Title: Tensor-Processing Units and the Density-Matrix Renormalization Group

Speakers: Martin Ganahl

Series: Machine Learning Initiative

Date: May 18, 2023 - 11:00 AM

URL: https://pirsa.org/23050036

Abstract: Tensor Processing Units are application specific integrated circuits (ASICs) built by Google to run large-scale machine learning (ML) workloads (e.g. AlphaFold). They excel at matrix multiplications, and hence can be repurposed for applications beyond ML. In this talk I will explain how TPUs can be leveraged to run large-scale density matrix renormalization group (DMRG) calculations at unprecedented size and accuracy. DMRG is a powerful tensor network algorithm originally applied to computing ground-states and low-lying excited states of strongly correlated, low-dimensional quantum systems. For certain systems, like one-dimensional gapped or quantum critical Hamiltonians, or small, strongly correlated molecules, it has today become the gold standard method for computing e.g. ground-state properties. Using a TPUv3-pod, we ran large-scale DMRG simulations for a system of 100 spinless fermions, and optimized matrix product state wave functions with a bond dimension of more than 65000 (a parameter space with more than 600 billion parameters). Our results clearly indicate that hardware accelerator platforms like Google's latest TPU versions or NVIDIAs DGX systems are ideally suited to scale tensor network algorithms to sizes that are beyond capabilities of traditional HPC architectures.

Zoom link: https://pitp.zoom.us/j/99337818378?pwd=SGZvdFFValJQaDNMQ0U1YnJ6NU1FQT09

Tensor-Processing Units and the Density-Matrix Renormalization Group

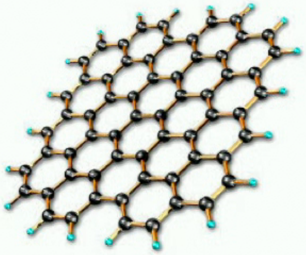
PRX Quantum 4, 010317, 2023

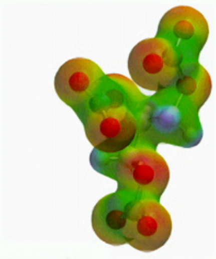
Martin Ganahl, Markus Hauru, Adam Lewis, Tomasz Wojno, Jackson Beall, Jae Yoo, Yijian Zou, Guifre Vidal

Perimeter Institute, May 18 2023

Martin Ganahl

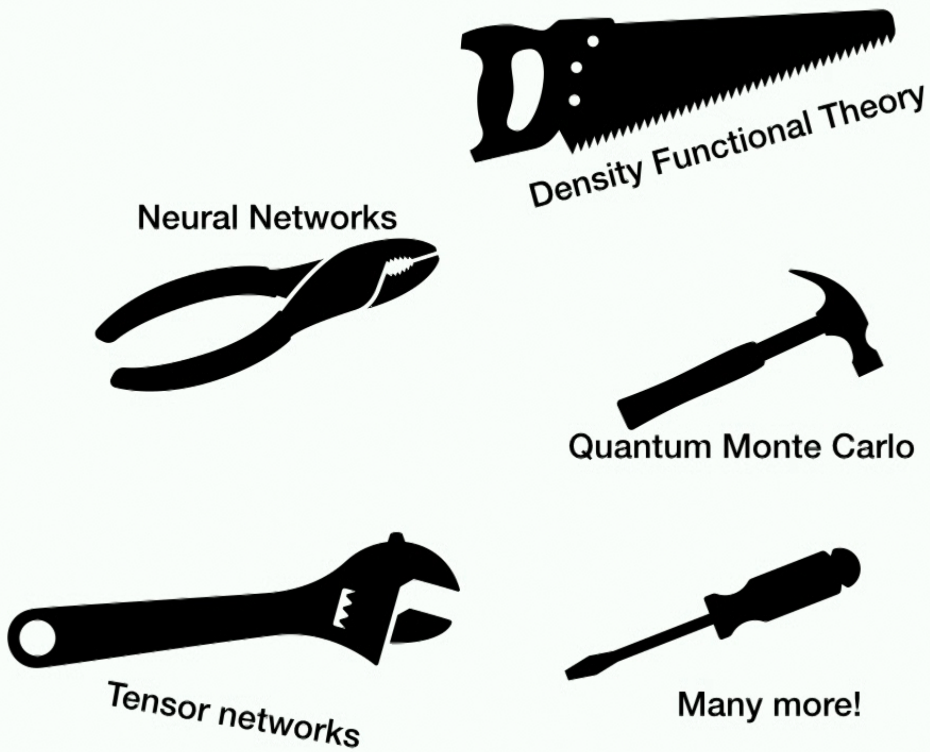
QUANTUM SIMULATION TOOLBOX





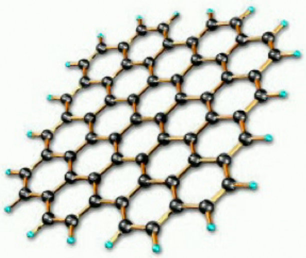
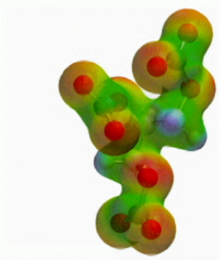
strongly correlated materials and molecules

infamously computationally challenging

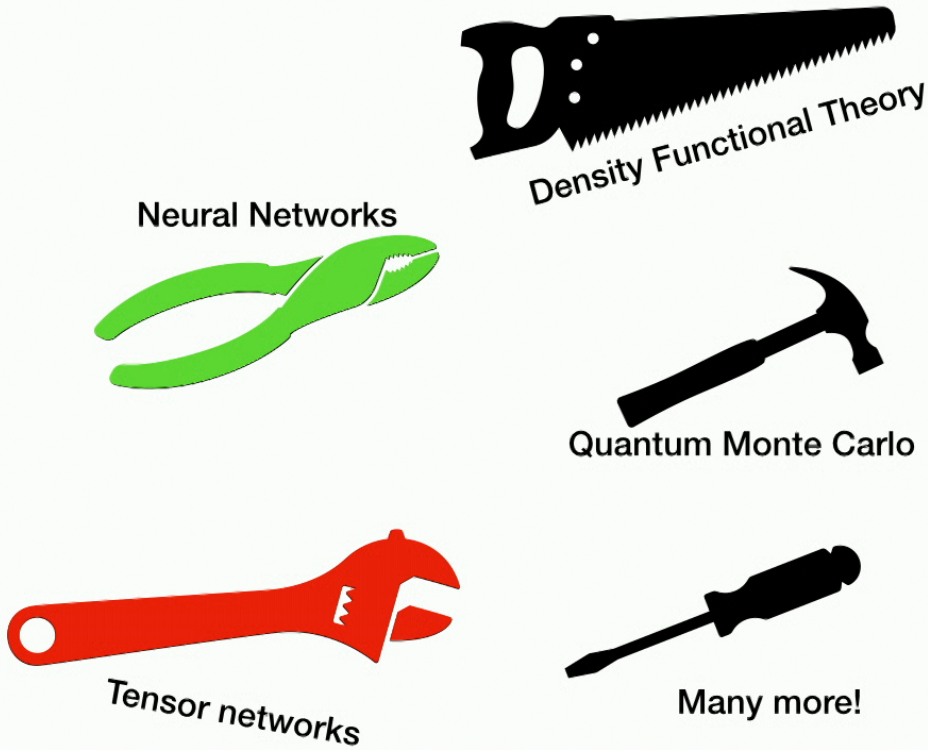


Martin Ganahl

QUANTUM SIMULATION TOOLBOX

strongly correlated materials and molecules



infamously computationally challenging

Benefits of quantum simulation on hardware accelerators

Martin Ganahl

HARDWARE ACCELERATORS AND QUANTUM SIMULATIONS

AI revolution fuelled by hardware development (GPU, TPU, FPGA) and vice-versa

Neural Networks

Success of neural networks depends critically on hardware accelerators

Martin Ganahl

HARDWARE ACCELERATORS AND QUANTUM SIMULATIONS

AI revolution fuelled by hardware development (GPU, TPU, FPGA) and vice-versa

Neural Networks

Success of neural networks depends critically on hardware accelerators

Tensor networks

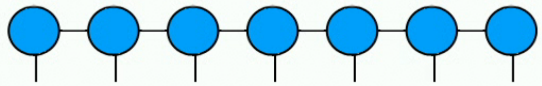
Computational tool for simulating strongly correlated quantum systems

Polynomial scaling, but can still become **computationally demanding**

Can benefit substantially from hardware accelerators (A. Menczer, **O. Legeza**, arXiv:2305.05581, Unfried, Hauschild & Pollmann 2023, ITensorGPU.jl by K. Hyatt)

Martin Ganahl

FAMOUS TN: THE DENSITY MATRIX RENORMALISATION GROUP

Algorithm for approximating ground-states of **local, 1d** Hamiltonians as matrix product states (MPS) 

Applications: condensed matter physics, **material science, quantum chemistry, machine learning**, even solving PDEs

DMRG is a **stepping stone** towards more sophisticated tensor networks methods (e.g. projected entangled pair states for 2d quantum systems)

Martin Ganahl

Overview

Brief overview of TPUs

Brief introduction to tensor networks

Showcase of results

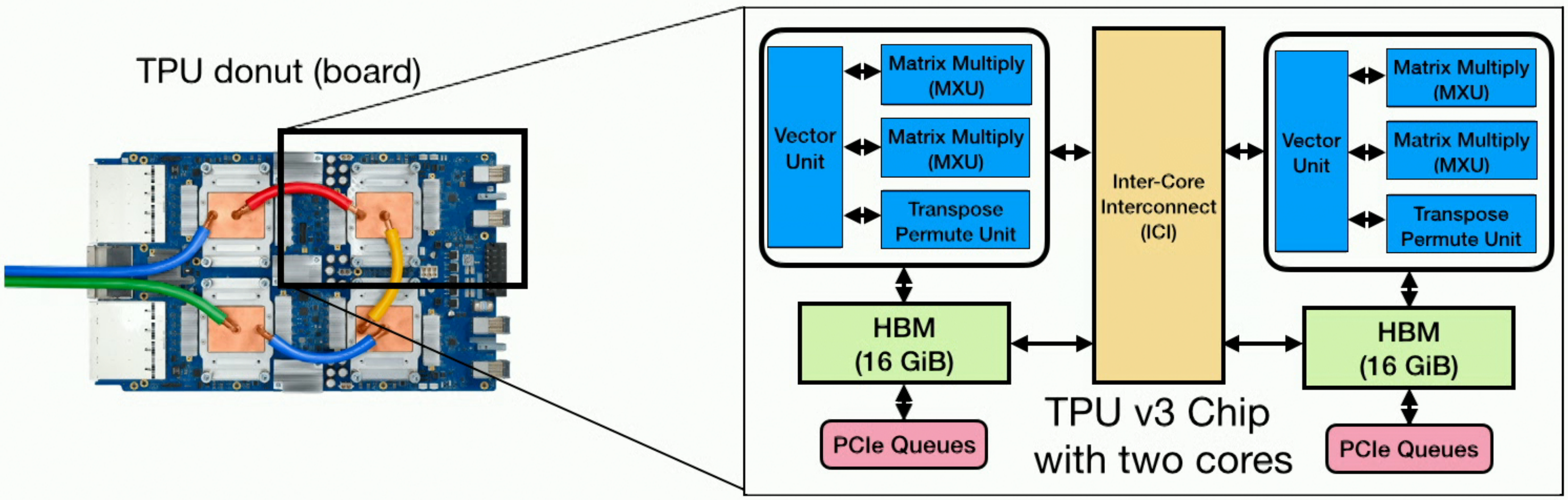
Martin Ganahl

WHAT ARE TENSOR PROCESSING UNITS?

- Custom-built ASICs (by Google) to support large-scale ML tasks
- Mostly used in Google's data-centers, but recently becoming accessible through Google Cloud Platform (GCP) (TPU-v4 just went online!)
- **Very** fast at matrix-multiplication
- **Useful for applications beyond ML, i.e. quantum simulation, tensor networks?**

Martin Ganahl

TPU-V3-BOARD AND TPU-V3 CHIP



- MXU: 128 x 128 **systolic array** (FLOPs in bfloat16, accumulation in float32)
- Each TPU core has access to **16 GB on-chip high-bandwidth memory**
- **Fast** Intercore Interconnect (ICI) links with bandwidth **656 Gbits/s**
- Peak Performance of a TPUv3 chip **~120/4 TFLOPS in bfloat16/float32** precision
- **A TPUv3-chip is roughly comparable to a modern GPU chip**

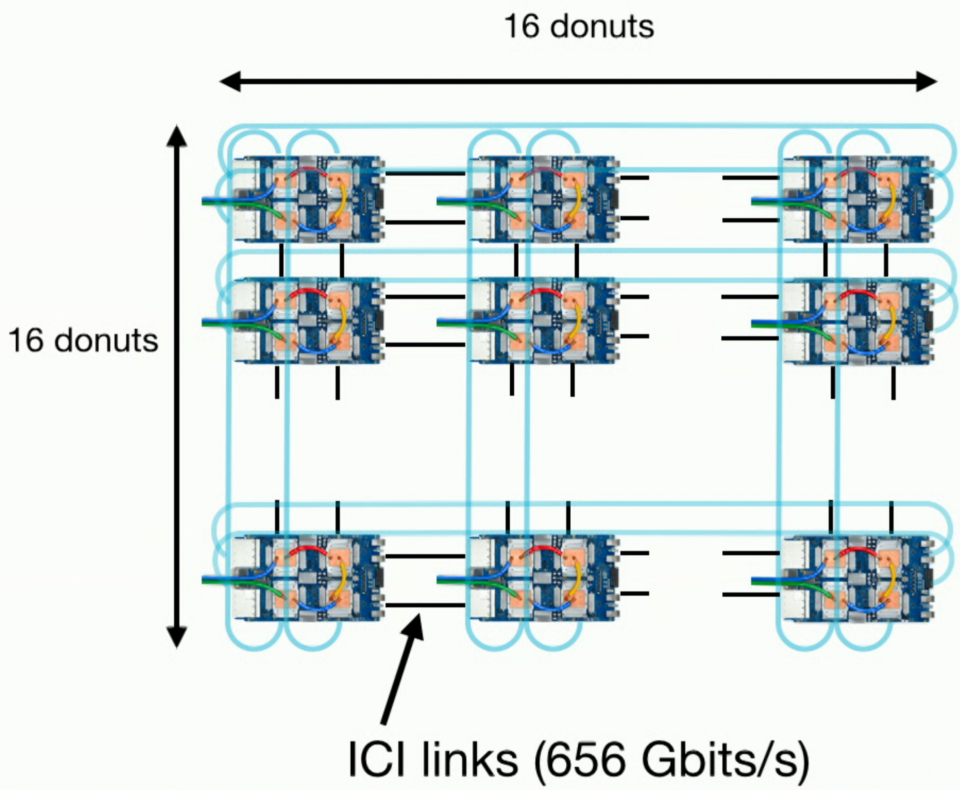
Martin Ganahl

WHAT ARE THE BENEFITS OF TPUS?

- **TPU chips can be coupled into a computing cluster (TPU pod)**
- **Pod sizes of up to 4096 (TPU-v4) TPU cores with ICI communication (NVidia's DGX has up to 8 A100 GPUs coupled via NVLink)**
- **Easily accessible through Google Cloud**

Martin Ganahl

WHAT IS A TPU POD?



- 2048 TPU cores
- 32 TB of non-shared(!) memory
- ~120 PetaFLOPS (bfloat16)

Martin Ganahl

TPUS - SOME CHALLENGES

No packages for large-scale linear algebra

Standard algos don't translate well to TPU architecture

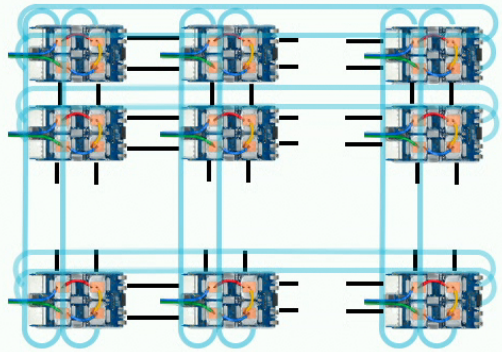
Double precision only at significant computational cost

Compilation times ("tracing") can become prohibitive

Martin Ganahl

TPU (V3) SUMMARY

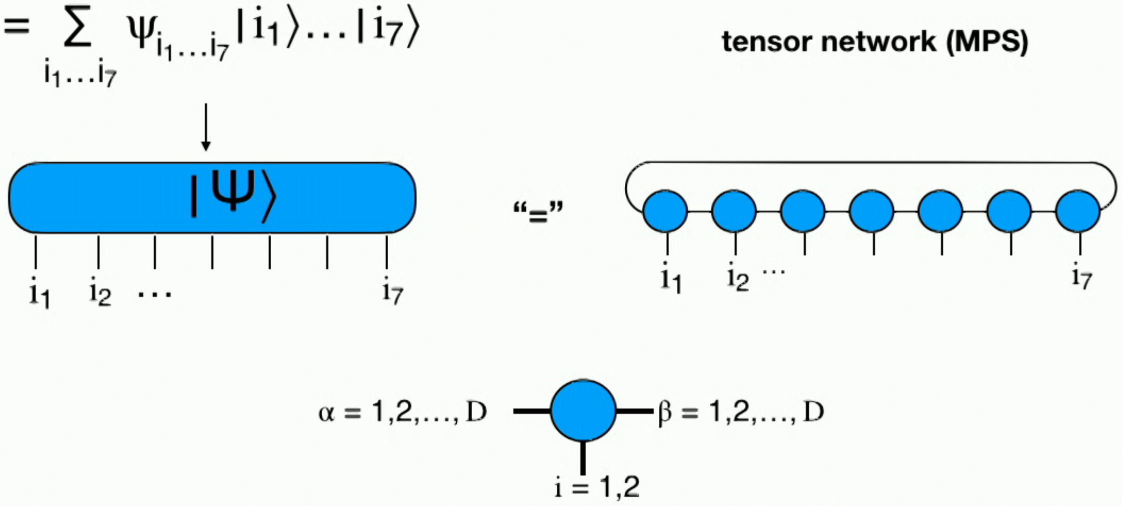
- A TPU pod is a **cluster** of tightly coupled accelerators
- 120 PetaFLOPS (bfloat16)
- extremely fast communication:
  42 Tbits worst-case bisection bandwidth
  (Infiniband for comparable setup has 6.4 Tbits)
- Each board is controlled by a separate CPU host
- TPU pod can be programmed using the SIMD paradigm (using python + JAX as high-level entry point)



Martin Ganahl

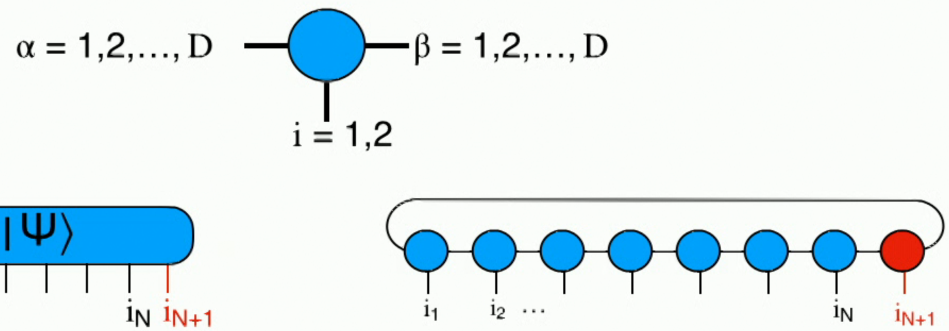
TENSOR NETWORKS

$$|\Psi_n\rangle = \sum_{i_1 \dots i_7} \psi_{i_1 \dots i_7} |i_1\rangle \dots |i_7\rangle$$



Martin Ganahl

TENSOR NETWORKS



2^N 2 complex numbers

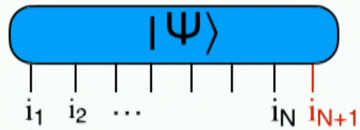
inefficient

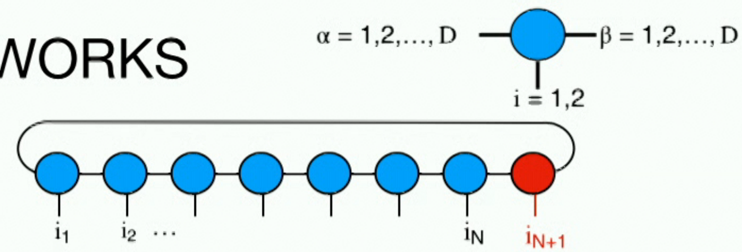
$\mathcal{O}(N)$ $+2D^2$ complex numbers

efficient

Martin Ganahl

TENSOR NETWORKS





2^N 2 complex numbers

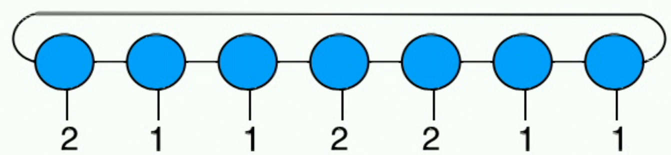
inefficient

$\mathcal{O}(N)$ $+2D^2$ complex numbers

efficient

Amplitudes

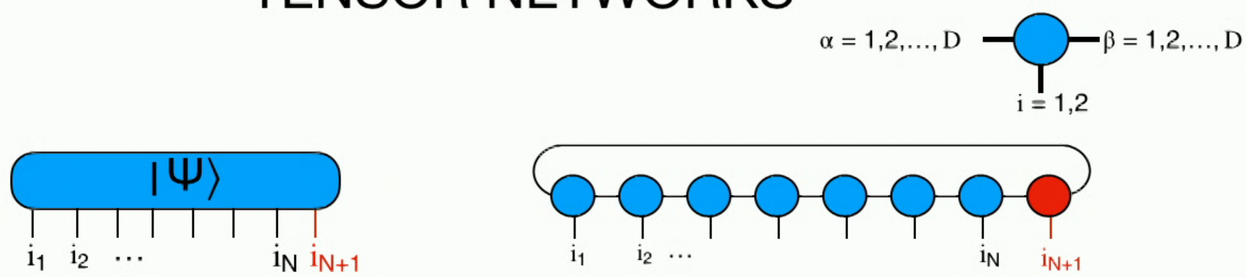
$\psi_{2112211} =$



Efficient for this network (MPS)

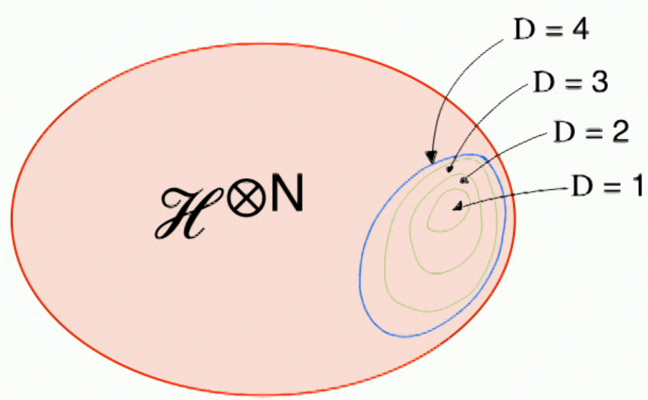
Martin Ganahl

TENSOR NETWORKS



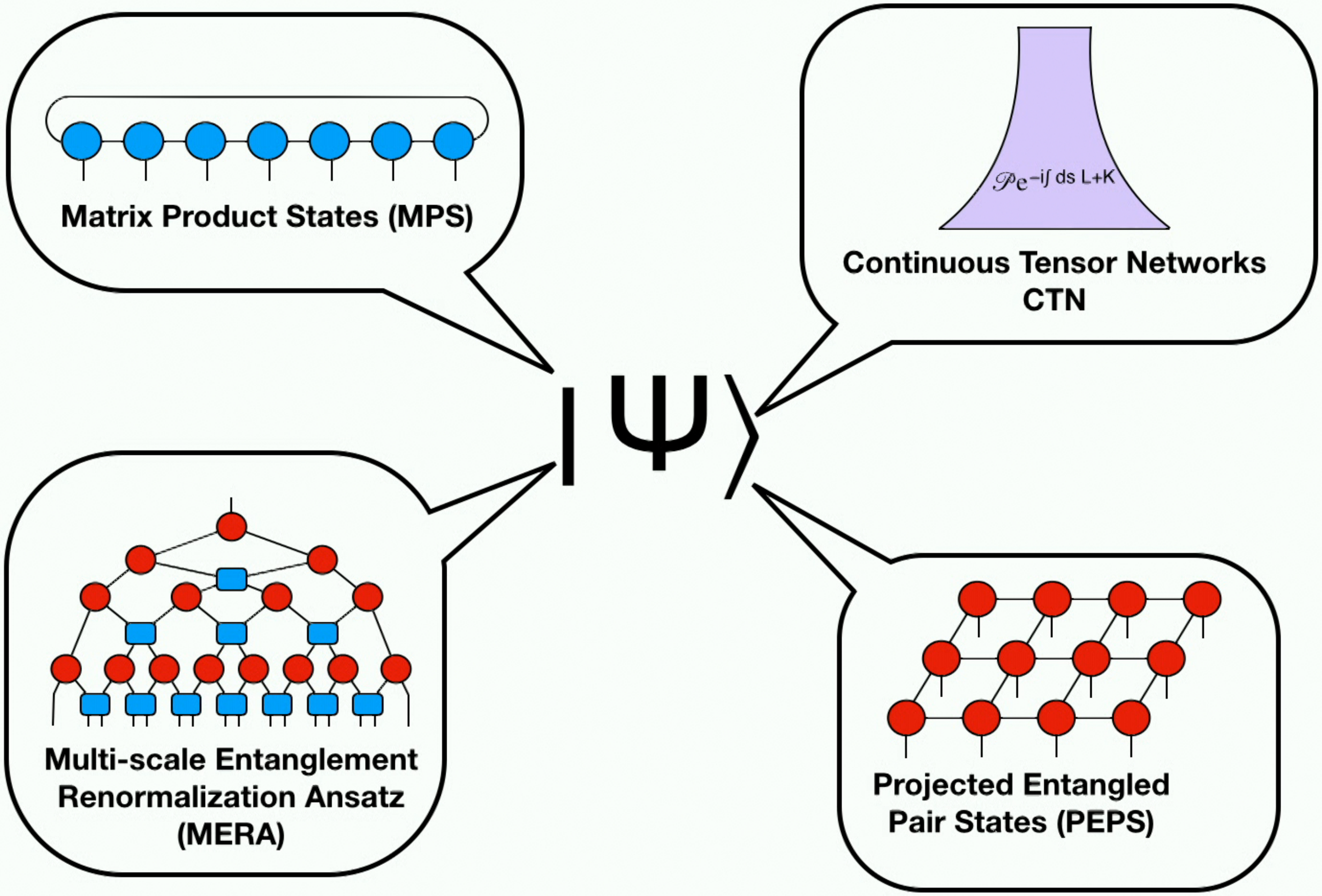
2^N 2 complex numbers $\mathcal{O}(N)$ +dD² complex numbers

inefficient **efficient**



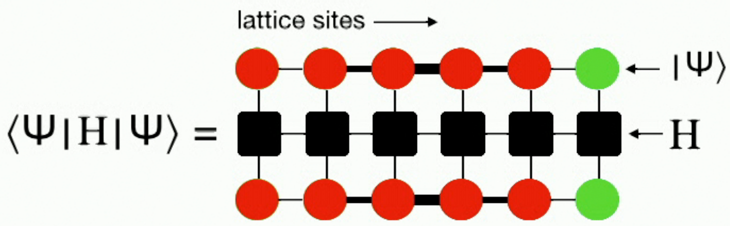
Martin Ganahl

OTHER TENSOR NETWORKS



Martin Ganahl

DMRG FROM 20000 FEET



"Alternating least squares optimisation"

DMRG requires a **1d geometry**

Scaling is ND^3, i.e. linear in number of lattice sites and cubic in the bond dimension

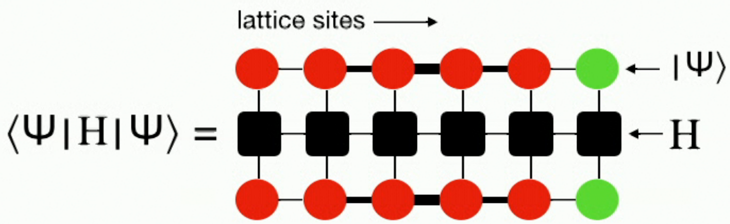
"Stronger" correlations require "larger" bond dimensions (i.e. more entangled ground states), e.g. for

- Quantum critical points
- **Long-range interactions**

Routinely applied to "quasi-2d" systems, e.g. long stripes, thin cylinders, a.s.o

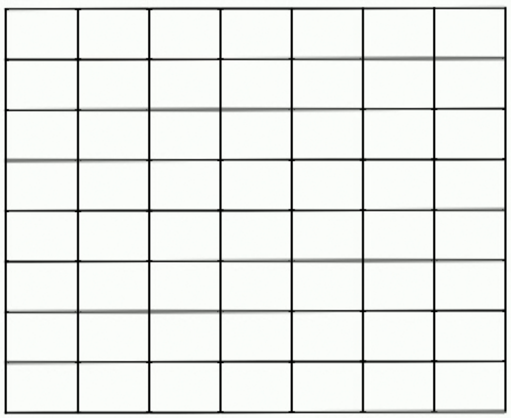
Martin Ganahl

DMRG FROM 20000 FEET



"Alternating least squares optimisation"

mapping 2d to 1d



DMRG requires a **1d geometry**

Scaling is ND^3, i.e. linear in number of lattice sites and cubic in the bond dimension

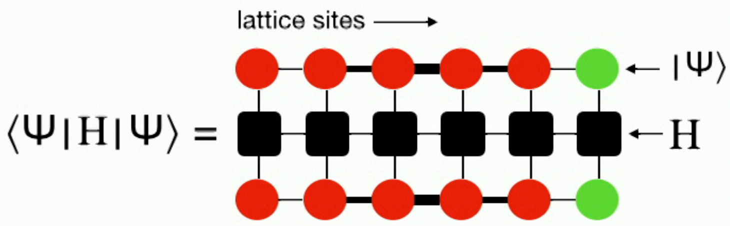
"Stronger" correlations require "larger" bond dimensions (i.e. more entangled ground states), e.g. for
- Quantum critical points
- **Long-range interactions**

Routinely applied to "quasi-2d" systems, e.g. long stripes, thin cylinders, a.s.o

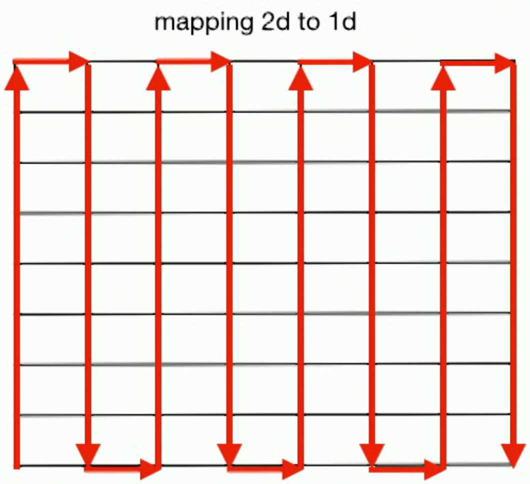
Stoudenmire & White, 2011

Martin Ganahl

DMRG FROM 20000 FEET



"Alternating least squares optimisation"



DMRG requires a **1d geometry**

Scaling is ND^3, i.e. linear in number of lattice sites and cubic in the bond dimension

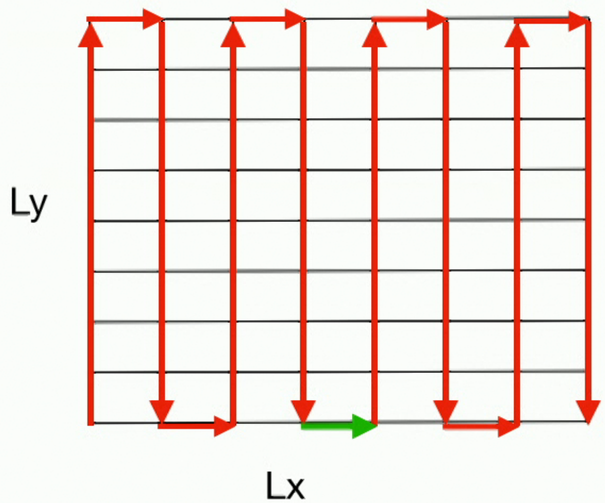
"Stronger" correlations require "larger" bond dimensions (i.e. more entangled ground states), e.g. for
- Quantum critical points
- **Long-range interactions**

Routinely applied to "quasi-2d" systems, e.g. long stripes, thin cylinders, a.s.o

Stoudenmire & White, 2011

Martin Ganahl

DMRG FOR 2D SYSTEMS



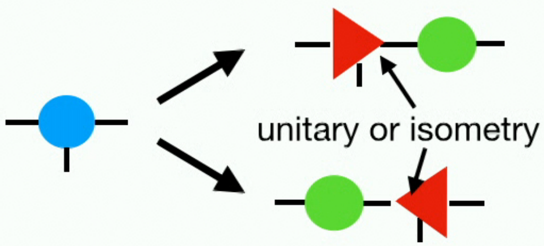
A nearest neighbour Hamiltonian on the square lattice becomes longer-ranged in the new representation!!!

Bond dimension needs to scale as D~exp(Ly) (in the best case)!!!

Martin Ganahl

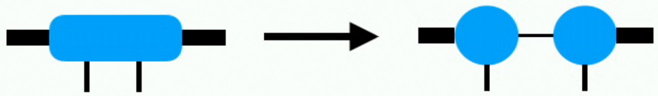
DMRG-THE CORE OPERATIONS

Orthogonalisation



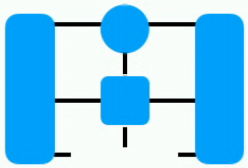
via SVD, QR or
polar factorisation

Rank-reduction ("truncation")



via SVD

Network contractions

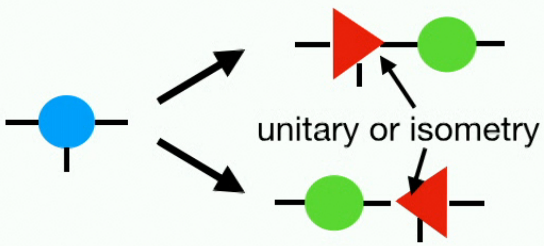


reduction to matrix
multiplications via
GEMM, DGEMM,…

Martin Ganahl

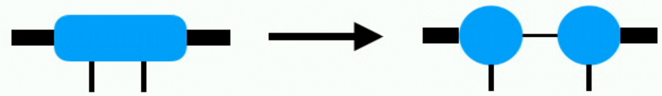
DMRG-THE CORE OPERATIONS

Orthogonalisation



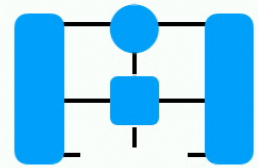
via SVD, QR or
polar factorisation

Rank-reduction ("truncation")



via SVD

Network contractions



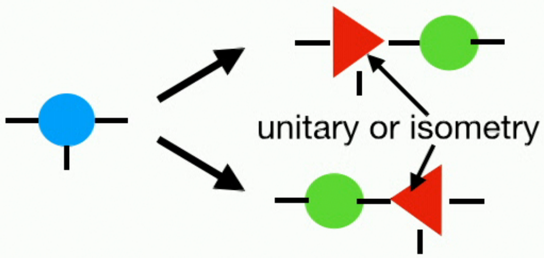
reduction to matrix
multiplications via
GEMM, DGEMM,…

DMRG, TD-DMRG, TEBD, TDVP, PEPS, MERA
Quantum Chemistry, ML-applications, condensed matter

Martin Ganahl

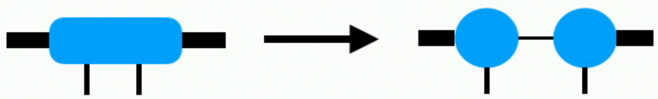
DMRG CORE OPERATIONS

Orthogonalisation



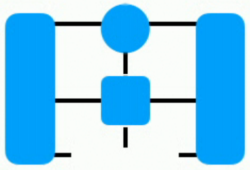
via SVD, QR or
polar factorisation

Rank-reduction ("truncation")



via SVD

Network contractions

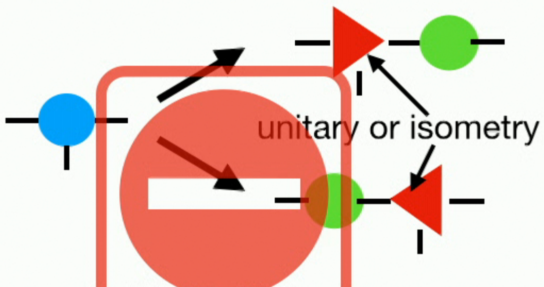


reduction to matrix
multiplications via
GEMM, DGEMM,…

Martin Ganahl

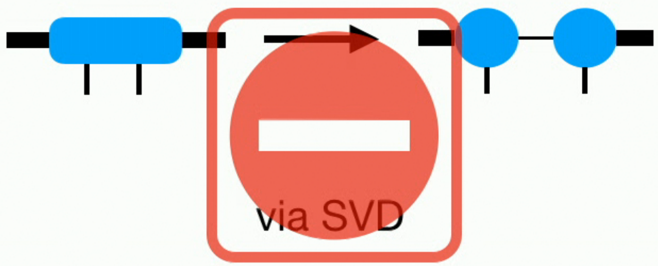
DMRG CORE OPERATIONS

Orthogonalisation



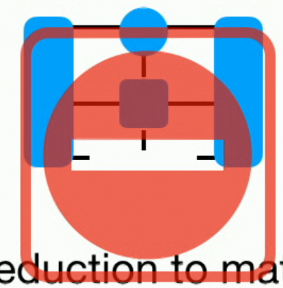
via SVD, QR or
polar factorisation

Rank-reduction ("truncation")



via SVD

Network contractions



reduction to matrix
multiplications via
GEMM, DGEMM,…

Martin Ganahl

DMRG ON TPUS - SOME CHALLENGES

Data distribution & tensor contractions

Scalable orthogonalisation methods

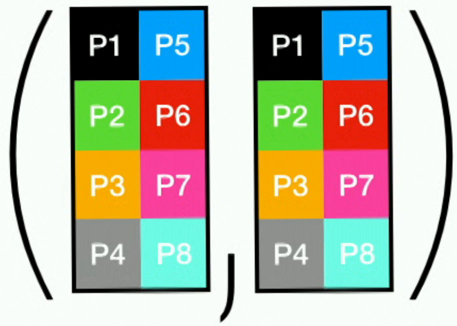
Scalable rank-reduction without SVD

Martin Ganahl

DATA DISTRIBUTION & TENSOR CONTRACTIONS

All DMRG tensors are of rank 3 -> treat them as a set of matrices

$$A^{i_n} = (A^0, A^1) =$$

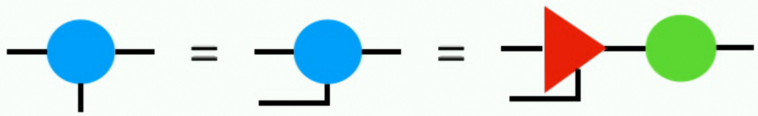


Each matrix is distributed across all available cores using a checkerboard pattern

Tensor contractions are reduced to loops over distributed matrix multiplications using SUMMA

Martin Ganahl

SCALABLE ORTHOGONALISATION



Back then, no scalable QR implementation was available!!
(That has changed now, see A. Lewis et al., arxiv:2112.09017)

Instead we use a **polar decomposition:** $A = UH$

Unitary — Positive semi-definite

Polar factor U can be obtained as the limit of a Newton-Schultz iteration $X_{n+1} = X_n \cdot (3/2 - (1/2)X_n^\dagger X_n)$

$$U = \lim_{n\to\infty} X_n$$

Martin Ganahl

SCALABLE RANK REDUCTION WITHOUT SVD

$M \ldots$ N by N matrix

Rank reduction via SVD: $M = USV \rightarrow \tilde{M} = U\tilde{S}V$ $\tilde{S}$: Truncate all values below δ

$$M = UH$$

$$M' \equiv H - \delta = U'H'$$

The unitary factor U' has eigenvalues {+1, -1} for eigenvalues of M' {>0, <0}

$P = (1 + U')/2$ Projects into the space of positive eigenvalues of U'

P is an N by N matrix for rank N/2, i.e. it has N/2 eigenvalues equal to 0

Using the subspace iteration one can find an orthogonal basis C for the column space of P s.t. $P = CC^\dagger$

C Is an N by m matrix, with m < N $\rightarrow \tilde{M} = UPH = UCC^\dagger H$

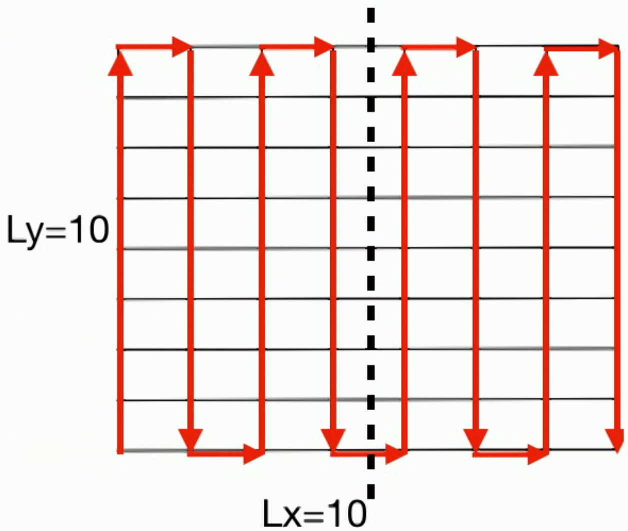
Martin Ganahl

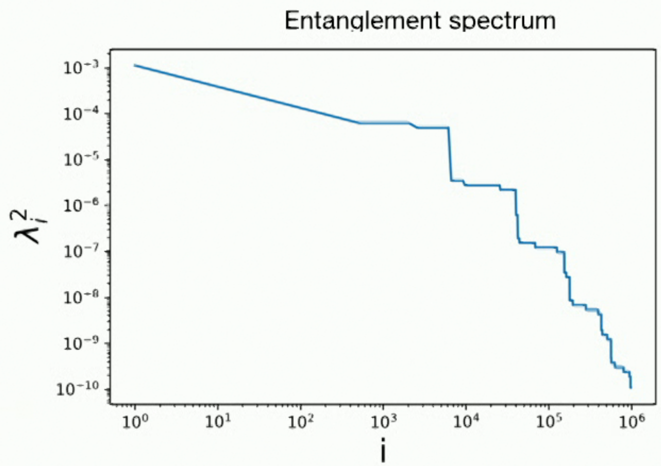
FREE FERMIONS ON 2D-SQUARE LATTICE

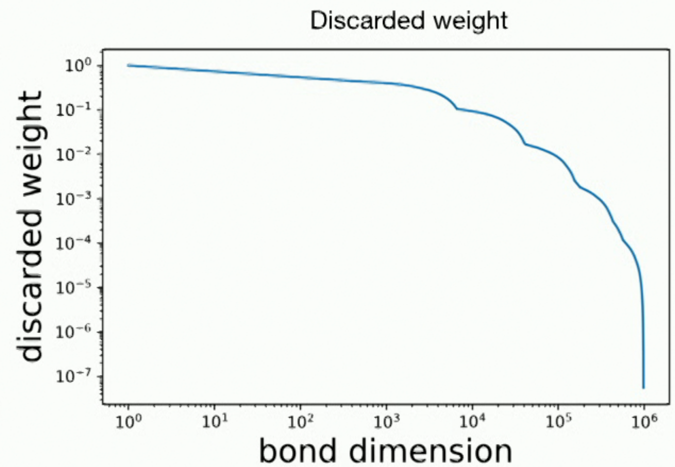
$$H_{sf} = -\sum_{\langle i,j\rangle} c_i^{\dagger} c_j + \mu \sum_i c_i^{\dagger} c_i \qquad \mu = 0$$ (Half filling)

Gapless excitations due to a 1d - fermi surface yield logarithmic corrections to area law in 2d

This is a very hard problem for DMRG!!





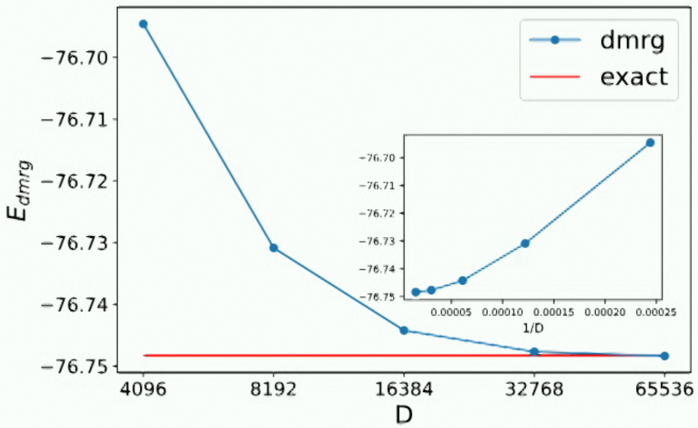


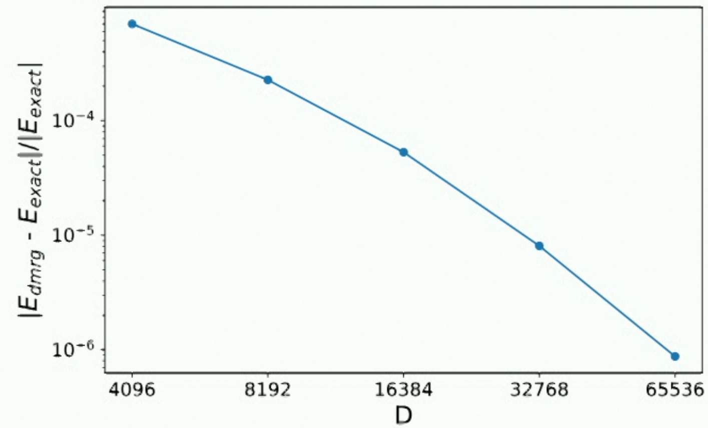
Martin Ganahl

FREE FERMIONS ON 2D-SQUARE LATTICE

$$H_{sf} = -\sum_{\langle i,j \rangle} c_i^\dagger c_j + \mu \sum_i c_i^\dagger c_i \quad \mu = 0$$ (Half filling)

Gapless excitations due to a 1d - fermi surface yield logarithmic corrections to area law in 2d





D = bond dimension

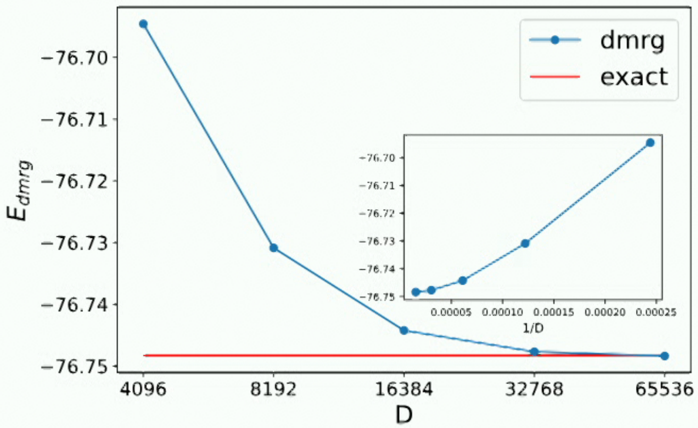
Martin Ganahl

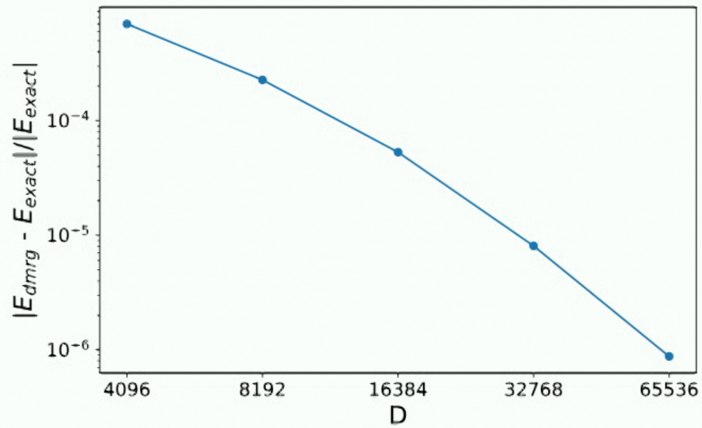
FREE FERMIONS ON 2D-SQUARE LATTICE

$$H_{sf} = -\sum_{\langle i,j \rangle} c_i^\dagger c_j + \mu \sum_i c_i^\dagger c_i \quad \mu = 0$$ (Half filling)

No symmetries utilised!

Gapless excitations due to a 1d - fermi surface yield logarithmic corrections to area law in 2d





D = bond dimension

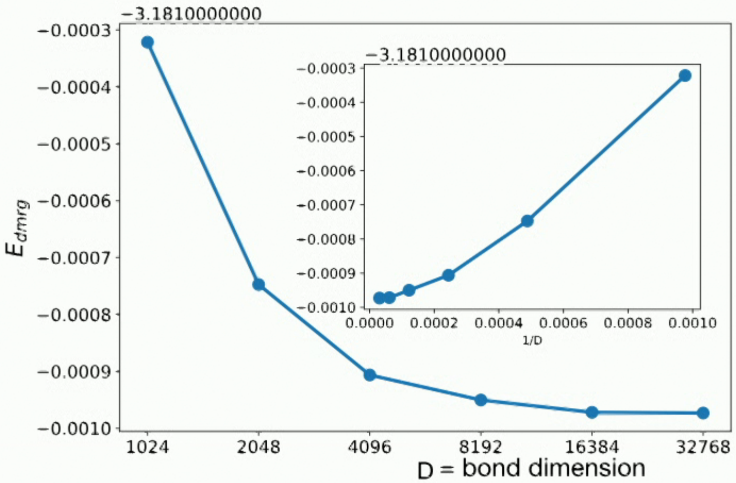
Martin Ganahl

TRANSVERSE-FIELD ISING MODEL

$$H_{TFI} = -\sum_{\langle i,j\rangle} \hat{\sigma}_i^z \hat{\sigma}_j^z + B \sum_i \hat{\sigma}_i^x$$

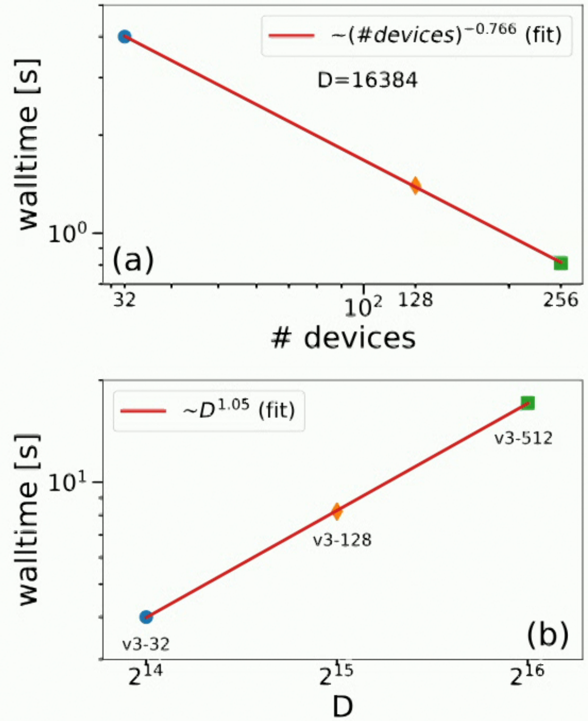
Critical point at $B \approx 3.0$

Robustly entangled ground-state, with entanglement obeying an area law



Martin Ganahl

String and weak scaling



Single optimisation step at D=2^{16} ~ 2 minutes on 1024 TPU cores

Summary & outlook

Tensor network algorithms scale extremely well to large-scale hardware accelerators

Large speedups at high utilisation achievable: DMRG with **D~65000 with no symmetries ~ 600B parameter optimization in less than a day**

Utilisation of symmetries will yield additional benefits (A. Menczer, O. Legeza, arXiv:2305.05581)

Other methods which would potentially benefit: PEPS, MERA & non-equilibrium methods

The importance of hardware accelerators for scientific computing will only increase in the future. Tensor network algorithms will very likely substantially benefit from this development!!

Large scale accelerators could make tensor network methods for higher dimensions even more competitive

Thank you for your attention!

Martin Ganahl