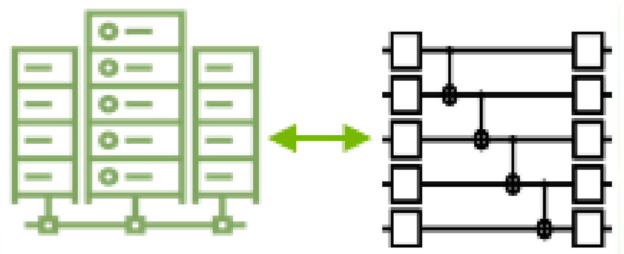
Towards the Quantum-Accelerated Supercomputer

Programming Future Heterogenous Quantum-Classical Supercomputing Architectures

Esperanza Cuenca-Gomez, Developer Relations Manager for Quantum Computing

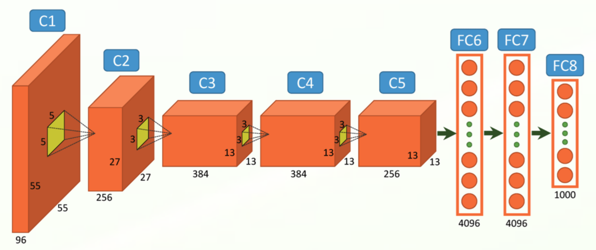
Forum Teratec 24, 29th of May of 2024

Accelerated Computing



Quantum Computing

Generative AI



Deep Learning

Scientific Computing

Quantum-Accelerated Supercomputing

Supercomputers are the foundation of Quantum R&D

Simulation

- Quantum computers are small and error-prone -> simulation is an essential tool
- **Today:** Powerful simulators enable algorithm and application R&D – new approaches (e.g. tensor networks)
- **Future:** Digital twins of quantum computers for design and architecture optimization

HPC Quantum Integration

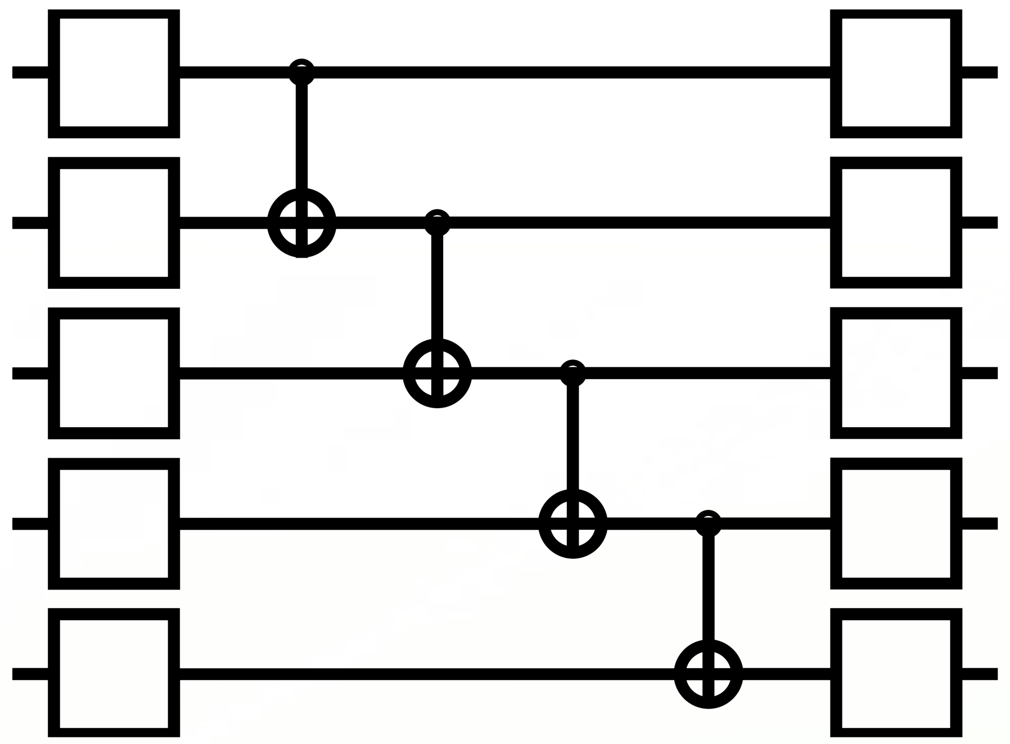
- Useful quantum computing will be hybrid
- **Today:** Enable domain scientists to start developing for QPUs, enable quantum researchers to use accelerated computing
- **Future:** quantum computers will integrate tightly with supercomputers as accelerators and be co-programmed

AI for Quantum

- Error correction, calibration, control, compilation are challenging computationally, real-time compute often needed
- Accelerated computing and AI can solve these problems
- **Today:** Enable AI research for all of the above
- **Future:** Hybrid Quantum+AI supercomputer with low-latency link

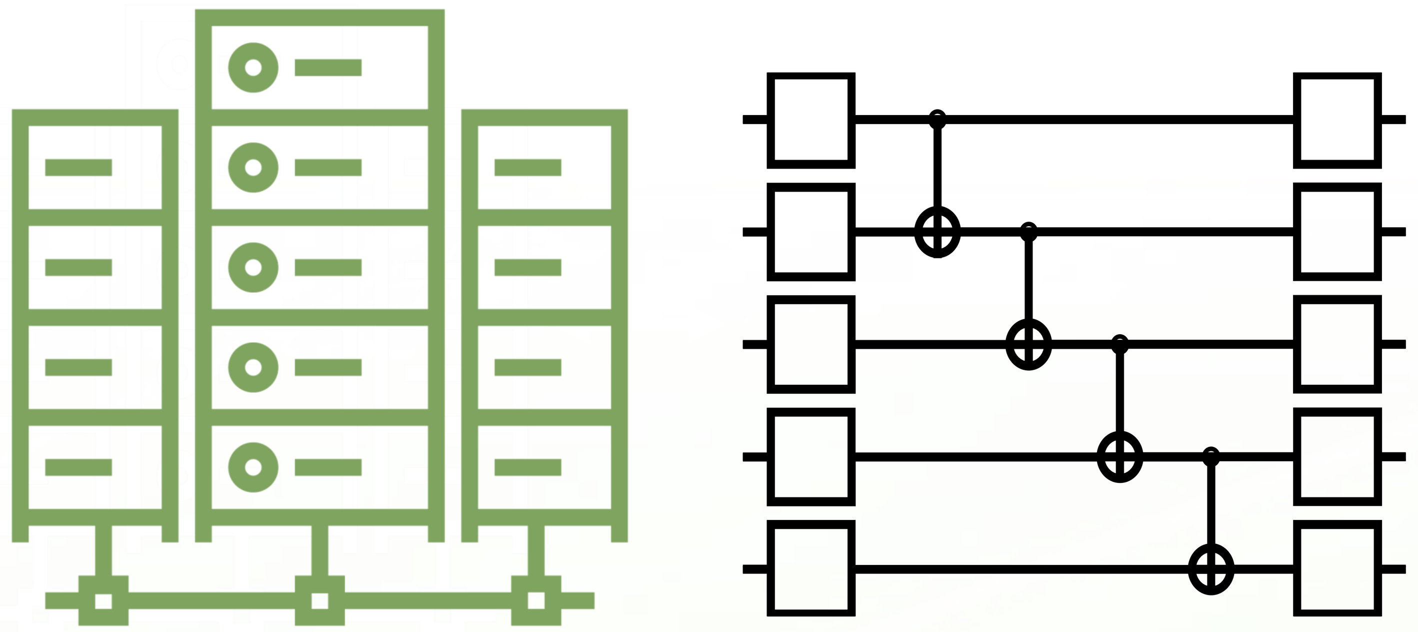
NVIDIA Quantum

Powering the Global Quantum Computing Community

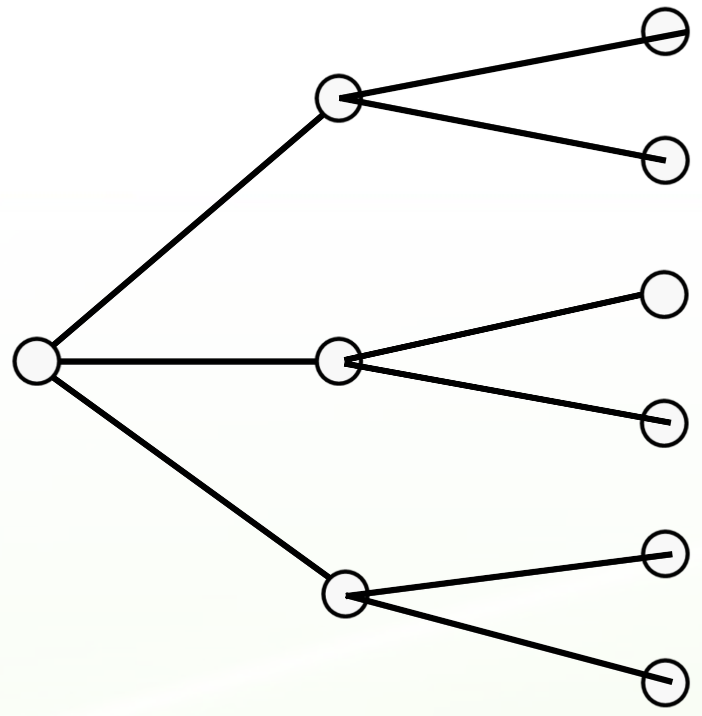


Simulation

Algorithm Design, Resource Estimation, QPU Design



HPC Quantum Integration



AI for Quantum

QEC, Calibration, Algorithms

CUDA-Q

Libraries | Programming Model | Tools | Infrastructure

cuQuantum

Quantum Simulation

DGX-Q

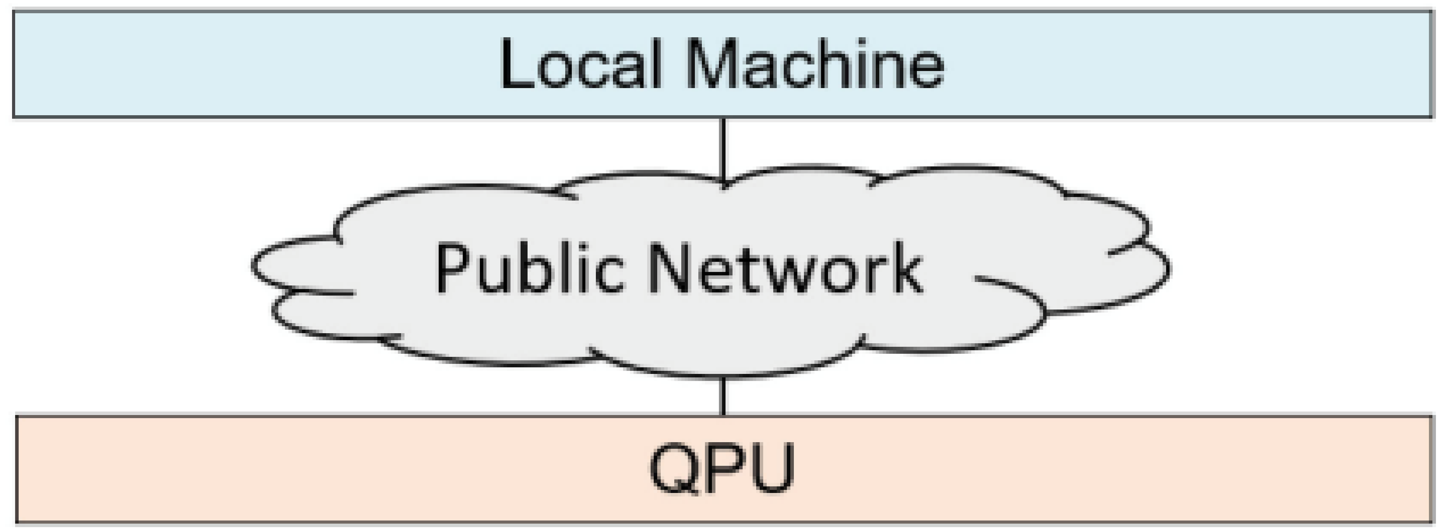
Quantum Integrated Computing

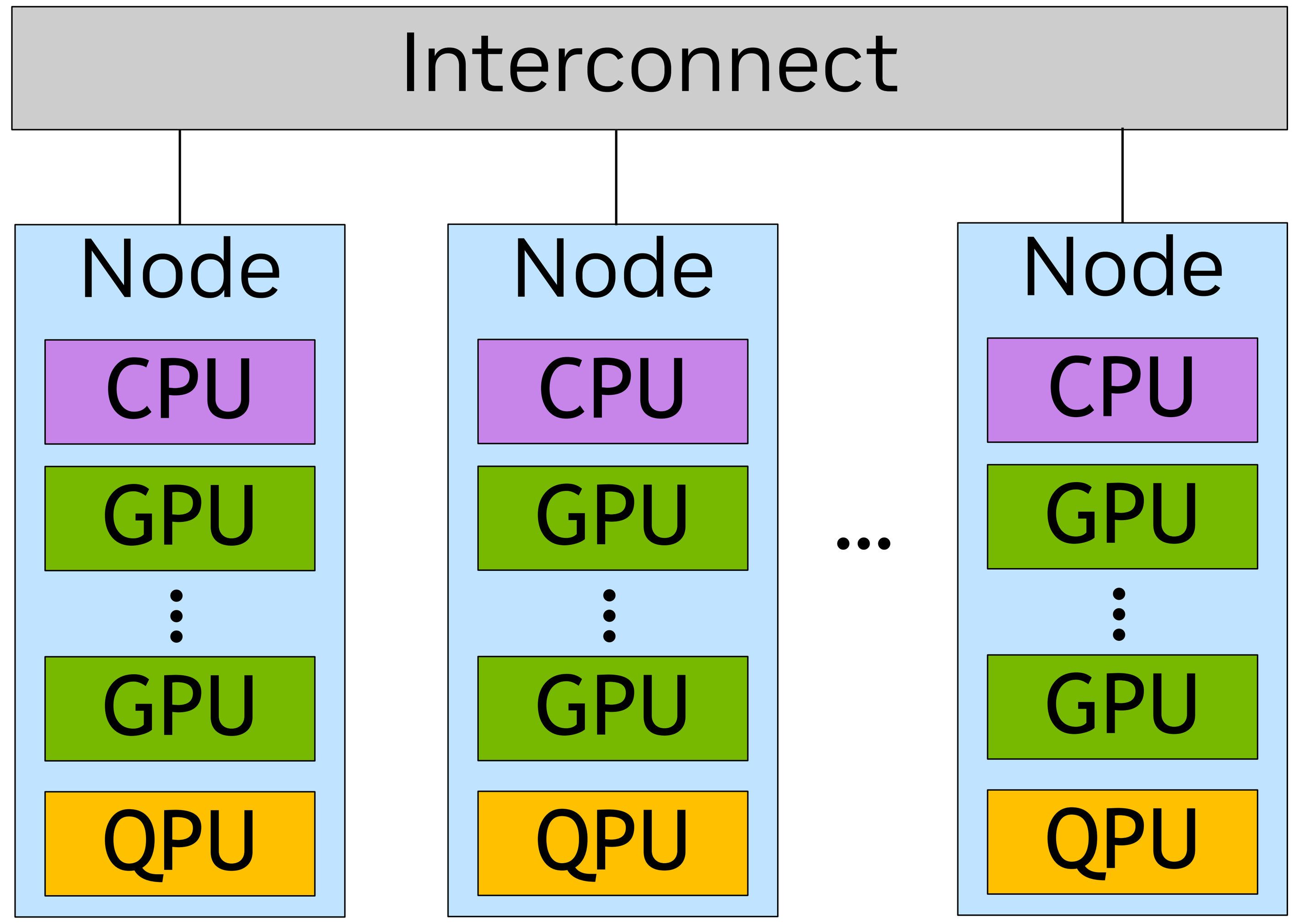
GPU Supercomputing

NVIDIA, HPC, and Quantum Computing

Integrate quantum computers seamlessly with the modern scientific computing ecosystem

- HPC centers are starting to integrate QPUs
- QPUs will accelerate certain workloads suitable for quantum acceleration
  - Quantum chemistry, materials simulation, AI
- CPUs and GPUs expected to enhance QPU performance
  - Classical preprocessing (circuit optimization) and postprocessing (error correction)
  - Optimal control and QPU calibration
  - Hybrid workflows
- Need the ability to seamlessly integrate these architectures





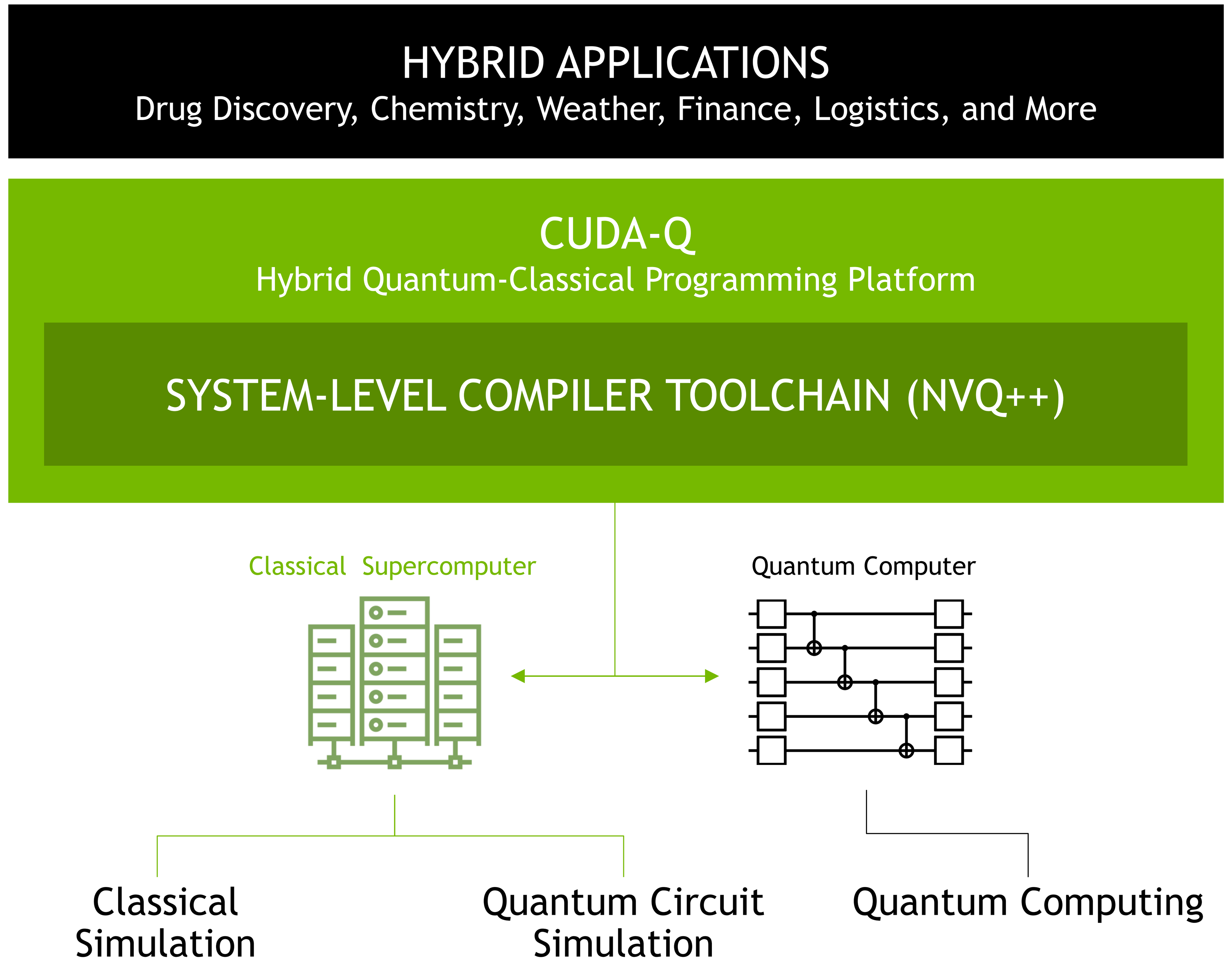
Introducing CUDA-Q

Platform for unified quantum-classical accelerated computing

- Programming model extending C++ and Python with ***quantum kernels***
- ***Open*** programming model, open-source compiler
  - https://github.com/NVIDIA/cuda-quantum
- ***QPU Agnostic*** – Partnering broadly including superconducting, trapped ion, neutral atom, photonic, and NV center QPUs
- ***Interoperable*** with the modern scientific computing ecosystem
- ***Retargetable*** - seamless transition from simulation to physical QPU

```
auto ansatz = [](std::vector<double> thetas) __qpu__ {
  cudaq::qvector q(3);
  x(q[0]);
  ry(thetas[0], q[1]);
  ry(thetas[1], q[2]);
  x<cudaq::ctrl>(q[2], q[0]);
  x<cudaq::ctrl>(q[0], q[1]);
  ry(-thetas[0], q[1]);
  x<cudaq::ctrl>(q[0], q[1]);
  x<cudaq::ctrl>(q[1], q[0]);
};

cudaq::spin_op H = ...;
double energy = cudaq::observe(ansatz, H, {M_PI, M_PI_2});
```



The CUDA-Q Stack

Platform for unified quantum-classical accelerated computing

CUDA-Q in Action

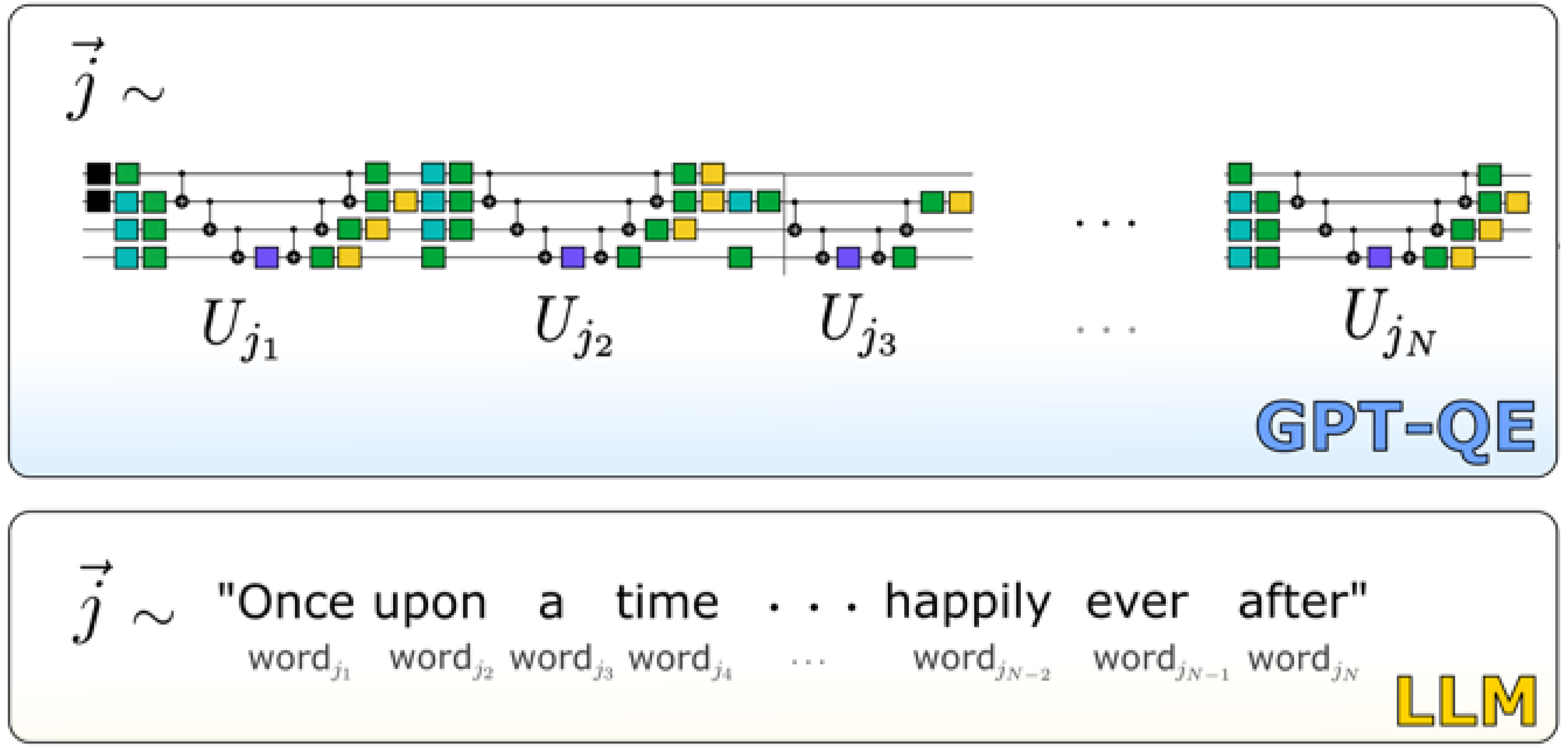
GPT-QE - University of Toronto and St. Jude Children’s Research Hospital with CUDA-Q

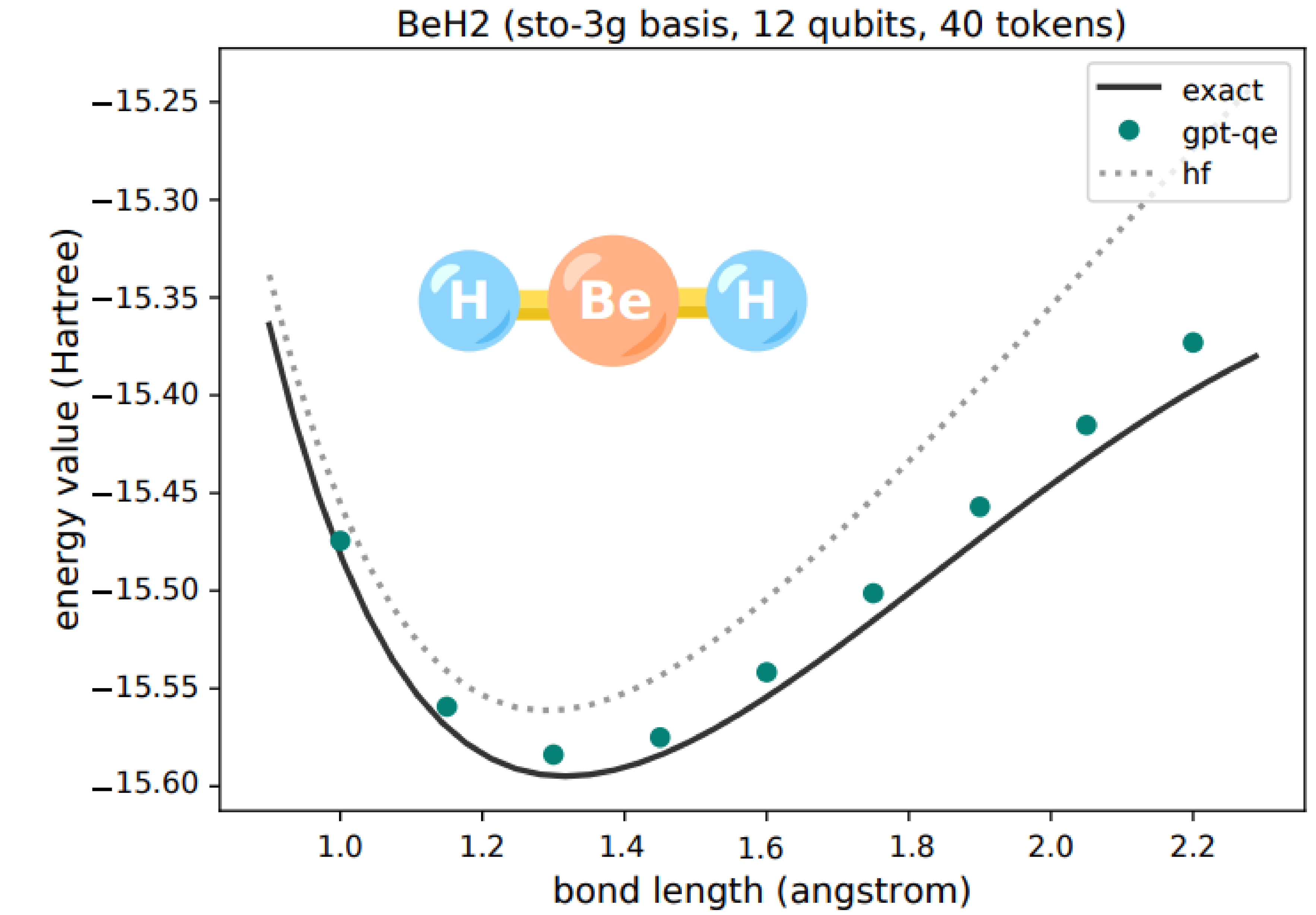
- Developed a novel Generative Pre-Trained Transformer-based (GPT) method for computing the ground-state energy of molecules of interest
- The first demonstration of a GPT-generated quantum circuit in the literature
- A powerful example of leveraging AI to accelerate quantum computing
- Executed using CUDA-Q on A100 GPUs on Perlmutter
- Opens the door to a wide variety of novel Generative Quantum Algorithms (GQAs) for drug discovery, materials science, and environmental challenges

UNIVERSITY OF TORONTO

St. Jude Children’s® Research Hospital

Finding cures. Saving children.





https://arxiv.org/pdf/2401.09253.pdf

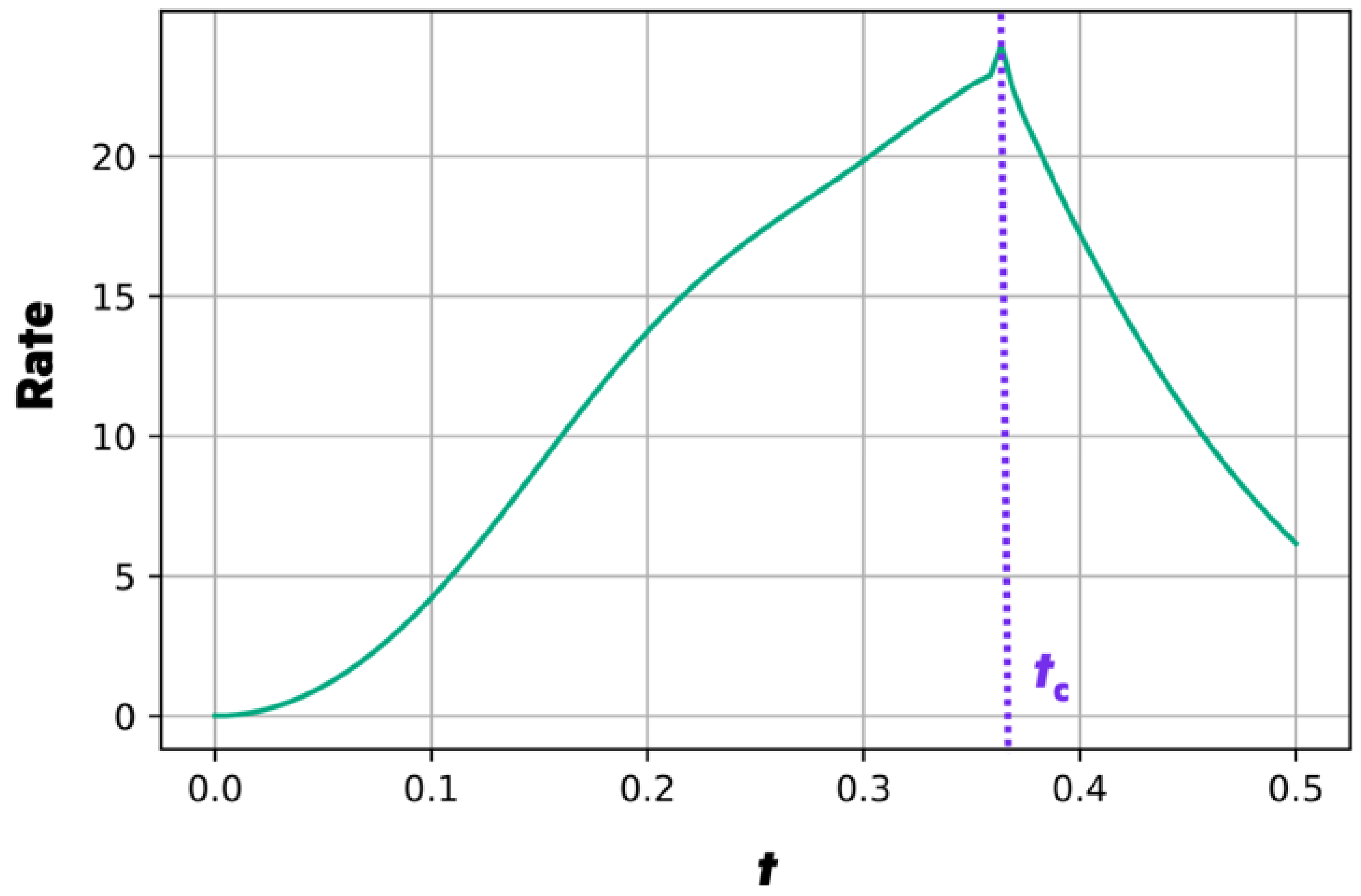
CUDA-Q in Action

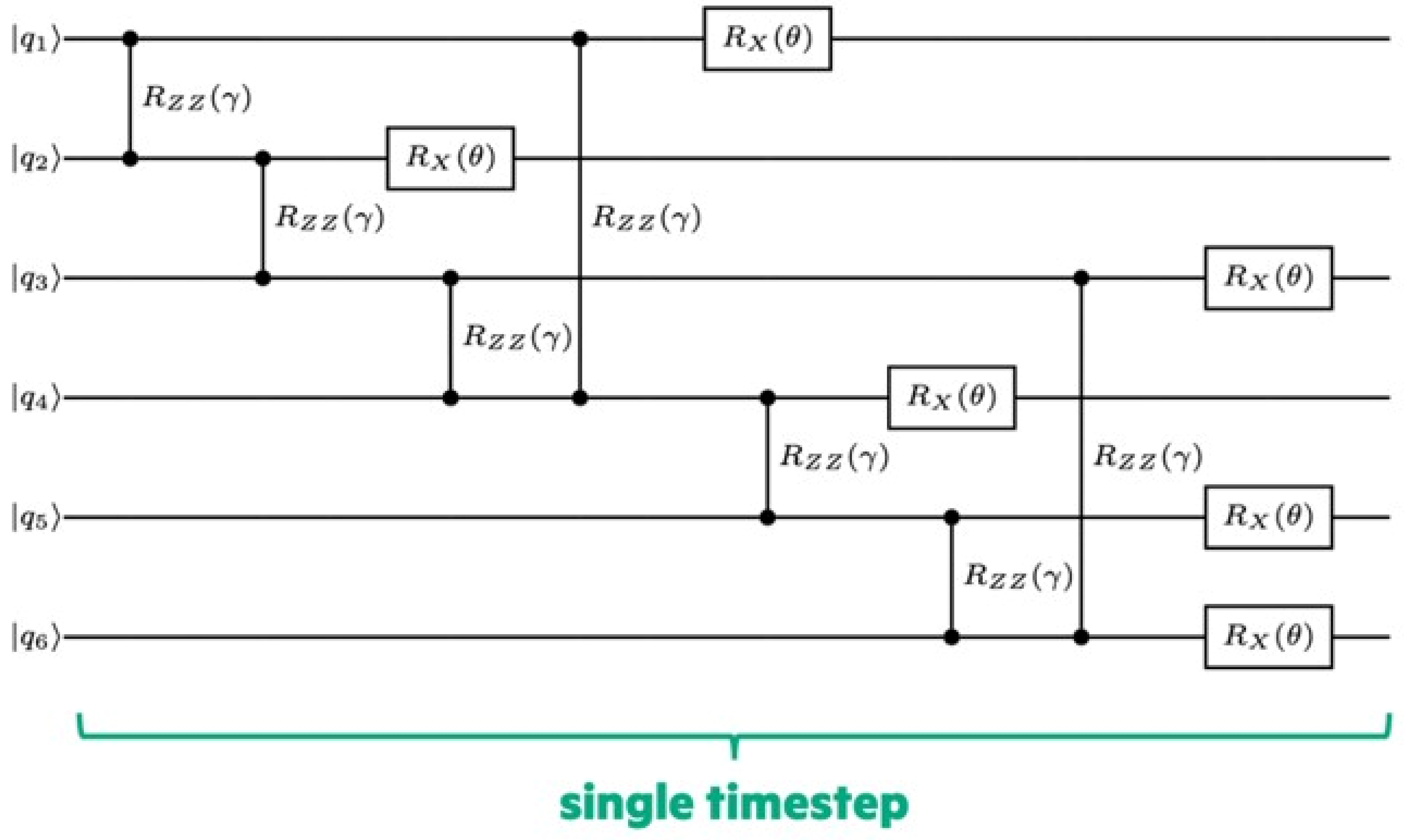
Speed-ups for time-evolution of the transverse field Ising model (TFIM)

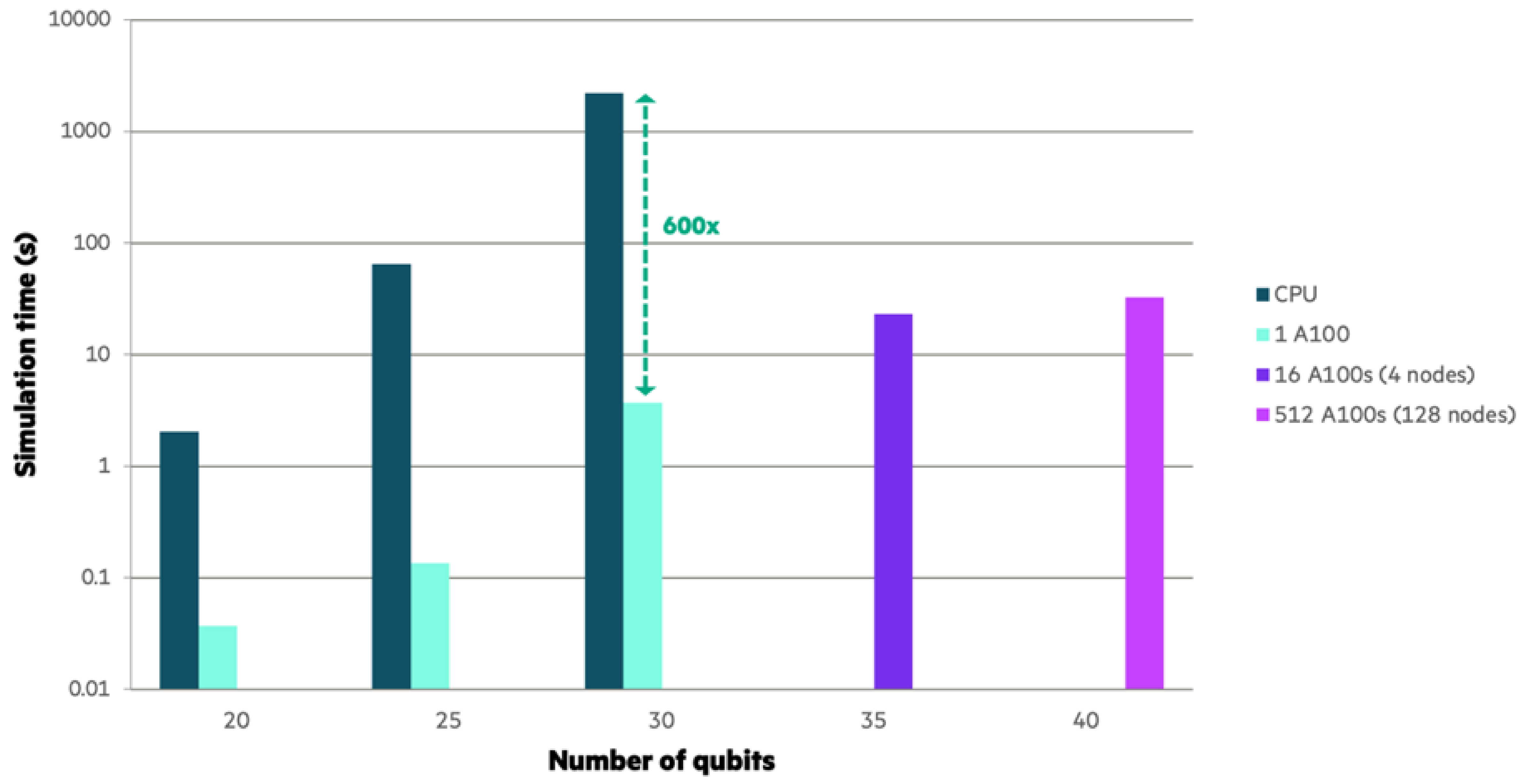
- Collaboration with Hewlett Packard Labs
- Study dynamical quantum phase transitions
  - Requires computation of overlap of initial state with time evolved state
- Leverage NVIDIA multi-node, multi-GPU simulation backend.
  - Distributed state-vector simulator
- 600x performance increase over multi-threaded CPU approaches

```
@cudaq.kernel()
def tfimEvolve(timeStep: float, params: list[float]):
    qubits = cudaq.qvector(40)
    ... Circuit, use input params ...

for time in range(finalTime):
    state = cudaq.get_state(tfimEvolve, time, params)
    overlaps.append(state.overlap(initialState)
```



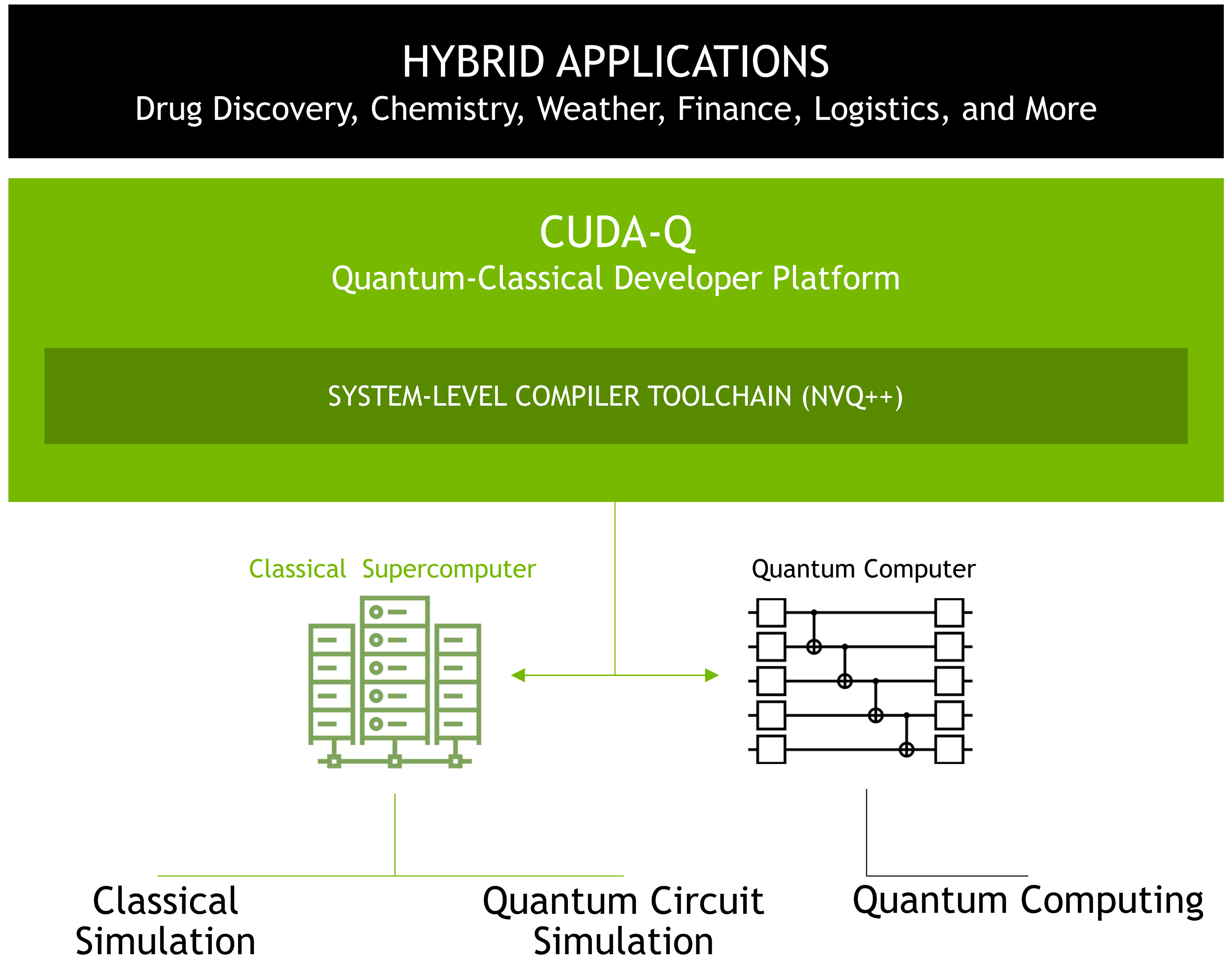




CUDA-Q: Now Available on GitHub and NGC

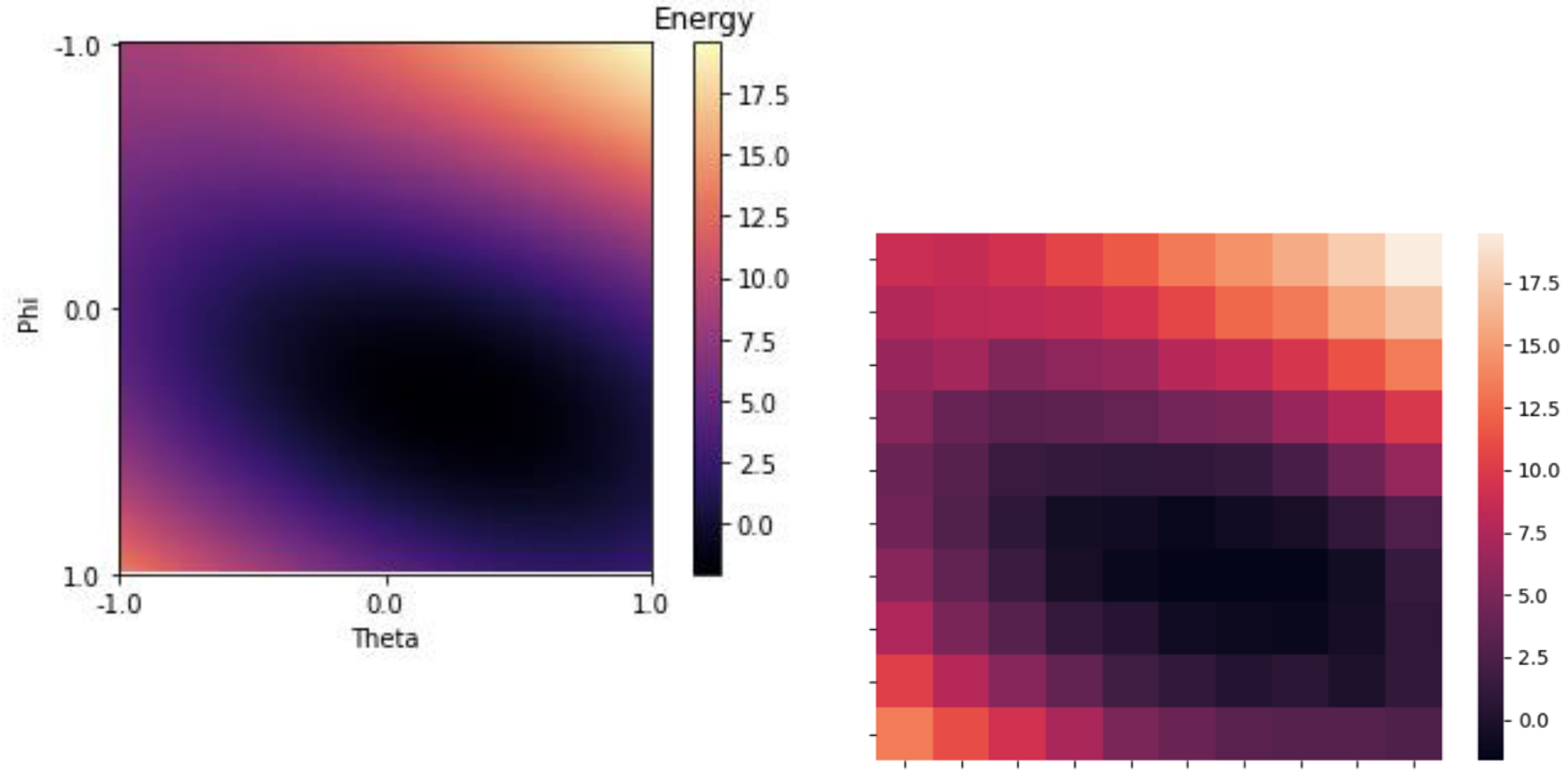
Seamlessly Target any Quantum Resource

CUDA-Q PLATFORM



CUDA-Q and Quantinuum

```
nvq++ -qpu=cuquantum vqe.cpp
```




```
nvq++ -qpu=quantinuum:h1 vqe.cpp
```

github.com/nvidia/cuda-quantum | https:// https://catalog.ngc.nvidia.com/orgs/nvidia/teams/quantum/containers/cuda-quantum

VQE: NVQ++/cuStateVec on A100 GPU vs Leading Pythonic Framework and SImulator with Thrust on A100 GPU

CUDA-Q: Now Available on GitHub and NGC

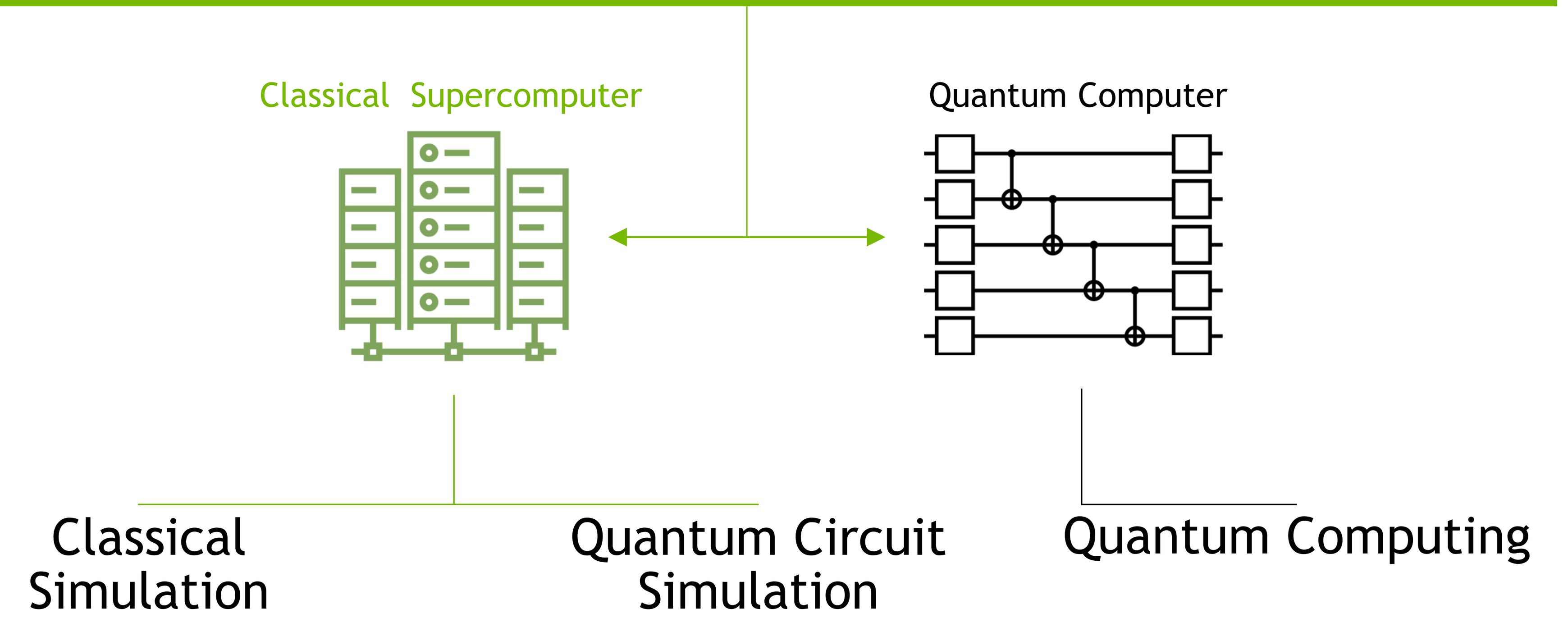
Adopted by Community’s Leaders to Enable Quantum-Accelerated Applications

CUDA-Q PLATFORM

HYBRID APPLICATIONS
Drug Discovery, Chemistry, Weather, Finance, Logistics, and More

CUDA-Q
Quantum-Classical Developer Platform

SYSTEM-LEVEL COMPILER TOOLCHAIN (NVQ++)



CUDA-Q QPU PARTNERS

ANYON
atom computing
IONQ
IQM
OQC
ORCA COMPUTING
PASQAL
QUANTUM BRILLIANCE
QUANTINUUM
QuEra COMPUTING INC.
rigetti
XANADU

… and growing

github.com/nvidia/cuda-quantum | https:// https://catalog.ngc.nvidia.com/orgs/nvidia/teams/quantum/containers/cuda-quantum

cuQuantum

Research the Quantum Computer of Tomorrow on the most Powerful Computer Today

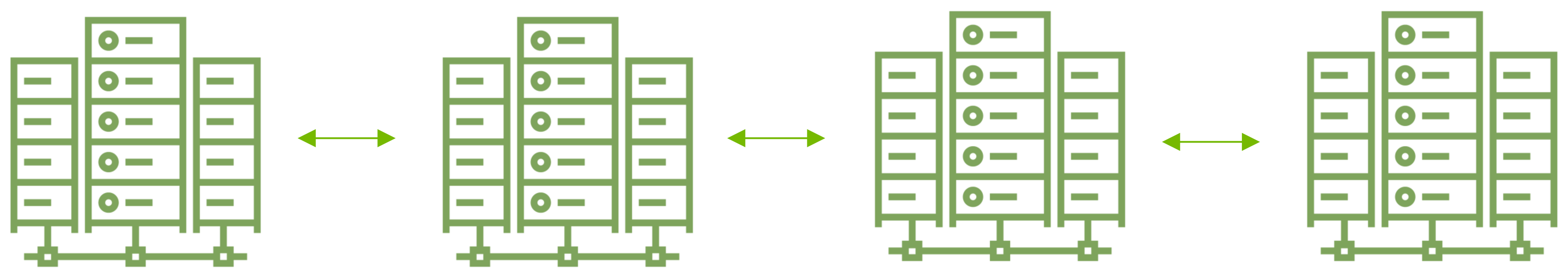
cuQuantum

Quantum Algorithms
Drug Discovery, Chemistry, Weather, Finance, Logistics, and More

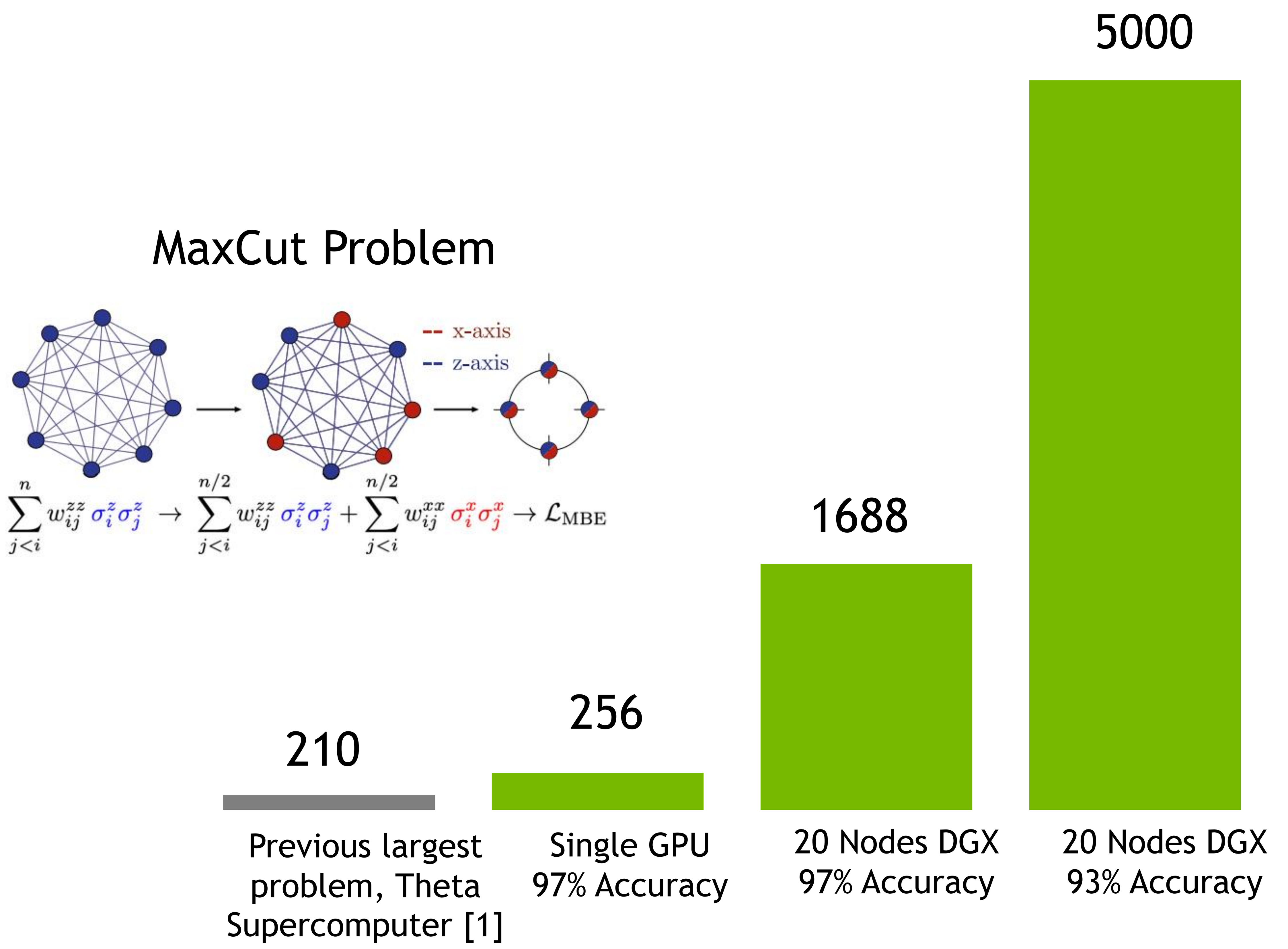
Quantum Simulation Frameworks
Cirq, Qiskit, Pennylane, TensorFlow Quantum, QIBO, …

cuQuantum
cuStateVec, cuTensorNet

GPU Supercomputing



Supercomputing Simulation at FTQC Scale



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

NVIDIA Quantum

Powering the Global Quantum Computing Community

160+

Quantum Partners

>90%

Largest Startups

>78%

QPUs Integrating CUDA Q

15/17

Leading Frameworks Accelerated

QUANTUM HARDWARE BUILDERS

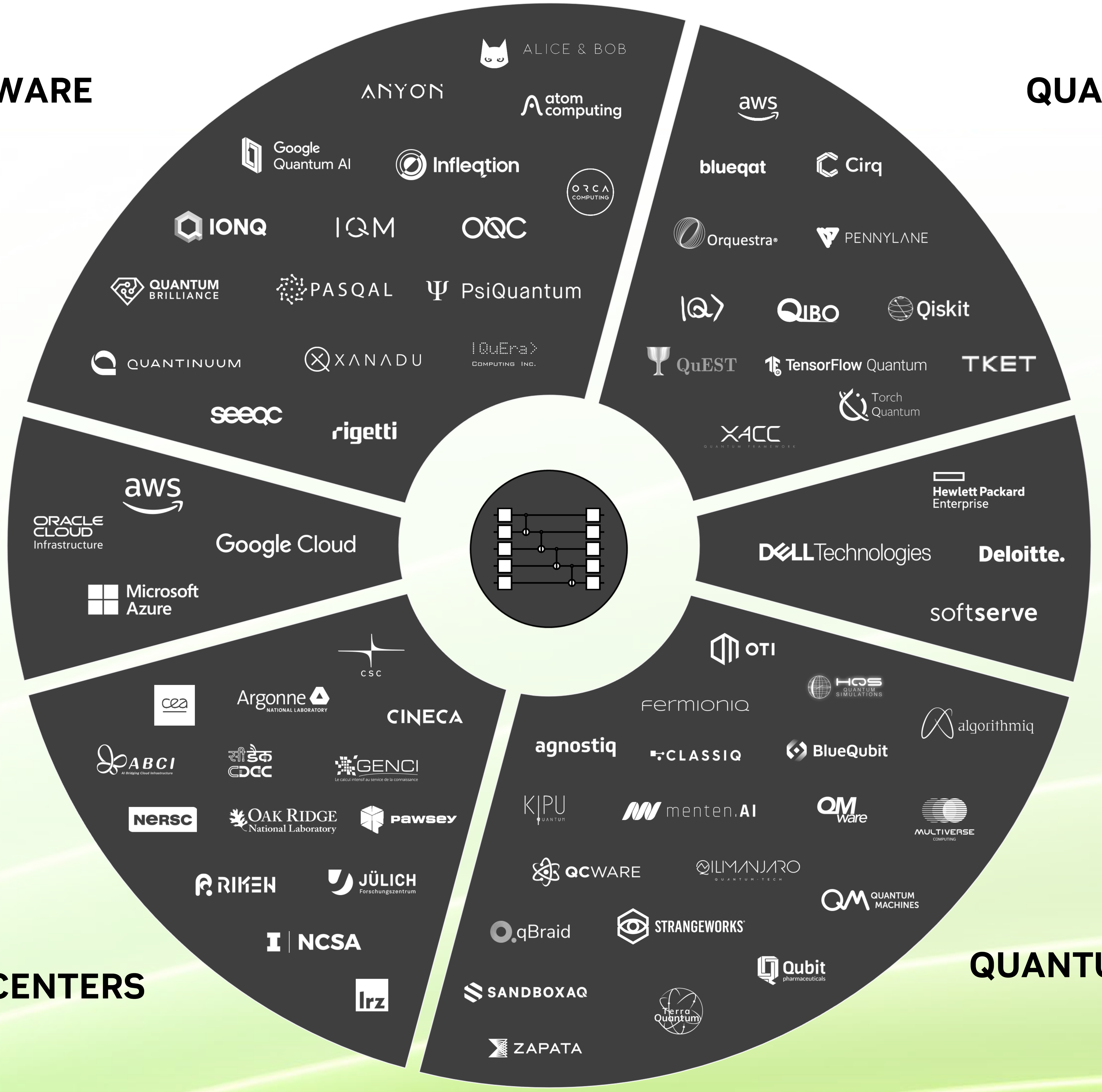
QUANTUM SIMULATION FRAMEWORKS

CLOUD SERVICE PROVIDERS

SYSTEMS BUILDERS & INTEGRATORS

RESEARCH CENTERS

QUANTUM SOFTWARE AND SYSTEMS



NVIDIA Quantum

Thank you

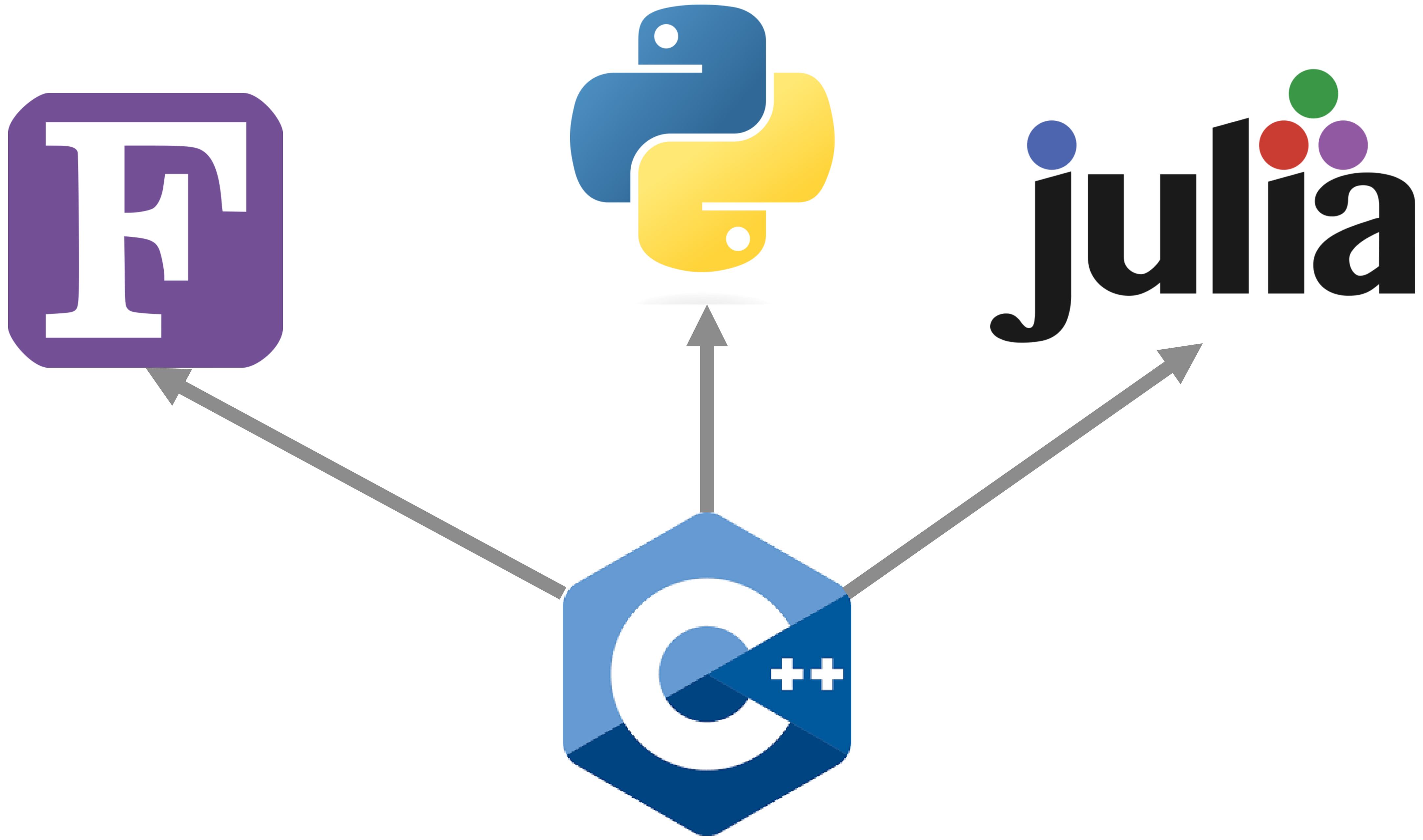
Annex

Further Information and Examples

Requirements for Programming the Hybrid Quantum-Classical Node

What can we learn from experience in the purely classical programming space?

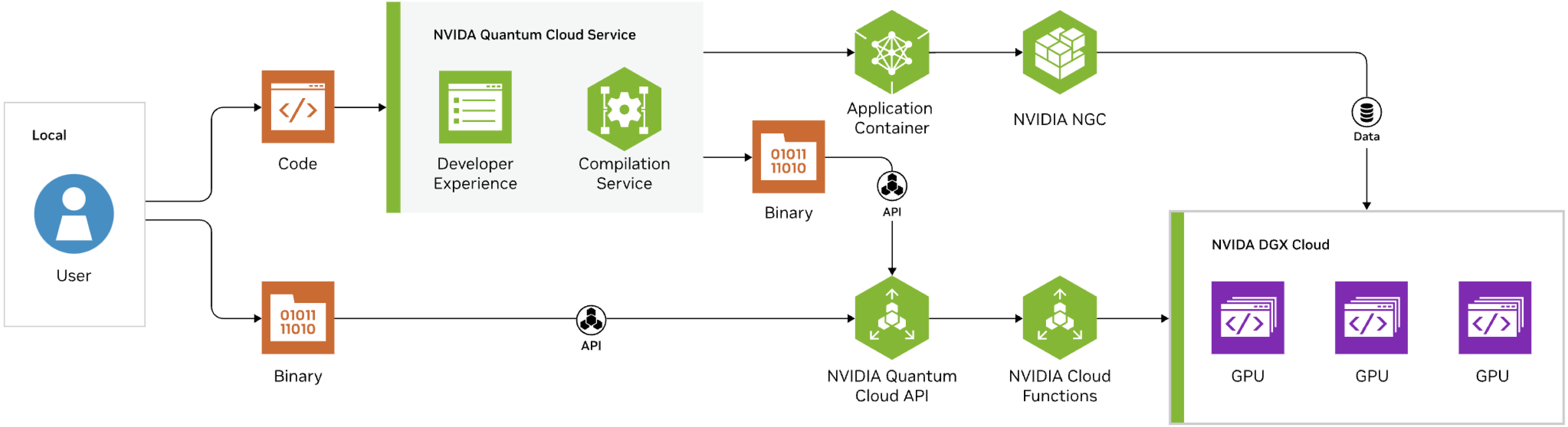
- Requirements
  - Performance
  - Familiar Programming Models
  - Integration with existing compilers and runtimes
- C++ as the Least Common Denominator for Programming Languages
  - Leads to optimal performance / for developers
  - Easily bind to high-level language controlapproaches
  - Most HPC applications are in C++ or Fortran
  - Most AI / ML frameworks are in Python, but APIs are often bound to performant C code (or JIT compiled)
- CUDA-like programming models
  - Cleanly separate device and host code



NVIDIA Quantum Cloud, cuPQC

NVIDIA Quantum Cloud

Quantum Supercomputing Everywhere, for Everyone



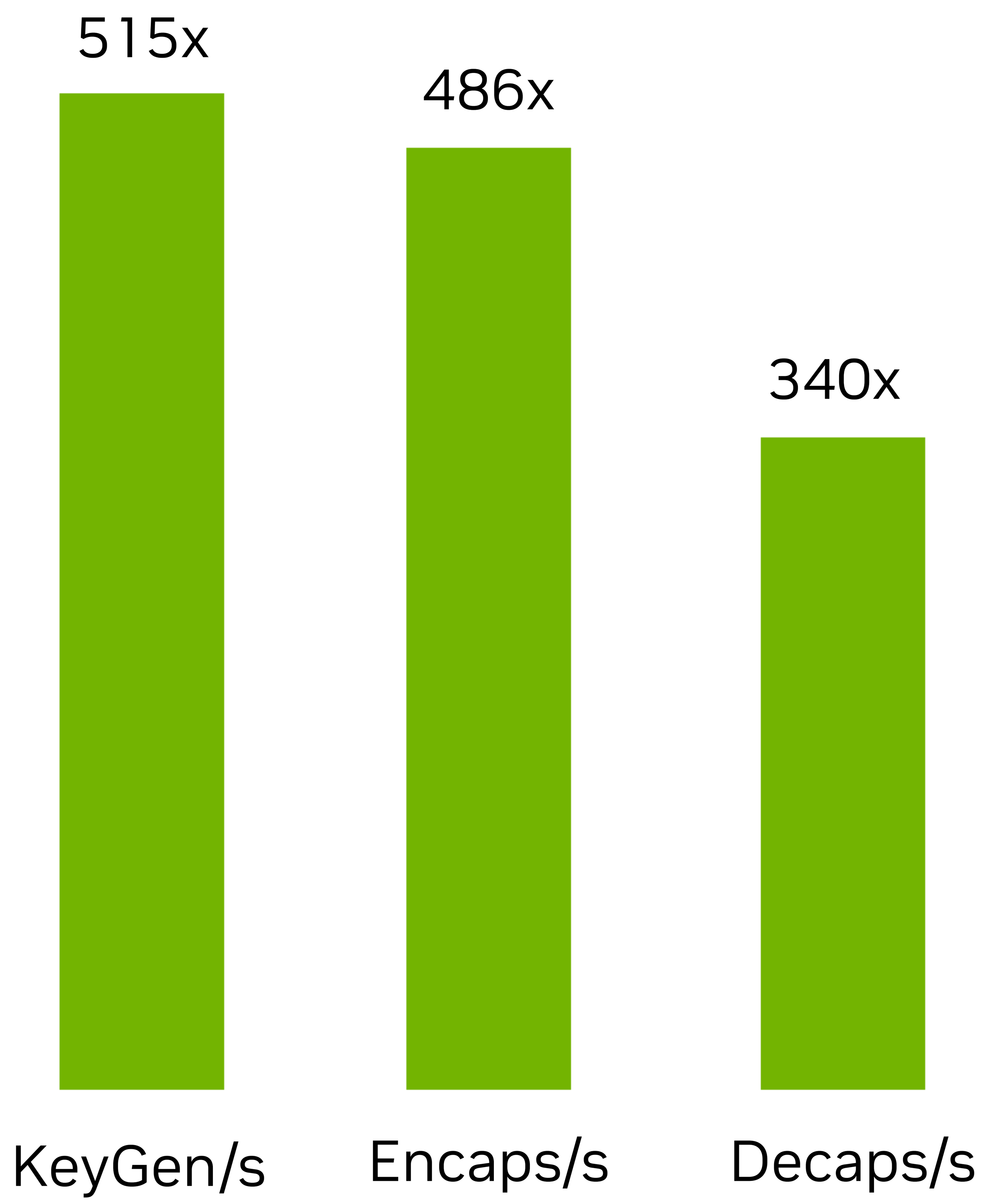
- Access to the most powerful Quantum Resources
- Develop locally, run any CUDA-Q app seamlessly in Cloud
- Call pre-built Quantum Cloud APIs from your application
- Run workloads on GPU Supercomputers
- Integrated ISV applications
- EA Available Today! Apply for Early Access at developer.nvidia.com/quantum-cloud-early-access-join

- QPU Partner Integrations coming soon
- Quantum Cloud Web Developer Portal coming soon

cuPQC

Primitives to Accelerate Quantum Safe Encryption

Kyber768, H100 speedup vs CPU
liboqs benchmarking suite with batching



- Accelerates all NIST Finalists
- Accelerates Kyber768 operations by 515x, 486x, and 340x over SotA
- Will be integrated with liboqs from Open Quantum Safe
- Private beta release available today

evolution OPEN QUANTUM SAFE PQ SHIELD QuSecure SANDBOXAQ™

DGX Quantum

First DGX-Q Deployment

Israeli Quantum Computing Center to Deploy World's First Tightly Integrated Quantum-Classical System

- IQCC to deploy DGX-Q, combining NVIDIA Grace Hopper Superchips with Quantum Machines OPX Quantum Control
- System to be connected to multiple QPUs with sub-microsecond latency from GPU to QPUs from ORCA, Quantware and Rigetti
- Enables research in Real-Time Accelerated Error Correction

QUANTUM MACHINES

ORCA COMPUTING

rigetti

"With DGX Quantum, QM-NVIDIA collaboratively develop a game-changing capability that's essential to reach quantum advantage. We are thrilled about this technology, which will enable quantum computer builders and researchers to unleash the next wave of massive performance improvements. Pioneers in quantum error mitigation and quantum error correction now have a brand-new playing field thanks to ultra low latency feedback and high throughput processing."

- Itamar Sivan, CEO of Quantum Machines

Some CUDA-Q Use Cases

NVIDIA

60 qubit chemical simulation for detecting toxic metals – MPS methods

BASF
We create chemistry

QML for automotive optimizations, 8 hours to 2 minutes on DGX A100

NVIDIA

Fastest quantum simulation in the world for portfolio optimizations on Polaris

JPMorgan Chase & Co.

NVIDIA

Quantum methods to advance the development of jet engines – largest CFD simulation in the world

ROLLS ROYCE