PROCEEDINGS OF SCIENCE

Estimation of global statistical significance of a new signal within the GooFit framework on GPUs

Adriano Di Florio[*]
Dipartimento Interateneo di Fisica di Bari and I.N.F.N.-Sezione di Bari
via G. Amendola 173, 70126 Bari, Italy
E-mail: adriano.diflorio@ba.infn.it

GPUs represent one of the most sophisticated and versatile parallel computing architectures that have recently entered in the HEP field. GooFit is an open source tool interfacing ROOT/RooFit to the CUDA platform that allows to manipulate probability density functions and perform fitting tasks. The computing capabilities of GPUs with respect to traditional CPU cores have been explored with a high-statistics pseudo-experiment method implemented in GooFit with the purpose of estimating the local statistical significance of an already known signal. The striking performance obtained by using GooFit on GPUs has been discussed in the previous edition (XII) of this conference. This method has been extended to situations when, dealing with an unexpected new signal, a global significance must be estimated. The LEE is taken into account by means of a scanning/clustering technique in order to consider, within the same background only fluctuation and anywhere in the relevant mass spectrum, any fluctuating peaking behaviour with respect to the background model. The presented results clearly indicate that the systematic uncertainty associated to the method is negligible and that the p-value estimation is not affected by the clustering configuration. A comparison with the evaluation of the global significance provided by the method of trial factors is also provided.

XIII Quark Confinement and the Hadron Spectrum - Confinement2018
31 July - 6 August 2018
Maynooth University, Ireland

[*] Speaker.

1. Introduction to `GooFit`

The word *GPU-accelerated 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. The computing capabilities are enhanced once a sequence of elementary arithmetic operations are performed in parallel on a huge amount of data. In the context of High Energy Physics (HEP) analysis application, `GooFit` is an under development open source data analysis tool, used in applications for parameter estimation, that interfaces ROOT [3]/RooFit [4] to the CUDA [5] parallel computing platform on nVidia's GPUs (it also supports OpenMP). `GooFit` acts as an interface between the MINUIT [6] minimization 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) minimization step on the *host side* (CPU) while the PDF/NLL is evaluated on the *device side* (GPU) [7]. Description and details about `GooFit` can be found elsewhere [1].

2. `GooFit` performances for Monte Carlo toys

Monte Carlo pseudo-experiments (MC toys) are used to estimate the probability (*p-value*) that background fluctuations would - alone - give rise to a signal as much significant as that seen in the data. To test the computing capabilities of GPUs with respect to CPU cores, a high-statistics MC toys technique was implemented both in ROOT/RooFit and `GooFit` frameworks [8] with the aim to estimate a *p-value* and specifically the local statistical significance of the structure observed by CMS close to the kinematical threshold of the $J/\psi\phi$ invariant mass in the $B^+ \to J/\psi\phi K^+$ decay [13].

The used hardware setup consists in two servers, one equipped with two nVidia TeslaK20 and 32 cores ($16+16$ by Hyper-Threading) and the other with one nVidia TeslaK40 and 40 ($20+20$) cores [14]. To efficiently run RooFit MC toys on the 72 CPUs available on the two servers hosting the GPUs, the PROOF-Lite [15] tool is used. On the other hand the *nVidia* Multi Process Service tool [16] allows the execution of - up to 16 - simultaneous processes on the same GPU acting as a scheduler and allowing a balanced full usage of the GPU. The optimized

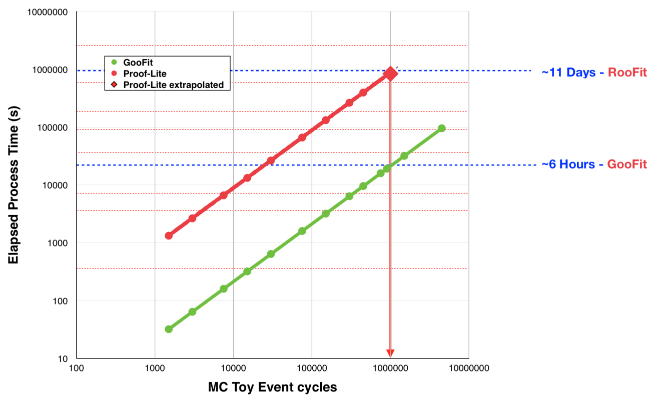

Figure 1: Comparison for the elapsed time employed with two TeslaK20 and one TeslaK40 together as a function of the number of MC toys; `GooFit`/MPS runs 48 concurrent processes while RooFit/PROOF-Lite runs on 72 CPUs. For 1M toys the red diamond point shows the extrapolated time (about 11days) for the RooFit application.

`GooFit` application running on GPUs has provided [8] [9] striking speed-up performances with respect to the RooFit application parallelized on multiple CPUs by means of PROOF-Lite tool. In particular, from the point of view of the end-user analyst having at its own disposal all the 72 CPU cores and the three GPUs, it has been measured that 1M of MC toys can be produced in about 11days with RooFit/PROOF-Lite and in about 6 hours only with `GooFit`/MPS (Fig.1). The extension of this method when a new unexpected signal is reconstructed will be presented in Section 4.

3. Exploring the applicability limits of Wilks theorem

By means of `GooFit` it has also been easier to explore the (asymptotic) behaviour of a likelihood ratio test statistic in different situations in which the Wilks theorem may apply or does not apply because its regularity conditions are not satisfied. The Wilks theorem [10] is often used to estimate the p-value associated to a new signal. Given two hypothesis, the *null* one, H_0, with ν_0 degrees of freedom (dof) and an *alternative* one, H_1, with ν_1 dof, any test statistic t, defined as a likelihood ratio $-2ln\lambda = -2ln(L_{H_0}/L_{H_1})$, or similarly (in the asymptotic limit) as a $\Delta\chi^2 = \chi^2_{H_0} - \chi^2_{H_1}$, approaches a χ^2 distribution with $\nu = \nu_1 - \nu_0$ dof, provided that the following regularity conditions hold:

1. H_0 and H_1 are nested (H_1 includes H_0),
2. while $H_1 \to H_0$, the H_1 parameters are well behaving (well defined and not approaching some limit),
3. asymptotic limit (namely in the enough large data sample regime).

Once this theorem can be applied, the p-value associated to the signal is $p = \int_{t_{obs}}^{\infty} \chi^2_{\nu_1-\nu_0}(t)dt$ and the use of pseudo-experiments to estimate the p-value is not needed (even if still suggested). When null hypothesis is background-only and the alternative is background plus signal, often the above conditions are not all satisfied, and the MC toys are mandatory. Indeed this is the case previously studied. The signal parameters in the model of H_1 hypothesis are: mass (m), width (Γ) and yield ($\mu \geq 0$); when $H_1 \to H_0$ the problem is that not only m and Γ are not well defined but also μ tends to the null limit. This explains the previous use of a MC toys technique. In general the distributions of a test statistic are not predictable and thus need to be extracted from pseudo-experiments. MC toys according to the previously discussed procedure and physics case have been generated for each of the following 4 cases: (i) m and Γ fixed, μ free; (ii) m and Γ fixed, μ free but constrained to be positive; (iii) m and Γ free, μ free; (iv) m and Γ free, μ free but constrained to be positive. The $\Delta\chi^2$ distributions for the four cases are shown in Fig. 2. The fourth case was the one studied so far (with much higher statistics).

3.1 Second case: m and Γ fixed, μ free but constrained to be positive

Let us consider the special case of the test statistic t_μ with the purpose to test $\mu = 0$ in a model where is assumed $\mu > 0$; rejecting the null hypothesis ($\mu = 0$) leads to the discovery of a new signal. In this case, following [11], the test statistic is $q_0 = -2ln\lambda(0)$ if the estimated signal strength $\hat{\mu} \geq 0$ while is null otherwise, with $\lambda(0)$ being the profile likelihood ratio for $\mu = 0$. The authors of [11] derive analytically that an asymptotic approximation for the PDF of the statistic q_0 under assumption of the background-only ($\mu = 0$) hypothesis is an equal mixture of a delta function at 0 and a chi-square distribution for one dof:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi q_0}}e^{-q_0/2}.$$

A fit to the test statistic distribution with a model consisting in a linear combination of a χ^2_ν function and a narrow step function at zero has been performed (Fig. 3), where the likelihood ratio distribution was obtained by the already discussed fit procedure but when fixing the values of mass and width parameters to the CMS estimates previously obtained, while leaving μ free. The best estimates obtained for the number of dof and the coefficient in front of the step function are $\hat{\nu} \simeq 0.992 \pm 0.001$ and $\hat{c} \simeq 0.507 \pm 0.001$, namely close to the approximate theoretical prediction. A chi-square test returns a 3.5% probability for this fit.

4. Clustering approach to address the LEE

When a new unexpected signal is reconstructed in HEP, the *global* significance of the associated peak needs to be estimated and the Look Elsewhere Effect [12] must be taken into account. This implies to consider, within the same background-only fluctuation and everywhere in the relevant mass spectrum, any random peaking behaviour with respect to the expected background model. Thus a scanning technique based on a clustering approach has been developed.

Beforehand a pseudo-data invariant mass distribution of 15K candidates in a generic region of interest, namely $[1,18]$GeV, has been generated according to a fictitious 7^{th} order polynomial background model on the top of which any desired amount of a *significant* signal, mimicked by a Voigtian model, can be artificially added close to 8GeV (as for instance in Fig.4). At this mass value a 60MeV mass resolution is considered. The fits to the pseudo-data distribution of Fig.4

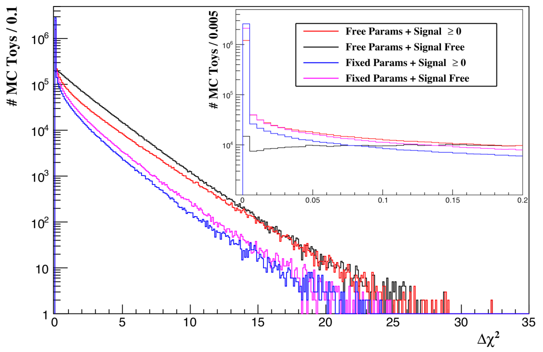

Figure 2: Different test statistic ($\Delta\chi^2$) distributions for the 4 cases discussed in the text, with the same number ($2M$) of MC toys.

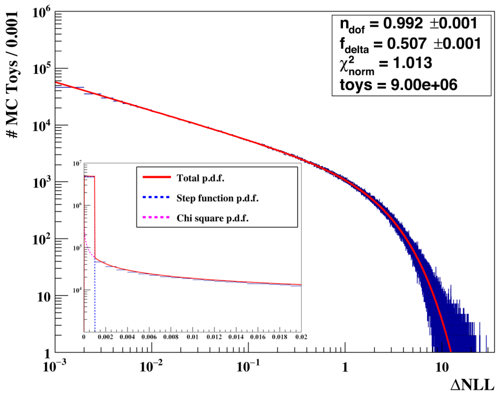

Figure 3: Fit to the ΔNLL distribution of case (2). Fit model has two components: a very narrow step function and a χ^2_ν.

are performed accordingly: the background-only model (the *Null Hypothesis* H_0) is a 7^{th} order polynomial function whereas the background+signal model (the *Alternative Hypothesis* H_1) is obtained by adding a Voigtian function. The resolution values in the latter are reasonably increased as a function of the increasing invariant mass while satisfying the 60MeV constraint at 8GeV. By performing the H_0 and H_1 fits, the (local) statistical significance of this peak is $Z\sigma = 5.5\sigma$ with Z approximately estimated by:

$$Z \simeq \sqrt{-2[ln(L_{H_1}) - ln(L_{H_0})]} \tag{4.1}$$

where L_{H_0} (L_{H_1}) is the likelihood evaluated for H_0 (H_1) hypothesis [17]. The MC toys method is configured as follows. As first step of each toy iteration, a distribution based on the background-only model is generated over the whole mass spectrum and the H_0 fit is performed. As a second step the clustering technique acts on each generated pseudo-experiment as follows:

1. search for a *seed* bin, namely for a bin whose content fluctuates more than $x\sigma$ strictly above the value of the background function in the center of that bin (σ is the statistical error associated to the considered bin).

2. Add any *side* bin to the *seed* bin if it holds a content that fluctuates more than $z\sigma$ strictly above the value of the background function in the center of that bin, otherwise the *seed* bin forms a 1-bin cluster.

3. Check also for *light* seeds, namely bins that fluctuate more than $y\sigma$ with $z < y < x$ and with at least a *side* bin fluctuating more than $z\sigma$. In case of positive result a cluster is formed.

In the third step, a series of independent H_1 fits is performed by cycling on the clusters collected in the clustering step. At the end of this step the fit with the best ΔNLL (the test statistic) is chosen. On the whole a ΔNLL distribution is obtained over all the processed MC toys.

A set of *baseline* clustering parameters $(x, y, z) = (2.25, 1.50, 1.00)$ has been chosen in order to satisfy two concurrent requirements: not missing any possible interesting fluctuation and avoiding selecting too many irrelevant fluctuations. This *baseline* configuration has been run on about 76M pseudo-experiments and the ΔNLL distribution is shown in Fig.5, with the superimposed red line indicating the ΔNLL_{data} value for the pseudo-data.

Table 1: Mean number of alternative hypothesis fits per toy ($< fit_{H_1} >$) and fraction of toys with no fit (f_{nofit}) for the three different clustering configurations described in the text.

Clustering configs.	$< fit_{H_1} >$	f_{nofit}
Tight (3.00, 1.75, 1.00)	2.2	∼10%
Baseline (2.25,1.50, 1.00)	4.5	∼1%
Loose (2.00, 1.25, 1.00)	6.6	0.1%

The global *p-value* is then estimated by:

$$p = \int_{\Delta NLL_{data}}^{\infty} f(\Delta NLL) d(\Delta NLL) \simeq \frac{9.820 \cdot 10^2}{7.584 \cdot 10^7} \simeq 1.295 \cdot 10^{-5} \tag{4.2}$$

This corresponds to the *global* statistical significance $Z\sigma = \Phi^{-1}(1-p)\sigma \simeq 4.22\sigma$, through the inverse function of the cumulative distribution of the standard Gaussian. As expected by considering the LEE, the *global* significance is relevantly lower than the estimated *local* one.

5. Evaluation of the possible systematic uncertainty

In order to test the behavior of the method and to estimate the possible systematic uncertainty associated to the clustering technique, three sets of configuration parameters, i.e. three values for the (x, y, z) parameters, have been carefully considered. After some tests with different cuts two further configurations are chosen besides the baseline clustering cuts: a set of tighter values $(3.00, 1.75, 1.00)$ and a set of looser values $(2.00, 1.25, 1.00)$. The Tab.1 reports details about these three clustering configurations such as the average number of H_1 fits per toy and the fraction of toys with no fit. These three configurations have been run on a same common set of 45M fluctuations and the three corresponding ΔNLL distributions are shown superimposed in Fig.6.

Table 2: Estimated *global* significances for the 3 clustering configurations with respect to different *local* significance values estimated by Eq.(4.1).

Local Significance	**4.0σ**	**4.5σ**	**5.0σ**	**5.5σ**	**6.0σ**
Tight (3.00, 1.75, 1.00)	2.21	2.91	3.58	4.22	4.87
Baseline (2.25,1.50, 1.00)	2.20	2.91	3.58	4.22	4.87
Loose (2.00, 1.25, 1.00)	2.20	2.91	3.58	4.22	4.87

By focusing on the region of interest for the estimation of the statistical significance, i.e. the tail of the ΔNLL distribution ($\Delta NLL > 20$), it is evident that there is no relevant difference (Fig.7) among the three configurations. This can be appreciated by inspecting, in Figure 6 and Figure 7, the normalized deviations $(x-y)/(x+y)$ of the other two distributions with respect to the baseline distribution. This is also confirmed by examining the estimated *global* significances for the *p-values*

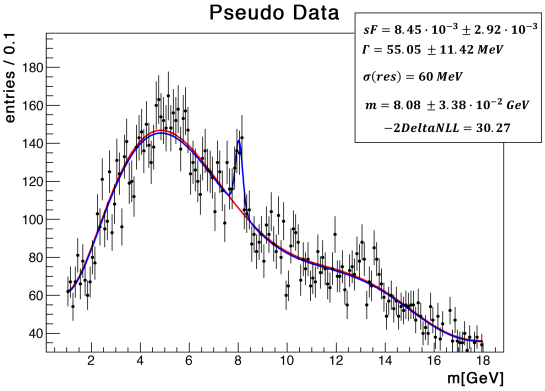

Figure 4: Simulated invariant mass distribution (pseudo-data). $H_0(H_1)$ fit is in red (blue); in the top right box the best values for the estimated parameters of the H_1 model are given.

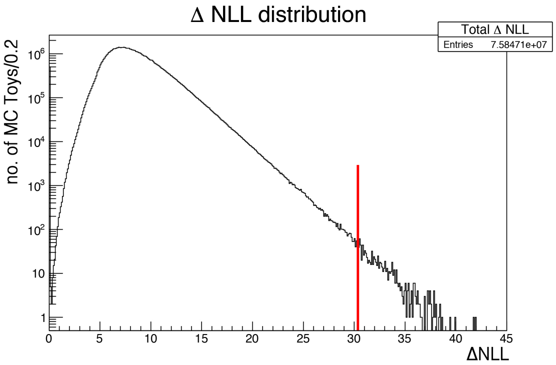

Figure 5: ΔNLL distribution for about 76M toys for the baseline configuration of clustering technique. The red line indicates the $\Delta NLL_{data} \simeq 30.27$ value for the pseudo-data distribution in Fig.4.

Table 3: Estimated *global* significances for the 3 clustering configurations with respect to different *local* significance values estimated by Eq.(4.1).

Local Sig.	**4.0σ**	**4.5σ**	**5.0σ**	**5.5σ**	**6.0σ**
GV method	2.09	2.82	3.48	4.10	4.71
MC Toys	2.20	2.91	3.58	4.22	4.87

corresponding to different values of *local* significances, as reported in Tab.3. It can be concluded that the systematic uncertainty on the *p-values* associated to the method is negligible.

5.1 Tail probabilities of the likelihood ratio test statistic by Gross and Vittels

In their 2010 paper [12], E. Gross and O. Vittels, proposed, among other results, a method to estimate an upper limit for the global p-value when the signal hypothesis (H_1) depends on s nuisance parameters ($\vec{\theta}$) that don't exist under the null hypothesis (H_0). In this section we will explore the applicability of their results to our usecase implementing the procedure described in [12] in `GooFit`. This method is based on counting the number of *upcrossings* of an arbitrary low level c_0 by the chosen test statistics $f(m)$ over the whole spectrum of m. The calculation of this factor acts as a correction to the Wilks' result for *local* significance, expressed as the tail probability of a chi-square distribution with s degrees of freedom:

$$P(q(\vec{\theta}) > c) \leq P(\chi^2_s > c) + < N(c_0) > (\frac{c}{c_0})^{s-1} e^{-(c-c_0)/2} \tag{5.1}$$

where c is the desired threshold for the test statistics, typically the value observed in the data. Thus we now apply this method to the usecase presented in the previous sections.In this example there are two nuisance parameters, the peak mass m and width Γ, and thus $\nu = 2$ degrees of freedom.

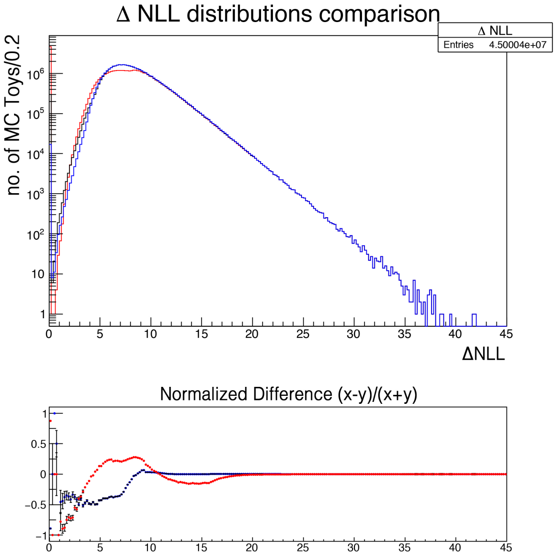

Figure 6: ΔNLL distributions for 45M of common fluctuations for the 3 configurations: baseline (black), tight (red) and loose (blue).

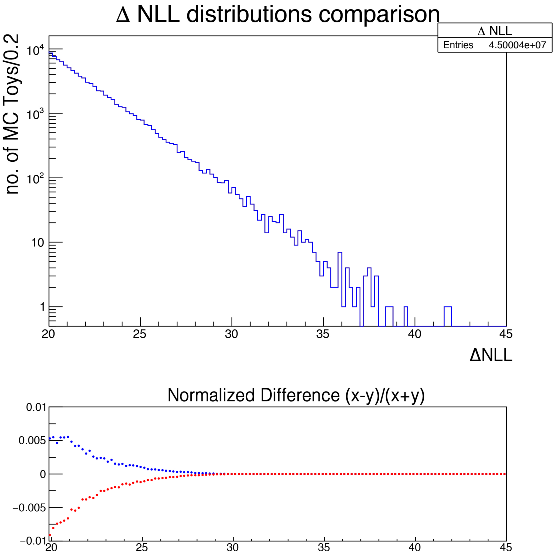

Figure 7: The same 3 ΔNLL distributions of Fig.6 once zoomed into the range 20.0-45.0 to inspect their tail behaviour.

We use a binned profile likelihood ratio as our test statistic, where the number of events in each bin is assumed to be Poisson distributed with an expected value $E(n_i) = \mu s_i(m;\Gamma) + (1-\mu)b_i$, where μ is the signal strenght parameter, or signal fraction. The chosen test statistics is the ΔNLL. As reference level we choose $c_0 = \nu - 1 = 1$. The procedure to estimate $\langle N(c_0)\rangle$ is set up as follows:

1. the mass distribution based on the background-only model H_0 the MC toys method is configured starting from the generation.
2. An H_1 fit is performed fixing the peak mass value to m; this fit is repeated $n = 1000$ times changing m and Γ in continuous steps in order to scan the whole mass spectrum.
3. At each mass point m the profile likelihood $f(m)$ is calculated and the distribution $f(m)$ along the mass spectrum is build (see Figure 8).
4. The number of *upcrossings* of $f(m)$ with respect to c_0 is thus readily calculated.

This cycle is performed for each pseudo experiment, an example is shown in Figure 8 together with the trend of $f(m)$. Therefore, this procedure to estimate $< N(c_0) >$ has has been implemented in *GooFit* for our sample pseudo-data and, after 10^4 toys, the result was: $< N(c_0) >= 7.3, \quad \sigma_{N(c_0)} = 2.4, \quad$ and $c_0 = 1.0$. Then the upper limit for $P(f(\hat{m}) > c)$ may be readily calculated and may be compared to the distribution $P_{MC}(f(\hat{m}) > c)$ built from MC toys in the *baseline* configuration:

$$P_{MC}(f(\hat{m}) > c) = \int_c^{\infty} f(\Delta NLL) d(\Delta NLL) \tag{5.2}$$

In Figure 9 and Table 4 the upper limit estimated with the G-V method is compared with the exact function extrapolated from the MC toys. The results are rather compatible.

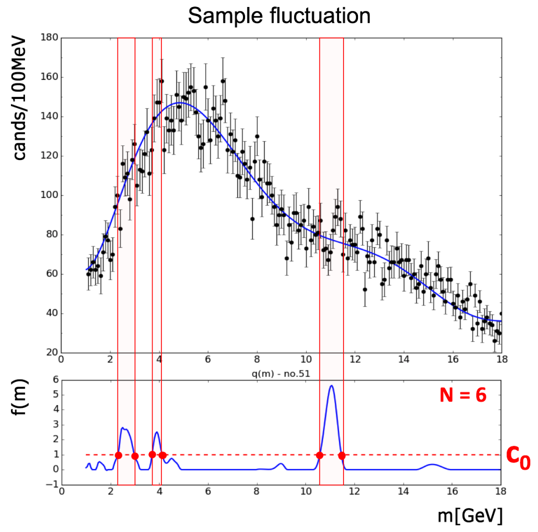

Figure 8: Sample fluctuation used to estimate $< N(c_0) >$. In blue the H_0 fit. In the panel below the $q(m)$ test statistics value as function of the peak mass. The red dashed line shows the c_0 threshold. For this specific sample $N(c_0) = 5$

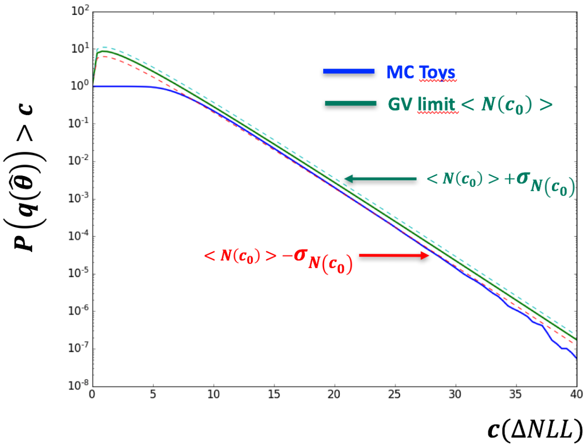

Figure 9: Estimated global significance with respect to different local significance. The upper limit estimated with the G-V method is compared with the exact function extrapolated from the baseline configuration MC toys.

Local Sig.	**4.0σ**	**4.5σ**	**5.0σ**	**5.5σ**	**6.0σ**
GV method	2.09	2.82	3.48	4.10	4.71
MC Toys	2.20	2.91	3.58	4.22	4.87

Table 4: The upper limit estimated with the G-V method is compared with the exact function extrapolated from the MC toys, with baseline clustering configuration

6. Conclusions

The results presented demonstrate the potentialities of GPU computing applied to the data analyses in HEP field. The `GooFit` MC toys application run by means of the MPS provides a striking speed-up with respect to the RooFit application parallelized on multiple CPUs by means of PROOF-Lite. In order 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 local statistical significance of an already known signal.

Also high-statistics pseudo-experiment method, based on a scanning and clustering approach, has been implemented and tested within `GooFit` framework with the purpose to estimate the *global* statistical significance of an unexpected new signal. The presented results clearly indicate that the systematic uncertainty associated to the method is negligible and that the *p-value* estimation is not affected by the clustering configuration. This kind of validation studies has been performed by exploiting the high performance of an optimized `GooFit` application running on GPU-equipped servers.

By means of the used application it has been also possible to explore the behaviour of a likelihood ratio test statistic in different situations in which the Wilks theorem may apply or does not apply because its regularity conditions are not satisfied. Also it has been possible to compare the MC toy results for *global* statistical significance estimation with the upper limit estimation method proposed in [12].

Aknowledgements

We are grateful for valuable support from G.Donvito, manager of the High Performance Cluster hosted by the ReCas [18] Data Center.

References

[1] Andreassen R, Meadows B T, de Silva M and Sokoloff M D *J. Phys.: Conf. Series* **513**, 052003 (2014)

[2] Schreiner H F et al. *arXiv:1710.08826* (2017)

[3] Brun R and Rademakers F *Nucl. Instrum. Meth.* **A 389**, 81-86 (1997)

[4] Verkerke W and Kirkby D P *eConf* **C0303241** MOLT007 (*Preprint* physics/0306116) (2003)

[5] `https://docs.nvidia.com/`; for these studies CUDA version 6.5(7.0) were used for Tesla K20(40) boards.

[6] James F and Roos M CERN Program Library routine D506 (1989) (long write-up)

[7] This can be checked by nVidia Visual Profiler (`nvvp`)).

[8] Pompili A and Di Florio A, *J. Phys. Conf. Ser.* **762**, 012044 (2016) [ACAT 2016]

[9] L Cristella, A Pompili, A Di Florio, *EPJ Web Conf.* **137** 11005 (2017) [XII QCHS]

[10] Wilks, S. S. *Annals Math. Statist.* **9**, 60-62 (1938)

[11] Cowan G, Cranmer K, Gross E and Vitells O *Eur. Phys. J.* **C71**, 1554 (2011)

[12] Gross E and Vitells O *Eur. Phys. J.* **C70**, 552 (2010)

[13] CMS Collaboration, *Phys. Lett.* **B 734**, 261 (2014)

[14] Tesla K20(K40) has 5(12)GB GDDRs; the 16(20) cores are E5-2640 v2 @ 2.0GHz with 64(256)GB of RAM

[15] `https://root.cern.ch/proof`; see for specific PROOF-Lite applications on multi-core servers: Barbone L, Donvito G and Pompili A, *J. Phys.: Conf. Series* **396** 042017 (2012)

[16] Multi Processes Server `https://docs.nvidia.com/deploy/pdf/CUDA_Multi_Process_Service_Overview.pdf`

[17] I.Narsky and F. C. Porter, *Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning*, Wiley, 2013.

[18] ReCas is a project financed by the italian MIUR (*PONa3_00052, Avviso 254/Ric.*); its web page is `http://www.recas-bari.it/index.php/en/`.