Recent advances in tensor network state methods

A journey from mathematical aspects towards industrial perspectives

Synergies among physics, chemistry, math and computer science

Örs Legeza

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

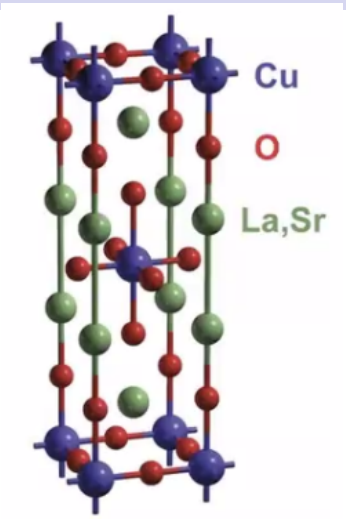
Institute for Advanced Study, Technical University of Munich, Germany

Parmenides Foundation, Pöcking, Germany

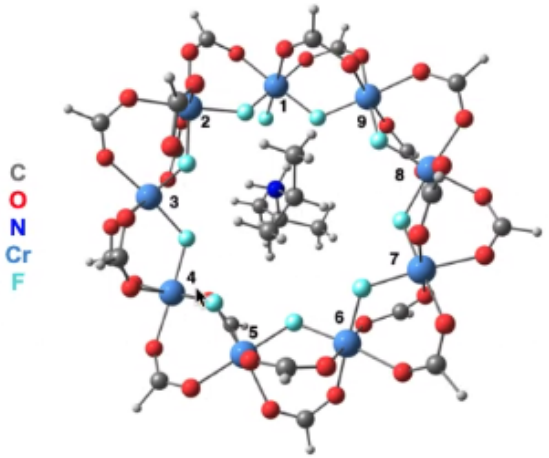
DYNAFLEX LTD, Budapest, Hungary

Academia-Industry Matching Event (AIME25)
Budapest 28.11.2025

Strong correlations between electrons used by nature and in new technologies



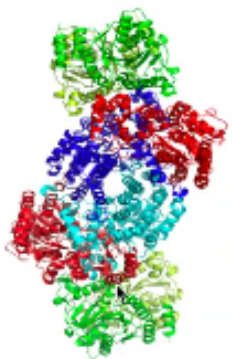
High T_c superconductors

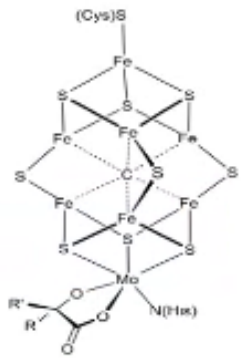


Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

Single molecular magnets (SMM)

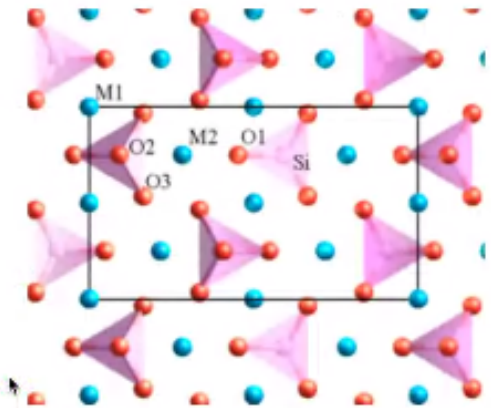
$$N_2 + 3\ H_2 \rightarrow 2\ NH_3$$





"FeMoco"

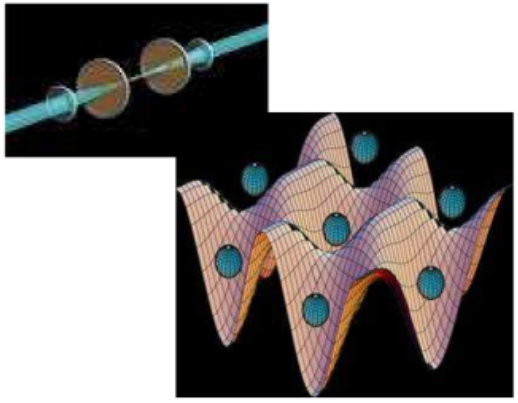
Nitrogen fixation



Battery technology

Experimental realizations: optical lattices

Numerical simulations: model systems



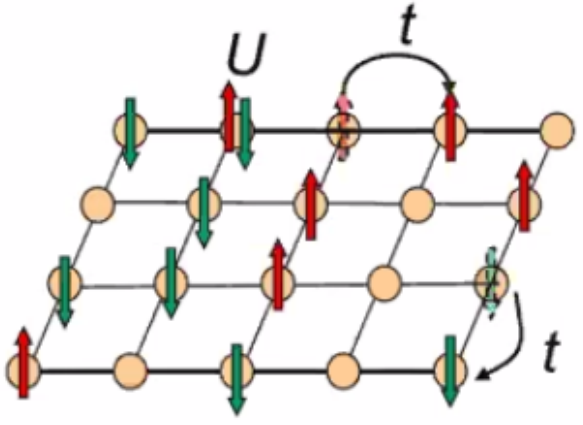
Atoms (represented as blue spheres) pictured in a 2D-optical lattice potential

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?

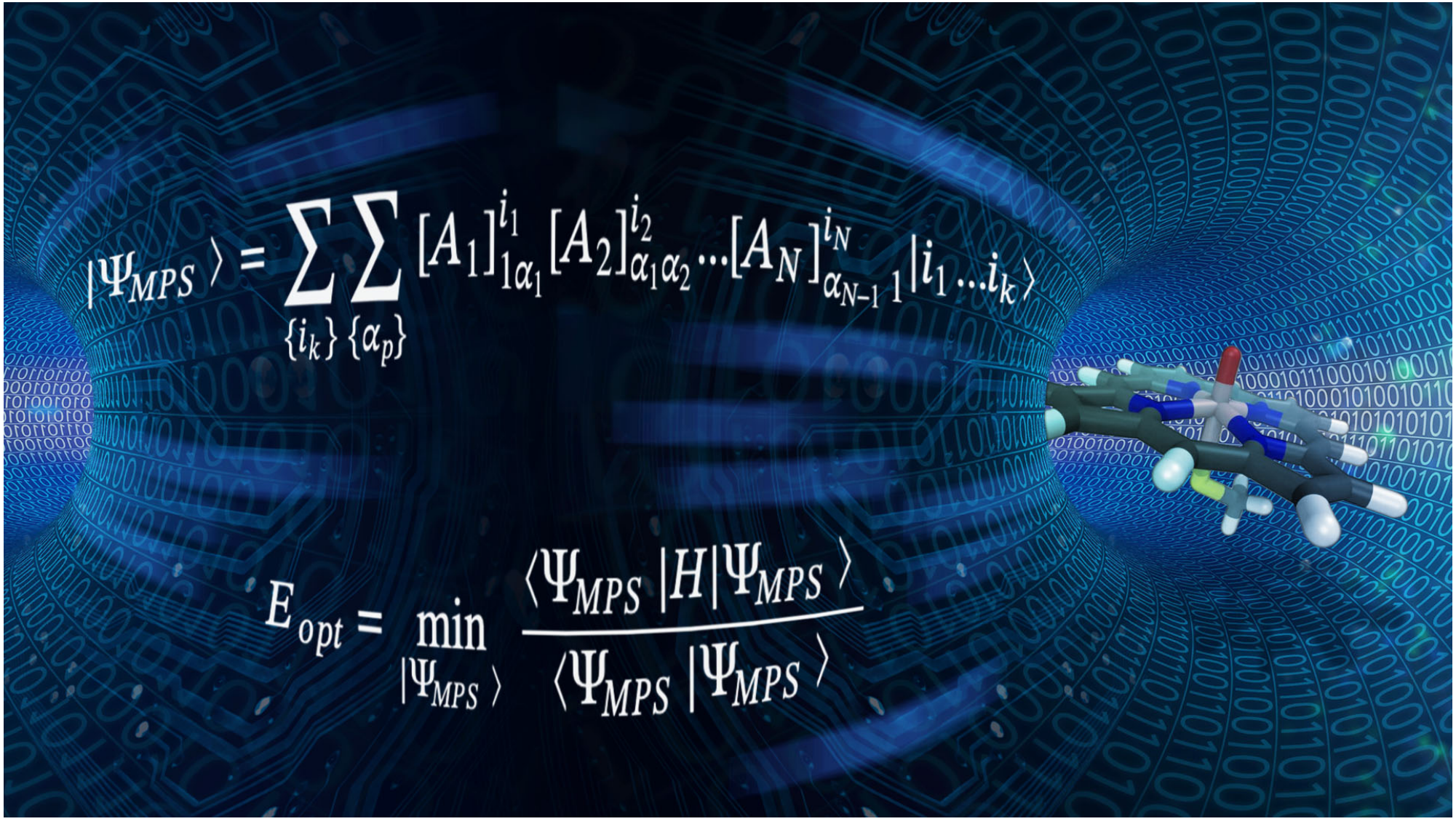
in collaboration with

- Our computer program package is used by more than 30 research groups worldwide for more than two decades in condensed matter physics, quantum chemistry, nuclear physics, quantum information theory, applied mathematics and computer science, etc...
- High-Performance Computing Center Stuttgart, Germany
- National Energy Research Scientific Computing Center (NERSC), USA

Recently there is also an increasing interest by industrial partners:

- NVIDIA, USA
- AMD, USA
- SandboxAQ, USA (Google startup)
- Riverlane LTD, UK
- Furukawa Electric Institute of Technology, Japan
- IBM, USA
- FACCTS, Germany
- Dynaflex LTD, Hungary

Wigner, PNNL, NVIDIA, SandboxAQ joint press release



https://www.pnnl.gov/news-media/collaboration-speeds-complex-chemical-modeling

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

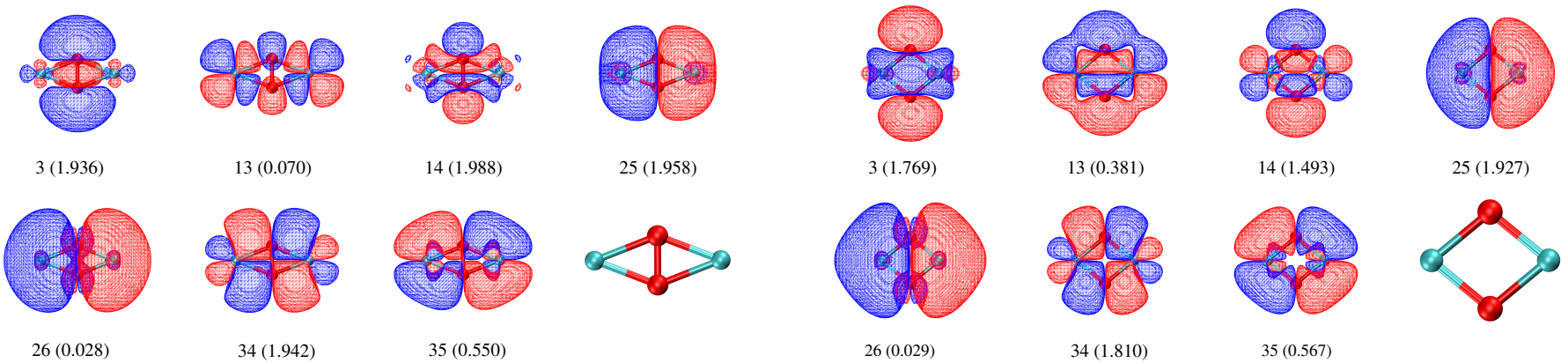
  • Symmetries: Abelian and non-Abelian quantum numbers, double groups, complex integrals, quaternion sym. etc

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

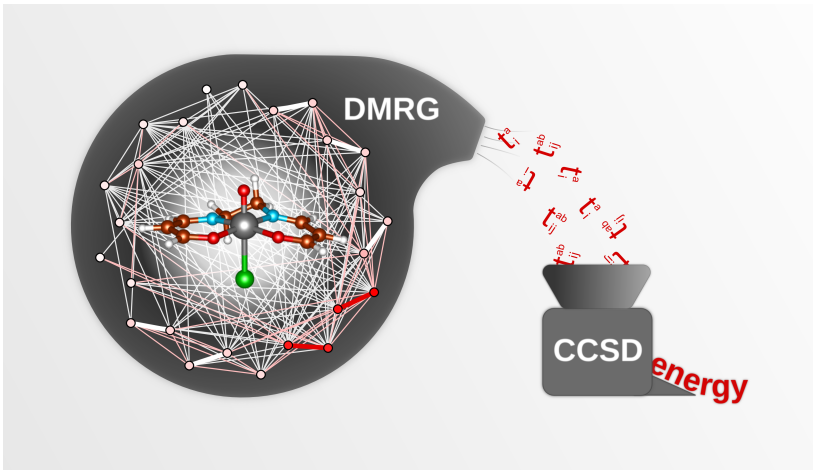
  • In ab inito DMRG the CAS size is: 70 electrons on 70 orbitals.

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

Quantum chemistry: modes are molecular orbitals (QC-DMRG) White, Martin (1999), Chan(2002), Ö.L.(2002), Reiher(2005), ...



- Combination with conventional methods
- post-DMRG methods



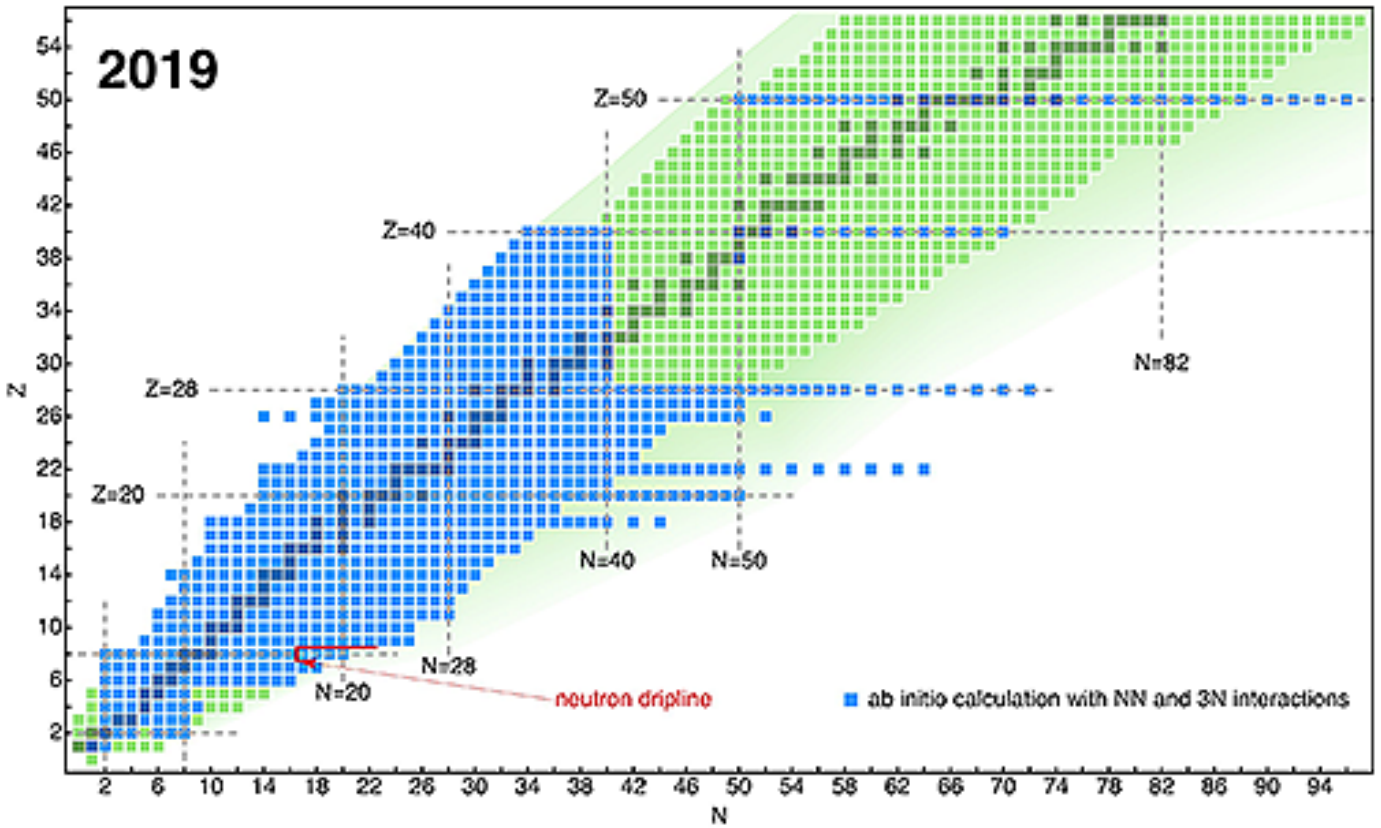
- Relativistic quantum chemistry: modes are spinors (4c-DMRG)

Knecht, Ö.L., Reiher (2014)

- electrons moving at relativistic speeds, close lying states and dynamical correlation, open d or f shells

Nuclear physics: modes are proton/neutron orbitals (JDMRG)

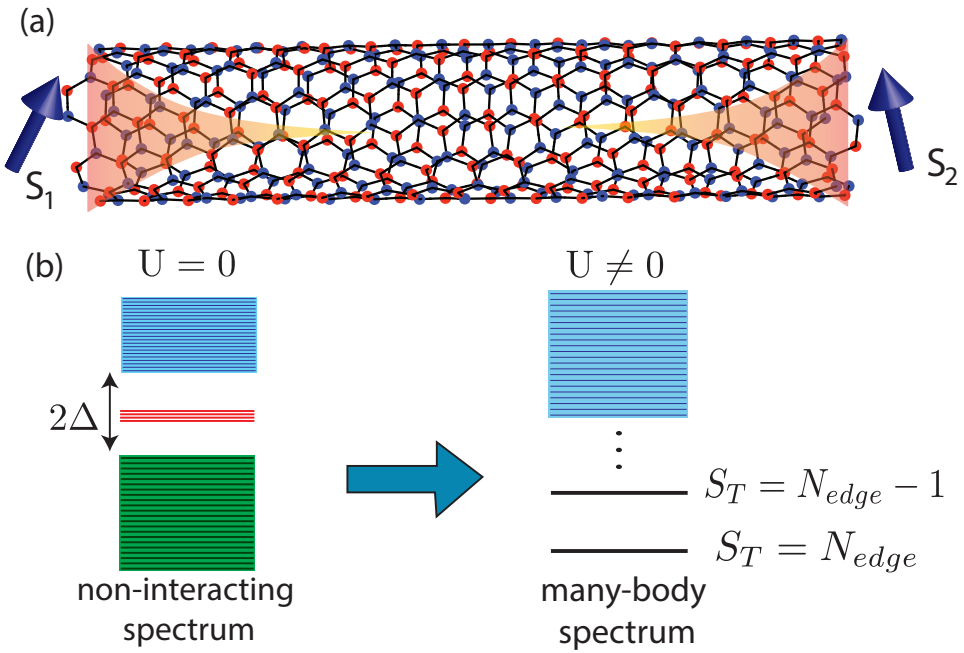
Dukelsky, Papenbrock, Pittel (2003), Ö.L., Veis, Dukelsky, Poves (2015)

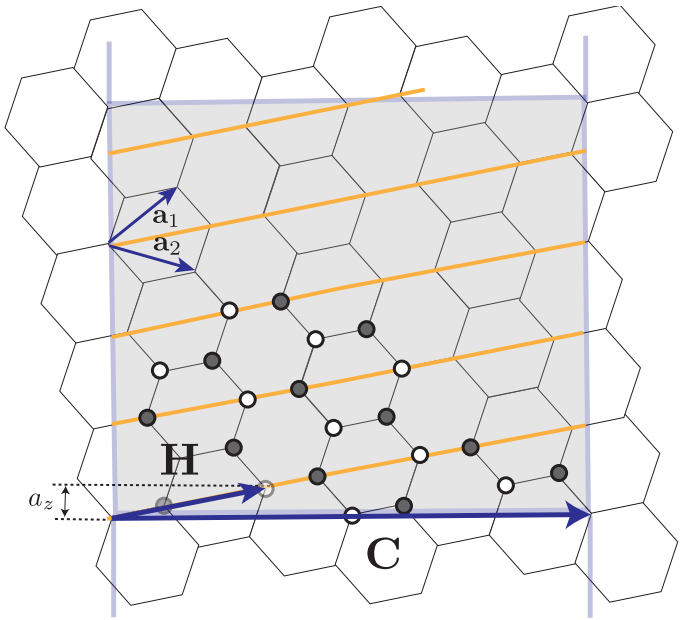


$$H = \sum_{\alpha} \varepsilon_{\alpha} c_{\alpha}^{\dagger} c_{\alpha} - \frac{1}{2} \sum_{\alpha\beta\gamma\delta} V_{\alpha\beta\gamma\delta} c_{\alpha}^{\dagger} c_{\beta}^{\dagger} c_{\delta} c_{\gamma} \, ,$$

- where $c_{\alpha}^{\dagger}$ and c_{α} creates and annihilates a particle with quantum numbers $\alpha = (n, l, j, m, \tau_z)$. $j \geq 1/2$, Isospin,
- no-core shell models
- effective Hamiltonian including parts of 3-body interactions

Real space: modes are lattice sites $\rightarrow$ new basis





$$H_0 = -\sum_{\mathbf{x},\mathbf{x}',s} t(\mathbf{x}-\mathbf{x}')c_s^\dagger(\mathbf{x})c_s(\mathbf{x}') \;\text{ with }\; \mathbf{r}=\mathbf{r}(\mathbf{x})=\mathbf{r}(\nu,\mathrm{l},\tau)\,,$$

- Construct and diagonalize the non-interacting part of the Hamiltonian and obtain the corresponding eigenfunctions $\phi_\alpha(\mathbf{r}) \equiv \phi_\alpha(\nu,\ell,\tau)$,
- Express the Coulomb interaction in this basis.
- For effective Coulomb interaction we use the so-called Ohno potential,

$$V(\mathbf{r}_1-\mathbf{r}_2) = \frac{e^2}{\epsilon_r}\frac{1}{\sqrt{(\mathbf{r}_1-\mathbf{r}_2)^2+\alpha^2}},$$

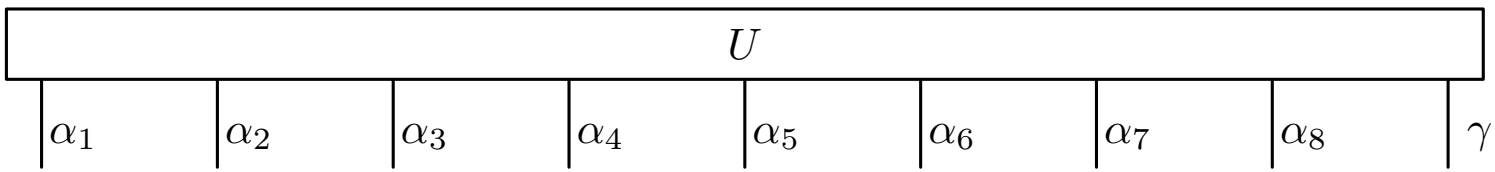
, Moca, Izumida, Dóra, Ö.L., Zaránd (2019)

Tensor product approximation

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

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

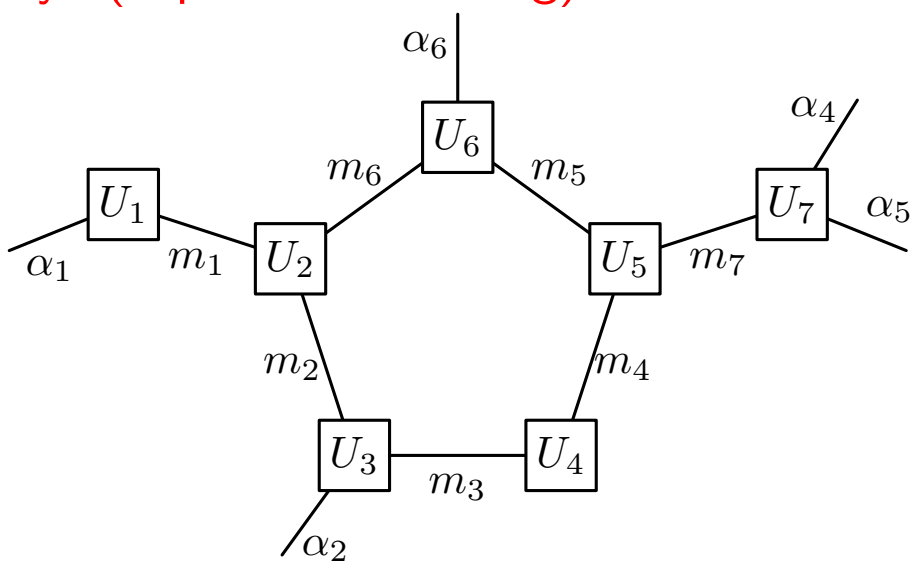
where $span\{|\alpha_i\rangle : \alpha_i = 1, \dots, q_i\} = \Lambda_i = \mathbf{C}^{q_i}$ and $\gamma = 1, \dots, m$.



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

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

A general tensor network representation of a tensor of order 5.



Matrix product state (MPS) representation / DMRG / TT

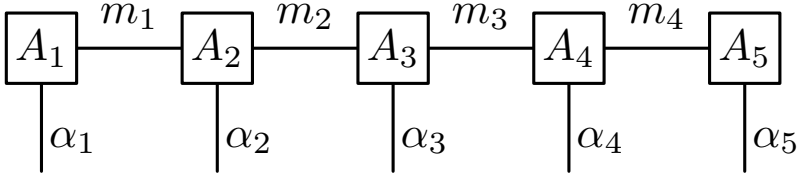
Exponential scaling $\rightarrow$ polynomial scaling

Affleck, Kennedy, Lieb Tagasaki (87); Fannes, Nachtergale, Werner (91), White (92)

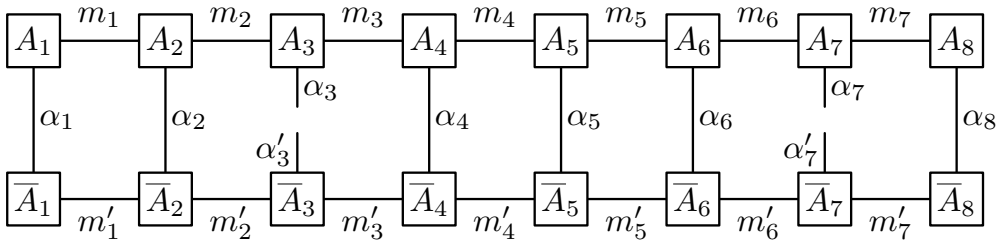
The tensor U is given elementwise as

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

We get d component tensors of order 2 or 3. Scaling: m^3.



Calculation of ρ_{ij} corresponds to the contraction of the network except at modes i and j.



von Neumann quantum information entropy, $s = -\sum_\alpha \lambda_\alpha^2 \ln \lambda_\alpha^2$.

Mutual information, $I = s_i + s_j - s_{ij}$.

Ö.L & Sólyom, (03), Rissler, Noack, White (06)

Single particle unitary mode transformation $U \in U(N)$

Krumnow, Veis, ÖL, Eisert 2015-2021

Friesecke, Werner, Kapas, Menczer, Ö.L. 2024

- New modes $\varphi'_i = \sum_j U_{ij}\varphi_j$, and $C' = G(U)^\dagger C$ where $G(U)$ is a unitary transformation on the space of many-body coefficient tensors.

- For time reversal symmetric case, C and the φ_i are real-valued and $U \in O(N)$ or, discarding an immaterial overall sign factor, $U \in SO(N)$.

- U can be parametrized as $U = e^A U_*$ with U_* an arbitrary fixed matrix in $SO(N)$ and A real and skew-symmetric, the parametrization being unique for U close to U_*.

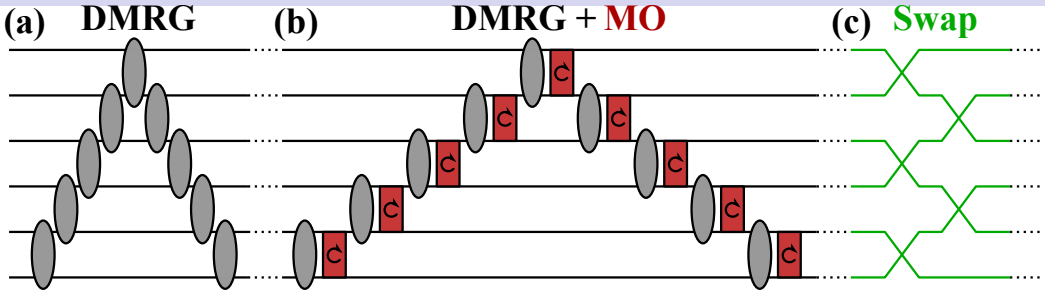
- Thus stationarity of a scalar function f on $SO(N)$ at U_* is equivalent to

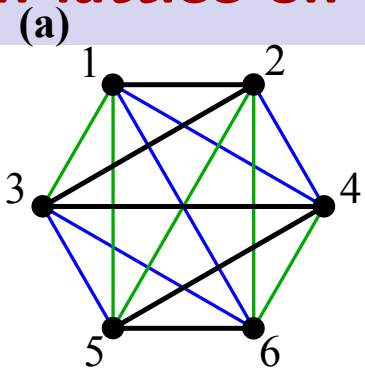
$$0 = \frac{d}{dt}\Big|_{t=0} f(e^{tA}U_*) = \mathrm{Tr}\,\frac{\partial f}{\partial U}(U_*)U_*^T A^T\,,\ \forall A^T = -A$$

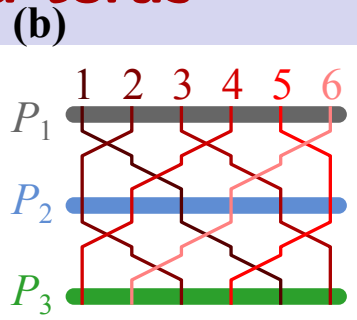
that is to say $\frac{\partial f}{\partial U}(U_*)U_*^T$ symmetric.

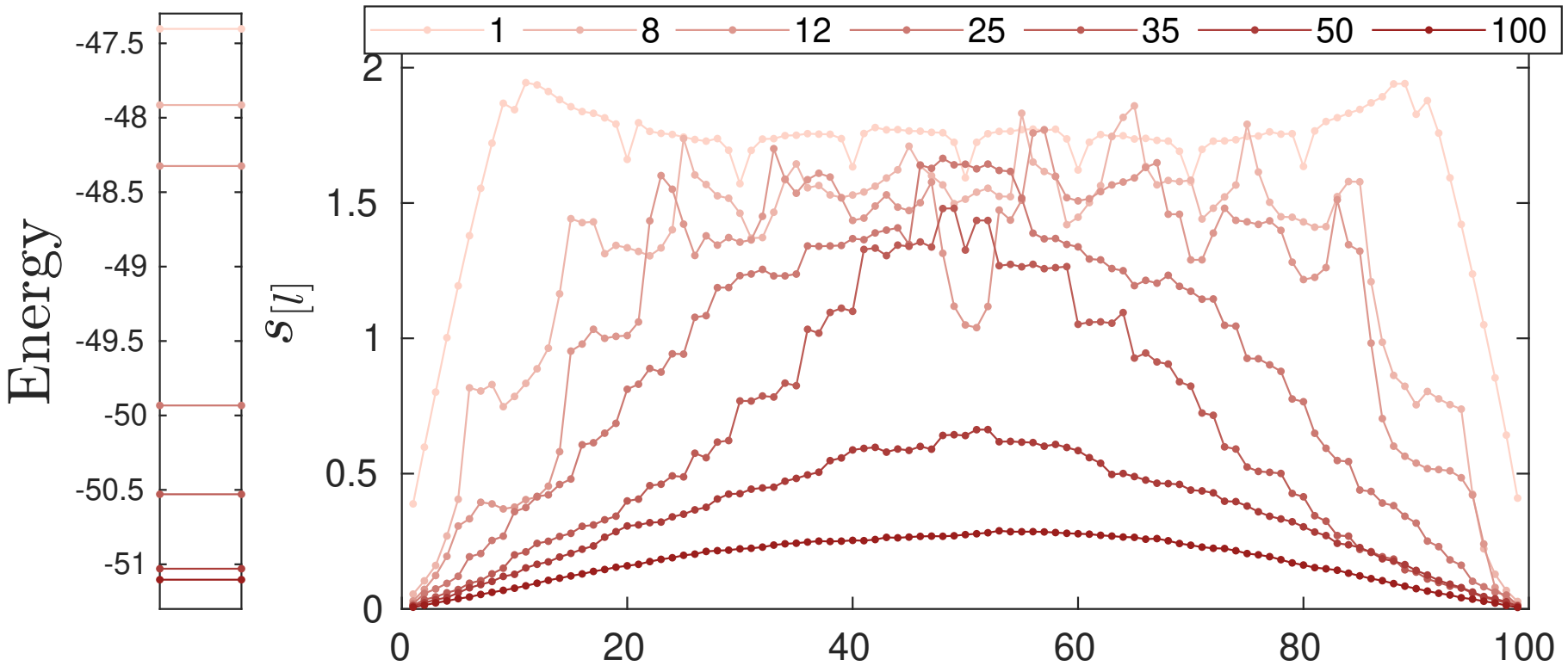
- Reduction to pairwise rotations: To achieve stationarity minimize f over all pairwise rotations

2-d spinless fermionic 10×10 quantum lattice on a torus





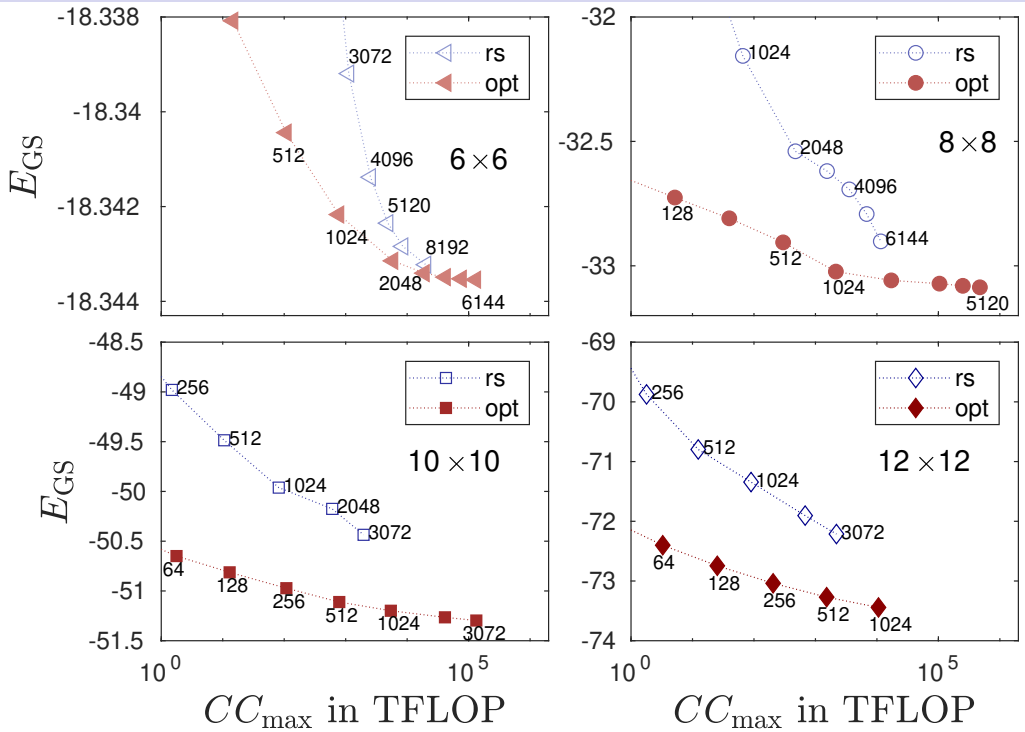




Two-qubit gate disentanglers via mode optimization

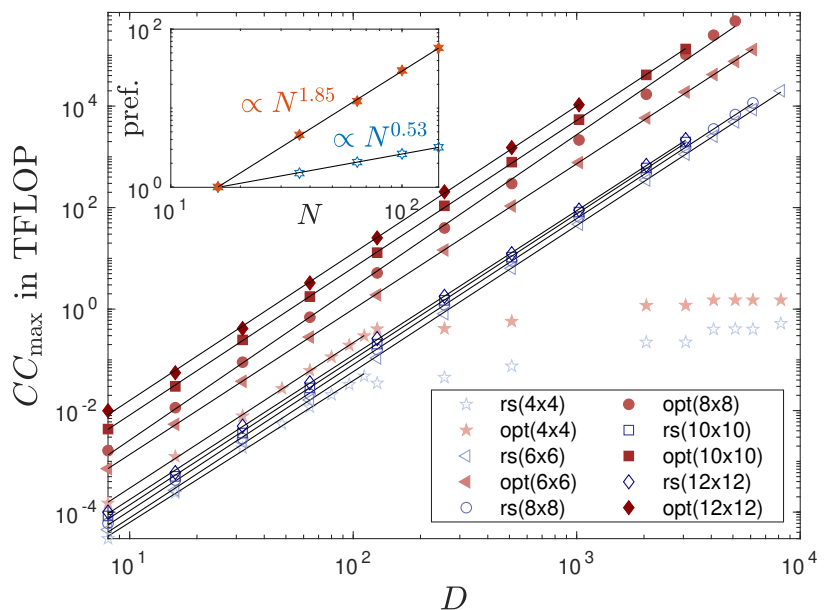
Computational complexity can be reduced by orders of magnitudes!

Direct connection to computer science: complexity in FLOPS



Menczer,Kapas,Werner,ÖL,2023

- Half-filled $N \times N$ Spinless model on a torus geometry
- $t = 1,\ t' = 0.4,\ V = 0.8$
- opt with $D = 80$
- Computational complexity in teraFLOPS



- The solid lines are first-order polynomial fits leading to exponents $\nu \simeq 3 \pm 0.2$
- inset: scaling of the prefactor as a function of system size N with fitted exponents 0.53 and 1.85 for the real space and for the optimized basis, respectively.

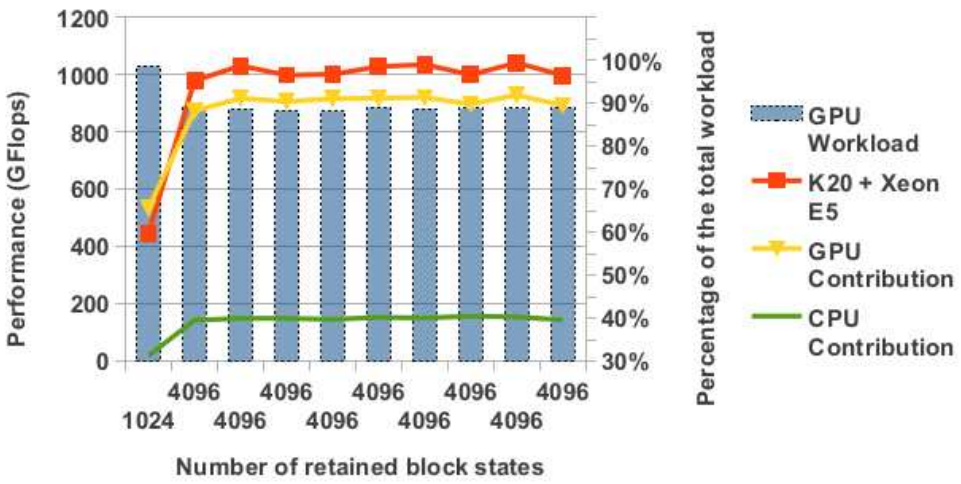
Towards exascale computations on supercomputers

• Underlying tensor and matrix algebra can be organized into several million of independent operations (tasks).

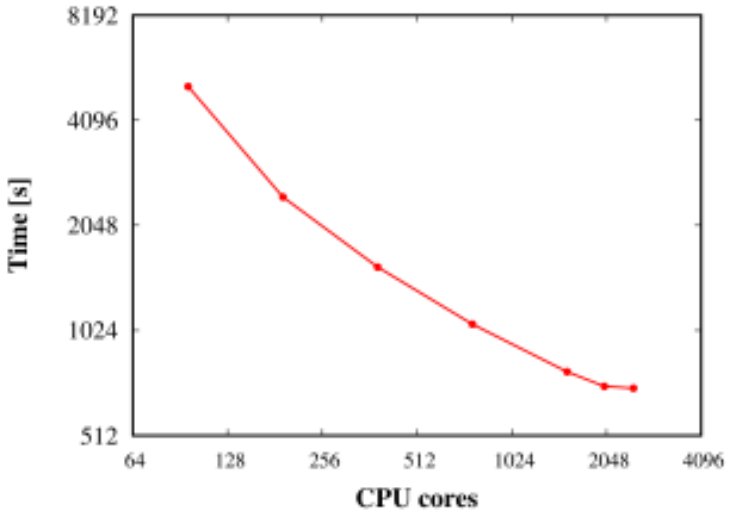
• Dense matrix operations are performed in parallel according to the so-called quantum number decomposed representations (sectors).

• parallelization over operators, sectors, positions

GPU: MPS and TNS
on kilo-processor architectures (K20):
Nemes, Barcza, Nagy, Ö.L., Szolgay, 2014



Massive parallelization
Brabec, Brandejs, Kowalski
Xanntheas, Ö.L., Veis (2020)



(a) Davidson procedure

FeMoco cluster
[CAS(113,76)]

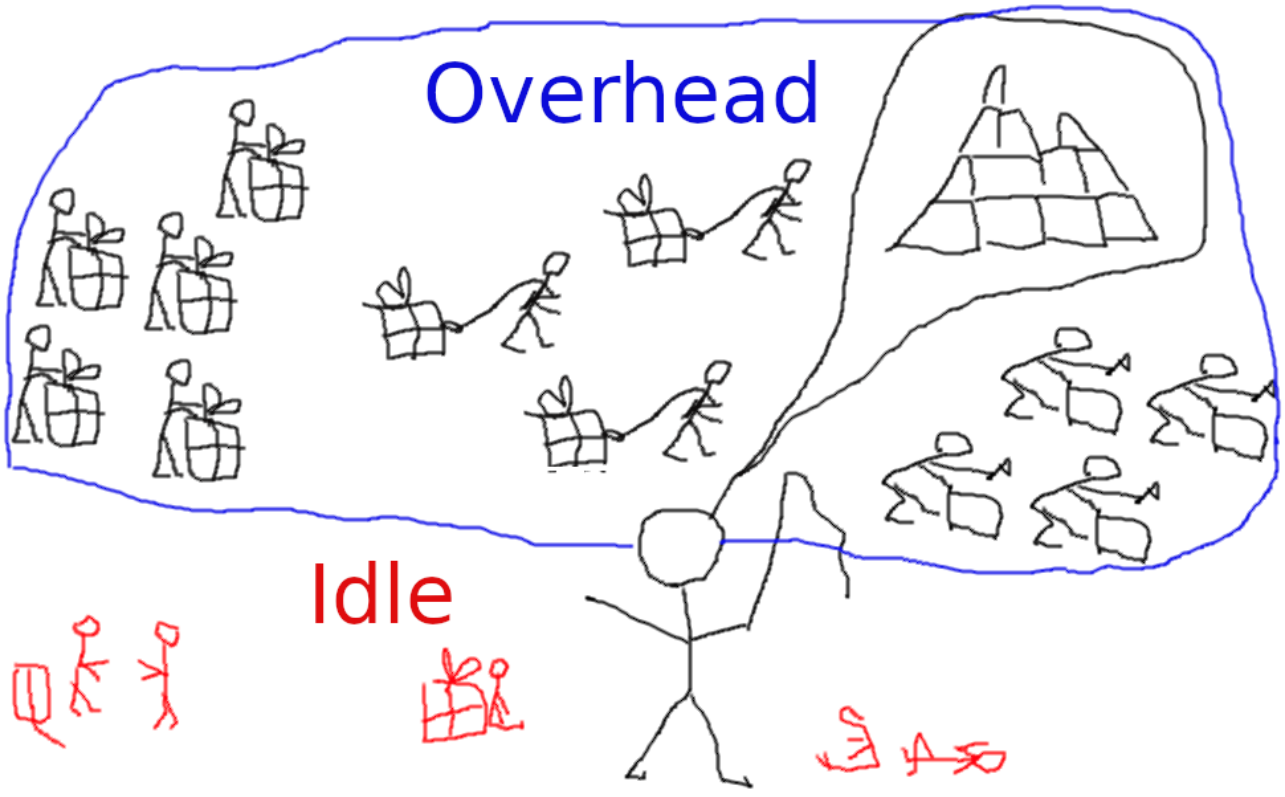
Centralized scheduling: unideal society

- Set of workers to generate tasks → Workers are threads
- Set of workers to transfer tasks → Transfer: IO communication
- Set of workers to execute tasks → CPU, GPU, FPGA units

- Central scheduler has to organize the full workflow, measure complexity of tasks, distribute tasks, check execution etc
- Central scheduler envisions the global aim & wants to accomplish it
- Tasks: several millions of independent tensor and matrix operations

Centralized scheduling: Huge overhead, units can be idle

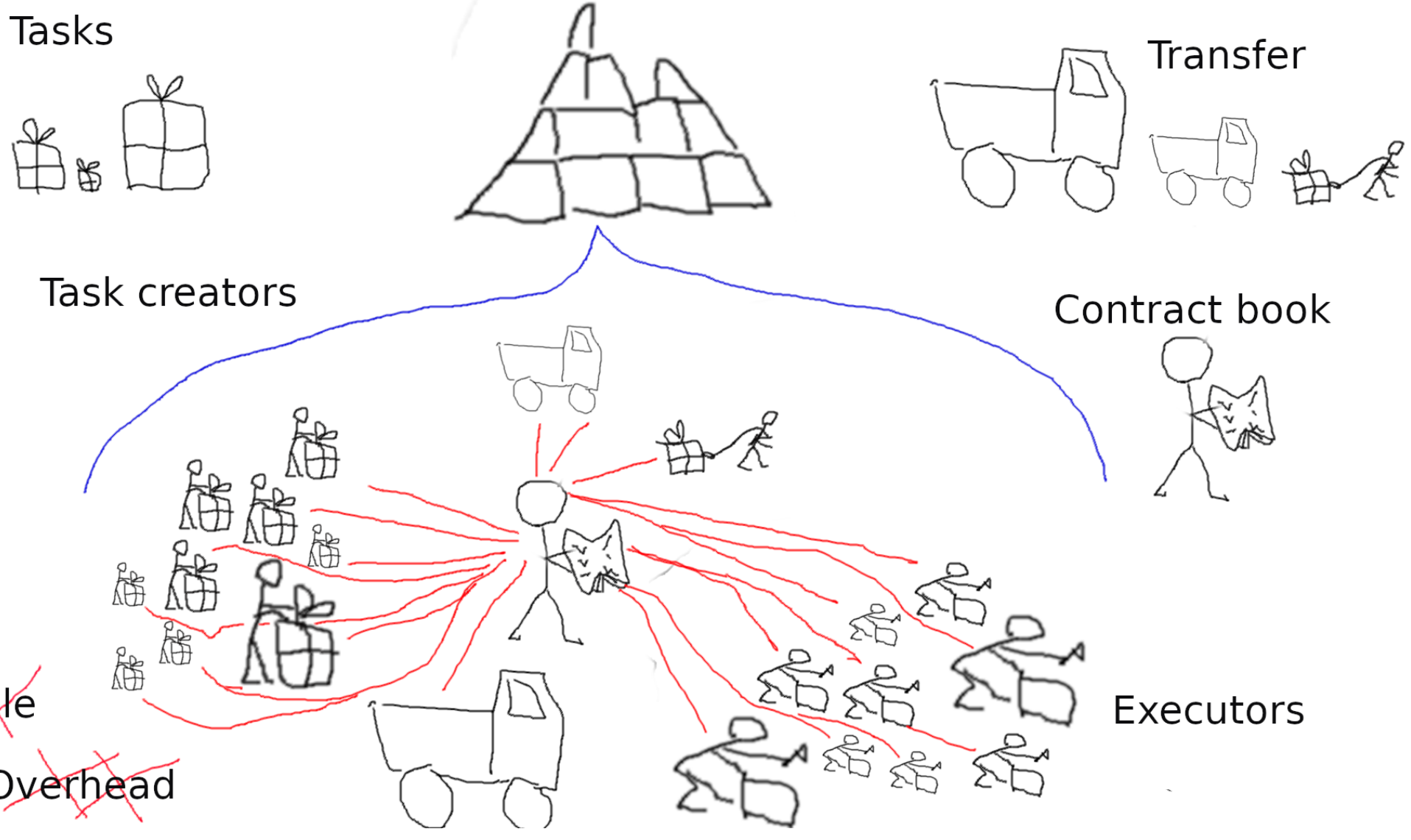
- Central scheduler performs lot of measurements, estimations, communication to rearrange tasks and workers → huge overhead



- Central scheduler cannot see everything in a given moment → workers can be idle
- Too much workload on scheduler → inefficient scheduling, tasks can pile up partially

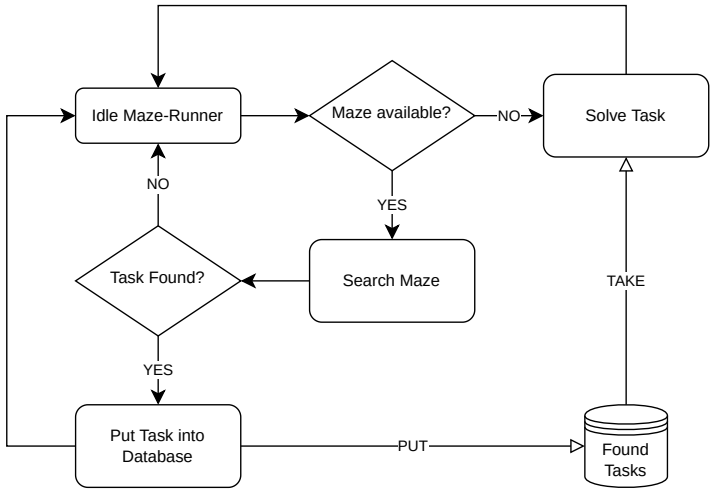
Self motivated workers → ideal "team-like" society

- Central unit: Contractor, contract book (only meta-data communicated, boolean-like bookkeeping flags)
- Everybody is motivated to achieve global aim

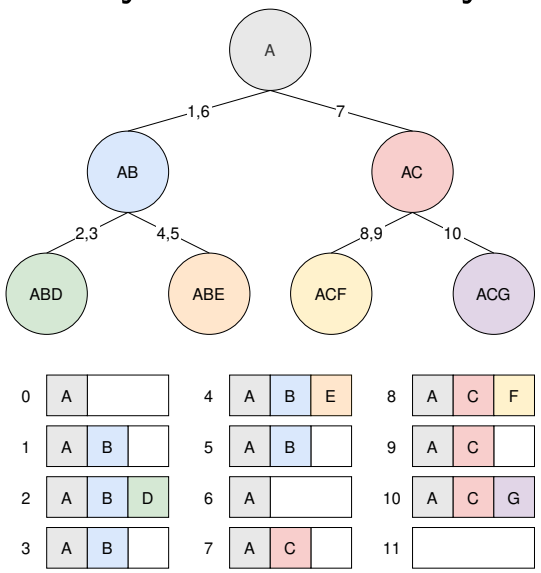


Novel algorithmic solutions A. Menczer, ÖL (2023)

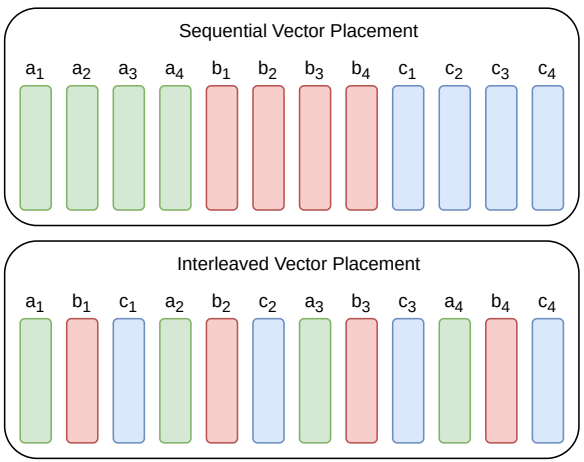
Life Cycle of a Maze-Runner Thread.



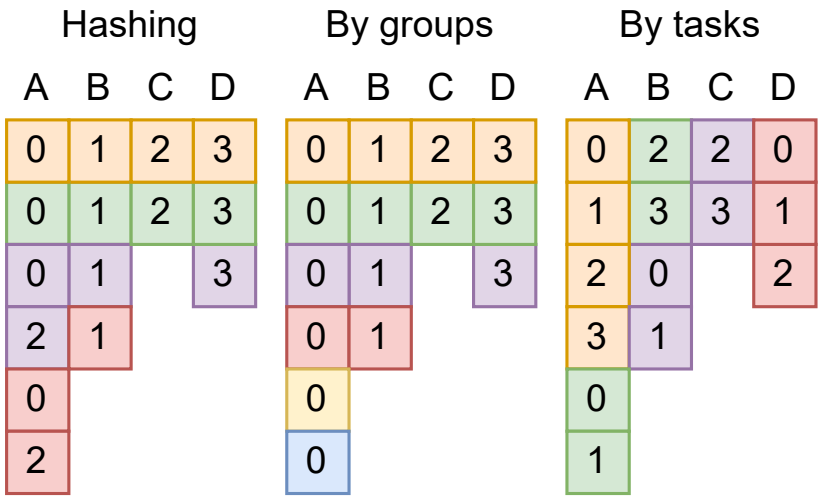
Graph theory based memory management



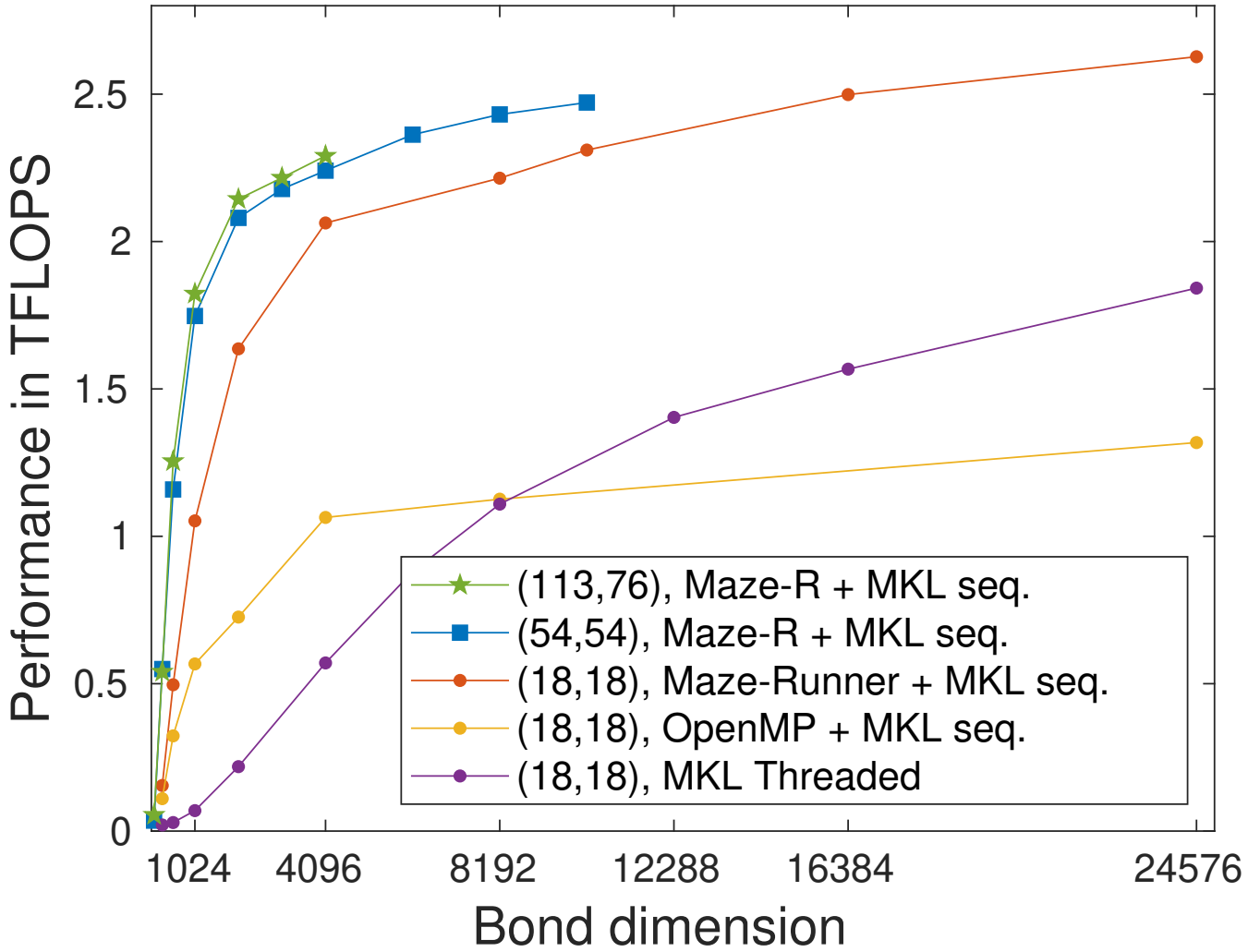
Strided Batched operations via data localization



Execution via hierarchy of tasks



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



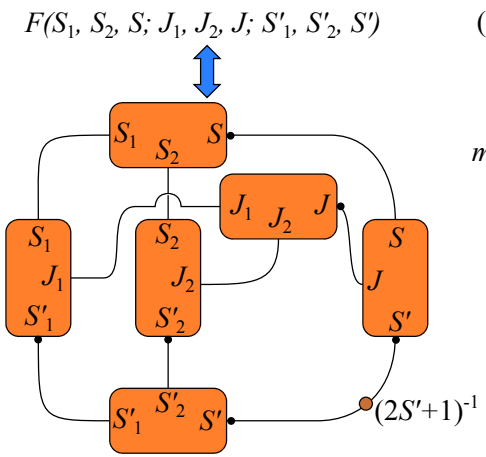
Performance measured in TFLOPS for the F_2 and FeMoco chemical systems for CAS(18,18) and CAS(54,54) orbitals spaces, respectively, as a function of the DMRG bond dimension on a dual Intel(R) Xeon(R) Gold 5318Y CPU system with 2×24 physical cores running at 2.10 Ghz.

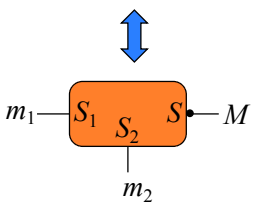
Boosting the effective performance via non-Abelian symmetries

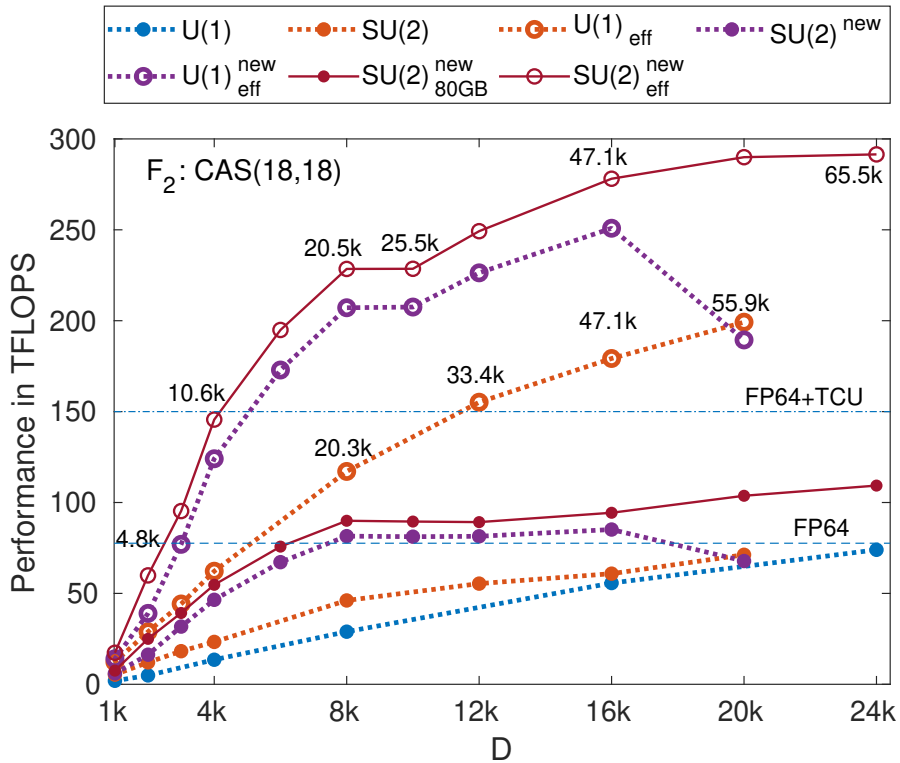
Benchmark on 8×A100 GPUs with 40GB VRAM. Menczer, Ö.L (2023), CAS(18,18)

$$[\![\mathbb{O}^{(\nu,L)} \otimes \mathbb{O}^{(\nu,k)}]\!]_{\gamma',\gamma} = \mathbb{O}^{(\nu,L)}\mathbb{O}^{(\nu,k)} F(S_\alpha, S_k, S_\gamma; S_L^{\mathrm{op}}, S_k^{\mathrm{op}}, S_L^{\mathrm{op}\prime}; S'_\alpha, S'_k, S'_\gamma),$$

where F equals the Wigner-9j symbol up to rescaling,



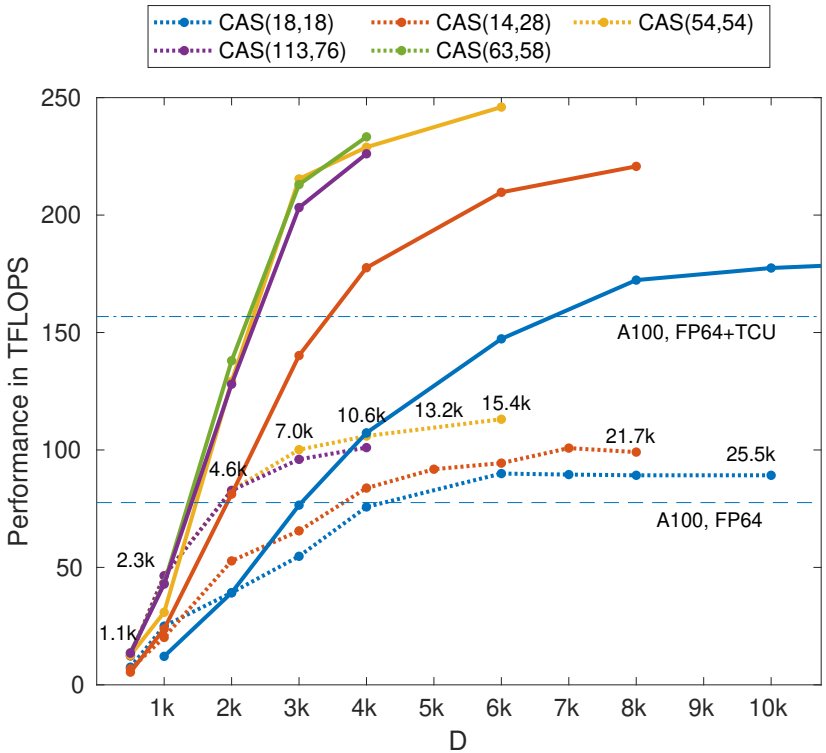


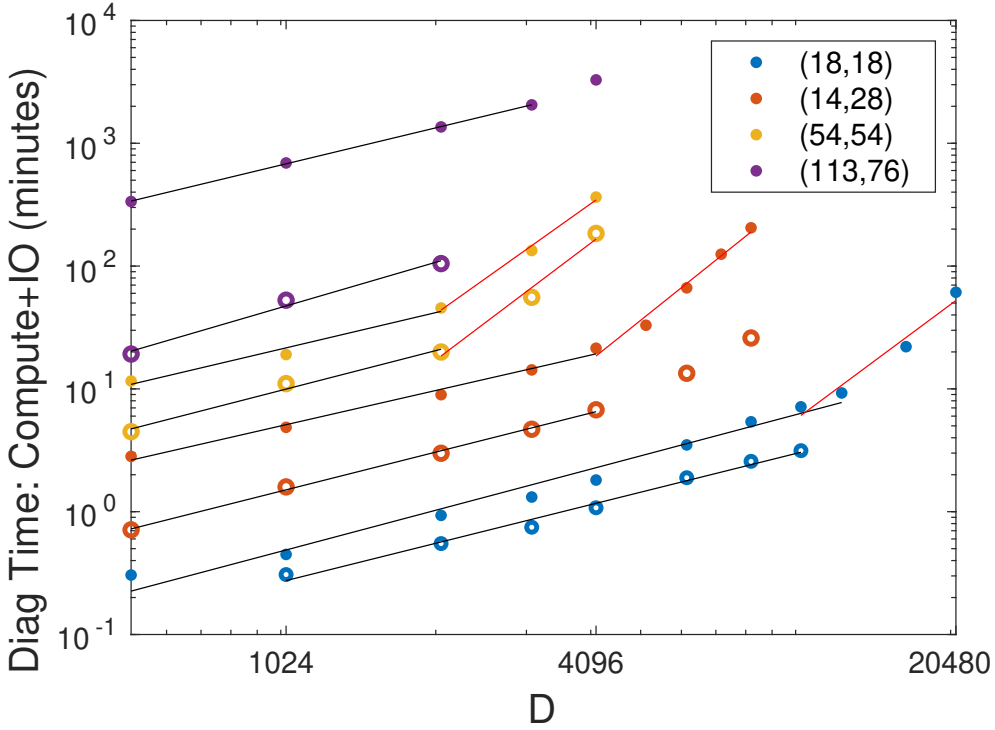


- New mathematical model for parallelization → felxibe scaling
- $D_{SU(2)} = 24576 \rightarrow D_{U(1)} = 2^{16} \rightarrow$ FCI solution
- We reached 108 TFLOPS > 76 TFLOPS of the FP64 limit of NVIDIA

→ utilization of highly specialized tensor core units (TCU)

Quarter petaflops on a single node $\sim$ 10000x speedup; $D^3 \to D$



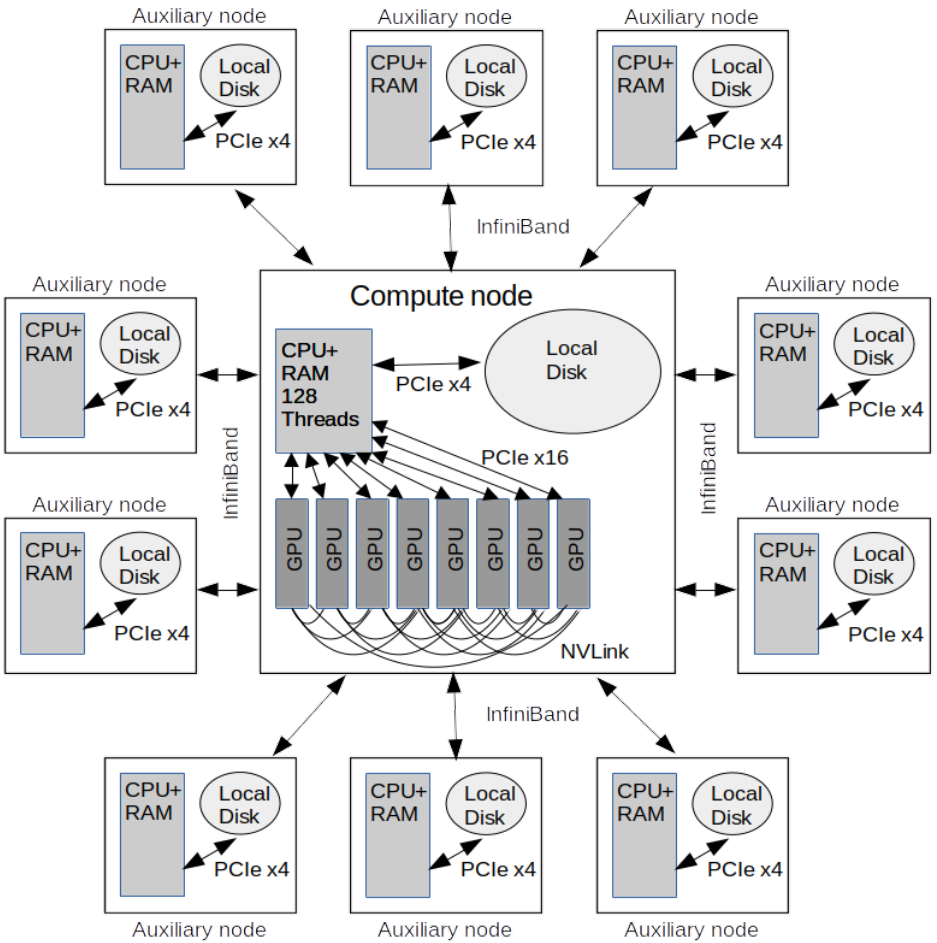


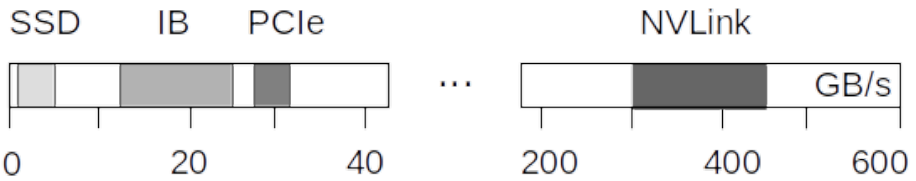
- NVIDIA DGX H100: 80x speedup wrt a single node with 128 cores
Testing performance up to $\sim$ 250 TFLOPS in collab with NVIDIA and SandboxAQ Menczer,Damme,Rask,Huntington,Hammond,Xantheas,Ganahl,ÖL

- New model to utilize NVIDIA D2D links. A. Menczer ÖL (unpublished 2023)

- Combination of our MPI and GPU kernels: full replacement of *boost library*, asynchronous IO, multiNode-multiGPU
$\to$ petascale computing. A. Menczer ÖL (unpublished 2023-2024)

Cost optimized TNS Menczer, ÖL 2024



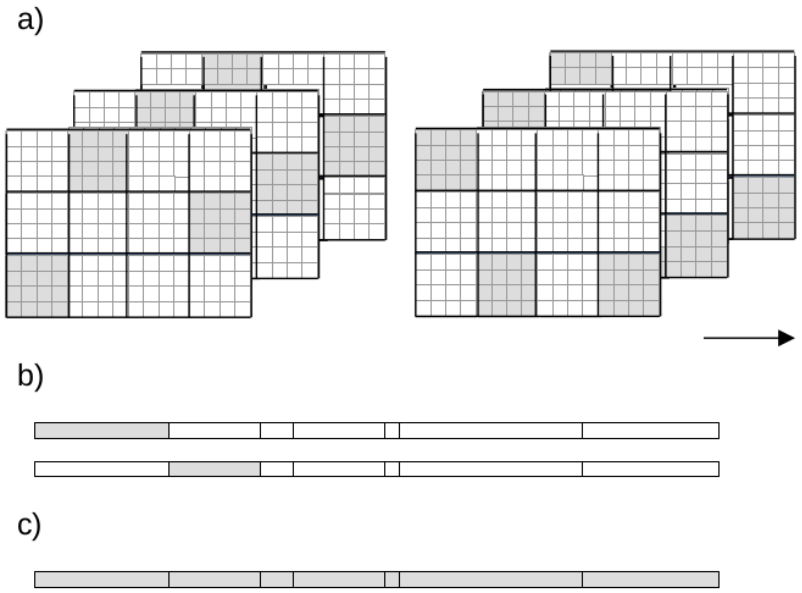


- DGX-H100 costs 100 USD/hour on Google Cloud

- Schematic plot of hardware topology illustrating the various communication channels (arrows), such as host to host (H2H), host to device (H2D) and device to host (D2H), and device to device (D2D), i.e., InfiniBand, PCI-E, and NVLink, accordingly.

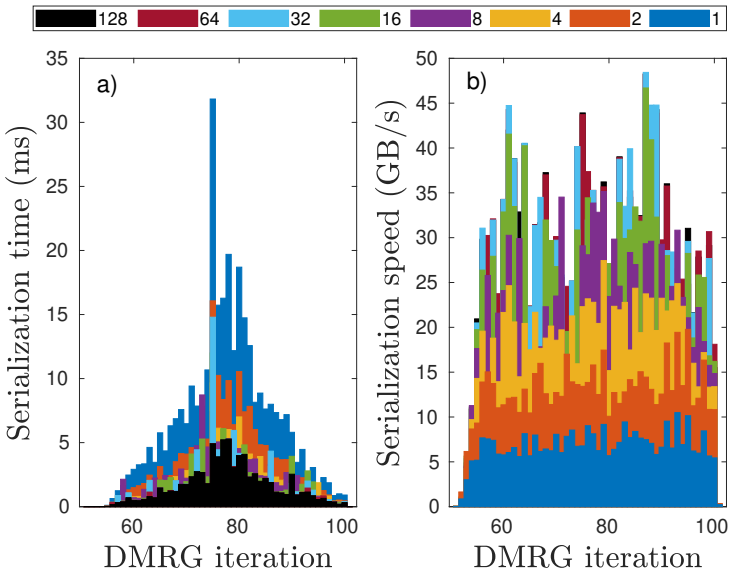
- The compute node is a very powerful and expensive unit surrounded by one or more cheap auxiliary nodes with minimal computational capacity, but with substantial amount of RAM

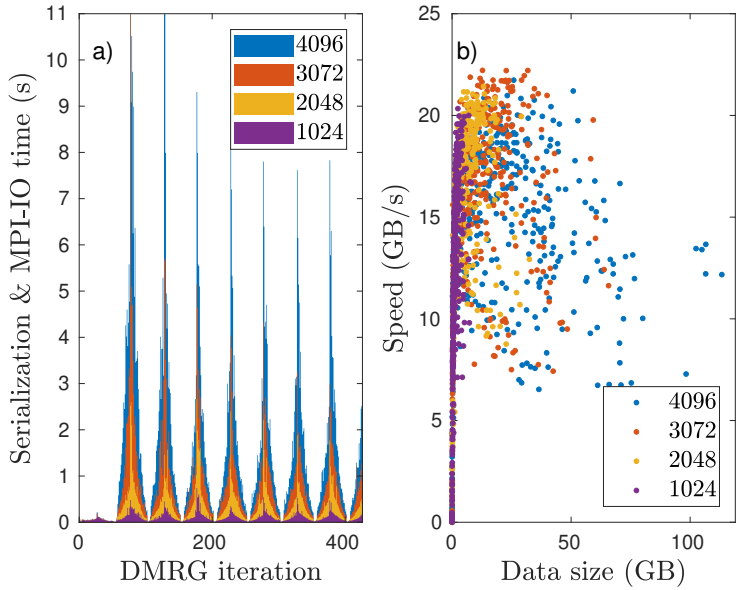
Precontracted network asynchronous MPI-IO Menczer, ÖL 2024



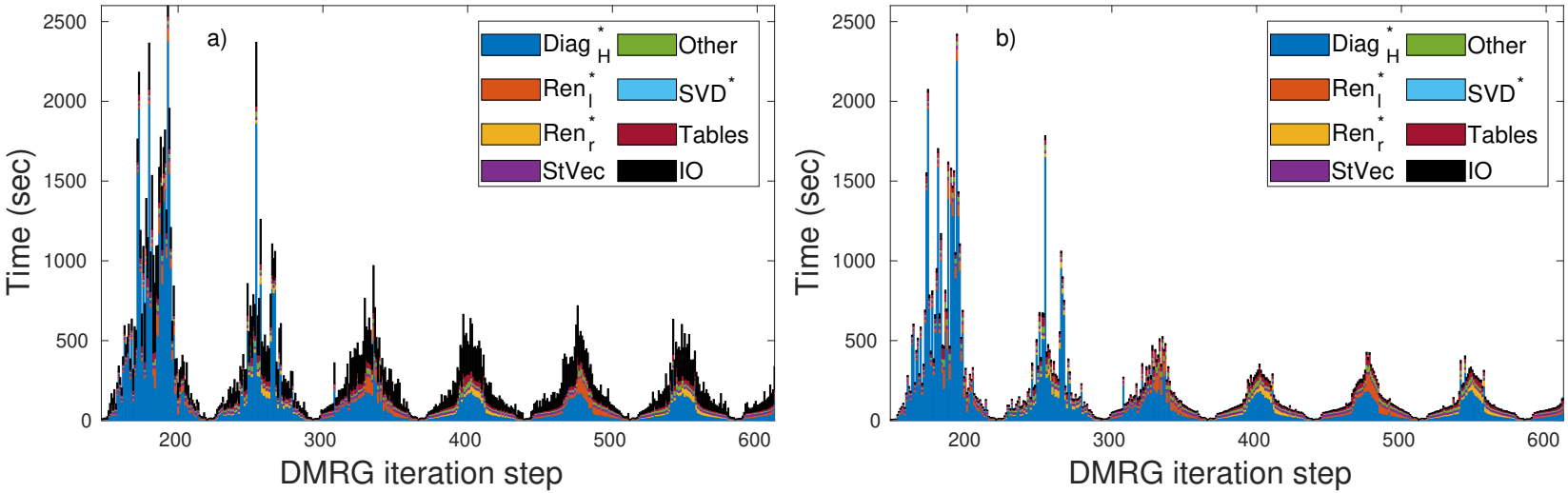
a) Schematic plot of quantum number based block sparse representation of matrices and tensors.
b) Skeleton of serialized data segments used during disk IO save procedure or MPI based communication.
c) Skeleton of serialized data segments filled completely with data when asynchronous save IO.

- FeMoco CAS(54,54), $D_{SU(2)} = 5120$, $D \simeq 17500$





Precontracted network asynchronous MPI-IO Menczer, ÖL 2024



- Decomposition of the total wall time as a function of DMRG iteration steps via synchronous IO operations (a) and via 10 auxiliary nodes (b) for the FeMoco CAS(113,76) model space using $D = 4096$ $SU(2)$ multiplets corresponding to largest $U(1)$ bond dimension values around $D_{U(1)} = 15400$.

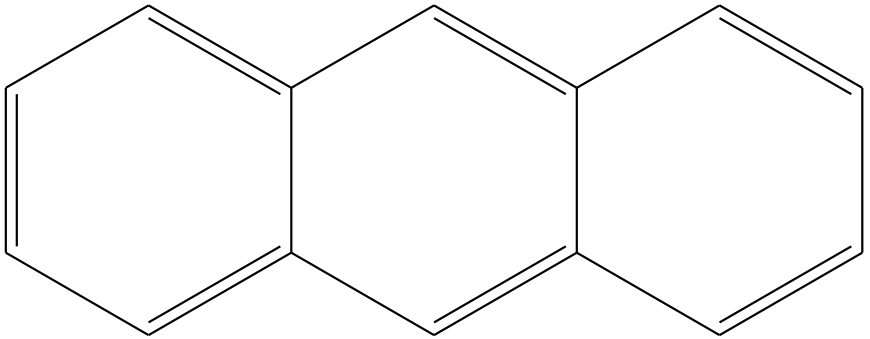
- The asterisks indicate functions converted to GPU already. The description of the legend is given in the main text. The first (warmup) sweep with $D = 512$ low bond dimension is not shown in the the plots.

- Currently, we save some 80-90 USD per hour.

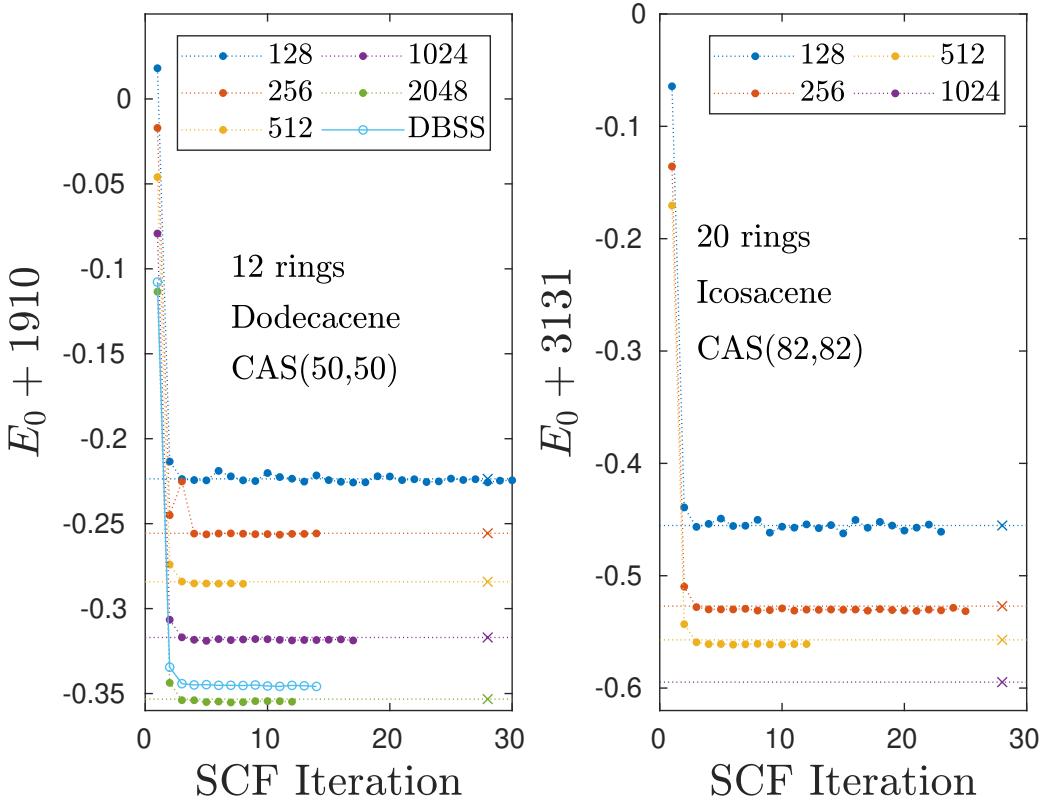
- Extensions using several powerful compute units is straighforward.

Spin adapted DMRGSCF on NVIDIA DGX-A100/H100

ÖL, Menczer,Ganyecz,Kapas,Werner,Hammond,Xantheas,Ganahl,Neese (JCTC 2025)



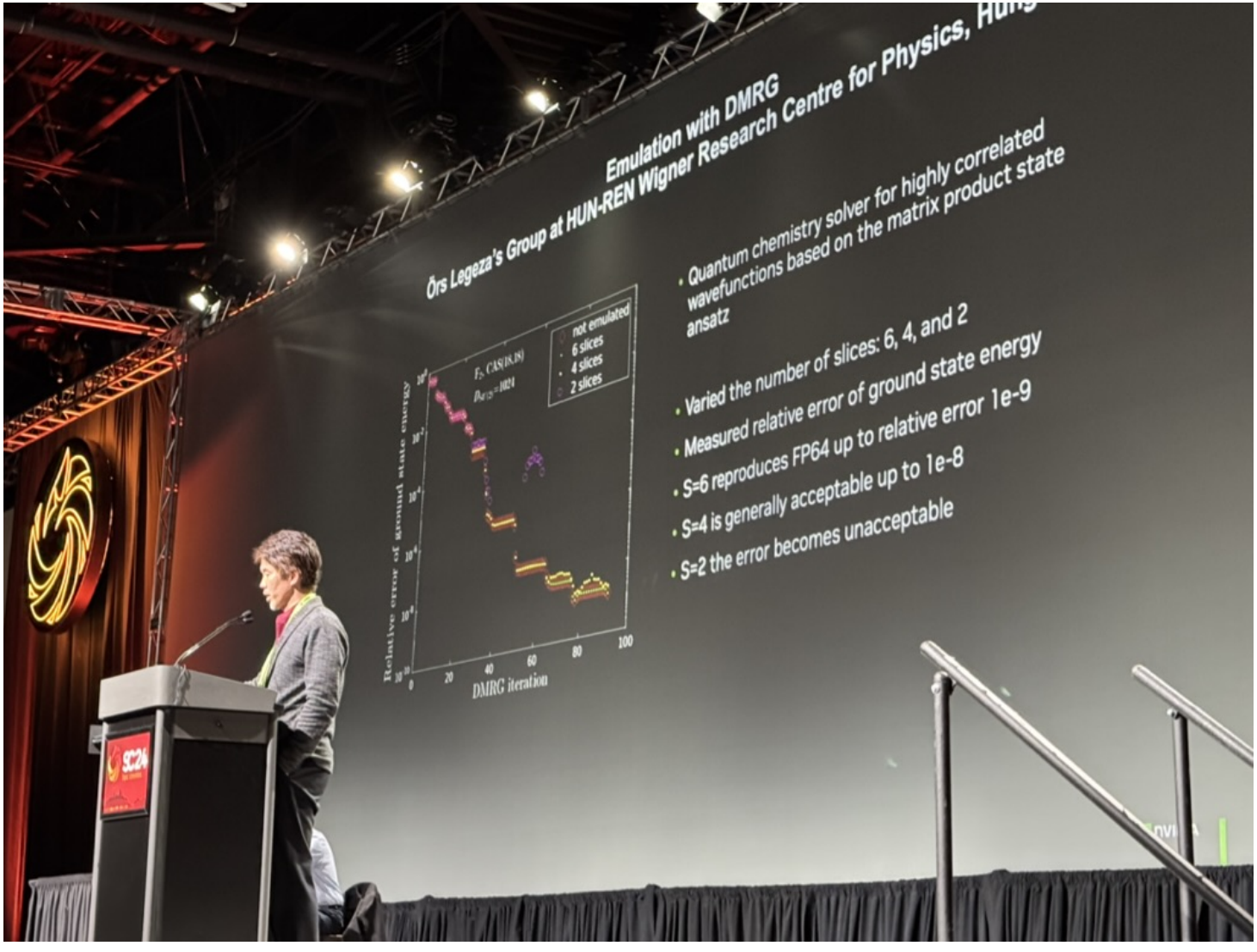
Polycyclic aromatic hydrocarbons



- For Heptacene, CAS(30,30), 25 DMRGSCF iterations with $D = 256$ using other codes took $\sim$ 7 days
$\rightarrow \sim$ 3.8 hours with our hybrid DMRG + ORCA
- Dodecacene CAS(50,50)
- Full orbital space:
328 electrons on 840 orbitals.
- Wall time: 13.3 hours ($D = 512$).

- Icosacene CAS(82,82)
- Full orbital space:
536 electrons on 1368 orbitals.
- Wall time: $\sim$ 1.2 days ($D = 512$).
- Use DMRG-TCC (DLPNO).

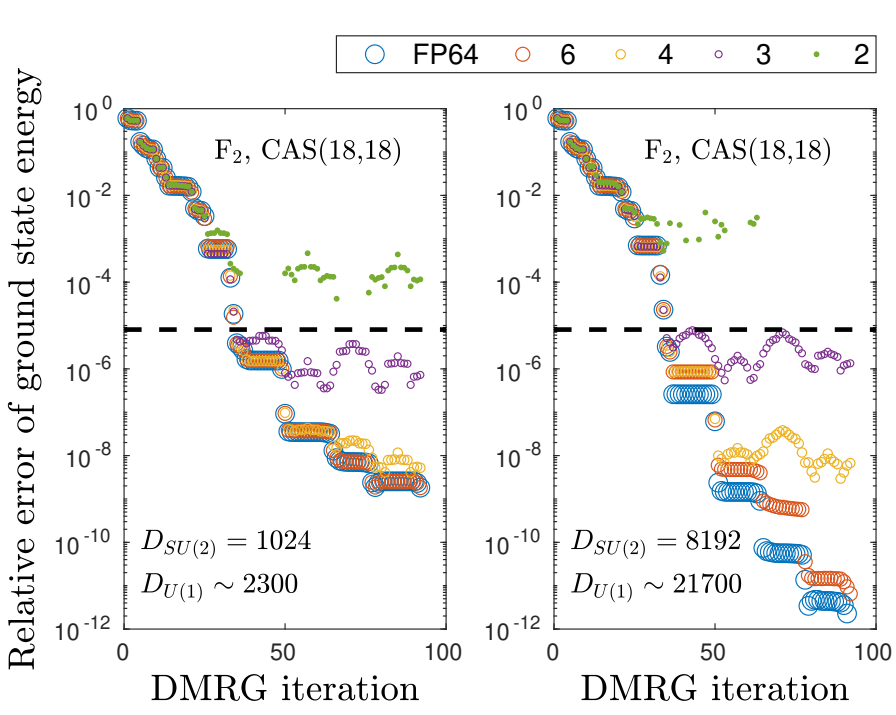
Mixed precision ab initio tensor network state methods adapted for NVIDIA Blackwell technology via emulated FP64 arithmetic (Rio Yokota at SC-2024, BoF TOP500)

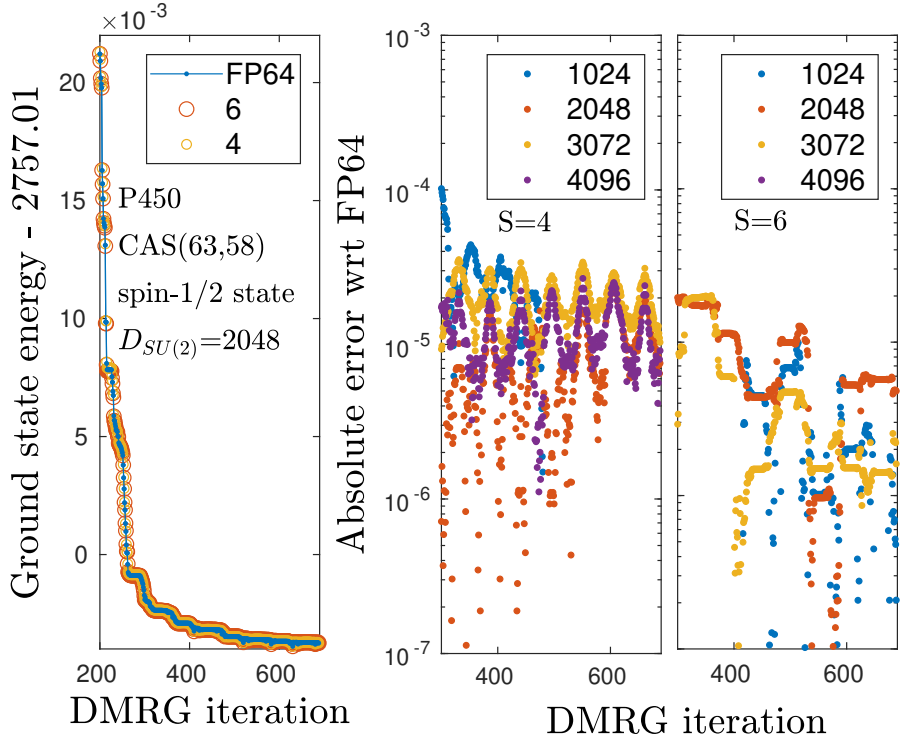


Mixed precision ab initio TNS methods adapted for NVIDIA Blackwell technology via emulated FP64 arithmetic

J. Gunnels, C. Brower, S. R. Bernabeu, J. Hammond, S. Xantheas, M. Ganahl, A. Menczer, Ö.L.

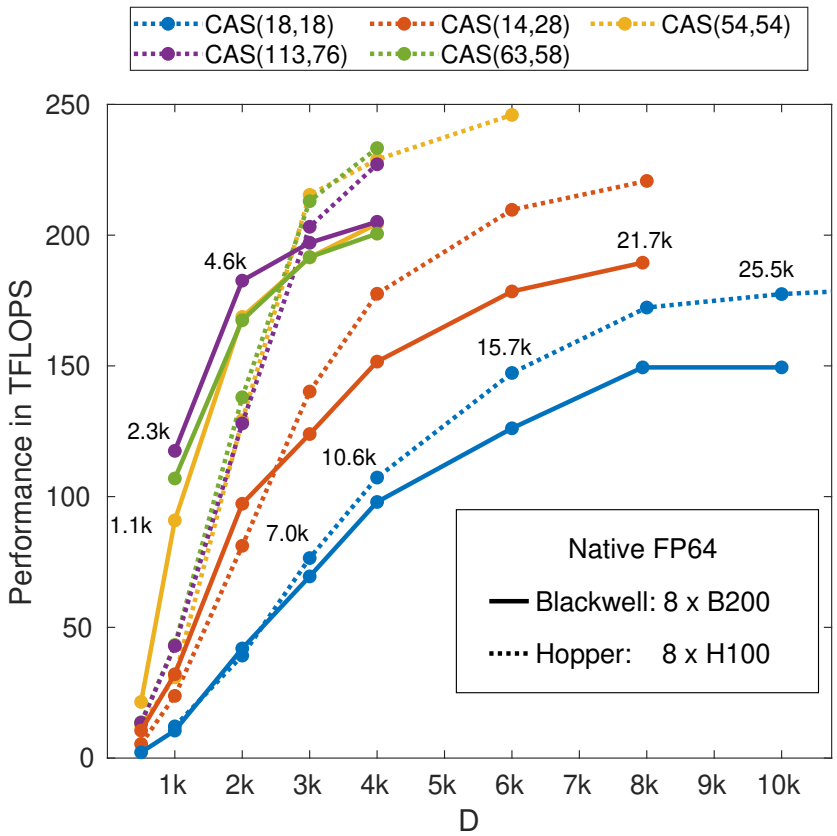
- Results obtained on DGX-B200 single node utilizing the Ozaki scheme
- Results obtained via early access utilizing a pre-release cuBLAS binary, and the data is subject to change.
- mantissa bit setting $\{15, 23, 31, 39, 47, 55\}$ for $S = 2, 3, 4, 5, 6, 7$ slices.

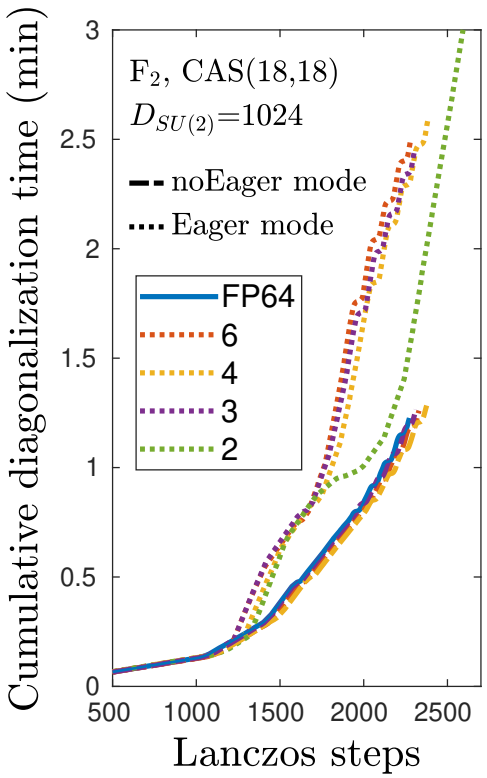


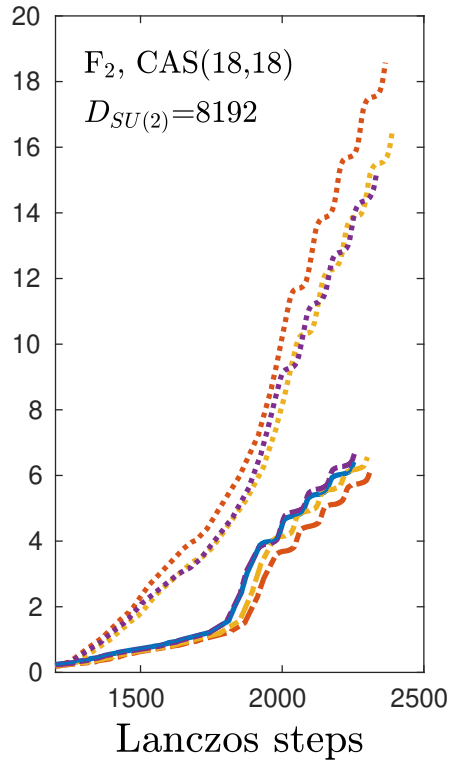


- Chemical accuracy, 1.6mHa, can be reached with 4, 6 slices

Performance assessment





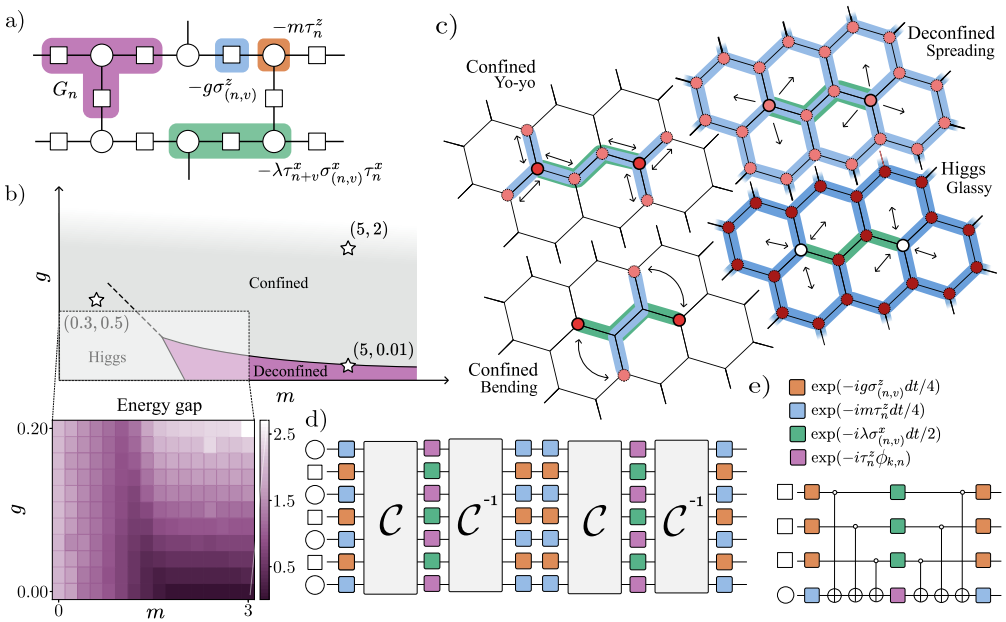


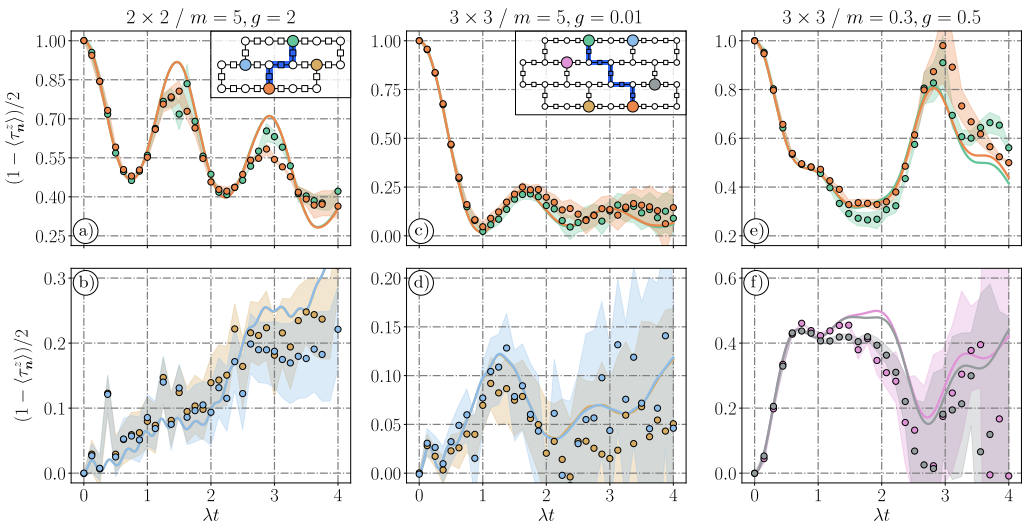
- Benchmark on DGX-H100 and DGX-B200 via native FP64
- Remark: with emulation we expect B200 to be faster in the near future (but currently don't have data)
- Most recent cuBLAS offers native FP64 and various options for emulation (native, emulated, mantissa bit setting, eager)

Simulation on IBM quantum chip vs DMRG/BUG

UBC,DIPC,IKERBASQUE,Wigner,IBM,CERN (2025)

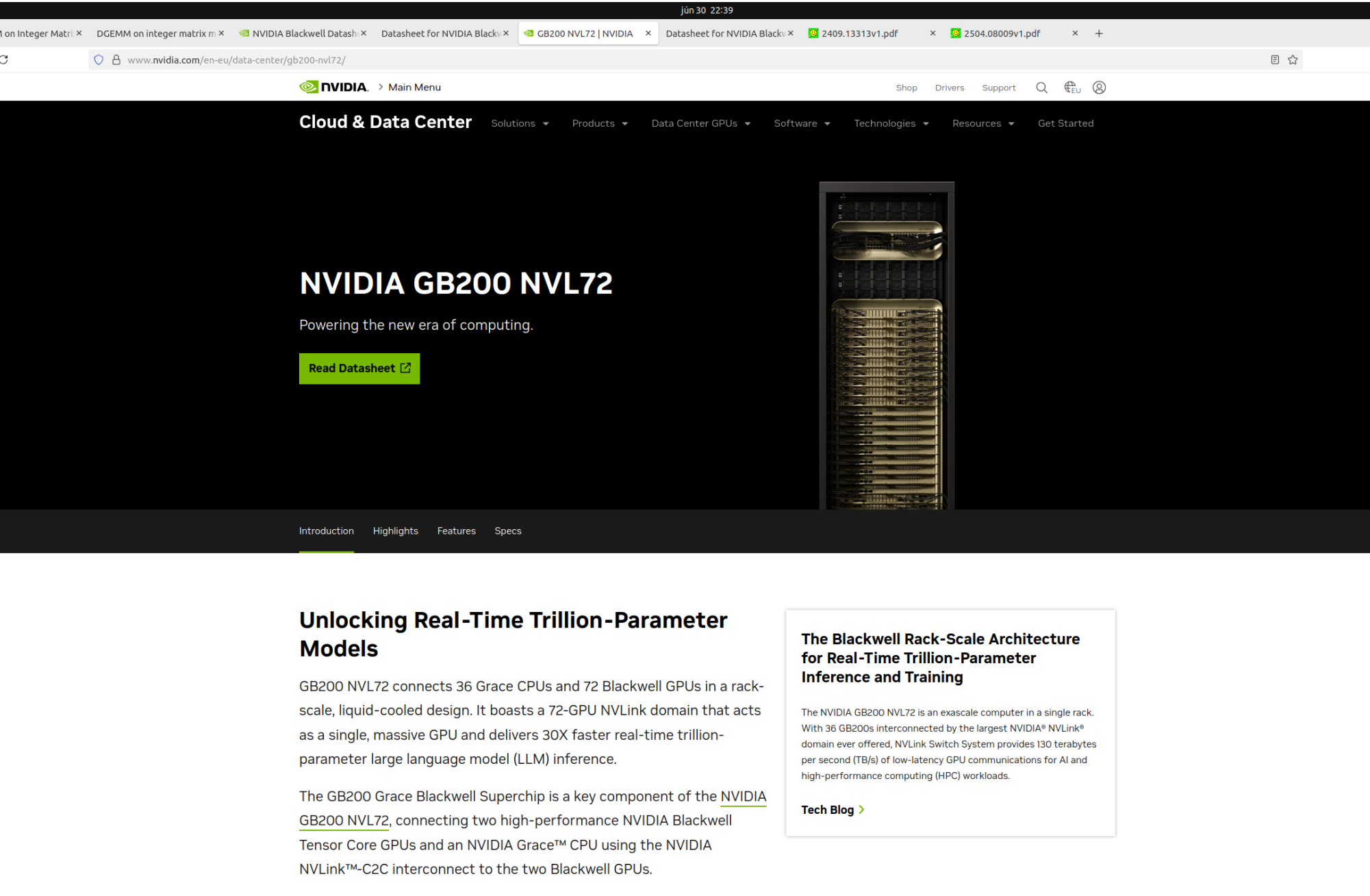
- Z_2-Higgs lattice gauge theory (hadronization, meson excit., topological effects)
- IBM superconducting quantum processor with up to 156 qubits
- Hexagonal quantum chip topology, error mitigation to reduce noise
- Basis Update Galerkin (BUG) novel TNS algorithm for time evolution





- Perfect agreement with simulation on real hardware up to 68 qubits
- For more qubits noise is too large on real hardware
- BUG can be used in quantum chemistry as well

New TNS benchmarks for quantum computing ???



Conclusion and near future on GH200, MI300, GB200 etc

- Tensor topologies together with proper basis representations are important for efficient data sparse representaion of the wavefunction
- Global and local mode optimization for tree-like TNS provides black-box tool to reduce computatinal complexity
- Long time evolution with adaptive mode transformation is a promising direction in collab. Eisert, Lübich
- Massive Parallelization multiNode-multiGPU $\rightarrow$ exascale computation
- Mixed precision TNS on specialized new hardware with lower energy consumption
- Future: New TNS benchmarks via NVIDIA GB200 NVL72 for quantum computing ??? Converting our code to AMD-MI300 (HIP, ROCm)
- $\rightarrow$ Simulation of realistic material properties in collab. Riverlane, Furukawa

Supports: Hungarian Academy of Sciences, the Hungarian National Research, Development and Innovation Office TKP2021-NVA-04, Quantum Information National Laboratory of Hungary, Alexander von Humboldt Foundation (Germany), Hans Fischer Senior Fellowship programme (IAS-TUM, Germany), SPEC, DOE, (PNNL, USA)