Search for matter-antimatter asymmetries in multi-body decays with GPUs

Shanzhen Chen, Jolanta Brodzicka, Marco Gersabeck, Chris Parkes, Andrew McNab

School of Physics and Astronomy, The University of Manchester, Oxford Road, Manchester M13 9PL, UK
E-mail: shanzhen.chen@cern.ch

Abstract. The LHCb experiment has recorded the world’s largest sample of charmed meson decays. The search for matter-antimatter asymmetries in charm sector uses large event samples and requires high precision analysis and thus intensive computing. This paper describes a powerful method to measure matter-antimatter asymmetries in multi-body decays where GPU systems have been successfully exploited. In this method, local asymmetries in the phase-space distributions were explored with an unbinned approach, and the parallelisation of GPU makes this approach usable in practice for the first time. The performance including GPUs on the Grid will be discussed in detail. With this new method, the world’s best sensitivities to particular decay channels have been achieved.

1. Introduction

The search for CP violation in charm is currently a highly active field of research. In the standard model (SM), CP violation in charm decays is very suppressed. However, many new physics models predict enhancements on CP violating phases in charm decays, beyond the $O(10^{-3})$ level generally predicted within SM [1, 2]. Large samples of charm events in relevant modes (> 100k events) have become available at the B-factories and the LHC. Thus, the studies of CP violation in charm decays provide a unique opportunity for probing New Physics beyond SM.

A promising avenue for observing CP violation effects is through the analysis of multi-body decays. Decays into three body final states can be analysed using a Dalitz plot [3] and higher body decays through generalisations of this approach. The rich structures of resonances and their interfering amplitudes allow the possibility to locate CP violating local regions within the phase space.

There are broadly two categories of methods used to search for CP violation in multi-body decays. The first is an amplitude analysis, which is a model dependent method. The other is to perform a model-independent search. Typically this latter approach has been performed with a binned $\chi^2$ approach to compare the phase space of the particle and anti-particle sample. This is often referred to as the $S_{CP}$ method [4]. A model independent unbinned statistical method known as the energy test was introduced in Ref. [5]. Ref. [6] applies this to Dalitz plot analyses and demonstrates in a simple model that this has significantly improved sensitivity to CP violation over the standard binned approach.
The energy test relies on a pairwise calculation that loops over all pairs of event, which leads to the number of calculations increasing quadratically with the sample size. The LHCb experiment has recorded charmed decays with a sample size exceeding any other experiment, leading to unprecedented precision in charmed meson $CP$ violation measurements. However, large samples require large amounts of calculations. Execution of all the calculations in a single Central Processing Unit (CPU) thread takes a prohibitively long time. Even though modern multi-core CPUs have realized thread-level parallelism, the number of parallel threads is still very limited.

A Graphics Processing Unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images for output to a display. Their highly parallel structure makes them more effective than CPUs for algorithms where large blocks of data are processed in parallel. The implementation of the energy test with GPUs successfully exploits its parallel calculation power and significantly reduces processing time, making the energy test feasible for the first time in $CP$ violation searches.

The first application of the energy test method to LHCb data is described in Ref. [7], and the world-best sensitivity on $CP$ violation in the $D^0 \rightarrow \pi^- \pi^+ \pi^0$ (charge conjugate decays are implied unless stated otherwise) decay channel is achieved.

2. Energy test

The energy test is an unbinned model-independent statistical method to search for time-integrated $CP$ violation in multi-body decays. This method relies on the comparison of the probability density functions of the two flavour conjugated samples in phase space.

A common choice of the three-body decay phase space is the Dalitz plot. Typically modern Dalitz plots use two of the three possible two-daughters invariant mass squares as axes since there are only two degrees of freedom in an all scalar three-body decay. In the absence of resonances, the allowed region of phase space in the Dalitz plot is uniformly populated with events, therefore any observation of structure reflects the dynamics of the decay. If enhanced populations are observed in certain mass regions, the presence of a resonance is indicated. The width and angular distribution can reveal the width and spin structures of the resonance, and distorted or twisted mass bands indicate the interference between resonances. These features of the Dalitz plot make it a perfect tool in the studies of three-body decays.

In the energy test, a test statistic $T$ is formed to realize the comparison, defined as

$$T \approx \frac{1}{n(n-1)} \sum_{i,j>i}^{n} \psi_{ij} + \frac{1}{\bar{n}(\bar{n}-1)} \sum_{i,j>i}^{\bar{n}} \bar{\psi}_{ij} - \frac{1}{n\bar{n}} \sum_{i,j}^{n,\bar{n}} \psi_{ij}. \quad (1)$$

Here $n$ and $\bar{n}$ are the number of events in the two flavour conjugated samples, the metric function $\psi_{ij} \equiv \psi(d_{ij}) = e^{-d_{ij}^2/2\sigma^2}$ is chosen to be Gaussian function with a tunable metric parameter $\sigma$, and for a three-body decay, $d_{ij} \equiv |\Delta \vec{r}_{ij}| = |(m_{2a}^2 - m_{2b}^2, m_{2b}^2 - m_{2c}^2, m_{2c}^2 - m_{2a}^2)|$. This $d_{ij}$ is the distance between two events in a three-body decay phase space, where the $a$, $b$, $c$ subscripts indicate the final-state particles. In our realisation, all three invariant masses are used in the calculation of distances between events. This does not add information, but avoids an arbitrary choice that could impact the sensitivity of the method to different $CP$ violation scenarios.

The three terms in the $T$ formula sum over the weighted distances among matter decay events, anti-matter decay events, and between events in different sets, respectively. For a $CP$ violating sample, events are more likely to gather at different regions in phase space for $CP$ conjugated decays. Therefore, the average distance in each flavour-tagged set is smaller than the average distance between the two different sets. Since the metric $\psi(d_{ij})$ is falling with increasing distance, a $CP$ asymmetry leads to a large $T$ value.
Figure 1. Energy test outputs for a simulated $D^0 \rightarrow \pi^- \pi^+ \pi^0$ sample. (a) Distribution of permutation $T$ values and nominal $T$ value for a simulated sample with 2% CP violation in the amplitude of the $\rho^+(770)$ resonance. Permutation $T$s are fitted with a GEV function and the nominal $T$ value is shown as a vertical line. (b) $T_i$ value distributions, and (c) local asymmetry significances are plotted on Dalitz plots for the same simulated sample. Figures reproduced from Ref. [7].

A permutation method is used to simulate samples without CP violation by randomly reassigning the flavour of each candidate. By comparing the nominal $T$ value observed in data to a distribution of $T$ values obtained from permutation samples, a $p$-value under the hypothesis of CP symmetry is obtained as the fraction of permutation $T$ values greater than the nominal $T$ value.

A statistical uncertainty of the $p$-value is obtained as the binomial standard deviation, reflecting the limited number of permutation tests. If large CP violation is observed, the nominal $T$ value tends to be lying outside the bulk of permutation $T$ values (see Fig. 1(a)). In this case, the permutation $T$ distribution can be fitted with a generalised extreme value (GEV) function [5], which is given as

$$f(T; \mu, \delta, \xi) = N \left[ 1 + \xi \left( \frac{T - \mu}{\delta} \right) \right]^{(-1/\xi)-1} \times \exp \left\{ - \left[ 1 + \xi \left( \frac{T - \mu}{\delta} \right) \right]^{-1/\xi} \right\},$$

with normalisation $N$, location parameter $\mu$, scale parameter $\delta$, and shape parameter $\xi$. The $p$-value from the fitted $T$ distribution can be calculated as the fraction of the integral of the function above the nominal $T$ value. The uncertainty on the fitted $p$-value is obtained by randomly resampling the fit parameters within their uncertainties, taking into account their correlations, and by extracting a $p$-value for each of these samplings.

The contribution of each single event from one flavour, $T_i$, and from the opposite flavour $\bar{T}_i$, to the total $T$ value is defined as

$$T_i = \frac{1}{2n(n-1)} \sum_{j \neq i}^{n} \psi_{ij} - \frac{1}{2n\pi} \sum_{j}^{\pi} \psi_{ij},$$

$$\bar{T}_i = \frac{1}{2n(n-1)} \sum_{j \neq i}^{n} \psi_{ij} - \frac{1}{2n\pi} \sum_{j}^{\pi} \psi_{ij}. \tag{3}$$

A permutation method is also used here to define the level of significance of each event.
The visualization of regions with significant asymmetries is obtained by plotting the asymmetry significance in terms of each event’s contribution on the Dalitz plot (Fig. 1(b)(c)).

3. Energy test with GPUs

The principal drawback of this unbinned statistical method is when calculating the $T$ value for a sample, the distances between every pair of events in this sample are considered, which leads to a significant computational time for large samples. Furthermore, a large number of permutations are required for the random flavour comparison samples to get a sufficient precision on the probabilistic interpretation of the $T$ value, for example to demonstrate that evidence ($>3\sigma$) for $CP$ violation was observed over one thousand permutations would be needed.

One solution to this time consuming problem is to implement highly parallelised graphics processors in the studies. The GPU, as a device dealing with millions of pixels in the monitor, is born to be good at fast simple operations. One way to turn the massive computational power of a modern graphics processor into general-purpose computing power is to use the Compute Unified Device Architecture (CUDA) [8] and Thrust library developed by NVIDIA. CUDA is a general purpose parallel computing platform and programming model that gives developers direct access to the virtual instruction set and memory of the parallel computational elements in CUDA GPUs. Thrust is a template library for CUDA C++ programming. Thrust provides Standard Template Library (STL) like interfaces for GPU programming, which contains parallel algorithms and data structures designed for high performance parallel computing. The Thrust features that used in the energy test are listed in Table. 1.

| Data structures       | Algorithms                  |
|-----------------------|-----------------------------|
| thrust::device_vector | thrust::sort                |
| thrust::host_vector   | thrust::transform_reduce     |
|                       | thrust::unary_fuction       |
|                       | thrust::copy                |

Thrust library provides two vector containers: host_vector and device_vector. As the their names suggest, host_vector is stored in host memory for program input and output while device_vector stored in GPU device memory for fast iteration. thrust::copy could transfer data from host to device, or result from device to host. The Gaussian function is formed via thrust::unary_fuction, and Gaussian function inputs are sorted via thrust::sort.

The core algorithm of the energy test is to loop over all events and calculate the interactions between every two events in phase space. For event $i$, the interactions to all other same sign and opposite sign events, $\sum_{j}^{n} \psi_{ij}/n - 1$ and $\sum_{j}^{n} \overline{\psi}_{ij}/n$ are calculated via

```
thrust::transform_reduce( dev_data_d0->begin(), dev_data_d0->end(),
                         gaussian_function, (double)0., plus<double>() );
```

and

```
thrust::transform_reduce( dev_data_d0bar->begin(), dev_data_d0bar->end(),
                         gaussian_function, (double)0., plus<double>() );
```

This transform_reduce function applies Gaussian function to every vector elements in parallel, and then returns the sum of all executed elements.
As shown in Table 2, we have tested energy test with GPUs on two different platforms, the Oakley computer farm of the Ohio Supercomputer Center and Manchester HEP GPU cluster. Both of the platforms work well and show good consistency. The parallelism of GPUs makes it possible to defeat the quadratic increasing calculation time and dramatically reduce computation time for datasets with $10^4$ to $10^6$ events. Figure 2 illustrate how averaged computing time per event in the energy test changes with sample size for CPUs and GPUs with less than $\sim 10^6$ events. However, there is still a limit on GPU’s parallelism. For samples with size beyond $\sim 10^6$ events, computing time increasing with sample size on GPUs turns to quadratic again.

| Name                        | Chip            | Cores | Clock [GHz] | RAM [Gb] |
|-----------------------------|-----------------|-------|-------------|----------|
| Oakley                      | NVIDIA M2070    | 448   | 1.15        | 6        |
| Manchester HEP GPU cluster  | NVIDIA K40c     | 2880  | 0.745       | 12       |

4. GPUs on Grid

Even via CUDA and Thrust, GPU computing power is easy to be exploited, an allocation of GPU resource is still needed to be done manually. Another way to manage the computing resource is to use the Grid.

To test the performance of GPUs on Grid, four NVIDIA Tesla K40c GPUs were installed in Manchester Tier2 Grid, which is part of the UK NorthGrid and Worldwide LHC Computing Grid (WLCG).

To get the direct access to the Grid GPU resource, ‘old fashioned’ gLite commands and Job Description Language (JDL) are used to submit jobs to Grid. gLite is a middleware of Grid computing, it provides framework for users to access the distributed computing and storage resources on Grid. JDL is the language used to describe jobs and the requirements of jobs, and to retrieve the output when the jobs are finished. JDL is a high-level language and is written in a file (called JDL file) consisting of lines having the format:

\[
\text{attribute} = \text{expression};
\]

A simple example of JDL file to submit the energy test job is shown below.

```plaintext
[Type = "Job";
JobType = "Normal";
Executable = "Script";
StdOutput = "std.out";
StdError = "std.err";
InputSandbox = {"Script", "EnergyTest.cu", "Data.root"};
```
The attributes Type and JobType specify the type of job, Executable specifies the command script to be run by the job, the attributes StdOutput and StdError define the names of the files which containing the standard output and standard error of the executable, InputSandbox and OutputSandbox list the files that will be copied from local computer to Grid as input and outputs from Grid to local computer. It is possible to specify a base Uniform Resource Identifier (URI) on a GridFTP server where the files listed in OutputSandbox will be copied by using OutputSandboxDestURI attribute.

Jobs could be submitted to GPU platforms on Manchester Tier2 Grid via gLite commands:

```
$> glite-ce-job-submit --autm-delegation -r ce01.tier2.hep.manchester.ac.uk:8443/cream-pbs-gpu <myJDLFile>
```

this command will returns a URL. To check the status of an submitted job and to retrieve a finished job, gLite commands

```
$> glite-ce-job-status <myURL>
```

and

```
$> glite-ce-job-output <myURL>
```
could be used.

5. Application of energy test in \( D^0 \rightarrow \pi^-\pi^+\pi^0 \) decay

The \( D^0 \rightarrow \pi^-\pi^+\pi^0 \) decay is a singly Cabibbo-suppressed neutral \( D \) meson decay. The amplitudes of this decay channel proceed through \( c \rightarrow udd \), which lead to the final states \( \rho^0\pi^0, \rho^+\pi^- \), and \( \rho^-\pi^+ \). These final states are common for both \( D^0 \) and \( \bar{D}^0 \). Time-integrated \( CP \) asymmetries in \( D^0 \) decays can have three components: direct \( CP \) violation in decays to specific states, indirect \( CP \) violation in \( D^0-\bar{D}^0 \) mixing, and indirect \( CP \) violation in interference of decays with and without mixing. Within SM, \( CP \) violation expected in \( D \) decay is small, however contributions from particles that are not described in the SM and participate in the loops of the penguin amplitude can enhance the \( CP \) violation effects expected within the SM [9]. Therefore \( CP \) violation in \( D^0 \rightarrow \pi^-\pi^+\pi^0 \) decay provides sensitivity to physics beyond SM.

A set of \( D^0 \rightarrow \pi^-\pi^+\pi^0 \) decay sample was collected during the LHCb 2012 run at a centre-of-mass energy of 8TeV, with the integrated luminosity \( L = 2fb^{-1} \). 663 × 10^3 candidates were selected with a purity of 85%.

A study of sensitivity with different types and amount of asymmetries is required in the interpretation of the energy test output. This sensitivity study is done with a simplified simulation of data generated according to the model described in Ref. [10] with the generator package Laura++ [11]. With 100 permutations, the studies of different \( CP \) asymmetry scenarios and their sensitivities are reported in Table 3.

With sensitivity known, the energy test is applied to the LHCb \( D^0 \rightarrow \pi^-\pi^+\pi^0 \) decay data. By counting the fraction of permutation \( T \) values above the nominal \( T \) value from data with 1000 permutations, a \( p \)-value of \((2.6 \pm 0.5) \times 10^{-2}\) for default \( \sigma (0.3GeV^2/c^4) \) is obtained. this results is consistent with the no \( CP \) violation hypothesis. The nominal \( T \) value for data and permutation \( T \) values for no \( CP \) asymmetry hypothesis, and the visualisation of significant regions on Dalitz plot are shown in Fig. 3.
Table 3. Overview of sensitivities to various CP violation scenarios. $\Delta A$ and $\Delta \phi$ denote, respectively, change in amplitude and phase of the resonance $R$.

| $R (\Delta A, \Delta \phi)$ | $p$-value (fit)          |
|-----------------------------|--------------------------|
| $\rho^0 (4\%, 0^\circ)$    | $3.3^{+1.1}_{-1.5} \times 10^{-4}$ |
| $\rho^0 (0\%, 3^\circ)$   | $1.5^{+1.4}_{-1.2} \times 10^{-3}$ |
| $\rho^+ (2\%, 0^\circ)$   | $5.0^{+8.8}_{-3.8} \times 10^{-6}$ |
| $\rho^+ (0\%, 1^\circ)$   | $6.3^{+3.3}_{-3.8} \times 10^{-4}$ |
| $\rho^- (2\%, 0^\circ)$   | $2.0^{+1.3}_{-0.9} \times 10^{-3}$ |
| $\rho^- (0\%, 1.5^\circ)$ | $8.9^{+22}_{-6.7} \times 10^{-7}$ |

Figure 3. Result from applying energy test to LHCb $D^0 \rightarrow \pi^- \pi^+ \pi^0$ decay data. (a) Permutation (no CP asymmetry) $T$ value distribution overlaid with a GEV fit function and the measured $T$ value shown as a vertical line. (b) Visualisation of local asymmetry significances. The positive (negative) asymmetry significance is set for the $D^0$ candidates which have positive (negative) contribution to the measured $T$ value. Figures reproduced from Ref. [7].

6. Conclusions

A novel unbinned model independent technique, which is known as the energy test, is implemented in the CP violation searches. With the help of the parallelism of GPUs, this approach is feasible for the first time. The energy test has extended the range of applicable methods for unbinned analysis of CP violation in multi-body modes, and the first physics results with the energy test is produced. In future work, this approach can be extended to higher body decays.

7. Acknowledgements

This work was supported in part by an allocation of computing time from the Ohio Supercomputer Center, Manchester HEP GPU cluster, and Manchester Tier2 Grid. We would like to express our gratitude to Mike Sokoloff and Sabah Salih for their help with GPU accounts in Ohio and Manchester.

References
[1] Artuso M, Meadows B and Petrov A A, 2008 Ann. Rev. Nucl. Part. Sci. 58 249-291.
[2] Antonelli M, et al., 2010 Phys. Rept. bf 494 197-414.
[3] Dalitz R H, 1953 Phil. Mag. 44 1068.
[4] Bediaga I, Bigi I I, Gomes A, Guerrer G, Miranda J and dos Reis A C, 2009 Phys. Rev. D 80 096006.
[5] Aslan B and Zech G, 2004 Nucl. Instrum. Meth. A 537 626.
[6] Williams M, 2011 Phys. Rev. D 84 054015.
[7] LHCb Collaboration, Aaij R, et al., 2015 Phy. Lett. B 740 158-167.
[8] NVIDIA Corporation, 2015, NVIDIA CUDA C programming guide, Version 7
[9] Grossman Y, Kagan A L and Nir Y, 2007 Phys. Rev. D 75 036008.
[10] BaBar Collaboration, Aubert B, et al., 2008 Phys. Rev. D 78 051102.
[11] Latham T, et al., Laura++ Dalitz plot fitting package, http://laura.hepforge.org/.