Tensor network state methods for molecular and nuclear systems on high performance computing infrastructures: a journey from academic 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

Many-Body Theory: Nuclear Physics Meets Quantum Chemistry
Mainz, 27.08.2026

in collaboration with

Our computer program package is used by more than 40 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 (HLRS), Germany
- Jülich Supercomputing Center (JSC), Germany
- National Energy Research Scientific Computing Center (NERSC), USA
- The Future of Computing Institute (FOCI), Rensselaer Polytechnic Institute (RPI), USA

Recently there is also an increasing interest by industrial partners:

- NVIDIA, USA
- IBM, USA
- Riverlane LTD, UK
- Furukawa Electric Institute of Technology, Japan
- SandboxAQ, USA (Google)
- AMD, USA
- Mathworks, USA
- Quantum Motion, UK & Australia
- Dynaflex LTD, Hungary

Wigner, PNNL, TUM, NVIDIA, SandboxAQ joint press release

October 28, 2024 | Feature

Collaboration Speeds Complex Chemical Modeling

Advanced graphics processing units lead to unprecedented calculation speed for highly correlated electronic structure calculations

Media Contact: PNNL News & Media Relations

SHARE:

A recent collaboration among researchers from HUN-REN Wigner Research Centre for Physics in Hungary and the Department of Energy's Pacific Northwest National Laboratory, along with industry collaborators SandboxAQ and NVIDIA, has achieved unprecedented speed and performance in efforts to

Related Content

APRIL 30, 2026 | News Release

PNNL Researchers to Speak About Genesis Mission at National Science Bowl

April 23, 2026 | News Release

AI Accelerators Deliver Accurate Models for Challenging Quantum Chemistry Calculations

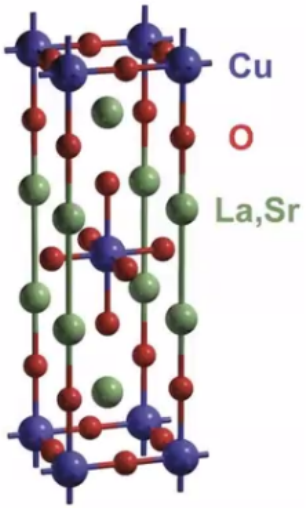
Graphic processing unit (GPU) AI accelerators show their worth in solving complex quantum chemical structures central to energy grand challenges

Karyn Hede, PNNL

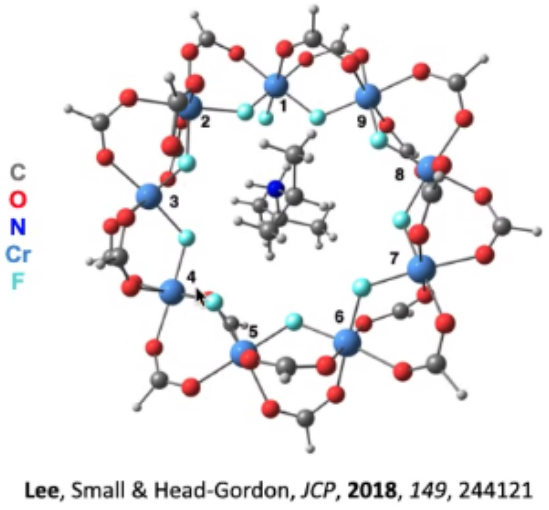
SHARE:

Related Content

Strong correlations between electrons used by nature and in new technologies



High-T_c superconductors

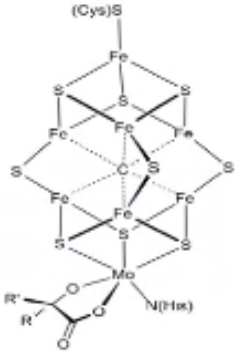


Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

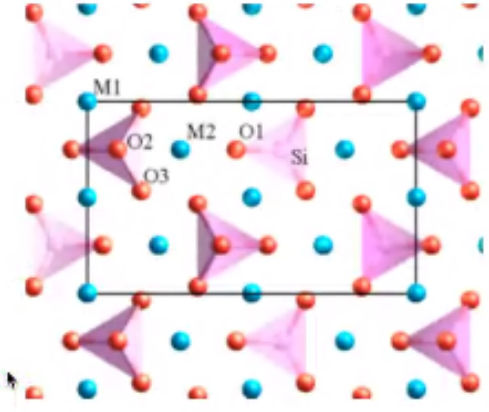
Single molecular magnets (SMM)

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

"FeMoco"



Nitrogen fixation



Battery technology

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

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

- 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 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 polynomials
- Major aim: to obtain the desired eigenstates of $\mathcal{H}$.

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

• Matrix ranks up to 100 000. Size of Hilbert space up to 10^9.

• In ab inito DMRG the CAS size is: 100 electrons on 100 orbitals.

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

• Embeddings: DMRG-SCF, DMRG-RAS, DMRG-TCC, ...

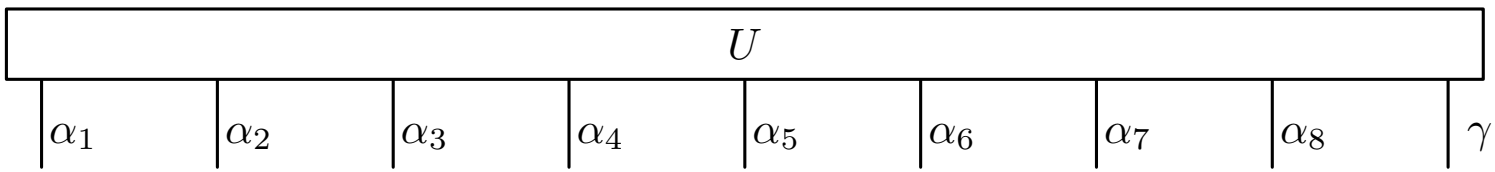
• Massively parallel implementations CPU/GPU$\rightarrow$ exascale on HPC

Tensor product approximation

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

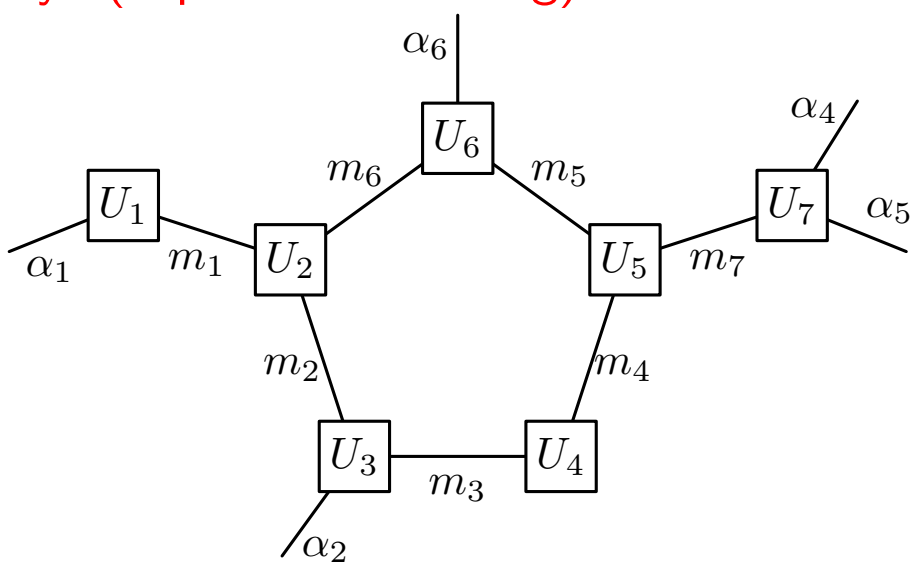
$$|\Psi_\gamma\rangle = \sum_{\alpha_1=1}^{q_1} \dots \sum_{\alpha_d=1}^{q_d} U(\alpha_1, \dots, \alpha_d, \gamma)\, |\alpha_1\rangle \otimes \cdots \otimes |\alpha_d\rangle \in \bigotimes_{i=1}^{d} \Lambda_i := \bigotimes_{i=1}^{d} \mathbf{C}^{q_i} \ ,$$

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



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

- $\alpha_i \in \{\uparrow, \downarrow\}$ or $\alpha_i \in \{0, \uparrow, \downarrow\uparrow\downarrow\}$
- We seek to reduce computational costs by parametrizing the tensors in some data-sparse representation using SVD.
- A general tensor network representation of a tensor of order 5.



Matrix product state (MPS) representation / DMRG / TT

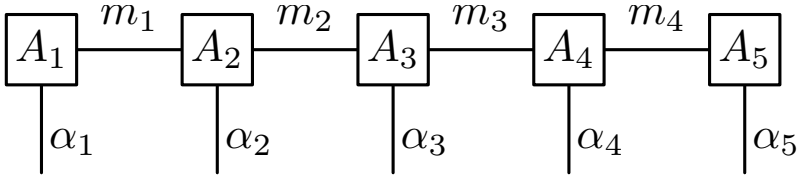
Exponential scaling $\rightarrow$ polynomial scaling

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

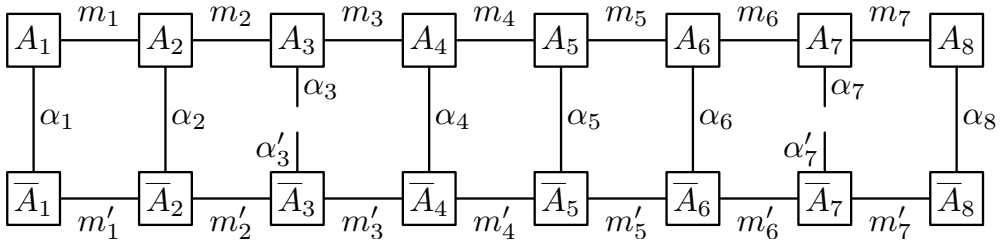
The tensor U is given elementwise as

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

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, $I = s_i + s_j - s_{ij}$. Block entropy: $S_{l \in \{1,2,3,\ldots\}}$

Vidal (02), Ö.L & Sólyom, (03), Rissler, Noack, White (06)

Global fermionic mode optimization via swap gates

Krumnow,Veis,ÖL,Eisert(2015-2016), Friesecke,Werner,Kapas,Menczer,ÖL(2024)

- Finding a parametrization with the minimum number of parameters for a given error margin is a task of utmost importance in modern quantum physics and chemistry
- Computatinal complexity $\sim$ block entropy area (BEA)

$$\mathcal{B}\alpha(C) = \sum_{\ell=1}^{N-1} S_\alpha(\rho_{1,2,\ldots,\ell})$$

$S_\alpha(\rho) = \frac{1}{1-\alpha} \ln(\mathrm{Tr}\,\rho^\alpha)$ is the Rényi entropy for some $0 < \alpha < 1$.

- New modes $\varphi_i' = \sum_j U_{ij}\varphi_j$,
- Thus stationarity of a scalar function f on $SO(N)$ at U_* is equivalent to

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

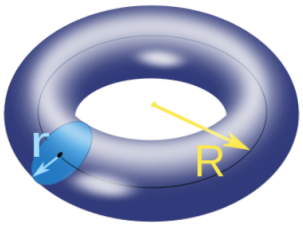
that is to say $\frac{\partial f}{\partial U}(U_*)U_*^T$ symmetric.

- Reduction to pairwise rotations: To achieve stationarity minimize f over all pairwise rotations $U_{ij}(\theta)$ using swap gates permutations.

- Mode transformation $\varphi_i' = \cos\theta\varphi_i + \sin\theta\varphi_j$, $\varphi_i' = -\sin\theta\varphi_i + \cos\theta\varphi_j$

2d-spinless Hubbard on a torus, 10×10 with $D = 80$

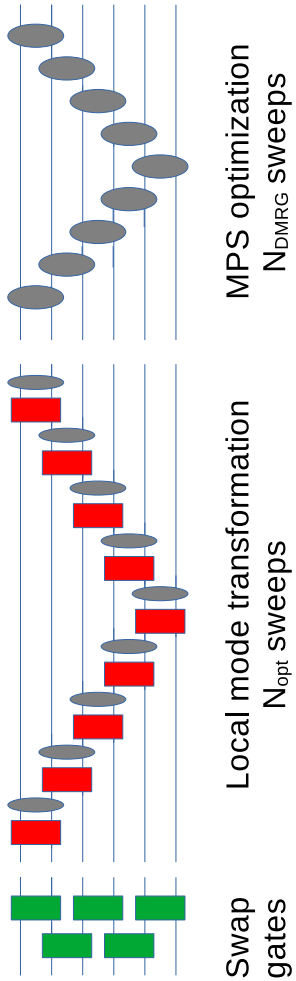
$$H = \sum_{\langle i,j \rangle} c_i^\dagger c_j + \sum_{\langle i,j \rangle} V n_i n_j,$$

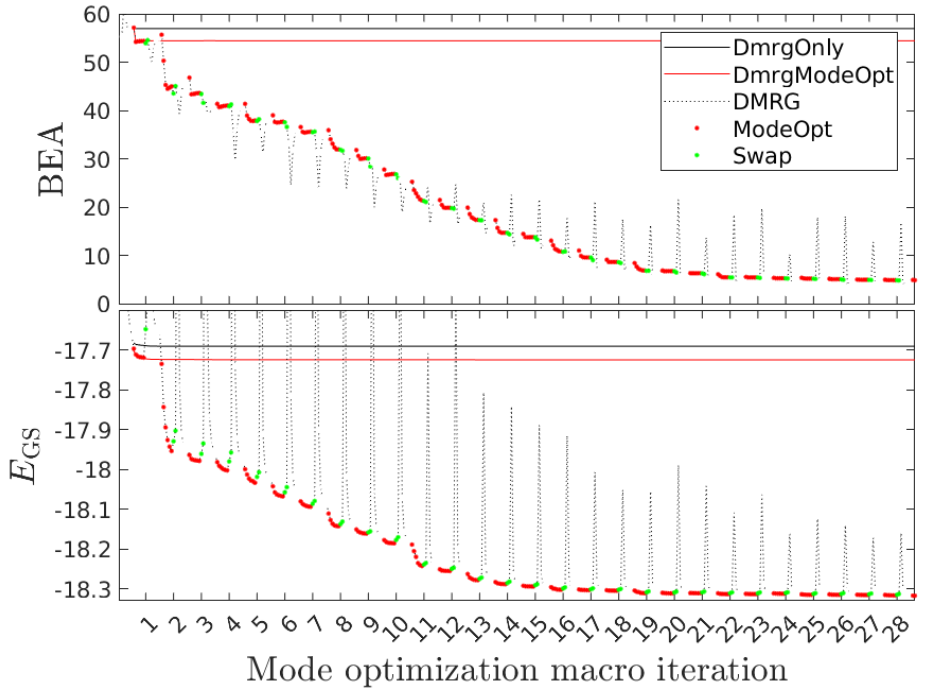


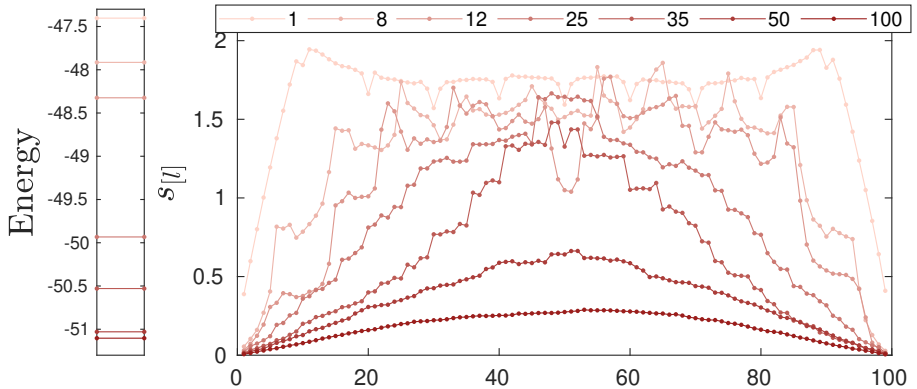
Joint optimization on the MPS and Grassman manifolds:

$$H = \sum_{i,j=1}^{n} t_{i,j} c_i^\dagger c_j +$$

$$\sum_{i,j,k,l=1}^{n} v_{i,j,k,l} c_i^\dagger c_j^\dagger c_k c_l,$$

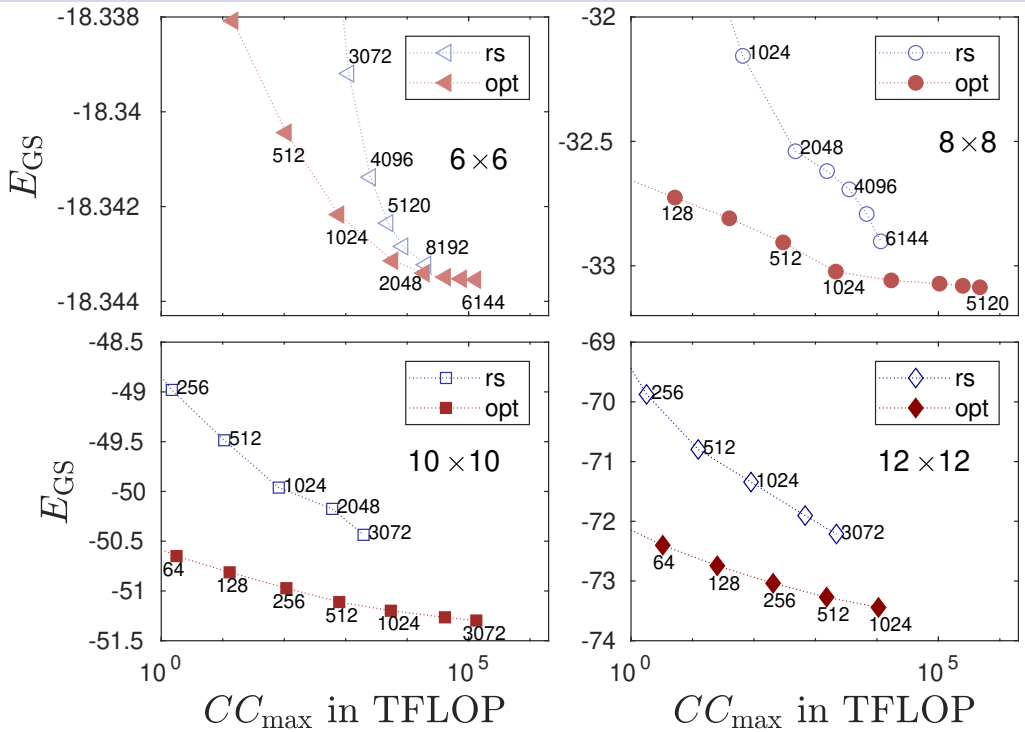






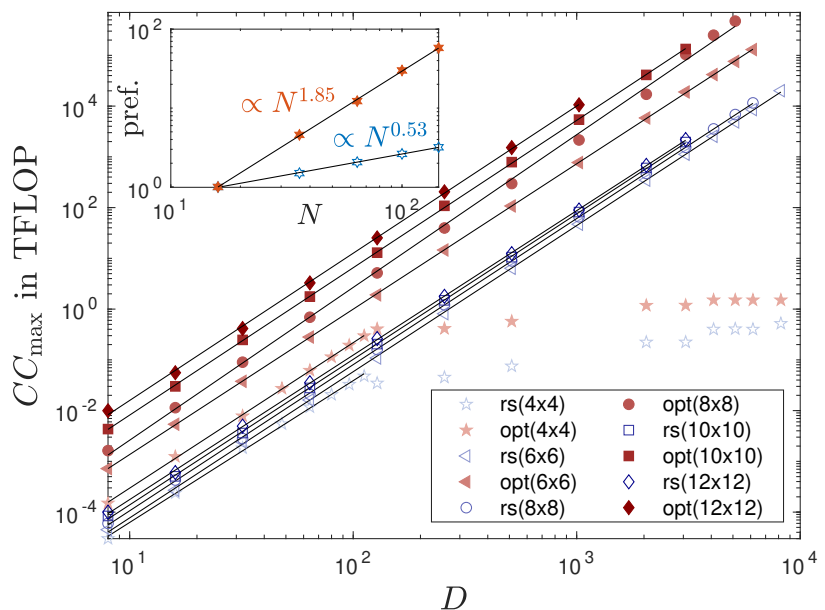
- Two-qubit gate disentanglers by minimizing half-Rényi entropy
- N/2 re-orderings of modes via two-qubit swap gate layer
- Computational complexity can be reduced by orders of magnitudes!

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.

Coupled cluster method with single and double excitations tailored by matrix product state wave functions

L. Veis, A. Antalik, F. Neese, Ö.L., J. Pittner (2016)

- Formally single reference theory, Fermi vacuum is a single determinant
- Split-amplitude ansatz

$$\Psi_{\mathrm{TCC}} = e^{T}\Psi_{\mathrm{ref}} = e^{T^{\mathrm{ext}}+T^{\mathrm{CAS}}}\Psi_{\mathrm{ref}}$$

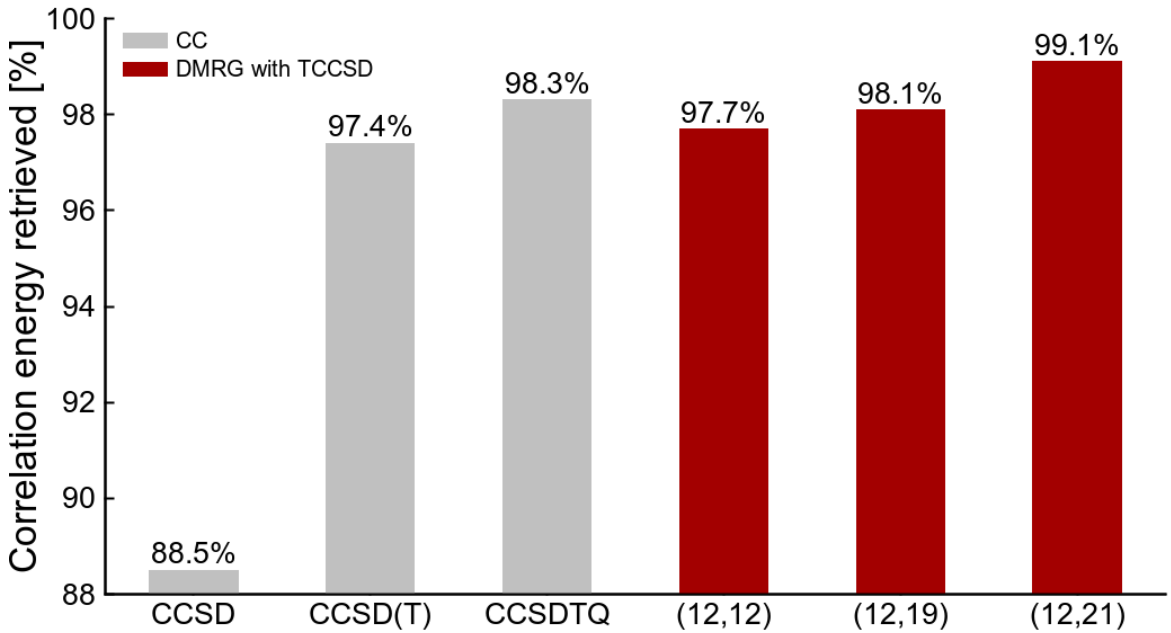
- T^{CAS}
  - amplitudes extracted from DMRG (CASCI) calculation
  - frozen during CC calculation
  - account for static correlation
- T^{ext}
  - determined through the usual CC
  - account for dynamic correlation

$$\begin{aligned}\Psi_{\mathrm{TCCSD}} &= e^{\left(T_1^{\mathrm{ext}}+T_2^{\mathrm{ext}}\right)} e^{\left(T_1^{\mathrm{CAS}}+T_2^{\mathrm{CAS}}\right)}\Psi_{\mathrm{ref}} \\ &\approx e^{\left(T_1^{\mathrm{ext}}+T_2^{\mathrm{ext}}\right)}\Psi_{\mathrm{CASCI}}\end{aligned}$$

- Requires only small modifications of the CC code

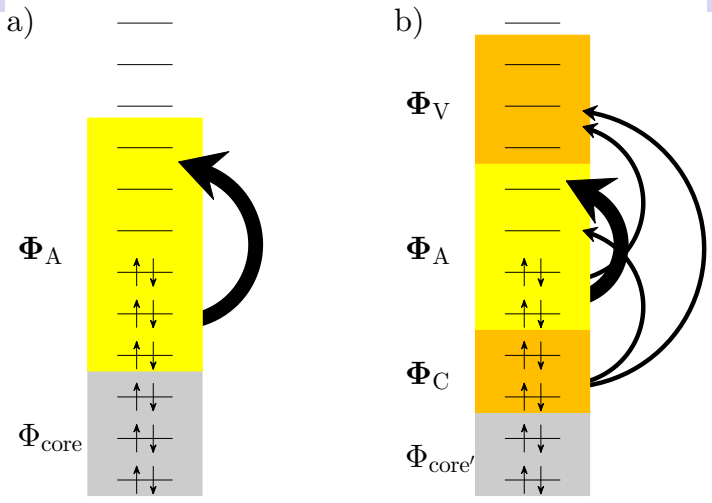
Chromium dimer – correlation energies

- Single-point calculation at 1.5 Å
- One-particle basis: RHF with Ahlrichs' SV basis set → (48e,42o)
- DMRG space selected based on $S^{(1)}$ profile
- DMRG performed with DBSS ($\epsilon_{\mathrm{tr}} \approx 10^{-7}$)
- Extrapolated DMRG by Olivares-Amaya et al. JCP 142, 034102, 2015 serves as a FCI benchmark

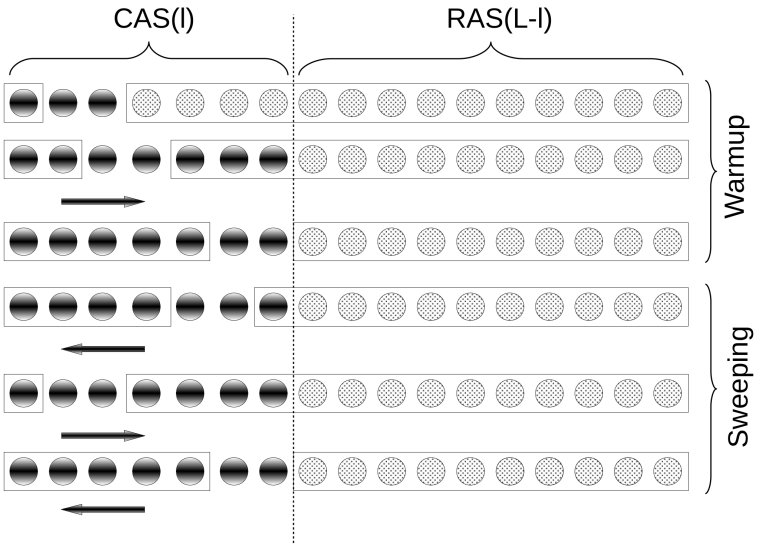


Error analysis: Faulstich, Laestedius, ÖL, Schneider, Kvaal: Quadratic error bound for a given CAS-EXT split.
Extensions: Jiri Pittner on similarity transformed TCCSD, Andrej Antalik on LPNO-TCCSD and Jan Brandejs on 4c-DMRG-TCCSD.

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 k.

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

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

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

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

Rigorous mathematical analysis Friesecke, Barcza, Ö.L. (2022)

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

$$H = H_0 + H' \text{ with}$$

$$H_0 = PHP + (E_0 + \Delta)Q, \quad H' = H - PHP - (E_0 + \Delta)Q$$

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, and E_0 is the CAS ground state energy, i.e. $E_0 = E^0_{\mathrm{CAS}}(\ell)$ with $\Delta > 0$.

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

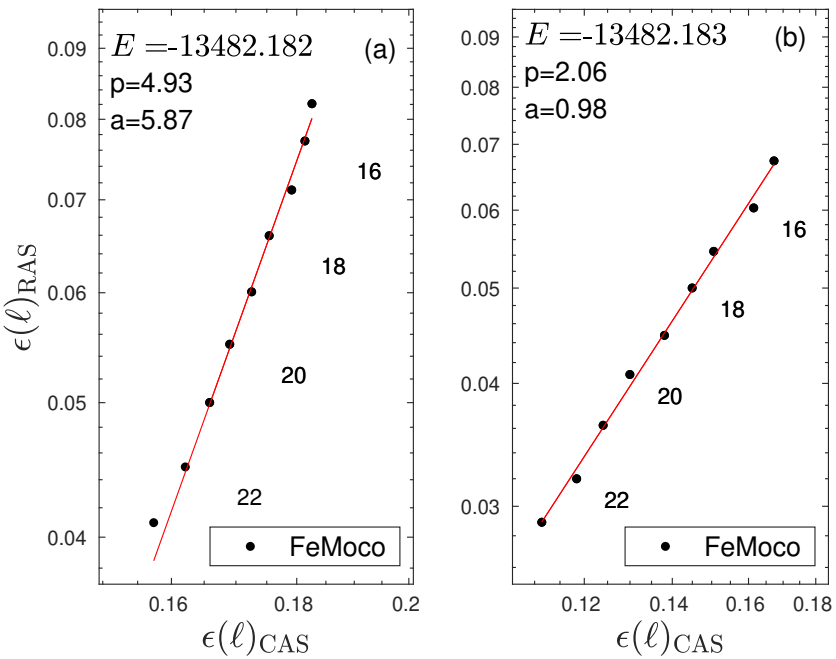
- DMRG-RAS can be considered an *embedding method*.
- Modified perturbation theory: optimal Δ is the *expected value of the spectral gap between the pure RAS eigenvalues and the CAS ground state*
- DMRG-RAS is a fully self-consistent method, and therefore capable of capturing more than the guaranteed perturbation contributions
- Example: Cr_2 CAS(12,68), $E^0(17,2) = -2086.8769$ is already below the CCSDTQ by 8×10^{-3}

- New extrapolation procedure to get FCI energy: DMRG-RAS-X:

$$E^0(\ell, 2) - E^{\mathrm{FCI}} = a\left(E^0_{\mathrm{CAS}}(\ell) - E^{\mathrm{FCI}}\right)^p \text{ for some } p > 1.$$

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=10240,NO)	-13482.1754
RAS(23,NO)	-13482.1544

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



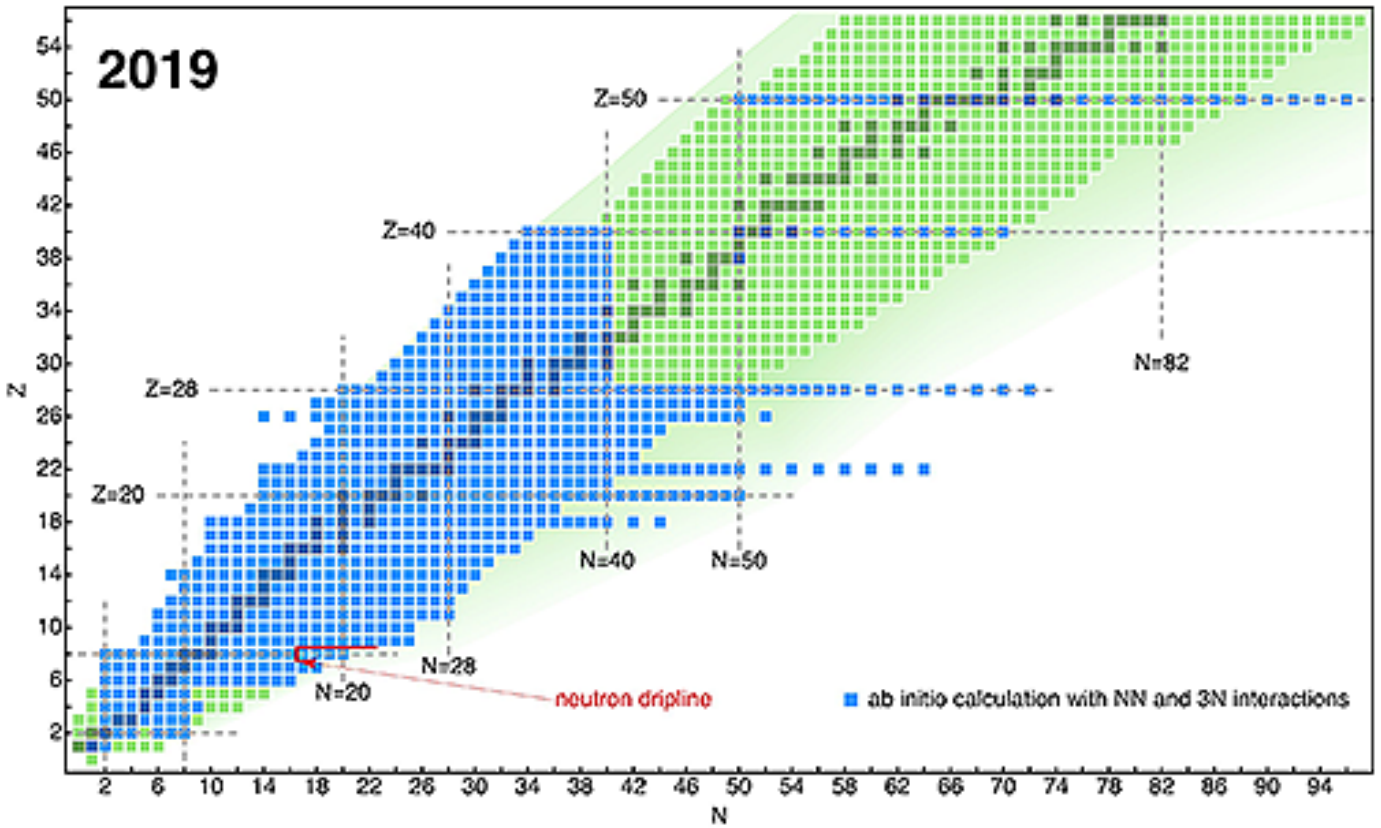
(a) Result of the DMRG-RAS-X for the FeMoco for the model space taken from Ref. Reiher(2007).

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

Produced on CPU-GPU for less than one day Friesecke, Barcza, ÖL (2023)

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

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



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

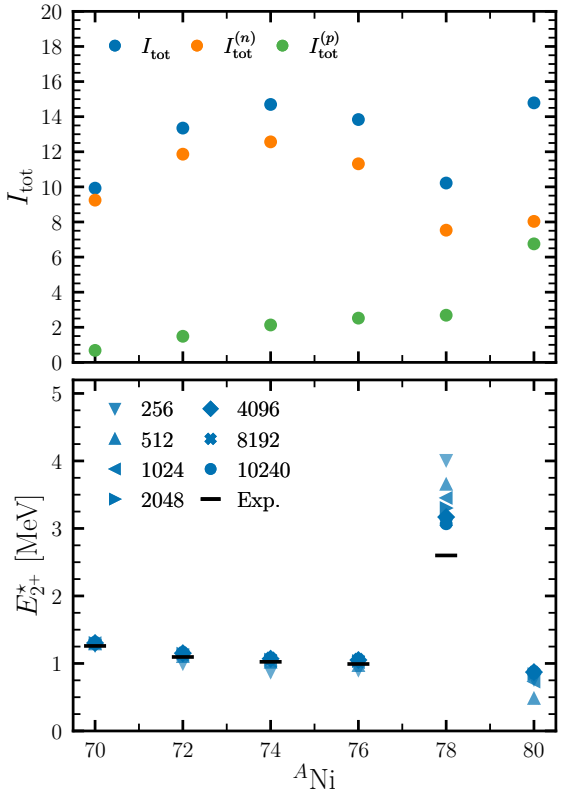
- where $c_{\alpha}^{\dagger}$ and c_{α} creates and annihilates a particle with quantum numbers $\alpha = (n, l, j, m, \tau_z)$. $j \geq 1/2$, Isospin,
- no-core shell models
- effective Hamiltonian including parts of 3-body interactions

In-medium similarity renormalization group & DMRG

- Starting with two-nucleon (NN) and three-nucleon (3N) interactions, the VS-IMSRG generates a valence-space-decoupled Hamiltonian restricted to an active space of limited size.

- Many-body operators of higher particle rank are truncated at the normal-ordered two-body level, defining the IMSRG(2) truncation.

- Neutron-rich nickel isotopes attract significant experimental attention, e.g., with the recent discovery of the doubly magic nature of ^{78}Ni



- Neutron, proton, and total entropies (top) and 2^+ excitation energies (bottom) along even-mass nickel isotopes.

- Entropies are calculated at bond dimension $M = 10240$ whereas for the excitation energies the bond dimension was varied between $M = 256 - 10240$.

Tichai,Knecht,Kruppa,Ö.L.,Moca,Schwenk,Werner,Zarand (2022)

- Current work: deformation, clustering

Spectroscopy of N=50 isotones with the VS-DMRG

Tichai,Kapás,Miyagi,Werner,Ö.L.,Schwenk,Zarand(2024)

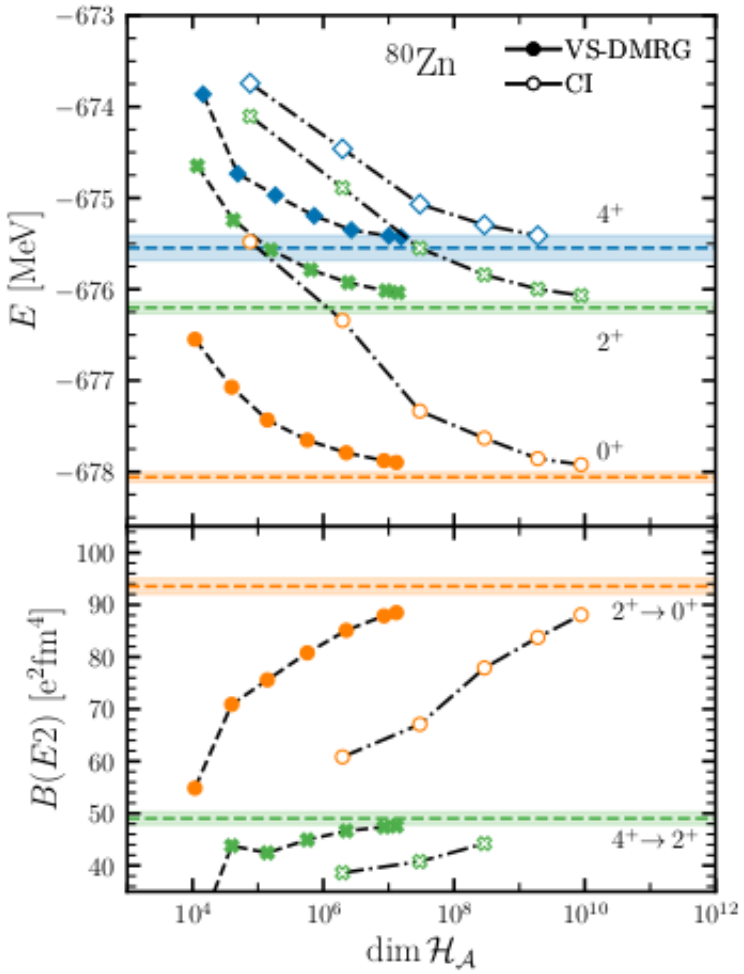


Figure 1: Convergence of VS-DMRG and CI energies (top panel) and $B(E2)$ transition strengths (bottom panel) of the lowest 0^+, 2^+, and 4^+ states and their transitions as a function of Hilbert-space dimension for ^{80}Zn. The largest CI dimension encountered at $T_{\max} = 7$ is given by dim $\mathcal{H}_\mathcal{A} = 10^{9.94} = 8.8 \times 10^9$. The VS-DMRG extrapolated energies and transition strengths (and their uncertainties) are shown by the horizontal bands, see text for details.

Current DMRG code for ^{80}Zn on NVIDIA Blackwell node:

- D=10240 takes 4-5 hours, $E = -677.885$
- D=20480 takes 13 hours (yesterday), $E = -677.9242$
- D=30720 running
- Excited states
- Expection values etc
- Looking for large scale interesting problems

Hunting for quantum advantage in electronic structure calculations is a highly non-trivial task, arXiv:2603.28648

ÖL, Menczer, Werner, Xantheas, Neese, Ganahl, Brower, Bernabeu, Hammond, Gunnels

- Due to developments over the past decades in both quantum computing and simulations on classical hardware, it is a serious challenge to identify a real-world problem where quantum advantage is expected to appear.

- Fe_4S_4 molecular cluster on a CAS(54,36) model space (72 qubits) was included in the *Quantum Advantage Tracker* webpage maintained by IBM and RIKEN. Our CPU-GPU DMRG solution takes $\sim$ 8 hours only.

- We propose new benchmark: CAS-SCF based orbital optimizations for [CAS(89,102)] size for the $Fe_5S_{12}H_5^4$ molecular system (204 qubits).

- It comprises twenty five open shell orbitals in its sextet ground state and an active space size of 331 electrons in 451 orbitals, i.e. 902 qubits.

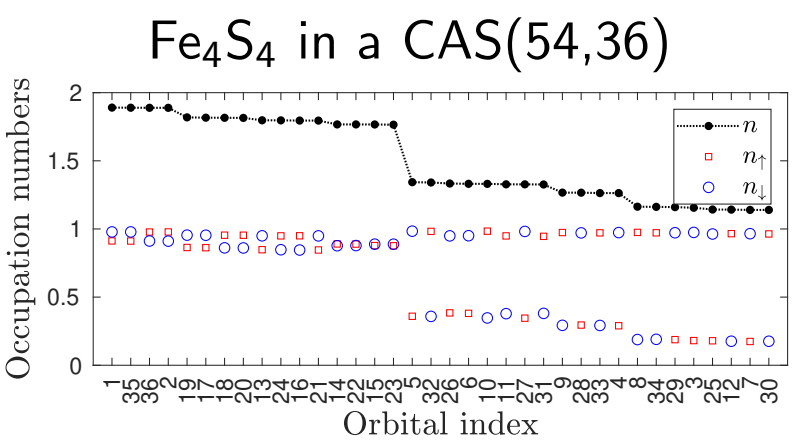
- Achieved via our mixed-precision spin-adapted *ab initio* Density Matrix Renormalization Group (DMRG) interfaced with the ORCA program on single node NVIDIA Blackwell GPU ($\sim$ 220 TFLOPS FP64).

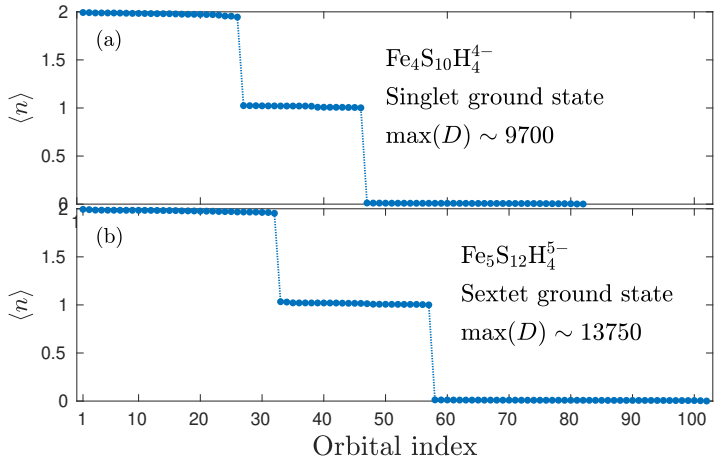
- Future: NVIDIA GB300 NVL72 up 576 GPUs and at JSC/RPI/Empire AI/Argonne NL $\sim 72/288/10,000/100,000$ GB300 GPUs.

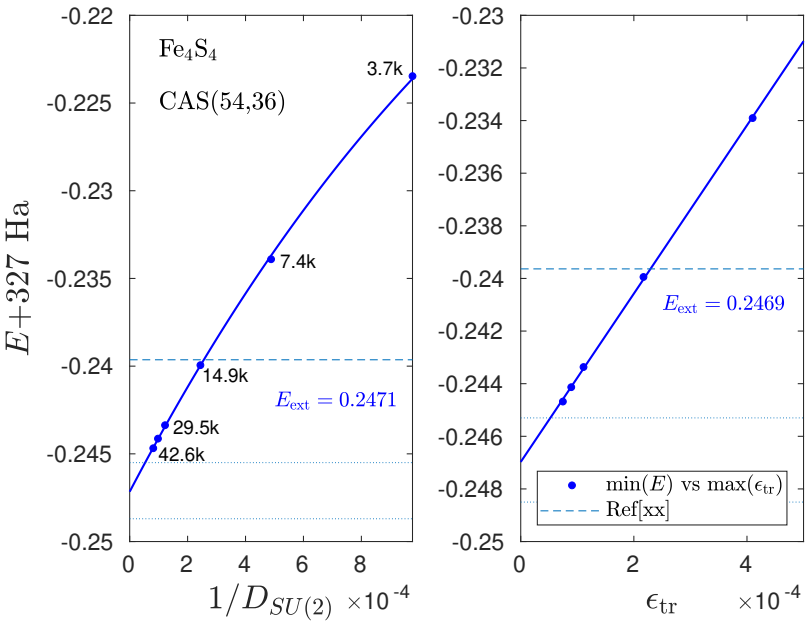
Classical vs quantum simulation? (DMRG-SCF on DGXB200)

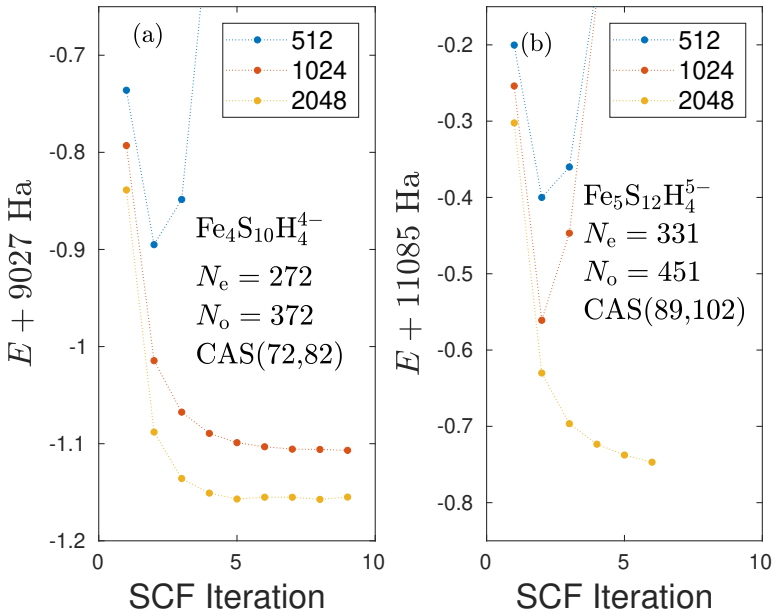
ÖL, Menczer,Werner,Xantheas,Neese,Ganahl,Brower,Bernabeu,Hammond,Gunnels (2026)

- Fe_4S_4 CAS(54,36) included quite recently in the Quantum Advantage Tracker webpage maintained by IBM and RIKEN.



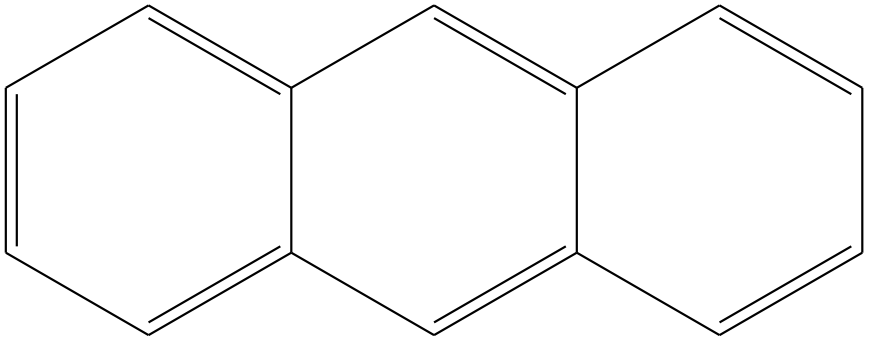




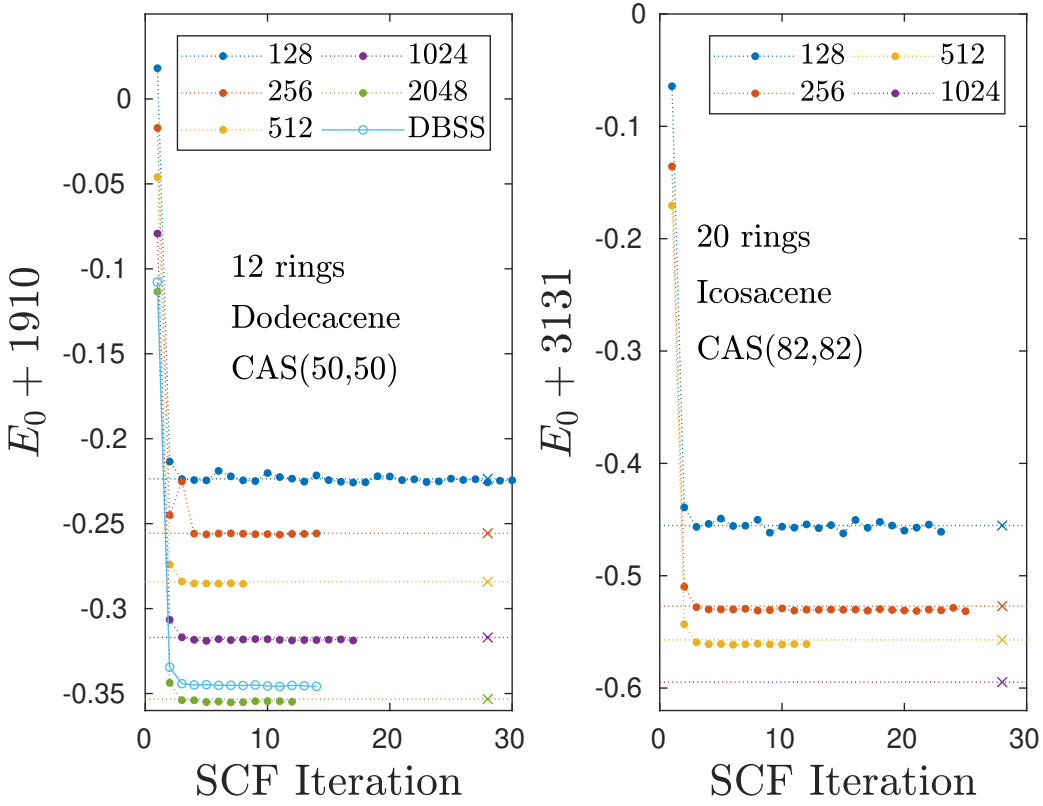


Self Consistent Field DMRG on NVIDIA DGX-H100/B200

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).

Introduction to BUG

General non-linear PDE

$$\dot{Y} = \Phi(Y) \qquad Y = Y_{ij} \in \mathbb{C}_{D_1 \times D_2}$$

Low-rank approximation

$$Y = USV^\dagger$$

$D_1 \times M$ $M \times M$ $M \times D_2$

Naive low-rank equations of motion

$$\dot{U} = \Phi(USV^\dagger)VS^{-1}$$
$$\dot{S} = U^\dagger\Phi(USV^\dagger)V$$
$$\dot{V}^\dagger = S^{-1}U^\dagger\Phi(USV^\dagger)$$

$\Longleftrightarrow$ Naive discretization is hard due to possibly large curvature (S^{-1})

- U and V are (partial) isometries, with orthonormalized columns
- Similar to SVD, but S is not necessarily diagonal

New variables: **No size increase!!**

$$X = US, \; Z = SV^\dagger$$

$$\dot{X} = \Phi(XV^\dagger)V$$
$$\dot{Z} = U^\dagger\Phi(UZ)$$

No S^{-1}

We can go through high curvature points without even noticing them!

Discretized scheme

$$\tilde{X} = X + \delta t\Phi(XV^\dagger)V$$
$$\tilde{Z} = Z + \delta tU^\dagger\Phi(UZ)$$
$$\hat{S} = S + \delta tU^\dagger\Phi(USV^\dagger)V$$

$\{\tilde{U}, \tilde{V}\}$ From QR decompositions of $\{\tilde{X}, \tilde{Z}\}$

$\tilde{S} = \tilde{U}^\dagger U\hat{S}V^\dagger\tilde{V}$ Basis transf.

[2] Gianluca Ceruti, Christian Lubich, and Dominik Sulz, SIAM Journal on Numerical Analysis 61, 194-222 (2023).

Quantum simulation on Q-chip (IBM Heron r2)

J. Cobos, J. Fraxanet, C. Benito, F. di Marcantonio, P. Rivero, K. Kapás, M. A. Werner, Ö. L., A. Bermudez, E. Rico

General Q-algorithms usually require arbitrary 2-qubit gates ⟺ Q-chip lattice grid

(swap gates are necessary)

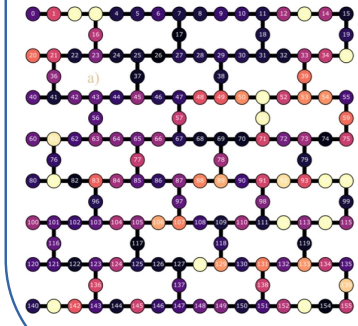
⇕

Noise limits circuit depth

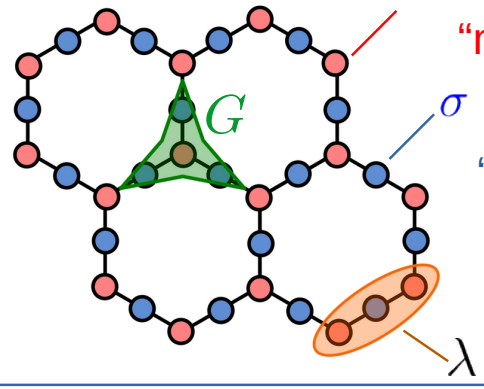
Idea: simulate a 2D lattice model compatible with the chip at hand!

(2+1)D Z_2 Higgs model on the honeycomb lattice

$$H = -m\sum_{\boldsymbol{n}}\tau^z_{\boldsymbol{n}} - g\sum_{(\boldsymbol{n},\boldsymbol{v})}\sigma^z_{(\boldsymbol{n},\boldsymbol{v})} - \lambda\sum_{(\boldsymbol{n},\boldsymbol{v})}\tau^x_{\boldsymbol{n}+\boldsymbol{v}}\sigma^x_{(\boldsymbol{n},\boldsymbol{v})}\tau^x_{\boldsymbol{n}}$$



⟺



$\tau \in \{\uparrow,\downarrow\} \Leftrightarrow N \in \{0,1\}$

"matter"

$\sigma \in \{\uparrow,\downarrow\} \Leftrightarrow E \in \{0,1\}$

"field"

Gauge invariance

$$G_n = \tau^z_n\sigma^z_{(n,1)}\sigma^z_{(n,2)}\sigma^z_{(n,3)}$$

$$[G_n, H] \equiv 0\ \forall n$$

"physical" subspace:

$$G_n|\Psi\rangle = +1|\Psi\rangle\ \forall n$$

Z2 Gauss law

$$N_n = \sum_v E_{(n,v)}\ \mathrm{mod}2$$

Absent plaquette term:

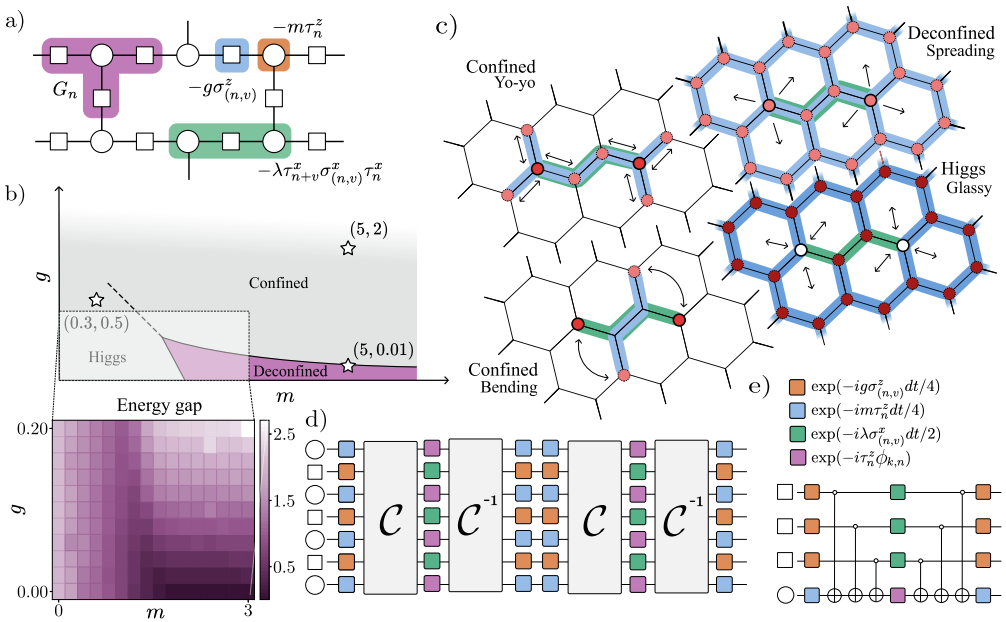
$$H_{\mathrm{P}} = J_{\mathrm{P}}\sum_{\mathrm{P}}\prod_{(n,v)\in\mathrm{P}}\sigma^x_{(n,v)}$$

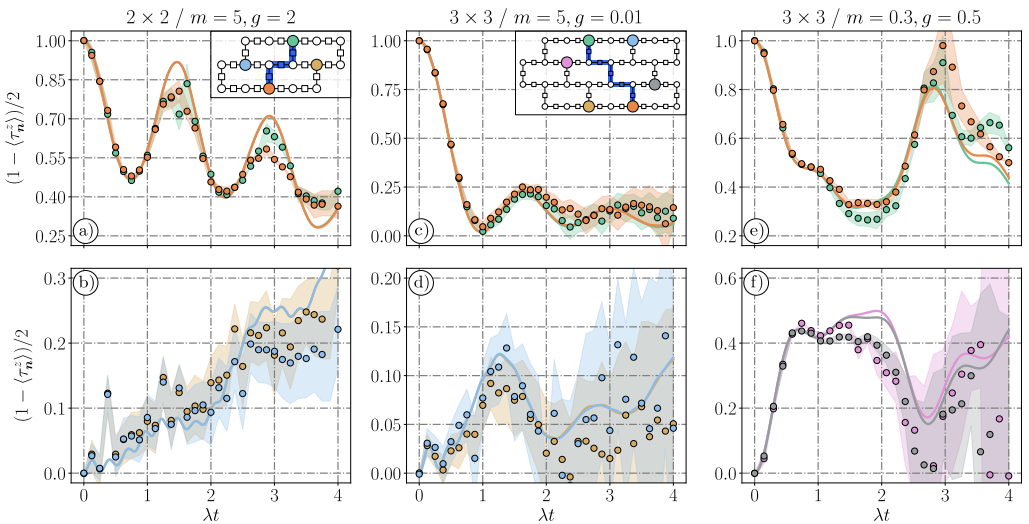
- Would generate direct field line fluctuations
- Is still generated perturbatively $J_{\mathrm{P}}^{\mathrm{eff}} \approx 0.25\frac{\lambda^6}{m^5}$

Simulation on IBM Heron r2 quantum chip vs DMRG/BUG

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

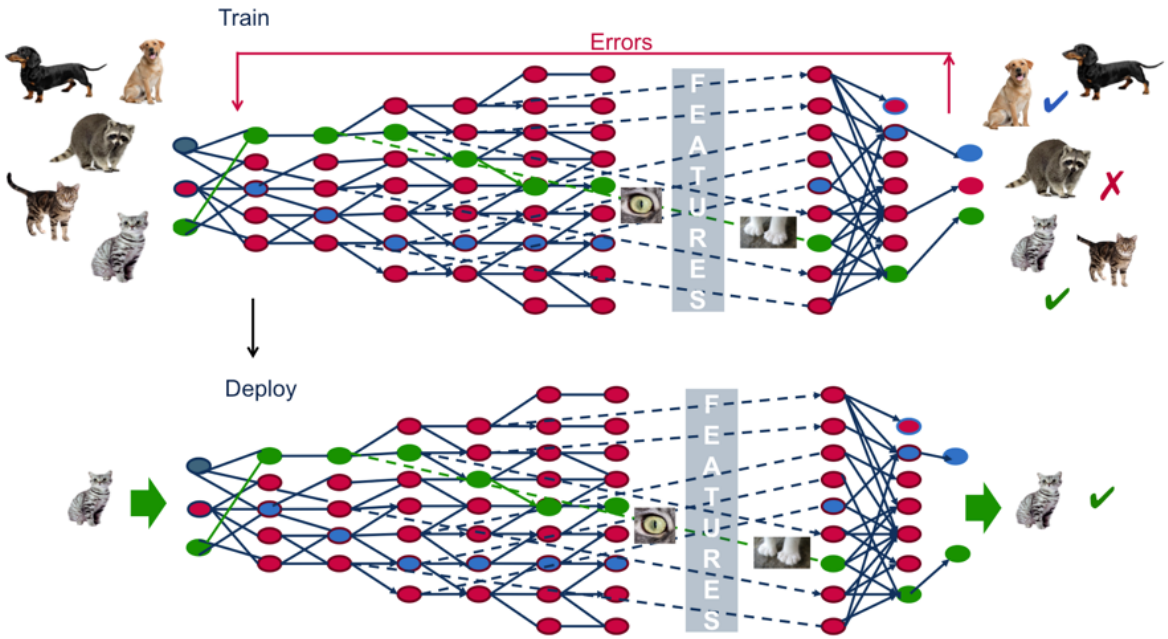
- Z_2-Higgs lattice gauge theory (hadronization, meson excit., topological effects)
- IBM superconducting quantum processor with up to 156 qubits
- Hexagonal quantum chip topology, error mitigation to reduce noise
- Basis Update Galerkin (BUG) novel TNS algorithm for time evolution





- 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

Machine learning, deep learning, robotics: TNS+Quantum $\rightarrow$ quantum inspired AI



- Quantum inspired machine learning!
- Better scaling, lower energy demand, explainable AI
- New sensors: machines begin to interact with our world!

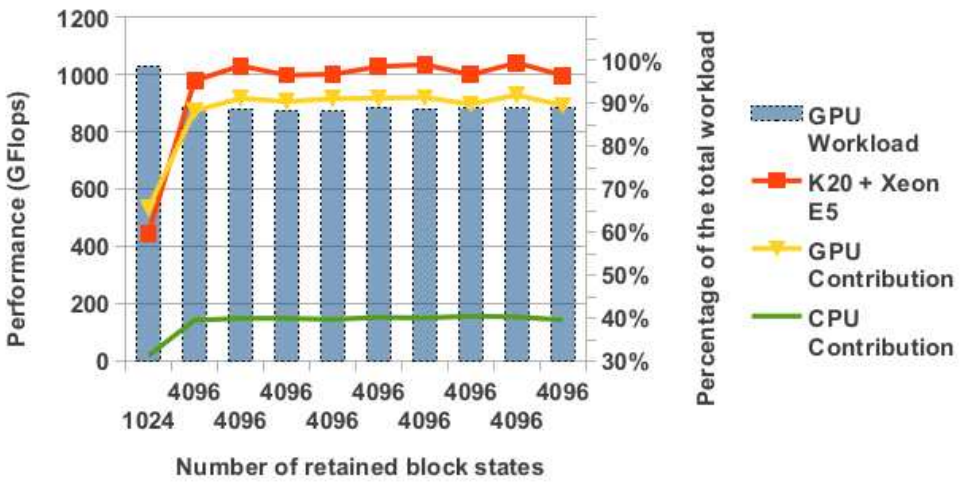
Towards exascale computations on supercomputers

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

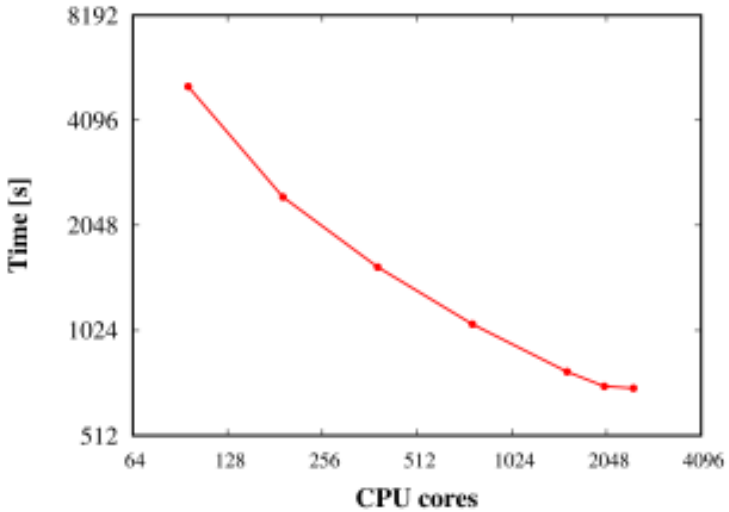
• Dense matrix operations are performed in parallel according to the so-called quantum number decomposed representations (sectors).

• 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)

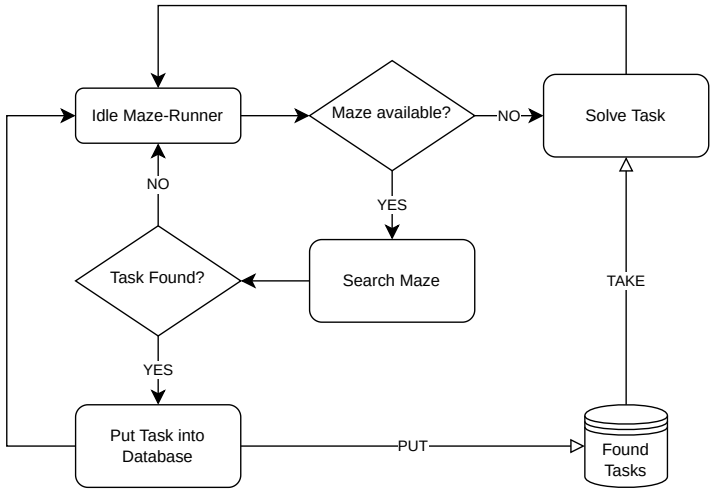


(a) Davidson procedure

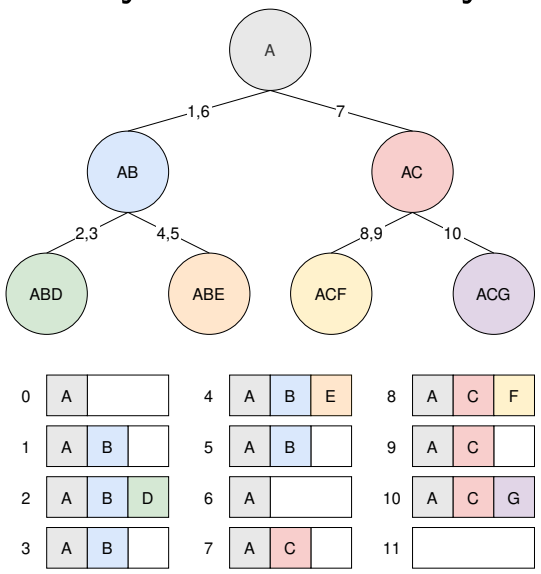
FeMoco cluster
[CAS(113,76)]

Novel algorithmic solutions & parallelization A.Menczer,ÖL(2023)

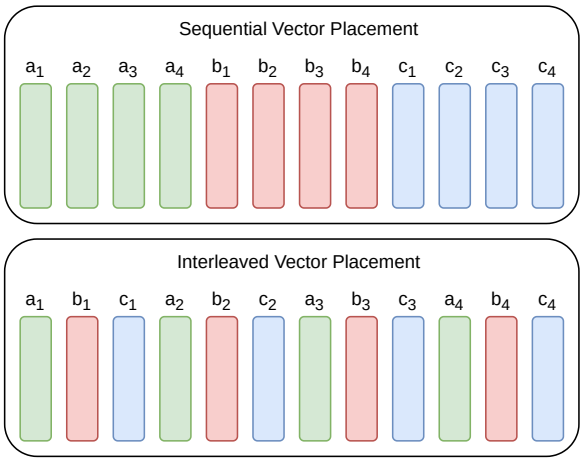
Life Cycle of a Maze-Runner Thread.



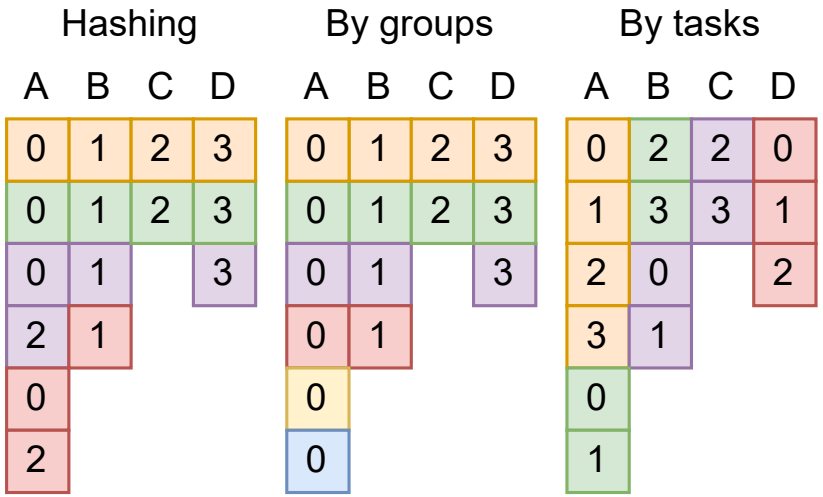
Graph theory based memory management



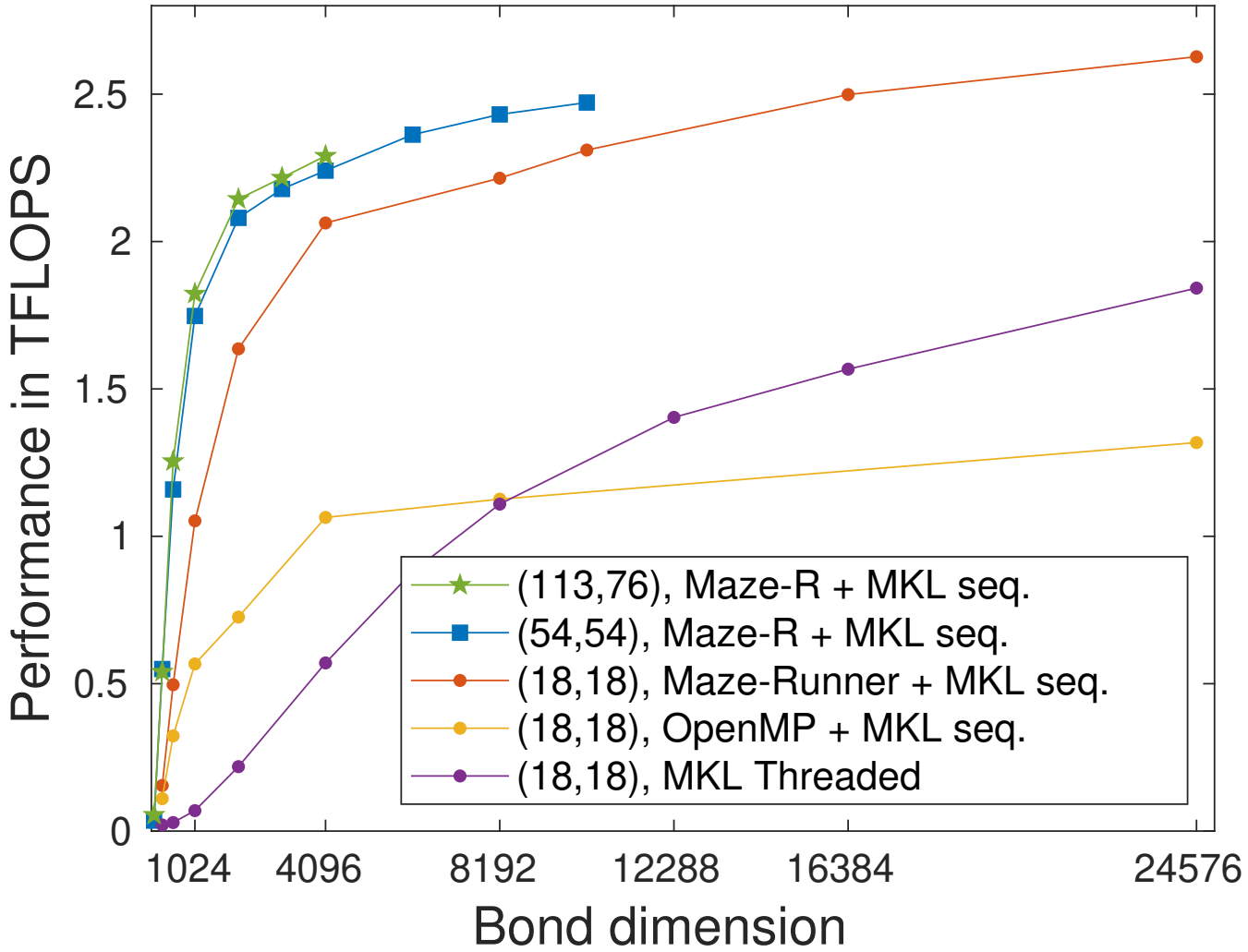
Strided Batched operations via data localization



Execution via hierarchy of tasks

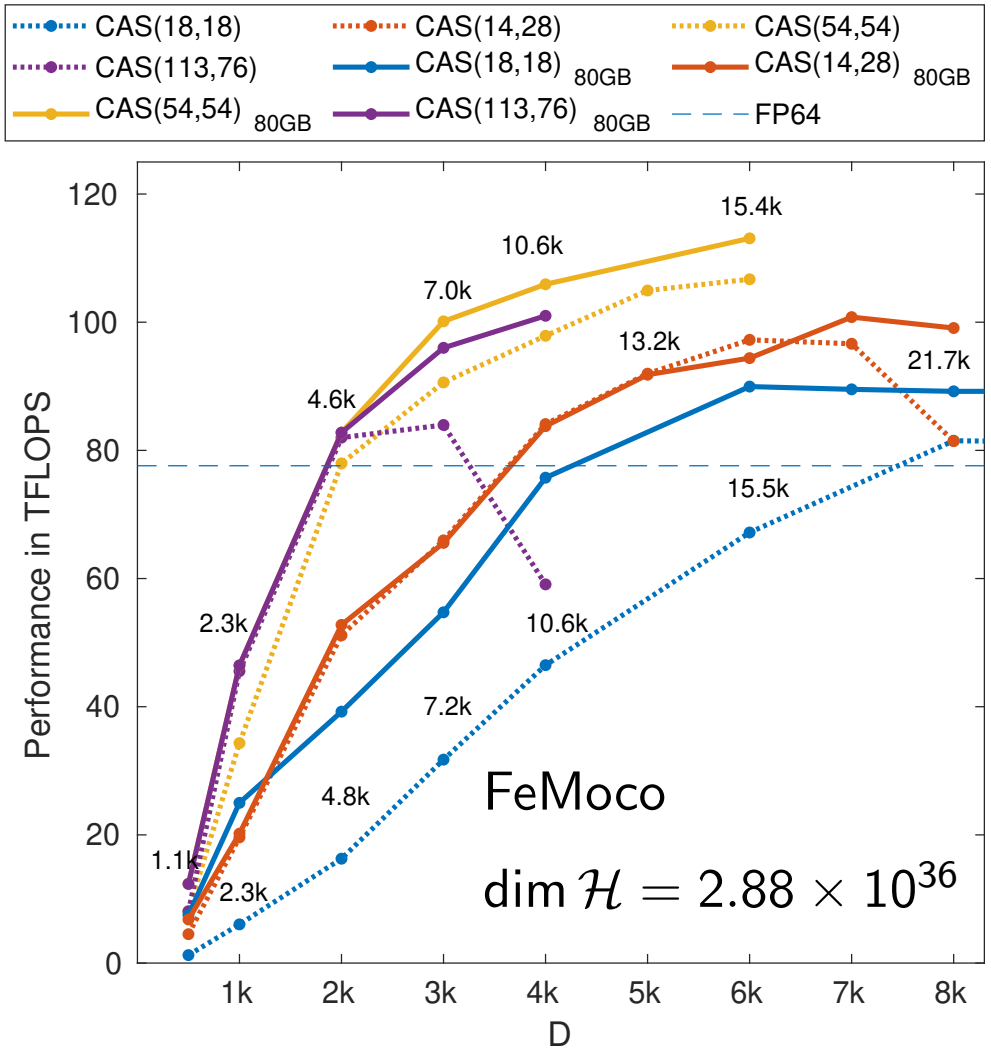


CPU only limit (for CAS(113,76) $\dim \mathcal{H} = 2.88 \times 10^{36}$)



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 performance via AI accelerators. Wall time: $D^3 \to D$



- A factor of 40 speedup compared to a single node with 128 cores
  $\to$ flexible scaling

- 116 TFLOPS $>$ 76 TFLOPS of the FP64 limit of NVIDIA $\to$ utilization of highly specialized tensor core units (TCU)

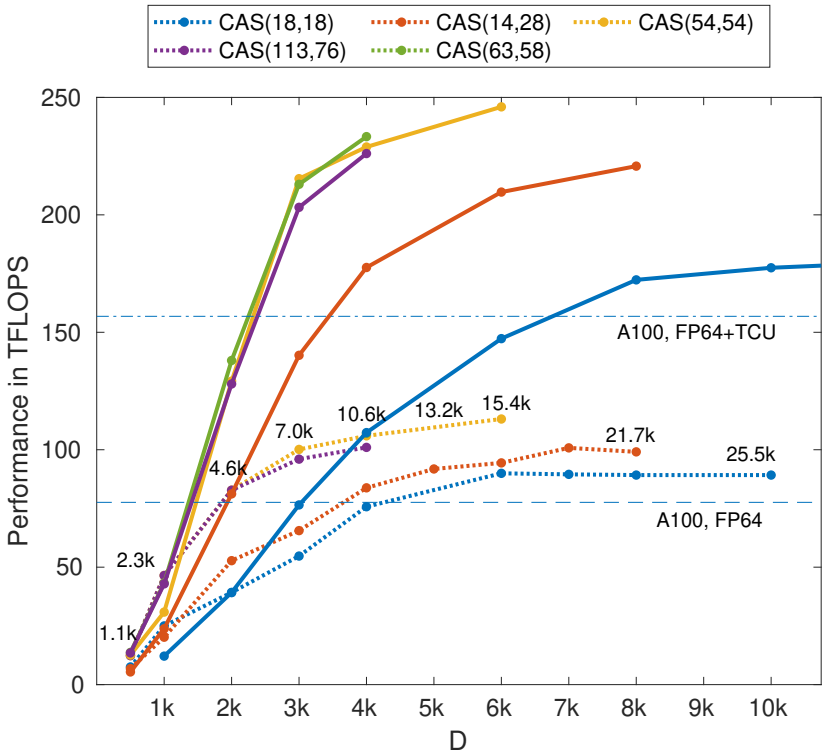
- Power consumption reduced by a factor of 5 to 8
  $\to$ Green DMRG

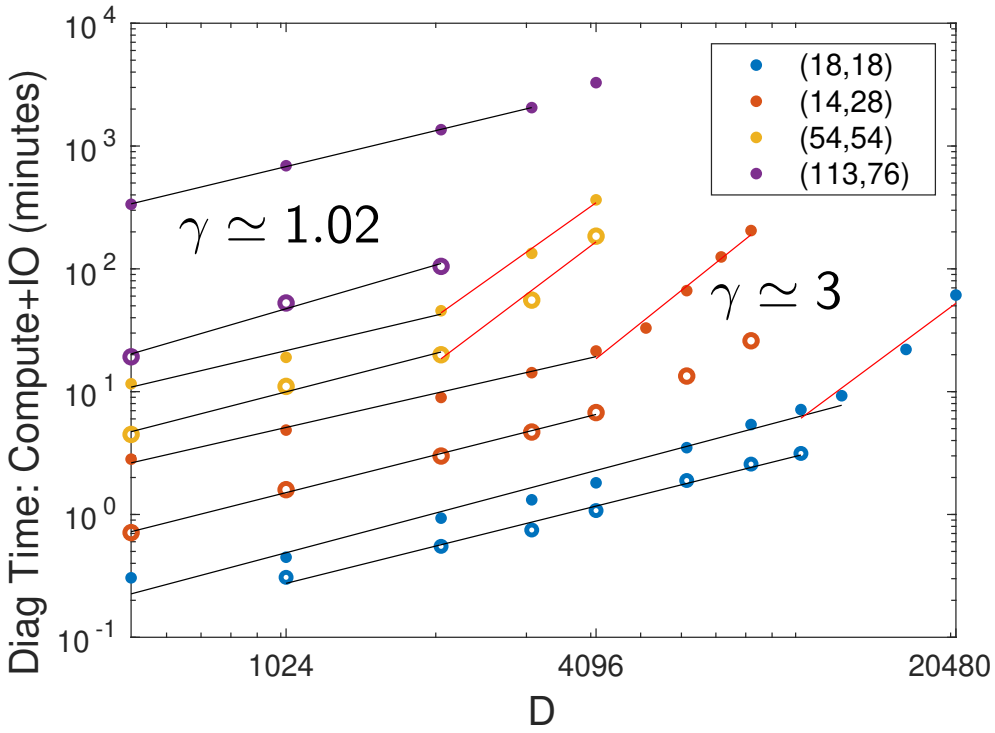
- NVIDIA DGX H100 and Grace Hopper GH200: Testing performance up to $\sim$ 240 TFLOPS in collab with NVIDIA and SandboxAQ

M. van Damme, A. Menczer, M. Ganahl, J. Hammond, Ö.L

- Combination of our MPI and GPU kernels: $\to$ petascale computing.

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

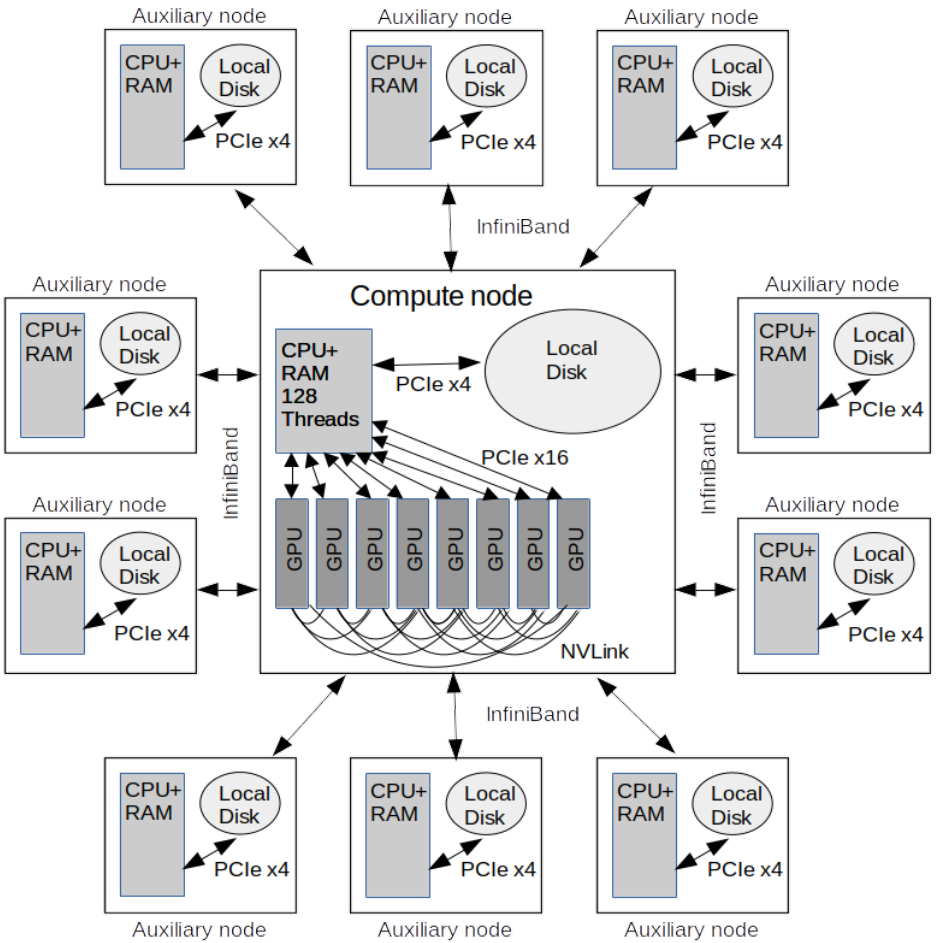
Testing performance up to $\sim$ 250 TFLOPS in collab with NVIDIA and SandboxAQ Menczer,Damme,Rask,Huntington,Hammond,Xantheas,Ganahl,ÖL

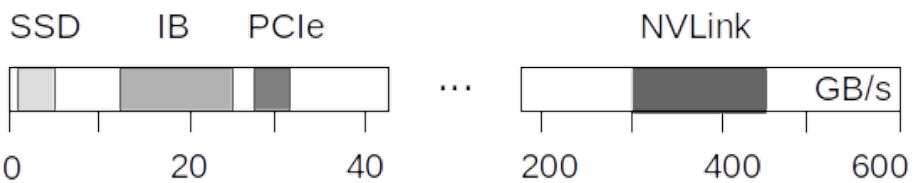
- New model to utilize NVIDIA D2D links. A. Menczer ÖL (unpublished 2023)

- 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 "green" TNS Menczer, ÖL 2024





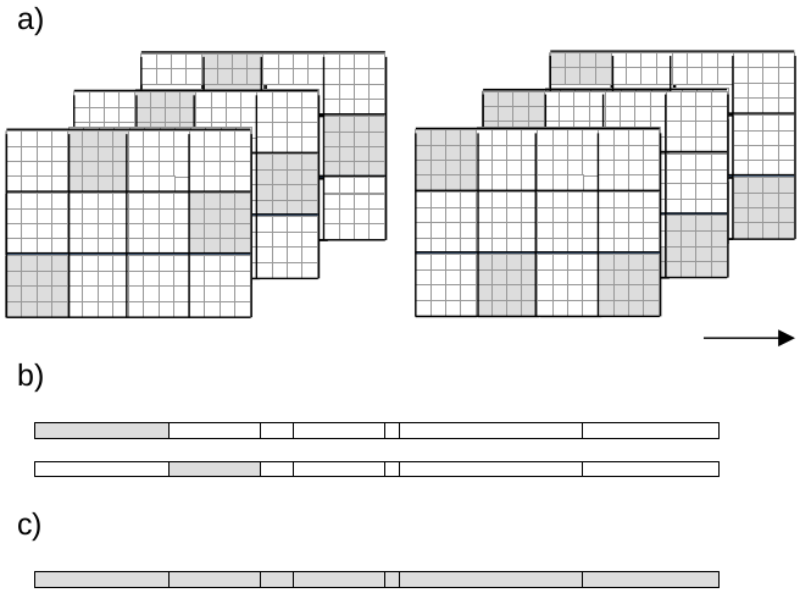
- DGX-H100 costs 100 USD/hour on Google Cloud

- 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

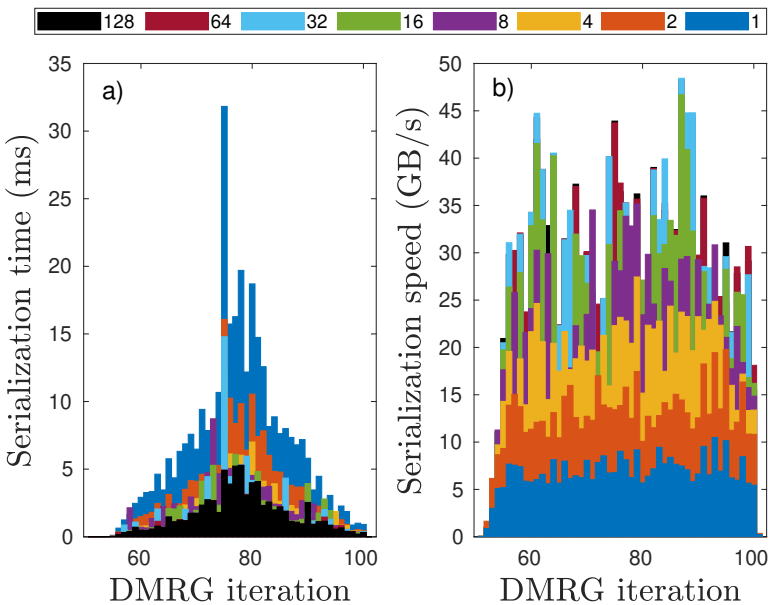
Precontracted network asynchronous MPI-IO Menczer, ÖL 2024

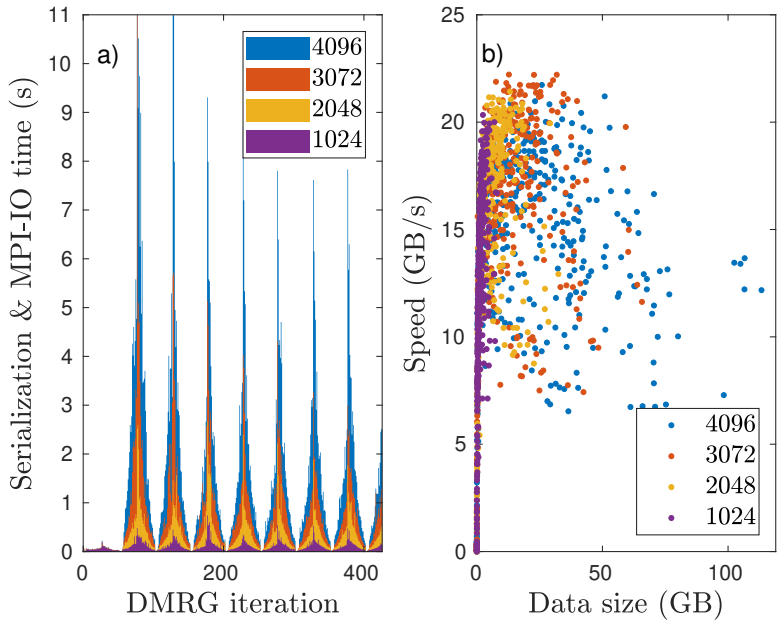
- Replacement of boost library with our in-house build MPI module



a) Quantum number based block sparse matrices and tensors.
b) Skeleton of serialized data segments via disk IO or MPI based communication.
c) Skeleton of serialized data segments filled completely with data when asynchronous save IO.

- We save 80-90 USD/h.





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

Emulating GEMM FP64 arithmetic via Ozaki's scheme

K. Ozaki, Y. Uchino, and T. Imamura

Performance GB200: 90 TFLOPS FP64TC $\leftrightarrow$ 20000 TFLOPS INT8TC

$\mathbb{F}$: a set of binary floating-point, *fl*, numbers (IEEE 754)

Matrices: $A \in \mathbb{F}^{p\times q}$ and $B \in \mathbb{F}^{q\times r}$ and $k \in \mathbb{N}$

Let nonsingular diagonal matrices $D_i \in \mathbb{F}^{p\times p}$ and $E_i \in \mathbb{F}^{r\times r}$ whose diagonal elements are powers of two.

$$A \approx D_1^{-1}D_1A_1 + D_2^{-1}D_2A_2 + \ldots D_k^{-1}D_kA_k$$

$$B \approx B_1E_1E_1^{-1} + B_2E_2E_2^{-1} + \ldots B_kE_kE_k^{-1}$$

where all elements in D_iA_i and B_iE_i for all $1 \leq i \leq k$ can be represented in INT8.

$(D_iA_i)(B_jE_j)$ can be computed without rounding error using INT8TC

Approximation $\hat{C} := fl\left(\sum_{i+j\leq k+1} D_i^{-1}((D_iA_i)(B_jE_j))E_j^{-1}\right)$

Step-1: splitting matrices into k-slices : $\mathcal{O}(pq) + \mathcal{O}(qr)$

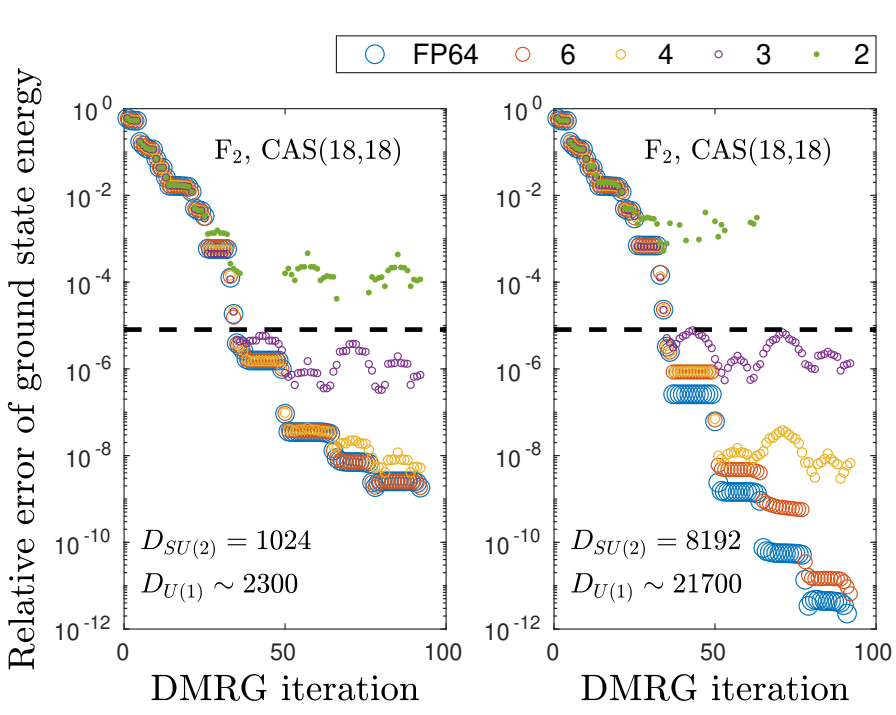
Step-2: $k(k+1)/2$ matrix products (INT8TC) : $k(k+1)pqr + \mathcal{O}(pr)$

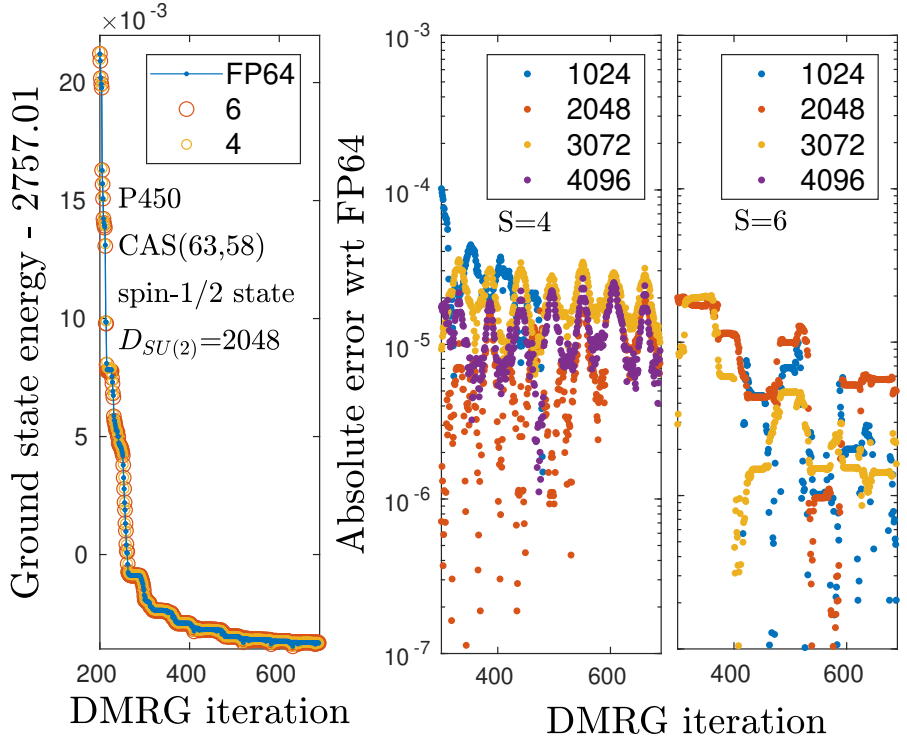
Step-3: reduction of matrix products (FP64) : $\mathcal{O}(pr)$

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

Unexploited power and future perspectives:

- data transfer I/O time between host and device can be eliminated by utilizing high-bandwidth low-latency devices such as NVIDIA Grace superchips or AMD MI300 systems.
- multiNode-multiGPU DMRG, but MPI-based protocols does not show reasonable scaling due to the limited bandwidth of Infiniband (note that double Infiniband has 25GB/s bandwidth).
- NVIDIA GB200 NVL72 connects up to 576 GPUs in a single NVLink domain with over 1 PB/s total bandwidth and 240 TB of fast memory.
- This, together with a revolutionary 1.8 TB/s of bidirectional throughput per GPU has the potential to push the performance of DMRG well into the PetaFLOPS regime.
- Further improvements in mixed-precision arithmetic could also accelerate DMRG calculations even more.
- Recent hardware such as B300 offers even more GPU memory, which is very beneficial for ab initio DMRG dealing with large data sets.

Conclusion and near future on GH200, MI450, GB300 etc

- Tensor network algorithms provide efficient data sparse representation to perform simulations on classical hardware
- Massive Parallelization multiNode-multiGPU → exascale computation
- Mixed precision TNS on specialized new hardware with lower energy consumption
- → Simulation of realistic material properties
- → Estimating quantum circuit depth
- → Quantum inspired AI
- → Hybrid classical quantum algorithms
- → RPI/Empire AI/Argonne: 288/10 000/100 000 GB300 GPUs
- → **TNS + HPC + Blue Ocean Strategy → Quantum CAD**

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)