

Distributed-Memory DMRG via Sparse and Dense Parallel Tensor Contractions

Ryan Levy^{*†}, Edgar Solomonik^{*‡}, and Bryan K. Clark^{*†}

^{*}Illinois Quantum Information Science and Technology Center (IQIST)

[†]Institute for Condensed Matter Theory and Department of Physics

[‡]Department of Computer Science

University of Illinois at Urbana-Champaign, IL 61801 USA

Email: {rlevy3, solomon2, bkclark}@illinois.edu

Abstract—The density matrix renormalization group (DMRG) algorithm is a powerful tool for solving eigenvalue problems to model quantum systems. DMRG relies on tensor contractions and dense linear algebra to compute properties of condensed matter physics systems. However, its efficient parallel implementation is challenging due to limited concurrency, large memory footprint, and tensor sparsity. We mitigate these problems by implementing two new parallel approaches that handle block sparsity arising in DMRG, via Cyclops, a distributed memory tensor contraction library. We benchmark their performance on two physical systems using the Blue Waters and Stampede2 supercomputers. Our DMRG performance is improved by up to 5.9X in runtime and 99X in processing rate over ITensor, at roughly comparable computational resource use. This enables higher accuracy calculations via larger tensors for quantum state approximation. We demonstrate that despite having limited concurrency, DMRG is weakly scalable with the use of efficient parallel tensor contraction mechanisms.

Index Terms—DMRG, tensor networks, tensor contractions, sparse tensors, quantum systems, Cyclops Tensor Framework

I. INTRODUCTION

One of the most successful optimization algorithms for 1D systems, the density matrix renormalization group (DMRG) [1]–[3], is celebrated for its speed and quality. By representing the Hamiltonian matrix as a series of tensor products, known as a matrix product operator (MPO), the ground state, i.e. minimal eigenvector of the Hamiltonian, can also be efficiently represented by tensor products in 1D. When applied to 2D systems, one must linearize the system, resulting in harder problems that scale exponentially as the width of the system. Large scale DMRG is then needed to effectively solve these problems. Achieving higher accuracy in DMRG enables better characterization of properties of fundamental quantum physical models of strongly correlated materials. However, high accuracy requires working with very large sparse tensors, necessitating the use of supercomputing resources.

The cost associated with accurate simulations of 2D systems have spurred efforts to improve parallelism within DMRG [4], [5]; so far the most practical improvements have been in the area of shared memory parallelism [6], [7] and the plurality of 2D DMRG papers use this mode of parallelism (if any parallelism at all). Two broad attempts at making a distributed-memory parallel DMRG have involved parallelizing the tensors themselves [8]–[13] or developing a new

numerical formulations of DMRG that allow for more concurrency by trading-off accuracy and compromising monotonicity of optimization. We demonstrate that effective use of distributed tensor contraction primitives suffices to accelerate the traditional DMRG algorithms with HPC resources and enable cost-effective high-accuracy calculations.

These improvements are consequential for the state of practice. A large-scale DMRG simulation can often take many weeks on a single node and is limited in accuracy by the available RAM on a machine. A massively parallel code which uses a distributed memory paradigm can overcome both these obstacles, accelerating the wall-clock time to achieve scientific results from weeks to days and reaching previously inaccessible wave function quality. However, parallelization is complicated by the need to exploit block-sparsity in tensors (due to symmetries in different systems) to minimize memory footprint and computational cost.

We present an implementation of the DMRG algorithm using the Cyclops Tensor Framework [14], allowing for a massively parallel code that has been observed to reach up to 3 TFlops/s on state of the art problems. We focus on finite 2D lattice models, which traditionally have an easily constructed Hamiltonian but require significant computational time to converge.

We introduce the formalism of tensor networks and provide a description of the DMRG algorithm in Section II. Section II describes prior work and tabulates past studies of parallel DMRG. Our own work is the first comparative study of DMRG parallelization approaches. In particular, In Section IV, we provide three approaches for managing block sparsity via sparse and dense distributed tensors. In Section V, we introduce two model problems characterizing different workloads: the 2D $J_1 - J_2$ Heisenberg model at $J_2 = 0.5$ (abbreviated *spins* throughout this work) and the triangular Hubbard model (abbreviated *electrons*). Our numerical experiments in Section VI demonstrates speed-ups of up to 99x in performance rate relative to a state-of-the-art single node code, with roughly the same resource efficiency.

II. DMRG BACKGROUND

We provide an overview of the DMRG algorithm and quantum number symmetries (specifically of $U(1)$ symme-

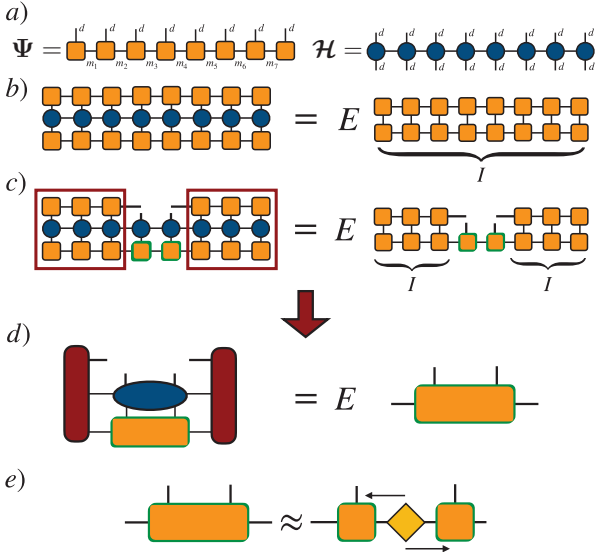


Figure 1: a) *left* the matrix product state (MPS) Ψ in orange as a tensor network and *right* the Hamiltonian H as a matrix product operator (MPO) in blue as a tensor network $\mathcal{H}$. Annotated are the physical indices d and the bond dimensions of the MPS m_i . b) The original problem $\langle \Psi | \mathcal{H} | \Psi \rangle = E \langle \Psi | \Psi \rangle$ with constraint shown in the tensor network representation. c) We select two adjacent sites to optimize simultaneously, shown in green. The optimization problem for these sites with normalization constraint is then recast as an eigenvalue problem. By exploiting an extra degree of freedom of the MPS, certain contractions can be reduced to the identity. d) The full optimization problem in b) is never directly used, instead an efficient representation is created by contracting all other sites into left and right environments. The two site tensor, shown in orange, is then optimized via a Davidson routine. e) After optimization, the order-4, two site tensor is split using SVD and (potentially) truncated to a bond dimension m_j . The singular values can then be absorbed either left or right, following the sweep direction, in order to retain a proper the orthogonal structure.

tries), which describe the sparsity structure of tensors in DMRG. More comprehensive reviews are available for tensor networks [15], DMRG [1]–[3] and quantum number symmetries [16]. At a high-level, given a Hermitian matrix represented as a 1D tensor network, the DMRG algorithm seeks to compute the eigenpair with the smallest eigenvalue (ground state energy) by using alternating optimization of a 1D tensor network that approximately represents the eigenvector.

A. Tensors, Tensor Networks, and Tensor Diagrams

The DMRG algorithm works with complex tensors. We denote an *order* N (with N *modes*) tensor of *dimensions* $s_1 \times \dots \times s_N$ as $\mathcal{T} \in \mathbb{C}^{s_1 \times \dots \times s_N}$ and its elements as $t_{i_1 \dots i_N}$. A *tensor network* $f_G(\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(M)})$ is described by a multi-graph $G = (V, E)$, where $V = \{\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(M)}\}$ and the

edges denote indices that define contraction between a pair of modes of two tensors or an uncontracted index (which we represent by a loop). If $E_i = \{e_{i1}, \dots, e_{im_i}\} \subseteq E$ is the collection of edges adjacent to vertex i and $L = \{l_1 \dots l_K\} \subseteq E$ is the set of loops, we can write the tensor network function as

$$w_{l_1 \dots l_K} = \sum_{e \in E} \prod_{j=1}^M t_{e_{i1} \dots e_{im_i}}^{(j)}.$$

Any such tensor contraction can be mapped to a matrix multiplication. If matrix multiplication is performed using the classical $O(n^3)$ algorithm, the cost of the contraction is given by the product of the dimensions of the tensor modes corresponding to all of the indices in E . We leverage the Einstein summation convention, omitting summation indices to describe tensor contractions, for instance we describe matrix multiplication as $c_{ij} = a_{ik} b_{kj}$.

A *tensor diagram* is a depiction of a tensor network f_G via the graph G , except that instead of loops, uncontracted edges correspond to edges that point into whitespace. Tensor diagrams provide a precise and intuitive way of expressing tensor networks, through which it is easier to see both the geometric structure as well as reason about contraction orderings, than via the algebraic expression of the tensor contraction. Tensor diagrams are widely used in tensor network literature; we refer the reader to [15] for a comprehensive introduction to their applications and interpretation.

B. Matrix Product States

The DMRG algorithm uses 1D tensor networks, namely the *matrix product state* (MPS) and the *matrix product operator* (MPO) [3]. These tensor networks are referred to as *tensor trains* in literature on tensor decompositions [17]. The MPS and MPO are used to represent a vector (the eigenvector guess in DMRG) and a matrix (the Hamiltonian in DMRG), respectively. Figure 1a provides the tensor diagram for an MPS (*left*) and displays the tensor diagram for an MPO (*right*).

The MPS is used as an approximation of the sought-after eigenvector in DMRG, also referred to as the *wave function*. The MPS tensor network contracts into an order N tensor $\Psi \in \mathbb{C}^{d \times \dots \times d}$ described by a set of N *sites*, each of which is represented by an order three tensor¹ $\mathcal{T}^{(j)}$ whose elements are $t_{i_j \sigma_j i_{j+1}}^{(j)}$. The MPS contracts to yield a tensor that corresponds to a *folding* of a vector $\text{vec}(\Psi) \in \mathbb{C}^{d^N}$,

$$\psi_{\sigma_1 \dots \sigma_N} = \sum_{i_1 \dots i_{N+1}} \prod_{j=1}^N t_{i_j \sigma_j i_{j+1}}^{(j)}.$$

The MPO represents the Hamiltonian matrix H as a tensor $\mathcal{H}$ of order $2N$ and factorizes it into a product of order 4 tensors,

$$h_{\sigma_1 \dots \sigma_N \nu_1 \dots \nu_N} = \sum_{i_1 \dots i_{N+1}} \prod_{j=1}^N h_{i_j \sigma_j \nu_j i_{j+1}}^{(j)}.$$

¹the first and last tensor would have one mode be of unit dimension by this convention

The *bond dimension* of an MPS or an MPO, which we denote by m and k respectively, is the maximum dimension of a mode of any site indexed by a contracted index i_j . We also refer to the j th bond dimension (range of index i_j) of an MPS as m_j . The *physical dimension* of an MPS or an MPO, which we denote by d , is the dimension of each mode of Ψ and $\mathcal{H}$ and in many contexts is a fixed small number like 2 or 4. The product of an MPO and an MPS $\mathcal{H}|\Psi\rangle$ can be represented exactly as an MPS with bond dimension kd .

MPOs with low bond dimension provide an exact representation of Hamiltonians for quantum lattice systems with local interactions [3]. The bond dimension generally grows with the number of terms in the Hamiltonian, which depends on the physical system being studied.

C. DMRG Algorithm

Given an MPO representation of a Hamiltonian H , the DMRG algorithm seeks to approximate its ground state and energy thereof by minimizing

$$\min_{\Psi \in \mathbb{C}^{d \times \dots \times d}, \langle \Psi, \Psi \rangle = 1} \frac{\langle \Psi | \mathcal{H} | \Psi \rangle}{\text{vec}(\Psi)^\dagger H \text{vec}(\Psi)}.$$

with Ψ approximated by an MPS. To do so the DMRG algorithm optimizes each site in an alternating manner. In particular it updates each site $\mathcal{T}^{(j)}$ of the MPS by solving a reduced quadratic optimization problem,

$$\min_{\mathcal{T}^{(j)} \in \mathbb{C}^{m_j \times d \times m_{j+1}}, \langle \mathcal{T}^{(j)}, \mathcal{T}^{(j)} \rangle = 1} \text{vec}(\mathcal{T}^{(j)})^\dagger \underbrace{\mathcal{Q}^\dagger H \mathcal{Q}}_{\mathcal{K}} \text{vec}(\mathcal{T}^{(j)}).$$

The projection $\mathcal{Q}$ is a matricization of a tensor $\mathcal{Q}$ defined by the contraction of all sites in the MPS except $\mathcal{T}^{(j)}$,

$$q_{\sigma_1 \dots \sigma_N i_j i_{j+1}} = \sum_{i_1 \dots i_{j-1} i_{j+2} \dots i_{N+1}} \prod_{k=1, k \neq j}^N t_{i_k \sigma_k i_{k+1}}^{(k)}.$$

The projected Hamiltonian $\mathcal{K} \in \mathbb{C}^{m_j d m_{j+1} \times m_j d m_{j+1}}$ is the matricization of a tensor $\mathcal{K} \in \mathbb{C}^{m_j \times d \times m_{j+1} \times m_j \times d \times m_{j+1}}$, so that

$$k_{i_j \sigma_j i_{j+1} l_j \nu_j l_{j+1}} = \sum_{\substack{\sigma_1 \dots \sigma_{j-1} \sigma_{j+1} \dots \sigma_N \\ \nu_1 \dots \nu_{j-1} \nu_{j+1} \dots \nu_N}} q_{\sigma i_j i_{j+1}}^\dagger h_{\sigma \nu} q_{\nu l_j l_{j+1}},$$

where we use vector notation for indices, e.g., $\sigma = \sigma_1 \dots \sigma_N$. In matrix form, we can write $\mathcal{Q}$ as $\mathcal{Q} = \mathcal{L} \otimes \mathcal{I} \otimes \mathcal{R}$. The left and right components of the MPS $\mathcal{L}$ and $\mathcal{R}$ may be orthogonalized $\mathcal{L}$ by performing a QR factorization of each site and maintaining orthogonality during the optimization process, which gives a QR factorization of $\mathcal{L}$ overall, after which the upper-triangular factor may be absorbed into the tensor $\mathcal{T}^{(j)}$. When $\mathcal{L}$ and $\mathcal{R}$ are both orthogonal, the MPS is said to be in a *canonical form* with *center site* j . A canonical form ensures that $\mathcal{Q}$ is an orthogonal projection and that any approximation (local error) to $\mathcal{T}^{(j)}$ amplifies overall error in the state minimally [18]. These components of the MPS are contracted with respective parts of the MPO $\mathcal{U}$ and $\mathcal{W}$, which include all sites in the MPO before and after the j th site. Within DMRG, these are combined with $\mathcal{L}$ and

$\mathcal{R}$, respectively, to form the *left* and *right environments*. For example, the left environment is given by

$$a_{i_j k_j l_j} = \sum_{\sigma_1 \dots \sigma_{j-1}} \sum_{\nu_1 \dots \nu_{j-1}} l_{\sigma_1 \dots \sigma_{j-1} i_j}^\dagger \underbrace{\left(\sum_{k_1 \dots k_{j-1}} \prod_{n=1}^{j-1} h_{k_n \sigma_n \nu_n k_{n+1}}^{(n)} \right)}_{u_{\sigma_1 \dots \sigma_{j-1} \nu_1 \dots \nu_{j-1} k_j}} l_{\nu_1 \dots \nu_{j-1} l_j}.$$

The right environment, $\mathcal{B}$ is formed similarly, and then the reduced Hamiltonian is defined by

$$k_{i_j \sigma_j i_{j+1} l_j \nu_j l_{j+1}} = \sum_{k_j, k_{j+1}} a_{i_j k_j l_j} h_{k_j \sigma_j \nu_j k_{j+1}}^{(j)} b_{i_{j+1} k_{j+1} l_{j+1}}.$$

This representation of $\mathcal{K}$ is dominated in size by $\mathcal{A}$ and $\mathcal{B}$ both of which have $m^2 k$ elements, while $\mathcal{K}$ is of size $m^4 d^2$. DMRG sweeps left-to-right and then back, extending the environments from center to center by contracting with the updated tensor. After forming the environments, the updated $\mathcal{T}^{(j)}$ may be obtained using an iterative method such as CG or Davidson's algorithm with $\mathcal{K}$ applied via the implicit form defined by the two environments and center MPO site $\mathcal{H}^{(j)}$. The form of each matrix vector product, which has overall cost $O(m^3 k d)$, is described by the tensor diagram in fig. 1(d).

Our implementation of the Davidson algorithm [26] is based on the ITensor library implementation [27], except without the use of preconditioning and with randomization to alleviate failed reorthogonalization. Alg. 1 outlines the approach in matrix form.

Algorithm 1 Davidson routine implementation. The effective matrix $\mathcal{A}$ given by the MPO and environment tensors, is shown in fig. 1 d).

```

1: Input: Tensor  $x_0 \in \mathbb{R}^{s_1 \times \dots \times s_4}$ 
2: Initialize  $v_0 = x_0$ ,  $v_0^A = \mathcal{A}x_0$ , matrix  $M$ 
3: for  $i=0,1,\dots$  do
4:   for  $j=0,\dots,i$  do
5:      $m_{ij} = m_{ji} = (v_j^A)^\dagger v_i$ 
6:   end for
7:   Diagonalize leading  $i \times i$  block of  $M$ , computing smallest eigenvalue/vector  $(\lambda, s)$ 
8:    $x = \sum_j s_j v_j$ ,  $q = \sum_j s_j v_j^A$ 
9:    $q = q - \lambda x$ 
10:  Check convergence based on norm of  $q$ 
11:  Orthogonalize  $q$  with all  $v_j$  via modified Gram-Schmidt
12:   $v_{i+1} = q$ ,  $v_{i+1}^A = \mathcal{A}v_{i+1}$ 
13: end for
14: return  $x/\|x\|$ 

```

In doing DMRG, we gradually increase bond dimension of the MPS, sweeping over all sites multiple times for each successive bond dimension choice. During the sweep, we use a subspace size of 2 in the Davidson routine. Additionally, while preconditioning accelerates convergence, we find that for the problems presented here, the additional memory and time cost is prohibitive compared to the cost of running more sweeps. This can be attributed to fact that each optimization subproblem is supplied a very good initial guess and need

System	Work	Method	Architecture	Maximum Bond Dim. (m)	Maximum Nodes
Heisenberg $J_1 - J_2$	this work	$U(1)$ DMRG	Distributed Memory	32 768	256
	Jiang, et al. [19]		NR ¹	12 000	NR ¹
	Wang, et al. [20]			12 000	
Triangular Hubbard	this work	$U(1)$ DMRG	Distributed Memory	32 768	256
	Shirakawa, et al. [21]		NR ¹	20 000	NR ¹
	Szasz, et al. [22]	$U(1) + k$ space iDMRG	Shared Memory	11 314	
Hubbard 1D Chain	Rincón, et al. [23]	$U(1)$ DMRG	Distributed Memory ²	1 000	8
$U - V$ Hubbard	Kantian, et al. [11], [12]		Distributed Memory	18 000	180 ³
Square Hubbard	Yamada, et al. [9], [24]	s -leg DMRG	Distributed Shared Memory	1 200	
Heisenberg 1D Chain	Vance, et al. [10]	$U(1)$ iDMRG	Distributed Memory ⁴	2 048	64
Heisenberg J_1	Stoudenmire, et al. [4]	Parallel $U(1)$ DMRG	Real-Space Parallel	2 000	10

¹ Not Reported, assumed single node shared memory architecture

² blocks are distributed but elements are not distributed over processors

³ we use the timing results published in ref. [11] which we expect to correlate with the publication ref. [12]

⁴ via PETSc and SLEPc to ref. [12]

Table I: Comparison of prior work on the systems of interest and other known parallel DMRG works on the lattice. The physical system are 2D cylinders unless noted otherwise. For completeness, we include two infinite DMRG (iDMRG) methods [1], [2], [25], which, while similar in some aspects to DMRG, is a different algorithm beyond the scope of this work.

not be solved to high accuracy for an intermediate bond dimensions.

A standard extension of optimizing a single site is to optimize two sites simultaneously. By contracting two neighboring sites and forming

$$x_{i_j \sigma_j \sigma_{j+1} i_{j+2}}^{(j,j+1)} = \sum_{i_{j+1}} t_{i_j \sigma_j i_{j+1}}^{(j)} t_{i_{j+1} \sigma_{j+1} i_{j+2}}^{(j+1)}$$

and then performing optimization over $x^{(j,j+1)}$.

After optimization, the new $x^{(j,j+1)}$ is decomposed back into two tensors via singular value decomposition (SVD) shown pictorially in fig. 1e. The number of singular values kept determines the new bond dimension of index i_{j+1} shared on the two order-3 tensors of the MPS $\mathcal{T}^{(j)}, \mathcal{T}^{(j+1)}$. The bond dimension at site j can increase exponentially from the two ends, so $m_j \leq \min(d^j, d^{N-j})$, but is generally capped at a particular value, $m \sim O(10^3 - 10^4)$. Error incurred due to truncation can be calculated from the sum of the truncated singular values. Our implementation removes all singular values below 10^{-12} . Note that the MPO bond dimension² $k \ll m$.

D. Quantum Numbers

An essential improvement is to decompose the tensors by global symmetry representations. Consider a global symmetry group G with element $g \in G$ and unitary operator representation $U_g \in \mathbb{C}^{d^{2N}}$. We will restrict to abelian groups, namely $U(1)$, although it is possible to consider non-abelian groups as well. Most Hamiltonians of interest respect a global symmetry, i.e. $[H, U_g] = 0 \forall g \in G$. Examples of the symmetries include the total magnetic spin S_z and/or particle number. This permits

the Hamiltonian, and subsequently the MPS and MPO, to be represented into a *block form* (see cartoon in fig. 3b). Once a given global symmetry is fixed in the MPS, the Hamiltonian and all local operators will respect this symmetry by Schur's lemma, and DMRG will respect the block form of the tensor networks.

Each order- r tensor $\mathcal{T}$ can be described in block form with a list $\{q^{(\ell)}\}$ of quantum number label tuples $q^{(\ell)} \in \mathbb{Z}^r$ where each tuple of labels correspond to an independent order- r tensor block $T_{q^{(\ell)}} \in \mathbb{C}^{d_1^{\ell} \times d_2^{\ell} \times \dots \times d_r^{\ell}}$. Each label $q_i^{(\ell)}$ in turn corresponds to a different degenerate space, and thus the size of each tensor block index is variable with a maximum size determined by the group symmetry element. With every block index set to the maximum size we can recover the original space of the tensor network.

Using quantum blocks induces block-sparsity through the tensors; the nature of the block sparsity (i.e. size, number, and structure of blocks) depends on the physics of the system (as seen in fig. 2a). Quantum blocks improves DMRG in various ways. When represented in a sparse or block-sparse format, the required memory can be decreased from $\prod_i^r d_i$ to $\sum_{\ell} \prod_i^r d_i^{\ell}$ per tensor. Contractions and SVD can now be performed over individual blocks. As the cost of these operations is often cubic in bond dimension, this is a non-trivial savings.

III. PRIOR WORK

Despite the requirement for large DMRG simulations in various physics problems (the DMRG algorithm is cited ≈ 6000 times), there has been little widespread use of a massive parallel implementation on the lattice. This situation differs in quantum chemistry applications of DMRG where there are higher costs associated with individual tensor contractions due to basis sets [8], [13].

²e.g. we consider $k \sim 30$ and $m > 4096$

In table I, we compare and contrast the prior work both on other parallel DMRG attempts, as well as serial DMRG simulations on the two prototypical physics systems we consider, the two-dimensional Heisenberg $J_1 - J_2$ and the triangular Hubbard system; these systems are further discussed in section V.

It is interesting to note that the plurality of simulations with the largest bond dimensions to date have primarily been on serial or shared memory machines. In particular, all studies of the two systems we are considering fall into this category. Bond dimensions have saturated around $m \sim 10000$ and are quickly being limited by the RAM required to store the necessary tensors and intermediate contractions. To avoid an extra factor of system size in RAM, the tensors for all but the two sites being worked on are often written to disk; this generates additional significant latency. At some point, even storing the tensors in RAM for a single optimization becomes impossible in the RAM on a single node. Our work is the first to exploit parallelism via general sparse and dense distributed tensor contractions. Our approach not only drastically improves the wall-clock time of the calculation over single-node execution, but gives access to bond dimensions inaccessible to the RAM on a single node (avoiding even the need to write to disk).

There are previous works that have implemented distributed-memory parallel DMRG and DMRG like algorithms. The cases where the standard DMRG algorithm has been parallelized with distributed memory include refs. [10], [12], [23] and massively-parallel shared memory in ref. [24].

Ref. [23] achieves parallelism over a one-dimensional system by distributing different quantum number blocks to different nodes. As the size of the largest block generically scales linearly with the bond dimension (see fig. 2a), this severely limits the maximum achievable problem size. Refs. [11], [12] develop a parallel distributed memory implementation to study the $U-V$ Hubbard model ($d = 4$). Using threading with Intel® Cilk™ Plus, they scale to a bond dimension of $m = 18000$ via block-sparse matrix contractions parallelizing blocks over processors and block elements over groups of processors. This approach was outlined in the appendix of a paper [12] describing new physics on the UV-model. Some scaling results of this effort are included in a related thesis [11]. In ref. [24], the authors use a slightly-modified DMRG algorithm; instead of optimizing two sites at a time, they optimize over $2W$ sites simultaneously where W is the width of their problem. This induces a cost of W over standard DMRG which they deal with in a parallel manner.

Other approaches instead parallelize alternatives or variants to DMRG. Ref. [10] parallelizes infinite DMRG [1], [2], [25], a variant of DMRG designed for translationally invariant MPS of uniform tensors $\mathcal{T} = \mathcal{T}^{(j)} \forall j$, by parallelizing contractions using the PETSc and SLEPc library. In ref. [4] a DMRG-like algorithm is proposed in which different nodes work on different ranges of sites. While this approach is shown to achieve good parallel scalability with over 10 nodes, each optimization is done in a way that is not consistent with the

tensors on other nodes, resulting in potential loss of accuracy and monotonicity in optimization.

Our work differs from these previous works on distributed DMRG both in scale as well as approach. We compute DMRG in the same way as the best sequential approach, preserving both the efficiency (i.e. same number of flops) as well as benefits of the standard algorithm. Unlike some prior approaches, we directly distribute each tensor (or quantum block of a tensor) over all nodes. This allows each processor to work simultaneously on each contraction and avoids load-balancing issues associated with different size quantum blocks on different nodes. In terms of scale, we use significantly more nodes (256), tackle difficult two-dimensional systems, and reach significantly bigger bond dimensions than previously. The closest prior work is ref. [12], which tackled an inherently different problem only achieving a factor of 2 lower in bond dimension. We also compare multiple different algorithms for performing tensor contractions as well as contrast two qualitatively different types of physical systems. Furthermore, the approach between the two works is different. All prior memory-distributed work has used a matrix/vector formalism for distributed computing while we work directly distributing higher-order tensors including throughout the intermediate steps of optimization.

IV. ALGORITHMS

Our parallelization³ is developed on top of the serial DMRG code `tensor-tools` and builds upon the Cyclops Tensor Framework [14] to leverage large scale computing resources. Cyclops handles the shared memory structure for the tensor, contraction, and provides a pass through to parallel linear algebra routines, *e.g.* SVD, from ScaLAPACK and the High-Performance Tensor Transpose library (HPTT) [28]. A variety of alternative tensor libraries exist, including ones focused on sparse tensor contractions [29]–[32] or distributed dense contractions [33]–[36] as Cyclops, as well as parallel block-sparse tensor contractions [32], [37], [38]. Development of DMRG using the last category (block-sparse libraries) would be promising future work, as they may more effectively exploit the sparsity structure of tensors we consider. However, their use may entail potential overheads both in implementation complexity, *e.g.*, unlike Cyclops, DBCSR requires manual specification of processor grids [39], as well as performance. Block-sparse tensor contraction libraries [32], [37], [38] have been previously only used for electronic structure methods with workloads that differ significantly from DMRG.

Our focus in this work is in the limit of large bond dimension where serial codes take significant time and the tensors become large enough that the gains of parallelization outweigh its overhead. DMRG, at fixed bond dimension, has linear scaling with system size and our parallelization approach preserves this scaling.

Algorithm	Flops	Memory		BSP cost for Davidson iteration	
		Davidson(M_D)	Environments	BSP supersteps	BSP comm cost
List	$O((m/q)^3 kd^2)$	$O((m/q)^2 kd^2)$		$O(N_b)$	$O(M_D/p^{2/3})$
Sparse-Sparse	$O((m/q)^3 kd^2)$	$O((m/q)^2 kd^2)$	$O(N(m/q)^2 k)$	$O(1)$	$O(M_D/p^{1/2})$
Sparse-Dense	$O(m^3 kd^2)$	$O(m^2 kd^2)$		$O(1)$	$O(M_D/p^{1/2})$

Table II: Complexity of each algorithm implementation, in terms of number of sites N , bond dimension $m = \sum_{\ell}^{N_b} b_{\ell}$ and number of blocks N_b , MPO bond dimension k , and physical dimension d . Here we are using a empirically motivated model where the ℓ th block has auxiliary dimension $b_{\ell} = \lfloor (m/q)r^{\ell} \rfloor$; the values $q = 4, r = 0.6$ for spins and $q = 10, r = 0.65$ for electrons are rough estimates of the parameters fit to our data.

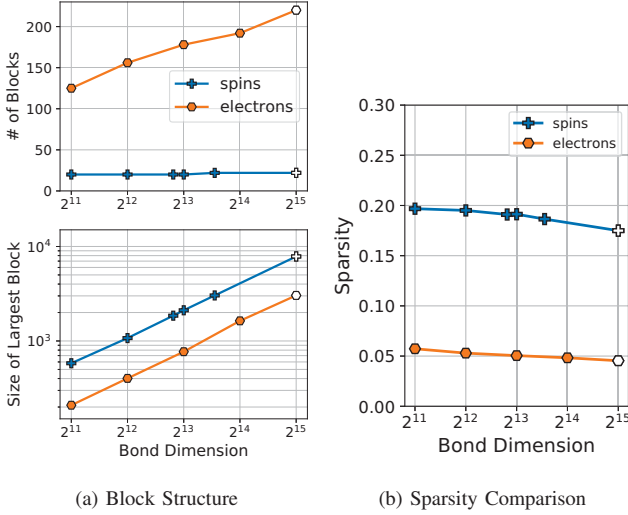


Figure 2: comparison of MPS (a) block structure and (b) sparsity for a representative MPS tensor. Open circle results are taken from MPS where only a small subset of sites (including the measured site) were at a given bond dimension. The largest block scales as $m^{0.94}$ for spins and $m^{0.97}$ for electrons.

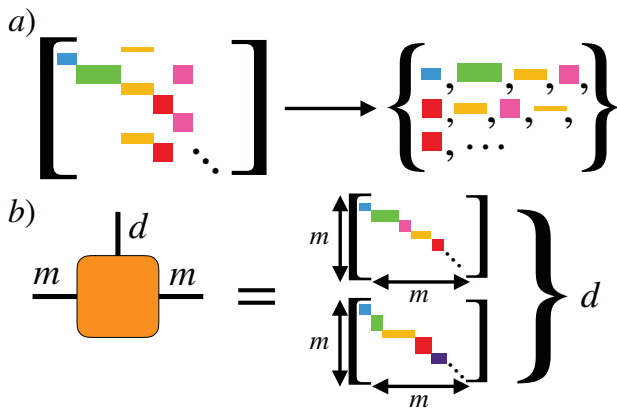


Figure 3: Cartoon of the tensor quantum block structure. a) For the list format, we deconstruct the many blocks into a list of distributed memory tensors. b) each tensor in an MPS has a special block diagonal structure, where the blocks form d $m \times m$ matrices.

Algorithm 2 Contraction of two tensor objects A and B to resultant tensor object C , each composed of a list of quantum number blocks

```

1: Input: Tensor objects A and B containing lists of quantum
   number blocks to represent tensors  $\mathcal{A} \in \mathbb{R}^{s_1 \times \dots \times s_{r_A}}$  and  $\mathcal{B} \in$ 
    $\mathbb{R}^{d_1 \times \dots \times d_{r_B}}$ , contracted index/indices mid
2:  $i_A = \text{getIndexLocations}(A, \text{mid})$ 
3:  $i_B = \text{getIndexLocations}(B, \text{mid})$ 
4:  $C = \text{new Tensor}(\text{order} = r_A + r_B - i_A.\text{size}())$ 
5:  $\text{CBlocks} = []$ ,  $\text{qToBlockIndex} = \{\}$ 
6: for ABlock in A.blocks do
7:    $q_A = \text{getQuantumNumberLabels}(\text{ABlock})$ 
8:   for BBlock in B.blocks do
9:      $q_B = \text{getQuantumNumberLabels}(\text{ABlock})$ 
10:    if  $q_A[i_A] \neq q_B[i_B]$  then continue
11:    end if
12:     $q_C = []$   $\triangleright$  build  $q_C$  from remaining labels
13:    for  $i \leftarrow 0, \dots, q_A.\text{size}()$  do
14:      if  $i$  not in  $i_A$  then  $q_C.\text{append}(q_A[i])$ 
15:      end if
16:    end for
17:    for  $i \leftarrow 0, \dots, q_B.\text{size}()$  do
18:      if  $i$  not in  $i_B$  then  $q_C.\text{append}(q_B[i])$ 
19:      end if
20:    end for
21:    if  $q_C$  is in  $\text{qToBlockIndex}$  then
22:       $\text{Cidx} = \text{qToBlockIndex}[q_C]$ 
23:       $\text{CBlocks}[\text{Cidx}] += \text{ABlock}.\text{contract}(\text{BBlock})$   $\triangleright$  CTF
24:    else
25:       $\text{CBlocks}.\text{append}(\text{ABlock}.\text{contract}(\text{BBlock}))$   $\triangleright$  CTF
26:       $\text{qToBlockIndex}[q_C] = \text{CBlocks}.\text{size}()$ 
27:    end if
28:  end for
29: end for

```

A. Tensor Structure

Our tensors have block-sparsity due to the quantum number structure of the matrix product state. We consider and test three different interfaces for representing this block-sparsity in memory which correspond to three different algorithms for contraction. The performance of each interface is dependent upon the structure of the quantum number blocks.

list algorithm: The first algorithm for tensor contraction, the *list* algorithm, stores each tensor as a set of memory distributed tensor blocks $\mathcal{T}_{q^{(\ell)}}$ for each tuple of quantum number labels $q^{(\ell)}$, initialized via a dense Cyclops tensor (cartoon shown

³source code can be found at <https://github.com/ClarkResearchGroup/tensor-tools>

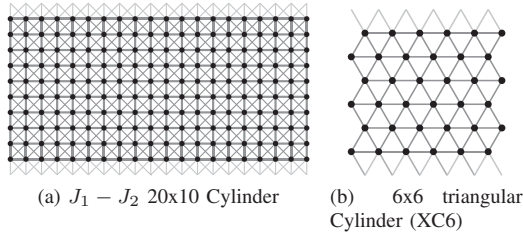


Figure 4: Lattice structure of the two benchmark systems.

in fig. 3a). To contract two tensors, the quantum number label structure must be analyzed to determine which blocks contract as well as the resulting block structure, shown in algorithm 2. All possible combination of blocks that have the same label along contracted indices are contracted together via parallel tensor contractions leaving the remaining labels that are not among the contracted indices as the resultant labels. The resultant tensor is then made up of a new set of distributed tensor blocks.

sparse-dense: Alternatively, tensors can be formed by combining all blocks into a single distributed tensor. To form the tensor, each quantum number label is mapped to a unique tensor index range of dimension of the quantum number label. Note that this tensor has non-trivial sparsity. All operations such as addition and contraction will produce the same output as with the list format, with a single contraction call. Tensors stored in this format have a higher memory cost, so that each MPS tensor now has storage cost dm^2 , the same as without quantum numbers. To conserve memory but exploit dense-dense tensor contraction performance, environment tensors, MPS, and MPO tensors are stored in a sparse format. Intermediate tensors of the Davidson routine are stored as dense, which we call the *sparse-dense* algorithm.

sparse-sparse: When a system has a quantum number with many relevant labels (e.g. particle number) or there are many combination of labels (e.g. two or more conserved quantities), we observe that the single tensor of blocks is quite sparse as shown in fig. 2b. Therefore we can store all intermediate tensors in a Cyclops sparse tensor, or the *sparse-sparse* algorithm. This has an additional overhead of keeping track of non-zero elements of each tensor along with determining output sparsity and the distributed distribution of elements. Knowledge of quantum number labels allows for pre-computation of the output sparsity, which can be provided to Cyclops to control memory consumption during contraction.

For all algorithms, the SVD portion of DMRG is performed via the list method. For sparse-sparse and sparse-dense this requires that the blocks are extracted from the single tensor and put into a temporary list format. As SVD is only defined on a order-2 tensor, the tensor indices are ‘wrapped’ to form an effective order-2 matrix with a row index and a column

index. Quantum numbers of the singular vector ‘tensors’ are used to calculate legal quantum numbers for the virtual index between the vectors and the singular values. Once this has been computed, a subset of blocks are reshaped into a matrix, grouped via similar quantum numbers along a row or column index, and decomposed. This produces a block structured tensor which can be reshaped into two order-3 tensors, and singular values absorbed into either the left or right tensor, along the direction of the sweep (see fig. 1e). We utilize a distributed SVD routine through ScaLAPACK [40], so as to minimize redistribution costs of moving data onto a single node to call a serial SVD routine. Finally any sparse structure is recovered by rearranging data in parallel from the list format into a single sparse tensor.

We include in table II complexity of each format using a simplified, empirically motivated block structure model. We also provide communication cost models for the most costly contractions in the method, which are quantified using the Bulk Synchronous Parallel (BSP) [41], [42], in terms of super-steps (number of global synchronizations) and communication cost (amount of data sent along the critical path of execution). The costs are based on the algorithms used by Cyclops, which have a cost that depends on available memory [14], [43]. We assume that enough memory is available to achieve the minimal possible communication when executing a block-wise contraction with all processors, but that no (at most a constant factor of) additional memory is available when all blocks are multiplied within a single tensor contractions. The analysis demonstrates that the choice of best method depends on the problem parameters (e.g., number of blocks), as there is a trade-off between synchronization and communication costs.

V. PHYSICAL SYSTEMS

Condensed matter systems fall broadly into two classes: electron systems for which the underlying itinerancy of the electron matters and spin-systems where the spins can be treated as stationary but interact with each other. We benchmark two challenging systems (one in each category) which have been heavily studied but for which there is not yet consensus about the underlying physics. The first, a spin system (called *spin* throughout the text) with $d = 2$ physical degrees of freedom, is the $J_1 - J_2$ Heisenberg model,

$$H = J_1 \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \mathbf{S}_i \cdot \mathbf{S}_j,$$

where $\mathbf{S}_i$ is a spin operator on site i , while $\langle i,j \rangle$ and $\langle\langle i,j \rangle\rangle$ iterate over sites on a 2D lattice that are, respectively, directly and diagonally adjacent. At $J_2/J_1 = 0.5$ there is disagreement about the phase coming both from multiple DMRG studies [19], [20], [44] as well as other tensor network approaches [45]–[47]. While we will not resolve that problem here, using microbenchmarks we will show that we can reach a bond dimension m significantly above current state-of-the-art simulations at comparable efficiency and significantly reduced wall-clock time.

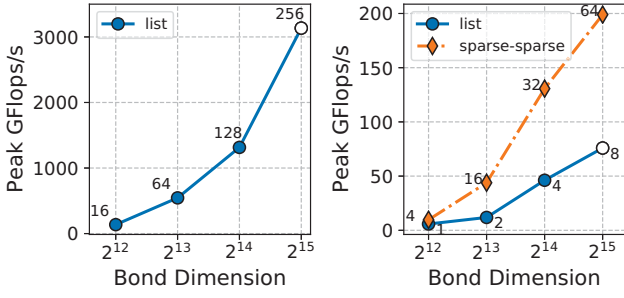


Figure 5: Summary of the peak performance of the spin (left) and electron (right) systems. GFlops/s are averaged over a sweep of 10 (1) sites for the spin (electron) system respectively. The largest spin and electron list result, denoted with an open marker, is estimated from half sweep data. Annotated are the number of nodes used.

In addition to looking at the spin system above, we also benchmark an electron system (called *electron*) with $d = 4$ physical degrees of freedom, the triangular Hubbard model,

$$H = -t \sum_{\langle i,j \rangle, \sigma} (c_{i\sigma}^\dagger c_{j\sigma} + h.c.) + U \sum_i n_{i\uparrow} n_{i\downarrow},$$

where $c_{i\sigma}^\dagger$ is an electron creation operator at site i with spin $\sigma = \{\uparrow, \downarrow\}$ and $n_{i\sigma} = c_{i\sigma}^\dagger c_{i\sigma}$ is the number operator of electrons with spin σ on site i . We set $t = 1$ and $U = 8.5$ and use N electrons with $N_\uparrow = N_\downarrow = N/2$. The Hubbard model simulations differ qualitatively from the Heisenberg simulations both in their physical content (being itinerant electrons instead of stationary spins) as well as their tensor structure. In particular, the physical degrees of freedom now contain four states per site (i.e $d = 4$) and, more importantly, there are two conserved global symmetries, spin and particle number. These two symmetries translate into two quantum numbers per label which significantly increases both the number of blocks and sparsity of blocks for the same bond dimension (see fig. 2a and fig. 2b). Different techniques also disagree on the phase diagram of this model [21], [22].

The Hamiltonian for both our systems is encoded as a MPO. The structure of this MPO is not unique. Because we are using ITensor to compare against, to ensure equity in this comparison, we use exactly the same MPO ITensor generates by directly using their AutoMPO functionality [27].

VI. NUMERICAL EXPERIMENTS

Numerical experiments are performed on Blue Waters and Stampede2 supercomputers. All reported Blue Waters timings use the Cray XE6 nodes with 64 GB of RAM and dual 8-“core” processors per node connected via Cray’s Gemini interconnect, while with Stampede2 we utilize only Knight’s Landing (KNL) nodes which have a 68 core processor, 96 GB of DDR4 RAM, 16 GB of MCDRAM, connected via an Intel Omni-Path interconnect. For all Stampede2 data and sparse algorithm data we use Intel’s MKL library which

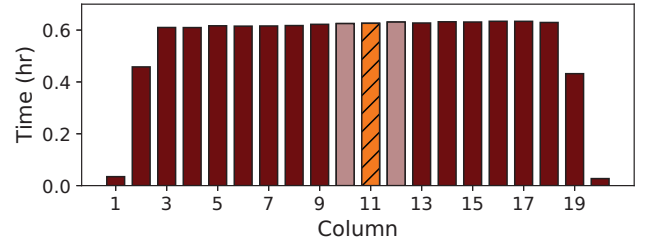


Figure 6: Time spent for each column of 10 sites for a full sweep at bond dimension $m = 8192$ for list spins. Spin benchmark optimizations are carried out on the middle 3 columns, denoted in light maroon and orange, which share similar timings to all non-edge columns in a full sweep. We report timing for only the hatched center column to ignore any additional edge effects.

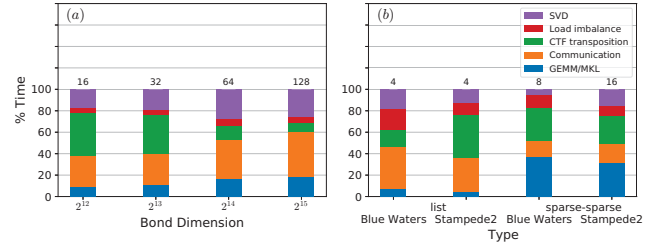


Figure 7: Breakdown of percentage of time spent for (a) spins on Blue Waters with 16 MPI processes/node and (b) electron systems at $m = 2^{14}$ on both Blue Waters (16 processes/node) and Stampede2 (64 processes/node). Annotated are the number of nodes used. Communication costs include MPI calls excluding those in SVD (ScaLAPACK `pdgesvd`); CTF transposition includes CTF mapping, transpose operations, and other small serial operations; and load imbalance is measured via `MPI_Barrier()` calls.

includes BLAS, batched BLAS, ScaLAPACK, and Sparse-Sparse routines. On Blue Waters for the list algorithm, we use Cray’s LibSci library for BLAS and ScaLAPACK routines.

In order to mitigate differences in MPS block sizes and distributions, we developed an interface to convert ITensor MPS data to a readable format for Cyclops. This enables us to have extremely close comparison to single node performance for all available problem sizes with comparable single node performance.

To calculate the number of flops, we measure FLOP operations using the built in Cyclops routines for the list method. The resulting measurement is then used as basis for ITensor, list, and sparse method performance rate calculations (flops/s). The peak GFlops/s performance rate is shown in fig. 5, where we obtain a maximum performance of 3.1 TFlops/s on Blue Waters and 198 GFlops/s on Stampede2.

In order to compare the single node performance of a shared memory architecture application, we benchmark ITen-

sor on a single node with 32 threads/node on Blue Waters and Stampede2. ITensor takes advantage of threaded BLAS and LAPACK routines; on Blue Waters we use the Cray LibSci library and on Stampede2 the Intel MKL library. To approximate timings for larger problem sizes than can fit on one node, we take the maximum performance rate (GFlops/s) and report the extrapolated numbers. On Blue Waters we use the Cray XE6 high memory nodes with 128 GB of RAM.

A half-sweep of a DMRG algorithm performs N optimizations (the other half-sweep optimizes sites in the opposite order). A typical MPS (in canonical form) has little variance in bond dimension across the system (modulo those near the edge). Therefore, the time per site should be largely uniform outside the first few and last sites; we validate this in fig. 6 comparing the time of each column (10 sites) over an entire sweep. For the spins system, instead of timing all sites, we therefore optimize the middle 3 columns (or 30 sites) of the lattice reporting the timing of the middle column; toward that end, we ensure the MPS has the reported bond dimension on those 30 sites but not necessarily outside of it. This is often done by an untimed sweep where we grow the bond dimension. All sweep times and GFlops/s reported for the Heisenberg model are from the middle 10 sites of these 30 site sweeps unless noted. For $m < 8192$ we additionally have an SVD cutoff of 10^{-9} and $m \geq 8192$ a cutoff of 10^{-12} . For the Hubbard model (electrons), we choose to benchmark a single DMRG step (the 15th and 16th sites) rather than the entire system.

A. Spins ($d = 2$)

On Blue Waters, where single node throughput is (comparatively) low, in figs 8, 9 and 10, we observe that list method is the fastest and most cost effective approach, performing significantly better than sparse-dense. We consider both the strong and weak scaling of the list method on Blue Waters. In looking at this scaling, the relevant ‘problem size’ to increase is the bond dimension, which dictates the accuracy of the DMRG approximation. We find in the strong scaling that the efficiency and speedup stays ideal only for a modest increase in the number of nodes (i.e. going from 2^3 to 2^4 nodes). Beyond this, we do not see significant gains in increasing the number of nodes at fixed bond dimension (see fig. 9) with an efficiency falling to approximately 60% under an additional doubling of the nodes.

Heuristically, after determining the maximum problem size on a single node, we observe in fig. 8a that doubling the number of nodes when doubling the bond dimension maintains good efficiency. We measure efficiency relative to single node execution of ITensor. We obtain near ideal efficiency at the largest node count, with bond dimension $m = 32768$. Note that doubling problem size does not double work but increases work/node by a factor of 8 and increases memory/node by a factor of 4.

As the problem size grows, the list algorithm is able to more efficiently use the computational resources compared to that of ITensor, despite growing communication costs. In

our performance breakdown in fig. 7(a) this is supported as when bond dimension is increased, there is a corresponding increase in local matrix-matrix multiplication (GEMM), signifying improvement in efficiency. A significant but not a dominant amount of time is spent in SVD and communication, which are expected parallel bottlenecks. We also consider peak relative efficiency in fig. 8b finding good peak performance both at small and large node counts, with a dependence on MPI processes per node which crossovers at $n = 2^6$ between 32/node and 16/node being preferred.

Finally, in fig. 10 we consider the relative time and cost of a variety of node counts, processes per node, bond dimensions, and algorithms. We find that we can achieve speedups of 5.9X to 99X of the wall clock time, as bond dimensions grow from $m = 4096$ to $m = 32768$, at a relative node cost that is 1.5 times that of running ITensor. In order to theoretically compare to problems which do not fit on a single node, we calculate times using the maximum single node performance rate. We find that by considering the Pareto optimal curve in fig. 10, which selects the best/minimum relative time for a given cost and bond dimension, on Blue Waters the curve entirely consists of list algorithm simulations.

B. Electrons ($d = 4$)

In this section, we now turn to the 6x6 triangular Hubbard Model with the larger local dimension, $d = 4$. The additional quantum number symmetry — particle number — implies significantly more blocks and hence greater sparsity (see fig. 2a). Due to the lower sparsity we benchmark the sparse-sparse tensor algorithm in addition to the list method. Although these both should cost roughly the same number of total flops, the overhead is different in list, we serially enumerate over quantum blocks contracting them; this has an overhead coming from contracting small tensors in a distributed way. On the other hand, the sparse-sparse approach contracts a single pair of tensors; unfortunately, extra overhead exists from sparse operations.

To determine time and flops/s, we again benchmark optimization of only one or two sites optimizations for each bond dimension. Finally, we construct the MPO with compression, where each order-4 tensor of $\mathcal{H}$ is truncated via SVD to a 10^{-13} cutoff, resulting in an MPO with a bond dimension $k = 26$. Similarly, the SVD truncation in the DMRG has an additional cutoff of 10^{-12} for $m < 16384$ and 0 otherwise.

To begin, we benchmark sparse-sparse strong scaling performance on Blue Waters and Stampede2 in fig. 12. We see non-smooth performance due to communication contraction mapping infrastructure of Cyclops. Despite this, there is nearly ideal or better than ideal strong scaling speedup at $m = 8192$ bond dimension. However, sparse format itself has a higher memory cost than the list format and thus on Stampede2 we find 4 nodes are the minimum rather than the 2 required on Blue Waters.

Looking at DMRG weak scaling in fig. 11a, we show that

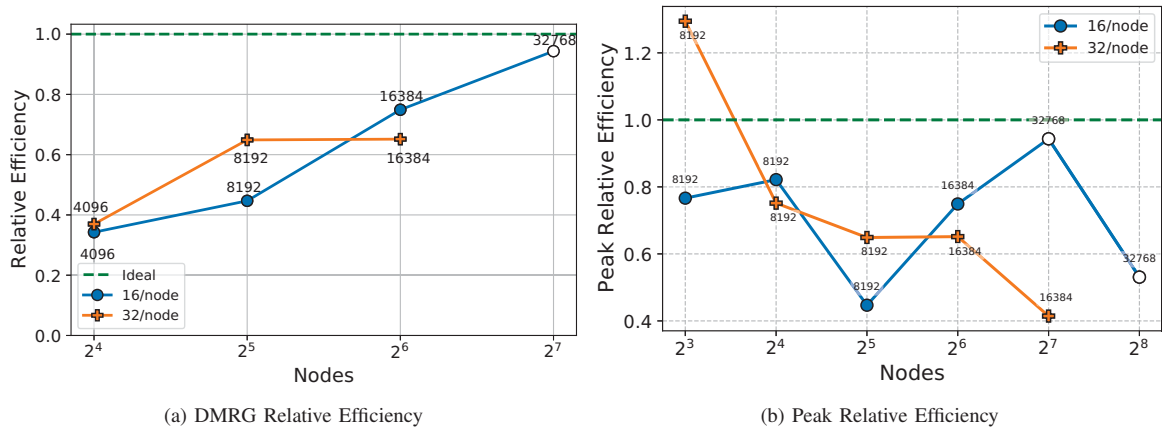


Figure 8: Weak scaling using fixed m /node (Left) and peak relative efficiency with respect to single node ITensor (Right) for list spins on Blue Waters. Bond Dimensions m are annotated on each point; relative efficiency is calculated by relative GFlops/s/node compared to ITensor on a single node at $m = 4096$. Peak relative efficiency is taken as the highest relative efficiency measured at a given node count. Open fill points are extrapolated from one half sweep.

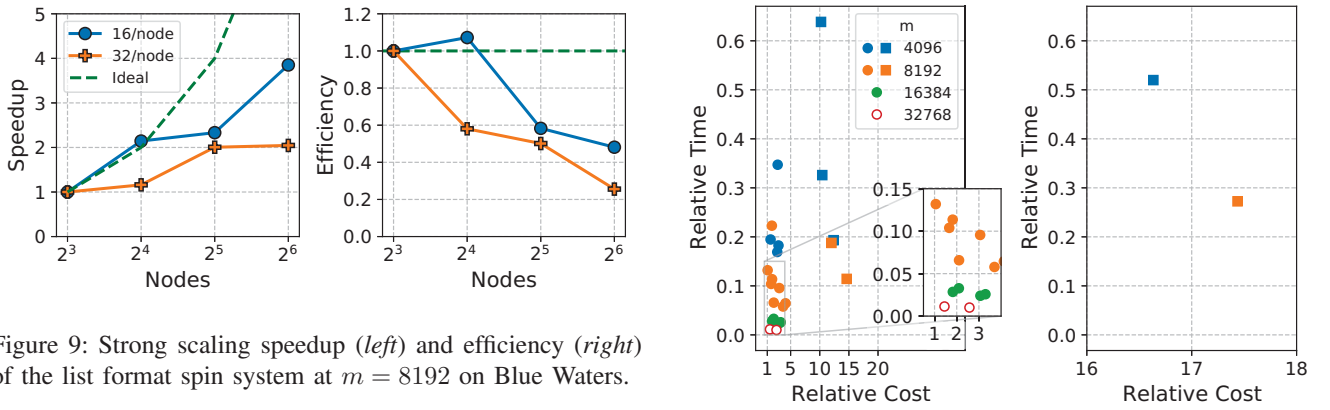


Figure 9: Strong scaling speedup (left) and efficiency (right) of the list format spin system at $m = 8192$ on Blue Waters.

we gain efficiency only at the largest problem sizes⁴. By comparing the relative time spent in fig. 7(b), we find that the sparse-sparse algorithm is able to take more advantage of sparse MKL calls, while the list method is dominated by communication and CTF transposition costs. These sparse MKL calls grow from approximately 14% time spent at $m = 4096$ to 52% time spent at $m = 32768$ for the sparse-sparse algorithm on Stampede2. There is an additional dependence on architecture, further exemplified via the peak relative efficiency in fig. 11b, as the sparse-sparse algorithm does not scale on Blue Waters, but is marginally better on Stampede2. The list algorithm is also more sensitive to overhead on each architecture: at the same node count Blue Waters has increased communication cost while Stampede2 has increased transposition costs.

The discrepancy extends to relative time and cost of the two systems as well (see fig. 13). For Blue Waters, the list

⁴efficiency at small node counts may also be improved with larger problem sizes

Figure 10: Spin system execution time and node hour cost relative to single node ITensor maximum performance rate (at $m = 4096$) using lists (circles) and sparse-dense (squares) on Blue Waters left and Stampede2 right by varying hyperparameters (node count and MPI processes/node); open points are extrapolated from a half sweep.

algorithm is the only method that is efficient in both cost and time, where the largest problem has nearly a speedup of 8X and at nearly the same performance rate a serial node (0.98X). For the sparse-sparse algorithm, we see much more expensive time-to-performance, with $m = 32768$ receiving a 14X performance rate speedup at 4.5X the relative cost.

On Stampede2, there is less of a drastic cost difference between algorithms. Using the list algorithm for $m = 16384$, there is a 2X performance rate improvement at a relative cost of 1.9X. The sparse algorithm exhibits a 3.9X speedup but at 8X the cost for our largest bond dimension. Time-to-performance can be lowered at low performance rate which

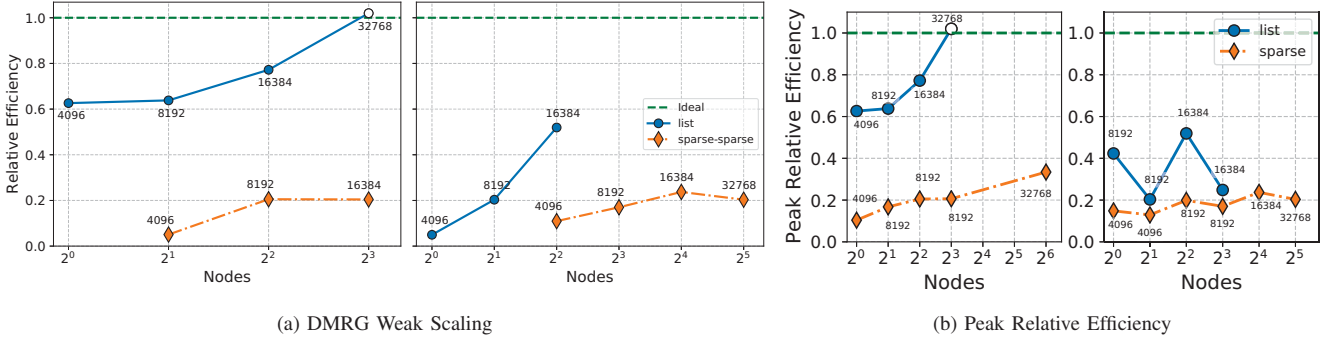


Figure 11: Weak scaling measurements for electrons compared to single node ITensor for (a) fixed m /node and (b) peak relative efficiency on Blue Waters (left plots of (a) and (b)) and Stampede (right plots). Relative efficiency is calculated by relative GFlops/s/node compared to single node ITensor at $m = 16384$ on Blue Waters and $m = 8192$ on Stampede2. Peak relative efficiency is taken as the highest relative efficiency measured at a given node count. Bond dimension m annotated at each point.

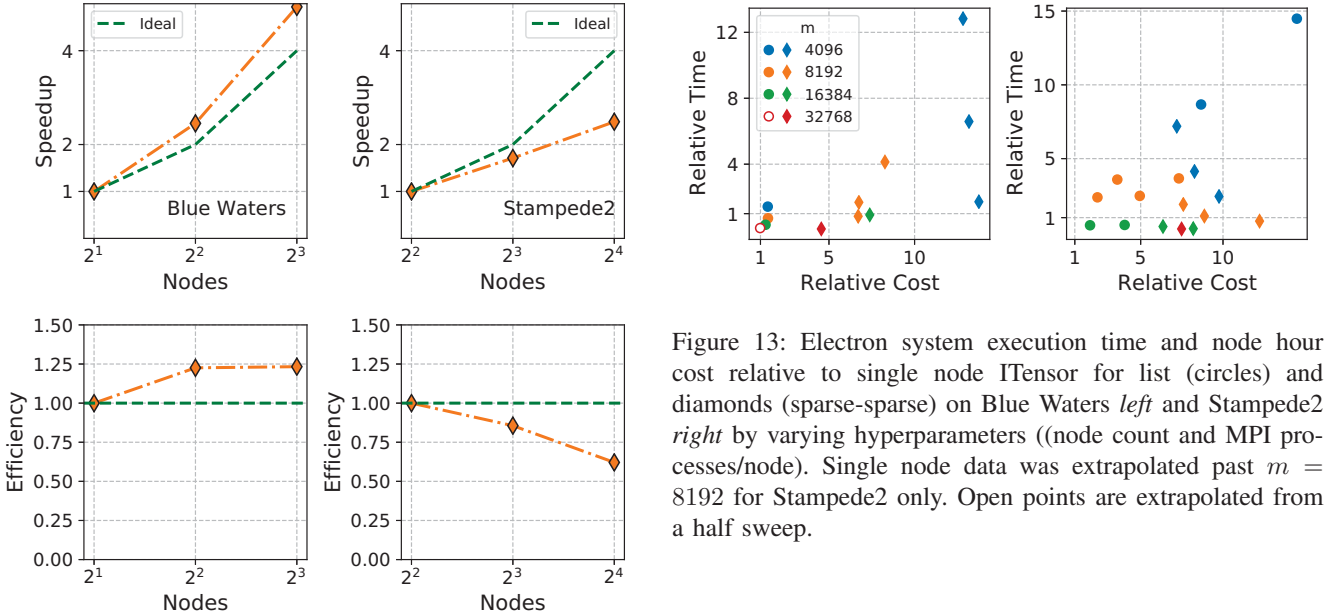


Figure 12: Strong scaling speedup (*top*) and efficiency (*bottom*) of Sparse-Sparse format for electrons at $m = 8192$ on Blue Waters (*left*) and Stampede2 (*right*)

leads to high costs.

VII. CONCLUSION

We develop a distributed memory implementation of the DMRG algorithm and benchmark on two contrasting systems. Drastically different block structures in the different systems require different tensor block algorithms to be efficient at scale. For few, large blocks, we find excellent weak scaling efficiency for large problems and exhibit speedups that are nearly at serial cost and reduce time-to-solution by 5.9X. For

Figure 13: Electron system execution time and node hour cost relative to single node ITensor for list (circles) and diamonds (sparse-sparse) on Blue Waters *left* and Stampede2 *right* by varying hyperparameters ((node count and MPI processes/node). Single node data was extrapolated past $m = 8192$ for Stampede2 only. Open points are extrapolated from a half sweep.

many blocks, our two methods produce faster time-to-solution but at a worse performance rate than a serial node. Despite these conflicting results, our implementation has an advantage over shared-memory models in that we can (weakly) scale to problems 64X greater in memory and 512X in complexity in real world applications. At up to 1.5X relative cost we were able to get up to 99X performance rate relative to an optimized single node code. This is crucial for solving the most complex problems with the high accuracy required to resolve conflicting results. These results should also extend to other DMRG-based algorithms and provide groundwork for future research on high performance implementations of tensor network methods. Particularly algorithms which do not need significant amounts of disk access are well suited for supercomputing applications at these large bond dimensions.

VIII. ACKNOWLEDGEMENTS

We thank Xiongjie Yu who was the primary developer (under the guidance of BKC) of the serial version of the `tensor-tools` code on which our parallel version was built. We acknowledge both the use of `ITensor` to compare against our `tensor-tools` code as well as the use/modification of their `AutoMPO` and lattice code directly to facilitate comparison. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. We used XSEDE to employ Stampede2 at the Texas Advanced Computing Center (TACC) through allocation TG-CCR180006. This research is also part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. BKC was supported by DOE de-sc0020165. ES was supported by NSF grant no. 1839204.

REFERENCES

- [1] S. R. White, “Density matrix formulation for quantum renormalization groups,” *Phys. Rev. Lett.*, vol. 69, no. 19, pp. 2863–2866, 1992.
- [2] —, “Density-matrix algorithms for quantum renormalization groups,” *Phys. Rev. B*, vol. 48, no. 14, pp. 10 345–10 356, 1993.
- [3] U. Schollwöck, “The density-matrix renormalization group in the age of matrix product states,” *Ann. Phys. (N. Y.)*, vol. 326, no. 1, pp. 96–192, 2011.
- [4] E. Stoudenmire and S. R. White, “Real-space parallel density matrix renormalization group,” *Physical Review B*, vol. 87, no. 15, p. 155137, 2013.
- [5] H. Ueda, “Infinite-size density matrix renormalization group with parallel Hida’s algorithm,” *Journal of the Physical Society of Japan*, vol. 87, no. 7, p. 074005, 2018.
- [6] M. Dolfi, B. Bauer, S. Keller, A. Kosenkov, T. Ewart, A. Kantian, T. Giamarchi, and M. Troyer, “Matrix product state applications for the alps project,” *Computer Physics Communications*, vol. 185, no. 12, pp. 3430–3440, 2014.
- [7] G. Hager, E. Jeckelmann, H. Fehske, and G. Wellein, “Parallelization strategies for density matrix renormalization group algorithms on shared-memory systems,” *Journal of Computational Physics*, vol. 194, no. 2, pp. 795–808, 2004.
- [8] G. K.-L. Chan, “An algorithm for large scale density matrix renormalization group calculations,” *The Journal of chemical physics*, vol. 120, no. 7, pp. 3172–3178, 2004.
- [9] S. Yamada, M. Okumura, and M. Machida, “Direct extension of density-matrix renormalization group to two-dimensional quantum lattice systems: Studies of parallel algorithm, accuracy, and performance,” *J. Phys. Soc. Japan*, vol. 78, no. 9, pp. 1–5, 2009.
- [10] J. Vance, “Large-Scale Implementation of the Density Matrix Renormalization Group Algorithm,” Ph.D. dissertation, SISSA, 2017. [Online]. Available: <https://iris.sissa.it/handle/20.500.11767/68070>
- [11] M. F. Dolfi, “Dilute systems with the density matrix renormalization group: from continuum to lattice models,” Ph.D. dissertation, ETH Zürich, 2019.
- [12] A. Kantian, M. Dolfi, M. Troyer, and T. Giamarchi, “Understanding repulsively mediated superconductivity of correlated electrons via massively parallel density matrix renormalization group,” *Physical Review B*, vol. 100, no. 7, Aug 2019. [Online]. Available: <http://dx.doi.org/10.1103/PhysRevB.100.075138>
- [13] J. Brabec, J. Brandeys, K. Kowalski, S. Xantheas, Ö. Legeza, and L. Veis, “Massively parallel quantum chemical density matrix renormalization group method,” 2020. [Online]. Available: <http://arxiv.org/abs/2001.04890>
- [14] E. Solomonik, D. Matthews, J. R. Hammond, J. F. Stanton, and J. Demmel, “A massively parallel tensor contraction framework for coupled-cluster computations,” *Journal of Parallel and Distributed Computing*, vol. 74, no. 12, pp. 3176–3190, 2014.
- [15] R. Orús, “A practical introduction to tensor networks: Matrix product states and projected entangled pair states,” *Ann. Phys. (N. Y.)*, vol. 349, pp. 117–158, 2014. [Online]. Available: <http://dx.doi.org/10.1016/j.aop.2014.06.013>
- [16] —, “Advances on tensor network theory: symmetries, fermions, entanglement, and holography,” *Eur. Phys. J. B*, vol. 87, no. 11, 2014.
- [17] I. V. Oseledets, “Tensor-train decomposition,” *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.
- [18] Y. Zhang and E. Solomonik, “On stability of tensor networks and canonical forms,” 2020.
- [19] H. C. Jiang, H. Yao, and L. Balents, “Spin liquid ground state of the spin- $\frac{1}{2}$ square J_1 - J_2 Heisenberg model,” *Phys. Rev. B*, vol. 86, no. 2, pp. 1–14, 2012.
- [20] L. Wang and A. W. Sandvik, “Critical level crossings and gapless spin liquid in the square-lattice spin-1/2 $J_1 - J_2$ Heisenberg antiferromagnet,” *Phys. Rev. Lett.*, vol. 121, p. 107202, Sep 2018. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.121.107202>
- [21] T. Shirakawa, T. Tohyama, J. Kokalj, S. Sota, and S. Yunoki, “Ground-state phase diagram of the triangular lattice Hubbard model by the density-matrix renormalization group method,” *Phys. Rev. B*, vol. 96, no. 20, p. 205130, nov 2017. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevB.96.205130>
- [22] A. Szasz, J. Motruk, M. P. Zaletel, and J. E. Moore, “Chiral spin liquid phase of the triangular lattice hubbard model: A density matrix renormalization group study,” *Phys. Rev. X*, vol. 10, p. 021042, May 2020. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevX.10.021042>
- [23] J. Rincón, D. J. García, and K. Hallberg, “Improved parallelization techniques for the density matrix renormalization group,” *Comput. Phys. Commun.*, vol. 181, no. 8, pp. 1346–1351, 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.cpc.2010.03.018>
- [24] S. Yamada, M. Okumura, T. Imamura, and M. Machida, “Direct extension of the density-matrix renormalization group method toward two-dimensional large quantum lattices and related high-performance computing,” *Jpn. J. Ind. Appl. Math.*, vol. 28, no. 1, pp. 141–151, 2011.
- [25] S. Östlund and S. Rommer, “Thermodynamic Limit of Density Matrix Renormalization,” *Phys. Rev. Lett.*, vol. 75, no. 19, pp. 3537–3540, nov 1995. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevLett.75.3537>
- [26] E. R. Davidson, “The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices,” *J. Comput. Phys.*, vol. 17, no. 1, pp. 87–94, jan 1975. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/0021999175900650>
- [27] <http://itensor.org/>.
- [28] P. Springer, T. Su, and P. Bientinesi, “HPTT: A High-Performance Tensor Transposition C++ Library,” in *Proceedings of the 4th ACM SIGPLAN International Workshop on Libraries, Languages, and Compilers for Array Programming*, ser. ARRAY 2017. New York, NY, USA: ACM, 2017, pp. 56–62. [Online]. Available: <http://doi.acm.org/10.1145/3091966.3091968>
- [29] D. Kats and F. R. Manby, “Sparse tensor framework for implementation of general local correlation methods,” *The Journal of Chemical Physics*, vol. 138, no. 14, pp. –, 2013.
- [30] S. Chou, F. Kjolstad, and S. Amarasinghe, “Format abstraction for sparse tensor algebra compilers,” *Proceedings of the ACM on Programming Languages*, vol. 2, October 2018.
- [31] F. Kjolstad, P. Ahrens, S. Kamil, and S. Amarasinghe, “Tensor algebra compilation with workspaces,” *International Symposium on Code Generation and Optimization*, February 2019.
- [32] E. Epifanovsky, M. Wormit, T. Kuš, A. Landau, D. Zuev, K. Khistyayev, P. Manohar, I. Kaliman, A. Dreuw, and A. I. Krylov, “New implementation of high-level correlated methods using a general block-tensor library for high-performance electronic structure calculations,” *Journal of Computational Chemistry*, 2013.
- [33] S. Hirata, “Tensor Contraction Engine: Abstraction and automated parallel implementation of configuration-interaction, coupled-cluster, and many-body perturbation theories,” *The Journal of Physical Chemistry A*, vol. 107, no. 46, pp. 9887–9897, 2003.

- [34] P.-W. Lai, K. Stock, S. Rajbhandari, S. Krishnamoorthy, and P. Sadayappan, "A framework for load balancing of tensor contraction expressions via dynamic task partitioning," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2013, pp. 1–10.
- [35] J. Nieplocha, R. J. Harrison, and R. J. Littlefield, "Global Arrays: A nonuniform memory access programming model for high-performance computers," *The Journal of Supercomputing*, vol. 10, pp. 169–189, 1996.
- [36] 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*. ACM, 2019, pp. 46–56.
- [37] J. A. Calvin, C. A. Lewis, and E. F. Valeev, "Scalable task-based algorithm for multiplication of block-rank-sparse matrices," in *Proceedings of the 5th Workshop on Irregular Applications: Architectures and Algorithms*, 2015, pp. 1–8.
- [38] C. Peng, J. A. Calvin, F. Pavosevic, J. Zhang, and E. F. Valeev, "Massively parallel implementation of explicitly correlated coupled-cluster singles and doubles using TiledArray framework," *The Journal of Physical Chemistry A*, vol. 120, no. 51, pp. 10 231–10 244, 2016.
- [39] I. Sivkov, P. Seewald, A. Lazzaro, and J. Hutter, "Dbcsr: A blocked sparse tensor algebra library," *arXiv preprint arXiv:1910.13555*, 2019.
- [40] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet *et al.*, *ScaLAPACK users' guide*. Siam, 1997, vol. 4.
- [41] L. G. Valiant, "A bridging model for parallel computation," *Communications of the ACM*, vol. 33, no. 8, pp. 103–111, 1990.
- [42] D. B. Skillicorn, J. Hill, and W. F. McColl, "Questions and answers about BSP," *Scientific Programming*, vol. 6, no. 3, pp. 249–274, 1997.
- [43] E. Solomonik, M. Besta, F. Vella, and T. Hoefer, "Scaling betweenness centrality using communication-efficient sparse matrix multiplication," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '17. New York, NY, USA: ACM, 2017, pp. 47:1–47:14.
- [44] S. S. Gong, W. Zhu, D. N. Sheng, O. I. Motrunich, and M. P. Fisher, "Plaquette ordered phase and quantum phase diagram in the spin-1/2 square J_1 - J_2 Heisenberg model," *Phys. Rev. Lett.*, vol. 113, no. 2, pp. 1–5, 2014.
- [45] R. Haghshenas and D. N. Sheng, "U(1)-symmetric infinite projected entangled-pair states study of the spin-1/2 square J_1 - J_2 Heisenberg model," *Phys. Rev. B*, vol. 97, no. 17, pp. 1–10, 2018.
- [46] W. Y. Liu, S. Dong, C. Wang, Y. Han, H. An, G. C. Guo, and L. He, "Gapless spin liquid ground state of the spin-1/2 J_1 - J_2 Heisenberg model on square lattices," *Phys. Rev. B*, vol. 98, no. 24, pp. 1–5, 2018.
- [47] D. Poilblanc and M. Mambrini, "Quantum critical phase with infinite projected entangled paired states," *Phys. Rev. B*, vol. 96, p. 014414, Jul 2017. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevB.96.014414>

Appendix: Artifact Description/Artifact Evaluation

SUMMARY OF THE EXPERIMENTS REPORTED

We run ITensor v2 (8acf334e) and our DMRG implementation built using the Cyclops Tensor Framework (versions 8054b8e5 for list and sparse-dense and fa6c020d for sparse-sparse), the HPTT library (v1 9fecf0c41) as described in our work, on the NCSA's Blue Waters supercomputer. We also run on XSEDE and TACC's Stampede2 KNL nodes, using ITensor v2 (561a6147 - these differences shouldn't affect runtime for our application) and Cyclops Tensor Framework Version 7b394902 for sparse-dense and fa6c020d for sparse-sparse. Blue Waters data collection occurred during Nov/Dec 2019 for list and Dec 2019, Jan-April 2020. Stampede2 data collection occurred during Dec 2019 for sparse-dense and Feb-April 2020 for sparse-sparse

ARTIFACT AVAILABILITY

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

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

Data Artifact Availability: There are no author-created data artifacts.

Proprietary Artifacts: No author-created artifacts are proprietary.

Author-Created or Modified Artifacts:

Persistent ID: 10.5281/zenodo.3880568

Artifact name: Initial Scripts for Distributed-Memory

↪ DMRG via Sparse and Dense Parallel Tensor

↪ Contractions (without Tensor-Tools library)

BASELINE EXPERIMENTAL SETUP, AND MODIFICATIONS MADE FOR THE PAPER

Relevant hardware details: Blue Waters Cray XE6 nodes with 2x AMD 6276 and 64GB RAM (and HiMem nodes with 2x AMD 6276 and 128 GB RAM); Stampede2 KNL nodes with 1x Intel Xeon Phi 7250, 96GB DDR4 with 16GB MCDRAM configured in cache mode

Compilers and versions: GCC 6.3.0/8.2.0, Intel 18.0.2/18.0.3.222

Libraries and versions: Cyclops Tensor Framework, HPTT, lib-sci/18.12.1

Key algorithms: DMRG, Davidson

URL to output from scripts that gathers execution environment information.

<https://gist.github.com/ryanlevy/19ccf94af1b61c88d37>

↪ a9ddb2d2d8ca44