Journal of Innovative Optical Health Sciences
Vol. 7, No. 3 (2014) 1450010 (6 pages)
© The Authors
DOI: 10.1142/S1793545814500102

World Scientific
www.worldscientific.com

Acceleration of optical coherence tomography signal processing by multi-graphics processing units

Xiqi Li*,†, Guohua Shi*,†, Ping Huang‡
and Yudong Zhang*,†,§
*The Key Laboratory on Adaptive Optics
Chinese Academy of Sciences
Chengdu 610209, P. R. China
†Institute of Optics and Electronics
Chinese Academy of Sciences
Chengdu 610209, P. R. China
‡Department of Ophthalmology
Peking University Third Hospital
Beijing 100191, P. R. China
§ydzh_ioe@163.com

Received 12 September 2013
Accepted 6 November 2013
Published 19 December 2013

A multi-GPU system designed for high-speed, real-time signal processing of optical coherence tomography (OCT) is described herein. For the OCT data sampled in linear wave numbers, the maximum processing rates reached 2.95 MHz for 1024-OCT and 1.96 MHz for 2048-OCT. Data sampled using linear wavelengths were re-sampled using a time-domain interpolation method and zero-padding interpolation method to improve image quality. The maximum processing rates for 1024-OCT reached 2.16 MHz for the time-domain method and 1.26 MHz for the zero-padding method. The maximum processing rates for 2048-OCT reached 1.58 MHz, and 0.68 MHz, respectively. This method is capable of high-speed, real-time processing for OCT systems.

Keywords: Optical coherence tomography; real-time signal processing; multi graphics processing units.

J. Innov. Opt. Health Sci. 2014.07. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

1. Introduction

Optical coherence tomography (OCT) is a non-invasive, cross sectional optical imaging technique that allows for high-resolution cross sectional tomography imaging of the internal microstructure in materials and biological systems involving back scattered or back-reflected samples.[1] Imaging speed has been a hotspot in OCT research. The acquisition speed of OCT has increased tremendously

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 3.0 (CC-BY) License. Further distribution of this work is permitted, provided the original work is properly cited.

J. Innov. Opt. Health Sci. 2014.07. Downloaded from www.worldscientific.com by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

since the development of Fourier domain OCT. High-speed spectral domain optical coherence for retinal imaging at 500 kHZ was achieved using a dual camera configuration.[2] Multimegahertz Fourier domain mode locking (FDML) based swept source (SS) OCT systems capable of acquiring high-resolution volumes at video-level speeds have been demonstrated, and 1.6 MHz systems have been presented. These systems have been used for clinical retinal imaging.[3] A 5 MHz OCT using an all-optical swept-source based on amplified dispersive Fourier transform (ADFT) and a 90 MHz OCT based on photonic time-stretch have been reported.[4,5] As the ultrahigh-speed OCT has extended, the demand for real-time signal processing of OCT data has increased. This is largely because of the interest in exploring the full potential of the technology. Recent advances in graphics processing units (GPUs) have provided a new means of parallel computation. The computation capability of the GPU is now much greater than that of the center processing units (CPU). Powerful GPUs can be handled easily using a compute unified device architecture (CUDA) program model.[6] The processing time of the FD-OCT system can be accelerated using this tool.

A real-time 4D signal processing and visualization process was performed using GPUs on a regular nonlinear-k Fourier domain OCT system. This previous team proposed a GPU accelerated linear spline interpolation (LSI) for λ-to-k re-sampling.[7] The maximum complete A-scan processing speeds were found to be 680 kHz for 1024 pixel OCT and 320 kHz for 2048-OCT. A new GPU-accelerated processing system showed an A-scan (16-bit, 2048 pixels) rate of > 2.24 MHz for a linear wave number output SS-OCT system and about 1.2 MHz for a nonlinear wave number output SD-OCT system.[8] Systems described in previous reports have reached considerable processing speeds, but there remain a few problems to be solved. Because the output of many swept sources of SS-OCT and the spectral of the SD-OCT are not consistent with linear wave numbers, so a re-sampling is needed. Because the k and λ values have an inverse relationship, specifically $k = 2\pi/\lambda$ and $\delta k = 2\pi\delta\lambda/\lambda^2$, the re-sampling must be performed nonlinearly. For this reason, zero-padding interpolation and with linear interpolation are commonly used in SD-OCT data re-sampling processes. This improves the signal to noise ratio (SNR).[9] A real-time processing system with zero-padding interpolation using multiple GPU has been proposed.[10] The computing speeds of 2048-OCT were found to be 38 and 69 kHz for single and dual GPUs, respectively. A real-time processing system with two GPUs has been reported.[11] The processing rate of nonuniform FFT (NUFFT) was not high enough because of the complexity of the NUFFT. A time-domain interpolation based on zero-padding has been reported.[12] This method can reduce processing time considerably and improve the SNR, as compared to the more commonly used zero-padding method.[12,13] Furthermore, the time-domain interpolation is suitable for parallel computing.[12–14] The data processing speed can reach 160 kHz for 1024-OCT and 70 kHz for 2048-OCT by a single GPU.[14]

In order to increase the processing rate, a real-time system with multiple GPUs was designed for real-time OCT signal processing.

2. Methods

An AMAX server was used to exploit the massive parallel computational power of the GPU. In this server, a Windows 7 operating system was run on two Intel Westmere X5690 CPUs, and four NVIDIA Tesla C2075 GPUs were placed in the PCI Express 2.0 slots. In this system, one GPU focused on image display and the other three were used as co-processors for signal processing. Figure 1 shows the setup of the system. The OCT signal processing project was built and compiled using Microsoft Visual Studio 2008 and CUDA 5.0.

Because no megahertz line rate OCT system was available, data acquired using a low-speed OCT system were used to evaluate the performance of the multi-GPU system. The data acquired in this way were set to 1024 pixels per A-scan and 2048 pixels per A-scan. They are here called 1024-OCT and

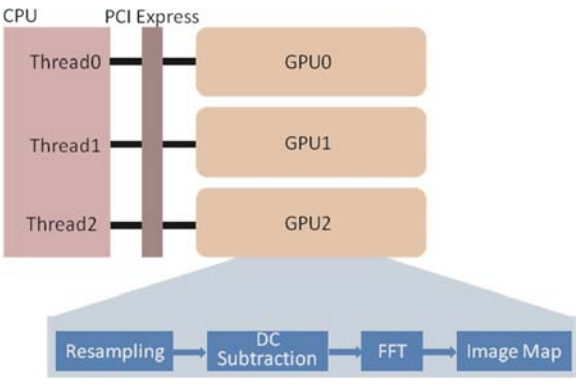

Fig. 1. Flowchart of multi-GPU signal processing.

J. Innov. Opt. Health Sci. 2014.07. Downloaded from www.worldscientific.com by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 1. Host to device bandwidth.

	GPU0	GPU1	GPU2	GPU0: GPU1	GPU0: GPU2
Bandwidth (GB/s)	4965.3	4961	4963.6	9929.2	14892.8

2048-OCT. The data width of both systems was 16 bit per pixel.

The system's bandwidth was test by the NVIDIA bandwidth test program in this setup. In this program, the host memory was allocated in the PINNED memory and transferred to the device memory in 32MB per batch. Table 1 shows the results. The results show that the bandwidth of the GPUs are almost the same, but the bandwidth would increase as the GPU increase. The data transfer rate is about 7.5 MHz for 1024-OCT and 3.7 MHz for 2048-OCT by using three GPUs. The processing speed should be increased until it can keep up with the transfer rate.

In order to make full use of the GPUs, the parallel programming language OpenMP was used to schedule the GPUs. Figure 1 shows the schedule. The program created three threads to control and schedule the three GPUs one by one. Using the schedule of the multiple threads, the acquired data were separated into three parts and then transferred to a different GPU. A multi-stream method was adapted to fully utilize the bandwidth of the PCI Express bus and reduce the time required for the transfer. Then the data were processed in parallel using these GPUs. Every GPU performed re-sampling, DC subtraction, the fast fourier transform (FFT) and image creation. The DC signal was obtained by averaging the data in every GPU to reduce the memory interactions between the GPUs. It can be seen as three different frames were processed. However, the final image was merged into a large field of view.

Three methods were used on the GPUs. The first was the Even k method. This method supposes that the data acquired are sampled linearly with respect to wave number. In this way, the image can be obtained by DC subtraction, FFT and direct image mapping. The second and the third methods suppose that the acquired data are sampled linearly with respect to wavelength. In this case, a nonlinear interpolation is needed to increase image quality. A time-domain interpolation method and zero-padding interpolation method were performed to re-sample the acquired data. In the time-domain interpolation method, a coefficient was calculated before the processing, and the re-sampled data were obtained by convoluting the coefficient and the signal data. In the zero-padding interpolation method, the padding time was set to four. This means that every A-scan signal was padded to four times its previous length after a forward FFT. This padding time struck a balance between image quality and processing speed. Then, an inverse FFT was performed to obtain the extended data. Finally, the re-sampled data were acquired through linear interpolation.

3. Results and Discussion

Then 1024-OCT and 2048-OCT were stimulated in order to show the computing capability of the multi-GPUs. All acquired volume data were divided into three batches for transfer and processing in the three GPUs. After the data were transferred to the GPUs, data processing was trigged immediately. The processing time was determined by noting the times at which the process began and ended.

At the beginning of the comparison process, the Even k, zero-padding and time-domain interpolation method was performed in a single GPU in both the 1024-OCT and 2048-OCT systems. Figure 2 shows the results.

Figure 2 shows that the computing capability increases as the batch lines increase. The processing rate of Even k@1024-OCT plateaued around 3.6 K and 4.2 K A-scans per batch. The plateau was observed when the stream processor was fully utilized and the latency was hidden by the pipeline of the stream processor of the GPU. Figure 3 shows the relationship between grid size and the time required for the background subtraction program. In this

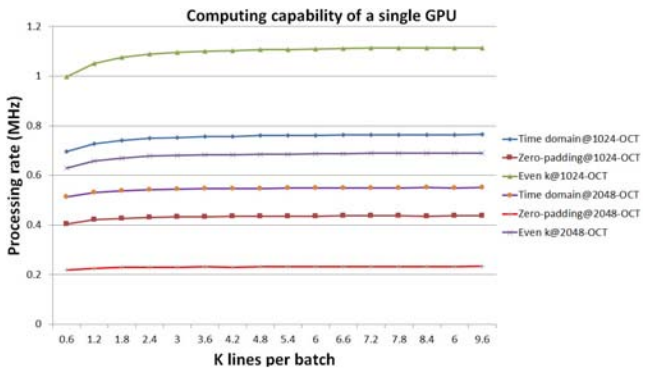

Fig. 2. Computing capability of a single GPU.

J. Innov. Opt. Health Sci. 2014.07. Downloaded from www.worldscientific.com
by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

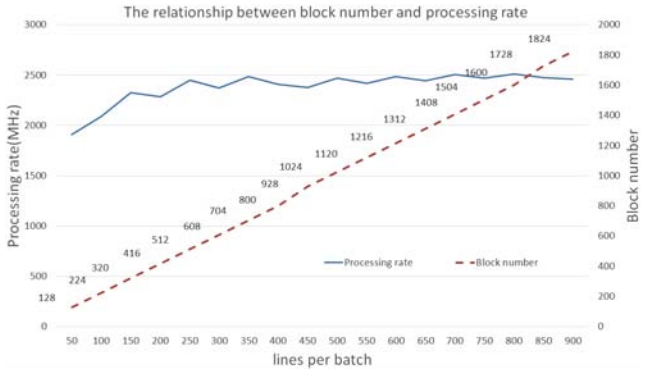

Fig. 3. Grid size and processing rate.

program, the acquired data were subtracted from the averaged data to produce the DC subtracted data. The block size was set to 32×32 to make full use of the single stream processing capability. The grid size was set by the size of the data height and the data width. As shown, the block number increased as the number of batch lines increased, but the processing rate plateaued around 250 lines per batch. This is because the computing capability was not fully utilized before that point. However, at that point, the grid size was 512, which was slightly greater than that of the GPU's stream processor. In this way, the threads can fully occupy the GPU and allow full use of the computing capability of the GPU.

By using a single GPU, the computing capability of the Even k reached 1.114 and 0.69 MHz for 1024-OCT and 2048-OCT, respectively. For the noneven k OCT systems, an additional re-sampling step was needed. The zero-padding interpolation and time-domain interpolation methods reached 0.437 and 0.766 MHz, respectively, for the 1024-OCT system. In the 2048-OCT system, these two methods were found to reach 0.23 and 0.55 MHz, respectively.

The processing rate of GPUs in groups of two and three was also estimated. Figure 4 shows the results, which indicate that these groups of GPUs exhibited higher processing rates than single GPUs. They also had plateau nodes when the batch increased, but these plateau points were larger than those of single GPUs. This is because the computing resources increased as the number of GPUs increased, requiring more lines per batch for full utilization of resources. Table 2 shows the plateaus of computing of the multi-GPU systems. The computing rates of the group of three GPUs reached 2.95 MHz for the Even k OCT setup. The computing rate of the zero-padding and time domain reached 1.26 and 2.16 MHz for the noneven k OCT setup, respectively.

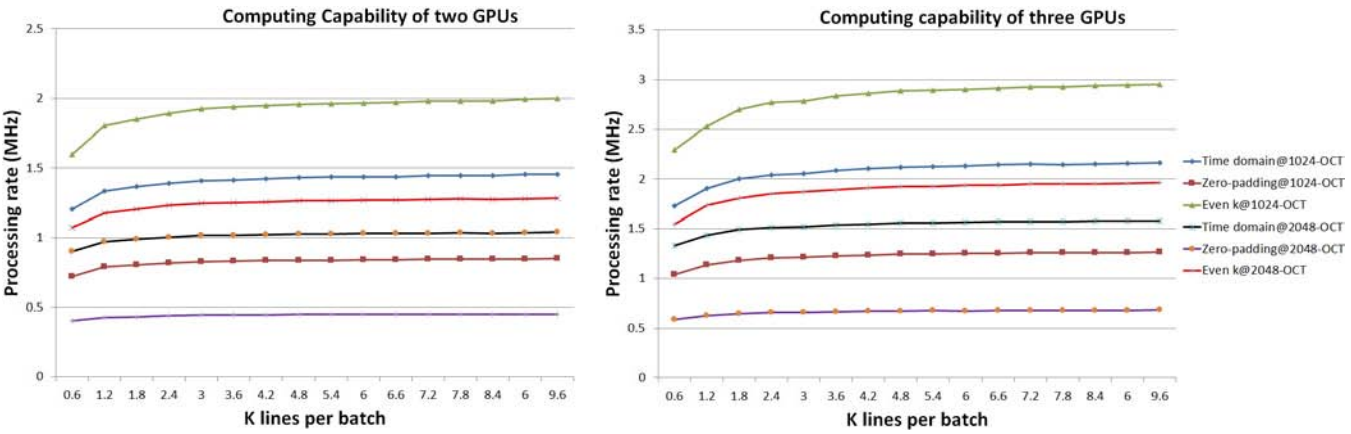

Fig. 4. Computing capability of two GPUs and three GPUs.

Table 2. Computing capability of multi-GPU systems.

	Even k		Zero-padding		Time domain	
	1024-OCT	2048-OCT	1024-OCT	2048-OCT	1024-OCT	2048-OCT
Single GPU	1.114 MHz	0.69 MHz,	0.437 MHz	0.23 MHz	0.766 MHz	0.55 MHz
Two GPUs	2.0 MHz	1.28 MHz	0.85 MHz	0.45 MHz	1.45 MHz	1.04 MHz
Three GPUs	2.95 MHz	1.96 MHz	1.26 MHz	0.68 MHz	2.16 MHz	1.58 MHz

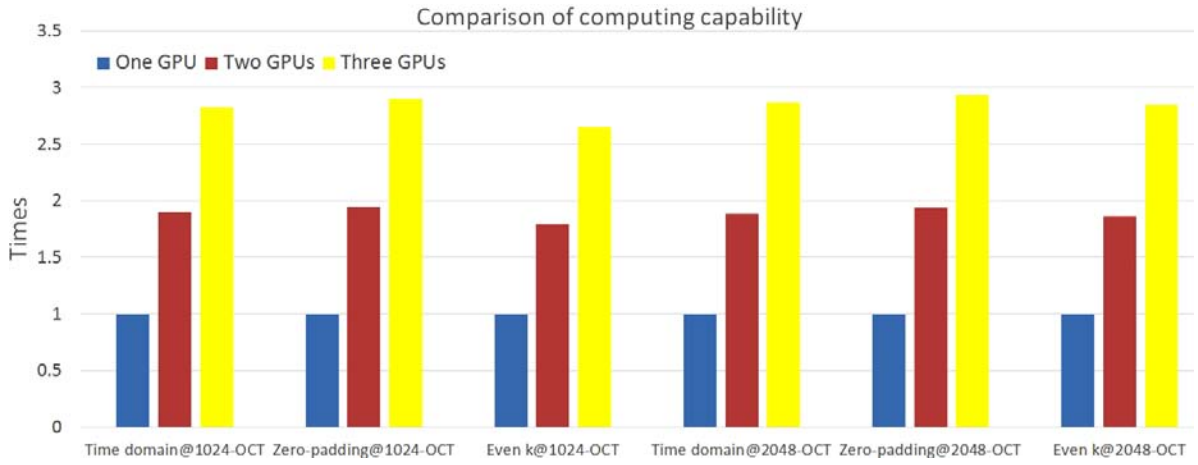

Fig. 5. Comparison of computing capability.

As indicated by the results described above, processing rates increase as the number of A-scans per batch increases. This indicates that, as the number of A-scans per batch increases, the GPUs utilization becomes more efficient. These results also show that the more GPUs there are, the greater the computing capability reached. Figure 5 shows a comparison of the capabilities of systems with different numbers of GPUs. In this figure, the computing capability of a single GPU served as the basis of computing capability of the three methods. The processing rates of systems with two and three GPUs processing rates were compared to this. The results showed that, as the number of GPUs increased, the computing capability also increased. However, this increase was not simply additive. This is because the schedules of the multi-GPU systems consumed additional resources and time. Different methods also showed increases of different magnitude. This was because different methods have different numbers of steps. An efficient, optimized model is needed.

The results show that the systems involving multiple GPUs had greater computing capability than the single-GPU system, but it is not possible to incorporate an unlimited number of GPUs into a single system. Another way to accelerate the computing capability is to use the latest and most advanced GPUs. The GPUs were NVIDIA Tesla C2075, which is not the most powerful GPU available. Every C2075 had a 448 stream processor, and its peak single-precision floating point performance was 1.03 T Flops. One previous study shows that a single GPU can process 2.24 MHz for 1024 even k setup.[8] The GPU used in this study was a GTX680, which has a 1536 stream processor. Its peak single-precision floating point performance is 3.09 T Flops. The latest Tesla K10 can reach 4.58T Flops. Because the computing capability of the GPUs increases in a linear fashion, about 10 MHz of computing capability can be expected with three GPUs. The latest GPUs have been updated to a PCI Express 3.0. A higher bandwidth can be obtained and the transfer speed can be increased.

4. Conclusion

In conclusion, this report displays the high speed of a multi-GPU system during OCT signal processing. The results show that the multi-GPUs system has bigger computing capability than a single GPU system. The computing resource increases as the GPUs increase, making a sustained A-scan processing rate of 2.95 MHz for even k signal processes which was recorded. Time-domain interpolation and zero-padding interpolation methods were used for uneven k OCT. Their processing rates reached 2.16 and 1.26 MHz, respectively. More and faster GPUs might facilitate even faster speeds. The multi-GPU systems may provide a more bandwidth than single-GPU systems, but they require complicated topology to make use of this advantage. Further optimization is needed.

Acknowledgments

We acknowledge the support from the union project of Peking University third hospital & Chinese Academy of Sciences (Grant No. 74490-04, Grant

J. Innov. Opt. Health Sci. 2014.07. Downloaded from www.worldscientific.com by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.

No. KJZD-EW-TZ-L03), the Sichuan Youth Science & Technology Foundation (Grant No. 13QNJJ0034), the West Light Foundation of the Chinese Academy of Sciences, the National Major Scientific Equipment program (Grant No. 2012YQ120080) and the National Science Foundation of China (Grant No. 6118082).

References

1. D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Pulifito, "Optical coherence tomography," *Science* **254**, 1178–1181 (1991).
2. L. An *et al.*, "High speed spectral domain optical coherence tomography for retinal imaging at 500,000 A-lines per second," *Biomed. Opt. Express* **2** (10), 2770–2783 (2011).
3. C. Blatter *et al.*, "Ultrahigh-speed non-invasive widefield angiography," *J. Biomed. Opt.* **17**(7), 070505 (2012).
4. J. Xu, C. Zhang, J. Xu *et al.*, 5 MHz all-optical swept-source coherence tomography based on amplified dispersive Fourier transform, *Novel Techniques in Microscopy*, Waikoloa Beach, Hawaii, US, NW5B (2013).
5. K. Goda, A. Fard, O. Malik *et al.*, "High-throughput optical coherence tomography at 800 nm," *Opt. Express* **20**(18), 19612–19614 (2012).
6. NVIDIA CUDA. Available at http://developer.nvidia.com/object/cuda.html.
7. K. Zhang, J. U. Kang, "Real-time 4D signal processing and visualization using graphics processing unit on a regular nonlinear-k Fourier-domain OCT system," *Opt. Express* **18**, 11772–11784 (2010).
8. Y. Jian, K. Wong, M. V. Sarunic, "Graphics processing unit accelerated optical coherence tomography processing at megahertz axial scan rate and high resolution video rate volumetric rendering," *J. Biomed. Opt.* **18**(2), 026002 (2013).
9. C. Dorrer, N. Belabas, J.-P. Likforman, M. Joffre, "Spectral resolution and sampling issues in Fourier-transform spectral interferometry," *J. Opt. Soc. Am. B* **17**, 1795–1802 (2000).
10. Y. Watanabe, S. Maeno, K. Aoshima, H. Hasegawa, H. Koseki, "Real-time processing for full-range Fourier-domain optical-coherence tomography with zero-filling interpolation using multiple graphic processing units," *Appl. Opt.* **49**(25), 4756–4762 (2010).
11. K. Zhang, J. U. Kang, "Graphics processing unit-based ultrahigh speed real-time fourier domain optical coherence tomography," *IEEE. Sel. Top.* **8**(4), 1270–1279 (2012).
12. Y. Zhang, X. Li, L. Wei, K. Wang, Z. Ding, G. Shi, "Time-domain interpolation for Fourier-domain optical coherence tomography," *Opt. Lett.* **34**, 18490–1851 (2009).
13. X. Li, G. Shi *et al.*, "Time-domain interpolation on graphics processing unit," *J. Innov. Opt. Health Sci.* **4**(01), 89–95 (2011).
14. X. Li, G. Shi *et al.*, "High-speed spectral domain optical coherence tomography signal processing with time-domain interpolation using graphics processing unit," *J. Innov. Opt. Health Sci.* **4**(03), 325–335 (2011).

J. Innov. Opt. Health Sci. 2014.07. Downloaded from www.worldscientific.com by 103.240.126.9 on 10/21/18. Re-use and distribution is strictly not permitted, except for Open Access articles.