2017

A Highly Accelerated Parallel Multi-GPU based Reconstruction Algorithm for Generating Accurate Relative Stopping Powers

Paniz Karbasi
Baylor University

Ritchie Cai
Baylor University

Blake Schultze
Baylor University

Hanh Nguyen
Baylor University

Jones Reed
Baylor University

See next page for additional authors

A Highly Accelerated Parallel Multi-GPU based Reconstruction Algorithm for Generating Accurate Relative Stopping Powers

Abstract

Low-dose Proton Computed Tomography (pCT) is an evolving imaging modality that is used in proton therapy planning which addresses the range uncertainty problem. The goal of pCT is generating a 3D map of relative stopping power measurements with high accuracy within clinically required time frames. Generating accurate relative stopping power values within the shortest amount of time is considered a key goal when developing an image reconstruction software. The existing image reconstruction softwares have successfully met this time frame and even exceeded this time goal, but require clusters with hundreds of processors. This paper describes a novel reconstruction technique using two graphics processing unit devices. The proposed reconstruction technique is tested on both simulated and experimental datasets and on two different systems namely Nvidia K40 and P100 graphics processing units from IBM and Cray. The experimental results demonstrate that our proposed reconstruction method meets both the timing and accuracy with the benefit of having reasonable cost and efficient use of power.

Disciplines

Engineering | Science and Technology Studies

Publication Details

Authors

Paniz Karbasi, Ritchie Cai, Blake Schultze, Hanh Nguyen, Jones Reed, Patrick Hall, Valentina Giacometti, Vladimir Bashkirov, Robert P. Johnson, Nick Karonis, Jeffrey Olafsen, Caesar Ordonez, Keith Schubert, and Reinhard W. Schulte

A Highly Accelerated Parallel Multi-GPU based Reconstruction Algorithm for Generating Accurate Relative Stopping Powers

Paniz Karbasi[1], Ritchie Cai[1], Blake Schultze[1], Hanh Nguyen[1], Jones Reed[1], Patrick Hall[1],
Valentina Giacometti[2],[6], Vladimir Bashkirov[6], Robert Johnson[3], Nick Karonis[4] Jeffrey Olafsen[5],
Caesar Ordonez[4] Keith E. Schubert[1],[6] *Senior Member, IEEE*, Reinhard W. Schulte[6] *Member, IEEE*

Abstract—Low-dose Proton Computed Tomography (pCT) is an evolving imaging modality that is used in proton therapy planning which addresses the range uncertainty problem. The goal of pCT is generating a 3D map of relative stopping power measurements with high accuracy within clinically required time frames. Generating accurate relative stopping power values within the shortest amount of time is considered a key goal when developing an image reconstruction software. The existing image reconstruction softwares have successfully met this time frame and even exceeded this time goal, but require clusters with hundreds of processors. This paper describes a novel reconstruction technique using two graphics processing unit devices. The proposed reconstruction technique is tested on both simulated and experimental datasets and on two different systems namely Nvidia K40 and P100 graphics processing units from IBM and Cray. The experimental results demonstrate that our proposed reconstruction method meets both the timing and accuracy with the benefit of having reasonable cost and efficient use of power.

[1] Paniz Karbasi, Ritchie Cai, Blake Schultze, Hanh Nguyen, Jones Reed, Patrick Hall and Keith Schubert are with the Department of Electrical and Computer Engineering, Baylor University, Waco, TX 76798 USA, email: Paniz_Karbasi@baylor.edu, Ritchie_Cai@baylor.edu, Blake_Schultze@baylor.edu, Hanh_Nguyen@baylor.edu, Jones_Reed@baylor.edu, Patrick_Hall@baylor.edu, Keith_Schubert@baylor.edu

[2] Valentina Giacometti is with the Center for Medical Radiation Physics, University of Wollongong, Wollongong, NSW, Australia, email: valentina8giacometti@gmail.com

[3] Robert Johnson is with the Santa Cruz Institute for Particle Physics, University of California, Santa Cruz, Santa Cruz, CA 95064, USA, email: rjohnson@ucsc.edu

[4] Nick Karonis and Caesar Ordonez are with the Department of Computer Science, Northern Illinois University, DeKalb, IL 60115 email: karonis@niu.edu, cordonez@cs.niu.edu

[5] Jeffrey Olafsen is with the Department of Physics, Baylor University, Waco, TX 76798 USA, email: Jeffrey_Olafsen@baylor.edu

[6] Valentina Giacometti, Vladimir Bashkirov (vabashkirov@llu.edu), Keith Schubert, and Reinhard Schulte (rschulte@llu.edu), are with Loma Linda University, Loma Linda, CA 92350

I. Introduction

Proton Computed Tomography (pCT) is a growing imaging technology in proton therapy planning. By addressing the range uncertainty problem, pCT images suggest more accurate treatment plans than X-ray CT images [1]. The pCT collaboration has developed a proton CT scanner including a silicon-based tracking system and a multi-stage scintillating energy detector for measuring the water equivalent path length (WEPL) of individual protons [2]. By using low-dose proton emission, calculating the individual proton's most likely path [3], and knowing their energy loss, large and sparse linear system of equations $Ax = b$ can be written where A is a $m \times n$ matrix containing the path data, b is the $m \times 1$ WEPLs vector, and x is the $n \times 1$ relative stopping powers (RSP) vector. Using the FBP image as the initial iterate, one can iteratively solve the system $Ax = b$ for generating the 3D map of the RSPs to be used in the treatment planning.

Various concepts ranging from advanced detector designs to appropriately selecting the preprocessing techniques and parameters of the reconstruction algorithm have a major impact on the quality and accuracy of the pCT images [4], [5]. Calculating accurate RSPs is of great importance, but to take advantage of the pCT systems in a clinical setup, there is also a great need for real-time algorithms that can process hundreds of millions of protons in concise time frames not more than a few minutes.

The pCT problem fits well within the Single Instruction Multiple Threads (SIMT) parallel programming paradigm of Graphics Processing Units (GPUs) since we treat each proton as an individual thread that can be processed in parallel. Although there is a great need for small pCT runs as a validation technique in clinics, when it comes to imaging an adult torso, because of the greatly increased problem size, the reconstruction time grows proportional to the four thirds power of the reconstructed object's volume. In order to achieve the time frames that are clinically meaningful (i.e. in under 5 minutes), we have proposed a reconstruction technique which takes advantage of systems with at least 2 GPUs and generates the 3D map of highly accurate RSP values within 40 seconds for 116 million and 79 seconds for 261 million proton histories. Moreover, based on the experimental results, our proposed reconstruction algorithm runs faster than the ex-

pected speedup on the Nvidia K40 GPUs which demonstrates it is a reasonable and economical alternative for the clinical pCT systems.

II. RELATED WORK

The rapid evolution of GPUs in the recent years has contributed to design and implementation of accelerated processing algorithms of many real world applications. Real-time reconstruction of the 3D maps of RSP measurements is a key factor that needs to be met in pCT imaging. In [6] it has been demonstrated that the reconstruction time can be reduced from 7 hours on a single machine to 53 seconds using a GPU cluster for a dataset of size 131 million protons. In [7] it has been further demonstrated that image reconstruction in pCT can be accelerated through a hybrid approach that uses both Message Passing Interface (MPI) and GPUs. Using this approach on the same cluster the reconstruction rumtime has been improved and reduced to 43 seconds for a similar size of dataset. In a recent study in [8], it is shown using the same approach as the first evaluation of the pCT software in [7], the execution time for generating accurate RSP values for a dataset of size 131 million protons is almost 30 seconds running on 60 processors (60 CPU cores + 60 GPUs). The method we have proposed in this paper, removes the cost of having a GPU cluster, and generates accurate RSPs with only a single computer and two P100 GPUs thus reducing the time to 40 seconds for a simulated CTP404 dataset of size 116 million protons.
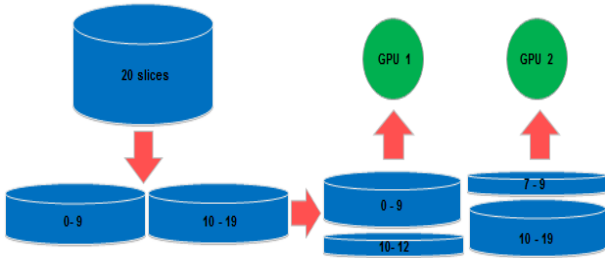

Fig. 1: Schematic drawing of Algorithm 1 with two GPUs

III. METHODOLOGY

In order to benefit from the Nvidia Tesla GPUs, we have developed a fast and easy to implement reconstruction algorithm that can run on systems with at least 2 GPUs with the advantage of eliminating the need of data transfers among the GPUs. The algorithm we have designed relies on the following two facts:

1) only a small fraction of protons intersect multiple slices along the vertical axis
2) within a slice, only a small portion of protons pass through and are used in reconstruction of that specific slice

Based on the first observation which is reported with details in Table I, we know that a proton's path does not have a significant deviation along the z axis or vertical slices of the reconstruction volume. This enables us to consider

an algorithm without having a significant concern about the GPU-to-GPU data transfers which could potentially add some considerable timing constrains.

The second observation allows us to split the protons among the available GPUs such that each GPU reconstructs a portion of the 3D image. Each GPU only needs some fraction of the protons and this lowers the amount of required memory and processing time per GPU leading to a faster reconstruction time overall.

The general structure of the proposed multi-GPU based reconstruction technique can be seen in Algorithm 1. The first step is the division of the reconstruction volume into several overlapping regions based on the number of available GPUs, while each region is assigned to a unique key. The overlap is set at twice the slices for a 96% coverage (Table I). In the second step, each proton is associated with the image region entered to and exited from. In the third step, based on the number of protons that pass through each image region, the required memory for each proton is allocated on each GPU. Lastly (fourth step), we perform a standard iterative solver [9] on each GPU. Finally, in the overlap of two regions, where there are two reconstructions of each slice, we select the reconstruction that is closest to the non-overlapped slices of its region. For example, in the reconstruction presented in this work, there are two regions, one containing slices $0-12$ and the other containing slices $7-19$. The overlap corresponds to slices $7-12$, and the final image will be made of slices $0-9$ from the first region and $10-19$ of the second. The high level description of Algorithm 1 is illustrated in Fig. 1.

Algorithm 1 multi-GPU based reconstruction algorithm

---

procedure MULTI-GPU BASED RECONSTRUCTION
    1: setupImageRegions()
    2: identifyProtonsPassingRegions()
    3: allocateMemoryForProtonsPerGPU()
    4: iterativelySolvePerGPU()
    5: selectSlice()

IV. EXPERIMENTS AND RESULTS

In this section, we compare the runtime of the proposed multi-GPU based reconstruction technique on three different systems using single and double GPUs, and compare the accuracy of the reconstructed RSP values generated by Algorithm 1 and the standard single GPU technique.

TABLE I: Percentage of protons passing through image slices along the z axis for the simulated and experimental CTP404 phantom with the 0.25 centimeters slice thickness

# of slices	Simulated CTP404	Experimental CTP404
1	32	42
2	43	36
3	19	15
4	4	3
≥ 5	2	4

A. Systems and Datasets

In order to analyze the performance of the proposed algorithm discussed in the previous section, we have tested the algorithm on three different platforms: 2 Nvidia K40s on a Xeon, Cray's 2 P100 GPUs, and IBM's P100 GPU.

Angle intervals of 4-degree per projection (90 projections for a full rotation scan) and continuous angle increments were used for the reconstruction of simulated and experimental CTP404 phantom datasets, respectively. Also, both of these datasets are composed of 20 slices in the vertical direction while the slice thickness is 0.25 cm.

The simulated data were obtained using the Geant4 based software simulation platform described in [10]. The simulated and experimental CTP404 data are composed of 116 million and 261 million proton histories, respectively, later reduced using data cleaning techniques (e.g. identifying and removing statistical outliers). The actual number of protons used by the iterative solver is 21 and 78 million protons for the simulated and experimental data respectively. When using the proposed reconstruction algorithm described in Algorithm 1, there are different number of protons passing through each image region which are reported in Table II.

Based on the results in Table II, protons are evenly distributed between the two image regions for the simulated data, while the top image region of the experimental data compasses twice the number of protons that pass through the bottom image region. Execution times reported in the next section heavily depend on the number of protons traversing the image regions and the distribution of data.

TABLE II: Number of protons (millions) in each image region used by the GPUs to perform the iterative solver

Data	Slices 0-12	Slices 7-19
Simulated CTP404	14	14
Experimental CTP404	62	33

B. Performance and Execution Time

In order to test the performance of our proposed reconstruction algorithm, we used two different block sizes 3200 and 320000 for the experimental data and 320000 for the simulated data. The block size is the number of protons to be processed in parallel based on the block iterative technique Diagonally-Relaxed Orthogonal Projections, (DROP [11]). When it comes to the performance of the iterative solver, in general, larger block sizes decrease the runtime of a single iteration but not necessarily generate accurate RSPs. Here we have reported the timings of both 3200 and 320000 for the experimental data for comparison purpose (see Tables V and VII).

Based on the timings in Table III, the performance of Algorithm 1 is about 1.75 times faster for the simulated data on K40 while based on Table V, the iterative solver with two GPUs is 1.3 times faster than the single-GPU iterative solver for the experimental data. The reason comes from the fact that the number of protons used by the top image region (slices $0-12$) is only 1.25 times less than the total number of protons used by the single-GPU iterative solver for experimental data

(Table II) and since slice number 9 comes from the top image region with 62 million protons, there is not a significant performance improvement with the iterative solver using two GPUs on the experimental data.

An important observation regarding the timings in Table III is that on P100 system, the runtime of the iterative solver is about 1.53 times faster with the proposed method and is similar to the expected speedup $\frac{21}{14}$ or 1.5. On the other hand, the speedup of the proposed method on K40 system is about 1.75, which is greater than the expected one. The reason comes from the difference between the memory sizes of K40 and P100 GPUs. In fact, the smaller memory of the K40 system in comparison to P100, leads to efficient usage of the cache on K40 leading to a faster runtime of a single iteration. The total runtime of the pCT software including the data reads and data cleaning techniques is reported in Tables IV and VI for the simulated and experimental data respectively.

TABLE III: Runtime (second) comparison of Algorithm 1 (step 4) with 1 iteration of the single-GPU iterative solver on simulated data with block size of 320000

	Current Algorithm	Proposed Algorithm
K40	13.3	7.6
P100	2.3 (IBM & Cray)	1.5 (Cray)

TABLE IV: Total reconstruction runtime (second) of simulated data converging after 5 and 3 iterations (Table III) of the iterative solver for single-GPU and Algorithm 1, respectively

	Current Algorithm	Proposed Algorithm
K40	154	96
P100	45 (IBM & Cray)	40 (Cray)

TABLE V: Runtime (second) comparison of Algorithm 1 (step 4) with 1 iteration of the single-GPU iterative solver on experimental data with block size of 3200

	Current Algorithm	Proposed Algorithm
K40	79.7	60.8
P100	10.2 (IBM & Cray)	7.1 (Cray)

TABLE VI: Total reconstruction runtime (second) of experimental data converging after 5 and 3 iterations (Table V) of iterative solver for single-GPU and Algorithm 1, respectively

	Current Algorithm	Proposed Algorithm
K40	521	311
P100	106 (IBM & Cray)	79 (Cray)

TABLE VII: Runtime (second) comparison of Algorithm 1 (step 4) with 1 iteration of the single-GPU iterative solver on experimental data with block size of 320000

	Current Algorithm	Proposed Algorithm
K40	36.8	31.2
P100	4.9 (IBM & Cray)	3.6 (Cray)

C. Accuracy and Image Quality

Simulated and experimental reconstructed images are shown in Fig. 2(a) and 2(b), respectively. The slices here presented, belong to the top image region (slice number 9 out of 20). The mean RSP values of the different inserts in these images are reported in Tables VIII and IX for the simulated and experimental data, respectively. The results in Tables VIII and IX show a very good agreement between the RSP values reconstructed using Algorithm 1 with three overlapping slices and the so-called true RPS [10].

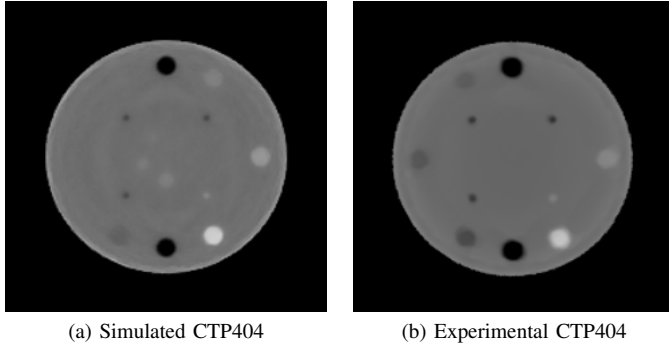| (a) Simulated CTP404 | (b) Experimental CTP404 |

Fig. 2: Reconstructed images of the CTP404 phantom

TABLE VIII: Reconstructed RSPs of the simulated CTP404

Insert	True RSP	1 GPU (% err.)	2 GPUs (% err.)
PMP	0.883	0.886 (0.33)	0.887 (0.45)
LDPE	0.980	0.986 (0.61)	0.988 (0.82)
Polystyrene	1.024	1.032 (0.78)	1.033 (0.87)
Acrylic	1.160	1.163 (0.25)	1.162 (0.17)
Delrin	1.359	1.349 (-0.73)	1.347 (-0.88)
Teflon	1.790	1.7895 (-0.03)	1.786 (-0.22)

TABLE IX: Reconstructed RSPs of the experimental CTP404

Insert	True RSP	1 GPU (% err.)	2 GPUs (% err.)
PMP	0.883	0.894 (1.24)	0.897 (1.59)
LDPE	0.980	0.989 (0.92)	0.990 (1.02)
Polystyrene	1.024	1.033 (0.87)	1.034 (0.97)
Acrylic	1.160	1.171 (0.94)	1.173 (1.12)
Delrin	1.359	1.345 (-1.03)	1.343 (-1.17)
Teflon	1.790	1.784 (-0.33)	1.781 (-0.50)

Finally, note that the RSP values converge only after three iterations of the iterative solver with Algorithm 1 running on a double GPU system, while it takes up to five iterations when running on a single GPU system.

V. CONCLUSION

Proton computed tomography is an evolving imaging technique currently under investigation to improve the accuracy of the proton treatment planning. Achieving the clinical timing constraints is one of the key goals for developing pCT softwares, especially when reconstructing large objects such as an adult torso.

In this paper, we proposed a fast and easy reconstruction technique that generates the 3D map of accurate RSP values in a very short amount of time with only two GPUs.

Based on the experiments, our proposed method surpasses the expected speedup on the Nvidia K40 GPUs which is a significant benefit that suggests an efficient and more economical hardware for a pCT system to be used in a clinical setup. Another important advantage of the proposed method, is that it can be easily be modified to be implemented on more than two GPUs, which we expect to further improve the timing constraints for exceedingly large datasets.

REFERENCES

[1] V. Bashkirov, R. Schulte, G. Coutrakon, B. Erdelyi, K. Wong, H. Sadrozinski, S. Penfold, A. Rosenfeld, S. McAllister, and K. Schubert, "Development of Proton Computed Tomography for Applications in Proton Therapy," in *APPLICATION OF ACCELERATORS IN RESEARCH AND INDUSTRY: Twentieth International Conference*, F. D. McDaniel and B. L. Doyle, Eds., vol. AIP Conference Proceedings Volume 1099. Fort Worth (Texas): American Institute of Physics, August 10-15 2008, pp. 460–463, iSBN: 978-0-7354-0633-9.

[2] R. Johnson, V. Bashkirov, G. Coutrakon, V. Giacometti, P. Karbasi, N. Karonis, C. Ordoñez, M. Pankuch, H.-W. Sadrozinski, K. Schubert *et al.*, "Results from a prototype proton-ct head scanner," *Physics Procedia*, vol. 90, pp. 209–214, 2017.

[3] R. Schulte, S. Penfold, J. Tafas, and K. Schubert, "A maximum likelihood proton path formalism for application in proton computed tomography," *Med. Phys.*, vol. 35, pp. 4849–4856, November 2008.

[4] B. Schultze, P. Karbasi, V. Giacomelli, T. Plautz, K. E. Schübert, and R. W. Schulte, "Reconstructing highly accurate relative stopping powers in proton computed tomography," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2015 IEEE*. IEEE, 2015, pp. 1–3.

[5] G. Dedes, L. De Angelis, S. Rit, D. Hansen, C. Belka, V. Bashkirov, R. Johnson, G. Coutrakon, K. Schubert, R. Schulte *et al.*, "Application of fluence field modulation to proton computed tomography for proton therapy imaging," *Physics in Medicine & Biology*, vol. 62, no. 15, p. 6026, 2017.

[6] K. L. Duffin, N. T. Karonis, C. E. Ordonez, M. E. Papka, G. Coutrakon, B. Erdelyi, E. C. Olson, and T. D. Uram, "An analysis of a distributed gpu implementation of proton computed tomographic (pct) reconstruction," in *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*. IEEE, 2012, pp. 166–175.

[7] N. T. Karonis, K. L. Duffin, C. E. Ordoñez, B. Erdelyi, T. D. Uram, E. C. Olson, G. Coutrakon, and M. E. Papka, "Distributed and hardware accelerated computing for clinical medical imaging using proton computed tomography (pct)," *Journal of Parallel and Distributed Computing*, vol. 73, no. 12, pp. 1605–1612, 2013.

[8] C. E. Ordoñez, N. Karonis, K. Duffin, G. Coutrakon, R. Schulte, R. Johnson, and M. Pankuch, "A real-time image reconstruction system for particle treatment planning using proton computed tomography (pct)," *Physics Procedia*, vol. 90, pp. 193–199, 2017.

[9] P. Karbasi, B. Schultze, V. Giacometti, T. Plautz, K. E. Schubert, R. W. Schulte, and V. A. Bashkirov, "Incorporating robustness in diagonally-relaxed orthogonal projections method for proton computed tomography," in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2015 IEEE*. IEEE, 2015, pp. 1–4.

[10] V. Giacometti, V. A. Bashkirov, P. Piersimoni, S. Guatelli, T. E. Plautz, H. F.-W. Sadrozinski, R. P. Johnson, A. Zatserklyaniy, T. Tessonnier, K. Parodi *et al.*, "Software platform for simulation of a prototype proton ct scanner," *Medical physics*, vol. 44, no. 3, pp. 1002–1016, 2017.

[11] S. N. Penfold, R. W. Schulte, Y. Censor, V. Bashkirov, S. Macallister, K. E. Schubert, and A. B. Rosenfeld, "Block-iterative and string-averaging projection algorithms in proton computed tomography image reconstruction," in *Biomedical Mathematics: Promising Directions in Imaging, Therapy Planning and Inverse Problems*, Y. Censor, M. Jiang, and G. Wang, Eds., The Huangguoshu International Interdisciplinary Conference. Madison, WI, USA: Medical Physics Publishing, 2010.