Available online at www.sciencedirect.com
ScienceDirect
Nuclear Instruments and Methods in Physics Research A 569 (2006) 341–345
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

GATE: Improving the computational efficiency

S. Staelens[a,*], J. De Beenhouwer[a], D. Kruecker[b], L. Maigne[c], F. Rannou[d], L. Ferrer[e], Y. D'Asseler[a], I. Buvat[f], I. Lemahieu[a]

[a]*UGent-ELIS, St-Pietersnieuwstraat, 41, B-9000 Gent, Belgium*
[b]*Institute of Medicine-Forschungszemtrum Juelich, D-52425 Juelich, Germany*
[c]*Département de Curiethérapie-Radiothérapie, Centre Jean Perrin, F-63000 Clermont-Ferrand, France*
[d]*Departamento de Ingenieria Informatica, Universidad de Santiago de Chile, Santiago, Chile*
[e]*INSERM U601, CHU Nantes, F-44093 Nantes, France*
[f]*INSERM U678 UPMC, CHU Pitié-Salpêtrière, F-75634 Paris, France*

Available online 11 September 2006

Abstract

GATE is a software dedicated to Monte Carlo simulations in Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET). An important disadvantage of those simulations is the fundamental burden of computation time. This manuscript describes three different techniques in order to improve the efficiency of those simulations. Firstly, the implementation of variance reduction techniques (VRTs), more specifically the incorporation of geometrical importance sampling, is discussed. After this, the newly designed cluster version of the GATE software is described. The experiments have shown that GATE simulations scale very well on a cluster of homogeneous computers. Finally, an elaboration on the deployment of GATE on the Enabling Grids for E-Science in Europe (EGEE) grid will conclude the description of efficiency enhancement efforts. The three aforementioned methods improve the efficiency of GATE to a large extent and make realistic patient-specific overnight Monte Carlo simulations achievable.

PACS: 87.53.Vb; 87.53.Wz; 87.57.−s

Keywords: Monte Carlo; GATE; Variance reduction; Cluster; Grid

1. Introduction

A new software, GATE [1], was designed as an upper layer for the Geant4 nuclear physics code and was tuned for use in nuclear medicine, more specifically to fulfill its role as a simulation platform for Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) incorporating all Geant4 features. An important disadvantage of GATE Monte Carlo simulations is the fundamental burden of computation time. This manuscript will describe three different techniques in order to improve the efficiency of those simulations. Firstly we will discuss the implementation of variance reduction techniques (VRTs), more specifically the incorporation of geometrical importance sampling. Secondly, the newly designed cluster version of the GATE software will be described. Finally, an elaboration on the deployment of GATE on the Enabling Grids for E-Science in Europe (EGEE) grid will conclude the description of the efficiency enhancement efforts.

2. Methods

Benchmark simulations were performed to estimate the simulation time for realistic PET and SPECT nuclear medicine setups. These benchmarks have been described in full detail in Ref. [1]. The computing time for the

*Corresponding author. Tel.: +32 9 264 66 28; fax: +32 9 264 66 18.
E-mail address: steven.staelens@ugent.be (S. Staelens).

0168-9002/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2006.08.070

PET benchmark averaged around 12 h on a 1.0 GHz processor. This corresponds to 852 generated and tracked events per second and 16 simulated coincidence detections per second. Calculation time for the SPECT case was 11 h on a 1 GHz processor, resulting in 417 generated and tracked events per second and 0.83 detections per second or 1.2 s per detection which is worse due to the collimator.

2.1. VRTs: methodology

Geometrical importance sampling is a VRT based on the crude criterion that only photons with a high detection chance should be tracked. Photons are increasingly split into exact copies with lowered weights as the distance to a detector decreases. Photon paths leading away from a detector are less likely to result in detection and therefore these photons are subject to Russian roulette in order to increase the simulation efficiency [2]. Geometrical importance sampling combined with Russian roulette introduces branches into the particle history. Simply adding all hits in a detector crystal would lead to a completely wrong simulated histogram. Therefore, a new track history has been developed within GATE [3]. It keeps a log of all tracks and their weights generated by GEANT4 and tracks are added where necessary to accomodate splitting and Russian roulette. The efficiency gain of importance sampling is inversely related to the sensitivity of the detector and despite the complex detangling and increased tracking overhead, it can result in a 5–15-fold increase over analog simulations. The incorporation of importance sampling has been validated extensively.

An efficiency comparison of this newly incorporated VRT with analog GATE simulations has been performed on a ^{67}Ga energy spectrum. Activity and acquisition time were kept constant while simulation time and number of detections are the parameters of interest. The efficiency will be defined as the number of detections/second times the quality factor (QF) which indicates the variation of the weights of the detected particles [4].

2.2. Running GATE on a cluster: methodology

In Monte Carlo simulations the amount of inter-process communication is small, and usually only at process start-up and termination. The parallelized simulations are made up of three steps: job splitting, the actual simulations (on a number of CPUs) and file merging. The most natural, simple and general scheme for splitting PET and SPECT simulations is the time-domain decomposition approach, in which the length of the experiment is split into a number of equally long smaller experiments. This approach does not involve any approximation nor simplification. Measures like random rates, scatter fractions, and system deadtime will be effectively the same as in a single-node run. The input to the job splitter are the GATE scripts, parameters and command-line options. A Random Number Generator (RNG) provides statistically independent seeds for the output which is a collection of non-parameterized (fully resolved) macro-files. The job splitter also provides a submit file for supported cluster platforms such as Condor and openMosix, to facilitate the startup of the simulation. Finally, a split file is generated that contains all information about the partitioned simulation to facilitate the merging of the output files. Since GATE does not allow any volume movement during data acquisition, virtual time slices are used for the time-domain decomposition. The output merger takes then finally as input the ROOT output files from the parallelized simulations and uses the split file to merge them. The implementation we describe is fully automatic and requires no interaction from the user [5].

To test the efficiency enhancement, the GATE benchmarks were run. The cluster used was based on openMosix and consisted of 38 nodes with 17 dual XEON 2.4 GHz processors and 21 dual XEON 2.8 GHz processors, each with 2 GB of memory. The benchmark series were executed using the cluster with an increasing number of CPUs. Efficiency was hereby defined as an acceleration factor (AF) based on Amdahl's law: $\mathrm{AF} = (T_s + T_p)/(T_s + T_p/p)$, where p is the number of CPUs, and T_s and T_p are the serial and parallel application times, respectively [6]. T_s was measured as the time necessary to process the GATE scripts that set up the simulations. T_p was calculated from the total simulation time and T_s. A more appropriate estimate (AE) of the acceleration factor can be made by inclusion of the merging time T_m into Amdahl's law: $(T_s + T_p)/(T_m + T_s + T_p/p)$.

2.3. Grid: methodology

EGEE [7] is part of the grid initiatives launched by the European Union to structure the grid infrastructures in Europe. GATE is a pilot biomedical application in this project. GATE simulations benefit from the geographically distributed grid computing resources to be parallelized thereby obtaining significant gain in computing time and are to be used in a near future in clinical routine for some specific applications. Currently, computing resources in EGEE, allocated to biomedical applications, consist of 1785 CPUs distributed on 23 geographic sites in Europe and Taiwan. In order to enable a transparent and interactive use of GATE applications on the grid, all the functionalities to run GATE simulations on distributed resources have been developed [8].

In order to show the advantage for the GATE simulations to partition the calculation on multiple processors, a simulation ran locally on a single processor Intel Xeon 3.06 GHz and was executed in parallel on two grid sites: Centre de Calcul of the Institut National de Physique Nucleaire et de Physique des Particules (CCIN2P3) and Laboratoire de Physique Corpusculaire Clermont-Ferrand (LPC). Splitting was done in 1, 10, 20,

50 and 100 subsimulations. The efficiency enhancement is defined here as a straightforward speed-up similar to AE (Section 2.2).

3. Results

3.1. VRTs

An efficiency test was performed by acquiring a ^{67}Ga-spectrum via the pulse height analyzer of a typical SPECT simulation study. Fig. 1 clearly shows a manifest reduction in variance. Table 1 shows the corresponding efficiency enhancement.

The efficiency of each simulation is calculated as described in Section 2.1. Dividing the efficiencies gives an indication of the relative efficiency enhancement, being 5.1 for the VRT case. When applying VRTs one should take into account that the statistics of the results are no longer Poissonian. Several noise restoration processes are described in literature, such as combining VRT with an intermediate Bernouilli experiment [9].

3.2. Running GATE on a cluster

The results of running GATE on a cluster are described in Fig. 2. For SPECT only a small deviation from the Amdahl prediction for linear scalability is observed, so the duration of a SPECT simulation basically drops with the number of CPUs if run on a cluster. For PET we see a large deviation from Amdahl's linear prediction. This can be explained as follows: the SPECT benchmark is a typical example of a low sensitivity system characterized by relatively small output file sizes for a long simulation time, whereas the PET setup resembles a high sensitivity system

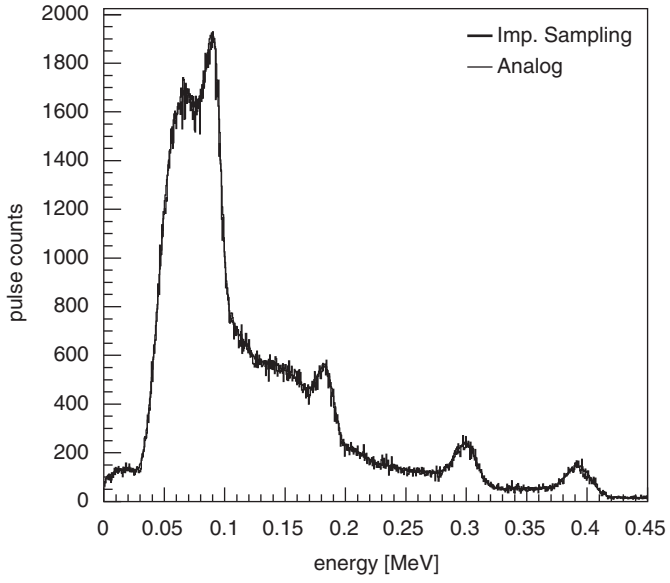

Fig. 1. ^{67}Ga spectra for simulations with (thick line) and without (thin line) importance sampling.

Table 1
VRT comparison

	Importance sampling	Analog
Activity	100 MBq	100 MBq
Acquisition time	30 s	30 s
Simulation time	5,600,000 s	1,954,000 s
Detections	5,165,891	325,920
QF	0.92	1

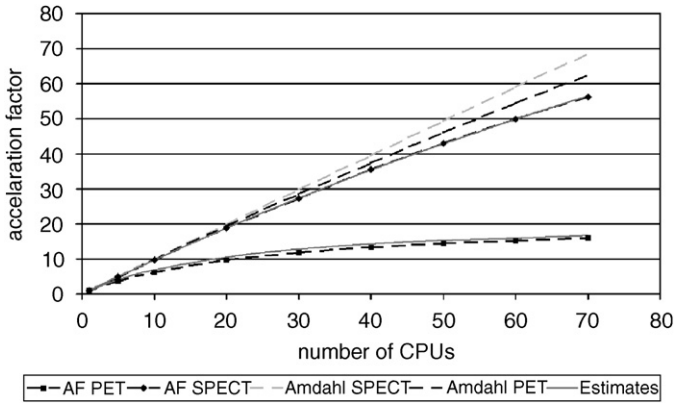

Fig. 2. Efficiency enhancements for the SPECT and PET experiments with the upper limits predicted by Amdahl's law and the estimates including the merging time.

with relatively high output volumes for a short simulation time. The percentual contributions of the output merger become larger for an increasing number of CPUs and this is a bottleneck for simulations with high data output sizes as shown in Fig. 3.

In a worst case scenario, high output, short duration simulation studies, can reach an optimum in cluster operation mode. That optimum can easily be calculated for the estimate (AE) incorporating the merging time for which we see a good agreement in Fig. 2.

3.3. Deployment on the grid

Table 2 illustrates the computing time in minutes of a GATE simulation running on a single processor Intel Xeon 3.06 GHz locally and the same simulation split in 10, 20, 50 and 100 jobs on the CCIN2P3 site. In this case the lowest computing time is obtained for 20 jobs running in parallel.

Concerning the same type of submission on the LPC site, Table 3 shows that the lowest computing time is obtained for the maximum splitting (100 partitions).

A gain factor of 28.2 is reached for this last submission that enables the user to run his simulation in 10 min instead of more than 4 h on a single processor. Performance loss and non-scalability is due to the queuing time that is really dependent on the policy of the site accepting the jobs. A possible bottleneck also remains in the network

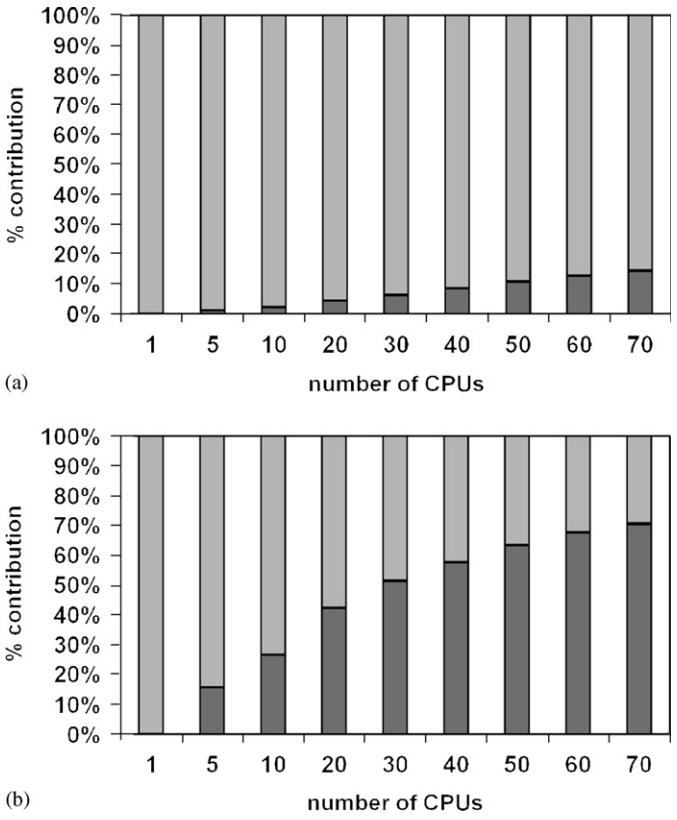

Fig. 3. Percentage of the total application time for the job splitter (black), the output merger (dark grey) and the actual simulation (light grey) as function of number of CPUs in (a) SPECT and (b) PET.

Table 2
Sequential versus grid computation time at CCIN2P3

Number of jobs	1 (local)	10	20	50	100
Computation time (min)	276.5	60.8	40.9	53.7	51.8
Efficiency		4.5	6.8	5.1	5.3

Table 3
Sequential versus grid computation time at LPC site

Number of jobs	1 (local)	10	20	50	100
Computation time (min)	276.5	37.5	22.6	13.9	9.8
Efficiency		7.4	12.2	19.9	28.3

connection with the remote site for analysis of the data. Therefore, there is a trade-off between job splitting and intrinsic grid time costs.

4. Discussion and conclusion

Classical ways to decrease Monte Carlo simulation time are the application of energy and path length cuts, as well as limiting the emission angle which are available in GATE by default as well as the use of parametrized voxels. Here, we presented the three additional implementations to improve the efficiency. For importance sampling, a maximal factor of 15 is expected by using the optimal splitting map, calculated by an inverse simulation. Recent work is ongoing for the incorporation of forced detection into GATE. The current implementation of importance sampling copies the time stamp which causes it currently to be limited to SPECT only thereby respecting the virtual clock philosophy for time management. In PET more complex sorting algorithms for coincidences, scatter fractions and random rates are needed which are mostly based on unique time window information. Moreover, in PET, gamma pairs are created which complicates the current track detangling implementation. For the cluster implementation, investigation is ongoing on how to parallellize the merger bottleneck. In Ref. [10] we already report on the use of this cluster software for the first time in research. Finally, the results obtained with grid are very encouraging and with the CPU resources increasing in the future, we can expect an increased efficiency enhancement factor. On the long term, incorporation in GATE of the fictituous cross-section method [11] and of precalculated detector response [12] will be studied.

GATE is still fundamentally slower compared to analytical simulators or dedicated packages but recent efforts in efficiency enhancement have narrowed the gap significantly and will do so even more in the future. Several cluster testing sites are running realistic GATE simulations already overnight, even for observer studies.

Acknowledgments

This work was supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT, Belgium), by the Fund for Scientific Research Flanders (FWO, Belgium) and by Ghent University.

References

[1] S. Jan, G. Santin, D. Strul, S. Staelens, et al., Phys. Med. Biol. 49 (19) (2004) 4543.
[2] M. Ljungberg, S. Strand, Comput. Methods Programs Biomed. 29 (1989) 257.
[3] J. De Beenhouwer, S. Staelens, M. Dressel, Y.D'Asseler, et al., Geometrical importance sampling and pulse height tallies in gate, in: Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, 2004, pp. 1349–1352.
[4] D. Haynor, R. Harrison, T. Lewellen, Med. Phys. 18 (5) (1991) 990.
[5] J. De Beenhouwer, D. Kruecker, S. Staelens, L. Ferrer, et al., Distributed computing platform for PET and SPECT simulations with GATE, in: Proceedings of the 2005 IEEE Medical Imaging Conference, Puerto Rico, 2005, pp. 2437–2440.
[6] G. Amdahl, Validity of single-processor approach to achieving large-scale computing capability, in: Proceedings of the AFIPS Conference, Reston, VA, 1967, pp. 483–485.
[7] EGEE, ⟨http://www.eu-egee.org/⟩.
[8] L. Maigne, D. Hill, P. Calvat, V. Breton, et al., Parallel Process. Lett. J. 14 (2) (2004) 177.
[9] A. Goedicke, B. Schweizer, S. Staelens, J. De Beenhouwer, Fast simulation of realistic SPECT projections using forced detection in

Geant4, in: Proceedings of the 3rd European Medical and Biological Engineering Conference, Praag, 2005, p. 6.
[10] S. Staelens, , K. Vunckx, D. Beque, Y. D'Asseler, et al., GATE simulations for optimization of pinhole imaging, Nucl. Instr. and Meth. A (2006), is press, doi:10.1016/j.nima.2006.08.071.
[11] I. Kawrakow, M. Fippel, Phys. Med. Biol. 45 (8) (2000) 2163.
[12] X. Song, W. Segars, Y. Du, B. Tsui, et al., Phys. Med. Biol. 50 (8) (2005) 1791.