Massively Parallel Tensor Network State Algorithms on Hybrid CPU-GPU Based Architectures

arXiv:2305.05581

Simulation of quantum lattice models,
nuclear shell models and
ab initio quantum chemistry hand in hand

Andor Menczer and Örs Legeza

Strongly Correlated Systems "Lendület" Research Group
Wigner Research Centre for Physics, Budapest, Hungary

Eötvös Loránd University, Budapest, Hungary

Fachbereich Physik, Philipps-Universität Marburg, Germany

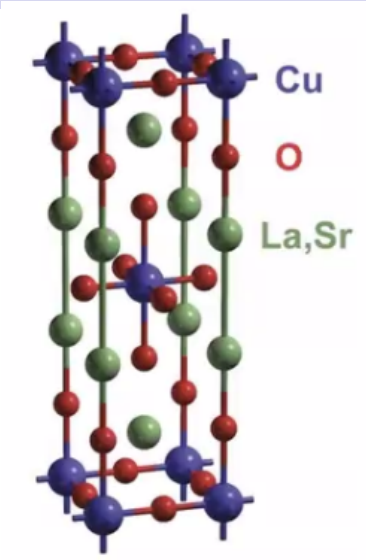
Institute for Advanced Study, Technical University of Munich, Germany

GPU-Day, Wigner RCP, 15.05.2023

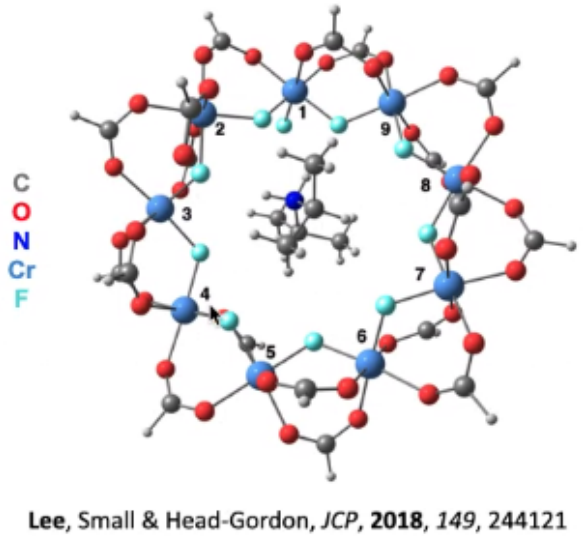
Topics to be covered

1. Motivations
2. Tensor product factorization (mathematically exact, loop free)
3. Novel solutions for efficient parallelization
   - Maze-Runners
   - Memory management: Data Dependency Trees
   - Strided Batched Matrix Multiplication for Summation
4. Benchmark results
   - CPU only limit
   - Hybrid CPU-multiGPU solution
   - Application of symmetries
5. Power consumption $\rightarrow$ Green DMRG
6. multiNode-multiGPU solution: Towards Exascale computing

Strong correlations between electrons $\rightarrow$ exotic materials



High T_c superconductors

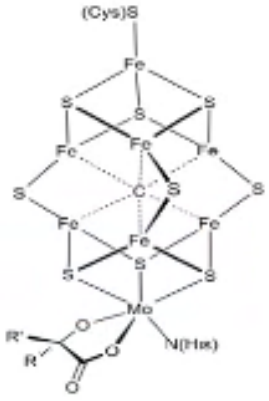


Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

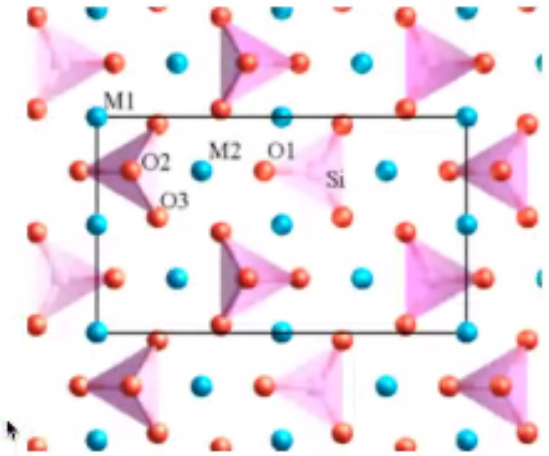
Single molecular magnets (SMM)

$$N_2 + 3\ H_2 \rightarrow 2\ NH_3$$

"FeMoco"



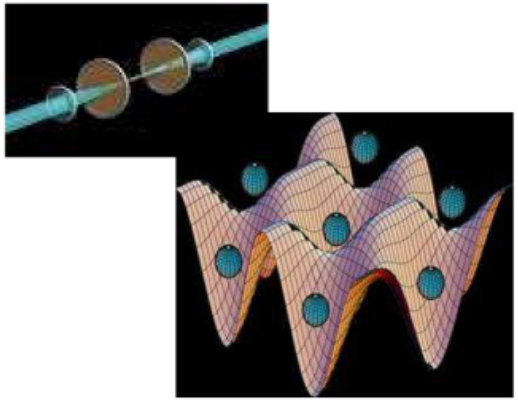
Nitrogen fixation



Battery technology

Experimental realizations: optical lattices

Numerical simulations: model systems



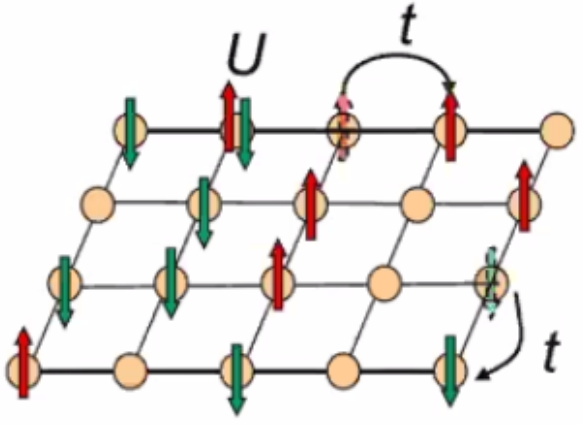
Atoms (represented as blue spheres) pictured in a 2D-optical lattice potential

Potential depth of the optical lattice can be tuned.

Periodicity of the optical lattice can be tuned.

Hubbard model: lattice model of interacting electron system

$$H = t \sum_{\langle i,j \rangle, \sigma} c^{\dagger}_{i,\sigma} c_{j,\sigma} + \frac{U}{2} \sum_{\sigma \neq \sigma'} \sum_{i} n_{i,\sigma} n_{i,\sigma'}$$



t hopping amplitude
U on-site Coulomb interaction
$\sigma \in \uparrow, \downarrow$ spin index

Classical or quantum computers?

Simulation of quantum systems on High Performance Comuting (HPC) infrastructure

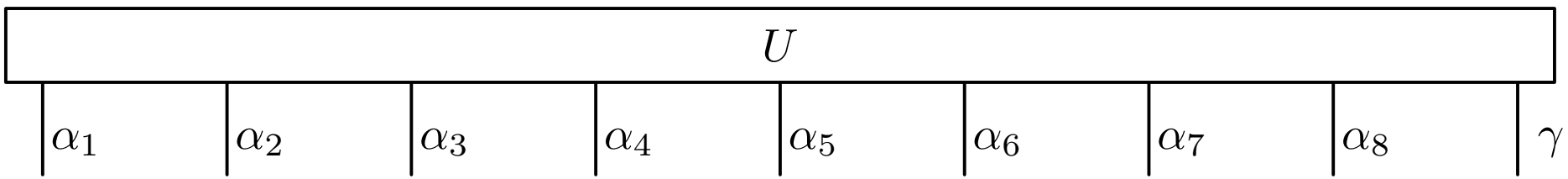
- Numerical (classical) simulation has become an important part of both basic and applied research.
- Enormous progress in High-Performance Computing (HPC) and the development of numerical algorithms $\rightarrow$ simulation of physical, chemical, biological, economical and ecological systems etc.
- For interacting quantum systems, however, a fundamental limitation emerges: the so-called curse of dimensionality.
- Computational effort scales exponentially with the dimension of the Hilbert space.
- There is no known universal "fix" for this problem.
- It is the interplay of quantum and classical simulation and the delicate divide between them that is the focus of massively parallelized tensor network state (TNS) algorithms designed for HPC infrastructures

Tensor product approximation

State vector of a quantum system in the discrete tensor product spaces

$$|\Psi_\gamma\rangle = \sum_{\alpha_1=1}^{n_1} \dots \sum_{\alpha_d=1}^{n_d} U(\alpha_1, \dots, \alpha_d, \gamma)\, |\alpha_1\rangle \otimes \cdots \otimes |\alpha_d\rangle \in \bigotimes_{i=1}^{d} \Lambda_i := \bigotimes_{i=1}^{d} \mathbf{C}^{n_i} \ ,$$

where $span\{|\alpha_i\rangle : \alpha_i = 1, \dots, n_i\} = \Lambda_i = \mathbf{C}^{n_i}$ and $\gamma = 1, \dots, m$.



- α is called 'physical' leg
- In a spin-1/2 model $\alpha_i \in \{\downarrow, \uparrow\}$.
- In a spin-1/2 fermionic model $\alpha_i \in \{0, \downarrow, \uparrow, \uparrow\downarrow\}$.

$\dim \mathcal{H}_d = \mathcal{O}(n^d)$ Curse of dimensionality!

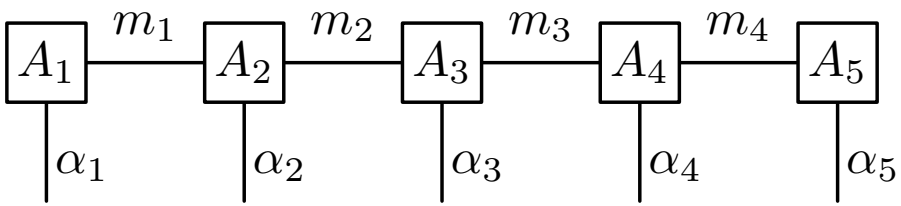
$\rightarrow$ need efficient data-sparse representation

Matrix product state (MPS) representation

The tensor U is given element-wise as

$$U(\alpha_1, \ldots, \alpha_d) = \sum_{m_1=1}^{r_1} \cdots \sum_{m_{d-1}=1}^{r_{d-1}} A_1(\alpha_1, m_1) A_2(m_1, \alpha_2, m_2) \cdots A_d(m_{d-1}, \alpha_d).$$

We get d component tensors of order 2 or 3.



A tensor of order 5 in Matrix Product State (MPS) representation also know as Tensor Train (TT). This yields a chain of matrix products:

$$U(\alpha_1, \ldots, \alpha_d) = \mathbf{A}_1(\alpha_1)\mathbf{A}_2(\alpha_2) \cdots \mathbf{A}_{d-1}(\alpha_{d-1})\mathbf{A}_d(\alpha_d)$$

with $[\mathbf{A}_i(\alpha_i)]_{m_{i-1}, m_i} := A_i(m_{i-1}, \alpha_i, m_i) \in \mathbb{C}^{r_{i-1} \times r_i}$.

Controlled truncation on m_i.

Redundancy:
$U(\alpha_1, \ldots, \alpha_d) = \mathbf{A}_1(\alpha_1)\mathbf{G}\mathbf{G}^{-1}\mathbf{A}_2(\alpha_2) \cdots \mathbf{A}_{d-1}(\alpha_{d-1})\mathbf{A}_d(\alpha_d)$

Affleck, Kennedy, Lieb Tagasaki (87); Fannes, Nachtergale, Werner (91), White(92), Römmer & Ostlund (94), Vidal (03); Verstraete(04); Oseledets & Tyrtyshnikov, 2009

TNS/DMRG provide state-of-the-art results in many fields

- General form of the Hamiltonian with one- and two-body interactions

$$\mathcal{H} = \sum_{ij\alpha\beta} T_{ij}^{\alpha\beta} c_{i\alpha}^{\dagger} c_{j\beta} + \frac{1}{2} \sum_{ijkl\alpha\beta\gamma\delta} V_{ijkl}^{\alpha\beta\gamma\delta} c_{i\alpha}^{\dagger} c_{j\beta}^{\dagger} c_{k\gamma} c_{l\delta} + \dots ,$$

- i, j, k, l label modes, $\alpha, \beta, \dots$ are color indices
- T_{ij} kinetic and on-site terms, V_{ijkl} two-particle scattering

$$V_{ijkl} = \int d^3x_1 d^3x_2 \Phi_i^*(\vec{x}_1) \Phi_j^*(\vec{x}_2) \frac{1}{\vec{x}_1 - \vec{x}_2} \Phi_k(\vec{x}_2) \Phi_l(\vec{x}_1)$$

- with appropriate choice of one-particle basis
- (DMRG): $\mathcal{O}(M^3 d^3) + \mathcal{O}(M^2 d^4)$
- Major aim is to obtain the desired eigenstates and measurable quantities

  • Symmetries: Abelian and non-Abelian quantum numbers, double groups, complex integrals, quaternion sym. etc

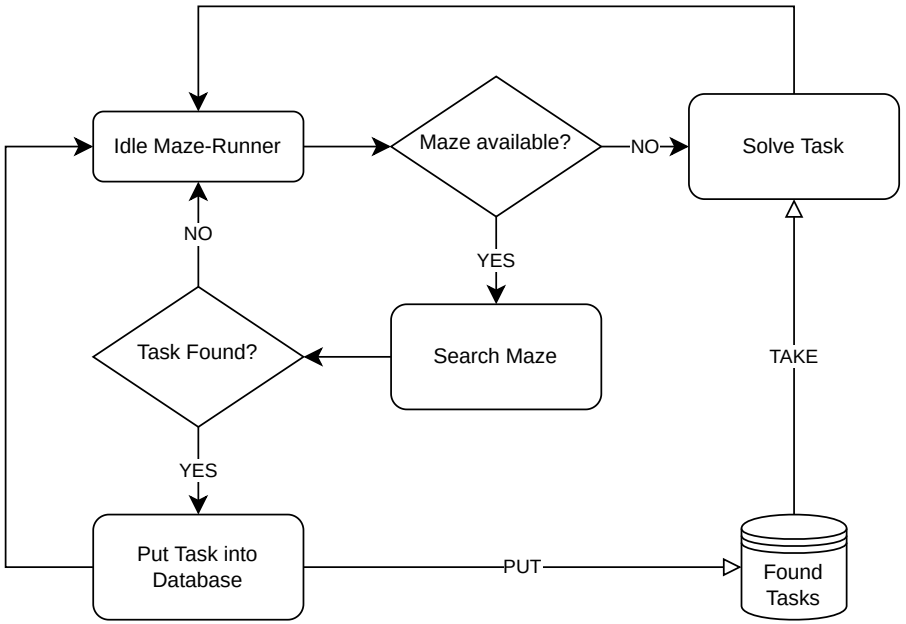
  • # of block states: 1 000 – 60 000. Size of Hilbert space up to 10^8.

  • In ab inito DMRG the CAS size is: 70 electrons on 70 orbitals.

  • 1-BRDM and 2-BRDM, finite temperature, dynamics

Efficient task processig: Maze-Runners

- In traditional producer-consumer models threads are casted into disjoint sets labeled as *producers* and *consumers*.
- Ideally, producer and consumer threads can run in parallel
- Instead of implementing high-complexity dynamic scheduling systems relying on task specific optimizations.

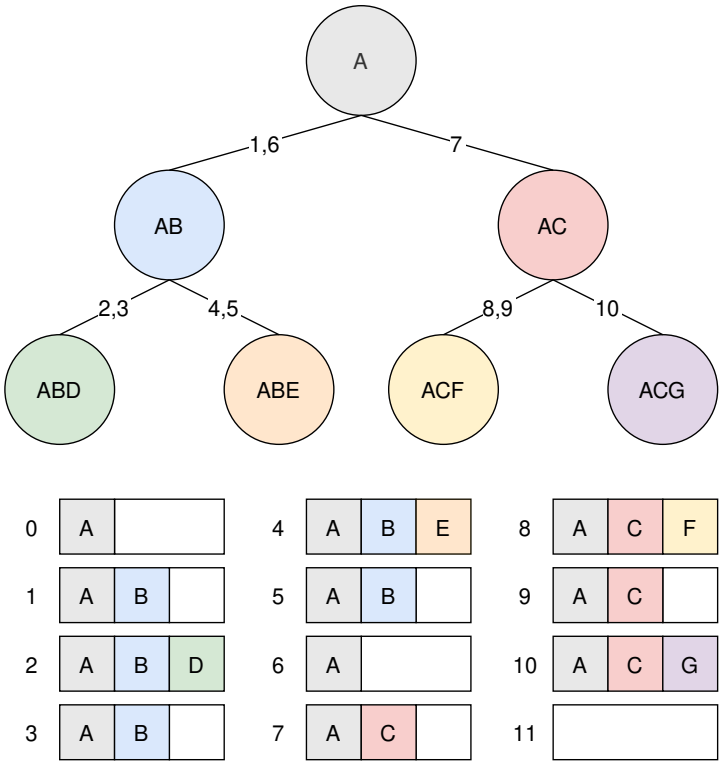


Life Cycle of a Maze-Runner Thread.

- Threads can be fed with tasks from any level of recursion.
- This ensures a magnitude of thread utilization not feasable with classical producer-consumer based pipelines.

Memory management: Data Dependency Trees

- Naive solution to memory management is to store all required data in memory at all times
- Usually datasets exceed the size of allocatable memory.
- Aim: IO to be hideable behind the parallely running computation



Buffering while Traversing the Data Dependency Tree.

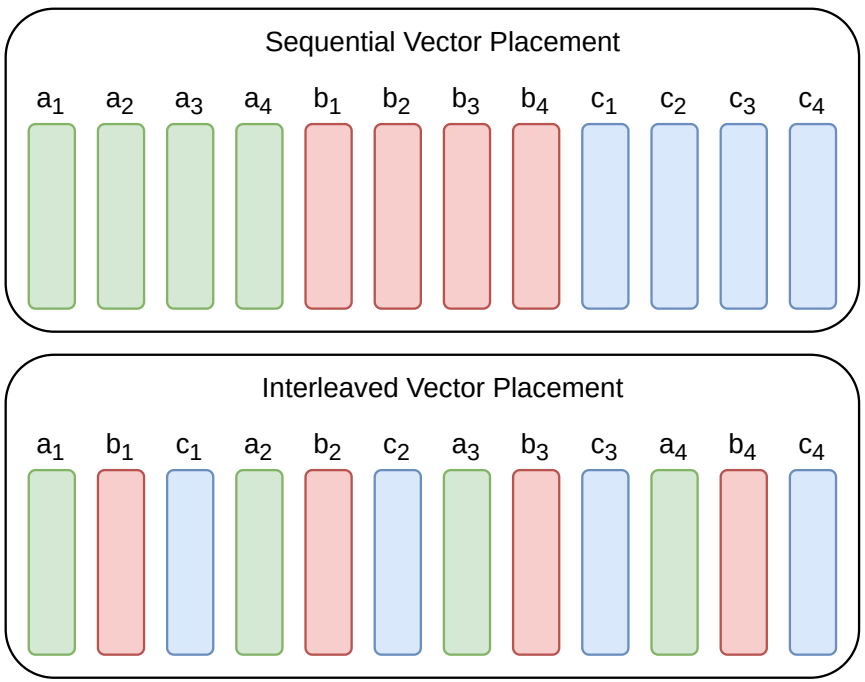
The numbers represent the order in which the vertices are visited.

The arrays show the buffer's content for each step.

- Gap-free, sequential write and read operations, no allocations and deallocations are required in the traditional sense.

Strided Batched Matrix Multiplication for Summation

- SIMD workloads have a tendency to perform poorly when bombarded with a high amount of small jobs.
- For aggregation of matrix multiplications, both Intel and NVIDIA has implemented solutions: Batched GEMM.



Vectorized output of a Strided Batched GEMM operation.

Normally, output vectors belonging to the same matrix are in a sequential order (top).

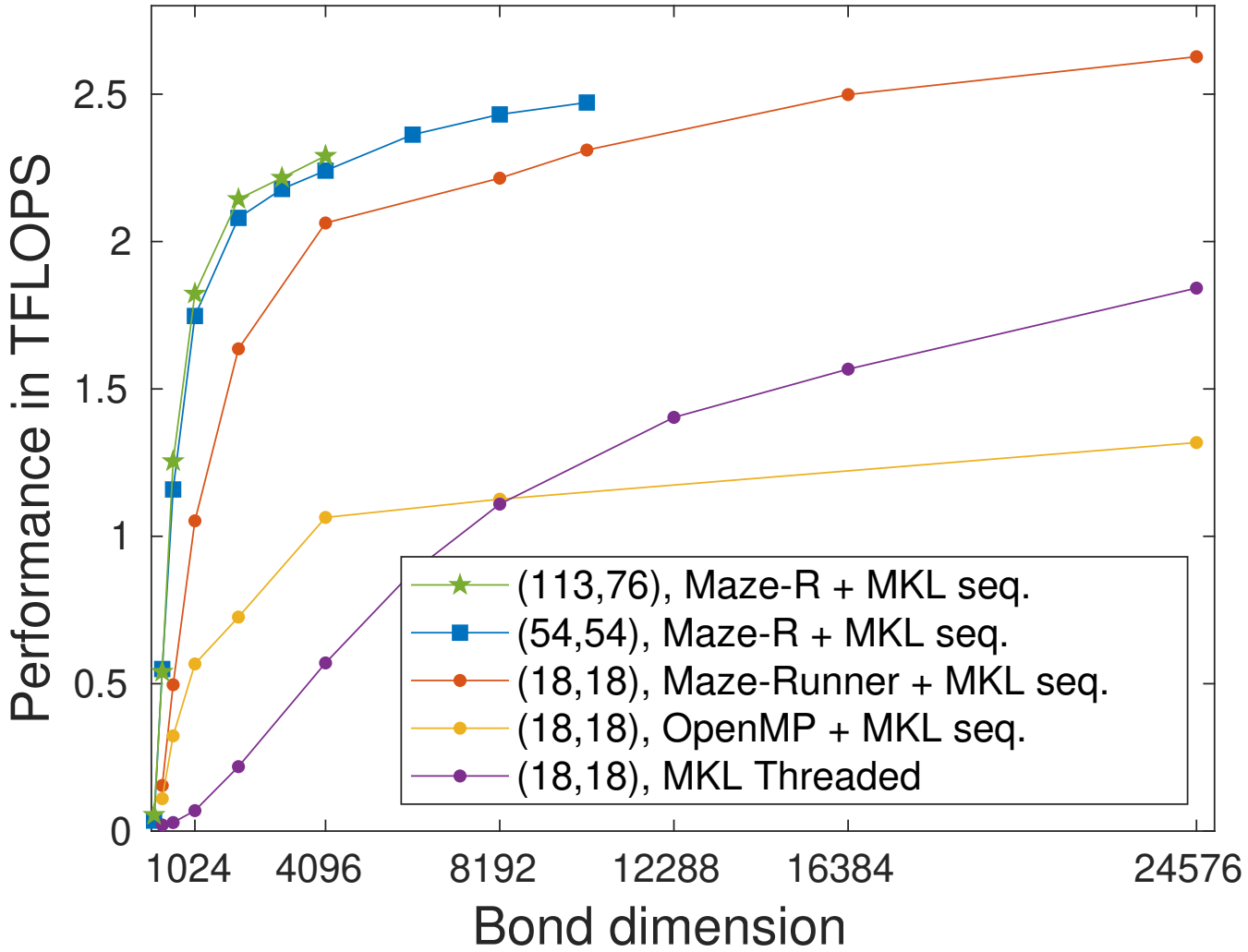
Interleaving the vectors of different matrices (bottom) is possible by altering the leading dimensions and stride values of the output matrices.

- We can perform batched type chained matrix multiplications without sum reduction at the end.

Properties of the TNS/DMRG algorithms

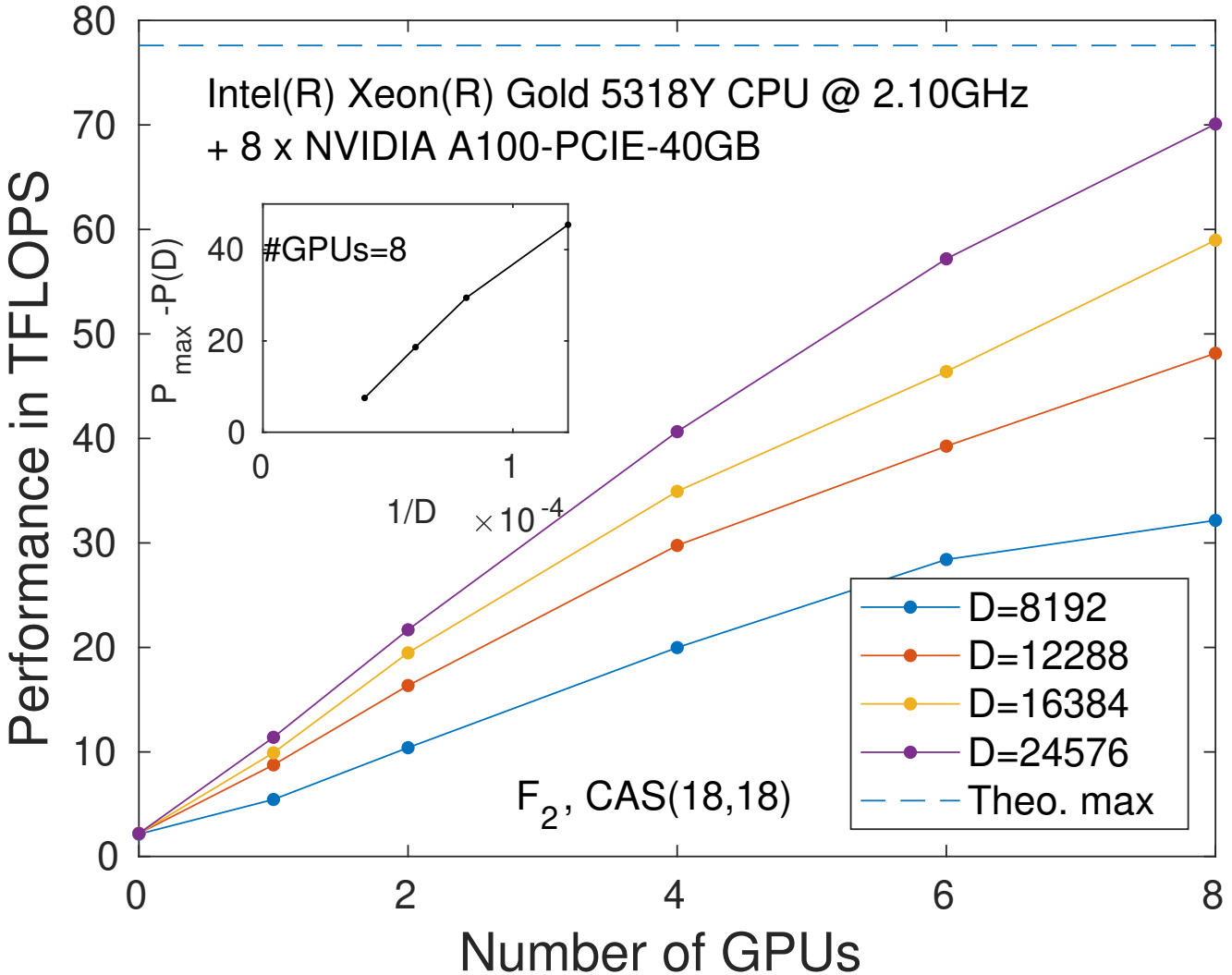
- Key aspect of TNS/DMRG: exponential scaling can be reduced to a polynomial form.
- Underlying tensor and matrix algebra can be organized into several million of independent operations (tasks).
- Dense matrix operations are performed in parallel according to the so-called quantum number decomposed representations (sectors).
- Full matrices, denoted as DMRG bond dimension, D, determines the accuracy of the calculations.
- The overall scaling of the DMRG is D^3N^4 where N stands for the system size.
- The memory requirement is proportional to D^2N^2.
- The iterative diagonalization of the effective Hamiltonian usually accounting for 85% of the total execution time.
- The renormalization step is responsible for 10% of the total execution time.

CPU only limit (for CAS(113,76) $\dim \mathcal{H} = 2.88 \times 10^{36}$)



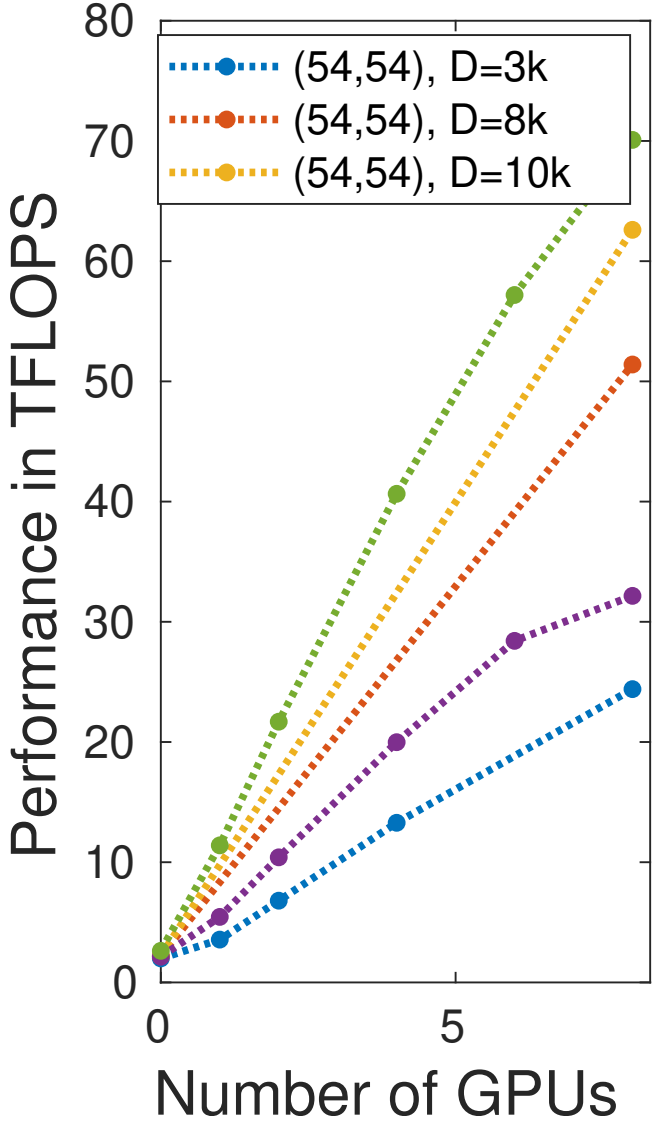
Performance measured in TFLOPS for the F_2 and FeMoco chemical systems for CAS(18,18) and CAS(54,54) orbitals spaces, respectively, as a function of the DMRG bond dimension on a dual Intel(R) Xeon(R) Gold 5318Y CPU system with 2×24 physical cores running at 2.10 Ghz.

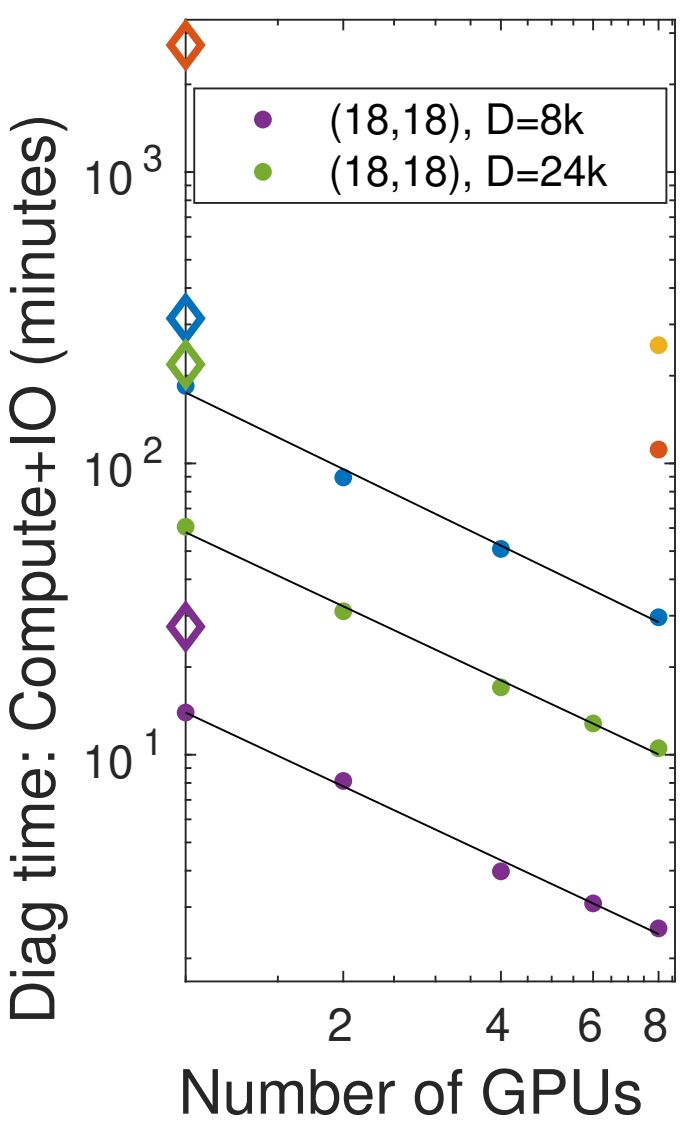
CPU-multiGPU



Performance measured in TFLOPS for the F_2 molecule, corresponding to CAS(18,18) as a function of the number of GPU devices for various fixed DMRG bond dimension values. Calculations have been performed on a dual Intel(R) Xeon(R) Gold 5318Y CPU system with 2x24 physical cores

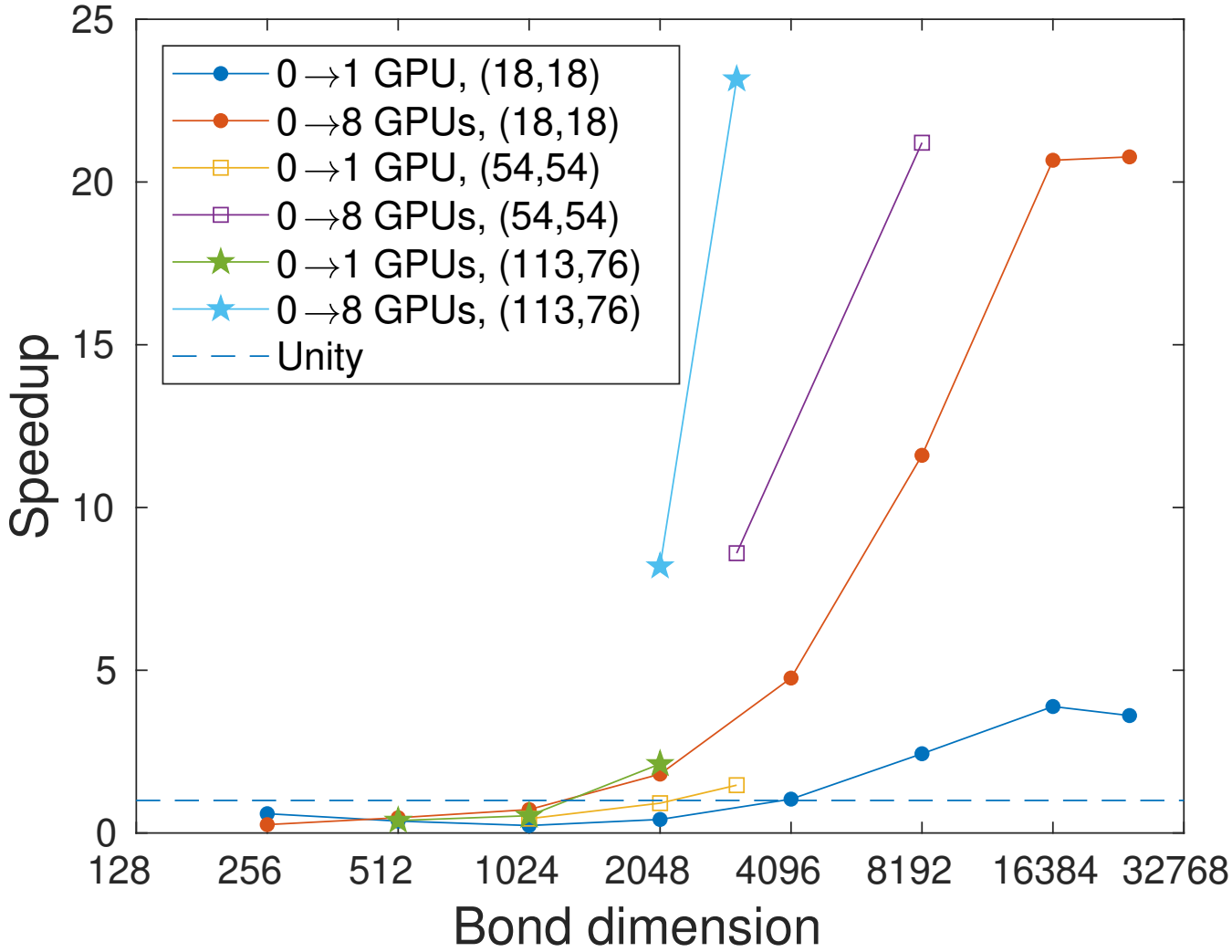
CPU-multiGPU: Perfomance and time vs number of GPUs





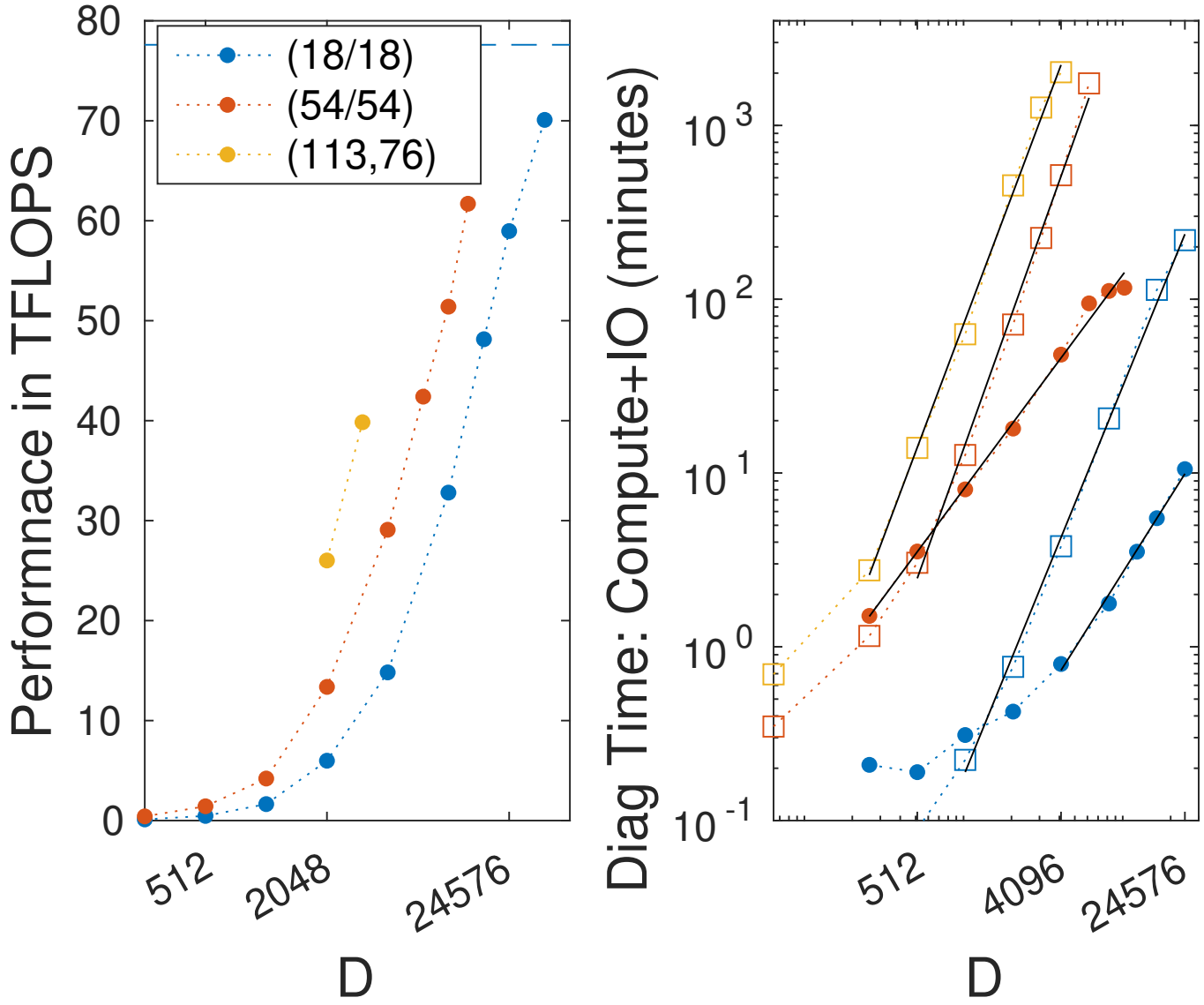
- Exponents: -0.8438, -0.8416 and -0.8729 (Ideal would be -1).

Speedup for selected data sets as a function of D



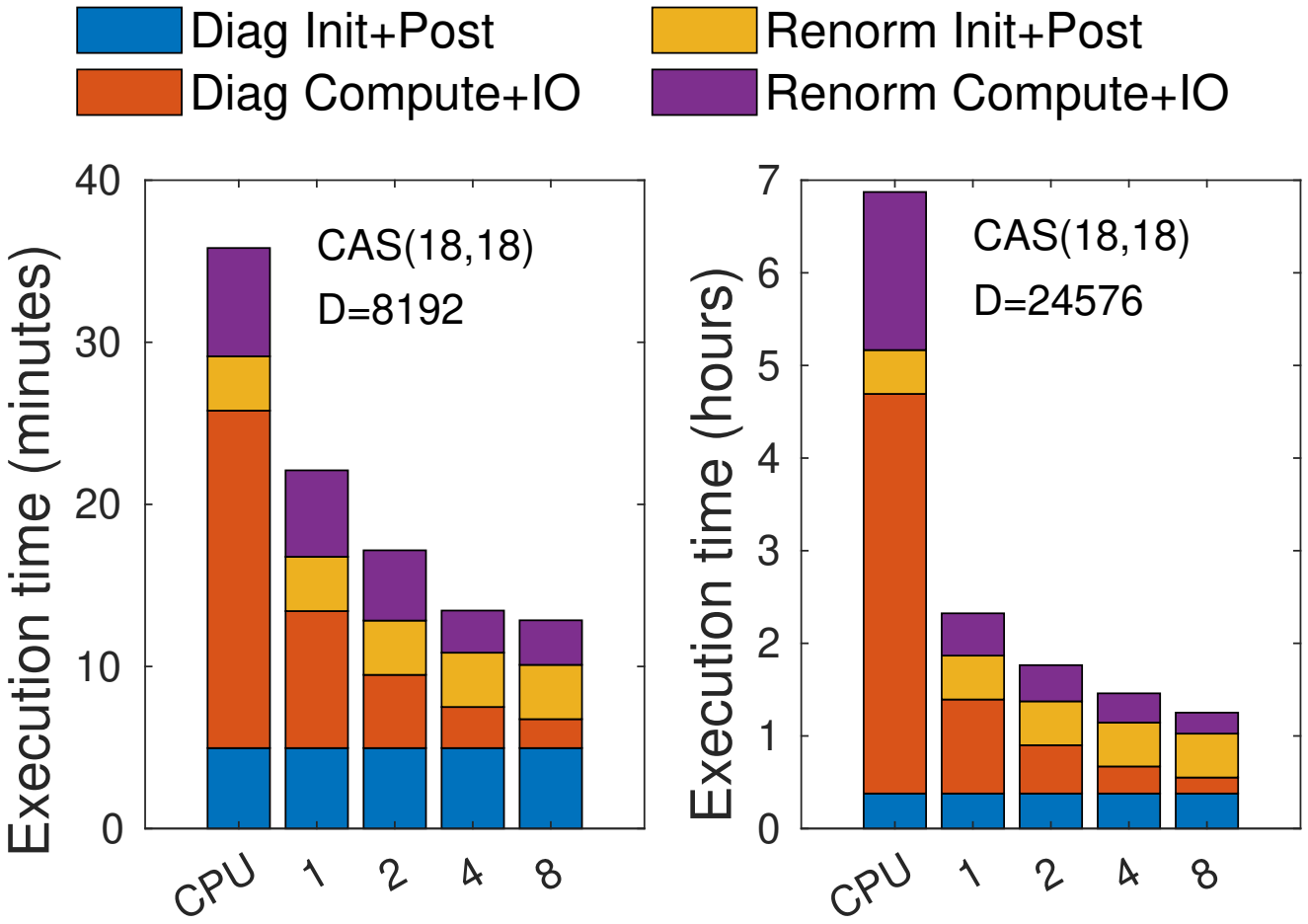
The dashed line at unity could be used to determine the minimal bond dimension for which the system dependent GPU accelerated solution becomes faster than the CPU only limit.

Performance and related total time for the 8 GPU accelerated diagonalization procedure: $D^3 \rightarrow D^{2.3} \rightarrow D^{1.25}$



Renormalization procedure: forming q-dits with "dits" > "bits"

- Execution time is split into two parts: The CPU-only and GPU accelerated parts — the latter together with its respective IO overhead
- Main bottleneck can be identified as the D2H CUDA kernels responsible for the retrieval of computed data on the devices



Power consumption of the TNS calculations $\rightarrow$ Green DMRG

- The power consumption of the TNS calculations are becoming one of the most important question due to high energy demands and costs.
- The thermal design power (TDP) for 2 $\times$ Intel(R) Xeon Gold 5318Y CPU is 2 $\times$ 165 Watts $\rightarrow$ 2.5 TFLOPS would lead to $\approx$ 7.5 GFLOPS/Watt.
- For an NVIDIA A100-PCIE-40GB device the TDP is 250 Watts.
- For our 8 card accelerated hybrid algorithm with 70 TFLOPS performance results in $\approx$ 30.04 GFLOPS/Watt.
- For a given calculation the cost of the energy demand arising from the processors can be reduced to one quarter of the original consumption.
- The energy consumption of the GPU devices fluctuates significantly, thus even a better ratio can be obtained.

Utilization of non-Abelian symmetries and more general tensor topologies

- Using more symmetries increases the number of sectors $\rightarrow$ big increase in the number of independent tasks
- For more general networks, the number of the independent tasks increases tremendously.
- The optimal number of GPU devices has not been reached at eight cards $\rightarrow$. An MPI based multiNode-multiGPU version is expected to further boost performance, introducing DMRG into the world of petascale computing.

Conclusions

- Using novel algorithmic developments we could reach maximum performance on CPU (2.5 TFLPS)
- Obtained linear scaling with number of GPUs for the hybrid CPU-multiGPU solution
- Almost reached theo. max performance with 8 cards (70 TFLOPS)
- Power law scaling with exponent $\simeq -0.87$: doubling GPUs halves execution time
- Exponent reduction: $D^3 \rightarrow D^{2.3} \rightarrow D^{1.25}$
- Power consumption: reduced by a fator of four or more
- Current work: multiNode-multiGPU $\rightarrow$ petascale computing

Supports: Lendület grant of the Hungarian Academy of Sciences, the Hungarian National Research, Development and Innovation Office, Quantum Information National Laboratory of Hungary, TKP-V04, Hans Fischer Senior Fellowship programme (IAS-TUM, Germany), SPEC, DOE, (PNNL, USA) Wigner Scientific Computing Laboratory (WSCLAB), the Eötvös Loránd University, the Governmental Information-Technology Development Agency supercomputer Komondor. HPE Apollo Hawk HLRS, Stuttgart.