A fast - Monte Carlo toolkit on GPU for treatment plan dose recalculation in proton therapy

M Senzacqua[1], A Schiavi[1], V Patera[1], S Pioli[2], G Battistoni[3], M Ciocca[4], A Mairani[4], G Magro[4] and S Molinelli[4].

[1] Sapienza - Università di Roma, Department of Basic and Applied Sciences for Engineering (SBAI), Via A. Scarpa 14, 00161 Roma, Italy
[2] INFN - Laboratori Nazionali di Frascati, Via Enrico Fermi 40, 00044 Frascati (Roma), Italy
[3] INFN - Sezione di Milano, Via Celoria, 16, 20133 Milano, Italy
[4] Fondazione CNAO, Strada Campeggi, 53 - 27100 Pavia, Italy

E-mail: martina.senzacqua@uniroma1.it

Abstract. In the context of the particle therapy a crucial role is played by Treatment Planning Systems (TPSs), tools aimed to compute and optimize the tratment plan. Nowadays one of the major issues related to the TPS in particle therapy is the large CPU time needed. We developed a software toolkit (FRED) for reducing dose recalculation time by exploiting Graphics Processing Units (GPU) hardware. Thanks to their high parallelization capability, GPUs significantly reduce the computation time, up to factor 100 respect to a standard CPU running software. The transport of proton beams in the patient is accurately described through Monte Carlo methods. Physical processes reproduced are: Multiple Coulomb Scattering, energy straggling and nuclear interactions of protons with the main nuclei composing the biological tissues. FRED toolkit does not rely on the water equivalent translation of tissues, but exploits the Computed Tomography anatomical information by reconstructing and simulating the atomic composition of each crossed tissue. FRED can be used as an efficient tool for dose recalculation, on the day of the treatment. In fact it can provide in about one minute on standard hardware the dose map obtained combining the treatment plan, earlier computed by the TPS, and the current patient anatomic arrangement.

1. Introduction

Charged Particle Therapy (CPT) is a radiation tumor treatment technique that uses protons or light ions. It aims to deliver a high precision treatment of solid tumors by exploiting the characteristic shape of the Bragg curve of charged hadrons: the dose deposition as a function of depth of traversed matter exhibits a sharp peak (the Bragg peak) at the end of the particle range. Thanks to its dose release profile accurate and efficient irradiation of the tumor can be obtained reducing the dose to the surrounding healthy tissue, thus achieving less post-irradiation complication probability with respect to the standard X-ray radiotherapy. According to recent statistics [1], at the end of 2014 more than 137,000 patients have been now treated worldwide with charged hadrons (about 86% with protons). The number of clinical centers dedicated to CPT currently in operation is 63 but by the end of 2018, 17 new proton therapy facilities will start to treat patients. The high spatial selectivity of CPT asks for stringent requirements on the accuracy that has to be achieved. The uncertainty on the position of the dose release in treatments can be due to different factors, such as uncertainties in particle range, the calibration

of the Computed Tomography (CT) images and also patient mis-positioning and organ motion during the treatment itself. Furthermore, it has to be considered that morphological changes may occurr between the CT and each of the several irradiation sessions of a CPT treatment, operated in different days. All the effects mentioned can contribute to a total uncertainty of the order of few millimeters. In order to preserve the intrinsic advantages of hadrontherapy, fast and accurate dose calculation tools are necessary to check, verify, and eventually correct, initial treatment planning. Softwares dedicated to the dose optimization and calculation are called Treatment Planning System (TPS) and different calculation algorithms are used and still in progress of study to improve performances both in terms of accuracy and time computing. The two main approaches of computation methods are Monte Carlo (MC) techinque and analythical pencil beam algorithm. The first method has high accuracy in dose calculation but the simulation time needed to obtain adequate statistics is still too long to permit use in clinical routine. The analythical TPS are quicker. The aim of the project FRED is to build an innovative TPS tool using Monte Carlo methods for simulation of proton energy deposition in tissues and exploiting the mosta advanced GPU devices. The implementation of Multiple Coulomb Scattering (MCS), energy fluctuation and nuclear interaction is described. In particular, we present the implementation of approximated algorithms that allow accurate and fast evaluation of theoretical models, whose exact resolution is a time consuming task. Results about linear dose profiles are shown in comparison with full MC code FLUKA [2]. Finally, performances in terms of time computing using different hardware architecture and software models are presented.

2. Transport methods

The transport of protons in tissues is simulated by a condensed-history Monte Carlo method. The patient region of interest is imaged by a Computed Tomography (CT) diagnostic test. The volume is divided into small parallelepipeds- or *voxels*- considered uniform in atomic composition and density. In crossing materials, each primary proton undergoes elastic and inelastic collisions with nuclei and electrons, as well as nuclear interactions that make it lose energy and change direction.

The particle is tracked from the source to the end of its range. The track is obtained as the sum of steps whose length is limited by different criteria:

- geometric limitator: each step can not cross the border of two adjacent voxels so the maximum length walkable is the distance from the particle position to the voxel border;
- energy limitator: the energy loss for each step can not exceed the 1% of the particle energy, in order to allow the approximation of constant mean energy loss along a single step;
- discrete event limitator: if along the step a nuclear inelastic interaction occurs, the end of the step is set to the interaction point.

The mean energy loss is computed following the method of *Fippel and Soukup* [3]. The ratio between the stopping power in the tissue S and in water S_w is expressed as a function of the proton kinetic energy T_p and the material mass density ρ. S_w is obtained from stopping power tables PSTAR[4] and it depends on T_p and the step length δz.

Energy straggling due to statistical fluctuations of number of collisons suffered by protons and of the energy transferred in each collision is reproduced using two matching approximations. In thick absorber regime the number of ionization events is high, and the distribution of energy loss can be considered gaussian. The standard deviation is computed through the following formula[5]:

$$\Omega^2 = 0.0855\rho\delta z\frac{T_{\max}}{\beta^2}\left(1-\frac{1}{2}\beta^2\right)\text{MeV} \tag{1}$$

Where $T_{\max} = \frac{2m_e\beta^2\gamma^2}{1+2\gamma m_e/m_p+(m_e/m_p)^2}$ is the maximum energy transferable in a collision to a single electron, where m_e and m_p are the mass of the electron and of the primary respectively. In case of very thin absorber, the distribution the distribution of energy loss is not gaussian anymore. The theory in thin absorber regime has been developed by Landau first[6] and then Vavilov[7]. The computation of analytic Vavilov solution is extremly time-consuming, for this reason we implemented our original algorithm to approximate the distribution. The method consists of fitting the Landau function with a logarithmic normal distribution L_N [1]. The parameters σ, θ and m are expressed as a function of the energy of the particle and the material properties ρ and δz. The logarithmic-normal approximation for thin absorbers shows a remarkable efficiency, in fact the accuracy in energy loss distribution reproduction is good and the sampling time is comparable to the gaussian thick aborbers.
For what concerns the Multiple Coulomb Scattering (MCS) the assumption of small angle approximation is done. This means that the angle θ between the direction of the proton before and after the step is considered small and can be computed through the quadratic sum of the projected angles θ_x and θ_y ($\theta = \sqrt{\theta_x^2 + \theta_y^2}$). The distribution of the projected angles is fitted with three different functions that can be set by the user according to his needs of accuracy and time costs. The three functions are, in increasing order of both accuracy and calculation time:

(i) Single gaussian approximation. The standard deviation of the distribution is computed using the Highland forumla[8]. This model well reproduces the core of the distribution, but it misses the tails due to large angle scattering;

(ii) Double gaussian approximation. A second gaussian is superimposed to the core one to cover partially the tails of the distribution. The standard deviation of the second broader gaussian is obtained by fitting histograms generated with FLUKA, and then indexed as a function of the particle energy and of the step length.

(iii) Gauss-Rutherford-like approximation. The name of the third model is due to the expression of the distribution superimposed to the core gaussian that is similar to the Rutherford hyperbola. The function is:

$$f_{GR}(\theta_{x,y}) = \frac{1-w}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{\theta_{x,y}}{\sigma_1}\right)^2\right] + w\frac{a}{(\theta_{x,y}+b)^c} \quad , \quad c \sim 2.0$$

where, $0 \leq w \leq 1$ is the weight of the Rutherford-like component, σ_1 is the Highland standard deviation and a, b and c are the parameters of the tail function. This model allows to reach an accuracy of the order of 0.001%, within 4 standard deviations of the scattering angle θ distribution.

The nuclear interactions are simulated using the cross sections provided by ICRU Report 63[9]. The atomic composition is recreated by using the conversion table provided by *Schneider et al.*[10] that indicates the atomic composition of biological tissues as a function of the Hounsfield Number. The nuclear elastic and inelastic cross sections are computed for each step. When a nuclear reaction occurs, the type, the multiplicity and the energy of secondary particles produced are extracted. Secondaries are treated differently according to their type:

- secondary protons and deuterons are tracked;
- neutrons are neglected and their energy is considered lost outside the region of interest;
- heavier particles such as alpha and light ions are stopped immediatly and their energy deposited in the voxel where the production occurs;

[1] $L_n(\lambda_v) = \frac{1}{(\lambda_v-\theta)\sqrt{2\pi}\sigma} \exp -\frac{(\ln(\frac{\lambda_v-\theta}{m}))}{2\sigma^2}$, with λ_v the Vavilov variable (see [7]) and σ, θ and m are the parameters of the function

3. Results

In figure 1 we show histograms of linear dose profiles for a pencil beam of 150 MeV protons in water, with an initial FWHM set to 0. FRED results (red line) are compared with Full Monte-Carlo FLUKA simulations. Particular attention has to be paid to the lateral dose profiles on the right. They refer to the dose distribution at a depth 90% of the BP position, that means around 14 cm in the water phantom. At that point, the beam lateral distribution is strongly changed with respect to the initial configuration and conditioned by all the process suffered by the particles. As it can be observed, even at this critical position there is a good agreement between FLUKA and FRED histograms. Also the tails are well reproduced, up to 3 orders of magnitude.

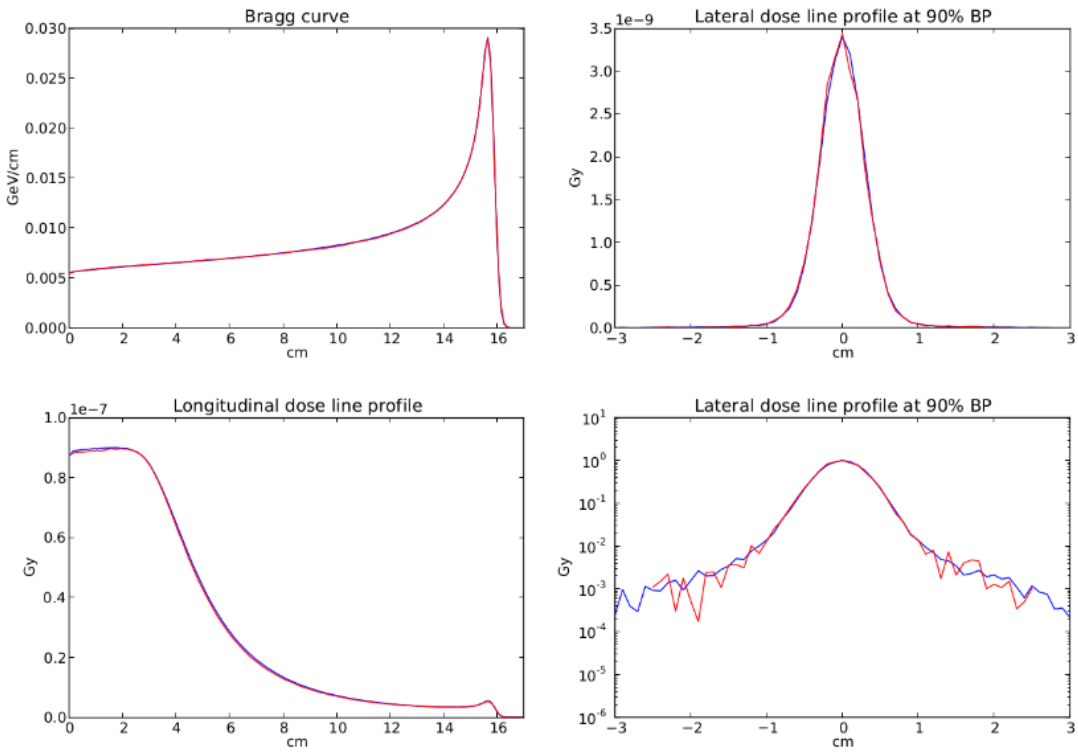

Figure 1. Linear dose profiles of a 150 MeV point section (FWHM=0.0) proton beam in water. In blue full-Monte Carlo data, in red FRED histograms. Upper left: longitudinal dose profile, also called *Bragg Peak*. On the bottom left: dose release inside the central voxels in the longitudinal direction. The decreasing trend of the longitudinal dose line profile testifies the emptying of the central voxels, due to scattering events that spread the beam. At the end of the path is visible a small peak representing the Bragg Peak. On the right: lateral dose profile about 1.5 cm before the Bragg Peak, both in linear (upper) and logarithmic (bottom) scale.

4. Time performances

In table 4 time performances for different architectures are displayed and compared with the time consumption of a simulation performed with the full Monte-Carlo tool FLUKA. As it can be observed, in the same running modality, FRED is 20 times faster than FLUKA on the same CPU hardware. This is due to the simplification of physics models implemented, and that's where the denomination *fast Monte-Carlo* comes from. The non-linear scaling of time with GPU number is due to the different types of GPU employed. As shown, running on GPU even

		Threads	primary/s	$\mu s/primary$
CPU	FLUKA	1	0.75K	1340
	FRED	1	15 K	68
	FRED	16	48 K	21
	FRED	32	80 K	12.5
GPU	FRED	1 GPU[1]	800 K	1.35
	FRED	2 GPU[2]	3500 K	0.3
	FRED	4 GPU[3]	20000 K	0.05

Table 1. Computing times for different hardware architectures.
[1] LAPTOP: MacBookPro(AMD Radeon R9 M370X).
[2] DESKTOP: Mac Pro (AMD FirePro D300).
[3] LINUX WorkStation with 4 NVIDIA GTX 980 GPUs

with a standard laptop the gain in terms of time is about three orders of magnitude with respect the full Monte Carlo and about 50 with respect to FRED single CPU modality.

5. Conclusions

The fast-Monte Carlo FRED showed a good agreement with full-Monte Carlo simulations in the computation of the dose distribution originated by proton beams passing through biological tissues. At the same time, the employement of GPU devices allows to significantly reduce the computation time. Thanks to its performances in terms of accuracy and calculation time, FRED represents an efficient tool that can be used in clinical routine for the recalculation of the treatment plan needed during the validation protocol. Furthermore, the short times implied in computation could allow the use of FRED in online dose monitoring techniques. By detecting the secondary products, in facts, the distal point of a treatment beam could be inferred. By taking into account the interactions undergone by the proton in the exit path using the information from a charged particle tracker detector and patient anatomic information available from a Computed Tomography scan, FRED can significantly improve the accuracy of the online dose monitoring device.

References

[1] http://www.ptcog.ch
[2] Böhlen T T, Cerutti F, Chin M P W, Fass A, Ferrari A, Ortega P G, Mairani A, Sala P R, Smirnov G and Vlachoudis V 2014 *Nuclear Data Sheets* **120** 211-14
[3] Fippel M and Soukup M 2004 *Med. Phys.* **31**(8) 2263-73
[4] Berger M J 1993 MD *Techn. Rep. NBSIR 4999 Nat. Inst. of Stand. and Techn.*
[5] Fippel M 2006 *New Techn. in Rad. Onc.* 197-206
[6] Landau L 1944 *J. Phys. (USRR)* **8** 201
[7] Vavilov P V 1957 *Sov. Phys. JETP* **5** 749
[8] Highland V L 1075 *Nucl. Instr. and Meth.* **129** 497-99
[9] 2000 *ICRU Report-63. Nuclear data for neutron and proton radiotherapy and radiation protection.*
[10] Schneider U, Pedroni E and Lomax A 1996 *Phys. Med. Biol* **41** 111-124