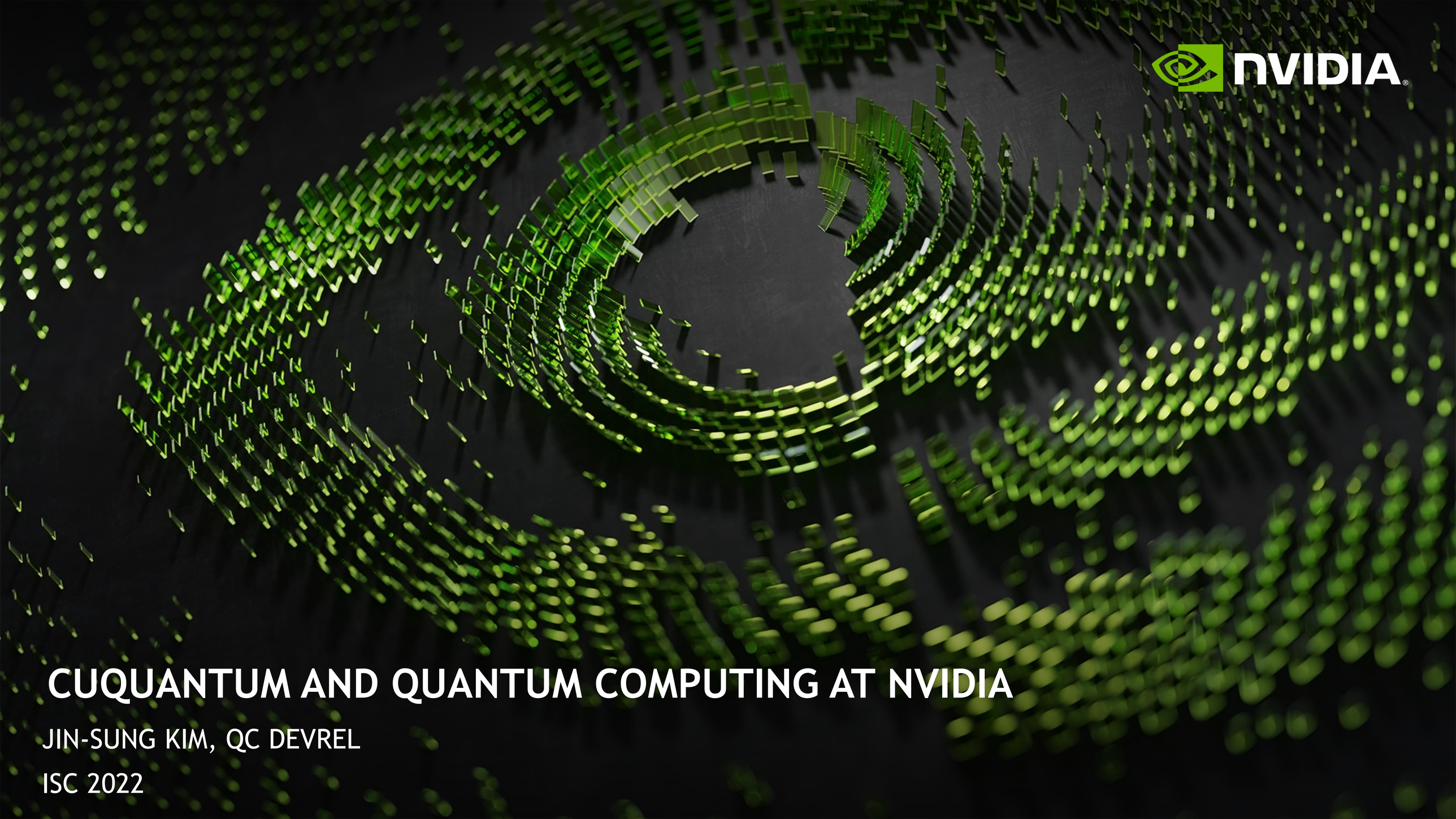
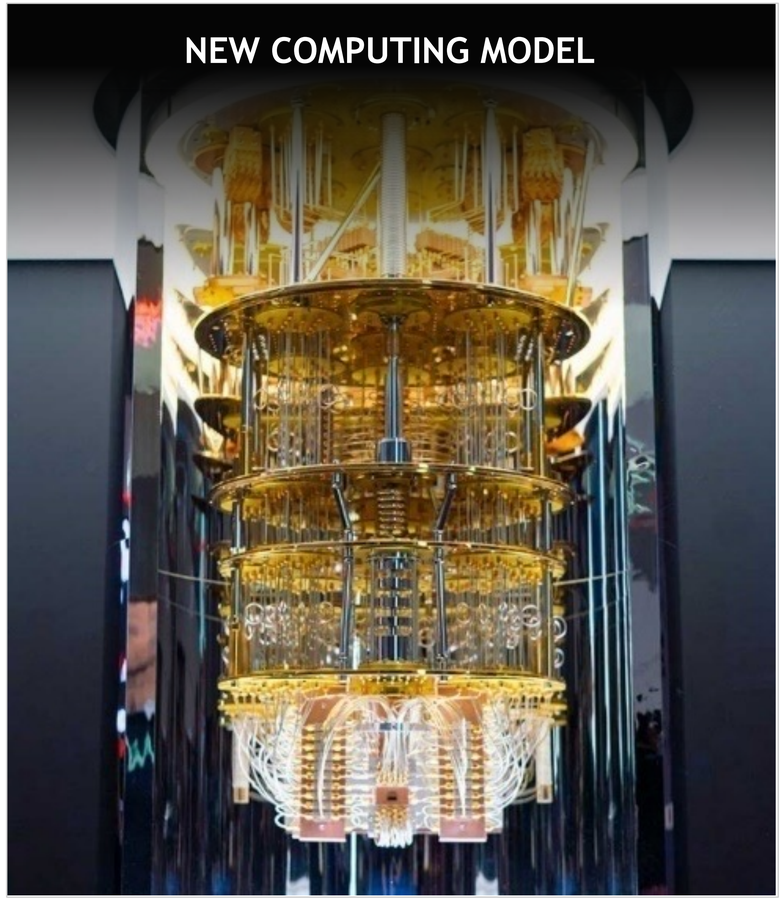


CUQUANTUM AND QUANTUM COMPUTING AT NVIDIA

JIN-SUNG KIM, QC DEVREL

ISC 2022

A NEW COMPUTING MODEL — QUANTUM



POTENTIAL USE CASES

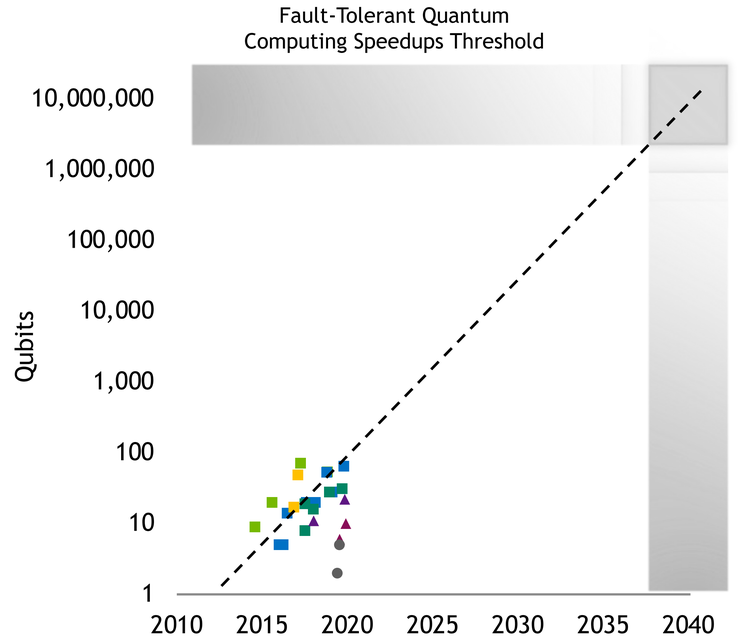
Computational Finance

Quantum Chemistry

Cryptography

Optimization

REQUIRES QUBITS SCALE TO DOUBLE EVERY YEAR

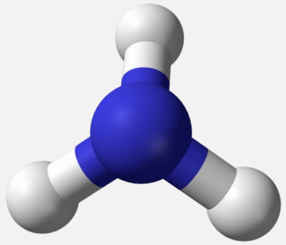


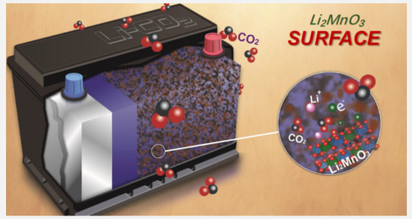
NEAR TERM APPLICATION POTENTIAL

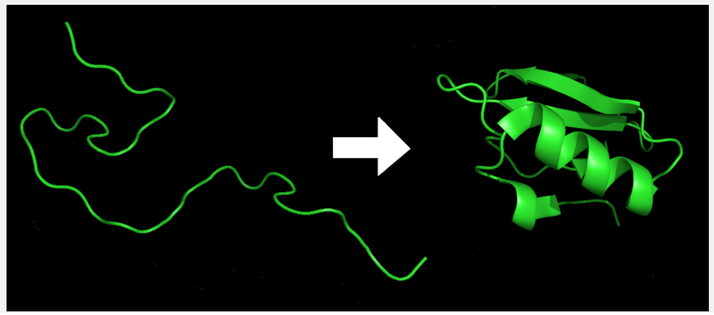
Applications with near term potential but quantum advantage is open question

CHEMISTRY, MATERIALS SCIENCE, DRUG DISCOVERY

- Ground state energy calculations
- Protein folding
- Variational Quantum Eigensolver

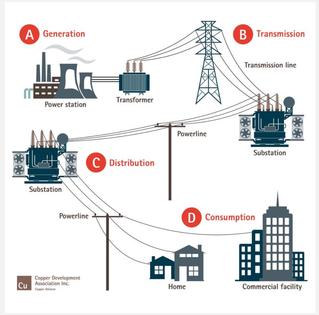


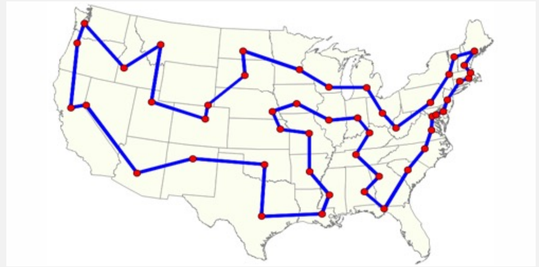




FINANCE, LOGISTICS, OPTIMIZATIONS

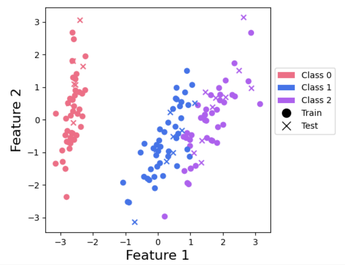
- Portfolio asset optimization
- Energy grid optimization
- Supply chain optimization

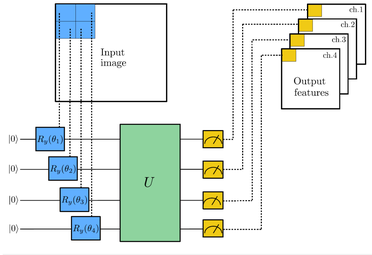


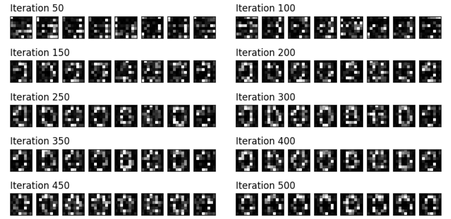


MACHINE LEARNING

- Classification problems
- Sampling problems







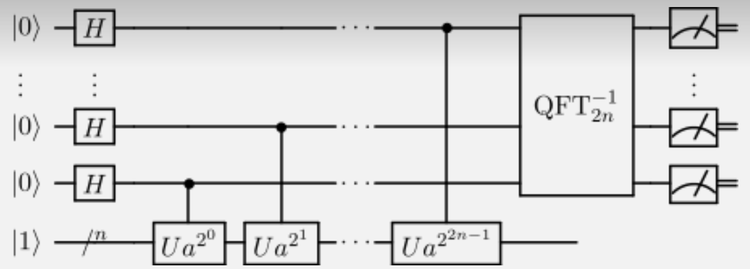
Images from PennyLane, Wikipedia, Lawrence Berkeley Lab, American Physical Society, Copper Development Association

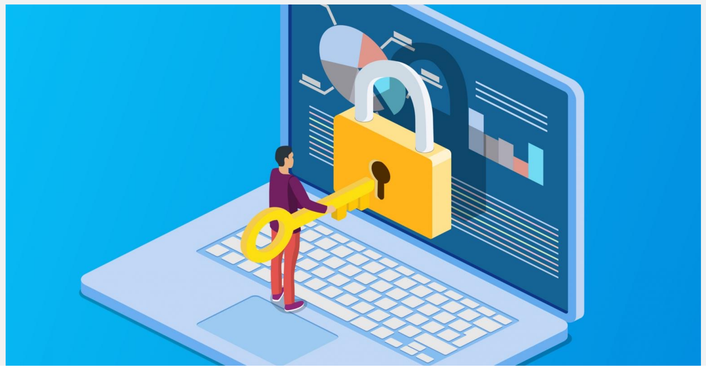
FAR TERM APPLICATIONS

Rigorous proofs of advantage
Many “perfect” qubits required however

SHOR’S ALGORITHM

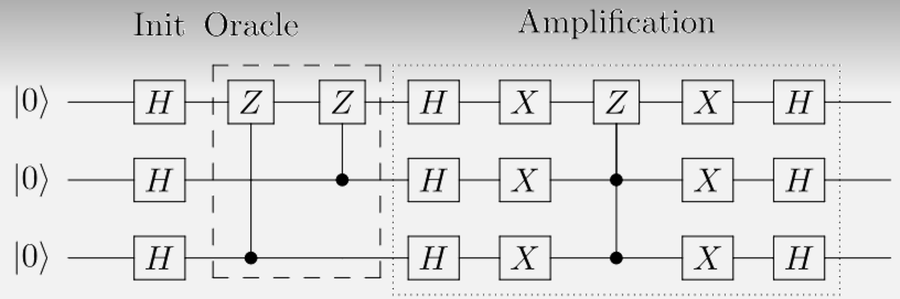
- Prime factorization of numbers - encryption
- Exponential speed-up

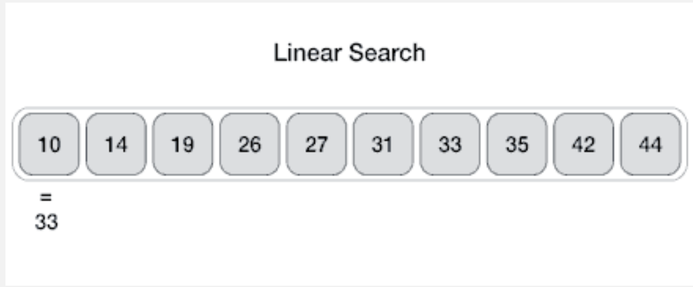




GROVER’S ALGORITHM

- Unstructured search
- Quadratic speed-up



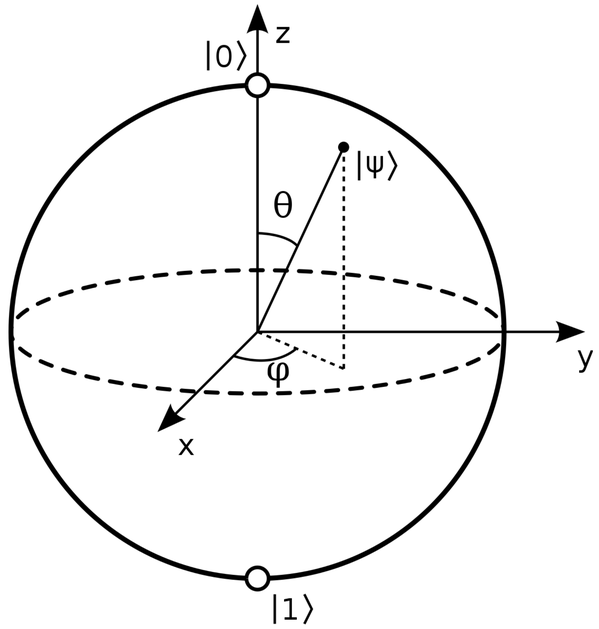


QUANTUM COMPUTING BASICS OPERATIONS

Superposition and Measurement

Bit

Qubit



$|\Psi\rangle = a|0\rangle + b|1\rangle = \begin{bmatrix} a \\ b \end{bmatrix}$

Measurement: wavefunction collapse

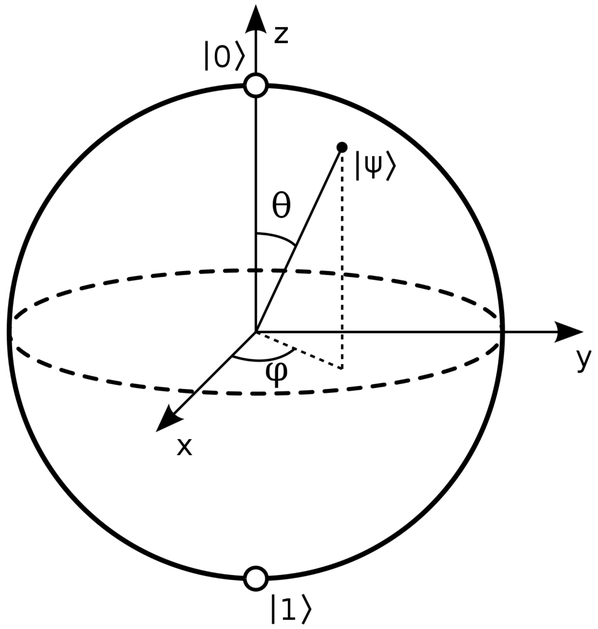
- measure only one state

$P_0 = |a|^2$

$P_1 = |b|^2$

QUANTUM COMPUTING BASICS OPERATIONS

Superposition and Measurement



0

0	1

1	0	1

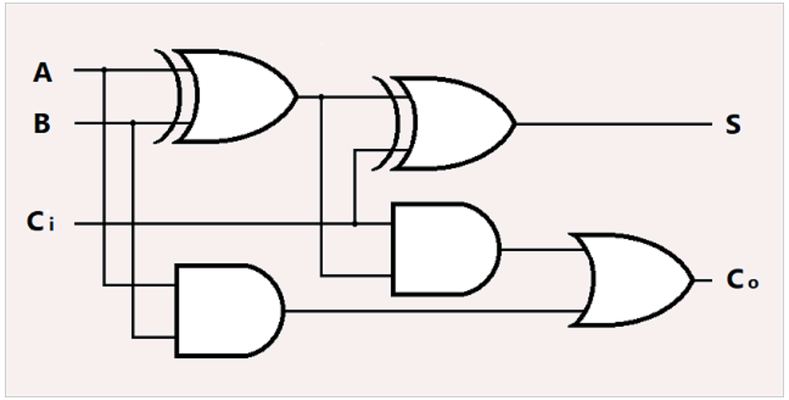
$$c_0|0\rangle + c_1|1\rangle = \begin{bmatrix} c_0 \\ c_1 \end{bmatrix}$$

$$c_{00}|00\rangle + c_{01}|01\rangle + c_{10}|10\rangle + c_{11}|11\rangle = \begin{bmatrix} c_{00} \\ c_{01} \\ c_{10} \\ c_{11} \end{bmatrix}$$

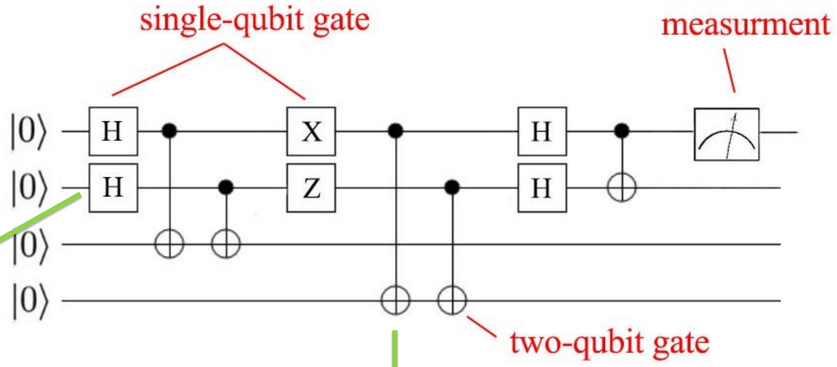
$$c_{000}|000\rangle + c_{001}|001\rangle + c_{010}|010\rangle + c_{011}|011\rangle + c_{100}|100\rangle + c_{101}|101\rangle + c_{110}|110\rangle + c_{111}|111\rangle = \begin{bmatrix} c_{000} \\ c_{001} \\ c_{010} \\ c_{011} \\ c_{100} \\ c_{101} \\ c_{110} \\ c_{111} \end{bmatrix}$$

QUANTUM CIRCUITS

Classical Circuit



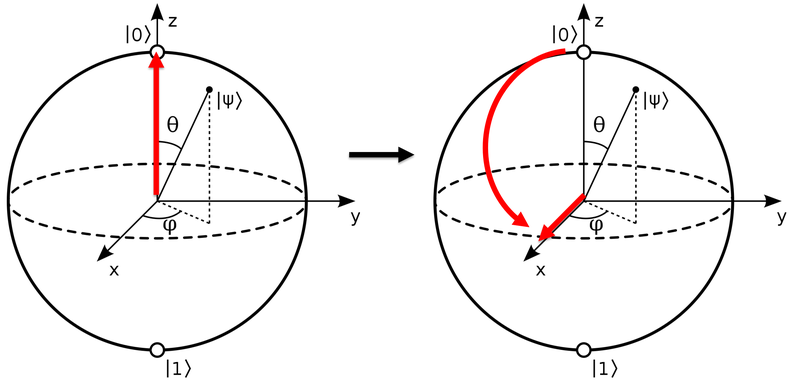
Quantum Circuit



Hadamard Gate:

Had$|0\rangle = |0\rangle + |1\rangle$

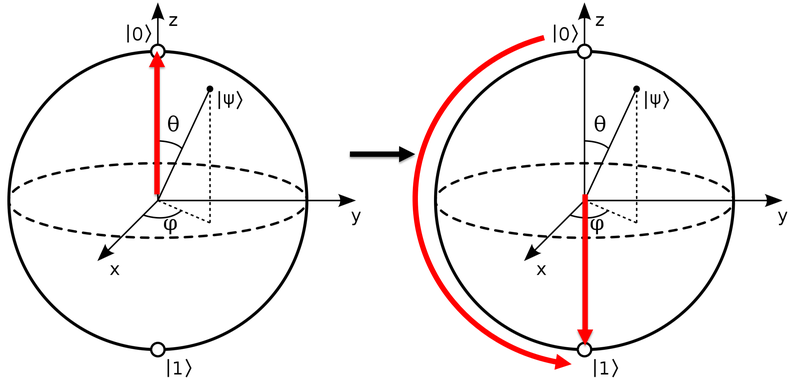
Had$|1\rangle = |0\rangle - |1\rangle$



CNOT Gate:

CNOT$|10\rangle = |11\rangle$

CNOT$|11\rangle = |10\rangle$

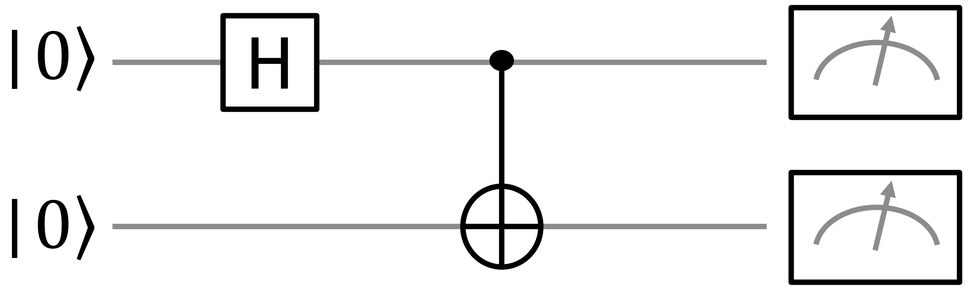


NVIDIA

QUANTUM ENTANGLEMENT

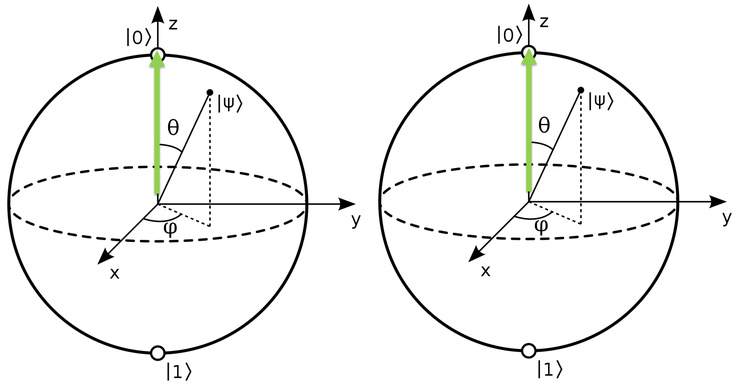
Hadamard Gate:
Had|0⟩ = |0⟩ + |1⟩
Had|1⟩ = |0⟩ - |1⟩

CNOT Gate:
CNOT|10⟩ = |11⟩
CNOT|11⟩ = |10⟩

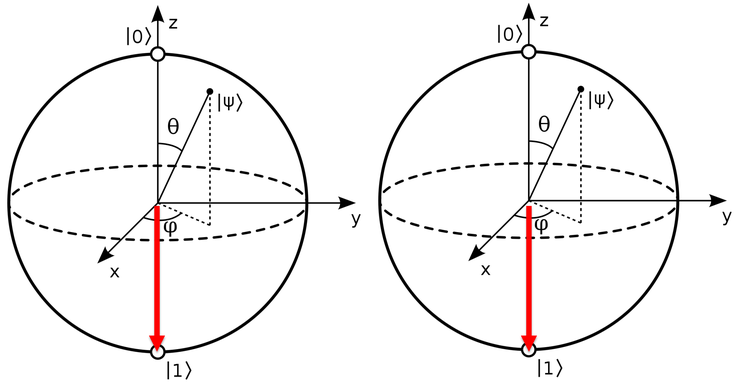


|00⟩ ➡ |00⟩ + |11⟩

|00⟩



|11⟩



LEADING QUBIT TECHNOLOGIES

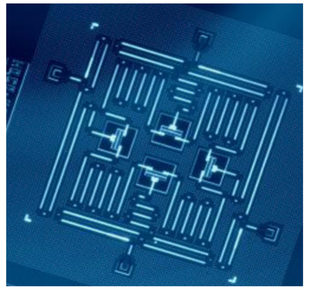
The challenge of engineering quantum hardware is to manipulate physical systems to implement superposition and entanglement

SUPERCONDUCTORS

- **Principle:** Superconducting circuits based on Josephson junctions
- **Strengths:** Gate error rates <1%
- **Weaknesses:** Qubits only hold state ~100μs, fixed connectivity, cross-talk

aws Google IBM

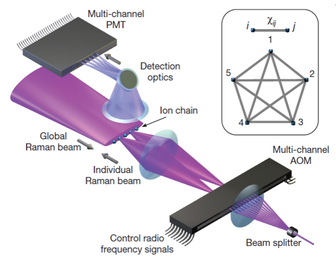
intel QuTech rigetti



ION TRAPS

- **Principle**: Ions in a vacuum, trapped & rotated by lasers
- **Strengths:** Long coherence time, all-to-all connectivity
- **Weaknesses:** Scalability, slow read-out

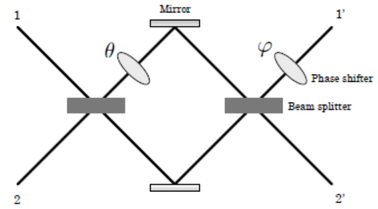
QUANTINUUM IONQ



SILICON PHOTONICS

- **Principle:** Store qubits as polarity of single photons, photonics for gates
- **Strengths:** Scalability, manufacturable
- **Weaknesses:** Photon sources/detectors, error rates, non-std computation model

PsiQuantum XANADU

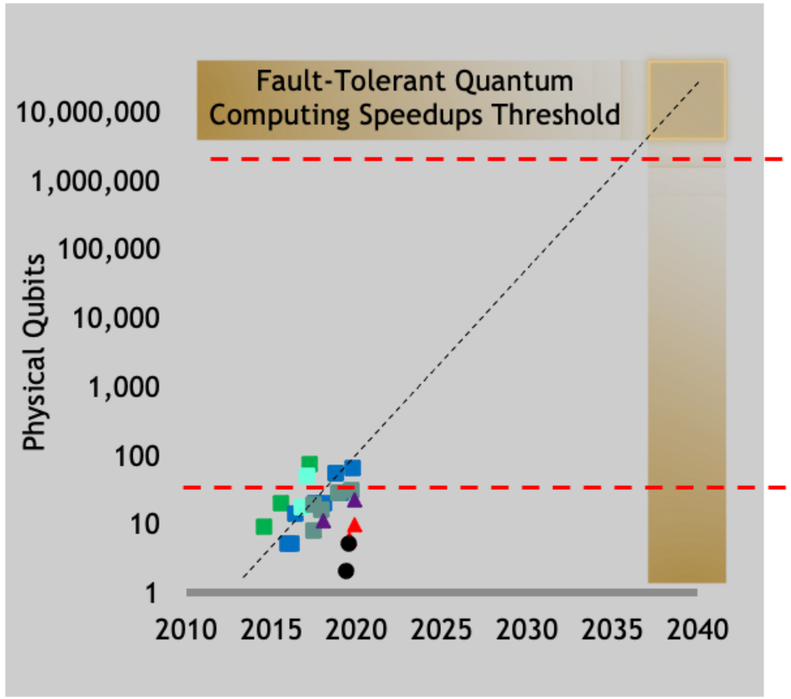


Other Approaches: Neutral Atoms, Quantum Dots, Topological Qubits, Diamond Vacancies

Practical QC is expected to require scaling these technologies to millions of qubits, error correction, and new quantum algorithms.

QC RESEARCH ROADMAP

Large improvements in qubit quantity & quality, error correction, needed for wide adoption



Fault-Tolerant QC Era:

1000:1-10000:1 redundancy for error-corrected *logical* qubits.
[Fowler 2012][Reiher 2016]

Exponential speedups on a limited set of applications with hundreds to thousands of logical qubits (millions of physical qubits).

Active Research: What are the best error correction algorithms?

Noisy Intermediate Scale Quantum (NISQ) Era:

Quantum gates are noisy, errors accumulate. Qubits lose coherence.

QC hardware will mitigate errors by using tens to hundreds of redundant physical qubits per logical qubit to mitigate errors.

Active Research: Will NISQs have quantum advantage on useful workloads?

Quantum Supremacy Threshold: Experimental confirmation of quantum speedup on a well-defined (not necessarily *useful*) problem.

Qubits and quantum gates are very noisy, hardware not very usable.

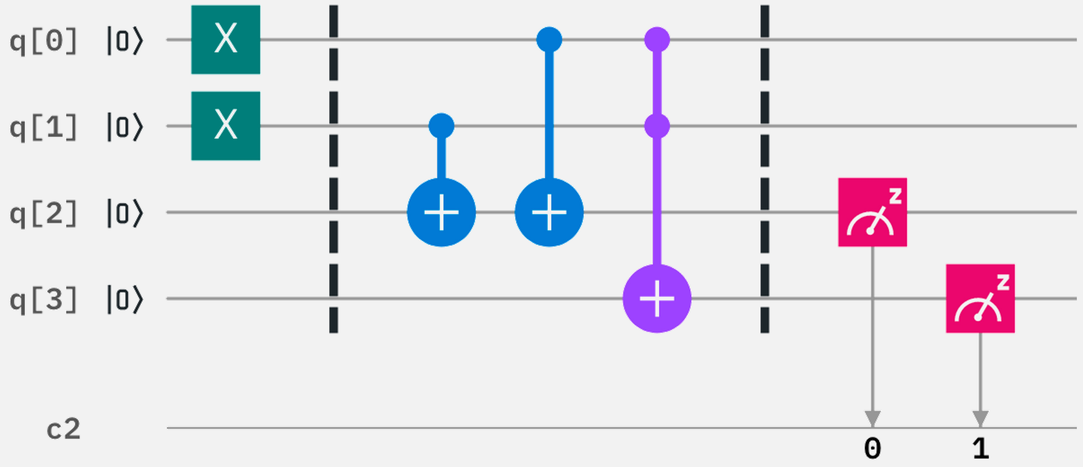
Active Research: Can this be simulated efficiently on GPU supercomputers?

GPU-BASED SUPERCOMPUTING IN THE QC ECOSYSTEM

Researching the Quantum Computers of Tomorrow with the Supercomputers of Today

QUANTUM CIRCUIT SIMULATION

Critical tool for answering today's most pressing questions in Quantum Information Science (QIS):

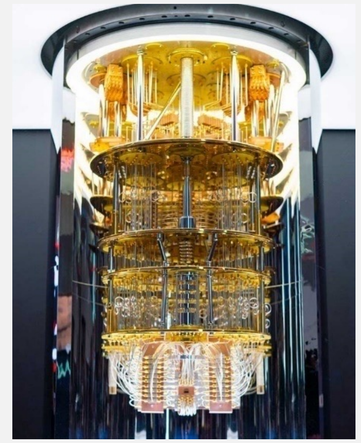


- What quantum algorithms are most promising for near-term or long-term quantum advantage?
- What are the requirements (number of qubits and error rates) to realize quantum advantage?
- What quantum processor architectures are best suited to realize valuable quantum applications?

HYBRID CLASSICAL/QUANTUM APPLICATIONS

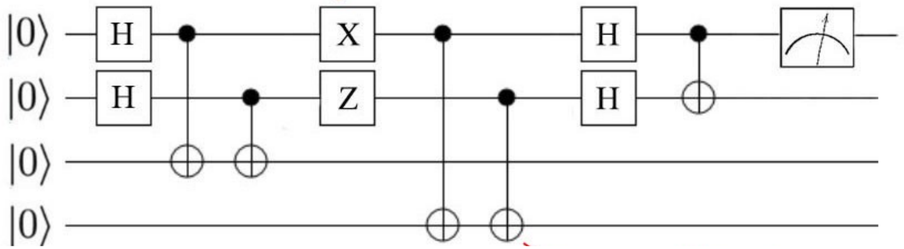
Impactful QC applications (e.g. simulating quantum materials and systems) will require classical supercomputers with quantum co-processors

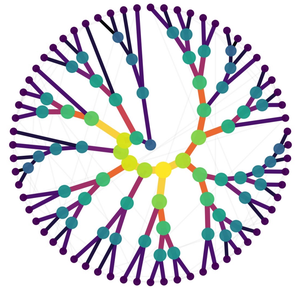
+



- How can we integrate and take advantage of classical HPC to accelerate hybrid classical/quantum workloads

TWO LEADING QUANTUM CIRCUIT SIMULATION APPROACHES





State vector simulation

“Gate-based emulation of a quantum computer”

- Maintain full 2^n qubit vector state in memory
- Update all states every timestep, probabilistically sample n of the states for measurement

Memory capacity & time grow exponentially w/ # of qubits - practical limit around 50 qubits on a supercomputer

Can model either ideal or noisy qubits

Tensor networks

“Only simulate the states you need”

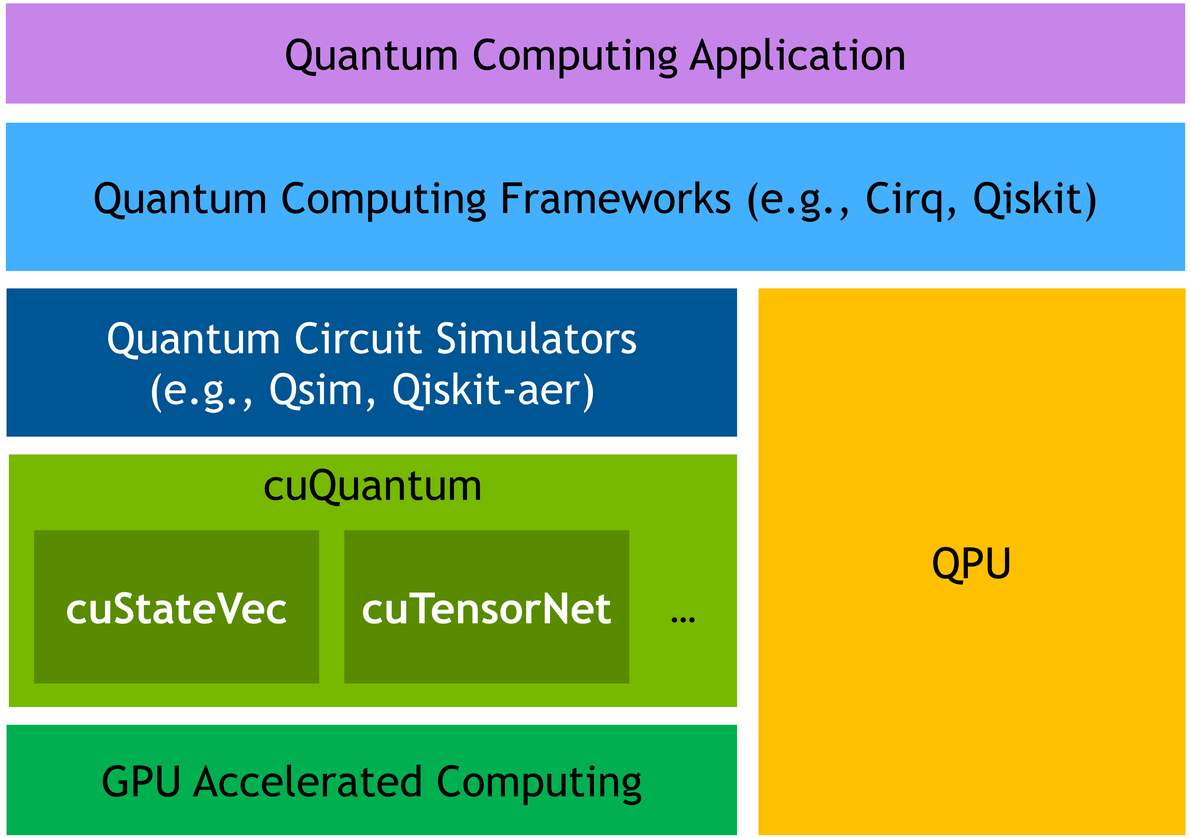
- Uses tensor network contractions to dramatically reduce memory for simulating circuits
- Can simulate 100s or 1000s of qubits for many practical quantum circuits

GPUs are a great fit for either approach

Tensor Network image from Quimb: https://quimb.readthedocs.io/en/latest/index.html

Introducing cuQuantum

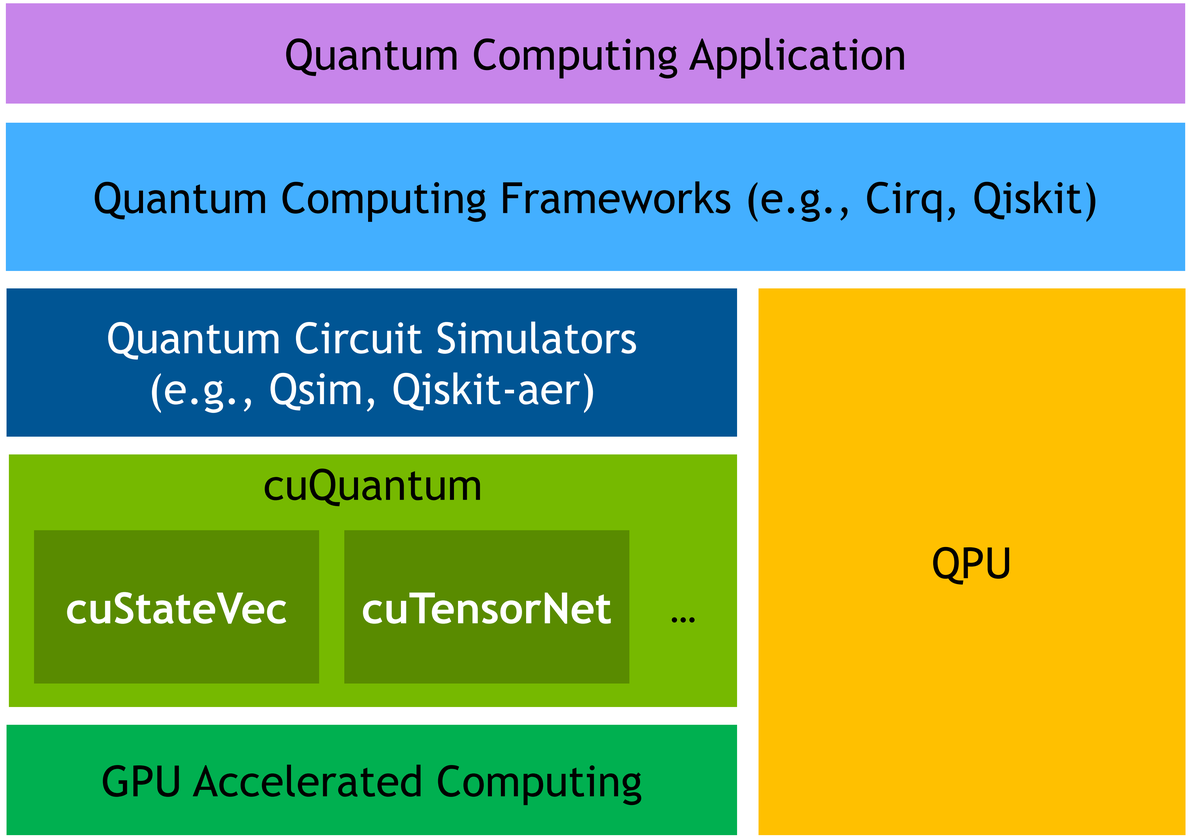
- cuQuantum is an SDK of optimized libraries and tools for accelerating quantum computing workflows
- cuQuantum is not a:
  - Quantum Computer
  - Quantum Computing Framework
  - Quantum Circuit Simulator



Introducing cuQuantum

- cuQuantum is a platform for quantum computing research
  - Accelerate Quantum Circuit Simulators on GPUs
  - Simulate ideal or noisy qubits
  - Enable algorithms research with scale and performance not possible on quantum hardware, or on simulators today
- General Access available now, integrated
  - Google Cirq
  - IBM Qiskit
  - Xanadu PennyLane
- DGX Quantum Appliance now available on NGC

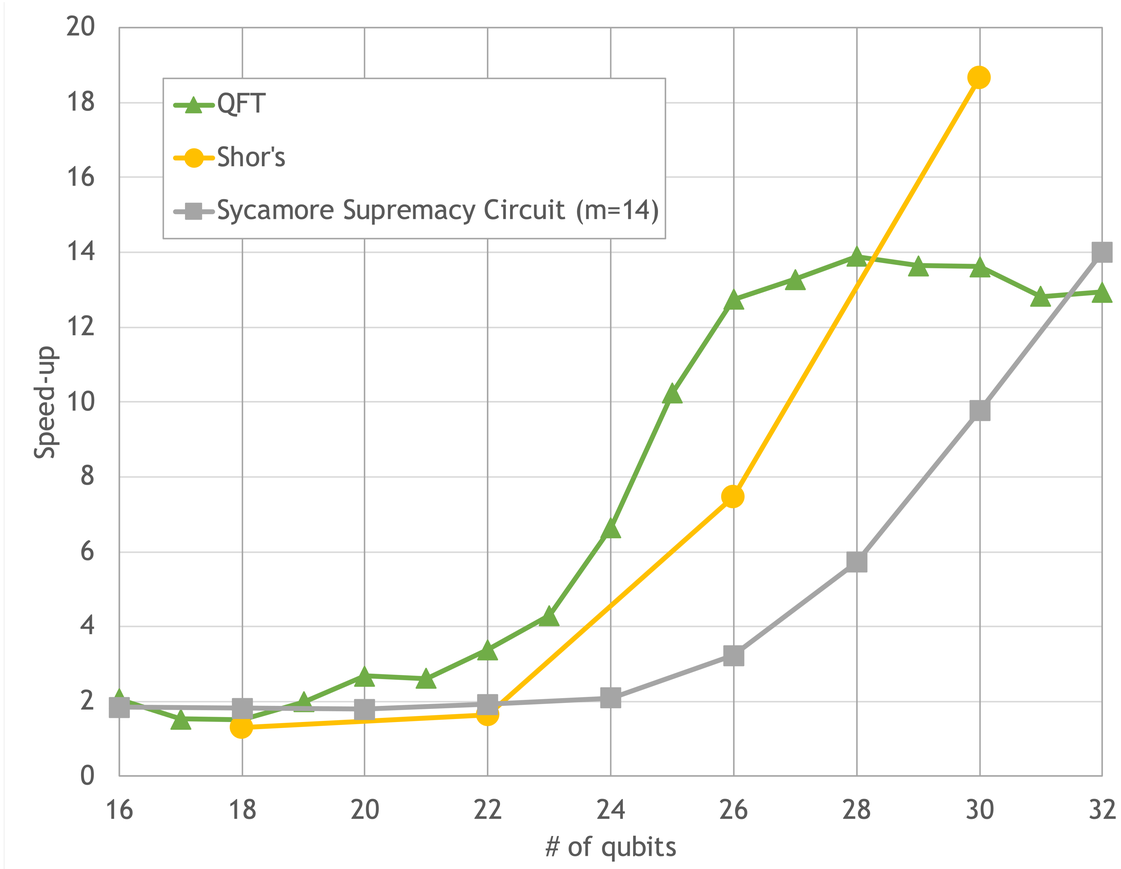
Cirq Qiskit Pennylane



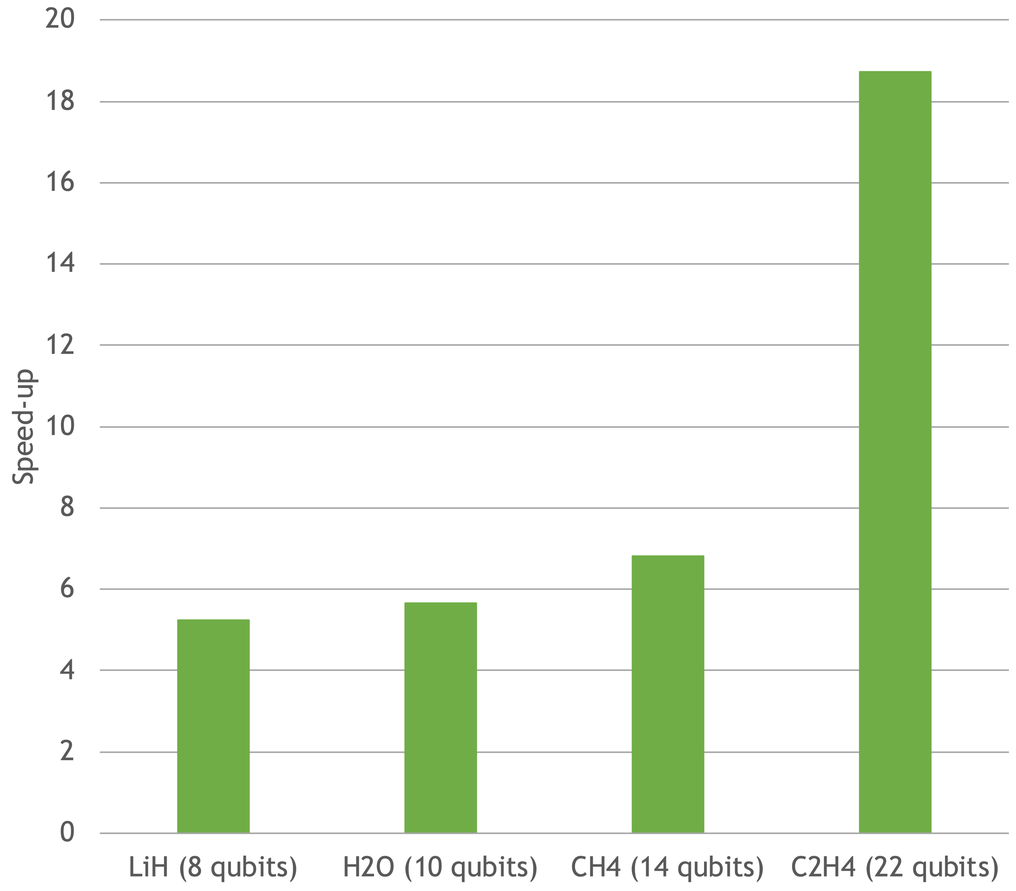
cuStateVec - SINGLE-GPU

PRELIMINARY PERFORMANCE OF Cirq/Qsim + cuStateVec ON THE A100

A100 80G vs 64 core CPU



VQE speed-up relative to single CPU



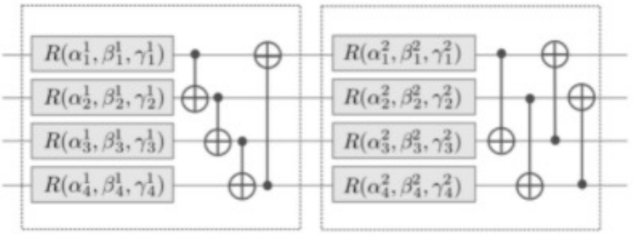
Benchmarks run using Cirq/Qsim with modifications to integrate cuStateVec
CPUs used were AMD EPYC 7742 with 64 cores
QFT circuit with 32 qubits and depth 63
Shor's circuit with 30 qubit and depth 15560 (integer factorized: 65)
Sycamore supremacy circuit m=14 with 7480 gates

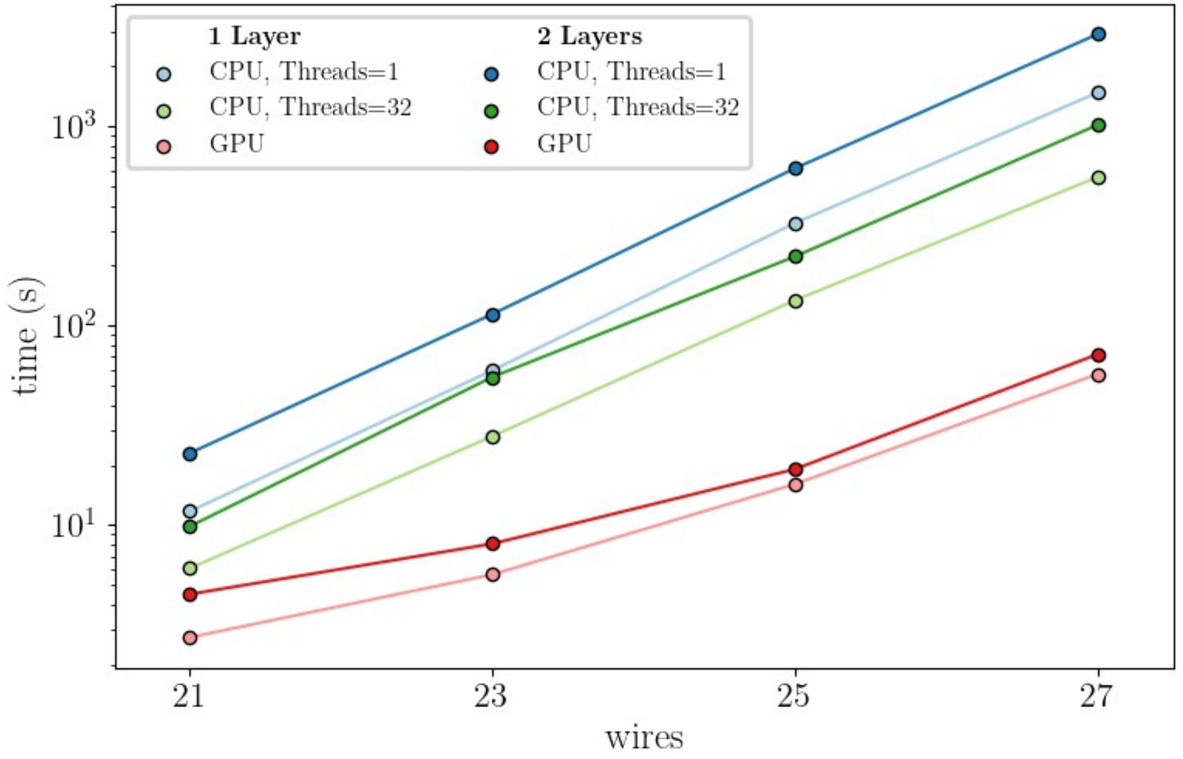
VQE benchmarks have all orbitals and results were measured for the energy function evaluation

NVIDIA

cuQuantum Support for Pennylane

- Leading open-source framework for quantum machine learning and quantum chemistry, built by Xanadu
  - Train quantum computers in the same way as neural networks
- New simulator *lightning.gpu*, with cuQuantum support, available now
- 10x speedup for QML circuits

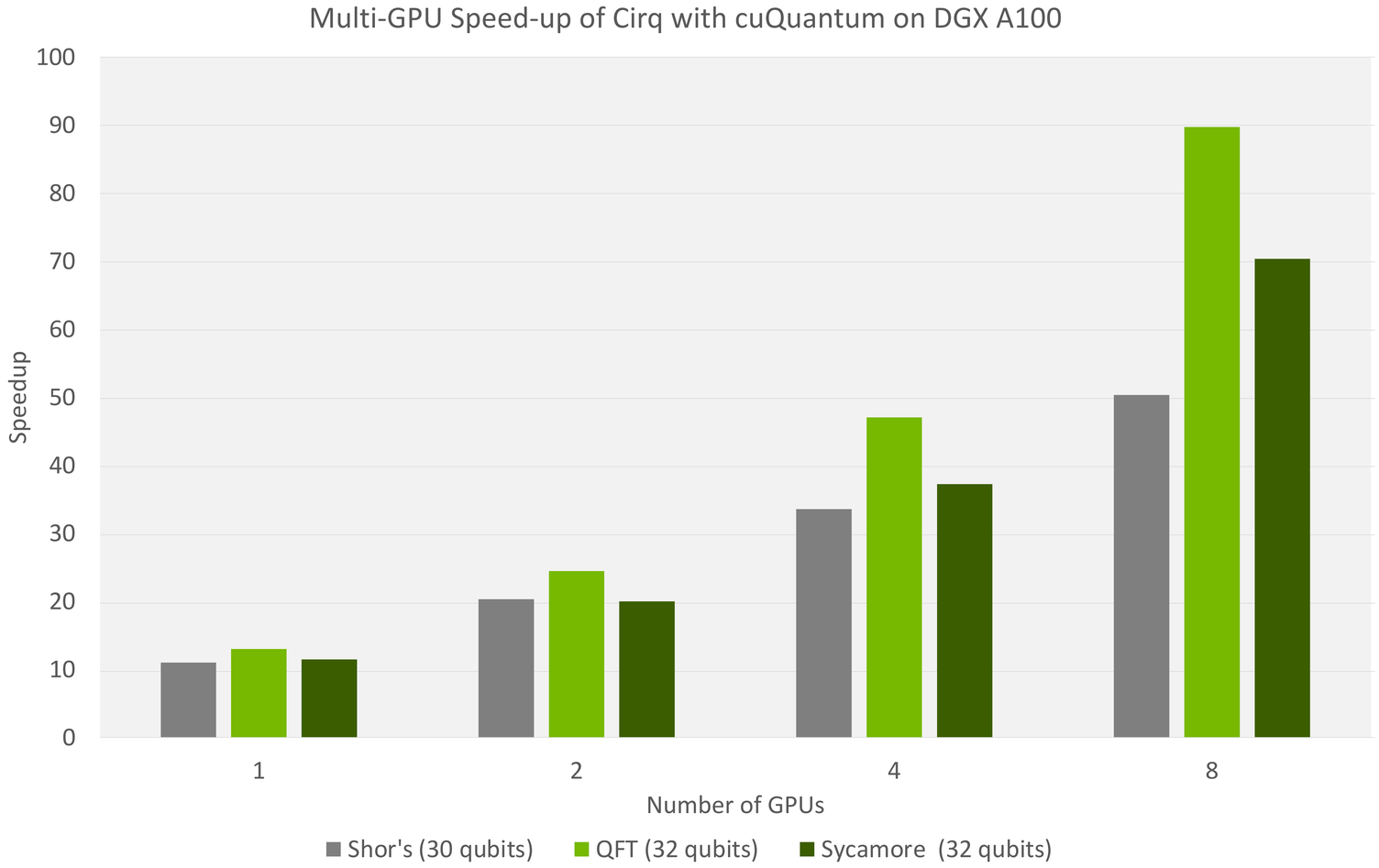




Announcing DGX Quantum Appliance

MULTI-GPU CONTAINER WITH CIRQ/QSIM/CUQUANTUM

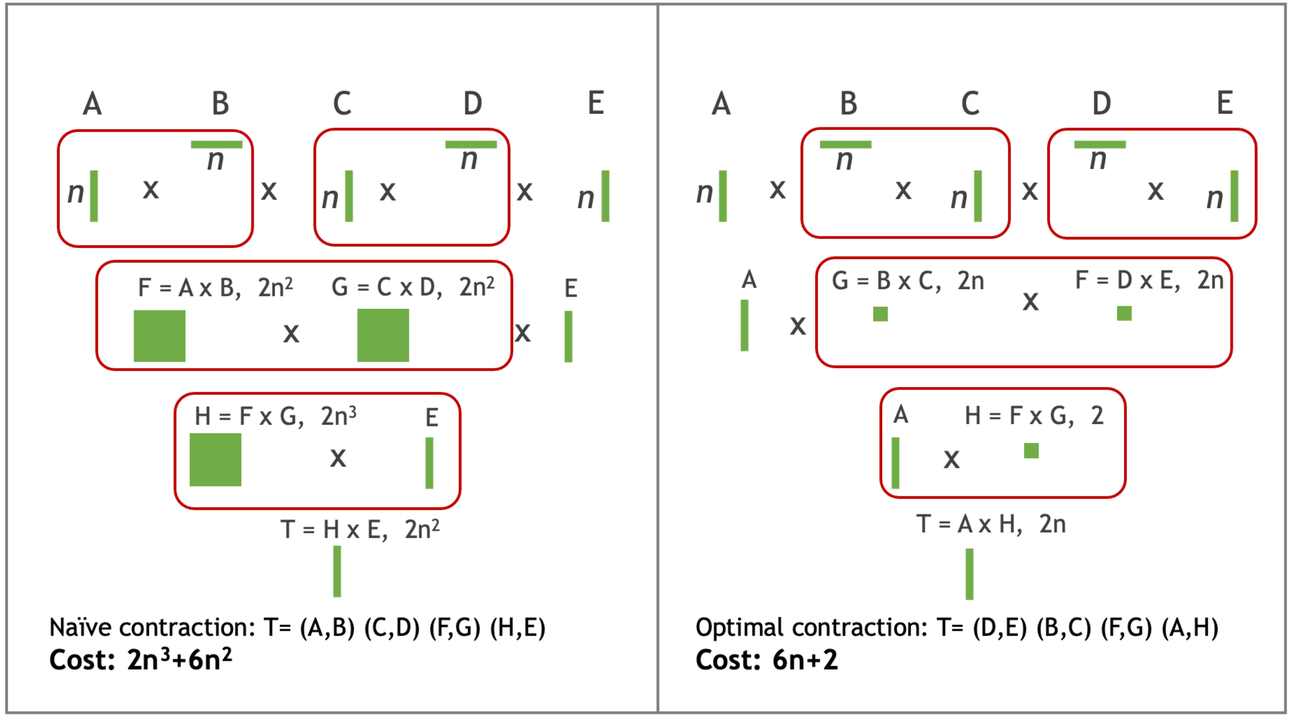
- Full Quantum Simulation Stack
- World class performance on key quantum algorithms
- Available February 2022

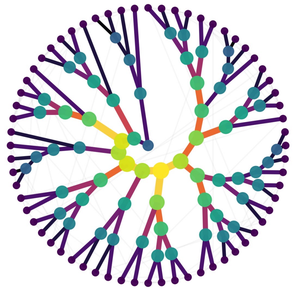


cuTensorNet

A LIBRARY TO ACCELERATE TENSOR NETWORK BASED QUANTUM CIRCUIT SIMULATION

- The cuTensorNet library initially will provide the following APIs:
  1. Given a tensor network definition calculate optimal contraction path subject to memory constraints and parallelization needs:
     - Hyper-optimization is used to find contraction path with lowest total cost (eg, FLOPS or time estimate)
     - Slicing is introduced to create parallelism or reduce maximum intermediate tensor sizes
  2. Given a contraction path for a Tensor Network calculate an optimized execution plan
     - Leverages cuTENSOR heuristics
  3. Execute the TN contraction
- cuTensorNet depends on the latest cuTENSOR library for executing all pairwise contractions for cuTENSOR

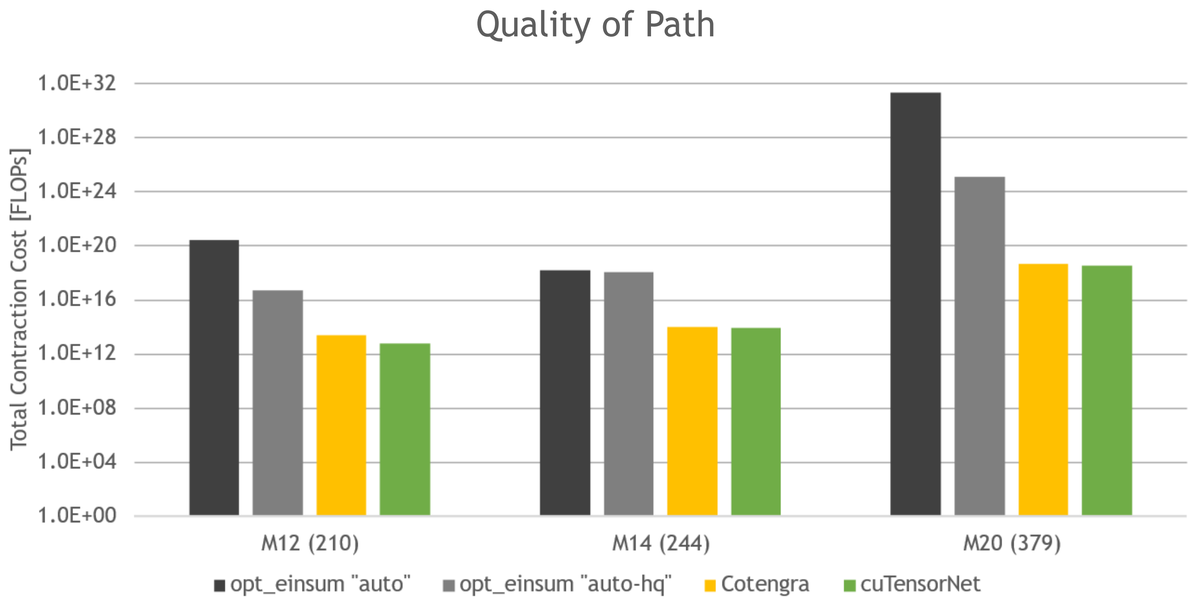


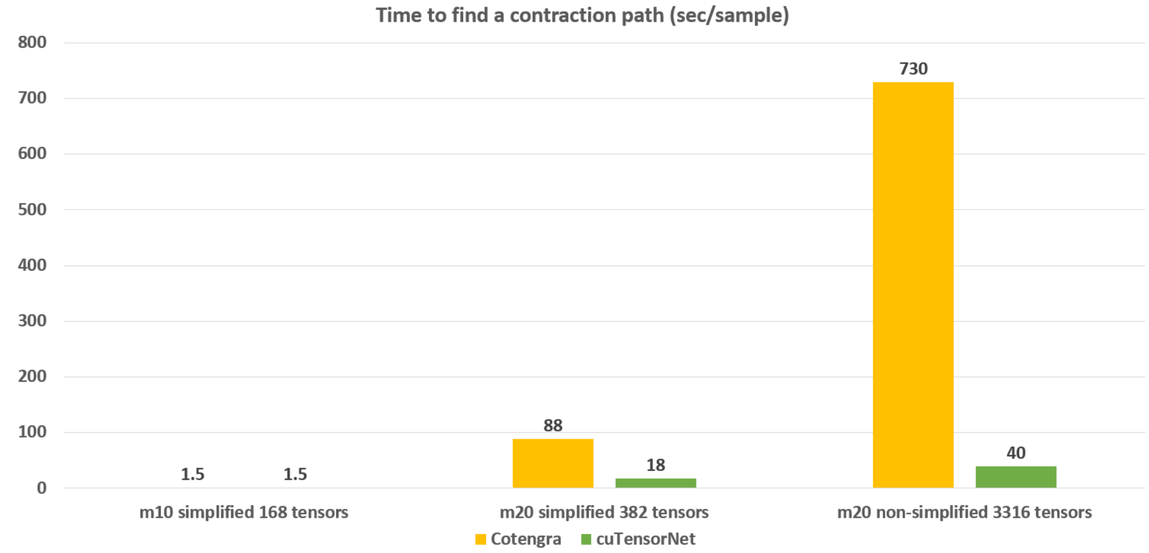


Tensor Network image from Quimb: https://quimb.readthedocs.io/en/latest/index.html

cuTensorNet

TENSOR NETWORK PATH OPTIMIZATION PERFORMANCE



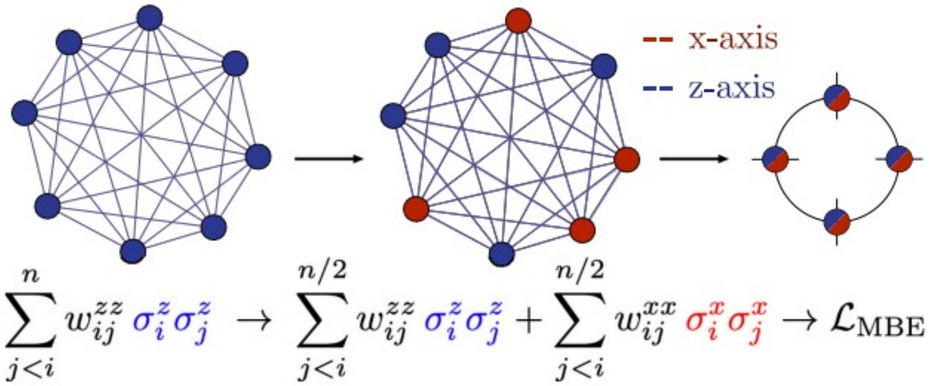


- cuTensorNet achieves SotA pathfinding results dramatically faster, and does better with more complex networks

[1] Gray & Kourtis, Hyper-optimized tensor network contraction, 2021 https://quantum-journal.org/papers/q-2021-03-15-410/pdf/
[2] opt-einsum https://pypi.org/project/opt-einsum/

Simulating MaxCut using Tensor Networks

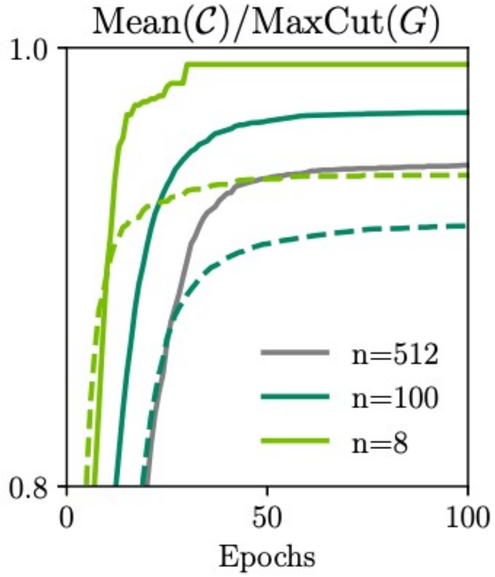
- Tensor Networks are a natural fit for MaxCut
  - Fried et. al. (2017) https://arxiv.org/abs/1709.03636
  - Huang et. al (2019) https://arxiv.org/pdf/1909.02559.pdf
  - Lykov et. al. (2020) https://arxiv.org/pdf/2012.02430.pdf



- Patti et. al.(2021): NVIDIA Research proposes a novel variational quantum algorithm
  - Based on 1D tensor ring representation
  - Multibasis encoding
  - Able to find accurate solution for 512 vertices (256 qubits) on a single GPU

  Paper: https://arxiv.org/pdf/2106.13304.pdf

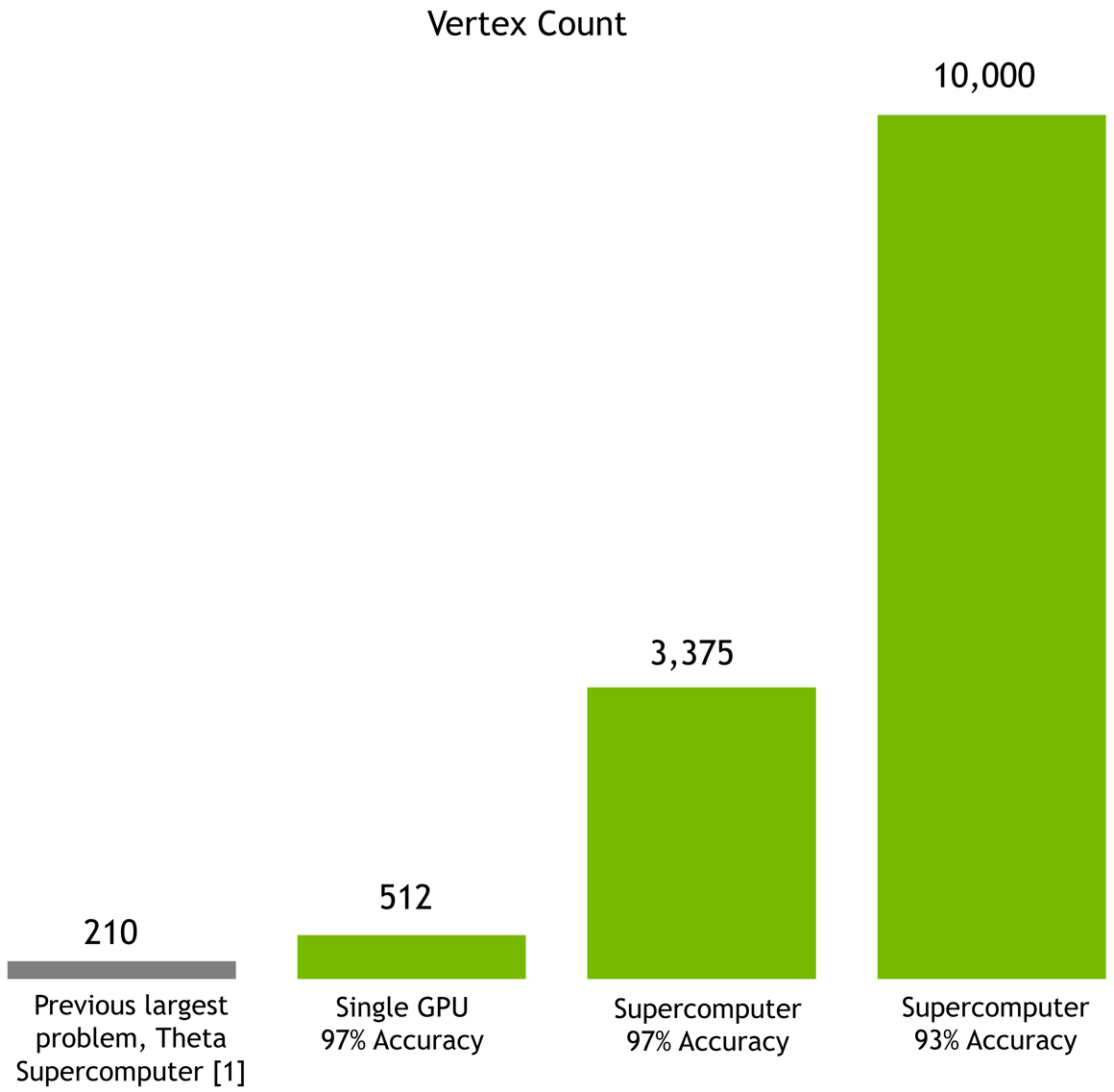
  Code: https://github.com/tensorly/quantum



Scaling to a Supercomputer

NVIDIA's Selene DGX SuperPOD based supercomputer

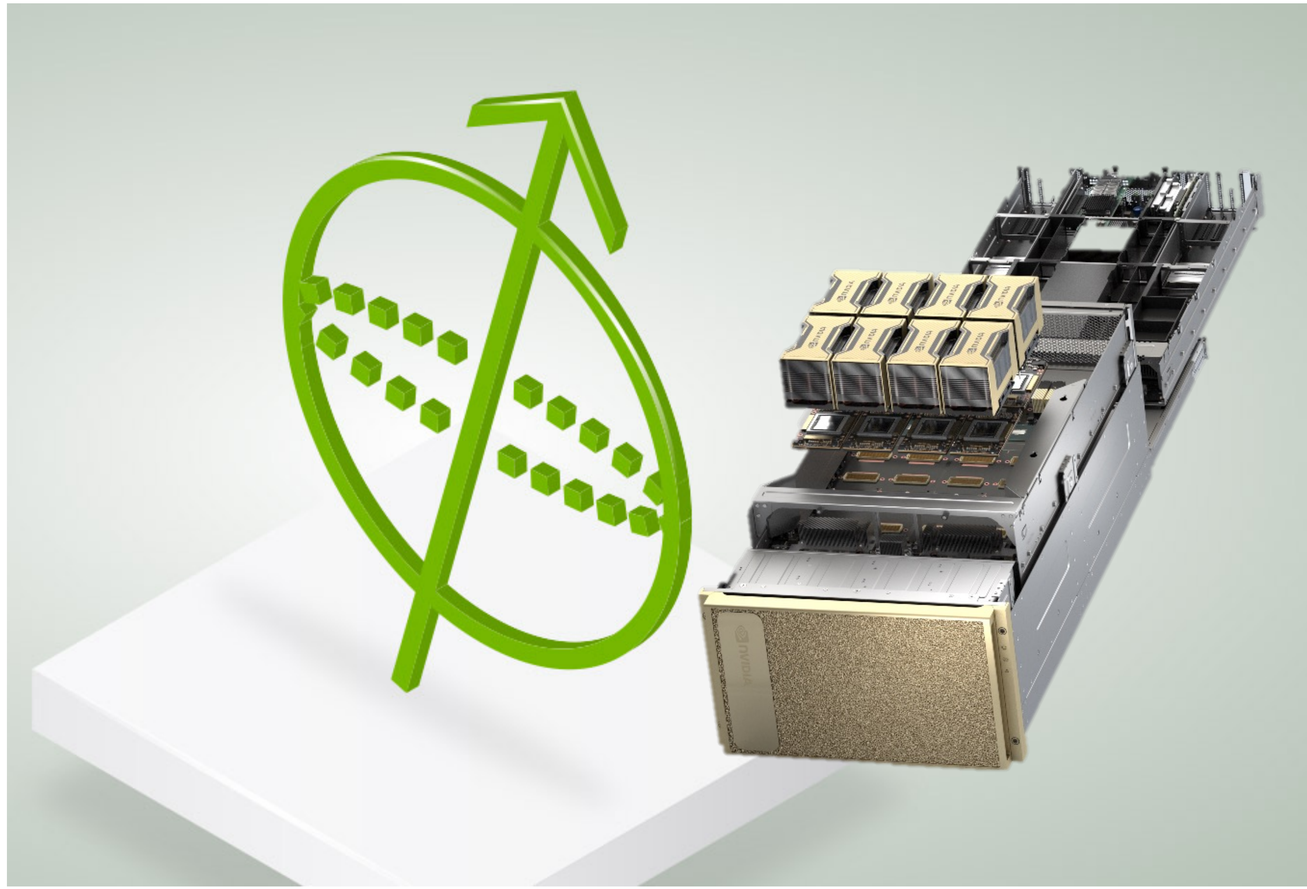
- Using NVIDIA's Selene supercomputer
- Solved a 3,375 vertex problem (1,688 qubits) with 97% accuracy
- Solved a 10,000 vertex problem (5,000 qubits) with 93% accuracy

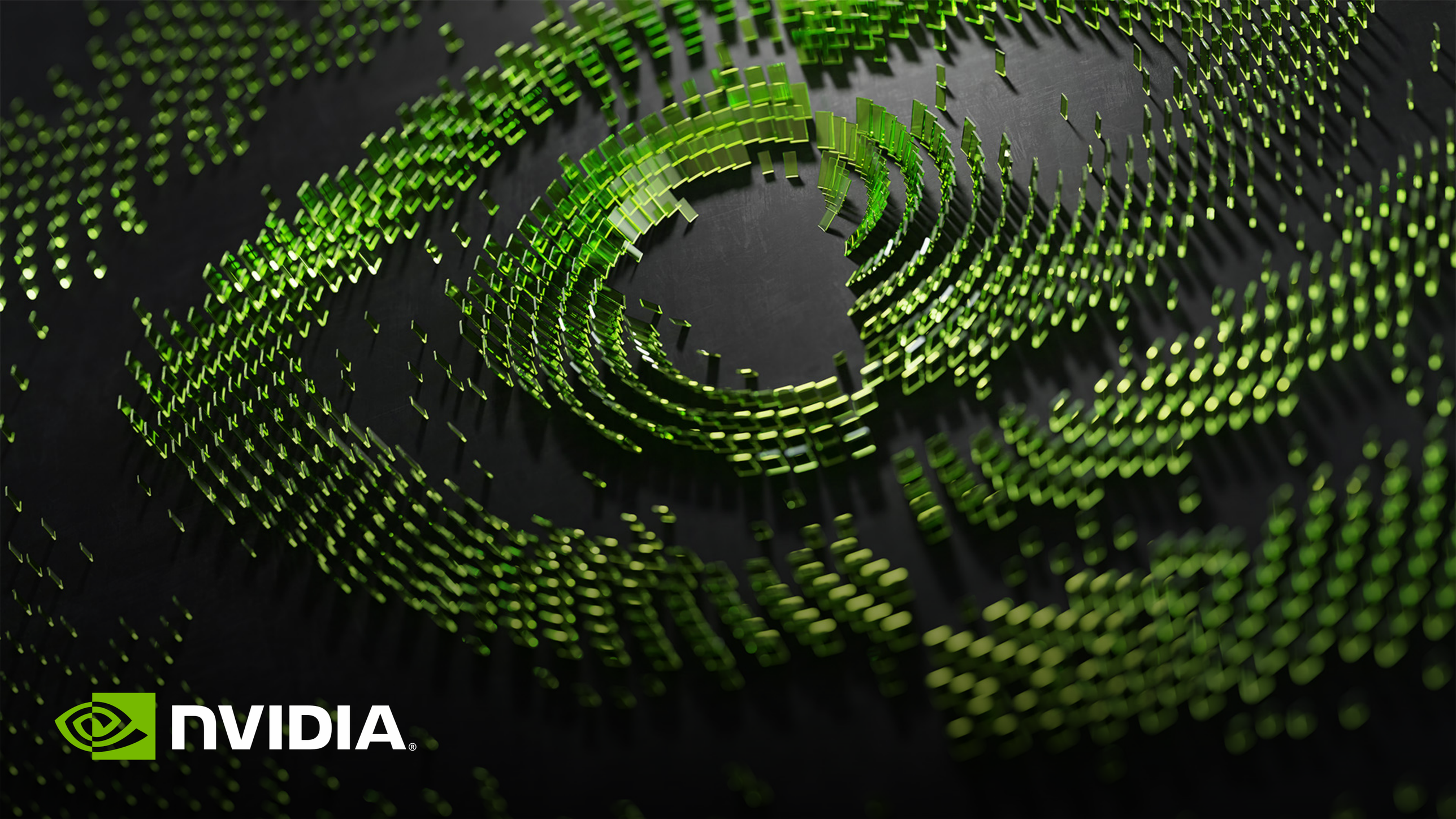


[1] Danylo Lykov et al, Tensor Network Quantum Simulator With Step-Dependent Parallelization, 2020
https://arxiv.org/pdf/2012.02430.pdf

WHAT TO EXPECT NEXT?

- General Access release of cuQuantum is now available for download at:
  - https://developer.nvidia.com/cuQuantum-downloads
- An NGC container is now available for the Cirq/Qsim Quantum Circuit Simulator accelerated with cuStateVec and optimized for multi-GPU execution on the DGX A100
- In 2022 we will scale to multiNode, continue to integrate with simulation frameworks and offer containers, and continue to optimize performance




NVIDIA