Towards exascale computation of quantum systems
Industrial perspectives

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 Stiftung, Pöcking, Germany

GPU-DAY

Budapest, 05.30.2024

in collaboration with

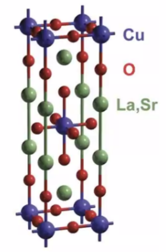
- more than 30 research groups worldwide from condensed matter physics, quantum chemistry, nuclear physics, quantum information theory, applied mathematics and computer science
- High-Performance Computing Center Stuttgart, Germany
- Pacific Northwest National Laboratory (PNNL), USA
- National Energy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

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

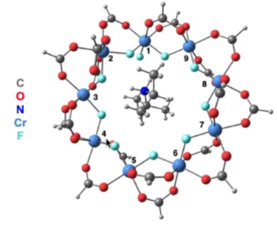
Recently there is also an interest by industrial partners.

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

Strong correlations between electrons used by nature and in new technologies



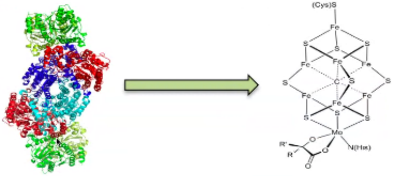
High T_c superconductors



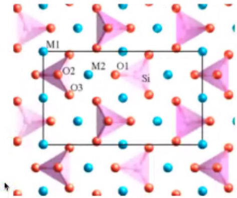
Lee, Small & Head-Gordon, *JCP*, **2018**, *149*, 244121

Single molecular magnets (SMM)

$N_2 + 3\,H_2 \rightarrow 2\,NH_3$ “FeMoco”



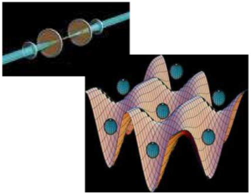
Nitrogen fixation



Battery technology

Experimental realizations: optical lattices

Numerical simulations: model systems



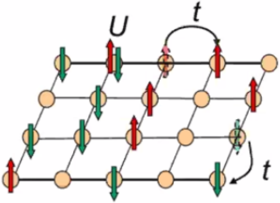
Atoms (represented as blue spheres) pictured in a 2D-optical lattice potential

Potential depth of the optical lattice can be tuned.

Periodicity of the optical lattice can be tuned.

Hubbard model: lattice model of interacting electron system

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

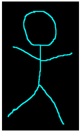


t hopping amplitude
U on-site Coulomb interaction
$\sigma \in \uparrow, \downarrow$ spin index

Classical or quantum computers?

Discrete basis, configuration space, superposition

Possible states of a person

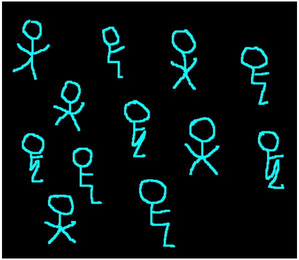

stands

sits

squats

Dimension of the local space $d = 3$

N persons in a room



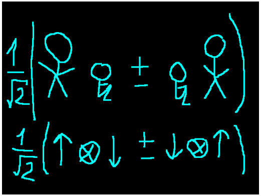
Dimension of the configuration space: 3^N, i.e., it scales exponentially

In quantum physics superposition is possible:

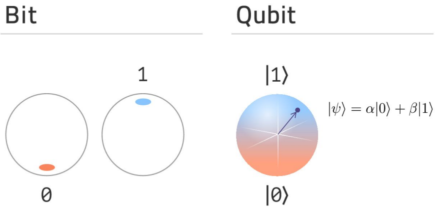
Ex. d=2 (two states allowed)

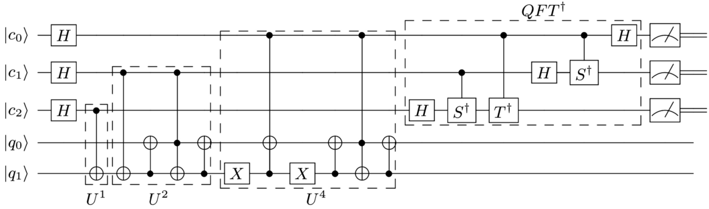
- Two persons (at position A and B).
- Four possible configurations.
- At position "A" person stands or squats with 50% probability.
- At position "B" person stands or squats with 50% probability.

Entangled state → quantum information (q-dits)


$\frac{1}{\sqrt{2}}\Big($ ± $\Big)$
$\frac{1}{\sqrt{2}}(\uparrow\otimes\downarrow \pm \downarrow\otimes\uparrow)$

Entanglement: quantum data processing





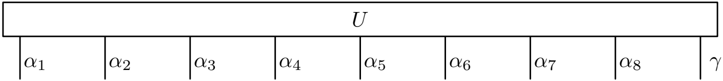
- Quantum computing: quantum supremacy or quantum advantage
- Quantum cryptography: secure communication
- Experimental realizations: quantum sensors (biomedical applications), unprecedented spatial resolution and sensitivity on atomic length scale

Tensor product approximation

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

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

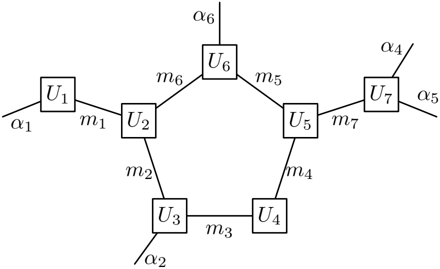
where $span\{|\alpha_i\rangle : \alpha_i = 1, \dots, q_i\} = \Lambda_i = \mathbf{C}^{q_i}$ and $\gamma = 1, \dots, m$.



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

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

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



Matrix product state (MPS) representation / DMRG / TT

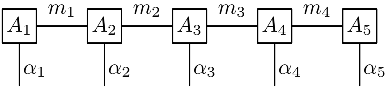
Affleck, Kennedy, Lieb Tagasaki (87); Fannes, Nachtergale, Werner (91), White (92),

Römmer & Ostlund (94), Vidal (03), Verstraete (04), Oseledets & Tyrtyshnikov, (09)

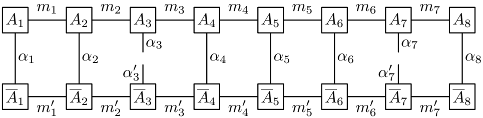
The tensor U is given elementwise as

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

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



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



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

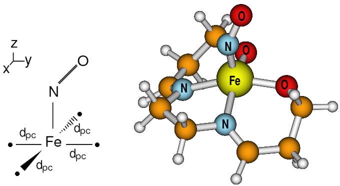
Mutual information, $I = s_i + s_j - s_{ij}$.

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

Interactions, entanglement and correlations

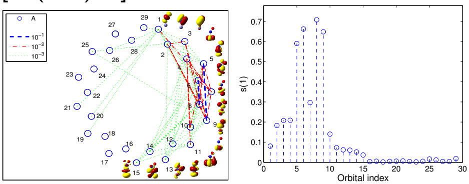
$$\mathcal{H} = \sum_{ij\alpha\beta} T_{ij}^{\alpha\beta} c_{i\alpha}^{\dagger} c_{j\beta} + \frac{1}{2} \sum_{ijkl\alpha\beta\gamma\delta} V_{ijkl}^{\alpha\beta\gamma\delta} c_{i\alpha}^{\dagger} c_{j\beta}^{\dagger} c_{k\gamma} c_{l\delta} \, ,$$

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



(a) $[\mathrm{Fe(NO)}]^{2+}$] (b) FeL(NO)

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



(a) Mutual information (b) Single orbital entropy

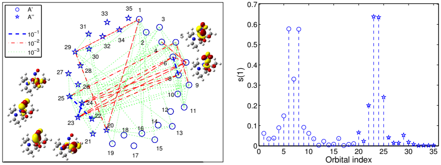
open d and f shells

Strongly correlated system

Effect of environment

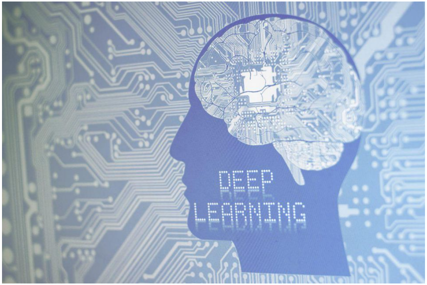
Boguslawski, Tecmer, Ö.L., Reiher (2012)

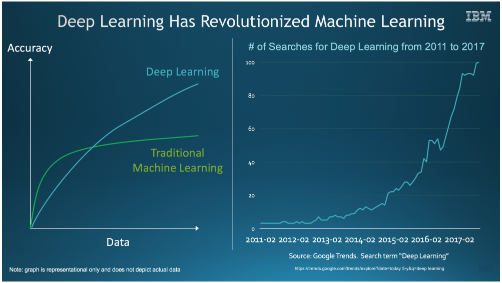
FeLNO



(a) Mutual information (b) Single orbital entropy

Deep learning, AI, ML, robotic

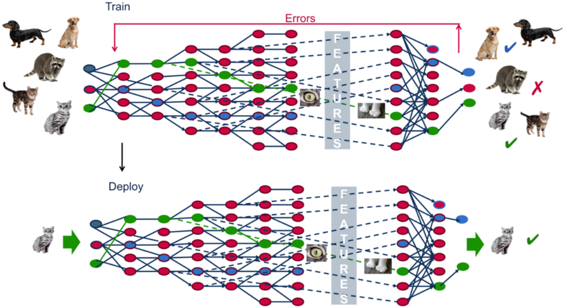




Note: graph is representational only and does not depict actual data

Source: Google Trends. Search term "Deep Learning"

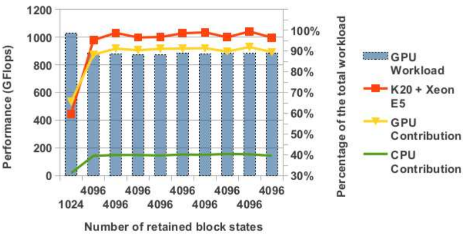
https://trends.google.com/trends/explore?date=today 5-y&q=deep learning



New sensors: machines begin to interact with our world!

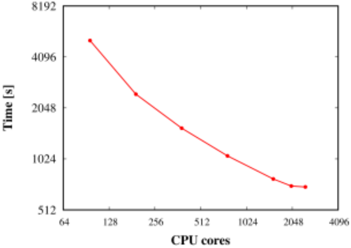
Towards exascale computations on supercomputers

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



K computer
Performance of over 10 Peta
floating point number operations per second
(10 Peta=10,000,000,000,000,000)

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



(a) Davidson procedure

FeMoco cluster
[CAS(113,76)]

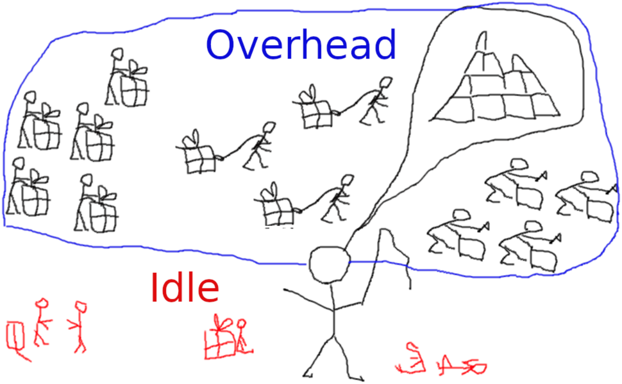
Centralized scheduling: unideal society

- 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

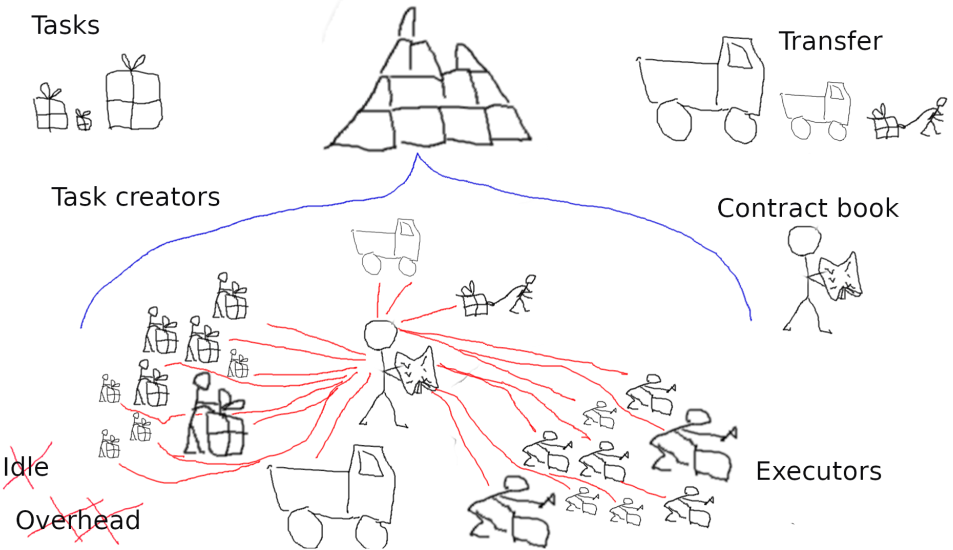
- Central scheduler performs lot of measurements, estimations, communication to rearrange tasks and workers → huge overhead



- 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



Tasks

Transfer

Task creators

Contract book

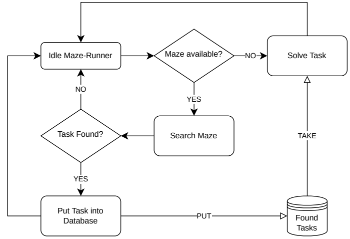
Executors

~~Idle~~

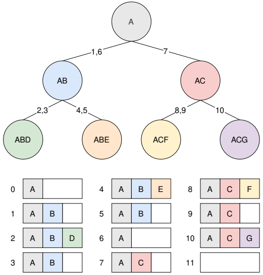
~~Overhead~~

Novel algorithmic solutions A. Menczer, ÖL (2023)

Life Cycle of a Maze-Runner Thread.

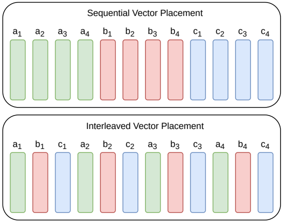


Graph theory based memory management

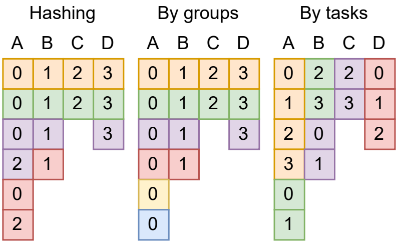


0	A		
1	A	B	
2	A	B	D
3	A	B	
4	A	B	E
5	A	B	
6	A		
7	A	C	
8	A	C	F
9	A	C	
10	A	C	G
11			

Strided Batched operations via data localization


Sequential Vector Placement
a_1 a_2 a_3 a_4 b_1 b_2 b_3 b_4 c_1 c_2 c_3 c_4
Interleaved Vector Placement
a_1 b_1 c_1 a_2 b_2 c_2 a_3 b_3 c_3 a_4 b_4 c_4

Execution via hierarchy of tasks



Hashing

A	B	C	D
0	1	2	3
0	1	2	3
0	1		3
2	1		
0			
2			

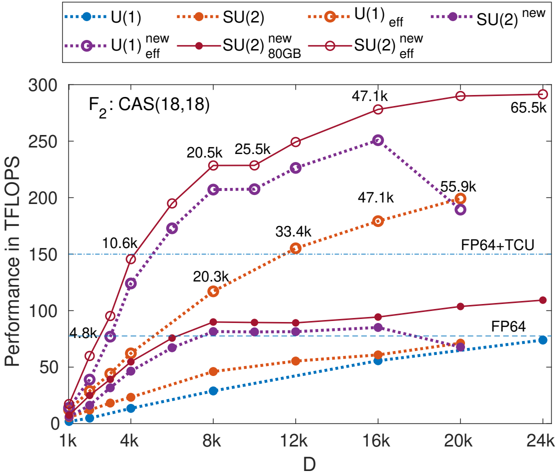
By groups

A	B	C	D
0	1	2	3
0	1	2	3
0	1		3
0	1		
0			
0			

By tasks

A	B	C	D
0	2	2	0
1	3	3	1
2	0		2
3	1		
0			
1			

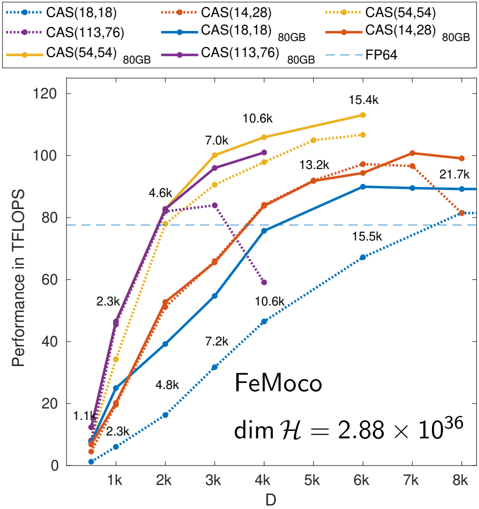
Boosting the effective performance via non-Abelian symmetries



A. Menczer, Ö.L (2023), CAS(18,18)

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

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



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

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

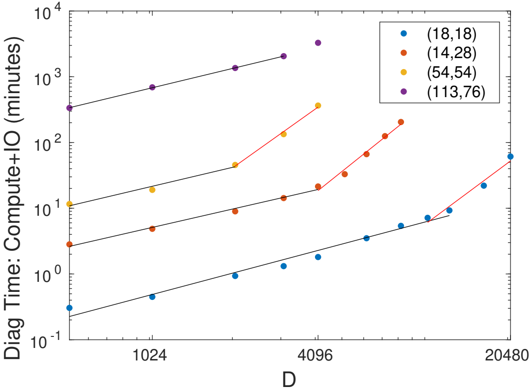
- Power consumption reduced by a factor of 5 to 8
  → Green DMRG

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

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

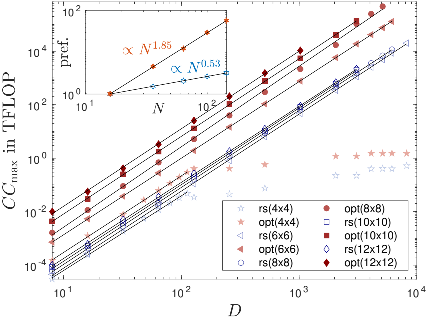
- Combination of our MPI and GPU kernels: → petascale computing.

Reducing D^3 scaling to linear scaling

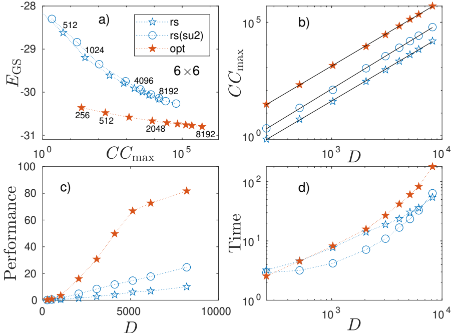


- New model to utilize NVIDIA D2D links. A. Menczer ÖL (unpublished 2023)
- NVIDIA DGX H100 and Grace Hopper GH200:
Testing performance up to ∼ 240 TFLOPS in collab with NVIDIA and SandboxAQ M. van Damme, A. Menczer, M. Ganahl, J. Hammond, Ö.L
- Combination of our MPI and GPU kernels:
multiNode-multiGPU → petascale computing. A. Menczer ÖL (unpublished

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

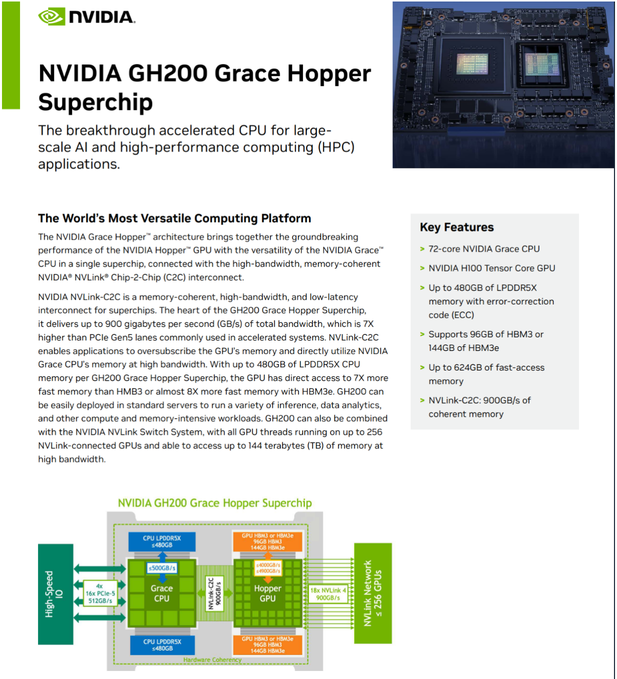


- The solid lines are first-order polynomial fits leading to exponents $\nu \simeq 3 \pm 0.2$
- inset: scaling of the prefactor as a function of system size N with fitted exponents 0.53 and 1.85 for the real space and for the optimized basis, respectively.



- 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

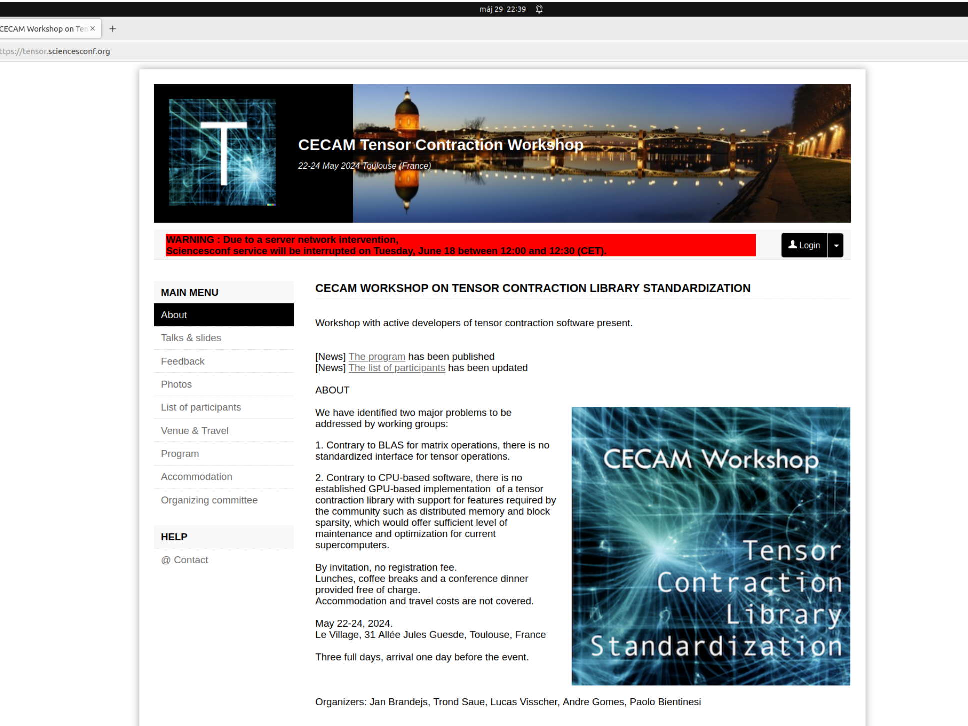
The NVIDIA Grace Hopper™ architecture brings together the groundbreaking performance of the NVIDIA Hopper™ GPU with the versatility of the NVIDIA Grace™ CPU in a single superchip, connected with the high-bandwidth, memory-coherent NVIDIA® NVLink® Chip-2-Chip (C2C) interconnect.

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

NVIDIA GH200 Grace Hopper Superchip



CECAM Tensor Contraction Workshop

22-24 May 2024 Toulouse (France)

**WARNING : Due to a server network intervention,
Sciencesconf service will be interrupted on Tuesday, June 18 between 12:00 and 12:30 (CET).**

CECAM WORKSHOP ON TENSOR CONTRACTION LIBRARY STANDARDIZATION

Workshop with active developers of tensor contraction software present.

[News] The program has been published
[News] The list of participants has been updated

ABOUT

We have identified two major problems to be addressed by working groups:

1. Contrary to BLAS for matrix operations, there is no standardized interface for tensor operations.

2. Contrary to CPU-based software, there is no established GPU-based implementation of a tensor contraction library with support for features required by the community such as distributed memory and block sparsity, which would offer sufficient level of maintenance and optimization for current supercomputers.

By invitation, no registration fee.
Lunches, coffee breaks and a conference dinner provided free of charge.
Accommodation and travel costs are not covered.

May 22-24, 2024.
Le Village, 31 Allée Jules Guesde, Toulouse, France

Three full days, arrival one day before the event.

Organizers: Jan Brandejs, Trond Saue, Lucas Visscher, Andre Gomes, Paolo Bientinesi

Conclusion

TNS methods + concepts of QIT + AI motivated hardware advances finally facilitates research on strongly correlated real enzymes and materials via exascale computation

Workshop: Recent progress on tensor network methods, April 22-25, 2024 TUM Institute for Advanced Study, Munich

Current work: DMRG+ORCA: 536 electrons on 1368 orbitals (Icosacene).

Supports: Lendület grant of the Hungarian Academy of Sciences, the Hungarian National Research, Development and Innovation Office TKP2021-NVA-04, 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 (TUM-IAS, Germany), SPEC, DOE, (PNNL, USA)

Thank you for your attention