The Journal of Chemical Physics

RESEARCH ARTICLE | JULY 10 2023

TAMM: Tensor algebra for many-body methods

Special Collection: High Performance Computing in Chemical Physics

Erdal Mutlu ; Ajay Panyala ; Nitin Gawande ; Abhishek Bagusetty ; Jeffrey Glabe ; Jinsung Kim ; Karol Kowalski ; Nicholas P. Bauman ; Bo Peng ; Himadri Pathak ; Jiri Brabec ; Sriram Krishnamoorthy

Check for updates

J. Chem. Phys. 159, 024801 (2023)
https://doi.org/10.1063/5.0142433

View Online
Export Citation
CrossMark

17 August 2023 00:38:49

500 kHz or 8.5 GHz?
And all the ranges in between.
Lock-in Amplifiers for your periodic signal measurements

Find out more
Zurich Instruments

AIP Publishing

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

TAMM: Tensor algebra for many-body methods

Cite as: J. Chem. Phys. 159, 024801 (2023); doi: 10.1063/5.0142433
Submitted: 13 January 2023 • Accepted: 14 June 2023 •
Published Online: 10 July 2023

View Online Export Citation CrossMark

Erdal Mutlu,[1,a] Ajay Panyala,[1,b] Nitin Gawande,[2] Abhishek Bagusetty,[3] Jeffrey Glabe,[1] Jinsung Kim,[4] Karol Kowalski,[5] Nicholas P. Bauman,[5] Bo Peng,[5] Himadri Pathak,[1] Jiri Brabec,[6] and Sriram Krishnamoorthy[7,c]

AFFILIATIONS

[1] Advanced Computing, Mathematics, and Data Division, Pacific Northwest National Laboratory, Richland, Washington 99354, USA
[2] Intel Corporation, Richland, Washington 99352, USA
[3] Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA
[4] School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
[5] Physical Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, USA
[6] J. Heyrovský Institute of Physical Chemistry, Academy of Sciences of the Czech Republic, 182 23 Prague 8, Czech Republic
[7] Google Inc., Mountain View, California 94043, USA

Note: This paper is part of the JCP Special Topic on High Performance Computing in Chemical Physics.
[a] **Author to whom correspondence should be addressed:** erdal.mutlu@pnnl.gov
[b] **Electronic mail:** ajay.panyala@pnnl.gov
[c] **Electronic mail:** sriram.krishnamoorthy@gmail.com

ABSTRACT

Tensor algebra operations such as contractions in computational chemistry consume a significant fraction of the computing time on large-scale computing platforms. The widespread use of tensor contractions between large multi-dimensional tensors in describing electronic structure theory has motivated the development of multiple tensor algebra frameworks targeting heterogeneous computing platforms. In this paper, we present Tensor Algebra for Many-body Methods (TAMM), a framework for productive and performance-portable development of scalable computational chemistry methods. TAMM decouples the specification of the computation from the execution of these operations on available high-performance computing systems. With this design choice, the scientific application developers (domain scientists) can focus on the algorithmic requirements using the tensor algebra interface provided by TAMM, whereas high-performance computing developers can direct their attention to various optimizations on the underlying constructs, such as efficient data distribution, optimized scheduling algorithms, and efficient use of intra-node resources (e.g., graphics processing units). The modular structure of TAMM allows it to support different hardware architectures and incorporate new algorithmic advances. We describe the TAMM framework and our approach to the sustainable development of scalable ground- and excited-state electronic structure methods. We present case studies highlighting the ease of use, including the performance and productivity gains compared to other frameworks.

© 2023 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0142433

17 August 2023 00:38:49

I. INTRODUCTION

Enabling highly-scalable computational environments that abstract and automate the development of complex tensor algebra operations is critical in order to advance computational chemistry toward more complex and accurate formulations capable of taking advantage of emerging exascale computing architectures and also serve as a foundation for a sustainable and portable electronic structure software stack. In particular, tensor contractions (TCs) are a universal language used in many areas of quantum mechanics to encode equations describing collective phenomena in many-body quantum systems encountered in quantum field theory, nuclear structure theory, material sciences, and quantum chemistry. Typically, contractions between

© Author(s) 2023

multi-dimensional tensors stem from the discretization procedures used to represent the Schrödinger equation in a finite-dimensional algebraic form. An important area where tensor contractions play a crucial role is electronic structure theory, where tensors describe basic interactions and parameters defining wave function expansions.

One of the most critical applications of TCs is the coupled-cluster (CC) formalism.[1–8] In the CC theory, the form of the complex TCs used to represent non-linear equations describing the correlated behavior of electrons in molecular systems also reflects the fundamental feature of the CC formalism referred to as the size-extensivity or proper scaling of the energy with the number of particles. For this reason, the CC formalism is one of the most accurate computational models used these days in computational chemistry.

The CC theory has assumed a central role in high-accuracy computational chemistry and still attracts much attention in theoretical developments and numerical implementations. In this effort, high-performance computing (HPC) and the possibility of performing TCs in parallel play an essential role in addressing steep polynomial scaling as a function of system size and extending CC formalism to realistic chemical problems described by thousands of basis set functions. To understand the scale of the problem, canonical CC formulations such as the ubiquitous CCSD(T) approach[9] (CC with iterative single and double excitations with non-iterative corrections due to triple excitations) involve contractions between four-dimensional tensors where ranges of thousands of entries can define each dimension. In order to alleviate efficient CC calculations on parallel platforms, several specialized tensor libraries have been developed over the last decade.

In the last few decades, significant effort has been expended toward enabling CC simulations for very large chemical systems to extend the applicability of the CC formalism further. In the reduced-scaling formulations, commonly referred to as the local CC methods,[10–18] one takes advantage of the local character of correlation effects to effectively reduce the number of parameters and the overall cost of CC iterations, allowing for calculations of systems described by 10 000–40 000 basis set functions. This paper discusses a new tensor library, Tensor Algebra for Many-body Methods (TAMMs), which provides a flexible environment for expressing and implementing TCs for canonical and local CC approximations. The development of TAMM functionalities for the ground-state formulations [canonical CCSD and CCSD(T) methods and local CC formulations] is supported by the NWChemEx project,[19] whereas excited-state formulations [CC Green's function (CCGF), Equation-of-Motion (EOM) CC, and time-dependent CC (TDCC)] are supported by the SPEC project.[20]

As HPC systems continue to evolve to include different types of accelerators, diverse memory hierarchy configurations, and varying intra- and inter-node topologies, there is a need to decouple the development of electronic structure methods from their optimizations for various platforms. TAMM enables the compact specification of various electronic structure methods that heavily rely on tensor operations while allowing independent yet incremental development of optimization strategies to map these methods to current and emerging heterogeneous platforms.

The rest of the paper is organized as follows. Section II briefly discusses other tensor algebra frameworks used for developing computational chemistry applications. Section III describes the details of our tensor algebra interface and the underlying constructs that are used to efficiently distribute tensor data and schedule and execute tensor operations on modern HPC platforms. Later in Sec. III C, we provide a feature comparison with other distributed tensor algebra frameworks. Section IV showcases multiple CC methods implemented using TAMM and performance results obtained using the OLCF Summit Supercomputer.[21] Section V demonstrates the performance of TAMM in comparison to other distributed tensor algebra frameworks.

II. REVIEW OF EXISTING INFRASTRUCTURE

Tensor-based scientific applications have been widely used in different domains, from scientific simulation to more recent machine learning (ML)-based scientific applications. Over the years, program synthesis and code generation have become the go-to solutions for such applications. The Tensor Contraction Engine (TCE),[22] which is used in the NWChem computational chemistry software package,[23] has been one such successful solution for automatically generating parallel code for various molecular orbital (MO) basis CC methods in Fortran 77. In later work, the TCE project[24] added support for optimizations on tensor expression factorization, optimized code generation for various hardware, and space-time trade-offs in order to improve and also implement more complex CC methods.

In a separate effort, the FLAME project[25] provided formal description support for describing linear algebra operations over matrices with support for the optimized implementation of these kernels for distributed memory systems. Later, various studies over-optimizing tensor algebra[26–28] have been proposed using the FLAME framework.

The Cyclops Tensor Framework (CTF)[29] was developed, aiming for a more efficient kernel implementation for tensor operations using concurrency. The framework focused on reducing the required communication in parallel executions of CC-based methods by using a dense triangular tensor representation. Solomonik and Hoefler extended CTF for general sparse computations.[30] Manzer *et al.* demonstrated the benefits of exploiting block sparsity in electronic structure calculations.[31] Neither approach includes notational support for the general block-structured nature of the sparsity that naturally occurs in electronic structure methods.

Epifanovsky *et al.*[32] developed an object-oriented C++ library called `libtensor` for efficient execution of post-Hartree–Fock electronic structure methods using a blocked representation of large-size dense tensors. In later work,[33] they optimized various operations using efficient memory management techniques that are thread-friendly and NUMA-aware. Unlike TAMM, libtensor does not offer a distributed infrastructure for tensor algebra operations and is restricted to shared memory systems.

Orca is a general quantum chemistry program involving implementations of reduced-scaling methods. In order to achieve the speedup, various approximations are employed, for example, density fitting, RIJ-COSX, or a local approach for NEVPT2 or CC methods. The C++ code employs message passing interface (MPI)-based parallelization schemes, and recently, in a pilot study, they implemented a scheme for generating 3- and 4-index integrals via

17 August 2023 00:38:49

© Author(s) 2023

accelerators.[34] Orca provides a wide range of computational chemistry methods, whereas TAMM provides a general distributed tensor algebra framework to enable the productive development of scalable computational chemistry applications.

The ExaTensor[35] library uses a domain-specific virtual processor concept to allow performance portable programming and execution of tensor operations on modern graphics processing unit (GPU) architectures. While the library is mostly focused on general GPU computation and the portability of such operations to different systems, they demonstrated its effectiveness on numerical tensor algebra workloads that mainly focus on quantum many-body computations on HPC platforms. The main focus of the ExaTensor library is to utilize GPUs for efficient tensor contractions. TAMM, on the other hand, provides a large set of supported tensor operations and utility routines. TAMM also provides special constructs allowing users to add new operations on tensors that are needed for the development of complete computational chemistry applications.

The TiledArray framework[36] is actively being developed for scalable tensor operations that support various computational chemistry methods.[37] It makes use of the Multiresolution Adaptive Numerical Environment for Scientific Simulation (MADNESS) parallel runtime,[38] which employs a high-level software environment for increasing both programmer productivity and code scalability in massively parallel systems. TiledArray employs a hierarchical view of sparsity coupled with explicit user-written loop nests to perform specialized operations over sparse tensors. TiledArray has several similarities with TAMM while also differing in several key aspects, which are explained later in Sec. III C.

DISTAL[39] is another recently developed distributed tensor algebra compiler that allows expressing computation on tensors as tensor algebra expressions. DISTAL lets users describe how the data and computation of a tensor algebra expression map onto a target machine, allowing separation of the specifications of data distribution and computation distributions. DISTAL lets users specialize computation to the way that data are already laid out or easily transform data between distributed layouts to match the computation. TAMM, on the other hand, offers generality by decomposing tensor algebra expressions into a series of distributed matrix multiplication and transposition operations. In a more recent work, the authors also introduced SpDISTAL,[40] which adds support for sparse data structures and allows users to define generic sparse tensors. SpDISTAL then generates code for executing operations over these tensors. Both DISTAL and TAMM have several similarities in the aspects of storage, computation, and scheduling. However, they also differ in certain key aspects, which are further elaborated in Sec. III C.

III. TAMM FRAMEWORK

This section provides a detailed explanation of our tensor algebra framework. Figure 1 shows the conceptual overview of our framework. TAMM provides a tensor algebra interface that allows users to describe tensor operations in a familiar mathematical notation, while underneath it employs efficient data distribution/management schemes as well as efficient operation execution on both central processing unit (CPU)-based and GPU-based hardware. Our framework leverages efficient libraries such as HPTT[41] for tensor index permutations on CPUs, LibreTT[42] for tensor index permutations on GPUs, and Global Arrays,[43] a partitioned global address space (PGAS) programming model for efficient data distribution.

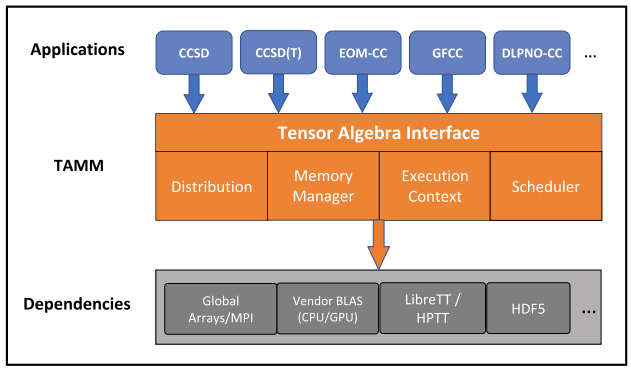

FIG. 1. Overview of tensor algebra for many-body methods (TAMM) framework.

A. Tensor algebra operations in TAMM

TAMM provides a flexible infrastructure for describing tensor objects using the general notion of an ***index space*** (simply an index range) that is used to describe the size of each dimension. TAMM also employs tiling capabilities, where users can define arbitrary or fixed size tiling over ***index spaces*** to construct ***tiled index spaces*** to represent a blocked representation of a tensor. This allows tensors to be constructed as a set of smaller blocks of data indexed by the Cartesian product of the number of tiles on each dimension. This notion of tiling enables efficient data distribution and efficient tensor operations (such as tensor contractions) that can better leverage the underlying hardware by utilizing the available cache system as well as the execution modules (i.e., GPUs and CPUs). TAMM's flexible tiling capabilities provide crucial support for load balancing and making effective trade-offs over data communication and efficient computation. These optimizations will be detailed in Subsection III B, where we describe the data communication and execution of tensor operations.

Figure 2 shows an example of ***index space***, ***tiled index space***, and ***tensor*** construction in the TAMM framework. Lines 2–4 show the construction of `IndexSpace` objects that represent a range of indices using a range constructor. Our framework also allows users to describe application-specific annotations over the subsections of these spaces. Lines 4–7 show an example of string-based annotation over subsections of the construction range. This allows users to easily access different portions of an index space, thereby enabling access to slices of a tensor and constructing new tensors using slices of the original index space. Similarly, users can encode ***spin*** information related to each dimension over the input tensors, allowing our run-time to allocate these tensors using a block-sparse representation. For the time being, we only support spin symmetry for representing the block sparsity where the zero-blocks are not stored. Users can define their own non-zero check functionality to be used during allocation and tensor operation execution for storage and computational benefits. Additional capabilities, such as restricted

17 August 2023 00:38:49

© Author(s) 2023

```
// Constructing index spaces
IndexSpace N{range(100)};
IndexSpace M{range(30)};
IndexSpace K{range(20)},
                {{" first ", {range(0, 10)}},
                 {"second", {range(10, 20)}}
                }};

// Tiling the index spaces
TiledIndexSpace tN{N, 10};
TiledIndexSpace tM{M, {10,20}};
TiledIndexSpace tK{K, 5};

// Constructing a tensor object
Tensor <double> A{tM, tK};
Tensor <double> B{tK, tN};
Tensor <double> C{tM, tN};

Tensor<double> ST{{tM, tN}, [] (const IndexVector& block_id,
      span<T> buf){
          /* user-defined logic to fill tensor blocks */ }};
```

FIG. 2. Source code example for `IndexSpace`, `TiledIndexSpace`, and `Tensor` construction using TAMM.

summations or point group and permutational symmetries, are not supported in the tensor specification. Point group symmetry is less concerning since TAMM targets large systems that often do not have such symmetry. However, developers can explicitly code these features using TAMM to construct the desired tensors. This is the case for methods such as CCSD(T), discussed in Sec. IV A 2. TAMM also supports specialized tensor construction, called ***lambda tensor***, where the user can provide a C++ lambda expression that specifies how a block of the tensor is computed, as shown in Line 19. This construct can be used to dynamically generate blocks of a tensor as needed. Note that these tensors are not stored in memory; they are read-only objects computed on the fly to be consumed in various tensor operations described later in this section. Similarly, TAMM also provides a special construct called ***view tensor*** to describe access to an existing tensor using C++ lambda expressions. The main use of view tensors is to define tensors of different shapes that can be used as a reference tensor for any operation as well as apply possible sparsity mapping/constraints on a dense tensor. Similar to lambda tensors, a view tensor does not allocate any additional storage but rather provides specialized access to the data available in the referenced tensor. We provide details of a use case for view tensors when discussing the DLPNO-CCSD method in Sec. IV B 2.

Lines 10–12 show the construction of `TiledIndexSpace` objects that represent a sliced index space that is used in tensor construction to have a blocked structure. Tiling can be applied as a fixed size (i.e., line 10) or as arbitrary tile sizes with full coverage of the index space (i.e., line 11). Finally, lines 15–17 in Fig. 2 show the construction of `Tensor` objects using tiled index spaces. Each `TiledIndexSpace` object used in the tensor constructor represents a dimension in the constructed tensors. In this example, tensor `A` is a 30 × 20 matrix with eight blocks, as denoted by `tM`, which consists of two tiles of sizes 10 and 20, while `tK` consists of four tiles that are

```
// construct tN labels
auto [a, b, c] = tN.labels<3>();
// construct tM labels
auto [i, j, k] = tM.labels<3>();
// construct tK labels
auto [l, m, n] = tK.labels<3>();
// construct subspaces of tK
auto [x, y, z] = tK(" first ").labels<3>();
```

FIG. 3. Source code example for constructing index labels.

size 5. While these lines construct a tensor object, the tensor is collectively allocated by all participating compute nodes in a subsequent operation using a ***scheduler***.

Another important concept in constructing tensor operations is ***index labels***, which allow specifying tensor operations in familiar mathematical notations and provide slicing capabilities over tensors by using the string-based subsections of the full index space. Labels are associated with ***tiled index spaces*** and used in the tensor operation syntax to describe the computation. Depending on the index spaces that the tensors are constructed on, users can specify string-based sub-spaces to define operations on different slices of the underlying allocated tensor object. Figure 3 shows examples of `TiledIndexLabel` construction using `TiledIndexSpace` objects. While lines 1–6 construct the labels over the full spaces, line 8 shows the label creation for the ***first*** portion of the `tK` index space (see construction on Fig. 2, line 4). These sub-spaces can then be used for specifying operations over sliced portions of the full tensor, as long as the index labels are from a subset of the original index space.

TAMM supports several tensor operations: tensor set, tensor addition, subtraction, scale, trace, transpose, general tensor contractions, inner and outer products, and reduction. TAMM supports different tensor data types and several mathematically valid operations between tensors of different data types (e.g., tensor contraction between a complex and real tensor). Figure 4 gives the grammar for allowed tensor operations' syntax in the TAMM framework. Each tensor operation syntax rule ($\langle$tensor − op$\rangle$) is composed of a left-hand side (lhs, $\langle$op − lhs$\rangle$) and a right-hand side (rhs, $\langle$rhs$\rangle$) operation. While ***lhs*** can only be a labeled tensor construct ($\langle$label − tensor$\rangle$), ***rhs*** can be of different types that result in different tensor operations, as follows:

- alpha value `[A(i,l) = alpha]`, which corresponds to a tensor set operation that assigns the corresponding alpha value to all elements of the tensor.
- labeled tensors `[A(i,l) += alpha * D(l,i)]` correspond to a tensor addition operation (with respect to the label permutation on tensor D) in Eq. (7) in Fig. 4.
- contraction of two labeled tensors `[C(i,a) += alpha * A(i,l) * B(l,a)]` updates the lhs with the tensor contraction results in Eq. (8) in Fig. 4.

Several additional operations on tensors are provided as utility routines in TAMM. Element-wise operations such as square, log, inverse, etc. are provided. Additional useful operations on tensors such as min, max, norm, etc. are also provided, including high-level routines to read from and write tensors to disk using parallel file I/O.

17 August 2023 00:38:49

© Author(s) 2023

$$
\begin{aligned}
\langle \text{tensor-op} \rangle &::= \langle \text{op-lhs} \rangle = \langle \text{op-rhs} \rangle && (1)\\
&\mid \langle \text{op-lhs} \rangle \mathrel{+}= \langle \text{op-rhs} \rangle && (2)\\
&\mid \langle \text{op-lhs} \rangle \mathrel{-}= \langle \text{op-rhs} \rangle && (3)\\
\langle \text{op-lhs} \rangle &::= \langle \text{label-tensor} \rangle && (4)\\
\langle \text{op-rhs} \rangle &::= \langle \text{alpha} \rangle && (5)\\
&\mid \langle \text{label-tensor} \rangle && (6)\\
&\mid \langle \text{alpha} \rangle * \langle \text{label-tensor} \rangle && (7)\\
&\mid \langle \text{alpha} \rangle * \langle \text{label-tensor} \rangle * \langle \text{label-tensor} \rangle && (8)\\
\langle \text{label-tensor} \rangle &::= \langle \text{tensor-id} \rangle \ (\ \langle \text{label-list} \rangle\ ) && (9)\\
\langle \text{alpha} \rangle &::= \textit{tensor value type} && (10)
\end{aligned}
$$

FIG. 4. Tensor operations grammar in extended Backus–Naur form.

```
// Constructing process group, memory manager, distribution
// to construct an execution context
ProcGroup pg = ProcGroup:: create_world_coll ();
ExecutionContext ec{pg, DistributionKind :: dense,
      MemoryManagerKind::ga};

// Constructing a scheduler for executing tensor operations
Scheduler sch{&ec};

// Tensor operations are queued for execution into the Scheduler
sch. allocate (A, B, C)
(A(i, 1) = 1.0)                          // SetOp – Grammar rule (5)
(B(i, 1) = A(i, 1))                      // AddOp – Grammar rule (6)
(B(i, 1) += -1.0 * A(i, 1))              // AddOp – Grammar rule (7)
(C(i, a) = 0.5 * A(i, 1) * B(1, a))      // MultOp – Grammar rule (8)
// All queued operations are executed.
. execute ();
```

FIG. 5. Source code example for executing the tensor operations using TAMM.

All these operations are defined using a parallel ***block_for*** construct provided by TAMM, allowing users to easily define custom element- and block-wise operations on a tensor. The ***block_for*** construct allows user-defined operations (using C++ lambda expressions) that are executed in parallel on distributed tensors.

Similar to tensor allocation, all other tensor operations that use and modify the distributed tensor data have to be performed via a ***scheduler*** that has the necessary information about the ***execution context*** of the operation. ***Execution context*** includes required information about the distribution scheme of the tensor data and the execution medium for the operations through ***distribution***, ***memory manager***, and ***process group*** constructs. Note that the tensor operations that are provided as utility routines do not use the scheduler but still require the ***execution context*** to perform the corresponding operation. TAMM employs a modular construction of these concepts to allow users to implement various distribution schemes as well as different distributed memory frameworks. Figure 5 shows the construction of a `Scheduler` object using the ***execution context*** and the execution of defined tensor operations. After creating a scheduler, users can directly queue tensor operations such as tensor allocation (line 12), tensor set and add operations (lines 13 and 14), and tensor contractions (line 15). Finally, all queued operations are executed using the ***execution context*** on a distributed system. The tensor operations, as well as the operations over the index spaces, are formally described in our previous work.[44] The syntax for expressing operations shown in lines 13–15 also indicates the productivity benefits that can be obtained by using TAMM. The operations expressed in these three lines are executed in parallel on CPUs, GPUs, or both on any distributed computing platform. Manually writing parallel code for these operations would lead to a significant increase in the number of source lines of code (SLOC). Extending such manual development of parallel code to a real application with a large number of such operations would only lead to a significant increase (orders of magnitude) in the SLOC count and development time, which would also make future improvements to such code infeasible.

This section summarizes the tensor algebra interface as an embedded domain-specific language (eDSL) in the TAMM framework. By implementing an eDSL, we were able to separate concerns for developing scientific applications. While the high-level tensor abstractions allowed domain scientists to implement their algorithms using a representation close to mathematical formulation, it also allowed framework developers to test various different optimization schemes on the underlying constructs (i.e., different data distribution schemes, operation execution on accelerators, use of different PGAS systems, etc.), which we will detail in the coming section.

B. Tensor distribution and operation execution in TAMM

TAMM leverages various state-of-the-art frameworks and libraries to achieve a scalable performance–portable implementation of tensor algebra operations on exascale supercomputing platforms through efficient data distribution and intra-node execution of tensor operation kernels on CPUs and GPUs. The default ***memory manager*** for tensor data distribution in TAMM is based on the Global Arrays framework, a Partitioned Global Address Space (PGAS) programming model that provides a shared memory-like programming model on distributed memory platforms. Global Arrays provides performance, scalability, and user productivity in TAMM by managing the inter-node memory and communication for tensors. A TAMM tensor is essentially a global array with a certain distribution scheme. We have implemented three tensor distribution schemes in TAMM. The first scheme computes an effective processor grid for a given number of processes. A dense tensor is then mapped onto the processor grid. The second scheme is a simple round-robin distribution that allocates equal-sized blocks in a round-robin fashion, where the block size is determined by the size of the largest block in the tensor. This distribution over-allocates memory and ignores sparsity. The third scheme allocates the tensor blocks in a round-robin fashion while taking block sparsity into account. By only allocating non-zero blocks in the tensor, it minimizes the memory footprint of overall computation, allowing bigger-sized problems to be mapped to the available resources.

TAMM uses the "Single Program Multiple Data (SPMD)" model for distributed computation. In this programming abstraction, each node has its own portion of tensors available locally

17 August 2023 00:38:49

© Author(s) 2023

as well as access to the remote portions via communication over the network. As a result, operations on whole tensors can result in access to remote portions of the tensor, with implied communication. More importantly, many operations (i.e., tensor contractions, addition, etc.) are implied to be collective as they involve the distributed tensors as a whole. While the tensor algebra interface follows a sequential ordering of tensor operations, we also tried to conceal the burden of thinking in a distributed manner while writing a scientific application. To avoid possible issues with operations on distributed tensors, TAMM is designed to represent tensors in terms of handles and requires tensor operations to be declared explicitly and executed using a ***scheduler***. Hence, any assignment performed on tensor objects will be a shallow copy as opposed to a deep copy, as a deep copy implies communication (message passing between nodes to perform the copy operation).

The computational chemistry methods are implemented as a sequence of operations over distributed tensors. Through this design, we were able to separate the specification of the operations from their implementations, allowing method developers to mainly focus on the computational chemistry algorithms while kernel developers can focus on the optimization of individual tensor operations. The execution of all tensor operations is managed by a scheduler. The TAMM scheduler employs a data flow analysis over the queued tensor operations to find the dependencies between each operation in order to load balance the computation and communication requirements of the overall execution. Using a levelized execution scheme, the scheduler is able to limit the global synchronizations to achieve load-balanced and communication-efficient schedules for the execution of all tensor operations. The dependency analysis over the high-level operations is based on a macro-operation graph. When two or more operations share the same tensor object and one of these operations updates the shared object, the operations are marked as conflicting operations that cannot be executed in parallel. This operation graph is used to construct a batch of operations that can be executed in parallel, minimizing the total number of global synchronizations required by the computation. The operations in these batches are then executed in an SPMD fashion. For instance, the canonical CCSD implementation in TAMM has 10 levels of operation batches that sum to over 125 tensor operations.

While TAMM hides the burden of choosing the best-performing schedule from the users through load-balanced scheduler execution, it allows users to control various aspects of the computation, such as data distribution, parallelization strategies, operation ordering, and execution medium (i.e., CPUs, GPUs). To achieve this, TAMM uses a modular structure to describe constraints imposed by the users to automate the construction of an execution plan for efficient execution of the computation. This allows users to incrementally optimize their code with minimal programming effort while keeping it readable, as opposed to a code generator-based solution. With these controls over the execution and distribution of the tensor data, users can choose from different optimizations at different granularities. For example, the user can increase the data locality by replicating frequently used small tensors on each distributed node or choosing from different distribution schemes for efficient tensor operation execution on various node topologies. Such an optimization can be implemented as a new data distribution scheme (i.e., SUMMA[45]) by extending the `Distribution` class. By simply using this distribution, users can enforce a specific distribution on tensors, which can optimize required data communication for specific operations. Another abstraction for optimization is the actual computation of the tensor operations. Users can define new operations by extending the `Op` class, which can then be scheduled to be executed. While TAMM supports the most common operations defined on tensors (i.e., addition and contraction), it also implements a `Scan` and `Map` operation that can be used to define various element/block-wise operations using lambda functions. Additionally, as a lower-level abstraction, users can also decide to describe new executions specific to a new architecture/accelerator on the kernel level (i.e., Double Precision General Matrix Multiply (DGEMM), GPU abstraction). While TAMM supports the main GPUs (i.e., Nvidia, AMD, and Intel) that will be available in upcoming HPC systems, users can also choose to implement new kernel-level abstractions for different hardware stacks [i.e., Field Programmable Gate Arrays (FPGAs) and Machine Learning/Deep Learning (ML/DL) accelerators].

To achieve highly optimized execution of tensor operations on a given platform, TAMM is designed to allow the use of multiple external libraries that provide optimized kernels for various operations. In addition to leveraging vendor-provided linear algebra libraries that are highly tuned for both CPUs and GPUs, TAMM also uses the following libraries: the HPTT[41] library for optimized tensor index permutations on CPUs; LibreTT[42] for enabling efficient index permutation on Nvidia, AMD, and Intel GPUs; the BLIS[46] library for efficient BLAS implementation in CPU; and finally, TensorGen[47] for generating optimized tensor contraction GPU kernels specialized for CC methods in computational chemistry.

While TAMM tries to hide the execution details from the end-user by employing high-level tensor algebra expressions, users can specify the medium on which they want the operations to be executed. Users can specify a particular operation to be executed on the CPU or GPU. All other execution-specific details like parallelization, CPU-to-GPU communication, and execution mechanisms are handled automatically by the TAMM infrastructure. While TAMM does not explicitly rely on OpenMP, it can leverage OpenMP for important operations like tensor transposition and GEMM via highly optimized external libraries such as HPTT and vendor BLAS. Setting the `OMP_NUM_THREADS` variable before running any TAMM-based application would enable OpenMP parallelization.

C. Comparison with other tensor algebra frameworks

In this section, we compare the functionality provided by TAMM with that provided by other distributed tensor algebra frameworks: TiledArray, CTF, DISTAL, and ExaTensor. Table I provides the key features and differences between these frameworks. While all these frameworks provide similar features in how tensors and tensor operations can be represented, TAMM is designed to be an extendable framework and hence differs from the remaining frameworks in the following key aspects discussed below:

- **Indexing and slicing:** Indexing and slicing capabilities in any tensor algebra framework are crucial to defining tensor operations. Indexing plays a key role in constructing tensor operations (see Fig. 5). Slicing is predominantly used in computational chemistry methods, allowing operations

17 August 2023 00:38:49

© Author(s) 2023

on isolated parts of tensors. Users can choose to write the tensor operations over slices of the tensors instead of having to copy the required portions first for such operations. While the tensor algebra frameworks being compared provide an indexing mechanism, they lack support for using slicing in the tensor operations. TA and CTF use a string-based solution for indexing but lack the capability to support sliced indexing. On the other hand, DISTAL and ExaTensor use object-based constructs to create and use labels in the computation. While DISTAL does not allow any slicing, ExaTensor lets users to use pre-defined sub-ranges for slicing. TAMM provides generic indexing and slicing capabilities through various index space operations, allowing users to define a subset of indices to create object-based labels (see Fig. 3) as well as string-based labels to construct tensor operations. Users can make use of slicing capabilities by using object-based labels over sub-ranges or by creating new slices on-the-fly or string-based labels to define operations over full tensors.

- **Sparsity:** Sparsity is widely used in several computational chemistry applications. All of the tensor algebra frameworks discussed in this section support some notion of sparsity for both data and computation. While DISTAL does not have direct support for sparsity, later work by the same authors, namely, SpDISTAL, incorporates different sparse data structures as well as specifications to describe the storage and computation of sparse tensors. SpDISTAL supports a generic specification for describing the dense/compressed format for each individual mode as well as pre-defined sparse data representations (i.e., CSR, CSC, etc.) for describing such tensors. Operations over these tensors are executed by generating custom kernels. To contrast, more application-oriented tensor libraries, such as TA, CTF, and ExaTensor, allow block-level sparsity. TA allows users to define non-zero and zero tiles that construct the tensors, whereas CTF uses pre-defined symmetry-based sparsity over the tensors. TAMM, on the other hand, incorporates sparsity by providing attributes over the tiled index spaces. By default, TAMM implements spin symmetry-related attributes to allocate block sparse tensors depending on the values of corresponding attributes on the tiles of the index space. Internally, TAMM analyzes the tensor blocks to determine if the corresponding block is non-zero to optimize storage and computation on block sparse tensors.
- **GPU Support:** It is crucial to effectively leverage the available resources on modern heterogeneous HPC platforms. TAMM currently enables tensor algebra operations (mainly tensor contractions) on multiple GPU architectures (AMD, Nvidia, and Intel). At the time of this writing, the remaining frameworks are limited to Nvidia GPUs (except ExaTensor, which has AMD GPU support). The unified GPU infrastructure design allows TAMM to be easily extended to add support for a new hardware accelerator. Support for the upcoming GPU architectures with performance portability was also explored using SYCL, a domain-specific, heterogeneous programming language. Using SYCL and oneAPI Math Kernel Library (oneMKL) interfaces,[48] perturbative triples contribution from the CCSD(T) formalism was demonstrated with large-scale simulations on different GPU architectures.[49]
- **Extendable backends:** TAMM currently uses Global Arrays (GAs), which is built on top of MPI, for managing all aspects of data distribution and communication. TAMM also has an experimental UPC++[50] backend that can alternatively be used instead of GA. TiledArray (TA) uses the MADNESS[38] parallel runtime as the backend; DISTAL uses the Legion[51] programming model and runtime system for distributing the data and computing; CTF and ExaTensor use MPI as the underlying parallel programming model. TAMM is designed to be easily extendable by implementing the methods in the ***process group*** and ***memory manager*** classes to support other parallel programming models. To the best of our knowledge, the remaining frameworks are not easily extendable to add support for additional parallel programming models.
- **Distribution:** A TAMM tensor is allocated with a distribution type specified via the ***execution context***. The distribution choices that are supported can be extended via the ***distribution*** class. DISTAL allows the specification of tensor distributions using a notation that allows the user to specify a process grid that the dimensions of a tensor can be mapped onto. Tensor distributions in TA and CTF are based on several parallel matrix multiplication algorithms (SUMMA, 2.5D, and 3D).[26] TA additionally gives users the option of specifying a process map onto which tiles are distributed. However, CTF cannot be easily extended to allow other distributions.
- **Scheduling and Execution:** DISTAL allows users to specialize computation to the way that data are already laid out or easily transform data between distributed layouts to match the computation, allowing fully customizable execution strategies for each operation. Several parallel matrix

TABLE I. Feature comparison with other distributed tensor algebra frameworks.

Frameworks	Indexing	Slicing	GPU support	Programming model
TAMM	Object- and string-based	Generic slicing	AMD, Intel, Nvidia	GA (extendable)
TA	String-based	No support	Nvidia	MADNESS
DISTAL	Object-based	No support	Nvidia	Legion
CTF	String-based	No support	Nvidia	MPI
ExaTensor	Object-based	Sub-range slicing	AMD, Nvidia	MPI

17 August 2023 00:38:49

© Author(s) 2023

multiplication algorithms from the literature[26] are expressible as schedules in DISTAL, which is currently not possible with any other framework in comparison. TA and CTF automatically execute tensor operations by dynamically redistributing data across operations if needed, but do not provide any way to customize the operation scheduling process. TAMM allows deferring execution for a collection of tensor operations. When the user explicitly calls the ***execute*** function, the TAMM scheduler analyzes the dependencies between the various operations and schedules them appropriately in order to load balance the computation and communication requirements of the overall execution across the collection of operations. However, TAMM does not provide any further customizations to the scheduling process. The trade-off between these frameworks is that TAMM, TA, and CTF fully automate the distribution process, while users must explicitly provide a schedule to distribute their computations with DISTAL.

In this section, we have provided a detailed explanation of the TAMM framework. TAMM leverages a modular infrastructure to enable the implementation of various optimizations on different levels of computation, from data distribution to execution schemes on different hardware. This design allowed us to implement different ***memory managers***, ***distribution*** schemes, and work distribution over different ***process groups*** without any major changes to the user-facing tensor algebra interface. Finally, we provided a feature-based comparison between TAMM and other distributed tensor algebra frameworks, namely, TiledArray, CTF, DISTAL, and ExaTensor.

IV. CC TAMM IMPLEMENTATIONS

In this section, we present case studies where TAMM was used to implement various scalable coupled-cluster (CC) methods for the latest HPC systems. While important, TAMM's primary contributions are not just the faster performing versions of these methods but also the ability to productively develop and explore new algorithms and apply those improvements to all existing and new applications implemented using TAMM. This includes improvements in intra-node execution (single core, OpenMP multicore, GPUs, etc.), data distribution strategies (e.g., replication, process group-based distribution), and parallel execution (compute partitioning and communication scheduling algorithms). It also allows the concurrent development of optimized equations, parallel algorithms, and optimized intra-node kernels by different teams through clearly defined interfaces. The variety of methods presented and their performance are evidence that this approach accelerates the development and exploration of CC methods.

A. Canonical methods

The canonical formulations of the Coupled-Cluster (CC) formalisms[1–6,8] are based on the exponential parametrization of the correlated ground-state wave function $|\Psi\rangle$,

$$|\Psi\rangle = e^{T}|\Phi\rangle, \tag{11}$$

where T is the cluster operator and the reference function $|\Phi\rangle$ in single-reference formulations is assumed to be represented by the Hartree–Fock Slater determinant. In practical realizations, it is assumed that $|\Phi\rangle$ provides a good approximation to the correlated ground state, $|\Psi\rangle$. The cluster operator T can be partitioned into its many body components, T_k,

$$T = \sum_{k=1}^{N} T_k, \tag{12}$$

defined as

$$T_k = \frac{1}{(k!)^2} \sum_{i_1,\ldots,i_k;a_1,\ldots,a_k} t_{a_1\ldots a_k}^{i_1\ldots i_k} a_{a_1}^{\dagger} \ldots a_{a_k}^{\dagger} a_{i_k} \ldots a_{i_1}, \tag{13}$$

where $a_p^{\dagger}$ (a_p) are the creation (annihilation) operators for an electron in the p-th state and indices $i_1, i_2, \ldots$ $(a_1, a_2, \ldots)$ refer to occupied (unoccupied) spin orbitals in the reference function $|\Phi\rangle$. The operators T_k, defined by the cluster amplitudes $t_{a_1\ldots a_k}^{i_1\ldots i_k}$, produce k-tuple excitations when acting on the reference function.

To define the equations needed for determining cluster amplitudes, we introduce wave-function expansion (11) into the Schrödinger equation by pre-multiplying both sides by e^{-T}. This procedure leads to an explicitly connected form of the energy-independent equations for amplitudes and energy, i.e.,

$$\langle \Phi_{i_1\ldots i_k}^{a_1\ldots a_k} | e^{-T} H e^{T} | \Phi \rangle = 0, \quad \forall_k, \forall i_1, \ldots, i_k, \quad \forall a_1, \ldots, a_k, \tag{14}$$

$$E = \langle \Phi | e^{-T} H e^{T} | \Phi \rangle, \tag{15}$$

where the electronic Hamiltonian H is defined as

$$H = \sum_{pq} h_q^p a_q^{\dagger} a_p + \frac{1}{4} \sum_{p,q,r,s} v_{rs}^{pq} a_r^{\dagger} a_s^{\dagger} a_q a_p, \tag{16}$$

where h_q^p and v_{rs}^{pq} are tensors representing interactions in the quantum system, and excited Slater determinants $|\Phi_{i_1\ldots i_k}^{a_1\ldots a_k}\rangle$ are defined as

$$|\Phi_{i_1\ldots i_k}^{a_1\ldots a_k}\rangle = E_{i_1\ldots i_k}^{a_1\ldots a_k} |\Phi\rangle. \tag{17}$$

In all CC formulations discussed here, to form Eq. (14) for cluster amplitudes, one needs to (1) find an efficient way for distributing and compressing tensors h_q^p, v_{rs}^{pq}, and $t_{a_1\ldots a_k}^{i_1\ldots i_k}$ across all nodes, (2) define efficient algorithms for partitioning TCs of multi-dimensional tensors across the parallel system, and (3) optimize communication between nodes to minimize the effect of the latency. To illustrate the scale of the problems, in brute force simulations, we have to store four-dimensional tensors $t_{a_1 a_2}^{i_1 i_2}$ and v_{rs}^{pq} that require storage proportional to $n_o^2 n_u^2$ and $(n_o + n_u)^4$, respectively, where n_o and n_u refer to the number of occupied and unoccupied orbitals in the reference function $|\Phi\rangle$. For CC simulations of the molecular systems defined by $n_o = 200$ and $n_u = 2800$, one needs to distribute data of the order of 150 TB. Therefore, to make these simulations possible, the support of sophisticated HPC algorithms and applied math is indispensable.

17 August 2023 00:38:49

© Author(s) 2023

1. Coupled cluster singles doubles (CCSD)

The CCSD formalism (CC with single and double excitations)[4] is one of the most popular CC approximations and is used in routine computational chemistry calculations as a necessary intermediate step toward more accurate CC models, such as the perturbative CCSD(T) formalism discussed in Sec. IV A 2, or excited-state or linear-response CC extensions in Sec. IV B 1. In the CCSD formalism, the cluster operator T is approximated as

$$T \simeq T_1 + T_2, \tag{18}$$

and the equations for cluster amplitudes are represented as

$$r_i^a = \langle \Phi_i^a | e^{-(T_1+T_2)} H e^{T_1+T_2} | \Phi \rangle, \tag{19}$$

$$r_{ij}^{ab} = \langle \Phi_{ij}^{ab} | e^{-(T_1+T_2)} H e^{T_1+T_2} | \Phi \rangle, \tag{20}$$

where the tensors r_i^a and r_{ij}^{ab} are commonly referred to as the residual vectors. Due to the large number of terms corresponding to complicated contractions between h_q^p/v_{rs}^{pq} and t_a^i/t_{ab}^{ij}, optimization of the expressions plays a crucial role. This is achieved by proper factorization by introducing the so-called recursive intermediates. For example,

$$\frac{1}{4} v_{mn}^{ef} t_{ef}^{ij} t_{ab}^{mn}, \tag{21}$$

term, which contributes to r_{ij}^{ab} in the naive approach, is characterized by $n_o^4 n_u^4$ numerical overhead. However, by introducing the intermediate tensor I_{mn}^{ij}, defined as (we assumed Einstein summation convention over repeated indices)

$$I_{mn}^{ij} = v_{mn}^{ef} t_{ef}^{ij}, \tag{22}$$

the term (21) can be given by the equation

$$\frac{1}{4} I_{mn}^{ij} t_{ab}^{mn}, \tag{23}$$

at the total numerical cost proportional to $n_o^4 n_u^2$.

Equation (23) can be expressed using the TAMM interface as follows:

$$\begin{aligned} &(I(m,\ n,\ i,\ j) = v(m,\ n,\ e,\ f) * t(e,\ f,\ i,\ j)), \\ &(r(a,\ b,\ i,\ j) += 0.25 * I(m,\ n,\ i,\ j) * t(a,\ b,\ m,\ n)). \end{aligned} \tag{24}$$

The TAMM code in (24) encapsulates the distributed execution of the tensor contractions on CPUs and GPUs, as well as the choice of the parallel programming model used by TAMM for managing distributed tensor data. Similarly, all other CCSD equations can be represented as tensor operations using the TAMM interface. As indicated in (24), the total number of lines of code is greatly reduced in comparison to hand-coded implementations, where one needs to explicitly manage data distribution using a particular programming model and also aspects of execution on various heterogeneous architectures. On the other hand, the TAMM representation of such equations makes it easy to develop and maintain scalable electronic structure methods (such as CCSD) that can be run on a variety of HPC platforms.

Another important technique that is useful in reducing the memory footprint of the CCSD approach is the Cholesky decomposition (CD) of the v_{rs}^{pq} tensor, i.e.,

$$v_{rs}^{pq} \simeq (pr|L)(L|qs) - (ps|L)(L|qr), \tag{25}$$

where L is an auxiliary index. The total memory required to store Cholesky vectors $(pq|L)$ needed to reproduce v_{rs}^{pq} with high accuracy is usually proportional to $(n_o + n_u)^3$.

We developed both open-shell and closed-shell implementations of the CCSD equations using TAMM. The use of TAMM ***index spaces*** for alpha and beta-occupied and virtual orbitals allowed us to represent the same theory optimized for specific problems easily. An example equation that is a contribution to the intermediate I in Eq. (24) may appear like the following:

$$\begin{aligned} &(I(m_a,\ n_a,\ i_a,\ j_a) \\ &\quad = v(m_a,\ n_a,\ e_b,\ f_b) * t(e_b,\ f_b,\ i_a,\ j_a)), \end{aligned} \tag{26}$$

where the added $_a$ and $_b$ notations respectfully refer to the corresponding indices for the alpha and beta sub-spaces. While the equations for the closed- vs open-shell implementations are different, the ease of writing them is the same due to the slicing capabilities on sub-spaces provided by TAMM.

As benchmark systems for testing the performance of TAMM implementations of the Cholesky-based CCSD,[52] we used two molecular systems previously used in studying the efficiency of the TCE implementations of the CC methods in NWChem.[53,54] The first benchmark system is the model of Bacteriochlorophyll (BChl) $MgC_{36}N_4O_6H_{38}$,[55,56] which plays an important role in understanding the mechanism of photosynthesis. In this process, the light is harvested by antenna systems and further funneled to BChl molecules, which initiate primary charge separation. The second benchmark system considered here is the β-carotene molecule, whose doubly excited states and triplet electronic states have recently been intensively studied in the context of singlet fission processes in carotenoid aggregates.[57–59] The β-carotene molecule consists of 96 atoms, while the BChl model contains 85 atoms, including a central magnesium atom. We use the cc-pVDZ basis for both systems and evaluate their performance on OLCF Summit.[21] Each Summit node contains two IBM POWER9 processors, each consisting of 22 cores and 512 GB of CPU memory. Each node is also equipped with six Nvidia Volta GPUs, each with 16 GB of memory, for a total of 96 GB of GPU memory per node.

Table II shows the CCSD performance compared to NWChem on 200 nodes of OLCF Summit. We measure the performance of the TAMM implementations against the TCE implementations in NWChem. The time per CCSD iteration is given in seconds. NWChem has a CPU-only implementation and uses 42 CPU cores on each node of Summit. TAMM-based Cholesky-CCSD uses only six MPI ranks per node, where each MPI rank is mapped to a single GPU. For these two molecular systems, we observe a 9–15× speedup with the TAMM-based Cholesky-CCSD implementation compared to the TCE CCSD method in NWChem. The CPU implementation of the CCSD tensor operations in NWChem comprises 11 314 source lines of code, whereas the Cholesky-CCSD implementation

17 August 2023 00:38:49

© Author(s) 2023

TABLE II. TAMM performance compared to NWChem on 200 nodes of OLCF Summit. Time per CCSD iteration is given.

Molecule	No. of atoms	No. of occupied Orbitals	No. of basis Functions	Time (s) NWChem[60]	Time (s) TAMM
BChl	85	171	852	1202	81
β-carotene	96	148	840	801	65

expressed using the TAMM framework is only 236 source lines of code. Since these 236 lines represent computation at a high level, they express both CPU and GPU computation. On the other hand, adding GPU capabilities to the NWChem CCSD code will only significantly increase the SLOC count and development time, which is why a GPU implementation for CCSD has not been attempted to date in NWChem. This clearly demonstrates the productivity benefits of expressing such computations in TAMM. Our implementation of the CCSD equations[52] is expressed similarly to the example in Fig. 5. CCSD is an example of how TAMM can be used to productively create an effective and efficient implementation of certain classes of computational chemistry methods. The remaining computational chemistry methods discussed in this paper are also implemented along similar lines.

2. Coupled cluster triples

The CCSD(T) formalism[9,61] is capable in many cases of providing the so-called chemical level of accuracy required in studies of chemical reactivity and thermochemistry. In the CCSD(T) approach, the perturbative correction due to triple excitations $[E^{(T)}]$ is added to the CCSD energy,

$$E^{\mathrm{CCSD(T)}} = E^{\mathrm{CCSD}} + E^{(T)}, \tag{27}$$

where

$$E^{(T)} = \sum_{\substack{i<j<k \\ a<b<c}} \frac{\langle\Phi|(T_2^\dagger V_N)|\Phi_{ijk}^{abc}\rangle\langle\Phi_{ijk}^{abc}|V_N T_2|\Phi\rangle}{\varepsilon_i+\varepsilon_j+\varepsilon_k-\varepsilon_a-\varepsilon_b-\varepsilon_c} + \sum_{\substack{i<j<k \\ a<b<c}} \frac{\langle\Phi|T_1^\dagger V_N|\Phi_{ijk}^{abc}\rangle\langle\Phi_{ijk}^{abc}|V_N T_2|\Phi\rangle}{\varepsilon_i+\varepsilon_j+\varepsilon_k-\varepsilon_a-\varepsilon_b-\varepsilon_c}, \tag{28}$$

where V_N is the two-body part of the electronic Hamiltonian in a normal product form, and $|\Phi_{ijk}^{abc}\rangle = a_a^\dagger a_b^\dagger a_c^\dagger a_k a_j a_i|\Phi\rangle$. The most expensive part of the CCSD(T) calculation, characterized by the $n_o^4 n_u^3 + n_o^3 n_u^4$ scaling, is associated with calculating the $\langle\Phi_{ijk}^{abc}|V_N T_2|\Phi\rangle$ term, which is defined as

$$\begin{aligned}\langle\Phi_{ijk}^{abc}|V_N T_2|\Phi\rangle = {} & v_{ma}^{ij}t_{bc}^{mk} - v_{mb}^{ij}t_{ac}^{mk} + v_{mc}^{ij}t_{ab}^{mk} - v_{ma}^{ik}t_{bc}^{mj} + v_{mb}^{ik}t_{ac}^{mj} \\ & + v_{mc}^{ik}t_{ab}^{mj} + v_{ma}^{jk}t_{bc}^{mi} - v_{mb}^{jk}t_{ac}^{mi} + v_{mc}^{jk}t_{ab}^{mi} - v_{ab}^{ei}t_{ec}^{jk} \\ & + v_{ac}^{ei}t_{eb}^{jk} - v_{bc}^{ei}t_{ea}^{jk} + v_{ab}^{ej}t_{ec}^{ik} - v_{ac}^{ej}t_{eb}^{ik} + v_{bc}^{ej}t_{ea}^{ik} \\ & - v_{ab}^{ek}t_{ec}^{ij} + v_{ac}^{ek}t_{eb}^{ij} - v_{bc}^{ek}t_{ea}^{ij}, \quad (i<j<k, a<b<c).\end{aligned} \tag{29}$$

Equation (29) can be separated into terms A_{ijk}^{abc} and B_{ijk}^{abc}, defined by contractions over occupied indices [A_{ijk}^{abc}; the first nine terms on the right hand side (rhs) of Eq. (29)] and terms corresponding to contractions over unoccupied indices [B_{ijk}^{abc}; remaining nine terms on the rhs of Eq. (29)], i.e.,

$$\langle\Phi_{ijk}^{abc}|V_N T_2|\Phi\rangle = A_{ijk}^{abc} + B_{ijk}^{abc}. \tag{30}$$

Analogously, the $\langle\Phi_{ijk}^{abc}|V_N T_1|\Phi\rangle$ term takes the form

$$\begin{aligned}\langle\Phi_{ijk}^{abc}|V_N T_1|\Phi\rangle = {} & +v_{ab}^{ij}t_c^k - v_{ac}^{ij}t_b^k + v_{bc}^{ij}t_a^k - v_{ab}^{ik}t_c^j + v_{ac}^{ik}t_b^j - v_{bc}^{ik}t_a^j \\ & + v_{ab}^{jk}t_c^i - v_{ac}^{jk}t_b^i + v_{bc}^{jk}t_a^i (i<j<k, a<b<c).\end{aligned} \tag{31}$$

While Eqs. (29) and (31) can be easily implemented with the tensor operations mentioned in Sec. IV A 1, the performance of the (T) correction, in terms of both speed and memory, is improved by encoding key symmetries by hand while still using the distributed tensor data structure provided by TAMM. This includes restricted index summation and permutational symmetries, which are not routine features in the current version of TAMM. In addition, the contractions in (T) are all fused so that the resulting output is a scalar instead of a six-dimensional intermediate tensor. Currently, TAMM does not support fusion across contractions writing to the same output tensor, but this capability is considered for the near future. Furthermore, the hand-coded kernels for (T) for various GPU architectures are required for the best performance. In addition to fusion, we are also considering incorporating optimized GPU kernels in TAMM for specific contractions such as the ones in (T). The scheduler can then detect such contractions and choose the corresponding optimized kernels for execution rather than executing the contractions through the current default pipeline.

Table III shows the TAMM-based NWChemEx triples correction (T) calculation[52] performance compared to NWChem on 512 nodes of the OLCF Summit. The time is given in seconds. NWChem has a GPU implementation of the triples correction and uses all 6 GPUs and 42 CPU cores on each node. The TAMM-based triples correction uses only six MPI ranks per node, where each MPI rank is mapped to a single GPU. For the BChl and β-carotene molecules, we observe a speedup of ~6–18× for the TAMM implementation compared to the TCE implementation in NWChem. The finer details of the TAMM-based triples correction implementation are detailed in Ref. 62.

TABLE III. NWChemEx-TAMM performance compared to NWChem for the perturbative correction in the CCSD(T) method on 512 nodes of the OLCF Summit.

Molecule	No. of atoms	No. of occupied Orbitals	No. of basis Functions	Time (s) NWChem[60]	Time (s) TAMM
BChl	85	171	852	18 285	1791
β-carotene	96	148	840	6 646	1164

17 August 2023 00:38:49

© Author(s) 2023

B. New methods

1. Equation-of-motion coupled cluster formalism

The equation-of-motion (EOM) methods[63–67] and closely related linear response (LR) CC formulations[68–70] can be viewed as excited-state extensions of the single-reference CC theory. In the exact EOMCC formalism, the wave function for the K-th excited state, $|\Psi_K\rangle$, is represented as

$$|\Psi_K\rangle = R_K e^T |\Phi\rangle, \tag{32}$$

where R_K is the excitation operator, which produces the K-th excited state when acting on the correlated ground-state in CC representations. The energy of the K-th state and the amplitudes defining the R_K operator can be calculated by solving the non-Hermitian eigenvalue problem

$$\bar{H} R_K |\Phi\rangle = E_K R_K |\Phi\rangle, \tag{33}$$

where the similarity transformed Hamiltonian $\bar{H}$ is defined as

$$\bar{H} = e^{-T} H e^T. \tag{34}$$

In the rudimentary EOMCCSD approximation (EOMCC with singles and doubles), the R_K and T operators are approximated as

$$R_K \simeq R_{K,0} + R_{K,1} + R_{K,2}, \tag{35}$$

$$T \simeq T_1 + T_2. \tag{36}$$

Similar to the CCSD method, the numerical scaling of the EOM-CCSD approach is dominated by $n_o^2 n_u^4$ terms, where n_o represents the number of occupied orbitals and n_u represents the number of unoccupied orbitals. The major computational complexity in the EOMCC method arises from tensor contractions involving various operators. The method involves two types of excitation operators: the CC operator T and the EOM operator R, besides the one-body and two-body components of the Hamiltonian operator. The EOMCCSD approach is a non-Hermitian eigenvalue problem, and because of the dimensionality of the problems, they are solved iteratively using a multi-root solver. Therefore, there is an additional cost for constructing and working with the iterative subspace. In terms of implementation, however, the same TAMM format used for writing various tensor contraction equations, as demonstrated while discussing Eq. (23), was followed for the EOMCC method.

The EOMCCSD method is usually employed in studies of excited states dominated by single excitations. It is also worth mentioning that LR-CC methods with singles and doubles lead to the same values of excited-state energies. However, in contrast to the EOMCCSD formalism, the LR-CCSD excitations are identified with the poles of frequency-dependent linear-response CC amplitudes.

Table IV shows the TAMM EOMCCSD calculation performance compared to NWChem on 200 nodes of the OLCF Summit. The time per EOMCCSD iteration is given in seconds. NWChem has a CPU-only implementation and uses 42 CPU cores on each of the 200 nodes of Summit. Just as with the previous calculations, the TAMM-based EOMCCSD uses six MPI ranks per node, where each MPI rank is mapped to a single GPU. The current TAMM implementation of the EOMCCSD approach has not been optimized at the same equation level as the CD-CCSD implementation. For example, in contrast to the CD-CCSD formalism, the EOMCCSD implementation fully represents two-electron integrals and uses a non-spin-explicit representation of the operators in the equations. While we plan to return to this implementation and incorporate the optimized (spin-explicit) equations, we already observe a significant speed-up over NWChem with this primitive implementation. CD-CCSD and EOMCCSD calculations are different algorithms for solving the corresponding problem, and the timing for an iteration of EOMCCSD is a sum of all of the steps in an interaction, which tends to be more computationally demanding than the Jacobi step in CD-CCSD. Nonetheless, one can see from Table IV that the TAMM EOMCCSD code is 2–3 times faster for β-carotene and BChl molecules.

TABLE IV. TAMM performance compared to NWChem on 200 nodes of the OLCF Summit. Time per EOMCCSD iteration is given.

Molecule	No. of atoms	No. of occupied Orbitals	No. of basis Functions	Time (s) NWChem[60]	Time (s) TAMM
BChl	85	171	852	2030	715
β-carotene	96	148	840	1170	540

2. DLPNO CCSD(T)

The development of reduced-scaling quantum chemical methods became a significant part of the recent research effort. In particular, for coupled cluster approaches, local domain-based methods have been introduced and successfully applied to large systems. Using even small computational clusters, the (DLPNO-CC)[14–18,71] formalism can be used for systems described by 10 000–40 000 basis set functions. The limit for system size can be further extended by utilizing parallel exascale architectures. However, to achieve this goal, several computational challenges associated with the data representation, distribution, and optimization of the DLPNO-CC equations (characterized by a large number of contractions and tensors involved in the underlying equations) have to be appropriately addressed. Our implementation is inspired by the work cited in Ref. 16. The basic idea is to carefully take advantage of the local character of correlational effects. The total CCSD energy can be seen as a sum of ij-pair specific contributions ε_{ij},

$$E^{\mathrm{CCSD}} = \sum_{ij} \varepsilon_{ij}. \tag{37}$$

In order to significantly reduce scaling, we are considering the following tasks for the NWChemEx DLPNO CCSD(T) implementation using TAMM:

1. Differentiate pairs with respect to their energy contributions ε_{ij} by sequential pre-screenings. Identify ij-pairs that (a) are negligible and immediately dropped, (b) can be evaluated at a lower-level model (MP2 level), and (c) can be evaluated at CC level.

17 August 2023 00:38:49

© Author(s) 2023

2. Find an optimal representation of the virtual space for each ij-pair in which the derived tensors (amplitudes, residuals, or integrals) are dense.
3. Find the optimal factorization of terms in amplitude equations.
4. Perform only those tensor contractions that lead to non-zero results (including integral transformation and amplitude equation evaluation).

In the local pair natural orbital CCSD method, the occupied orbitals are localized (for example, by the Pipek–Mezey or Foster–Boys algorithms[72,73]), and the virtual space is constructed specifically for a given occupied orbital pair. First, the virtual orbitals are transformed to a local basis of projected atomic orbitals (PAOs) $|\tilde{\mu}\rangle$ as $|\tilde{\mu}\rangle = (1 - \sum_i |i\rangle\langle i|)|\mu\rangle$,[10,72,73] which are *a priori* local and can be easily used to form domains corresponding to local occupied orbitals. In the next step, pair natural orbitals (PNOs) are constructed.[15,16,77–80] Therefore, in Eqs. (19) and (20), instead of virtual index $a, b, \ldots$ we will get pair-specific PNOs $\tilde{a}_{ij}$, $\tilde{b}_{ij}$, $\ldots$ where we assume that the size of PNO space $N(\mathrm{PNO}(ij)) \ll N(\mathrm{MO}_{\mathrm{virt}})$. The PNO spaces are obtained from pair density matrices $\mathbf{D}^{ij}$,

$$\mathbf{D}^{ij} = \tilde{\mathbf{T}}^{ij}\mathbf{T}^{ij\dagger} + \tilde{\mathbf{T}}^{ij\dagger}\mathbf{T}^{ij}, \tag{38}$$

where $\mathbf{T}^{ij}$ are MP2 amplitudes in PAO basis

$$T^{ij}_{\tilde{\mu}\tilde{\nu}} = \frac{(i\tilde{\mu}|j\tilde{\nu})}{\varepsilon_{\tilde{\mu}} + \varepsilon_{\tilde{\nu}} - f_{ii} - f_{jj}}, \tag{39}$$

where f_{ii} and f_{jj} are occupied orbital energies, and $\varepsilon_{\tilde{\mu}}$ or $\varepsilon_{\tilde{\nu}}$ are PAO orbital energies obtained by diagonalization of the Fock matrix transformed to PAO space. The transformation matrices $\mathbf{d}^{ij}$ transforming orbitals from PAO space to PNO space, and occupation numbers n^{ij} correspond to eigenvectors and eigenvalues of the diagonalized matrix $\mathbf{D}^{ij}$,

$$\mathbf{D}^{ij}\mathbf{d}^{ij} = n^{ij}\mathbf{d}^{ij}. \tag{40}$$

The introduction of ij-specific PNO spaces, which are mutually non-orthogonal, leads to more complicated expressions in the amplitude equations because we also need to involve overlap matrices between different PNO spaces. The matrix transforming ij-PNO space to kl-PNO space $S^{ij;kl}_{\tilde{a}_{ij}\tilde{b}_{kl}}$ is defined as

$$\mathbf{S}^{ij;kl} = \mathbf{d}^{ij\dagger}\mathbf{S}^{\mathrm{PAO}}\mathbf{d}^{kl}, \tag{41}$$

where $\mathbf{S}^{\mathrm{PAO}}$ is the PAO overlap matrix. Therefore, when considering the PNO space, the term in Eq. (21) can take the form

$$\frac{1}{4} v^{\tilde{e}_{mn}\tilde{f}_{mn}}_{mn} t^{ij}_{\tilde{e}_{ij}\tilde{f}_{ij}} t^{mn}_{\tilde{a}_{mn}\tilde{b}_{mn}} S^{mn;ij}_{\tilde{a}_{mn}\tilde{a}_{ij}} S^{mn;ij}_{\tilde{b}_{mn}\tilde{b}_{ij}} S^{mn;ij}_{\tilde{e}_{mn}\tilde{e}_{ij}} S^{mn;ij}_{\tilde{f}_{mn}\tilde{f}_{ij}}. \tag{42}$$

However, it is also possible to transform the e, f indices from ij- to mn-PNO space first and then perform the contraction of the integral with the first amplitude. The complex structure of these terms significantly expands the possibilities of how they can be factorized. The cost is also affected by the integral evaluation, which depends on the available memory. We can pre-compute not only $v^{\tilde{e}_{ij}\tilde{f}_{ij}}_{ij}$ type integrals but also some mixed PNO space integrals (only those that will be needed).

In our work, we assume that we will utilize a larger number of nodes, so we will have more memory and computing effort available. In that case, we can afford tighter thresholds, leading to larger domains and PNO spaces, which means higher precision in the recovery of the correlation energy. For the implementation of DLPNO formulations in TAMM, we employed a code generator implemented in Python that transforms the canonical equations by the transformation rules for various spaces (i.e., PNO, PAO, etc.). Using a code generator allowed us to systematically convert equations while automatically applying an operation cost minimization algorithm for single-term optimization. We anticipate that these kinds of code generators on the high-level equations will enable trying different transformations and automatically choosing the best performing one. While this implementation is in the early stages, we were able to validate the correctness of the generated code using our infrastructure to directly compare the results with their canonical counterparts.

The perturbative correction (T) described in Sec. IV A 2 is in the local version evaluated using the same equations with some differences.[81] While converged T_1 and T_2 amplitudes are obtained in their PNO space, the virtual indices in terms A^{abc}_{ijk} or B^{abc}_{ijk} are represented in triple natural orbitals (TNOs), which are computed from triplet density matrices obtained from three pair density matrices, $\mathbf{D}^{ijk} = 1/3(\mathbf{D}^{ij} + \mathbf{D}^{ik} + \mathbf{D}^{jk})$. In order to perform contractions of integrals and amplitudes in Eqs. (29) and (31), we need to involve transformation matrices between PNO and TNO spaces, $\mathbf{S}^{ij;klm}$, which are computed the same way as in Eq. (41) where $\mathbf{d}^{kl}$ is replaced by $\mathbf{d}^{klm}$, transforming PAO orbitals to TNO space. Similar to the DLPNO CCSD implementation, we leveraged the TAMM framework's dense tensor infrastructure to represent the perturbation correction implementation with block-sparse computation. Using the PNO representation of the amplitudes from the CCSD implementation, we implemented the DLPNO formulation of the canonical equations.

For the development of DLPNO-based methods in NWChemEx, we utilized specialized view tensor construction to avoid redundant storage of the transformation tensors, especially for the transformation required for the occupied pair indices. Figure 6 shows an example use case specific to DLPNO CCSD equations. In this example, we are using a view tensor over the large `Sijkl` transformation tensors. Depending on the pair indices used within the equations, various kinds of transformations are required. In this example, the constructed `Sijki` tensor (line 17) has values from `Sijkl` if the first index of the first pair matches the second index of the second pair. If not, the corresponding indices are set to zero. In this case, there was no need to translate the blocks or update the tensor since the tensor was used as read-only. However, users can provide special constraints on both the translation of the blocks and access to these blocks. By using this feature, we were able to reduce the memory footprint of DLPNO methods and add additional constraints to leverage sparsity within the computation. Since the TAMM allows mixed usage of specialized tensors with other tensor types in all supported operations, the representation and execution patterns of the DLPNO CCSD equations in the TAMM did not have to change in such scenarios.

17 August 2023 00:38:49

© Author(s) 2023

```
// Translate lambda for translating BlockId to reference tensor
// No translation is needed for this case
auto no_translate = [](IndexVector blockid) -> IndexVector {
  return blockid;
};

// Lambda for specifying how to put values to the reference tensor
// For read-only usage leave it empty
auto put_copy_func = [](const BlockSpan<T>& in, BlockSpan<T>&
     out, const IndexVector& blockid) -> void {};

// Lambda for specifying how to get values from the reference tensor
auto get_S_is_si = [=](const BlockSpan<T>& from_span,
     BlockSpan<T>& to_span, const IndexVector& blockid) -> void{
    // Only copy the values where first index of first pair
    // and last index of second pair matches
};

Tensor<T> Sijki{Sijkl, {is, si, a_is, a_si}, no_translate,
     get_S_is_si, put_copy_func};
```

FIG. 6. View the tensor construction for Sijkl transformation tensors in DLPNO CCSD equations.

3. Time-dependent coupled-cluster method

In addition to the stationary and frequency-dependent formulations of CC theory, one could witness significant progress on our end in developing a time-dependent CC method for simulating real-time dynamics.

The TDCC method has been studied in the context of the CC linear-response theory,[68,69,82,83] x-ray spectroscopy and Green's function theory,[85–89] nuclear physics,[90–92] condensed matter physics,[93] and quantum dynamics of molecular systems in external fields.[94–99] These studies have also initiated an intensive effort toward understanding many aspects of the TDCC formalism, including addressing fundamental problems such as the form of the action functional, the form of the time-dependent molecular orbital basis, the physical interpretation of time-dependent orbitals, various-rank approximations of the cluster operator, and the numerical stability of time integration algorithms. One of the milestone achievements in developing a time-dependent CC formalism was Arponen's action functional for the bi-variational coupled cluster formalism[93] and the following analysis by Kvaal considering time-varying orbitals.[95]

The time-dependent Schödinger equation (TDSE),

$$i\partial_\tau|\Psi(\tau)\rangle = H|\Psi(\tau)\rangle, \tag{43}$$

is an initial value problem, which has a formal solution $|\Psi(\tau)\rangle = e^{-iH\tau}|\Psi\rangle$, with the initial condition, $|\Psi(0)\rangle = |\Psi\rangle$. An efficient way of finding approximate solutions of the TDSE for many-electron systems is associated with the utilization of the time-dependent CC (TD-CC) Ansatz,

$$|\Psi(\tau)\rangle = e^{T(\tau)}|\Phi\rangle. \tag{44}$$

Upon plugging TD-CC Ansatz into Eq. (43) and projecting onto the excited determinants, one obtains

$$i\partial_\tau\langle\Phi^{a_1\cdots a_k}_{i_1\cdots i_k}|T(\tau)|\Phi\rangle = \langle\Phi^{a_1\cdots a_k}_{i_1\cdots i_k}|\overline{H}(\tau)|\Phi\rangle, \tag{45}$$

where $T(\tau)$ is a time-dependent cluster operator, $|\Phi\rangle$ is a time-independent reference wave function in our consideration, and $\overline{H}(\tau) = e^{-T(\tau)}He^{T(\tau)}$ is the similarity transformed time-dependent Hamiltonian, which is the generator for the time evolution.

The similarity transformed CC Hamiltonian is non-Hermitian. Therefore, the computation of observables in this framework requires a bi-variational approach,[65,93] i.e., both the bra and ket states must be varied independently. The expectation value of any observable in the TDCC framework is defined as

$$\langle O\rangle = \langle\Phi|(1+\Lambda(\tau))\overline{O}(\tau)|\Phi\rangle, \tag{46}$$

where Λ is a de-excitation operator. For obtaining equations for the Λ, we write down the TDSE, $-i\partial_\tau\langle\Psi'(\tau)| = \langle\Psi'(\tau)|H$, for the bra vector, with $H = H^\dagger$,

$$-i\partial_\tau\langle\Phi|(1+\Lambda(\tau))e^{-T(\tau)} = \langle\Phi|(1+\Lambda(\tau))e^{-T(\tau)}H. \tag{47}$$

Multiplying by $e^{T(\tau)}$ from the right and right projecting to the excited determinants obtains

$$-i\partial_\tau\langle\Phi|\Lambda(\tau)|\Phi^{i_1\cdots i_k}_{a_1\cdots a_k}\rangle = \langle\Phi|(1+\Lambda(\tau))\overline{H}(\tau)|\Phi^{i_1\cdots i_k}_{a_1\cdots a_k}\rangle. \tag{48}$$

The right-hand side of Eq. (45) is the same as the stationary CC amplitude equations and hence implemented using TAMM, similar to how it is implemented for the stationary theory, as shown in Sec. IV A 1. However, the time-dependent CC amplitudes are naturally complex since the underlying equation we are solving is the TDSE. Furthermore, one-body and two-body Hamiltonian matrix elements are real-valued quantities. TAMM support for mixed complex-real tensor operations helps avoid the explicit handling of the real and imaginary data separately. For example, a tensor contraction for the TDCC method,

$$(R(a,b,i,j) += 0.5 * v(a,b,c,d) * t(c,d,i,j)), \tag{49}$$

can be expressed exactly the same way as in the stationary CC theory, as shown in Eq. (24). However, in this case, R and t are complex-valued, while v is a real-valued quantity.

We have propagated the time-dependent CC amplitudes using a first-order Adams–Moulton method, which is an implicit time-propagator. This algorithm requires copying and swapping of the real and complex parts of time-dependent amplitudes in each micro-iteration. TAMM does not directly support swapping the real and imaginary parts of each element of a complex tensor. However, in this case, we used the *block_for* construct of TAMM to perform the swap in parallel. A specialized C++ lambda expression, as shown in Fig. 7, is provided to the *block_for*, which simply specifies that the real and imaginary parts of each element in a given block need to be swapped and copied to (or updated) the corresponding element in the destination tensor block.

The TDCC method performs time propagation for a sufficiently long duration to obtain well-resolved spectra. However, obtaining computing time for such an extended period in a single run is unlikely when using shared computing resources. In addition, it is highly beneficial to track simulations to adjust and optimize various parameters in between different runs of the same simulation. The ability to checkpoint and restart simulations, hence, plays a significant role in the efficient utilization of computing resources.

17 August 2023 00:38:49

© Author(s) 2023

```
// src, dest tensors are given
auto complex_swap_lambda = [&](const IndexVector& blockid) {

  // assert TensorType is complex
  const auto dsize = dtensor.block_size(blockid);
  span<TensorType> sbuf{stensor.access_local_buf(), dsize};
  span<TensorType> dbuf{dtensor.access_local_buf(), dsize};

  for(auto i = 0; i < dsize; i++) {
    TensorType val(sbuf[i].imag(), sbuf[i].real());
    if(update) dbuf[i] += val;
    else dbuf[i] = val;
  }
};

block_for(execution_context, dtensor(), complex_swap_lambda);
```

FIG. 7. Parallel element-wise swap and copy operation between complex tensors.

```
tamm::write_to_disk (tensor, tensor_filename);
tamm::read_from_disk(tensor, tensor_filename);
```

FIG. 8. Parallel file I/O operations in TAMM.

The TAMM library provides parallel file I/O operations on tensors, which helps with checkpointing and restarting the TDCC calculations. We simply use the TAMM interface to write and read tensor data to and from the disk, as shown in Fig. 8. The underlying TAMM infrastructure handles the parallelization aspects of these I/O operations.

The TAMM infrastructure has been utilized to implement the time evolution of the $T(\tau)$ operator. We have considered singles (S) and doubles (D) excitation approximations in our implementation. Our implementation uses Cholesky-decomposed two-body electron repulsion matrix elements,[100,101] and we exploit spin-explicit formalism to evaluate tensor contractions of matrix elements of various operators.

4. Coupled cluster Green's function

The correlated formulations of Green's function methods are indispensable elements of the computational infrastructure needed not only to calculate ionization potentials, electron affinities, and excitation energies but also as quantum solvers for various embedding formulations. The CC formalism provides a natural platform for the development of the one-body Green's function and the introduction of high-rank correlation effects.[102–105] Without loss of generality, here we will focus only on the retarded part of Green's function (the advanced part can be developed in an analogous way), defined by matrix elements $G^R_{pq}(\omega)$,

$$G^R_{pq}(\omega) = \langle \Psi_g | a_q^\dagger (\omega + (H - E_0) - i\eta)^{-1} a_p | \Psi_g \rangle, \tag{50}$$

where E_0 is the corresponding ground-state energy for the N-electron system, η is a broadening factor, and $|\Psi_g\rangle$ represents the ground-state of the N-electron system. In CC, we are using different parametrizations for the bra ($\langle\Psi_g|$) and ket ($|\Psi_g\rangle$) wave functions,[65,93] i.e.,

$$\langle \Psi_g | = \langle \Phi | (1 + \Lambda) e^{-T}, \tag{51}$$

$$|\Psi_g\rangle = e^{T} |\Phi\rangle, \tag{52}$$

which leads to the following form of retarded part of the CC Green's function (CCGF):[102–107]

$$G^R_{pq}(\omega) = \langle \Phi | (1 + \Lambda) \overline{a_q^\dagger} (\omega + \bar{H}_N - i\eta)^{-1} \overline{a_p} | \Phi \rangle. \tag{53}$$

The similarity transformed operators here, $\overline{A}$ ($A = H, a_p, a_q^\dagger$), are defined as $\overline{A} = e^{-T} A\, e^{T}$ (the $\overline{H}_N$ stands for a normal product form of $\overline{H}$). The numerically feasible algorithms for calculating (53) employ ω-dependent auxiliary operators $X_p(\omega)$,

$$\begin{aligned} X_p(\omega) &= X_{p,1}(\omega) + X_{p,2}(\omega) + \cdots \\ &= \sum_i x^i(\omega)_p a_i + \sum_{i<j,a} x^{ij}_a(\omega)_p a_a^\dagger a_j a_i + \cdots, \quad \forall_p, \end{aligned} \tag{54}$$

that satisfy equations

$$(\omega + \overline{H}_N - i\eta) X_p(\omega) |\Phi\rangle = \overline{a_p} |\Phi\rangle. \tag{55}$$

Using these operators, matrix elements can be expressed in a simple form

$$G^R_{pq}(\omega) = \langle \Phi_g | (1 + \Lambda) \overline{a_q^\dagger} X_p(\omega) | \Phi_g \rangle. \tag{56}$$

In our implementation of CCGF formalism, we approximate Λ operator by $T^\dagger$. The main numerical effort associated with constructing a retarded CC Green's function is associated with the need to solve a large number of independent linear equations, which in turn can contribute to efficient parallel schemes utilizing multiple levels of parallelism. Our CCGF algorithm consists of an outer loop that checks for the convergence of the entire calculation. We refer to iterations of this loop as levels. Each level mainly consists of two loops, which are the most computationally intensive part of the entire calculation. The first loop goes over the frequencies ($\omega's$) sampled using the adaptive midpoint refinement strategy.[108] The second loop goes over all the orbitals ($p's$), and the CCGF singles and doubles equations are solved for all (ω, p) pairs for a given level in this loop. Since the bulk of the computation here is associated with high-dimensional tensor contractions, we express the tensor operations similar to the cases in Fig. 5 and Eq. (24), even though the tensor operations in CCGF involve a mix of complex and real data types. This demonstrates the uniform operation representation provided by TAMM for operations on mixed data types.

We refer to the list of all (ω, p) pairs in a given level as CCGF tasks, each of which can be executed independently. Each task solves the CCGF singles and doubles equations until convergence for a given (ω, p) pair. All the p orbitals for a given ω are divided across process groups that are set up at the beginning of each level. Currently, the user is responsible for explicitly creating MPI process groups and using them to construct the TAMM

17 August 2023 00:38:49

© Author(s) 2023

process group and ***execution context*** objects, which are, in turn, used for executing each task. Each TAMM ***process group*** now contains a subset of the resources (processes) provided to the CCGF calculation. In our CCGF implementation, the size of a process group for computing each task is determined automatically for a given problem size (i.e., number of occupied and virtual orbitals) and the resources (total number of nodes, processes, and GPUs per node) provided for that run. All process groups are created to be the same size. We are actively working on providing abstractions in TAMM for expressing computation to be divided across process groups without the user having to create them manually using MPI. Once the ***execution context*** objects representing different process groups are created, all CCGF tasks are distributed across the available process groups and executed using different execution contexts.

The resulting tensors from each task are stored on disk using the parallel file I/O routines provided by TAMM for writing and reading distributed tensors to and from disk, as shown in Fig. 8. This enables restarting the CCGF calculation at any point in the calculation. If the set of tasks in a given level is not completed, the subsequent CCGF run upon restart will create process groups only for the remaining tasks and execute them. Another feature provided by TAMM used in CCGF is batched parallel file I/O operations. Several tensors for each task are written to disk as part of the CCGF calculation. The number of tensors written to disk grows with increasing CCGF problem size and the number of frequencies (ω) to be solved. Even though each tensor is read/written using parallel file I/O, the operation uses a small subset of available nodes/processes to perform the read/write operation for a single tensor. This is because the entire set of available resources used for the CCGF calculation is not required to read/write even large tensors from/to disk. This leads to resource under-utilization since most of the resources are idle while a subset of them are performing the parallel file I/O operation on one distributed tensor at a time. To address this issue, TAMM provides high-level batched tensor I/O routines that can read several hundred tensors from disk concurrently by automatically dividing the total available resources into smaller process groups. Since all tensors that need to be read/written are not necessarily of the same size, variable-sized process groups are dynamically created based on the sizes of each tensor in a given list of tensors. Each of these process groups reads/writes a tensor using parallel file I/O within the group. Several tensors are read/written concurrently by different process groups, leading to significantly better resource utilization and an overall improvement in the total time spent in file IO. The user would express such an operation in TAMM, as shown in Fig. 9. The operation is expressed exactly the same way as one would express it for reading or writing a single tensor (e.g., Fig. 8), with the only obvious difference being that a list of tensor handles and corresponding filenames needs to be provided in this case. Reference 109 describes the highly scalable CCGF implementation developed using TAMM in detail, along with the performance and scalability analysis at the OLCF Summit. The implementation is also publicly available.[110]

This section provides an overview of some important CC methods and discusses key aspects of their implementations using TAMM. While frameworks such as TiledArray, CTF, DISTAL, and ExaTensor provide similar features in how tensor operations in CC methods can be represented, TAMM differs from them, as explained in Sec. III C. To the best of our knowledge, these are the first complete scalable distributed implementations of the discussed CC methods that can be run on different GPU architectures while also using different parallel programming models.

```
tamm::write_to_disk ( tensor_list , tensor_filenames_list );
tamm::read_from_disk( tensor_list , tensor_filenames_list );
```

FIG. 9. Batched parallel file I/O operations in TAMM.

V. PERFORMANCE COMPARISON WITH OTHER TENSOR ALGEBRA FRAMEWORKS

We have provided feature comparisons with other distributed tensor algebra frameworks in Sec. III C, where key library features such as tensor-construction and tensor-operation specification, hardware support, and the underlying parallel programming model backends (for distributed data management) are compared. This section details the performance comparison results with other distributed tensor algebra frameworks, namely, TiledArray (TA) and Cyclops Tensor framework (CTF). We did not consider comparisons with ExaTensor since the development page[111] states that the library has pending performance issues as well as numerous problems with existing MPI-3 implementations at the time of this writing.

All our experiments were performed on the National Energy Research Scientific Computing Center (NERSC) Perlmutter supercomputer. The GPU partition was used for all the experiments. Each node in this partition has a 64-core AMD EPYC 7763 CPU, 256 GB of DDR4 DRAM, and 4 Nvidia A100 GPUs, each with 40 GB of HBM2e RAM. Each GPU node is connected to four NICs, and the GPU nodes are connected using the Slingshot 11 interconnect. The TAMM experiments were configured to run with four MPI ranks per node and one MPI rank per GPU. The TiledArray experiments were configured with four MPI ranks per node (one MPI rank per GPU) and two threads per rank since that is the TiledArray developer's recommended configuration for Perlmutter. The CTF implementation provided the best performance when using either 8 or 12 MPI ranks per node (2 or 3 MPI ranks per GPU), depending on the problem size and the number of nodes used in the experiment. For a given dimension size (N) and node count, we ran CTF using both 8 and 12 MPI ranks per node and chose the best timing out of the two. All codes were compiled with GCC 11.2, CUDA 11.7, and cray-mpich 8.1.

We compared the performance of a tensor contraction shown in Eq. (57), which is one of the most expensive operations in many CC calculations. For benchmarking purposes, each dimension of the 4D tensors in Eq. (57) is chosen to be of the same size (N), resulting in a total size of (N^4) double-precision floating-point elements for each tensor. The input tensors are filled with random data in each case. We used the readily available TA and CTF codes[29,36] for benchmarking Eq. (57) and ensured that the ordering of the indices used in the contraction is consistent across the different benchmark codes. The TA benchmark allows specifying the number of blocks for each dimension. For a given dimension size (N) and node count, we ran the TA benchmark several times using a varying number of blocks

17 August 2023 00:38:49

© Author(s) 2023

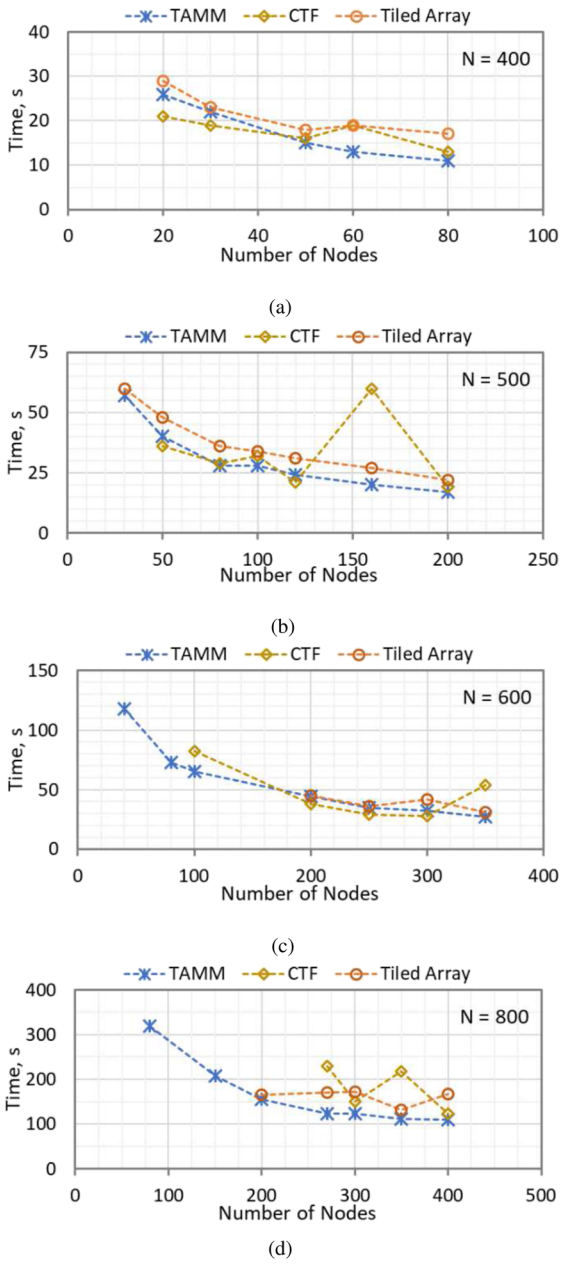

FIG. 10. Scaling of 4D tensor contraction with TAMM and comparison with two other implementations.

along each dimension in each run and chose the timings for the best performing case for comparing against TAMM,

$$C(a,b,c,d) += A(a,b,m,n) * B(m,n,c,d). \quad (57)$$

The performance results for TAMM, TiledArray, and CTF implementations of Eq. (57) are shown in Fig. 10. Plots for $N = 400$, 500, 600, and 800 are also shown on different Perlmutter node counts. With an increasing number of nodes in order to strong scale for a given problem size (N), TAMM demonstrated consistent performance improvements. In contrast, TA and CTF exhibited some oscillating behavior. The TAMM implementation strong scales consistently for a given problem size (N) and provides competitive performance compared to TA and CTF, especially for large problem sizes and node counts.

Table V shows the minimum number of nodes required on Perlmutter to run each of the three implementations for the tensor contraction shown in (57) for large problem sizes ($N = 600$ and $N = 800$). As N increased, the required minimum number for running the benchmark in the cases of TA and CTF was much higher than expected in comparison to TAMM. At the time of this writing, CTF crashed due to segmentation faults on node counts less than 100 for $N = 600$ and node counts less than 270 for $N = 800$. Similarly, TA also required a minimum of 40 nodes for $N = 600$ and 200 nodes for $N = 800$, whereas TAMM required only 25 and 80 nodes, respectively, for those cases.

We also compare the performance of the contraction in Eq. (57) with COSMA[112] by representing the contraction as a distributed matrix-multiply operation. COSMA is a state-of-the-art distributed matrix-multiply library that is capable of running on GPUs. Equation (58) shows the equivalent of Eq. (57), represented as a distributed matrix-multiply operation in COSMA,

$$C(i,j) += A(i,k) * B(k,j). \quad (58)$$

We use the COSMA provided distributed matrix-multiply miniapp.[113] The size of each dimension for the COSMA benchmark was chosen as N^2 in order to match the overall sizes for each tensor [i.e., (N^4)] in Eq. (57). Table VI shows the values for N considered in the comparison. At the time of this writing, we observe that COSMA, like the TiledArray and CTF implementations, also requires a certain minimum number of nodes, which is significantly higher than what is required by the corresponding TAMM implementation for a given problem size (N). We observe that the TAMM performance is nearly identical to COSMA for $N = 400$ and about 10% better than COSMA for $N = 500$. For larger problem sizes, $N = 600$ and $N = 800$, the COSMA implementation ran out of memory on 300 and 400 nodes, respectively. The DISTAL[39] work also provides a performance comparison with COSMA. The DISTAL authors mention that COSMA generally achieved about 10% higher performance in comparison to DISTAL, using up to 256 nodes of the Lassen supercomputer. Since the code for DISTAL is not in the public domain at the time of this writing, a direct performance comparison with DISTAL is currently not possible.

TABLE V. Minimum number of Perlmutter nodes required to run the dense 4D tensor contraction shown in Eq. (57).

	Number of nodes		
Dim size (N)	TAMM	TA	CTF
600	25	40	100
800	80	200	270

17 August 2023 00:38:49

© Author(s) 2023

TABLE VI. Performance comparison with COSMA.

Dim size (N)	Node count	TAMM	COSMA
300	50	6	5
400	100	9	9
500	100	28	OOM[a]
500	200	17	19
600	300	33	OOM[a]
600	400	28	OOM[a]

[a]Did not run to completion due to out-of-memory errors.

CCSD benchmark: We implemented our Cholesky-CCSD equations as tensor expressions in TA and CTF to further investigate the performance on a real application. We ran calculations of Ubiquitin-DGRTL74,84 molecule with the aug-cc-pVDZ basis (O = 146, V = 1096, 7810 Cholesky vectors) on up to 350 nodes of Perlmutter using the best performance parameters for TA and CTF. With TAMM, for a single CCSD iteration on 240 nodes, we observed 80% speedup over CTF while TA ran out of memory. On 350 nodes, we observed nearly identical timing with CTF and 2.5x speed-up over TA.

VI. CONCLUSION

We have introduced and discussed the Tensor Algebra for Many-body Methods framework that enables scalable and performance-portable implementations of important computational chemistry methods on modern large-scale heterogeneous high-performance computing systems. We described the TAMM framework in detail by introducing a tensor algebra interface that provides a high-level representation of the tensor algebra operations as an embedded domain-specific language. This interface enables the separation of concerns between scientific application development and high-performance computing development efforts. The domain scientist or scientific application developer can focus on the method development instead of the performance optimization details, whereas the HPC developers focus on the underlying algorithms and optimizations. Later, we presented our modular infrastructure that allows the implementation of different optimizations on tensor data distribution, execution, and scheduling of tensor operations for efficient execution on modern heterogeneous HPC platforms. We also compared the features of TAMM against several other distributed tensor algebra frameworks. Through various case studies, we showcased the performance and productivity benefits of using the TAMM framework for implementing complete ground- and excited-state electronic structure methods that are expressed as operations on tensors. Finally, we evaluated the performance of TAMM against other distributed tensor algebra frameworks and demonstrated its competitiveness and scalability on large problem sizes and node counts on NERSC Perlmutter.

ACKNOWLEDGMENTS

This research was supported by the Exascale Computing Project (No. 17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration, and by the Center for Scalable Predictive Methods for Excitations and Correlated Phenomena (SPEC), which is funded by the U.S. Department of Energy (DoE), Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences as part of the Computational Chemical Sciences (CCS) program at Pacific Northwest National Laboratory (PNNL) under Grant No. FWP 70942. PNNL is a multi-program national laboratory operated by Battelle Memorial Institute for the U.S. DoE under Contract No. DE-AC06-76RLO-1830. N.P.B. also acknowledges support from the Laboratory Directed Research and Development (LDRD) Program at PNNL. We gratefully acknowledge the computing resources provided and operated by the Joint Laboratory for System Evaluation (JLSE) at Argonne National Laboratory, the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract No. DE-AC02-06CH11357, and the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract No. DE-AC05-00OR22725.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Erdal Mutlu: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Ajay Panyala**: Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Nitin Gawande**: Software (supporting). **Abhishek Bagusetty**: Software (supporting). **Jeffrey Glabe**: Writing – review & editing (supporting). **Jinsung Kim**: Software (supporting). **Karol Kowalski**: Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Nicholas P. Bauman**: Methodology (supporting); Software (supporting); Writing – review & editing (supporting). **Bo Peng**: Methodology (supporting); Writing – review & editing (supporting). **Himadri Pathak**: Methodology (supporting); Software (supporting); Writing – review & editing (supporting). **Jiri Brabec**: Methodology (supporting); Software (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Sriram Krishnamoorthy**: Software (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]F. Coester, Nucl. Phys. **7**, 421 (1958).
[2]F. Coester and H. Kümmel, Nucl. Phys. **17**, 477 (1960).
[3]J. Čížek, J. Chem. Phys. **45**, 4256 (1966).
[4]G. D. Purvis and R. J. Bartlett, J. Chem. Phys. **76**, 1910 (1982).
[5]J. Paldus, J. Čížek, and I. Shavitt, Phys. Rev. A **5**, 50 (1972).
[6]J. Paldus and X. Li, Adv. Chem. Phys. **110**, 1 (1999).
[7]T. D. Crawford and H. F. Schaefer, Rev. Comput. Chem. **14**, 33 (2000).
[8]R. J. Bartlett and M. Musiał, Rev. Mod. Phys. **79**, 291 (2007).

17 August 2023 00:38:49

© Author(s) 2023

[9]K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, Chem. Phys. Lett. **157**, 479 (1989).
[10]C. Hampel and H. J. Werner, J. Chem. Phys. **104**, 6286 (1996).
[11]M. Schütz, J. Chem. Phys. **113**, 9986 (2000).
[12]M. Schütz and H.-J. Werner, Chem. Phys. Lett. **318**, 370 (2000).
[13]F. Neese, F. Wennmohs, and A. Hansen, J. Chem. Phys. **130**, 114108 (2009).
[14]C. Riplinger and F. Neese, J. Chem. Phys. **138**, 034106 (2013).
[15]P. Pinski, C. Riplinger, E. F. Valeev, and F. Neese, J. Chem. Phys. **143**, 034108 (2015).
[16]C. Riplinger, P. Pinski, U. Becker, E. F. Valeev, and F. Neese, J. Chem. Phys. **144**, 024109 (2016).
[17]F. Pavosevic, P. Pinski, C. Riplinger, F. Neese, and E. F. Valeev, J. Chem. Phys. **144**, 144109 (2016).
[18]M. Saitow, U. Becker, C. Riplinger, E. F. Valeev, and F. Neese, J. Chem. Phys. **146**, 164105 (2017).
[19]See https://nwchemex-project.github.io/NWChemEx/ for detailed information about NWChemEx project.
[20]See https://spec.labworks.org/home for detailed information about SPEC project.
[21]O. R. L. C. Facility, Oak Ridge Leadership Computing Facility, 2020.
[22]S. Hirata, J. Phys. Chem. A **107**, 9887 (2003).
[23]M. Valiev *et al.*, Comput. Phys. Commun. **181**, 1477 (2010).
[24]G. Baumgartner *et al.*, Proc. IEEE **93**, 276 (2005).
[25]J. A. Gunnels, F. G. Gustavson, G. M. Henry, and R. A. van de Geijn, ACM Trans. Math. Software **27**, 422 (2001).
[26]M. D. Schatz, R. A. van de Geijn, and J. Poulson, SIAM J. Sci. Comput. **38**, C748 (2016).
[27]M. D. Schatz, T. M. Low, R. A. van de Geijn, and T. G. Kolda, SIAM J. Sci. Comput. **36**, C453 (2014).
[28]J. Poulson, B. Marker, R. A. van de Geijn, J. R. Hammond, and N. A. Romero, ACM Trans. Math. Software **39**(13), 1 (2013).
[29]E. Solomonik, D. Matthews, J. R. Hammond, J. F. Stanton, and J. Demmel, "Domain-specific languages and high-level frameworks for high-performance computing," J. Parallel Distrib. Comput. **74**, 3176 (2014).
[30]E. Solomonik and T. Hoefler, arXiv:1512.00066 (2015).
[31]S. Manzer, E. Epifanovsky, A. I. Krylov, and M. Head-Gordon, J. Chem. Theory Comput. **13**, 1108 (2017).
[32]E. Epifanovsky *et al.*, J. Comput. Chem. **34**, 2293 (2013).
[33]K. Z. Ibrahim, S. W. Williams, E. Epifanovsky, and A. I. Krylov, "Analysis and tuning of libtensor framework on multicore architectures," in *2014 21st International Conference on High Performance Computing (HiPC)* (IEEE, 2014), pp. 1–10.
[34]F. Neese, F. Wennmohs, U. Becker, and C. Riplinger, J. Chem. Phys. **152**, 224108 (2020).
[35]D. I. Lyakh, Int. J. Quantum Chem. **119**, e25926 (2019).
[36]J. A. Calvin and E. F. Valeev, "TiledArray: A general-purpose scalable block-sparse tensor framework" (2019); https://github.com/valeevgroup/tiledarray.
[37]C. Peng, J. A. Calvin, F. Pavošević, J. Zhang, and E. F. Valeev, J. Phys. Chem. A **120**, 10231 (2016).
[38]R. Harrison *et al.*, SIAM J. Sci. Comput. **38**, S123 (2016).
[39]R. Yadav, A. Aiken, and F. Kjolstad, "Distal: The distributed tensor algebra compiler," in *Proceedings of the 43rd ACM SIGPLAN International Conference on Programming Language Design and Implementation, PLDI 2022* (Association for Computing Machinery, New York, NY, 2022), pp. 286–300.
[40]R. Yadav, A. Aiken, and F. Kjolstad, "SpDISTAL: Compiling distributed sparse tensor computations," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage And Analysis, SC'22* (IEEE Press, 2022).
[41]P. Springer, T. Su, and P. Bientinesi, *in Proceedings of the 4th ACM SIGPLAN International Workshop on Libraries, Languages, and Compilers for Array Programming* (ACM, 2017), p. 56.
[42]V. Anisimov, LibreTT - Tensor transpose library (2021); https://github.com/victor-anisimov/Libretт.
[43]J. Nieplocha *et al.*, Int. J. High Perform. Comput. Appl. **20**, 203 (2006).
[44]E. Mutlu, K. Kowalski, and S. Krishnamoorthy, "Toward generalized tensor algebra for ab initio quantum chemistry methods," in *Proceedings of the 6th ACM SIGPLAN International Workshop on Libraries, Languages and Compilers for Array Programming* (Association for Computing Machinery, 2019), pp. 46–56.
[45]R. A. Van De Geijn and J. Watts, Concurrency: Pract. Exper. **9**, 255 (1997).
[46]F. G. Van Zee and R. A. van de Geijn, ACM Trans. Math. Software **41**, 1 (2015).
[47]J. Kim *et al.*, "A code generator for high-performance tensor contractions on GPUs," in *IEEE/ACM International Symposium on Code Generation and Optimization, CGO 2019, Washington, DC, USA, February 16–20, 2019* (IEEE, 2019), pp. 85–95.
[48]oneAPI, oneAPI math kernel library (onemkl) interfaces, https://github.com/oneapi-src/onemkl, 2020.
[49]A. Bagusetty, A. Panyala, G. Brown, and J. Kirk, "Towards cross-platform portability of coupled-cluster methods with perturbative triples using SYCL," in *2022 IEEE/ACM International Workshop on Performance, Portability and Productivity in HPC (P3HPC)* (IEEE, 2022), pp. 81–88.
[50]Y. Zheng, A. Kamil, M. B. Driscoll, H. Shan, and K. Yelick, "UPC++: A PGAS extension for C++," in *2014 IEEE 28th International Parallel and Distributed Processing Symposium* (IEEE, 2014), pp. 1105–1114.
[51]M. Bauer, S. Treichler, E. Slaughter, and A. Aiken, "Legion: Expressing locality and independence with logical regions," in *SC '12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (IEEE, 2012), pp. 1–11.
[52]B. Peng, Nicholas Bauman and S. Krishnamoorthy, Nwchemex coupled cluster (2022); https://github.com/NWChemEx-Project/CoupledCluster.
[53]K. Kowalski, S. Krishnamoorthy, R. M. Olson, V. Tipparaju, and E. Apra, "Scalable implementations of accurate excited-state coupled cluster theories: Application of high-level methods to porphyrin-based systems," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis* (IEEE, 2011), pp. 1–10.
[54]H.-S. Hu, K. Bhaskaran-Nair, E. Aprà, N. Govind, and K. Kowalski, J. Phys. Chem. A **118**, 9087 (2014).
[55]M. Rätsep, Z.-L. Cai, J. R. Reimers, and A. Freiberg, J. Chem. Phys. **134**, 024506 (2011).
[56]A. F. Collings and C. Critchley, *Artificial Photosynthesis: From Basic Biology to Industrial Application* (John Wiley and Sons, 2007).
[57]M. B. Smith and J. Michl, Chem. Rev. **110**, 6891 (2010).
[58]C. Wang, D. E. Schlamadinger, V. Desai, and M. J. Tauber, ChemPhysChem **12**, 2891 (2011).
[59]C. Wang, C. J. Berg, C.-C. Hsu, B. A. Merrill, and M. J. Tauber, J. Phys. Chem. B. **116**, 10617 (2012).
[60]E. Apra *et al.*, J. Chem. Phys. **152**, 184102 (2020).
[61]J. F. Stanton, Chem. Phys. Lett. **281**, 130 (1997).
[62]J. Kim *et al.*, "Scalable heterogeneous execution of a coupled-cluster model with perturbative triples," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (IEEE, 2020), pp. 1–15.
[63]J. Geertsen, M. Rittby, and R. J. Bartlett, Chem. Phys. Lett. **164**, 57 (1989).
[64]D. C. Comeau and R. J. Bartlett, Chem. Phys. Lett. **207**, 414 (1993).
[65]J. F. Stanton and R. J. Bartlett, J. Chem. Phys. **98**, 7029 (1993).
[66]P. Piecuch and R. J. Bartlett, "EOMXCC: A new coupled-cluster method for electronic excited states," in *Advances in Quantum Chemistry* (Academic Press, 1999), Vol. 34, pp. 295–380.
[67]K. Kowalski and P. Piecuch, J. Chem. Phys. **115**, 643 (2001).
[68]H. J. Monkhorst, Int. J. Quantum Chem. **12**, 421 (1977).
[69]H. Koch and P. Jørgensen, J. Chem. Phys. **93**, 3333 (1990).
[70]H. Koch, O. Christiansen, P. Jørgensen, A. M. Sanchez de Merás, and T. Helgaker, J. Chem. Phys. **106**, 1808 (1997).
[71]A. Altun, M. Garcia-Ratés, F. Neese, and G. Bistoni, Chem. Sci. **12**, 12785 (2021).
[72]J. M. Foster and S. F. Boys, Rev. Mod. Phys. **32**, 300 (1960).
[73]J. Pipek and P. G. Mezey, J. Chem. Phys. **90**, 4916 (1989).
[74]S. Vijay-Kumar, C. E. Bugg and W. J. Cook, J. Mol. Biol. **194**, 531 (1987).
[75]M. Schütz and H.-J. Werner, J. Chem. Phys. **114**, 661 (2001).

17 August 2023 00:38:49

© Author(s) 2023

[76]M. Schütz, Phys. Chem. Chem. Phys. **4**, 3941 (2002).
[77]C. Edmiston and M. Krauss, J. Chem. Phys. **42**, 1119 (1965).
[78]R. Ahlrichs and W. Kutzelnigg, Theor. Chim. Acta **10**, 377 (1968).
[79]Y. Guo, K. Sivalingam, E. F. Valeev, and F. Neese, J. Chem. Phys. **144**, 094111 (2016).
[80]M. Schwilk, Q. Ma, C. Köppl, and H.-J. Werner, J. Chem. Theory Comput. **13**, 3650 (2017).
[81]C. Riplinger, B. Sandhoefer, A. Hansen, and F. Neese, J. Chem. Phys. **139**, 134101 (2013).
[82]D. Mukherjee and P. K. Mukherjee, Chem. Phys. **39**, 325 (1979).
[83]E. Dalgaard and H. J. Monkhorst, Phys. Rev. A **28**, 1217 (1983).
[84]P. Pasikowski, M. Cydzik, A. Kluczyk, P. Stefanowicz, and Z. Szewczuk, Biomol. Concepts **1**, 67 (2010).
[85]K. Schönhammer and O. Gunnarsson, Phys. Rev. B **18**, 6606 (1978).
[86]D. R. Nascimento and A. E. DePrince III, J. Phys. Chem. Lett. **8**, 2951 (2017).
[87]J. J. Rehr *et al.*, J. Chem. Phys. **152**, 174113 (2020).
[88]F. D. Vila, J. J. Kas, J. J. Rehr, K. Kowalski, and B. Peng, Front. Chem. **9**, 734945 (2021).
[89]F. D. Vila, K. Kowalski, B. Peng, J. J. Kas, and J. J. Rehr, J. Chem. Theory Comput. **18**, 1799 (2022).
[90]P. Hoodbhoy and J. W. Negele, Phys. Rev. C **18**, 2380 (1978).
[91]P. Hoodbhoy and J. W. Negele, Phys. Rev. C **19**, 1971 (1979).
[92]D. A. Pigg, G. Hagen, H. Nam, and T. Papenbrock, Phys. Rev. C **86**, 014308 (2012).
[93]J. Arponen, Ann. Phys. **151**, 311 (1983).
[94]C. Huber and T. Klamroth, J. Chem. Phys. **134**, 054113 (2011).
[95]S. Kvaal, J. Chem. Phys. **136**, 194109 (2012).
[96]T. B. Pedersen and S. Kvaal, J. Chem. Phys. **150**, 144106 (2019).
[97]H. E. Kristiansen, Ø. S. Schøyen, S. Kvaal, and T. B. Pedersen, J. Chem. Phys. **152**, 071102 (2020).
[98]T. Sato, H. Pathak, Y. Orimo, and K. L. Ishikawa, J. Chem. Phys. **148**, 051101 (2018).
[99]H. Pathak, T. Sato, and K. L. Ishikawa, J. Chem. Phys. **152**, 124115 (2020).
[100]B. Peng and K. Kowalski, Chem. Phys. Lett. **672**, 47 (2017).
[101]B. Peng and K. Kowalski, J. Chem. Theory Comput. **13**, 4179 (2017).
[102]M. Nooijen and J. G. Snijders, Int. J. Quantum Chem. **44**, 55 (1992).
[103]M. Nooijen and J. G. Snijders, Int. J. Quantum Chem. **48**, 15 (1993).
[104]M. Nooijen and J. G. Snijders, J. Chem. Phys. **102**, 1681 (1995).
[105]L. Meissner and R. J. Bartlett, Int. J. Quantum Chem. **48**, 67 (1993).
[106]K. Bhaskaran-Nair, K. Kowalski, and W. A. Shelton, J. Chem. Phys. **144**, 144101 (2016).
[107]B. Peng and K. Kowalski, Mol. Phys. **116**, 561 (2018).
[108]R. Van Beeumen *et al.*, J. Chem. Theory Comput. **13**, 4950 (2017).
[109]B. Peng, A. Panyala, K. Kowalski, and S. Krishnamoorthy, Comput. Phys. Commun. **265**, 108000 (2021).
[110]A. P. Bo Peng, K. Kowalski, and S. Krishnamoorthy, GFCCLib: Green's function coupled cluster library(2020); https://github.com/spec-org/gfcc.
[111]D. I. Lyakh, ExaTENSOR: A tensor algebra library for distributed HPC systems, https://github.com/ORNL-QCI/ExaTENSOR.
[112]G. Kwasniewski *et al.*, "Red-blue pebbling revisited: Near optimal parallel matrix-matrix multiplication," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC'19* (Association for Computing Machinery, New York, NY, 2019).
[113]G. Kwasniewski *et al.*, COSMA communication-optimal matrix-multiplication (2019); https://github.com/eth-cscs/COSMA.

17 August 2023 00:38:49

© Author(s) 2023