CUDA-Q - tensornet-mps backend

Tobias Valentin Bauer

Nvidia CUDA-Q training, LRZ - 18.02.2026

Outline

Tensornetworks and quantum computing - a cat-and-mouse game?

- *"Scientists Say They Used Classical Computers to Outperform Google's Sycamore QC"* - Quantum Insider
- *We argue that an ideal simulation of the same task can be performed on a classical system in 2.5 days and with far greater fidelity.* - Edwin Pednault, John Gunnels, and Jay Gambetta (IBM).
- *In our 2019 paper, we said that classical algorithms would improve. [...] We don't think this classical approach can keep up with quantum circuits in 2022 and beyond.* - Sergio Boixo, (Google Quantum AI)

Outline

1. Introduction
   - Kicker
   - What you'll be learning
   - TNs and quantumsimulation
2. Theory
   - Foundations and Preliminaries
   - Entanglement and maximal bond dimension
   - Expectation of a prepared state
   - Sampling with TN based quantumsimulation
3. Exercise session
4. Literature

What you'll be learning in this section

- Entanglement and maximal bond dimension
- How can we calculate expectation of a state with TNs?
- How can we simulate shots with TNs?
- Acceleration and use cases of TN-based quantum simulation
  → *Practical session*
- **Goal:** Give you a more nuanced understanding for what TN backends are capabale of and what their boundaries are.

Outline

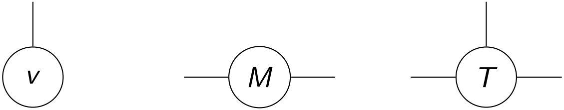
Outline

Memory demand of statevector simulation

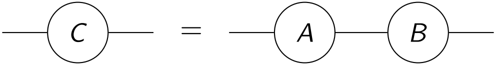
$$|\psi\rangle = \sum_{i=1}^{2^n} \alpha_i |i\rangle, \text{(n-qubit quantum state)}$$

- Single precision: 4 byte for real- and 4 byte for complex part of α_i
- $\Rightarrow 2^n \times 8$ byte memory-footprint for n qubits.
- $n = 31$ qubits:
$$2^{31} \times 8 \approx 17 \text{ GiB}$$
- $n = 32$ qubits:
$$2^{32} \times 8 \approx 34 \text{ GiB}$$
- $n = 33$ qubits:
$$2^{33} \times 8 \approx 64 \text{ GiB}$$

Tensornetwork notation



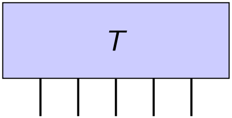
Tensor contraction



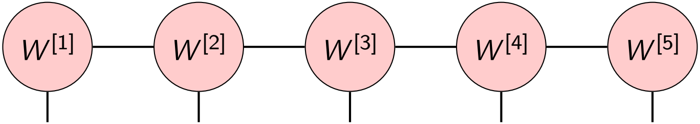
Einstein summation notation

$$C_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj} = A_{ik} B_{kj}$$

Quantum states as tensors



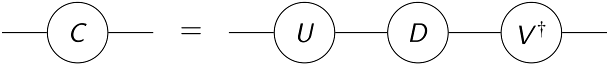
- $|\psi\rangle = T_{i_1,\dots,i_n}|i_1 \dots i_n\rangle;\ i_j \in \mathbb{F}_2^{\otimes n};\ T_{i_1 \dots i_n} \in \mathbb{C}$



- $|\psi\rangle = T_{\sigma_1,\dots,\sigma_5}|\sigma_1 \dots \sigma_5\rangle = \sum_{\vec{\alpha},\vec{\sigma}} W^{[1]}_{\sigma_1\alpha_1} W^{[2]}_{\sigma_2\alpha_2} W^{[3]}_{\sigma_3\alpha_3} W^{[4]}_{\sigma_4\alpha_4} W^{[5]}_{\sigma_5\alpha_5}|\sigma\rangle$

How tensornetworks can help us in quantum simulation

Matrix SVD (graphical tensor notation)



Explicit notation

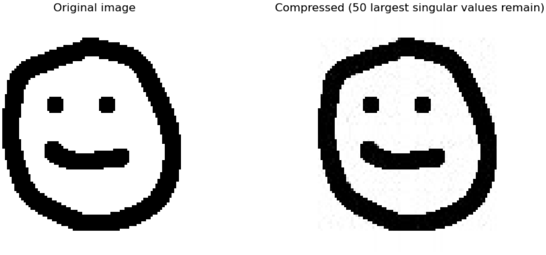
- $C = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix}$

  $= \begin{pmatrix} u_{00} & u_{10} \\ u_{01} & u_{11} \end{pmatrix} \begin{pmatrix} D_0 & \\ & D_1 \end{pmatrix} \begin{pmatrix} v_{00} & v_{10} \\ v_{01} & v_{11} \end{pmatrix}^\dagger$

- $\begin{pmatrix} u_{00} \\ u_{01} \end{pmatrix} D_0 \begin{pmatrix} v_{00} \\ v_{01} \end{pmatrix}^\dagger$ is the best approximation of C in Frobenius-norm assuming $D_0 \dots D_1$ and rows/columns of U, V are ordered from biggest to smallest.

How tensornetworks can help us in quantum simulation

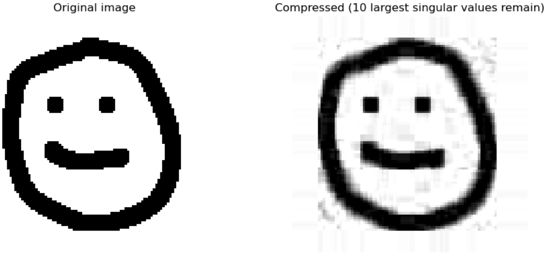
$$\begin{pmatrix} u_{00} & \cdots & u_{0,99} \\ \vdots & \ddots & \vdots \\ u_{99,0} & \cdots & u_{99,99} \end{pmatrix} \begin{pmatrix} D_0 & & 0 \\ & \ddots & \\ 0 & & D_{99} \end{pmatrix} \begin{pmatrix} v_{00} & \cdots & v_{0,99} \\ \vdots & \ddots & \vdots \\ v_{99,0} & \cdots & v_{99,99} \end{pmatrix}^{\dagger}$$



Delete rows/columns corresponding to 50 smallest singular values.

How can tensornetworks help us in quantum simulation

$$\begin{pmatrix} u_{00} & \cdots & u_{0,99} \\ \vdots & \ddots & \vdots \\ u_{99,0} & \cdots & u_{99,99} \end{pmatrix} \begin{pmatrix} D_0 & & 0 \\ & \ddots & \\ 0 & & D_{99} \end{pmatrix} \begin{pmatrix} v_{00} & \cdots & v_{0,99} \\ \vdots & \ddots & \vdots \\ v_{99,0} & \cdots & v_{99,99} \end{pmatrix}^\dagger$$



Delete rows/columns corresponding to 90 smallest singular values.

How can tensornetworks help us in quantum simulation

$$\begin{pmatrix} u_{00} & \cdots & u_{0,99} \\ \vdots & \ddots & \vdots \\ u_{99,0} & \cdots & u_{99,99} \end{pmatrix} \begin{pmatrix} D_0 & & 0 \\ & \ddots & \\ 0 & & D_{99} \end{pmatrix} \begin{pmatrix} v_{00} & \cdots & v_{0,99} \\ \vdots & \ddots & \vdots \\ v_{99,0} & \cdots & v_{99,99} \end{pmatrix}^\dagger$$

Original image

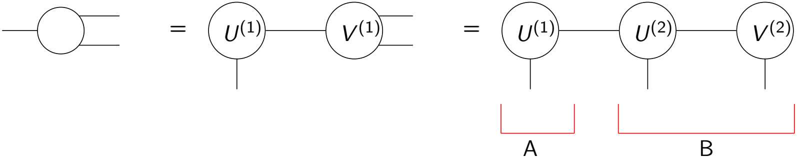
Compressed (95 smallest singular values remain)

Delete rows/columns corresponding to 5 largest singular values.

Outline

Entanglement and maximal bond dimension

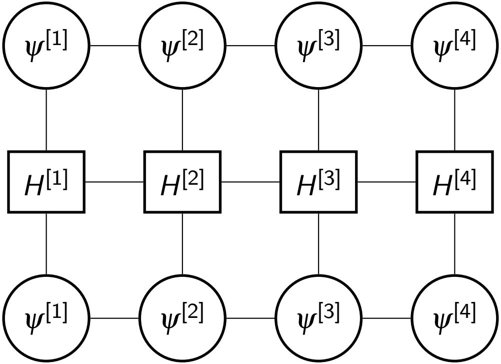
$$\psi_{s_1,(s_2,s_3)} = U^{(1)}_{s_1,\alpha_1} V^{(1)}_{\alpha_1,s_2,s_3} = U^{(1)}_{s_1,\alpha_1} U^{(2)}_{\alpha_1,s_2,\alpha_2} V^{(2)}_{\alpha_2,s_3}$$



- **Schmidt-Decomposition** $|\psi\rangle = \sum T_{\vec{\sigma}} |\vec{\sigma}\rangle = \sum_{i=1}^{r} s_i |\vec{\delta}_i\rangle |\vec{\gamma}_i\rangle$
- **Entanglement entropy** $S_{A/B} = -\sum_{i=1}^{r} s_i^2 \ln(s_i^2)$, maximal for $s_i = r^{-\frac{1}{2}} \forall i \in [r]$ [Del; Nie]

Outline

Expectation in tensornetwork-backends



Outline

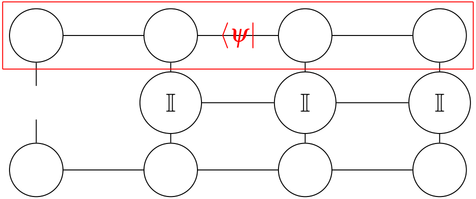
Sampling in tensornetwork-backends

- Probability for projective measurements for measuring an eigenvalue m associated to a projector P_m: $p(m) = \langle\psi|P_m|\psi\rangle$
- Chain rule of probability

$$p(s_1, s_2, s_3, \ldots, s_N) = p(s_1)\, p(s_2 \mid s_1) \cdots p(s_N \mid s_1, s_2, \ldots, s_{N-1}).$$

- Step 1)

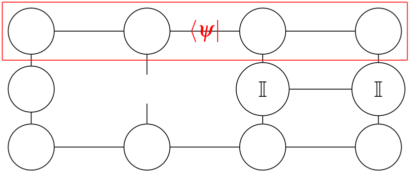
$$p(s_1) = \langle\psi|s_1\rangle\langle s_1|\bigotimes^{n-1}\mathbb{I}|\psi\rangle =$$



Sampling in tensornetwork-backends

- Step 2) Sample s_1 according to the probability distribution function obtained in Step 1)
- Step 3)

$$p(s_2|s_1) = \langle\psi|s_1\rangle\langle s_1|\otimes|s_2\rangle\langle s_2|\bigotimes^{n-1}\mathbb{I}|\psi\rangle =$$



Sampling in tensornetwork-backends

- **Take-home message:** For every shot, we need to do some tensornetwork contractions.
- This scaling with shot-number makes tensornet simulations of shots consistently underperform in a scope where statevector simulators are still viable [SMP25].

Literature I

[Del] J. von Delft. *Tensor Networks for Many-Body Physics 2023*. Lecture Notes.

[Nie] M. Nielsen. *Quantum Computing and Quantum Information*.

[SMP25] G. Schieffer, S. Markidis **and** I. Peng. *Harnessing CUDA-Q's MPS for Tensor Network Simulations of Large-Scale Quantum Circuits*. 27 **january** 2025. arXiv: 2501.15939 [quant-ph]. URL: http://arxiv.org/abs/2501.15939 (**urlseen** 09/01/2026). prepublished.