Memory-Centric 3D Image Reconstruction with Hierarchical Communications on Multi-GPU Node Architecture

Mert Hidayetoğlu and Wen-mei W. Hwu (Advisor)
University of Illinois at Urbana-Champaign, IL 61810, USA

ABSTRACT

X-ray computed tomography is a commonly used technique for non-invasive imaging at synchrotron facilities. Iterative tomographic reconstruction algorithms are often preferred for recovering high quality 3D volumetric images from 2D X-ray images, however, their use has been limited to small/medium datasets due to their computational requirements. In this paper, we propose a high-performance iterative reconstruction system for terabyte(s)-scale 3D volumes. Our design involves three novel optimizations: (1) optimization of (back)projection operators by extending the 2D memory-centric approach to 3D; (2) performing hierarchical communications by exploiting "fat-node" architecture with many GPUs; (3) utilization of mixed-precision types while preserving convergence rate and quality. We extensively evaluate the proposed optimizations and scaling on the Summit supercomputer. Our largest reconstruction is a mouse brain volume with 9K×11K×11K voxels, where the total reconstruction time is under three minutes using 24,576 GPUs, reaching 65 PFLOPS: 34% of Summit's peak performance.

ACM Reference Format:
Mert Hidayetoğlu and Wen-mei W. Hwu (Advisor). 2020. Memory-Centric 3D Image Reconstruction with Hierarchical Communications on Multi-GPU Node Architecture . In *The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC20), November 16–19, 2020, Atlanta, GA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3295500.3356220

1 INTRODUCTION

X-ray computed tomography (XCT) is used regularly for micrometer- and nanometer-scale resolution non-invasive 3D imaging. Applications involve nanoelectronics, materials research, cell morphology, and medical imaging. Although direct methods based on Fourier-slice theorem are commonly used for image reconstruction with low computational complexity, they require requires long scanning time and are highly prone to noise in the measurement. On the other hand, iterative image reconstruction is robust to measurement noise and allows fast scanning. However, computational requirements of iterative XCT have made their use exception rather than the rule. As a remedy, we optimize iterative XCT on GPUs with a novel memory-centric algorithm design: MemXCT. This poster presentation summarizes these optimizations involving a pseudo-Hilbert ordering algorithm and a 3D multi-stage buffering technique that overcomes inefficiencies due to irregular data accesses.

SC20, November 16–19, 2020, Atlanta, GA, USA

ACM ISBN 978-1-4503-6229-0/19/11…$15.00
https://doi.org/10.1145/3295500.3356220

Each generation of X-ray light source yields brighter beam that allows scanning larger objects with higher data acquisition rate. To be specific, brilliance triples every 18 months. On the other hand, processing rate (arguably) doubles every 18 months in accordance with the Moore's law. Henceforth, traditional iterative XCT approaches does not address the performance gap between data acquisition and data processing. Although MemXCT implements a load-balanced partitioning of processing domains, scaling of large image reconstruction has an inevitable communication bottleneck, especially the low bandwidth between computational nodes. To overcome this bottleneck, we develop a hierarchical communication and data reduction strategy that exploits the fat-node (multi-GPU) architecture of state-of-the-art leadership class supercomputers.

2 MEMORY-CENTRIC ALGORITHM DESIGN

Previous high-performance iterative XCT implementations perform essential projection and backprojection operations on-the-fly by numerically simulating X-ray propagation through the object [1],[2]. These operations are not only redundant (because of duplicated operations in each iteration) but also inefficient due to irregular memory access patterns (because GPU architecture is optimized for regular data processing.) Memory-centric approach, on the other hand, simulates and pre-processes the X-ray propagation once and stores the projection and backprojection data structures explicitly as sparse matrices in memory. These sparsity patterns of these matrices are localized by ordering the 3D input (measurement) and output (image) domains using a stacked pseudo-Hilbert ordering of each slice of the input (sinogram) and output (tomogram). Even though memory-centric approach has a higher memory complexity, our results show it is faster by 7×–50× compared to traditional compute-centric approach [3].

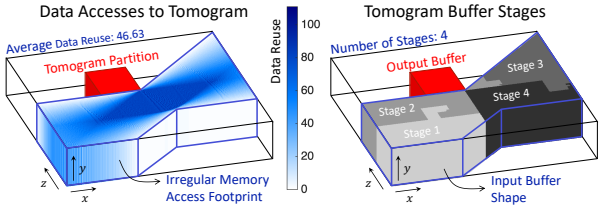

Figure 1: Multi-Stage Input Buffering.

3 PERFORMANCE OPTIMIZATIONS

The naive multiplication of sparse matrices suffer from irregular data access. MemXCT preprocess the memory access footprint of each thread block and stages the irregular data access at shared memory. To overcome large memory footprints with small shared memory, each block stages multiple buffers. The multi-stage buffering strategy provides data reuse from shared memory and minimizes high-latency global memory access.

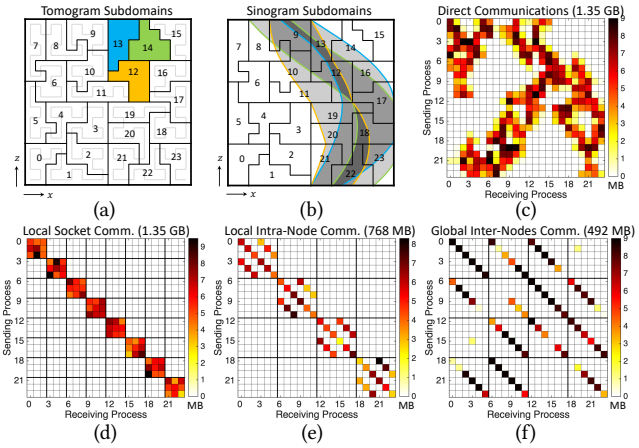

Figure 2: Hierarchical communication patterns.

To overcome the memory-bandwidth bottleneck, we improve the arithmetic intensity by minibatching many tomograms and reuse the sparse matrix elements from register. Fig. 1(a) shows irregular memory access footprint on a tomogram minibatch. In this case, the average data reuse from shared memory is 46.63. Fig. 1(b) shows the four buffer stagings to cover the corresponding memory footprint.

4 HIERARCHICAL COMMUNICATION & DATA REDUCTION

To explain the hierarchical communication strategy, Fig. 2 (a) and (b) shows the partitioning of tomogram and sinogram domains among 24 GPUs (each GPU processes a single partition) and communication footprint between them. That is, the partial data produced by tomogram partitions 12–14 are shown in Fig. 2 (b) with their corresponding colors. The corresponding sparse communication matrix is shown in Fig. 2 (c). Direct (naive) communications performs the required sparse communications and then each GPU computes the total data by a summation of the overlapping partial data. As shown in the figure, GPU 18 receives partial data produced by the highlighted GPUs 12–14 and reduces the overlapping region. In direct communications, the total communicated data is 1.35 GB.

Hierarchical communication strategy exploits the multi-GPU node architecture by reducing data locally wherever possible to minimize the low-bandwidth communications between nodes. To explain, Fig. 2(d) shows the local high-bandwidth communications between three GPUs which are fully-connected. Then the data is reduced according to the communication footprints. Likewise, Fig. 2(e) shows the intra-node communications and reduction between six GPUs inside a node. Finally, Fig. 2(f) shows the inter-node communications and reductions. As the figure suggest, even though the total communicated data is increased, communicated data between nodes is decreased down to 492 MB.

5 NUMERICAL RESULTS

We implement MemXCT with double, single, and mixed precisions. To demonstrate the effect of each optimization, Fig. 3 shows the breakdown of the end-to-end reconstruction times of `Shale` and `Charchoal` datasets [4] on four nodes (24 GPUs) and 128 nodes (768 GPUs) of Summit supercomputer, respectively. These results show that optimized SpMM reduces kernel execution time significantly in all cases. The performance of the kernel is measured at 75 TFLOPS for Shale on four nodes (24 GPUs) and 2.38 PFLOPS for Charcoal on 128 nodes (768 GPUs). As shown in Fig. 10, execution time is dominated by communication for most of the cases. Hierarchical communications reduce the communication time by more than 50% in all cases. Our largest reconstruction involves a 6.56-TB `brain` dataset which is reconstructed within three minutes on 4,096 nodes (24,576 GPUs) reaching 65 PFLOPS: 34% of Summit's peak performance.

6 CONCLUSION

This short paper summarizes the system MemXCT: a memory-centric approach for iterative XCT for 3D image reconstruction. The proposed optimizations overcome the irregular data accesses with a novel multi-stage buffering algorithm and provide higher arithmetic intensity that efficiently employs GPU. To provide scaling, communications are optimized by exploiting the multi-GPU node architecture of leadership-class supercomputers. We extensively demonstrate MemXCT scaling up to tens of thousands of GPUs, which is portable to next-generation exascale computing systems as well as the next-generation imaging systems.

REFERENCES

[1] T. Bicer, D. Gürsoy, V. D. Andrade, R. Kettimuthu, W. Scullin, F. D. Carlo, and I. T. Foster, "Trace: A high-throughput tomographic reconstruction engine for large-scale datasets," *Advanced Structural and Chemical Imaging*, vol. 3, p. 6, Jan 2017.

[2] X. Wang, A. Sabne, P. Sakdhnagool, S. J. Kisner, C. A. Bouman, and S. P. Midkiff, "Massively parallel 3d image reconstruction," in *International Conference for High Performance Computing, Networking, Storage and Analysis*, p. 3, ACM, 2017.

[3] M. Hidayetoğlu, T. Biçer, S. G. De Gonzalo, B. Ren, D. Gürsoy, R. Kettimuthu, I. T. Foster, and W.-m. W. Hwu, "MemXCT: Memory-centric X-ray CT reconstruction with massive parallelization," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–56, 2019.

[4] M. Hidayetoğlu, T. Biçer, S. G. De Gonzalo, B. Ren, V. De Andrade, D. Gürsoy, R. Kettimuthu, I. T. Foster, and W.-m. W. Hwu, "Petascale XCT: 3D image reconstruction on multi-gpu nodes," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, p. accepted, 2020.

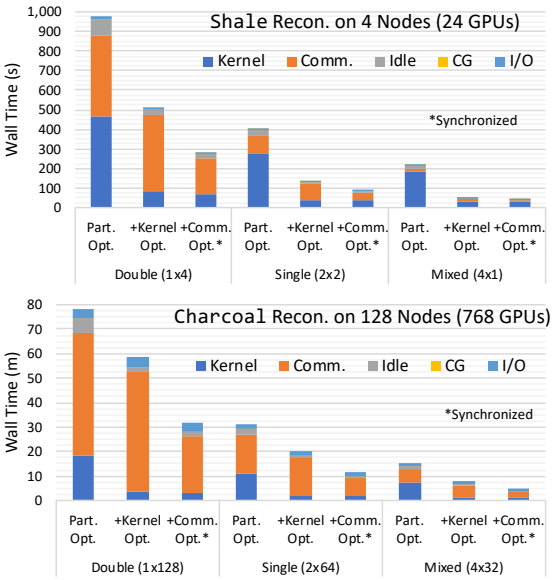

Figure 3: Breakdown of end-to-end reconstruction times.