Real-time triggering in HEP using GPUs

R. Fantechi[1], V. Innocente[1], G. Lamanna[2],
F. Pantaleo[1], M. Sozzi[2]
(CERN – INFN Pisa)

March 20th , 2013

Introduction

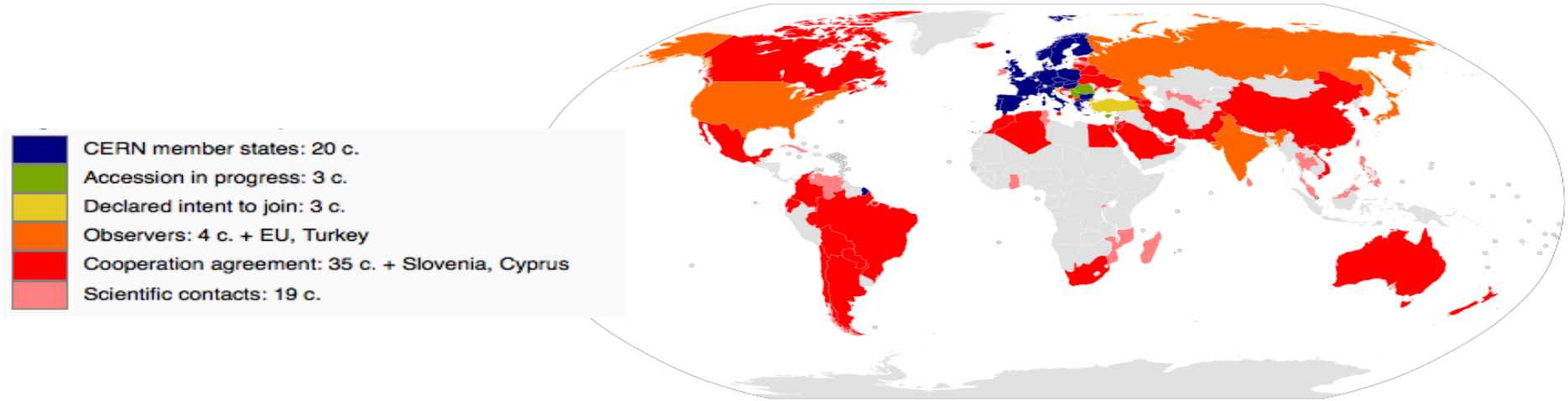
CERN

European Organization for Nuclear Research

Founded in 1954 by 12 countries.

2013: 20 member states

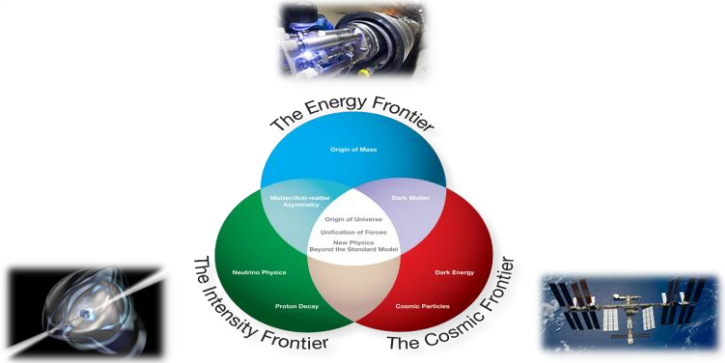
More than 10,000 users all around the world



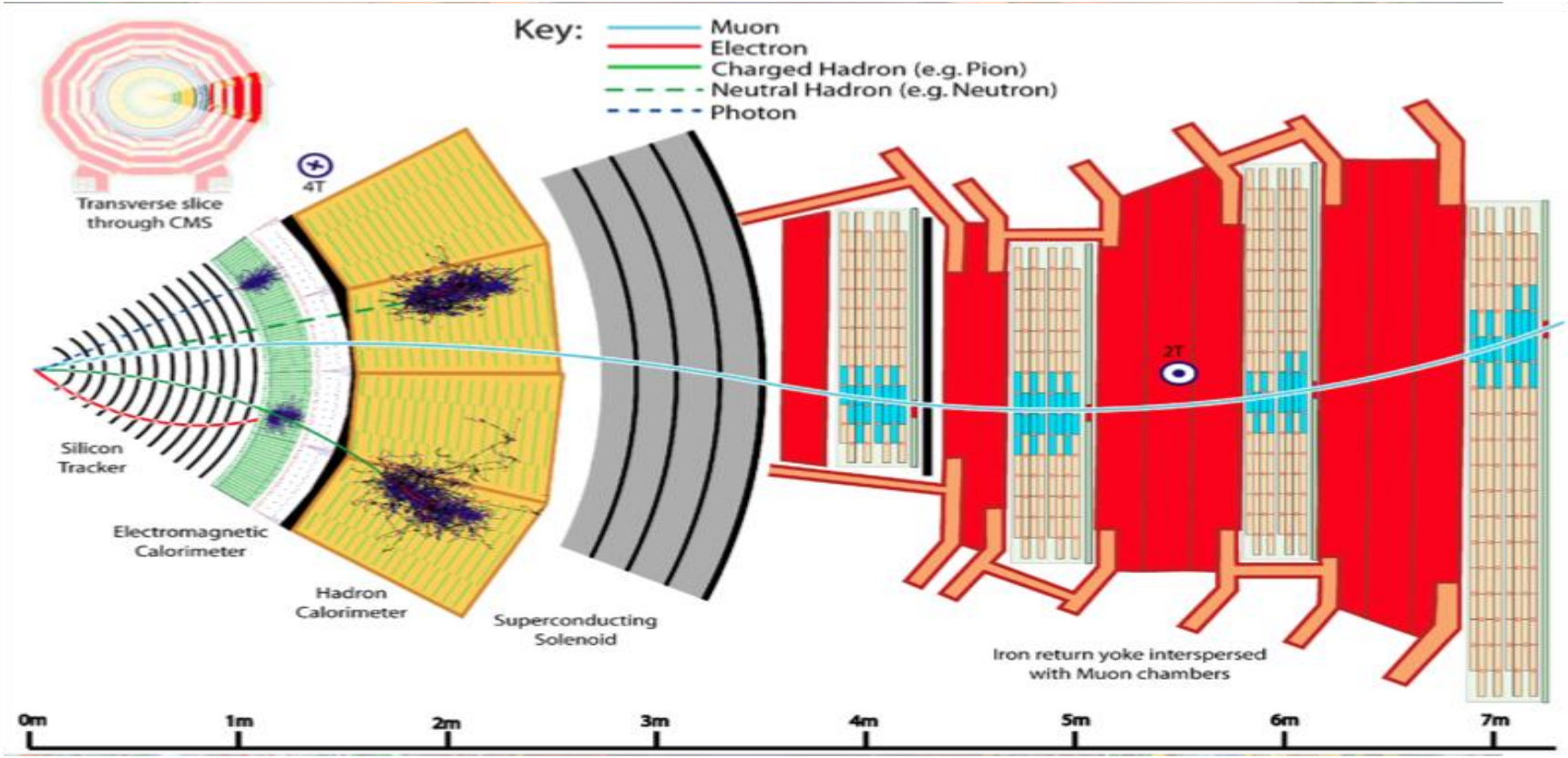
Physics Frontiers

Only by exploring these 3 frontiers we can find the answers to the most fundamental questions of the mankind:

- What is the universe made of?
- What are the rules that govern its evolution?
- What are its origins?
- What is its destiny?

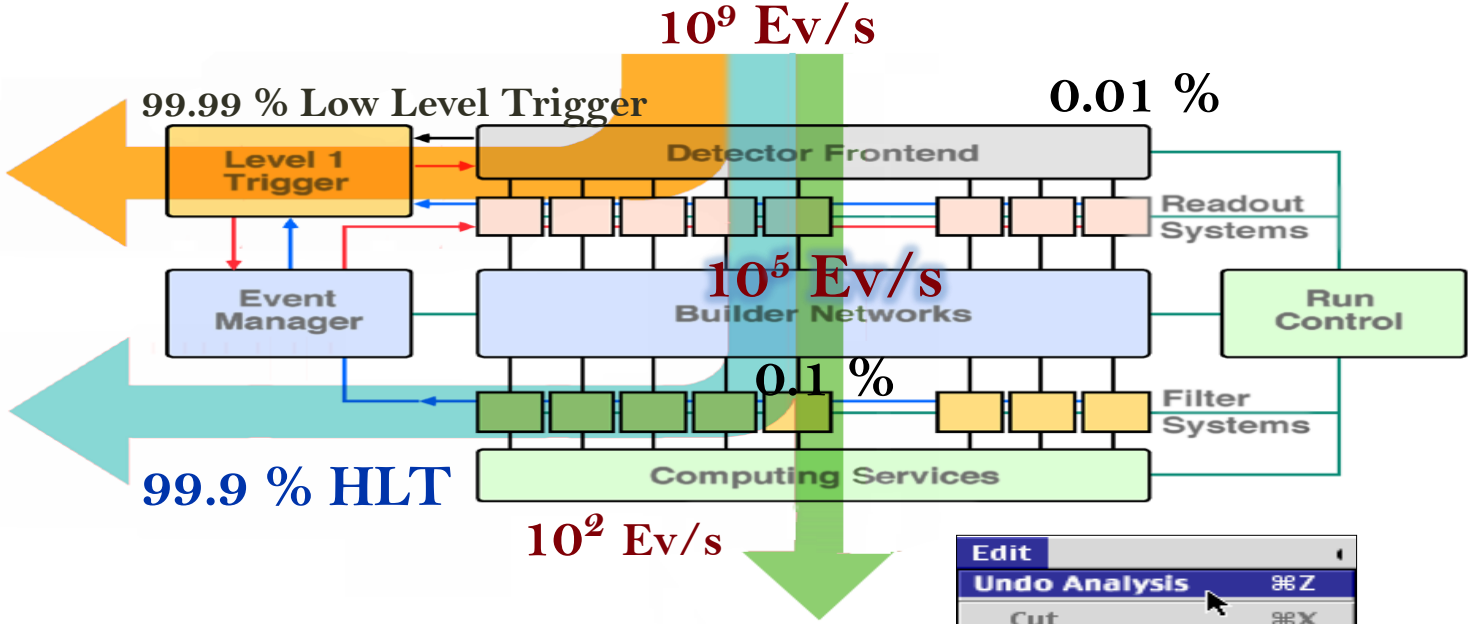


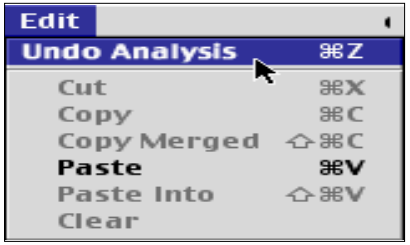
Detector "onion" structure



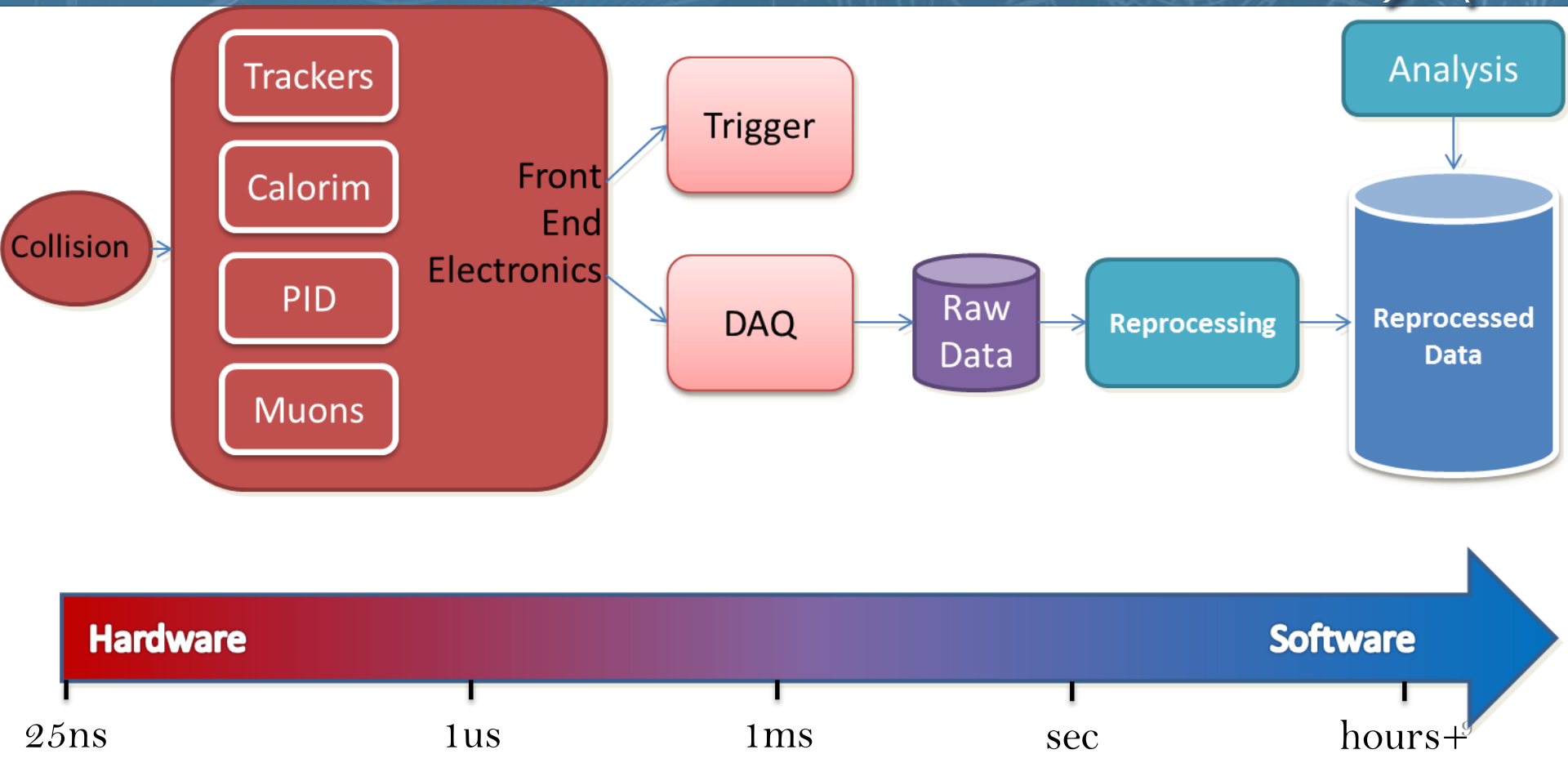
TRIGGER

Event Selection Flow

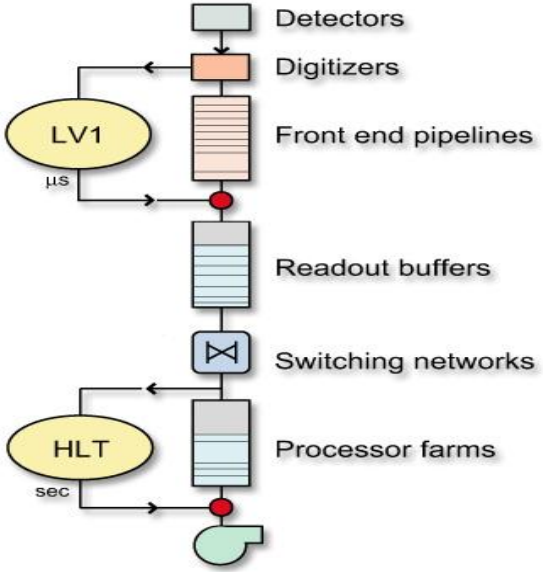




Detector structure



Low Level Trigger



- **Time needed for decision $\Delta t_{dec} < 1$ ms**
- Particle rate O(10MHz)
- Need pipelines to hold data
- Need fast response

- Backgrounds are huge
- High rejection factor

- Algorithms run on local, coarse data
- Ultimately, determines the physics

HEP needs

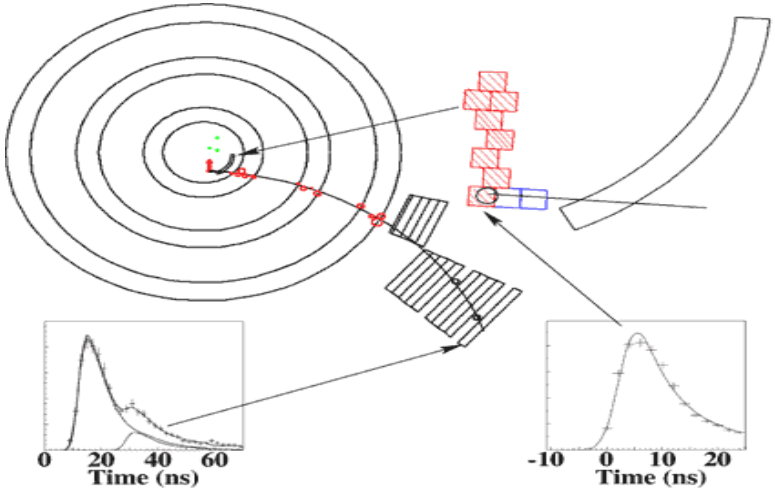
- High Energy Physics detectors needs:
  - Processing large amounts of information
  - **Very short time response**
  - Complex structures, many sensors with topographic information
  - Efficient processing of data
- Huge benefits expected from the use of multi- and many-cores techniques
- Where are we ?

NA62

NA62 – The Goal

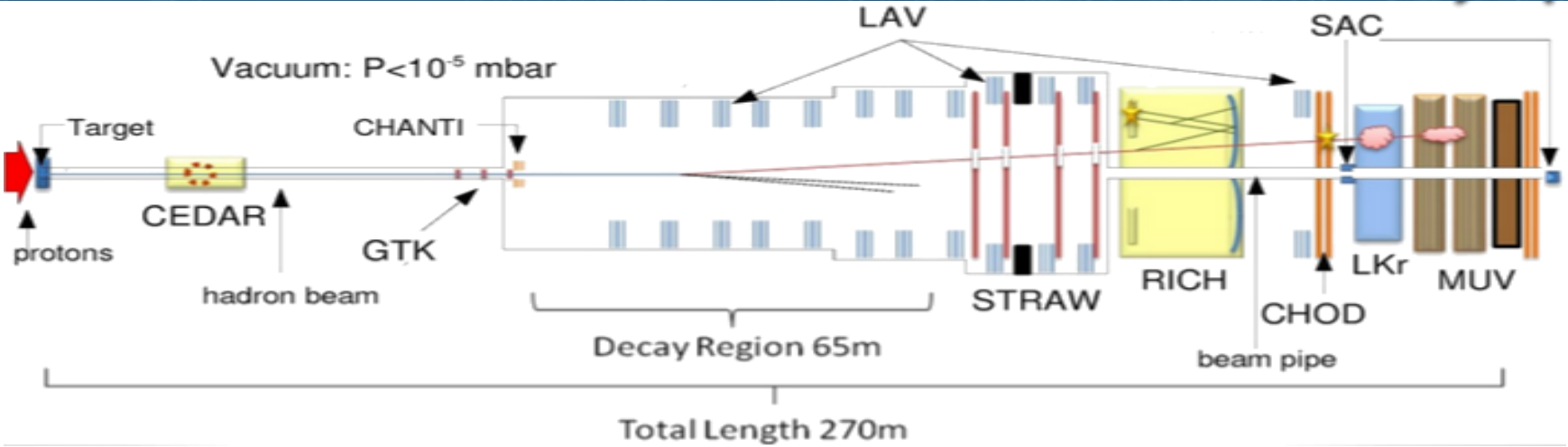
Precision measurement of the ultra-rare decay process $K^+ \rightarrow \pi^+ \nu\nu$

First observed at Brookhaven National Labs 1997-2001 (see Figure)
Extremely sensitive to any unknown new particles, even way beyond the reach of direct experimental searches at new and forthcoming accelerators



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**

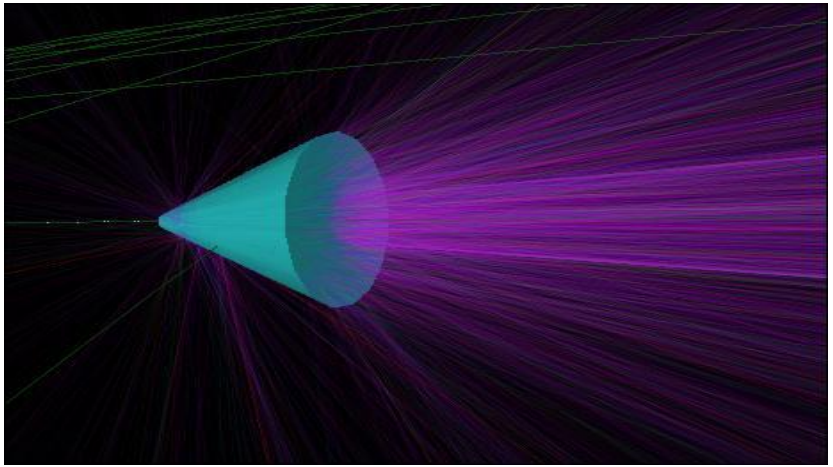
Experimental Technique



- Kaons decay in-flight from an unseparated 75 GeV/c hadron beam, produced with 400 GeV/c protons from SPS on a fixed berilium target
- ~800 MHz hadron beam with ~6% kaons
- The pion decay products in the beam remain in the beam pipe
- **Goal:** measurement of **O(100)** events in two years of data taking with % level of systematics
- Present result (E787+E949): 7 events, total error of ~65%.

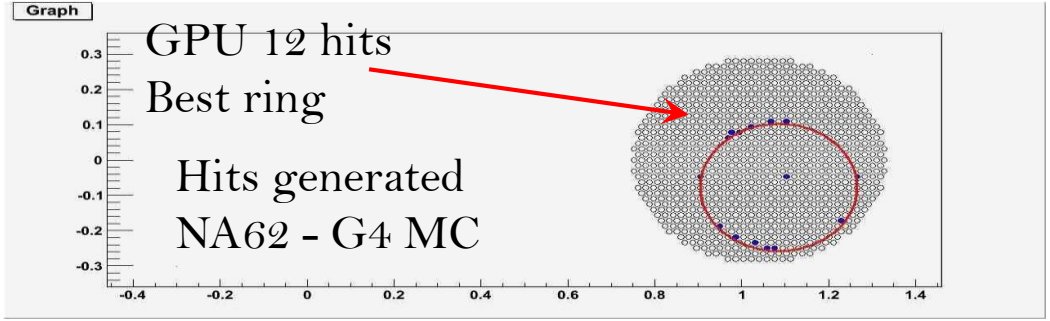
NA62 RICH Real-time Trigger

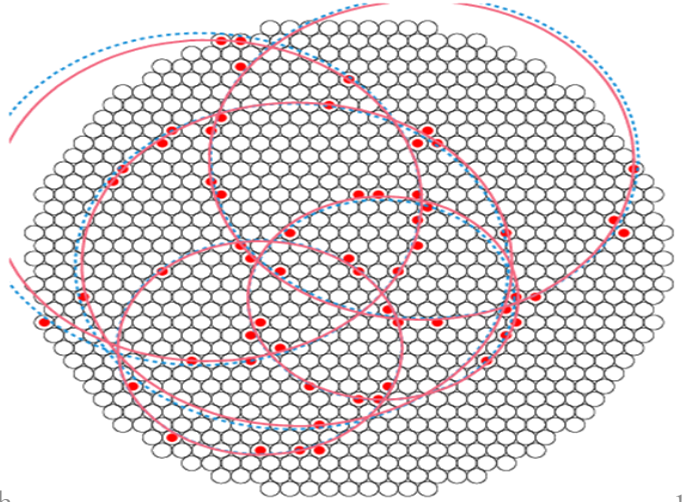
Cherenkov Radiation



Ring Reconstruction

- Natively built for pattern recognition problems
- First attempt: ring reconstruction in RICH detector.





It's a **pilot project**, very promising R&D!

Data Flow

Read data directly from Network Interface buffers using PF_RING DNA

Filling data structures of arrays, waiting for a good quantity of events to sustain the throughput (d_1)

Multiple threads transfer this data to GPU Memory on different streams

Multiple threads launch kernels on different streams

Concurrently transfer the results to the NIC ring buffers and to the frontend electronics (d_2)

Max time O(100us)

Data received from the NaNet onboard ethernet interface

Data is streamed onto the PCI Express bus towards a GPU ring buffer

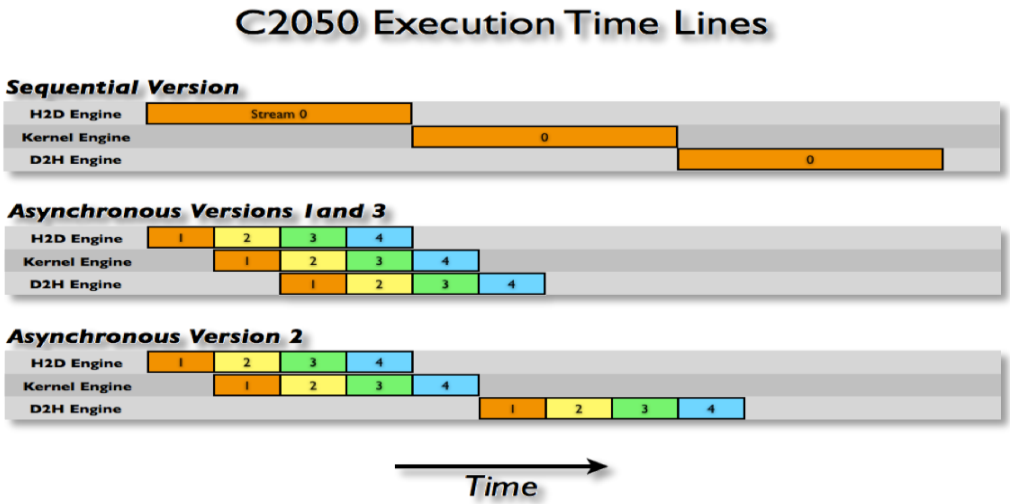
Once a buffer slot is completely filled the CPU calls the kernel to be executed by the GPU

The result is then copied on the host memory

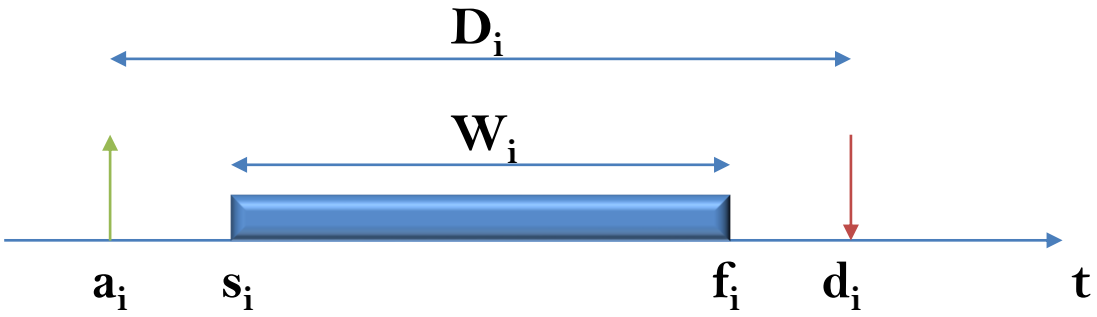
See Alessandro Lonardo's talk – S3286

Stream Scheduler

- Exploit the instruction-level parallelism (i.e. pipelining streams)
- This is usually done by interlacing one stream instructions with another stream ones
- This cannot be done in real-time without the introduction of other **unknown** latencies
- CPU thread-level parallelism

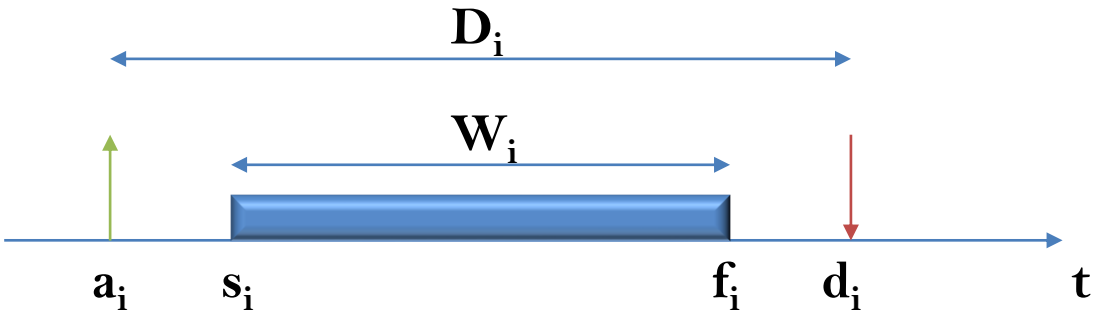


Task Parameters



a_i	arrival time
s_i	start time
W_i	worst-case execution time
d_i	absolute deadline
f_i	finishing time

Lateness



- The scheduler must be aware of the lateness defined as $L_i = f_i - di$
- It has to be partially preemptive: if a packet is late it must be given higher priority, if too late it has to be ignored.
  - cudamemcpy and kernel execution cannot be stopped
- The more precise the time synchronization with the detector, the more precise the lateness (PTP).

NA62 RICH Tests

Hardware configuration (1/2)

First Machine

- GPU: NVIDIA Tesla C2050
  - 448 CUDA cores @ 1.15GHz
  - 3GB GDDR5 ECC @ 1.5GHz
  - CUDA CC 2.0 (Fermi Architecture)
  - PCIe 2.0 (effective bandwidth up to ~5GB/s)
- CPU: Intel® Xeon® Processor E5630 (released in Q1'10)
  - 2 CPUs, 8 physical cores (16 HW-threads)

Hardware configuration (2/2)

Second Machine

- GPU: NVIDIA GTX680
  - 1536 CUDA cores @ 1.01GHz
  - 2GB GDDR5 ECC @ 1.5GHz
  - CUDA CC 3.0 (Kepler Architecture)
  - PCIe 3.0 (effective bandwidth up to ~11GB/s)
  - CUDA Runtime v4.2, driver v295.20 (Feb '12)
- CPU: Intel® Ivy Bridge Processor i7-3770 (released in Q2 '12)
  - 1 CPUs, 4 physical cores (8 hw-threads) @3.4GHz

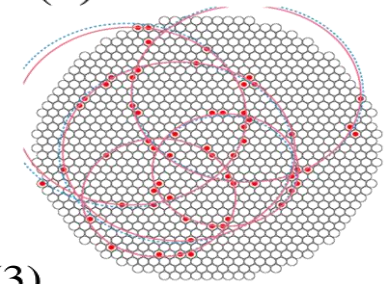
Crawford Algorithm

Consider a circle of radius R, centered in (x_0, y_0) and a list of points (x_i, y_i).
The following relations exist:

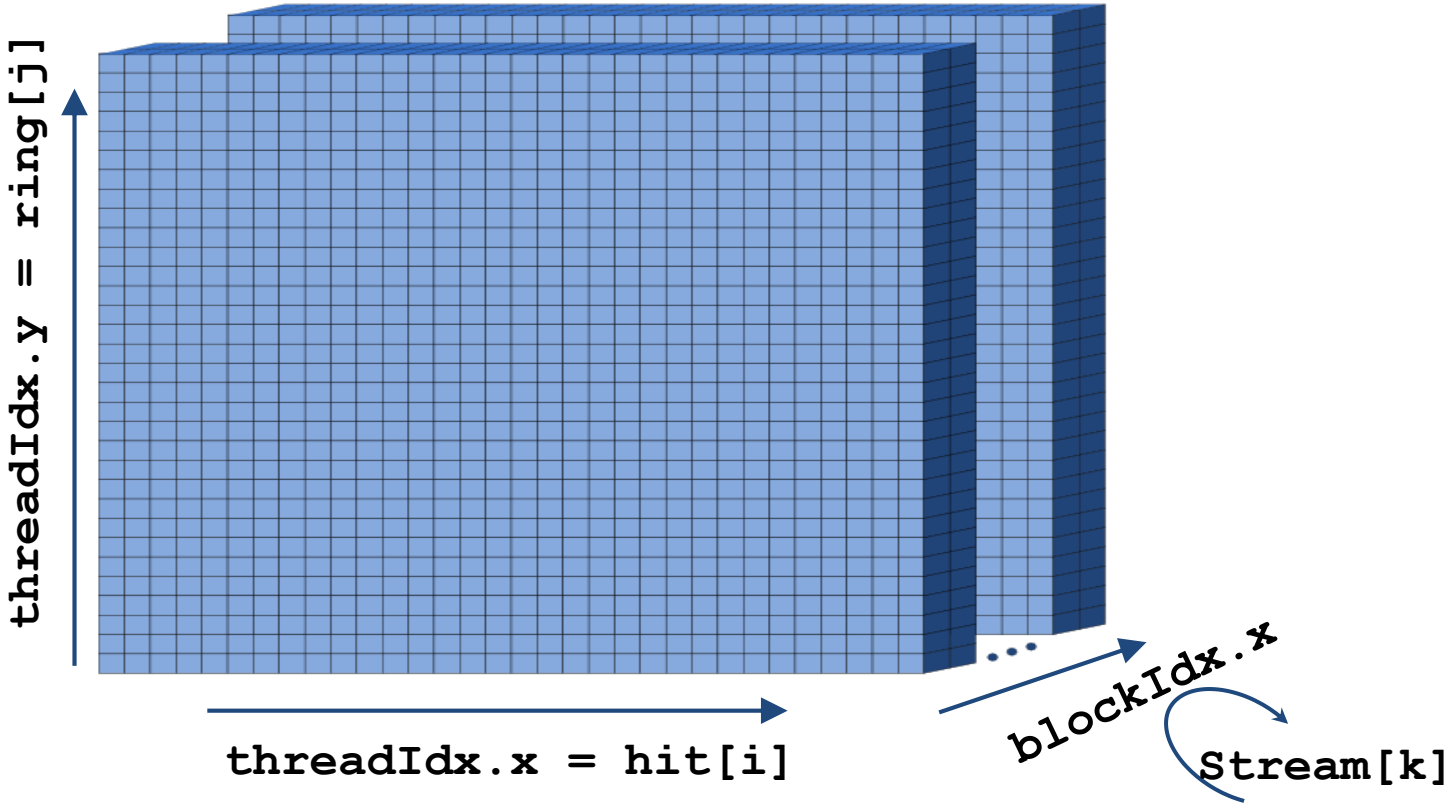
$$x_0^2 + y_0^2 - R^2 = \frac{1}{N}\{2x_0 \sum x_i + 2y_0 \sum y_i - \sum x_i^2 - \sum y_i^2\}. \qquad (1)$$

$$x_0\{\sum x_i^2 - \frac{(\sum x_i)^2}{N}\} + y_0\{\sum x_i y_i - \frac{\sum x_i \sum y_i}{N}\} = \frac{1}{2}\{\sum x_i^3 + \sum x_i y_i^2 - \sum x_i \frac{\sum x_i^2 + \sum y_i^2}{N}\}, \qquad (2)$$

$$x_0\{\sum x_i y_i^2 - \frac{\sum x_i \sum y_i}{N}\} + y_0\{\sum y_i^2 - \frac{\sum y_i^2}{N}\} = \frac{1}{2}\{\sum x_i^2 y_i + \sum y_i^3 - \sum y_i \frac{\sum x_i^2 + \sum y_i^2}{N}\}. \qquad (3)$$



GPU grid organization

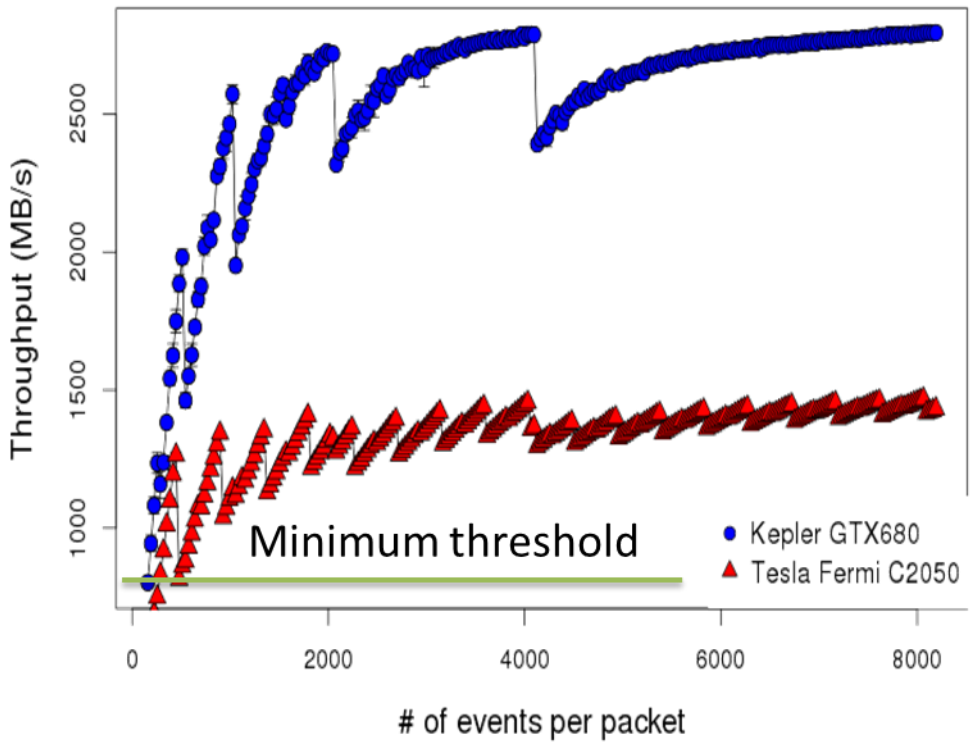


Results - Throughput

Saturation plateau (**1.4GB/s** and **2.7GB/s**)

The right choice of packet dimension is not unique.
It depends on the maximum latency we don't want to exceed and on the input rate of events.

Considering that the maximum rate per sub-detector (@10MHz particles rate) for NA62 experiment is ~**800MB/s**, I would consider the throughput test **PASSED**

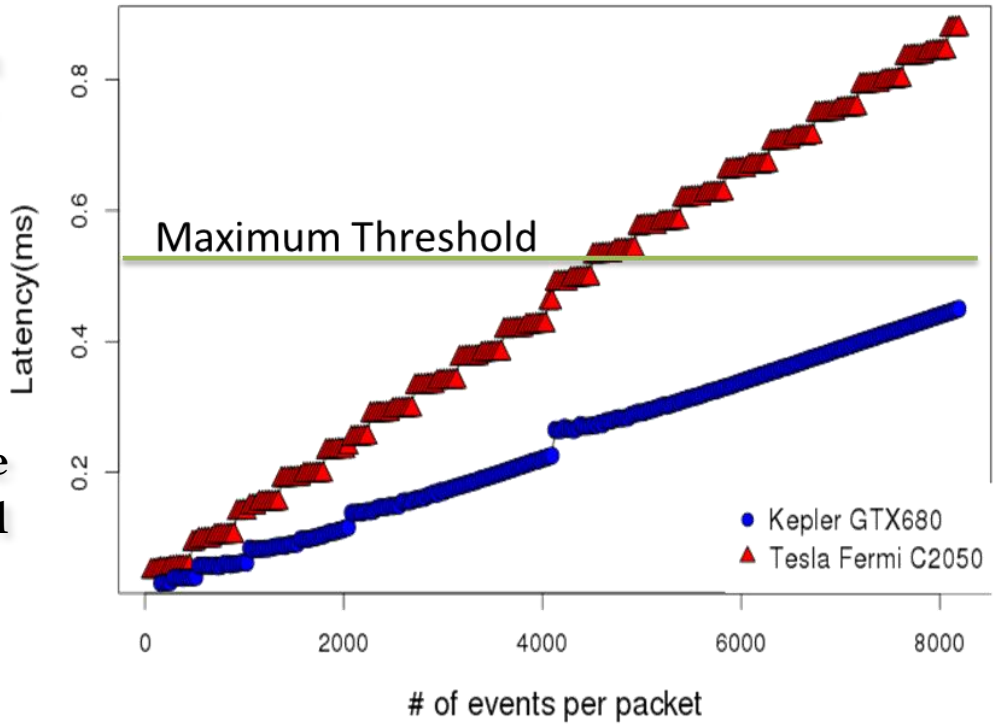


Results - Latency

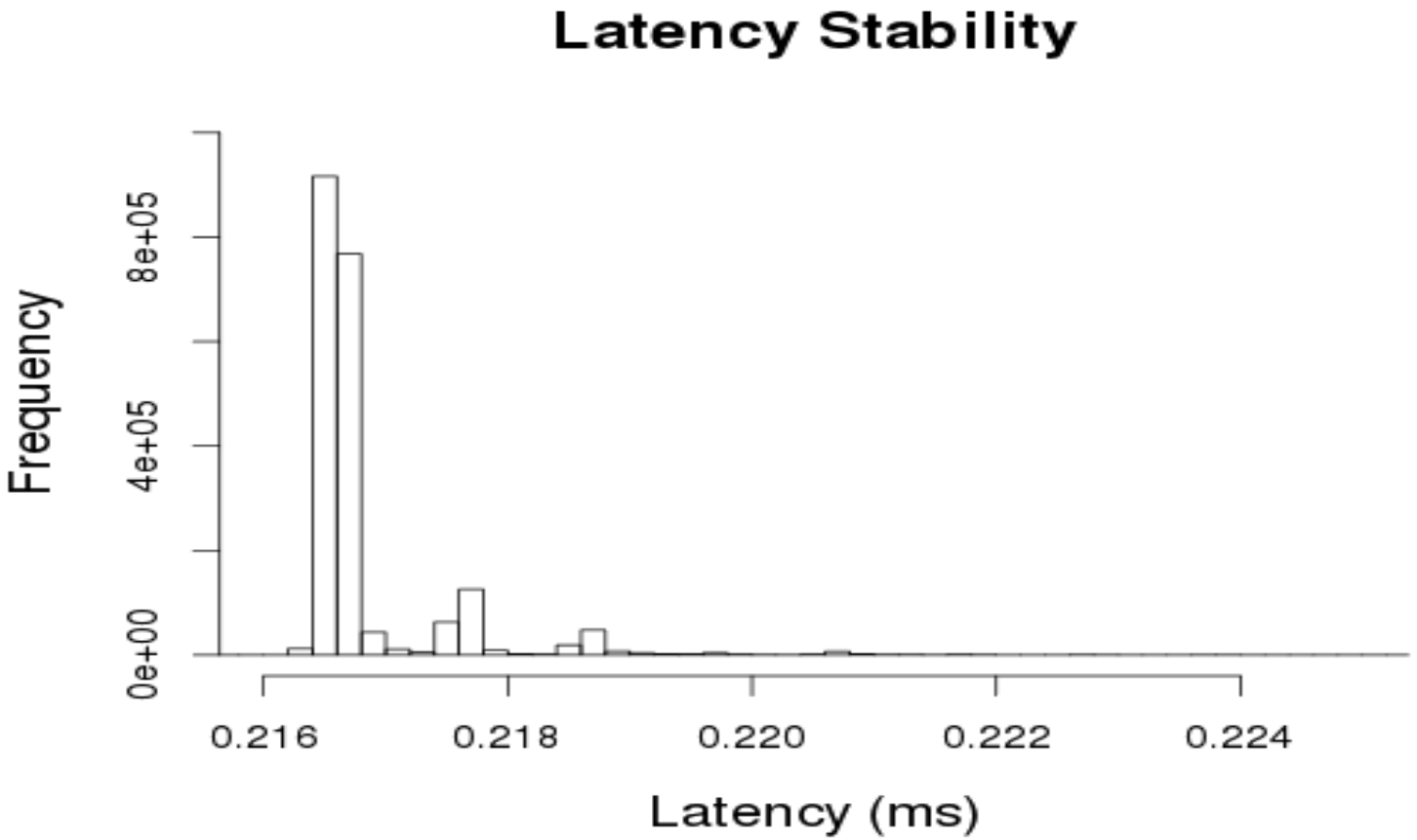
Latency pretty stable wrt event size.

- A lower number of event inside a package is better to achieve a low latency.
- A larger number of event guarantees a better performance and a lower overhead.

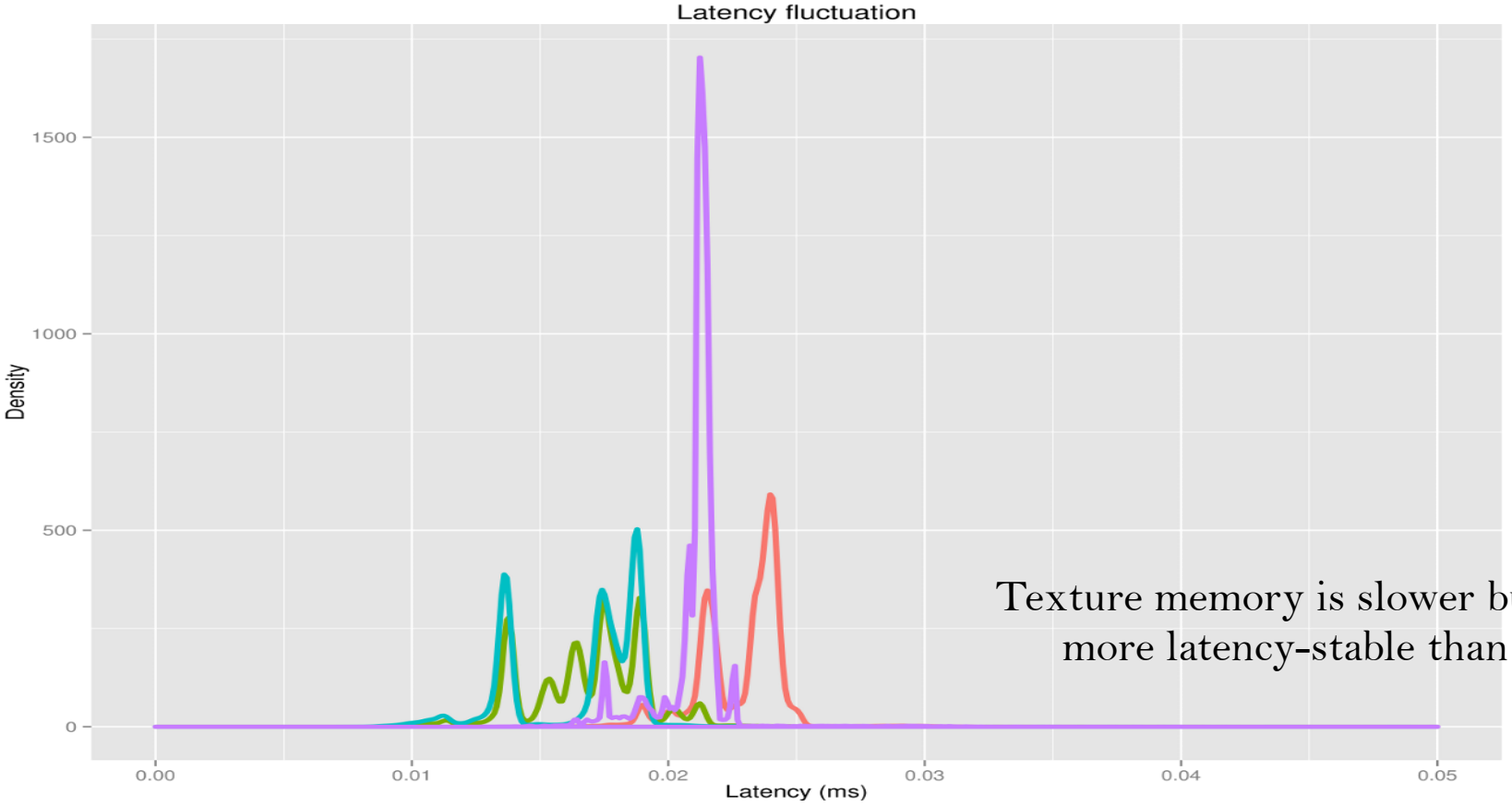
The choice of the packet size depends on the technical requirements.



Results - Latency Stability



Memory Latency Stability



Texture memory is slower but way more latency-stable than the others

Conclusion

- GPUs seem to represent a good opportunity, not only for analysis and simulation applications, but also for more “hardware” jobs.
- Replacing custom electronics with fully programmable processors to provide the maximum possible flexibility is a reality not so far in the future.

Questions?