Monte Carlo–based fluorescence molecular tomography reconstruction method accelerated by a cluster of graphic processing units

Guotao Quan, Hui Gong, Yong Deng, Jianwei Fu, and Qingming Luo
Huazhong University of Science and Technology, Wuhan National Laboratory for Optoelectronics, Britton Chance Center for Biomedical Photonics, Wuhan 430074, China

Abstract. High-speed fluorescence molecular tomography (FMT) reconstruction for 3-D heterogeneous media is still one of the most challenging problems in diffusive optical fluorescence imaging. In this paper, we propose a fast FMT reconstruction method that is based on Monte Carlo (MC) simulation and accelerated by a cluster of graphics processing units (GPUs). Based on the Message Passing Interface standard, we modified the MC code for fast FMT reconstruction, and different Green's functions representing the flux distribution in media are calculated simultaneously by different GPUs in the cluster. A load-balancing method was also developed to increase the computational efficiency. By applying the Fréchet derivative, a Jacobian matrix is formed to reconstruct the distribution of the fluorochromes using the calculated Green's functions. Phantom experiments have shown that only 10 min are required to get reconstruction results with a cluster of 6 GPUs, rather than 6 h with a cluster of multiple dual opteron CPU nodes. Because of the advantages of high accuracy and suitability for 3-D heterogeneity media with refractive-index-unmatched boundaries from the MC simulation, the GPU cluster-accelerated method provides a reliable approach to high-speed reconstruction for FMT imaging. *© 2011 Society of Photo-Optical Instrumentation Engineers (SPIE).* [DOI: 10.1117/1.3544548]

Keywords: fluorescence molecular tomography; reconstruction; Monte Carlo; cluster of graphics processing units.

Paper 10513R received Sep. 17, 2010; revised manuscript received Dec. 29, 2010; accepted for publication Jan. 3, 2011; published online Feb. 15, 2011.

1 Introduction

As fluorescence labeling technology develops,[1,2] an increasing number of biologists want to observe the distribution of fluorescence targets *in vivo*. Therefore, fluorescence molecular tomography (FMT)[3,4] has been developed. Because FMT can be used to quantitatively reveal the fluorescence marker's fluorescence yield[5,6] and lifetime[7] *in vivo*, FMT is a promising, small-animal imaging method for drug development and cancer research.[8,9]

The reconstruction algorithm is the core technology of FMT. The algorithm can be summarized as follows.[4,10] First, the distribution of the photon density of fluorescence with a different source, which is called Green's function, is calculated. Then, by applying the Fréchet derivative, a Jacobian matrix is formed with the calculated Green's functions. Finally, the distribution of the fluorescence yield is inversely calculated from the Jacobian matrix by an optimization algorithm. The first step in the algorithm is called the forward problem, which investigates the propagation of light in tissue; the second and third steps are called the inverse problem, which reconstructs the distribution of the fluorescence yield. A traditional reconstruction algorithm has been proposed based on a diffusion approximation (DA) and using the finite-element method (FEM).[11–14] The traditional method is very fast, but it is valid only when the reduced scattering coefficients of the tissue are far greater than the absorption coefficients. Although the FMT reconstruction method based on the high order of the radiative transport equation (RTE) is suitable for heterogeneous media with complex distributions of optical coefficients, low computational efficiency limits its application. For example, Joshi et al. proposed a radiative transport-based frequency-domain fluorescence tomography with the most accurate angular discretization of RTE, but it requires more than 3.5 h on a 16-node Beowulf cluster.[15]

Monte Carlo (MC) methods were introduced in FMT. They are used to calculate the Green's functions with which the Jacobian matrix is formed to reconstruct the distribution of the fluorescence yield. Because MC methods are the gold standard for simulating the light propagation in tissue and are valid for all kinds of tissue,[16–18] they are suitable for small-animal imaging, which has a complex distribution of optical coefficients. Kumar et al. proposed a reconstruction method based on a MC method that reconstructs the distribution of the fluorescence lifetime in time-domain fluorescence tomography for small-animal imaging. A phantom experiment showed that this method can clearly separate two fluorochromes with 6-mm spacing.[19] Zhang et al. also introduced MC methods into the reconstruction of fluorescence tomography for small-animal imaging.[20] However, because of the low computational efficiency and the large number of Green's functions requiring calculation, more than 6 h are required to reconstruct the result, even with a parallel computing cluster of central processing units (CPUs).

Address all correspondence to: Hui Gong, Huazhong University of Science and Technology, Wuhan National Laboratory for Optoelectronics, Britton Chance Center for Biomedical Photonics, Wuhan 430074, China. E-mail: huigong@mail.hust.edu.cn.

1083-3668/2011/16(2)/026018/8/$25.00 © 2011 SPIE

Graphics processing units (GPUs) have been introduced in MC simulations to accelerate the simulation of the propagation of light in tissue when the distribution of the optical coefficient is already determined. The GPU is the core of a graphics card. In the past, all software ran on CPU; GPUs were only used in image processing and output until general-purpose computing on graphics processing units were proposed and unleashed the power of GPUs for parallel computation. The peak floating-point operations per/s (flop/s) of marketavailable graphics hardware can reach 1000 Gflop/s, which is 10 times higher than that of a Harpertown CPU with a 3.2-GHz frequency.[21] By using the power of the GPU for parallel computation, MC methods, which simulate the propagation of light in tissue, can be accelerated to speeds that are 100 to 1000 times faster than those obtained with the CPU only. For example, Alerstam et al. used a single GPU to accelerate an MC simulation by a factor of 1000 for multilayered tissues;[22] Lo et al. proposed a multi-GPU-accelerated MC simulation for photodynamic therapy treatment;[23] Fang and Boas sped up an MC simulation with a single GPU by a factor of 100 to 300 for complex 3-D turbid media;[24] and Fang also released their code named Monte Carlo eXtreme (MCX).[25]

Until now, the FMT reconstruction method based on DA with FEM is still limited in high-scattering tissue; however, many kinds of tissues in small animals do not satisfy the condition of high-scattering, such as rabbit muscle, colon adenocarcinomas,[26] the cerebral spinal fluid layer in the brain, and cysts in the human breast.[24] FMT reconstruction methods based on the high orders of both RTE and MC are suitable for heterogeneity media with a complex distribution of optical coefficients, but poor computational efficiency has tremendously limited their applications. GPU can tremendously accelerate the speed of MC; however, it has not yet been used in FMT reconstruction. Furthermore, single GPU-accelerated FMT reconstruction cannot satisfy the demand for fast reconstruction because of the large number of source-detector pairs.

In this paper, we propose a fast MC-based FMT reconstruction, which is suitable for 3-D heterogeneity media with refractive-index-unmatched boundaries and a complex distribution of optical coefficients. Phantom experiments with either high or low-scattering are used to demonstrate the accuracy and speed of the method. By analyzing the reconstructed localization and concentration of the fluorochromes, we compare the accuracy of our method with a traditional method based on DA with FEM. By the load-balancing strategy, we optimize the performance of the GPU cluster for faster reconstruction and compare with that based on CPU and single GPU.

2 Method

2.1 *Theory of FMT*

In this section, a brief introduction to the theory of FMT reconstruction based on MC simulations accelerated by a GPU cluster will be given.

The distribution of fluorescence photons in tissue can be expressed as follows:[3,7]

$$\phi_f(R_d, R_s) = \int g^{\lambda em}(R_d, r)x(r)\phi^{\lambda ex}(r, R_s)dr, \tag{1}$$

where $\phi_f(R_d, R_s)$ is the photon density of fluorescence at R_d with a source R_s, $g^{\lambda em}(R_d, r)$ is Green's function with a source at r (whose wavelength is λem) and a detector at R_d, $\phi^{\lambda ex}(r, R_s)$ is the excitation photon density with a source at R_s (whose wavelength is λex) and a detector at r and $x(r)$ can be expressed as follows:

$$x(r) = \eta\mu_{\alpha f}(r)\frac{1 - j\omega\tau(r)}{1 + (\omega\tau(r))^2}, \tag{2}$$

where $\eta\mu_{\alpha f}(r)$ is fluorescence yield, η is the quantum efficiency, $\mu_{\alpha f}(r)$ is the absorption coefficient of the fluorochrome, $\tau(r)$ is the lifetime of the fluorochrome at r,[11] and ω is the modality frequency of the light source. Because our system is a continuous wavelength system, therefore in this paper, $\omega = 0$ and $x(r) = \eta\mu_{\alpha f}(r)$.

When the light source is a narrow collimation laser, the light source can be expressed as a δ function, so $\phi^{\lambda ex}(r, R_s) = g^{\lambda ex}(r, R_s)$ (Refs. 7 and 27), and, according to reciprocity,[28] $g^{\lambda em}(R_d, r) = g^{\lambda em}(r, R_d)$. Equation (1) can be expressed as follows:

$$\phi_f(R_d, R_s) = \int g^{\lambda em}(r, R_d)x(r)g^{\lambda ex}(r, R_s)dr. \tag{3}$$

Applying the Fréchet derivative and combining Eq. (3) yields the Jacobian function:[4,7]

$$J(R_d, R_s) = \frac{\partial[\phi_f(R_s, R_d)]}{\partial x} = \int g^{\lambda em}(r, R_d)g^{\lambda ex}(r, R_s)dr, \tag{4}$$

where $J(R_d, R_s)$ is the Jacobian function with a source at R_s and a detector at R_d.

The number of reconstructed fluorescence yields of FMT was determined by the number of discrete points (Np). To reduce the ill-posed number of coefficients for FMT, the measurement points referring to the number of sources (Ns) multiplied by the number of detectors (Nd) must be greater than the number of reconstructed fluorescence yield, which is equal to Np. Therefore, the number of source-detector pairs is very large. When there are Ns sources, Nd detectors, and Np discrete points in the experiment, the Jacobian matrix can be expressed as follows:

$$J = \begin{bmatrix} J\left(R_d^1, R_s^1\right) \\ \cdots\cdots \\ J\left(R_d^i, R_s^j\right) \\ J\left(R_d^{Nd}, R_s^{Ns}\right) \end{bmatrix}. \tag{5}$$

In this paper, the Green's function in Eq. (4) was calculated by the kernel of MCX. According to Eq. (5), there are $Ns + Nd$ Green's functions to be calculated to construct a $(Ns \times Nd) \times Np$ Jacobian matrix.

Combining the Jacobian matrix in Eq. (5), and according to Tikhonov regularization,[29,30] the distribution of the fluorescence yield can be reconstructed. Tikhonov regularization can be expressed as

$$\min(||Jx - y||^2 + \lambda||\Gamma x||), \tag{6}$$

where J is the Jacobian matrix shown in Eq. (5), Γ is the Tikhonov matrix (generally, $\Gamma = I$ where I is the identity matrix) and $\lambda = 10^{-20}$, where λ is the relaxation factor, which was determined by the L-curve method.[31] The variable y is the photon density of fluorescence in the experiment, and $x = \eta\mu_{af}$ represents the reconstructed fluorescence yield.[11] The conjugate gradient method was used to solve Eq. (6).[32]

2.2 *Acceleration by a GPU Cluster*

To further improve the speed of the reconstruction method based on MC simulations, with the goal of satisfying the demand for fast reconstruction in FMT, a GPU cluster was constructed with MPICH2, which is a high-performance and widely portable implementation of the MPI standard (both MPI 1.0 and MPI 2.0). The entire calculation task of the Green's functions needed by the inverse problem was distributed to the GPU cluster. The hardware configuration is listed in Table 1. Three personal computers in a local area network equipped with a total of six GPUs of the G200 framework (which supports both double and single precision) were used to construct a GPU cluster. A flowchart of the GPU cluster-accelerated FMT reconstruction method based on MC simulations is shown in Fig. 1.

The processes in Fig. 1 can be briefly summarized as follows:

(1) The host computer receives the file containing the position and direction of the sources and the detectors, the average optical coefficients, and the phantom size from the experiment. The computer then transmits this information to each computing node in the GPU cluster.

(2) The CPU compute nodes receive all of the experimental information from the host computer and distribute the tasks of Green's function computation into different GPUs. Because the calculation time for Green's functions is different for different GPUs, the time consumption is determined by the GPU that would have the lowest performance if the computing tasks were equally distributed into each GPU of the cluster. Therefore, a load-balancing strategy is used to achieve maximum parallel efficiency. All tasks are distributed into GPUs according to the following equation:

$$N_i = T_{\text{task}} * \frac{P\_\text{GPU}_i}{\sum_{i}^{N_{\text{GPU}}} P\_\text{GPU}_i}, \tag{7}$$

where N_i is the number of computing tasks for Green's functions for the i'th GPU and T_{task} is the total number of Green's function tasks. P_GPU_i is the processing power of the i'th GPU, which is listed in Table 1. N_{GPU} is the total number of GPUs. The load-balancing of CPU is not considered because the main compute task in the compute node of CPU is to calculate the Green's function by GPU, so CPU is rarely used.

(3) According to Eq. (7), the start and end indices of the computing task (i_{start}, i_{end}) for each GPU are calculated. The MC simulation is used to calculate the Green's function from i_{start} to i_{end}, and the fluences produced by MC are saved in a file named i.mc2. The MC simulation refers to MCX.

(4) When the entire list of Green's functions has been calculated by the GPU cluster, all of the files that record the results of the Green's functions are transformed to the host computer to construct the Jacobian matrix required for reconstruction.

(5) Tikhonov regularization is used to reconstruct the distribution of the fluorescence yield. The expression is shown in Eq. (6).

To ensure the validity of the Green's function calculation, the kernel of the MC simulation refers to MCX with the fast math library of compute unified device architecture (CUDA) and without the atomic compile option. The random-number generators (RNGs) for the MC simulation use a floating-point based RNG with a chaotic logistic map, the size of the logistic lattice is 5, this RNG has a higher computational efficiency as the optimization of 32 bit floating-point operations.[24] In addition, all of the codes based on the CPU were compiled in Visual Studio 2008 with the -02 (MaxSpeed) option on a Windows Server 2008 R2 X64 system. The codes based on the GPU were compiled with the NVCC 3.0 (beta) which is a compiler in CUDA.

3 Phantom Experiments

In this section, two phantom experiments for both low- and high-scattering media are used to demonstrate the accuracy and speed of MC-based FMT reconstruction accelerated by a GPU cluster.

3.1 *Reconstruction Accuracy*

Four glass tubes (2.2-mm in diameter) with different concentrations of Dir-Boa fluorescence dye[33] were placed in a rectangular box ($13.5\times40.5\times60$ mm) with a mixed solution of Intralipid and India ink. The concentration in the four glass tubes ranged from 1.8 μmol/L to 1.2 μmol/L. Two phantom experiments with different kinds of optical coefficients for the mixed solution of Intralipid and India ink were conducted. The glass tubes in the phantom experiment are heterogeneous media because the optical coefficients of the solution of Dir-Boa fluorescence dye (high-absorption, low-scattering) are different from those for the mixed solution of Intralipid and India ink, and the refractive index of glass is also mismatched with the mixed solution.

Table 1 The hardware of the GPU cluster.

	CPU	Graphic card	Processing power (peak) Gflop/s (Ref. 40)
Computer 1	Pentium(R) Dual-Core E5300	NVIDIA GTX 295 with two GPUs	1788.48
Computer 2	Intel(R) Core™ i7 920	NVIDIA GTX 295 with two GPUs	1788.48
Computer 3	Intel(R) Xeon(R) X5570	NVIDIA Quadro FX4800 with one GPU	693.50
		NVIDIA Tesla C1060 with one GPU	933.12

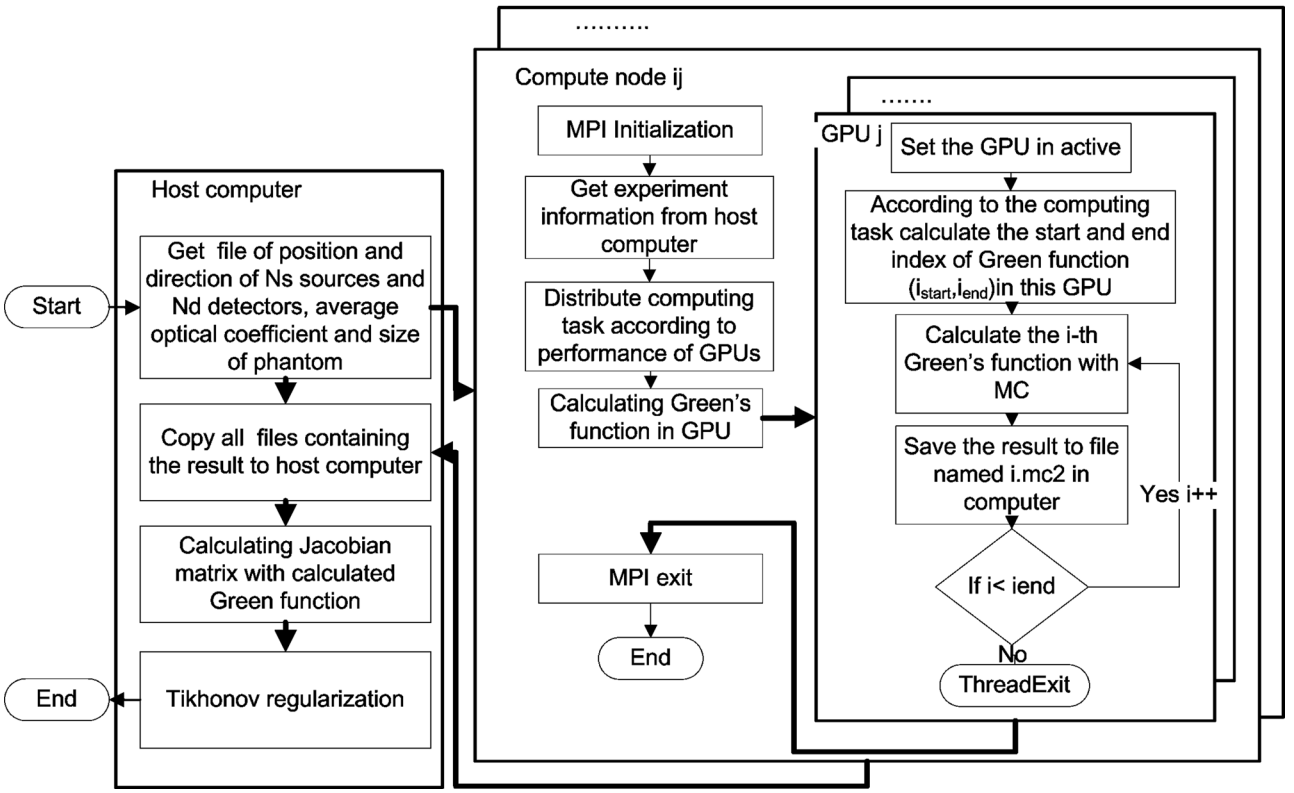

Fig. 1 A flowchart of the GPU cluster-accelerated FMT reconstruction method based on MC simulations.

The heterogeneity of the fluorescence solution is considered in the calculation of the Green's function for MC-based FMT reconstruction but not in that based on DA with FEM because it is not accurate for DA when $\mu_a > \mu_s'$.

Case 1. The absorption coefficient of the first experiment was 0.23 mm^{-1}, and the reduced scattering coefficient was 0.68 mm^{-1}. In this case, $\mu_\alpha \approx \mu_s'$, which indicates a low-scattering case. In theory, the DA will be invalid. However, the MC simulation was accurate.

Case 2. The absorption coefficient of the second experiment was 0.07 mm^{-1}, and the reduced scattering coefficient was 1.8 mm^{-1}. In this case, $\mu_\alpha \ll \mu_s'$, which indicates the high-scattering case. In theory, the DA should be as accurate as the MC simulation.

These two types of optical coefficients refer to the muscle tissue and the skin of a rabbit at a wavelength of 790 nm.[26] A dual-modality imaging system combined with micro-computer tomography (micro-CT) and FMT [34] was then used to obtain the fluorescence images. A micro-CT slice was used to prove the localization accuracy of the reconstruction.

The reconstruction area (1.35×40.5×20 mm) was smaller than the real size of the phantom because the fluorescence signal only appeared in that area. The spatial resolution of the FMT is about 1 mm;[35] therefore, by applying a 1-mm^3 discrete accuracy, the reconstruction area was separated into 14×41×20 voxels. In this experiment, there were 99 sources and 220 detectors, so 319 Green functions were calculated. Therefore, the amount of data from the experiment (99×220 = 21,780) was larger than the number of reconstructed fluorescence yield, which was equal to the number of voxels (14×41×20 = 11,480). As a result, the ill-posed FMT reconstruction could be relieved.

Six GPUs were used to calculate the Green's functions, and 10^5 photon moves were simulated at each thread of the GPU. There were 2560 threads in total and 256 threads in each block of the GPU. We repeated this calculation 12 times to increase the total number of simulation photons and to avoid the kernel launch time-out error.[25] In each calculation of a Green's function, there were 1.36×10^7 photons to be simulated on average, and 4.3410^9 photons were simulated for the total reconstruction process.

The reconstruction method based on a DA with the FEM refers to NIRFAST[36,37] and TOAST.[27,38] By applying a 1-mm^3 mesh, the reconstruction area was discrete, with 12,054 points and 50,100 tetrahedrons.

Tikhonov regularization was solved with a conjugate gradient method that was stopped at the 18th iteration, when the difference between two consecutive iterations was less than 10^{-6}. The reconstructed distribution of the fluorescence yield is shown in Fig. 2.

The average value of the reconstructed fluorescence coefficient of the four fluorochromes for both case 1 and case 2 was calculated from Figs. 2(b), 2(c), 2(e), and 2(f). These values were calculated with different FMT reconstruction methods based on both MC simulations and DA. Because the fluorescence yield was in direct proportion to the concentration of the fluorochromes, the reconstructed concentrations of four fluorochromes were calculated through a linear fit of the fluorescence yield against the real concentration of the fluorochromes. The results are shown in Fig. 3.

The error between the linear fit and the real reconstructed concentration was used to determine the coefficient of the linear fit, which refers to the linearity of the reconstructed

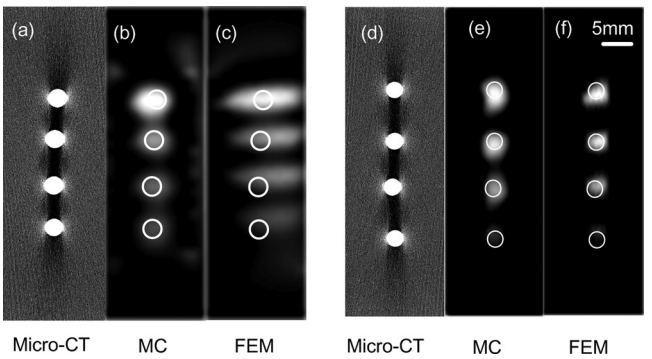

Fig. 2 Reconstruction slices. (a), (b), and (c) show the reconstruction slices for case 1. (d), (e), and (f) show the reconstruction slices for case 2. Micro-CT refers to the central reconstruction slice with Micro-CT. MC refers to the central reconstruction slice for the FMT reconstruction method based on MC simulations accelerated by a GPU cluster. FEM refers to the central reconstruction slice for the FMT reconstruction method based on a DA with the FEM. The position and size of the white circles in (b) and (c) were determined by the Micro-CT slice in (a), and the white circles in panels (e) and (f) were determined by the Micro-CT slice in panel (d).

concentration. The resulting values are listed in Table 2. The sum of squares error (SSE) refers to the error between the linear fit and the reconstructed concentration, and *R*-squared indicates the linearity of the fit. Under ideal conditions, the SSE is zero and *R*-squared is 1.

3.2 *Acceleration Performance*

In this section, the effect of the load-balancing method based on the calculation power of GPU is compared with equal-load-balancing. The acceleration performance for different numbers of GPUs in the GPU cluster is compared for both case 1 and case 2.

We distribute all the tasks into GPU clusters with two different balancing methods: load-balancing with Eq. (7) and equal loading, which equally distributes the task into GPUs. We use the experimental data for case 1 and case 2. The configuration of the MC and GPU is mentioned in Sec. 3.2, and the reconstruction time is recorded for comparison. The results are shown in Table 3.

Table 2 The coefficients of the linear fit.

	Case 1		Case 2	
	FEM	MC	FEM	MC
SSE	0.0657	0.0031	0.0093	0.0055
R-squared	0.7076	0.9837	0.9543	0.9755

From Table 3, we find that load-balancing based on Eq. (7) increases the computational efficiency.

We separately tested the performance of the FMT reconstruction method based on MC simulations with a GPU cluster containing one, two, four, or six GPUs. In every thread of the GPU, 10^5 photon moves were simulated. There were 7680 threads in total in computer 1 and computer 2, but there were 6144 threads in total in computer 3. Every block in the GPU contained 256 threads. Each GPU in a GTX295 graphics card contains 30 independent multiprocessors (MPs), which contain 16384 registers, and MCX occupies 54 registers at each thread. Therefore, 256 threads could be run simultaneously in each MP. For the entire GPU, $30\times256 = 7680$ threads could be run simultaneously. However, the GPU on the NVIDIA Quadro FX4800 graphics card of computer 3 only contained 24 MPs, so $24\times256 = 6144$ threads in total could be run simultaneously. With these settings, the computing power of a graphics card with a G200 framework can be fully unleashed. Each calculation of a Green's function was repeated 12 times to avoid kernel launch time-out errors[25] and to increase the total number of simulated photons.

For each calculation of a Green's function, there were 4×10^7 simulated photons on average and 1.3×10^{10} in total for the entire reconstruction. Because the total number of photons in the reconstruction process was only determined by the total number of calculated Green's functions, the photon moves in each thread, the total number of threads, and the different numbers of GPUs in the GPU cluster did not influence the total number

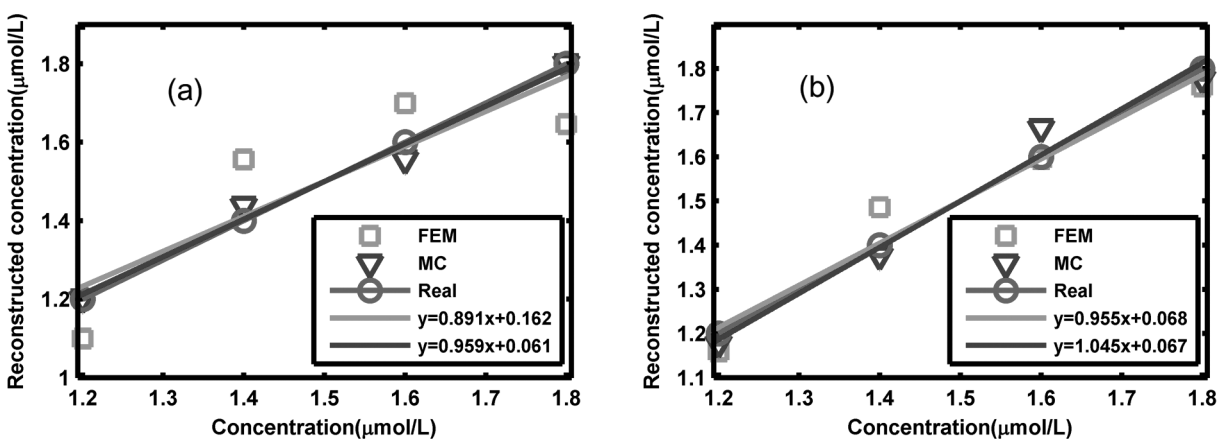

Fig. 3 Reconstruction of the concentration of fluorochromes. (a) shows the reconstructed concentration versus the real concentration of four different fluorochromes for case 1. The concentrations were reconstructed by the FMT reconstruction method based on both MC simulations and DA. (b) represents the case 2 experimental conditions. The linear fit function is also shown in (a) and (b).

Table 3 Performance of load-balancing.

	Case 1		Case 2	
Number of GPUs	Equal load(s)	Load-balancing(s)	Equal load(s)	Load-balancing(s)
4	867	704	1133	920
6	585	465	697	554

of simulation photons. However, there were small differences in the total number of photons when the GPU of the Quadro FX4800 in computer 3 was used because the total number of available threads for the Quadro FX4800 was only 6144.

The influence of the number of GPUs in the GPU cluster on the amount of time consumed for the FMT reconstruction method based on MC simulations is shown in Fig. 4. The time data for one and two GPUs in the GPU cluster came from the computing node of computer 1. Four GPUs refers to the computing nodes of computers 1 and 2, and six GPUs in the GPU cluster refers to computing nodes of computers 1, 2, and 3.

We also compared the acceleration ratio between the speeds of the CPU code and the GPU cluster code. The speed of the CPU code was estimated by the tMCimg code,[39] which was compiled with –O2(maxspeed) in Visual Studio 2008. Because it takes a long time to use the tMCimg in FMT reconstruction, we only tested the speed of tMCimg for one calculation of a Green's function. This speed was approximately equal to the speed for the whole reconstruction because the Green's function calculation consume 95% of the time needed for the total reconstruction. The results showed that for case 1, the calculation speed of a Green's function based on CPU code was 75 photons/ms, and for case 2, the speed was 7.6 photons/ms. The performance using different numbers of GPUs in the GPU cluster is shown in Fig. 5, and the acceleration ratio compared with the CPU is shown in Table 4.

4 Discussion and Conclusion

For the low-scattering case ($\mu_a \approx \mu_s'$), Figs. 2(a), 2(b) and 2(c), and Fig. 3(a) illustrate that the GPU-cluster-accelerated FMT reconstruction method based on MC simulations accurately reconstructed the localization and concentration of fluorochromes. The traditional reconstruction method based on a DA with the FEM failed in this case. This superiority of our method in reconstructing the concentration is quantitatively demonstrated in Table 2.

In the experiment for the high-scattering case ($\mu_\alpha \ll \mu_s'$), which has typically been reconstructed with the traditional reconstruction method based on a DA with the FEM, the GPU cluster–accelerated FMT reconstruction method based on MC simulations also accurately reconstructed the localization and concentration of fluorochromes. Figures 2(d), 2(e), 2(f), Fig. 3(b), and Table 2 show that, in the high-scattering case, the two methods achieve almost identical results. The only disparity is a very small difference in the linearity of the reconstructed concentration because the heterogeneity of the fluorescence solution is considered in the MC-based FMT reconstruction method accelerated by GPU clusters and the size of the fluorescence solution is small, which cannot cause a large error in the calculation of the Green's function. This small difference indicates the advantage of our method for heterogeneous media with refractive-index-unmatched boundaries.

The GPU cluster-accelerated FMT reconstruction method based on MC simulations can accurately reconstruct the localization and concentration of fluorochromes for heterogeneous media under both low- and high-scattering conditions. In contrast, the traditional FMT reconstruction method based on a DA with the FEM was only valid for the case in which $\mu_\alpha \ll \mu_s'$.

Figure 4 shows that the FMT reconstruction method based on MC simulations accelerated by a single GPU for case 1 and case 2 requires 44 and 49 min to finish the whole process, which does not satisfy the demand for fast reconstruction in FMT. However, the FMT reconstruction method based on MC simulations accelerated with six GPUs for case 1 and case 2 requires only 7.7 and 9.2 min, which is the same as the time required by the FMT reconstruction method based on a DA with the FEM.[12] We also found that the more GPUs in the GPU cluster, the less time was

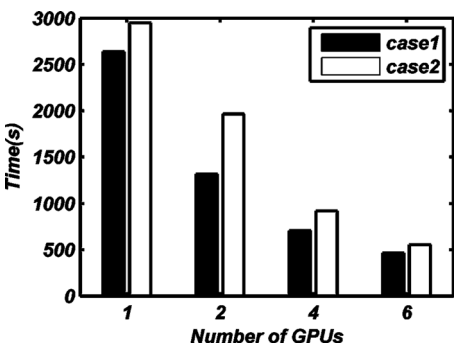

Fig. 4 The time performance of different numbers of GPUs in the GPU cluster for case 1 and case 2.

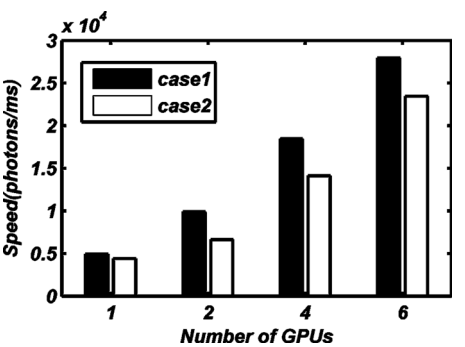

Fig. 5 The speed performance of different numbers of GPUs in the GPU cluster for case 1 and case 2.

Table 4 The acceleration ratio for different numbers of GPUs.

Number of GPUs	1	2	4	6
Acceleration ratio in case 1	66x	132x	246x	373x
Acceleration ratio in case 2	580x	871x	1859x	3088x

required for the FMT reconstruction based on MC simulations, which predicts that a GPU cluster with about 15 Nvidia GTX 480 graphic cards (about 3 times faster than Nvidia GTX 295) may require only 1 min to finish the entire FMT-reconstruction process. Note, however, that the larger reduced scattering coefficient required more processing time in the FMT reconstruction based on MC simulations because a larger reduced scattering coefficient means that the photons need more steps to escape from the tissue, which requires more time to calculate.

The more GPUs in the GPU cluster, the faster the speed of the FMT reconstruction method based on MC simulations. Compared with the CPU code, the FMT reconstruction method based on MC simulations accelerated with 6 GPUs was 3088 times faster for case 2, but only 373 times faster for case 1. This difference is caused by the speed of the FMT reconstruction method based on MC simulations decreasing more quickly for the CPU code than for the GPU cluster when the reduced scattering coefficient increased.

In summary, with the MPI standard and load-balancing method, we propose a fast FMT reconstruction method which is based on MC simulation and accelerated by a GPU cluster. Through two phantom experiments on both high- and low-scattering media, we prove that this method can accurately and rapidly reconstruct the fluorochrome localization and concentration for 3-D heterogeneous media with refractive-index-unmatched boundaries and complex distributions of optical coefficients.

Acknowledgments

We gratefully acknowledge the financial support of the National Major Scientific Research Program of China (Grant No. 2011CB910400) and the National Natural Science Foundation of China (Grant No. 60828009).

References

1. R. Y. Tsien, "Building and breeding molecules to spy on cells and tumors," *Febs. Letters.* **579**(4), 927–932 (2005).
2. M. Funovics, R. Weissleder, and C. H. Tung, "Protease sensors for bioimaging," *Anal. Bioanal. Chem.* **377**(6), 956–963 (2003).
3. V. Ntziachristos, "Fluorescence Molecular Imaging," *Annu. Rev. Biomed. Eng.* **8**(1), 1–33 (2006).
4. A. B. Milstein, S. Oh, K. J. Webb, C. A. Bouman, Q. Zhang, D. A. Boas, and R. P. Millane, "Fluorescence optical diffuse tomography," *Appl. Opt.* **42**(16), 3081–3094 (2003).
5. D. Y. Churmakov, I. V. Meglinski, and D. A. Greenhalgh, "Amending of fluorescence sensor signal localization in human skin by matching of the refractive index," *J. Biomed. Opt.* **9**(2), 339–46 (2004).
6. D. Y. Churmakov, I. V. Meglinski, S. A. Piletsky, and D. A. Greenhalgh, "Analysis of skin tissues spatial fluorescence distribution by the Monte Carlo simulation," *J. Phys. D* **36**(14), 1722–1728 (2003).
7. A. T. Kumar, S. B. Raymond, G. Boverman, D. A. Boas, and B. J. Bacskai, "Time resolved fluorescence tomography of turbid media based on lifetime contrast," *Opt. Express.* **14**(25), 12255–12270 (2006).
8. X. Montet, J. L. Figueiredo, H. Alencar, V. Ntziachristos, U. Mahmood, and R. Weissleder, "Tomographic fluorescence imaging of tumor vascular volume in mice," *Radiology* **242**(3), 751–758 (2007).
9. G. Zacharakis, H. Kambara, H. Shih, J. Ripoll, J. Grimm, Y. Saeki, R. Weissleder, and V. Ntziachristos, "Volumetric tomography of fluorescent proteins through small animals in vivo," *Proc. Natl. Acad. Sci. U.S.A.* **102**(51), 18252–18257 (2005).
10. S. R. Arridge and J. C. Schotland, "Optical tomography: forward and inverse problems," *Inverse Probl.* **25**(12), 1–59 (2009).
11. X. A. Cong and G. Wang, "A finite-element-based reconstrution method for 3-D fluorescence tomography," *Opt. Express* **13**(24), 9847–9857 (2005).
12. B. Ralf Schulz, A. Ale, A. Sarantopoulos, M. Freyer, E. Soehngen, M. Zientkowska, and V. Ntziachristos, "Hybrid System for Simultaneous Fluorescence and X-ray Computed Tomography," *IEEE Trans. Med. Imaging* **29**(2), 465–473 (2009).
13. D. Kepshire, N. Mincu, M. Hutchins, J. Gruber, H. Dehghani, J. Hypnarowski, F. Leblond, M. Khayat, and B. W. Pogue, "A microcomputed tomography guided fluorescence tomography system for small animal molecular imaging," *Rev. Sci. Instrum.* **80**(4), 043701 (2009).
14. Y. Tan and H. Jiang, "Diffuse optical tomography guided quantitative fluorescence molecular tomography," *Appl. Opt.* **47**(12), 2011–2016 (2008).
15. A. Joshi, J. C. Rasmussen, E. M. Sevick-Muraca, T. A. Wareing, and J. McGhee, "Radiative transport-based frequency-domain fluorescence tomography," *Phys. Med. Biol.* **53**(8), 2069–2088 (2008).
16. B. C. Wilson and G. Adam, "A Monte Carlo model for the absorption and flux distributions of light in tissue," *Med. Phys.* **10**(6), 824–830 (1983).
17. T. Li, "MCVM: Monte Carlo Modeling of Photon Migration In Voxelized Media," *J. Innov. Opt. Health Sci.* **3**(2), 91–102 (2010).
18. L. Wang, S. L. Jacques, and L. Zheng, "MCML—Monte Carlo modeling of light transport in multi-layered tissues," *Comput. Methods Programs Biomed.* **47**(2), 131–46 (1995).
19. A. T. N. Kumar, S. B. Raymond, A. K. Dunn, B. J. Bacskai, and D. A. Boas, "A time domain fluorescence tomography system for small animal Imaging," *IEEE Trans. Med. Imaging* **27**(8), 1152–1163 (2008).
20. X. Zhang, C. Badea, M. Jacob, and G. A. Johnson, "Development of a noncontact 3-D fluorescence tomography system for small animal in vivo imaging," *Proc. SPIE* **7191**, 71910D (2009).
21. *NVIDIA CUDA™ Programming Guide.* 2009; 2.3: Available from: http://developer.download.nvidia.com/compute/cuda/2_3/toolkit/docs/NVIDIA_CUDA_Programming_Guide_2.3.pdf.
22. E. Alerstam, T. Svensson, and S. Andersson-Engels, "Parallel computing with graphics processing units for high-speed Monte Carlo simulation of photon migration," *J. Biomed. Opt.* **13**(6), 060504 (2008).
23. W. C. Y. Lo, T. D. Han, J. Rose, and L. Lilge, "GPU-accelerated Monte Carlo simulation for photodynamic therapy treatment planning," *Proc. SPIE* **7373**, 737313 (2009).
24. Q. Q. Fang and D. A. Boas, "Monte Carlo Simulation of Photon Migration in 3-D Turbid Media Accelerated by Graphics Processing Units," *Opt. Express.* **17**(22), 20178–20190 (2009).
25. Q. Fang, *Monte Carlo eXtreme (MCX)*, 2009; 0.2: Available from: http://mcx.sourceforge.net/cgi-bin/index.cgi?Home.
26. J. F. Beek, P. Blokland, P. Posthumus, M. Aalders, J. W. Pickering, H. J. C. M. Sterenborg, and M. J. C. vanGemert, "In vitro double-integrating-sphere optical properties of tissues between 630 and 1064 nm," *Phys. Med. Biol.* **42**(11), 2255–2261 (1997).
27. S. R. Arridge, "Optical tomography in medical imaging," *Inverse Probl.* **15**(2), R41–R93 (1999).
28. S. R. Arridge and M. Schweiger, "Photon-measurement density functions. Part 2: Finite-element-method calculations," *Appl. Opt.* **34**(34), 8026–8037 (1995).
29. B. W. Pogue, T. O. McBride, J. Prewitt, U. L. Osterberg, and K. D. Paulsen, "Spatially variant regularization improves diffuse optical tomography," *Appl. Opt.* **38**(13), 2950–2961 (1999).
30. A. N. Tikhonov, A. S. Leonov, and A. G. Yagola, *Nonlinear Ill-Posed Problems (Applied Mathematics & Mathematical Computation)*, Springer, Berlin (1997).

31. D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, "Tikhonov regularization and the L-curve for large discrete ill-posed problems," *J. Comput. Appl. Math.* **123**(1–2), 423–446 (2000).
32. T. Rubaek, P. M. Meaney, P. Meincke, and K. D. Paulsen, "Nonlinear microwave imaging for breast-cancer screening using Gauss-Newton's method and the CGLS inversion algorithm," *IEEE Trans. Antenn. Propag.* **55**(8), 2320–2331 (2007).
33. Z. Zhang, W. Cao, H. Jin, J. F. Lovell, M. Yang, L. Ding, J. Chen, I. Corbin, Q. Luo, and G. Zheng, "Biomimetic Nanocarrier for Direct Cytosolic Drug Delivery," *Angew. Chem. Int. Ed.* **48**(48), 9171–9175 (2009).
34. X. Yang, H. Gong, G. Quan, Y. Deng, and Q. Luo, "Combined system of fluorescence diffuse optical tomography and micro-CT for small animal imaging," *Rev. Sci. Instrum.* **81**(5), 054304 (2010).
35. E. E. Graves, J. Ripoll, R. Weissleder, and V. Ntziachristos, "A submillimeter resolution fluorescence molecular imaging system for small animal imaging," *Med. Phys.* **30**(5), 901–11 (2003).
36. H. Dehghani, *NIRFAST.* 2008; Available from: http://www.dartmouth.edu/~nir/nirfast/index.php.
37. H. Dehghani and D. T. Delpy, "Near-infrared spectroscopy of the adult head: effect of scattering and absorbing obstructions in the cerebrospinal fluid layer on light distribution in the tissue," *Appl. Opt.* **39**(25), 4721–9 (2000).
38. M. Schweiger and S. Arridge, *TOAST.* 2008; Available from: http://web4.cs.ucl.ac.uk/research/vis/toast/license.html.
39. D. A. Boas, J. P. Culver, J. J. Stott, and A. K. Dunn, "Three dimensional Monte Carlo code for photon migration through complex heterogeneous media including the adult human head," *Opt. Express.* **10**(3), 159–170 (2002).
40. *Comparison of Nvidia graphics processing units.* Available from: http://en.wikipedia.org/wiki/Comparison_of_Nvidia_graphics_processing_units.