A High-Performance and Energy-Efficient CT Reconstruction Algorithm For Multi-Terabyte Datasets

SAND2013-3799C

Edward S. Jimenez, Laurel J. Orr, and Kyle R. Thompson

Sandia National Laboratories
PO Box 5800
Albuquerque,NM 87185
{*esjimen,ljorr,krthomp*}*@sandia.gov*

I. Introduction

Industrial Computed Tomography (CT) is an indirect 3D imaging technique that typically consists of datasets that are orders-of-magnitude larger compared to medical-scale datasets due to detectors with many pixels (usually 6-16 million pixels), many more projections, larger reconstruction volumes, high energy x-rays, or any combination thereof [1]–[3]. Although many GPU-based approaches to medical CT Reconstruction could be applied to industrial-scale CT datasets, the time required to reconstruct the volume could still be unreasonable due to bottlenecks created by the much larger dataset and computational requirement. Furthermore, these bottlenecks could become much more exaggerated by the GPU kernel design as well as host I/O capacity in both the host memory and storage. This work presents a flexible and portable multi-GPU reconstruction algorithm that exhibits high-performance and energy efficiency on a wide range of reconstruction tasks applicable to industrial applications.

II. Approach

Our approach centers around an irregular kernel design. We not only exploit the GPU's massively multi-threaded architecture and fast memory, but also:

- *Memory uploads* - Data uploads are maximized instead of minimized to accommodate more slices simultaneously.
- *Host Pinned Memory* - Pinned host memory will allow for faster data upload to the devices.
- *GPU Cache hit-rate maximization* - An irregular approach dramatically improves the GPU cache hit-rate thus maximizing computational performance and a reduction of wasted GPU clock cycles.
- *Dynamic Task Allocation* - Varying GPU tasks with respect to location in the volume will improve load-balancing between the GPUs as well as benefit the irregular computation.
- *Instruction Ordering* - Kernel design was implemented such that any register latency from write backs are amortized by assigning instructions that are independent of the inaccessible register.
- *Resource Maximization* - The kernel is designed independent of the GPU model and specifications, this will allow the kernel to execute optimally across a wide variety of GPUs. Resource maximization will query the resources available on the GPU and ensure that all compute cores are utilized as well as all device memory.

III. Implementation

- We present two implementations of our Large-Scale GPU-based Reconstruction Algorithm
  - The first is a lockstep multi-GPU Approach (LA) where the GPUs perform compute and storage tasks in lockstep.
  - The second is a modularized approach (MA) where the GPUs perform all tasks independently of one another, the only synchronization step is where more sinogram data needs to be read. Additionally, a CPU thread controlling a GPU is not responsible for storage write tasks, this task is offloaded to a separate CPU thread that is independent of GPU tasks, thus allowing compute and storage tasks to be performed simultaneously regardless of the number of GPUs connected to the host.
  - The implementations were written using a hybrid environment of C, CUDA 5.0, and OpenMP 2.0.

IV. Evaluation

- This two-page summary will look at the performance of a single system.
- The system is a Supermicro Server with dual Intel Xeon Processors @ 2.0 GHz (Octo-core with Hyper-Thread), 512 GB RAM, 8 disk RAID0 array, and 8 Tesla M2090 GPUs (6GB GDDR5 and 512 Streaming Processors each).
- The full paper will also evaluate performance on standard workstations, desktop systems, notebook, and a 5-node heterogeneous cluster.
- Performance will be measured against two datasets. The first is a 4000^3 voxel volume reconstruction from 1800 16 megapixel projections which is representative of the larger-end of current reconstructions. The second is a teravoxel (1 trillion voxels) volume reconstruction from 10,000 100 megapixel projections, which was selected as a stress-test as well as a potential future-sized dataset.
- Each system will run both GPU-based algorithms, a Naive single slice GPU-based algorithm (i.e. a GPU "Port" of CPU-based approaches), and a CPU-based multi-threaded reconstruction algorithm that is widely used in the national laboratory complex.

V. Preliminary Results

The preliminary results presented are a subset of results that will be presented in the final version. The server systems was chosen to give the reader a clear understanding of the portability of the code and its scalability characteristics by varying the number of GPUs available to the system. Additionally, Fermi-class graphics processors were used for the preliminary results, a section in the results for the final paper will focus on Kepler-class graphics processors.

A. 64 Gigavoxel Dataset

Table I presents performance results of the 8-GPU Server on the 64 gigavoxel reconstruction task. Many works have shown that GPU-based CT reconstruction significantly outperform CPU-based implementations, so this is to be expected [4]–[6]. We also show a very significant improvement in reconstruction performance from a Naive reconstruction kernel design to the synchronized irregular approach and even further improvement when a modularized approach is implemented with the irregular approach. It should be noted here that the difference between the LA and MA implementations is simply a redesign of the CPU-based thread assignments so that I/O tasks such as host memory transfers and storage tasks are handled by dedicated CPU threads and threads controlling CPU kernel launches solely focus on feeding data to its assigned GPU.

TABLE I
64 GIGAVOXEL DATASET - SERVER PERFORMANCE

Algorithm	Time(minutes)	Speedup	Energy(kWh)
CPU	2046.71	N/A	17.92
Naive-8GPU	552.53	3.70x	12.47
LA-8GPU	37.21	55x	0.98
MA-8GPU	27.71	73.86x	0.76

B. Teravoxel Dataset

Table II shows performance values on the synthetic teravoxel (1 trillion voxels) data set and even more dramatic improvement in performance metrics where again the MA implementation is superior. It should be noted that for the CPU-based reconstruction, only a subvolume was reconstructed and the overall values were extrapolated, the subvolume reconstructed consisted of 10 billion voxels located in the center slices of the volume as this is typically the region that algorithms perform best due to coalesced memory reads and reduced necessary data for a full reconstruction.

TABLE II
TERAVOXEL DATASET - SERVER PERFORMANCE

Algorithm	Time(hours)	Speedup	Energy(kWh)
CPU	2576	N/A	1362.43
Naive-8GPU	114.4	22.5x	164.34
LA-8GPU	23.2	111x	46.96
MA-8GPU	20	128.8x	38.72

C. Energy Consumption

Also presented in tables I, and II is the energy consumption of each algorithm. No previous work could be found on energy metrics of reconstruction algorithms. These values present relevant information since large-scale reconstructions will require non-trivial amounts of energy to complete due to the amount of data and computational complexity. For the 64 gigavoxel dataset, we see a 23.5x energy consumption improvement and a 35.2x improvement for the energy consumption for the teravoxel dataset. Improvement in both time and energy implies a very significant improvement in the energy-delay product [7].

D. Scalability

Table III shows scalability performance of the MA, LA, and Naive algorithms. The Naive approach exhibits negative scalability for many GPUs most likely due to the bandwidth pressure on the PCI-E bus whereas the irregular approaches exhibit similar scalability characteristics to each other with the lockstep (LA) approach slightly outperforming the modularized approach (MA) even though the modularized approach consistently completes the reconstruction task faster than the LA implementation. It is clear that the system is approaching its bandwidth limitation and if more than 8 GPUs were available then both irregular approaches would most likely exhibit identical behavior.

TABLE III
TERAVOXEL DATASET - SCALABILITY

GPUs	MA	LA	Naive
1	1x	1x	1x
2	1.86x	1.87x	*
4	3.25x	3.29x	*
6	4.32x	4.38x	1.72x
8	4.96	5.28x	1.58x

VI. CONCLUSION

- Our portable GPU-based reconstruction algorithm performs well on a wide range of systems with varying numbers of graphics processors. These algorithms presented improved computational performance over CPU-based implementations and other GPU-based approaches.
- A modularized approach improves performance and provides better support to GPU resource utilization although scalability may be impacted.
- Intelligent algorithm design in GPU kernels has shown that significant energy savings can be achieved, again, not only compared to CPU-based methods, but also against Naive approaches to GPU kernel design.
- Green computing is typically focused on innovative hardware design, but should also focus more on software design.

VII. ACKNOWLEDGEMENTS

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

REFERENCES

[1] S. Izumi, S. Kamata, K. Satoh, and H. Miyai, "High energy x-ray computed tomography for industrial applications," *Nuclear Science, IEEE Transactions on*, vol. 40, no. 2, pp. 158 –161, apr 1993.

[2] H. H. Barrett and K. J. Myers, *Foundations of Image Science*. Wiley-Interscience, 2004.

[3] E. S. Jimenez, L. J. Orr, and K. R. Thompson, "An Irregular Approach to Large-Scale Computed Tomography on Multiple Graphics Processors Improves Voxel Processing Throughput," in *Workshop on Irregular Applications: Architectures and Algorithms*, ser. The International Conference for High Performance Computing, Networking, Storage and Analysis, Nov. 2012.

[4] F. Xu and K. Mueller, "Ultra-fast 3d filtered backprojection on commodity graphics hardware," in *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, april 2004, pp. 571 – 574 Vol. 1.

[5] F. Xu and K. Mueller, "Accelerating popular tomographic reconstruction algorithms on commodity pc graphics hardware," *Nuclear Science, IEEE Transactions on*, vol. 52, no. 3, pp. 654 – 663, june 2005.

[6] K. Mueller, F. Xu, and N. Neophytou, "Why do commodity graphics hardware boards (GPUs) work so well for acceleration of computed tomography?" in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 6498, Feb. 2007.

[7] R. Gonzalez and M. Horowitz, "Energy dissipation in general purpose microprocessors," *Solid-State Circuits, IEEE Journal of*, vol. 31, no. 9, pp. 1277–1284, 1996.