

Accelerating Multigrid-based Hierarchical Scientific Data Refactoring on GPUs

Jieyang Chen, Lipeng Wan, Xin Liang*, Ben Whitney, Qing Liu[†], David Pugmire, Nicholas Thompson, Jong Youl Choi, Matthew Wolf, Todd Munson[‡], Ian Foster[§], Scott Klasky

Oak Ridge National Laboratory, Oak Ridge, TN, USA

* Missouri University of Science and Technology, Rolla, MO, USA

[†] New Jersey Institute of Technology, Newark, NJ, USA

[‡] Argonne National Laboratory, Lemont, IL, USA

[§] University of Chicago, Chicago, IL, USA

{chenj3, wanl, whitneybe, pugmire, thompsonna, choij, wolfm, klasky}@ornl.gov
xlang@mst.edu qliu@njit.edu {tmunson, foster}@anl.gov

Abstract—Rapid growth in scientific data and a widening gap between computational speed and I/O bandwidth make it increasingly infeasible to store and share all data produced by scientific simulations. Instead, we need methods for reducing data volumes: ideally, methods that can scale data volumes adaptively so as to enable negotiation of performance and fidelity tradeoffs in different situations. Multigrid-based hierarchical data representations hold promise as a solution to this problem, allowing for flexible conversion between different fidelities so that, for example, data can be created at high fidelity and then transferred or stored at lower fidelity via logically simple and mathematically sound operations. However, the effective use of such representations has been hindered until now by the relatively high costs of creating, accessing, reducing, and otherwise operating on such representations. We describe here highly optimized data refactoring kernels for GPU accelerators that enable efficient creation and manipulation of data in multigrid-based hierarchical forms. We demonstrate that our optimized design can achieve up to 250 TB/s aggregated data refactoring throughput—83% of theoretical peak—on 1024 nodes of the Summit supercomputer. We showcase our optimized design by applying it to a large-scale scientific visualization workflow and the MGARD lossy compression software.

Index Terms—Multigrid, Data refactoring, GPU

I. INTRODUCTION

With the dawn of the big data era, managing the massive volume of data generated by data-intensive applications becomes extremely challenging, particularly for scientific simulations [1, 2] running on leadership-class high-performance computing (HPC) systems and experiments running on federated instruments and sensor platforms. For instance, the XGC dynamic fusion simulation code [3, 4] from the Department of Energy (DOE)’s Princeton Plasma Physics Laboratory can generate 1 PB every 24 hours when running on the DOE’s fastest supercomputers, and may soon generate 10 PB per day. The Square Kilometer Array (SKA) [5] plans to generate data at 1 PB/s within 10–20 years. Few storage systems can keep up with such data rates. Moreover, even if all data could be stored, the high costs of processing them with standard multi-pass analysis routines often lead to significant degradation in overall scientific productivity [6, 7].

There are no universal solutions to the many technical and domain-specific challenges of managing the overwhelming

amount of complex and heterogeneous scientific data, and current approaches are usually passive and based on rules of thumb. For instance, scientists may decimate in time by reducing output data frequency by some arbitrary factor (e.g., writing one of every 1000 simulation steps). Although such approaches can effectively reduce the amount of data written to storage, they increase the risk of missing novel scientific discoveries, as discarded data may contain important features. Moreover, the limited capacity of fast storage such as parallel file systems means that data are eventually moved to slower storage, such as archival storage systems. For example, on Oak Ridge National Laboratory (ORNL)’s Summit supercomputer, data can only be kept on the parallel file system for 90 days before it is either moved to archival storage systems such as HPSS [8] or permanently deleted. Once moved to archival storage, it can take weeks or longer to retrieve for analysis.

In plotting a course to address these data management challenges, it is important to remember the perspective of the end user, the domain scientist, for whom a dataset is valued not for its size in bytes but for the scientific information it contains. A dataset need not be especially large to capture some feature of interest, and in fact the most valuable insights often come from just a small portion of the original data. A domain scientist seeking to answer a question using a dataset would ideally be able to retrieve only the smallest subset or reduced representation of the data necessary to answer the question to the desired level of accuracy. It is challenging to support this workflow with existing data compression techniques given the great variety of analyses a scientist might need to run, since the data would need to be compressed and stored separately for each analysis and each level of accuracy required, resulting in high computational and storage costs.

Data refactoring is the capability of building a data representation in a hierarchical form such that a reader can easily, efficiently, and transparently access data at varying degrees of fidelity. To enable this capability, new algorithms such as multigrid-based hierarchical data refactoring [9–11] have recently been developed by the applied mathematics community. That data refactoring approach models a dataset with a series of hierarchically organized *coefficient classes*,

such that an approximation of the original data with a specified fidelity can be reconstructed by using different numbers of coefficient classes. We call the process of building coefficient classes *decomposition* and the process of reconstructing data from coefficient classes *recomposition*.

Hierarchical data refactoring gives both data producers (e.g., scientific simulations) and consumers (e.g., data analysis routines) the flexibility to store, transport, and access data to satisfy space and/or accuracy requirements. For example, data sharing between two coupled scientific applications [12] can be optimized by intelligently moving coefficient classes through multi-tiered-storage systems (e.g., storage systems containing non-volatile memory, magnetic disks, and tapes) [13, 14] and/or networks based on available capacity and bandwidth. In Figure 1, simulation data are refactored into five coefficient classes and then shared with data analysis routines via multi-tiered-storage systems and networks. When accuracy can be estimated based on the number of selected coefficient classes, users can control the accuracy of the reconstructed data while storing and reading the data. If user-defined accuracy requirements indicate that information encoded in the first four coefficient classes are enough for subsequent data analyses, then the fifth coefficient class can be ignored. Then, the four coefficient classes can be intelligently shared over the storage systems and network based on their size, available bandwidth/capacities, and accuracy requirements from data analysis routines. In the figure, Data Analysis Routine 1 needs only two coefficient classes to achieve desired accuracy, while Routine 2 needs four. The ability to choose a reduced number of coefficient classes allows users to reduce data movement costs substantially.

As great as the benefits of reduction in data movement and management costs may be, if the decomposition and recomposition routines are too expensive, then the total process is less useful in production. The use of Graphics Processing Units (GPUs) for scientific computations that can be adopted to the streaming execution model has increased significantly due to the high parallel computational power and memory throughput of GPUs. As the algorithms involved in multigrid-based hierarchical data refactoring are highly parallelizable, using GPUs to accelerate its routines is attractive. Also, we anticipate that if used with merging GPU communication technologies [15, 16] (e.g., NVLink, GPUDirect RDMA, etc.), GPU data refactoring would be greatly beneficial for speeding up data sharing for both CPU- and GPU-based scientific applications.

We focus here on accelerating the two major routines, decomposition and recomposition, in multigrid-based data refactoring on GPUs and evaluating the benefit for producer and consumer applications. Although the multigrid-based algorithms are naturally parallelizable, achieving good performance requires carefully designed parallel algorithms together with deep optimizations for GPU architectures. Our specific contributions, and the sections in which they are described, are as follows.

In §III, we describe the first multigrid-based data refactoring routines for modern GPU architectures, and present systematic

TABLE I: NOTATION USED IN ALGORITHMS, FORMULATIONS, FIGURES

| Symbol | Description |
|-------------------|---|
| u | Function represented by the original data. |
| N_l | Nodes at grid level l . |
| C_l | Coefficients at grid level l . |
| V_l | Function space with respect to N_l . |
| Q_l | The L^2 projection onto V_l . |
| Π_l | The piecewise linear interpolant in space V_l . |
| $a \rightarrow b$ | b is calculated using a . |

optimizations for multigrid-based data refactoring at three levels: instruction level, kernel level, and program-structure level. These optimizations can balance both minimizing memory footprint and improving memory access efficiency.

In §IV, we demonstrate our design by implementing the state-of-the-art non-uniform multi-dimensional multigrid-based data refactoring algorithms of Ainsworth et al. [9–11], and show that our methods perform well on both a consumer-class desktop and the Summit supercomputer, achieving $145\times$ and $14\times$ speedups compared with state-of-the-art CPUs and GPUs, and 250 TB/s throughput on 1024 Summit nodes.

In §V, we use two common scenarios in scientific computing to showcase our work: 1) reducing data movement costs between simulations and in situ visualization applications; and 2) speeding up lossy compression for scientific data.

II. BACKGROUND

A. Theory of multigrid-based hierarchical data refactoring

The multigrid-based hierarchical data refactoring developed by Ainsworth et al. supports nonuniformly-spaced structured multidimensional data, commonly found in scientific computations, by using hierarchical representations to approximate data. Specifically, they decompose data from fine grid representation to coarse grid representation in an iterative fashion, with a global correction to account for the impact of missing grid nodes in each iteration. If we use functions to represent the discrete values continuously, the decomposition from fine grid level l to coarser grid level $l-1$ can be formulated with the notation in Table I as follows,

$$\underbrace{Q_{l-1}u}_{\text{Projection onto } V_{l-1}} = \underbrace{Q_l u}_{\text{Projection onto } V_l} - \underbrace{(I - \Pi_{l-1})Q_l u}_{\text{Coefficients}} + \underbrace{(Q_{l-1}u - \Pi_{l-1}Q_l u)}_{\text{Corrections}} \quad (1)$$

where the piecewise linear function u takes the same values as the original data for each node; $Q_{l-1}u$ and $Q_l u$ are the function approximations of u at levels $l-1$ and l , respectively; $(I - \Pi_{l-1})Q_l u$ is the difference between the values of the fine grid nodes at level l and their corresponding piecewise linear approximations; and $(Q_{l-1}u - \Pi_{l-1}Q_l u)$ is the global correction. According to Eq. (1), two major steps are involved at each level of the multigrid decomposition: 1) compute coefficients for the current multigrid level l ; and 2) compute the global correction and add it to the nodes in the next coarse grid (level $l-1$). In what follows, we introduce how to compute coefficients and corrections.

1) *Compute coefficients*: The coefficients store the difference between the data approximated by nodes at levels l (i.e., N_l) and $l-1$ (i.e., N_{l-1}) before corrections are added. Since N_{l-1} is contained in N_l , its nodes have the same values in

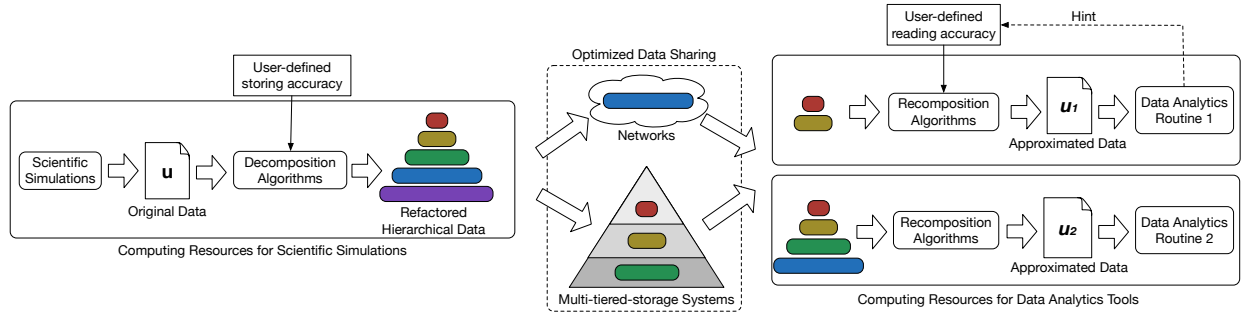


Fig. 1: Example of hierarchical data refactoring helping optimize data movement in scientific workflows by intelligently moving each coefficient class across networks and/or multi-tiered-storage systems, based on available capacity and bandwidth.

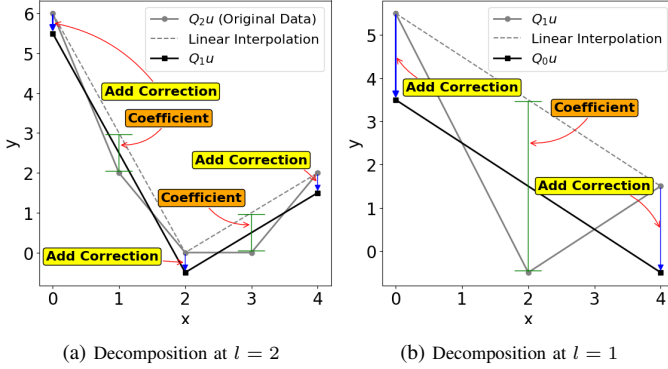


Fig. 2: Example of decomposing a 1D dataset produced from discretizing a quadratic function: $y = x^2 - 5x + 6$

both levels; thus the nonzero differences only occur on nodes in $N_l \setminus N_{l-1}$. Figure 2 shows how coefficients are calculated along one dimension through linear interpolation. It can be generalized to multi-dimensional cases easily by using multi-linear interpolations for approximation.

2) *Compute correction*: Ainsworth et al. prove that the correction is the orthogonal projection of the calculated coefficients at grid level l onto V_{l-1} [9]; thus, adding the correction to the next coarse grid better approximates data in the current grid. To explain, we first define z_{l-1} as the correction for grid at level $l-1$. From Eq. (1), we have that:

$$z_{l-1} - \underbrace{(I - \Pi_{l-1})Q_l u}_{\text{Coefficients}} = -(Q_l - Q_{l-1})u \in V_{l-1}^\perp \quad (2)$$

If we apply L^2 projection at grid level $l-1$ (i.e., Q_{l-1}) to both sides of Eq. 2, it leads to a zero function since it belongs to V_{l-1}^\perp . Also, since z_{l-1} is in V_{l-1} , $Q_{l-1}z_{l-1} = z_{l-1}$. So, we can see that z_{l-1} is the orthogonal projection of the coefficients onto V_{l-1} . Namely, $Q_{l-1}(I - \Pi_{l-1})Q_l u = z_{l-1}$.

The correction can thus be computed by solving a variational problem: find $z_{l-1} \in V_{l-1}$ such that $(z_{l-1}, v_{l-1}) = ((I - \Pi_{l-1})Q_l u, v_{l-1})$ for all $v_{l-1} \in V_{l-1}$. Then, z_{l-1} can be found by solving linear systems $M_{l-1}z_{l-1} = f_{l-1}$ where M_{l-1} is a tensor product of the mass matrices [17] of each dimension, i.e., $M_{l-1} = M_{l-1}^1 \otimes M_{l-1}^2 \cdots \otimes M_{l-1}^d$, where d is the number of dimensions and f_{l-1} is the load vector, which can be calculated using: $f_{l-1} = R_l M_l \text{vec}(C_l)$, where R_l is a transfer matrix that converts basis functions from V_l to V_{l-1} and C_l is the coefficient matrix at level l , which consists of computed coefficients at $N_l \setminus N_{l-1}$ and zeros at N_{l-1} .

Overall decomposition/recomposition process: Figure 3 illustrates this process on a 5×5 2D dataset. The original data is on the left, and the refactored representation is on the right. The decomposition process moves from left to right (i.e., from finest to coarsest grid) and involves four steps: computing coefficient and computing correction (II.A.1 and II.A.2) for each of the two levels. For multi-dimensional data, the computation of correction is done by working on each dimension in a prescribed order [18]; in this 2D example, it proceeds first along the rows and then along the columns. Recomposition moves from right to the left: i.e., from coarsest to finest grid. There are again four total stages, but these occur in the reverse order. The approximation of the original data is produced after recomposition. Based on how coefficients are omitted in recomposition, an error bound on data approximation can be computed [10].

B. Existing GPU-based data refactoring

The state-of-the-art MGARD [19] GPU-based data refactoring system redesigns original serial algorithms to expose high parallelism to suit the many-core architecture of modern GPUs. It achieves $O(n^3)$ thread concurrency for computing coefficients and $O(n^2)$ thread concurrency for computing corrections, and applies node reordering such that each kernel can take advantage of coalesced memory accesses. Theoretically, with large inputs, these levels of thread concurrency are more than enough to fully occupy GPU cores that can help achieve high data refactoring throughput. However, performance evaluation shows that it still suffers from underutilized memory throughput, achieving less than 10% of theoretical peak.

III. DESIGNING GPU-ACCELERATED DATA REFACTORING

We next discuss the design of our GPU-accelerated multigrid-based hierarchical data refactoring method. We first focus on the optimizations for each computing kernel involved in data refactoring. We classify the computing patterns into three categories and propose three general kernel designs for GPUs. Following the efficient kernel designs, we discuss optimizations to help each of the kernels efficiently work together so that their performance can be maximized. Finally, we discuss design details about how to use heuristic auto tuning to maximize the refactoring throughput.

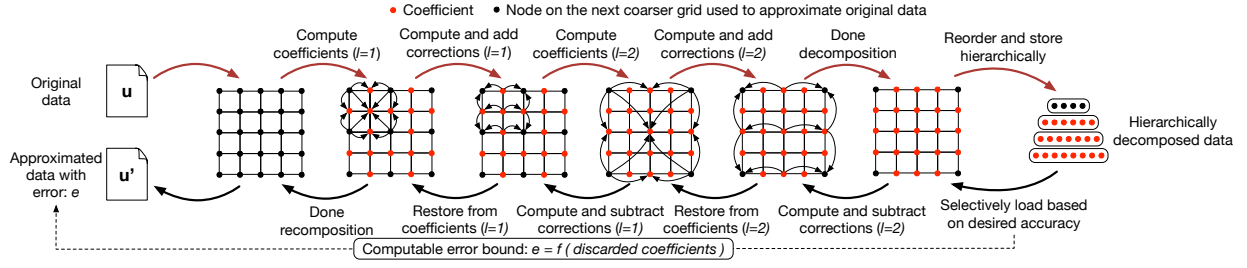


Fig. 3: Multigrid-based data refactoring: Decomposition (left to right); recomposition (right to left).

A. Designing optimized GPU multigrid kernels

Decomposition and recomposition each involve three major steps: 1) computing coefficients; 2) mass-transfer matrix multiplication; and 3) correction solver. Based on their computation pattern, we can classify them into three categories: *grid processing style*; *linear processing style*; and *iterative processing style*. We design kernels dedicated for each processing style.

1) *Grid processing kernel (GPK)*: Grid processing style has the characteristic of processing data in a grid-wise fashion. Namely, it processes nodes within the domain of a grid in a certain resolution level (e.g., N_l) or between neighboring levels (e.g., N_l and N_{l-1}). In the multigrid-based data refactoring, the calculation of coefficients follows the grid processing style. The major calculation is to compute the interpolation at nodes in $N_l \setminus N_{l-1}$ using nodal values in N_{l-1} . Parallelization can favor either interpolation operations (i.e., parallelism $\propto O(N_l \setminus N_{l-1})$) or accessing nodal values (i.e., parallelism $\propto O(N_l)$). The former can lead to a less thread divergence, while the latter can achieve a higher memory access efficiency. The computation of coefficients is a memory bound operation, as its time complexity is $O(n)$. Therefore, it is essential to optimize in favor of memory access efficiency instead of computation. This is also chosen in the state-of-the-art GPU data refactoring [19].

The key strategy they used to optimize for memory access is to use shared memory to cache a block of data for processing, of which the nodes values are loaded/stored in a coalesced-friendly fashion. However, we identify that keeping efficient data movement on memory bound computations is not enough to achieve good performance. The level of thread divergence in a computation can still make a great impact on the overall performance and sometimes it can wrongly convert computation from memory bound to compute bound. The reason is threefold: 1) high degrees of thread divergence can great increase the total cycle cost in computation; 2) variable floating point operation counts caused by different interpolation types further brings workload imbalance which leads to longer idling cycles; 3) as shown in Figure 4, some thread blocks also need to calculate coefficients in the ghost region, which exacerbate the effect of thread divergence.

However, we found that keeping efficient memory access patterns is not exclusive with having low thread divergence. In designing our GPK, we propose to decouple memory access and computation on nodal values in terms of thread-node assignment through a thread reassignment strategy. Specifically, we use two different thread-node assignments for load-

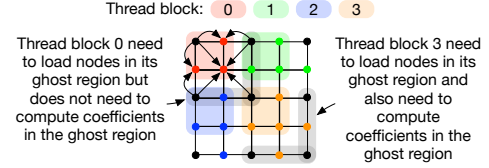


Fig. 4: The workload of computing coefficients is distributed among 4 thread blocks. Calculating coefficients in the corresponding ghost regions is needed for some thread blocks (e.g., thread blocks 1, 2, and 3).

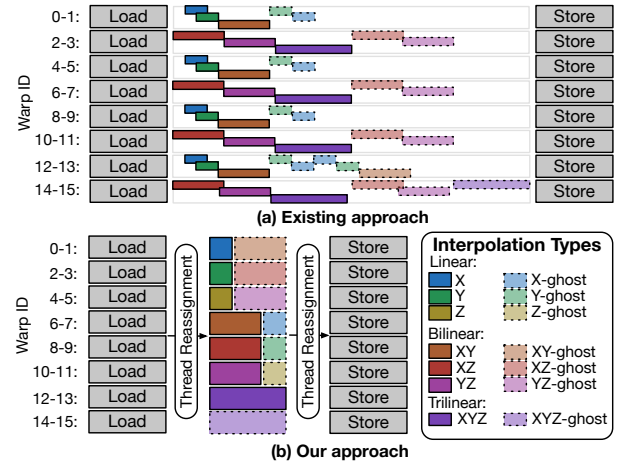


Fig. 5: Conceptual flow of a thread block with $8 \times 8 \times 8$ threads (16 warps) calculating coefficients using the existing and our grid processing kernel GPK. The thread reassignment strategy allows GPK to greatly reduce thread divergence.

ing/storing nodal values and computing interpolations such that we maintain efficient coalesced memory access pattern while having one warp process the same type of interpolations along the same dimension.

Figure 5 shows the conceptual execution flow of one thread block using the existing approach and our proposed GPK when computing coefficients. As nodes in N_l need to be shared with neighbors during interpolation operations, we let each thread block coordinate work on a block of data and use shared memory as a scratch space. We organize threads such that threads in the same warp load values that are consecutive in memory to achieve efficient coalesced memory access patterns. For computing, we apply a thread re-assignment strategy to achieve divergence-free execution. Algorithm 1 shows how we calculate the thread-interpolation operation assignments that minimize thread divergence. It is easy to see that the reassignment processing brings negligible computing overhead.

Algorithm 1: Thread re-assignment strategy

```

1 Function InterpolationType():
2    $B_x, B_y, B_z \leftarrow$  Thread block size
3    $x, y, z \leftarrow$  Thread local indexes within thread block
4    $\text{lane\_id}, \text{warp\_id} \leftarrow x, y, z$ 
5    $T \leftarrow 8$  //total num. of warp group (4 for 2D)
6    $\text{group\_id} \leftarrow \text{warp\_id} / (B_x \times B_y \times B_z) / T$ 
7   switch( $\text{group\_id}$ )
8     case 0: Linear-x(main) and Bilinear-xy(ghost)
9     case 1: Linear-y(main) and Bilinear-xz(ghost)
10    case 2: Linear-z(main) and Bilinear-yz(ghost)
11    case 3: Bilinear-xy(main) and Linear-x(ghost)
12    case 4: Bilinear-xx(main) and Linear-y(ghost)
13    case 5: Bilinear-yz(main) and Linear-z(ghost)
14    case 6: Trilinear-xyz(main)
15    case 7: Trilinear-xyz(ghost)
16 return

```

2) *Linear processing kernel (LPK)*: The linear processing style computes stencil operations on elements in vectors along one dimension in a grid. In multigrid-based data refactoring, when multiplying the mass and transfer matrices with computed coefficients, the computations become stencil operations, as the matrices are defined as:

$$M_{ij} = \begin{cases} 2(h_i + h_{i+1}) & \text{if } i = j \\ h_i & \text{if } |i - j| = 1 \\ 0 & \text{else} \end{cases}$$

$$R_{ij} = \begin{cases} 1 & \text{if } i = j/2 \\ r_{j-1} & \text{if } i = (j-1)/2 \\ 1 - r_j & \text{if } i = (j+1)/2 \\ 0 & \text{else} \end{cases}$$

where h_i is the spacing between the i^{th} node and the $i+1^{\text{th}}$ node and $r_i = h_i / (h_i + h_{i+1})$. As shown in Figure 6(a), each value of each node needs to be computed using the original values of its neighbors, which means it cannot update its stored value unless all neighbors have finishing using its original value for computation. Such data dependencies present a dilemma for kernel design: common out-of-place designs (i.e., element-wise parallelism) bring high parallelism but also high memory footprint; on the other hand, in-place design (i.e., vector-wise parallelism), used in [19], sacrifices the opportunity to exploit intrinsic parallelism.

To eliminate this dilemma, we design a novel linear processing kernel (LPK) with four optimizations. First, we change the original computation from in-place to out-of-place to achieve finer-grain parallelism. Second, we merge the mass and transfer matrices to reduce computational costs. We call the new matrix *mass-trans*, which is defined as:

$$K_{ij} = \begin{cases} (2 + r_{j-2})h_{j-1} + (1 + r_j) & \text{if } i = j/2 \\ (2r_{j-2} + 1)h_{j-1} + 2r_{j-2}h_{j-2} & \text{if } i = (j-1)/2 \\ (3 - 2r_j)h_{j+1} + 2r_{j+1}h_{j+1} & \text{if } i = (j+1)/2 \\ r_{j-2}h_{j-2} & \text{if } i = (j-2)/2 \\ (1 - r_j)h_{j+1} & \text{if } i = (j+2)/2 \\ 0 & \text{else} \end{cases}$$

Third, we use shared memory to cache a tile of nodes to allow sharing of coefficients (input) between different threads, so as to reduce total accesses to global memory. Finally, to reduce extra memory footprint we use a kernel fusion

technique to fuse the operation of copying coefficients with the multiplication of the mass-trans matrix with coefficients along the first dimension. By eliminating the need to store a copy of the computed coefficient in the workspace, we avoid a large increase in the overall memory footprint.

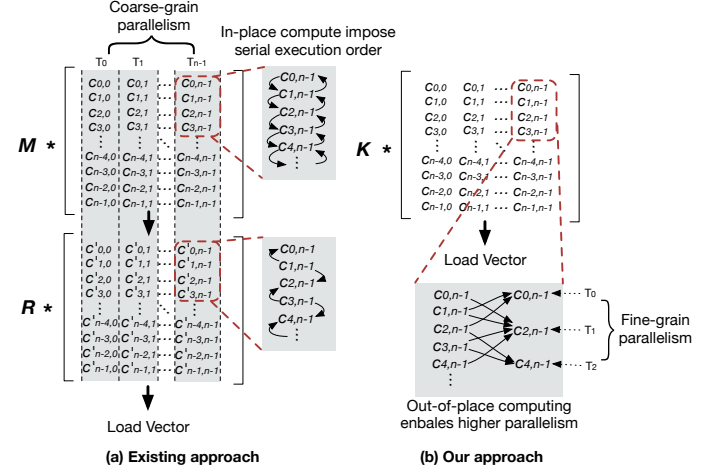


Fig. 6: The conceptual workflow of mass and transfer multiplication using existing approach and proposed approach. Through optimizations, our approach achieves finer grain parallelism.

3) *Iterative processing kernel (IPK)*: The iterative processing style has the characteristic of processing nodes in a grid that contains strong data dependencies such that nodes have to be processed iteratively in a certain order. In multigrid-based data refactoring, the correction solver needs to solve for the corrections. We use the Thomas algorithm [20], which needs a forward and a backward pass on the load vector. Since the load vectors along one dimension can be solved independently, they can be solved in parallel. This level of parallelization is well exploited in [19]. Specifically, they assign each thread to handle the solving process of one load vector independently. Although this brings high thread concurrency with divergence free execution, it actually suffers from inefficient memory accesses for two reasons: first when solving vectors on leading dimension full coalesced memory access cannot be achieved (actual achieve efficiencies are about only 12% and 25% for single and double precision data); second, compared with GPK and LPK, IPK only has $O(n^2)$ degrees of thread concurrency, which may bring less on-the-fly memory accesses to fully utilize the memory bandwidth.

To address this issue, we proposed a novel processing kernel, IPK, that can guarantee efficient coalesced memory access patterns with high concurrent memory accesses. We first parallelize the vectors by assigning a batch to a thread block. Since the update of each node depends on its neighboring elements, we use shared memory as scratch space to avoid polluting the un-processed nodes. Specifically, we let each thread block iteratively work on a segment of load vectors at a time until the whole vector is updated. Thus, as shown in Figure 7, during the computation we divide the elements in the vectors into six regions: 1) the processed region stores updated elements (gray); 2) the main region consists of elements that the current iteration is working on (green); 3)

and 4) due to dependence on the neighboring elements, the original values of elements in the two ghost regions (red and cyan) are needed to update the elements in the main region; 5) for better streaming processor utilization, we pre-fetch data needed for the next iteration (purple); and 6) we mark the unprocessed region as in block. The regions move forward as the computation proceeds. One challenge in designing the algorithm is to simultaneously consider maximizing coalesced global memory access patterns, minimizing bank conflict in accessing shared memory, and minimizing thread divergence. We use a dynamic data-thread assignment strategy [21–25] to optimize both the accessing and computation of coefficients.

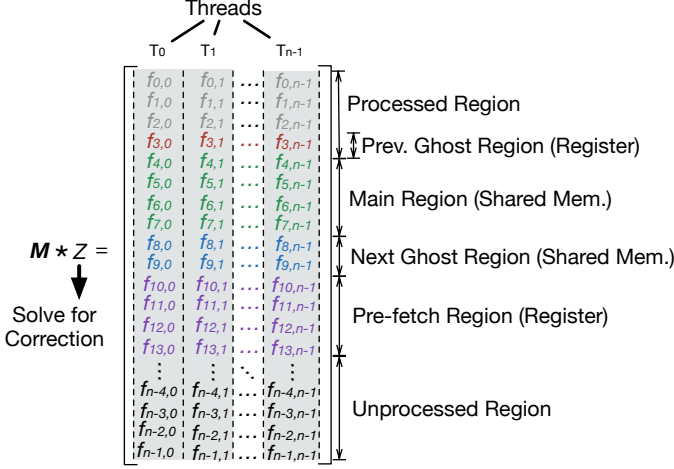


Fig. 7: Correction solver designed following iterative processing kernel (IPK). The node vectors are partitioned into six regions during processing. The use of shared memory ensures efficient coalesced memory accesses regardless of which dimension it is processing. Data prefetching further increases concurrency on memory accesses.

B. Overall algorithms

Figure 8 shows how we use our optimized kernels to build data refactoring routines for multi-dimensional data on GPUs. For each level, the computed coefficients are also used for correction calculations. This process involves altering the values of coefficients. So, to preserve the values of previously computed coefficients, the correction is computed in a workspace. In the state-of-the-art design [19], the computed coefficients are first copied to the workspace before they are used for computing corrections, which limit the design’s capability to do out-of-place computing unless using extra memory space. Our optimization merges the copy of coefficients, with the first mass-trans matrix multiplication, so that it enables out-of-place computation. We further extend out-of-place mass-trans matrix multiplication for processing other dimensions, which improves parallelism with a slight increase in memory footprint for the workspace. In the state-of-the-art design, the workspace is of size $m \times n \times k$ and in our design its size is $(m+1) \times (n+1) \times k$, where m , n , and k are the three dimensions of the input data. The recomposition process is the opposite so we omit showing its process due to the page limit.

C. Heuristic Performance Auto Tuning

When launching each proposed kernel, choosing the execution parameters is important for achieving good performance,

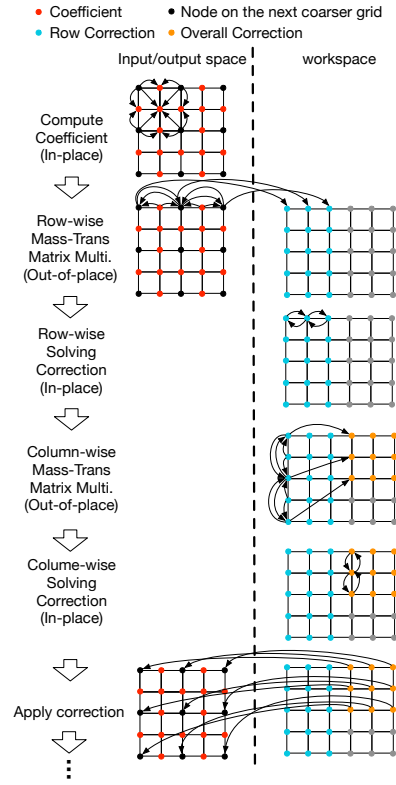


Fig. 8: Overall decomposition process with optimizations.

since even with the optimized design the parameters can still greatly impact the efficiency of memory accesses, warp divergence, context switch overhead, etc. Auto tuning is an effective approach for searching the optimum configurations. However, brute force search can be expensive and thus impractical. Thus, we propose to use a heuristic auto tuning approach guided by theoretical performance models for our GPU data refactoring. We first build performance models for the three kernels we proposed. Among all tunable execution parameters, we find that the size of the thread block (B_x, B_y, B_z) plays an important role in determining each kernel’s performance. Since we eliminate the majority of the thread divergence and inefficient computations, we assume the memory load/store takes the majority time, so we only consider the total amount of memory transactions with their efficiency. The estimated execution time of each kernel can be modeled as:

$$T_{\text{GPK}} = \lceil B_x + 1/(S/L) \rceil \cdot (S/L) \cdot (B_y + 1) \cdot (B_z + 1) \cdot \lfloor N/B_x \rfloor \cdot \lfloor N/B_y \rfloor \cdot \lfloor N/B_z \rfloor \cdot 2L \cdot (1/\text{Peak Mem. Band.})$$

$$T_{\text{LPK}} = (\lceil B_x/(S/L) \rceil \cdot S/L + 2S/L) \cdot B_y \cdot B_z \cdot \lfloor N/B_x \rfloor \cdot \lfloor N/B_y \rfloor \cdot \lfloor N/B_z \rfloor \cdot 2L \cdot (1/\text{Peak Mem. Band.})$$

$$T_{\text{IPK}} = (\lceil G/(S/L) \rceil \cdot S/L + \lceil B_x/(S/L) \rceil \cdot S/L \cdot \lfloor N/B_x \rfloor) \cdot B_y \cdot B_z \cdot \lfloor N/B_y \rfloor \cdot \lfloor N/B_z \rfloor \cdot 2L \cdot (1/\text{Peak Mem. Band.})$$

TABLE II: RANKING OF ESTIMATED PERFORMANCE OF SEVEN TYPICAL THREAD BLOCK SIZE CONFIGURATIONS; ACTUAL BEST IN RED.

| B_z | B_y | B_x | GPK | LPK | IPK |
|-------|-------|-------|----------|----------|----------|
| 2 | 2 | 2 | 7 | 7 | 7 |
| 4 | 4 | 4 | 6 | 6 | 1 |
| 4 | 4 | 8 | 4 | 5 | 2 |
| 4 | 4 | 16 | 2 | 4 | 3 |
| 4 | 4 | 32 | 1 | 3 | 4 |
| 2 | 2 | 64 | 5 | 2 | 5 |
| 2 | 2 | 128 | 3 | 1 | 6 |

where S is the number of bytes per memory transaction, (32 in our test GPU); L is bytes per float (4 for single, 8 for double); and G is the dimension of the next ghost region, set to S/L so that ghost data can fit into exactly one memory transaction and do not consume too much shared memory. Table II shows the ranking of estimated performance using seven typical thread block size configurations. Numbers in red represent the actual best configuration as determined by profiling. We can see our performance model can help up predict relationship between different configuration in terms of performance with relative high accuracy. It helps us narrow down the searching space for auto tuning. For instance, in our following evaluation, we only let the auto tuning search and pick among the estimated top three configurations to save time.

IV. EXPERIMENTAL EVALUATION

We evaluate our work on two GPU-enabled platforms. Each node of the **Summit** supercomputer at ORNL is equipped with 6 NVIDIA **Volta** GV100 GPUs with 16 GB memory on each GPU and two 22-core (of which 21 cores/socket are accessible for computation) IBM POWER9 CPUs with 512 GB memory. **Turing** is a GPU-accelerated desktop with an NVIDIA RTX 2080 Ti GPU with 11 GB of memory and one 8-core Intel i7-9700K CPU with 32 GB of memory.

A. Evaluation methodology

We use datasets from a Gray–Scott reaction–diffusion simulation [26, 27]. Each node in the input grid data is represented as single or double precision floating point values. Note that our data refactoring algorithms have deterministic computation time complexity regardless of the values in the chosen dataset, so it will yield the same performance for any dataset with the same dimensions and size. For simplicity, we let each dimension have the same size in our experiments.

We evaluate five different data refactoring implementations.

- **SOTA-GPU**: We use the state-of-the-art GPU data refactoring in the MGARD lossy compression software [19] as our GPU baseline. Its design includes two performance tuning parameters: thread block size and number of CUDA streams. In our evaluation, we use the best performance achieved by hand tuning those parameters.
- **SOTA-CPU**: We use the state-of-the-art CPU data refactoring implemented in the MGARD lossy compression software [19], parallelized with MPI for a fair comparison, as our CPU baseline.
- **OPT**: Our GPU data refactoring, which uses our novel grid/linear/iterative processing kernels (i.e., GPK, LPK, and IPK) but not auto tuning.
- **OPT+AT**: OPT plus auto tuning.

B. Evaluation on kernels

We first show the performance improvement we achieve from accelerating the three major operations in data refactoring on GPUs. Figure 9 shows speedups achieved on the three operations on the two GPU platforms with both single and double precision inputs. The input size is $513 \times 513 \times 513$. For single precision input, with the thread-level load-compute decoupled design, coefficient calculation with GPK outperforms the existing design by $4.9\times$ and $6.9\times$ on Summit and Turing GPUs, respectively. For mass-transfer matrix multiplication, with higher thread concurrency and data dependency free calculation, LPK achieves $6.3\times$ and $4.1\times$ speedups on Summit and Turing GPUs, respectively. For correction solver, IPK triples the performance on Summit and doubles the performance on Turing with the same level of thread concurrency as the state-of-the-art design, thanks to the more efficient memory access patterns. Also, leveraging our heuristic auto tuning capability, the optimum configurations can be selected automatically, yielding additional $1.2\text{--}4.9\times$ speedups compared with choosing one configuration for all kernels and input sizes.

C. Evaluation on data refactoring on a single GPU

Figure 10 shows the end-to-end data refactoring throughput achieved on a single GPU with different input sizes. (As decomposition and recomposition are symmetric processes, they have identical performance.) To see how close the achieved data refactoring throughput is to the theoretical peak throughput, we estimate the theoretical peak by dividing the achievable single pass throughput with the accumulated number of passes on the entire input data over the data refactoring process. (The achievable single pass throughput is the maximum throughput achievable when data are read and stored on GPU memory once. We measured it through a specially designed benchmark kernel that simultaneously reads and writes the same amount of data from and to the GPU memory without computation.)

The accumulated number of passes is calculated by summing the number of passes for all decomposition levels: $\text{passes per level} \times \frac{1}{1-\frac{1}{8}}$. $\text{passes per level} = 1(\text{coefficient calculation}) + 1(\text{copy to workspace}) + 5.25(\text{correction calculation}) + 0.125(\text{apply correction})$.

The theoretical peaks for Summit and Turing GPUs are 49.8 GB/s and 32.0 GB/s, respectively, for both single and double precision data. The state-of-the-art GPU data refactoring methods that we use as our baseline achieve only up to 10.4% of the theoretical peak throughput; our optimized GPU data refactoring achieves up to 83.8% of theoretical peak.

D. Evaluation on multi-node performance at scale

To show the potential of GPU-accelerated data refactoring in large-scale scientific applications, we conduct a weak scaling test on Summit. Here we parallelize the workload by assigning each GPU or CPU core an equal-sized data partition and perform decomposition and recomposition independently. Due to the nature of multigrid-based data refactoring, parallelizing the workload in this way brings near linear speedups with negligible impact on decomposition and recomposition results. We assign each GPU or CPU core to one MPI process and

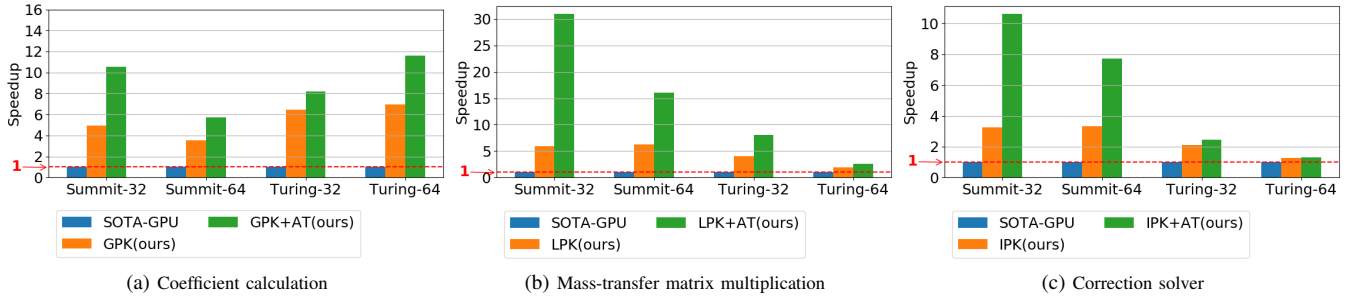


Fig. 9: Speedups achieved through using our proposed processing kernels compared with the state-of-the-art GPU designs. 32 and 64 represent single and double precision input.

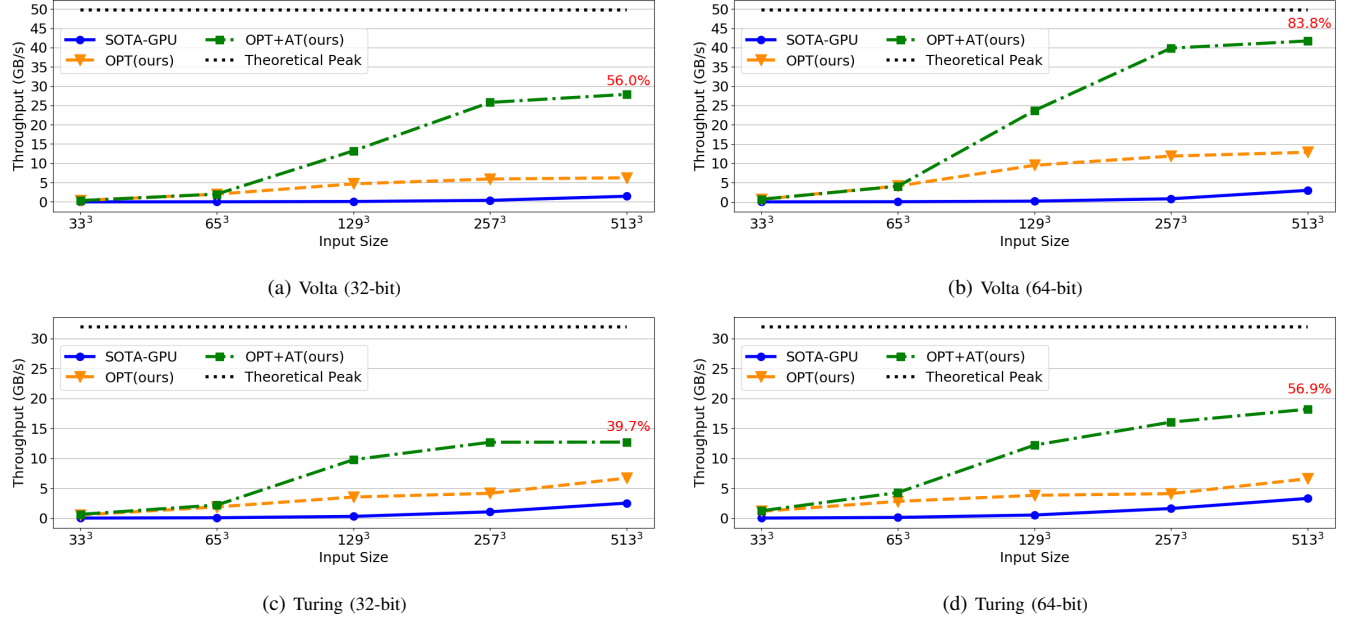


Fig. 10: Data refactoring throughput on a single GPU

perform data refactoring on 1 GB of simulation data. For each computing node, we use the total available number of GPUs and CPU cores. We scale the number of nodes up to 1024 in our tests on Summit. As shown in Figure 11, our optimized GPU data refactoring method achieves much greater throughput than state-of-the-art GPU and CPU designs. For example, we need only four computing nodes to achieve 1 TB/s data throughput, whereas state-of-the-art GPU and CPU designs require 64 and 512 nodes, respectively. With 1024 nodes (i.e., 6144 Volta GPUs), we achieve up to 250.26 TB/s aggregated data refactoring throughput.

V. SHOWCASE

Data refactoring algorithms were designed to offer much greater flexibility when managing large scientific data than the traditional methods. With well-designed data management, data can be shared between scientific applications more intelligently with a large reduction in I/O costs. However, inefficient data refactoring routines can diminish the benefits brought by data refactoring itself. Here we use two examples to show the benefits of GPU-based data refactoring over the CPU designs.

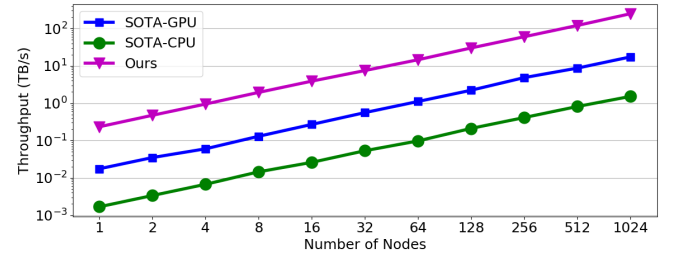
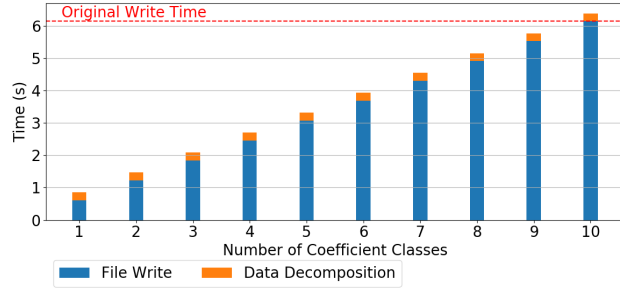


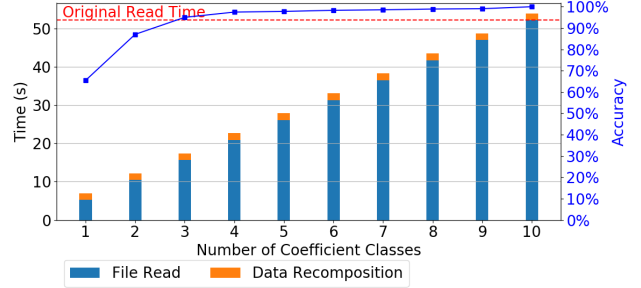
Fig. 11: Aggregated data refactoring throughput at scale on Summit. 6 GPUs or 42 CPU cores are used per computing node, with each GPU or CPU core handling 1 GB in double precision.

A. Visualization workflow

First we show how our GPU optimizations can make data refactoring effective when used for I/O cost reduction in scientific workflows that rely on file-based data sharing. Figure 12 shows the cost of writing and reading a 4 TB simulation data file using 4096 and 512 processes using the state-of-the-art ADIOS I/O library [28] on Summit with GPU-accelerated data refactoring enabled. By writing or reading fewer coefficient classes, we can see immediate cost reduction in file write and read. When our efficient GPU-accelerated data refactoring is

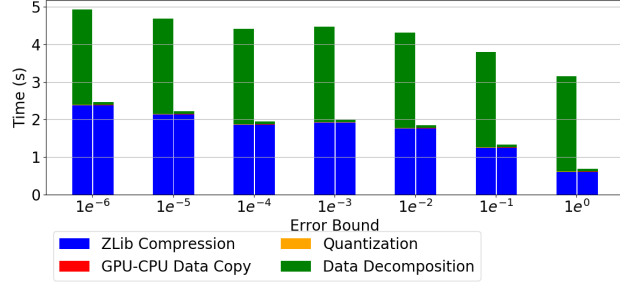


(a) Write simulation data

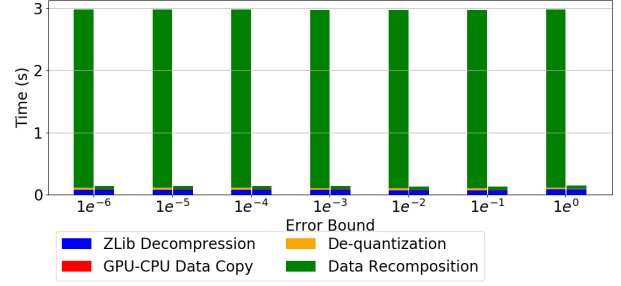


(b) Read simulation data and visualize

Fig. 12: Showcase 1: Data refactoring in scientific visualization workflow



(a) Compression



(b) Decompression

Fig. 13: Showcase 2: MGARD lossy compression using CPU (left bars) vs. GPU (right bars)

used, we can see this reduction in the cost of file write and read can be effectively translated into a reduction in the total I/O cost. Although multigrid-based data refactoring allows us to encode the most important information in the data with a few coefficient classes, it would not reduce the total I/O cost unless those coefficient classes can be efficiently computed or used for data recovery. For example, in our experiments we achieve $\sim 95\%$ accuracy for a chosen feature in the visualization result (i.e., the total area of the iso-surfaces [29, 30]) with only three out of ten coefficient classes. This can be effectively translated into $\sim 66\%$ I/O cost reduction.

B. Lossy compression

Multigrid-based hierarchical data refactoring can also be used as a preconditioner in scientific lossy compression software. As one of the key components in lossy compression workflows, it is important to have efficient data refactoring in order to make fast lossy compression possible. We showcase how our GPU-accelerated data refactoring can help improve the performance of lossy compression workflows in the MGARD lossy compression software. MGARD is a CPU-based lossy compressor with three components in its workflow: multigrid-based data refactoring, quantization, and entropy encoding. Figure 13 shows the time breakdown of the each component in MGARD [19] when data refactoring remains on the CPU (left bars) or is off-loaded to the GPU (right bars). In our test, besides the data refactoring process, we also off-load the quantization and de-quantization processes to the GPUs, since it can help reduce the GPU-CPU data transfer cost. The entropy encoding stage (ZLib lossless compression) is kept on the CPU. We can see that our GPU-accelerated data

refactoring can greatly reduce the overall execution time of the lossy compression workflows.

VI. RELATED WORK

Multigrid-based data refactoring shares some similarities with multigrid solvers, such as the use of multiple interlocking grids. But while multigrid solvers aim to accelerate the solving of linear systems, multigrid-based data refactoring aims to reconstruct scientific data progressively with hierarchical representations. This difference in focus leads to fundamental differences in both algorithms and optimization that prevent direct translation of GPU optimizations.

From an algorithmic perspective, although data refactoring and multigrid solvers have some operations in common, data refactoring composes these operations in a unique way. Further, the correction used in data refactoring is designed specifically for the orthogonal projection, while the correction in multigrid solvers is used to generate the fine grid solution. From a GPU optimization perspective: optimizations for data refactoring need to consider handling large-volume scientific data, which means we need to consider not only limited GPU memory but also cases where refactoring process might share resources with original scientific computations on GPUs. So, it is essential to optimize for low memory footprint as well as performance. Although part of the kernels used in data refactoring share similar computation patterns to those found in multigrid solvers, it is challenging to leverage existing work directly to achieve good parallelism and memory footprint balance in data refactoring. For example, state-of-the-art GPU refactoring [19] uses a parallelization technique proposed by Basu et al. [31], which only use coarse grain vector-wise parallelism, which can cause lower performance for data

refactoring. Although fine-grain parallelism has been achieved in previous works [32–36], it generally brings high memory footprint and it would require considerable effort to apply the optimizations to different algorithms.

VII. CONCLUSION

We have presented optimized data refactoring kernels that allow for use of GPUs to accelerate multigrid-based hierarchical refactoring for scientific data. We evaluated our designs on two platforms, including the leadership-class Summit supercomputer at ORNL, and showed that our GPU version can speed up data refactoring by up to $145\times$ and $14\times$ compared with state-of-the-art CPU and GPU designs, respectively, and can achieve 250 TB/s throughput using 1024 nodes on Summit. We also showcased our work using a large-scale scientific visualization workflow and the MGARD lossy compression technique. Together, these results demonstrate that scientists have another opportunity for dealing with their high data throughput requirements. Inline refactoring of scientific data can offer performance improvements and temporal fidelity that can benefit a number of science scenarios.

ACKNOWLEDGMENT

This work was made possible by support from the Department of Energy’s Office of Advanced Scientific Computing Research, including via the CODAR and ADIOS Exascale Computing Project (ECP) projects. This research used resources of the Oak Ridge Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

REFERENCES

- [1] F. Alexander *et al.*, “Exascale applications: Skin in the game,” *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2166, 2020.
- [2] L. Wan *et al.*, “Data management challenges of exascale scientific simulations: A case study with the Gyrokinetic Toroidal Code and ADIOS,” in *The 10th International Conference on Computational Methods*, ser. ICCM’19, 2019.
- [3] S. Ku *et al.*, “Full-f gyrokinetic particle simulation of centrally heated global ITG turbulence from magnetic axis to edge pedestal top in a realistic tokamak geometry,” *Nuclear Fusion*, vol. 49, no. 11, 2009.
- [4] C.-S. Chang *et al.*, “Numerical study of neoclassical plasma pedestal in a tokamak geometry,” *Physics of Plasmas*, vol. 11, no. 5, 2004.
- [5] R. Taylor *et al.*, “Science with the Square Kilometer Array: Motivation, key science projects, standards and assumptions,” *arXiv preprint astro-ph/0409274*, 2004.
- [6] L. Wan *et al.*, “Comprehensive Measurement and Analysis of the User-Perceived I/O Performance in a Production Leadership-Class Storage System,” in *IEEE 37th International Conference on Distributed Computing Systems*, ser. ICDCS ’17, 2017, pp. 1022–1031.
- [7] L. Wan *et al.*, “Analysis and Modeling of the End-to-End I/O Performance in OLCF’s Titan Supercomputer,” in *IEEE 19th International Conference on High Performance Computing and Communications*, ser. HPCC ’17, 2017, pp. 1–9.
- [8] “HPSS,” www.hpss-collaboration.org/. Accessed: 10/2020.
- [9] M. Ainsworth *et al.*, “Multilevel techniques for compression and reduction of scientific data—the univariate case,” *Computing and Visualization in Science*, vol. 19, no. 5–6, pp. 65–76, 2018.
- [10] M. Ainsworth *et al.*, “Multilevel techniques for compression and reduction of scientific data—the multivariate case,” *SIAM J. Scientific Computing*, vol. 41, no. 2, 2019.
- [11] M. Ainsworth *et al.*, “Multilevel techniques for compression and reduction of scientific data—quantitative control of accuracy in derived quantities,” *SIAM J. Scientific Computing*, vol. 41, no. 4, 2019.
- [12] I. Foster *et al.*, “Online data analysis and reduction: An important co-design motif for extreme-scale computers,” *International Journal of High-Performance Computing Applications*, vol. in press, 2020.
- [13] L. Wan *et al.*, “SSD-Optimized Workload Placement with Adaptive Learning and Classification in HPC Environments,” in *30th International Conference on Massive Storage Systems and Technology*, ser. MSST ’14, 2014.
- [14] L. Wan *et al.*, “Optimizing Checkpoint Data Placement with Guaranteed Burst Buffer Endurance in Large-Scale Hierarchical Storage Systems,” *Journal of Parallel and Distributed Computing*, vol. 100, pp. 16–29, 2017.
- [15] A. Li *et al.*, “Tartan: evaluating modern gpu interconnect via a multi-gpu benchmark suite,” in *2018 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2018, pp. 191–202.
- [16] A. Li *et al.*, “Evaluating modern gpu interconnect: Pcie, nvlink, nv-sli, nvswitch and gpudirect,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 1, pp. 94–110, 2019.
- [17] “Mass matrix computation in the finite element method,” <https://demonstrations.wolfram.com/MassMatrixComputationInTheFiniteElementMethod>.
- [18] X. Liang *et al.*, “Optimizing multi-grid based reduction for efficient scientific data management,” *arXiv preprint arXiv:2010.05872*, 2020.
- [19] *MGARD Lossy Compression Software*, 2020 (accessed April 21, 2020). [Online]. Available: <https://github.com/CODARcode/MGARD>
- [20] K. E. Atkinson *et al.*, *Elementary numerical analysis*. Wiley New York, 1985.
- [21] J. Chen *et al.*, “TSM2: optimizing tall-and-skinny matrix-matrix multiplication on GPUs,” in *ACM International Conference on Supercomputing*, 2019, pp. 106–116.
- [22] C. Rivera *et al.*, “Tsm2x: High-performance tall-and-skinny matrix-matrix multiplication on gpus,” 2020.
- [23] J. Chen *et al.*, “Online algorithm-based fault tolerance for cholesky decomposition on heterogeneous systems with gpus,” in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2016, pp. 993–1002.
- [24] J. Chen *et al.*, “Fault tolerant one-sided matrix decompositions on heterogeneous systems with gpus,” in *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2018, pp. 854–865.
- [25] J. Chen *et al.*, “Gpu-abft: Optimizing algorithm-based fault tolerance for heterogeneous systems with gpus,” in *2016 IEEE International Conference on Networking, Architecture and Storage (NAS)*. IEEE, 2016, pp. 1–2.
- [26] J. E. Pearson, “Complex patterns in a simple system,” *Science*, vol. 261, no. 5118, pp. 189–192, 1993.
- [27] “Gray-Scott Simulation Code,” <https://github.com/pnorbert/adiosvm/tree/master/Tutorial/gray-scott>, [Online; accessed 2019].
- [28] Q. Liu *et al.*, “Hello ADIOS: The challenges and lessons of developing leadership class I/O frameworks,” *Concurrency and Computation: Practice and Experience*, vol. 26, no. 7, pp. 1453–1473, 2014.
- [29] J. Chen *et al.*, “Understanding performance-quality trade-offs in scientific visualization workflows with lossy compression,” in *2019 IEEE/ACM 5th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-5)*. IEEE, 2019, pp. 1–7.
- [30] I. Yakushin *et al.*, “Feature-preserving lossy compression for in situ data analysis,” in *49th International Conference on Parallel Processing-ICPP: Workshops*, 2020, pp. 1–9.
- [31] P. Basu *et al.*, “Compiler-based code generation and autotuning for geometric multigrid on GPU-accelerated supercomputers,” *Parallel Computing*, vol. 64, no. C, 2017.
- [32] N. Bell *et al.*, “Exposing fine-grained parallelism in algebraic multigrid methods,” *SIAM J. Scientific Computing*, vol. 34, no. 4, 2012.
- [33] K. Esler *et al.*, “GAMPACK (GPU accelerated algebraic multigrid package),” in *13th European Conference on the Mathematics of Oil Recovery*, 2012.
- [34] J. Sebastian *et al.*, “GPU accelerated three dimensional unstructured geometric multigrid solver,” in *International Conference on High Performance Computing and Simulation*, ser. HPCS ’14, 2014.
- [35] C. Richter *et al.*, “Multi-GPU acceleration of algebraic multi-grid preconditioners for elliptic field problems,” *IEEE Transactions on Magnetics*, vol. 51, no. 3, 2015.
- [36] M. A. Clark *et al.*, “Accelerating lattice QCD multigrid on GPUs using fine-grained parallelization,” in *SC’16*, 2016, pp. 68:1–68:12.