CATEGORY: ASTRONOMY & ASTROPHYSICS - AA07

POSTER P5327

CONTACT NAME Michele Mastropietro: michele.mastropietro@oa-roma.inaf.it

GPU TECHNOLOGY CONFERENCE

Maximum Likelihood Estimation on GPUs: Leveraging Dynamic Parallelism

M. Mastropietro[1], D. Bastieri[2,3], A. Pigato[2], A. Madonna[1,2],
S. Amerio[3], D. Lucchesi[3], L.A. Antonelli[1] & G. Lamanna[4]

1. *Rome Observatory, INAF, Rome, Italy*
2. *CUDA Research Center, University of Padova, Italy*
3. *Dept. Physics and Astronomy, Univ. Padova and INFN, Padova, Italy*
4. *LAPP, Laboratoire d'Annecy-le-Vieux de physique des particules, Annecy, France*

Abstract

The estimation of the Maximum Likelihood (MLE) is the most robust algorithm used in gamma-ray astronomy but, particularly if used in conjunction with *unbinned* analysis, uses a huge amount of computing resources. Typically, the estimation of the maximum is left to a single-thread minimizer, like MINUIT, running on a CPU while providing a call-back function that may estimate the likelihood on the GPU. We propose an alternative to the MINUIT package, that leverages Levenberg-Marquardt algorithm and Dynamic Parallelism and runs entirely on GPUs.

Maximum Likelihood Approach

The Maximum Likelihood Approach (MLA) in High-Energy Astrophysics allows to estimate the model that most *likely* produced the data that were collected. The use of a MLA in scientific analysis dates back to the very first definition of *likelihood* by Fisher (1925), as he originally suggested that it could be used for parameter estimation. Neyman & Pearson (1928) then devised the likelihood ratio test, to compare the null hypothesis against an alternative one, but only when Wilks (1938) established an analytical expression, asymptotically exact, for this ratio, the so-called Wilks' theorem, the MLA started to be widely employed in science. A comprehensive treatment of the likelihood and the MLA may be found, among other, in Edwards (1972).

The importance of Wilks' theorem may be appreciated when analyzing the data collected by *Fermi* LAT (Atwood 2009) over 2 years (the 2FGL catalog: Nolan 2012). As can be seen in the picture below, Fig. 1, photons cluster around *candidate* sources, whose likelihood can be established via Wilks' theorem. Given the paucity of high-energy photons compared, for instance, to optical ones, data follow Poisson statistics, whose behavior in connection with the likelihood was originally described by Cash (1979). The extension to the MLA was first described by Mattox (1996), whose approach we follow here. The likelihood has also been used to reconstruct the parameters of the events collected by HESS[1] as described by de Naurois (2009). As a test case for our implementation, we will use the MLA used by the *Fermi* LAT Collaboration[2].

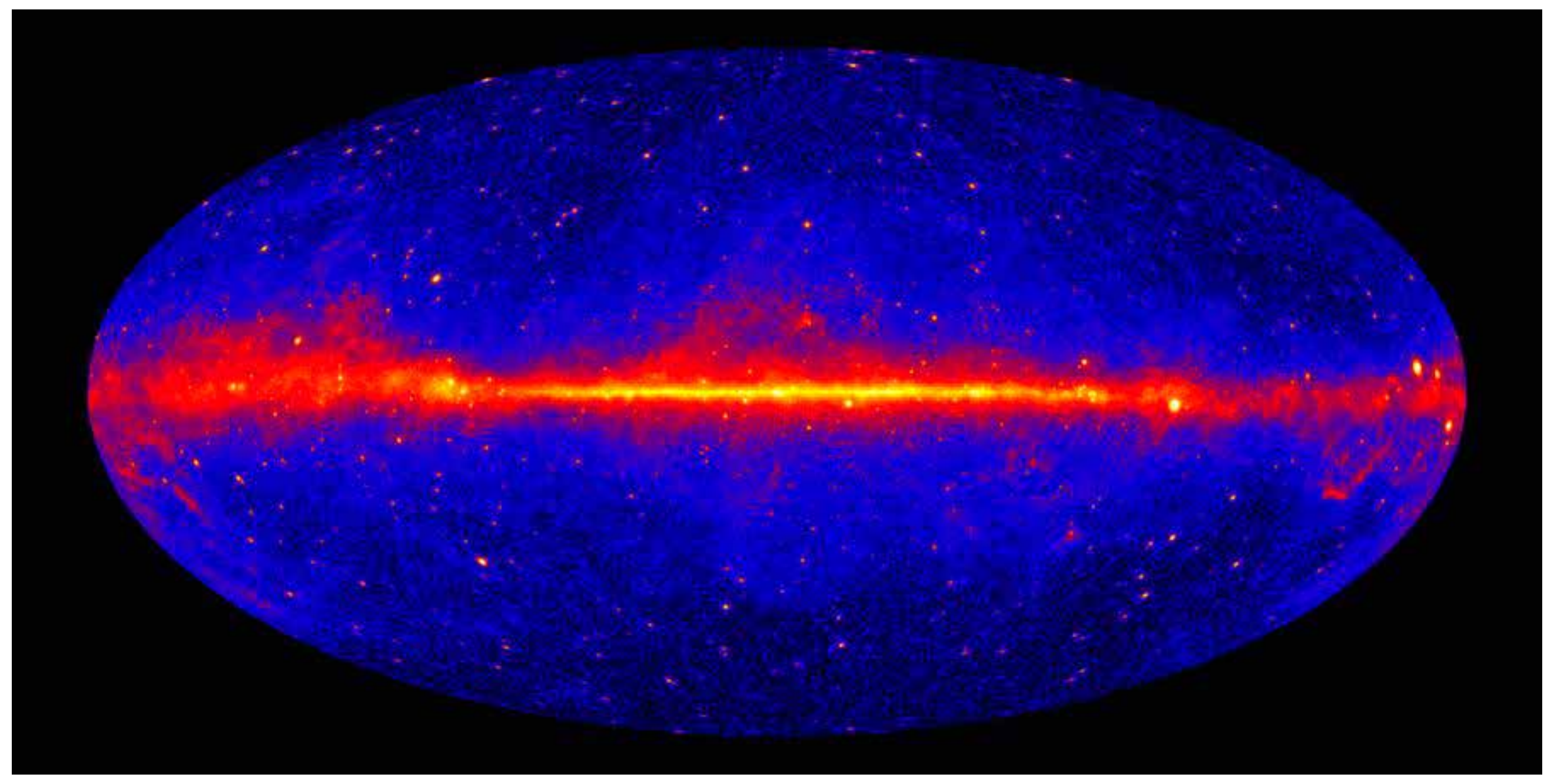

Figure 1. A plot in galactic coordinates of data collected by *Fermi* LAT in two years. Photons are denser going from blue to red to yellow to white.

[1] Info about HESS may be found at http://www.mpi-hd.mpg.de/hfm/HESS

[2] See http://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone

Implementation

The whole analysis chain of *Fermi* LAT develops mainly along the lines shown in the picture below (Fig. 2). The most time-consuming items, see Fig. 3 (left), are the evaluation of the *livetime-cube* and of the *likelihood*. Whereas the livetime cube had already been ported under GPU a while ago[3], the likelihood evaluation was ported to GPU by A. Pigato as a task in his MSc thesis (2013).

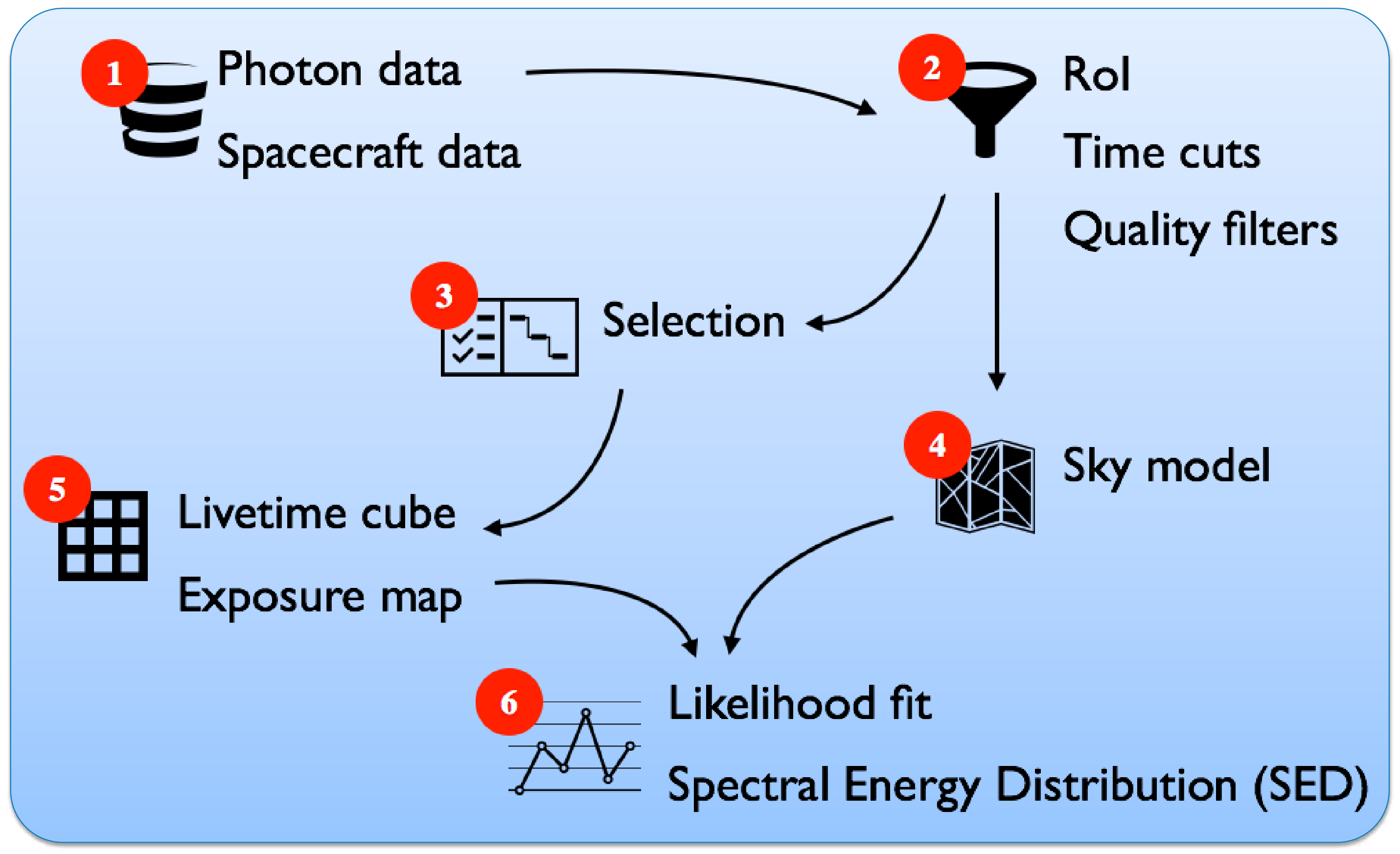

Figure 2. The analysis chain of the *Fermi* LAT Collaboration. The most time-consuming items are the *livetime cube* calculation (item #5) and the evaluation of the likelihood, invoked by the external minimizer to evaluate the model parameters that best fit the data and to compute the *TS* of the source (#6).

Levenberg-Marquardt algorithm for maximum-likelihood evaluation

The goal is to find a set of model parameters $\boldsymbol{x} = \{\alpha_k\}$ for which the function $L(\boldsymbol{x})$ is minimized:

$$L(\boldsymbol{x}) = -\sum_{i=1}^{n_{ph}} \log f_i(\boldsymbol{x}) = -\sum_{i=1}^{n_{ph}} \log \sum_{j=1}^{n_s} J_j(E_i; \{\alpha_k\})$$

$J_j(E_i; \{\alpha_k\})$ is the flux densities corresponding to each source j, emitting a photon of energy E_i. To solve the problem we adopted the Levenberg-Marquardt algorithm (LMA) which can be thought as a combination of gradient descent and Gauss-Newton minimization methods.

At each iteration it is necessary to solve a linear system to evaluate the step $\boldsymbol{h}$ needed to converge towards the local minimum following the iteration scheme: $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \boldsymbol{h}$. The system

$$\boldsymbol{A}\,\boldsymbol{h} = -\boldsymbol{g}$$

is solved via Cholesky decomposition which is the best method to solve systems with symmetric and positive definite matrices of coefficients.
Matrix $\boldsymbol{A}$ is defined as:

$$\boldsymbol{A} = \boldsymbol{H} + \lambda \cdot \operatorname{diag}(\boldsymbol{H}), \qquad \text{where } H_{p,q} = -\sum_{i=1}^{n_{ph}} \frac{1}{f_i^2} \frac{\partial f_i}{\partial x_p} \frac{\partial f_i}{\partial x_q} \approx \nabla^2 L(\boldsymbol{x})$$

The components of the gradient vector $\boldsymbol{g}(\boldsymbol{x}) = \nabla L(\boldsymbol{x})$ are:

$$g_k = -\sum_{i=1}^{n_{ph}} \frac{1}{f_i(\boldsymbol{x})} \frac{\partial f_i(\boldsymbol{x})}{\partial x_k}$$

[3] http://www.nvidia.com/content/cuda/spotlights/gpu-accelerated-astronomy.html

The parameter λ is updated (usually increased or decreased by an order of magnitude) at each iteration, depending on how well the function L can be approximated by a linearized model. Large values of λ bring the algorithm closer to the gradient descent method whereas small values of λ bring it closer to the Gauss-Newton one.

Kernel implementations

We assumed the flux densities to be in the form $J_j(E_i; k, b) = kE^{-b}$ (power law). We implemented two kernels to compute the value $L(\boldsymbol{x})$ and the vector gradient $\boldsymbol{g}(\boldsymbol{x})$. In both kernels we used the strategy to parallelize on bins of photons. The inner summation on the sources (index j) is done with a for-loop inside the kernel, whereas the summation on the photons (index i) is done on the GPU after each thread corresponding to the index i returns the value of $f_i(\boldsymbol{x})$. The kernel implementing LMA launches at each iteration the value and the gradient kernels, going forward in the process. This is possible by leveraging dynamic parallelism offered by devices of compute capability 3.5 or higher.

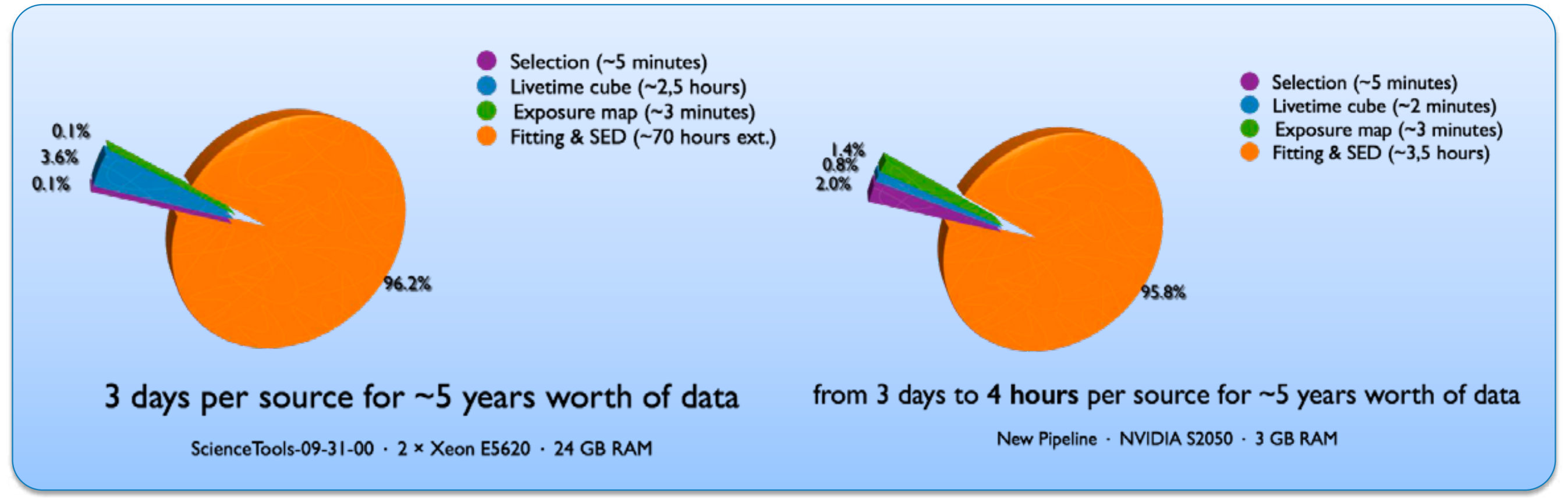

Figure 3. Comparison of execution times for the entire *Fermi* LAT analysis chain. CPU only (on the left) compared to the case when the livetime and the actual likelihood data sums are computed on the GPU (on the right).

Conclusions

We envisaged two different strategies for implementing the MLA to search for the model that best fits collected data: in the first one, we exploit the versatility of minimizing package like MINUIT developed at CERN. MINUIT needs two callback functions from the user, the actual function to be minimized and its gradient with respect to the parameters array. Typically, the likelihood, actually its logarithm, consists of a huge sum over the events, the pixels or some event templates, which could be run in a much faster way on a GPU. In the case of *Fermi* LAT likelihood computation, this has lead to an average execution acceleration of ~20×. The second strategy is to devise our own minimizer that executes directly on the GPU. It is not as versatile as MINUIT, but adapting the Levenberg-Marquardt algorithm to the computation of the maximum likelihood has been proven feasible. The net gain should be that the information about the parameters are not sent back and forth from the CPU to the GPU, giving us some more room to gain further acceleration. Preliminary tests applied to the case where spectral model were fixed to plain power laws showed a further gain of 2-3×. Further studies will be needed, above all in order to check whether the versatility we loose is worth the acceleration we obtain.

REFERENCES

Atwood W.B. et al., "The Large Area Telescope on the Fermi Gamma-Ray Space Telescope Mission", *The Astrophysical Journal* 697, 1071–1102 (2009)
Cash W., "Parameter Estimation in Astronomy through Application of the Likelihood Ratio", *The Astrophysical Journal* 228, 939–947 (1979)
de Naurois M. & Rolland L., "A high performance likelihood reconstruction of gamma-rays for IACTs", *Astroparticle Physics* 32, 231–252 (2009)
Edwards A.W.F., "Likelihood", Cambridge Univ. Press 1972.
Fisher R.A., "Statistical Methods for Research Workers", Oliver and Boyd, Edinburgh 1925.
James F., "MINUIT Tutorial", Procs. of "*1972 CERN Computing and Data Processing School*", Pertisau, Austria, rev. 2004.
Mattox J.R. et al., "The Likelihood Analysis of EGRET Data", *The Astrophysical Journal* 461, 396–407 (1996)
Neyman J. & Pearson E.S., "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I", *Biometrika* 20A, 175–240 (1928)
Nolan P.L. et al., "Fermi Large Area Telescope Second Source Catalog", *Astrophysical Journal Supplement* 199, 31 (2012)
Pigato A., "The Imprint of New Physics in AGN Cutoffs", MSc Physics (supervisor: D. Bastieri), Padova 2013.
Wilks S.S., "The large-sample distribution of the likelihood ratio for testing composite hypotheses", *Annals of Math. Statistics* 9, 60–62 (1938)