NCHC Tech Sharing - NVIDIA Modulus and cuQuantum

Dr. CK Lee | Scientist of NVAITC Taiwan @ NVIDIA

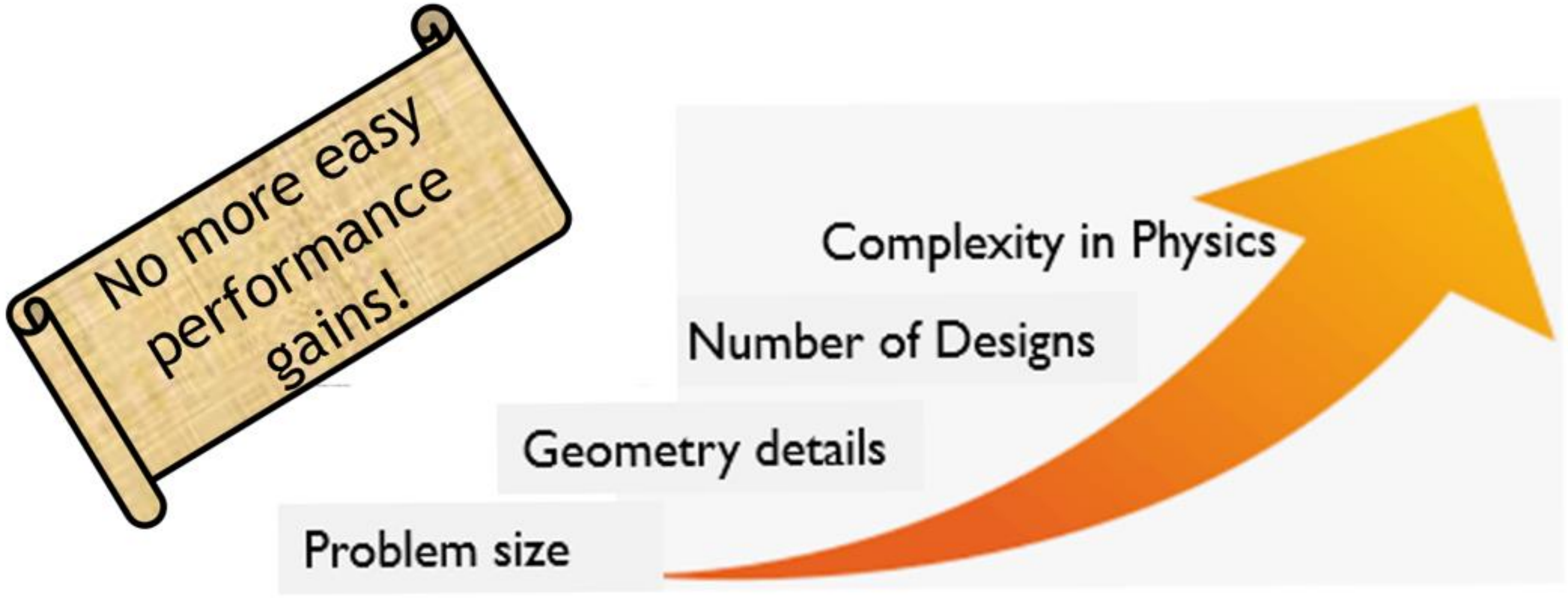
Outline

How AI Technology to Enable Net Zero - Modulus

What's inside CUDA Quantum

SATURATING PERFORMANCE IN TRADITIONAL HPC

Simulations are getting larger & more complex



Traditional simulation methods are:

- Computationally expensive
- Demand ever-increasing resolution
- Plagued by domain discretization techniques
- Not suitable for data-assimilation or inverse problems

SOLVING PDES WITH NEURAL NETWORKS

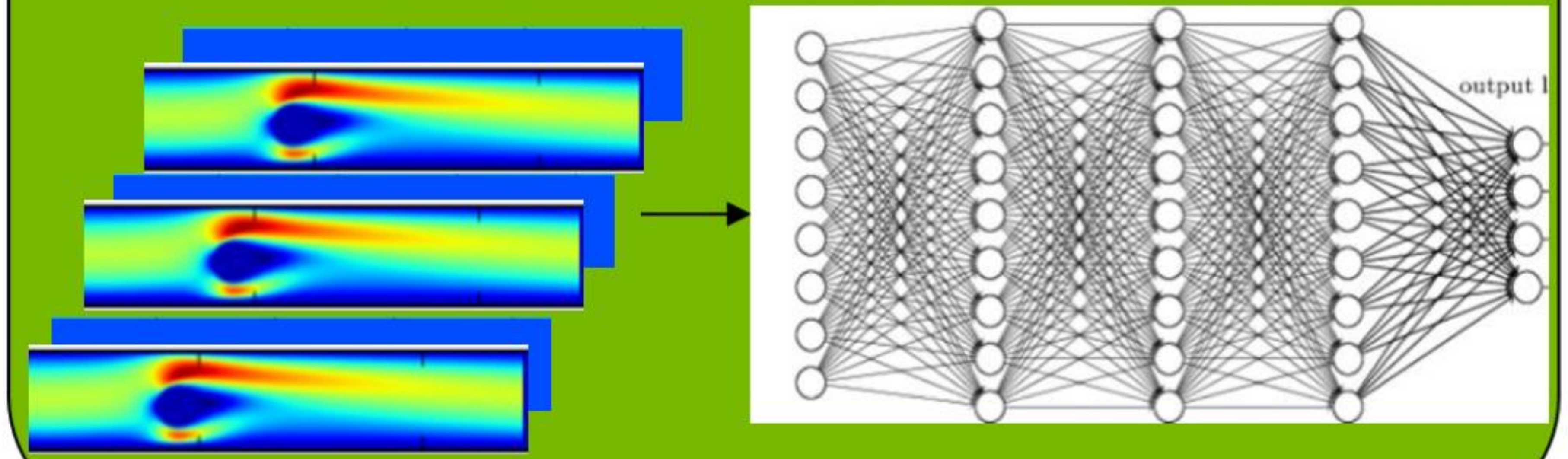
A **Data Driven Neural Network** requires training data

A **Physics Driven Neural Network** solver does **NOT** require training data

Supervised Data Driven Approach

Simulation Dataset

Neural Network



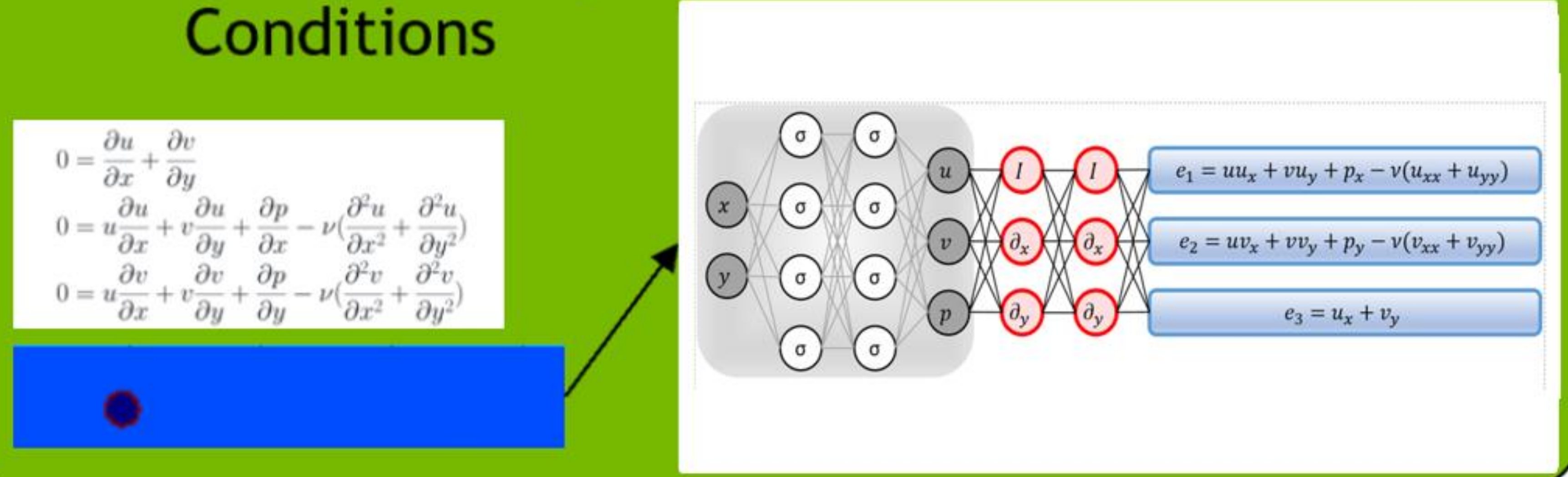
N layers

Unsupervised Physics Driven Approach

PDE and Boundary Conditions

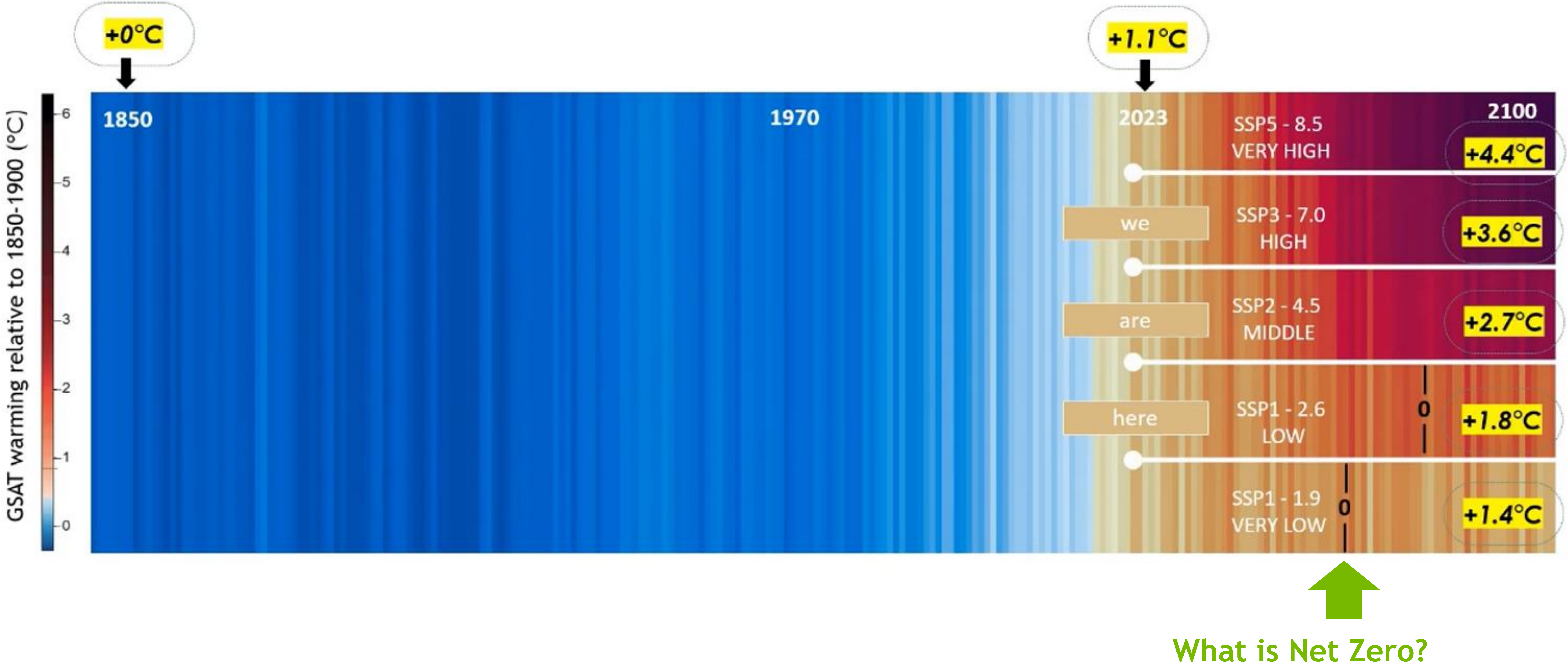
Neural Network

$$0 = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$$
$$0 = u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + \frac{\partial p}{\partial x} - \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$
$$0 = u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + \frac{\partial p}{\partial y} - \nu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right)$$



m*N layers (for m^{th} order PDE)

Acting Against Climate Change, and its Challenges



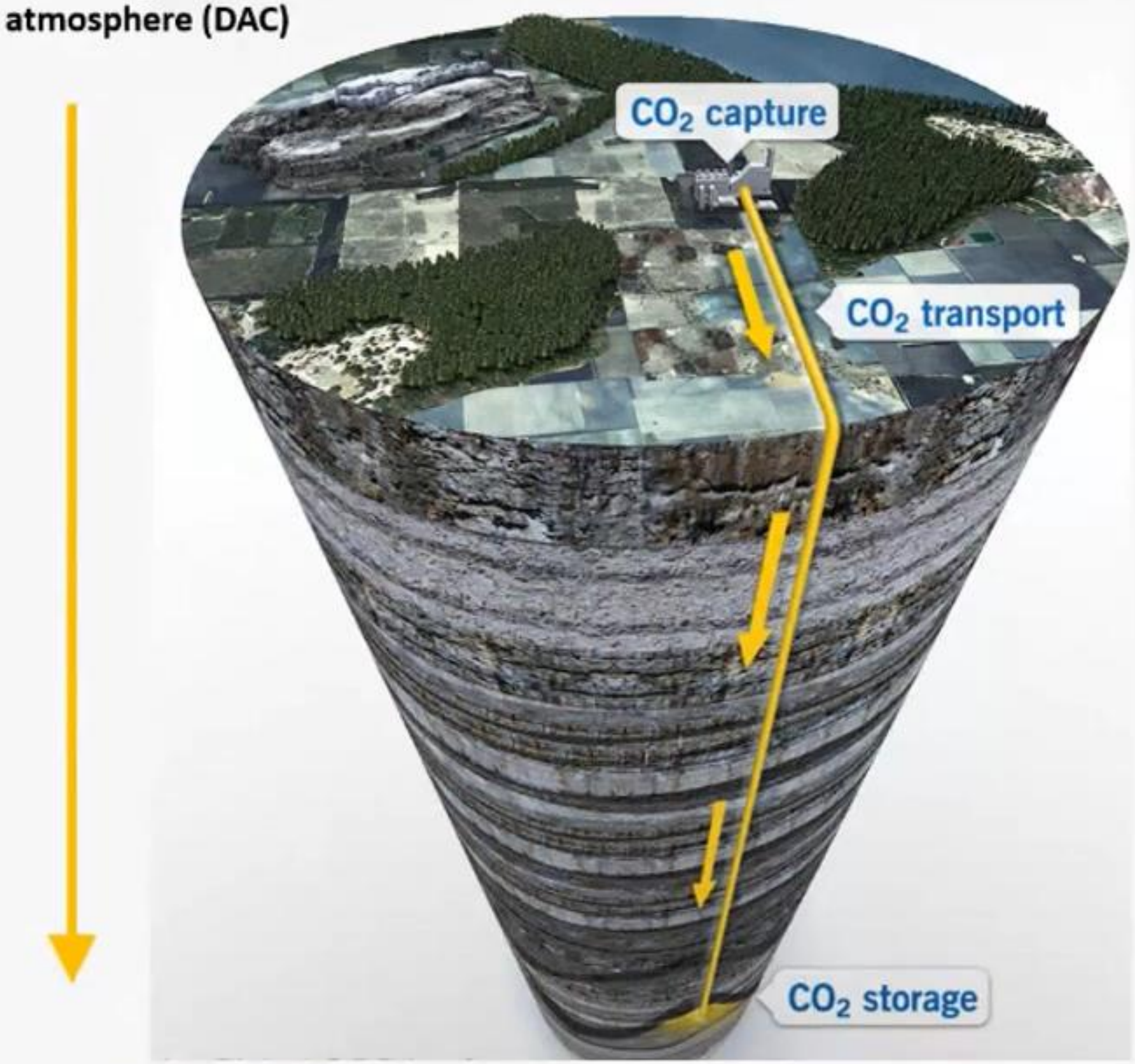
Source: www.ipcc.ch

The Role of Carbon Capture and Storage to Achieve Net Zero

- IEA reports the need to store 1200 Mt CO_2/year by 2030
- The current storage = 10 Mt CO_2/year
- No problem related to storage capacity, which is estimated at several thousand Gtones of CO_2
- Challenges remain to prove duration and reliability of storage for which numerical modeling is key
- Time counts: each year gained by modeling CCS faster through Artificial Intelligence, is less CO_2 in the atmosphere!

Capture CO_2 from:

- fossil fuel combustion
- heavy industry processes
- bioenergy combustion
- atmosphere (DAC)



Whatever is the capture technique,
=> storing the captured CO_2 is needed

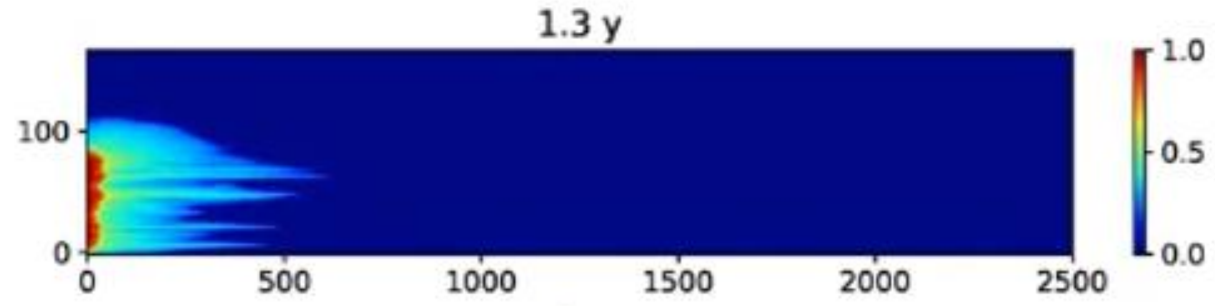
Provided by the Global CCS Institute

CLIMATE CHANGE MITIGATION: MODELING CO_2 STORAGE

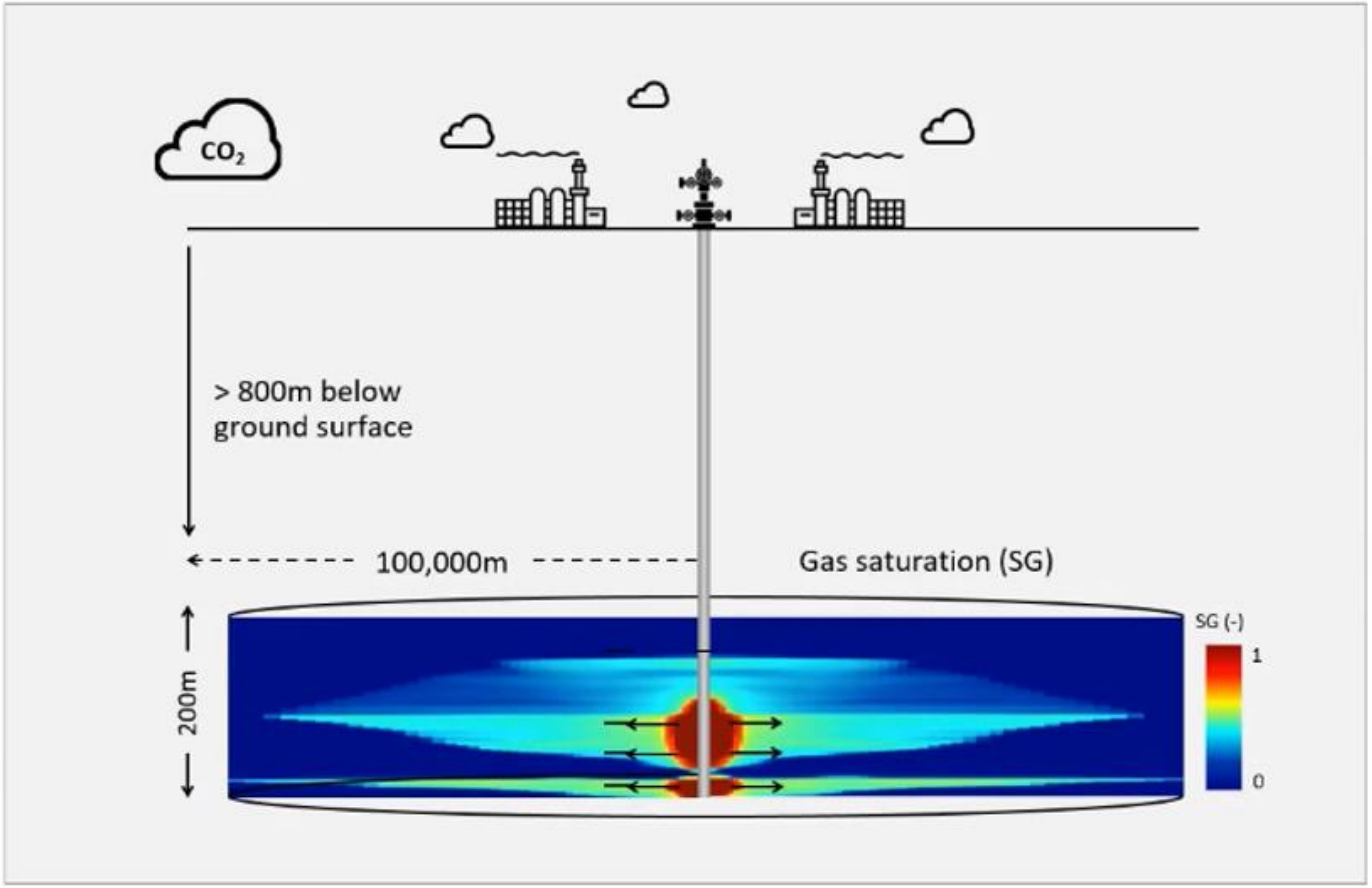
A study of challenging subsurface process

Inject CO_2 into subsurface

1. Store CO_2 in a saline formation
2. High pressure storage medium
3. Multi phase salt water and CO_2 interaction
4. High resolution details
5. Storage plans for decades to centuries



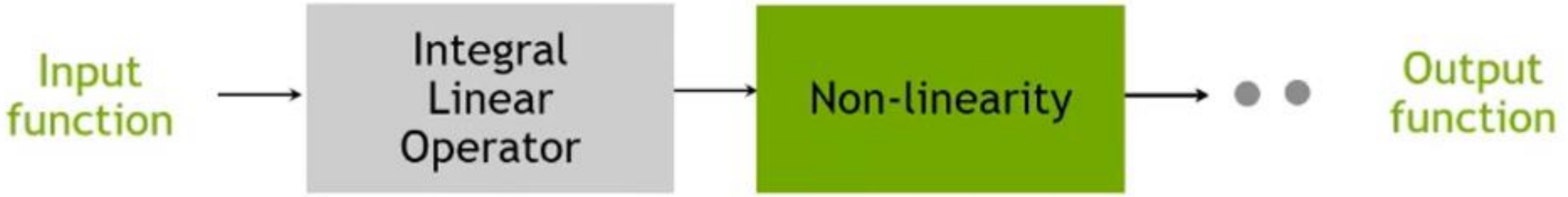
Highly costly and slow using traditional simulations



NVIDIA

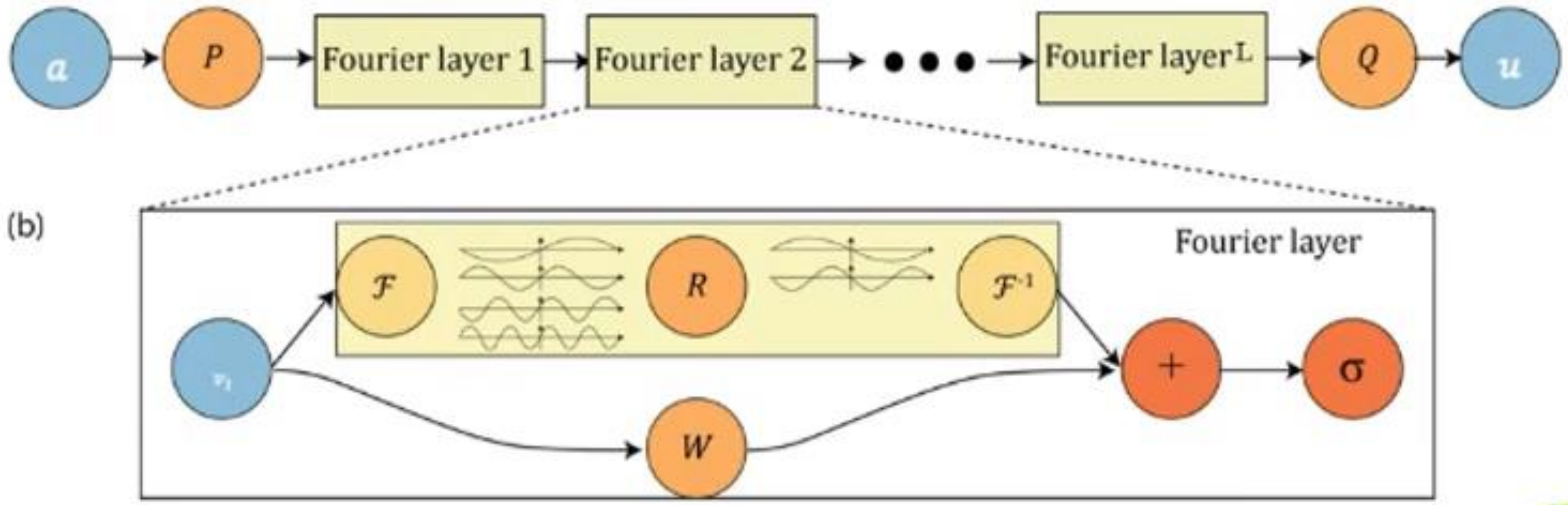
NUERAL OPERATORS: DEEP LEARNING ON FUNCTION SPACES

Neural operators learn maps between functions



$$\int \kappa(x, y)\, v(y)\, dy$$

- Integral operator outputs functions.
- Integral operator is discretization invariant.
- Fourier neural operators (FNO)



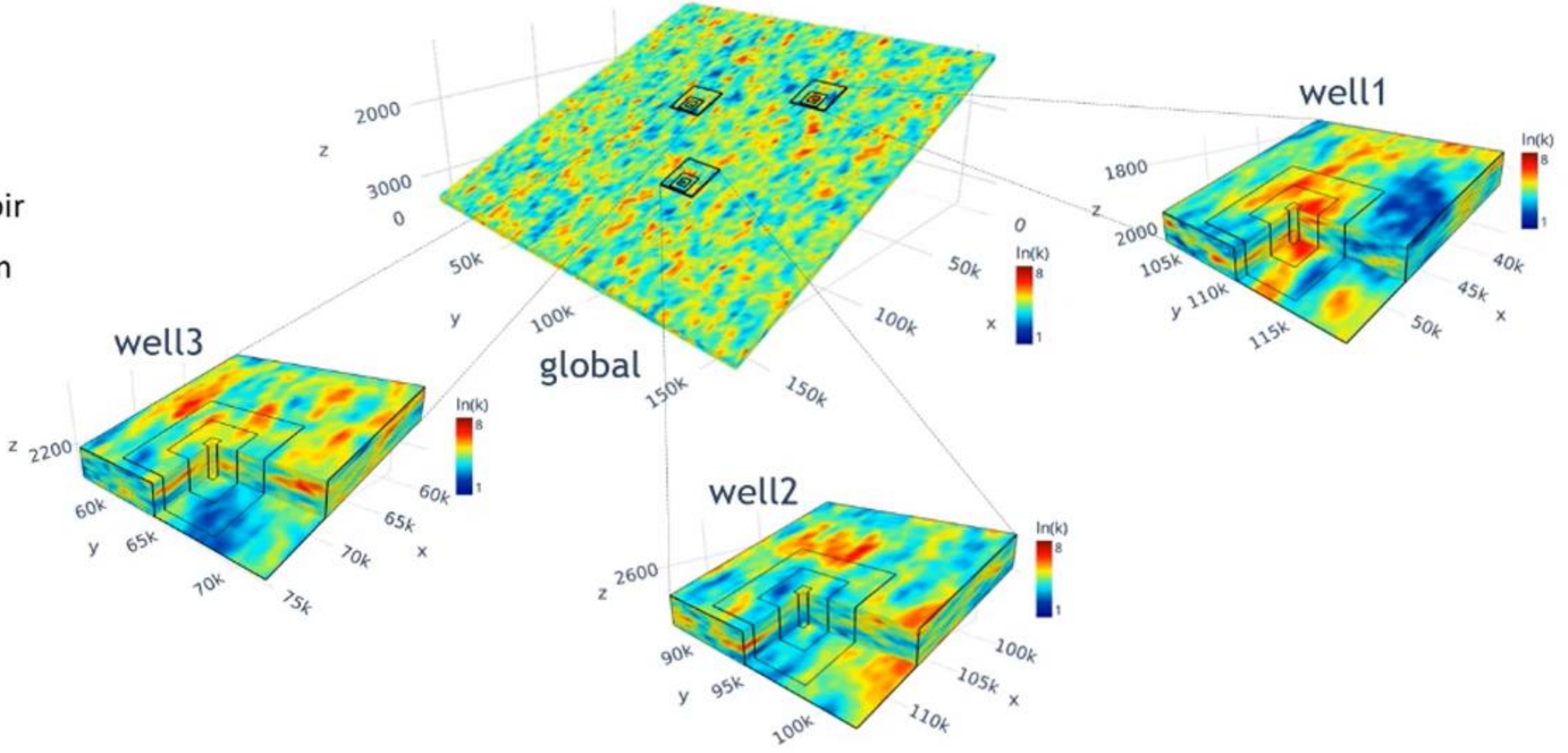
NVIDIA

FOUR-DIMENSIONAL CCS MODELING WITH AI (FNO)

Four-dimensional large scale 200km spatiotemporal CCS

Real-world CCS

1. Multi-well CCS
2. Dipped reservoir
3. Inhomogeneous reservoir
4. High resolution up to 1m
5. Multi detailed physics



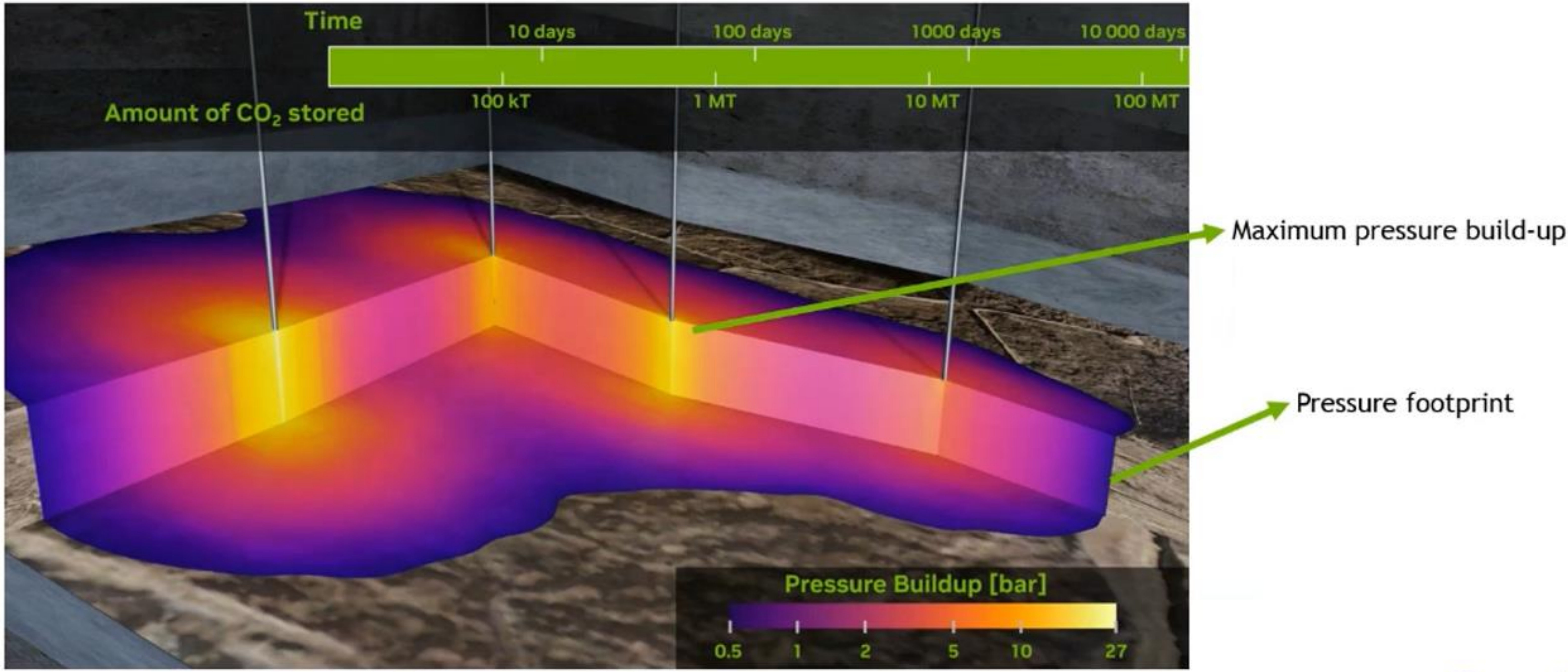
Permeability Heat Map

NVIDIA

NESTED FNO AND ACCURATE INFERENCE

Inference in subsurface Earth

Accurate estimation of engineering and scientific quantities



… and Receive Useful, Visual & Statistical Guidance?

From a Highly Interactive Future Climate Information System, at High Resolution, that Serves Society…

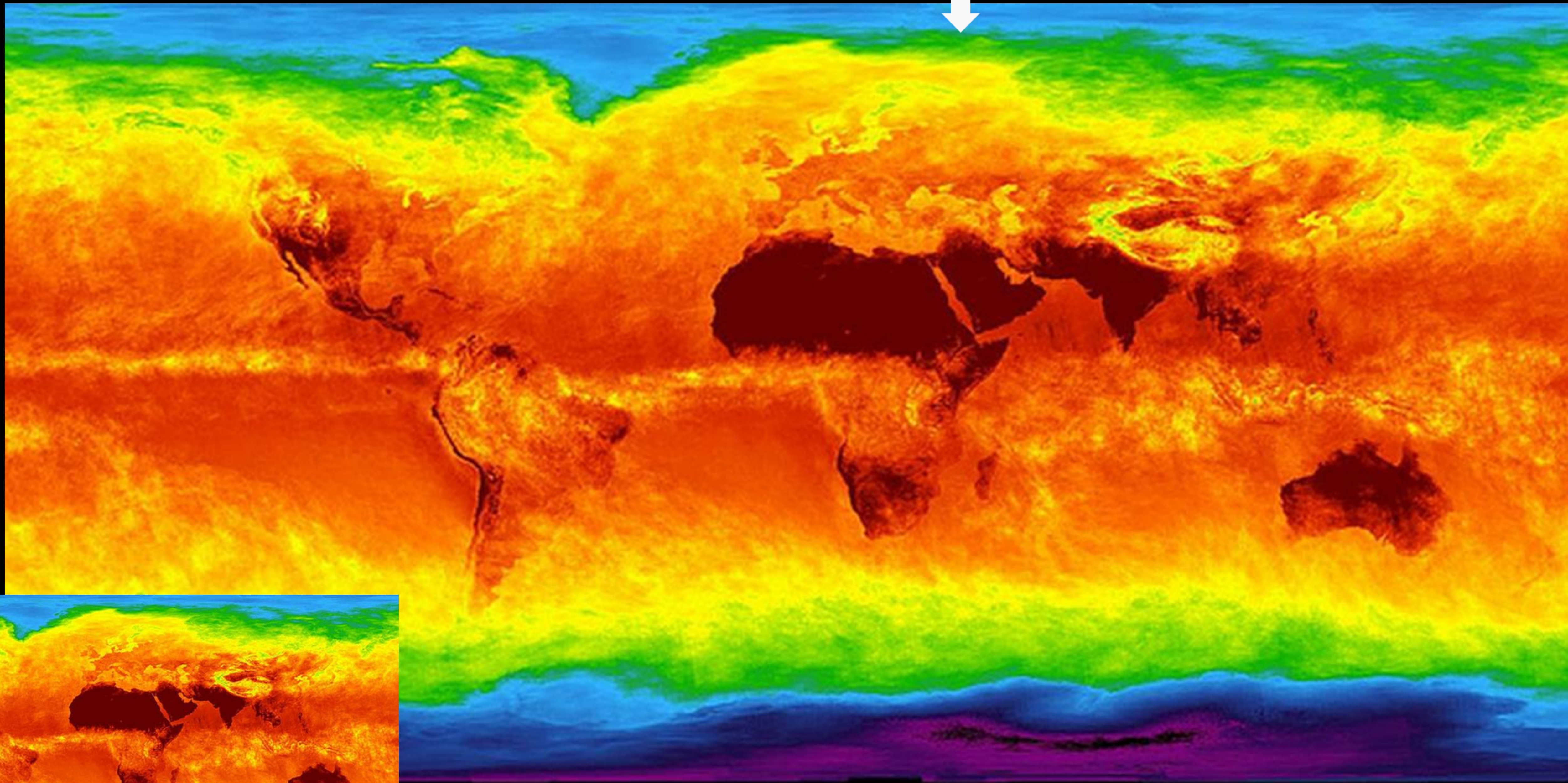
Earth-2 Mission #1

Interacting with Climate Predictions at Low Latency.

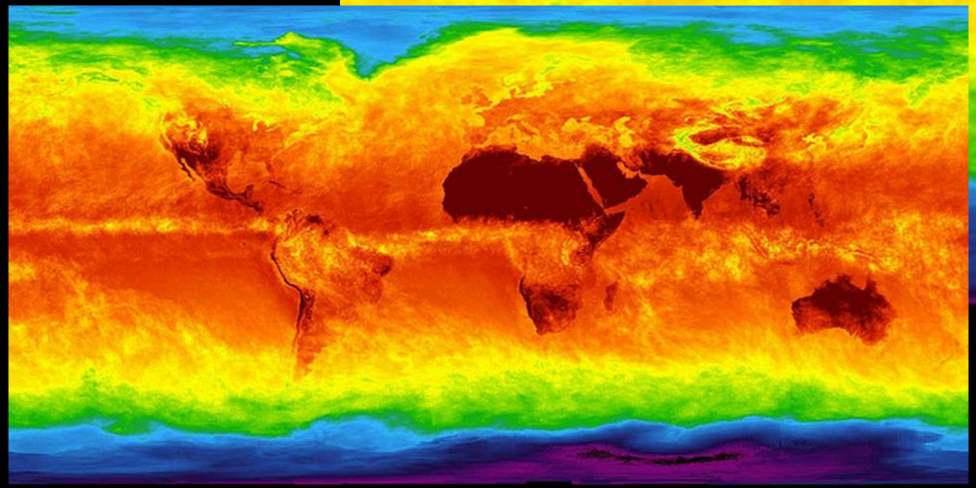
Earth-2 Mission #2

Achieving Next-Gen Climate Predictions using Hybrid Physics, Machine Learning & HPC.

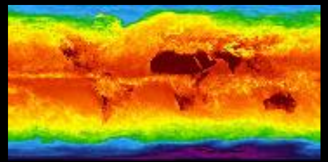
FourCastNet: NV's DDWP, first to be trained at ambitious 0.25-deg global resolution



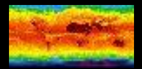
FourCastNet, Pathak et al. (2022), 0.25°, ~1,000,000 Pixels, ViT+AFNO



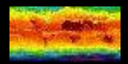
GNN, Keisler et al. (2022), 1°, 64,000 Pixels, Graph Neural Networks



DLWP, Weyn et al. (2020). 2°, 16K pixels, Deep CNN on Cubesphere/(2021) ResNet

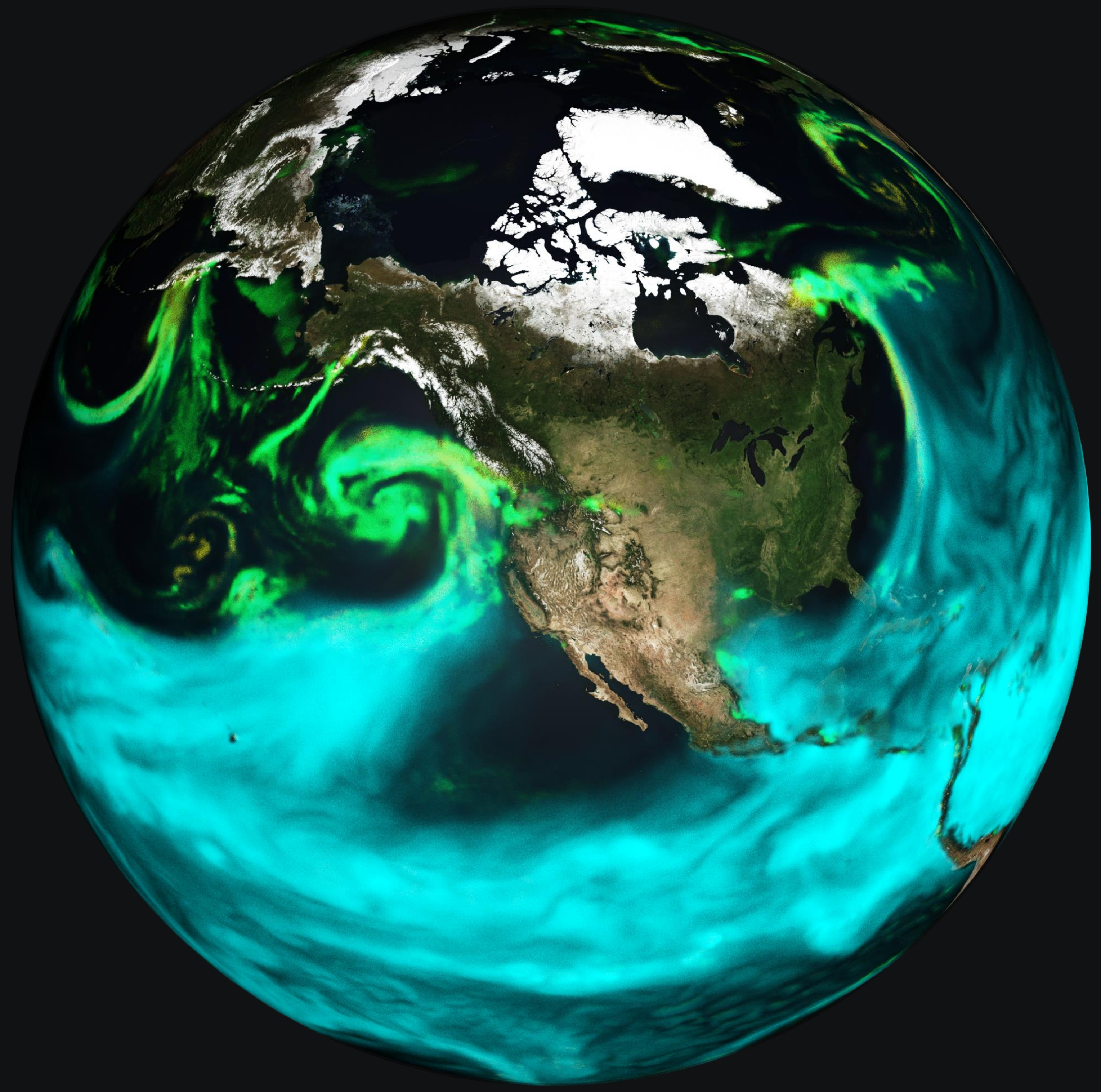


Weyn et al. (2019), 2.5° N.H only, 72x36, 2.6k pixels, ConvLSTM



WeatherBench, Rasp et al. (2020). 5.625°, 64x32, 2K pixels, CNN

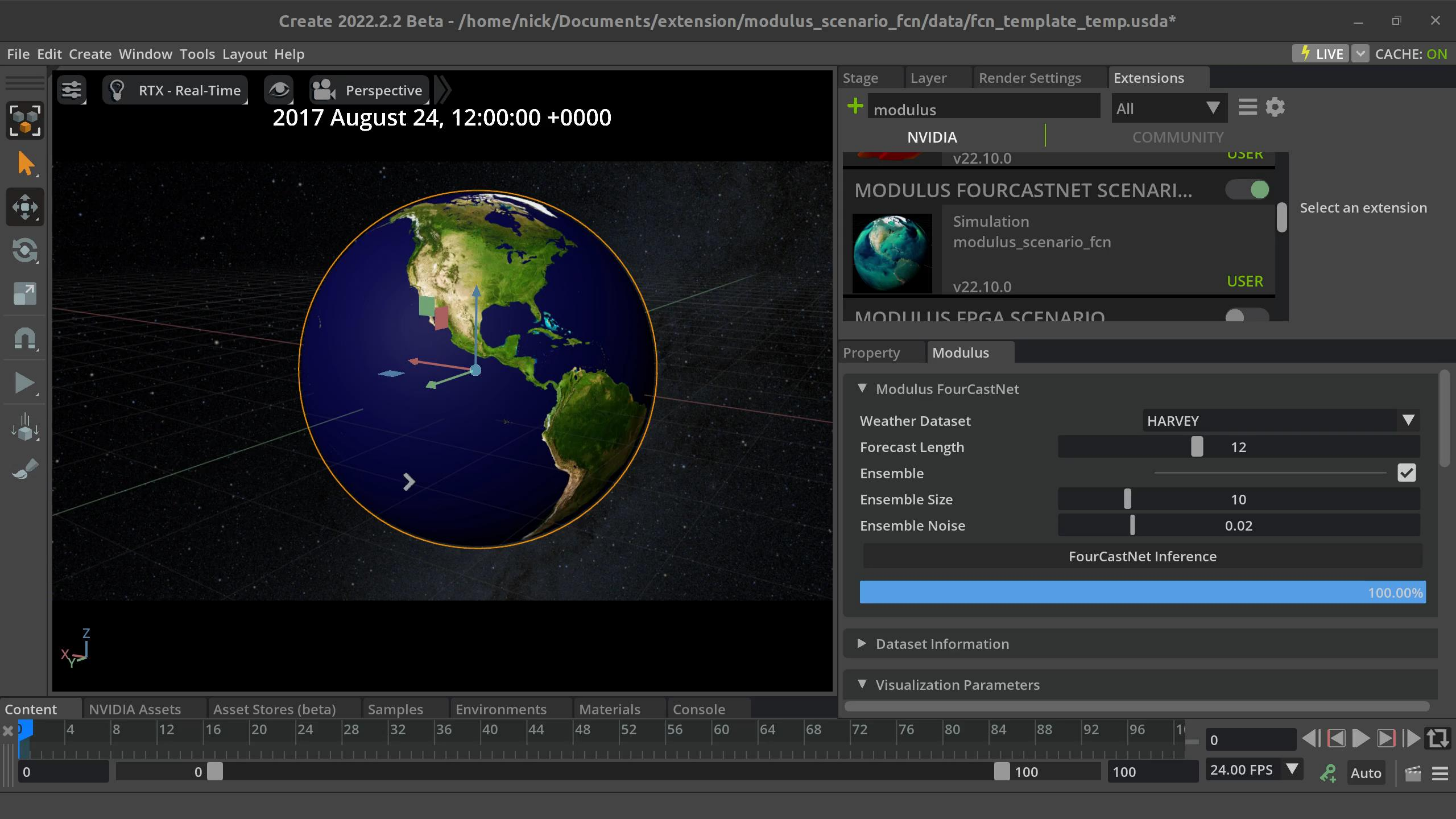
Deuben & Bauer (2018), 6°, 60x30, 1.8K pixels, MLP



Weather Simulation with FourCastNet

Fully data-driven weather prediction.

Scope	Global, Medium Range
Model Type	Full-Model AI Surrogate
Architecture	AFNO (Adaptive Fourier Neural Op.)
Resolution:	25km
Training Data:	ERA5 Reanalysis
Initial Condition	GFS / UFS
Inference Time	0.25 sec (2-week forecast)
Speedup vs NWP	$O(10^4\text{-}10^5)$
Power Savings	$O(10^4)$



Create 2022.2.2 Beta - /home/nick/Documents/extension/modulus_scenario_fcn/data/fcn_template_temp.usda*
File Edit Create Window Tools Layout Help
LIVE
CACHE: ON
RTX - Real-Time
Perspective
2017 August 24, 12:00:00 +0000
Stage
Layer
Render Settings
Extensions
modulus
All
NVIDIA
COMMUNITY
v22.10.0
MODULUS FOURCASTNET SCENARI...
Simulation
modulus_scenario_fcn
v22.10.0
USER
Select an extension
Property
Modulus
Modulus FourCastNet
Weather Dataset
HARVEY
Forecast Length
12
Ensemble
Ensemble Size
10
Ensemble Noise
0.02
FourCastNet Inference
100.00%
Dataset Information
Visualization Parameters
Content
NVIDIA Assets
Asset Stores (beta)
Samples
Environments
Materials
Console
24.00 FPS
Auto

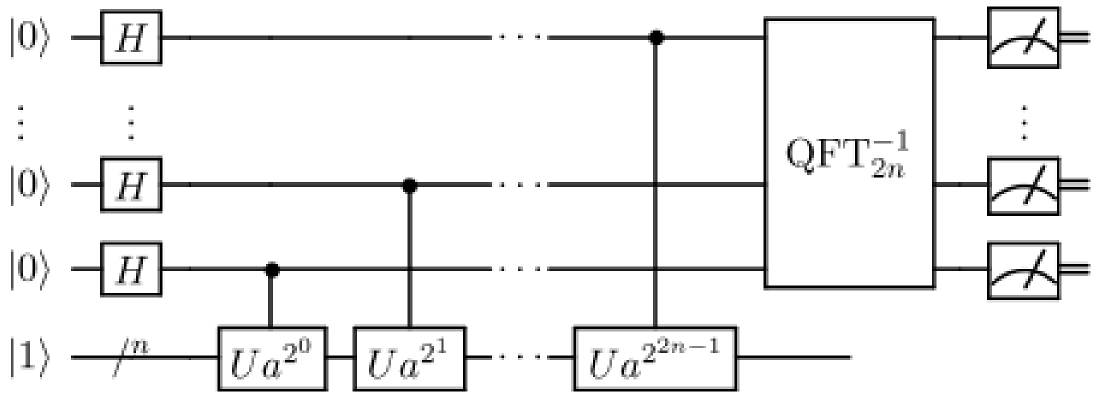
cuQuantum & CUDA Quantum

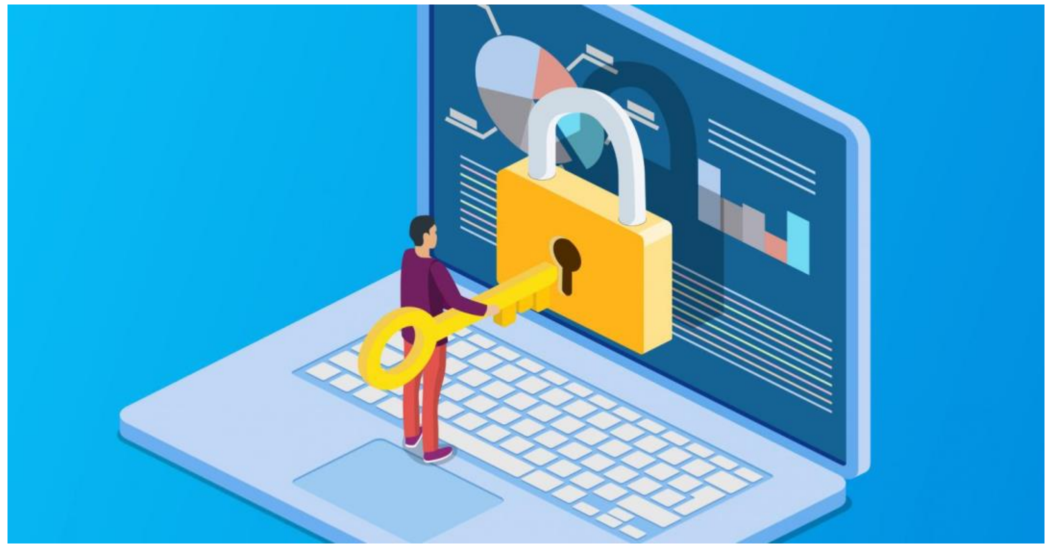
FAR TERM APPLICATIONS

Rigorous proofs of advantage, many “perfect” qubits required

SHOR’S ALGORITHM

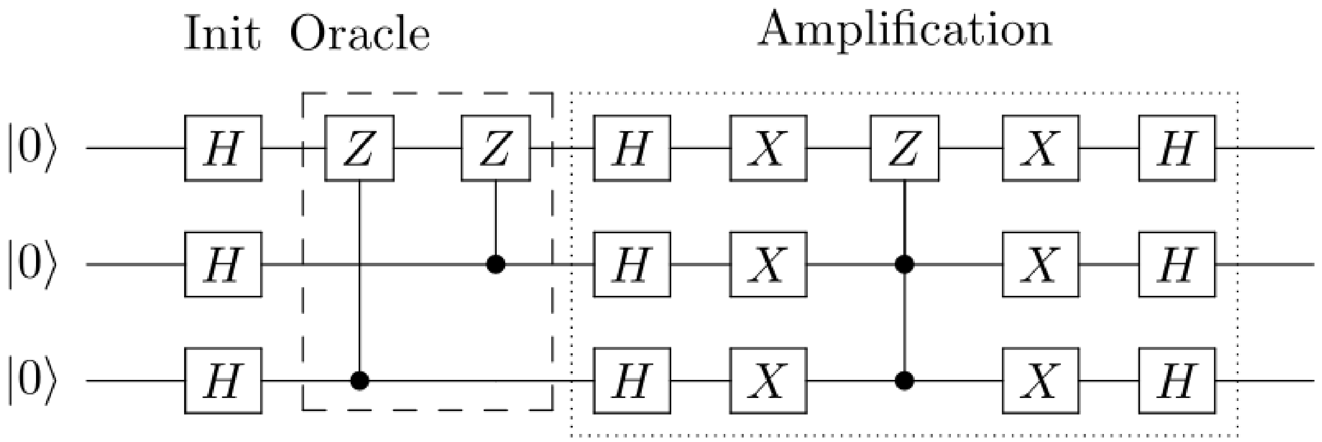
- Prime factorization of numbers - encryption
- Exponential speed-up

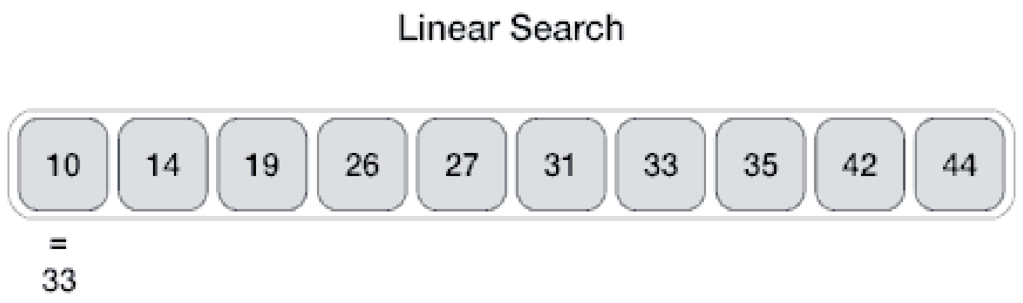




GROVER’S ALGORITHM

- Unstructured search
- Quadratic speed-up



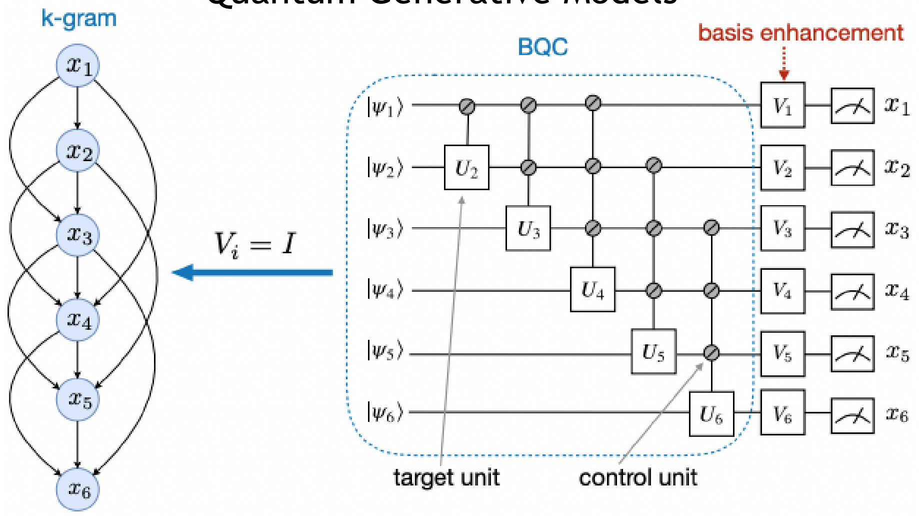


POTENTIAL NEAR-TERM QUANTUM USE-CASES

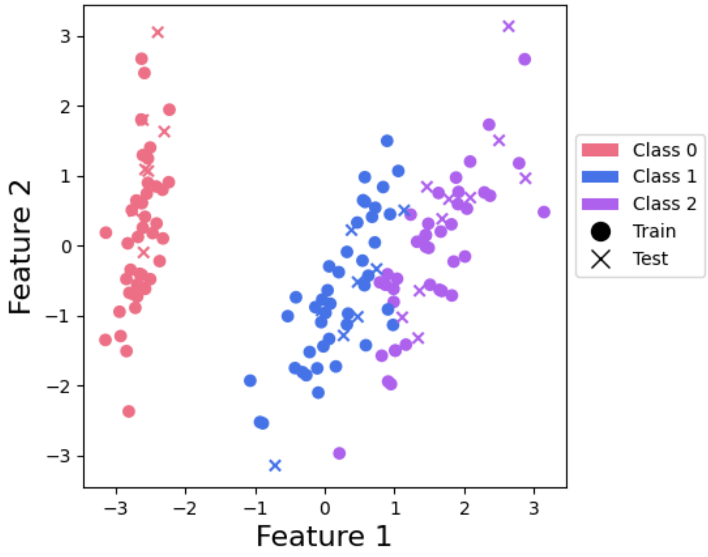
Applications with near-term potential, but quantum advantage is an open question

Quantum Machine Learning

Quantum Generative Models



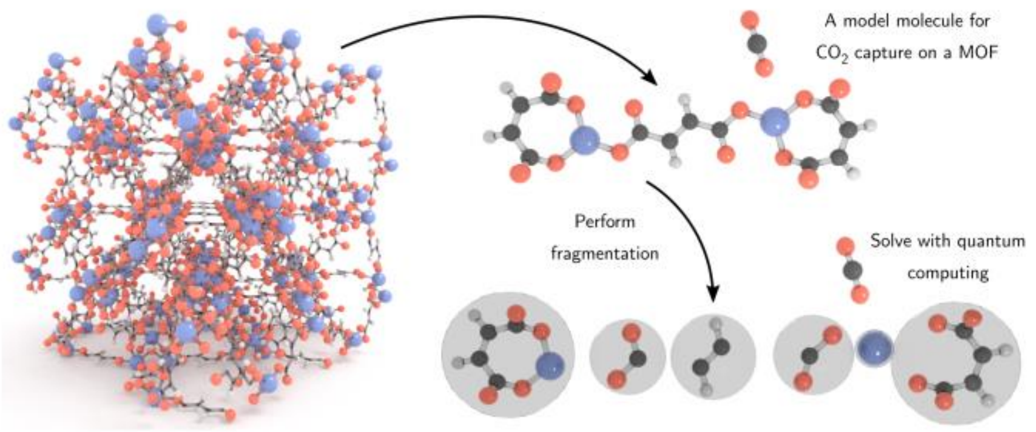
Quantum Support Vector Machine



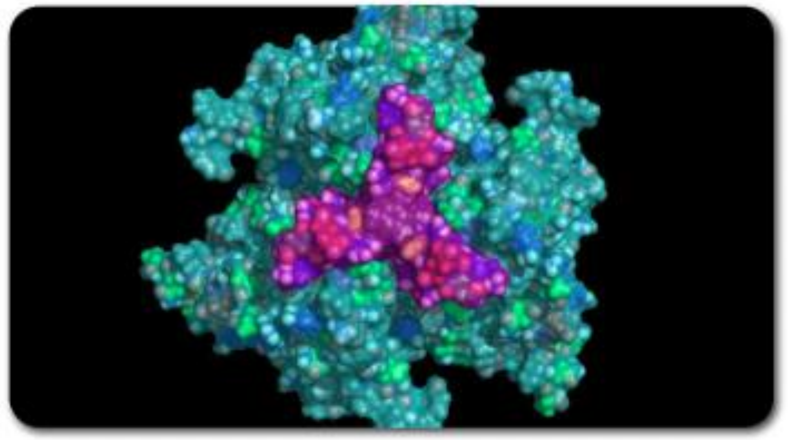
Gao, et al, Phys. Rev. X **12**, 021037
Pennylane.ai

Quantum Chemistry

Variational Quantum Eigensolver for carbon capture



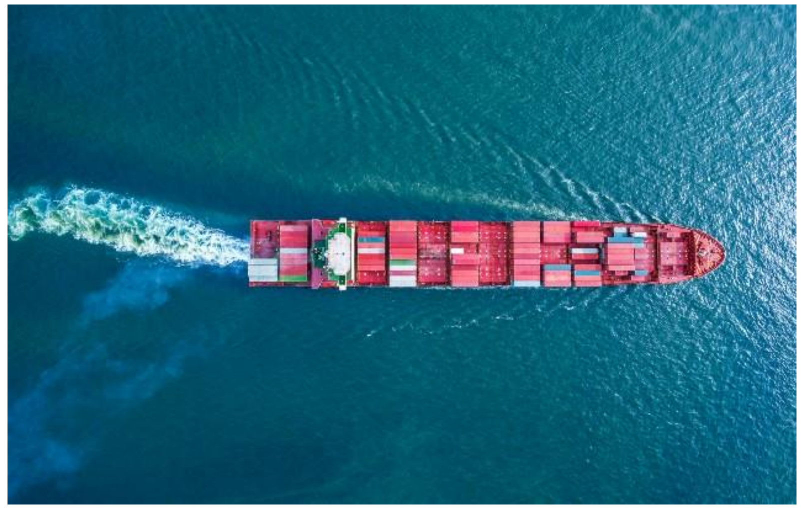
Protein folding



Greene-Diniz, et al, arXiv:2203.15546,
Menten.ai

Combinatorial Optimization

QAOA for resource allocation



Logistics optimization

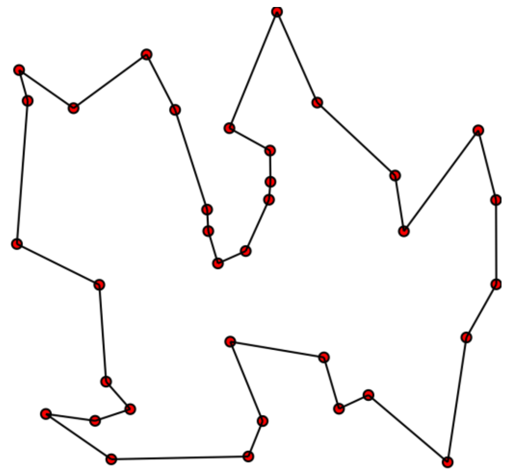


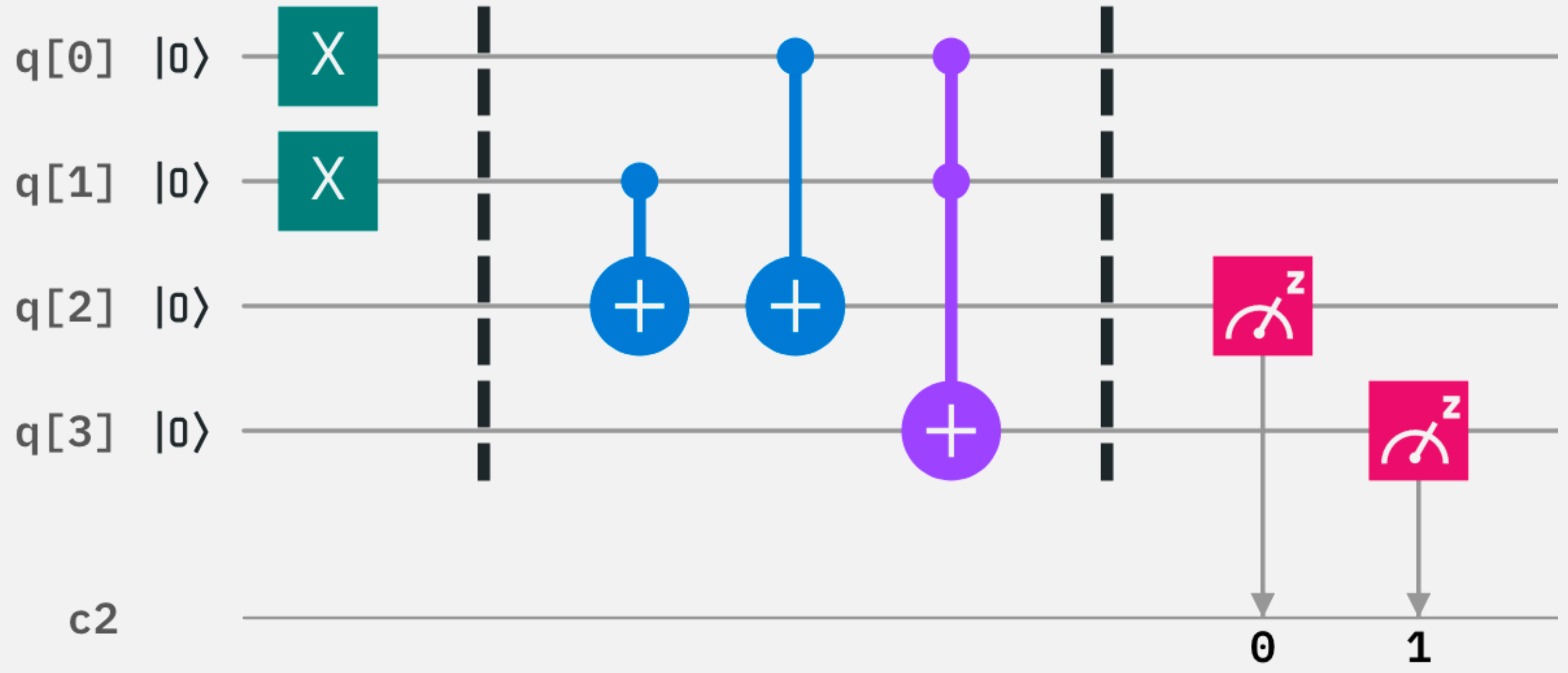
Image from ibm.com
Wikipedia.com

GPU-BASED SUPERCOMPUTING IN THE QUANTUM COMPUTING 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?

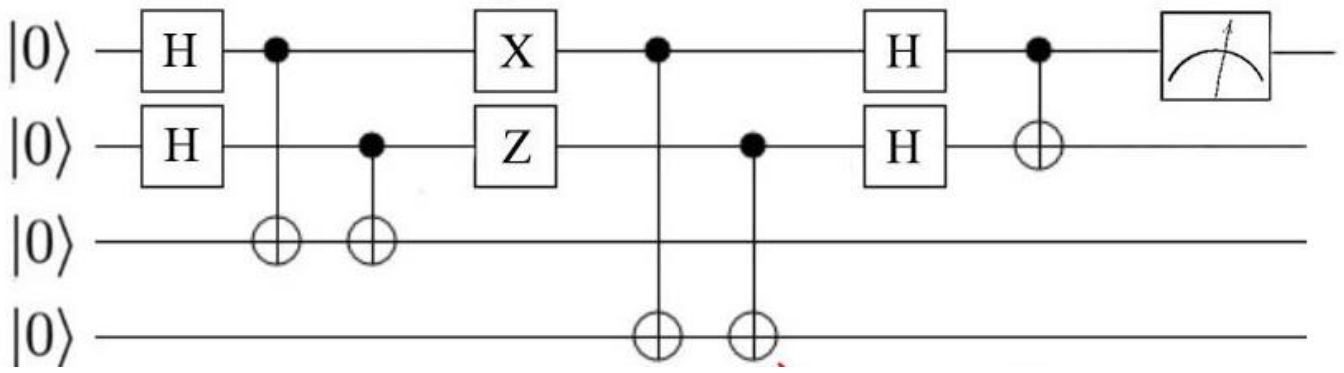
HYBRID CLASSICAL/QUANTUM APPLICATIONS

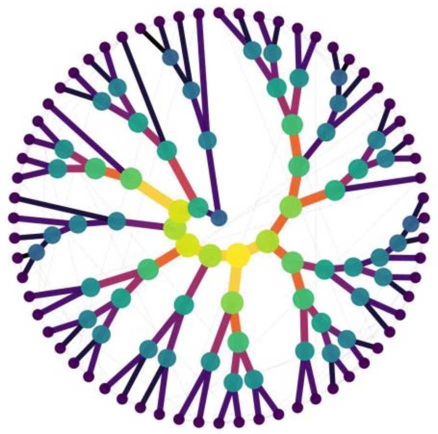
Impactful QC applications (e.g. simulating quantum materials and systems) will require classical supercomputers with quantum co-processors

+

- How can we integrate and take advantage of classical HPC to accelerate hybrid classical/quantum workloads?
- How can we allow domain scientists to easily test coprogramming of QPUs with classical HPC systems?
- Can we take advantage of GPU acceleration for circuit synthesis, classical optimization, and error correction decoding?

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

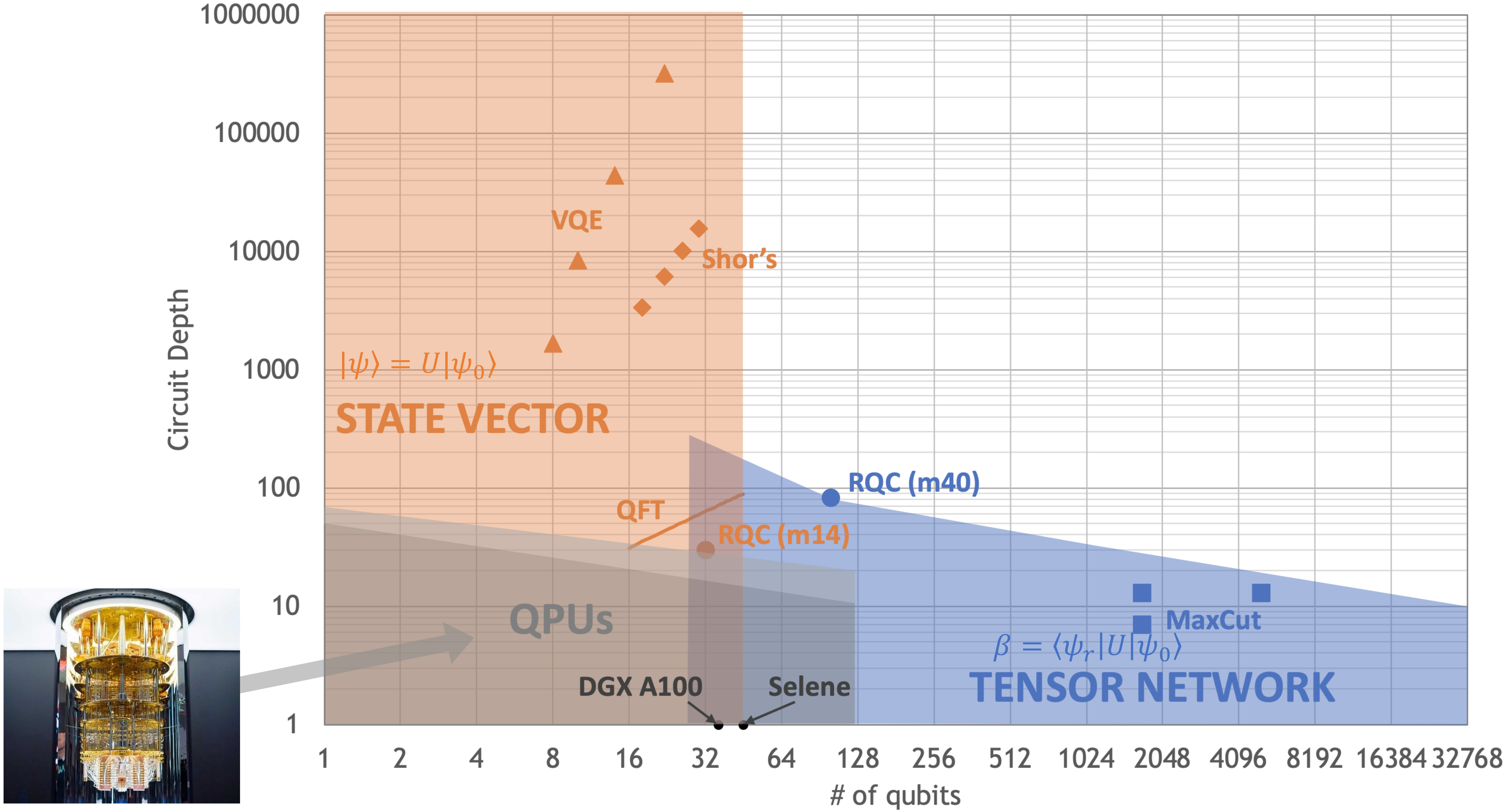
“Only simulate the states you need”

- 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

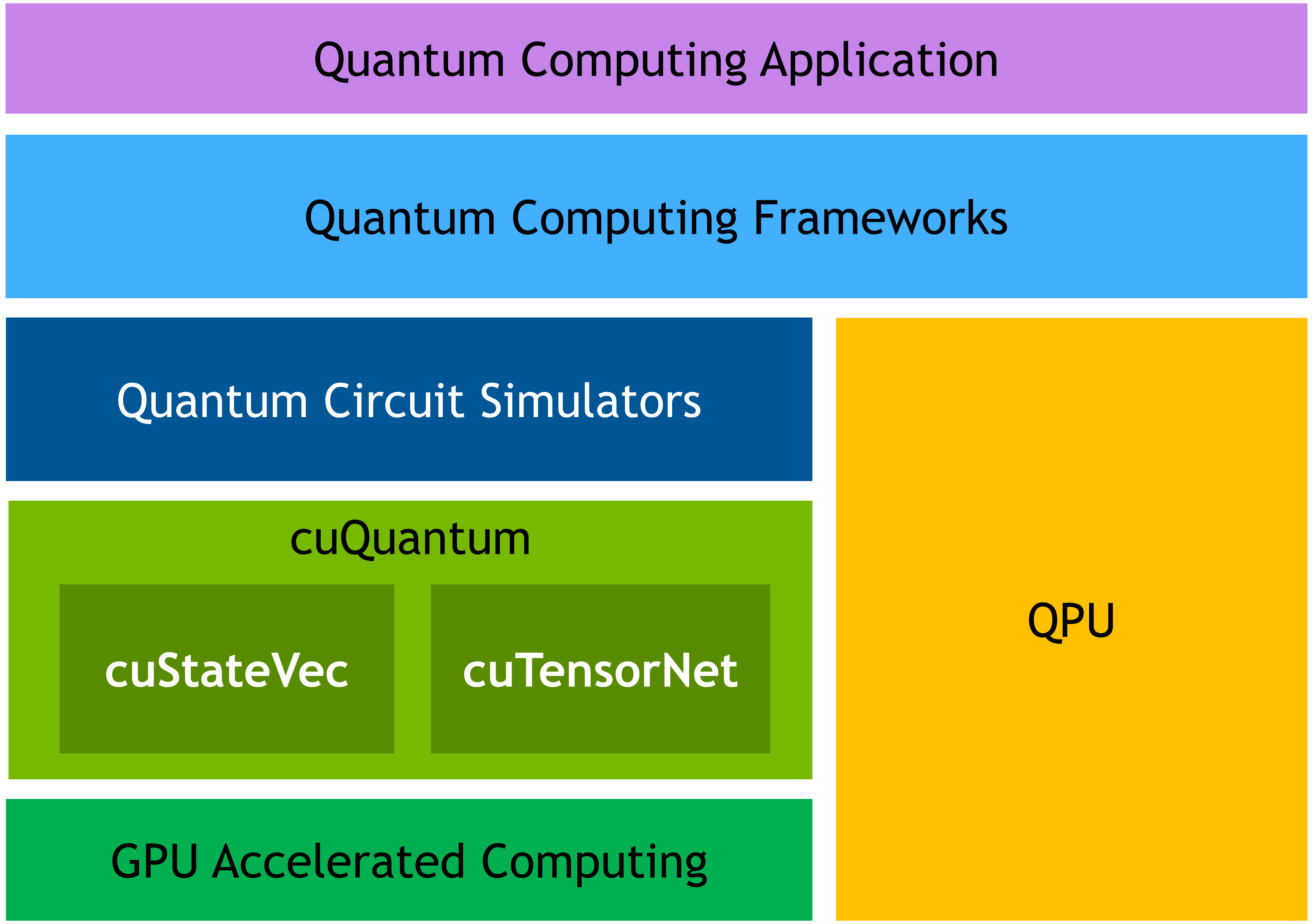
POWERFUL SIMULATIONS WITH cuQuantum

Researching & Developing the Computers of Tomorrow Requires Powerful Simulations Today



cuQuantum OVERVIEW

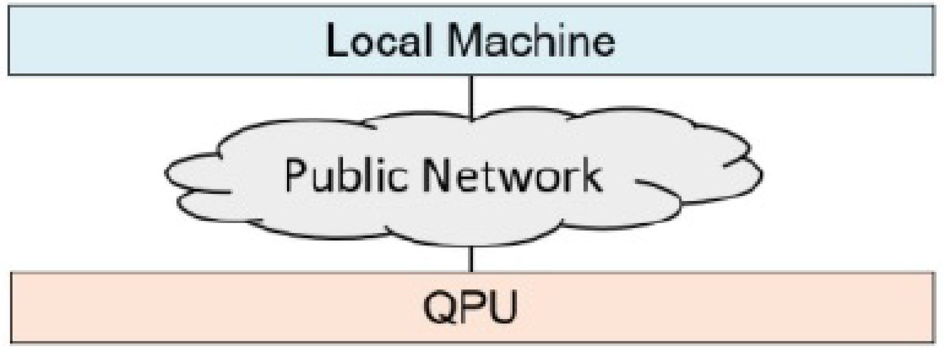
- **Platform for quantum computing research**
  - Accelerate Quantum Circuit Simulators on GPUs
  - Simulate ideal or noisy qubits
  - Enable algorithms research with scale and performance not possible on quantum hardware, or on simulators today
- **cuQuantum General Access available now**
  - Integrated into leading quantum computing frameworks Cirq, Qiskit, and Pennylane
  - C and Python APIs
  - Available today at developer.nvidia.com/cuquantum



Motivation behind CUDA Quantum

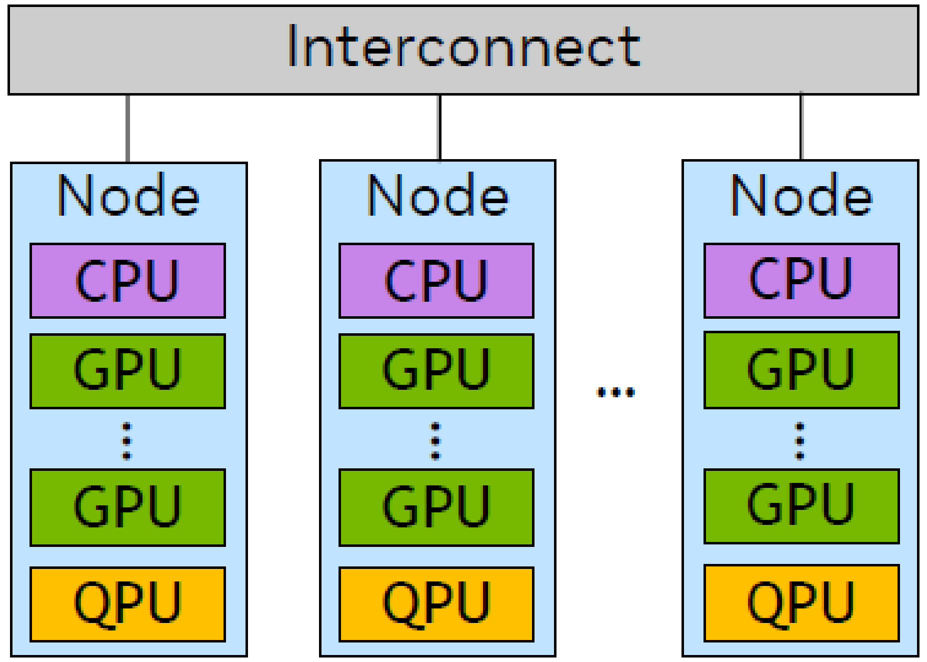
Integrate quantum computers seamlessly with the modern scientific computing ecosystem

- HPC centers and many other groups worldwide are focused on the integration of quantum computers with classical computers/supercomputers
- We expect quantum computers will accelerate some of today's most important computational problems and HPC workloads
  - Quantum chemistry, Materials simulation, AI
- We also expect CPUs and GPUs to be able to enhance the performance of QPUs
  - Classical preprocessing (circuit optimization) and postprocessing (error correction)
  - Optimal control and QPU calibration
- Want to enable researchers to seamlessly integrate CPUs, GPUs, and QPUs
  - Develop new hybrid applications and accelerate existing ones
  - Leverage classical GPU computing for control, calibration, error mitigation, and error correction



Quantum Programming Today

Great for early experimentation.



Where we need to get...

Quantum Programming with NVIDIA

Hybrid quantum-classical applications at scale.

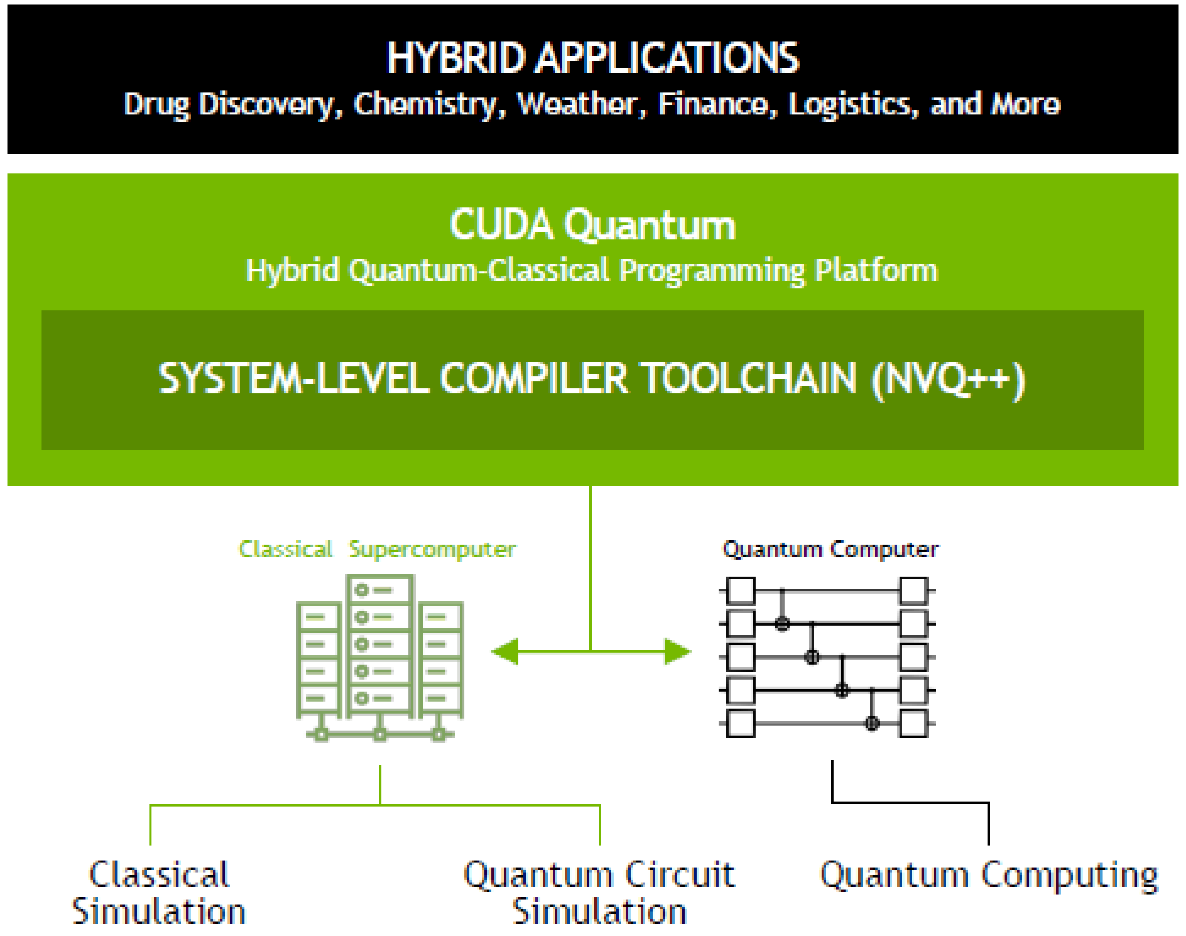
Figure adapted from:
Quantum Computers for High-Performance Computing. Humble, McCaskey, Lyakh, Gowrishankar, Frisch, Monz. IEEE Micro Sept 2021. 10.1109/MM.2021.3099140

Introducing CUDA Quantum

Platform for unified quantum-classical accelerated computing

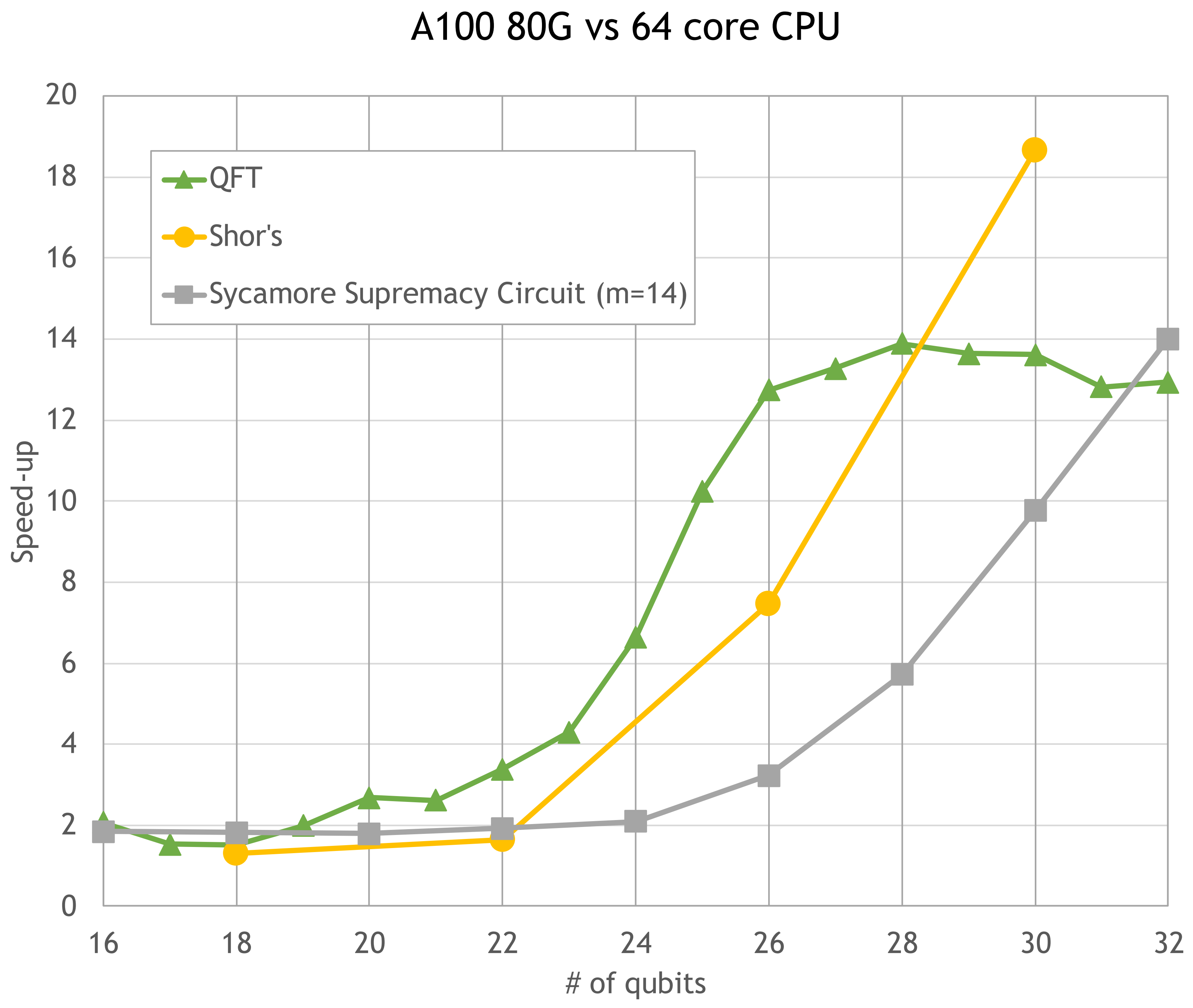
- Programming model extending C++ and Python with quantum kernels
- Open programming model, open-source compiler
  - https://github.com/NVIDIA/cuda-quantum
- QPU Agnostic – Partnering broadly including superconducting, trapped ion, neutral atom, photonic, and NV center QPUs
- Interoperable with the modern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

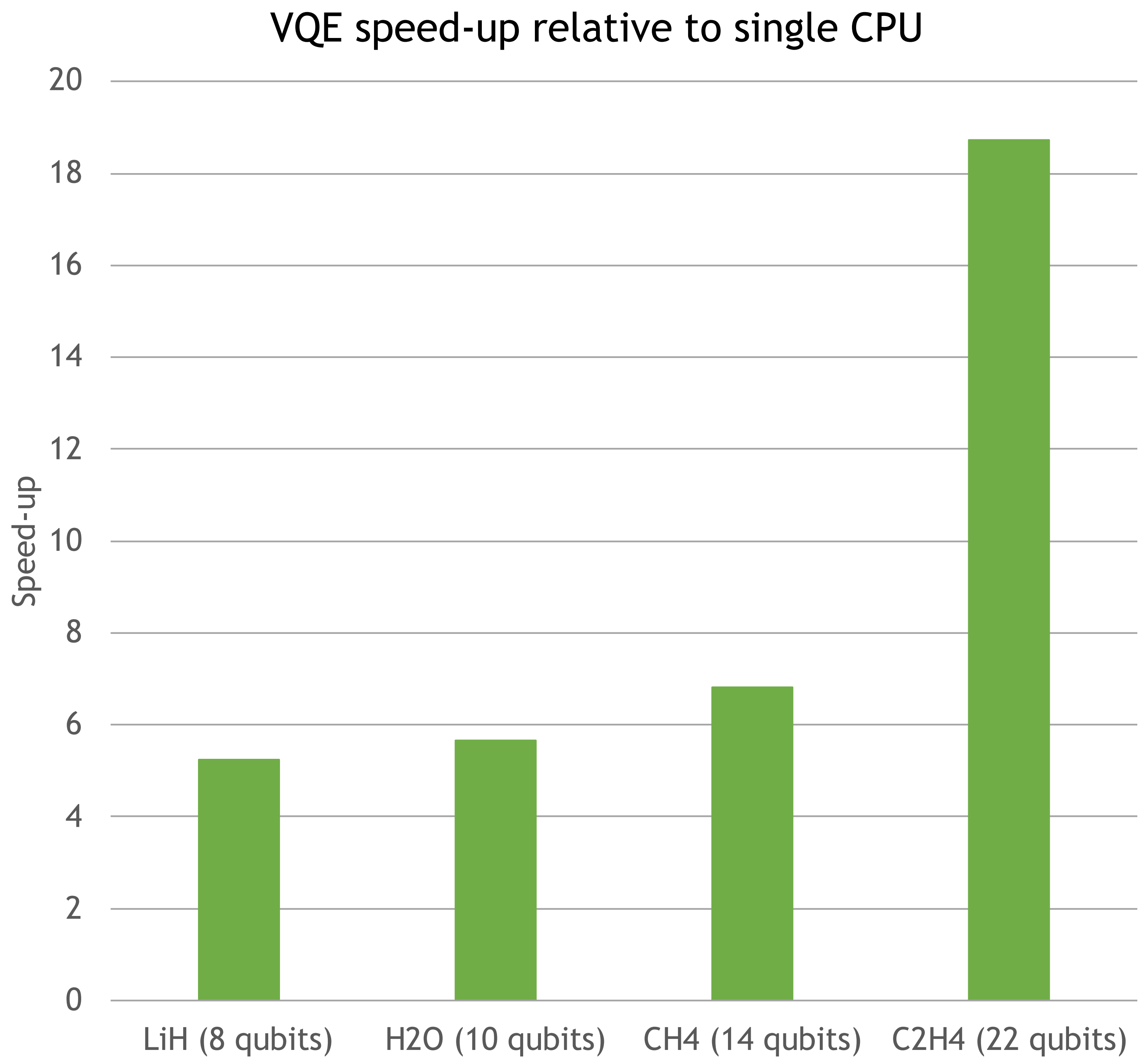
```
auto ansatz = [](std::vector<double> thetas) __qpu__ {
  cudaq::qreg<3> q;
  x(q[0]);
  ry(thetas[0], q[1]);
  ry(thetas[1], q[2]);
  x<cudaq::ctrl>(q[2], q[0]);
  x<cudaq::ctrl>(q[0], q[1]);
  ry(-thetas[0], q[1]);
  x<cudaq::ctrl>(q[0], q[1]);
  x<cudaq::ctrl>(q[1], q[0]);
};

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



cuStateVec - SINGLE GPU PERFORMANCE

PRELIMINARY PERFORMANCE OF cirq/qsim + cuStateVec ON THE A100





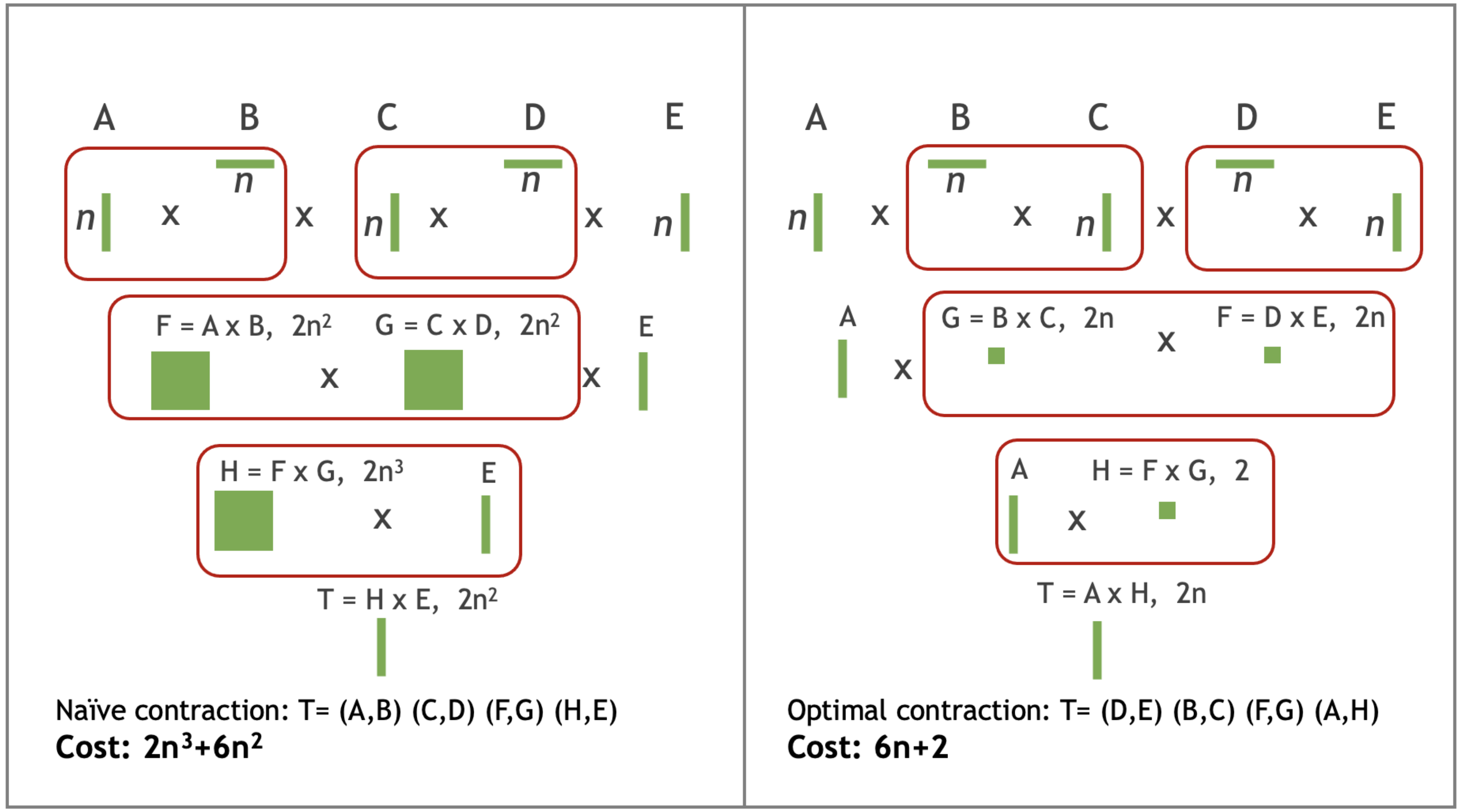
Benchmarks run using cirq/qsim with modifications to integrate cuStateVec
CPUs used were AMD EPYC 7742 with 64 cores
QFT circuit with 32 qubits and depth 63
Shor's circuit with 30 qubit and depth 15560 (integer factorized: 65)
Sycamore supremacy circuit m=14 with 7480 gates

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

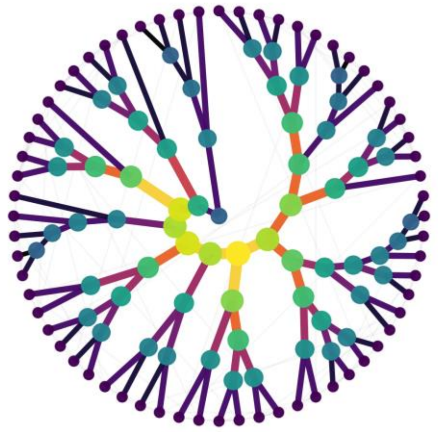
cuTensorNet

A LIBRARY TO ACCELERATE TENSOR NETWORK BASED QUANTUM CIRCUIT SIMULATION

- For many practical quantum circuits, tensor networks enable scaling of simulation to 100s or 1000s of qubits
- cuTensorNet provides APIs to:
  1. Convert a circuit written in Cirq or Qiskit to a tensor network
  2. Calculate an optimal path for the contraction
     - 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
  3. Calculate an execution plan and execute the TN contraction
     - Leverages cuTENSOR heuristics



https://developer.nvidia.com/blog/scaling-quantum-circuit-simulation-with-cutensornet/

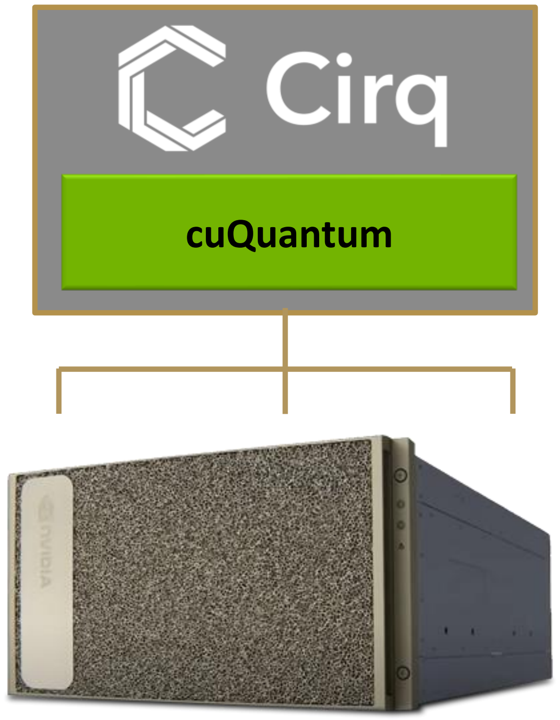


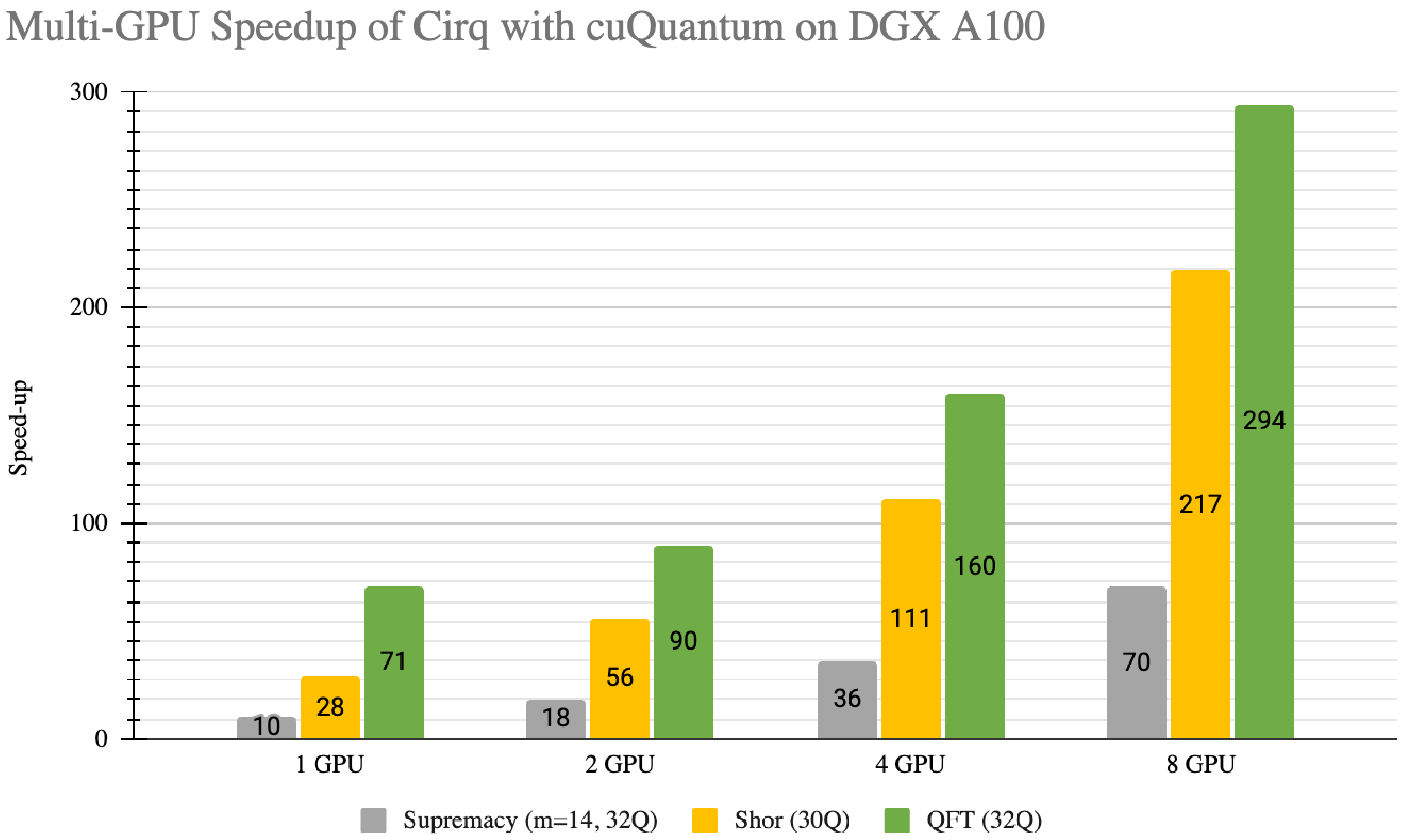
Tensor Network image from Quimb: https://quimb.readthedocs.io/en/latest/index.html

DGX cuQuantum Appliance

MULTI-GPU CONTAINER WITH cirq/qsim + cuQuantum

- Full Quantum Simulation Stack with a cirq/qsim frontend
- World class performance on key quantum algorithms
- Available Now on NGC





JOIN NVIDIA DEVELOPER PROGRAM TODAY

Supporting the Community That's Changing the World

LEARN MORE ABOUT DEVELOPER PROGRAM AND JOIN NOW.
https://developer.nvidia.com/developer-program

GET NVIDIA OPTIMIZED CONTAINERS, MODELS AND MORE.
https://developer.nvidia.com/downloads

EXPLORE NVIDIA INCEPTION AND APPLY TODAY.
https://www.nvidia.com/en-us/startups


NVIDIA