# Novel approaches in material science and ab initio quantum chemistry via massively parallel tensor network state methods on Hybrid CPU-GPU based HPC architectures

## Boosting effective performance via Wigner-Eckhart theorem

Örs Legeza

in collaboration with

Gero Friesecke, Gergely Barcza, Andor Menczer

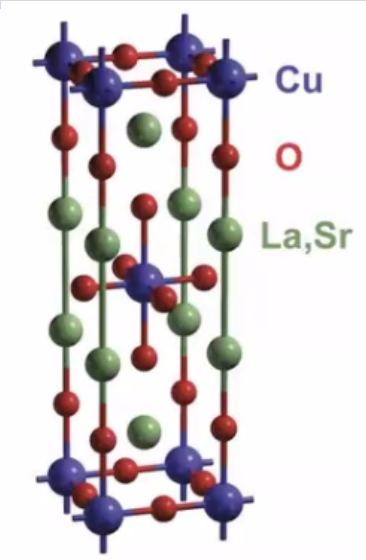
Strongly Correlated Systems "Lendület" Research Group
Wigner Research Centre for Physics, Budapest, Hungary

Institute for Advanced Study, Technical University of Munich, Germany

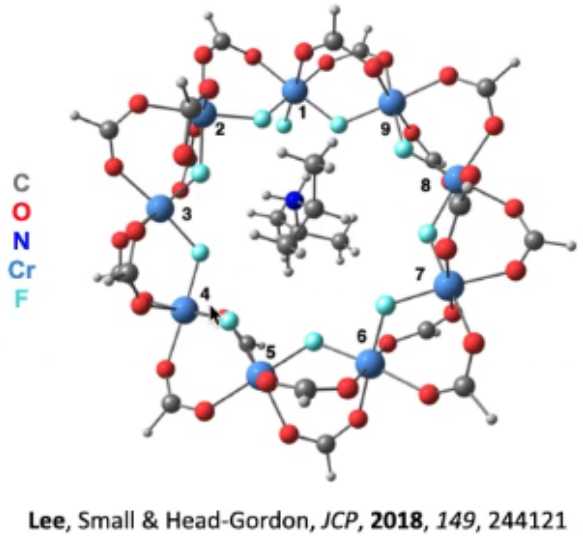
Wigner 121 Scientific Symposium

Budapest, 18-20 September 2023

# Strong correlations between electrons $\rightarrow$ exotic materials




High $T_c$ superconductors

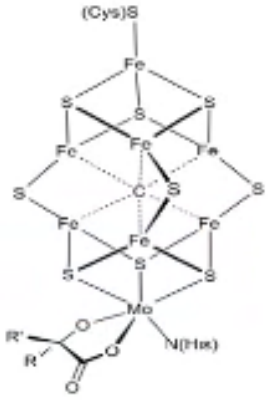



Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

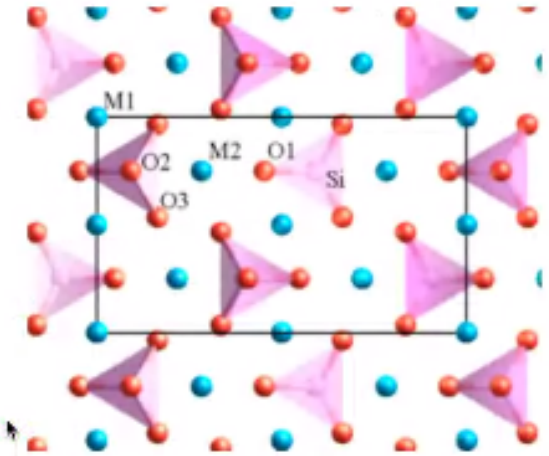
Single molecular magnets (SMM)

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

"FeMoco"



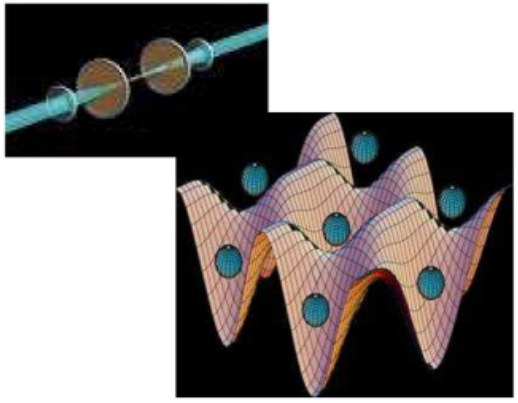

Nitrogen fixation




Battery technology

# Experimental realizations: optical lattices

# Numerical simulations: model systems



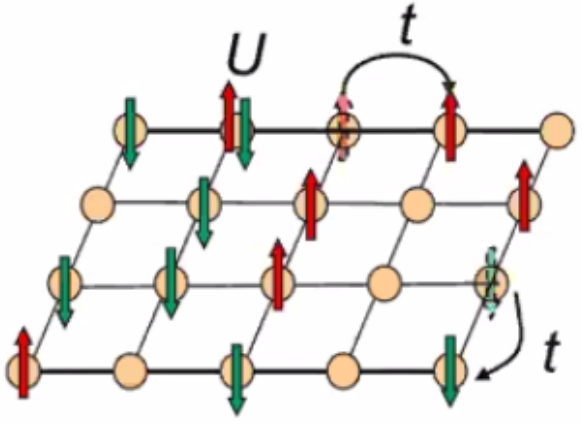
Atoms (represented as blue spheres) pictured in a 2D-optical lattice potential

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 via classical computation

Problem:

- Ever growing demand for efficient simulation of quantum systems via classical computation
- A fundamental limitation emerges: the so-called curse of dimensionality, that is, the computational effort scale exponentially with the system size

Solution:

- 1) Searching for algorithms to reduce the exponential scaling by controlled approximations
- 2) Fully taking advantage of modern High-Performance Computing (HPC) infrastructures

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

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

- $T_{ij}$ kinetic and on-mode terms, $V_{ijkl}$ two-particle scatterings
- We consider usually lattice models in real space (DMRG)
- In quantum chemistry modes are electron orbitals (QC-DMRG)
- In UHF QC spin-dependent inetractions (UHF-QCDMRG)
- In relativistic quantum chemistry modes are spinors (4c-DMRG)
- In nuclear problems modes are proton/neutron orbitals (JDMRG)
- In k-space modes are momentum eigenstates (k-DMRG)
- For particles in confined potential modes $\rightarrow$ Hermite polynoms
- Major aim: to obtain the desired eigenstates of $\mathcal{H}$.

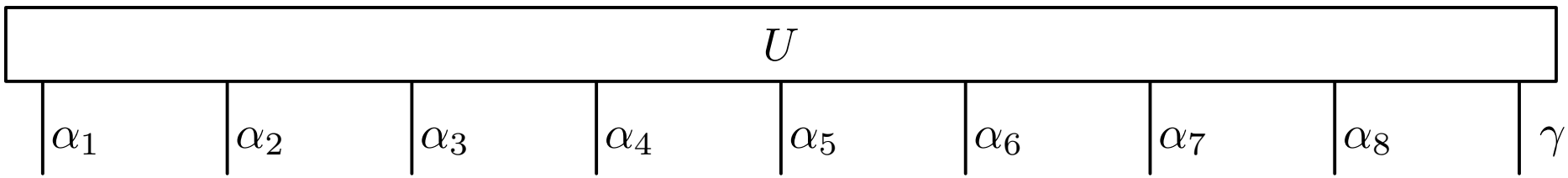
- Symmetries: Abelian and non-Abelian quantum numbers, double groups, complex integrals, quaternion sym. etc
- # of block states: 1 000 – 60 000. Size of Hilbert space up to $10^8$.
- In ab inito DMRG the CAS size is: 70 electrons on 70 orbitals.
- 1-BRDM and 2-BRDM, finite temperature, dynamics
- Massively parallel implementations CPU/GPU$\rightarrow$ exascale on HPC

# Tensor product approximation

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

$$|\Psi_\gamma\rangle = \sum_{\alpha_1=1}^{q_1} \dots \sum_{\alpha_d=1}^{q_d} U(\alpha_1, \dots, \alpha_d, \gamma)\, |\alpha_1\rangle \otimes \cdots \otimes |\alpha_d\rangle \in \bigotimes_{i=1}^{d} \Lambda_i := \bigotimes_{i=1}^{d} \mathbf{C}^{q_i} \ ,$$

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




- In a spin-1/2 model $\alpha_i \in \{\downarrow, \uparrow\}$.
- In a spin-1/2 fermionic model $\alpha_i \in \{0, \downarrow, \uparrow, \uparrow\downarrow\}$.

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

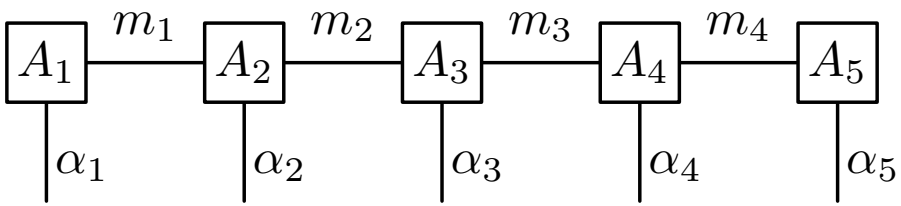
- We seek to reduce computational costs by parametrizing the tensors in some data-sparse representation.

# Matrix product state (MPS) representation / DMRG / TT

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

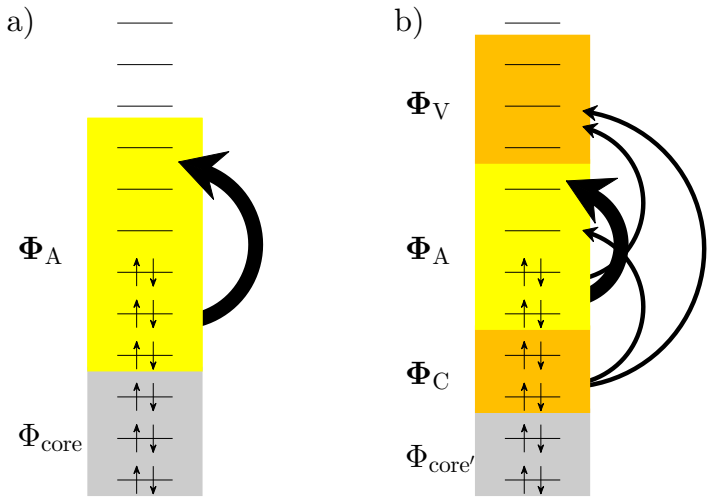
Controlled truncation on $m_i$.

Redundancy:

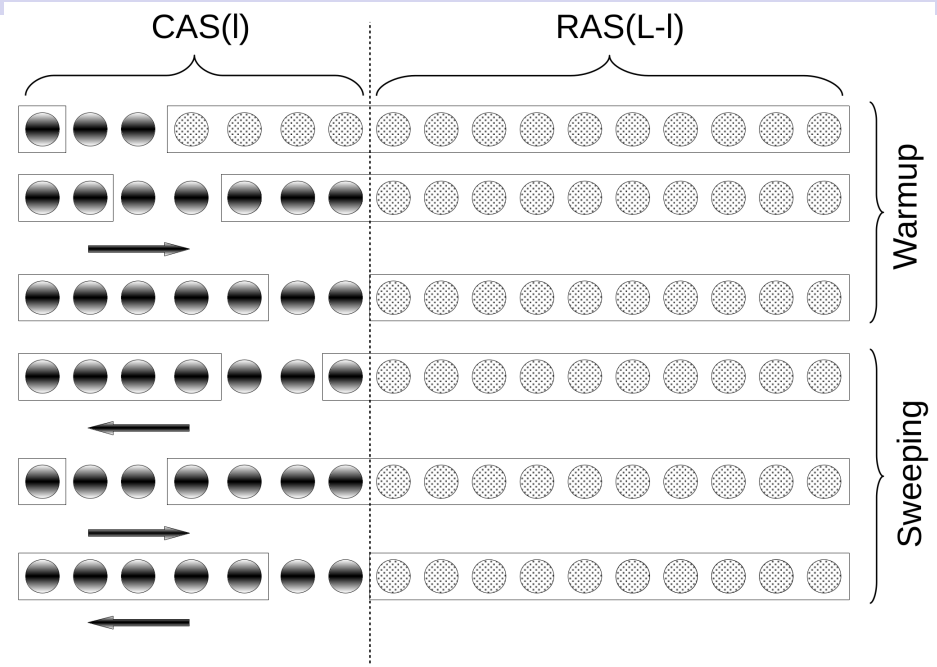
$$U(\alpha_1, \ldots, \alpha_d) = \mathbf{A}_1(\alpha_1)\mathbf{G}\mathbf{G}^{-1}\mathbf{A}_2(\alpha_2) \cdots \mathbf{A}_{d-1}(\alpha_{d-1})\mathbf{A}_d(\alpha_d)$$

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

# Restricted active space DMRG Barcza,Werner,Zaránd,Ö.L.,Szilvási(2021)



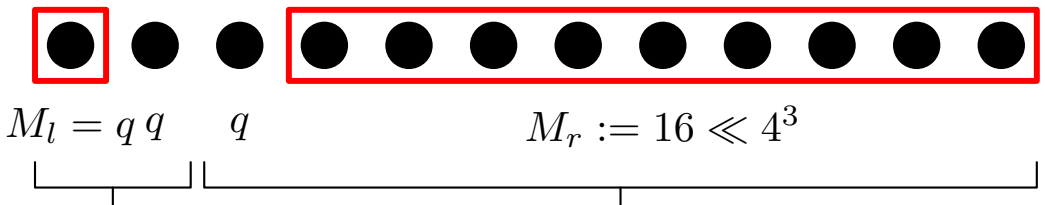

Schematic illustration of the CAS and RAS concepts.




DMRG-RAS scheme

• In the RAS scheme, in addition to active orbitals some virtual (V) and core (C) orbitals can also be excited with restrictions: the maximal number of particle excitations in these orbitals is $r$.

• Implementation through the dynamically extended active space (DEAS) procedure. ÖL, J. Sólyom, 2003, (similar appr. by Larsson et al 2022)



# Ground state energy of $C_2$ frozen-core cc-pVTZ (L=58)

- DMRG-RAS is an embedding methodi, i.e.,

$$H = \underbrace{PHP}_{H_{\mathrm{CAS}\to\mathrm{CAS}}} + \underbrace{QHP}_{H_{\mathrm{CAS}\to\mathrm{RAS}}} + \underbrace{PHQ}_{H_{\mathrm{RAS}\to\mathrm{CAS}}} + \underbrace{QHQ}_{H_{\mathrm{RAS}\to\mathrm{RAS}}}$$

| method | energy (Ha) | $\Delta_{\mathrm{E}}$ (%) |
|---|---|---|
| CI-SDTQ | -75.7765 | 97.8 |
| CC-SD[a] | -75.7496 | 90.8 |
| CC-SD(T)[a] | -75.7832 | 99.5 |
| CC-SDT[a] | -75.7810 | 99.0 |
| CC-SDTQ[a] | -75.7845 | 99.9 |
| NEVPT2(8)[a] | -75.7540 | 91.9 |
| RAS-SD-DMRG(8,$M = 5051$) | -75.7704 | 96.2 |
| RAS-SD-DMRG(14,$\chi = 10^{-6}$) | -75.7809 | 99.0 |
| RAS-SD-DMRG(18,$\chi = 10^{-6}$) | -75.7836 | 99.6 |
| CAS-DMRG($\chi = 10^{-6}$) | -75.7849 | 99.9 |
| CAS-DMRG($M = 4096$) | -75.7850 | 100.0 |

- Similar performance measured along the PES for $d \leq 5$.
- Spectroscopic constants agree with FCIQMC data up to 3 digits.

# Rigorous mathematical analysis of the error dependence

**Friesecke, Barcza, Ö.L. (2022)**

N-electron Hilbert space for the DMRG-RAS method:

$$\mathcal{H}(\ell, k) = \mathcal{H}_{\mathrm{CAS}}(\ell) \bigoplus \mathcal{H}_{\mathrm{RAS}}(L - \ell, k)$$

$$E^0(\ell, k) = \min_{\Psi \in \mathcal{H}(\ell,k) \,:\, \langle \Psi, \Psi \rangle = 1} \langle \Psi,\, H\Psi \rangle,$$

partitioning of the full Hamiltonian into a reference Hamiltonian associated with the CAS energy and a remainder:

$$\begin{aligned} H &= H_0 + H' \text{ with} \\ H_0 &= PHP + (E_0 + \Delta)Q \\ H' &= H - PHP - (E_0 + \Delta)Q \end{aligned}$$

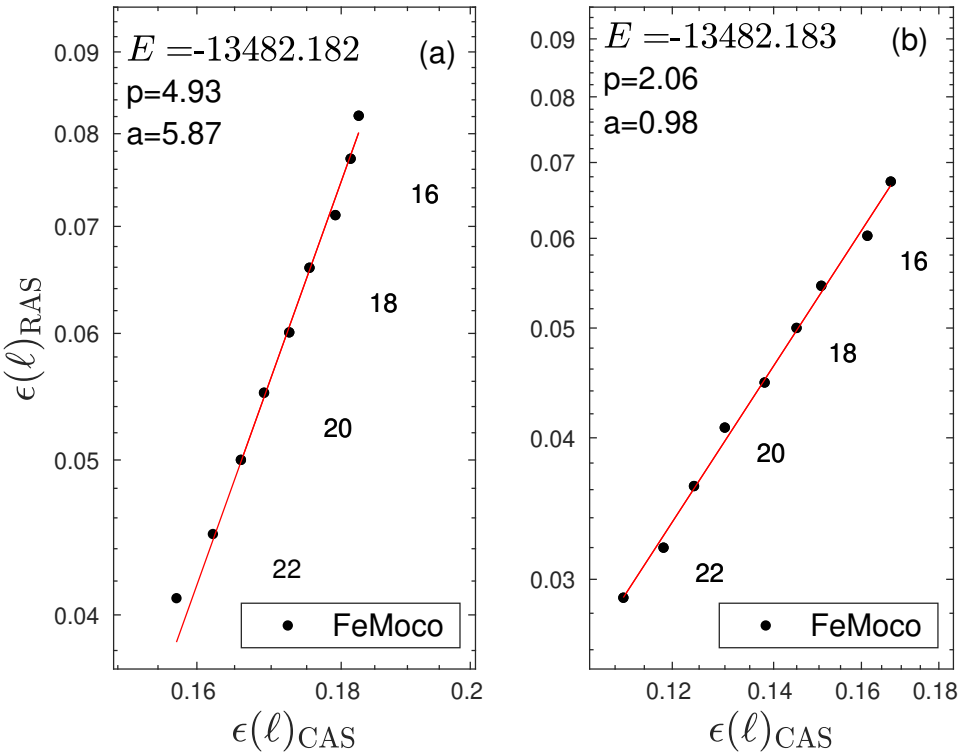
where $P$ is the projector of $\mathcal{H}$ onto the CAS Hilbert space $\mathcal{H}_{CAS}$, $Q = I - P$ is the projector onto the RAS Hilbert space, $E_0$ is the CAS ground state energy, i.e.

$$E_0 = E^0_{\mathrm{CAS}}(\ell),$$

and $\Delta > 0$ is a parameter to be chosen later.

| Method | Ground state energy |
|---|---|
| i-FCIQMC-RDME | -13482.17495(4) |
| i-FCIQMC-PT2 | -13482.17845(40) |
| sHCI-VAR | -13482.16043 |
| sHCI-PT2 | -13482.17338 |
| DMRG | -13482.17681 |
| DMRG(D=8192) | -13482.1718 |
| DMRG(D=10240,NO) | -13482.1754 |
| RAS(23) | -13482.1421 |
| RAS(23,NO) | -13482.1544 |

Non-extrapolated ground state energies obtained by various methods for the FeMoco in CAS(54,54) orbital space.




(a) Result of the DMRG-RAS-X for the FeMoco for the model space taken from Ref. Reiher(2007).

(b) The same but for the natural orbital basis.

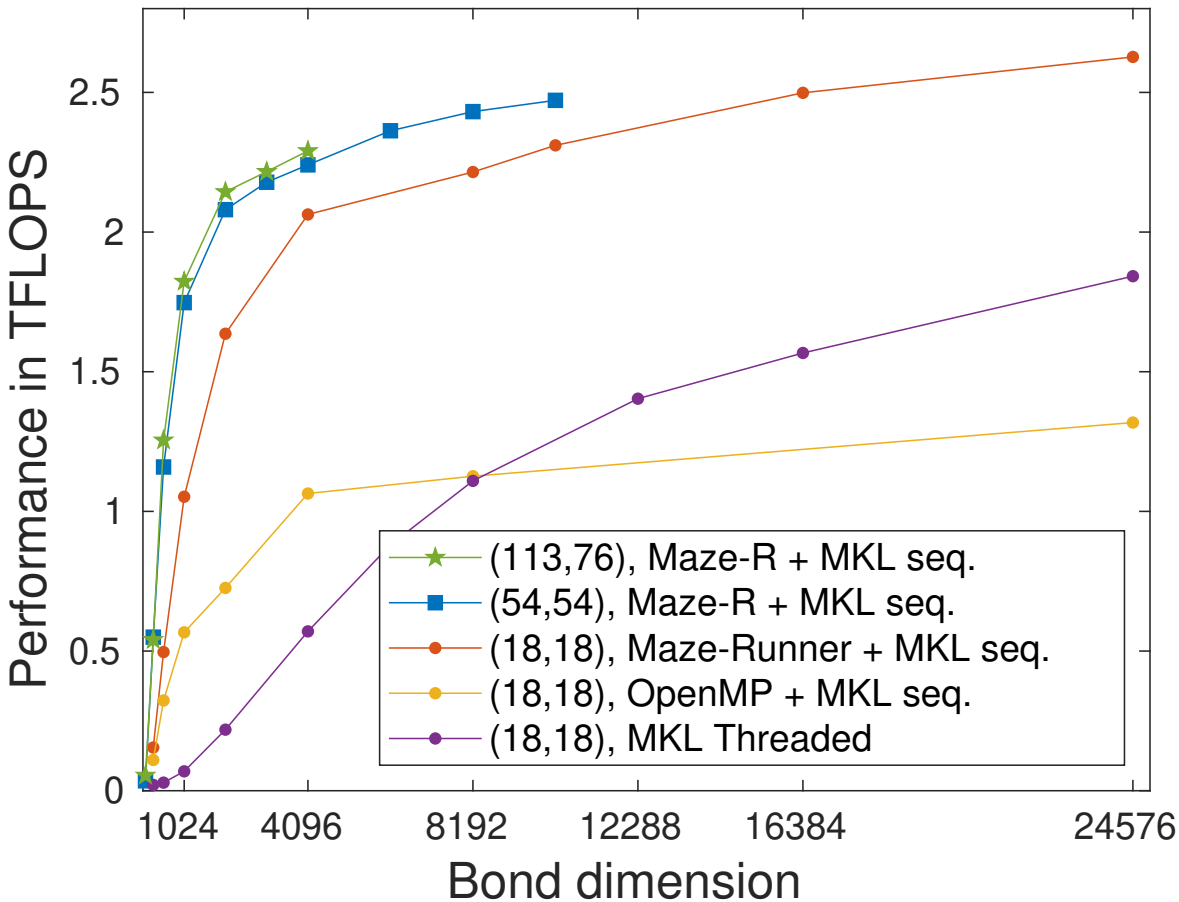
Produced on CPU-GPU for less than one day

Friesecke, Barcza, ÖL (2023)

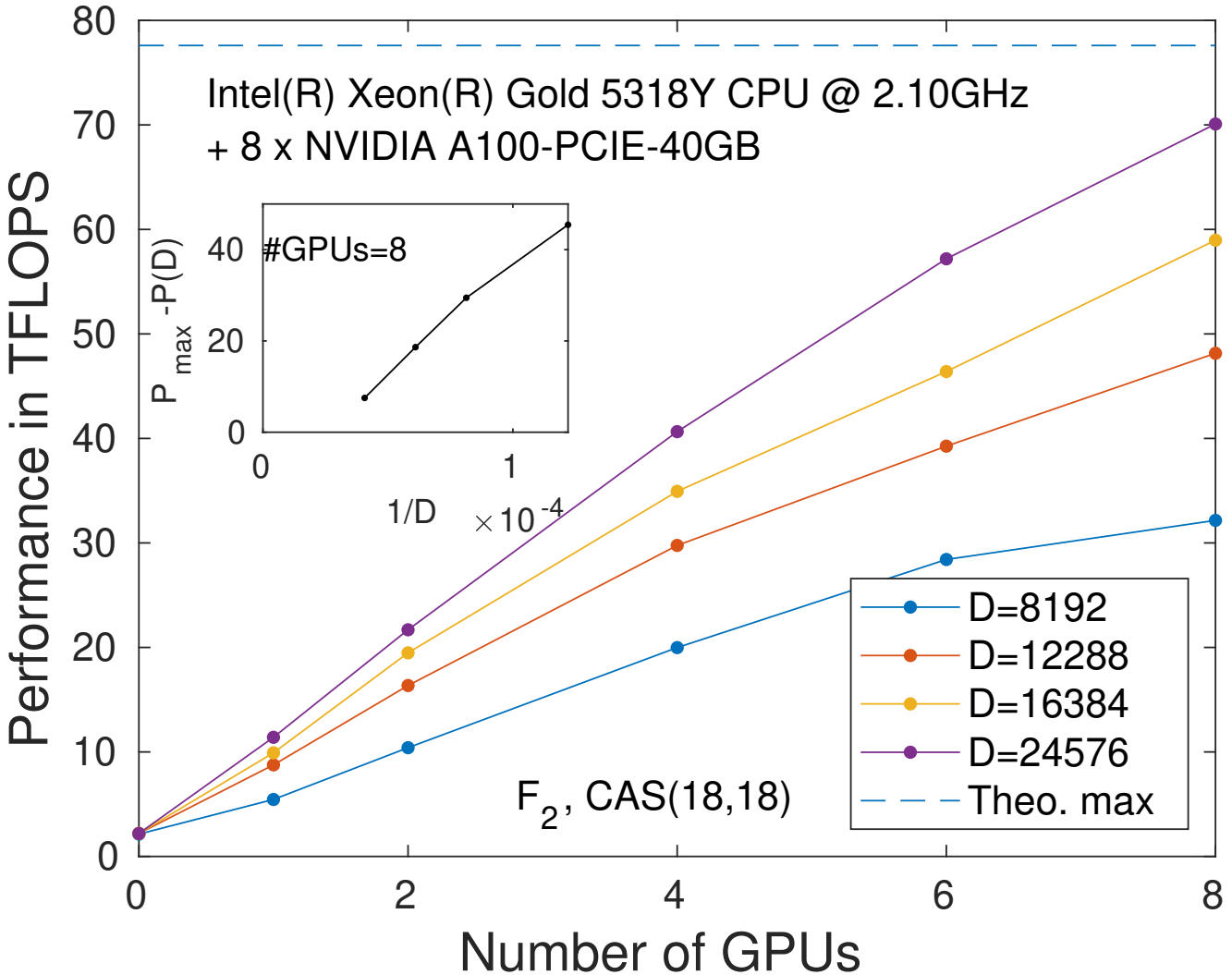
# CPU only limit (for CAS(113,76) $\dim \mathcal{H} = 2.88 \times 10^{36}$)

A. Menczer, ÖL (2023)

- Novel algorithmic (producer-consumer) model for parallelization.
- Novel gap-free, sequential write and read operations, no allocations and deallocations in the traditional sense.
- Strided batched type chained matrix multiplications without sum reduction at the end.



# 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

# Boosting effective performance via Wigner-Eckhart theorem

- Large-scale tensor operations substituted with multi-million independent vector and matrix operations
- The matrices and tensors are decomposed into smaller components (sectors) based on quantum numbers
- Non-Abelian symmetries in HPC framework is a highly non-trivial task as it requires a more delicate mathematical framework based on Wigner-Eckhart theorem leading to correction factors:

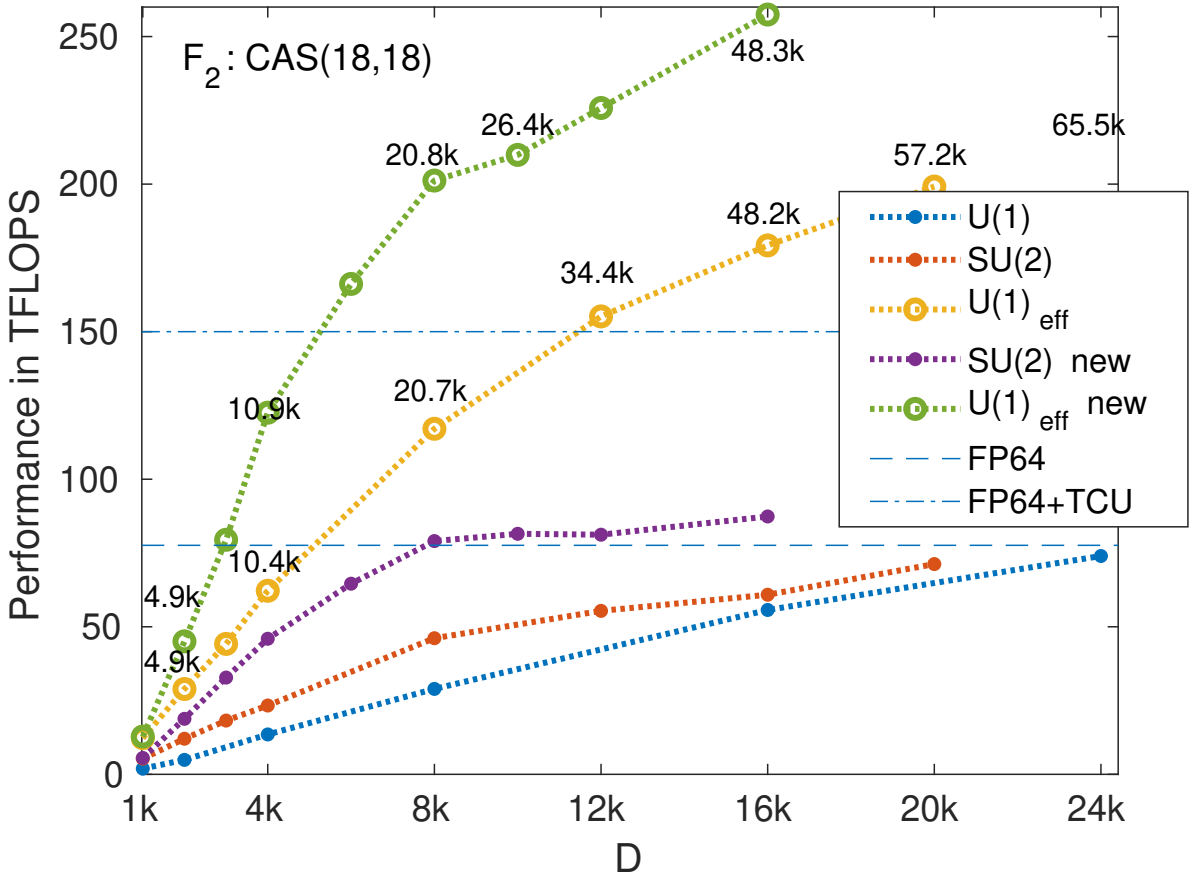
$$\tilde{C} = \sqrt{(2j_1'+1)(2j_2'+1)(2j+1)(2k+1)} \times W_{9j}(j_1, j_2, j, k_1, k_2, k, j_1', j_2', j') \tag{1}$$

where

$$W_{9j}(j_1, j_2, j, k_1, k_2, k, j_1', j_2', j') = \sum_{x=x_{\min}}^{x_{\max}} (-1)^{2x}(2x+1) \cdot W_{6j}(j_1, j_2, j, k, j', x) \times W_{6j}(k_1, k_2, k, j_2, x, j_2') \cdot W_{6j}(j_1', j_2', j', x, j_1, k_1), \tag{2}$$

- **Wigner-9j and Wigner-6j tensors**

# Boosting the effective performance via non-Abelian symmetries

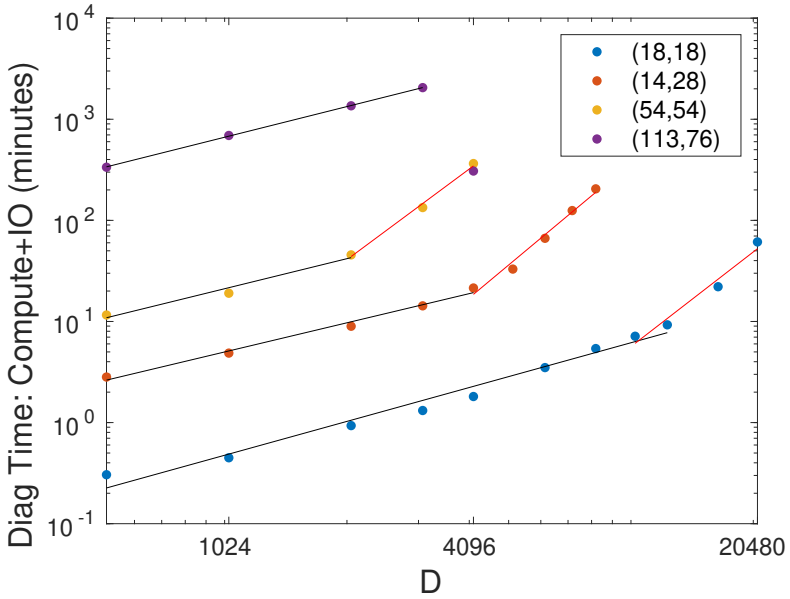



A. Menczer, Ö.L (unpublished, 2023), similar results for FeMoco(113,76)

- New mathematical model for parallelization $\rightarrow$ felxibe scaling
- We reached 108 TFLOPS $>$ 76 TFLOPS of the FP64 limit of NVIDIA $\rightarrow$ utilization of highly specialized tensor core units (TCU)
- Estimated effective $U(1)$ performance is about 250-500 TFLOPS.

# Dramatic reduction is scaling exponents: $D^3 \rightarrow D^{0.98}$

- New mathematical model for parallelization
- Computational burden of parallelization is marginal and evenly distributed among workers
- An adaptive buffering technique is used to dynamically match the level of data abstraction
- The non-Abelian symmetry related tensor algebra based on Wigner-Eckhart theorem is fully detached from the conventional tensor network layer




| System | CAS | $\gamma_1$ | $\gamma_2$ |
|---|---|---|---|
| $F_2$ | (18,18) | 1.11 | 3.10 |
| $N_2$ | (14,28) | 0.96 | 3.3 |
| FeMoco | (54,54) | 0.98 | 2.97 |
| FeMoco | (113,76) | 1.01 | - |

Table: Fitted exponents for the eight GPU accelerated diagonalization step.

# Conclusion

- Tensor topologies together with proper basis representations are important for efficient data sparse representaion of the wavefunction
- DMRG-RAS is variational, free of uncontrolled method errors and has the potential to outperfom conventional methods for strongly correlated molecules
- DMRG-RAS-X can provide the "missing digit"
- Our new mathematical model for massive parallelization via MPI and NVIDIA-DGX $\rightarrow$ multiNode-multiGPU exascale computation
- Current work: utilization of NVIDIA interlinks for further speedup
- We hope to offer: simulation of realistic material properties

Supports: Hungarian Academy of Sciences, the Hungarian National Research, Development and Innovation Office, Quantum Information National Laboratory of Hungary, European Research Area(ERA), Hans Fischer Senior Fellowship programme (IAS-TUM, Germany), SPEC, DOE, (PNNL, USA)