doi:10.1017/S1431927622011564
Microsc. Microanal. 28 (Suppl 1), 2022

Efficient Memory Storage and Linear Parallel Scaling for Large-Scale Electron Ptychography

Xiao Wang[1*], Debangshu Mukherjee[1], Aristeidis Tsaris[1], Mark Oxley[2], Olga Ovchinnikova[1] and Jacob Hinkle[1]

[1.] Computational Sciences & Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
[2.] Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
* Corresponding author: wangx2@ornl.gov

Electron ptychography is a microscopic imaging method that reconstructs an object from a set of spatially overlapping coherent electron diffraction pattern measurements. Since the image resolution in ptychography is limited by the extent of the diffraction patterns as opposed to electron lenses, ptychographic imaging sets the current records for the highest microscope resolution [1, 2]. Despite its promise in delivering unprecedented image resolution, ptychographic imaging requires enormous memory and is limited by the size of the diffraction data sets [3-5]. This memory requirement, in turn, constrains the achievable image resolution for ptychography and hinders real-time reconstruction. Unfortunately, ptychographic reconstructions often fail to guide data acquisition because the reconstructions are too slow due to the memory constraints and the significant number of computations needed to process the data.

To address the memory constraint and slow reconstruction issues, this paper presents a halo gradient exchange method for large-scale 3D ptychographic reconstruction. The halo gradient exchange method decomposes the image to be reconstructed into spatially contiguous tiles in the shape of squares. Similarly, it also decomposes probe locations into groups in the same pattern. Then, it equally distributes the tiles and the probe locations among Graphical Processing Units (GPUs). When a probe location is large and is not fully circumscribed by the tile, such as with defocused datasets, each tile is extended with halos so that an extended tile covers the range of all probe locations in the group assigned to the GPU. Figure 1(a) shows an example scan pattern for 4 overlapped probe locations, and each probe location is colored differently. Figure 1(b) shows that the image to be reconstructed and the 4 probe locations are equally distributed among 4 GPUs into tiles, and each tile is extended with halos to cover the probe locations. With the above design, each GPU, thereby, only receives a small portion of data needed for image reconstruction, achieves efficient storage of ptychography data in the GPU memory, and significantly reduces the GPU memory footprint. Then, the halo gradient exchange method lets each GPU perform independent but parallel reconstruction for its assigned extended tile until the algorithm converges. Finally, the GPUs abandon the halos but stitch together their tiles, and the stitched result is the same as the ground truth reconstruction without any decomposition.

An issue with the above design, however, is that the reconstructed tiles across GPUs conflict with each other, especially for voxels near the edge of the tiles. In the example of Figure 1(c), the purple shaded region should be reconstructed from both the yellow and the red probe locations without any decomposition. With the image tiles and probe locations decomposition, the same color region now has two but conflicting copies, one with GPU 1 and one with GPU 2, and each copy is reconstructed from incomplete diffraction data (either the yellow or the red probe location alone). To address this conflict, the halo gradient exchange method uses image gradients to synchronize reconstructions for different

tiles across GPUs by adding image gradients from each GPU's extended tile to its conflicting copies at other GPUs. Then, the GPUs synchronize their tiles and resolve conflicts by updating the tiles with the added image gradients.

The synchronization overhead of adding image gradients from one GPU to all other GPUs that have conflicts can be overwhelming. To minimize synchronization overhead, the halo gradient exchange method creates three private memory buffers for each GPU, a send buffer, a receive buffer and an accumulated gradients buffer. The accumulated gradients buffer keeps track of the gradient changes when the GPU updates its extended tile. Then, the gradient changes stored in the accumulated gradients buffer are transferred to the send buffer, which in turn transfers the accumulated gradients to the receive buffers at other remote GPUs that have a conflict. As the GPUs send and receive accumulated gradients through the send and receive buffers in the backend to resolve the conflicts, the GPUs are performing reconstructions in parallel for their extended tiles in the forefront. With this mechanism, the communication cost across GPUs can be hidden and a linear parallel scaling can be achieved on a small cluster of GPUs.

We perform evaluations on two datasets, a small lead titanate ($PbTiO_3$) simulated dataset with 462 probe locations and a 1024 by 1024 by 100 reconstructed volume, and a large $PbTiO_3$ simulated dataset with 4158 probe locations and a 3072 by 3072 by 100 reconstructed volume. For the small dataset, the halo gradient exchange method diminishes the required memory footprint by 21 times from 8.3 GB to 0.39 GB on 24 GPUs. The reconstruction time for the halo gradient exchange method is 18 times faster than the state-of-the-art parallel ptychographic reconstruction algorithm, PtyGer [4, 5], on the same number of GPUs with a parallel scaling efficiency close to 100%. For the large dataset, the halo gradient exchange method diminishes the required memory footprint by 34 times from 75.6 GB (extrapolated) to 2.2 GB on 24 GPUs.

This paper proposes a novel gradient-based image decomposition method for 3D ptychographic reconstruction and distribute diffraction measurements and image tiles among GPUs, so that the memory footprint for each GPU is significantly reduced and data are efficiently stored in GPU's memory. In addition, this paper also presents an asynchronous communication framework that uses extra memory buffers to hide communication cost across GPUs in the background while simultaneously performing reconstructions across GPUs, so that the communication cost across GPUs is minimized and linear parallel scaling can be achieved on a cluster of GPUs.

https://doi.org/10.1017/S1431927622011564 Published online by Cambridge University Press

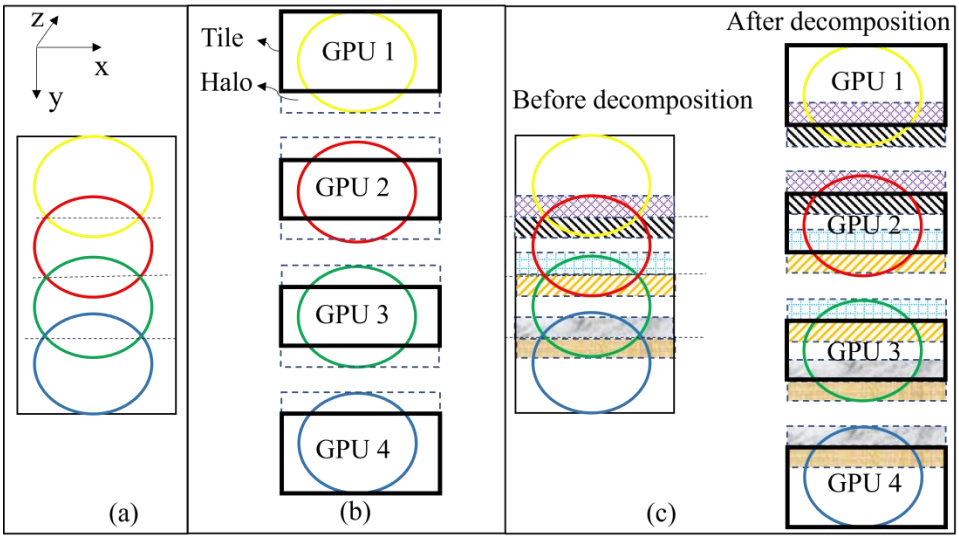

Figure 1. (a) An example scan pattern with 4 overlapped probe locations with a different color for each. (b) Each GPU is assigned with a tile and a probe location. Each tile is also extended with halos to cover the entire range of the probe location. (c) Reconstructions among GPUs conflict with each other, such as the same color shaded regions in the figure, because each GPU only has partial diffraction measurements.

References:

[1] Y Jiang et al., Nature **559** (2018), p. 343. DOI: 10.1038/s41586-018-0298-5
[2] Z Chen et al., Science **372** (2021), p. 826. DOI: 10.1126/science.abg2533
[3] YSG Nashed et al., Optics Express **22** (2014), p. 32082. DOI: 10.1364/OE.22.032082
[4] X Yu et al., In Proceedings of the ACM International Conference on Supercomputing (2021), p. 354. DOI: 0.1145/3447818.3460380
[5] X Yu et al., (2021). arXiv:cs.DC/2106.07575

https://doi.org/10.1017/S1431927622011564 Published online by Cambridge University Press