Università degli Studi di Bari

Dipartimento Interateneo di Fisica "M. Merlin"

PhD School in Physics XXXI Cycle

Activity report on the of the first year of doctoral school.

Phd Student: Adriano Di Florio

Supervisor: Dr. Alexis Pompili

17th November 2016

In the present report, with respect to the first year of Doctoral School in Physics, the research work is presented in section 1 and the didactic work is presented in section 2.

1 Research work

During the first year the research work has mainly focused on two topics:

1. the study of GPU computing applications within studies of charmonium-like exotic states
2. the development and study of new tracking algorithm within the CMS collaboration on heterogeneous computing systems

1.1 GPU computing applications within studies of charmonium-like exotic states

1.1.1 Introduction to `GooFit`

The word heterogeneous computing refers to an enhancement of application performances that can be obtained by offloading compute-intensive portions to the GPU, while the remaining code still runs on the CPUs. In the context of High Energy Physics (HEP) analysis application, `GooFit` is an under development open source data analysis tool,used in applications for parameters'

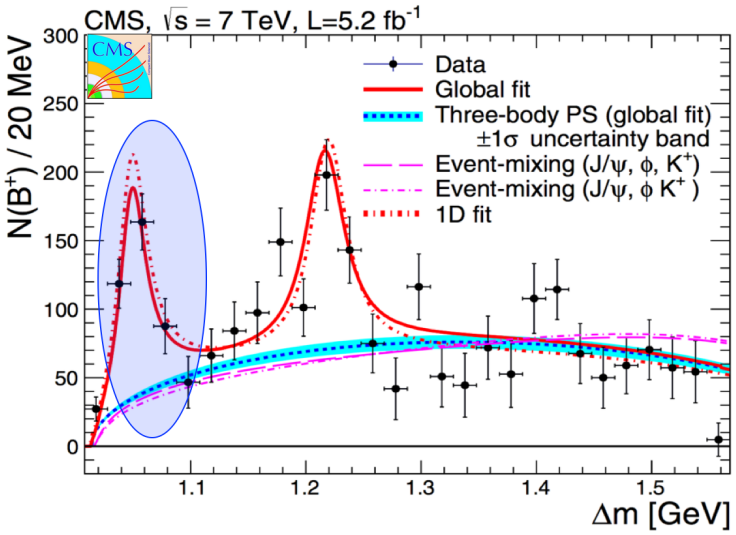

Figura 1: Fits to the background-subtracted $J/\psi\phi$ invariant mass in the $B^+ \to J/\psi\phi K^+$ decay [CMS, PLB 734 (2014) 261]; as first case of study the significance of the left peak has been estimated via MC toys with `GooFit`.

estimation, that interfaces ROOT/RooFit to the CUDA parallel computing platform on nVidia's GPUs (it also supports OpenMP). `GooFit` acts as an interface between the MINUIT minimisation algorithm and a parallel processor which allows a Probability Density Function (PDF) to be evaluated in parallel. Fit parameters are estimated at each negative-log-likelihood (NLL) minimisation step on the *host side* (CPU) while the PDF/NLL is evaluated on the *device side* (GPU).

1.1.2 Statistical significance estimation by MC toys with GooFit

As first use case to test the computing capabilities of GPUs with respect to traditional CPU cores, a high-statistics toy Monte Carlo technique has been implemented both in ROOT/RooFit and `GooFit` frameworks with the purpose to estimate the statistical significance of the structure observed by CMS close to the kinematical boundary of the $J\psi\phi$ invariant mass in the three-body decay $B^+ \to J\psi\phi K^+$ [CMS, PLB 734 (2014) 261]. The optimised `GooFit` application running on GPUs has provided striking speed-up performances with respect to the RooFit application parallelised on multiple CPUs by means of PROOF-Lite tool.

The ongoing next step is the extension of the `GooFit` MC toys significance estimation method to situations with a new unexpected signal and a global significance must be estimated. In this case the Look Elsewhere

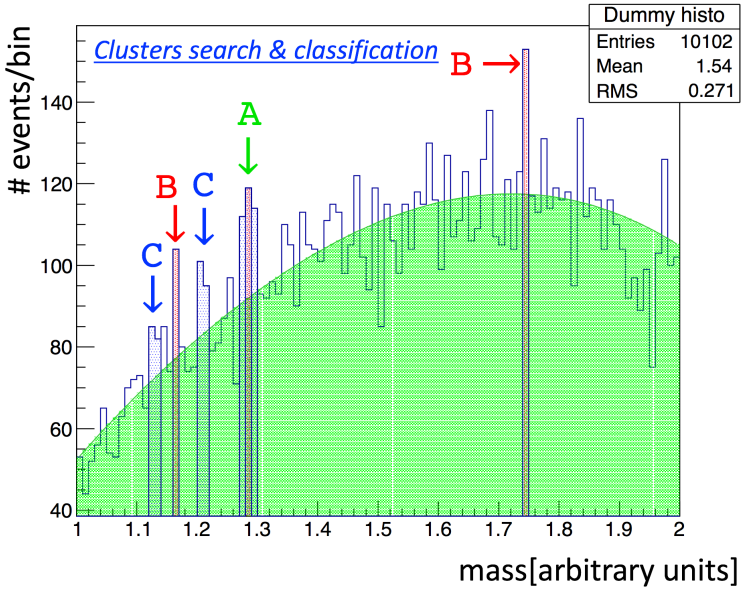

Figura 2: A dummy distribution generated form an hypothetical polynomial background to which has been applied the clustering scanning technique described in the text.

Effect must be taken into account and a scanning technique needs to be implemented in order to consider relevant peaking behaviour with respect to the background model everywhere in the mass spectrum within the same fluctuation. Thus a scanning technique has been developed on the basis of a clustering approach and has been designed with the aim to satisfy two concurrent requirements:

A not missing any possible interesting fluctuation;

B not selecting too many irrelevant fluctuations since this would heavily increase the computing time needed to execute all the related fits.

The spectrum scanning steps are configured as follows:

1. For each MC Toy iteration a distribution based on the background p.d.f. model is generated in the range of the whole mass spectrum via `Hit or Miss` procedure. The number of events of the generated distribution is fixed to the number of events found in the data.

2. The *Null Hypothesis* fit is performed with the background function only (the same used to generate the data) in order to set up the clustering procedure.

3. Search for a *seed*, which is is defined as a bin whose content fluctuates more than $x\sigma$ strictly above the value of the background function in the center of that bin (where σ is the statistical error of the considered bin).

4. Check if the seed's side bins show a content that fluctuates more than $y\sigma$ strictly above the value of the background function in the center of that bin. In case of positive result the side bin(s) is(are) attached to the seed (*A-type* cluster). In case of negative result the seed bin is taken alone (*B-type* cluster).

5. Check also for "light" seeds: bins that fluctuates more than $z\sigma$ with $z < x$ and with at least a side bin fluctuating more than $y\sigma$. In case of positive result a cluster is formed (*C-type* cluster).

For each fluctuation found with this clustering technique, a set of fits is performed changing the parameters' range and starting values. The three parameters x (single seed threshold), y (side bin threshold) and z (additional sided seed threshold) has been tuned in order to meet both the requirements. Then recently the procedure has been configured to run on the new *Recas HPC Cluster* and the application is being run with different clustering configurations to be able to estimate the systematic uncertainty on the p-value estimation related to the method itself. All these studies on the estimation of the local and global statistical significance by MC toys with GooFit will be reported in a CMS Analysis Note [AN-2015-334].

1.1.3 Amplitude analysis fit of $B^0 \to J/\psi K^+\pi^-$

Traditional *Dalitz Plot* analyses deal with 3-body decays without any vector state as a daughter. For example many analyses of B o D meson decays at the *B-factories* dealt with π and K, either charged and/or neutral. In this case the decay amplitudes are calculated in a 2-dimensional parameter space, namely the *Dalitz Plot* space itself. In 3-body decays with vectors in the final state the decay amplitude has to be calculated on *n-dimensional* parameter space within the helicity formalism. Among the search for *tetraquark* exotic states in CMS, there are some decay channels needing an *amplitude analysis*:

- $B^0 \to J/\psi K^+\pi^-$, to search for $Z(4430)$, $Z_c(4240)$ and $Z_c(3900)$ states;
- $B^0 \to J/\psi\phi K^+$, to search for $Y(4140)$ and other structures in $J/\psi\phi$;

The efforts during the first year have been focussed on the first decay channel. Assuming the only intermediate 2-body states are the K^*s, in this case the parameter space is four dimensional described as:

$$\Phi = (m^2_{K\pi}, m^2_{\psi\pi}, \theta_\psi, \phi_{\psi K^*})$$

where (see Figure 3)

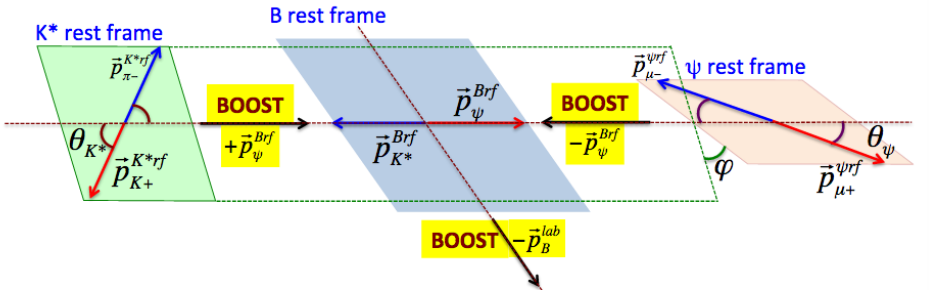

Figura 3: $B^0 \to J/\psi K^*$ decay spatial representation with the different rest frame, momenta and angles between rest frame planes.

- θ_ψ is the ψ helicity angle, i.e. angle between the and μ mometum in ψ rest frame;
- $\phi_{\psi K^*}$ is the angle between ψ and K^* decay planes;

The *amplitude analysis* fit strategy requires to check if fitting the data with a p.d.f. based only on the K*s is able to reproduce the real distributions without the need of extra Z states; if the fit will be not satisfactory a contribution for the $Z(4430)$, $Z_c(4240)$ and $Z_c(3900)$ have to be taken into account. First of all the fitting procedure have to be validated:

1. generating events are generated (via `RooFit`) according to a chosen p.d.f. model, setting the parameters to the Belle's paper amplitude analysis parameters values [Phys.Rev. D90 (2014) 112009];
2. carrying out the *amplitude analysis* fit and verifying that the fit estimates are compatible with the values given in input.

Only the most relevant intermediate K^*s resonances [$B^0 \to J/\psi K^*$] are considered and they contribute to the p.d.f. with 28 fit parameters (one *absolute value* and one *phase* for each helicity amplitude; one amplitude for each spin-0 K^*, three for each K^* with spin greater than zero). Such a complex fit requires high computational capabilities and it requires very long fitting times if carried out with `RooFit` was requiring very long fitting times. Thus the whole fitting code is being ported on `GooFit` to run it on GPUs. A first result has already been achieved: the `GooFit` fit takes only 10 minutes performing over 1000 `MIGRAD` calls with alle the 28 parameters. On the other hand `RooFit` needs one hour to fit only a four parameters p.d.f. The porting and the testing of the fitting procedure to `GooFit` is ongoing.

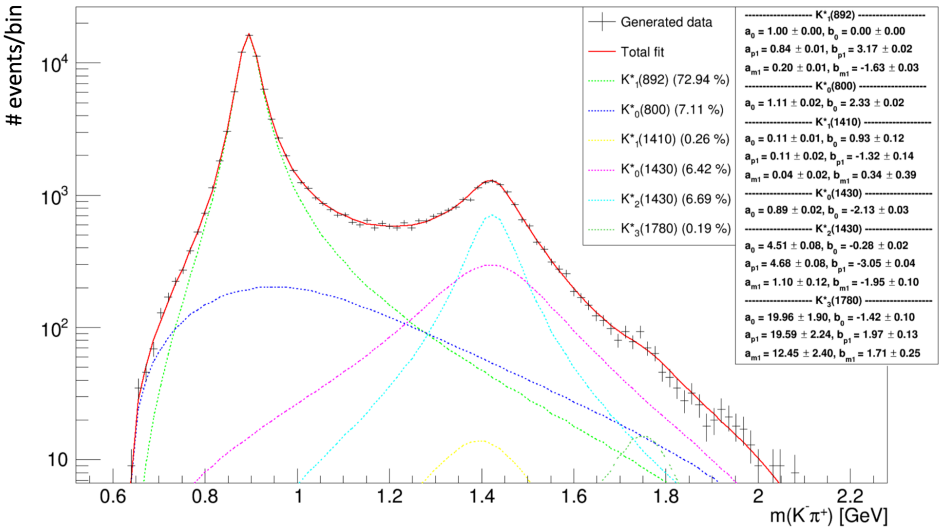

Figura 4: Projection on $m^2_{K\pi}$ spectrum of the four dimensional generated data and fitting function.

1.1.4 Inclusive search of the production of the Y(4140) state

Hadronic spectroscopy has experienced a renaissance in the last decade thanks to the experimental findings at B-factories, Tevatron and recently at the LHC. In the last ten years a new wide zoology of charmonium-like states (the so called X,Y,Z), decaying to conventional charmonium, has been observed in spite of these states being above the open-charm threshold(s). They do not fit into the charmonium spectrum predicted by potential models in the contest of conventional quark model of hadrons. Experimental and theoretical pictures are far from being clear and there are opportunities at LHC to confirm these states and study their properties.

One of the neutral charmonium-like XYZ meson states is the Y(4140) state. The CDF collaboration at the Tevatron (FermiLab) reported the first evidence for the $Y(4140)$ state in 2009 and confirmed it later in 20011 with higher statistics. the Y(4140) state was appearing as an intermediate state in the decay $B^+ \to Y(4140)K^+ \to J/\psi\phi K^+$, close the $J/\Psi\phi$ threshold. The CMS Collaboration observed (with a statistical significance $> 5\sigma$) the $Y(4140)$ peaking structure in the $J/\psi\phi$ invariant mass spectrum. Recently (2015) D0 collaboration found evidence of prompt and non prompt direct production of this state, in p-pbar collisions ($p\bar{p} \to Y(4140) + X$ with $Y \to J/\psi\phi$), by studying inclusively the J/ mass spectrum. Moreover, during last year, LHCb collaboration circulated a paper draft, with the first amplitude analysis of the $B^+ \to J/\psi\phi K^+$, observing even four $J/\psi\phi$structures each with a significance $> 5\sigma$. Among them the $Y(4140)$ is the closest to the

kinematical threshold and is best described as a $D^s\bar{D}_s^*$cusp.

Uno dei mesoni neutri *charmonium-like* è lo stato $Y(4140)$. La collaborazione CDF al Tevatron (FermiLab) ha riportato la prima evidenza della $Y(4140)$ nel 2009, riconfermandola nel 2011 come stato intermedio nel decadimeno $B^+ \to Y(4140)K^+ \to J/\Psi\phi K^+$. La collaborazione CMS ha osservato una struttura compatibile (con una significatività statistica maggiore di 5σ) nello spettro di massa invariante $J/\Psi\phi$ misurando $m_Y \approx 4148MeV$ e $\Gamma_Y = 28MeV$. Anche la collaborazione D0 al Tevatron ha ricercato lo stato Y(4140), rilevandolo e misurandone massa e larghezza simili nello stesso canale di decadimento del mesone B^+. Inoltre la collborazione ha trovato evidenza della produzione diretta di questo stato nelle collisioni $p\bar{p}$ ($p\bar{p} \to Y + altro$ con $Y \to J/\Psi\phi$), studiando lo spettro inclusivo di massa $J/\Psi\phi$. Nel corso di questo anno la collaborazione LHCb collaboration circulated a paper draft [9], with the first amplitude analysis of the B+ J/K+, observing four J/ structures each with a significance $> 5\sigma$. The lightest of them is the is the Y(4140) and is best described as a DsDs* cusp even if a resonant interpretation is also possible with mass consistent with, but width much larger than, previous measurements of the claimed Y(4140) state. Until now 8 TeV *Run I* data have been explored and the future plan is to use 13 TeV *Run II* data.

1.2 New tracking algorithm on heterogeneous computing systems

During *Run II* and *Run III*, the increased luminosity with the consequent increased pile-up will pose significant new challenges for the CMS detector, in particular for the reconstruction of tracks that will be heavily affected by the increased track density. The quest of significantly reducing the 40 MHz data rate delivered by proton-proton collisions to the detectors, together with the retention of physics signal potentially interesting for searches of new physics phenomena led to the evaluation of modern multi-cores and many-cores architectures for the enhancement of the existing High-Level Trigger (HLT). The primary goal of the HLT is to apply a specific set of physics selection algorithms on the events read out and accept the events with the most interesting physics content. By its very nature of being a computing system, the HLT relies on technologies that have evolved extremely rapidly but that cannot rely anymore on an exponential growth of frequency guaranteed by the manufacturers. Indeed the online farm consists of about 20000 CPU Xeon cores and a single event is assigned to a single logical core. Also now, with a pile up average of 50, not all the tracks are reconstructed for all the events at the HLT and this will be even more difficult at higher pile-up at higher luminosities. The High Luminosity LHC (HL-LHC) is a project to increase the luminosity of the Large Hadron Collider to $5 \cdot 10^3 4 cm^{-2} s^{-1}$ and

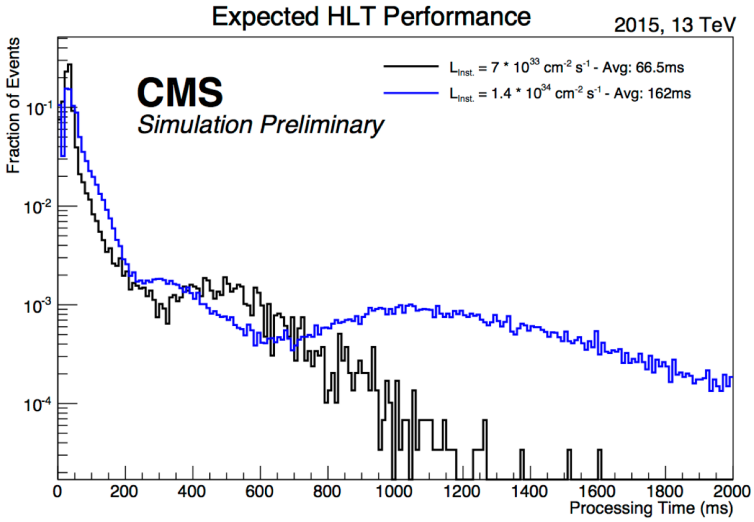

Figura 5: CMS HLT simulated track reconstruction processing time at 13 TeV with a pileup between $40 - 50$ at different instantaneous luminosities. In black $L_{inst} = 7 \cdot 10^3 3 cm^{-2} s^{-1}$, average processing time of 66.5ms; in blue $L_{inst} = 1.4 \cdot 10^3 4 cm^{-2} s^{-1}$, average processing time of 162ms)

the CMS experiment is planning a major upgrade in order to cope with an expected average number of overlapping collisions per bunch crossing of 140.

Graphics Processing Units (GPUs) are massively parallel architectures that can be programmed using extensions (such as CUDA) to the standard C and C++ languages. In a synchronous system GPUs are proved to be highly reliable and show a deterministic time response even in branch divergences. These two features allow GPUs to be perfectly suited to run pattern recognition algorithms on detector data in a trigger environment. From the physics perspective, such an enhancement of the trigger capabilities would allow inclusion of new tracking triggers and the selection of events that are currently not recorded efficiently. Moreover nowadays GPUs are becoming more widely used in scientific computing and so they are getting cheaper and better supported by the manufacturers. Thus rethinking of tracking algorithms in parallel could be a future-proof solution to the present and future track reconstruction issues explained above. During the first year part of the research work has been collaborating with the LHC CMS tracking software group in order to develop a new version of the tracking algorithm for the silicon pixel tracker based on a parallelised version of *Cellular Automaton*. A *Cellular Automaton* is a mathematical model used to describe the evolution of discrete complex systems. It consists of a finite graph of interconnected *cells* and the number of values that the state of each of these

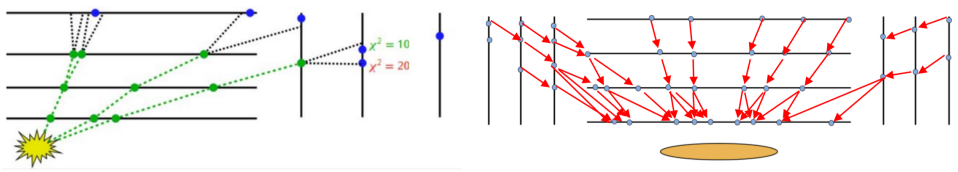

Figura 6: Comparison between the current quadruplet seeds creation procedure based on the propagation of a triplet to the fourth layer (left) and the *Cellular Automaton* method which create all the possible hit pairs from pairs of adjacent layers and the joins compatible pairs that share hits checking their compatibility.

cells can assume is finite. The state that a *cell* will assume depends on the states of the neighbouring cells and on its previous state: this dependence defines the rules of evolution of the graph. Given as initial state (i.e. the graph with its connections between neighbouring cells, and the set of rules of evolution), the system is let free to evolve for a given number of generations. In the current algorithm the quadruplets seeds for building tracks are created from triplets propagation in three main steps starting from the creation of an hit pair, adding then in two steps a consistent third hit and a consistent fourth hit on the outer layers of the pixel tracker. The new under development *Cellular Automaton* based algorithm creates instead all the possible doublets or cells between all the possible layer pairs in order to have as much as data as possible to be fed to the *Cellular Automaton* itself running on a GPU node. The latest tests have been quite encouraging since they shows that the new parallelised version of the algorithm both retains and sometimes exceeds the present procedure efficiency and fake rejection performance (see Figures 7 and 8) and also shows much better timing performance(see Figures 9). From Figure 8 is apparent that the new procedure ,at some point, is performing slightly worse than the former triplet propagation method. Even though, given the low occupancy of the GPU node during track reconstruction, this kind of problem do not affect the timing performance of the *Cellular Automaton*, one of the next development will be adding a further step in the tracking workflow chain a further doublets filtering step based on matching hits pixel clusters with machine learning and pattern recognition techniques.

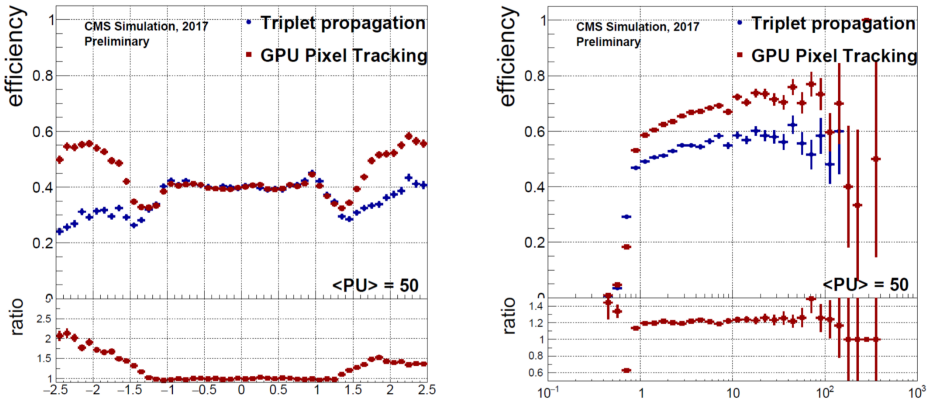

Figura 7: Physics performance for a 13 TeV $t\bar{t}$ simulated event at pileup 50: track reconstruction efficiency with respect to track p_t and η.

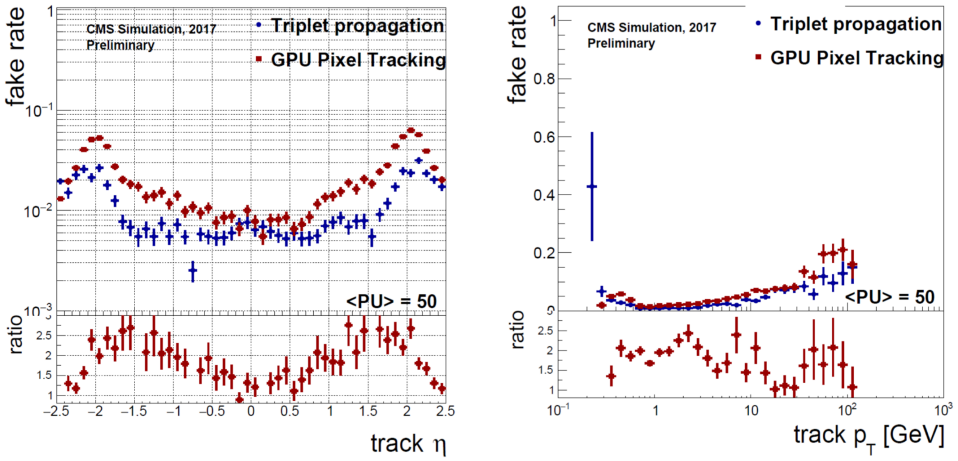

Figura 8: Physics performance for a 13 TeV $t\bar{t}$ simulated event at pileup 50: fake track reconstruction with respect to track p_t and η.

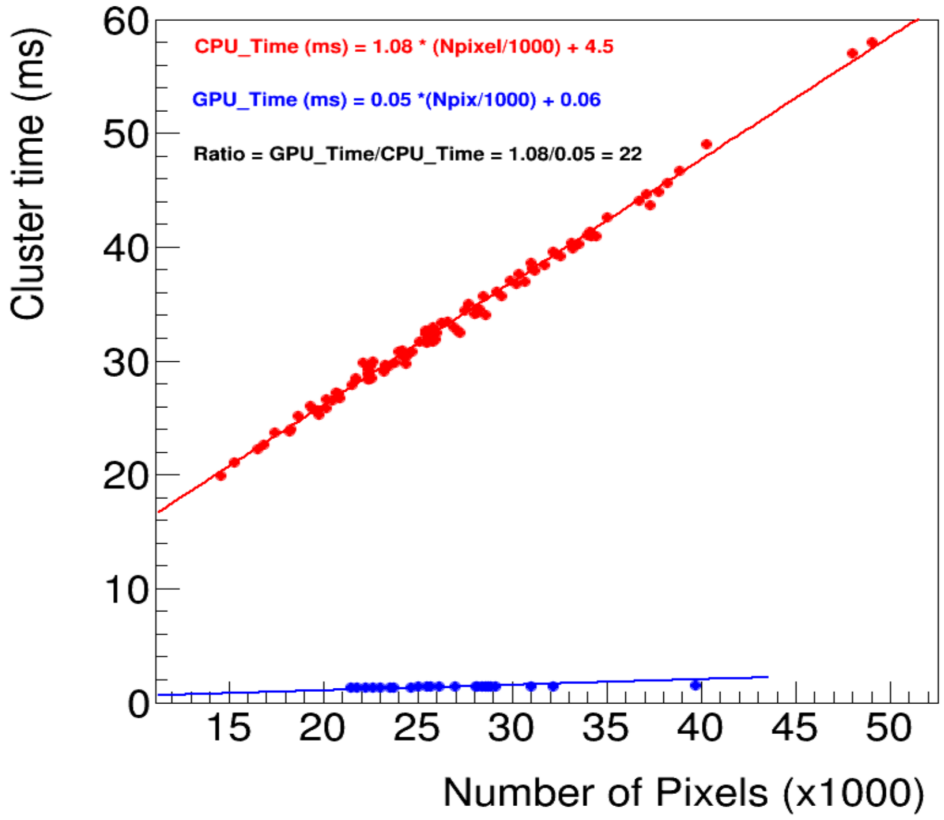

Figura 9: Timing performance comparison of the present CPU based triplet propagation algorithm (running on *Intel 477K* CPU cores) and the GPU based *Cellular Automaton* algorithm (running on a *Tesla K40* GPU board).

2 Didactic work

Here is reported the didactic work. Schools, PHD courses, conferences and workshops attend during the first year are reported

PHD Courses

- Course of "Management and knowledge of European research model and promotion of research results"
- Course of "Inroduction and advanced C++ programming"
- Course of "Statistical and computational model of data analysis"
- Course of Inglese "How to prepare a technical speech in English"
- Course of "Standard model and beyond"
- Course of "Programming with Python"
- Course of "Gaseous Detectors"

Schools attended

- CERN School of Computing 2016 (Mol, Belgio, SKC-CEN, 28 Agosto - 9 Settembre 2016)
- XXVIII Seminario Nazionale di Fisica Nucleare e Subnucleare "Francesco Romano" (Otranto, 3-10 Giugno 2016)
- Corso intensivo di programmazione CUDA di schede grafiche (Dipartimento Interateneo di Fisica di Bari, 11-13 Maggio 2016)

Workshop and conferences:

- Riunione Commissione Calcolo e Reti INFN (Roma, 5 Luglio 2016)
- 22nd International Conference on Computing in High Energy and Nuclear Physics, CHEP 2016 (San Francisco, 10-14 Ottobre 2016)
- Giornate di Studio sul Piano Triennale INFN (Catania, 3-4 Dicembre 2015)
- WLCG Workshop (San Francisco, 8-9 Ottobre 2016)
- CMS Physics Week (CERN, 8-12 Febbraio 2016)

Adriano Di Florio