

Petascale XCT: 3D Image Reconstruction with Hierarchical Communications on Multi-GPU Nodes

Mert Hidayetoğlu, Tekin Bicer[†], Simon Garcia de Gonzalo[‡], Bin Ren[§], Vincent De Andrade[†],
Doga Gursoy[†], Raj Kettimuthu[†], Ian T. Foster[†], and Wen-mei W. Hwu

University of Illinois at Urbana-Champaign, USA

[†]Argonne National Laboratory, IL, USA,

[‡]Barcelona Supercomputing Center, Spain

[§]College of William & Mary, VA, USA

Abstract—X-ray computed tomography is a commonly used technique for noninvasive 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 $9K \times 11K \times 11K$ voxels, where the total reconstruction time is under three minutes using 24,576 GPUs, reaching 65 PFLOPS: 34% of Summit’s peak performance.

I. INTRODUCTION

Synchrotron light source facilities around the world help tens of thousands of researchers every year carry out extremely challenging experiments and ground-breaking research. X-ray computed tomography (XCT) is one of the widely used imaging modalities at synchrotron light sources for imaging materials, samples and biological specimens in 3D with high temporal and spatial resolution, ranging from several micrometers down to sub-20 nanometer resolution. The high-energy synchrotron X-ray sources, such as the Advanced Photon Source (APS) at Argonne National Laboratory (ANL), enable imaging thick specimens that can yield massive amounts of measurements, exceeding tens of GB/s rates and producing TBs-scale data per experiment [1]. For example, imaging a single adult mouse brain of a few centimeters diameter at μm resolution requires a “tiled” tomography experiment that produces more than 1.7 TB ($9K \times 11K$ images with $4.5K$ angles) measurement data. Further, the reconstruction of such data generates more than 4.3 TB 3D volumetric image (with $9K \times 11K \times 11K$ voxels) [2]. High-performant scalable solvers are needed to reconstruct these large experimental datasets, especially considering that advancements in domain sciences, such as neuroscience, require imaging and reconstruction of many of these samples.

Due to experimental constraints, such as extreme conditions (high radiation dose) or physical limitations (vibrations or

drifts during data acquisition), produced data can be far from the ideal and can consist of noisy images with undesired artifacts. These imperfect measurements can adversely impact the reconstruction process resulting in unsatisfactory output. The ability to mitigate such noise and artifacts are important considerations when it comes to choosing between the two broad categories of reconstruction approaches: analytical methods and iterative solvers. While analytical methods, such as filtered-backprojection, are typically fast algorithms, they produce sub-optimal reconstructions with imperfect (noisy) measurement data. In contrast, iterative tomographic reconstruction solvers enable integration of advanced regularizers and models, and iteratively reach for a solution by solving an optimization problem. They also provide an essential framework for incorporating imperfections into the model and therefore, mitigate these artifacts and achieve the desired resolution. However, such improved reconstruction comes at the expense of significantly higher computational cost.

Iterative reconstruction solvers have so far been used for small/medium scale tomography datasets mostly due to the aforementioned computational costs. Different parallelization methods have been developed to ease this cost, from naive data-parallel approaches that process different slices of the measurement data (sinograms) in parallel to advanced in-slice and memory-centric approaches [3]–[5]. For parallel beam geometry, sinogram-based reconstruction methods exploit data parallelism by reconstructing each 2D slice (sinogram) independently. After the reconstruction is done, all reconstructed slices (tomograms) are gathered to a 3D volume. This approach provides reasonable execution time for most smaller datasets, but larger datasets require more aggressive parallelization. In-slice parallelization techniques improve the speed of single sinogram reconstruction by distributing parts of a sinogram to several processes, but introduce synchronization and communication overheads during iterations since the processes that are involved in the processing of the same sinogram need to combine their intermediate results. The MemXCT advanced memory-centric parallelization technique [4] mitigates the synchronization and communication overheads by using memoization, and provides efficient sparse matrix representations and communication patterns.

While the MemXCT parallelization technique provides sig-

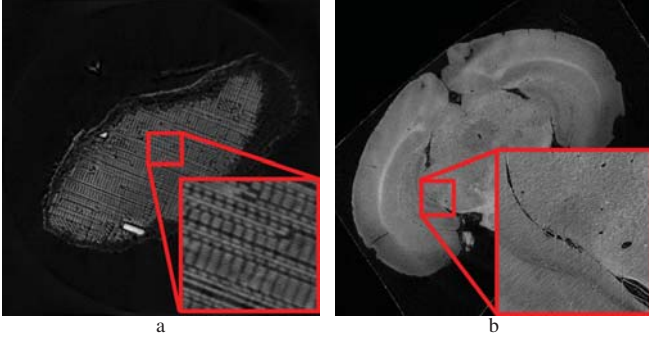


Fig. 1: (a) Tilted slice of an IC chip reconstruction and (b) horizontal slice of a mouse brain reconstruction.

nificant performance improvement compared to its alternatives, reconstruction of full-sized volumes of extreme-scale samples still requires long processing time. For example, authors in [4] report that reconstruction of a single mouse brain sinogram (a slice of the measurement data) requires 10 secs using 256K-cores at Theta supercomputer at ANL. The full reconstruction of the sample (9K sinograms) requires more than 25 hours with the whole supercomputer. Fig. 1(a) and Fig. 1(b) show reconstructions of a medium-scale integrated circuit (IC) and a large-scale mouse brain tomograms (slices of 3D images), respectively. The reconstruction quality is extremely important to distinguish features for these samples (transistors and wires for IC, and blood vessels and myelinated axon tracts for brain sample), therefore iterative solvers are preferred method for reconstruction. Note that a typical science study will likely require the full reconstruction of many samples.

Large-scale GPU resources, such as the Summit supercomputer at the Oak Ridge National Laboratory (ORNL) provide opportunities to apply iterative reconstruction algorithms to extremely large tomography datasets. However, further optimizations and advanced parallelization techniques must be used to achieve maximum efficiency on such resources. In this paper, we introduce novel optimizations for state-of-the-art memory-centric iterative reconstruction approaches to enable reconstruction of extremely large tomography datasets. Specifically, we make the following contributions:

- We introduce improved data partitioning and parallelization techniques that extend 2D MemXCT *data parallelism* with 3D *batch parallelism*. Our approach enables optimized SpMM operations that perform common computational kernels on different voxels (fusing), while performing memoization of irregular data accesses at 3D space (compared to 2D plane in MemXCT).
- We present a hierarchical communication strategy that exploits multi-GPU node architecture. Our approach leverages asynchronous multi-level data reduction to minimize communication overhead between nodes.
- We propose mixed-precision implementation that performs computations using single (or double) precision operations, whereas it stores and communicates data in half precision

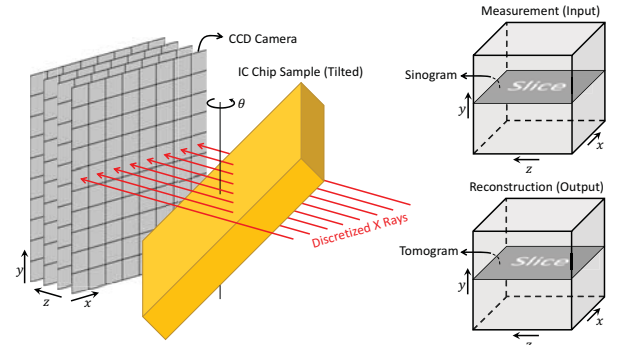


Fig. 2: An experimental setup that shows tomographic data acquisition.

for reduced memory and communication footprint.

- We provide a comprehensive evaluation of our optimizations and system with four real-world tomography datasets using up to 24K NVIDIA V100 GPUs and demonstrate sustained peta-scale performance. In particular, we reconstruct a mouse brain volume with $9K \times 11K \times 11K$ voxels within 3 minutes using the whole Summit supercomputer, reaching 65 PFLOPS throughput (or 34% of Summit's theoretical peak performance).

II. BACKGROUND

In this section, we briefly explain tomographic data acquisition with synchrotron light sources and the basics of iterative reconstruction process.

A. Tomography Experiments

During a tomography experiment, a sample is placed on top of a rotation stage and exposed to X-ray beams. As X-ray beams travel through sample, the photons are attenuated by the sample according to Beer-Lambert law [6], [7]. The attenuated beams are, then, measured at the detector and an X-ray *projection* of the sample is recorded at the detector, as illustrated in Fig. 2. This process is repeated for different rotational views of the sample, θ , with the aim of meeting Crowther criterion [8]. Consequently, this process generates a set of projections, p_θ , where, for example, $\theta = \{0^\circ, 1^\circ, 2^\circ, \dots, 179^\circ\}$ degrees.

Iterative tomographic reconstruction aims to solve an optimization problem. The following cost function is often used:

$$\hat{x} = \underset{x \in C}{\operatorname{argmin}} \{ \|y - Ax\|_2^2 + R(x) \} \quad (1)$$

Here, y is the measurement data, x represents the estimated versions of the unknown object, C is a constraint on x and $R(x)$ is a regularizer function. A is the system matrix or forward operator that depicts how the X-ray beams intersect the voxels in each rotational view and thus the relationship between x and y , i.e. the sinograms and the object, respectively. The solution \hat{x} is the version of the estimated object that minimizes the cost function.

A generic optimization solver requires computing the gradients and updating of the object based on those gradients in an iterative fashion. First, the *forward operator*, A , is applied

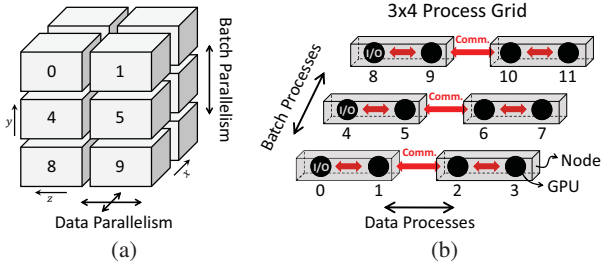


Fig. 3: (a) 3D domain partitioning and (b) assignment on GPUs.

to previously estimated x_i object, then the result is subtracted from sinograms, $y - Ax_i$, and residual r_i is computed based on a norm, e.g., Euclidean norm as in equation 1. Next, r_i is *back projected* on x_i to compute the gradients. Finally, the estimated object x_i is updated based on the computed gradients and the new object x_{i+1} derived for next iteration. In this work, we consider parallel beam geometry for experiments at the synchrotrons and that all beams travel perpendicularly to the axis of rotation, which enables independent reconstruction of slices in 3D object volume from their corresponding sinograms. We also perform an optimized version of Siddon's algorithm for accurate forward and back projection operators [9].

B. Challenges and Opportunities

The sparse system matrix A in Eq. 1 represents the incident voxels of each ray, and therefore its memory footprint can be large. Forward and backprojection operators perform irregular accesses to A and x_i while computing residuals and updating objects. These operations introduce significant overhead during the iterations. Further, if the A matrix cannot fit into available memory, incident rows and columns need to be computed repetitively before each forward and backprojection operation. Furthermore, if some A and x elements involved in an inner product calculation performed by a process are in the memory of another process, that missing information must be communicated.

Prior work MemXCT addresses communication and irregular data access overheads using multi-level Hilbert ordering to maximize the likelihood that all system matrix A elements involved in an inner product of the sparse matrix-vector multiplication are in the same partition [4]. It also avoids repetitive computation of A with memoization while partitioning large memory footprint to many processes. However, the MemXCT approach processes each sinogram independently of others, hence the optimizations are performed only within 2D slices.

In this work, we identify and exploit the following optimization opportunities in 3D:

- Assuming a ray, $u_{i,j}$, where i and j are the location of the ray in the y and x dimensions, respectively, all rays $u_{*,j}$ trace the same voxels in their respective slices for all rotational views. This property can be used to *fuse* rays along the y dimension so that the A elements can be reused when processing these rays.
- MemXCT approaches typically utilize single data type throughout their execution. This can introduce unnecessary overhead during data movement. Mixed precision data types

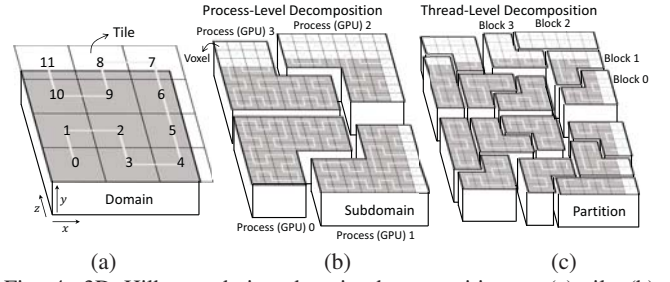


Fig. 4: 3D Hilbert-ordering domain decomposition at (a) tile (b) process, i.e., GPU, and (c) thread, i.e., block, levels.

can minimize data transfer volume while meeting precision requirements of computation.

- Direct communication between reconstruction processes can create redundant overhead. Reducing data between co-located processes can significantly improve data transfer efficiency.

III. ALGORITHM DESIGN & OPTIMIZATIONS

In this section, we present five categories of optimizations that exploit the opportunities identified in Section II.

A. Partitioning and Parallelization

We propose a data partitioning and parallelism arrangement strategy for both input (sinogram) and output (tomogram) data cubes as depicted in Fig. 3(a). The proposed strategy combines two principal parallelism modalities: *batch parallelism* and *data parallelism*.

Batch parallelism takes advantage of the fact that there is no data dependency between the slices in the y direction. Therefore, tomogram and sinogram slices are partitioned equally among *batch processes*¹ and reconstructed in an embarrassingly-parallel fashion. However, because batch parallelism does not involve partitioning of data in the x and z dimensions, the per-process memory footprint of slices may not fit within the GPU memory of batch processes.

Data parallelism solves this problem by partitioning slices among *data processes* in the x - z plane along with their corresponding data structures. As a result, per-process memory footprint is reduced so that it can fit within GPU memory. Nevertheless, data parallelism comes with a communication overhead because of dependency between data partitions in the x and z dimensions. The detailed design of the proposed strategy consists of the following components to minimize communication and maximize data reuse.

1) *Hilbert-Ordering Domain Decomposition*: In order to partition a batch of slices in the x - z plane while maintaining data locality, we extend the MemXCT 2D Hilbert-ordering domain decomposition to 3D, where it is applied to all tomogram and sinogram slices in the y direction. Fig. 3 and Fig. 4 depict decomposition at the process and thread levels: First, both tomogram and sinogram domains are tiled into square patches and ordered with pseudo-Hilbert ordering. These patches are partitioned equally among processes, where each partition

¹In the context of this paper, each process corresponds to a GPU.

corresponds to a *subdomain* as shown in Fig. 4(b). Each subdomain is processed by a single GPU.

Fig. 3(a) shows that a domain is partitioned into 12 subdomains: two in the x direction, two in the z direction, and three in the y direction. Fig. 3(b) depicts assignment of these 12 subdomains to a GPU grid with 12 processes. The GPU grid consists of six nodes and each node contains two GPUs. Since communications within node have a higher bandwidth than that of between nodes, data processes processing adjacent subdomains are placed in the same node whenever possible.

Then each subdomain is partitioned among GPU thread blocks as shown in Fig. 4(c). Data connectivity and locality of partitions provided by the Hilbert-ordering domain decomposition are essential for optimized SpMM design (Sec. III-B) and hierarchical communications (Sec. III-D) – two of the major contributions of this paper.

2) *Batch Processing Pipeline*: To overlap MPI communication and GPU computation, we partition each batch into smaller *I/O batches* that are processed sequentially, i.e., one I/O batch is reconstructed at a time. As discussed in Sec. III-B, optimized SpMM fuses multiple slices in a I/O batch into *minibatches*. Processing of these minibatches are pipelined by overlapping MPI communications and GPU computations as explained in Sec. III-E. To achieve sufficient overlapping, there needs to be at least a few minibatches in an I/O batch.

3) *Optimal Partitioning Strategy*: In order to minimize the communication overhead of data parallelism, it is better to minimize partitioning of the 3D data cube in the x - z dimension; only until per-process memory footprint fits into GPU memory. Then batch partitioning should take over in the y dimension with no additional overhead. The level of batch parallelism is limited by the total number of slices in the y direction and the need to fuse slices so that the sparse matrix A can be reused from register by the optimized SpMM.

TABLE I: Computational Complexity

	Per Process	Total
Comput.	$MN^2/P_b P_d + MN/P_b \sqrt{P_d}$	$MN^2 + MN\sqrt{P_d}$
Memory	$N^2/P_d + N/\sqrt{P_d}$	$N^2 P_b + N P_b \sqrt{P_d}$
Comm.*	$MN/P_b \sqrt{P_d}$	$MN\sqrt{P_d}$

M row channels, N column channels, P_b batch processes, P_d data processes.

*Latency and contention terms are omitted.

4) *Computational Complexity*: The computational cost of a projection depends on size of the 2D detector grid depicted in Fig. 2. Here, M and N are the number of row and column channels in the detector grid. Since each discretized ray measured by a detector propagates through $\mathcal{O}(N)$ voxels, each projection takes $\mathcal{O}(MN^2)$ time. Parallel rays have the same trajectory in all slices, and therefore it is sufficient to store a single sparse matrix with $\mathcal{O}(N^2)$ nonzeros and reuse it from memory for all M slices. The total memory footprint is increased by a factor of batch processes (P_b) because of the sparse matrix duplication. On the other hand, data parallelization partitions the sparse matrix into P_d data processes, where each process computes a partial projection. The geometrical shapes of the partial data is shown in Fig. 7(b). The partial

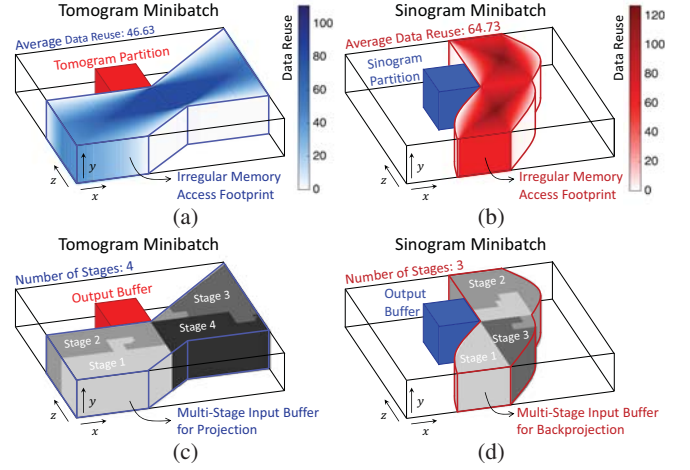


Fig. 5: (a, b) Tomogram and sinogram partitions and respective data access footprints, and (c, d) multi-stage buffer shapes.

data of a processes scales with $MN/\sqrt{P_d}$ because the cross-section of each subdomain on the detector halves only when P_d is quadrupled. To obtain the total projection, the partial data is communicated among all data processes to be reduced at the receiving data processes. This additional communication and reduction is the overhead of data parallelism. Table I summarizes per-process and total computation, memory, and communication complexities of the proposed 3D partitioning.

B. XCT-Optimized SpMM Design

By considering memory access patterns specific to XCT, we optimize the proposed mixed-precision SpMM kernel via a 3D input/output buffering algorithm. This subsection discusses these optimizations and explains our CUDA implementation given in Listing 1: The same implementation is used for both projection and backprojection.

1) *3D Input Buffering*: A naive SpMV implementation suffers not only from irregularity but also redundant accesses to tomogram and sinogram data because each voxel is (irregularly) accessed more than once by different threads. As a remedy, our optimized kernel stages data accesses in the GPU shared memory via 3D input buffers. Fig. 5 illustrates the basic idea of our optimization by depicting memory access footprints for a $256 \times 256 \times 50$ minibatch; (a) shows tomogram voxels accessed by a sinogram partition and (b) shows sinogram voxels accessed by a tomogram partition, where each partition is computed by a single thread block. In 3D input buffering, each thread block loads the highlighted input data (Fig. 5: (a) and (b)) to shared memory once, and reuses it (irregularly) from shared memory. As a result, memory accesses are reduced by the factor of data reuse shown in the figure, i.e., darker shade represents higher data reuse from shared memory. With these reuses, the GPU performance becomes limited by the memory bandwidth consumed when reading A and A^T from memory (see the roofline analysis in Sec. IV-C1). We overcome this memory-bandwidth bottleneck by reusing A elements from registers as we discuss next.

2) *Register Reuse*: Even with input buffering, the optimized SpMV utilizes only less than two percent of the GPU’s theoretical peak compute throughput because of its low arithmetic intensity (FLOPS/byte). This is because each fused multiply add (FMA) requires two data elements from memory: (1) shared-memory *index* of the visited voxel by X-ray and (2) intersection *length* of the ray going through the voxel. To increase the arithmetic intensity, we propose to reuse index and value data from registers. That is, many SpMVs in a minibatch are fused as $AX = B$, where each column of X and B corresponds to a slice of tomogram and sinogram values in the minibatch, respectively. As a result, the arithmetic intensity of the operation increases by the fusing factor, a.k.a., minibatch size. However, the fusing factor cannot be arbitrarily large because fusing imposes register pressure as discussed next.

3) *3D Output Buffering*: To provide data reuse from register, each thread loads one index and one length at a time and then reuses this pair from register for all corresponding output voxels in the y direction of the minibatch. To accumulate the partial sums, each thread allocates an output buffer (`acc` in Line 10 of Listing 1). Line 28 shows how a thread reuses the index and value pairs to access its corresponding data in the 3D input buffer. However, since each streaming multiprocessor (SM) has a limited number of registers that are used by all threads, enlarging the minibatch size, i.e., `FFACTOR`, can increase the number of registers required by each thread and hence elevate register pressure on the SMs. When threads collectively use more registers than available in SM, register contents are spilled to slower memory and hamper GPU performance. Results show that GPUs are able to obtain 34% of their theoretical compute throughput by carefully tuning the minibatch size as shown in Sec. III-B

Listing 1: Optimized SpMM Mixed-Precision Kernel

```

1  //MATRIX STRUCTURE
2  struct matrix{ unsigned short ind; half len; };
3  //PROJECTION KERNEL
4  __global__ void kernel_project(half *y, half *x, matrix *mat,
5                                int *displ, int numrow, int numcol,
6                                int *buffdispl, int *buffmap,
7                                int *mapdispl, int *mapnz,
8                                int bufsize ){
9      extern __shared__ half shared[];
10     float acc[FFACTOR] = 0.0;
11     int wind = threadIdx.x*WARP_SIZE;
12     for(int buff = buffdispl[blockIdx.x]; buff <
13         buffdispl[blockIdx.x+1]; buff++){
14         int mapoffset = mapdispl[buff];
15         for(int i = threadIdx.x; i < mapnz[buff]; i += blockDim.x){
16             int ind = buffmap[mapoffset+i];
17             #pragma unroll
18             for(int f = 0; f < FFACTOR; f++){
19                 shared[f*bufsize+i] = x[f*numcol+ind];
20             }
21             __syncthreads();
22             int warp = (buff*blockDim.x+threadIdx.x)/WARP_SIZE;
23             for(int n = displ[warp]; n < displ[warp+1]; n++){
24                 matrix mat = indval[n*WARP_SIZE+wind];
25                 float len = __half2float(mat.len);
26                 #pragma unroll
27                 for(int f = 0; f < FFACTOR; f++){
28                     acc[f] += __half2float(shared[f*bufsize+mat.ind])*len;
29                 }
30             }
31             __syncthreads();
32         }
33         int row = blockIdx.x*blockDim.x+threadIdx.x;
34         if(row < numrow){
35             #pragma unroll
36             for(int f = 0; f < FFACTOR; f++){
37                 y[f*numrow+row] = __float2half(acc[f]);
38             }
39         }
40     }

```

4) *Multi-Stage 3D Buffering*: Each SM has a limited capacity of shared memory (96 KB for V100 GPUs), which is not enough for large memory access footprint of a 3D partition. Therefore we provide access to input data in multiple stages. Fig. 5: (c) and (d) show four-stage and three-stage bufferings for projection and backprojection, respectively. Each stage loads 96 KB data whose shapes are governed by Hilbert ordering as described in Sec. III-A1. Line 12 in Listing 1 shows the iterations over stages, where only a portion of the 3D input buffer is loaded by mapping `buffmap`, and only a partial value of the 3D output buffer is computed in each stage. Because each staging has a synchronization overhead (seen in Lines 21 and 30), fewer number of stages is desirable. The 3D buffering strategy increases the number of stages since memory footprint of input buffers grow with increasing minibatch size (in the y direction). We mitigate increasing memory footprint using mixed precision implementation as explained in the following subsection.

C. Mixed-Precision Implementation

Mixed-precision implementation stores and communicates data in half precision (16-bit) and performs all arithmetic operations with single precision (32-bit) through in-core data conversions as in highlighted in Listing 1 (red). This approach reduces memory and communication footprint as well as provides a higher GPU throughput by moving data in half-precision while maintaining numerical accuracy by performing FMAs in single-precision. Several issues need to be addressed while using mixed precision implementation: (1) Half-precision has a lower range and quantization, and therefore it can suffer from overflow and underflow. (2) When reading sparse matrix A from memory, each warp of 32 threads accesses only 64 bytes of data at a time, which only utilizes half of the 128-byte GPU cache-line and therefore results in sub-optimal cache utilization. We address these issues with *adaptive normalization* and *data packing*, respectively.

1) *Adaptive Normalization*: We avoid the half-precision underflows by artificially increasing the voxel size in the tomogram. Since the nonzero sparse matrix values represent the travel lengths of incident X-rays per voxel, increasing the voxel sizes results in larger half-precision values and hence avoids underflows. On the other hand, the overflows are handled by maintaining input and output data in double or single precision before performing kernel operations. We accomplish this by normalization and denormalization of input and output data before and after type castings, respectively. The (de)normalization factor is adaptively changed in each iteration with respect to the max-norm of the evolving input vector to prevent overflows while minimizing underflows.

2) *Data Packing*: The underutilization of GPU cache line not only wastes memory bandwidth but also introduces latency. As a remedy, we pack both index and length data elements into a single structure of four bytes. This structure is shown in Line 2 of Listing 1: It consists of an unsigned two-byte integer that represents the voxel index in shared-memory, and a two-byte half-precision floating point number representing the length.

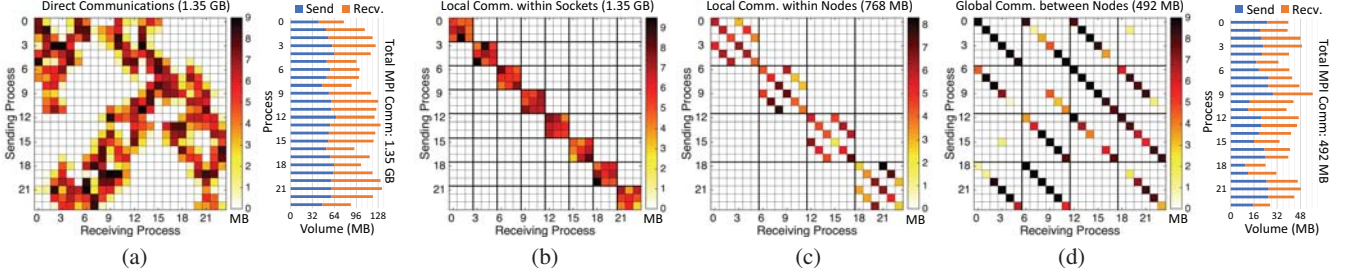


Fig. 6: (a) Direct communication and load balancing. Three-level hierarchical communications: (b) Socket-level communication; (c) Node-level communication; (d) Global communication and load balancing.

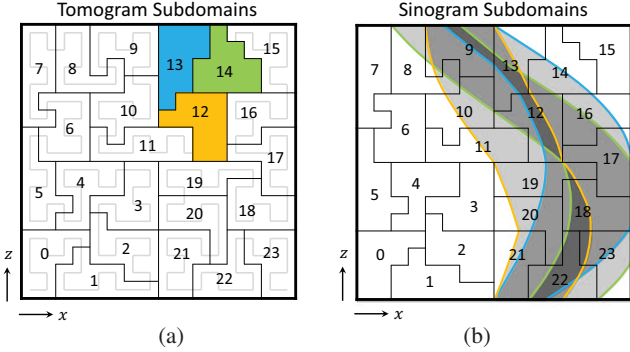


Fig. 7: Tomogram (a) and sinogram (b) subdomains. Partial data footprint of tomogram subdomains 12–14 are shown in (b).

As a result, each warp of 32 threads utilizes the full 128-byte cache line.

D. Hierarchical Communications

Data must be partitioned when solving large problems in order to fit per-process memory footprint within available GPU memory. However, data partitioning requires communication and reduction of partial data, as elaborated in Sec. III-A4. Consequently, image reconstruction time is dominated by communication overhead for large problems. As a solution, we propose a novel hierarchical communication with partial data reduction, another major contribution of this paper. This strategy is designed for multi-GPU node architectures where high-bandwidth connections between *peer* GPUs are exploited. This section describes direct communication of partial data (the baseline), the main idea behind local communication and reduction of partial data, our communication hierarchy, and its efficient implementation.

1) *Direct Communications*: Fig. 7 shows partitioning of (a) tomogram data and (b) sinogram data into 24 subdomains. The subdomain shapes are governed by Hilbert-ordering domain decomposition (Sec. III-A1). In projection, each process computes only a partial sinogram data which needs to be communicated and reduced to find total sinogram data. As an example, Fig. 7(b) shows the partial data footprints of processes 12–14 on sinogram subdomains, e.g., process 12 sends its corresponding partial data to processes 8–13 and 18–23. The resulting communication matrix is shown in Fig. 6(a). Communication volume between two processes depends on

the amount of overlapped area between sender’s partial data footprint and receiver’s sinogram subdomain. As a result, the communication is sparse and irregular. Following the communication, receiving processes *reduce* (sum) the overlapping partial data coming from sender processes, which completes projection operation. This description is also valid for backprojection as it is a transpose of projection.

Fig. 6(a) also shows communication volume of each process and total amount of communicated partial data. Large problems communicate a large portion of data through slow interconnect between nodes, which constitutes a dominating bottleneck. We relieve this bottleneck by reducing partial data locally within nodes and communicating the reduced data between nodes, as we shall discuss next.

2) *Local Reduction of Partial Data*: The proposed hierarchical communication exploits high-bandwidth connections between GPUs within nodes by a *local* communicator and reduction of the overlapping partial data among sender processes. To explain, we consider processes 12–14 (Fig. 7(a)) located in the same node. Overlapping portions of their partial data (Fig. 7(b)) are communicated and reduced within the node, rather than communicating the original partial data among nodes directly. Then, the reduced partial data within the node is sent to receivers by a *global* communication, where they are further reduced to find total sinogram values. The locality of subdomains provided by Hilbert ordering is essential in this hierarchical communication because spatially-local subdomains yield more overlapping data and more local partial reductions. Similarly, fatter nodes with more highly-connected GPUs yield more local reduction of partial data.

3) *The Three-Level Hierarchy*: This paper considers Summit’s node architecture, where each node has two CPU *sockets*. Each socket is connected to three GPUs that are densely-connected with high-bandwidth interconnect. The CPU sockets within a node are connected with a slower data bus. Each node is connected to an even slower infiniband network. Although, we explain our hierarchical communication and reduction in the context of the Summit architecture, the method is general and applicable to other node architectures with different number of sockets and GPUs.

Since the data bus between sockets is slow, we do not perform local reduction directly among six GPUs within a node. Instead, we first perform a socket-level communication and reduction among three GPUs within a socket. Then, an

additional node-level communication and reduction among six GPUs within a node. Finally, a global communication and reduction among all GPUs between nodes is performed.

Fig. 6(b) presents the socket-level communication matrix as a block-diagonal structure because each GPU talks only to three GPUs (including itself) in the same socket. The socket-level reduction reduces the original 1.35 GB partial data down to 768 MB (43% reduction). Then, the reduced partial data is further reduced within nodes among six GPUs. Fig. 6(c) shows the intra-node local communication matrix. After the node-level communication, the partial data is further reduced to 492 MB (36% additional reduction). Finally, GPUs send reduced partial data only among nodes as shown in Fig. 6(d). As a result, the total data communicated among slowly-connected nodes is reduced by 64% compared to direct communication.

4) *Efficient Implementation*: To implement the proposed three-level hierarchical communication, we leverage `MPI_Comm_split` to define local communicators within sockets and nodes, and a global communicator among nodes, respectively. To prevent data from being staged through the CPU during local communications, we use CUDA inter-process communication (IPC) capability, i.e., communicating data directly from GPU to GPU bypassing CPU. Furthermore, local communications are overlapped using CUDA streams. Then local reductions are performed on GPUs. Global communications are implemented with MPI with CPU staging through pinned buffers. We employ non-blocking `MPI_Issend` and `MPI_Irecv` for overlapping global communications. The partial data is then moved back to the GPUs at the receiving processes for global reductions.

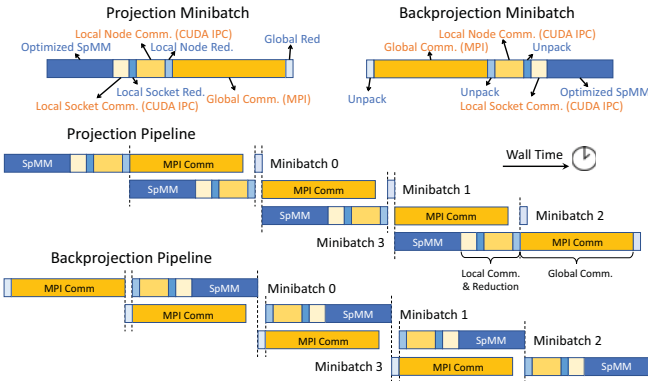


Fig. 8: Overlapping four minibatches in the batch processing pipeline.

E. Communication Overlapping

Even when communication time is reduced by applying hierarchical communication strategy, iterative solution time remains bounded by MPI communications for large problems (see Fig. 11). To further alleviate this communication bottleneck, we propose an overlapping strategy that exploits multiple slices in a reconstruction batch which is only possible in 3D image reconstruction. That is, global (MPI) communication of a minibatch is overlapped with local operations of another minibatch involving the optimized SpMM kernel,

local reduction kernels, and local (socket-level and node-level) communications. Fig. 8 depicts overlapping of four minibatches during projection and backprojection operations. Projection overlaps global MPI communication of a minibatch with local operations of the next minibatch. On the other hand, backprojection overlaps local operations with next minibatch's MPI communication. This strategy is the most effective when kernel time and communication time are comparable.

IV. EXPERIMENTAL RESULTS

This section introduces an extensive evaluation of our approaches. We first evaluate the performance of our optimizations and designs on Summit supercomputer using four real-world experimental datasets. Then, we investigate strong and weak scaling properties of our system. Finally, we compare convergence of our optimizations with four different levels of precision.

A. Experimental Setup

1) *Summit Supercomputer*: All experiments are conducted on Summit supercomputer at computing facility of ORNL. Summit consists of 4,608 IBM AC922 nodes with two POWER9 CPUs and six NVIDIA V100 GPUs per node. Nodes are connected via dual-raid fat-tree network. Each CPU in the node is densely-connected to three GPUs with Nvidia NVlinks, where each link has 50 GB/s one-way and 100 GB/s bidirectional bandwidth. We refer to each CPU and its corresponding densely-connected GPUs as a *socket* in this paper. Each node consists of two sockets. Sockets are connected via an X bus with 64 GB/s bidirectional theoretical bandwidth. Each CPU has 22 cores and 250 GB memory and each GPU has 80 SM and 16 GB memory.

TABLE II: Datasets and Memory Footprints

Sample	Measurement Data Cube ($K \times M \times N$)	I/O Data Footprint	Memory Footprint
Shale Rock	$1501 \times 1792 \times 2048$	52.1 GB	120 GB
IC Chip	$1210 \times 1024 \times 2448$	36.7 GB	139 GB
Activated Charcoal	$4500 \times 4198 \times 6613$	1.23 TB	2.82 TB
Mouse Brain	$4501 \times 9209 \times 11283$	6.56 TB	10.9 TB

2) *Datasets Used for Experiments*: Table II characterizes the four datasets used in our evaluation². Dimensions are given in $K \times M \times N$, where K denotes the number of projections and M and N denote the number of vertical and horizontal channels in the 2D detector grid, respectively. All measurements follow parallel beam scan geometry as described in Sec. II. The corresponding I/O data and memory footprints are given for all datasets for single precision. All measurements are obtained by experiments at APS, ANL. The Shale and Charcoal datasets are open, while Chip and Mouse are proprietary [2], [10], [11]. Even though IC Chip and Shale have similar dimensions, we prefer to provide computational performance for the freely available Shale for benchmarking purposes. Nevertheless, we use IC Chip for the convergence

²In the rest of the paper, we refer to these datasets as Shale, Chip, Charcoal, and Brain, according to their order in Table II.

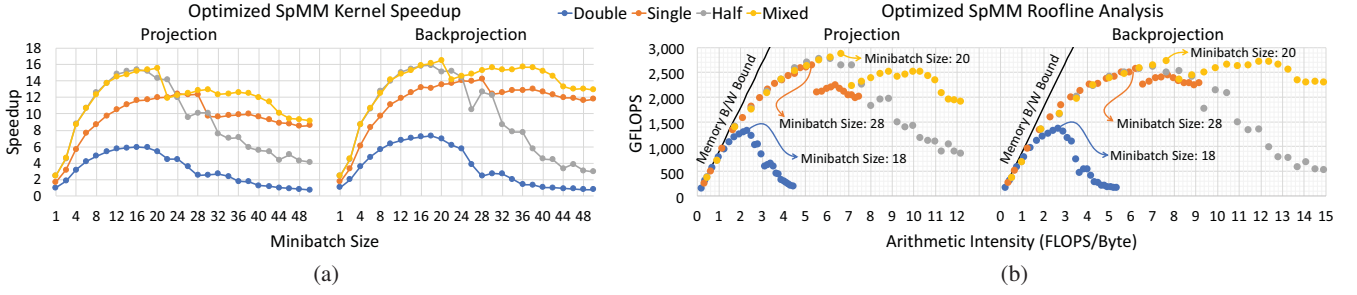


Fig. 9: (a) Speedup and (b) per-GPU performance of XCT-Optimized SpMM kernel as a function of the slice fusing factor, i.e. minibatch size. Speedup is based on double-precision projection with no optimization (fusing factor 1).

TABLE III: Overall Reconstruction Speedup

Part. Opt.	Prec.	Shale on Four Nodes			Charcoal on 128 Nodes		
		Part.*	Recon.	Speed.	Part.*	Recon.	Speed.
Part. + Kernel Opt.	Double	1×(4×6)	979 s	1×	1×(128×6)	78.4 m	1×
	Single	2×(2×6)	405 s	2.42×	2×(64×6)	31.3 m	2.51×
	Mixed	4×(1×6)	215 s	4.56×	4×(32×6)	15.1 m	5.20×
Part. + Kernel + Comm. Opt.	Double	1×(4×6)	513 s	1.91×	1×(128×6)	58.4 m	1.34×
	Single	2×(2×6)	134 s	7.30×	2×(64×6)	20.4 m	3.85×
	Mixed	4×(1×6)	51.1 s	19.2×	4×(32×6)	8.0 m	9.78×
Part. + Kernel + Comm. Opt.	Double	1×(4×6)	218 s	4.49×	1×(128×6)	27.0 m	3.00×
	Single	2×(2×6)	76.5 s	12.79×	2×(64×6)	10.0 m	7.87×
	Mixed	4×(1×6)	42.2 s	23.19×	4×(32×6)	4.30 m	18.19×

*Total Number of Partitions = Batch Nodes × (Data Nodes × Partitions per node). Data partitions per node is set to six because each node consists of six GPUs.

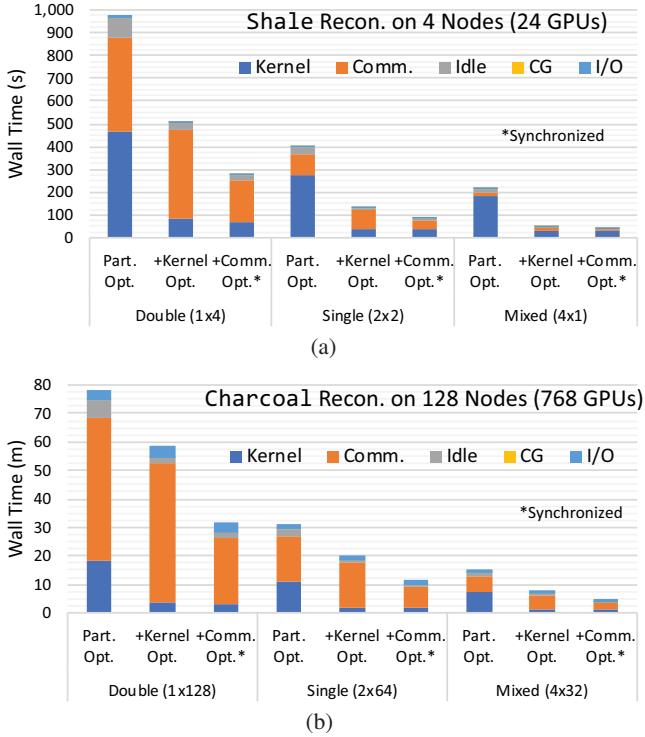


Fig. 10: Breakdown of end-to-end reconstruction times for (a) Shale and (b) Charcoal. *Overlapping does not applied (i.e., communications are synchronized) to measure each portion's time.

results (Sec. IV-F) because it is a numerically challenging case with contaminating noise.

B. Overall Reconstruction Performance

This subsection evaluates the overall reconstruction speedup with our optimizations. Table III reports end-to-end reconstruction time for Shale on four nodes and Charcoal

on 128 nodes. These are the minimum number of nodes required to fit the corresponding memory footprints of double-precision reconstructions. The first three rows report the result without optimized SpMM, hierarchical communications, or communication overlapping. We only optimize the baseline implementation so that it uses the optimal combinations of batch and data parallelism according to the level of precision used for storing the data in memory. That is, lower precision representations shrink the memory footprint, allowing more batch parallelization and less data partitioning. As explained in Sec. III-A3, data partitioning comes with communication overhead and thus it is desirable to employ more batch parallelism and less data partitioning. In this example, as shown in Table III, we perform partitioning in node granularity, where each node handles six data partitions (one per GPU). The batch nodes partition slices in the 3D reconstruction domain (data duplication, no communication) and data nodes partition each slice (data partitioning, communication). The next three rows apply kernel optimizations via optimized SpMM implementation (Sec. III-B). The last three rows apply hierarchical communications (Sec. III-D) and overlapping (Sec. III-E). As seen in Table III, overall speedup over double-precision baseline is 23.19× and 18.19× for Shale and Charcoal reconstructions, respectively.

To further study the effect of each optimization, Fig. 10 shows the breakdown of the end-to-end reconstruction time. The communications are synchronized (i.e., not overlapped) in order to accurately measure the time taken by each activity. However, we do not synchronize GPU kernels and thus idle time reflects the corresponding load imbalance. 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. Sec. IV-C and Sec. IV-D further analyze the optimized SpMM and communication optimizations in detail.

C. Optimized SpMM Performance

Fig. 9(a) shows optimized SpMM speedup with varied minibatch sizes: Performance increases in all cases with larger minibatching due to the aforementioned register reuse

TABLE IV: Communicated Data* and Effective System Bandwidth

	Prec.	Socket-Level Comm.		Node-Level Comm.		Global Comm.		Memcpy B/W
		Data	B/W	Data	B/W	Data	B/W	
Direct	Double	N/A		N/A		36.6 TB	1.61 TB/s	35.2 TB/s
	Single					18.3 TB	1.61 TB/s	34.9 TB/s
	Mixed					9.16 TB	1.59 TB/s	34.6 TB/s
Hierar.	Double	36.6 TB	174 TB/s	21.4 TB	21.3 TB/s	15.2 TB	1.58 TB/s	34.9 TB/s
	Single	18.3 TB	170 TB/s	10.7 TB	22.8 TB/s	7.58 TB	1.55 TB/s	34.5 TB/s
	Mixed	9.16 TB	164 TB/s	5.35 TB	23.5 TB/s	3.79 TB	1.49 TB/s	33.6 TB/s

*Per projection (and backprojection). Fig. 11 involves 30 projections and 31 backprojections.

provided by our optimized SpMM. However, the kernel performance stagnates when minibatch size reaches around 16, and then it drops with larger minibatch sizes. This is due to a combination of register pressure and synchronization overhead introduced by large minibatch sizes as discussed in Sec. III-B. Nevertheless, minibatch sizes of 18, 28, 16, and 20 provide a maximum of $6.47\times$, $7.77\times$, $6.30\times$, and $6.58\times$ kernel speedup compared to no minibatching with double, single, half, and mixed precision, respectively. Overall, mixed precision achieves the best performance of $15.66\times$ speedup over the double precision baseline.

1) *Roofline Analysis*: To analyze the optimized SpMM performance further, Fig. 9(b) shows the roofline analysis plot, where the horizontal axis shows the arithmetic intensity, i.e., the number of floating-point operations per byte accessed from memory, and the vertical axis shows the GFLOPS performance per GPU. Each data point in the figure corresponds to one data point in Fig. 9(a), where increasing minibatch size increases the arithmetic intensity thanks to the data reuse from registers (Sec. III-B). Lowering precision naturally increases arithmetic intensity due to the less number of bytes per data element. Half and mixed precisions yield the same arithmetic intensity. The memory bandwidth bound (maximum performance possible) with respect to the theoretical 900 GB/s memory bandwidth of V100 GPU is included in Fig. 9(b). This figure shows that the performance with small arithmetic intensities is well-bounded by the memory bandwidth and starts to diverge when intensity increases. This is mainly due to the synchronization overhead of multi-stage input buffering.

Double and half precision performance start to degrade for minibatch sizes larger than 18 because of the register spilling as a consequence of the pressure incurred (see Sec. III-B3). On the other hand, single and mixed precisions do not experience register spilling up to a minibatch size of 50. This is mainly because register usage is well optimized for single precision unrolled FMAs (Line 26 of Listing 1). Nevertheless, performance drops significantly at minibatch sizes of 28 and 20 for single and mixed precisions, respectively. This sharp drop in performance may be due to a change in compiler optimization strategy to prevent register spilling by sacrificing performance under a high register pressure. However, `nvcc` is proprietary, preventing us from further investigation or verification. 16 is set as the minibatch size for all our experiments since the slices count is often a multiplication of 16.

2) *Comparison with cuSPARSE*: We compare our optimized SpMM performance with `cusparseSpMM`, which uses

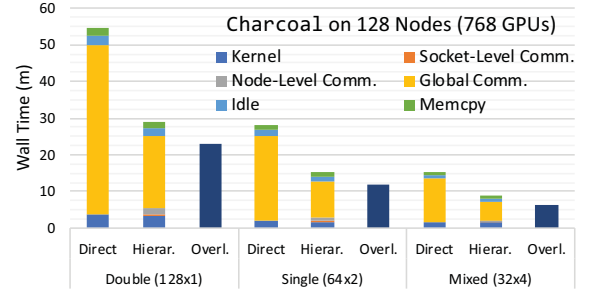


Fig. 11: Communication time breakdown, Charcoal on 128 nodes.

`cuSPARSE` library³. In order to find the best performance of both implementations, we aggressively try all minibatch sizes up to 50 and select the best ones. Our results show that our optimizations provide $1.53\times$ to $2.38\times$ speedup for double and single precision types, respectively. Unfortunately, `cusparseSpMM` is not able to converge to a solution with half precision type. Therefore we cannot make a direct comparison for other types. We are investigating the reason of this.

D. Communication Performance

To investigate communication optimizations further, Fig. 11 shows the breakdown of communication time for Charcoal reconstruction on 128 nodes. Communication time with direct and hierarchical communication strategies (Sec. III-D), and total time with communication overlapping (Sec. III-E) are investigated. Fig. 11 shows that hierarchical communication reduces the total communication time by 52%. Communication overlapping offers an additional speedup of 21% to 29% in total execution time. As opposed to the example in Fig. 8, Charcoal reconstruction has less overlapping opportunities because the global communication dominates the execution, bounding the overlapping efficiency.

Table IV shows the amount of communicated data in TB and corresponding effective system bandwidth in different levels of the communication hierarchy. Communication data is naturally smaller in all cases with lower precision. Effective system bandwidth is calculated by dividing communication data amount by respective communication time shown in Fig. 11. Table IV shows that the effective bandwidth within each socket is about $100\times$ faster than that among nodes. Similarly, the effective bandwidth among sockets is $15\times$ faster than that among nodes. This high bandwidth within nodes alleviates the local communication overhead for hierarchical communications as seen in Fig. 11. Overall, hierarchical communication reduces the communication among nodes by 58% for all precision types.

E. Scaling Performance

To investigate the scaling properties of the proposed application, we perform two strong and one weak scaling experiments with all optimizations applied. Communication overlapping is disabled to correctly measure individual timings. All experiments perform 30 CG iterations with mixed precision.

³<https://developer.nvidia.com/cusparse>

1) *Strong Scaling*: First, we scale 128 slices of *Shale* up to 128 nodes. Fig. 12(a) shows the scaling results. This experiment demonstrates the limited scalability when there is a small number of slices. This experiment scales up to only 128 nodes where each node reconstructs only one slice. In the base case, there are eight minibatches (each contains 16 slices) so the solution can scale efficiently only up to eight nodes (see III-A3). We need to reduce the size of the minibatches after eight nodes to achieve more parallelism by working on finer-granularity. However, SpMM performance drops due to reduced data reuse from registers. As a result, we scale up to 128 nodes (one slice per node) but the performance is suboptimal due to the bottleneck of reduced SpMM performance as shown in Fig. 12(a).

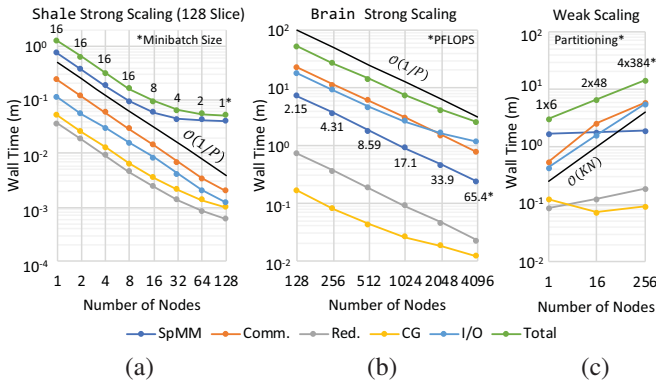


Fig. 12: Strong and weak scaling results: Shale and Brain datasets.

Fig. 12(b) shows the strong scaling result of *Brain* reconstruction from 128 nodes (the minimum number of nodes that *Brain* fits) to 4096 nodes (89% of Summit). As opposed to the previous experiment, there are 9209 slices in *Brain* (see Table II) that allow us to scale without compromising from SpMM performance. As a result, the total reconstruction time is in agreement with $\mathcal{O}(1/P)$ curve, where P is the number of nodes. The I/O performance starts to degrade at 1024 nodes because of the contention that parallel I/O puts into the file system. Nevertheless, thanks to our proposed optimizations, **we reconstruct 3D Brain dataset in 2.5 minutes on Summit**. To the best of our knowledge, this is the largest XCT image reconstruction in near-real time. The optimized SpMM kernel performance reaches 65.4 PFLOPS on 24576 GPUs (six GPUs per node): 34% of Summit's theoretical peak performance.

2) *Weak Scaling*: To investigate the weak scaling properties, we take *Shale* as the base dataset, and then synthetically double each dimension in the aforementioned measurement data cube ($K \times M \times N$). Each time we double all dimensions, and the nominal computation (excluding the parallelization overhead) increases by $16\times$ (see Table I). Accordingly, we increase the number of nodes $16\times$ each time we double measurement dimensions. As discussed in Table I, the memory footprint increases by $8\times$ at each step of the scaling. We apply the suggested optimal partitioning strategy in Sec. III-A3 by partitioning data structures among eight nodes and slices

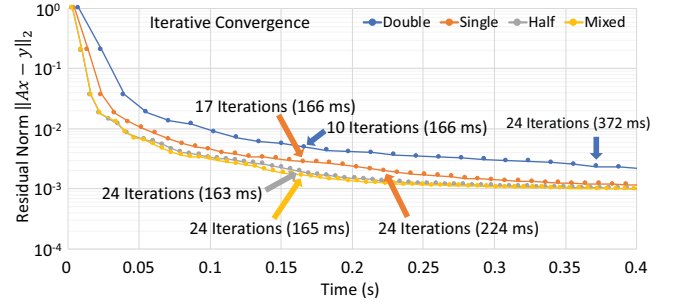


Fig. 13: Convergence speed for *Chip* with various precisions.

between two nodes. Fig. 12(c) shows the weak scaling results. Since nominal computation per node remains the same, SpMM time remains constant accordingly. However, communication and I/O time increases and becomes the bottleneck for large problems. I/O could be further optimized either by a custom design or a high-performance library, which is beyond the scope of this work.

F. Convergence Performance

To investigate iterative convergence rates, we reconstruct a slice from *Chip* dataset. Since this dataset is noisy, iterative solution experiences noise *overfitting* after 24 iterations and even though the residual norm shrinks further, measurement noise manifests into the image reconstruction. To prevent that, we terminate the iterative solution after 24 iterations. No serious convergence problem is observed with reduced precisions. This is mainly because the numerical noise floor is well below the measurement noise level—this applies to other datasets as well. Fig. 13 shows the (relative) residual norm for a single slice with respect to the execution time with double, single, half, and mixed precisions. Mixed (and half) precision performs 24 iterations in 165 ms, while single- and double-precision implementations complete the same number of iterations with 224 and 372 ms, respectively.

V. RELATED WORK

Tomographic reconstruction has been researched extensively over the years. Iterative reconstruction algorithms usually show better image quality compared to analytical approaches [12]–[16] and have been used to accommodate limitations on experimental dataset, e.g. noisy measurements, missing angles and others [17]–[19].

Many parallelization techniques have been developed to ease the computational requirements of iterative algorithms, including distributed computing methods using multi-core [20]–[23]. However, they found limited applicability for large-scale dataset due to the computational requirements.

Many-core architectures, such as GPUs, have been widely used for reconstruction small to medium dataset [24]–[26] and received significant attention in recent years [4], [5], [27]–[29]. Medical imaging is one of the areas that extensively utilizes advanced tomographic reconstruction techniques. Since limiting dose exposure to patients is crucial, iterative reconstruction approaches are widely used to accommodate noisy

measurements [30]–[32]. Many-core systems can deliver the computational throughput required by the reconstruction tasks, however their limited memory and (host-device) communication cost can introduce significant overhead [33]. In contrast, our solution provides mixed precision computations and multi-level reduction techniques to ease communication cost.

Synchrotron radiation facilities can generate small to extreme-scale data using variety of imaging modalities, including tomography [34]–[36]. High performance software infrastructures have been built to handle tomographic reconstruction workflows that require large-scale compute resources [37]–[42]. The end-to-end performance of these systems heavily relies on underlying reconstruction engines and can directly benefit from the optimizations proposed in our work.

VI. CONCLUSION

In this work, we introduce novel optimization techniques for MemXCT approach that exploit 3D volume properties and multi-GPU node architecture during (back)projection operators. Specifically, our simultaneous data and batch partitioning scheme provides configurable volume distribution among processes and enables reconstruction of extremely large tomography data; XCT-Optimized SpMM design considers the sparse matrix structure and efficiently provides GPU utilization by reusing data from shared-memory and registers; hierarchical communications performs extra intra-node communications to minimize the inter-node communication; and finally, effective use of mixed-precision types further reduces memory footprint and communication volume while maintaining the reconstruction quality. The algorithm design and optimizations described in this paper enable an $11k \times 11k \times 9k$ 3D brain image reconstruction under three minutes using 24,576 GPUs on Summit supercomputer, reaching 65 PFLOPS: 34% of Summit's peak performance.

ACKNOWLEDGMENTS

This material was partially supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences, under Contract DE-AC02-06CH11357. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. We thank Narayanan (Bobby) Kasthuri from UChicago/Argonne, and Rafael Vescovi and Ming Du from Argonne for sharing the mouse brain dataset. The mouse brain, IC chip, and activated charcoal data were collected by Vincent De Andrade at beamline 32-ID, Advanced Photon Source, Argonne National Laboratory. This work is supported by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR). This research is based in part upon work supported by the Center for Applications Driving Architectures (ADA), a JUMP Center co-sponsored by SRC and DARPA. This material is supported by funding

from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433.

REFERENCES

- [1] "APS Science 2018," Tech. Rep. ANL-18/40, ISSN 1931-5007, Advanced Photon Source, Argonne National Laboratory, January 2019.
- [2] R. Vescovi, M. Du, V. de Andrade, W. Scullin, D. Gürsoy, and C. Jacobsen, "Tomosaic: efficient acquisition and reconstruction of teravoxel tomography data using limited-size synchrotron X-ray beams," *Journal of Synchrotron Radiation*, vol. 25, pp. 1478–1489, Sep 2018.
- [3] T. Bicer, D. Gürsoy, R. Kettimuthu, F. De Carlo, G. Agrawal, and I. T. Foster, "Rapid tomographic image reconstruction via large-scale parallelization," in *Euro-Par 2015: Parallel Processing*, pp. 289–302, Springer, 2015.
- [4] 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.
- [5] X. Wang, V. Sridhar, Z. Ronaghi, R. Thomas, J. Deslippe, D. Parkinson, G. T. Buzzard, S. P. Midkiff, C. A. Bouman, and S. K. Warfield, "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*, pp. 1–23, 2019.
- [6] A. Beer, "Bestimmung der absorption des rothen lichts in farbigen flüssigkeiten," *Ann. Physik*, vol. 162, pp. 78–88, 1852.
- [7] J.-H. Lambert, *JH Lambert,... Photometria, sive de Mensura et gradibus luminis, colorum et umbrae*. sumptibus viduae E. Klett, 1760.
- [8] R. A. Crowther, D. DeRosier, and A. Klug, "The reconstruction of a three-dimensional structure from projections and its application to electron microscopy," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 317, no. 1530, pp. 319–340, 1970.
- [9] R. L. Siddon, "Fast calculation of the exact radiological path for a three-dimensional ct array," *Medical Physics*, vol. 12, no. 2, pp. 252–255, 1985.
- [10] F. De Carlo, D. Gürsoy, D. J. Ching, K. J. Batenburg, W. Ludwig, L. Mancini, F. Marone, R. Mokso, D. M. Pelt, J. Sijbers, et al., "Tomobank: a tomographic data repository for computational x-ray science," *Measurement Science and Technology*, vol. 29, no. 3, p. 034004, 2018.
- [11] M. Du, R. Vescovi, K. Fezzaa, C. Jacobsen, and D. Gürsoy, "X-ray tomography of extended objects: a comparison of data acquisition approaches," *JOSA A*, vol. 35, no. 11, pp. 1871–1879, 2018.
- [12] 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.
- [13] K. Mohan, S. Venkatakrishnan, J. Gibbs, E. Gulsoy, X. Xiao, M. De Graef, P. Voorhees, and C. Bouman, "Timbir: A method for time-space reconstruction from interlaced views," *Computational Imaging, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.
- [14] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pp. 945–948, IEEE, 2013.
- [15] P. P. Bruyant, "Analytic and iterative reconstruction algorithms in spect," *Journal of Nuclear Medicine*, vol. 43, no. 10, pp. 1343–1358, 2002.
- [16] J. A. Fessler, M. Sonka, and J. M. Fitzpatrick, "Statistical image reconstruction methods for transmission tomography," *Handbook of medical imaging*, vol. 2, pp. 1–70, 2000.
- [17] S. Aslan, V. Nikitin, D. J. Ching, T. Bicer, S. Leyffer, and D. Gürsoy, "Joint pycho-tomography reconstruction through alternating direction method of multipliers," *Optics express*, vol. 27, no. 6, pp. 9128–9143, 2019.
- [18] V. Nikitin, S. Aslan, Y. Yao, T. Biçer, S. Leyffer, R. Mokso, and D. Gürsoy, "Photon-limited ptychography of 3d objects via bayesian reconstruction," *OSA Continuum*, vol. 2, no. 10, pp. 2948–2968, 2019.
- [19] D. J. Ching, M. Hidayetoğlu, T. Biçer, and D. Gürsoy, "Rotation-as-fast-axis scanning-probe x-ray tomography: the importance of angular diversity for fly-scan modes," *Applied optics*, vol. 57, no. 30, pp. 8780–8789, 2018.

- [20] J. Treibig, G. Hager, H. G. Hofmann, J. Hornegger, and G. Wellein, "Pushing the limits for medical image reconstruction on recent standard multicore processors," *International Journal of High Performance Computing Applications*, 2012.
- [21] J. Agulleiro and J.-J. Fernandez, "Fast tomographic reconstruction on multicore computers," *Bioinformatics*, vol. 27, no. 4, pp. 582–583, 2011.
- [22] M. Jones, R. Yao, and C. Bhole, "Hybrid MPI-OpenMP programming for parallel OSEM PET reconstruction," *Nuclear Science, IEEE Transactions on*, vol. 53, pp. 2752–2758, Oct. 2006.
- [23] C. Johnson and A. Sofer, "A data-parallel algorithm for iterative tomographic image reconstruction," in *7th Symposium on the Frontiers of Massively Parallel Computation*, pp. 126–137, Feb 1999.
- [24] A. Sabne, X. Wang, S. J. Kisner, C. A. Bouman, A. Raghunathan, and S. P. Midkiff, "Model-based iterative CT image reconstruction on GPUs," in *22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pp. 207–220, ACM, 2017.
- [25] W. van Aarle, W. J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K. J. Batenburg, and J. Sijbers, "The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography," *Ultra-microscopy*, vol. 157, pp. 35–47, 2015.
- [26] X. Li, Y. Liang, W. Zhang, T. Liu, H. Li, G. Luo, and M. Jiang, "cuMBIR: An efficient framework for low-dose x-ray CT image reconstruction on GPUs," in *International Conference on Supercomputing*, pp. 184–194, ACM, 2018.
- [27] X. Yu, H. Wang, W.-c. Feng, H. Gong, and G. Cao, "Gpu-based iterative medical ct image reconstructions," *Journal of Signal Processing Systems*, vol. 91, no. 3-4, pp. 321–338, 2019.
- [28] P. Chen, M. Wahib, S. Takizawa, R. Takano, and S. Matsuoka, "ifdk: a scalable framework for instant high-resolution image reconstruction," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–24, 2019.
- [29] M. Hidayetoglu, C. Pearson, I. El Hajj, L. Gurel, W. C. Chew, and W. Hwu, "A fast and massively-parallel inverse solver for multiple-scattering tomographic image reconstruction," in *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 64–74, 2018.
- [30] D. Lee, I. Dinov, B. Dong, B. Gutman, I. Yanovsky, and A. W. Toga, "CUDA optimization strategies for compute-and memory-bound neuroimaging algorithms," *Computer Methods and Programs in Biomedicine*, vol. 106, no. 3, pp. 175–187, 2012.
- [31] C.-Y. Chou, Y.-Y. Chuo, Y. Hung, and W. Wang, "A fast forward
- projection using multithreads for multirays on GPUs in medical image reconstruction," *Medical Physics*, vol. 38, no. 7, pp. 4052–4065, 2011.
- [32] B. Jang, D. Kaeli, S. Do, and H. Pien, "Multi GPU implementation of iterative tomographic reconstruction algorithms," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pp. 185–188, IEEE, 2009.
- [33] T. B. Jablin, P. Prabhu, J. A. Jablin, N. P. Johnson, S. R. Beard, and D. I. August, "Automatic CPU-GPU communication management and optimization," in *ACM SIGPLAN Notices*, vol. 46, pp. 142–151, ACM, 2011.
- [34] D. Gürsoy, T. Biçer, A. Lanzirotti, M. G. Newville, and F. De Carlo, "Hyperspectral image reconstruction for x-ray fluorescence tomography," *Optics express*, vol. 23, no. 7, pp. 9014–9023, 2015.
- [35] D. J. Duke, A. B. Swantek, N. M. Sovis, F. Z. Tilocco, C. F. Powell, A. L. Kastengren, D. Gürsoy, and T. Biçer, "Time-resolved x-ray tomography of gasoline direct injection sprays," *SAE International Journal of Engines*, vol. 9, no. 1, pp. 143–153, 2016.
- [36] D. Gürsoy, T. Biçer, J. D. Almer, R. Kettimuthu, S. R. Stock, and F. De Carlo, "Maximum a posteriori estimation of crystallographic phases in x-ray diffraction tomography," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 373, no. 2043, p. 20140392, 2015.
- [37] T. Bicer, D. Gursoy, R. Kettimuthu, I. T. Foster, B. Ren, V. De Andrede, and F. De Carlo, "Real-time data analysis and autonomous steering of synchrotron light source experiments," in *2017 IEEE 13th International Conference on e-Science (e-Science)*, pp. 59–68, IEEE, 2017.
- [38] R. J. Pandolfi, D. B. Allan, E. Arenholz, L. Barroso-Luque, S. I. Campbell, T. A. Caswell, A. Blair, F. De Carlo, S. Fackler, A. P. Fournier, et al., "Xi-cam: a versatile interface for data visualization and analysis," *Journal of synchrotron radiation*, vol. 25, no. 4, pp. 1261–1270, 2018.
- distributed resources," *Journal of synchrotron radiation*, vol. 23, no. 4, pp. 997–1005, 2016.
- [40] D. Morozov and T. Peterka, "Block-parallel data analysis with diy2," in *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 29–36, 2016.
- [41] O. Yildiz, J. Ejarque, H. Chan, S. Sankaranarayanan, R. M. Badia, and T. Peterka, "Heterogeneous hierarchical workflow composition," *Computing in Science Engineering*, vol. 21, no. 4, pp. 76–86, 2019.
- [42] Z. Liu, T. Bicer, R. Kettimuthu, and I. Foster, "Deep learning accelerated light source experiments," in *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*, pp. 20–28, IEEE, 2019.

Appendix: Artifact Description/Artifact Evaluation

SUMMARY OF THE EXPERIMENTS REPORTED

We ran our optimized iterative reconstruction system on Summit supercomputer with IBM spectrum MPI 10.3.1.2-20200121. Our iterative reconstruction is based on conjugate gradient solver. We also used the default compiler on Summit (xl, 16.1.1-5). We used four different tomography datasets: a mouse brain (bio)sample, an integrated circuit (IC) sample, a charcoal sample (under Datasets/Mosaic), and a shale sample (under Datasets/Round-Robin). Charcoal and shale samples can be downloaded from TomoBank, (<https://tomobank.readthedocs.io>). The other two datasets, mouse brain and IC are not publicly available.

ARTIFACT AVAILABILITY

Software Artifact Availability: Some author-created software artifacts are NOT maintained in a public repository or are NOT available under an OSI-approved license.

Hardware Artifact Availability: There are no author-created hardware artifacts.

Data Artifact Availability: Some author-created data artifacts are NOT maintained in a public repository or are NOT available under an OSI-approved license.

Proprietary Artifacts: There are associated proprietary artifacts that are not created by the authors. Some author-created artifacts are proprietary.

BASELINE EXPERIMENTAL SETUP, AND MODIFICATIONS MADE FOR THE PAPER

Relevant hardware details: Summit supercomputer, GPU cluster, 2 IBM POWER9 CPUs and 6 NVIDIA Volta per node, 4608 nodes, 250PB 2.5TB/s GPFS

Operating systems and versions: Red Hat Enterprise Linux version 7.4

Compilers and versions: xl v16.1.1-5, cuda v10.1.243

Libraries and versions: IBM Spectrum MPI 10.3.1.2-20200121

Key algorithms: conjugate gradient

Input datasets and versions: Charcoal and shale sample datasets can be downloaded from TomoBank, (<https://tomobank.readthedocs.io>).

ARTIFACT EVALUATION

Verification and validation studies: We ran our tests several times on Summit cluster and presented the performance results. We measured strong and weak scaling efficiencies. Our results were consistent in each run.

Accuracy and precision of timings: MPI timings (with and without barriers) are used for timings. Several runs are performed.

Quantified the sensitivity of results to initial conditions and/or parameters of the computational environment: We used real-world tomography datasets, therefore our initial conditions were set by the experimental data. Since this is iterative reconstruction, the convergence is sensitive to number of iterations. We used heuristic early termination to determine number of iterations.

Controls, statistics, or other steps taken to make the measurements and analyses robust to variability and unknowns in the system. Our computing resources were explicitly allocated during the system evaluation; therefore, there should be minimal noise in terms of computation performance evaluation. The network resources were shared within cluster, thus there might be some network contention that we couldn't profile. However, all tests were run several times and no anomalies had been observed.