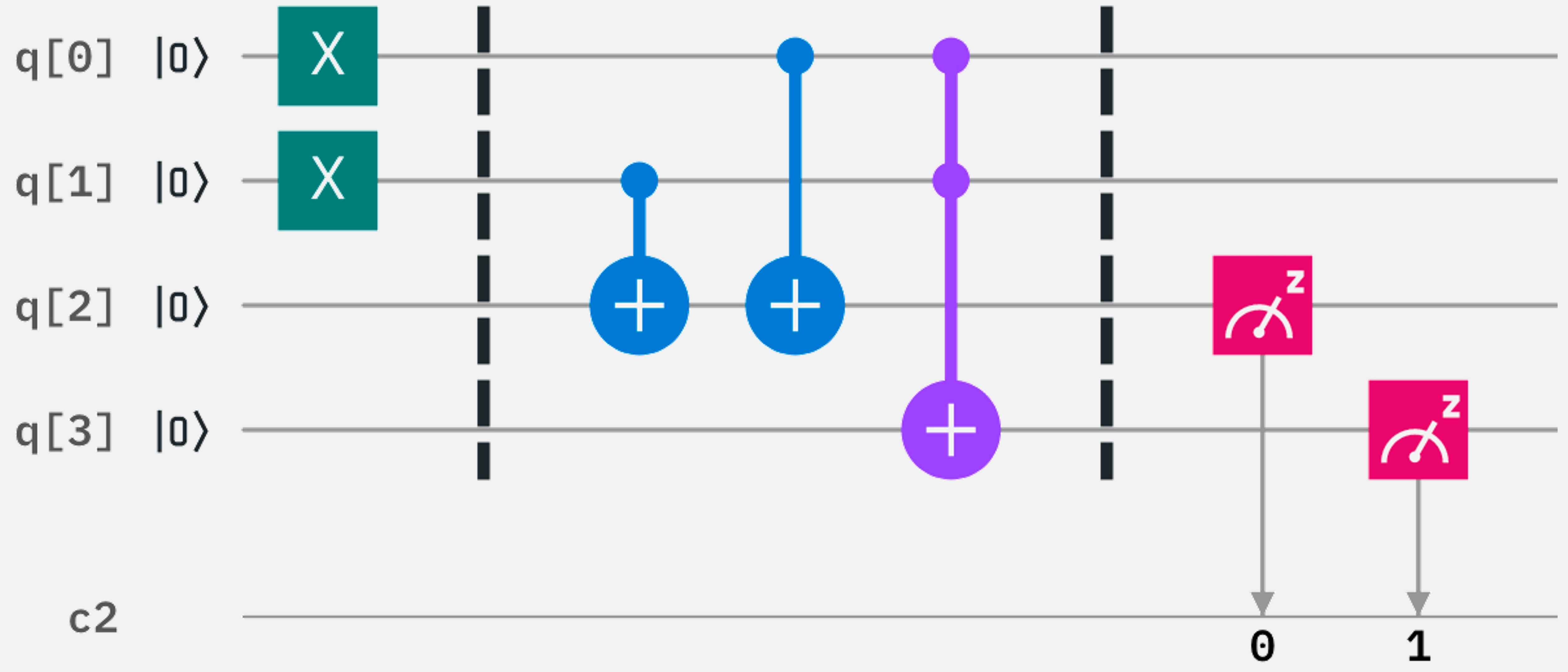
NVIDIA Quantum: cuQuantum and QODA

GPU Supercomputing and Quantum

Researching the Quantum Computers of Tomorrow with the Supercomputers of Today

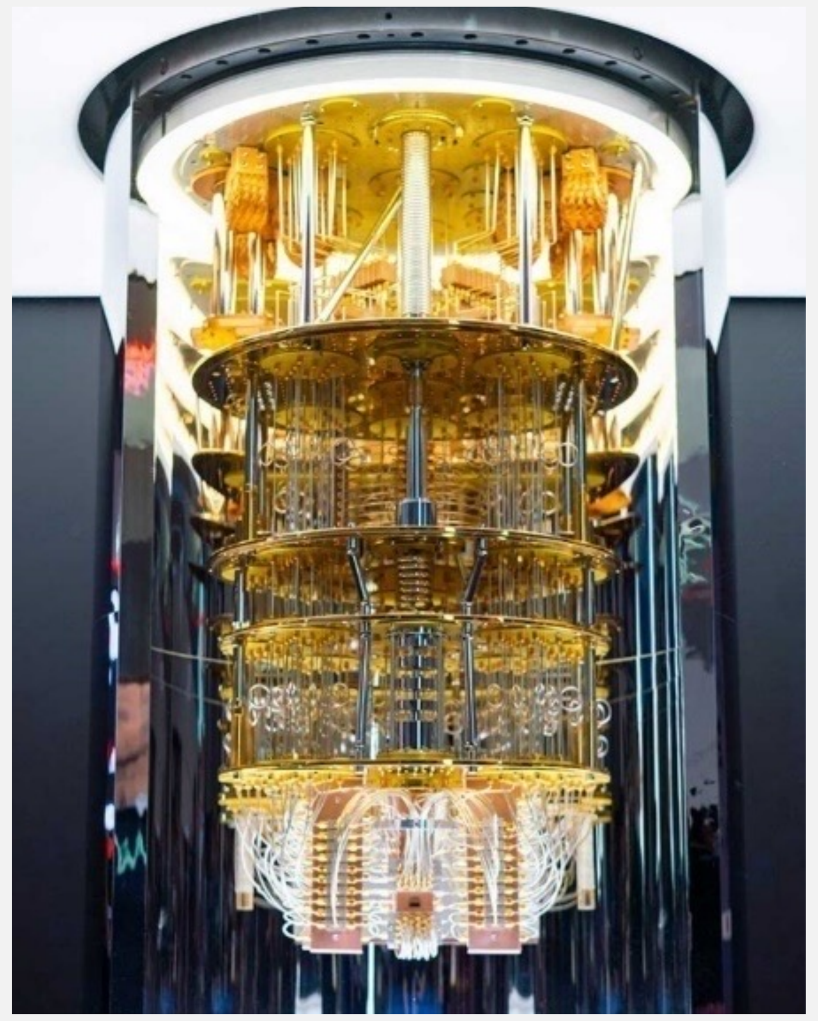
QUANTUM SIMULATION
CUQUANTUM



- Develop algorithms at scale of valuable quantum computing
- Discover use cases with quantum advantage
- Design and validate future hardware

HYBRID QUANTUM-CLASSICAL COMPUTING
QODA

+

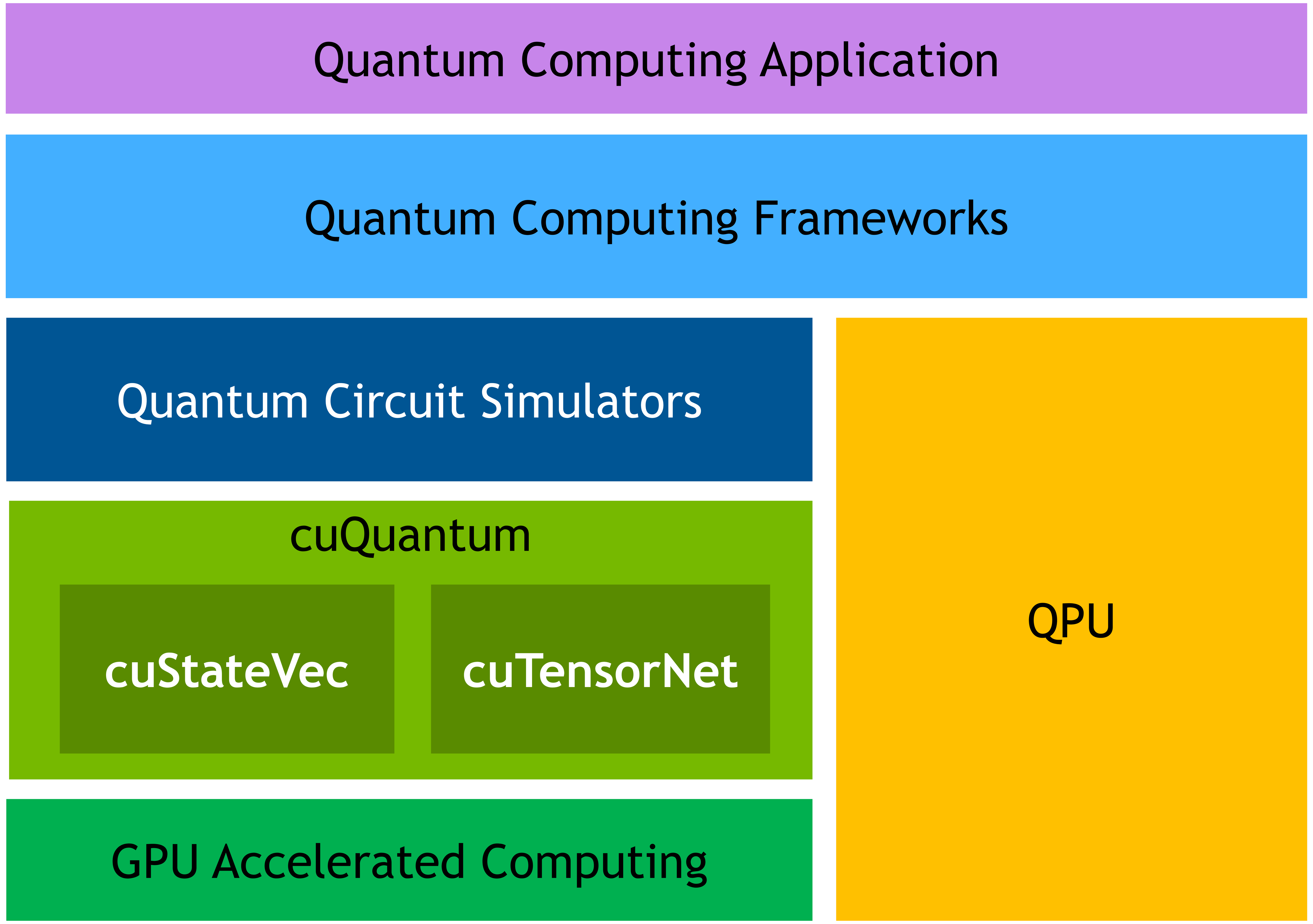


- Develop quantum applications by integrating quantum into leading accelerated applications
- Build a platform that is familiar to domain scientists
- Unparalleled performance and scientific productivity using the best resource for the task

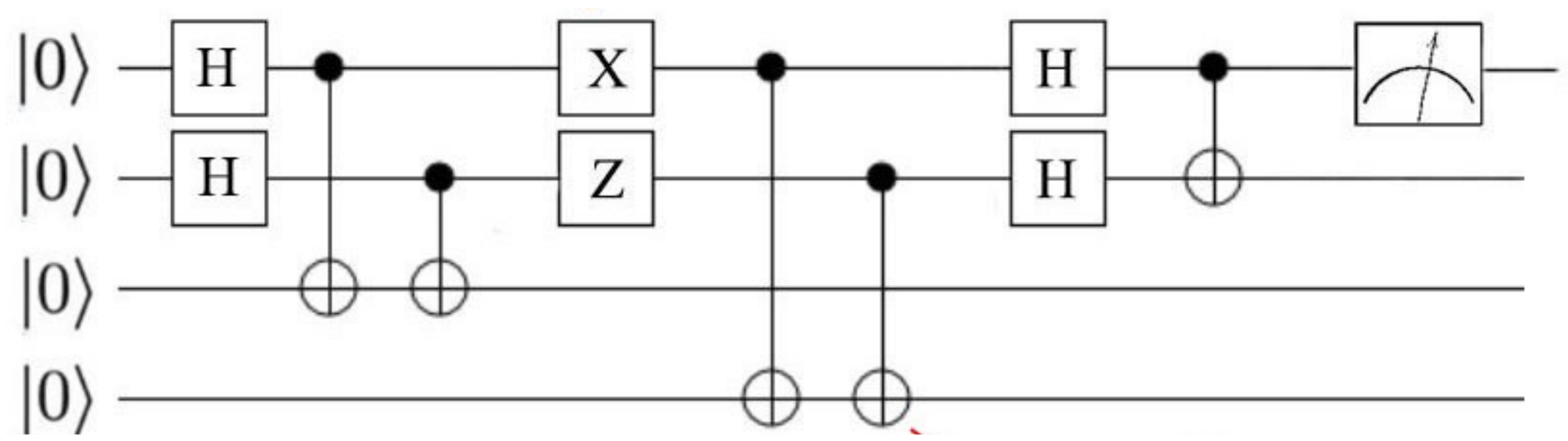
Supercomputing Scale Quantum Circuit Simulation with cuQuantum

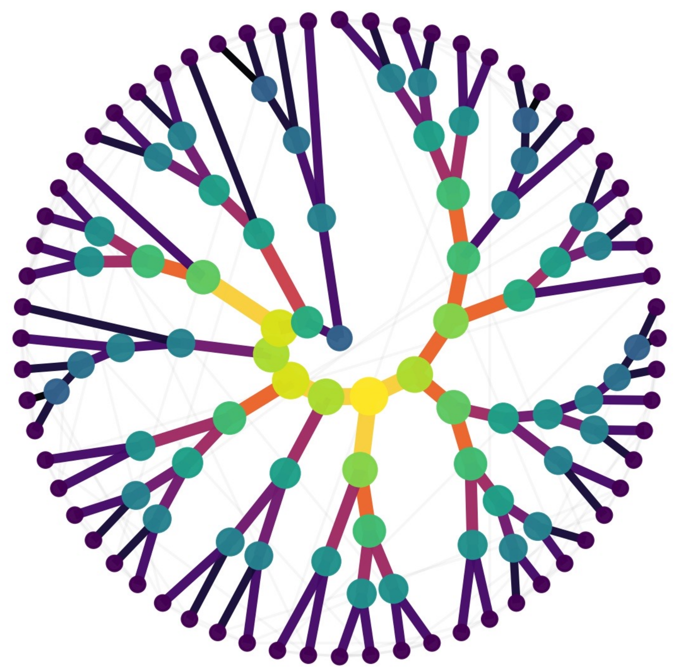
cuQuantum
Deployed on Perlmutter now!

- **SDK for Quantum Circuit Simulation**
  - 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
- **cuQuantum available now**
  - Integrated into leading quantum computing frameworks Cirq, Qiskit, and Pennylane
  - C and Python APIs
  - Available today at developer.nvidia.com/cuquantum



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

Researching & developing the quantum applications of tomorrow requires powerful simulations today

NVIDIA

DGX cuQuantum Appliance

Deployed on Perlmutter

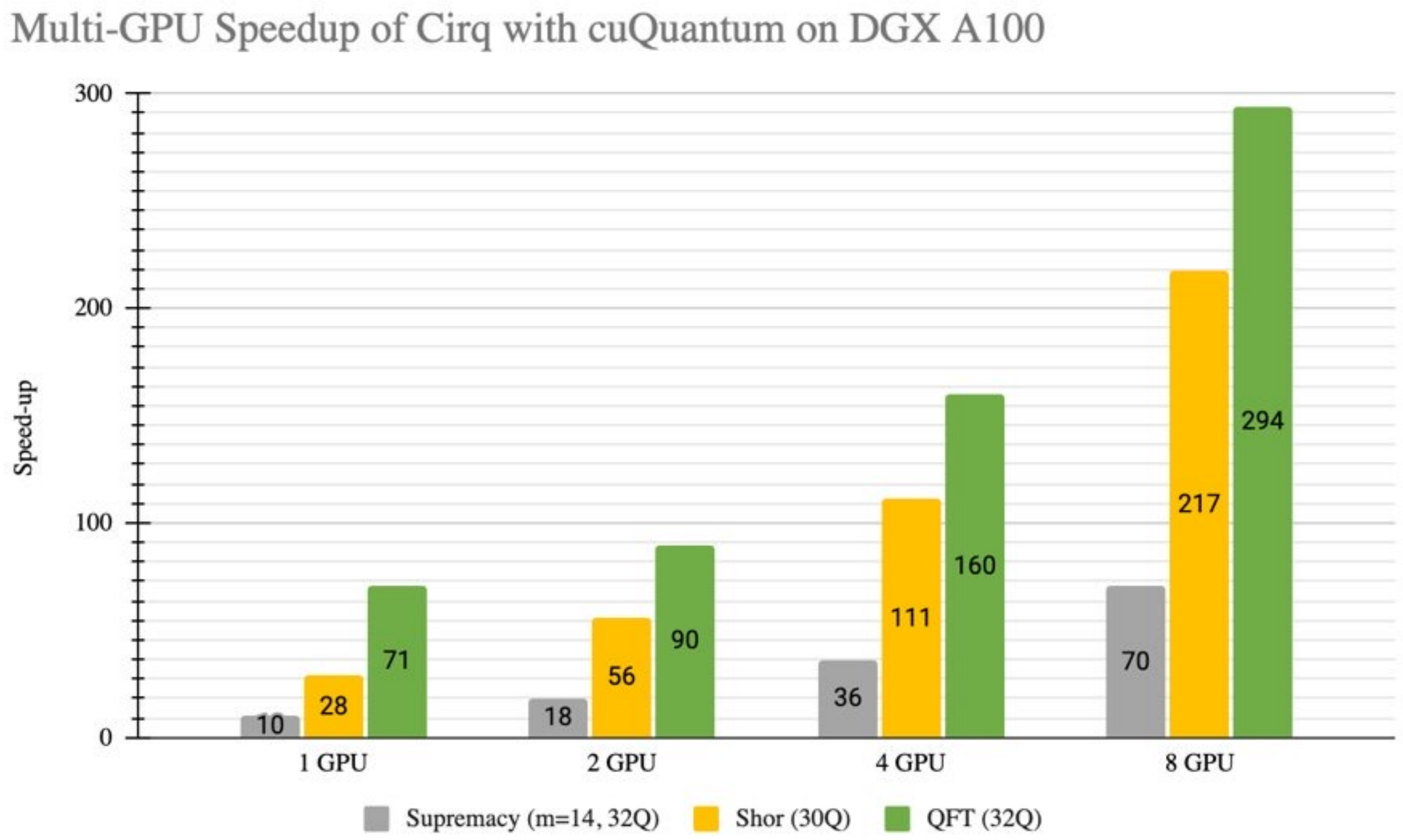
Fully integrated quantum simulation solution

- State-of-the-art *performance*
- Unmatched simulation *scale*

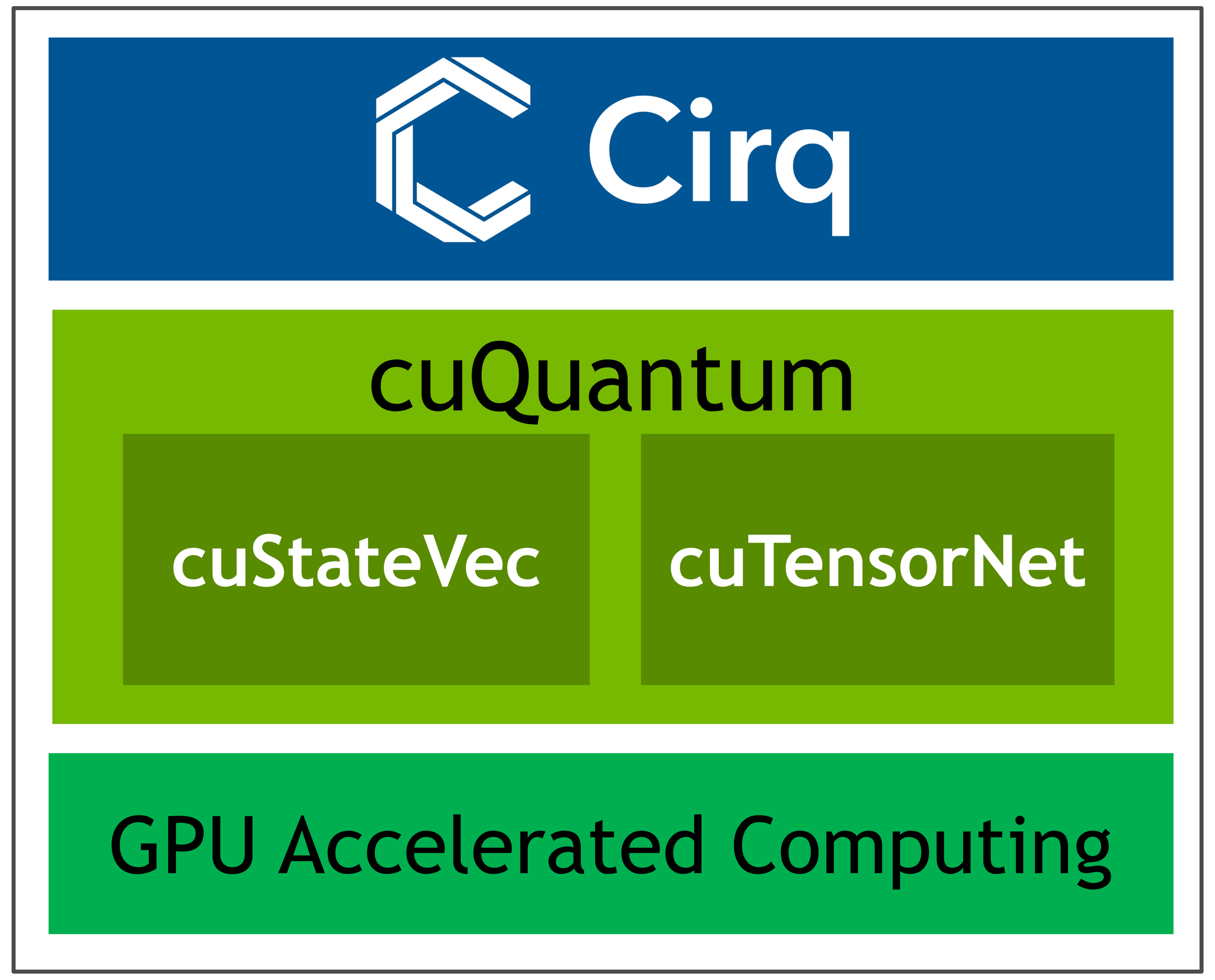
Reduce the simulation time by ***orders of magnitude***

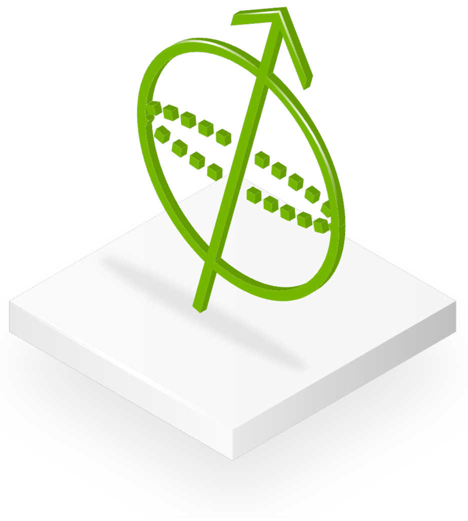
Simulate ***thousands*** of perfect or noisy qubits

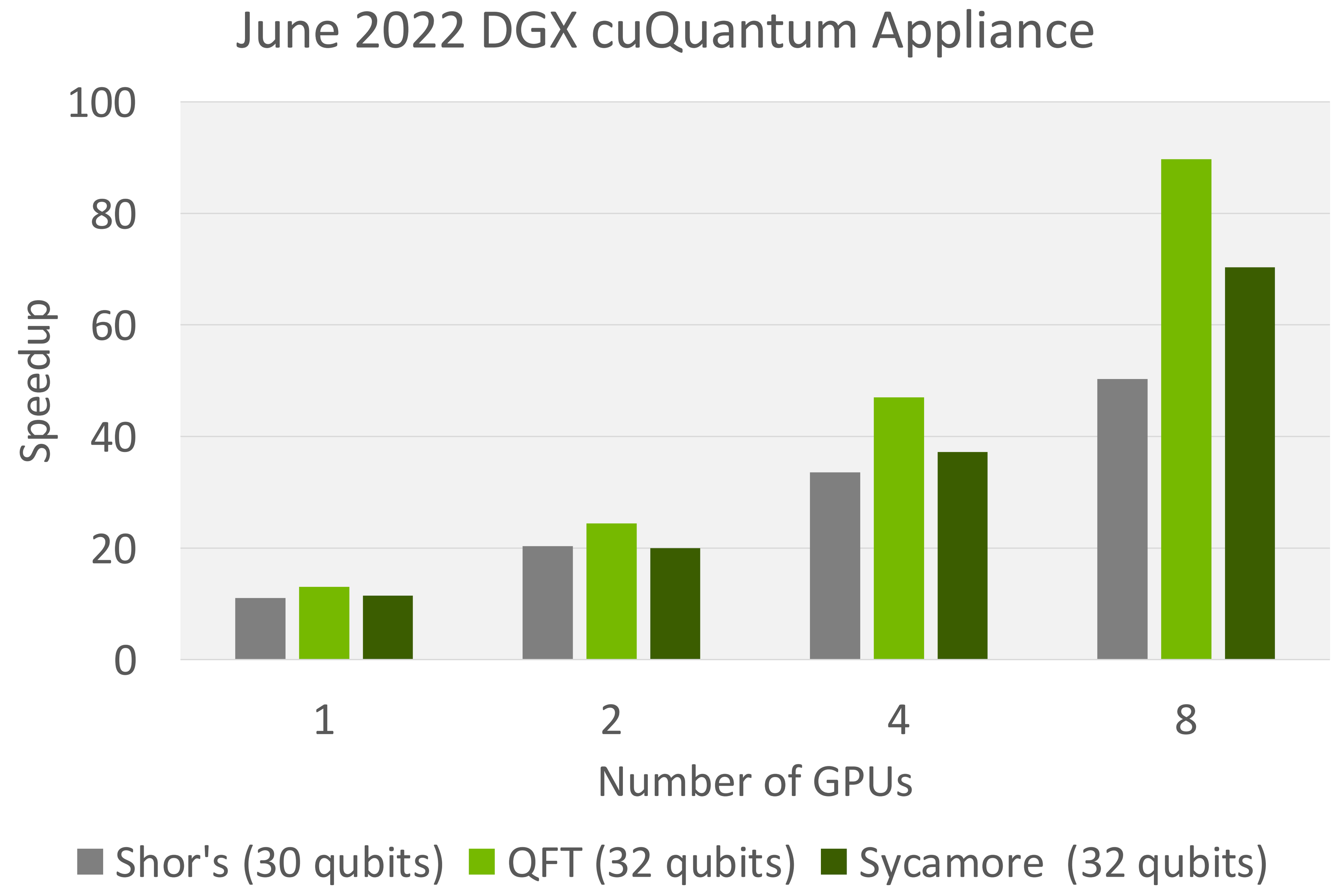
Up to 4.4X Perf since June



DGX cuQuantum Appliance







DGX cuQuantum Appliance

Coming in Q4: Qiskit Integration with multi-node, multi-GPU support

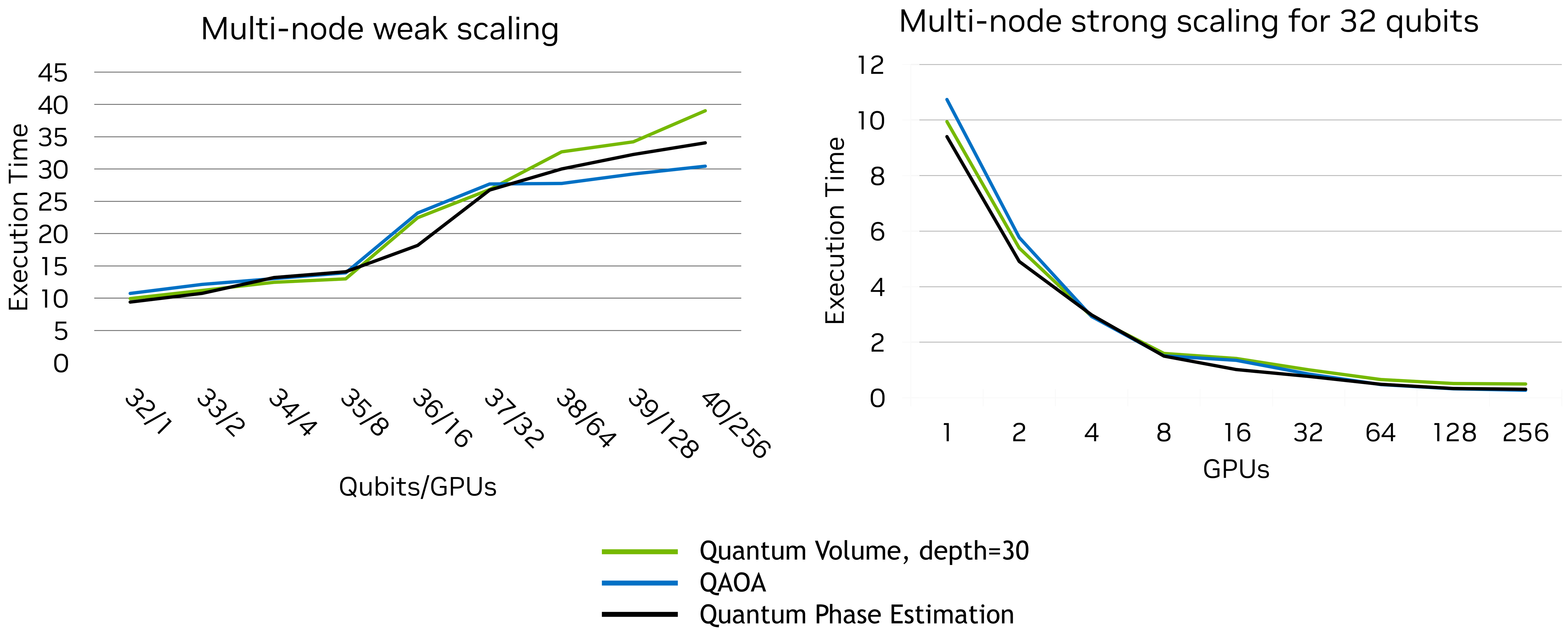
Fully integrated quantum simulation solution

- State-of-the-art *performance*
- Unmatched simulation *scale*

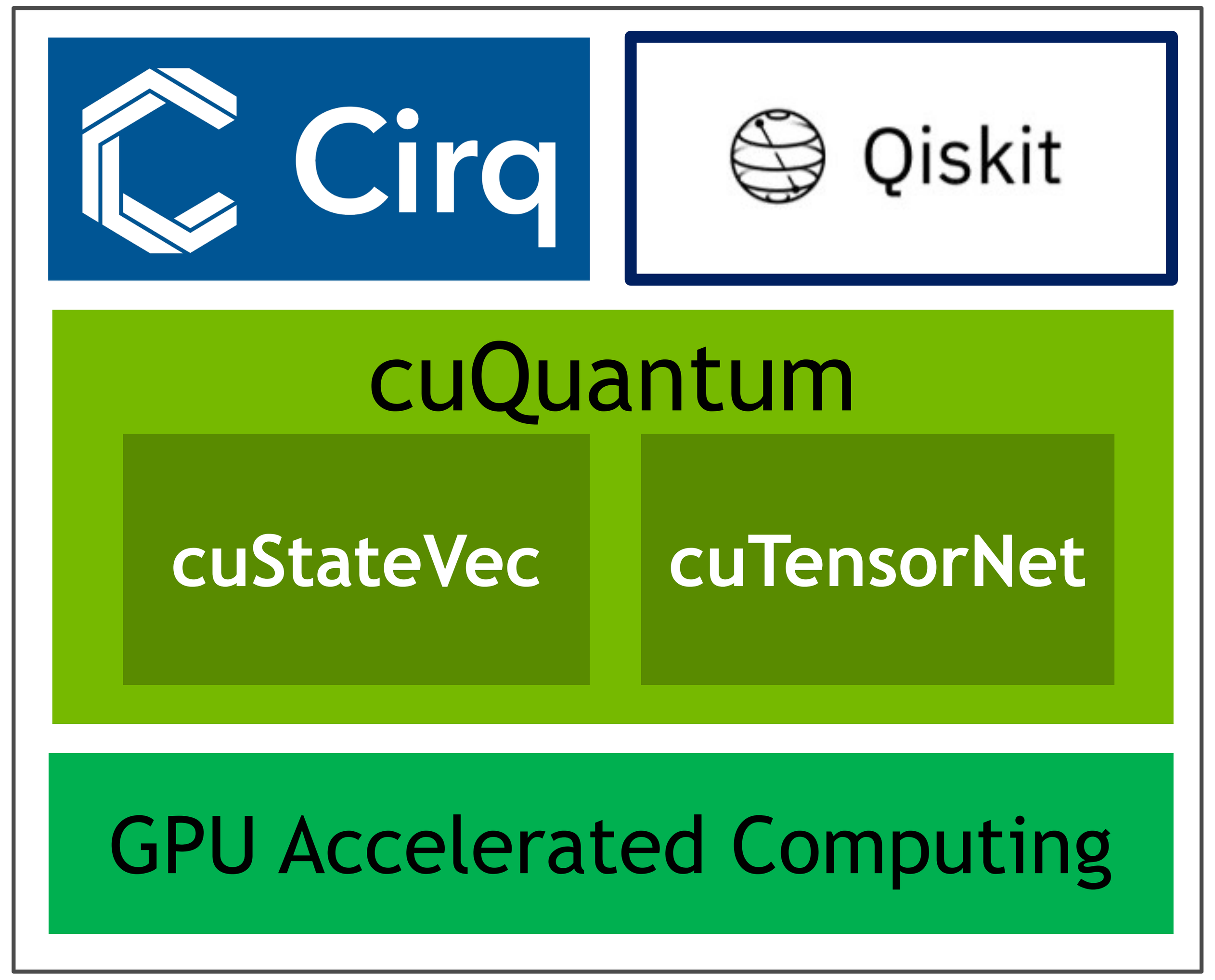
Reduce the simulation time by *orders of magnitude*

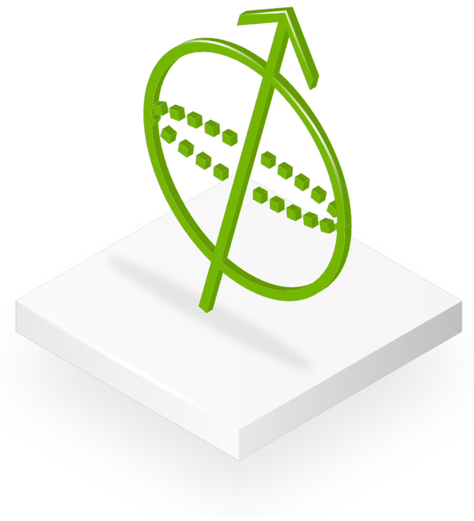
Simulate *thousands* of perfect or noisy qubits

Trivially Scale Quantum Algorithms with Industry Leading Performance

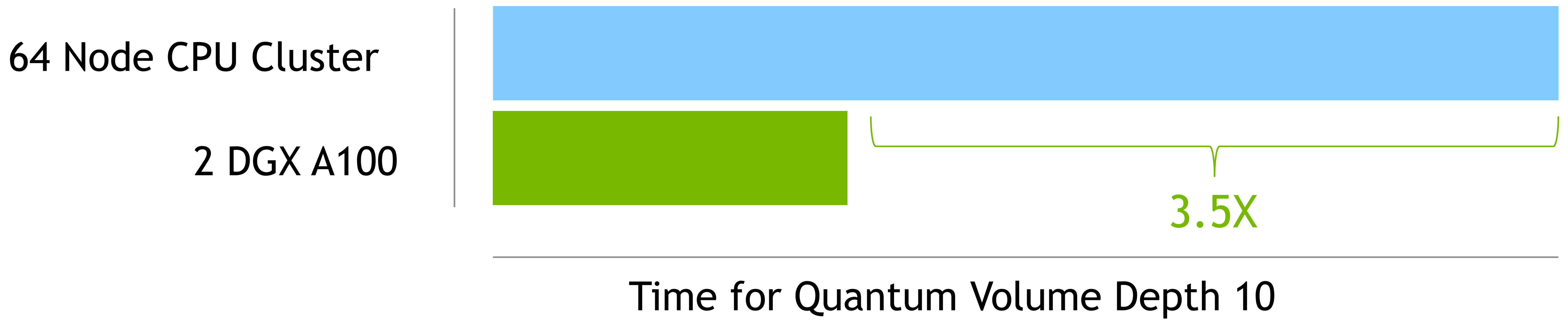


DGX cuQuantum Appliance





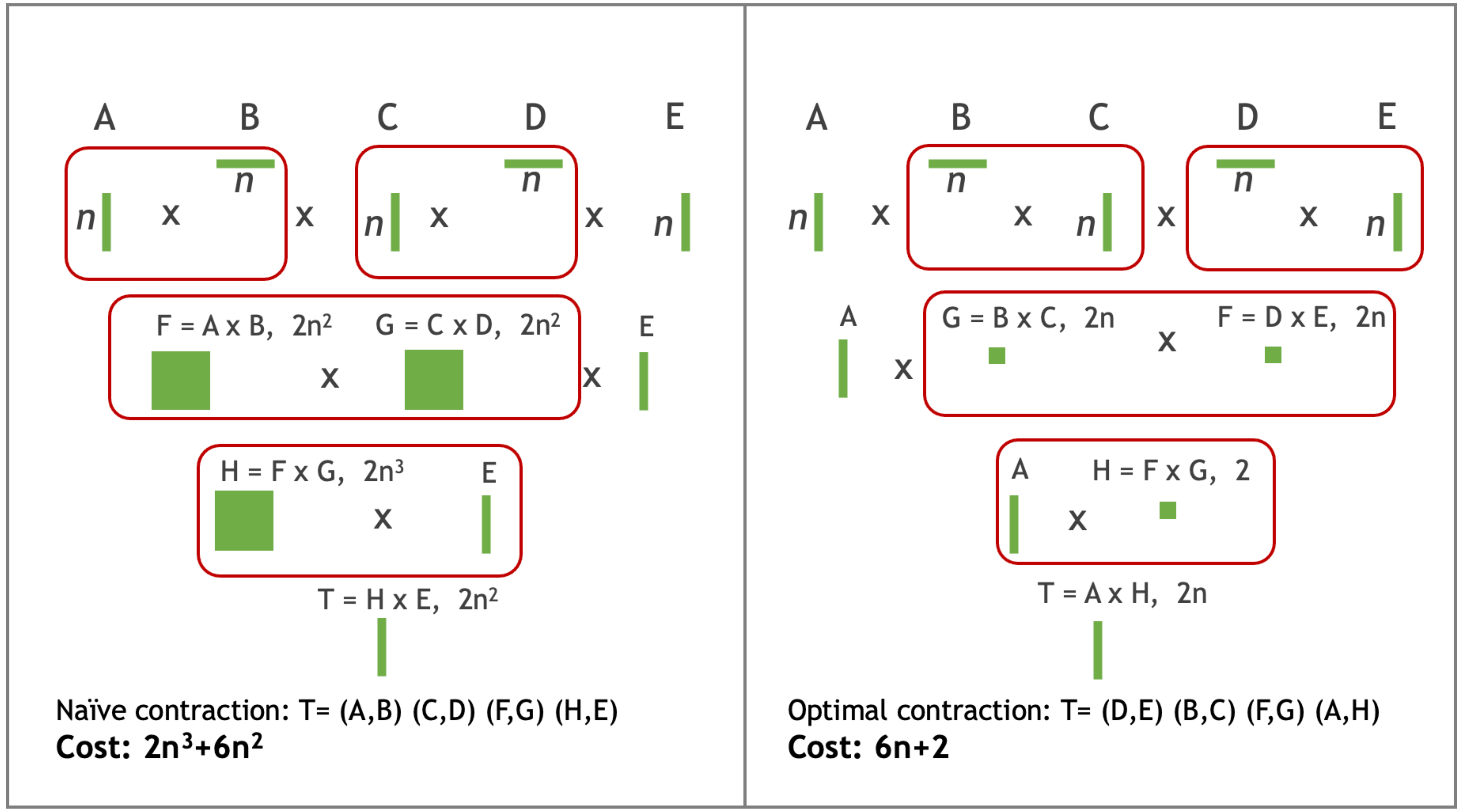
Record breaking performance
2 DGX A100 vs previous best on 64 node CPU cluster

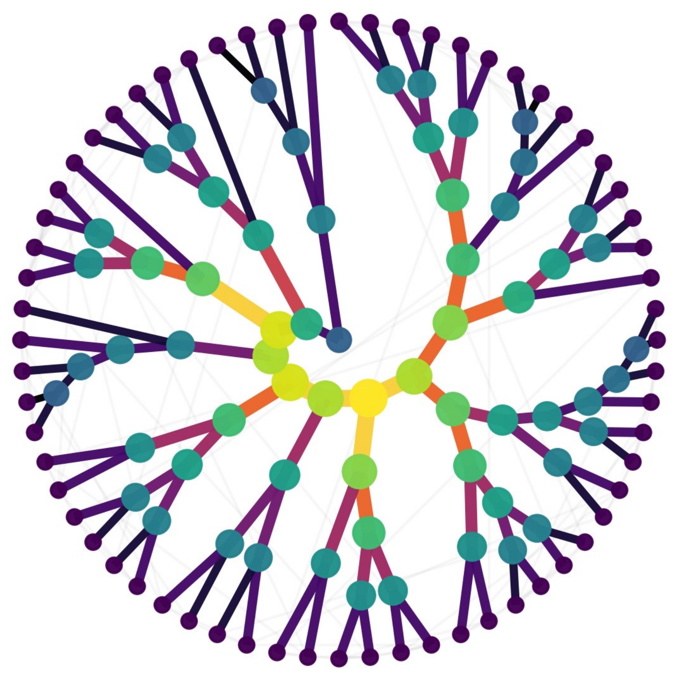


cuTensorNet

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

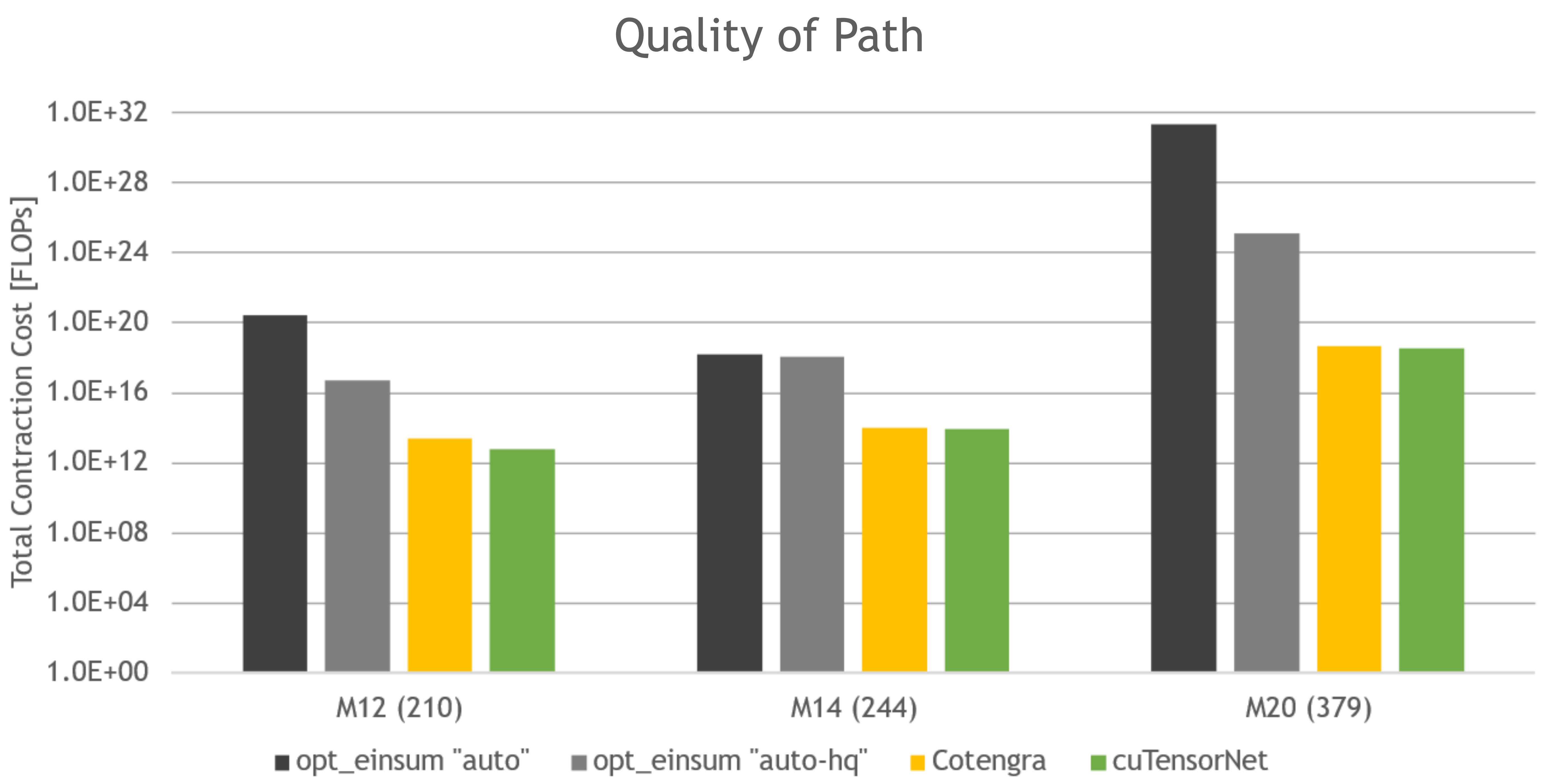


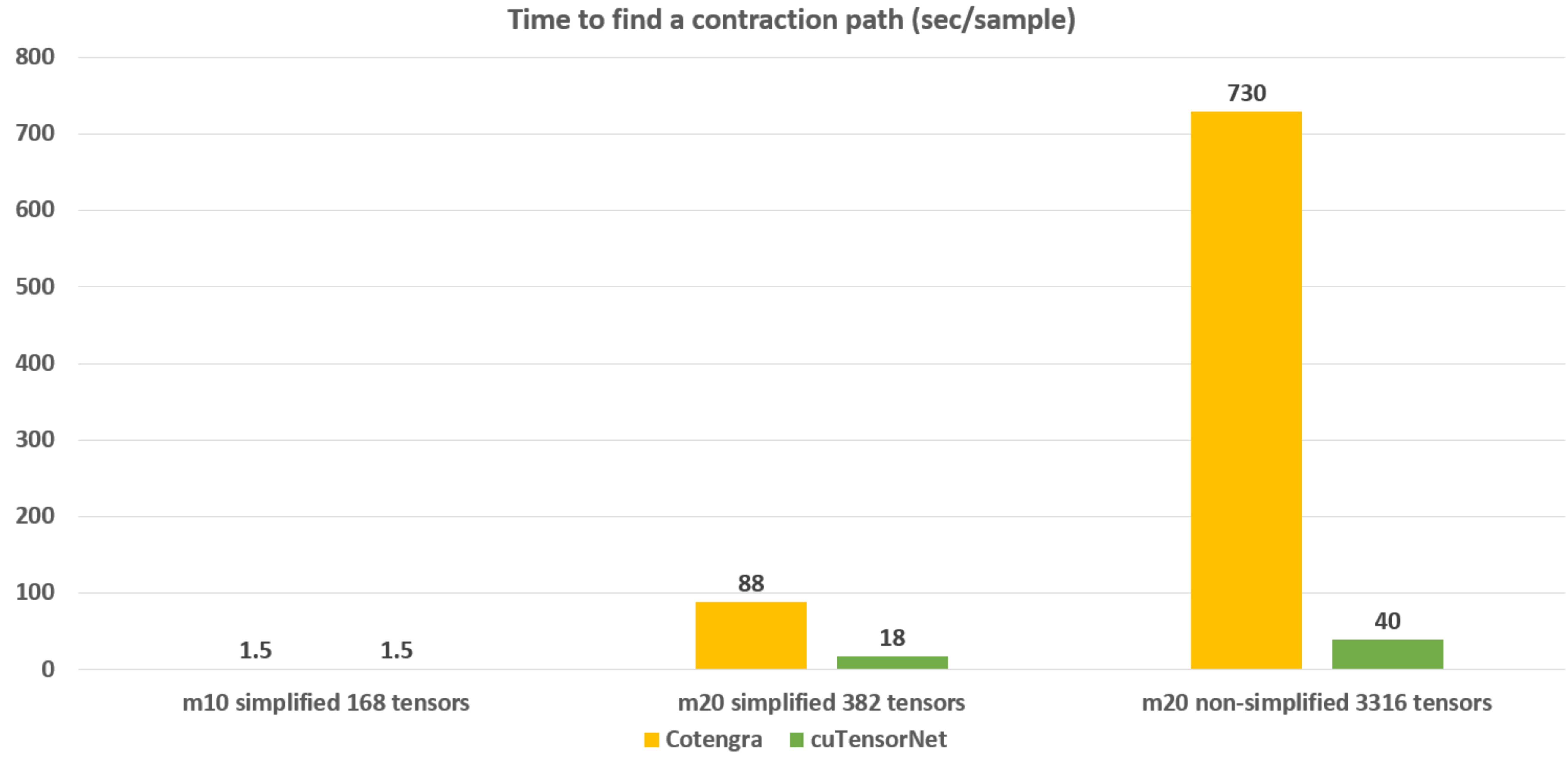


Tensor Network image from Quimb: https://quimb.readthedocs.io/en/latest/index.html

cuTensorNet

TENSOR NETWORK PATH OPTIMIZATION PERFORMANCE





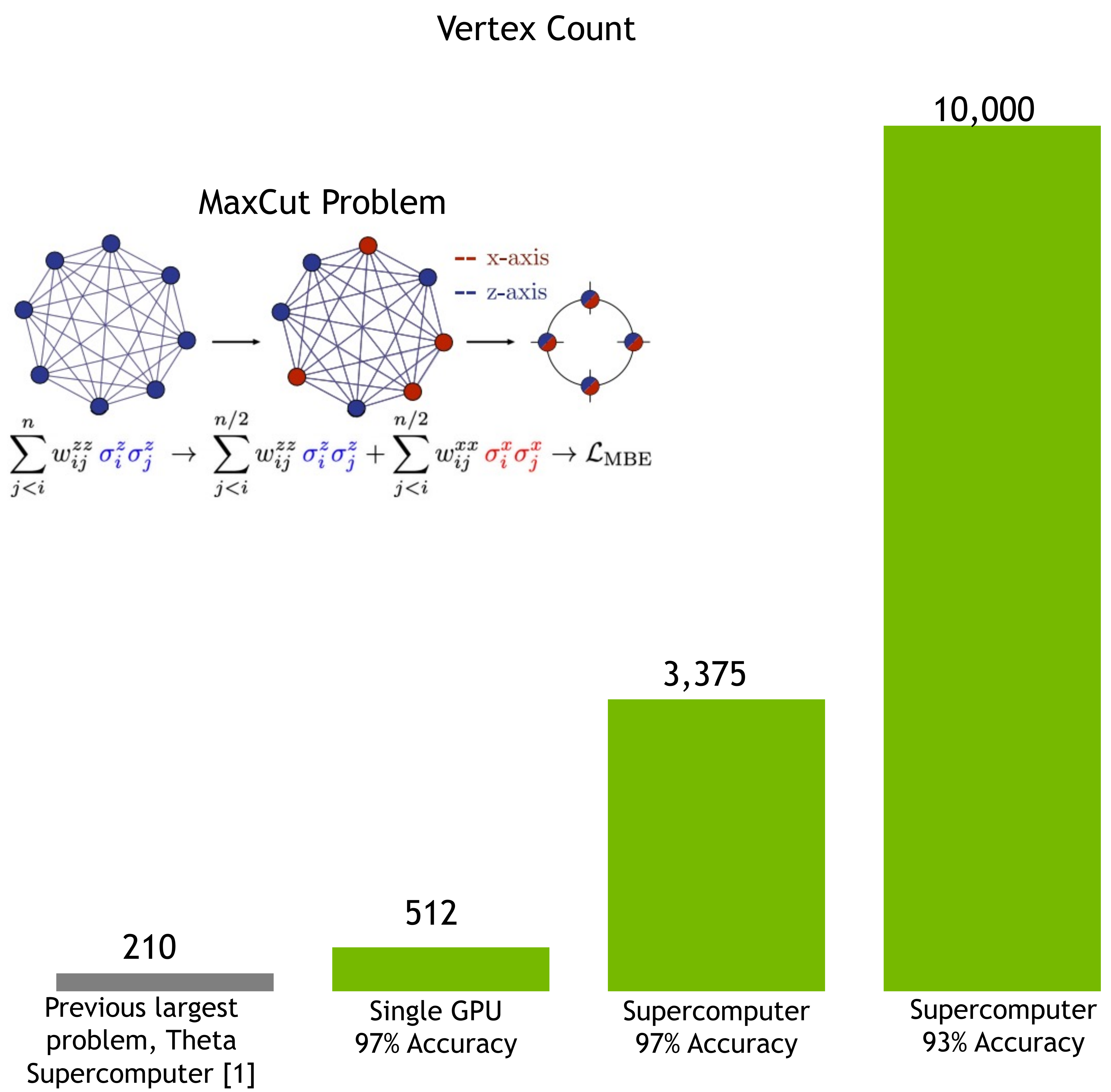
- cuTensorNet achieves SotA pathfinding results dramatically faster, and does better with more complex networks

[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/

Scaling Simulations to a Supercomputer

NVIDIA's Selene DGX SuperPOD based supercomputer

- Using 20 nodes of the Selene Supercomputer with tensor network method simulation
- 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

NVIDIA cuQuantum Ecosystem

aws
BMW
Deloitte.
FUJIFILM
IBM
Google Quantum AI
ORACLE
softserve

CLASSIQ
ColdQuanta
IONQ
PASQAL
PsiQuantum
QCWARE
QUANTINUUM
QUANTUM BRILLIANCE
rigetti
seeqc
XANADU
ZAPATA

bluegat
Cirq
Orquestra®
PENNYLANE
QIBO
Qiskit
XACC QUANTUM FRAMEWORK

Argonne NATIONAL LABORATORY
JÜLICH Forschungszentrum
NERSC
OAK RIDGE National Laboratory

Summary

- cuQuantum available today: https://developer.nvidia.com/cuquantum-sdk
  - Supports statevector and tensor network methods
  - Simulate noisy or perfect qubits
  - Integrated in all major quantum circuit simulation frameworks (Cirq, Qiskit, PennyLane…)
- DGX cuQuantum Appliance available on Perlmutter today as well as for download: https://catalog.ngc.nvidia.com/orgs/nvidia/containers/cuquantum-appliance
  - Multi-GPU supported today
  - Q4 release will include Qiskit

NVIDIA