EPJ Web of Conferences **153**, 04010 (2017) DOI: 10.1051/epjconf/201715304010
ICRS-13 & RPSD-2016

Modeling of Radiotherapy Linac Source Terms Using ARCHER Monte Carlo Code: Performance Comparison for GPU and MIC Parallel Computing Devices

Hui Lin[1], Tianyu Liu[1], Lin Su[2], Bryan Bednarz[3], Peter Caracappa[1], X. George Xu[1*]

[1]*Nuclear Engineering Program, Rensselaer Polytchnic Institute, Troy, NY, USA, 12180*
[2]*Radiation Oncology, John Hopkins University, Baltimore, MD , USA, 21218*
[3]*Department of Medical Physics, University of Wisconsin, Madision, WI, 53705*

Abstract. Monte Carlo (MC) simulation is well recognized as the most accurate method for radiation dose calculations. For radiotherapy applications, accurate modelling of the source term, i.e. the clinical linear accelerator is critical to the simulation. The purpose of this paper is to perform source modelling and examine the accuracy and performance of the models on Intel Many Integrated Core coprocessors (aka Xeon Phi) and Nvidia GPU using ARCHER and explore the potential optimization methods. Phase Space-based source modelling for has been implemented. Good agreements were found in a tomotherapy prostate patient case and a TrueBeam breast case. From the aspect of performance, the whole simulation for prostate plan and breast plan cost about 173s and 73s with 1% statistical error.

1 Introduction

As a modern radiotherapy technique, Intensity modulated radiation therapy (IMRT) is designed to deliver superior dose conformity and uniformity compared to traditional three-dimensional conformal therapy [1]. However, few treatment planning systems (TPS) employ sophisticated dose calculation algorithms such as Monte Carlo (MC) method. Approximate algorithms are widely used in many TPSs, such as convolution/superposition method, are known to be fast but only able to generate approximate dose results in cases when the treatment site involves complex and heterogeneous tissue structures [2]. In comparison, Monte Carlo method for dose calculations in radiation therapy is well recognized for its accuracy. The issue of lengthy calculation times used to be the main bottleneck of MC method being widely applied in the medical physics community as a clinically feasible approach.

Recently, with the development of MC codes optimized for radiotherapy calculations as well as the availability of much faster and affordable computer accelerator technologies, such as the general-purpose Graphical Processing Units (GPU) by NVIDIA and AMD and the Xeon Phi coprocessors by Intel, decent acceleration factors have been observed compared to conventional CPU-based calculations. These significant advances have led to the clinical use of MC algorithms at some treatment centers and the promised availability of MC photon/electron planning modules among several commercial treatment planning vendors. With this in mind, we developed the ARCHER Monte Carlo code that takes advantage of a new class of parallel computing devices — GPU and Xeon Phi coprocessors [3-5]. Our motivation for developing the ARCHER code is to perform a thorough performance evaluation of modern computing devices in the context of Monte Carlo simulations, while providing the community a practical, efficient, and reliable dosimetry tool that can utilize all those devices. Previously we have showed that, by utilizing the GPU, ARCHER was able to finish a treatment planning dose calculations for Tomotherapy in about 70 seconds using single-precision floating point while the same plan took the GEANT4 code 10 CPU hours [6]. This paper demonstrates the latest effort to develop and apply the source modelling of Radiotherapy Linac photon source in the ARCHER GPU and Xeon Phi-based MC dose engine.

2 Methods and Materials

2.1 Accelerator Treatment Head Modelling

Generally, source modelling in the context of MC treatment planning contains multiple beam components. Energy, location and direction distributions of each source particle are explicitly expressed for each component of the accelerator treatment head. Over the years, some routes have been employed including direct use of phase space data from the MC treatment head simulations, measurement-based models and virtual, multi-source models from the treatment head simulation. Among them, phase space sources have the potential to provide the most accurate physical descriptions of the source, and they

* Corresponding author: xug2@rpi.edu

have been accepted as inputs by many well-known MC dose calculation tools such as EGSnrc and GEANT4 [7,8].

The simulation of a radiotherapy Linac treatment head could be divided into several methods. The first and most straightforward method is to perform the simulation of the patient-dependent phase-space sources. In ARCHER, two commonly used phase-space file formats defined in EGS code and the International Atomic Energy Agency (IAEA) [9] are supported. In this method, the phase space files were generated with the effects of secondary collimators. When the patient-dependent phase space files are available, this method is efficient. However, in most cases, the generation of patient-dependent phase spaces depends on other Monte Carlo codes like BEAMnrc, which is time-consuming and clinically impractical. In the following section, we described a method utilizing patient-independent phase space files.

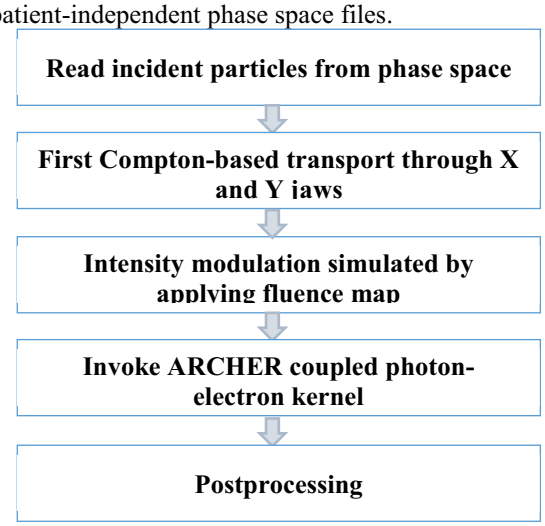

Fig. 1. Workflow of ARCHER Linac source modelling.

Fig. 1 demonstrated the workflow of the photon Linac source modelling in ARCHER. The namely patient-independent phase space file is generated by performing the simulation of the patient-independent structures of treatment head and storing a phase-space file at a plane just below the fixed elements of the accelerator head. The advantage of this approach is that once this part of phase space files is scored, it could then be reused as often as necessary. Then ARCHER takes care of the modelling of beam collimation and intensity modulation. For the simulation of secondary jaws, First Compton-based approximate transport was implemented. The assumption was built on the fact that the principle photon interaction in tungsten from 0.5 to 5MeV is the Compton interaction. The location of the interaction as the photon traverses thickness t is randomly sampled over [0,t) assuming exponential attenuation. The interaction probability is evaluated using the ratio of the Compton and total attenuation coefficients as shown in Equation (1), where μ indicates the attenuation coefficient of photons and t indicates the jaws thickness as particle transverses through.

$$f(t) = Ce^{-\mu t}\mu dt \qquad (1)$$

As Equation (2) demonstrated, the particle weight was then modified based on scattered photons' energy and direction, while the emerged Compton electrons are not continued.

$$w_f = w_i(1 - e^{-\mu(E)t})\frac{\mu_c}{\mu}e^{-\mu(E')t'} \qquad (2)$$

Intensity modulation is achieved by supplying fluence map for each treatment field. Phase-space particles that pass through the jaw openings are projected to the MLC plane. A weighting factor determined by the particles position in the fluence map is then assigned to the particle. If the particle already carries a weighting factor in the patient-independent phase-space, the two factors are multiplied. The final factor is carried throughout the subsequent transport process and is used to adjust the dose deposition accordingly.

2.2. Helical Tomotherapy accelerator

Helical tomotherapy is a sophisticated intensity-modulated radiation therapy (IMRT) delivery modality developed at the University of Wisconsin-Madison and commercialized by Tomotherapy, (now owned and distributed by Accuray Inc). The Hi-Art Tomotherapy system utilizes a unique mechanical structure that resembles a helical CT scanner. Different from the traditional IMRT, the treatment head in a Tomotherapy system is mounted on a slip ring gantry and can rotate continuously around the isocenter. During the treatment, the head rotates continuously while the couch is translated concurrently, delivering the dose in a helical manner. In our model, all particles defined in the phase space files lie on a cylindrical surface formed by the gantry rotation and couch translation—such that these particles can be used in the next-stage MC simulations involving patient-specific parameters of the treatment plan. Contaminant electrons are not included in PSFs because a previous study had showed that the electron contamination from a helical tomotherapy system could be ignored without introducing measurable error. After benchmarking the code, one clinical prostate case is calculated.

2.3. Electron-Photon Coupled Transport Kernel

The main physics models implemented in ARCHER were similar to production MC code DPM. The original sequential DPM code was developed for fast dose calculations in radiotherapy treatment planning [11]. It aims for simulating coupled photon-electron transport with a set of approximations valid for the energy range considered in radiotherapy. Specifically, for photon transport, we considered photoelectric effect, Compton scattering, and pair production. Rayleigh scattering is disregarded in this study because it has negligible impact on dose distribution for photon energy range used for radiotherapy. The photon transport is handled by using the Woodcock tracking method which eliminates frequent

boundary checking for photon transport in heterogeneous geometries [12]. As for the electron transport, DPM implements a condensed history technique. Step-by-step simulation is used for inelastic collisions and bremsstrahlung emission involving energy losses above certain cutoffs. It also employs new transport mechanics and electron multiple scattering distribution functions to allow long transport steps. These mechanisms enable the electron cross a few heterogeneity boundaries in one step and hence increase the simulation efficiency. The continuous slowing down approximation is employed for energy losses below some preset energy thresholds. Positron transport is treated as electron and two photons are created at the end of the positron path to account for the annihilation process. The accuracy of DPM has been demonstrated to be within for both clinical photon and electron beams [13]. In ARCHER, this DPM-based physics kernel is coupled with other modules, such as the Constructive Solid Geometry (CSG), DICOM decoder and phase space file manager, to enable accurate, fast, and patient-specific dose calculations.

2.4. Code Design and micro-optimization

ARCHER uses "symmetric execution" model where the CPU and MIC work concurrently. The code is written in hybrid MPI-OpenMP. Three types of processes were created: root process, I/O process and compute process. The root process is launched on the host CPU and acts as an overall coordinator. Its job is to distribute the task and perform tally reduction. The I/O process, also on the host, is used to read the phase space files from the hard drive to the system memory and send them to the compute process. The compute process resides in the MIC coprocessor and performs transport simulation in parallel. For a heterogeneous system with 1 CPU and n MICs, we used 1 root process, and n I/O processes paired with n compute processes. On each MIC, the compute process spawned a total of 240 threads to fully saturate the hardware resource. To hide I/O and data transfer with computation, the double buffering method was implemented, which lets the I/O process to read the next phase space file and copy it to the compute process before the simulation using the current phase space file is completed.

To ensure the reliability of our MIC and GPU codes, we decided to use double-precision for all calculations. However, before the Pascal architecture, NVIDIA GPUs don't support FP64 natively. NVIDIA has offered some workarounds, but is very slow. To solve this problem, we implemented this fast warp-based floating point 64 atomic add algorithm. According to our test, our implementation is ~600x faster than Nvidia's original workarounds.

For CPU and MIC, localized vectorization and software-based data prefetch were implemented. Although MC is known to have many branch statements which is hard to vectorize on global scale, local vectorization for tight for-loops still exist. Thus for CPU, SSE4 is used to vectorize the instruction set, and IMCI is the instruction set for vectorization on the MIC.

The current MICs use legacy processors that are subject to stall when the data are being loaded from the DRAM, resulting in long latency. The data prefetch method is able to reduce this latency by loading data several loops ahead so that they are readily available in the cache when actually needed.

2.5. Phase Space Source Implementations

For the prostate plan, the size of the phase space file is 6.4 GB, which is too large to load into Xeon Phi's memory. Another concern is of efficiency. It takes at least tens of seconds to transfer a large amount of particle data from a hard drive to Xeon Phi memory. Although tens of seconds is relatively short for CPU-based MC simulations, it consumes a large portion of the total computational time for Xeon Phi-based MC dose calculations, posing a bottleneck to further improve efficiency. To circumvent the above problems, the entire PSF file is partitioned into smaller pieces based on their type and energy and then simulated in batches. Typically, the batch number is set to be 12, reducing the sub-PSF to a more manageable size of 530MB. A sorting of particles from the phase-space source has also been employed before launching them for dose calculations. Specifically, the code works as following: the CPU launches the particles sequentially and the CPU reads the first sub-PSF and copies the data to the Xeon Phi memory. The Xeon Phi kernel is then launched asynchronously to perform the MC calculations. The CPU reads in the next sub-PSF immediately after the first kernel is launched. In this way, the simulation and PSF reading are performed concurrently, thus enhancing the code execution efficiency.

2.6. Patient CT Phantom

The CT dataset is converted into mass density and material composition using the Hounsfield unit (HU)-to-density conversion curve. Using a commercial treatment planning system at UW–Madison, the conversion is calibrated in accordance with recommendations by Verhaegen and Devic [14]. In these simulations, four materials, each having densities specified by the HU, are used for the patient phantom: water, dry air, compact bone (defined by ICRU), and lung (defined by ICRP) (15). Although the code is capable of more materials, these four materials provide enough accuracy for the MC investigation. The prostate phantom consists of 260×256×108 voxels, with a voxel size of $0.1875\times0.1875\times0.3cm^3$.

3 Results

3.1. Commissioning Fields Validation

Fig. 2 illustrates the percentage depth dose (PDD) between simulated phase space source modelling and measurement in 60×60×40cm^3 water phantom for Varian TrueBeam 6MV beam at SSD 100 cm. The dose was normalized to different scale for better visualization. The voxel size of the phantom is 0.4×0.4×0.4cm^3 for field size of 4×4, 10×10, and 40×40cm^2. A good agreement was observed for all fields. The dose difference was better than 2% except in build-up region where the difference is up to 5.87%.

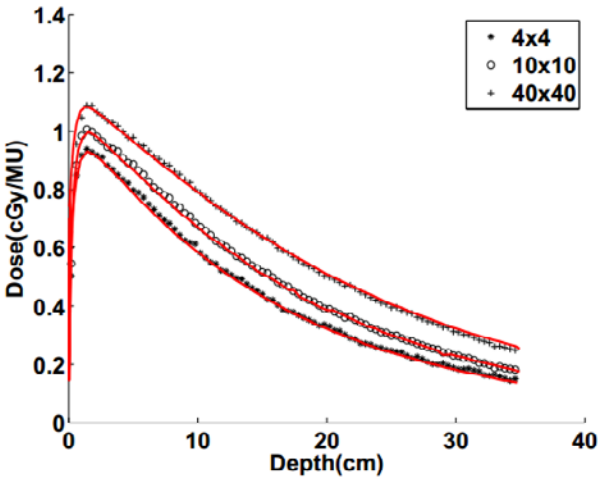

Fig. 2. Experimental (red solid lines) and ARCHER calculated depth dose curves. The Monte Carlo dose was averaged over 3×3 voxels.

Fig. 3 illustrates the comparison of cross profile at depth 2.5 cm for field sizes 4×4, 10×10, 40×40cm^2 between our simulation and measurement at SSD 100 cm. It is shown that the profiles at depth 2.5 cm matched well in all fields.

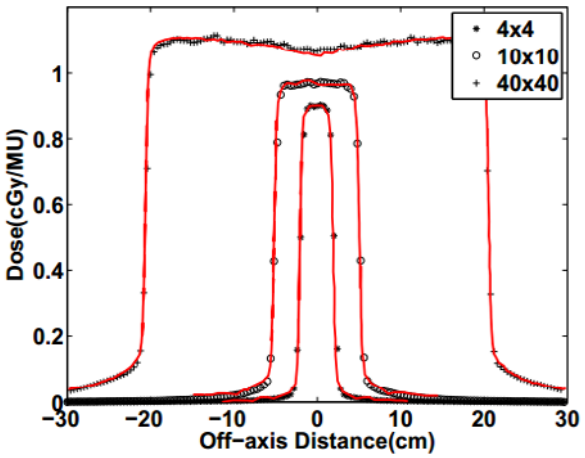

Fig. 3. Experimental (red solid lines) and ARCHER calculated dose profiles for 2.5 cm depth. The Monte Carlo dose was averaged over three voxels with a total length of 1.2 cm.

3.2. Dosimetric Results for Clinical Cases

One realistic Tomotherapy and TrueBeam plan were selected for this study. They are optimized in commercial TPS and then simulated with phase space based source modelling. The calculation results of the prostate tomotherapy case and the TrueBeam breast treatment using ARCHER were shown in Fig. 4 and 5. For the voxel doses in the PTV, relative statistical error (1σ) is kept to be about 1% considered in this study. The simulated dose distribution was shown in the first column. For comparison, the dose results from DPM are provided in the second column. It could be observed that these two sets of isodose lines overlapped with each other at most regions for both cases.

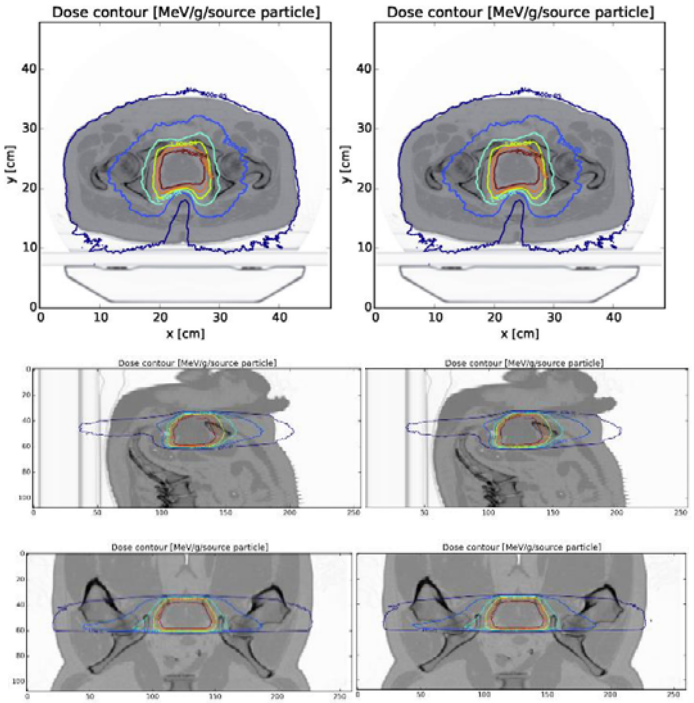

Fig. 4. Isodose distribution of a Tomotherapy plan between ARCHER simulated results (left) and DPM results (right). The isodose curves were shown in the first column, with transverse, coronal and sagittal views in three columns, respectively.

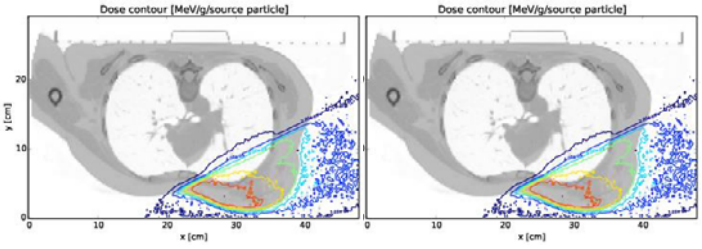

Fig. 5. Isodose distribution of a TrueBeam 6MV breast treatment plan between ARCHER simulated results (left) and DPM results (right).

3.3. Performance Studies

For fast MC simulations using Phase Space as source input, it is found that reading PSF from the hard drive costs a large fraction of code execution time. On the other hand, using the batch simulation explained before, file reading works concurrently with kernel execution,

thus avoiding explicit file read time. With this improvement, the preprocess time includes loading 600 million particles and CT phantom is found to be only about 3s for the prostate case. Table 1 summarizes the performance results for comparison among CPU, Xeon Phi and GPU. To be noted, all the results were calculated under double-precision, and no aggressive compiler options were used since that may trade the accuracy for speed. It could be seen that by employing Intel Xeon Phi coprocessors and GPUs, comparing with Intel x5650 CPU, about 9-12 times speedup has been achieved.

Table 1. Timing comparisons among CPU, MIC and GPU for Tomotherapy and TrueBeam cases.

Processor	No. of particles	Total wall time [sec]	
		Tomotherapy	TrueBeam
1 X5650 CPU(one thread)	600 million	2057	758
1 5110P MIC	600 million	187	81
3 5110P MICs	600 million	84	40
1 Kepler K40 GPU	600 million	173	73
2 Kepler K40 GPU	600 million	89	39

4 Discussions

The treatment head model is crucial to the Monte Carlo dose simulation in external radiotherapy. Some previous studies have shown good agreements between simulation and measurement using MC simulated phase space as the head model. As the most accurate way of head model, direct and full MC simulation of accelerator head components can provide us a phase space file which contained nearly realistic particles' fluence distribution for the rest phantom simulation. The GPU implementation using Phase space file [16] were also carried out in several different methods.

In this work, we took the concept of approximate transport through patient-dependent collimators, where the jaws used First-Compton based model and then intensity modulated by using the fluence distribution. This model is efficient to implement on MIC/GPU while its accuracy is validated through the comparison between the simulation result and measurement of several open fields range from 4×4 cm^2 to 40×40 cm^2.

As for speed, compared to a highly optimized CPU implementation, the MIC and GPU have shown comparable performance gain of 9 to 12 times when simulating the real clinical cases. The reasons we optimized CPU code are mainly for code portability and fair comparison. To compare the accelerator's performance with CPU, the sequential code should be optimized, multi-thread parallelized. Additionally, although single precision floating point operations are faster than double precision, it's also more error-prone, since it could result in dose underestimate, especially for the voxels surrounding a radiation source. This importance of using double-precision for radiotherapy dose calculation was proven by Magnoux et al. in their paper [18].

5 Conclusions

In this study, we have successfully developed and extended ARCHER to Xeon Phi-based fast and accurate dose calculation engine for radiotherapy. The simulated results were validated with experimental measurement and one clinical TomoTherapy treatment plan of the prostate and one TrueBeam treatment plan of the breast were studied to further benchmark against DPM to show the clinical utility. Optimization techniques have been applied, includes sorting the source particles on-the-fly before transport to Xeon Phi and GPU based on particle type and energy, and batch method transferring the Phase Space source to reduce the preprocess time. The future work includes the development of explicit approximate transport of linac MLCs and extension to other energy levels and other popular vendor machines such as Varian TrueBeam Flattening Filter Free machine.

Acknowledgement

This research is funded by the US National Institute of Biomedical Imaging and Bioengineering (R01EB015478). We would also like to thank Intel and Nvidia for the generous hardware donation.

References

1. T. Bortfeld. Phys. Med. Biol.51, R363–R379 (2006)
2. A. Fogliata, E. Vanetti, D. Albers, C. Brink, A. Clivio, T. Knöös, G. Nicolini, and L. Cozzi. Phys. Med. Biol.52, 1363–1385 (2007)
3. XG Xu, T. Liu, L. Su, et al. Annals of Nuclear Energy (2014)
4. T Liu, XG Xu, CD Carothers. Annals of Nuclear Energy (2014)
5. XG Xu, T Liu, L Su, FB Brown et al. Trans. Am. Nucl. Soc (2013)
6. L Su, Y Yang, B Bednarz, XG Xu et al. Medical physics (2014)
7. I. Kawrakow and D. W. O. Rogers. NRCC Report No. PIRS-701 (2011)
8. S. Agostinelli, J. Allison, K. E. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, and G. Barrand. Nucl. Instrum. Methods, Phys. Res A506(3), 250–303 (2003)
9. Capote R. Radiother. Oncol.84S217 (2007)
10. Schmidhalter D, Manser P, Frei D, Volken W and Fix M K. Med. Phys.37492–504 (2010)
11. Sempau J, Wilderman S J and Bielajew A F. Physics in Medicine and Biology 45 2263-91 (2000)
12. Woodcock E, Murphy T, Hemmings P, et al. Applications of Computing Methods to Reactor Problems: Argonne National Laboratorie s Report) pp ANL -7050 (1965)

13. N. Tyagi, A. Bose, and I. J. Chetty. Med. Phys.31, 2721–2725 (2004)
14. F. Verhaegen and S. Devic. Phys. Med. Biol.50, 937–946 (2005)
15. I. Kawrakow and B. R. B. Walters. Med. Phys.33, 3046–3056 (2006)
16. Townson, Reid W., et al. Physics in medicine and biology 58.12 (2013)
17. Siebers, Jeffrey V., et al. Physics in medicine and biology 47.17 :3225 (2002)
18. Magnoux, Vincent, et al. Physics in medicine and biology 60.13 : 5007 (2015)