https://doi.org/10.2352/EI.2024.36.12.HPCI-188

Accelerated Laminographic Image Reconstruction using GPUs

***Bin Ma*[1], *Viktor Nikitin*[2], *Dong Li*[1], *Tekin Bicer*[2]**
[1] ***University of California, Merced, California, USA***
[2] ***Argonne National Laboratory, Lemont, Illinois, USA***

Abstract

Laminography is a specialized 3D imaging technique optimized for examining flat, elongated structures. Laminographic reconstruction is the process of generating 3D volume from a set of 2D projections that are collected during the laminography experiment. Iterative reconstruction techniques are typically the preferred computational method for generating high-quality 3D volumes, however, these methods are computationally demanding and therefore can be infeasible to apply to large datasets. To counteract these challenges, we require state-of-the-art computational methods that can efficiently utilize high-performance computing resources such as GPUs. In this work, we investigate the integration of the Unequally Spaced Fast Fourier Transform (USFFT) with two optimization methods: the Alternating Direction Method of Multipliers (ADMM) and the Conjugate Gradient (CG). The usage of USFFT addresses non-uniform sampling issues typical in laminography, while the combination of ADMM and CG introduces robust regularization techniques that enhance image quality by preserving edges and reducing noise. We further accelerated the iterative algorithm of USFFT by preprocessing the image into the frequency domain. Compared to the original algorithm, the optimized USFFT method achieved a 1.82x speedup. By harnessing heterogeneous computing and parallel computing with both CPU and GPU, our approach significantly accelerates the reconstruction process while keeping the quality of the generated images. We evaluate the performance of our methods using real-world datasets collected at 32-ID beamline at Advanced Photon Source using Argonne Leadership Computing Resources.

Introduction

Laminography is an imaging technique primarily used for detecting layered objects such as semiconductor wafers, circuit boards, and mouse brain [1–4]. In a laminography setup, the sample under investigation is placed on a stage that is both tilted and capable of rotation. This arrangement is a key aspect of laminography, as it allows for the detailed imaging of layers within the sample. When the X-ray is applied, the tilting of the stage ensures that the X-ray beam intersects with the layers of the object at an oblique angle. This angle is particularly advantageous for imaging flat, layered structures, as it enhances the visibility of features within these layers on 2D projections.

The principle behind laminography is based on the idea that by tilting and rotating the sample while it is being exposed to X-rays, different layers within the sample can be selectively focused upon. This results in clearer and more detailed images of the internal structures. The technique is, for example, highly effective for detecting faults and defects in layered objects, such as delamination, cracks, and foreign materials embedded within the layers.

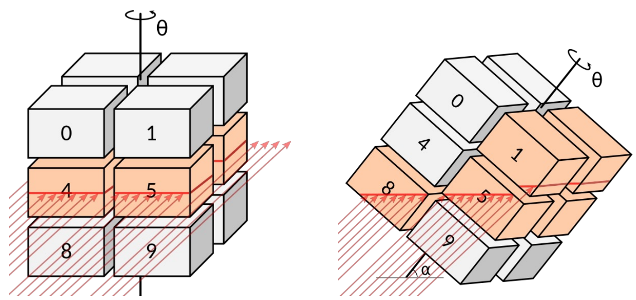

Figure 1: Experimental setups for (a) tomography and (b) laminograpy.

Although laminography is a powerful imaging technique for layered structures, it still faces several challenges in terms of reconstruction operations. For example, the limited-angle data and overlapping structures can complicate the reconstruction task. Further, state-of-the-art parallelization methods for image reconstruction techniques cannot be applied to laminography out-of-box [5–7]. For instance, while parallel beam geometry enables the straightforward application of data parallelization on 3D volume slices (or projection sinograms) during tomography experiments [2, 8], the tilted angle of the sample in laminography results in X-ray propagation paths that intersect multiple slices in 3D volume as illustrated in Fig. 1. This intersection, compared to tomographic reconstruction, complicates the parallelization of computationally intensive forward and back projection operations during iterative reconstruction process.

The Fast Fourier Transform (FFT) method, particularly when augmented with regularization techniques, stands out as an advanced solution for addressing a range of challenges in data-intensive fields such as medical imaging, signal processing, and computational physics [9]. Its inherent computational efficiency makes it suitable for handling large datasets, potentially alleviating issues related to data size and offering faster reconstructions. The regularization techniques employed alongside FFT further enhance its effectiveness. Regularization involves introducing additional information or constraints to counteract the ill-posed nature of certain problems, such as those encountered in limited-angle tomography. By incorporating these constraints, FFT-based methods can yield more accurate and stable solutions, even in cases where traditional methods might struggle.

However, the deployment of 3D FFT for reconstruction purposes brings its own set of challenges. One of the primary concerns is the computational load. The process of performing 3D FFT reconstruction requires executing numerous 2D FFTs on large data arrays, which can be computationally demanding. This demand extends not only to processing power but also to memory requirements, as large datasets necessitate substantial memory for both storage and efficient processing.

Moreover, the efficacy of FFT in parallel computing environments hinges on the careful partitioning and distribution of data. The parallelization of FFT operations, a common approach to handle large-scale problems, requires careful planning and execution. The data must be partitioned and distributed in a manner that aligns with the nature of FFT operations and their data access patterns to ensure accurate and efficient reconstruction. This requirement poses a significant challenge in terms of both algorithm design and implementation.

Related Work

There is extensive literature on 3D image reconstruction algorithms for (m/n)CT datasets [10–12]. The recent studies also include high-performance computing techniques for reconstructing large-scale 3D volumes [7, 13, 14]. These techniques not only enable the scalable and efficient usage of advanced computing resources [15, 16] but also provide opportunities to deal with extremely large problem sizes via scientific workflows that can federate experimentation and supercomputer-scale computational resources [17–20]. Complementary to these efforts, AI/ML techniques have also been incorporated into the reconstruction operations to ease the computational requirements of different reconstruction tasks and enhance the quality of generated images [21–30]. Laminography can be seen as an extension of the tomography imaging technique, which introduces a tilted angle to the rotation stage. This property enables isolation and visualization of specific planes within an object, bypassing the limitations imposed by traditional tomography when dealing with objects of considerable depth or layered compositions [9]. Laminography is also the preferred method when the rotation angle during the experimentation is limited [31].

The existing calculation methods of 3D image reconstruction mainly focus on the spatial domain and frequency domain. Spatial domain methods focus on the manipulation and reconstruction of images directly, where the value of each point (pixel or voxel) represents intensity or color. Techniques like Filtered Back Projection (FBP), Algebraic Reconstruction Technique (ART), and Iterative Reconstruction (IR) fall under this category [10]. These methods typically involve operations such as back-projection, iterative refinement, and algebraic computations to reconstruct the image from projections.

In addition to spatial methods, frequency-based methods are also widely used in 3D image reconstruction. Frequency domain methods involve working with the Fourier Transform (FFT)[32] or related transforms of the image data. The Fourier slice theorem, which states that the 1D Fourier transform of a projection of a 2D object is equivalent to a slice of the 2D Fourier transform of the object, is the cornerstone of this method. The integration of FFT brings a mathematical shortcut that significantly accelerates the computation of the projections required for reconstruction.

These image reconstruction algorithms are computationally expensive[33]. The introduction of accelerators, such as GPUs, has made it possible to parallelize these algorithms and make them execute efficiently, thereby dramatically reducing computation time [5, 6, 34].

Objective

In this work, we focus on efficiently solving the 3D laminographic reconstruction problem with a Fourier-based iterative method. Considering the time-intensive computational demands of laminography reconstructions, which can span extended time periods, our study aims to truncate this processing duration by integrating state-of-the-art methods with the prowess of modern hardware and accelerator capabilities.

We focus on every facet of the reconstruction pipeline, which includes algorithmic improvements, optimizing data I/O interactions with storage drives, enhancing the efficiency of CPU-GPU data transfers, and improving computational throughput on the GPU. The overarching objectives are to alleviate computational bottlenecks during the reconstruction of large-scale laminography datasets, decrease prolonged multi-day processing times to hours or minutes, and keep the accuracy of the reconstruction operations and the quality of the outputs.

Methodology and Experimental Design

In this section, we provide detailed information about our advanced iterative algorithm tailored for USFFT computations on GPU, categorized into ”forward” and ”adjoint” phases[9]. At the same time, we perform the regularization of image reconstruction and complete its calculation on the CPU. This methodology was realized within a heterogeneous, high-performance CPU-GPU computational environment.

Iterative Reconstruction in Frequency Domain

The forward phase of the reconstruction algorithm is a critical component in generating a refined image during iterations. This phase involves a sequence of Fourier transforms, each serving a specific purpose in the image reconstruction process. Initially, a one-dimensional inverse unequally spaced Fourier transform (USFFT1D) is executed. This step transforms frequency-domain data back into a spatial or time domain along one axis, laying the groundwork for subsequent transformations. Following this, a 2-dimensional (2D) unequally spaced inverse Fourier transform is performed (USFFT2D). This step extends the transformation process to two dimensions, further reconstructing the image from its frequency-domain representation. The final step in the forward phase is a direct 2D Fourier transform. This final transformation produces the recovered images for comparison with the original spatial-domain data, and it sets the stage for the computations in the adjoint phase. After these transformations, the algorithm calculates the disparity between the newly generated image and the original image, which identifies areas of divergence and guides the adjustments needed in subsequent iterations to achieve a more accurate reconstruction.

The adjoint phase streamlines image reconstruction by iteratively optimizing it. It first converts spatial discrepancies found in the forward phase into the frequency domain, aligning them with the original data for easier adjustment. Then, it applies a 2D unequally spaced Fourier transform, followed by a 1D transform, to generat a 3-dimensional tensor. This tensor is indicative of the gradient required for the updates in the upcoming iterations. The gradient provides directional guidance for the algorithm, indicating how the image reconstruction should be adjusted to more closely align with the original image.

To enhance the robustness and efficiency of our algorithm,

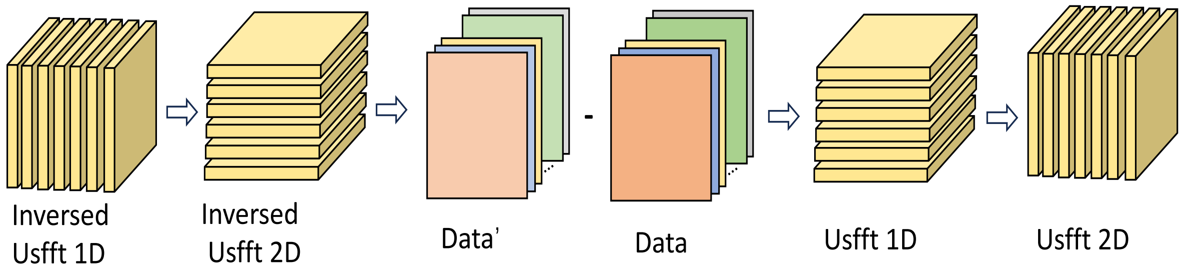

Figure 2: "Chunking" strategy for USFFT1D and USFFT2D computations in our 2-step reconstruction method.

Algorithm 1 3-step FFT update.

Input: 3D reconstruction matrix **M**, X-ray data D
Output: Gradient g
for $i = 0$ to $titer$ **do**
  forward:
    $D_{hat} = FFT2d(USFFT2D(USFFT1D(M)))$
  adjoint:
    $g = USFFT1D(USFFT2D(FFT2D(D_{hat} - D)))$
  $alpha, d = CG(g, g_{prev})$
  $M = update(alpha, d, g, M)$
end for

Algorithm 2 2-step FFT update.

Input: 3D reconstruction matrix **M**, X-ray data D' in frequency domain
Output: Gradient g
for $i = 0$ to $titer$ **do**
  forward:
    $D'_{hat} = USFFT2D(USFFT1D(M))$
  adjoint:
    $g = USFFT1D(USFFT2D(D'_{hat} - D'))$
  $alpha, d = CG(g, g_{prev})$
  $M = update(alpha, d, g, M)$
end for

we have integrated a differential regularization framework that exploits the inherent differences between neighboring image frames. This innovative approach is underpinned by a robust computational strategy that utilizes the parallelism of multi-threaded CPU architectures, thus ensuring both high accuracy and computational speed. We have conducted an overlap between the differential computations performed on the CPU and the frequency-domain computations carried out on the GPU. This strategy facilitates a seamless heterogeneous parallel computing environment via overlapping the computations and provides overall computational efficiency (and throughput) during reconstruction.

To find the best hyper-parameter for gradients and differences computation, we employ the conjugate gradient (CG) descent method. The CG is predicated on its proven efficacy in navigating the complex landscape of optimization problems with precision and speed. Upon identifying the optimal step size, we integrate the differential updates with the gradient updates on the reconstructed images. This integration ensures each update contributes effectively to the enhancement of the image reconstruction quality.

GPU Memory Limitation

The conventional methods have significant limitations while processing large volumes of raw image datasets due to the constrained memory capacities especially with accelerators such as GPUs. To address this problem, we adopted a "chunking" strategy, allowing for segmented and efficient 1D and 2D Fourier transformations as shown in Figure 2. Specifically, our methodology exploited the 3D row-based partitions for the execution of 1D Fourier Transform Fast (FFT) operations. Simultaneously, for 2D FFT operations, we employ column-based partitions. This approach is designed to ensure that 1D and 2D FFT operations on the GPU are executed without encountering prohibitive memory limitations. Furthermore, our partitioning strategy is instrumental in achieving a seamless integration and synchronization between data transfer and computation on GPUs.

Scaling Up Computational Resources

The chunking strategy we employed significantly enhances the scalability of our USFFT computation operators across multiple GPUs within a single node. This approach involves dividing the data along the direction of chunk segmentation, proportionately distributing it according to the number of GPUs, as shown in Figure 3. Through this method, every GPU leverages the combined efficiency of chunking and pipeline computation.This strategic division and distribution of data also ensure that each GPU can operate independently and concurrently, optimizing the use of available computational resources. By aligning the data partitioning with the physical architecture of the GPUs, we mitigate potential bottlenecks in data transfer, thereby maximizing the throughput of the entire system.

The adaptability of our chunking strategy to scale up across multiple GPUs within a single node also highlights its flexibility in a variety of computational environments and it effectively increases the utilization of computing resources. Whether dealing with a small cluster of GPUs or a large-scale high-performance computing (HPC) environment, our approach can be tailored to meet the specific needs and constraints of the available hardware.

Optimization on Reversible Algorithm

Another key contribution to our work is that we have refactored and improved the state-of-the-art reconstruction algorithm. Specifically, we changed the original three-step FFT reconstruction method, as shown in Algorithm 1, to an enhanced two-step method, as shown in Algorithm 2. A necessary preparatory step

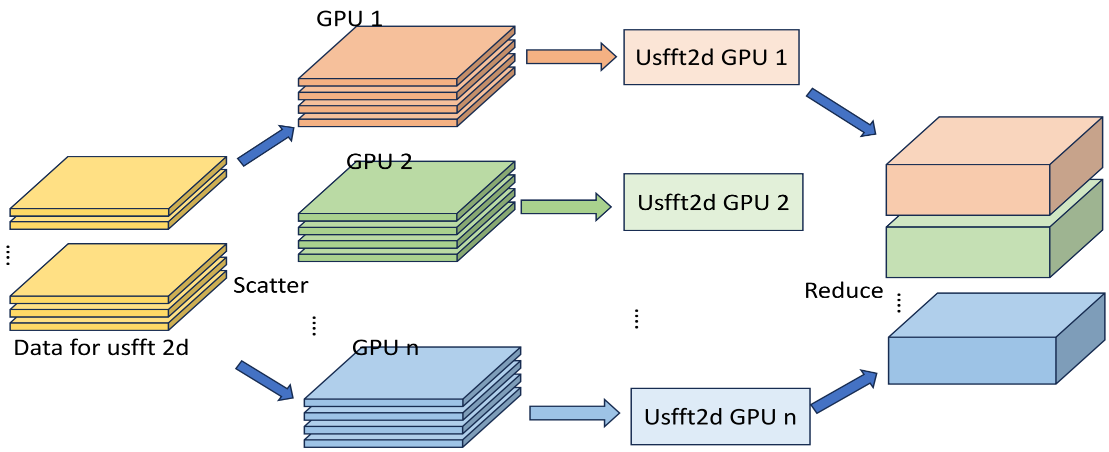

Figure 3: Single node multi-GPU scale up for USFFT2d

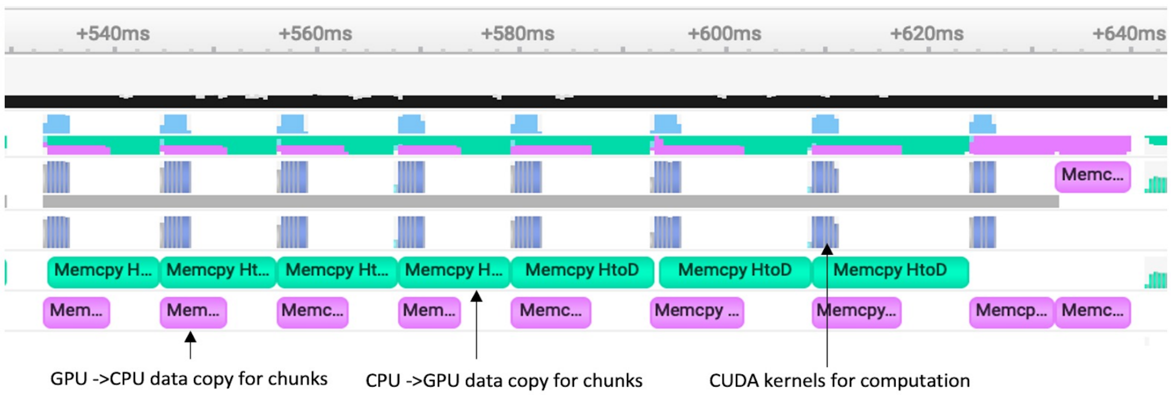

Figure 4: Computation process profiled by the NVIDIA Nsight. Note the overlap between computation and data transfers.

is the overall mapping of all images onto the frequency domain to ensure the integrity of the optimized algorithm compared to the original version. This change eliminates the additional FFT operation overheads that exist in the original version. Our empirical evaluations revealed a significant increase in computational throughput – 1.82x speedup per iteration – which shows the overhead introduced due to the redundant FFT computations in both the forward and adjoint phases.

Results and Discussion

We evaluated our method with a real-world laminography dataset of a mouse brain sample. This sample was collected at the 32-ID beamline at APS. The dataset size is 13.2 GB, which comprises 750 total projections, each with a resolution of 1536 x 4608 pixels. To expedite the evaluation of our methods, we've employed a 4x4 binning technique on these original 2D projections. This preprocessing step allows for a more rapid analysis while maintaining the necessary level of detail for accurate reconstruction. The corresponding reconstructed 3D volume has 384x1152x1152 dimensions, which results in a dataset size of 3.8 GB.

The reconstructions are performed on a high-end compute node at Polaris supercomputer at Argonne Leadership Computing Facility. The compute node consists of one AMD EPYC 7543P CPU and four NVIDIA A100 GPUs. The CPU has 32 cores and 512GB DDR memory, while each GPU has 40 GB memory.

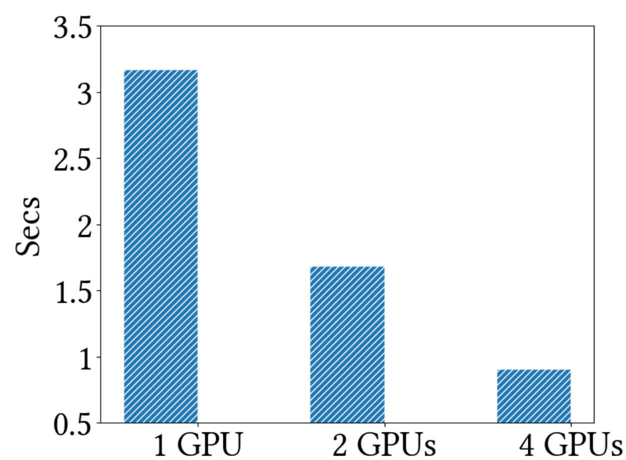

Figure 5: Strong scalability for USFFT2d operator.

GPUs are connected with NVLink interconnect.

Our evaluation presents an insightful comparison between two distinct reconstruction methods applied to the mouse brain dataset, with a focus on computational efficiency and optimization techniques. It shows that the reconstruction of the sampled mouse brain dataset can be converged in 14 and 25 mins using our (enhanced) two-step and (original) three-step reconstruction methods, respectively, on a single GPU. We also overlap the data transfer and GPU computation via our data partitioning/chunking scheme, as previously shown in Figures 2 and 3, which provides 1.96X speedup compared to without such optimization.

Since our performance profiling revealed that USFFT2d takes the most time among the FFT operators – reaching 70% of all FFT calculations – we also evaluated the (strong) scalability of this operator on 1, 2, and 4 GPUs (within a single computing node). As illustrated in Figure 5, the USFFT2d demonstrates exceptional scalability with our chunking strategy, achieving a close to linear speed-up of 1.8x with 2 GPUs and 3.5x with 4 GPUs, when compared to a single GPU setup. Through the implementation of the chunking strategy, we have realized efficient scalability and parallel computing capabilities. We are currently working on extending our parallelization method to multi-node multi-GPUs settings so that we can reconstruct even larger 3D volumes.

Conclusion

In this work, we developed a parallel, GPU-optimized reconstruction algorithm for laminography. We have developed an innovative two-step reconstruction technique specifically designed to streamline processing by eliminating unnecessary Fast Fourier Transform (FFT) calculations, thereby enhancing computational speed and efficiency. We propose a data partitioning and chunking scheme that strategically enables the simultaneous execution of host-device communication and GPU-based computations. Lastly, we present a cutting-edge GPU-accelerated differential regularization method. Our method improves the image quality while preserving edges that are critical for accurate interpretation of image content. Our method also helped noise reduction which eliminate unwanted signals within the reconstructed images. Our single-GPU optimizations provide 1.82x speedup compared to the unoptimized version.

Acknowledgements

This research used resources of the Advanced Photon Source (APS) and Argonne Leadership Computing Facility (ALCF), U.S. Department of Energy (DOE) Office of Science user facilities, and is based on work supported by Laboratory Directed Research and Development (Project Number: 2023-0104) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. DOE under Contract No. DE-AC02-06CH11357.

References

[1] S. L. Fisher et al. "Laminography in the lab: imaging planar objects using a conventional x-ray CT scanner". In: *Measurement Science and Technology* 30.3 (Feb. 2019), p. 035401.

[2] T. Bicer et al. "Trace: a high-throughput tomographic reconstruction engine for large-scale datasets". In: *Advanced Structural and Chemical Imaging* 3.1 (Jan. 2017). ISSN: 2198-0926.

[3] K. J. Batenburg et al. "3D imaging of semiconductor components by discrete laminography". In: *Stress induced phenomena and reliability in 3D microelectronics: AIP Conference Proceedings*. Vol. 1601. 2014, pp. 168–179.

[4] L. Helfen et al. "Synchrotron-radiation computed laminography for high-resolution three-dimensional imaging of flat devices". In: *physica status solidi (a)* 204 (2007).

[5] M. Hidayetoglu et al. "Petascale XCT: 3D Image Reconstruction with Hierarchical Communications on Multi-GPU Nodes". In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society. 2020, pp. 510–522.

[6] X. Wang et al. "Massively parallel 3D image reconstruction". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2017, pp. 1–12.

[7] P. Chen et al. "Scalable FBP decomposition for cone-beam CT reconstruction". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2021, pp. 1–16.

[8] T. Bicer et al. "Rapid tomographic image reconstruction via large-scale parallelization". In: *European Conference on Parallel Processing*. Springer. 2015, pp. 289–302.

[9] V. Nikitin et al. *Laminography as a tool for imaging large-size samples with high resolution*. 2024.

[10] A. C. Kak et al. *Principles of computerized tomographic imaging*. SIAM, 2001.

[11] S. V. Venkatakrishnan et al. "Plug-and-play priors for model based reconstruction". In: *2013 IEEE global conference on signal and information processing*. IEEE. 2013, pp. 945–948.

[12] K. A. Mohan et al. "TIMBIR: A method for time-space reconstruction from interlaced views". In: *IEEE Transactions on Computational Imaging* 1.2 (2015), pp. 96–111.

[13] M. Hidayetoğlu et al. "MemXCT: Memory-centric x-ray CT reconstruction with massive parallelization". In: *International Conference for High Performance Computing, Networking, Storage and Analysis*. 2019, pp. 1–56.

[14] X. Wang et al. "Consensus equilibrium framework for super-resolution and extreme-scale CT reconstruction". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2019, pp. 1–23.

[15] T. Bicer et al. "Real-time data analysis and autonomous steering of synchrotron light source experiments". In: *IEEE 13th International Conference on e-Science (e-Science)*. IEEE. 2017, pp. 59–68.

[16] T. Bicer et al. "Tomographic Reconstruction of Dynamic Features with Streaming Sliding Subsets". In: *2020 IEEE/ACM 2nd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP)*. IEEE. 2020, pp. 8–15.

[17] R. Vescovi et al. "Linking scientific instruments and computation: Patterns, technologies, and experiences". In: *Patterns* 3.10 (2022), p. 100606. ISSN: 2666-3899.

[18] J. Chung et al. "Scistream: Architecture and toolkit for data streaming between federated science instruments". In: *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing*. 2022, pp. 185–198.

[19] T. Bicer et al. "Optimization of tomographic reconstruction workflows on geographically distributed resources". In: *Journal of Synchrotron Radiation* 23.4 (2016), pp. 997–1005.

[20] T. Bicer et al. "High-performance ptychographic reconstruction with federated facilities". In: *Smoky Mountains Computational Sciences and Engineering Conference*. Springer. 2021, pp. 173–189.

[21] S. Venkatakrishnan et al. "Algorithm-driven Advances for Scientific CT Instruments: From Model-based to Deep Learning-based Approaches". In: *arXiv preprint arXiv:2104.08228* (2021).

[22] C. Benmore et al. "Advancing AI/ML at the Advanced Photon Source". In: *Synchrotron Radiation News* 35.4 (2022), pp. 28–35.

[23] B. Sun et al. "SOLAR: A Highly Optimized Data Loading Framework for Distributed Training of CNN-based Scientific Surrogates". In: *arXiv preprint arXiv:2211.00224* (2022).

[24] J. Liu et al. "Large Scale Caching and Streaming of Training Data for Online Deep Learning". In: *Proceedings of the 12th Workshop on AI and Scientific Computing at Scale using Flexible Computing Infrastructures*. 2022, pp. 19–26.

[25] Z. Wu et al. "Deep learning-based low-dose tomography reconstruction with hybrid-dose measurements". In: *2020 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC) and Workshop on Artificial Intelligence and Machine Learning for Scientific Applications (AI4S)*. IEEE. 2020, pp. 88–95.

[26] S. Aslan et al. “Joint ptycho-tomography reconstruction through alternating direction method of multipliers”. In: *Optics express* 27.6 (2019), pp. 9128–9143.

[27] Z. Liu et al. “Deep learning accelerated light source experiments”. In: *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*. IEEE. 2019, pp. 20–28.

[28] Z. Liu et al. “TomoGAN: low-dose synchrotron x-ray tomography with generative adversarial networks: discussion”. In: *JOSA A* 37.3 (2020), pp. 422–434.

[29] S. Aslan et al. “Joint ptycho-tomography with deep generative priors”. In: *Machine Learning: Science and Technology* 2.4 (2021), p. 045017.

[30] A. V. Babu et al. “AI-assisted automated workflow for real-time x-ray ptychography data analysis via federated resources”. In: *Electronic Imaging* 35 (2023), pp. 1–6.

[31] F. Xu et al. “Comparison of image quality in computed laminography and tomography”. In: *Opt. Express* 20.2 (Jan. 2012), pp. 794–806.

[32] J. Fessler et al. “A min-max approach to the multidimensional nonuniform FFT: application to tomographic image reconstruction”. In: *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*. Vol. 1. 2001, 706–709 vol.1.

[33] J. A. Fessler. “Iterative Tomographic Image Reconstruction Using Nonuniform Fast Fourier Transforms”. In: 2002.

[34] P. Chen et al. “IFDK: A scalable framework for instant high-resolution image reconstruction”. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2019, pp. 1–24.

Author Biography

Bin Ma is a PHD student in University of California, Merced. His research interests in heterogeneous memory systems and their applications to deep learning system and HPC applications.

Dong Li is an associate professor at University of California, Merced. He is the director of Parallel Architecture, System, and Algorithm Lab (PASA) and a co-director of High Performance Computing Systems and Architecture Group at UC Merced. Previously (2011-2014), he was a research scientist at the Oak Ridge National Laboratory (ORNL). Before that, he earned a PhD in computer science from Virginia Tech. He is an associate editor for IEEE Transactions on Parallel and Distributed Systems (TPDS). Dong’s research focuses on high performance computing (HPC), and maintains a strong relevance to computer systems (especially systems for large-scale AI/ML).

Tekin Bicer is a computer scientist in Data Science and Learning division at ANL. He has expertise in HPC, large-scale parallel and distributed systems, and AI/ML methods, with a special focus on large-scale x-ray image analysis problems. He received his Ph.D. from Computer Science and Engineering Department at Ohio State University in 2014, where he studied large-scale computing systems that address data management and analysis problems on high-end clusters and cloud compute resources.

License