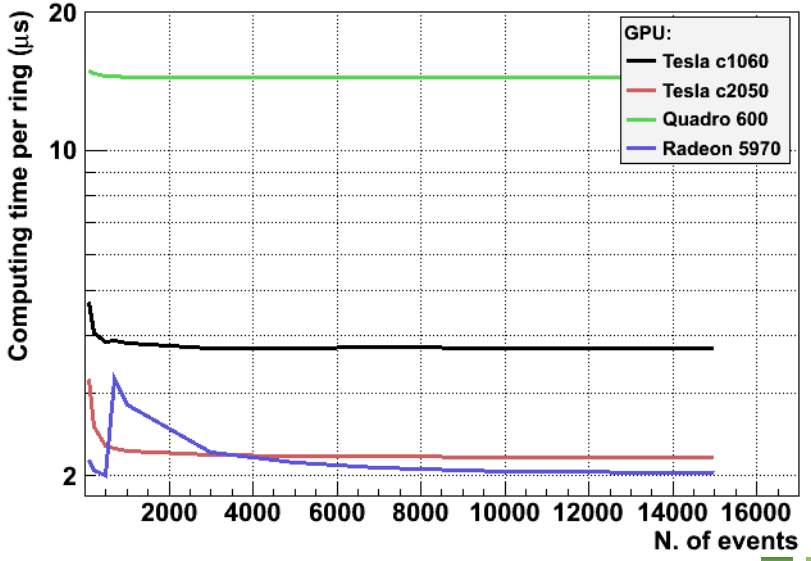
- MATH: Translation of the ring to centroid. In this system a least square method can be used. The circle condition can be reduced to a linear system , analitically solvable, without any iterative procedure.

Processing time

- Using Monte Carlo data, the algorithms are compared on Tesla C1060
- For packets of >1000 events, the MATH algorithm processing time is around 50 ns per event

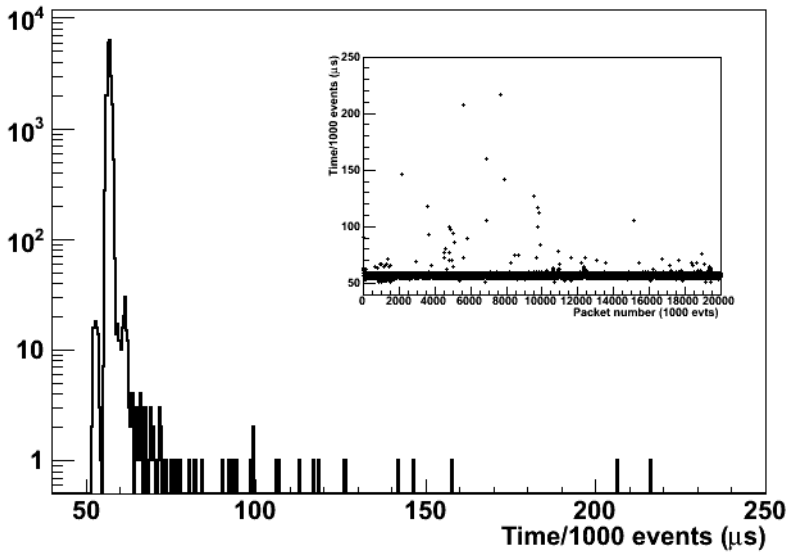


- The performance on DOMH (the most resource-dependent algorithm) is compared on several Video Cards
- The gain due to different generation of video cards can be clearly recognized.

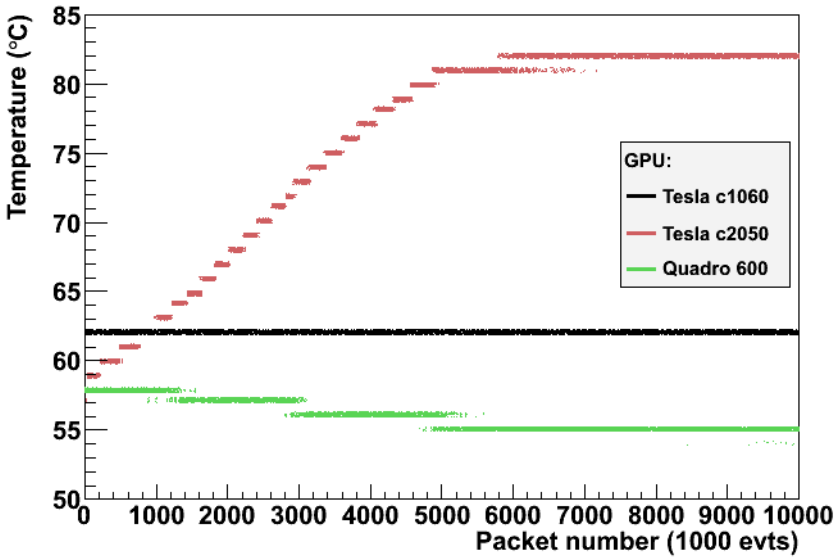


Processing time stability

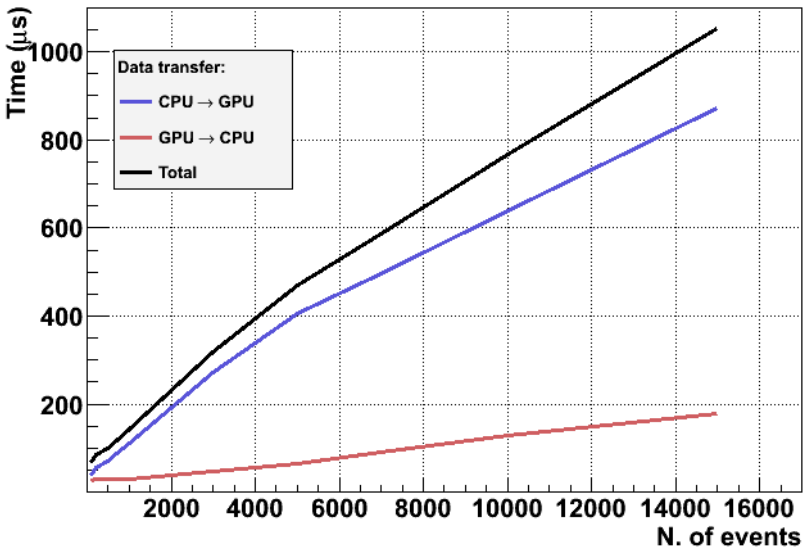
- The stability of the execution time is an important parameter in a synchronous system
- The GPU (Tesla C1060, MATH algorithm) shows a "quasi deterministic" behavior with very small tails.

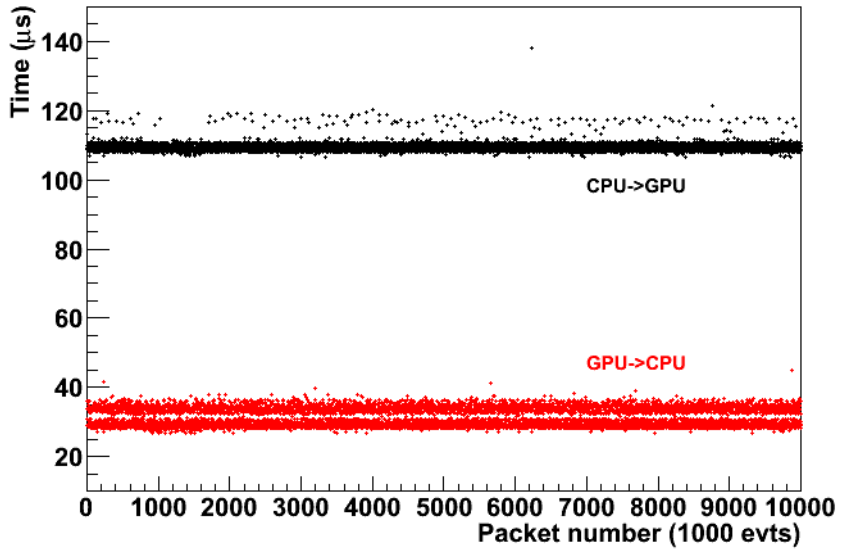


- The GPU temperature, during long runs, rises in different way on the different chips, but the computing performances aren't affected.



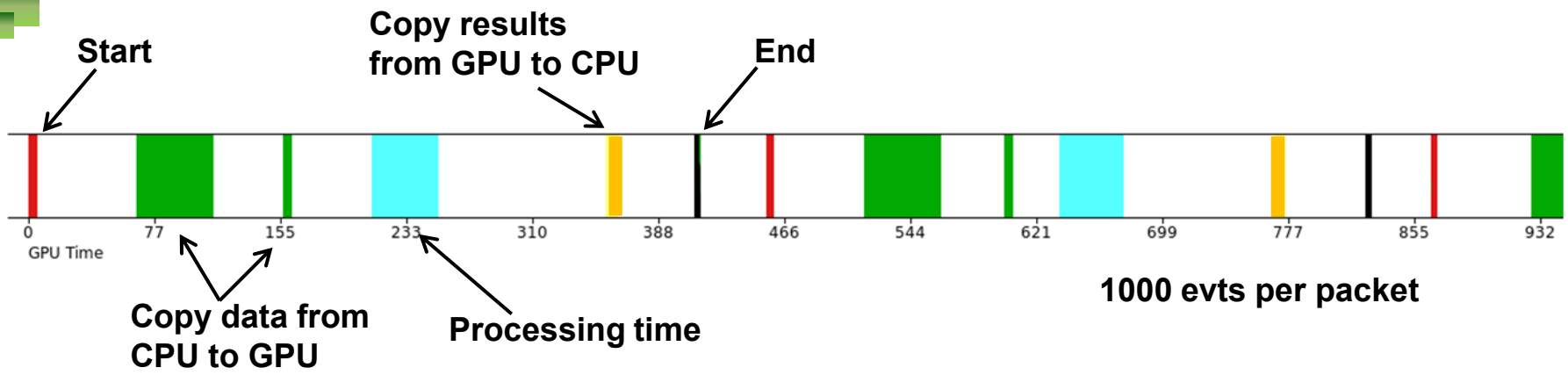
Data transfer time



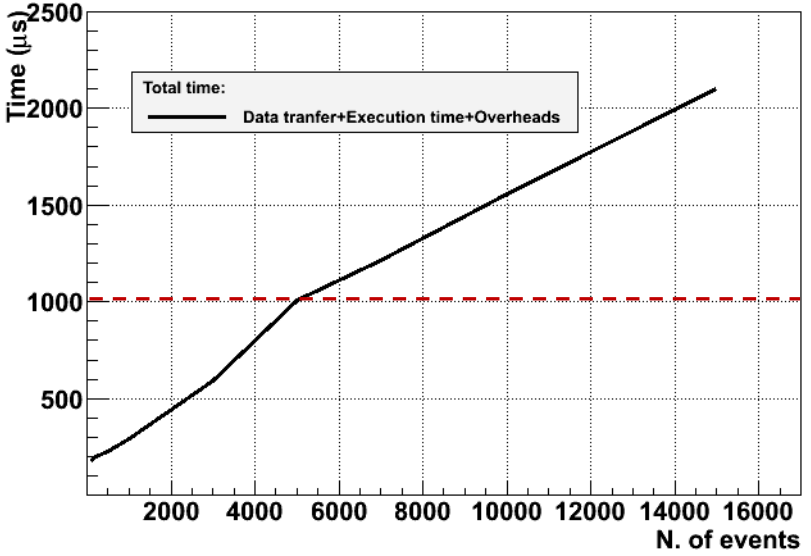


- The data transfer time significantly influence the total latency
- It depends on the number of events to transfer
- The transfer time is quite stable (double peak structure in GPU→CPU transfer)
- Using page locked memory the processing and data transfer can be parallelized (double data transfer engine on Tesla C2050)

Total GPU time to process N events

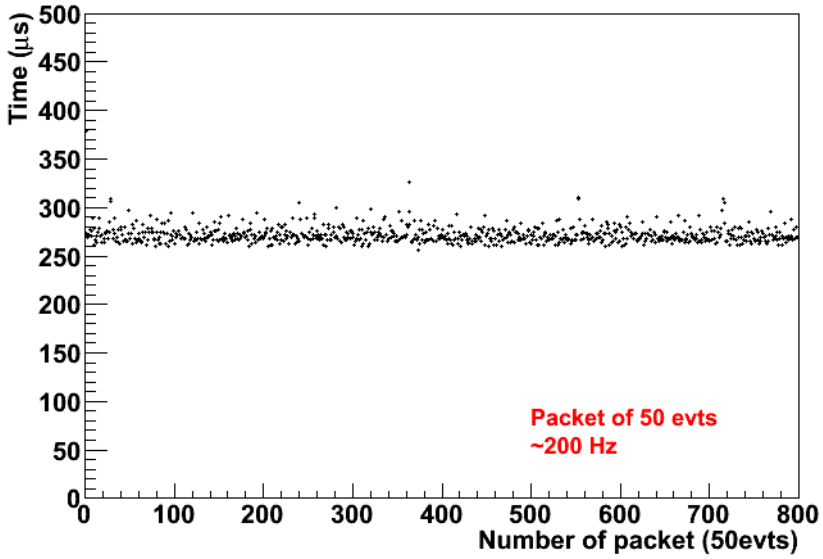


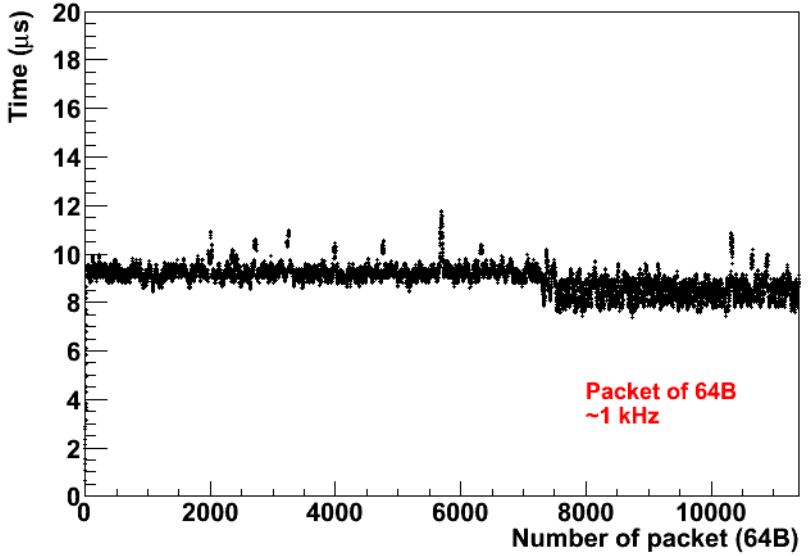
- The total latency is given by the transfer time, the execution time and the overheads of the GPU operations (including the time spent on the PC to access the GPU): for instance ~300 us are needed to process 1000 events



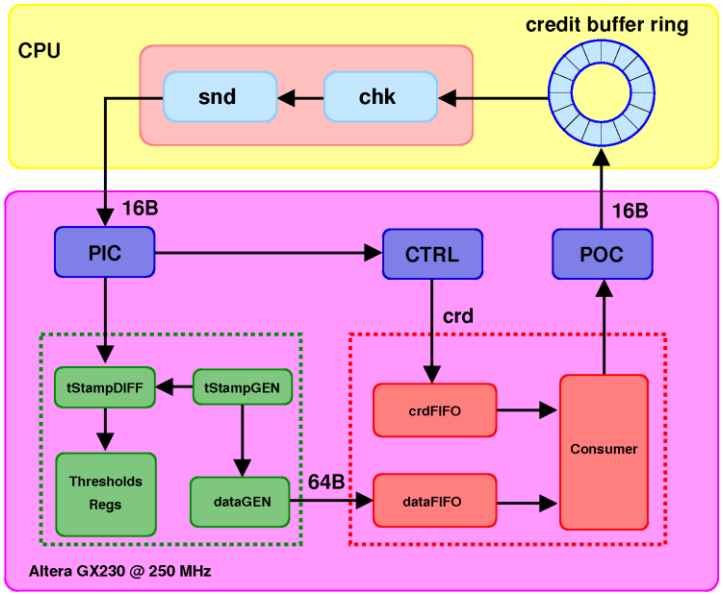
GPU Latency measurement

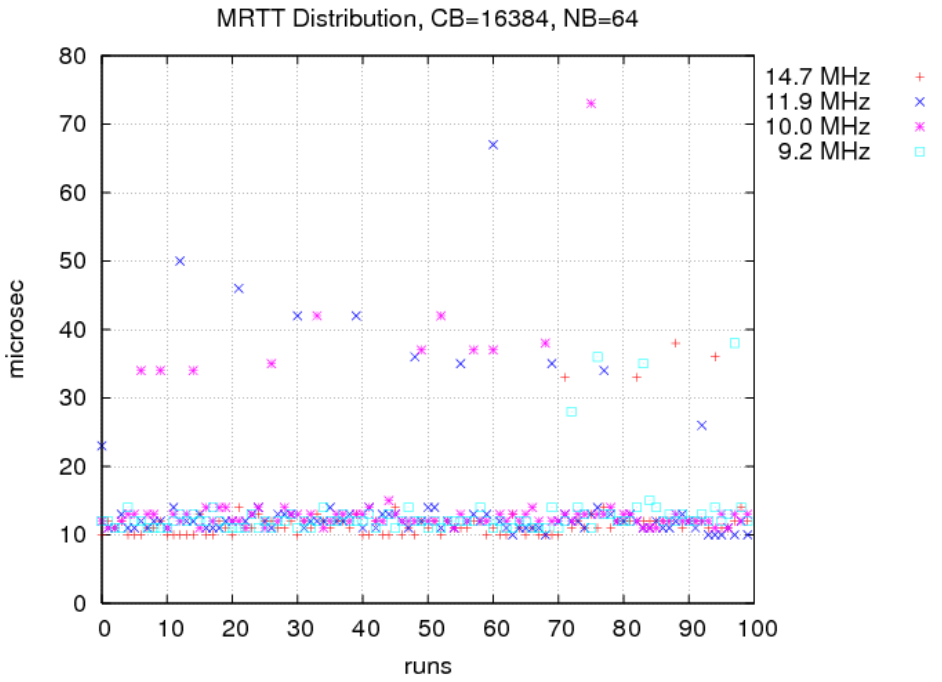
- Other contributions to the total time in a working system:
  - Time from the moment in which the data are available in user space and the end of the calculation: O(50 us) additional contribution. (plot 1)
  - Time spent inside the kernel, for the management of the protocol stack :~10 us for network level protocols, could improve with Real Time OS and/or FPGA in the NIC (plot 2)
  - Transfer time from the NIC to the RAM through the PCI express bus





Data transfer time inside the PC

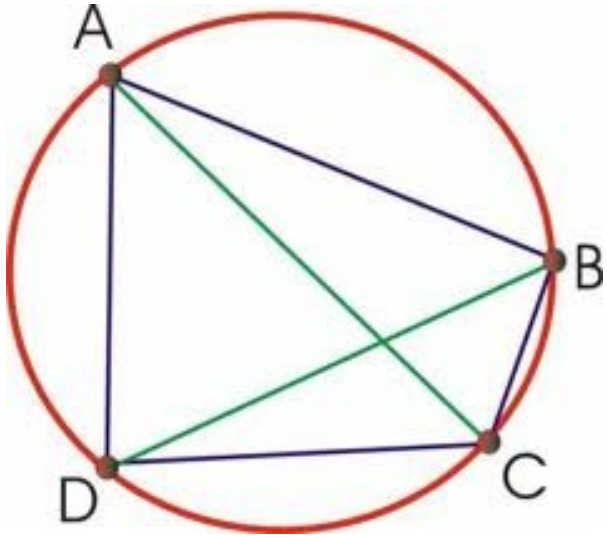




- ALTERA Stratix IV GX230 development board to send packets (64B) through the PCI express
- Direct measure (inside the FPGA) of the maximum round trip time (MRTT) at different frequencies
- No critical issues, the fluctuations could be reduced using Real Time OS

GPU@L1 RICH trigger

- After the L0: ~50% 1 track events, ~50% 3 tracks events
- Most of the 3 tracks, which are background, have max 2 rings per spot
- Standard multiple rings fit methods aren't suitable for us, since we need:
  - Trackless
  - Non iterative
  - High resolution
  - Fast: ~1 us (1 MHz input rate)

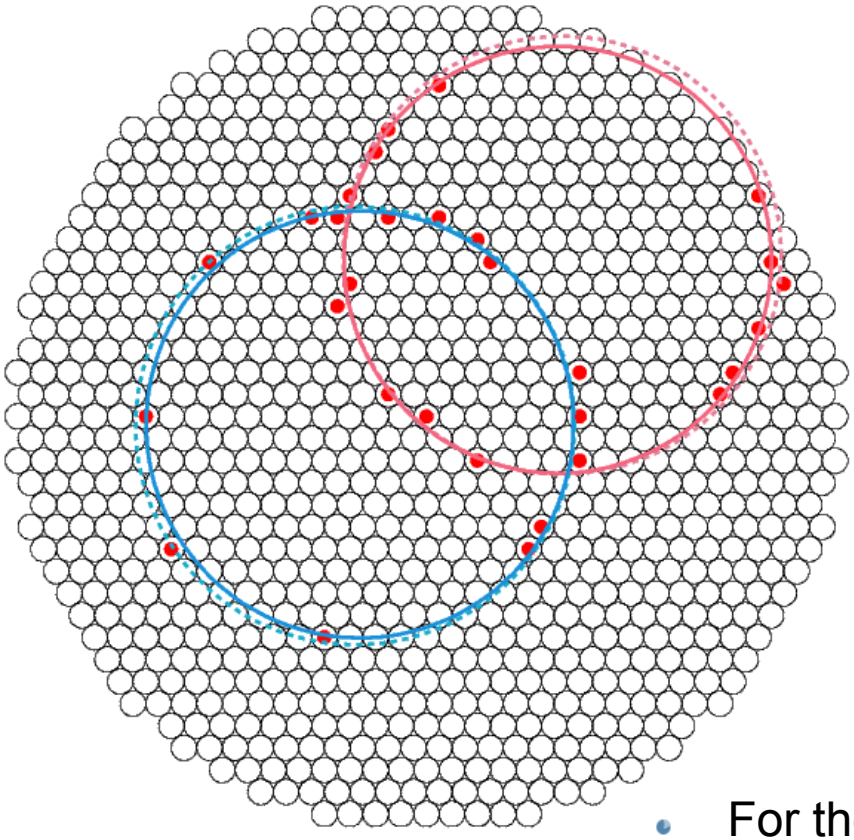


- New approach→ use the Ptolemy's theorem (from the first book of the Almagest)

"A quadrilater is cyclic (the vertex lie on a circle) if and only if is valid the relation:

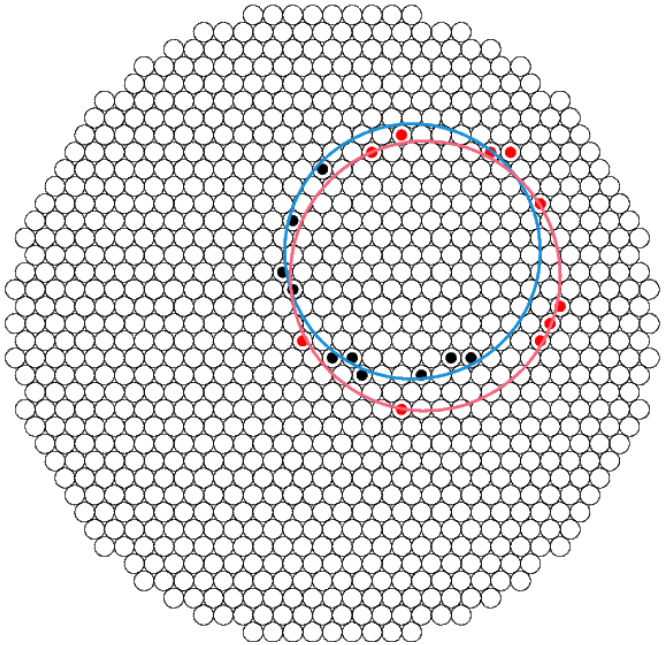
$$AD*BC+AB*DC=AC*BD$$ "

Almagest algorithm description



- Select a triplet randomly (1 triplet per point = N+M triplets in parallel)
- Consider a fourth point: if the point doesn't satisfy the Ptolemy theorem reject it
- If the point satisfy the Ptolemy theorem, it is considered for a fast algebraic fit (i.e. math, riemann sphere, tobin, …). Continue with the next point.
- Each thread (one thread for each hit) converges to a candidate center point. Each candidate is associated to Q quadrilaters contributing to his definition
- For the center candidates with Q greater than a threshold, the points at distance R (the candidate radius) are considered for a more precise re-fit. All the other points are associated to the second ring

Almagest algorithm results



- The real position of the two generated rings is:

1→ (6.65, 6.15) R=11.0

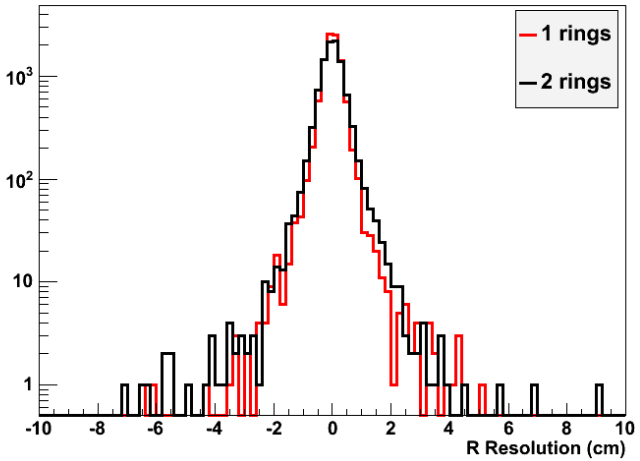
2→ (8.42,4.59) R=12.6

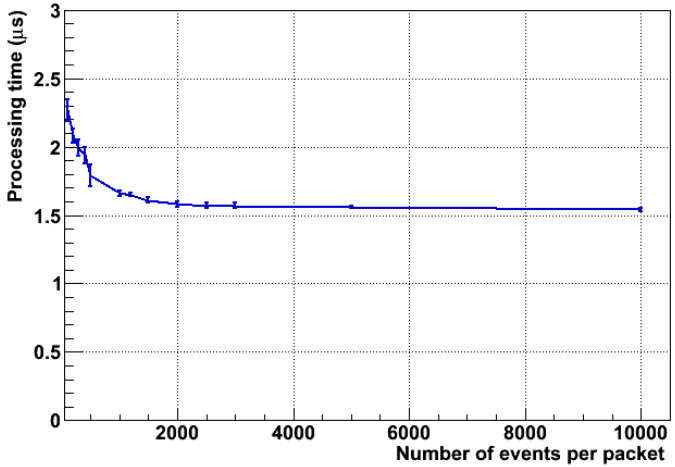
- The fitted position of the two rings is:

1→ (7.29, 6.57) R=11.6

2→ (8.44,4.34) R=12.26

- Fitting time on Tesla C1060: 1.5 us/event





Conclusions

- The GPUs can be used in trigger system to define high quality primitives
- In asynchronous software levels the use of video cards is straightforward, while in synchronous hardware levels issues related to the total latency have to be evaluated: in our study no showstoppers evident for a system with 10 MHz input rate and 1 ms of allowed latency
- First application: ring finding in the RICH of NA62
  - ~50 ns per ring @L0 using a parallel algebraic fit (single ring)
  - ~1.5 us per ring @L1 using a new algorithm (double ring)
- Further applications: Online track fitting (Kalman filter, cellular automaton, …), CHOD online corrections, …
- A full scale "demonstrator" will be ready for the technical runs next spring

References:

i) IEEE-NSS CR 10/2009: 195-198

ii) Nucl.Instrum.Meth.A628:457-460,2011

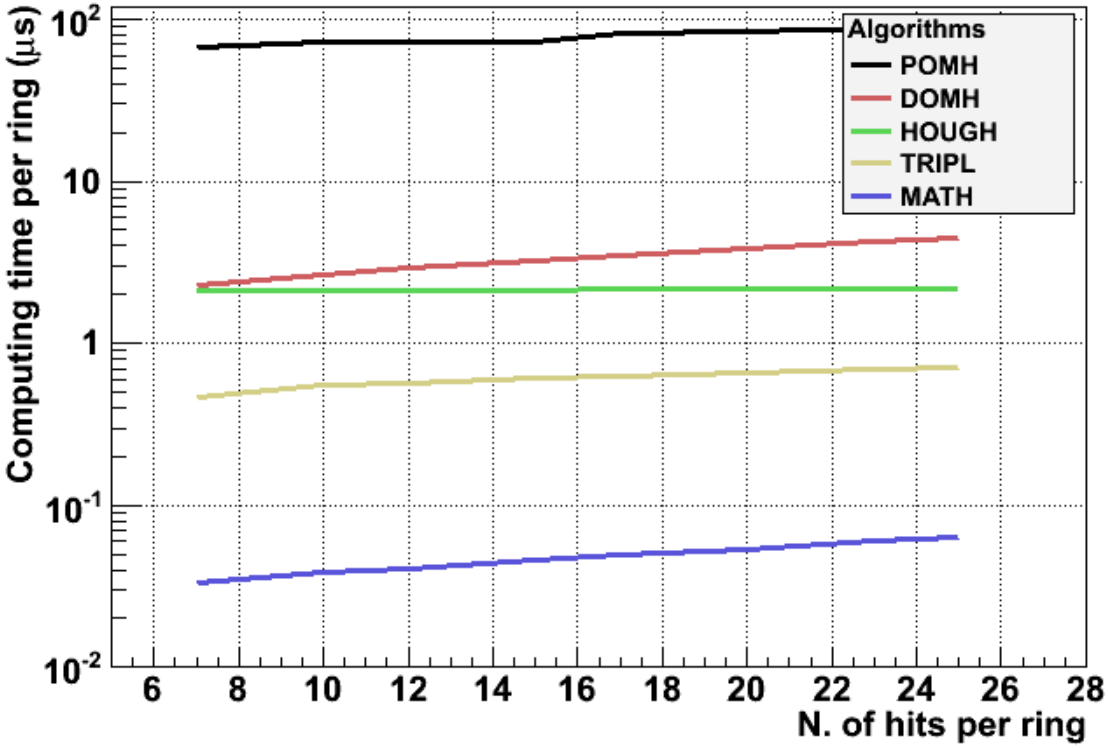
iii) Nucl.Instrum.Meth.A639:267-270,2011

iv) "Fast online triggering in high-energy physics experiments using GPUs" submitted to NIM

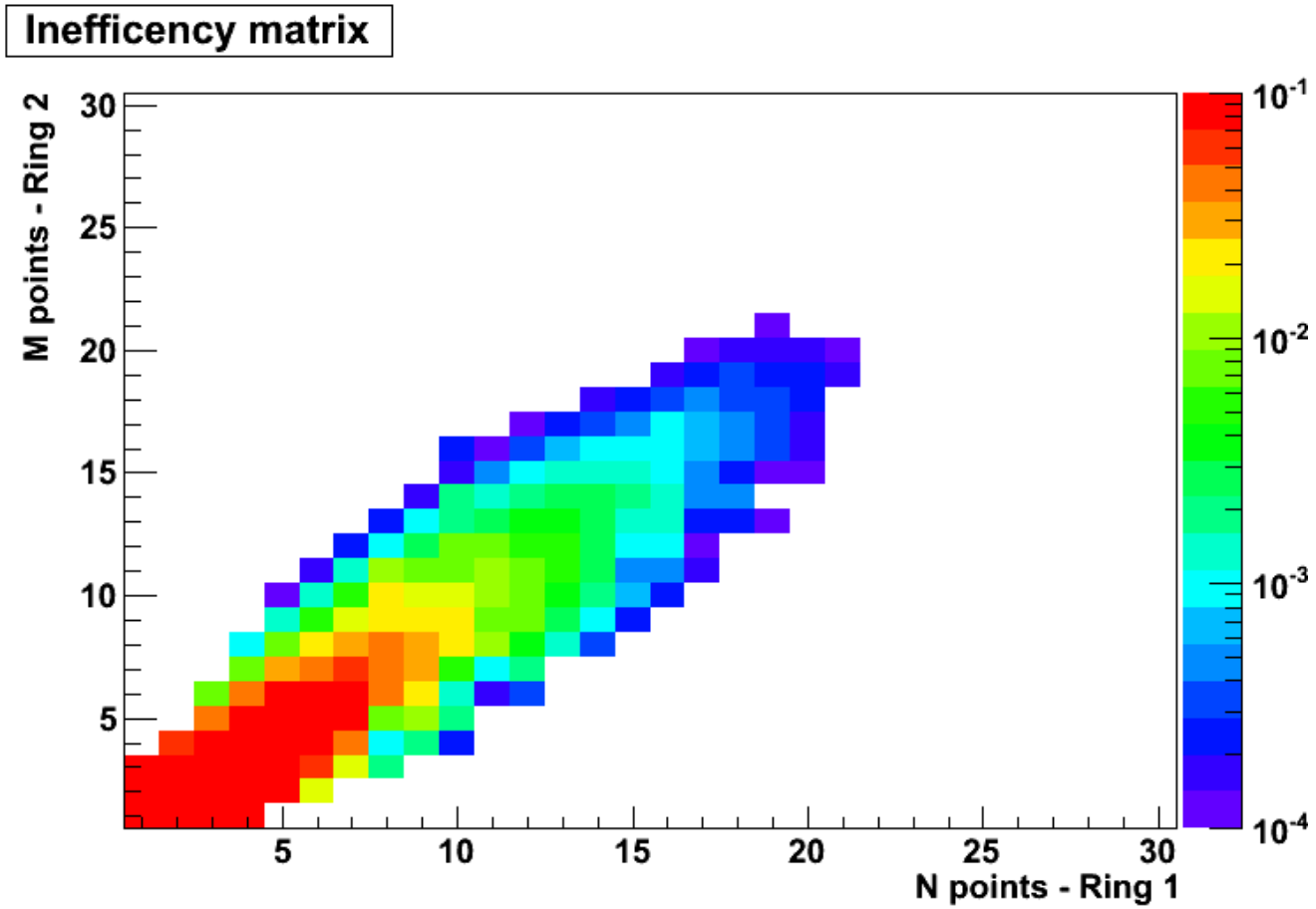
SPARES

Dependence on number of hits

- The execution time depends on the number of hits
- Different slope for different algorithms: depending on the number and the kind of operation in the GPU



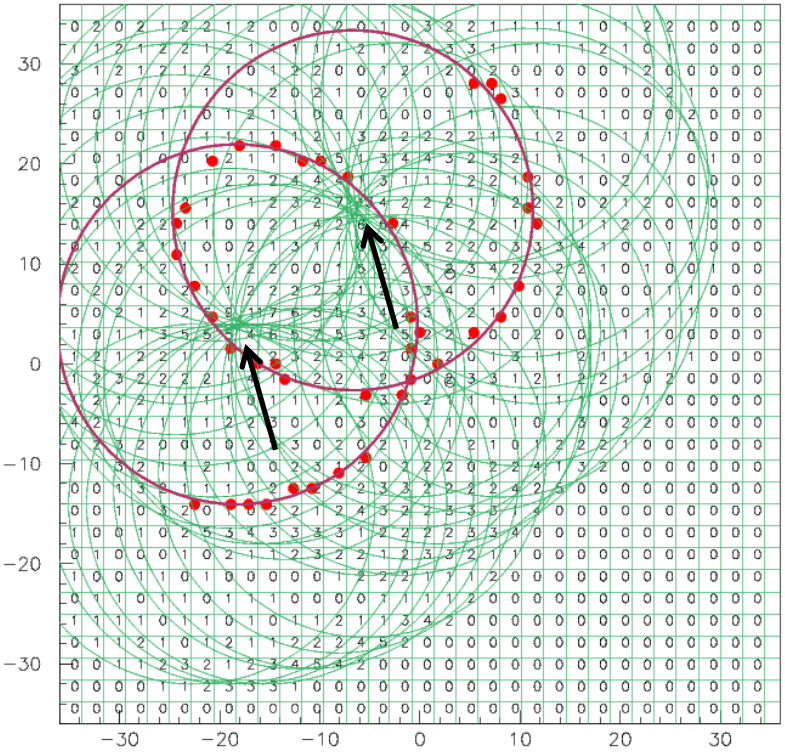
Almagest "a priori" inefficiency



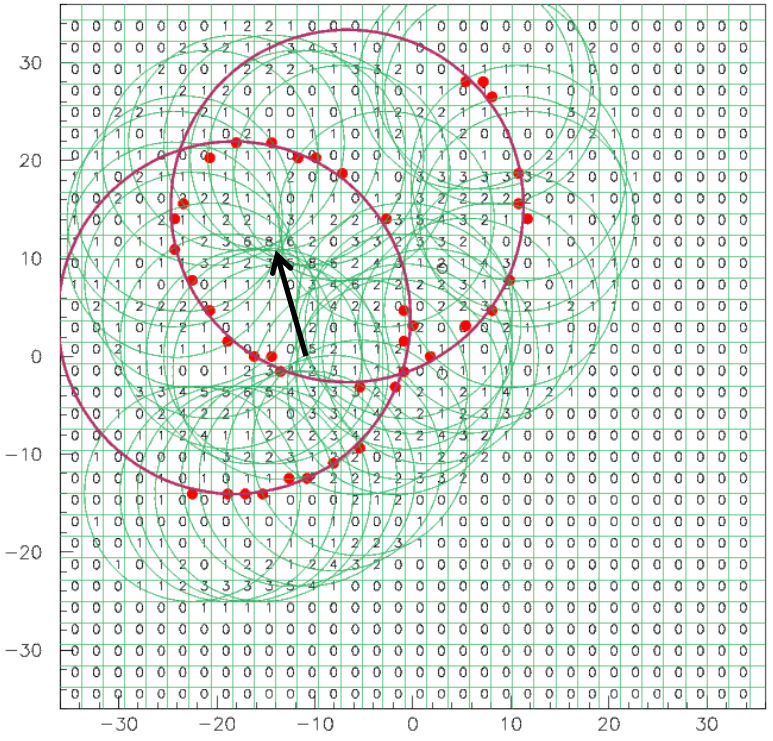
- In the N+M triplets randomly considered, at least one triplet have to lie on one of the two rings
- If none triplet is good the fit doesn't converge: inefficiency
- The inefficiency is negligible either for high number of hits or for "asymmetrics" rings (N≠M)

Hough fit

18 cm test rings



12 cm test rings



- The extension of the Hough algorithm isn't easy
- Fake points appear at wrong test radius value → several seeds
- The final decision should be taken with a hypothesis testing, looping on all the possible seeds (for instance: the correct result is given by the two seeds that use all the hits) → complicated and time expensive

Trigger bandwidth