Simulating quantum systems on high performance computing infrastructures

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

ELTE Bolyai Kollégium

Budapest, 05.08.2025

in collaboration with

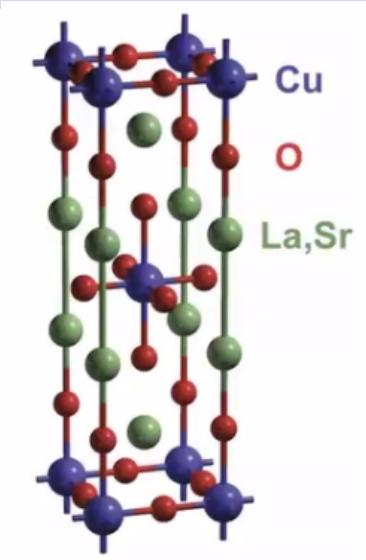
- more than 30 research groups worldwide from condensed matter physics, quantum chemistry, nuclear physics, quantum information theory, applied mathematics and computer science
- High-Performance Computing Center Stuttgart, Germany
- Pacific Northwest National Laboratory (PNNL), USA
- National Energy Research Scientific Computing Center (NERSC), USA

Our computer program package is used by more than 30 research groups worldwide for more than two decades.

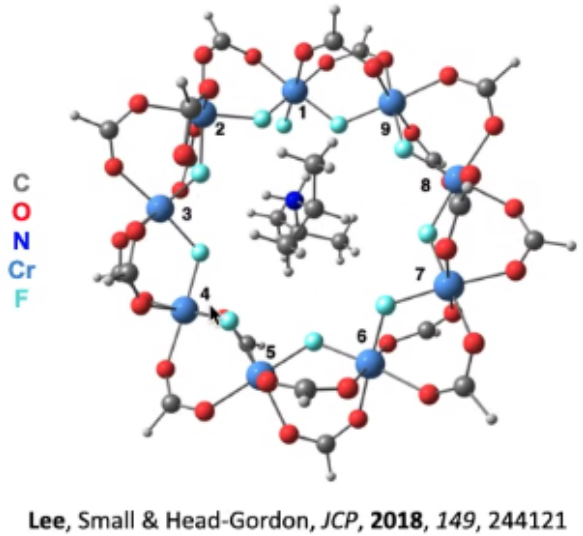
Recently there is also an interest by industrial partners.

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

Strong correlations between electrons $\rightarrow$ exotic materials



High T_c superconductors

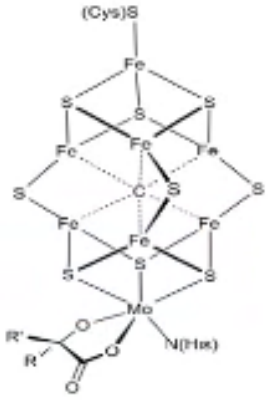


Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

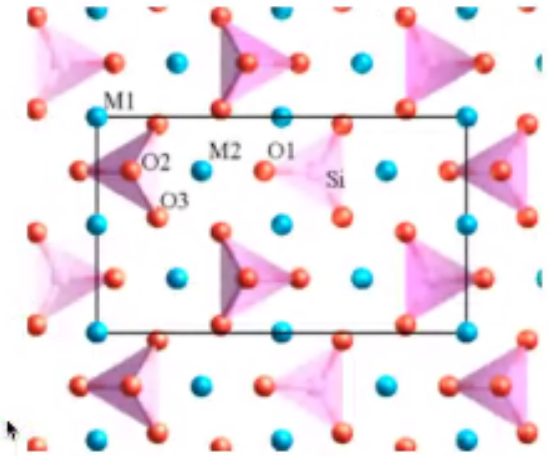
Single molecular magnets (SMM)

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

"FeMoco"



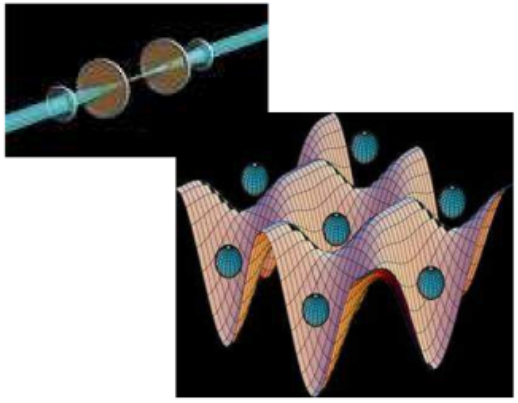
Nitrogen fixation



Battery technology

Experimental realizations: optical lattices

Numerical simulations: model systems



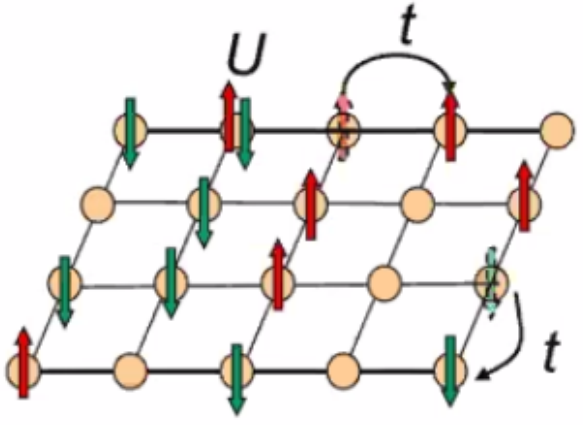
Atoms (represented as blue spheres) pictured in a 2D-optical lattice potential

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

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



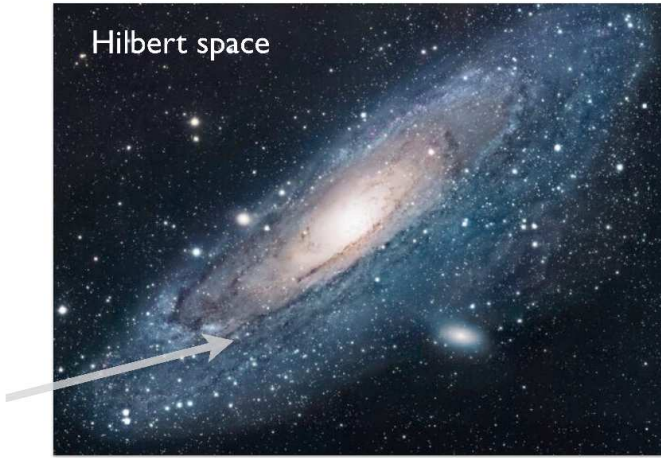
t hopping amplitude
U on-site Coulomb interaction
$\sigma \in \uparrow, \downarrow$ spin index

Classical or quantum computers?

Ultimate need for High Performance Computing

- Simulation of quantum systems on classical computers scales exponentially with system size.

- Solving the problem is like findig a star in the galaxy.



Tensor network states novel algorithms can provide efficient simulations on classical computers, but we need massively parallelized codes on HPCs.

Main research and algorithmic optimization tasks:

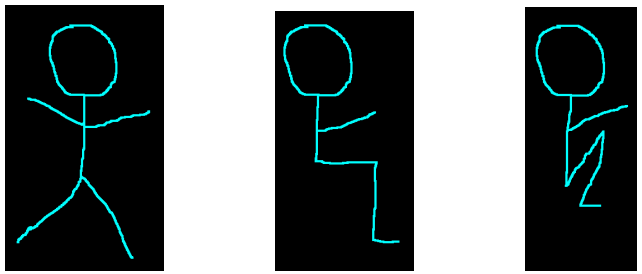
- New mathematical models to reduce computational complexity
- Connection to quantum information theory to reduce computational complexity
- Applications to strongly correlated quantum systems
- New mathematical models for hybrid CPU and GPU parallelization

Motivations, open problems from computational point of view

- Open d and f shell electron systems, transition metals, heavy elements, molecular magnets, extended periodic systems
- Efficient treatment of static (strong) and dynamic (weak) correlations, strongly correlated (multi-reference) systems, reaction paths, transition states, avoided crossings, embedding methods
- Automatic selection of active space
- Choice of basis
- Efficient treatment of relativistic effects
- Efficient treatment of excited states
- Dynamics, time evolution
  coupling to environment, dissipative systems, entanglement barrier
- Finite temperature
- Beyond Born-Oppenheimer approximation
- Correlation structures, clusterization, chemical bonding, multipartite entanglement

Discrete basis, configuration space, superposition

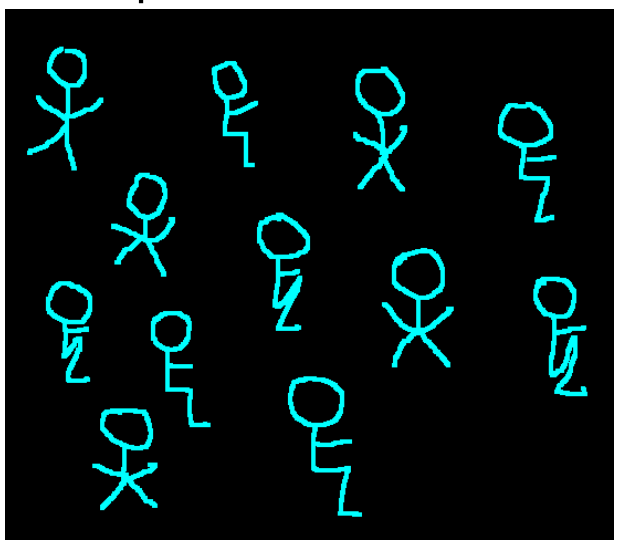
Possible states of a person



stands sits squats

Dimension of the local space $d = 3$

N persons in a room



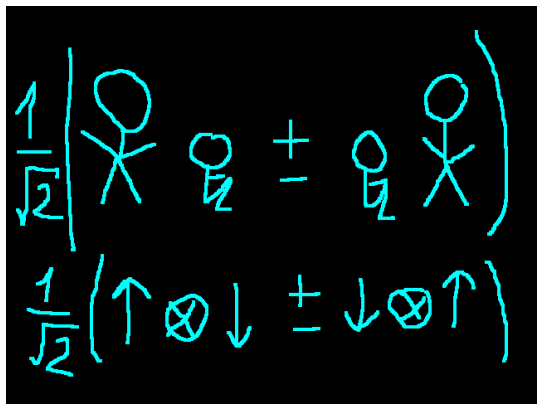
Dimension of the configuration space: 3^N, i.e., it scales exponentially

In quantum physics superposition is possible:

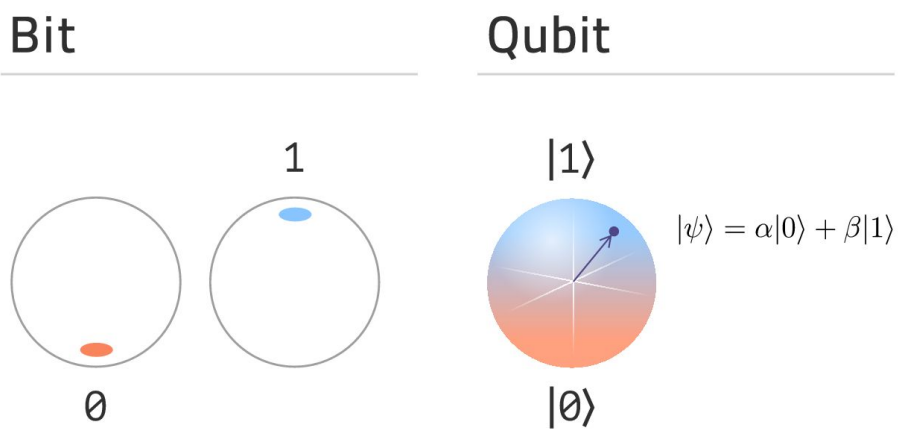
Ex. d=2 (two states allowed)

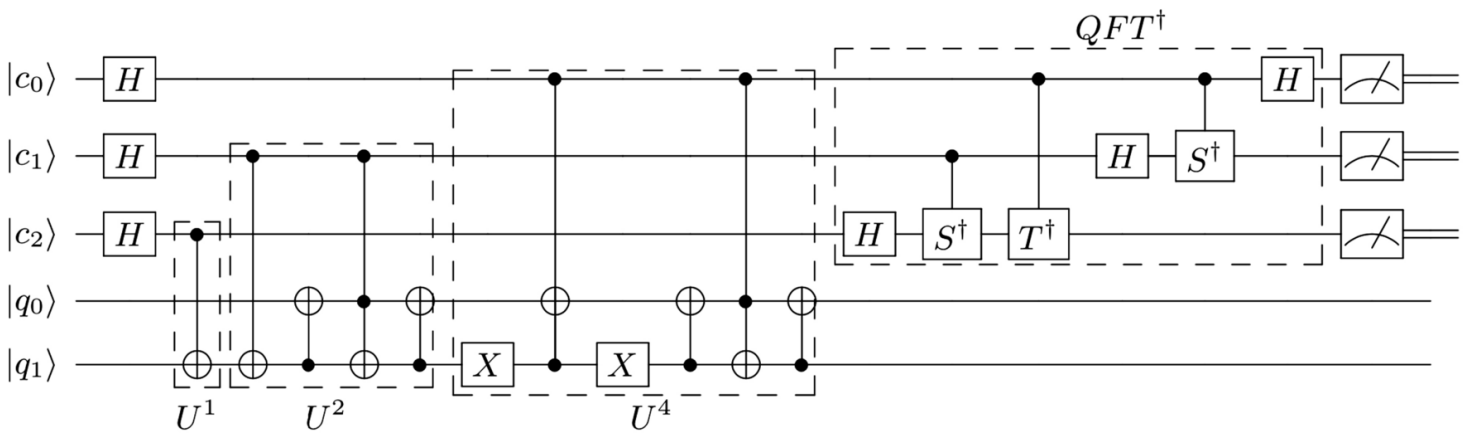
- Two persons (at position A and B).
- Four possible configurations.
- At position "A" person stands or squats with 50% probability.
- At position "B" person stands or squats with 50% probability.

Entangled state $\rightarrow$ quantum information (q-dits)



Entanglement: quantum data processing





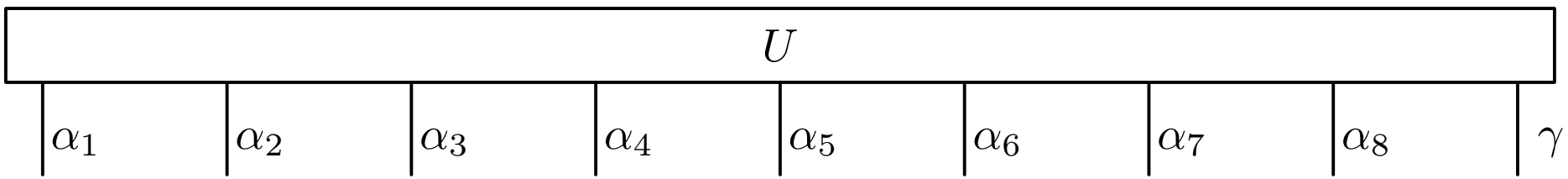
- Quantum computing: quantum supremacy or quantum advantage
- Quantum cryptography: secure communication
- Experimental realizations: quantum sensors (biomedical applications), unprecedented spatial resolution and sensitivity on atomic length scale

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 $\alpha_i \in \{0, \downarrow, \uparrow, \uparrow\downarrow\}$.

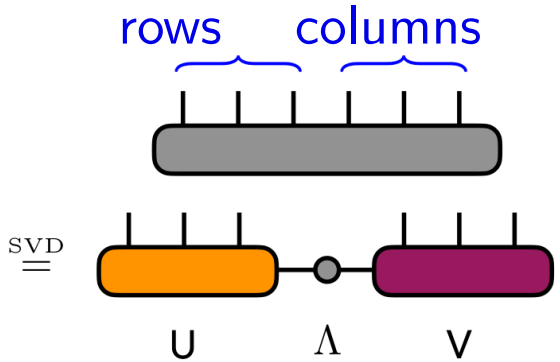
$\dim \mathcal{H}_d = \mathcal{O}(n^d)$ Curse of dimensionality!

$\rightarrow$ need efficient data-sparse representation

Matrix representation of the wave function

• We separate indices α_i into two groups (bipartite representation), so U can be written in matrix form

$$[U]^{\alpha_{i+1},\ldots,\alpha_d}_{\alpha_1,\ldots,\alpha_i} \in \mathbb{C}^{n_1\cdots n_i} \otimes \mathbb{C}^{n_{i+1}\cdots n_d}$$



Schmidt-decomposition or
singular value decomposition (SVD)

$$|\Psi_\gamma\rangle = \sum_a |\Psi\rangle = \sum_{a=1}^{m} \lambda_a |u^a_{\alpha_1,\alpha_2,\ldots,\alpha_i}\rangle |v^a_{\alpha_1,\alpha_2,\ldots,\alpha_i}\rangle,$$

where Λ is a diagonal matrix (λ_a) and

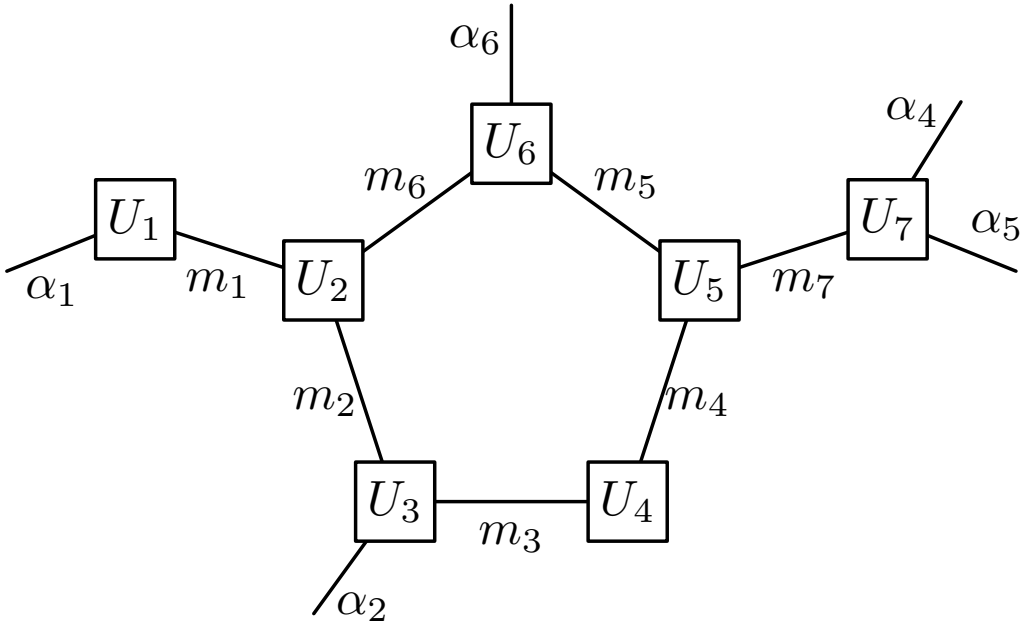
$$m = \min(n_1 \cdots n_i, n_{i+1} \cdots n_d)$$

If $m = 1$ then $|\Psi\rangle$ is a product state, otherwise it is entangled.

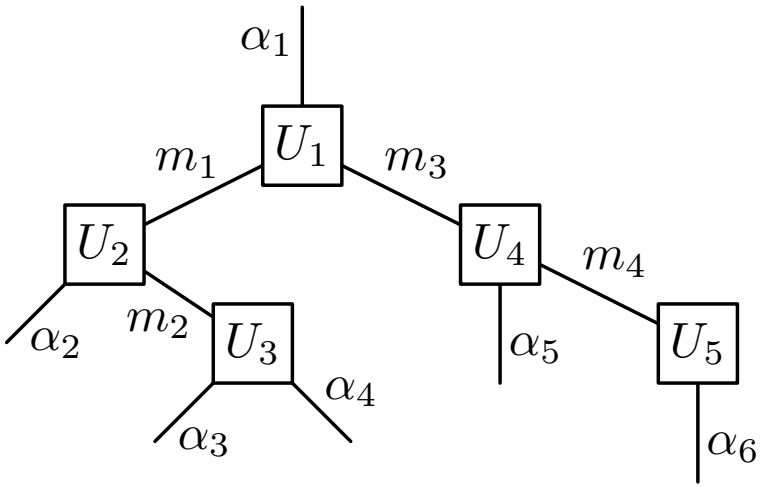
In practice we keep $m < \min(n_1 \cdots n_i, n_{i+1} \cdots n_d)$ based on a priory set error margin, thus truncation leads to an approximate solution.

It is a major task to determine the scaling of m (for example area law).

Tensor product representation



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



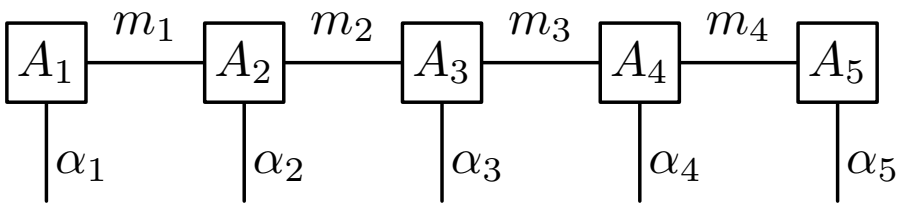
An arbitrary example of a tensor tree (loop free).

Matrix product state (MPS) representation of DMRG

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:

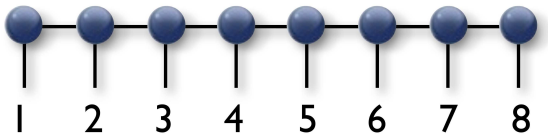
$$U(\alpha_1, \ldots, \alpha_d) = \mathbf{A}_1(\alpha_1)\mathbf{G}\mathbf{G}^{-1}\mathbf{A}_2(\alpha_2) \cdots \mathbf{A}_{d-1}(\alpha_{d-1})\mathbf{A}_d(\alpha_d)$$

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

Novel TNS algorithms for simulations on HPC architectures

1D MPS

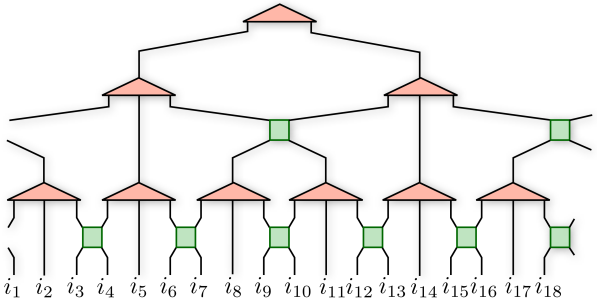
Matrix-product state



White, Östlund, Rommer

1D MERA

Multi-scale entanglement renormalization ansatz

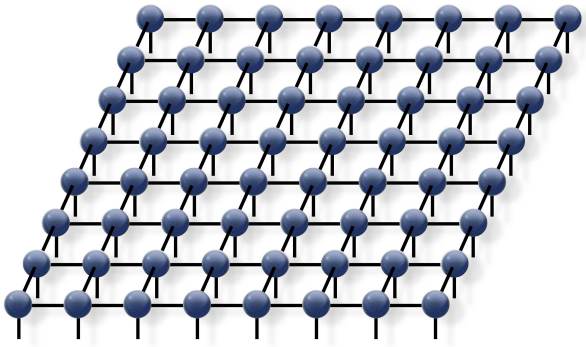


1D TTNS

Tree tensor network state

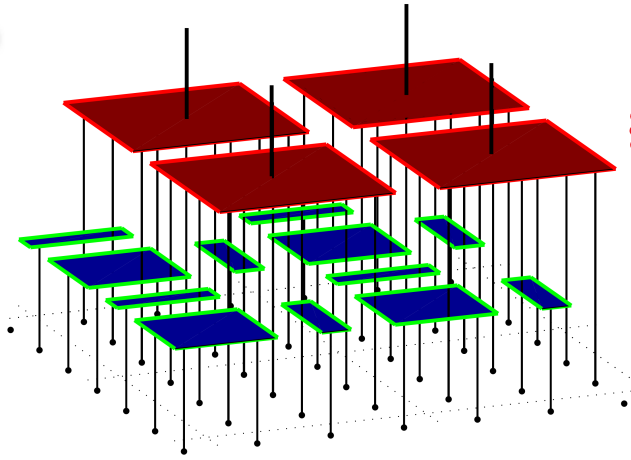
Vidal, Corboz

2D PEPS



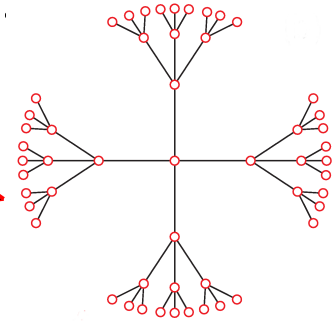
Verstraete, Cirac Jordan, Orus, Vidal

2D Mera



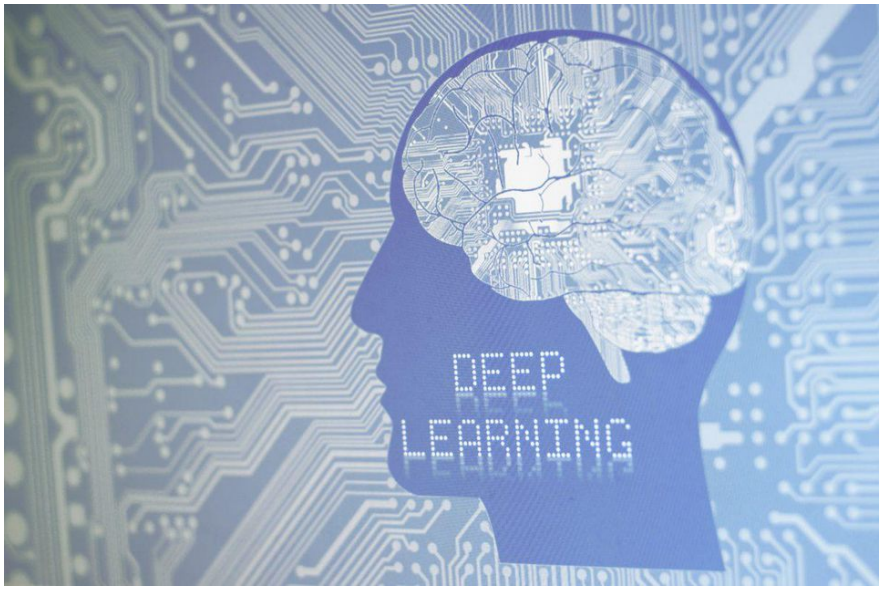
Vidal, Evenbly

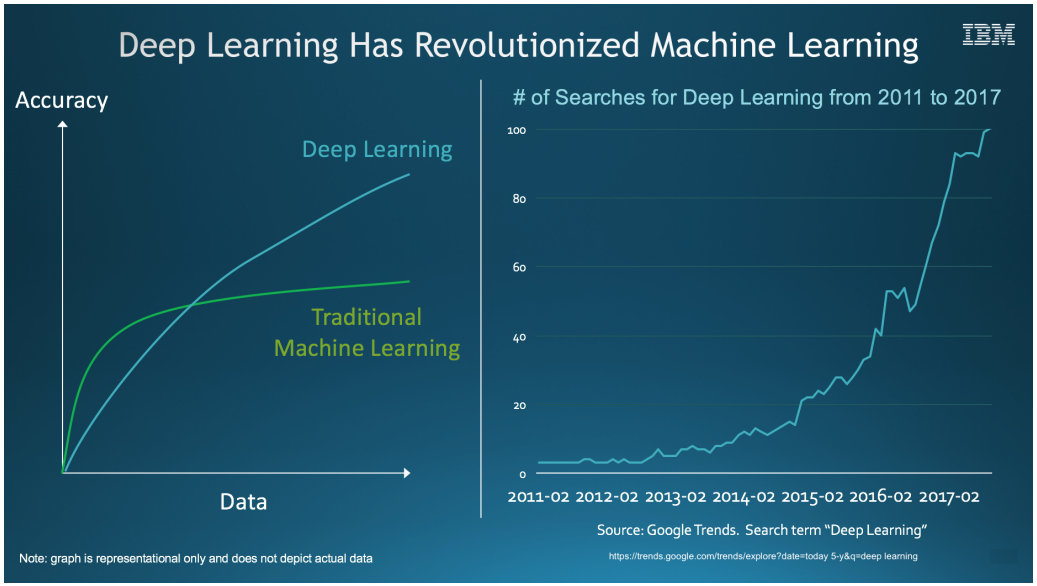
2D Tree-

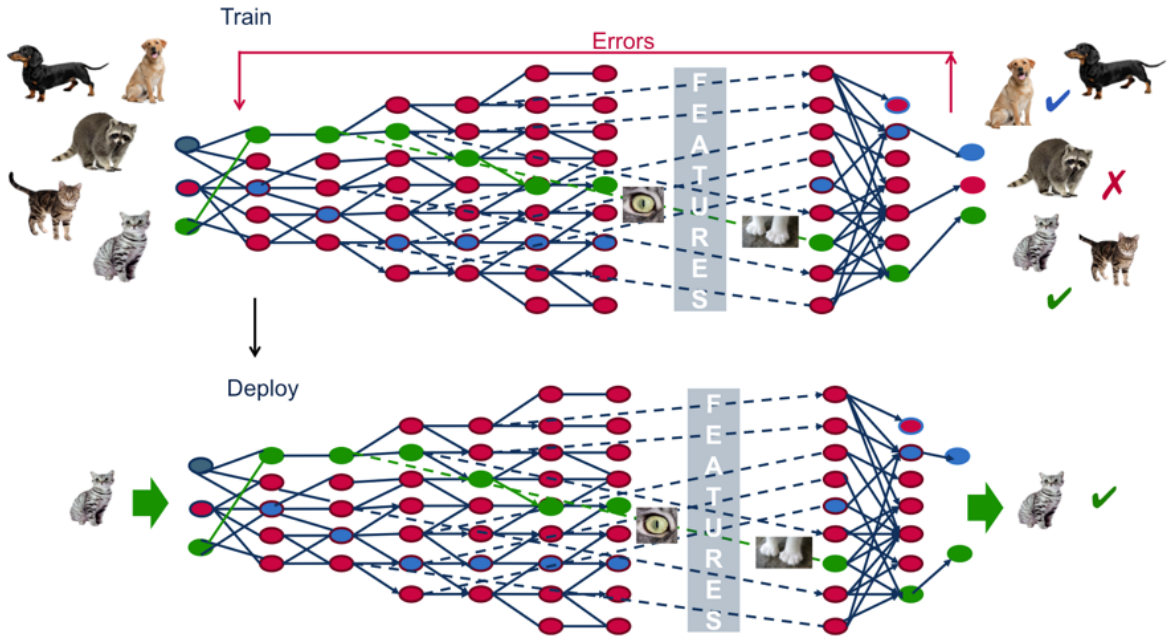


Vidal, Corboz, Verstraete, Murg, Legeza, Noack

Deep learning, AI, ML, robotic







New sensors: machines begin to interact with our world!

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: 100 electrons on 100 orbitals.
- 1-BRDM and 2-BRDM, finite temperature, dynamics
- Massively parallel implementations CPU/GPU$\rightarrow$ exascale on HPC

Main properties and questions

- Polynomial scaling (DMRG): $\mathcal{O}(M^3 d^3) + \mathcal{O}(M^2 d^4)$
- CAS-like multireference method suitable for strongly correlated els.
- Variational
- Size-consistent (with appropriate choice of active space)
- DMRG can be combined with standard methods
  $\rightarrow$ DMRG-TCCSD, DMRG-NEVPT2
  to optimize orbitals $\rightarrow$ DMRG-SCF

Three main questions

- How to choose optimal rank?
  $\rightarrow$ Dynamic Block State Selection (DBSS) (2003)
- How to choose active space and optimal network structure?
  $\rightarrow$ Based on the entanglement and correlation patterns (2003,2015)
- How to choose optimal basis?
  $\rightarrow$ Fermionic mode transformation (2016)

$\rightarrow$ concepts of quantum information theory

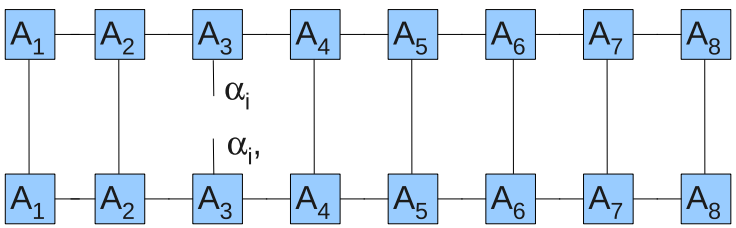
One- (ρ_i) and two-mode ($\rho_{i,j}$) reduced density matrix

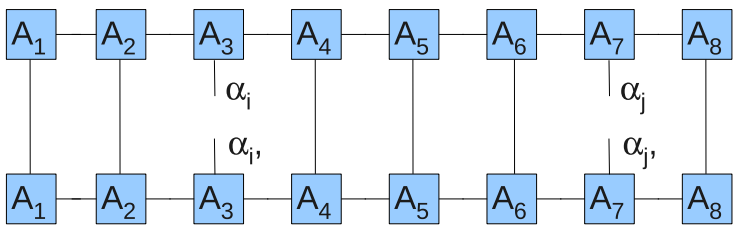
$$|\psi\rangle = \sum_{\alpha_1,\ldots,\alpha_N} C_{\alpha_1,\ldots,\alpha_N} |\alpha_1 \ldots \alpha_N\rangle \ ,$$

- $\rho_{i,j}$ is calculated by taking the trace of $|\Psi\rangle\langle\Psi|$ over all local bases except for α_i and α_j, the bases of modes i and j, i.e.,

$$\rho_{i,j}([\alpha_i, \alpha_j], [\alpha_i', \alpha_j']) = \sum_{\alpha_1,\ldots,\cancel{\alpha_i},\ldots,\cancel{\alpha_j},\ldots,\alpha_N} C_{\alpha_1,\ldots,\alpha_i,\ldots,\alpha_j,\ldots,\alpha_N} \ C^*_{\alpha_1,\ldots,\alpha_i',\ldots,\alpha_j',\ldots,,\alpha_N} \ .$$

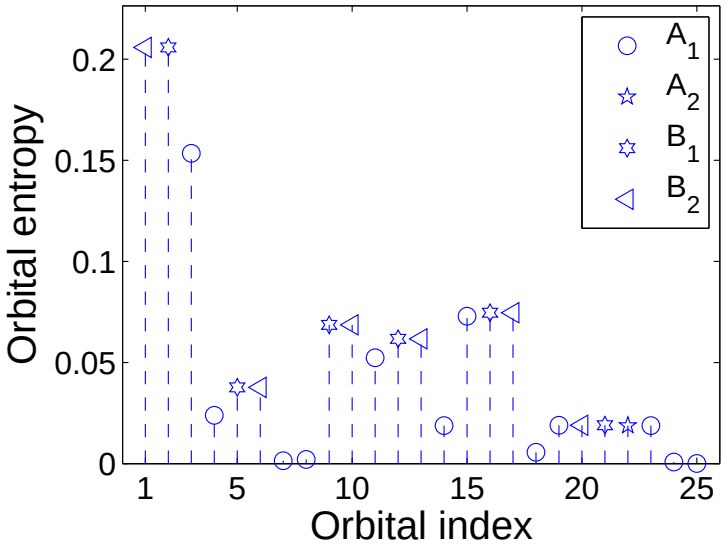
- In the MPS representation, calculation of ρ_{ij} corresponds to the contraction of the network except at modes i and j.

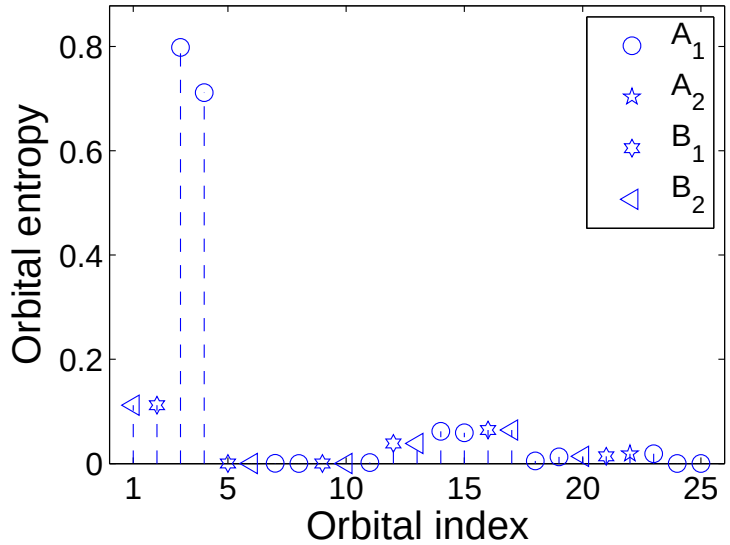


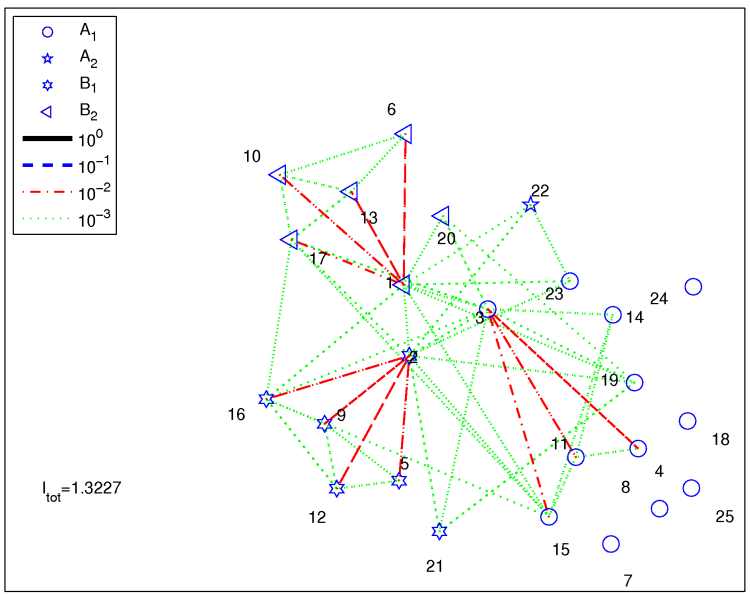


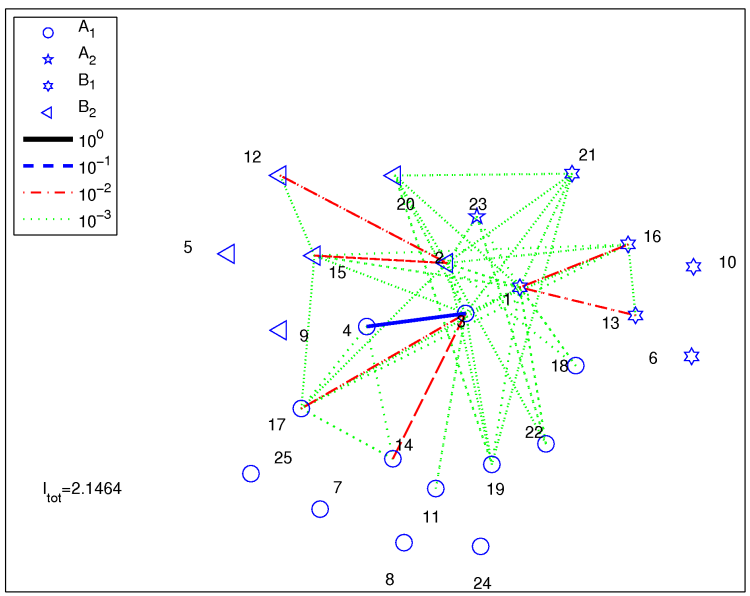
- von Neumann quantum information entropy, $s = -\sum_\alpha \lambda_\alpha^2 \ln \lambda_\alpha^2$.
- Mutual information, $I = s_i + s_j - s_{ij}$.

Selection of active space, multiply connected networks









LiF at r=3.05 LiF at r=13.7

Ö.L., Sólyom (2003), Rissler, White, Noack, ECP (2006), Murg, Verstraete, Schneider, Nagy, Ö.L. (2013)
Stein, Reiher (2016), Golub, Antalik, Veis, Brabec (Machine Learning, 2020)

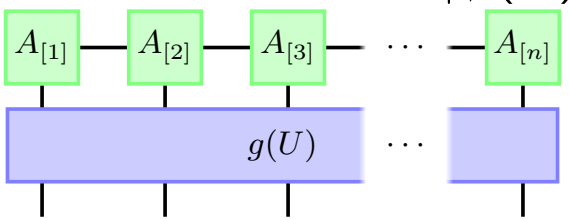
Redefinition of the fermionic modes by a linear transformation

Krumnow, Veis, Ö.L., Eisert, 2014-2016

- Linear transformations of a set of fermionic annihilation operators $\{c_i\}$ to a new set $\{d_i\}$ satisfying the canonical anti-commutation relations:

$$c_i = \sum_{j=1}^{Np} U_{i,j} d_j, \quad p \text{ denotes the number of different fermion species}$$

- Under this change of basis a state vector $|\psi(U)\rangle = G(U)|\psi(\mathbb{1})\rangle$



- Denoting the Hamiltonian written in terms of the transformed modes by $H(U) = G(U)^\dagger H G(U)$, we are interested in the solutions of

$$(U_{\text{opt}}, |\psi_{\text{opt}}\rangle) = \operatorname{argmin}_{\substack{U \in U(Np), \\ |\psi\rangle \in \mathcal{M}_{D_{\max}}}} \langle\psi|H(U)|\psi\rangle.$$

- The global basis change is composed of local unitaries solutions of

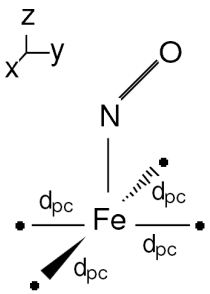
$$U_{\text{opt}}^{\text{loc}} = \operatorname{argmin}_{U \in V} f_j\big(|\psi(\mathbb{1}_j \oplus U \oplus \mathbb{1}_{N-j-2})\rangle\big),$$

cost function $f_j^{(1)}(|\psi\rangle) = ||\Sigma_\psi^j||_1$ where Σ_ψ^j denotes the Schmidt spectrum of $|\psi\rangle$ for a bipartiting cut between sites j and $j+1$.

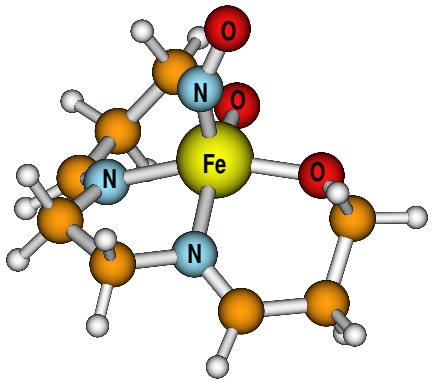
Interactions, entanglement and correlations

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

Applications in condensed matter physics, quantum chemistry, nuclear physics, relativistic effects, etc

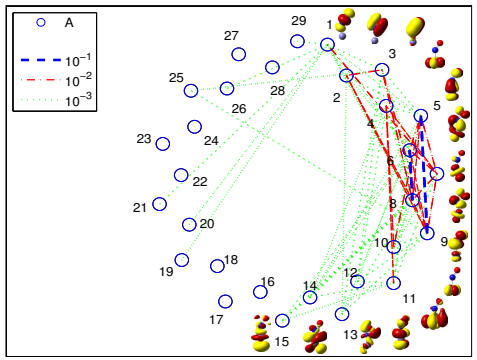


(a) $[Fe(NO)^{2+}]$

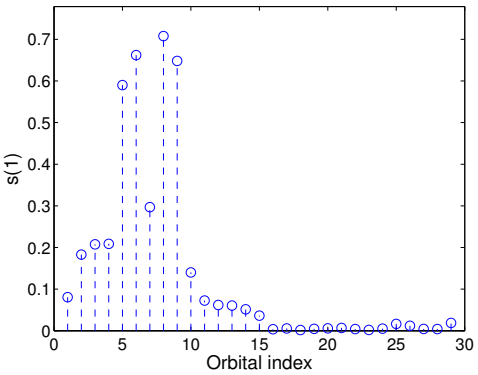


(b) FeL(NO)

$[Fe(NO)^{2+}]$

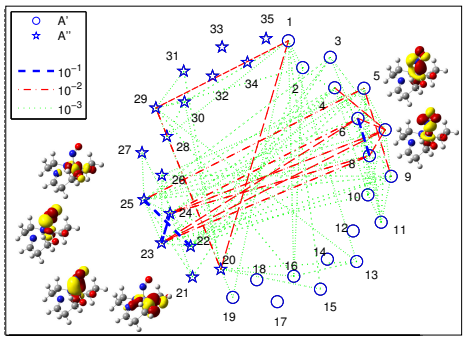


(a) Mutual information

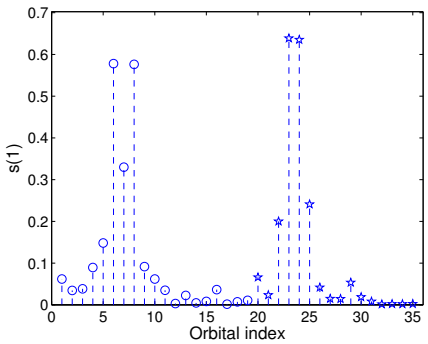


(b) Single orbital entropy

FeLNO



(a) Mutual information



(b) Single orbital entropy

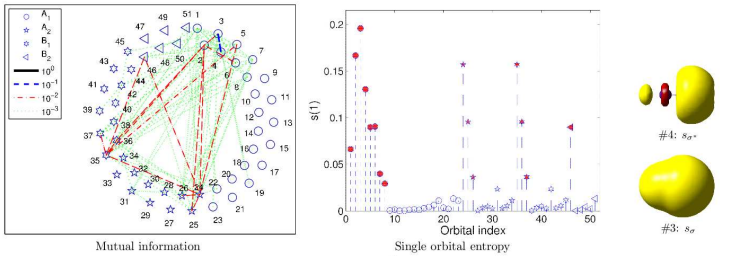
Strongly correlated system

Effect of environment

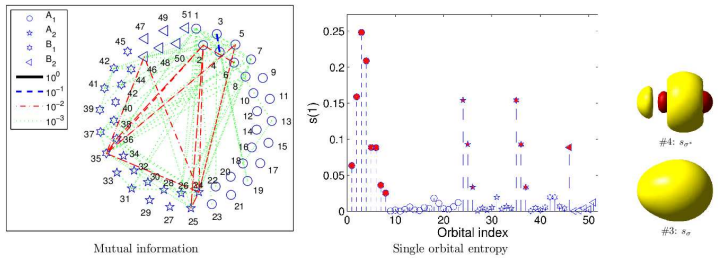
Boguslawski, Tecmer, Ö.L., Reiher (2012)

Chemical bond forming and breaking vs Entanglement

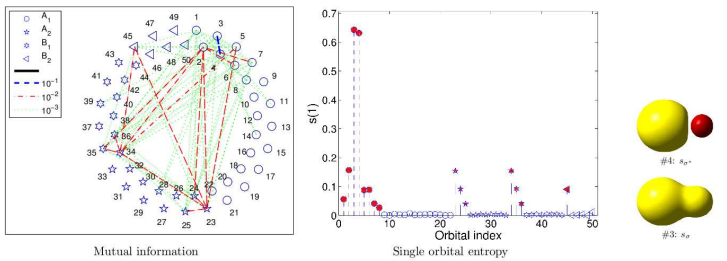
Boguslawski, Tecmer, Barcza, Ö.L., Reiher, 2013



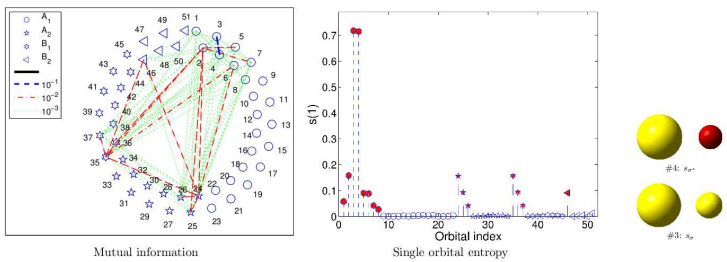
(a) $d_{CsH} = 2.5$Å



(b) $d_{CsH} = 3.5$Å

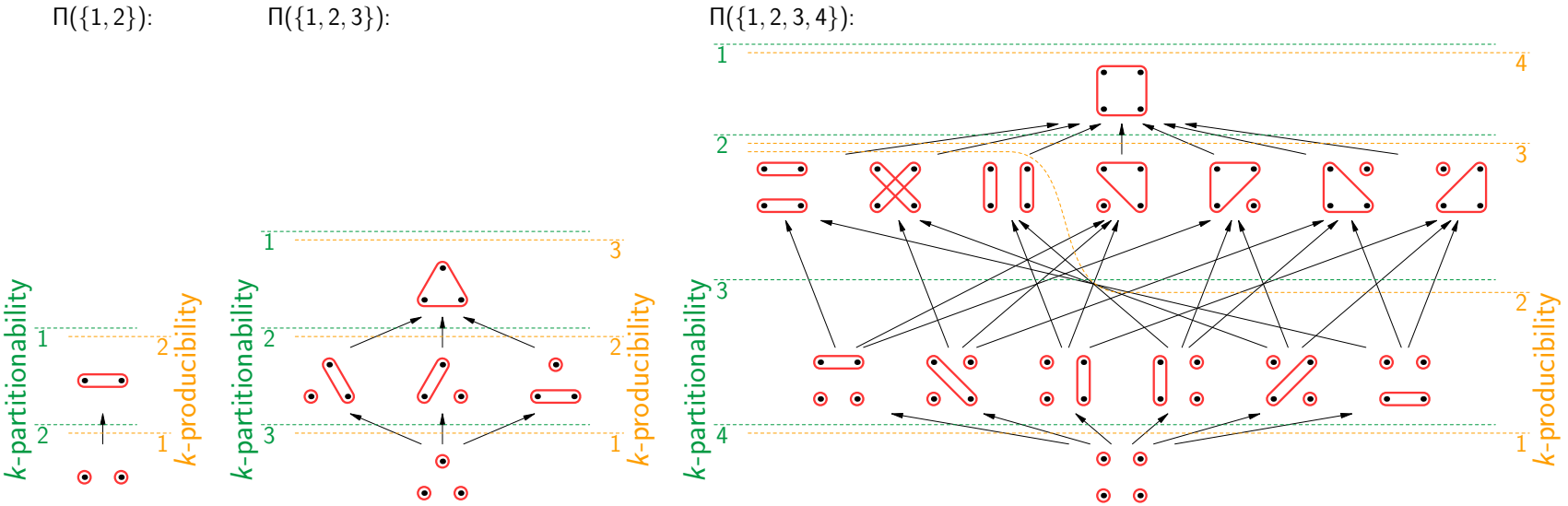


(c) $d_{CsH} = 5.5$Å



(d) $d_{CsH} = 7.0$Å

Multiorbital correlations Sz. Szalay (2015)



- partitions of the system:
  $\xi = \{X_1, X_2, \ldots, X_{|\xi|}\} \equiv X_1|X_2|\ldots|X_{|\xi|} \in \Pi(L)$
- refinement: $\upsilon \preceq \xi$ def.: $\forall Y \in \upsilon, \exists X \in \xi : Y \subseteq X$
- ξ-**correlation** (ξ-mutual information):

$$C_\xi(\varrho) = \min_{\sigma \in \mathcal{D}_{\xi-\text{uncorr}}} D(\varrho||\sigma) = \sum_{X \in \xi} S(\varrho_X) - S(\varrho)$$

- multipartite monotonicity: $\upsilon \preceq \xi \Leftrightarrow C_\upsilon \geq C_\xi$

k-partitionability-correlation and k'-producibility correlation:

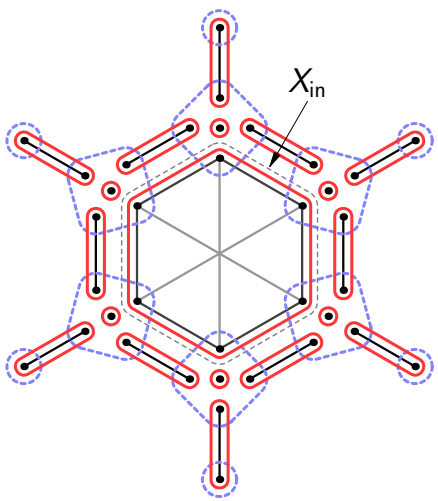
$$C_{k\text{-part}}(\varrho_L) = C_{\boldsymbol{\mu}_k}(\varrho_L) = \min_{|\mu|>k} C_\mu(\varrho_L), \quad C_{k'\text{-prod}}(\varrho_L) = C_{\boldsymbol{\nu}_{k'}}(\varrho_L) = \min_{\forall N \in \nu : |N| < k'} C_\nu($$

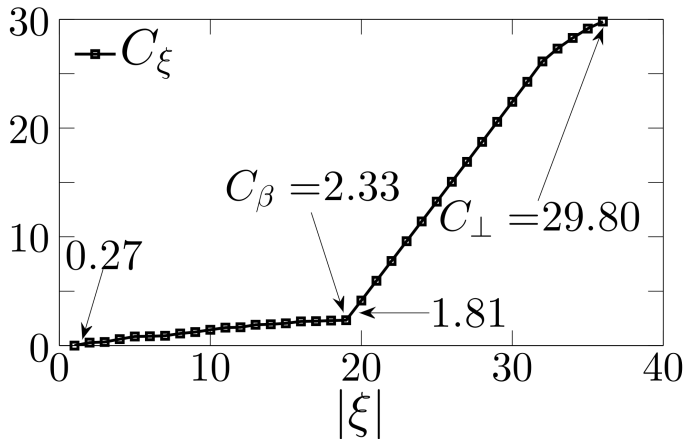
Example (aromatic system): C_6H_6 (benzene)

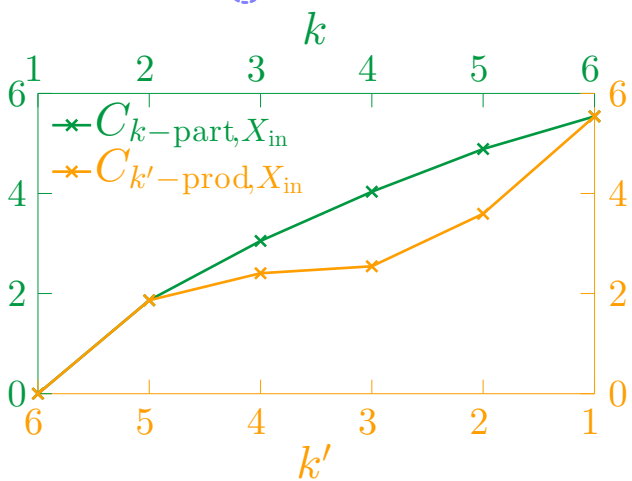
Szalay, Barcza, Szilvási, Veis, Ö.L (2017)

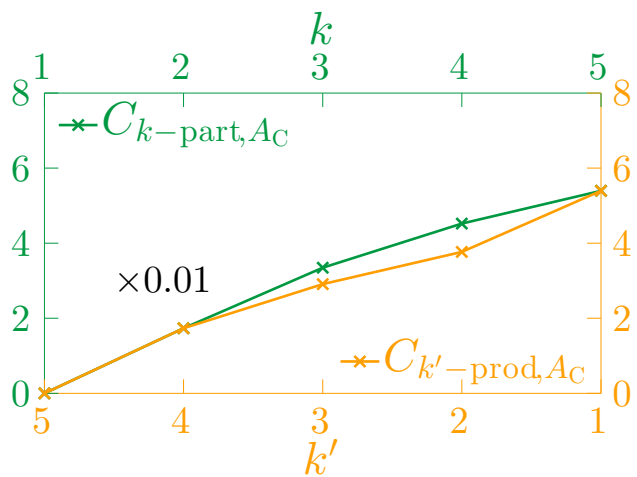
"atoms": $\alpha = A_1|A_2|\ldots|A_{|\alpha|}$, "bonds": $\beta = B_1|B_2|\ldots|B_{|\beta|}$

$$\sum_{A\in\alpha} C_{\perp,A}(\varrho_A) + C_\alpha(\varrho_M) = \sum_{B\in\beta} C_{\perp,B}(\varrho_B) + C_\beta(\varrho_M) = C_\perp(\varrho_M)$$

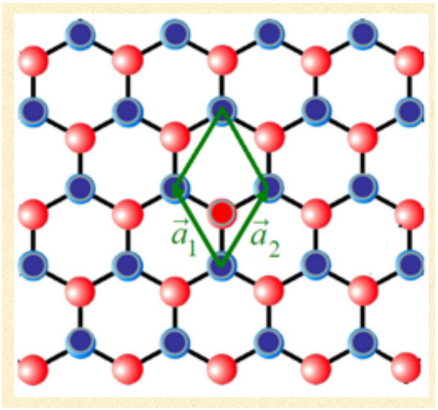


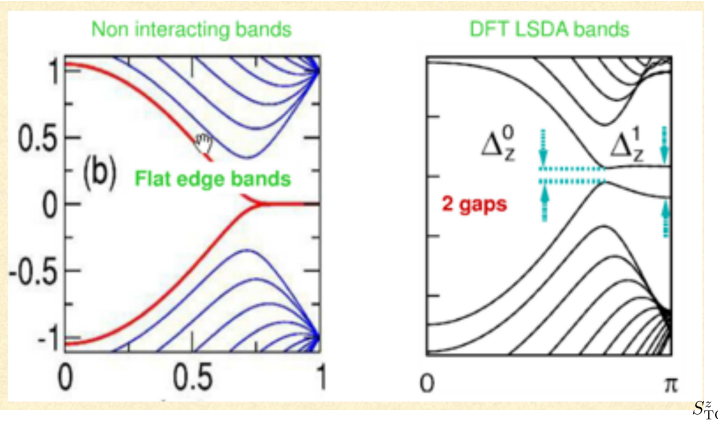




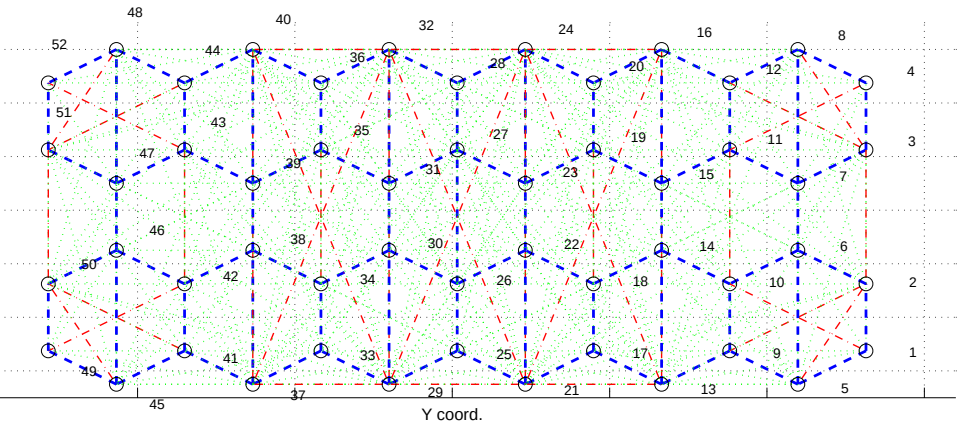


Example on graphene nanoribbons I. Hagymási, Ö.L (2016)





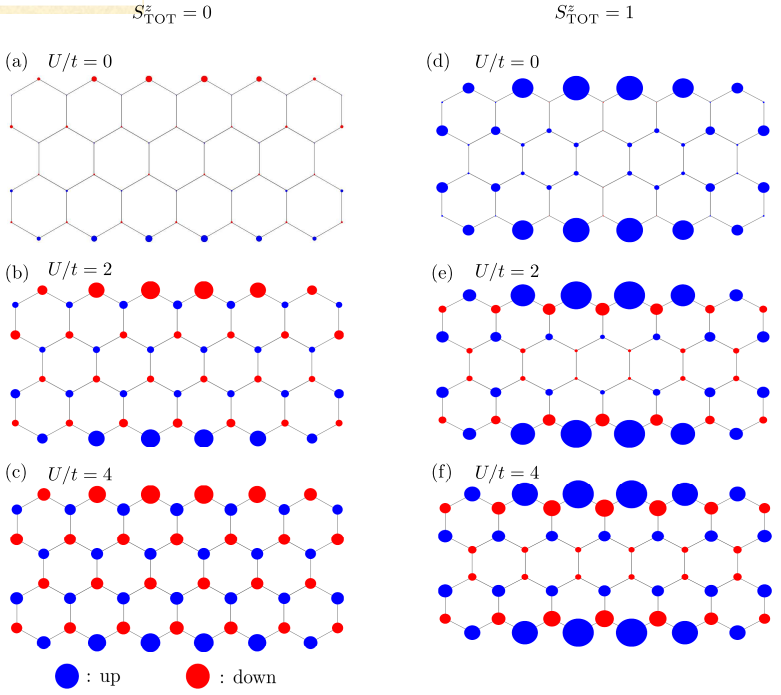
- Flat bands disappear when interaction is switched on
- Need TNS due to strong quantum fluctuations



- DMRG $D = 20000$, $U = 2$

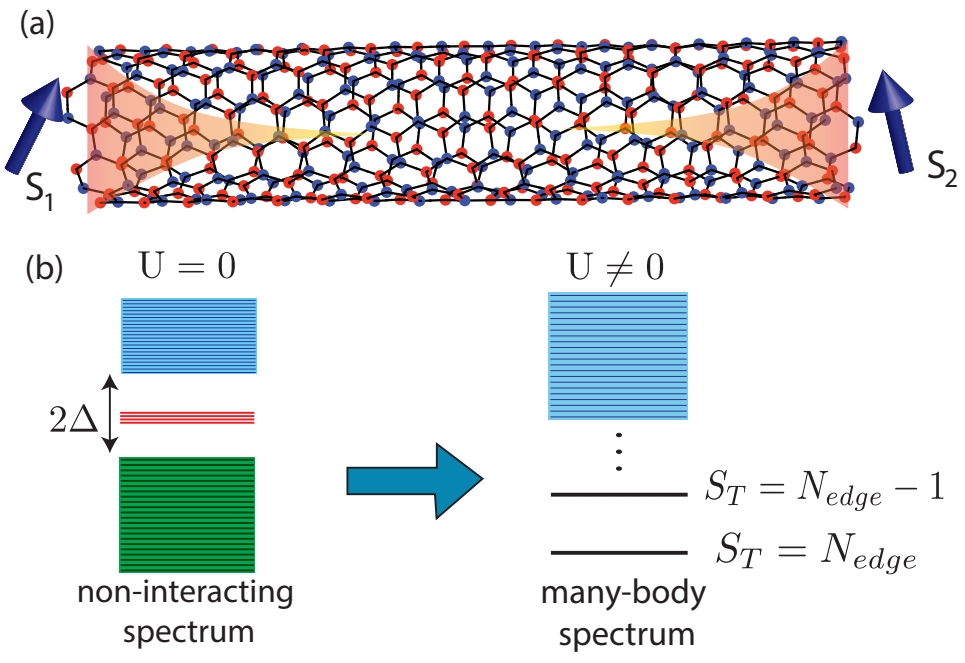
Problem revisited with modetransformation, monitoring emerging modes for zigzag, armchair, periodic BC etc.. Fraction of D is needed

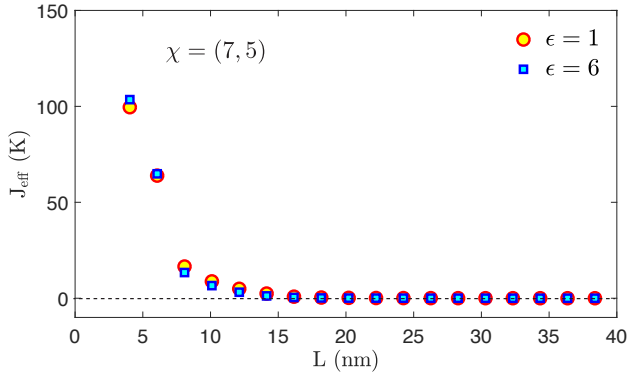
Mate, Vizkeleti, Szalay, Hagymasi, Ö.L.

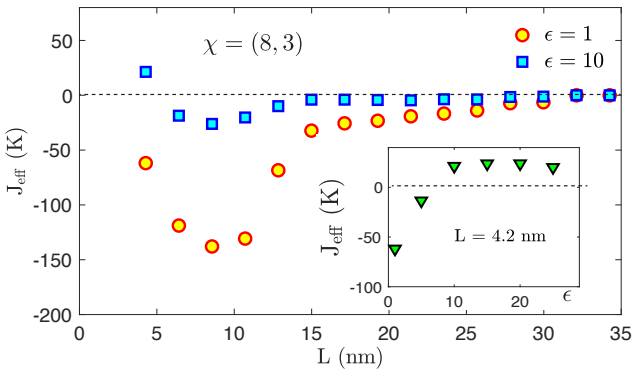


(left) S_i^z for the ground state in the presence of a pinning magnetic field at the bottom zigzag

Topologically protected, correlated end spin formation in carbon nanotubes Moca, Izumida, Dóra, Ö.L., Asboth, Zaránd (2019)

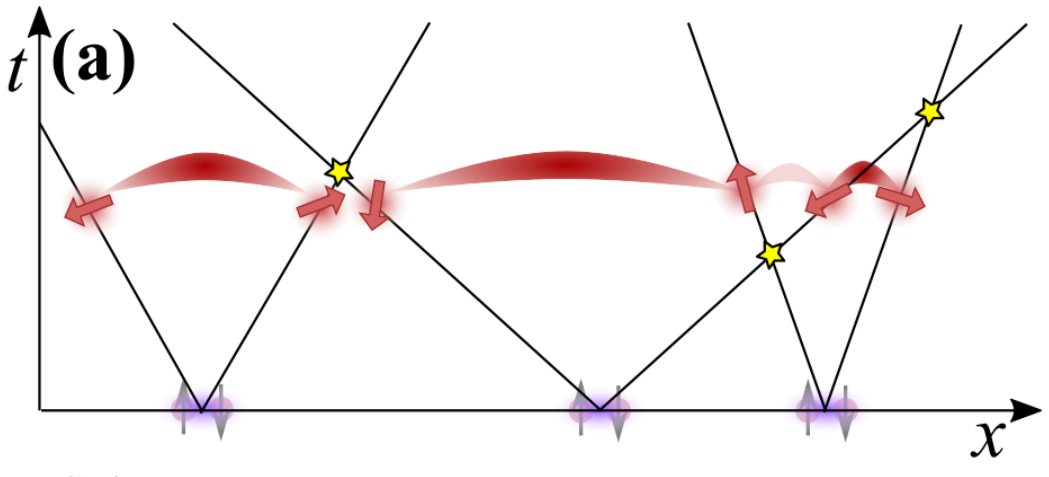


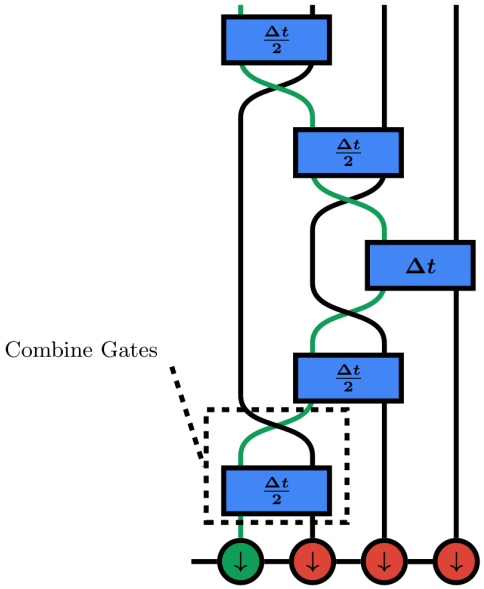




- $S_1 = S_2 = \frac{N_{\text{edge}}}{2}$
- Topological nanotubes spontaneously form double dot devices, which may provide a platform for quantum computation.
- Sign of the interaction can be changed by changing the dielectric constant of the environment.
- Coupling between ferromagnetic edge states is length and chirality dependent

Long time evolution Krumnow, Eisert, Ö.L. (2019)





- At time $t = 0$ we perturb the system.
- After the quench the quasiparticles collide with each other.
- There are different time-evolution methods for MPS which are currently in use to solve the time-dependent Schrödinger equation (TDSE).
- application of $\hat{U}(\delta_t) = e^{-i\delta_t \hat{H}}$, i.e. , $|\psi(t)\rangle \rightarrow |\psi(t+\delta_t)\rangle$
- time-evolving block decimation (TEBD), *MPOW*$^{\mathrm{I,II}}$, Krylov, time-dependent variational principle (TDVP)
- each has advatages and disadvateges.
- TDVP $\rightarrow$ general non-local Hamiltonians (quantum chemsitry)

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

Kinoshita, Hino, Bartlett, JCP (2005), Veis, Antalik, Neese, Ö.L., Pittner (2016)

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

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

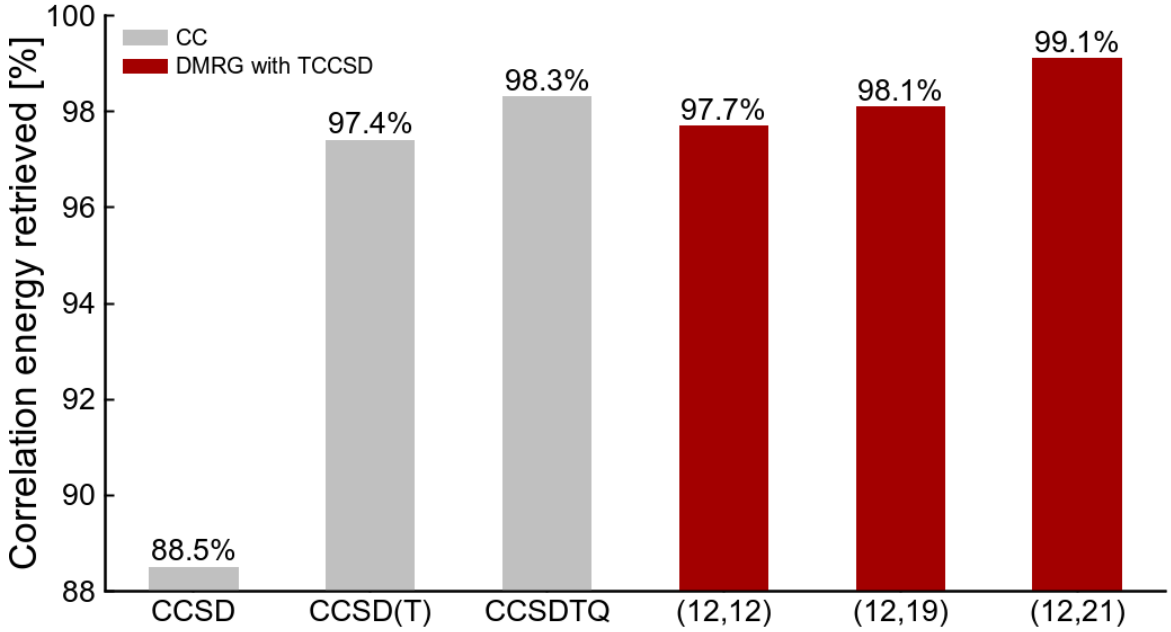
- $\mathcal{T}^{\mathrm{CAS}}$
  - amplitudes extracted from DMRG (CASCI) calculation
  - frozen during CC calculation
  - account for static correlation
- $\mathcal{T}^{\mathrm{ext}}$
  - determined through the usual CC
  - account for dynamic correlation

$$\begin{aligned} |\Psi_{\mathrm{TCCSD}}\rangle &= e^{\left(\mathcal{T}_1^{\mathrm{ext}}+\mathcal{T}_2^{\mathrm{ext}}\right)} e^{\left(\mathcal{T}_1^{\mathrm{CAS}}+\mathcal{T}_2^{\mathrm{CAS}}\right)} |\Psi_{\mathrm{ref}}\rangle \\ &\approx e^{\left(\mathcal{T}_1^{\mathrm{ext}}+\mathcal{T}_2^{\mathrm{ext}}\right)} |\Psi_{\mathrm{CASCI}}\rangle \end{aligned}$$

- Requires only small modifications of the CC code

Chromium dimer – correlation energies

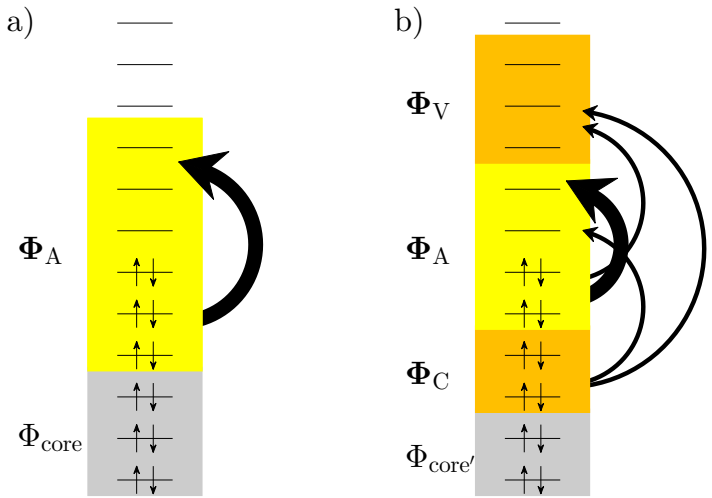
- Single-point calculation at 1.5 Å
- One-particle basis: RHF with Ahlrichs' SV basis set → (48e,42o)
- DMRG space selected based on $S^{(1)}$ profile
- DMRG performed with DBSS ($\epsilon_{\mathrm{tr}} \approx 10^{-7}$)
- Extrapolated DMRG by Olivares-Amaya et al. JCP 142, 034102, 2015 serves as a FCI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



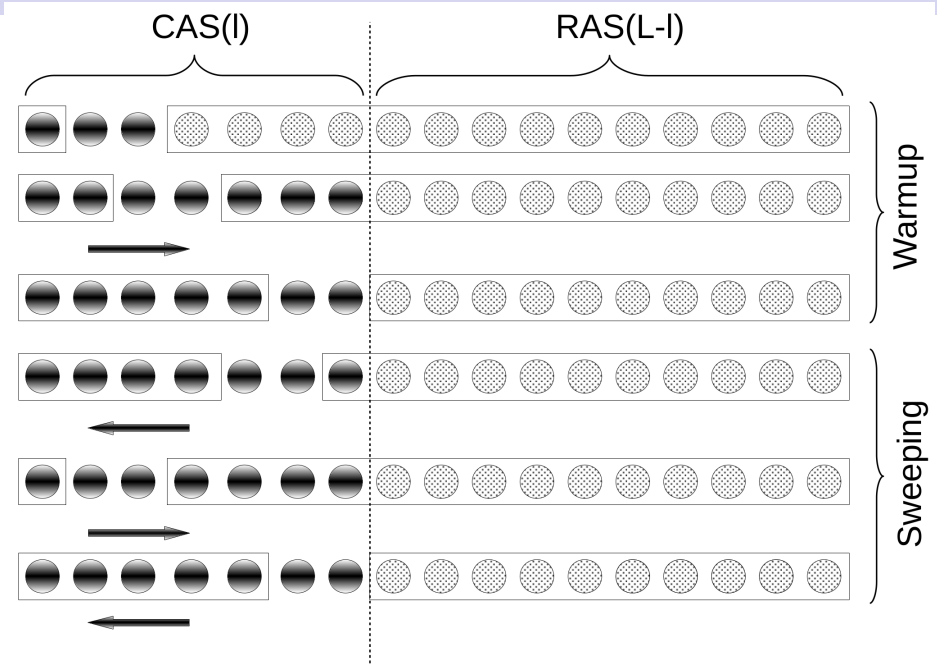
• DMRG-TCC has a quadratic error bound, Faulstich, Laestadius, Ö.L., Schneider, Kvaal(2019) but optimal CAS-EXT split only numerically

• Extensions: similarity transformed TCCSD, LPNO-TCCSD, DLPNO-TCCSD, TCCSDtq, 4c-DMRG-TCCSD, excited states.

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



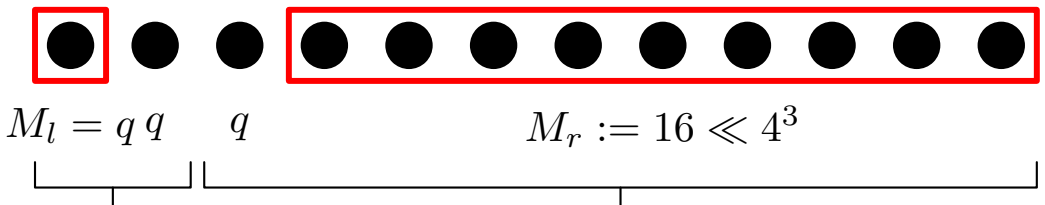
Schematic illustration of the CAS and RAS concepts.



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.,

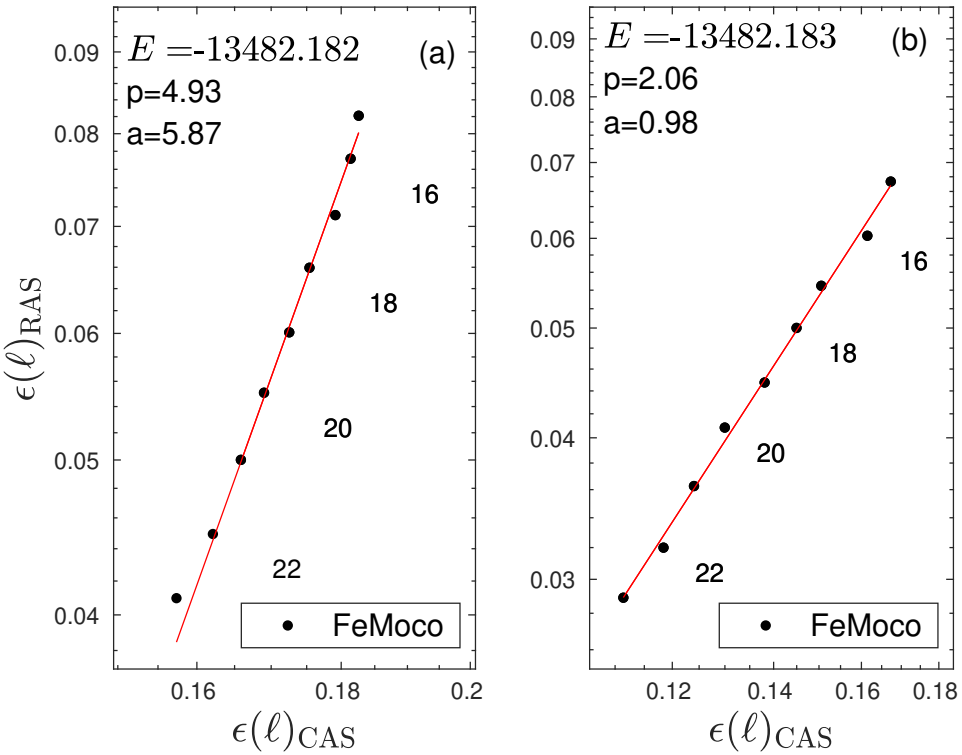
$$H = \underbrace{PHP}_{H_{\mathrm{CAS}\to\mathrm{CAS}}} + \underbrace{QHP}_{H_{\mathrm{CAS}\to\mathrm{RAS}}} + \underbrace{PHQ}_{H_{\mathrm{RAS}\to\mathrm{CAS}}} + \underbrace{QHQ}_{H_{\mathrm{RAS}\to\mathrm{RAS}}}$$

method	energy (Ha)	Δ_{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.

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)

Ab initio theory of negatively charged boron vacancy qubit in hBN Ivády, Barcza, Thiering,Li, Hamdi, Chou, Ö.L., Gali (2019)

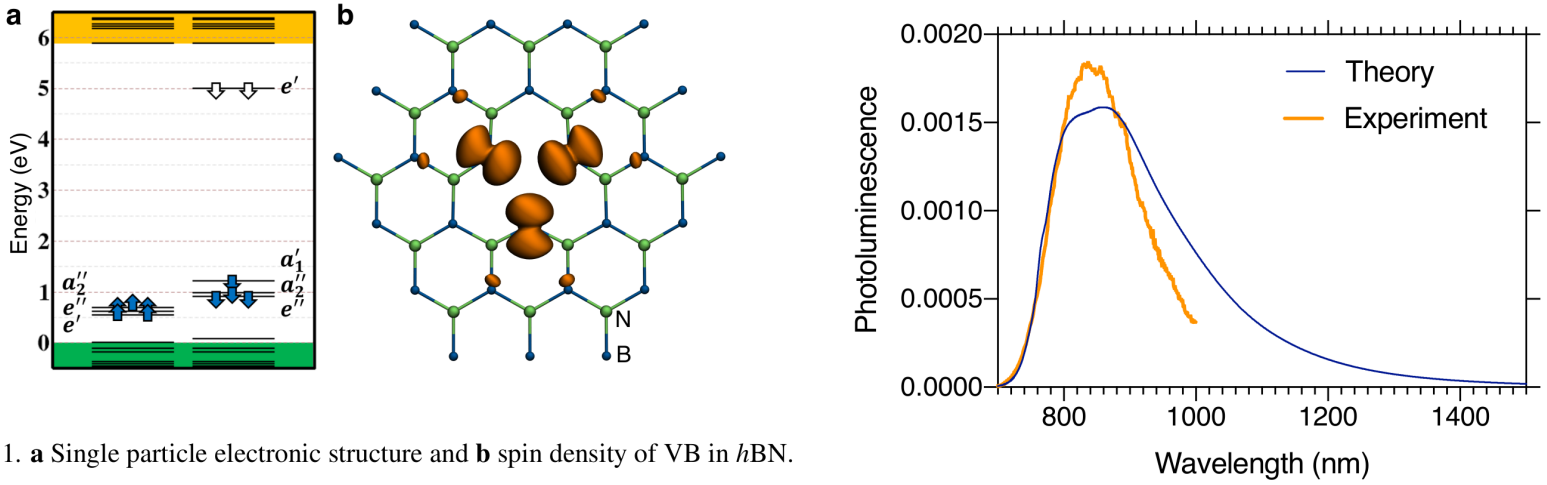


FIG. 1. **a** Single particle electronic structure and **b** spin density of VB in *h*BN.

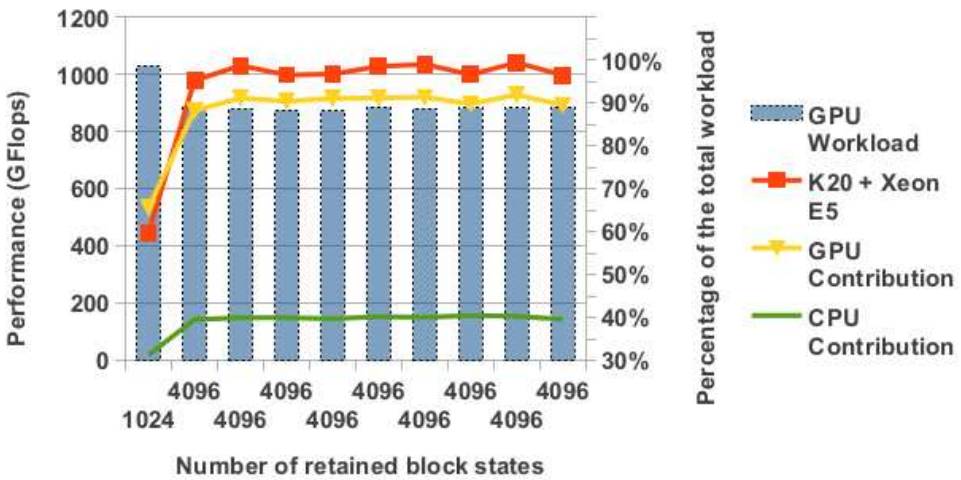
- Novel combiation of DFT and DMRG for extended systems to treat excited states
- DMRG on top of plane-wave Kohn-Sham orbitals
- ab initio results explain magneto-optical properties of VB in hBN.

Highly tunable magneto-optical response from MgV color centers in diamond Pershin, Barcza, Ö.L., Gali (2021)

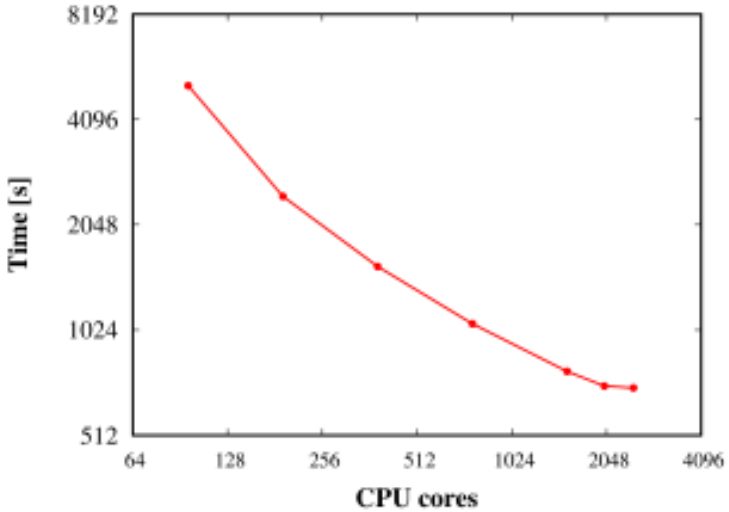
- Potential of magnesium-vacancy (MgV) in diamond to operate as a qubit by computing the key electronic and spin properties with robust theoretical methods: Unprecedented control over the magneto-optical response from a qubit by modulating the operational conditions.

Towards exascale computaions on supercomputers

GPU: MPS and TNS
on kilo-processor architectures:
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)]

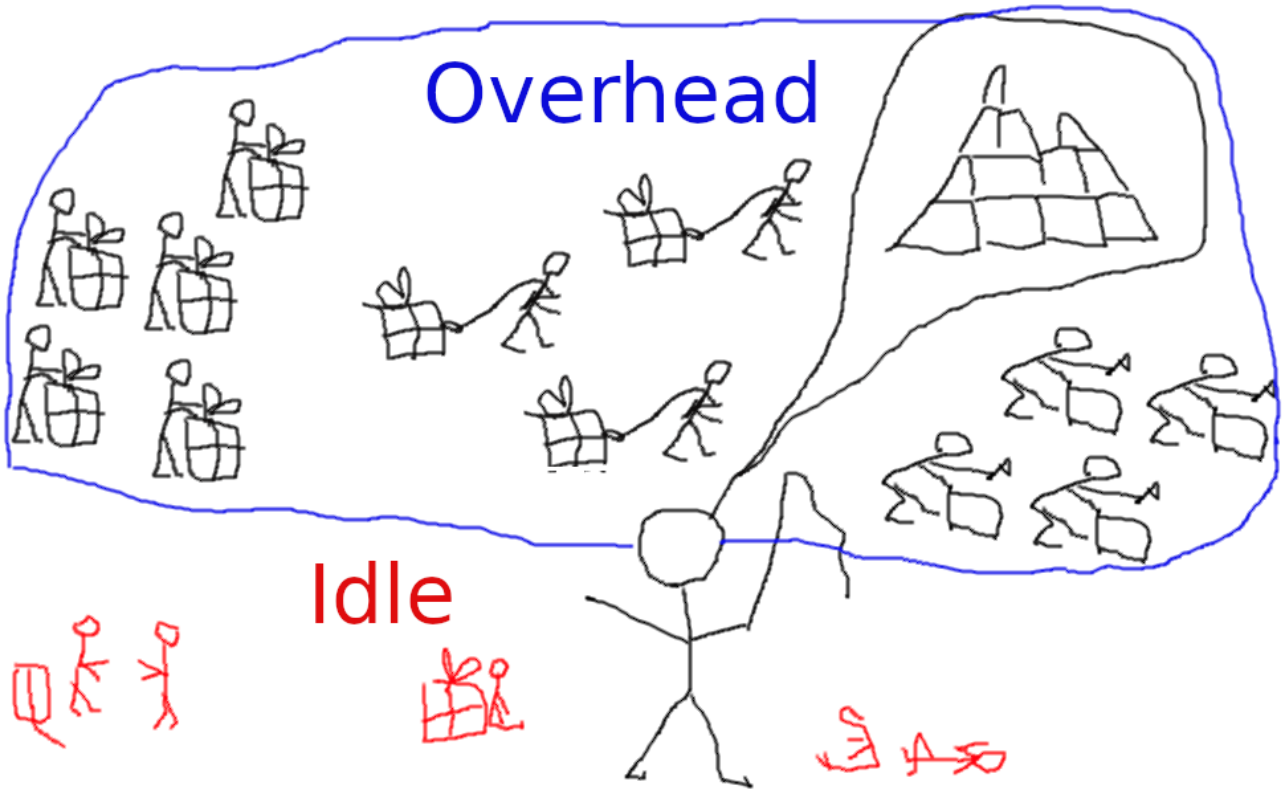
Centralized scheduling: unideal society

- 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

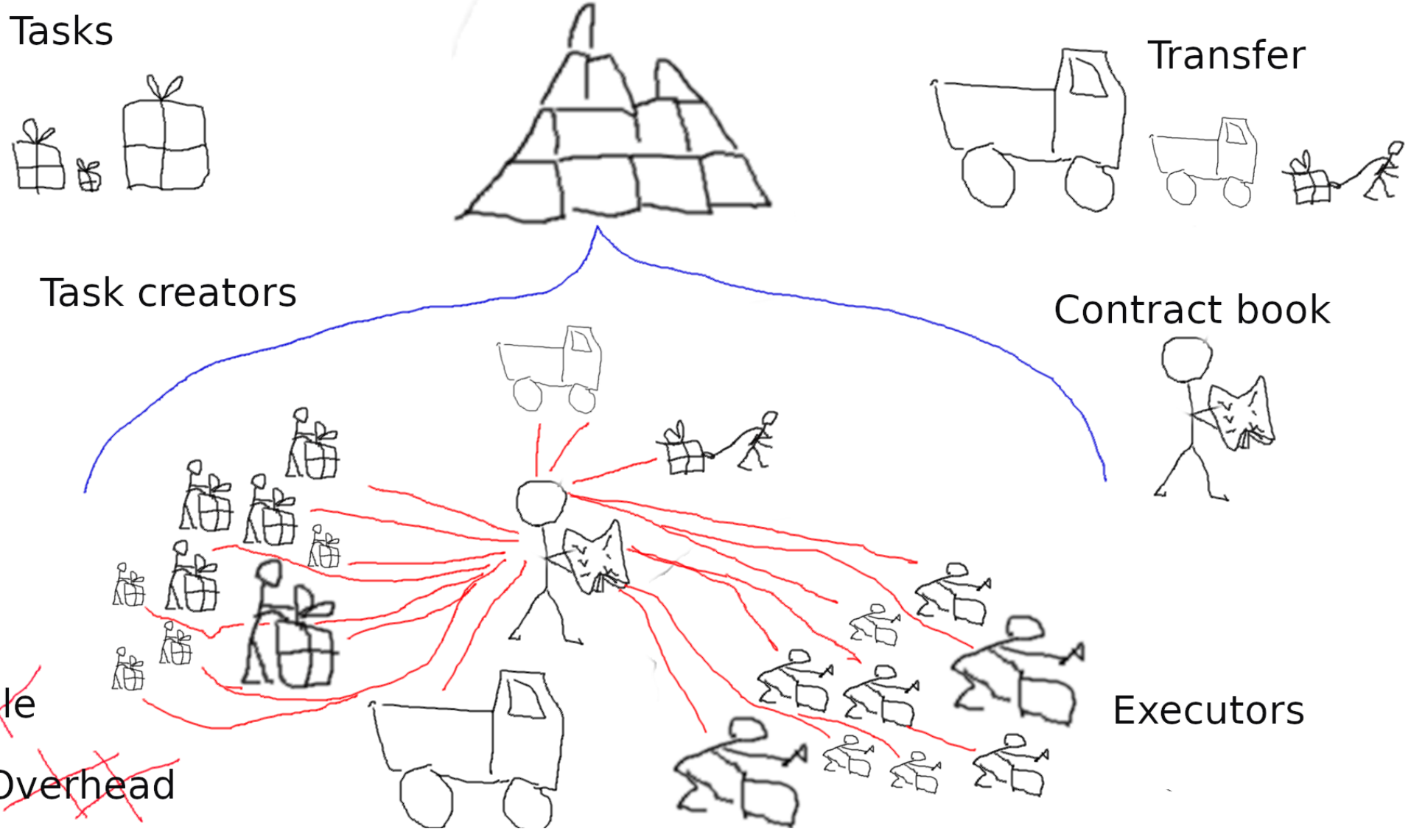
- Central scheduler performs lot of measurements, estimations, communication to rearrange tasks and workers → huge overhead



- 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

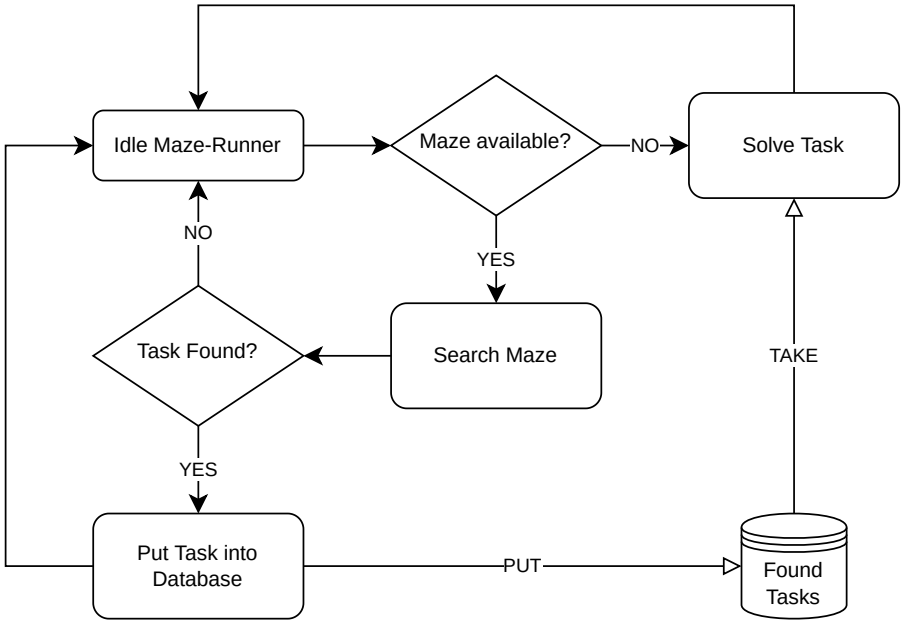
- Central unit: Contractor, contract book (only meta-data communicated, boolean-like bookkeeping flags)
- Everybody is motivated to achieve global aim



Efficient task processig: Maze-Runners Menczer, ÖL (2023)

Nemes, Barcza, Nagy, Ö.L., Szolgay (2014)

- 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.

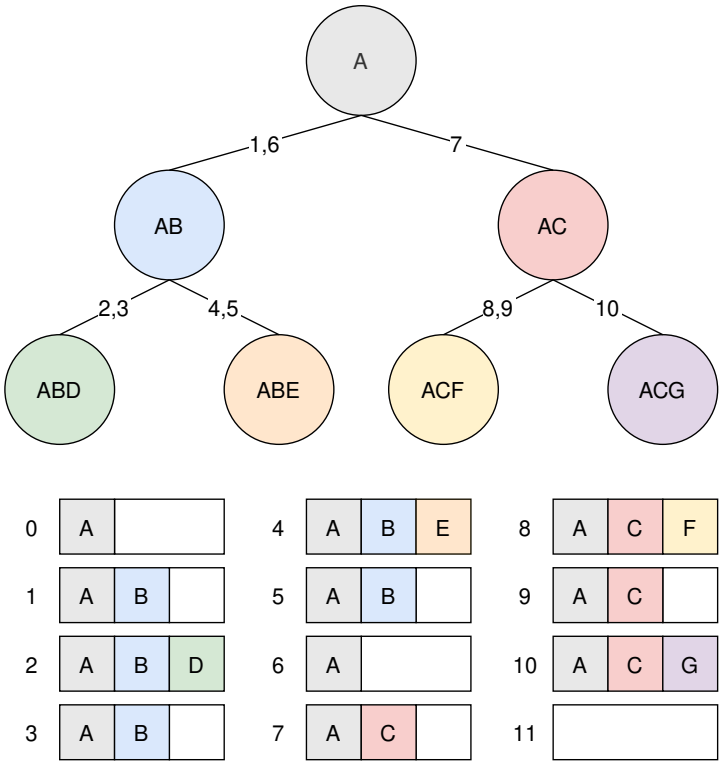


Life Cycle of a Maze-Runner Thread.

- 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.

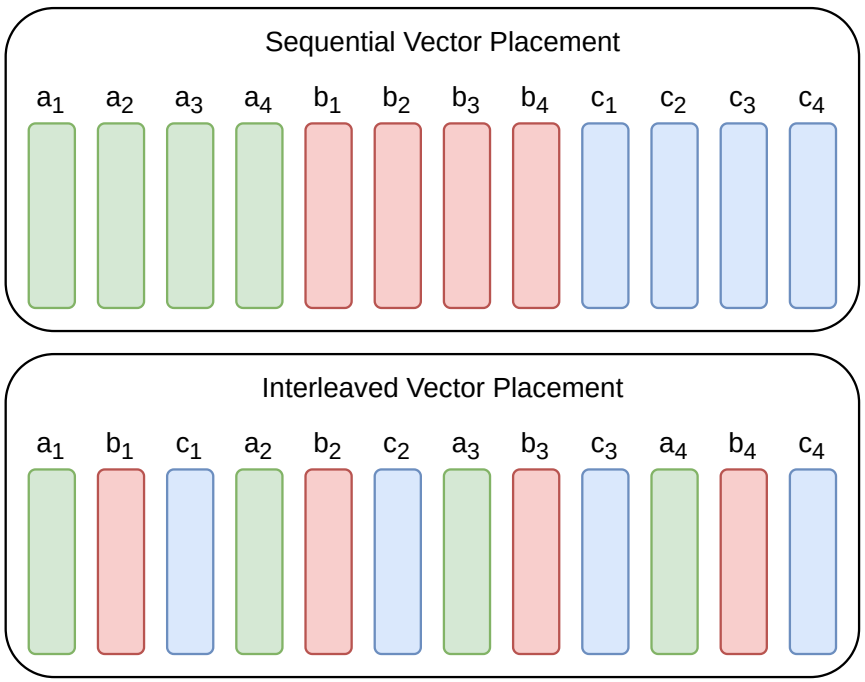
The numbers represent the order in which the vertices are visited.

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.



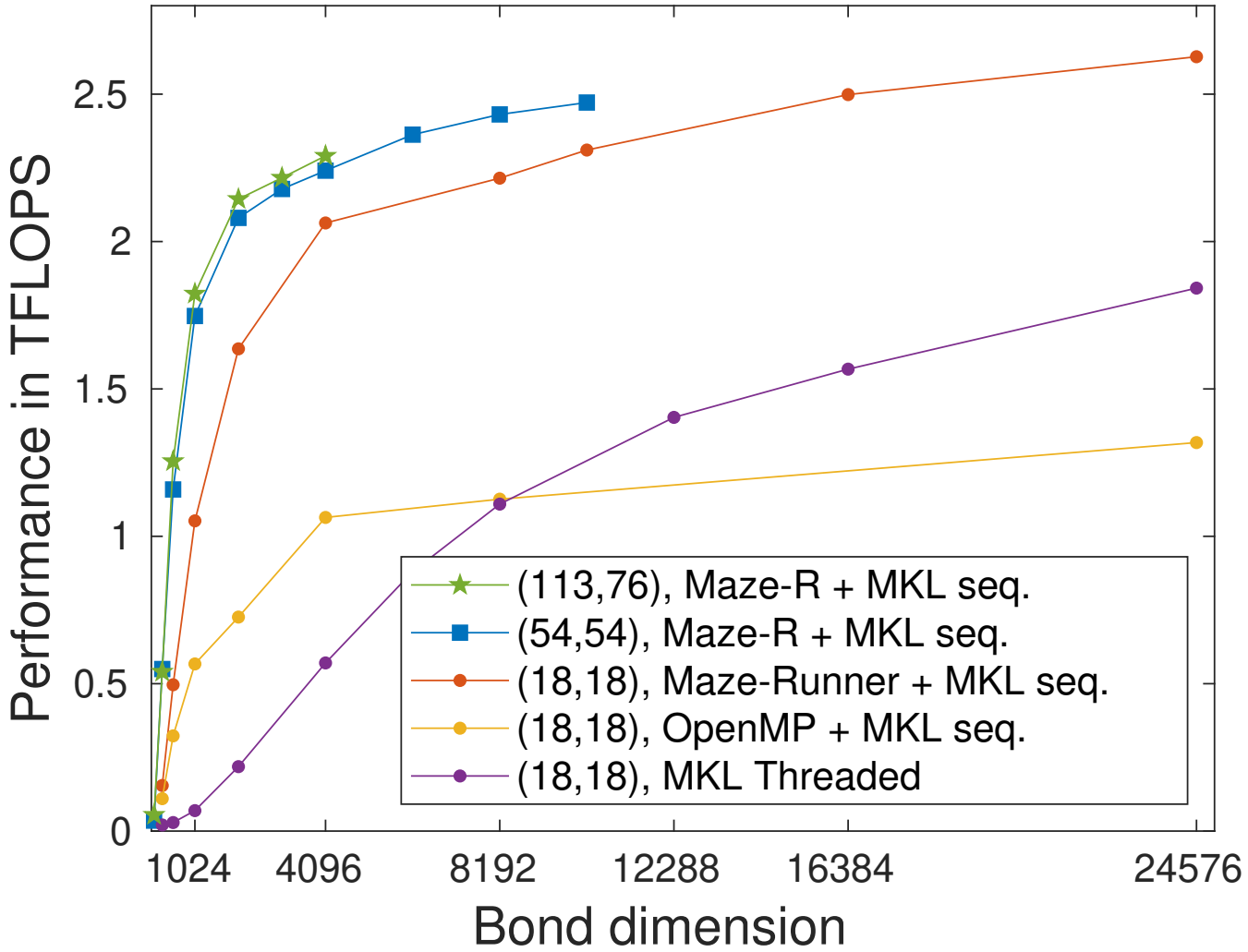
Vectorized output of a Strided Batched GEMM operation.

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.

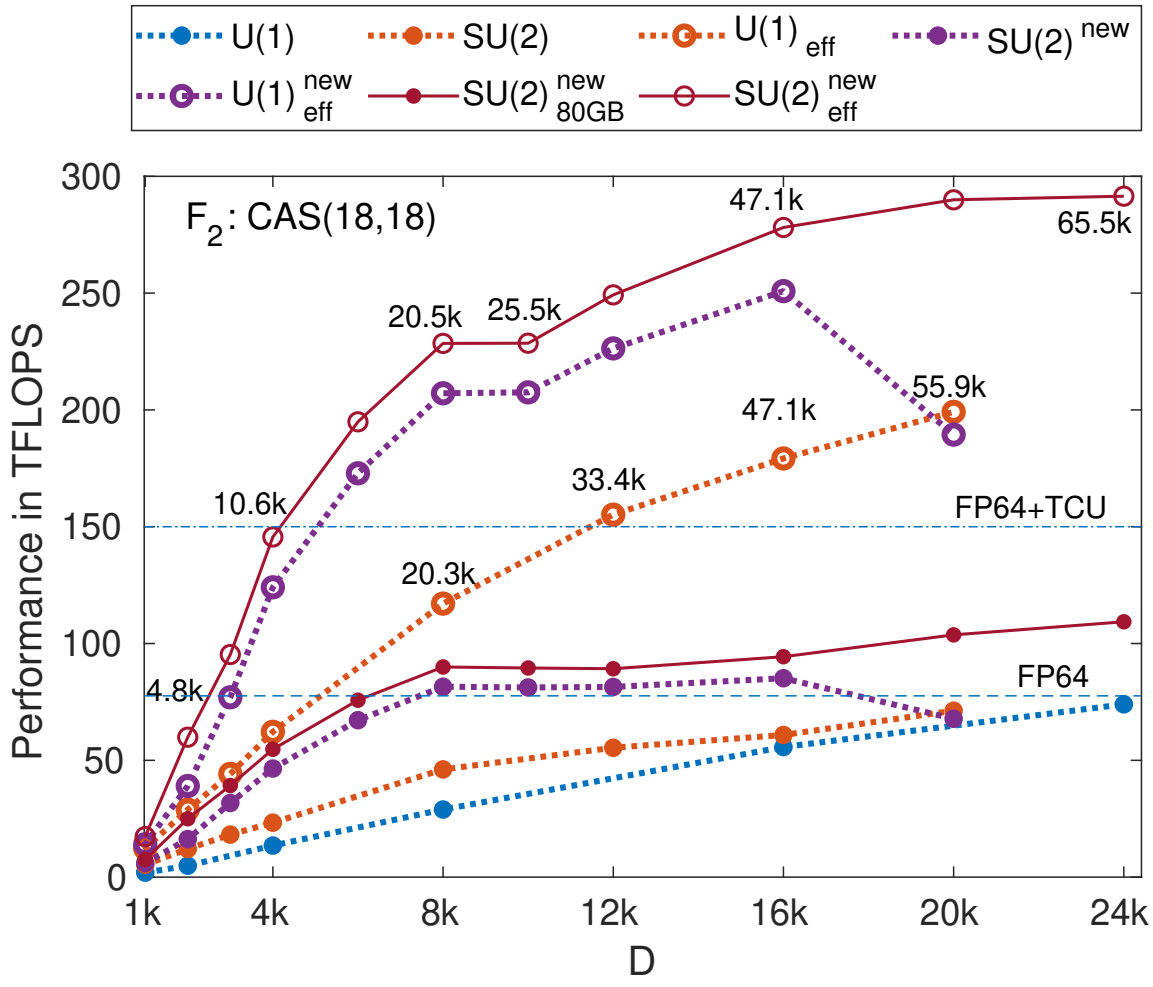
- We can perform batched type chained matrix multiplications without sum reduction at the end.

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.

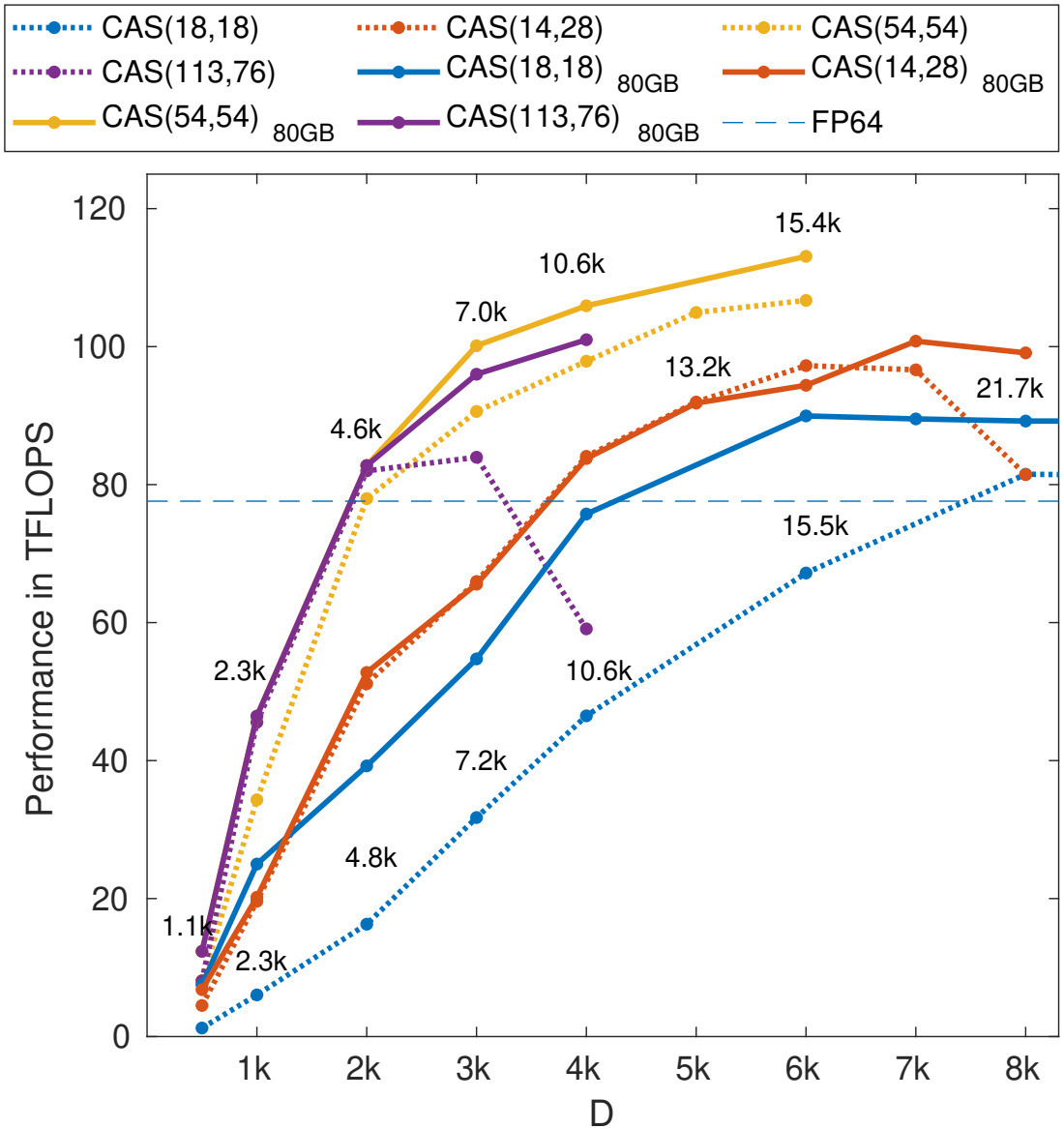
Boosting the effective performance via non-Abelian symmetries



A. Menczer, Ö.L (2023), CAS(18,18)

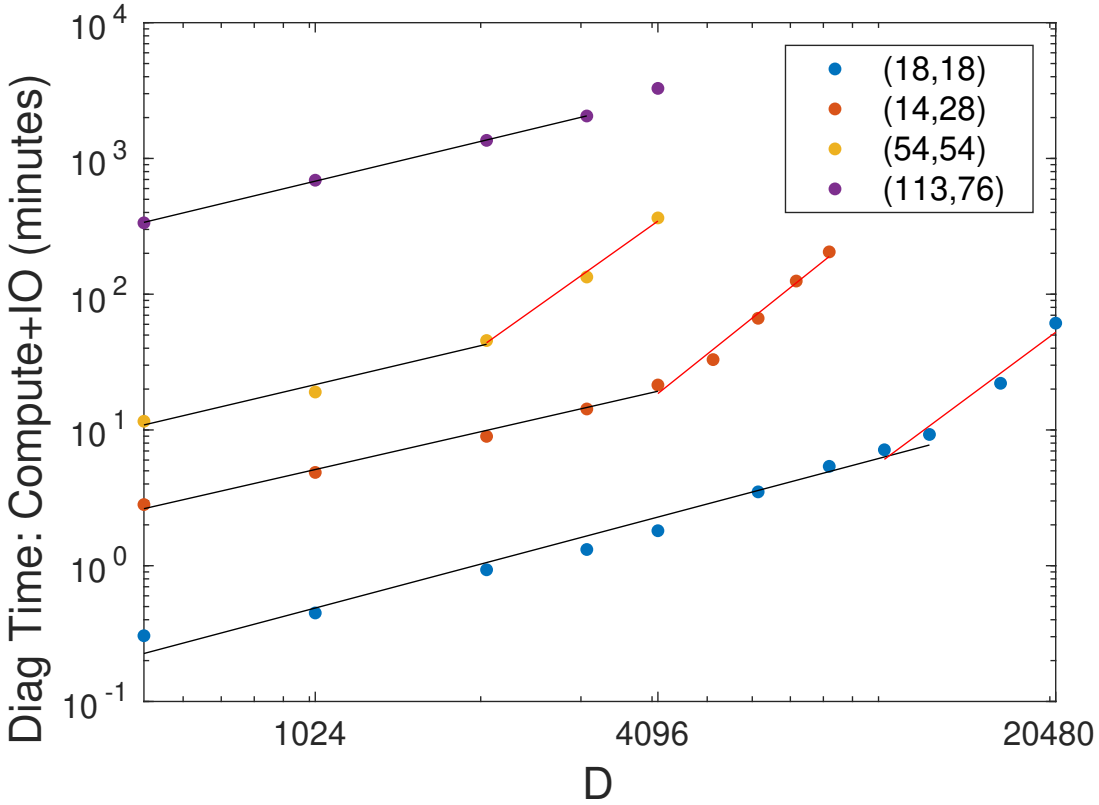
- New mathematical model for parallelization $\rightarrow$ felxibe scaling
- $D_{SU(2)} = 24576 \rightarrow D_{U(1)} = 2^{16} \rightarrow$ FCI solution

Utilization of highly specialized tensor core units (TCU)



- Large CAS spaces A. Menczer, Ö.L (2023), FeMoco(113,76)
- We reached 116 TFLOPS $>$ 76 TFLOPS of the FP64 limit of NVIDIA

Reducing D^3 scaling to linear scaling

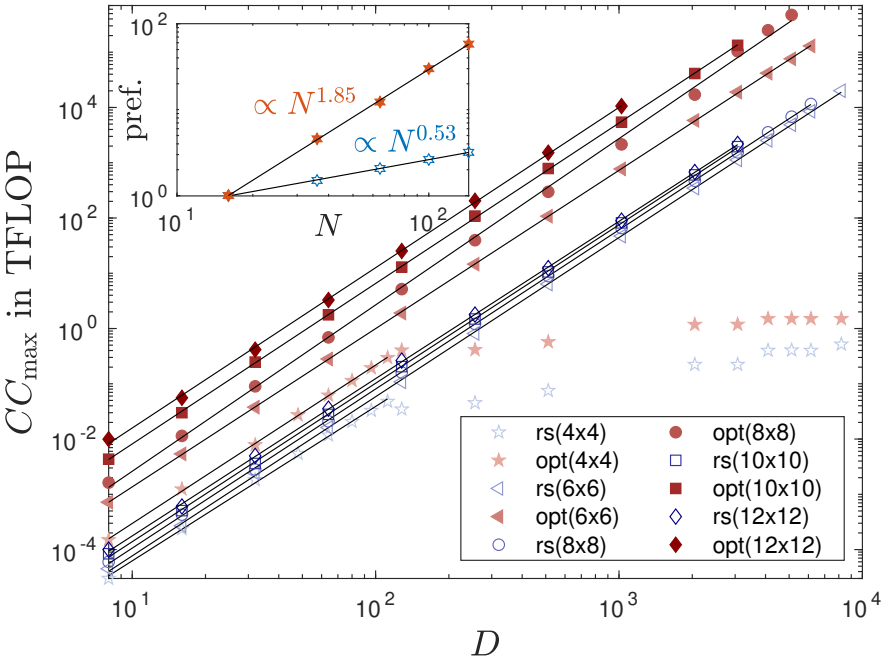


- New model to utilize NVIDIA D2D links. A. Menczer ÖL (unpublished 2023)

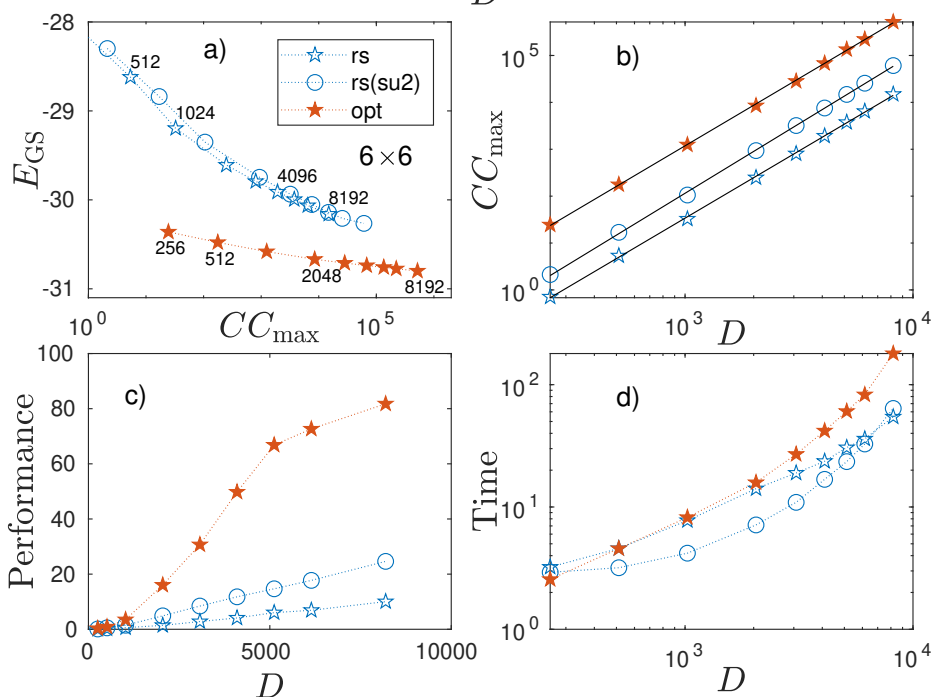
- NVIDIA DGX H100 and Grace Hopper GH200:
Testing performance up to $\sim$ 240 TFLOPS in collab with NVIDIA and SandboxAQ M. van Damme, A. Menczer, M. Ganahl, J. Hammond, Ö.L

- Combination of our MPI and GPU kernels:
multiNode-multiGPU $\rightarrow$ petascale computing. A. Menczer ÖL (unpublished

Maximum computational complexity for 2D $t - t' - V$ model



- 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.



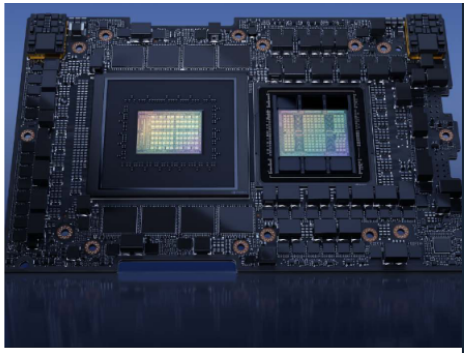
- Half-filled 6×6 Hubbard model at $U = 4$ on a torus geometry
- Performance in TFLOPS
- Time in minutes

Our TNS/DMRG code will be used as one of the benchmarks

NVIDIA

NVIDIA GH200 Grace Hopper Superchip

The breakthrough accelerated CPU for large-scale AI and high-performance computing (HPC) applications.



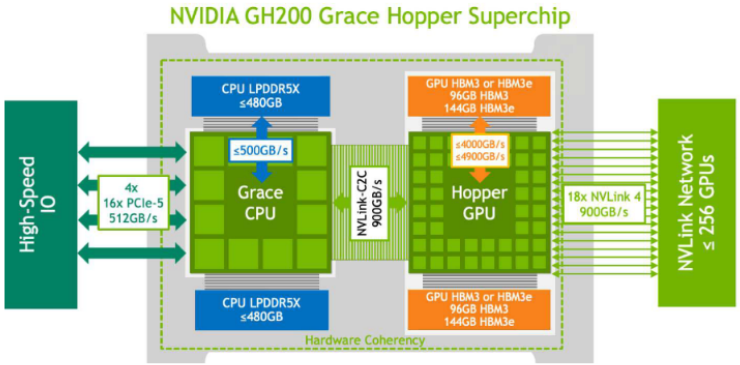
The World's Most Versatile Computing Platform

The NVIDIA Grace Hopper™ architecture brings together the groundbreaking performance of the NVIDIA Hopper™ GPU with the versatility of the NVIDIA Grace™ CPU in a single superchip, connected with the high-bandwidth, memory-coherent NVIDIA® NVLink® Chip-2-Chip (C2C) interconnect.

NVIDIA NVLink-C2C is a memory-coherent, high-bandwidth, and low-latency interconnect for superchips. The heart of the GH200 Grace Hopper Superchip, it delivers up to 900 gigabytes per second (GB/s) of total bandwidth, which is 7X higher than PCIe Gen5 lanes commonly used in accelerated systems. NVLink-C2C enables applications to oversubscribe the GPU's memory and directly utilize NVIDIA Grace CPU's memory at high bandwidth. With up to 480GB of LPDDR5X CPU memory per GH200 Grace Hopper Superchip, the GPU has direct access to 7X more fast memory than HMB3 or almost 8X more fast memory with HBM3e. GH200 can be easily deployed in standard servers to run a variety of inference, data analytics, and other compute and memory-intensive workloads. GH200 can also be combined with the NVIDIA NVLink Switch System, with all GPU threads running on up to 256 NVLink-connected GPUs and able to access up to 144 terabytes (TB) of memory at high bandwidth.

Key Features

- 72-core NVIDIA Grace CPU
- NVIDIA H100 Tensor Core GPU
- Up to 480GB of LPDDR5X memory with error-correction code (ECC)
- Supports 96GB of HBM3 or 144GB of HBM3e
- Up to 624GB of fast-access memory
- NVLink-C2C: 900GB/s of coherent memory



Startseite
Forschung +
Studium +
Department −
Aktuelles & Events −
Veranstaltungen
Geschichte
Talent Management & Diversity +
Outreach Aktivitäten +
Firmen-Kooperationen +
Gründungen
Statistikberatung
Parabelrutsche
Personen +

Startseite › Department › Aktuelles & Events

Workshop: Recent progress on tensor network methods, April 22-25, 2024

TUM Institute for Advanced Study, Munich

The aim of the workshop is to bring together condensed matter physicists, mathematicians and theoretical chemists to continue the exploration of this active and growing field of research and to stimulate further developments of tensor network methods.

Foto: Andreas Heddergott / TUM

The focus will be on innovative ideas for moving beyond current limits of quantum many body simulations despite the major challenges of high-dimensionality and accuracy. Topics include the interplay between modes, rank truncation and network topology, hybridization with other approaches, and parallelization.

Organizers

- Thomas Barthel, Duke University
- Gero Friesecke, Technical University of Munich
- Henrik Larsson, University of California, Merced
- Örs Legeza, TUM-IAS Hans Fischer Senior Fellow & Wigner Research Centre for Physics, Budapest

The workshop is open to young researchers, who are furthermore encouraged to apply for presenting a contributed talk or poster on registration. Some funding for young researchers may be available on request.

The workshop starts on **Monday, 22nd April at 9am** and finishes on **Thursday, 25th April 2024 at approx. 1pm.**

Registration

Please register here by **March 15th 2024**

Department of Mathematics

Technische Universität München
TUM School of Computation, Information and Technology

Boltzmannstraße 3
85748 Garching

Conclusion

- Tensor topologies together with proper basis representations are important for efficient data sparse representaion of the wavefunction
- Local mode transformation: MPS/TNS based black-box tool to improve basis
- Long time evolution with adaptive mode transformation is a promising direction
- Combination of TNS with other (conventional) methods can exploit benefits of the individual methods
- Massive Parallelization
- $\rightarrow$ Simulation of realistic material properties

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