GPUs for statistical data analysis in HEP: a performance study of *GooFit* on GPUs vs. *RooFit* on CPUs

Alexis Pompili and Adriano Di Florio
(on behalf of the CMS Collaboration)

Dipartimento Interateneo di Fisica di Bari and I.N.F.N.-Sezione di Bari,
via Amendola 173, I-70126 Bari, Italy

E-mail: `alexis.pompili@ba.infn.it`

Abstract. 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 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 $B^+ \to J\psi\phi K^+$. GooFit is a data analysis open tool under development that interfaces ROOT/RooFit to CUDA platform on nVidia GPU. The optimized GooFit application running on GPUs hosted by servers in the Bari Tier2 provides striking speed-up performances with respect to the RooFit application parallelised on multiple CPUs by means of PROOF-Lite tool. The considerably resulting speed-up, while comparing concurrent GooFit processes allowed by CUDA Multi Process Service and a RooFit/PROOF-Lite process with multiple CPU workers, is presented and discussed in detail. By means of GooFit it has also been 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.

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`[1] is an under development open source data analysis tool, used in applications for parameters' estimation, that interfaces ROOT[2]/RooFit[3] to the CUDA[4] parallel computing platform on nVidia's GPUs (it also supports OpenMP). `GooFit` acts as an interface between the MINUIT[5] 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)[6]. Description and details about `GooFit` can be found elsewhere[1]. In this study a comparison between RooFit and `GooFit` performances is presented when a huge amount of pseudo-experiments need to be fitted several times with the aim to estimate a p-value and thus the statistical significance of a new signal reconstructed from the data. 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[7].

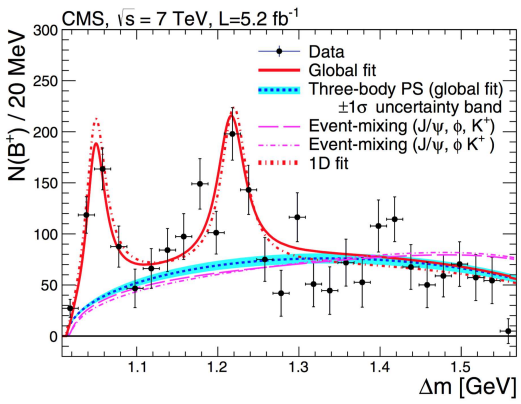

Figure 1. Fits to the background-subtracted $J/\psi\phi$ invariant mass in the $B^+ \to J/\psi\phi K^+$ decay[8]; the significance of the left peak is estimated in this study.

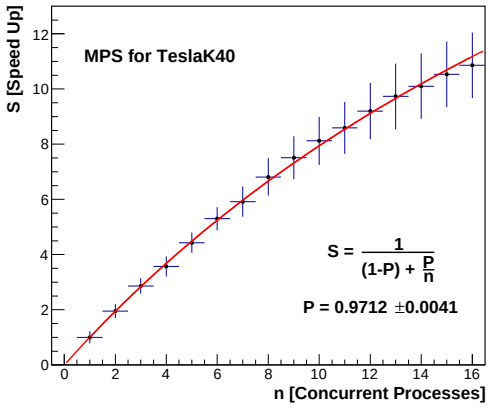

Figure 2. Amdhal's fit to speed-up perfomances of `GooFit`/MPS for a GPU TK40 and up to 16 CPUs.

2. Pseudo-experiments for p-value estimation and `GooFit` performances

In order to test the computing capabilities of GPUs with respect to CPU cores a high statistic toy MC technique has been implemented both in `GooFit` and RooFit frameworks to estimate the local 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 $B^+ \to J/\psi\phi K^+$[8]. The found fit parameters (Fig. 1) are compatible with the state Y(4140) observed for the first time by CDF[9].

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. A single toy fit cycle consists in the following sequence of steps:

1) generation of fluctuated background binned distribution according to the 3-body phase-space model (the number of entries are fixed to that in the data thus ignoring Poisson fluctuations);
2) a Binned Maximum Likelihood fit (BML) is done with phase-space model (null hypothesis);
3) 8 BML fits are performed by adding to the phase-space a voigtian model truncated to account for the kinematical threshold (alternative hypothesis); the gaussian resolution function has fixed width ($2MeV$) and the signal yield is constrained to be positive. For each bin the PDF value is estimated by integration over the bin since the signal is steep with respect to the bin size. The 8 fits differ by the starting values (2 masses and 4 widths) within the region of interest defined from CDF values[9] (no need to take the Look-Elsewhere-Effect (LEE) into account).
4) For each fit a $\Delta\chi^2$ value is calculated with respect to the null hypothesis fit and the best value is chosen among the 8 alternative fits. The $\Delta\chi^2$ (the test statistic) distribution is obtained over the whole sample of MC toys.

The speed-up for a `GooFit` single process with respect to a RooFit one is 64(48) for a TK40(20). The `GooFit` process does not exploit the whole computational power of a GPU ($\simeq$ 70%); the Multi Process Server (MPS) allows the execution of up to 16 simultaneous processes on the same GPU acting as a scheduler and allowing a balanced full use of GPU capabilities (each process uses one shared GPU and one exclusively assigned CPU). On the other hand PROOF-lite, a dedicated version of PROOF optimized for single multi-core machines[10], is used to efficiently run RooFit toys in parallel on the 72 CPUs available on the two servers. The speed-up of MPS (PROOF-Lite) with respect to one `GooFit` (RooFit) process as a function of the number of independent processes (workers) follows Amdhal's Law (Fig. 2); in both cases the parallelizable fraction is $\simeq 0.97$ i.e. the serial overhead is $\simeq 3\%$ of the application execution.

A first performance comparison can be obtained on the server hosting the two TeslaK20; when

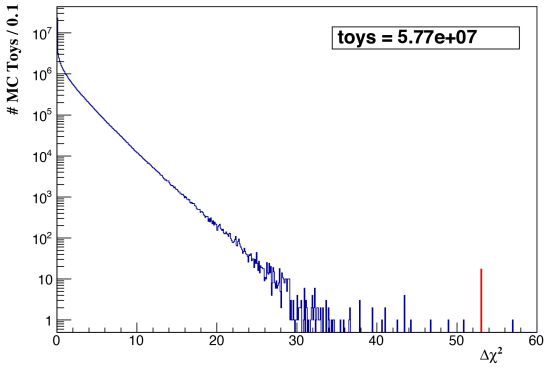

Figure 3. Final $\Delta\chi^2$ distribution with one toy (shown in Fig. 4) exceeding the value observed in the data ($\simeq 53.0$) marked by a red tick.

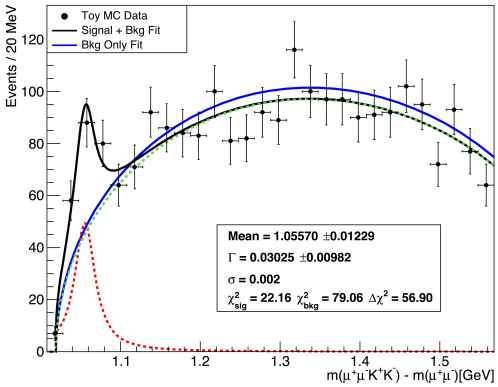

Figure 4. Bkg-only and bkg+signal fits to the MC toy characterized by a $\Delta\chi^2 \simeq 56.9$.

using 2 `GooFit`/MPS jobs on the 2 GPUs and 30 CPUs (each running 15 processes) against one PROOF-Lite job using 30 CPUs, the speed-up is $\simeq 45$ and pretty stable, as expected, with respect to a strongly varying number of MC toys.

A second comparison can be carried out on both the servers hosting both types of GPUs as a function of the number of produced toys, but limited to 16 independent processes because of the MPS limit for the single TK40. Specifically the comparison is between one RooFit/PROOF-Lite job using 16 workers (on 16 CPU cores) and one `GooFit`/MPS job running 16 simultaneous processes on a single TK40 or TK20. When using a TK40(TK20) the speed-up has been measured to be $\simeq 60$ (40); the speed-up representing the gain within the same micro-architecture (TK40 vs TK20) is $\simeq 1.5$. All speed-up values are stable with the number of toys.

A third performance comparison can be done from the end-user's point of view. Assuming the analyst has at his own disposal the 2 servers used in these studies equipped with 3 GPUs and 72 CPU cores, it has been measured that $1M$ of MC toys can be produced in ~ 11 days with RooFit/PROOF-Lite and ~ 6 hours with `GooFit`/MPS. To get a signal statistical significance $> 5\sigma$ a p-value $< 3 \cdot 10^-7$ is needed, namely at least $3.3M$ of MC toys are needed. However, as in the case under investigation, the significance estimation may require many more MC toys.

The final $\Delta\chi^2$ distribution, $f(\Delta\chi^2)$ is shown in Fig. 3; the MC toys production was stopped once a fluctuation with $\Delta\chi^2 > \Delta\chi^2_{obs}$ was found (Fig. 4). The p-value is estimated by:

$$P = \int_{\Delta\chi^2_{obs}}^{\infty} f(\Delta\chi^2) d(\Delta\chi^2) \simeq (57.7 \cdot 10^6)^{-1} \simeq 1.73 \cdot 10^{-8} \quad (1)$$

This corresponds to the statistical significance $Z\sigma = \Phi^{-1}(1-P)\sigma \simeq 5.52\sigma$, through the inverse function of the cumulative distribution of the standard gaussian, that is compatible with the lower limit of 5σ quoted in [8] on the basis of $50.5M$ of MC toys obtained by means of RooFit.

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[11] 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 predictible and can 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: (1) m and Γ fixed, μ free; (2) m and Γ fixed, μ free but constrained to be positive; (3) m and Γ free, μ free; (4) m and Γ free, μ free but constrained to be positive. The $\Delta\chi^2$ distributions for the four cases are shown in Fig. 5. The fourth case was the one studied so far (with much higher statistics). The first two cases have also been studied and are going to be discussed hereafter.

3.1. First case: m and Γ fixed, μ free (either positive or negative)

Let us consider a likelihood ratio test statistic $t_\mu = -2ln\lambda(\mu)$, where μ is the `strength parameter`, as the basis of the statistical test. This can be a test of $\mu = 0$ for purposes of establishing the existence of a signal process.

In this first case, following [12], the PDF of the test statistic $f(t_\mu|\mu) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{t_\mu}}e^{-t_\mu/2}$ asymptotically approaches a $\chi^2_{\nu=1}$ distribution; this is in agreement with the Wilks theorem, with the difference of dof being one and represented by μ.

A fit to the test statistic distribution with a χ^2_ν model has been performed, 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 (alternatively they could have been fixed to CDF values), while leaving μ free. The best estimate obtained for the number of dof is $\hat{\nu} \simeq 1.014 \pm 0.001$, thus close to the approximate theoretical prediction; the goodness of fit is checked using a chi-square test that returns a 11.8% probability.

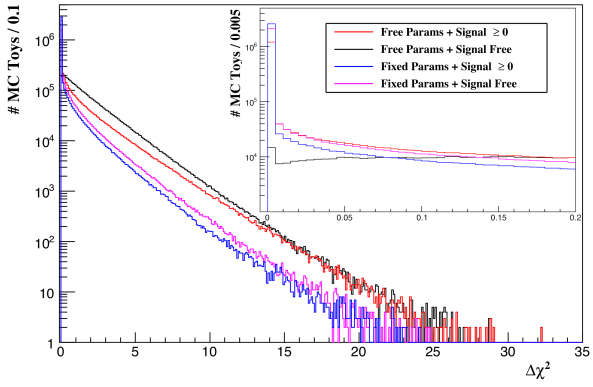

Figure 5. Different test statistic ($\Delta\chi^2$) distributions for the 4 cases discussed in the text, with the same number ($2M$) of MC toys.

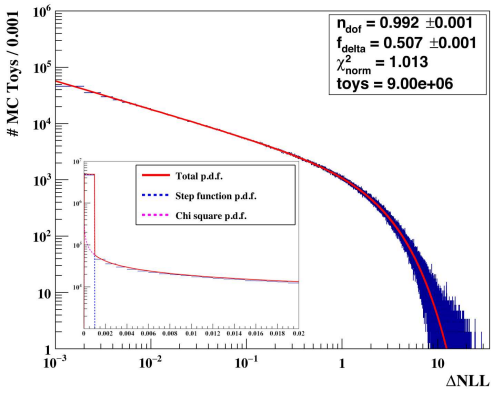

Figure 6. Fit to the ΔNLL distribution of case (2). Fit model has two components: a very narrow step function and a χ^2_ν.

3.2. 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 we assume $\mu > 0$; rejecting the null hypothesis ($\mu = 0$) leads to the discovery of a new signal. In this case, following [12], 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 [12], derive analitically 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. 6), 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. Future developments

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. The method can be extended to situations with an unexpected signal and a global significance must be estimated: the LEE must be considered and a scanning technique must be implemented in order to consider relevant peaking behaviour with respect to the background model everywhere in the mass spectrum within the same fluctuation. This would increase the execution time of a cycle of fits performed on a single fluctuation. Different scan configurations should be tried to evaluate the systematic uncertainty associated to the scan. The RooFit-based approach would be unbearable for the long required processing time and the use of `GooFit` would be mandatory.

Aknowledgements

We are grateful for valuable suggestions and feedback from B.A.Hittle, M.D.Sokoloff and our CMS colleagues T.Dorigo, F.Pantaleo, L.Cristella and G.Donvito. Code development/testing and results were obtained on the IT resources of ReCas[13] hosted by the Bari Tier-2. Support was also provided by the italian *Project 20108T4XTM* of type `PRIN-2010/2011`.

References

[1] Andreassen R, Meadows B T, de Silva M, and Sokoloff M D 2014 *J. Phys.: Conf. Series* **513** 052003. The development of `GooFit` is supported by US NSF grants (*NSF-1005530*, *NSF-1414736*). `GooFit`'s source code lives in the repository `https://github.com/GooFit` that includes a manual and examples. A valuable tutorial can be found at http://indico.cern.ch/conferenceDisplay.py?confId-235992.

[2] Brun R and Rademakers F 1997 *Nucl. Instrum. Meth.* **A 389** 81-86

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

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

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

[6] An intense memory transfer takes place between CPU and GPU (checked by nVidia Visual Profiler (`nvvp`)).

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

[8] CMS Collaboration 2014 *Phys. Lett.* **B 734** 261

[9] CDF Collaboration 2009 *Phys. Rev. Lett.* **102** 242002 and 2011 (arXiv:1101.6058)

[10] Barbone L, Donvito G and Pompili A 2012 *J. Phys.: Conf. Series* **396** 042017

[11] Wilks S S 1938 *Ann. Math. Stat.* **9** 60-62

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

[13] 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/`.