Real-time use of GPUs in High-Energy Physics experiments

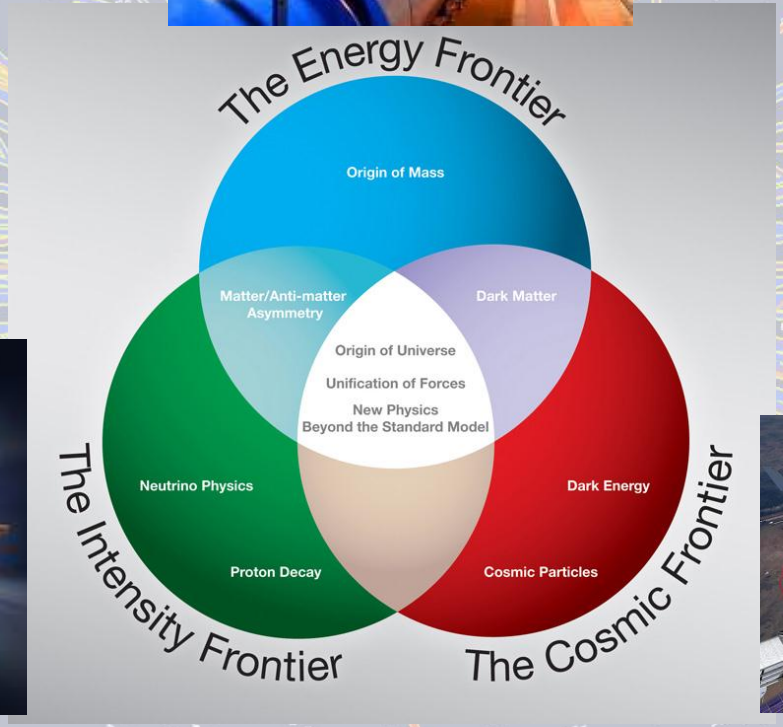
Marco S. Sozzi

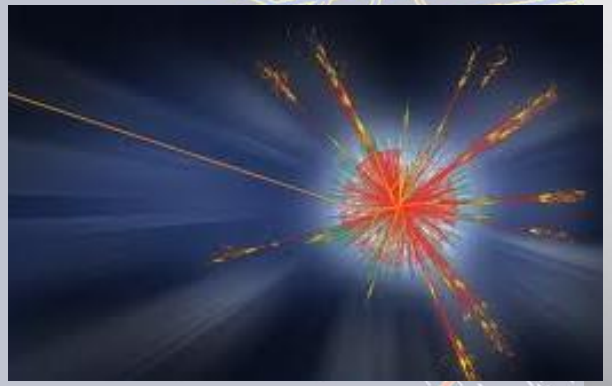
University of Pisa
Istituto Nazionale di Fisica Nucleare
CERN

With:
G. Lamanna, J. Pinzino, F. Pantaleo *(Pisa U. and CERN)*

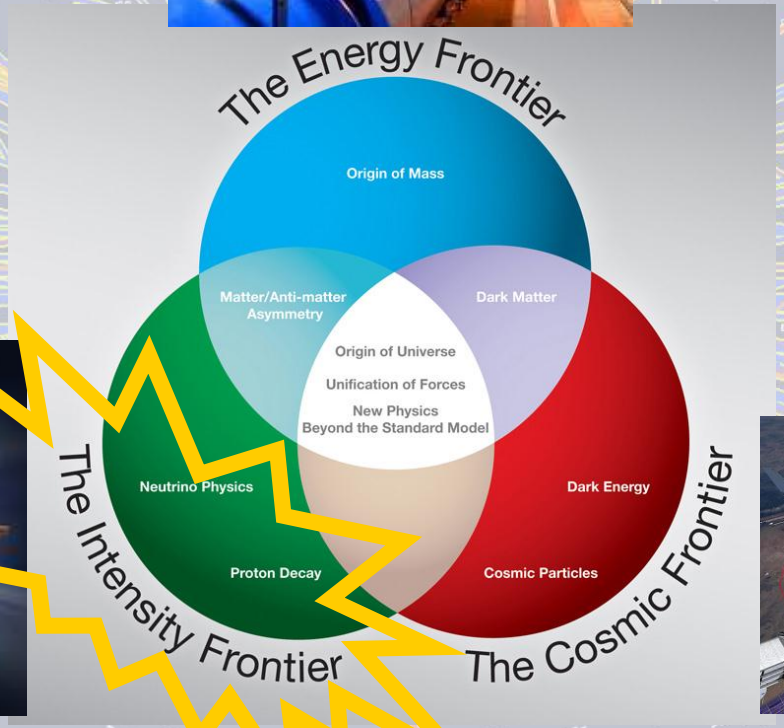
GPU TECHNOLOGY CONFERENCE

The frontiers of physics

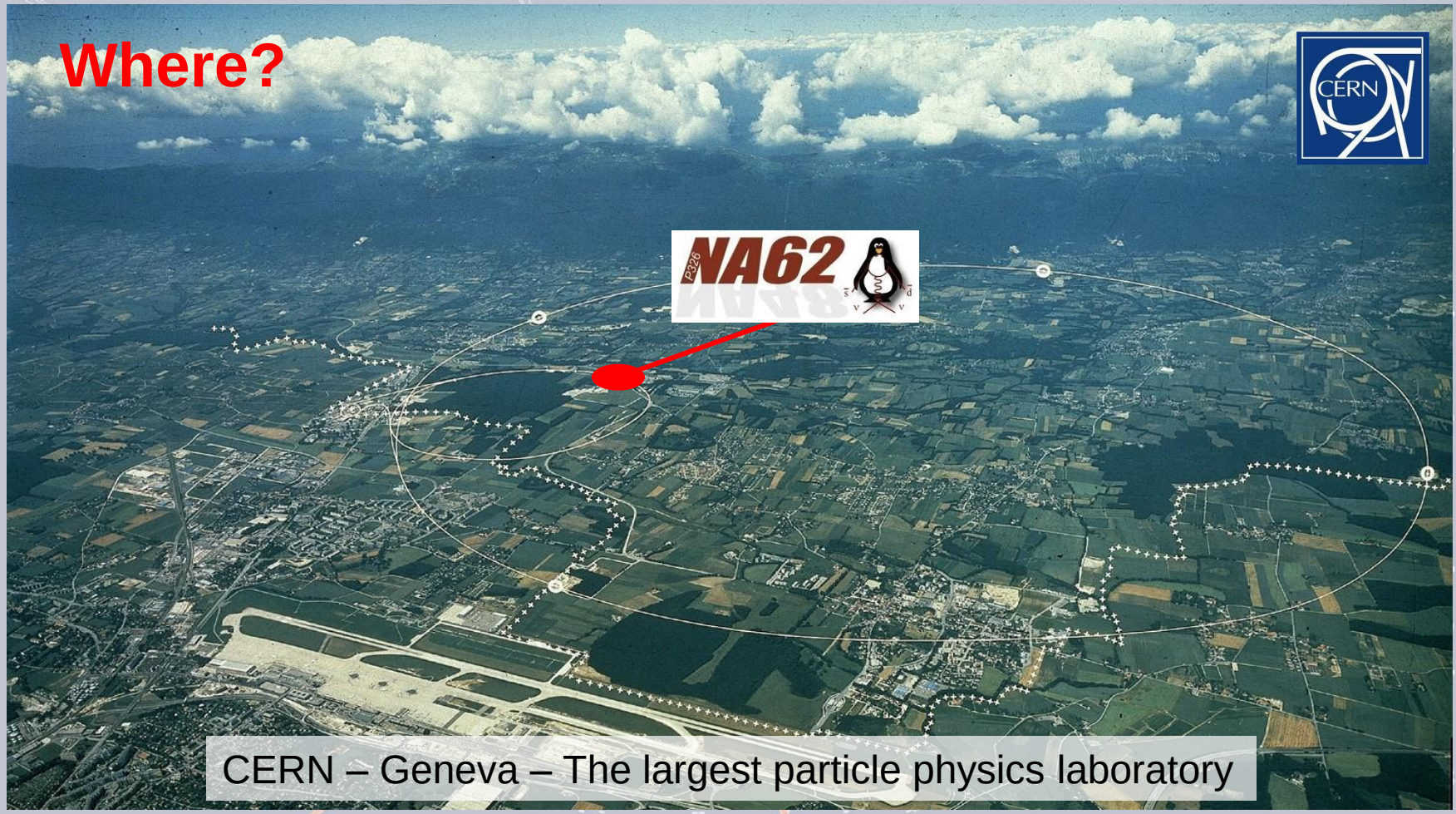




The frontiers of physics



Where?


CERN
NA62
P326
CERN – Geneva – The largest particle physics laboratory

The CERN NA62 experiment

155 people

from 27 institutions

Belgium, Bulgaria, Czech Republic, Germany, Italy, Mexico, Romania, Russia, Slovakia, Switzerland, United Kingdom, USA

Approved 2007 – In preparation – First tests 2012 – First data taking 2014

NA62 experiment - the goal

Precision measurement of the ultra-rare decay process $K^+ \rightarrow \pi^+ \nu\nu$

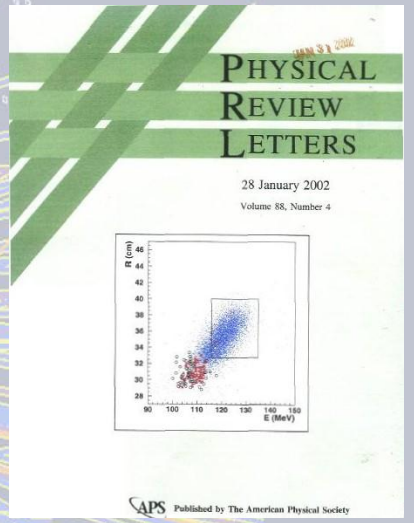
First observed at Brookhaven National Labs 1997-2001

Why?

Extremely sensitive to any unknown ***new particles***,
even ***way beyond*** the reach of direct experimental searches
at new and forthcoming accelerators

But…

Very intense primary proton beam: **10^{13} protons/s** onto solid target
Very intense secondary beam: **10^{9} particles/s**
Many (uninteresting) events: **10^{7} decays/s**

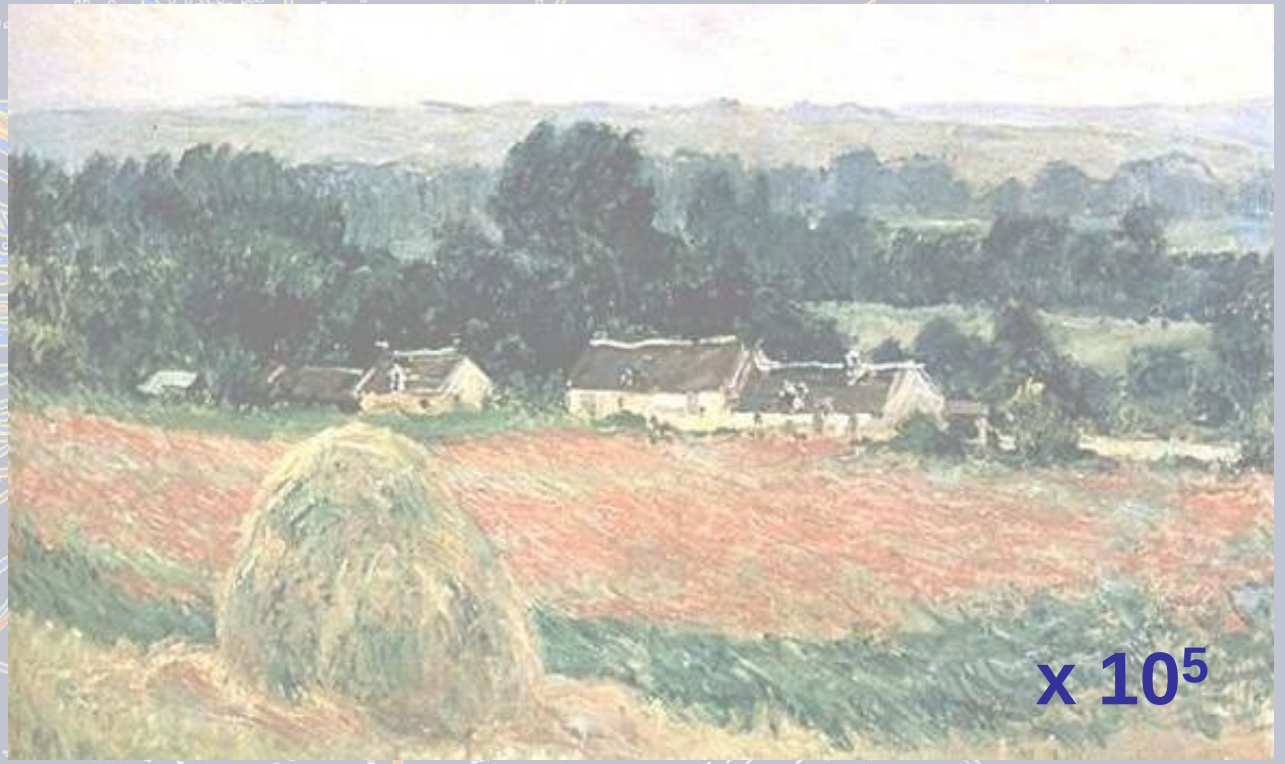


The problem

How rare is ultra-rare? **1 target event in 10^{10} particle decays**

Aim to collect about 100 events in 2-3 years of data-taking

Record digital data from detectors and quickly filter it down to manageable levels



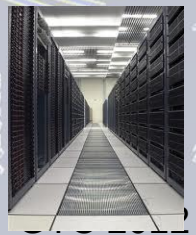
Triggering in HEP experiments

Multi-level triggering reduces the amount of data to manageable levels in a set of successive stages

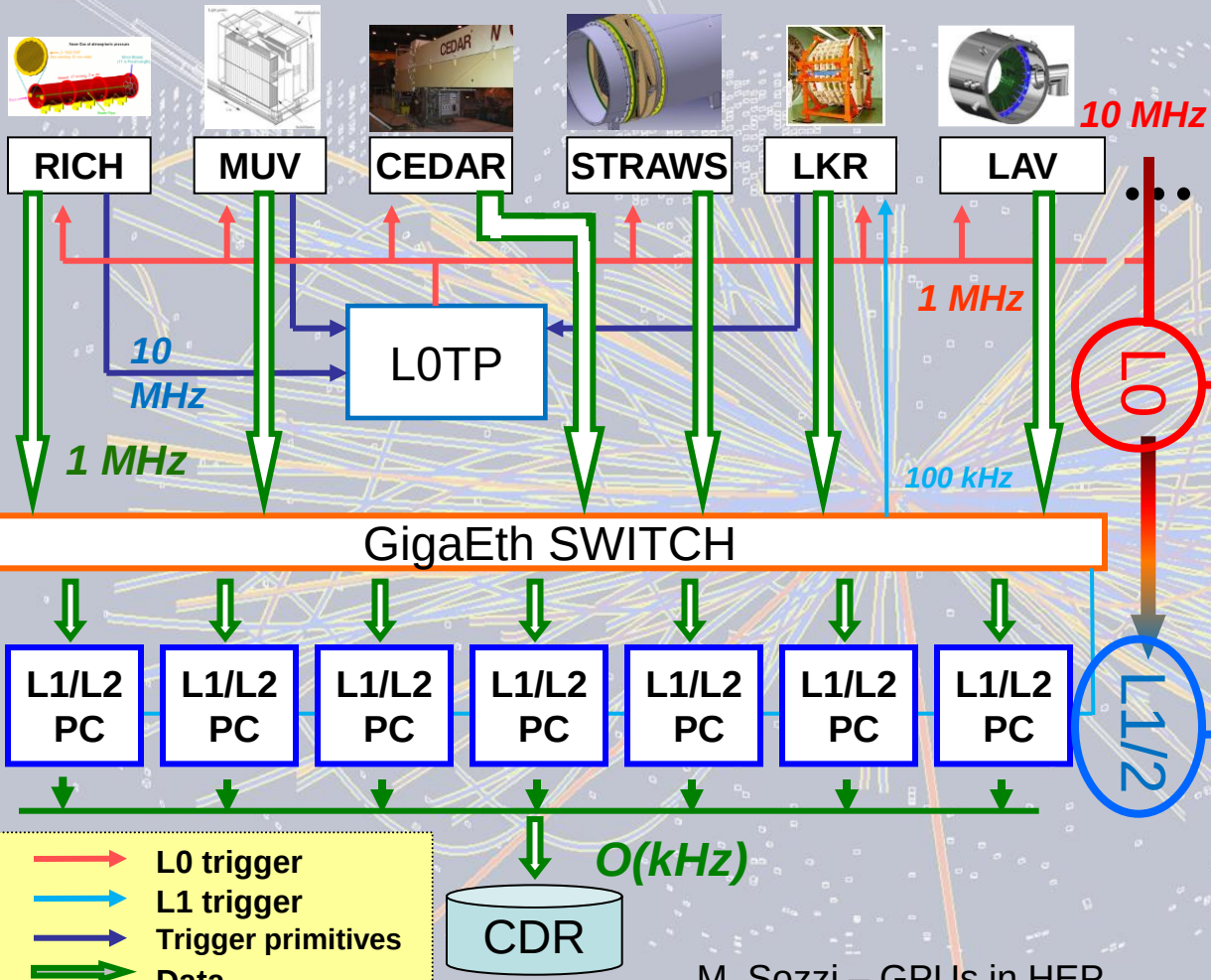
Lower level: custom hardware, simple patterns in reduced information from a few fast detectors;
►move full detector data from buffers to PC memories if OK

Higher levels: hierarchical chain of simplified reconstruction algorithms on some detectors, specific features of interesting events, using switched computing farms;
►store detector data to permanent storage if OK

Offline analysis: full reconstruction algorithms
►reduced data samples for further physics analysis



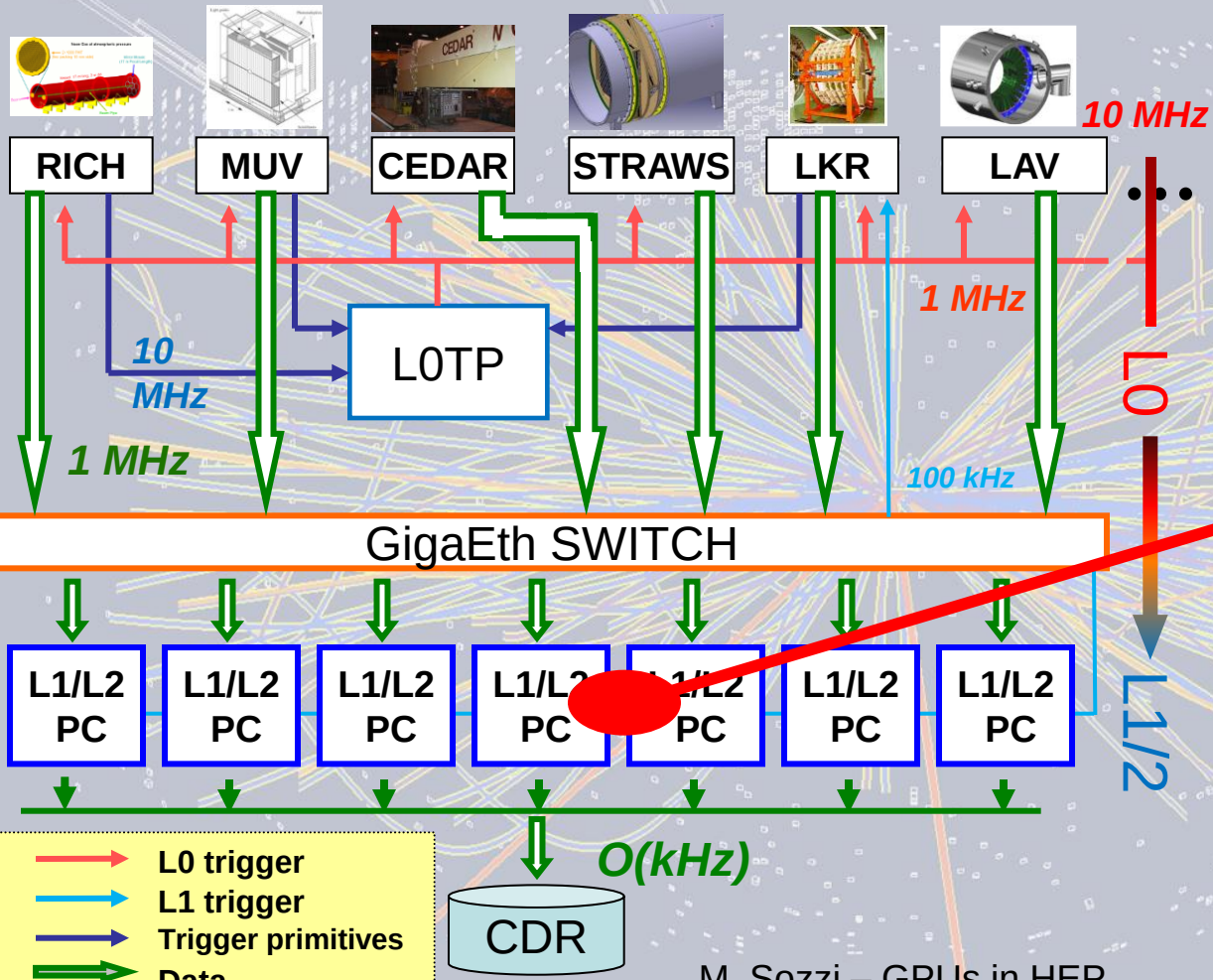
NA62 Trigger and Data Acquisition



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

Where GPUs? The easy spot



The use of GPUs at the software trigger levels (L1/2) is straightforward:

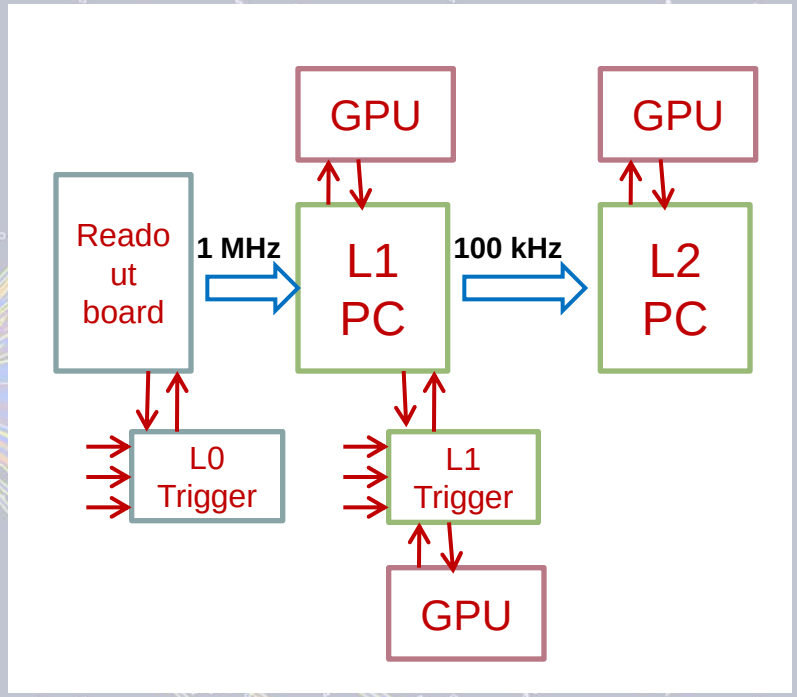
exploit the additional GPU computing power to reduce the number of PCs in the farms

GPUs in high-level trigger stages

Parallelization:

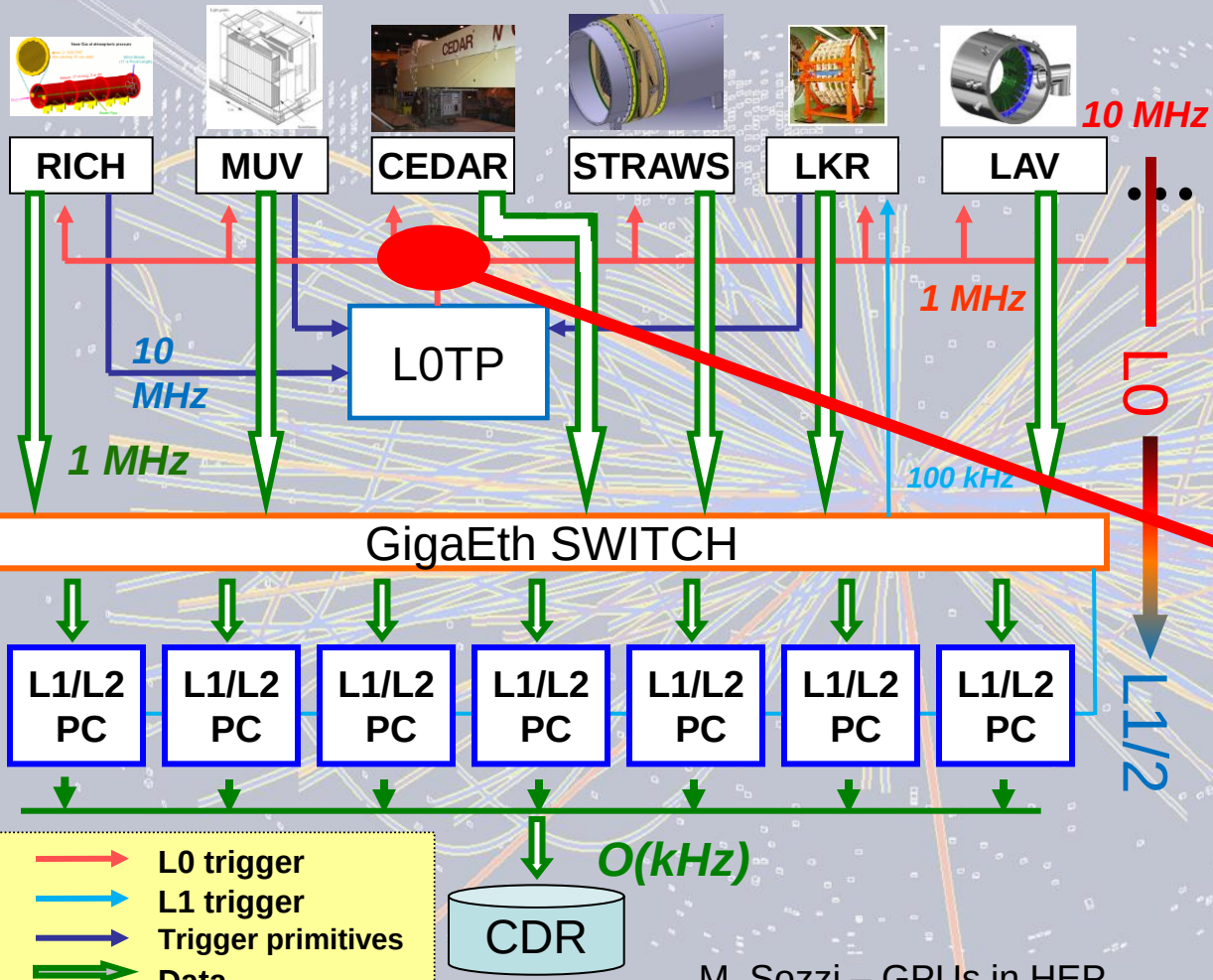
►Some event classification algorithms can be parallelized, but

►Intrinsic parallelization of independent events analysis



Work in progress, expect large performance boost for parallelizable algorithms. Not the focus of this talk.

Where GPUs? The real challenge



The use of GPUs at the first trigger level (L0) is much more challenging

(thus quite a more significant paradigm-shift)

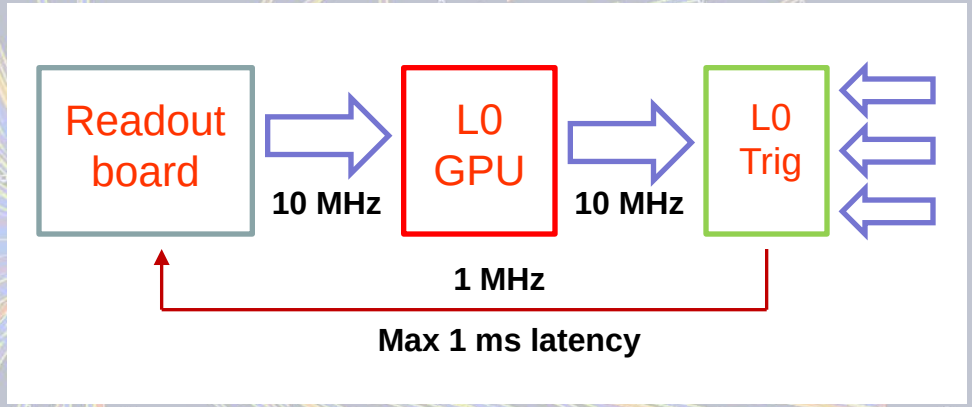
GPUs at lowest trigger stage

The issues:

▲Fixed and small latency: maximum memory size of the L0 circular event buffers

▲Deterministic behavior: synchronous trigger

▲Very fast algorithms: high rate

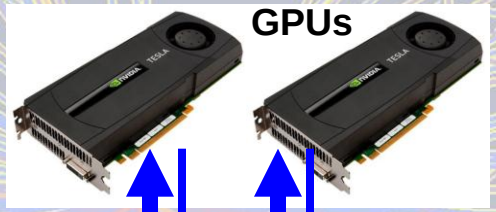


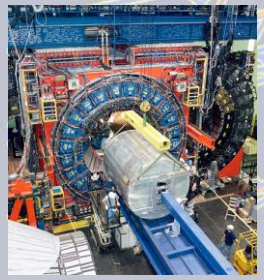
- Absolute performance measurements (rather than boost factor wrt CPU algorithms)
- No data exchange between units

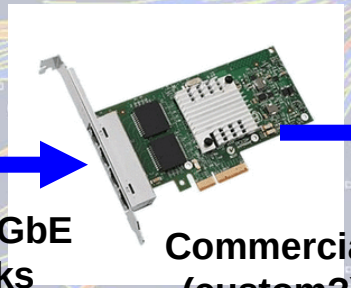
The picture

Try to replace the lowest level trigger with a GPU-based system
Use standard off-the-shelf components as much as possible for

- ease of programming
- scalability
- easy upgrade
- cost effectiveness





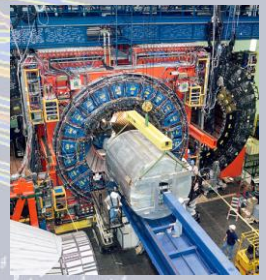


1-10 GbE links

Commercial (custom?) NIC

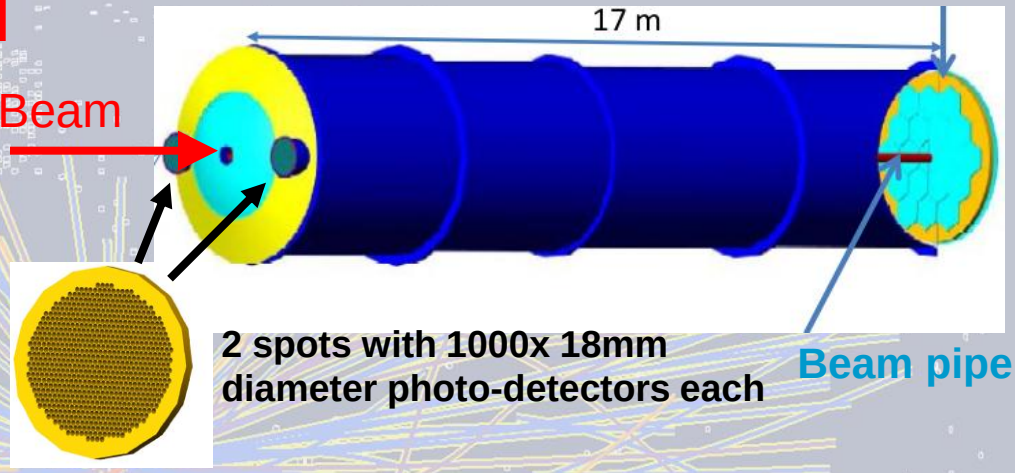
Linux PC

Commercial NIC (or custom adapter)



First case study: RICH

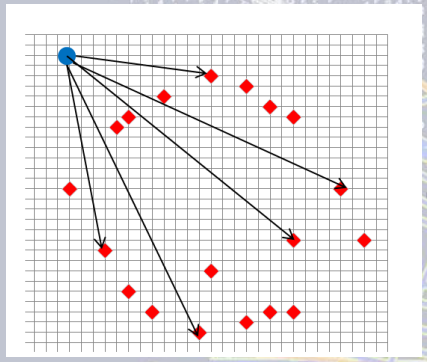
Pattern-finding in ring-imaging Čerenkov detector



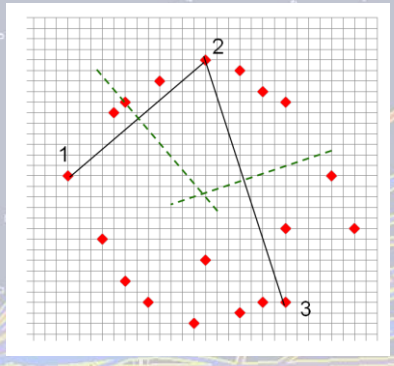
Passing relativistic particle produces cone-shaped light flash in Neon-filled vessel, resulting in partial circular hit pattern on photo-detectors

- 100 ps time resolution
- 10 MHz particle rate
- 20 photons (hits) on average
- Information reduced to ~40 byte/event in FPGAs

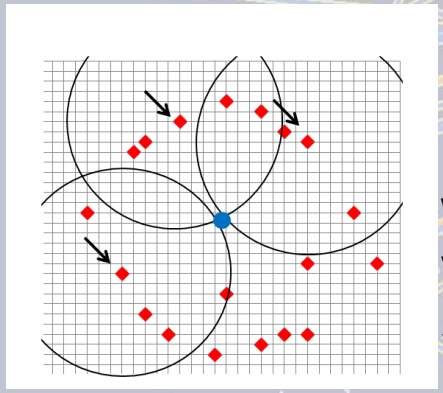
Ring-finding algorithms



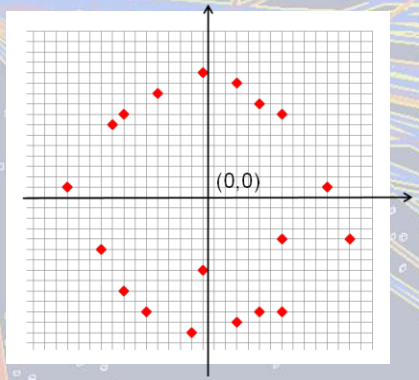
DOMH/POMH: evaluate hit distances from all centers on a grid and perform voting



TRIPL: average hit centers from randomly selected triplets of hits



HOUGH: iterated binned Hough transform with stepping radii and voting in 3D parameter space



MATH: linearized non-iterative least-squares method, by translating origin to centroid of hits

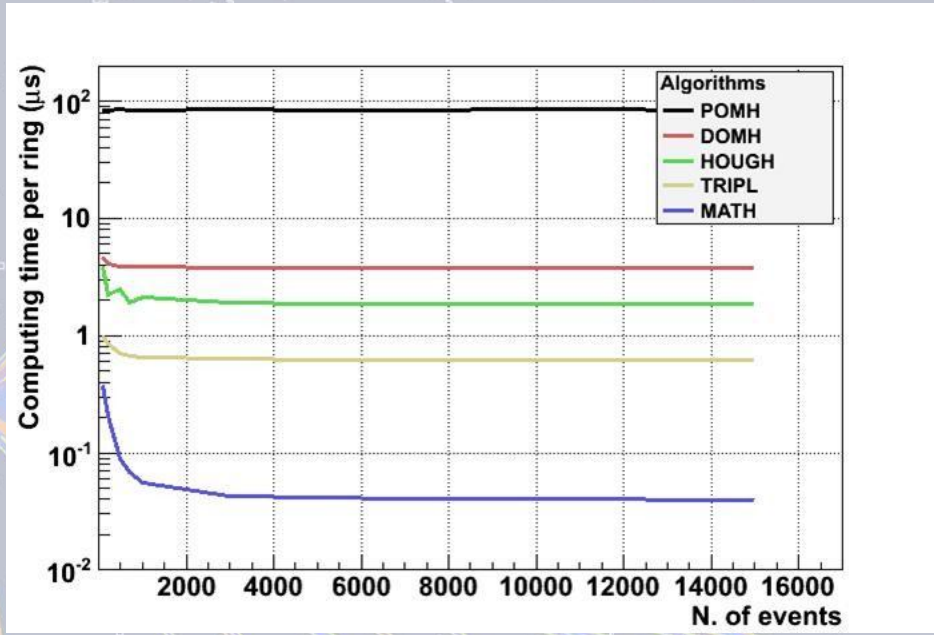
GPU performance

Test bench with single-ring simulated data running on NVIDIA Tesla C1060

~1000 events per set are enough

The fastest algorithm requires only 50 ns per event (ring)

Performances were compared on different GPU devices from different vendors and different generations (incl. NVIDIA Quadro 600, Tesla C1060, Tesla C2050)

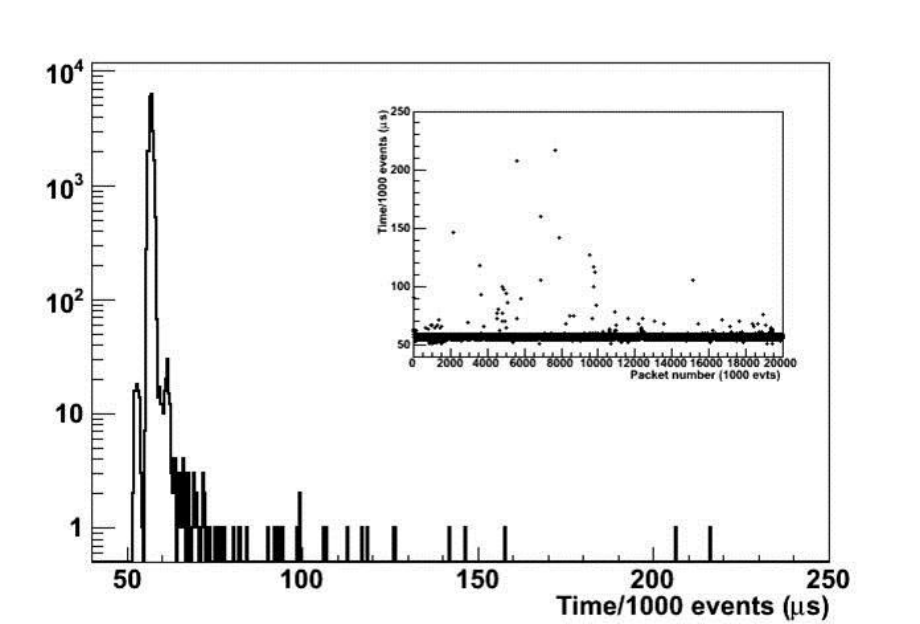


Processing time stability

Processing time stability is paramount in this hard-real time application

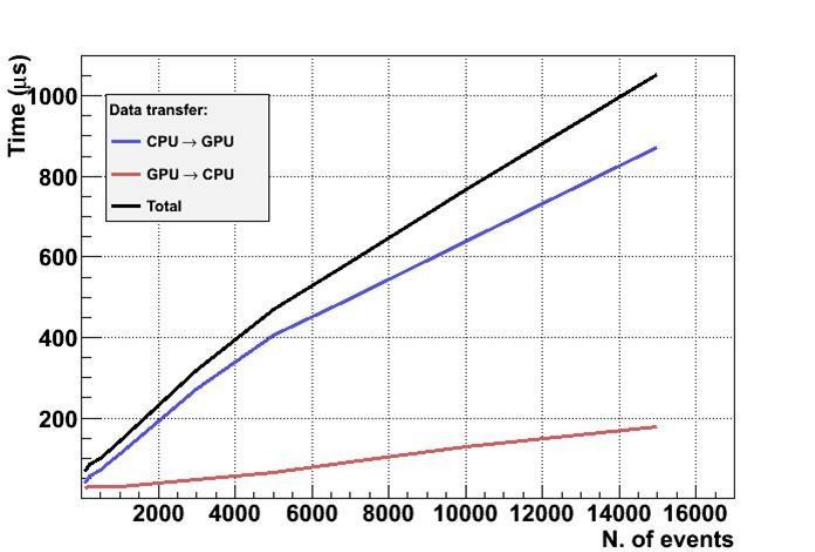
MATH algorithm on Tesla C1060 shows sufficiently small tails in time distribution

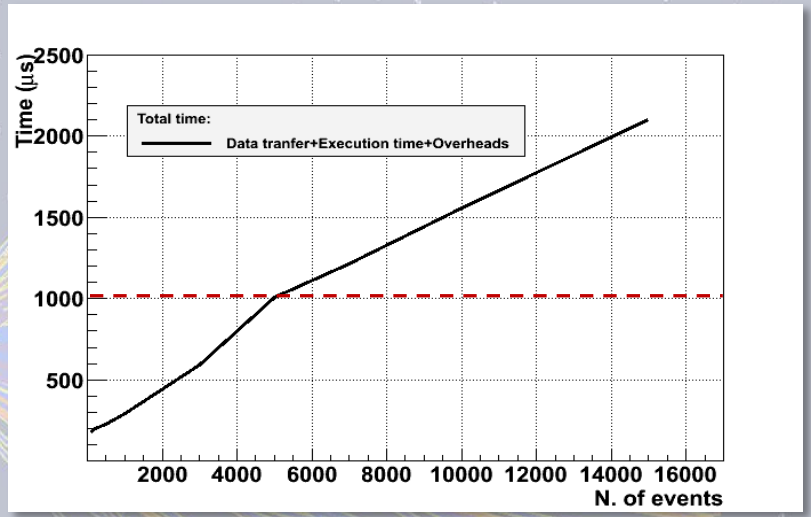
Study of temperature dependence in continuous long runs (hours): processing times not affected despite ~20°C GPU temperature rise

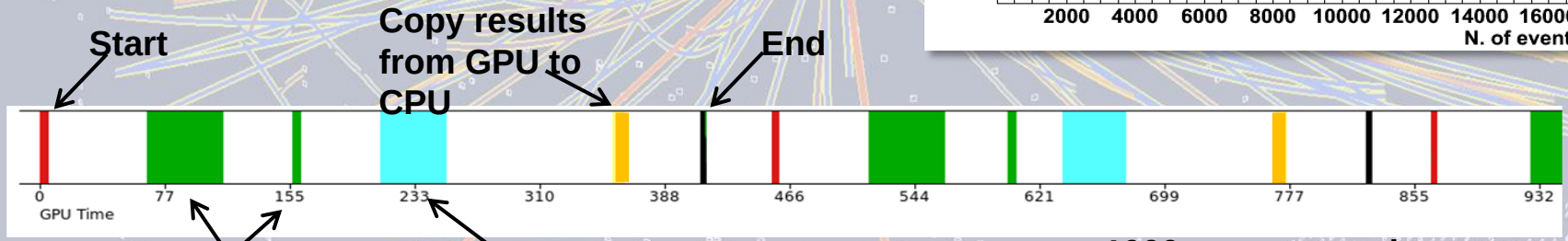


Data transfer times

High throughput in (~600 MB/s), small out. Page-locked memory and streaming.

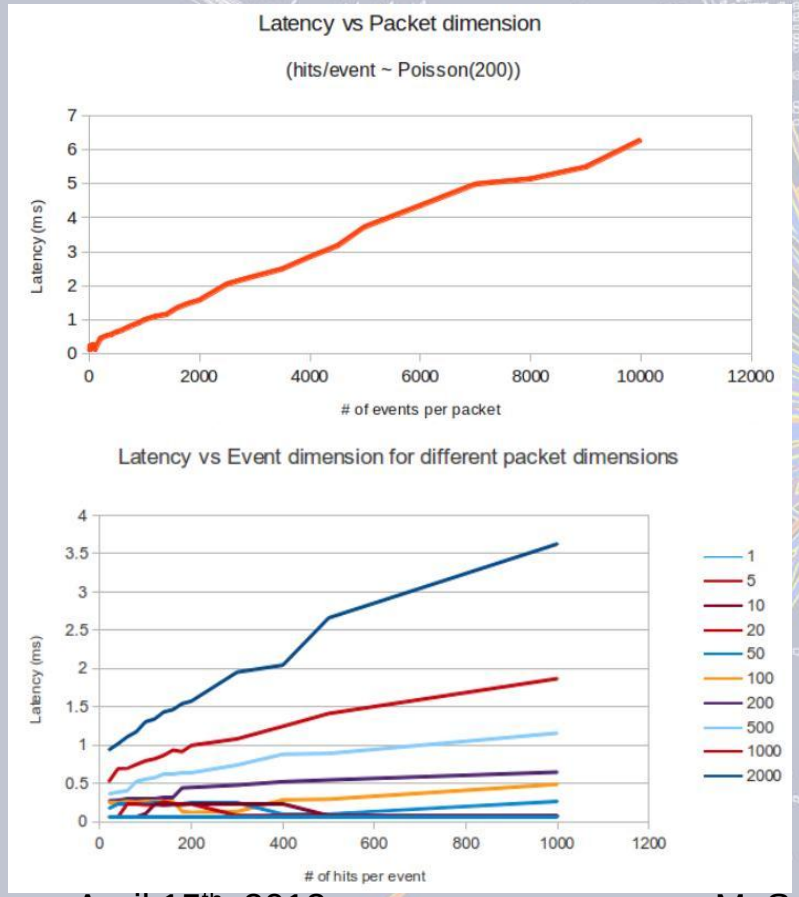


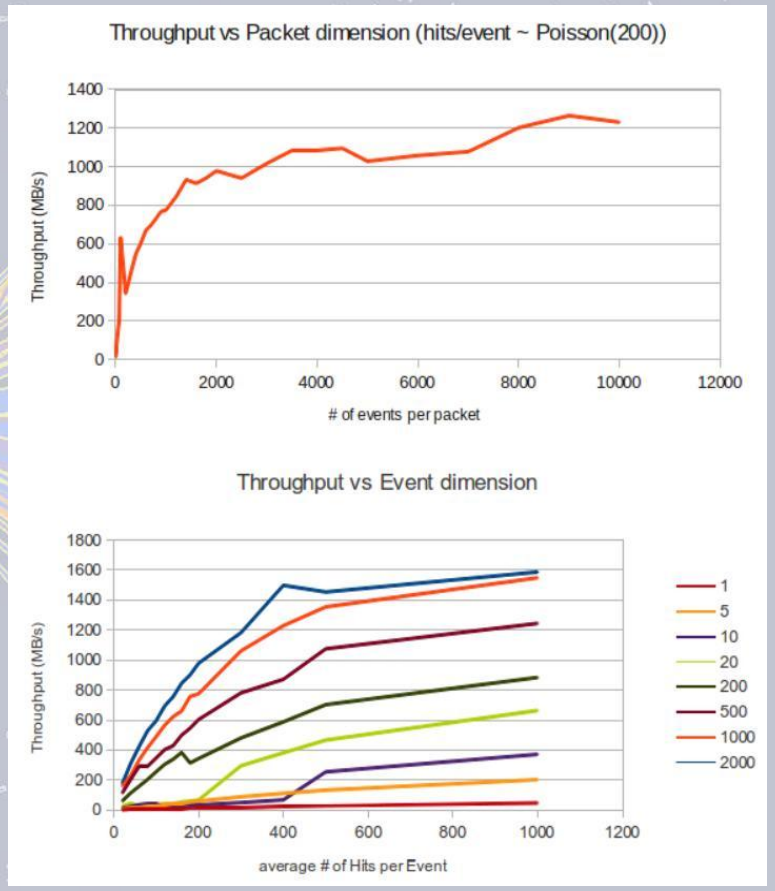




Packet size

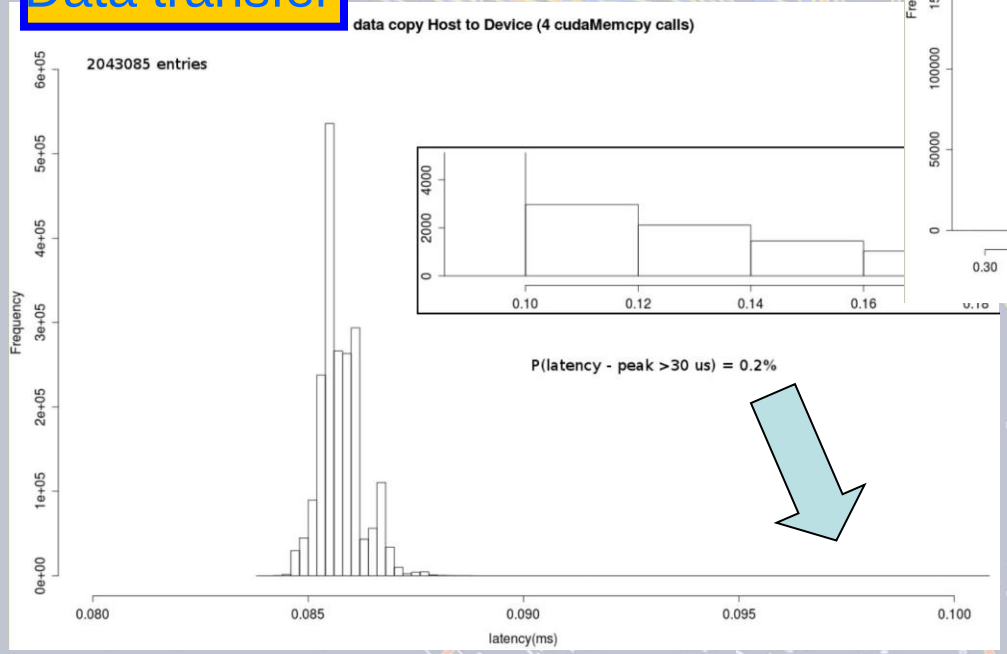
Latency vs. throughput compromise

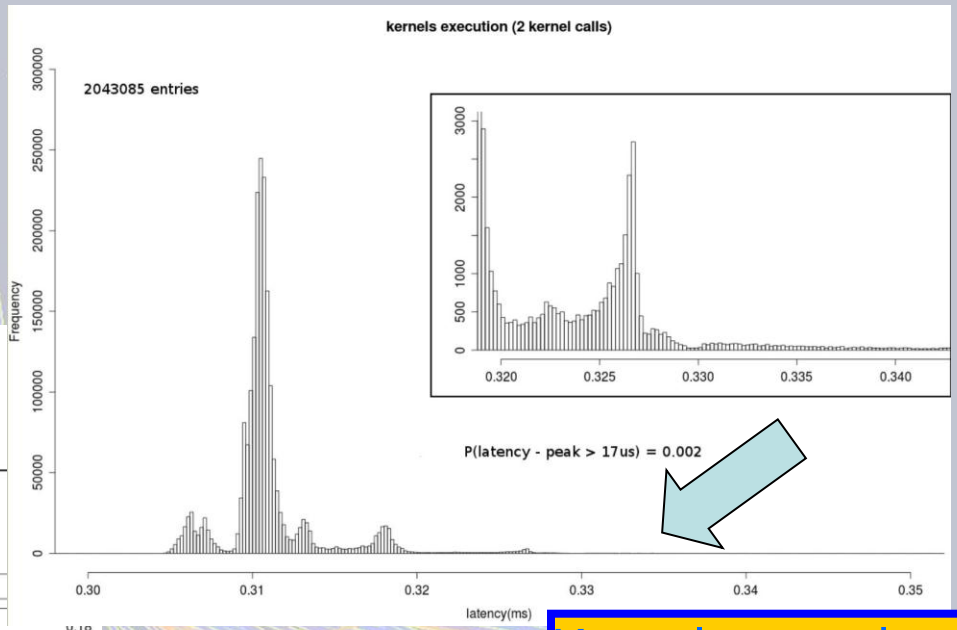




Time tails

Data transfer



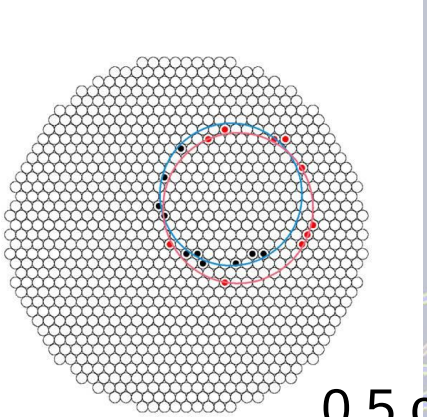


Kernel execution

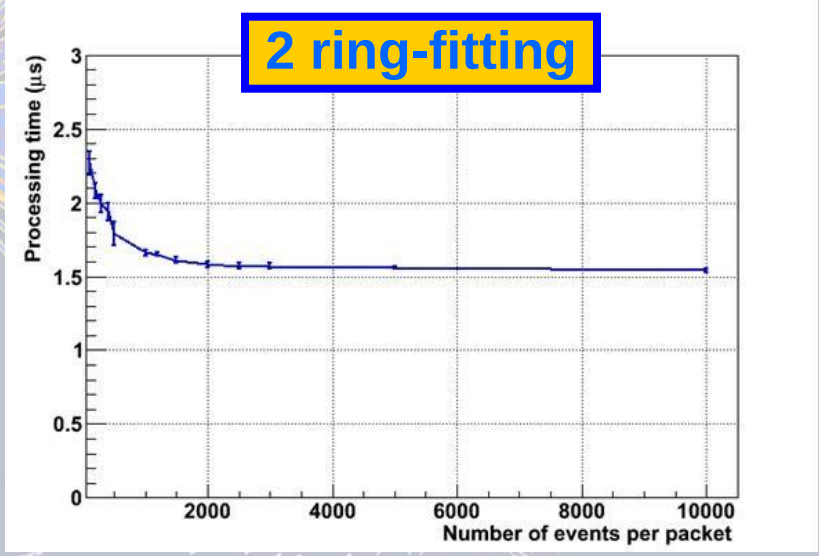
Multi-ring pattern finding

Rejection of multi-ring patterns
Standard algorithms not suitable.
Requirements:
- Seedless
- Non-iterative
- Fast

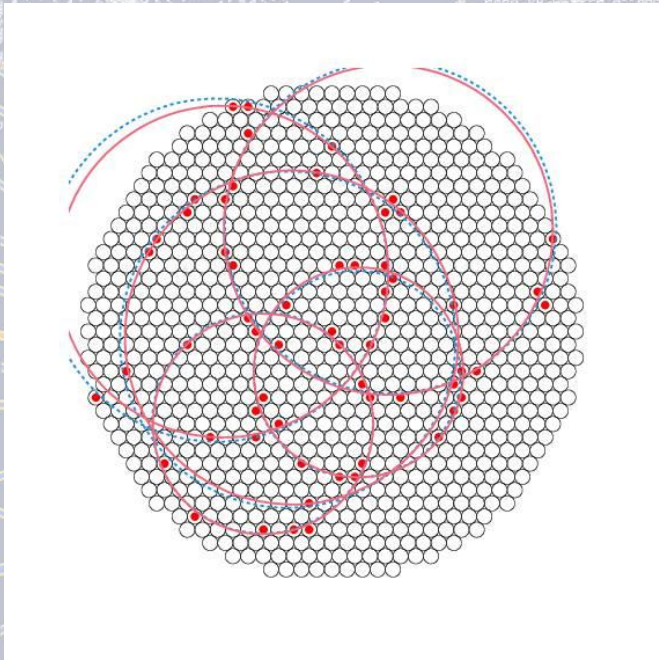
New algorithm based on Ptolemy's theorem on cyclic polygons: hits lying on a ring identify candidates, those above threshold are subject to further fit

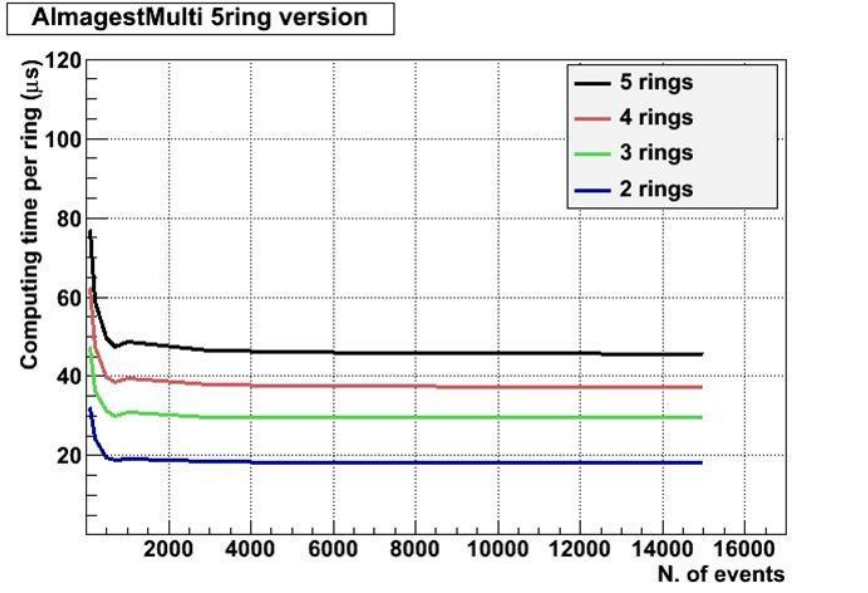


0.5 cm accuracy



Multi-ring parallel search



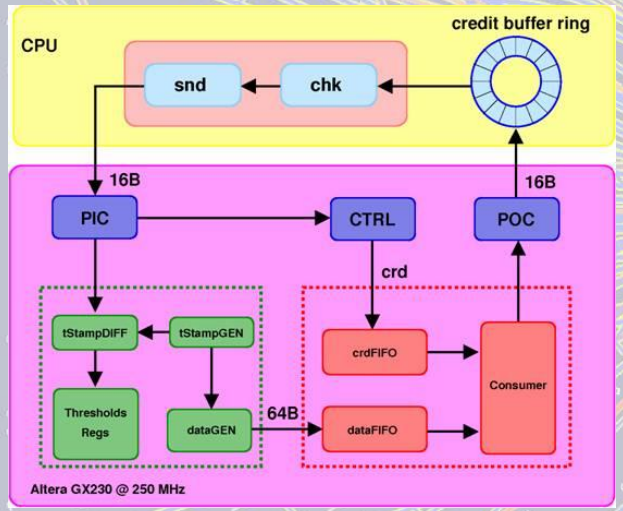


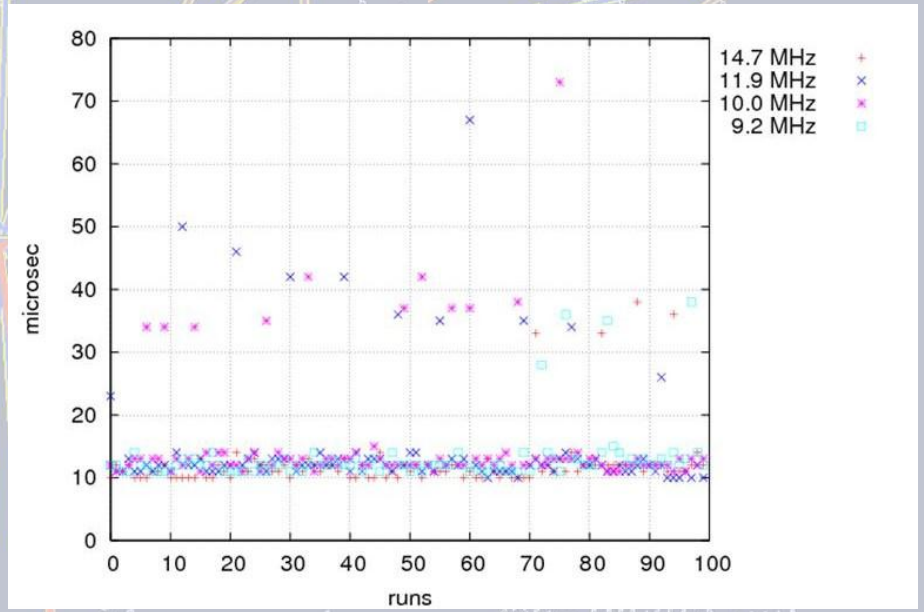
~90% efficiency on many-ring events

[2-3 rings: 2ms on standard double-core PC]

Host response fluctuations

Absolute maximum round-trip time (MRTT) of transfer on PCIe 2.0 measured from Altera Stratix IV GX FPGA to Linux PC (user space application) and back for 64B packets



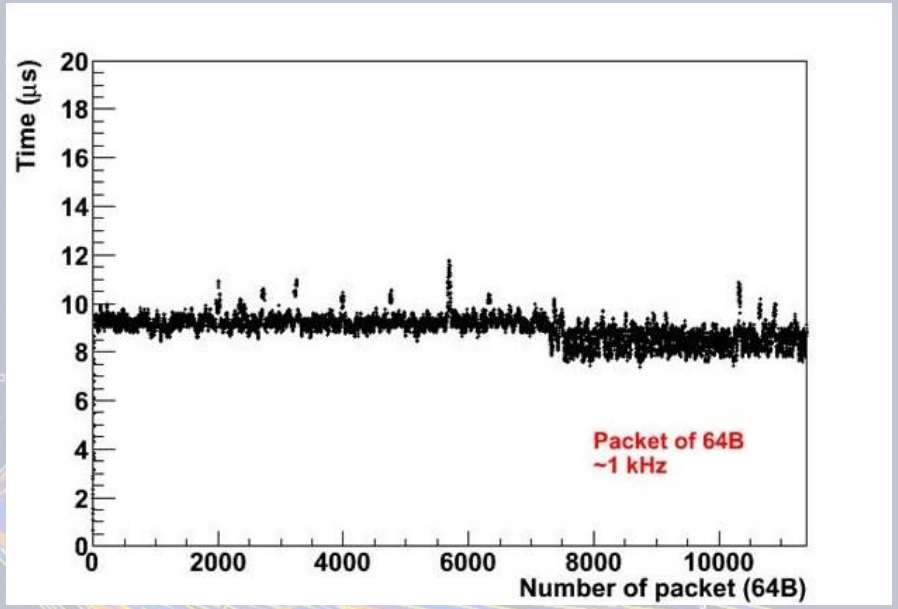


Response time controllable at the tens of µs level

F. Schifano, M. Pivanti ***(Ferrara Univ.)***

NIC to host

On a PC running standard Linux the traversal of the network stack to user space is a major source of latency AND latency fluctuations



Use PF_RING sockets (**L. Deri**, NTOP) for high-speed packet capture with Direct NIC Access to user-space: 1.1 M packets/s (on 1 GbE adapter) reachable

Other issues

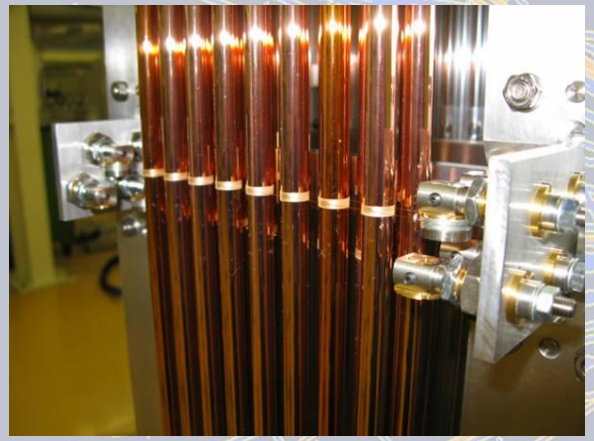
In order to use CPU/GPU in hard real-time, the host must have knowledge of "experiment time" with a precision equal to a small fraction of the maximum latency

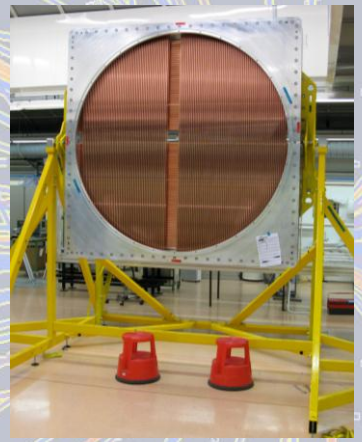
This synchronization issue could be handled by PTP (IEEE 1588) messages (1 μs accuracy) handled by NIC and detector hardware. Implementation test in progress.

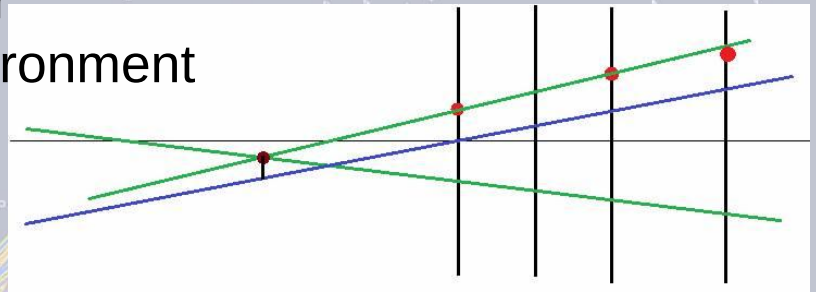
Use of real-time Linux kernels was considered:
it would improve host response times,
but so far it looks unnecessary

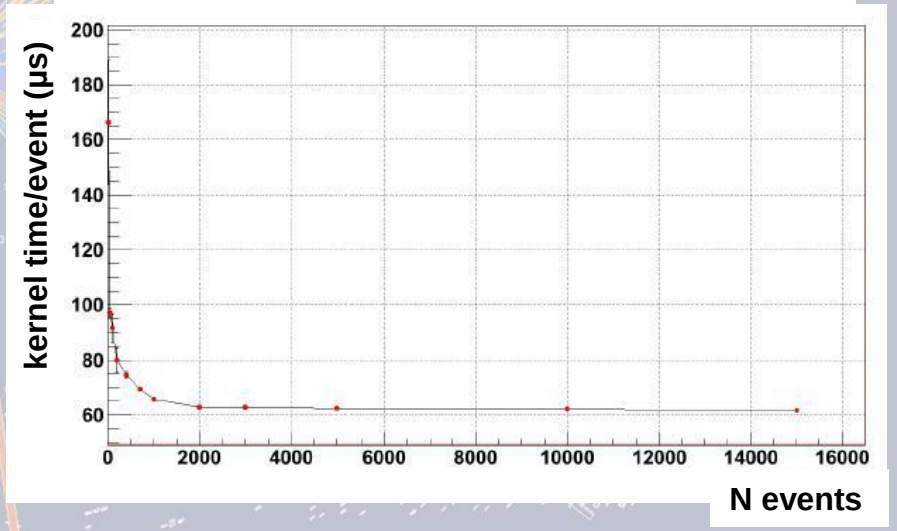
Second case study: STRAWS

Converging track rejection in a 4-chamber magnetic spectrometer in a high-rate environment









Different time-scale (L1 trigger) with 1 MHz input rate and no strict latency requirements

Summary and perspectives

An ongoing exploration of GPU use in a **hard real-time** environment for a high-performance triggering system in a CERN experiment is being carried on.

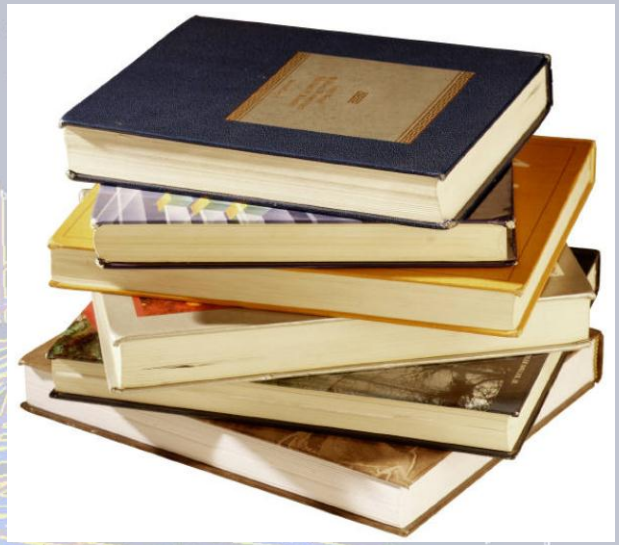
Very short **latencies** and the need to limit their **fluctuations** pose a significant challenge on several aspects of such a system

Results are encouraging with no fundamental show-stoppers

A complete **demonstrator system** is going to be installed and tested in the first low-rate test run of NA62 in **2012**

Data-taking starting in **2014** might profit significantly from the inclusion of a complementary full-scale GPU-based trigger system.

Further information



- IEEE-NSS Conf. Record 10/2009,195
- Nucl. Instrum. Meth. Phys. Res. A 628 (2011) 457
- Nucl. Instrum. Meth. Phys. Res. A 639 (2011) 267
- “Fast online triggering in high-energy physics experiments using GPUs” Nucl. Instr. Meth. Phys. Res. A 662 (2012) 49