Future computing:
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

GPU Day 2026 (AOP2026)
Budapest 28.05.2026

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

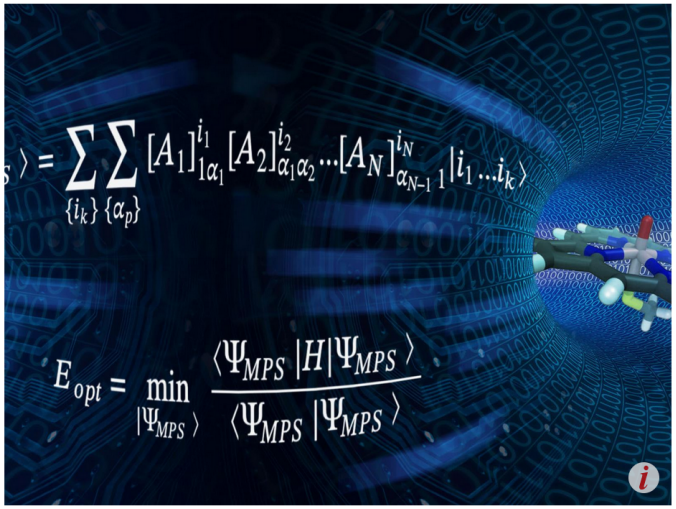
Wigner, PNNL, TUM, NVIDIA, SandboxAQ joint press release

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:

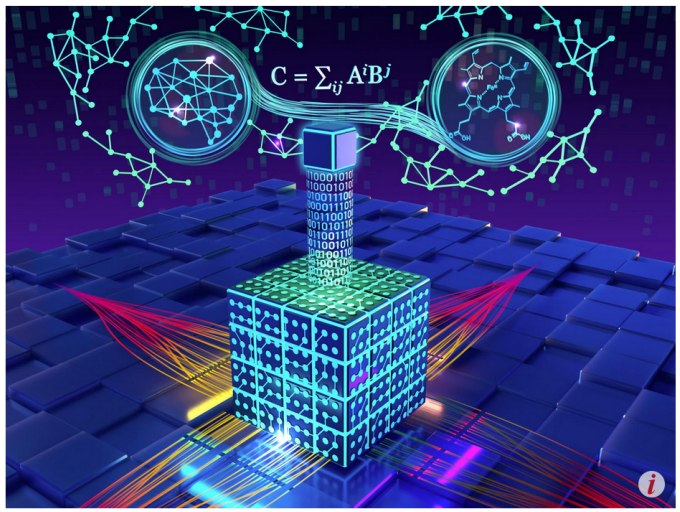
A recent collaboration among researchers from HUN-REN Wigner Research Centre for Physics in Hungary and the Department of Energy's Pacific Northwest National Laboratory, along with industry collaborators SandboxAQ and NVIDIA, has achieved unprecedented speed and performance in efforts to

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

APRIL 30, 2026 | News Release

PNNL Researchers to Speak About Genesis Mission at National Science Bowl

Related Content

Market aspects: technology transfer from academy to industry

RED OCEANS are all the industries in existence today – the known market space. In red oceans, industry boundaries are defined and accepted, and the competitive rules of the game are known.

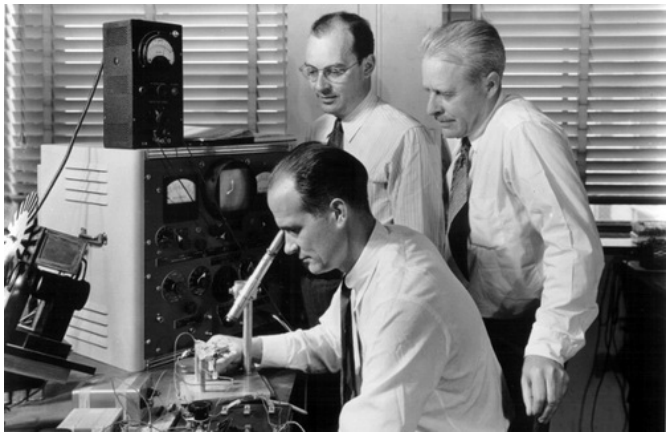
Here, companies try to outperform their rivals to grab a greater share of existing demand. As the market space gets crowded, profits and growth are reduced. Products become commodities, leading to cutthroat or 'bloody' competition. Hence the term red oceans.

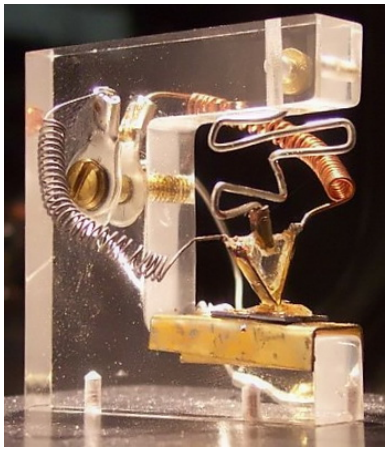
BLUE OCEANS, in contrast, denote all the industries not in existence today – the unknown market space, untainted by competition. In blue oceans, demand is created rather than fought over. There is ample opportunity for growth that is both profitable and rapid.

In blue oceans, competition is irrelevant because the rules of the game are waiting to be set. A blue ocean is an analogy to describe the wider, deeper potential to be found in unexplored market space. A blue ocean is vast, deep, and powerful in terms of profitable growth.

Quantum physics in everyday life?

• The first transistor was demonstrated in 1947, by three Bell Labs researchers: Walter Brattain, John Bardeen, and William Shockley.





• This sentence could have been uttered then: Transistor in everyday life!

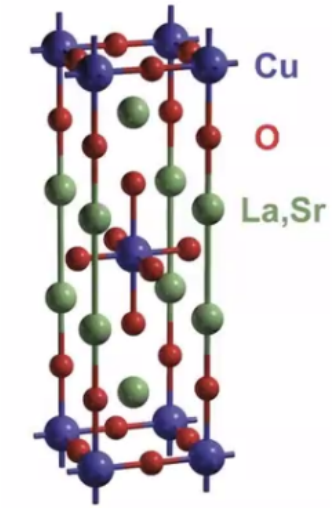
• Maybe everyone would have thought them crazy. And today?

• Our lives today are almost unimaginable without computing technology (Smartphone vs Apollo-13).

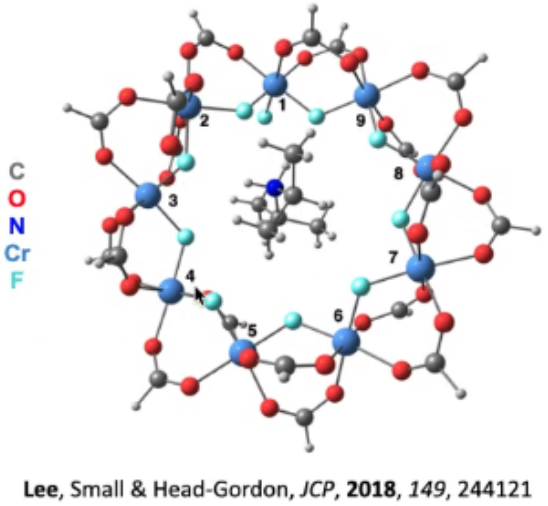
Simultaneously a shift from Blue Ocean to Red Ocean Strategy.

• Devices based on quantum effects could provide a new solution: quantum computers.

Strong correlations between electrons used by nature and in new technologies



High-T_c superconductors

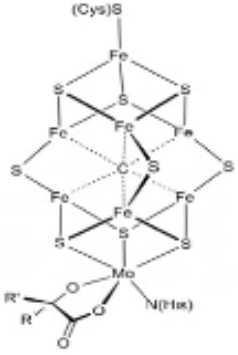


Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

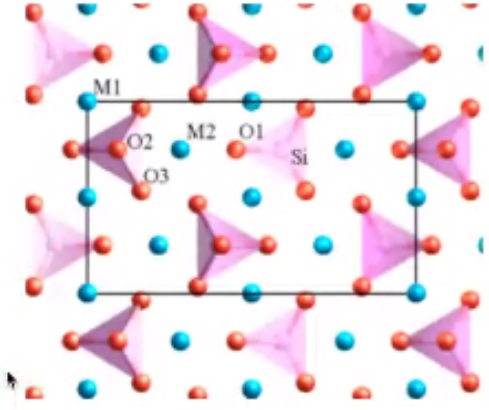
Single molecular magnets (SMM)

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

"FeMoco"



Nitrogen fixation



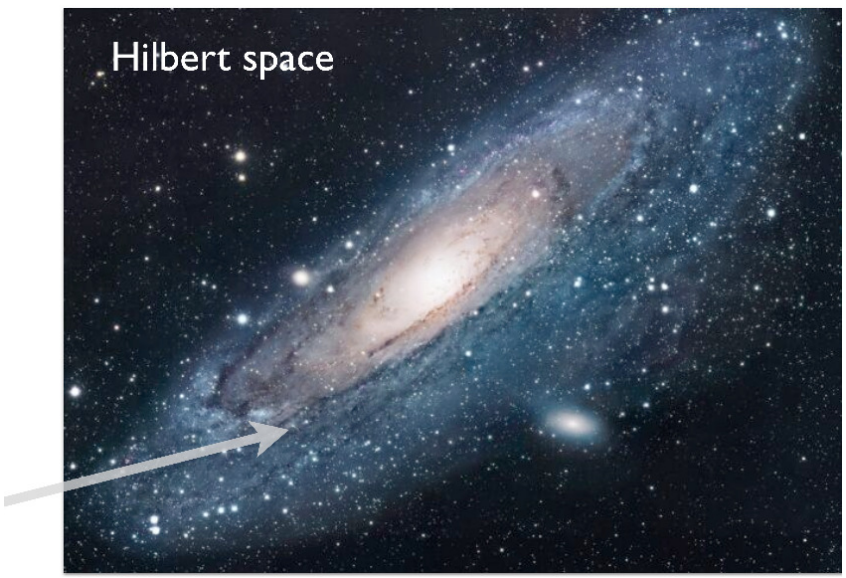
Battery technology

Classic Simulation: Quantum CAD !?

Motivation and goal: In the engineering world, expensive models are not built every time, but rather with computer programs.

R. P. Feynman (1985): Simulation of quantum systems on classical computers scales **exponentially** with the size of the system, while on a quantum computer it would scale polynomially.

The difficulty of the task is like finding a star in the universe.



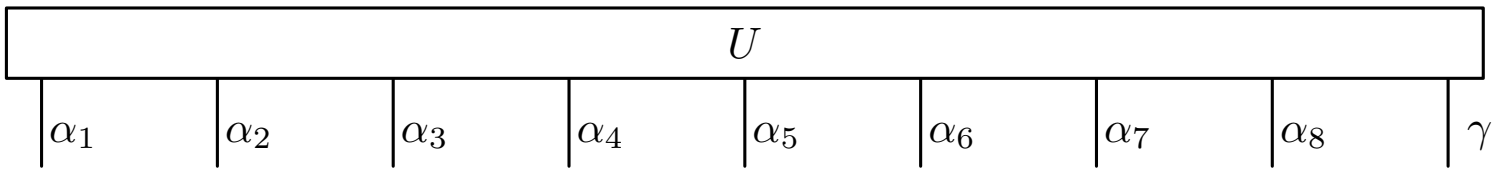
Intermediate solution: development of appropriate mathematical algorithms $\rightarrow$ **tensor factorization: polynomial scaling**.

Tensor product approximation

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

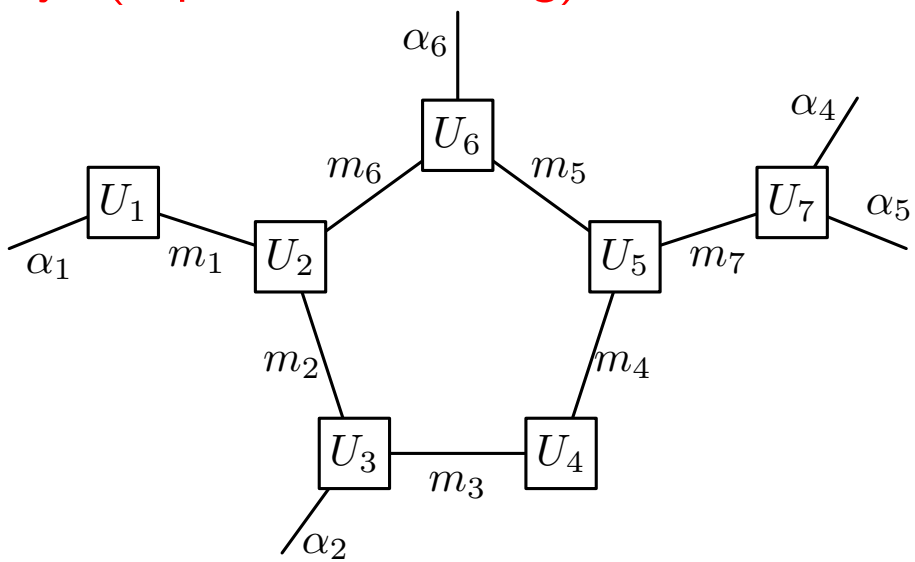
$$|\Psi_\gamma\rangle = \sum_{\alpha_1=1}^{q_1} \dots \sum_{\alpha_d=1}^{q_d} U(\alpha_1, \dots, \alpha_d, \gamma)\, |\alpha_1\rangle \otimes \cdots \otimes |\alpha_d\rangle \in \bigotimes_{i=1}^{d} \Lambda_i := \bigotimes_{i=1}^{d} \mathbf{C}^{q_i} \ ,$$

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)

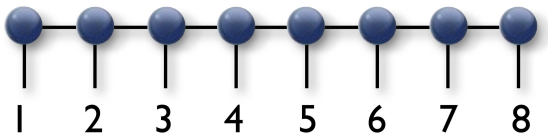
- $\alpha_i \in \{\uparrow, \downarrow\}$ or $\alpha_i \in \{0, \uparrow, \downarrow\uparrow\downarrow\}$
- We seek to reduce computational costs by parametrizing the tensors in some data-sparse representation using SVD.
- A general tensor network representation of a tensor of order 5.



tensor network state (TNS) methods

1D MPS

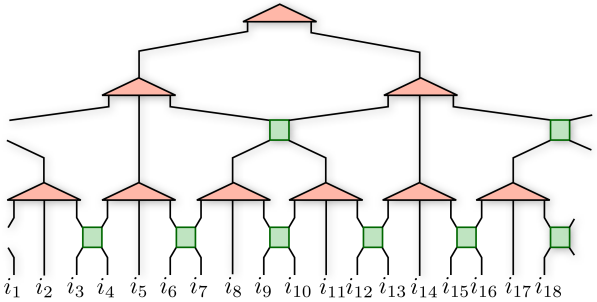
Matrix-product state



White, Östlund, Rommer

1D MERA

Multi-scale entanglement renormalization ansatz

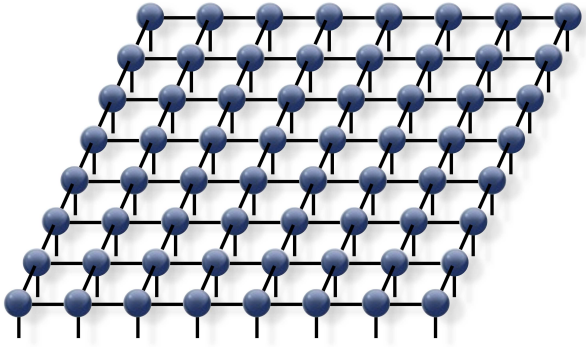


1D TTNS

Tree tensor network state

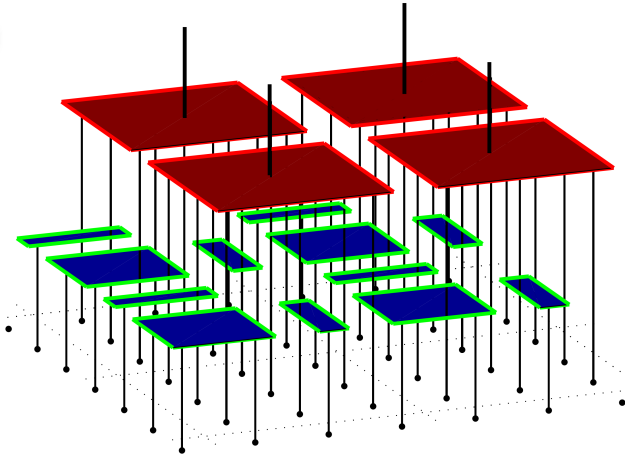
Vidal, Corboz

2D PEPS



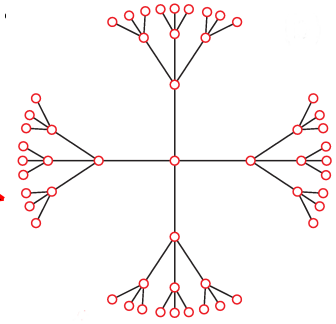
Verstraete, Cirac Jordan, Orus, Vidal

2D Mera



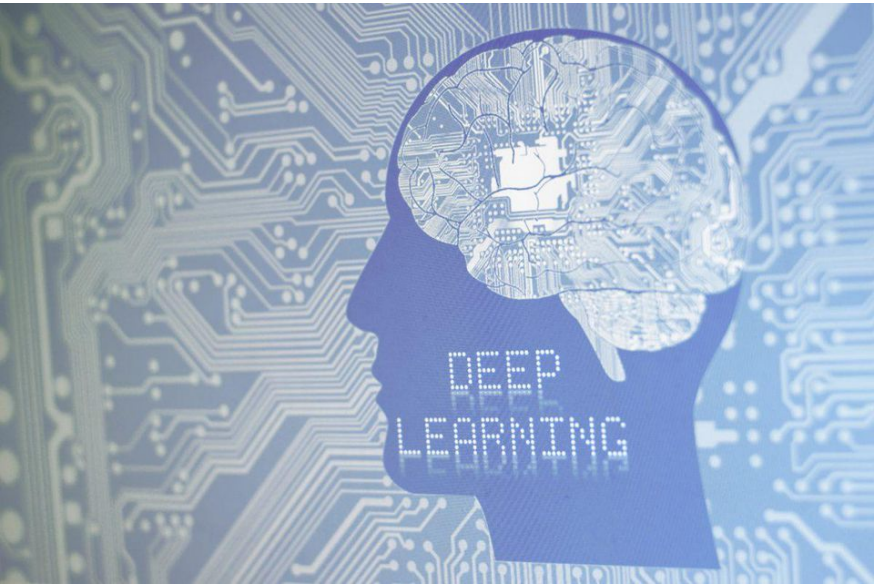
Vidal, Evenbly

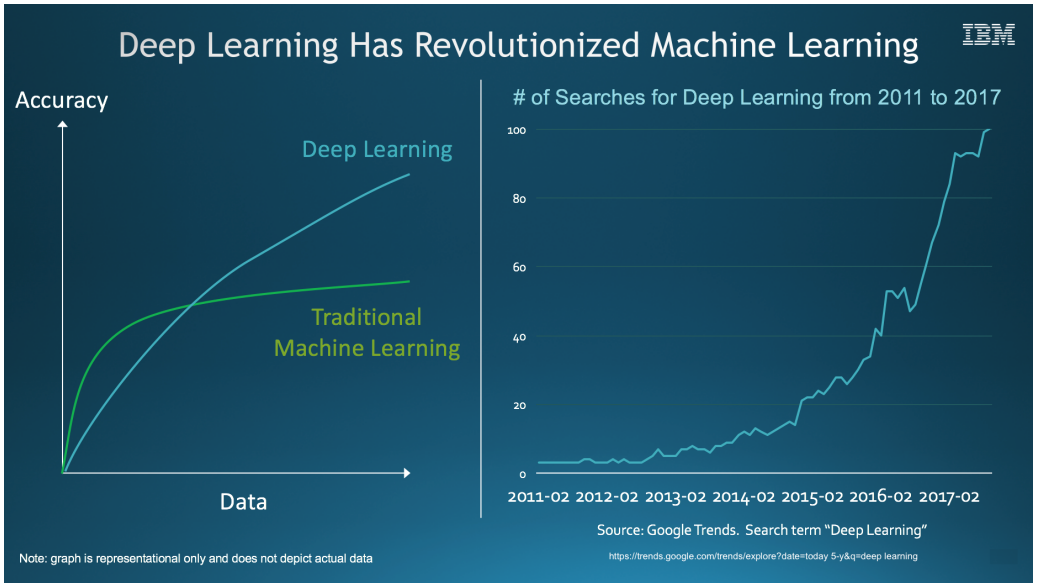
2D Tree-

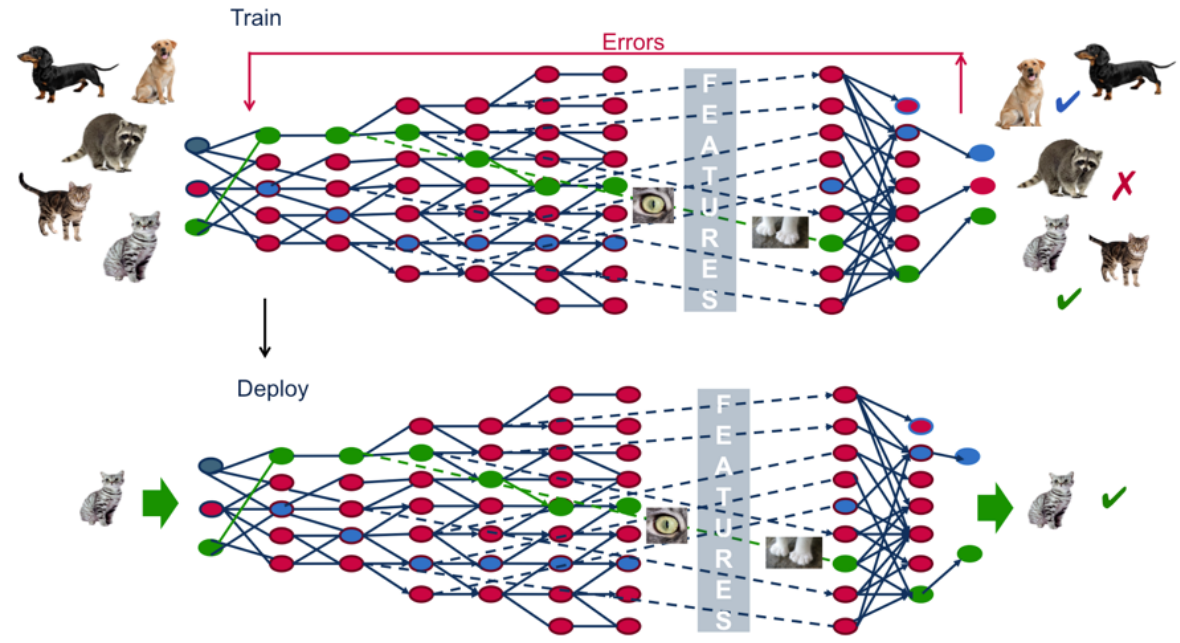


Vidal, Corboz, Verstraete, Murg, Legeza, Noack

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!

High performance computing (HPC)

- FLOPS (Floating-point Operations Per Second): a measure of computer performance, calculating how many floating-point calculations (math with decimal numbers) a processor can perform per second.
- Imagine flops: how many times you can blink per second.

DDMRG simulations: ACM Gordon Bell Award in 2012

- Number of processors: 88,128
- Total memory capacity: over 1 petabyte
- Computing capacity: 10.51 petaflops (10^{15})

2024

- Computing capacity: 1 exaflops (10^{18})

The new technology is around us !!!

NVIDIA › Main Menu

Shop Drivers Support

Cloud & Data Center Solutions Products Data Center GPUs Software Technologies Resources Get Started

NVIDIA GB200 NVL72

Powering the new era of computing.

Read Datasheet

Introduction Highlights Features Specs

Unlocking Real-Time Trillion-Parameter Models

GB200 NVL72 connects 36 Grace CPUs and 72 Blackwell GPUs in a rack-scale, liquid-cooled design. It boasts a 72-GPU NVLink domain that acts as a single, massive GPU and delivers 30X faster real-time trillion-parameter large language model (LLM) inference.

The GB200 Grace Blackwell Superchip is a key component of the NVIDIA GB200 NVL72, connecting two high-performance NVIDIA Blackwell Tensor Core GPUs and an NVIDIA Grace™ CPU using the NVIDIA NVLink™-C2C interconnect to the two Blackwell GPUs.

The Blackwell Rack-Scale Architecture for Real-Time Trillion-Parameter Inference and Training

The NVIDIA GB200 NVL72 is an exascale computer in a single rack. With 36 GB200s interconnected by the largest NVIDIA® NVLink® domain ever offered, NVLink Switch System provides 130 terabytes per second (TB/s) of low-latency GPU communications for AI and high-performance computing (HPC) workloads.

Tech Blog ›

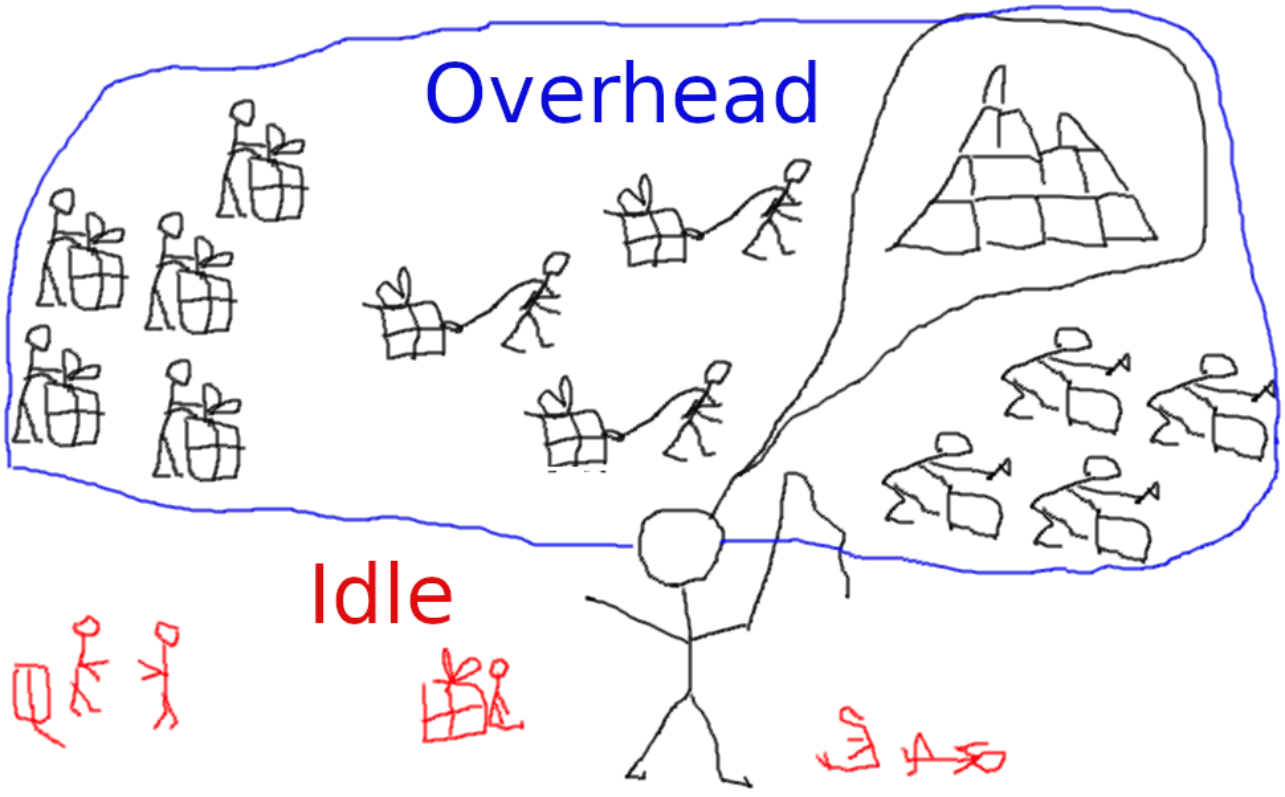
Centralized scheduling: non-ideal society

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

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

Centralized scheduling: Huge overhead, units can be idle

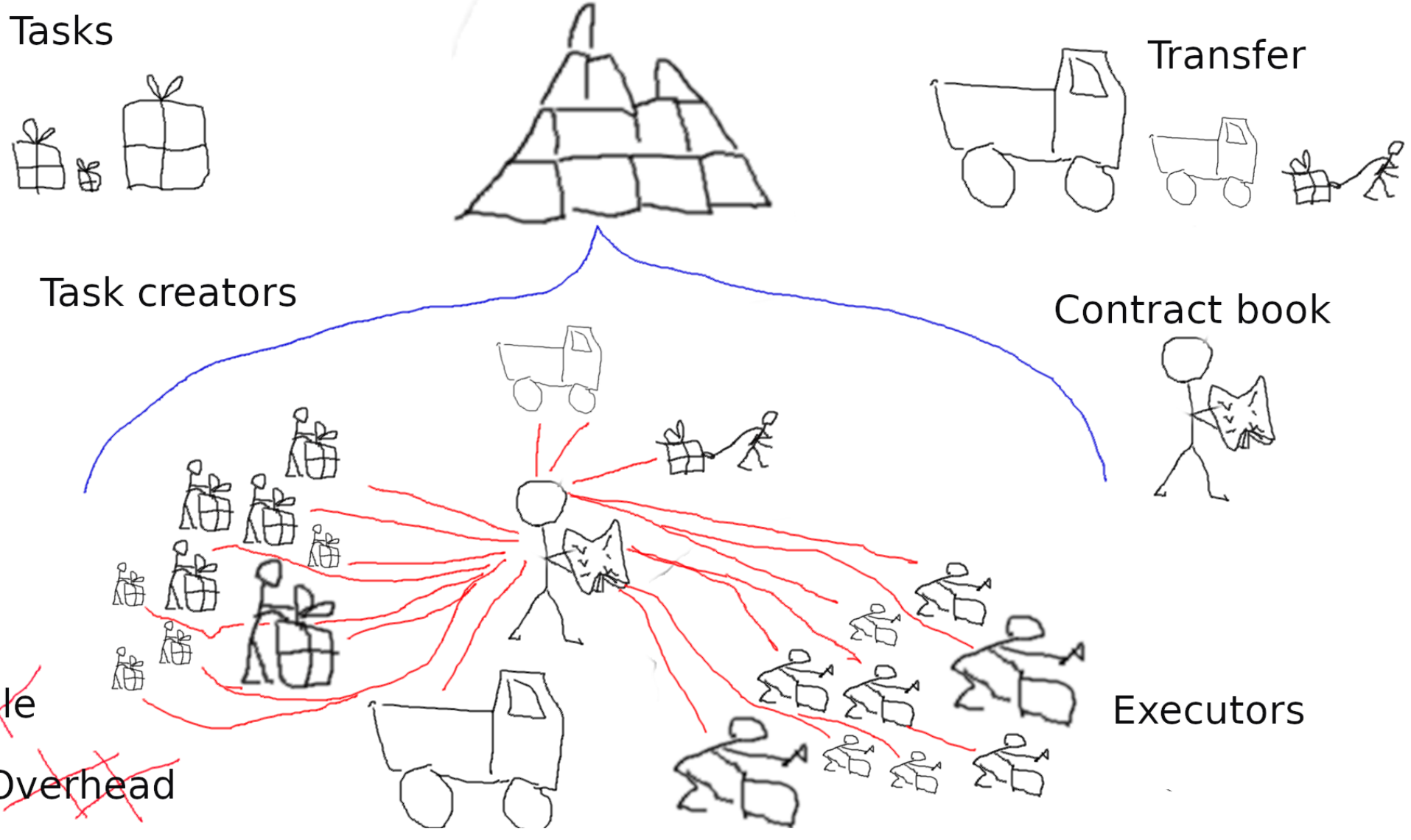
- Central scheduler performs lot of measurements, estimations, communication to rearrange tasks and workers → huge overhead



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

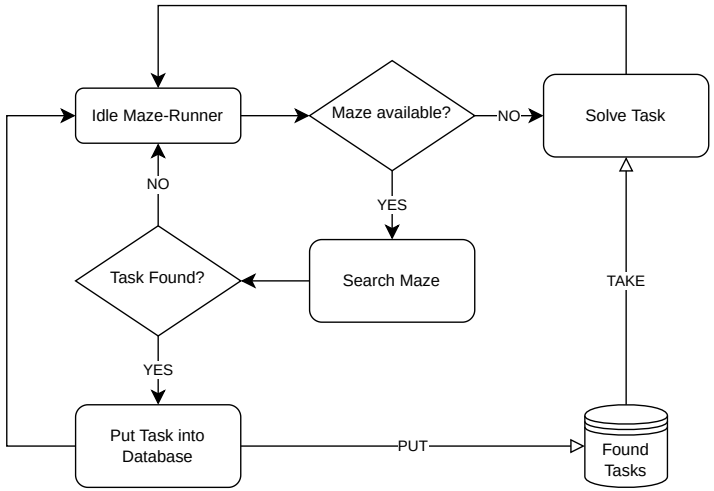
Self motivated workers → ideal "team-like" society

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

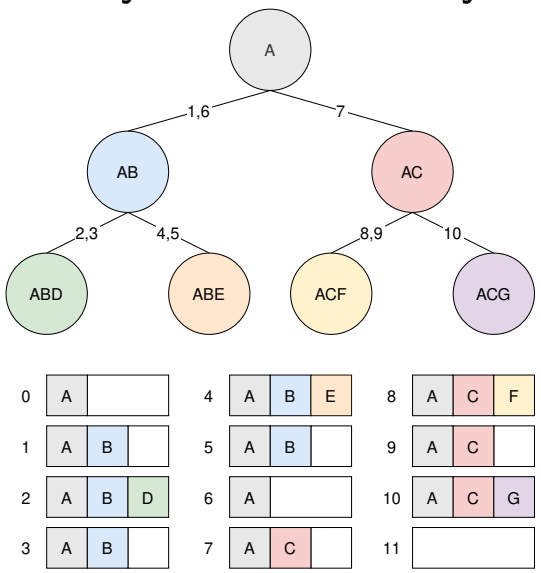


Novel algorithmic solutions & parallelization A.Menczer,ÖL(2023)

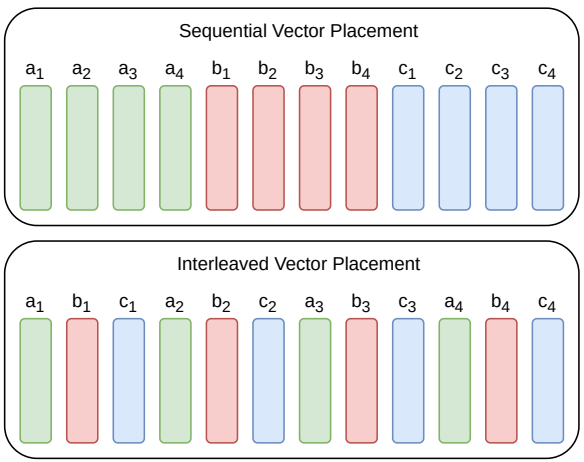
Life Cycle of a Maze-Runner Thread.



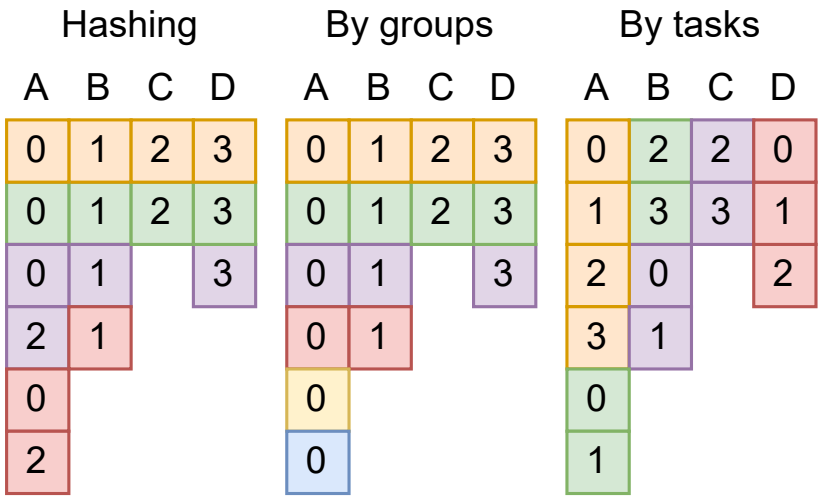
Graph theory based memory management



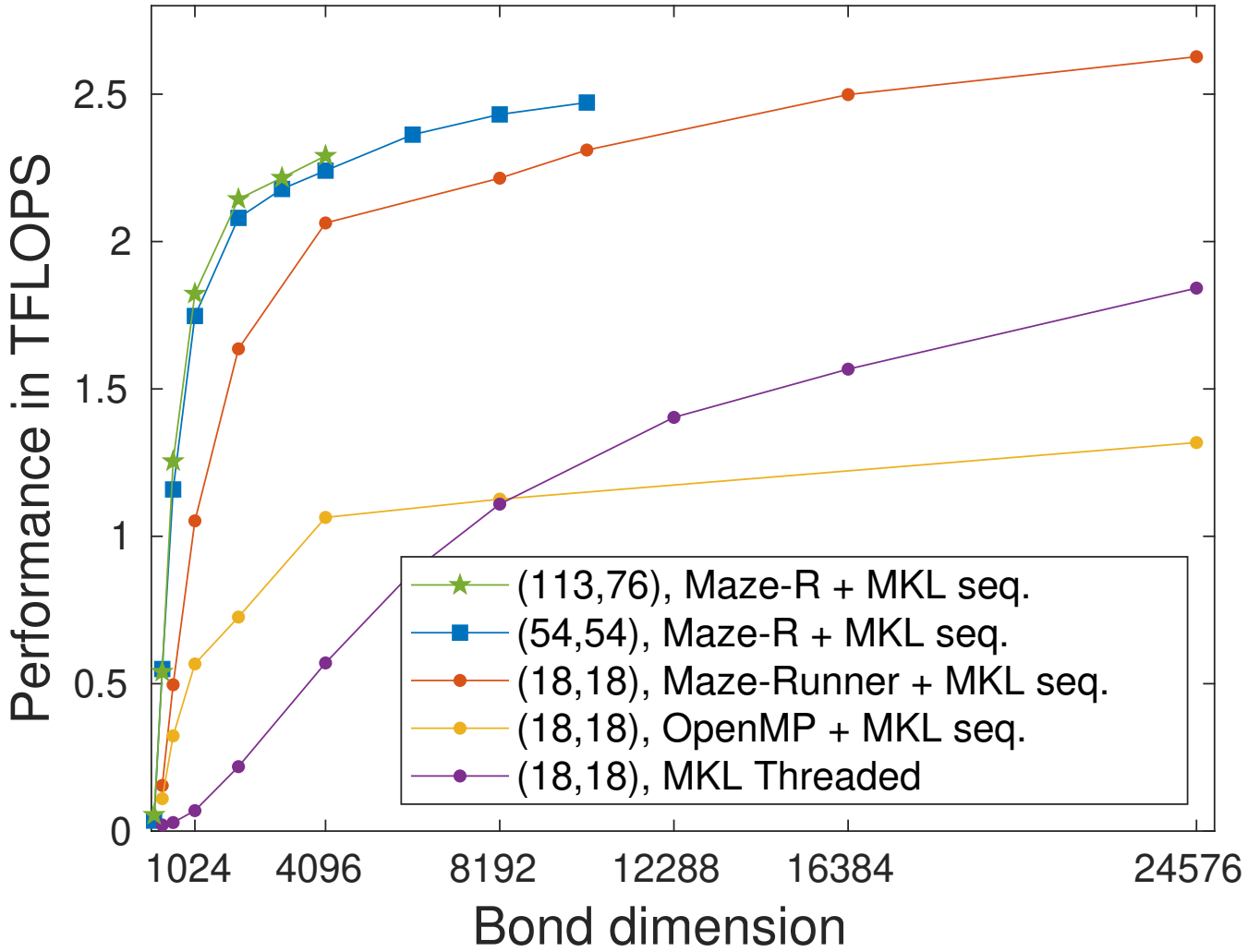
Strided Batched operations via data localization



Execution via hierarchy of tasks



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



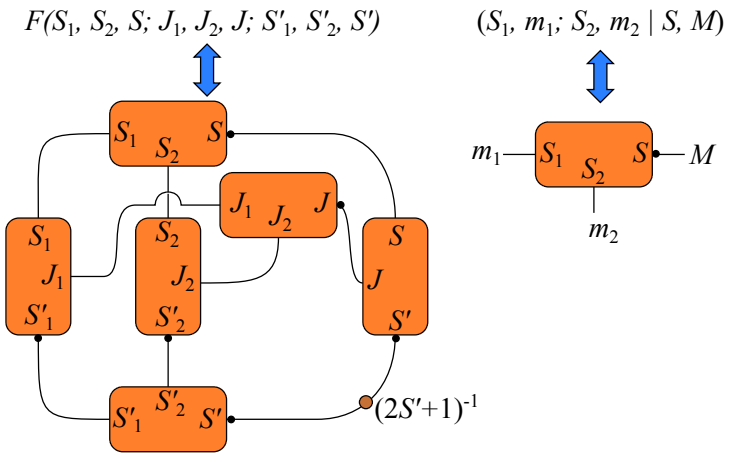
Performance measured in TFLOPS for the F_2 and FeMoco chemical systems for CAS(18,18) and CAS(54,54) orbitals spaces, respectively, as a function of the DMRG bond dimension on a dual Intel(R) Xeon(R) Gold 5318Y CPU system with 2×24 physical cores running at 2.10 Ghz.

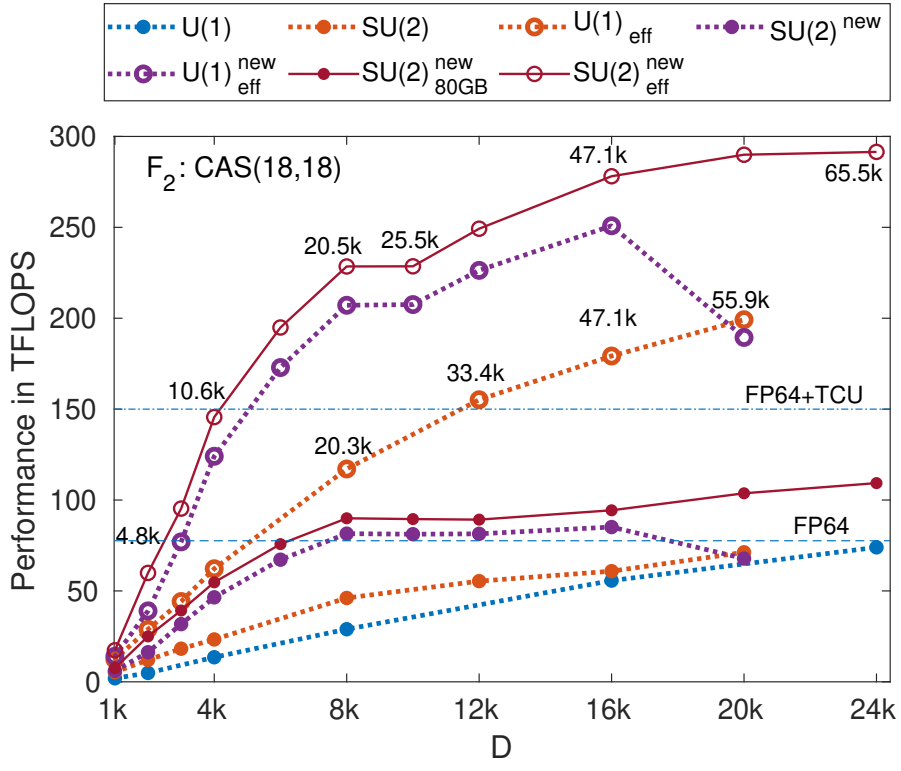
Boosting the effective performance via non-Abelian symmetries

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

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

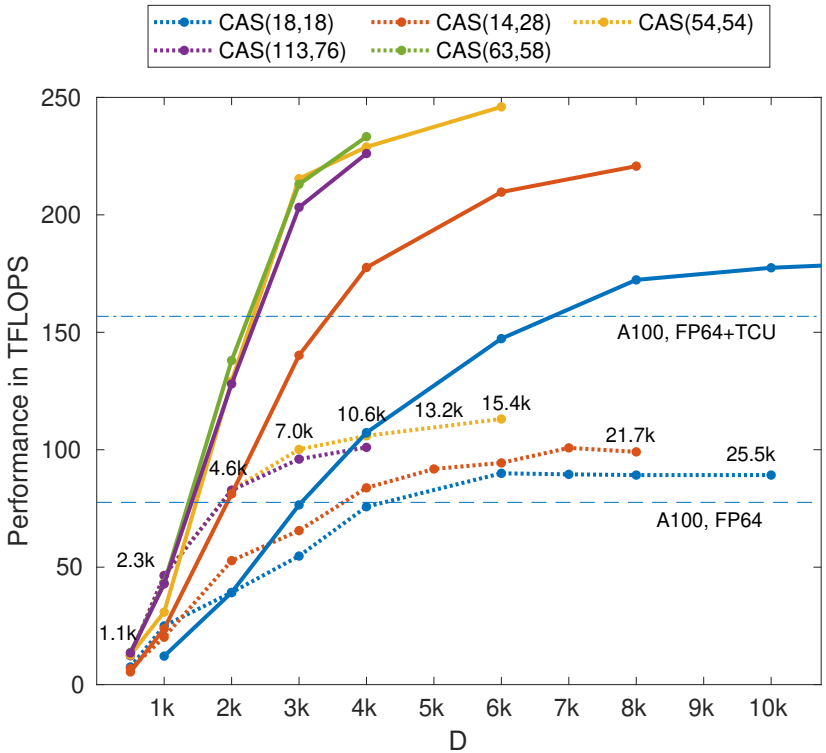
where F equals the Wigner-9j symbol up to rescaling,

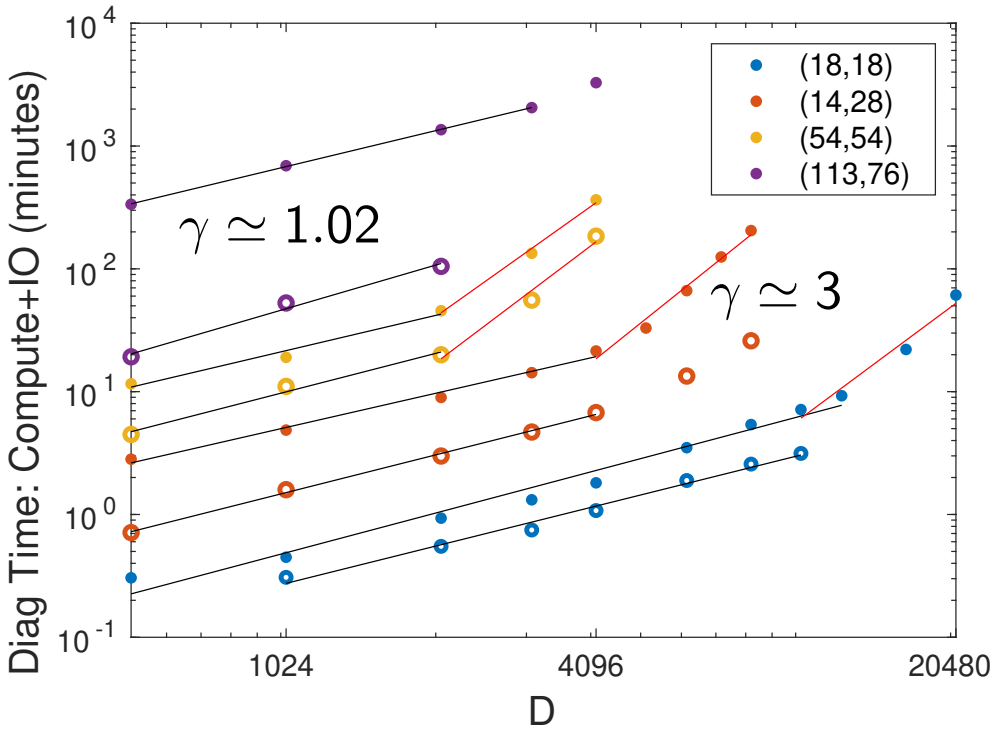




- New mathematical model for parallelization → flexible scaling
- $D_{SU(2)} = 24576 \rightarrow D_{U(1)} = 2^{16} \rightarrow$ FCI solution

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)

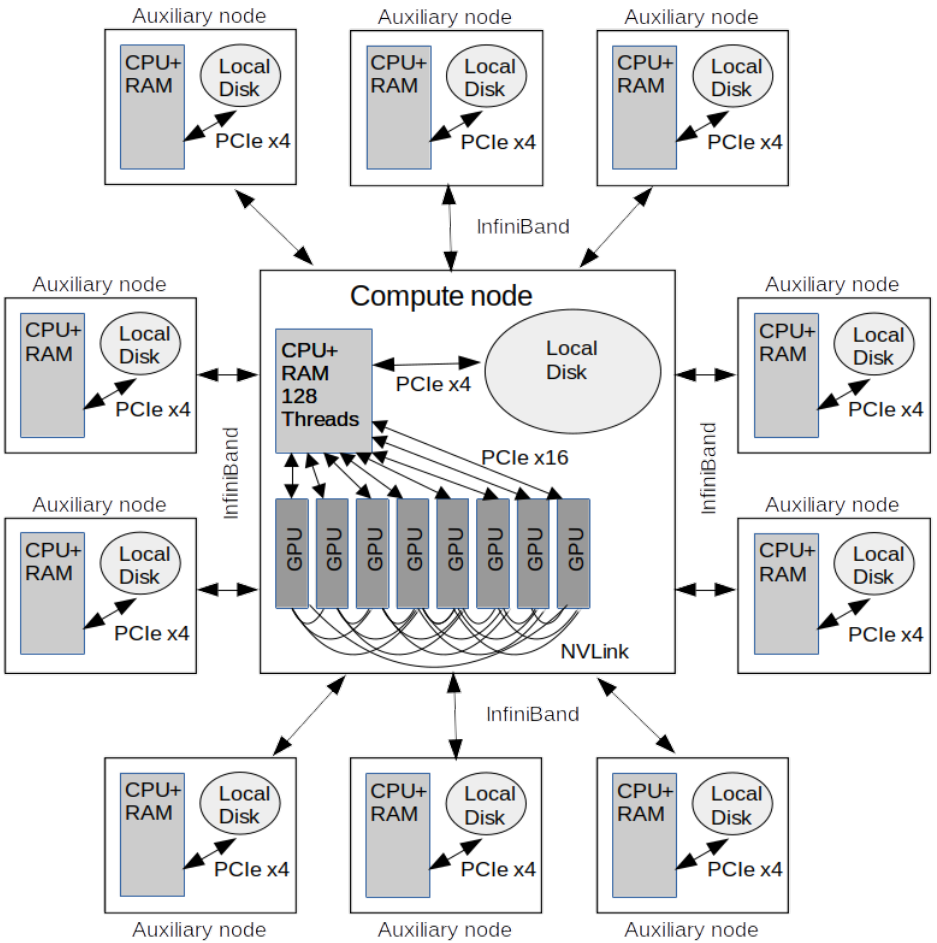
- Combination of our MPI and GPU kernels: full replacement of *boost library*, asynchronous IO, multiNode-multiGPU

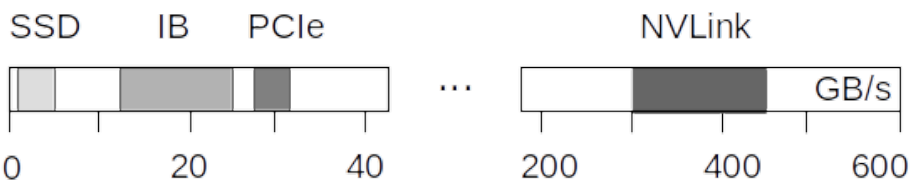
$\to$ petascale computing. A. Menczer ÖL (unpublished 2023-2024)

Power consumption of the TNS calculations $\rightarrow$ Green DMRG

- The power consumption of the TNS calculations are becoming one of the most important questions due to high energy demands and costs.
- The thermal design power (TDP) for 2 $\times$ Intel(R) Xeon Gold 5318Y CPU is 2 $\times$ 165 Watts $\rightarrow$ 2.5 TFLOPS would lead to $\approx$ 7.5 GFLOPS/Watt.
- For an NVIDIA A100-PCIE-40GB device the TDP is 250 Watts.
- For our 8 card accelerated hybrid algorithm with 70 TFLOPS performance results in $\approx$ 30.04 GFLOPS/Watt.
- For a given calculation the cost of the energy demand arising from the processors can be reduced to one quarter of the original consumption.
- The energy consumption of the GPU devices fluctuates significantly, thus even a better ratio can be obtained.

Cost optimized "green" TNS Menczer, ÖL 2024





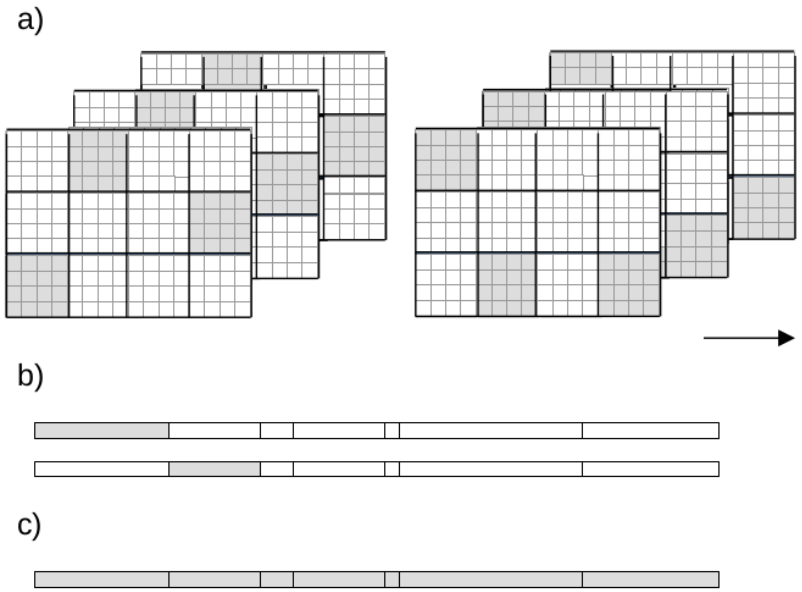
- DGX-H100 costs 100 USD/hour on Google Cloud

- 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

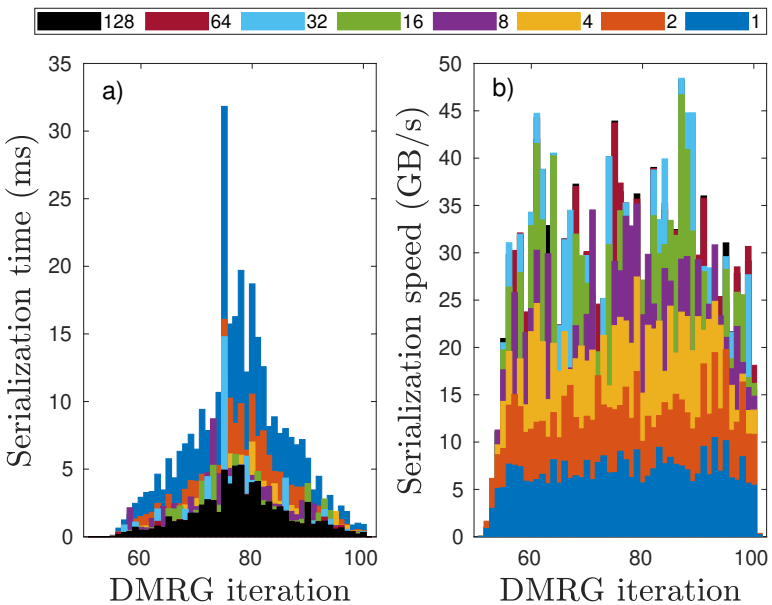
Precontracted network asynchronous MPI-IO Menczer, ÖL 2024

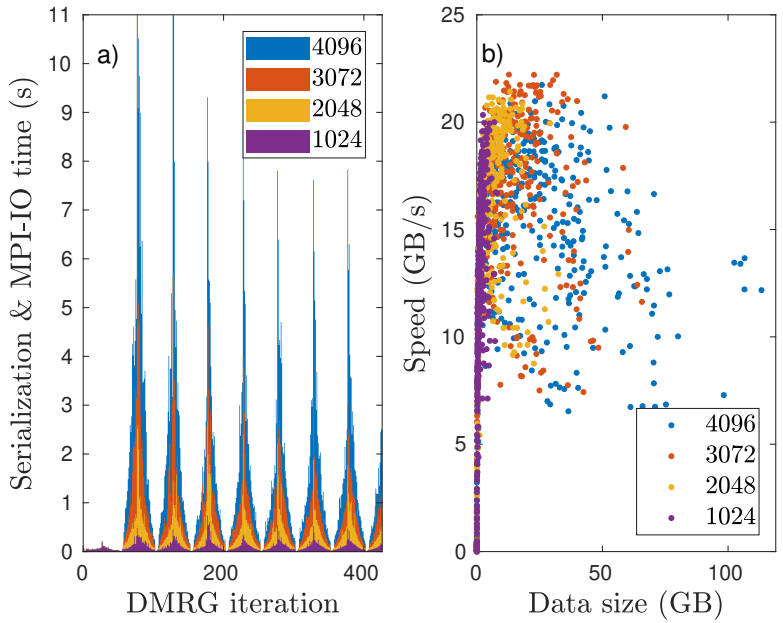
- Replacement of boost library with our in-house build MPI module



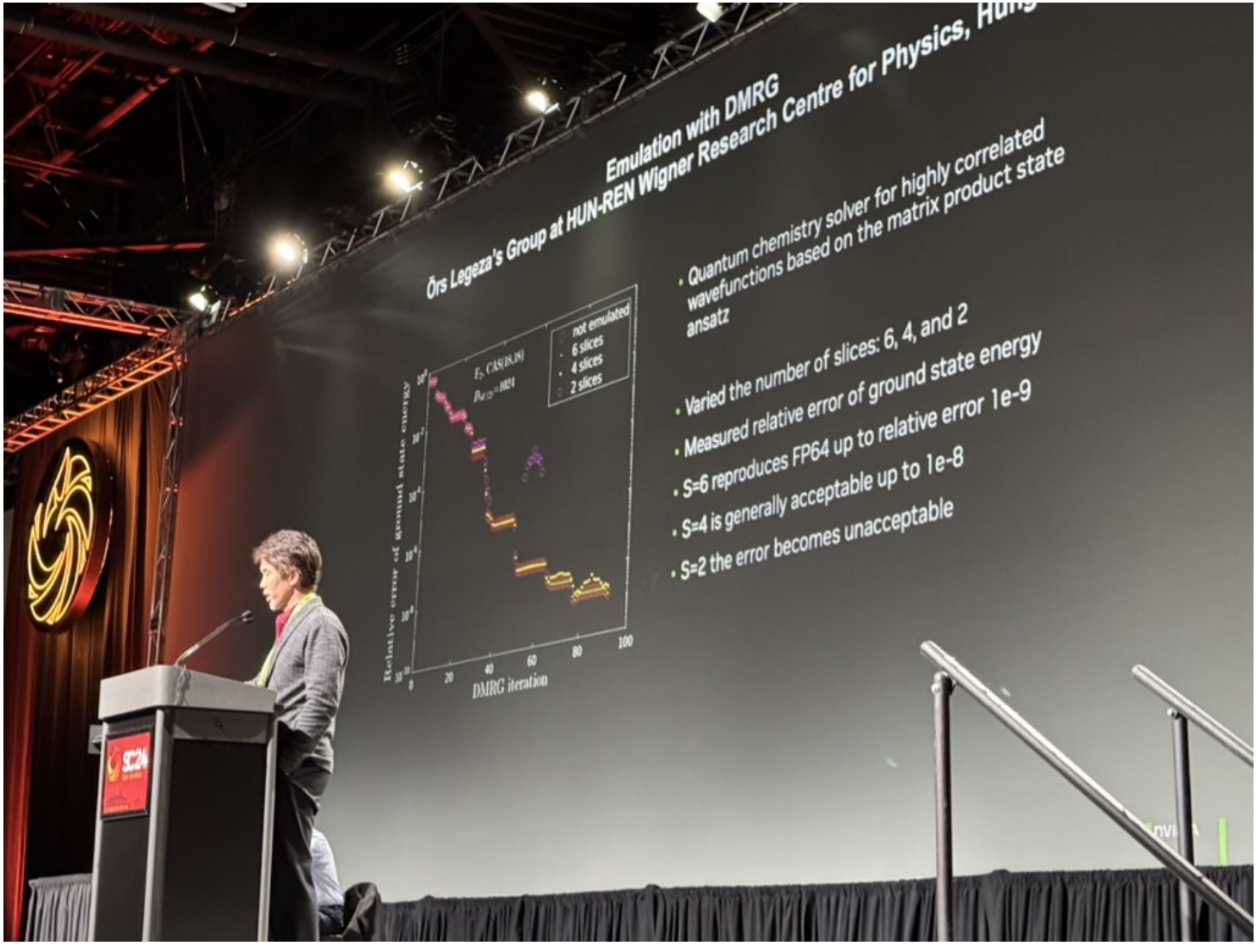
a) Quantum number based block sparse matrices and tensors.
b) Skeleton of serialized data segments via disk IO or MPI based communication.
c) Skeleton of serialized data segments filled completely with data when asynchronous save IO.

- 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.

$$A \approx D_1^{-1} D_1 A_1 + D_2^{-1} D_2 A_2 + \ldots D_k^{-1} D_k A_k$$

$$B \approx B_1 E_1 E_1^{-1} + B_2 E_2 E_2^{-1} + \ldots B_k E_k E_k^{-1}$$

where all elements in $D_i A_i$ and $B_i E_i$ for all $1 \leq i \leq k$ can be represented in INT8.

$(D_i A_i)(B_j E_j)$ can be computed without rounding error using INT8TC

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

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

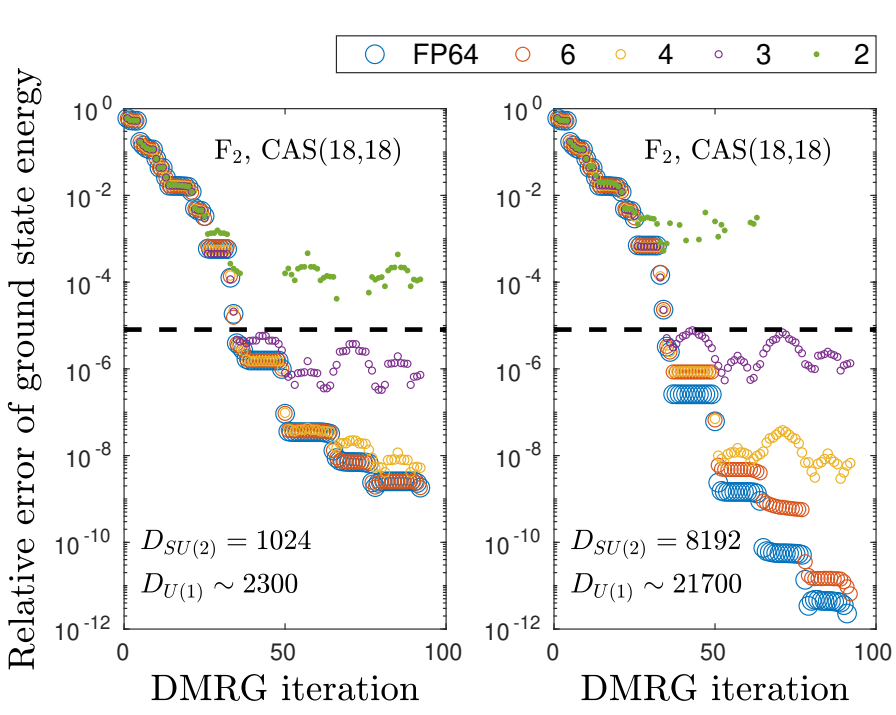
Step-2: $k(k+1)/2$ matrix products (INT8TC) : $k(k+1)pqr + \mathcal{O}(pr)$

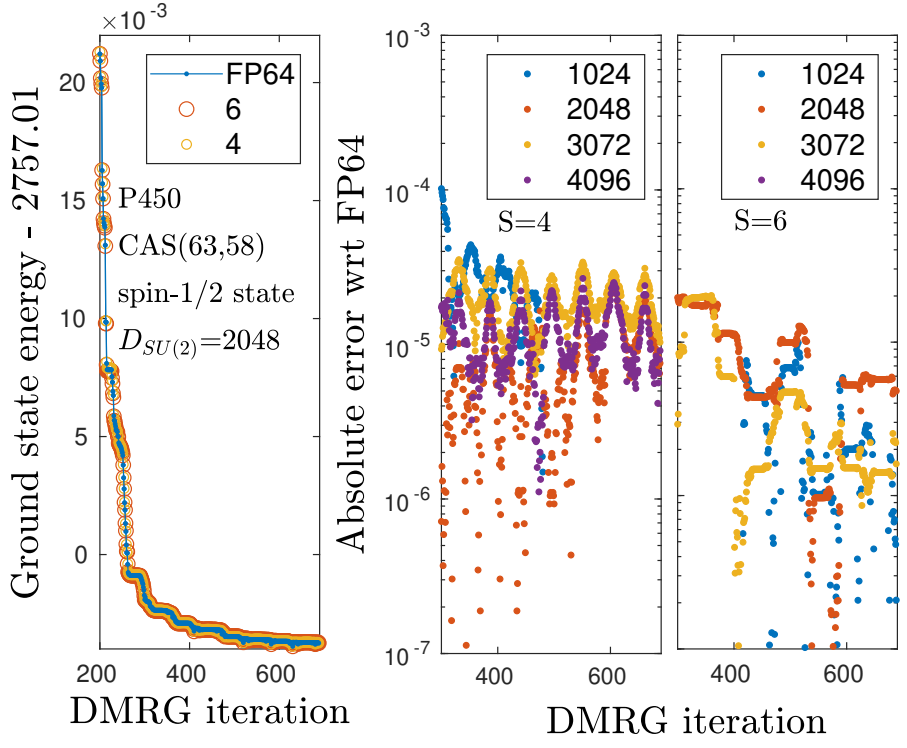
Step-3: reduction of matrix products (FP64) : $\mathcal{O}(pr)$

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.



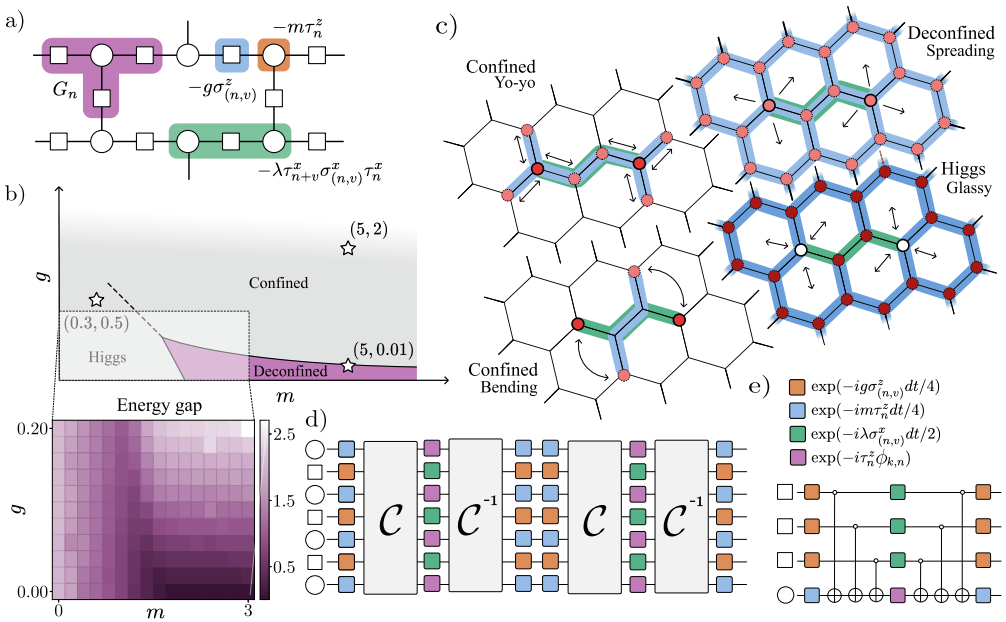


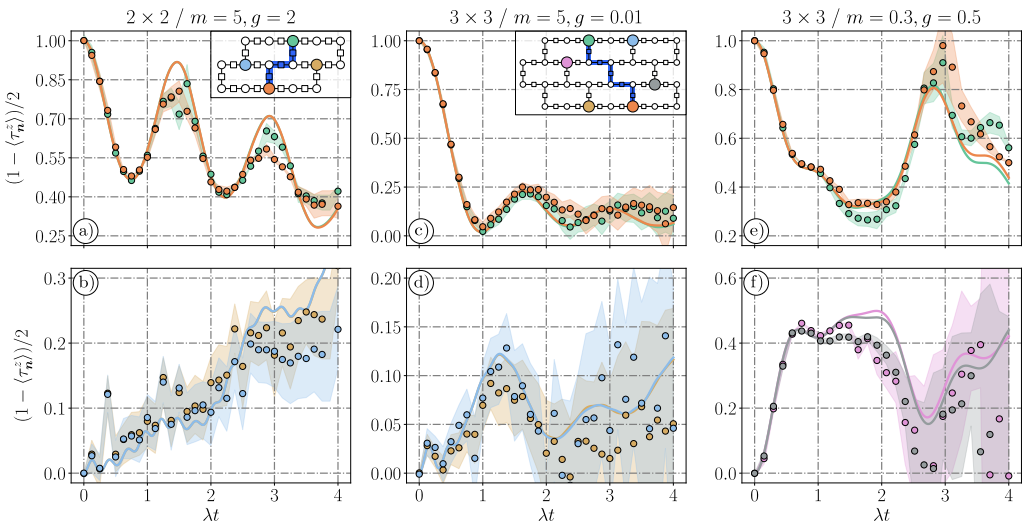
- Chemical accuracy, 1.6mHa, can be reached with 4, 6 slices

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

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.

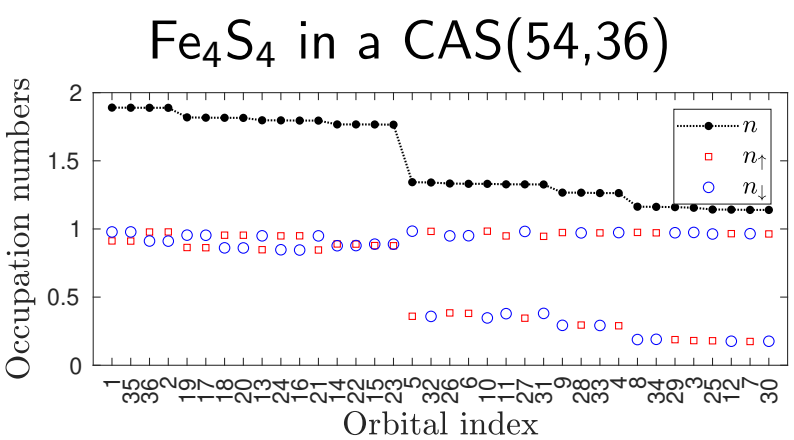
- Achieved via our mixed-precision spin-adapted *ab initio* Density Matrix Renormalization Group (DMRG) interfaced with the ORCA program on single node NVIDIA Blackwell GPU ($\sim$ 220 TFLOPS FP64).

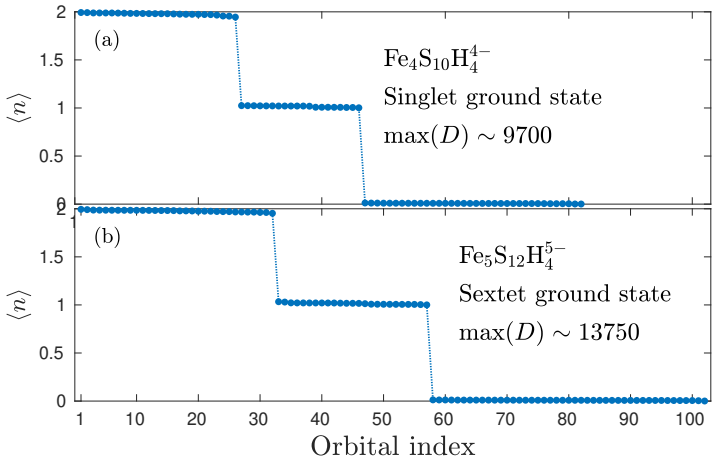
- Future: NVIDIA GB300 NVL72 up 576 GPUs and at JSC/RPI/Empire AI/Argonne NL $\sim 72/288/10,000/100,000$ GB300 GPUs.

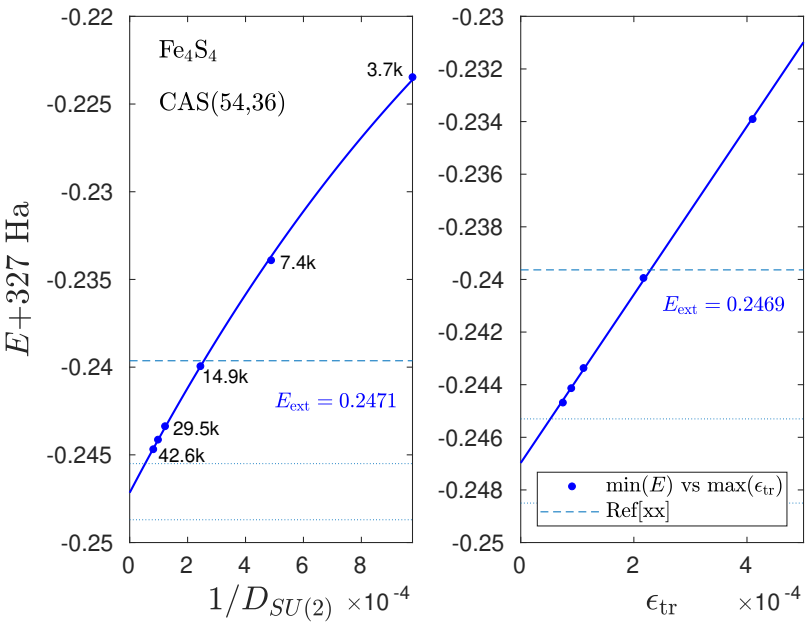
Classical vs quantum simulation? (DMRG-SCF on DGXB200)

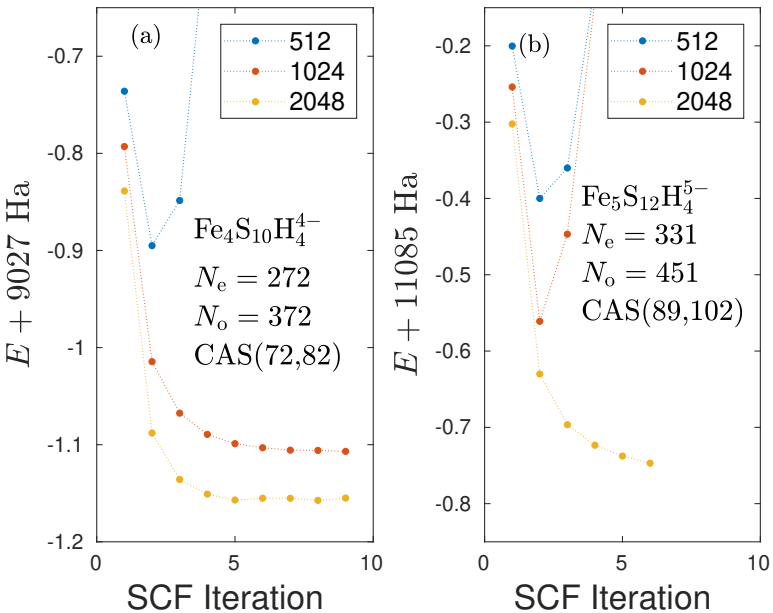
ÖL, Menczer,Werner,Xantheas,Neese,Ganahl,Brower,Bernabeu,Hammond,Gunnels (2026)

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









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 $\rightarrow$ exascale computation
- Mixed precision TNS on specialized new hardware with lower energy consumption
- $\rightarrow$ Simulation of realistic material properties
- $\rightarrow$ Estimating quantum circuit depth
- $\rightarrow$ Quantum inspired AI
- $\rightarrow$ Hybrid classical quantum algorithms
- $\rightarrow$ RPI/Empire AI/Argonne: 288/10 000/100 000 GB300 GPUs
- $\rightarrow$ **TNS + HPC + Blue Ocean Strategy $\rightarrow$ 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)