Towards exascale computation via tensor network state algorithms

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

Örs Legeza

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

Institute for Advanced Study, Technical University of Munich, Germany

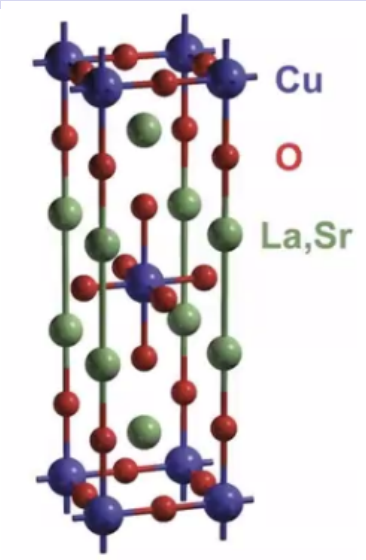
Jülich, 06.12.2023

in collaboration with

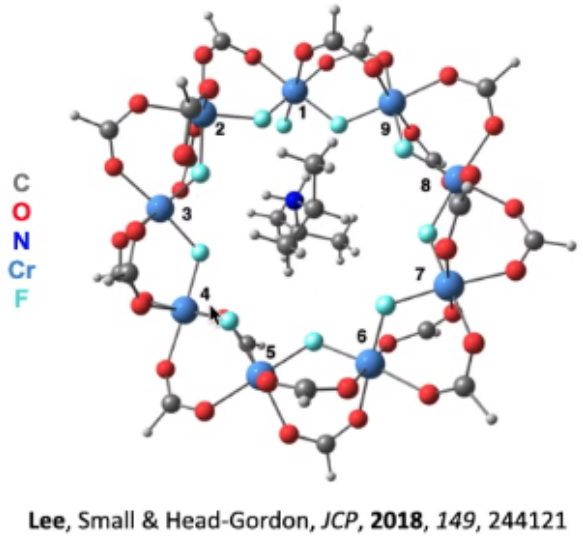
- more than 30 research groups worldwide from condensed matter physics, quantum chemistry, nuclear physics, quantum information theory and computer science
- High-Performance Computing Center Stuttgart, Germany
- Pacific Northwest National Laboratory (PNNL), USA
- National Energy Research Scientific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

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

Strong correlations between electrons $\rightarrow$ exotic materials



High T_c superconductors

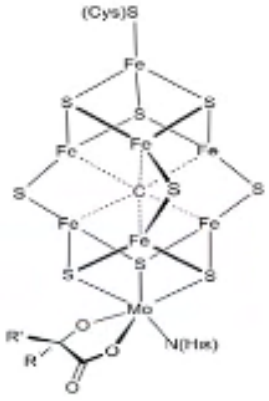


Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

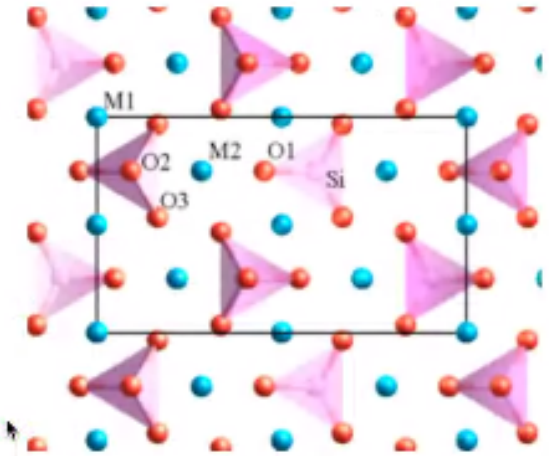
Single molecular magnets (SMM)

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

"FeMoco"



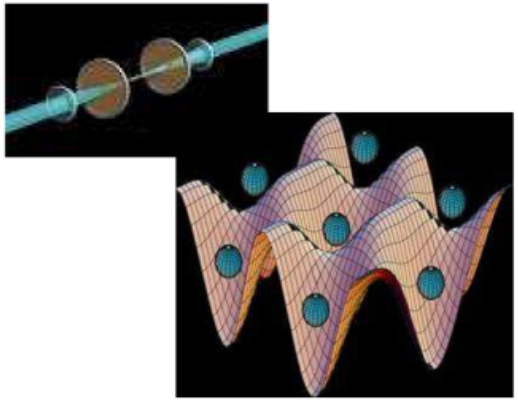
Nitrogen fixation



Battery technology

Experimental realizations: optical lattices

Numerical simulations: model systems



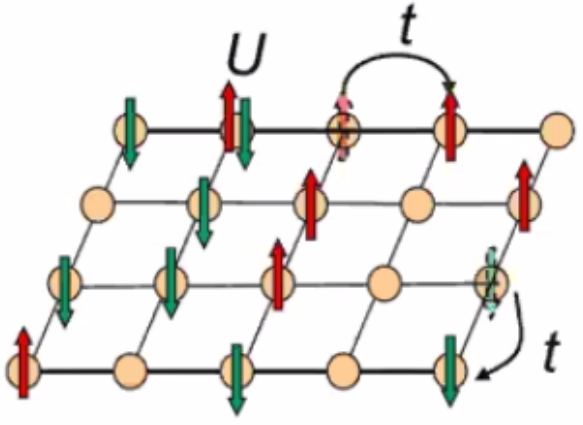
Atoms (represented as blue spheres) pictured in a 2D-optical lattice potential

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

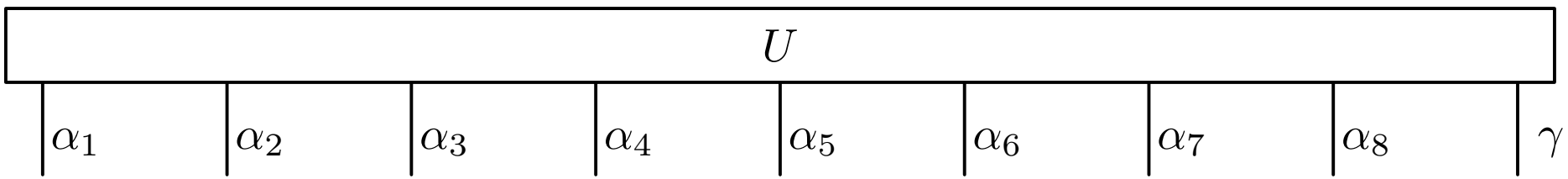
Classical or quantum computers?

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 $\gamma = 1, \dots, m$.

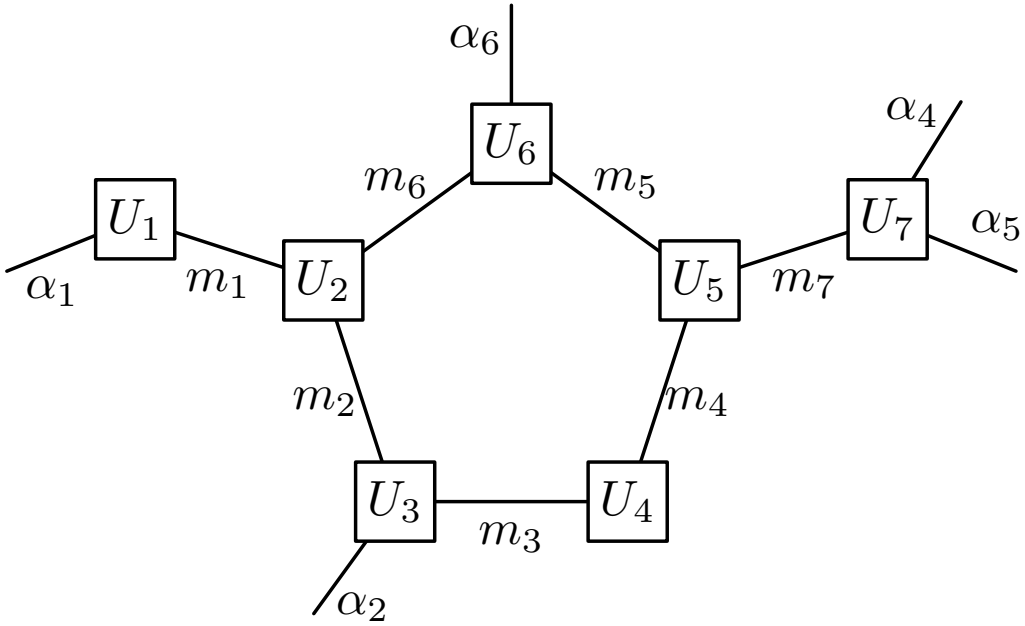


- α is called 'physical' leg
- In a spin-1/2 model $\alpha_i \in \{\downarrow, \uparrow\}$.
- In a spin-1/2 fermionic model $\alpha_i \in \{0, \downarrow, \uparrow, \uparrow\downarrow\}$.

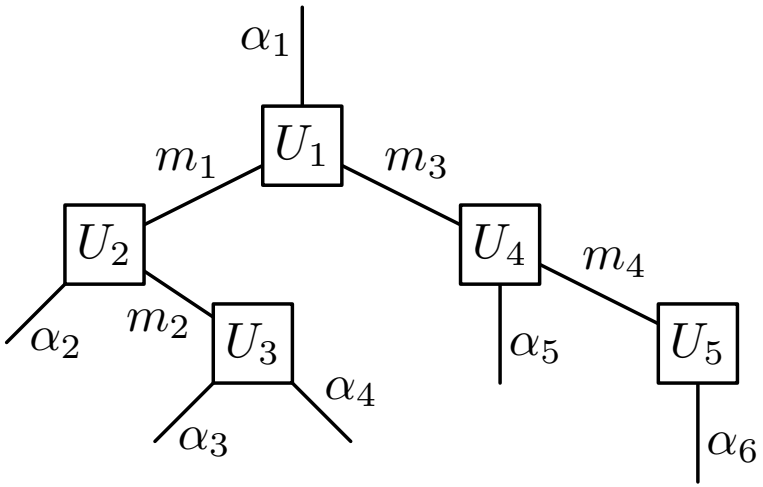
$\dim \mathcal{H}_d = \mathcal{O}(n^d)$ Curse of dimensionality!

$\rightarrow$ need efficient data-sparse representation

Tensor product representation



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

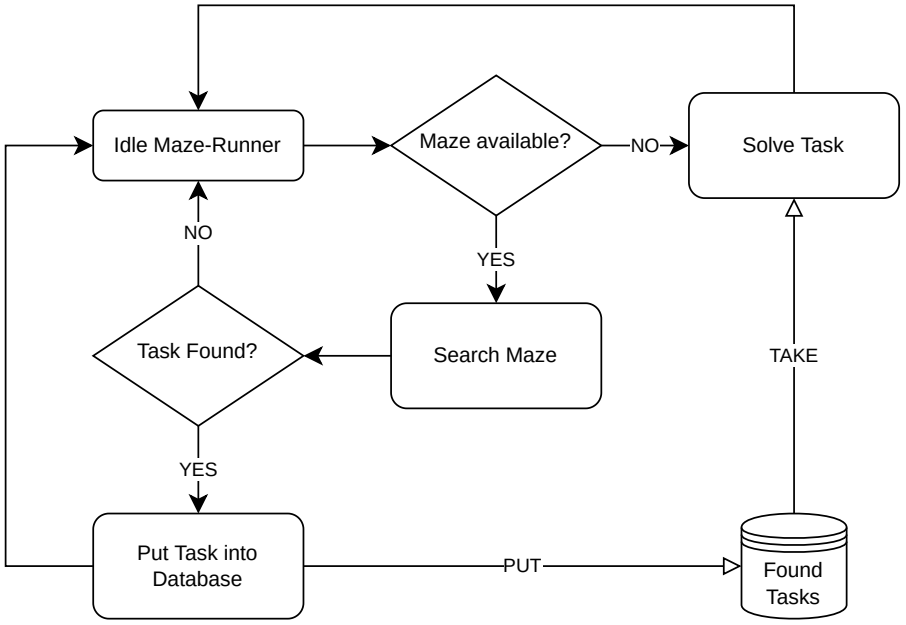


An arbitrary example of a tensor tree (loop free).

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.

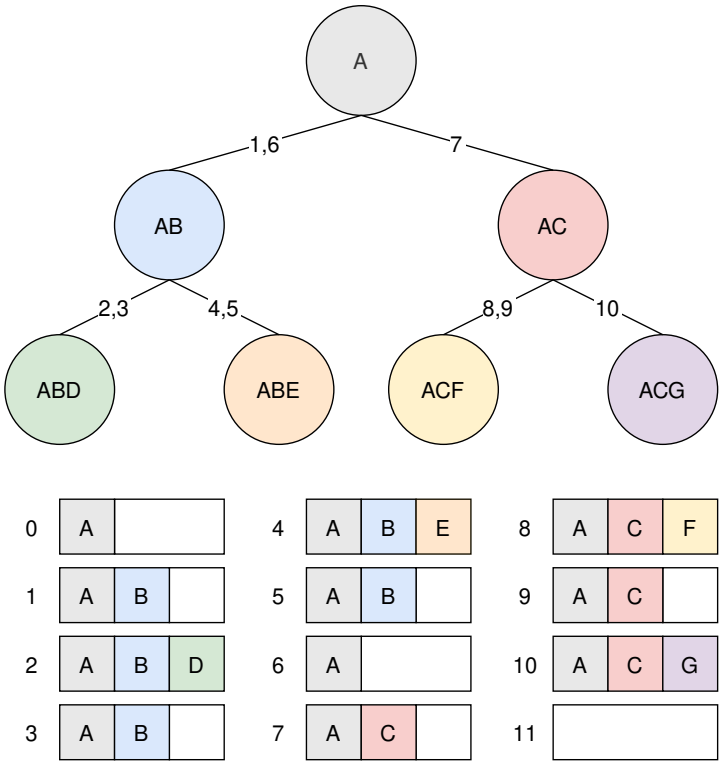


Life Cycle of a Maze-Runner Thread.

- 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.

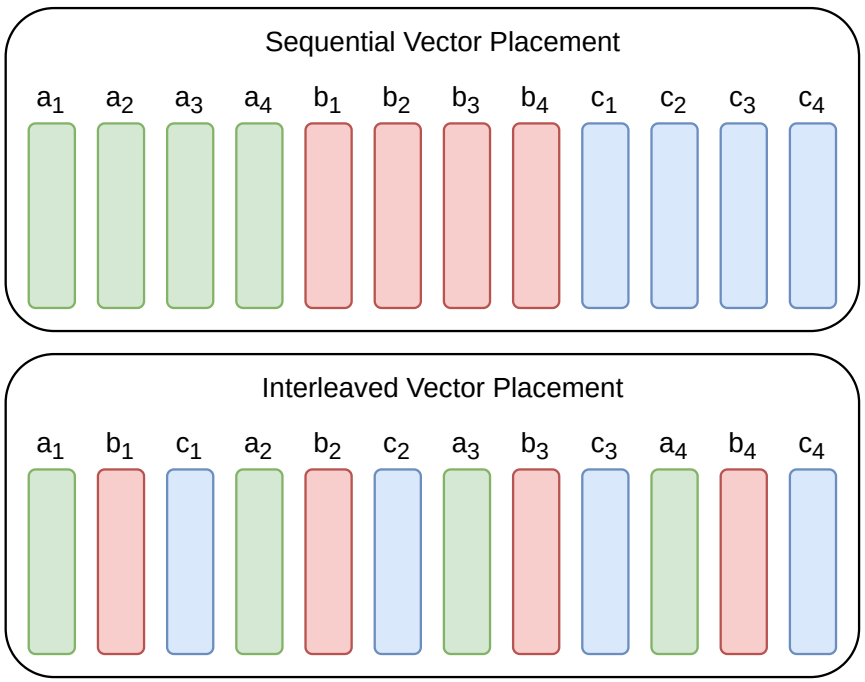
The numbers represent the order in which the vertices are visited.

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.



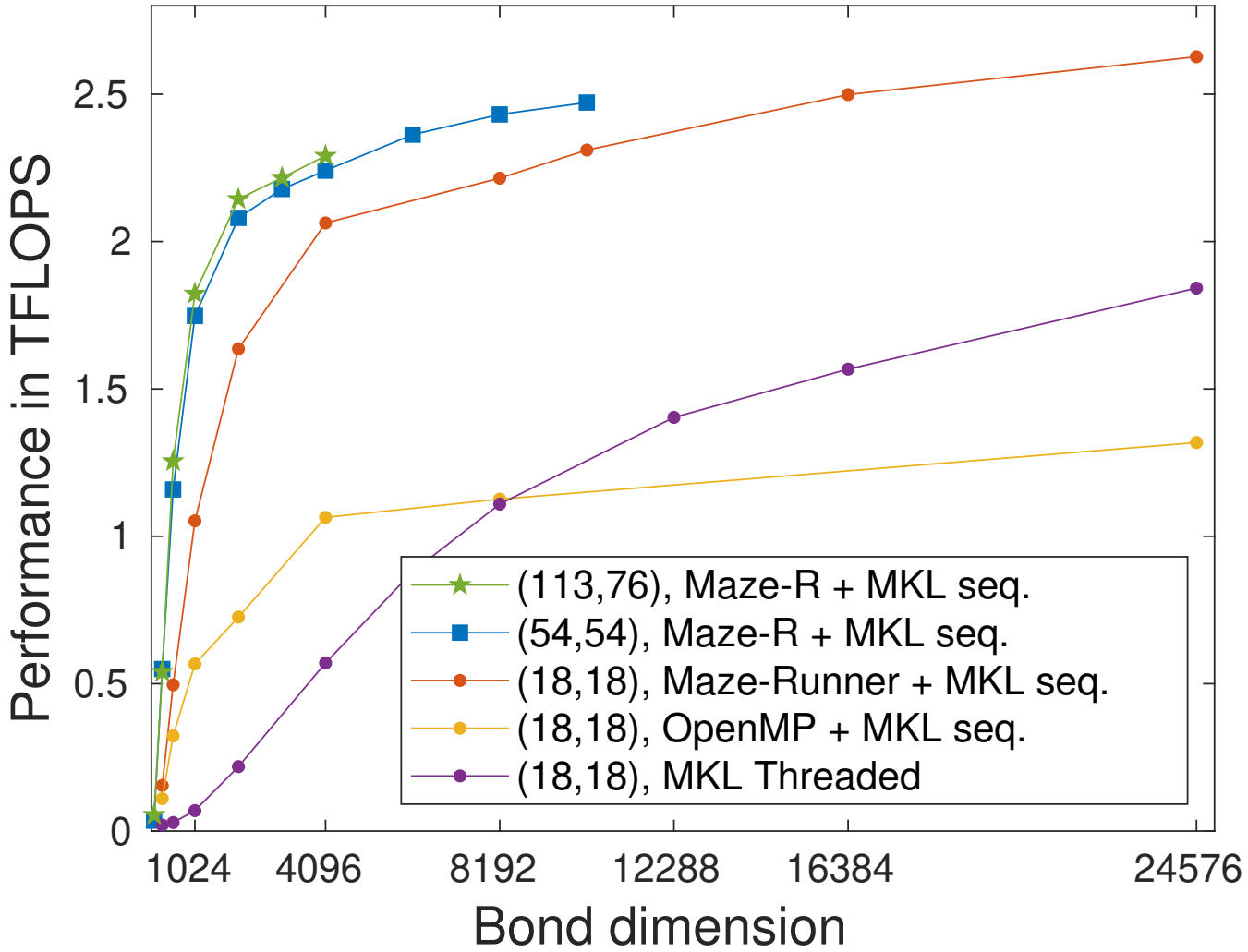
Vectorized output of a Strided Batched GEMM operation.

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.

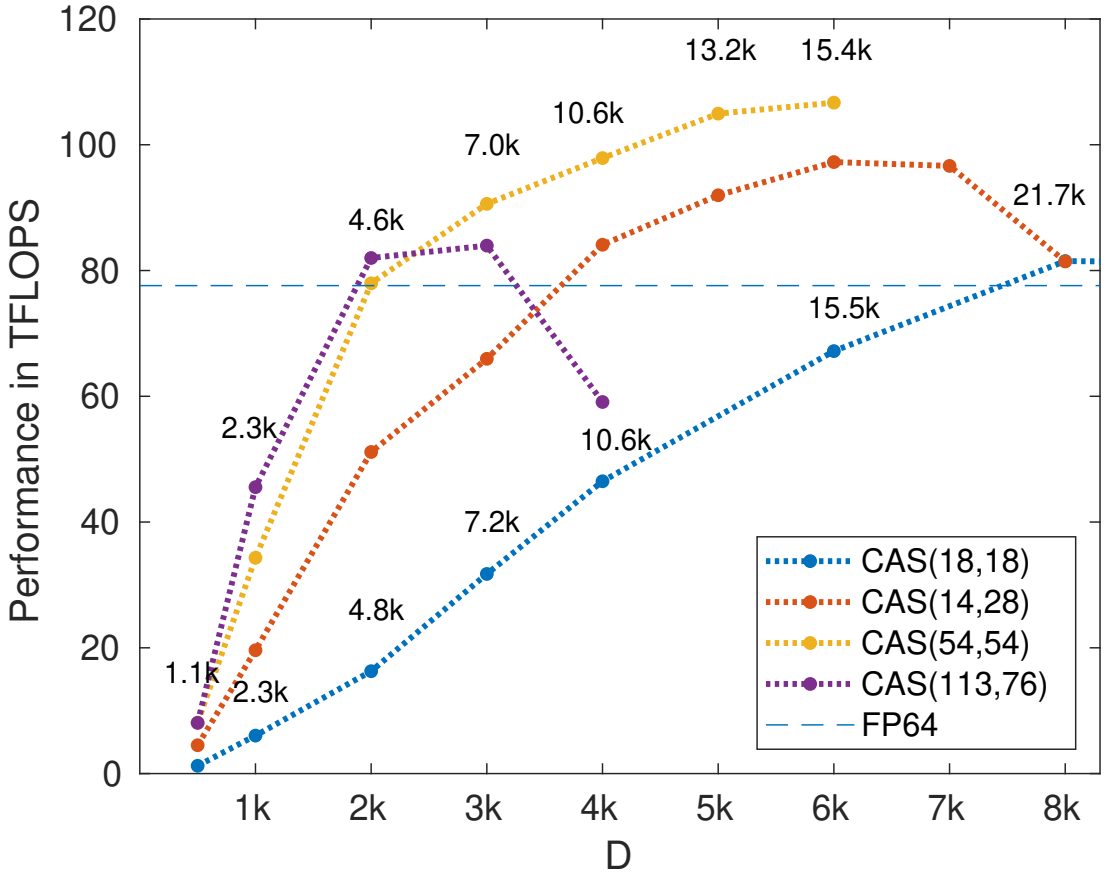
- We can perform batched type chained matrix multiplications without sum reduction at the end.

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

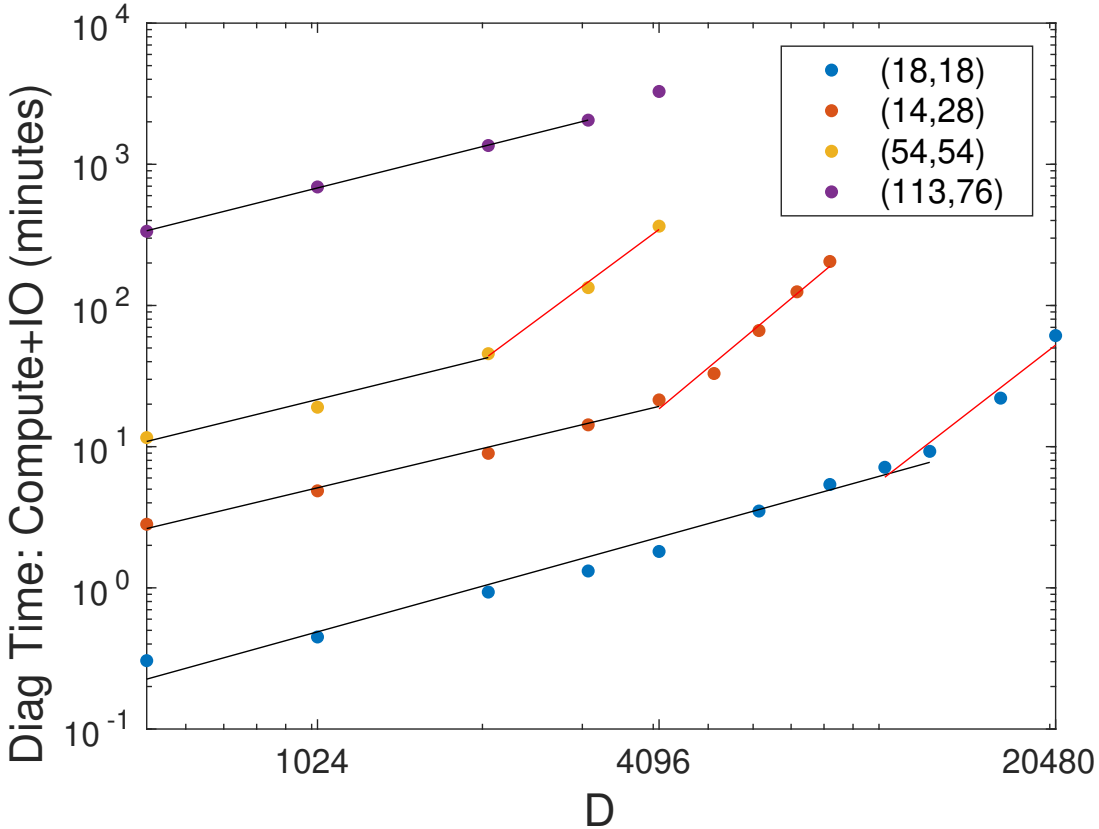


A. Menczer, Ö.L (2023), FeMoco(113,76)

- New mathematical model for parallelization $\rightarrow$ felxibe scaling
- $D_{SU(2)} = 24576 \rightarrow$ FCI solution
- We reached 110 TFLOPS $>$ 76 TFLOPS of the FP64 limit of NVIDIA

$\rightarrow$ utilization of highly specialized tensor core units (TCU)

Boosting the effective performance via non-Abelian symmetries



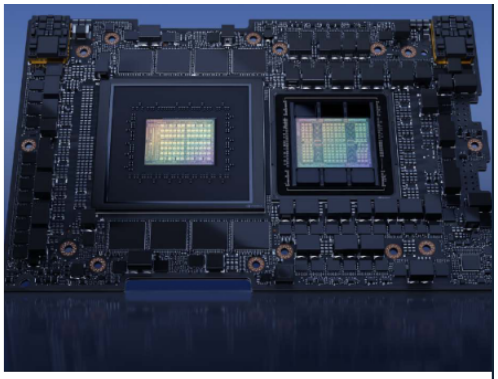
- New model to utilize NVIDIA D2D links. A. Menczer ÖL (unpublished 2023)

- 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

NVIDIA GH200 Grace Hopper Superchip

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



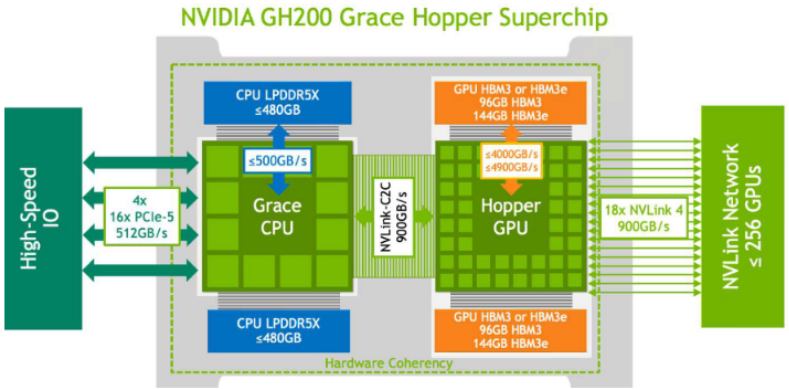
The World's Most Versatile Computing Platform

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



Conclusion

- Underlying tensor and matrix algebra can be organized into several million of independent operations (tasks)
- Tensor networks methods are ideal for parallelization $\rightarrow$ felxible scaling
- Application of two-qbit gates and general network structure $\rightarrow$ simulation of quantum computing
- AI and deep learning can be formulated via tensor network methods
- Capturing strong correltaions $\rightarrow$ a universal simulator for material properties, chemical reactions, quantum information theory etc
- multiNode-multiGPU $\rightarrow$ exascale computation
- Would be open for and be happy about collaborations

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)