ABCI グランドチャレンジ 成果報告会

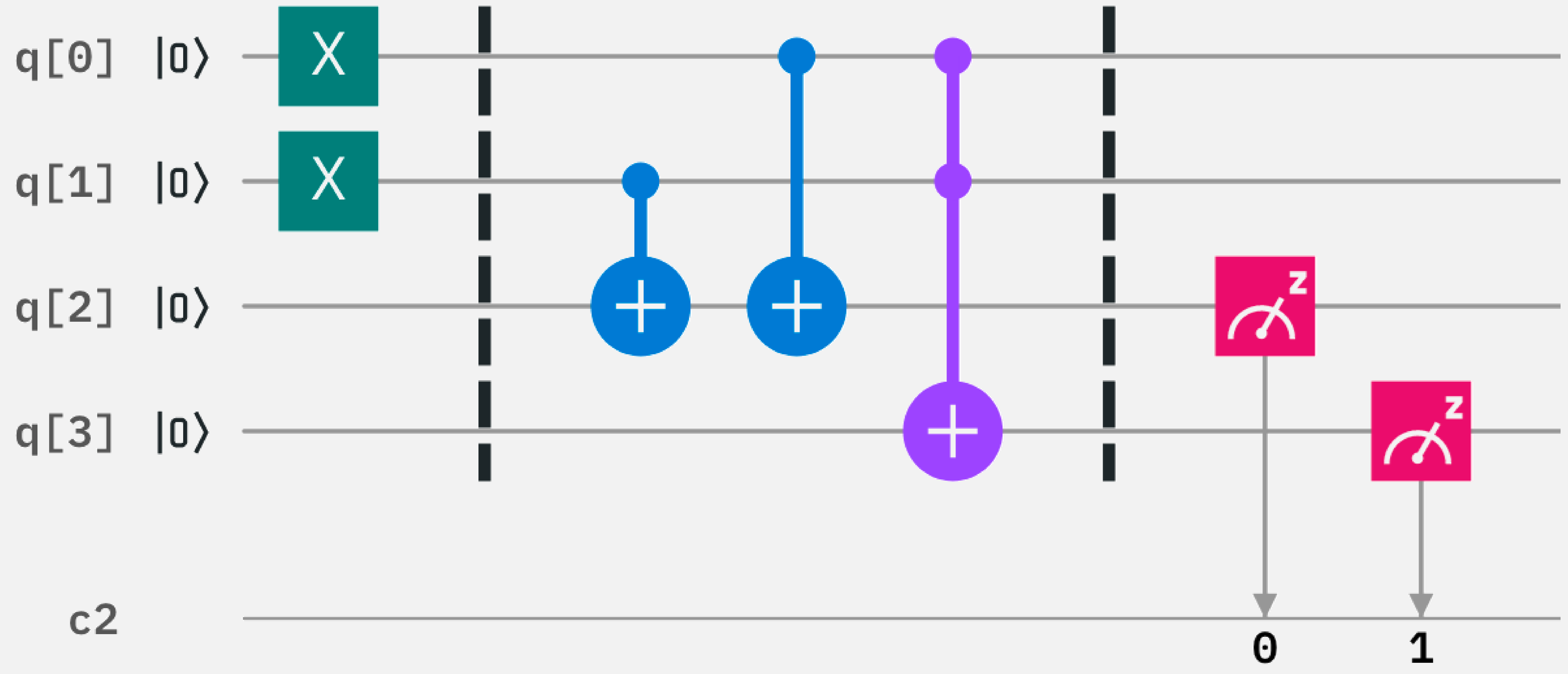
森野 慎也, Principal Software Engineer, Quantum Computing, CUDA Math Libraries Team, NVIDIA, 1/20/2023

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?

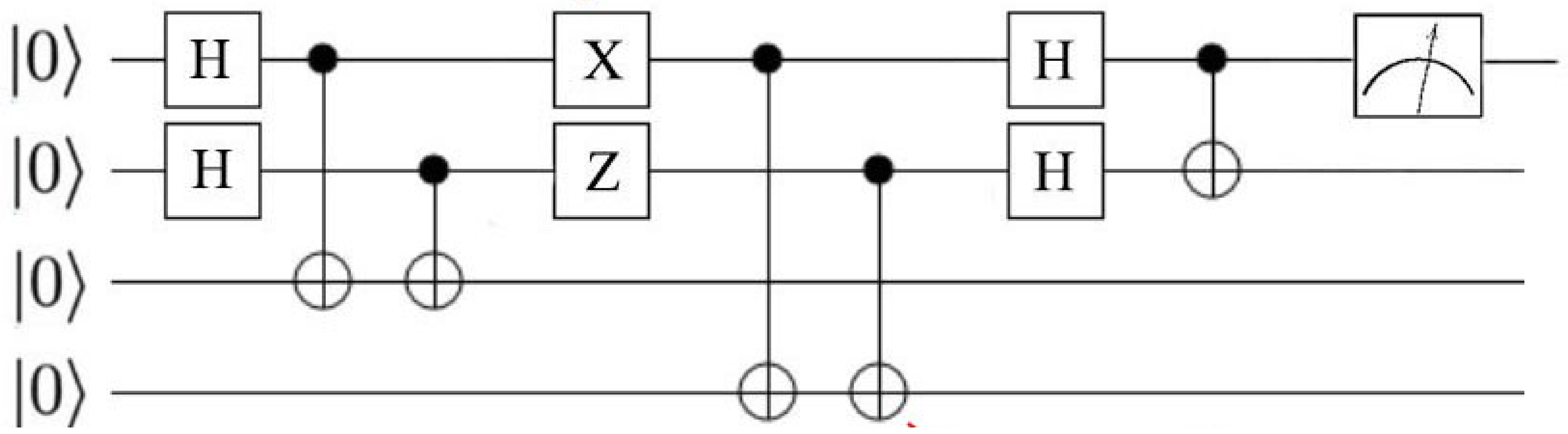
HYBRID CLASSICAL/QUANTUM APPLICATIONS

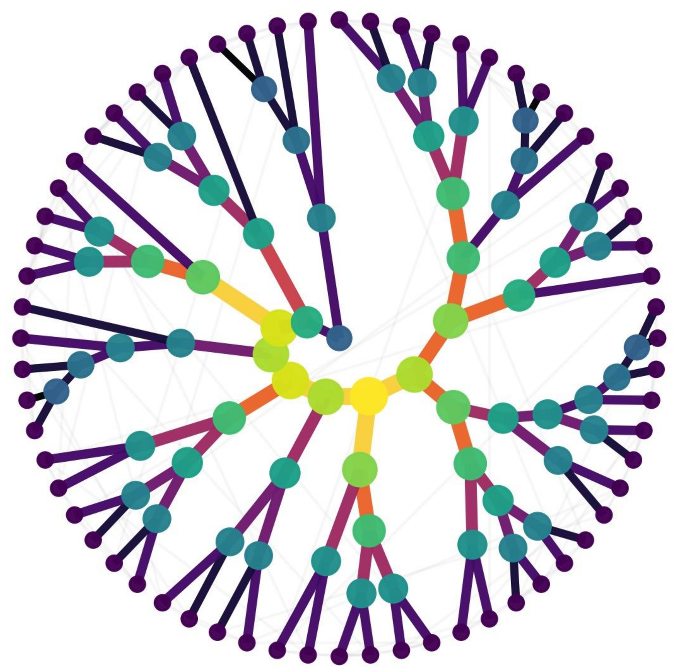
Impactful QC applications (e.g., simulating quantum materials and systems) will require classical supercomputers with quantum co-processors

+

- 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

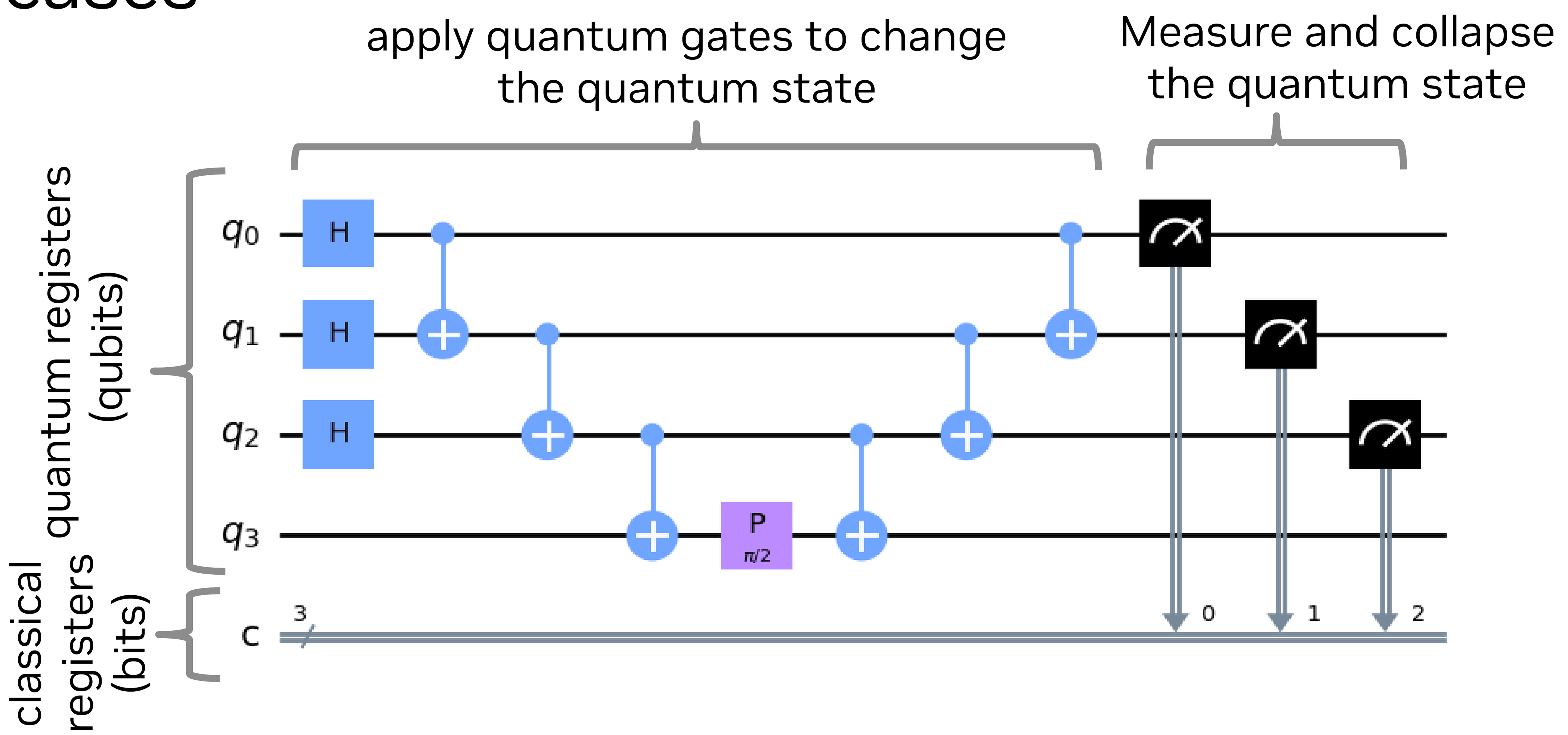
GPUs are a great fit for either approach

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

State vector simulation with NVIDIA cuQuantum

libcustatevec

- Library of a set of C-APIs specifically designed for state vector simulators to cover common use cases

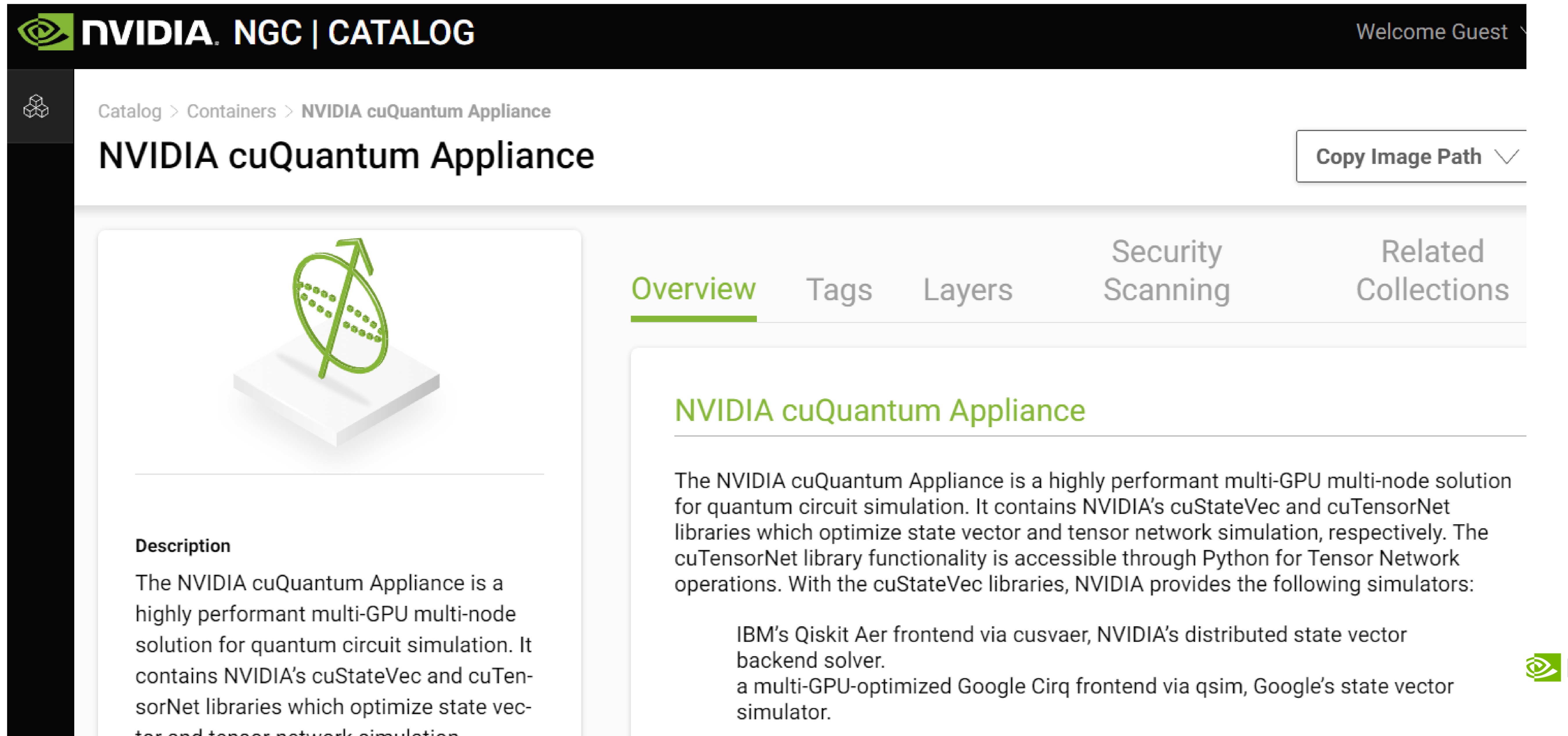



```
custatevecStatus_t
custatevecApplyMatrix(custatevecHandle_t         handle,
                      void*                      sv,
                      cudaDataType_t             svDataType,
                      const uint32_t             nIndexBits,
                      const void*                matrix,
                      cudaDataType_t             matrixDataType,
                      custatevecMatrixLayout_t   layout,
                      const int32_t              adjoint,
                      const int32_t*             targets,
                      const uint32_t             nTargets,
                      const int32_t*             controls,
                      const uint32_t             nControls,
                      custatevecComputeType_t    computeType,
                      void*                      extraWorkspace,
                      size_t                     extraWorkspaceSizeInBytes);
```

NVIDIA cuQuantum Appliance

- Multi-GPU multi-node solution for quantum circuit simulation. Available on NGC as a docker container
- cuStateVec
  - Integrated to Cirq/qsim and Qiskit/Qiskit-Aer
- cuTensorNet
  - Python for Tensor Network operations

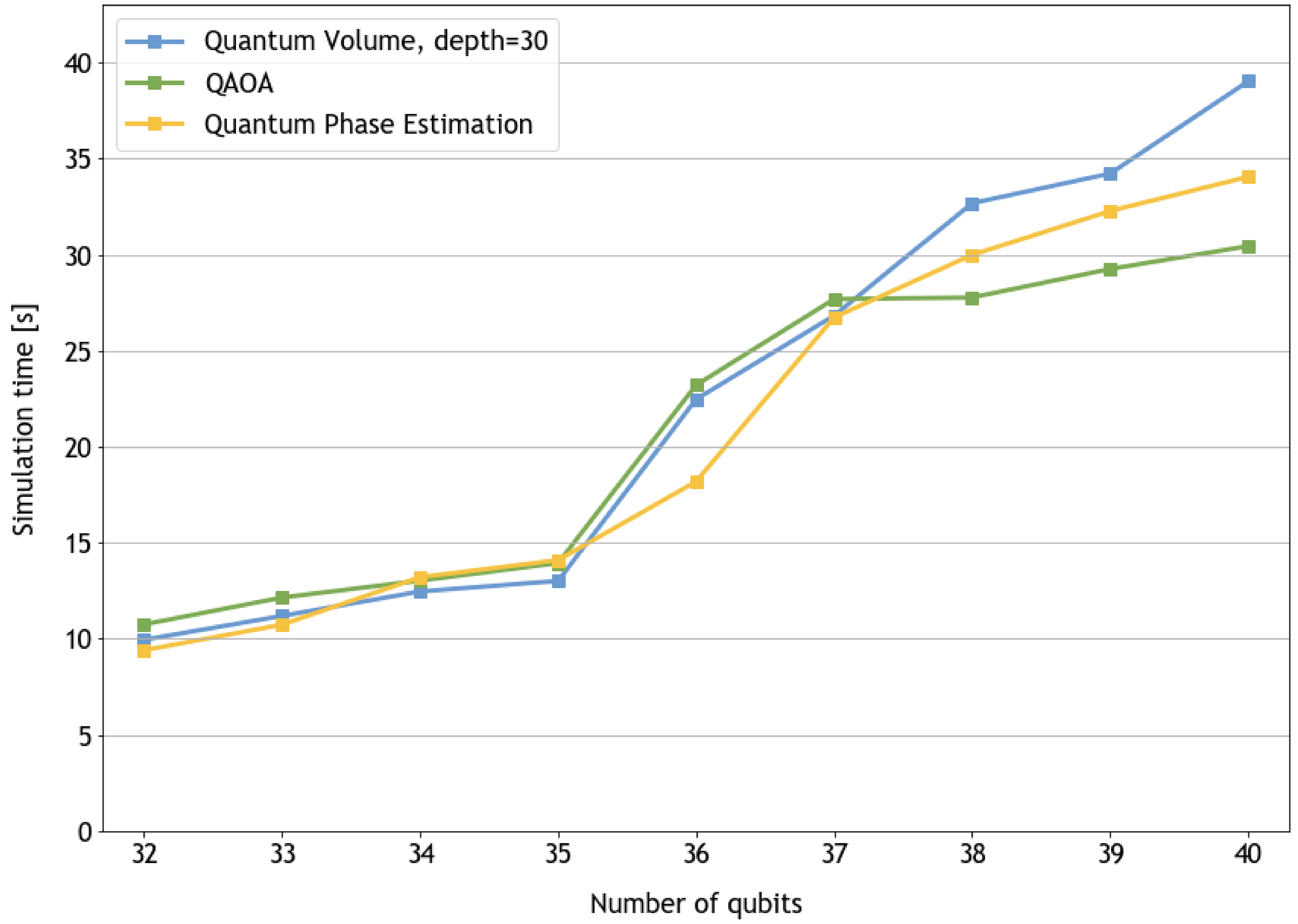
[Link to the NGC Page]



NVIDIA cuQuantum Appliance 22.11

Qiskit/Qiskit-Aer

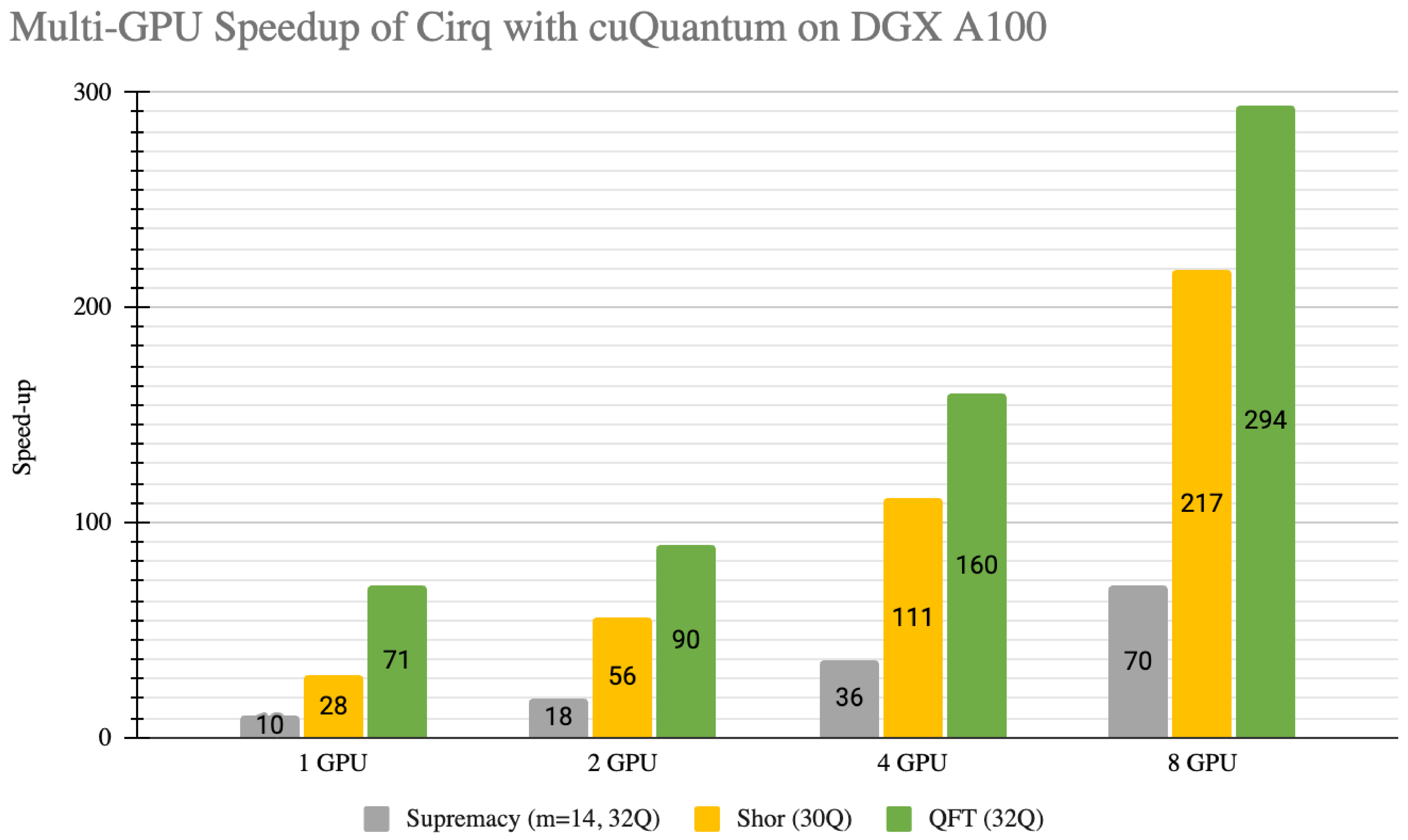
- Full Quantum Simulation Stack with a Qiskit/qsim frontend
- New package, cusvaer, extends Qiskit-Aer capability to run “**multi-node**” state vector simulation
- Prototype was executed during “**ABCI grand challenge**”



[NVIDIA technical blog:
Achieving Supercomputing-Scale Quantum Circuit Simulation with the NVIDIA cuQuantum Appliance]

Cirq/qsim

- Full Quantum Simulation Stack with a cirq/qsim frontend
- qsim is extended to run “**multi-GPU**” state vector simulation



State vector simulation

- Basic simulation method
  - Simulators are provided in quantum computing frameworks
- Exact simulation

State vector

- Represents quantum state as a vector of complex numbers. Its length is 2^{n_qubits},
- Requires a huge amount of memory
  - Assuming complex 128

    20 qubit, 16 MiB

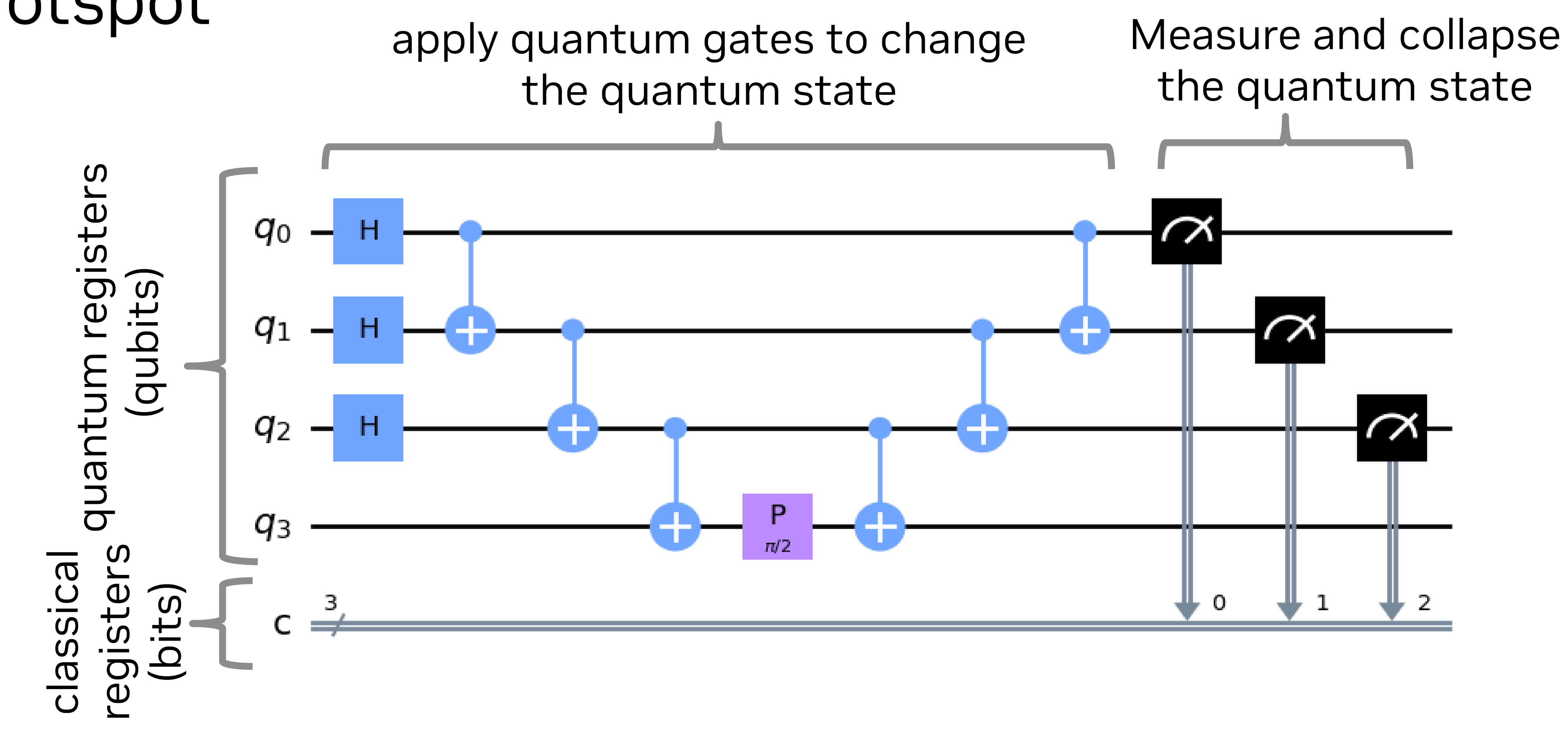
    30 qubit, 16 GiB

    40 qubit, 16 TiB

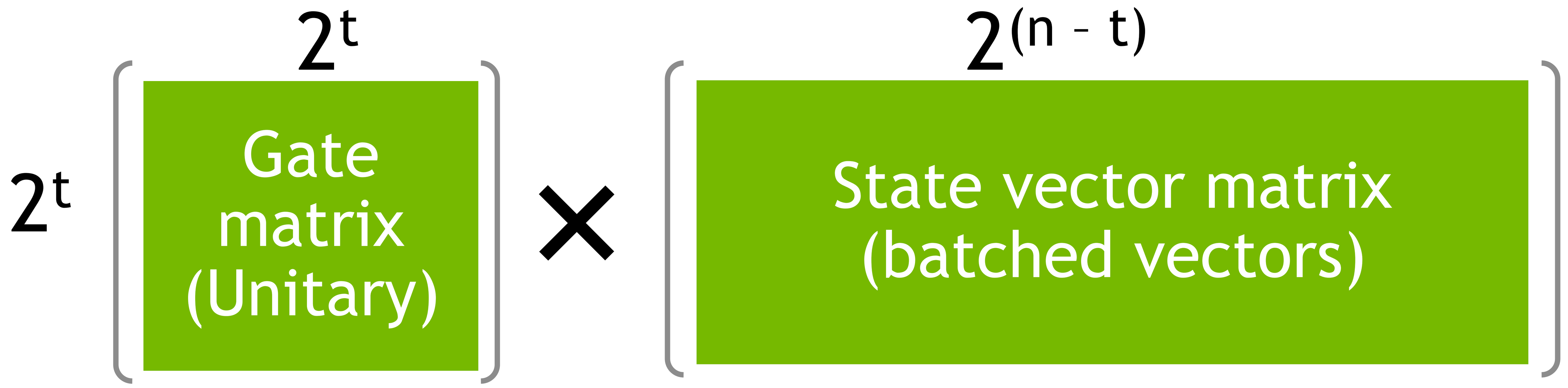
    50 qubit, 16 PiB
- Multiple GPUs/nodes increase the number of qubits

Gate application

- Hotspot



- In-place matrix-matrix multiplication



t: Number of target qubits
n: Number of qubits

Why GPU ?

High single device performance

- Fast gate application for small gate matrices
    - High memory bandwidth
- Fast gate application for large gate matrices
    - High floating-point performance

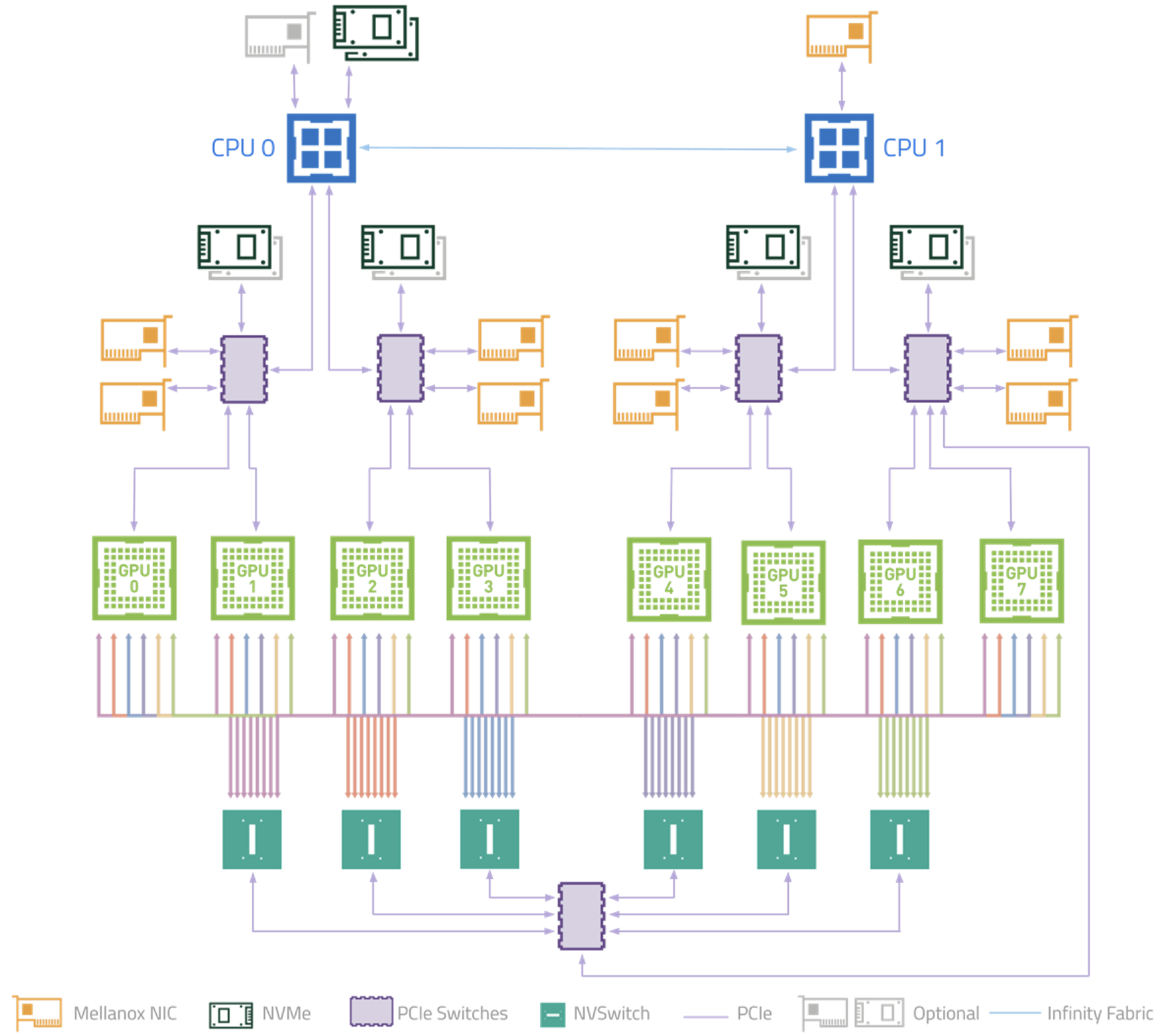
NVIDIA A100 Specification

	NVIDIA A100 (80G)	NVIDIA A100 (40G)
Memory Bandwidth	1.6 TB/sec	2.0 TB/sec
FP32 Peak Performance	19.5 TFLOPS	
FP64 Peak Performance	19.5 TFLOPS	

Fast GPU-interconnect (DGX/HGX A100)

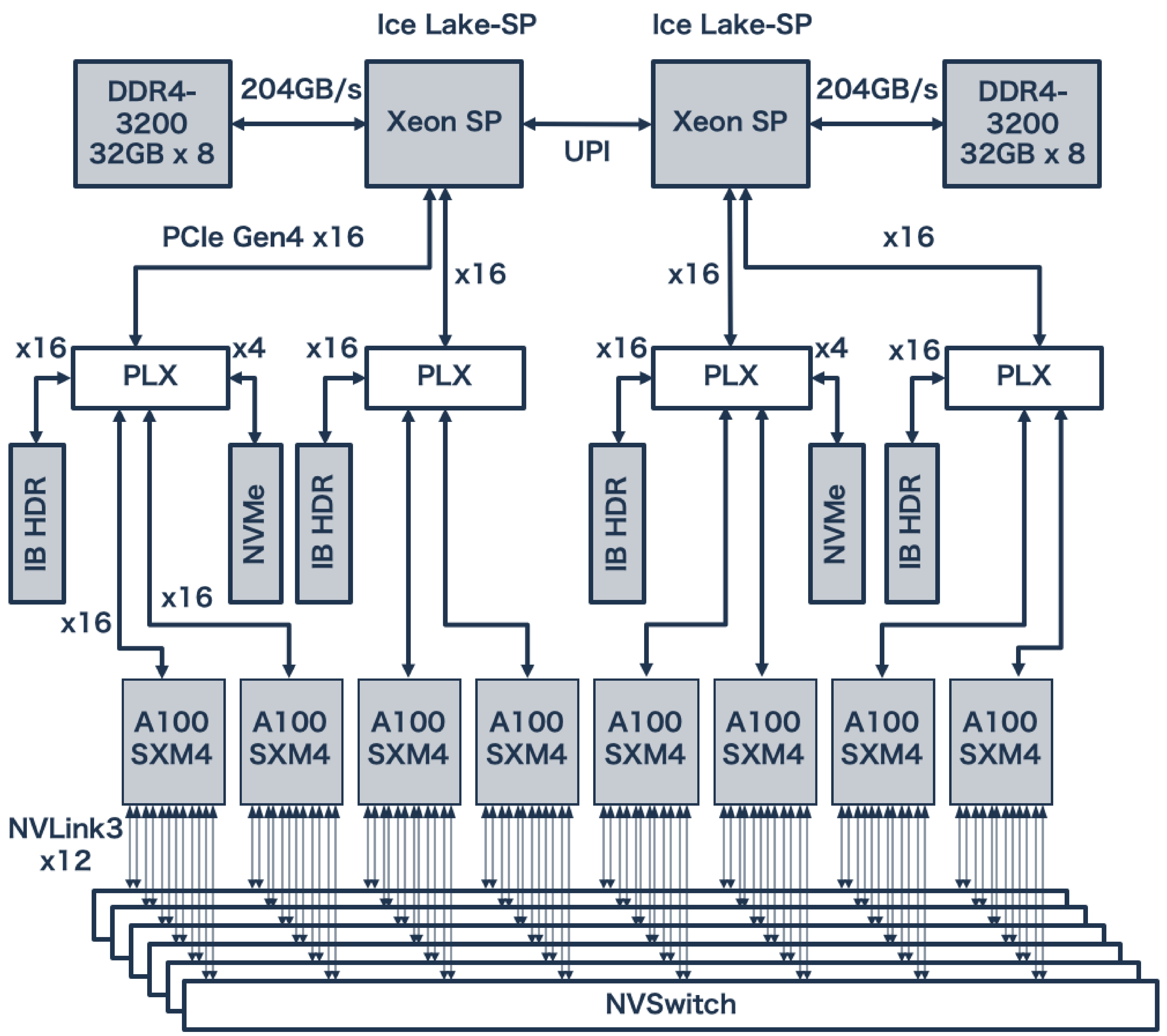
- Fast data exchange between distributed state vectors in multiple GPUs
- NVLink bandwidth: 600 GB/s (bidirectional)
- NVSwitch bisection bandwidth : 2.4 TB

NVIDIA DGX A100 System



Distribute simulation to multiple nodes

ABCI Compute Nodes (A)



120 Compute Nodes (A) that form in total **960 NVIDIA A100 GPU** accelerators.

FUJITSU PRIMERGY GX2570 M6（1 server in 4U）	
CPU	Intel Xeon Platinum 8360Y Processor (54 MB Cache, 2.4 GHz, 36 Cores, 72 Threads) ×2
GPU	NVIDIA A100 for NVLink 40GiB HBM2 ×8
Memory	512GiB DDR4 3200MHz RDIMM
Local Storage	2.0TB NVMe SSD (Intel SSD DC P4510 u.2) ×2
Interconnect	**InfiniBand HDR (200Gbps) ×4**

High-Speed Interconnect

•Compute Nodes (A) can communicate with each other in **full-bisection bandwidth**.

Ref: https://abci.ai/en/about_abci/computing_resource.html

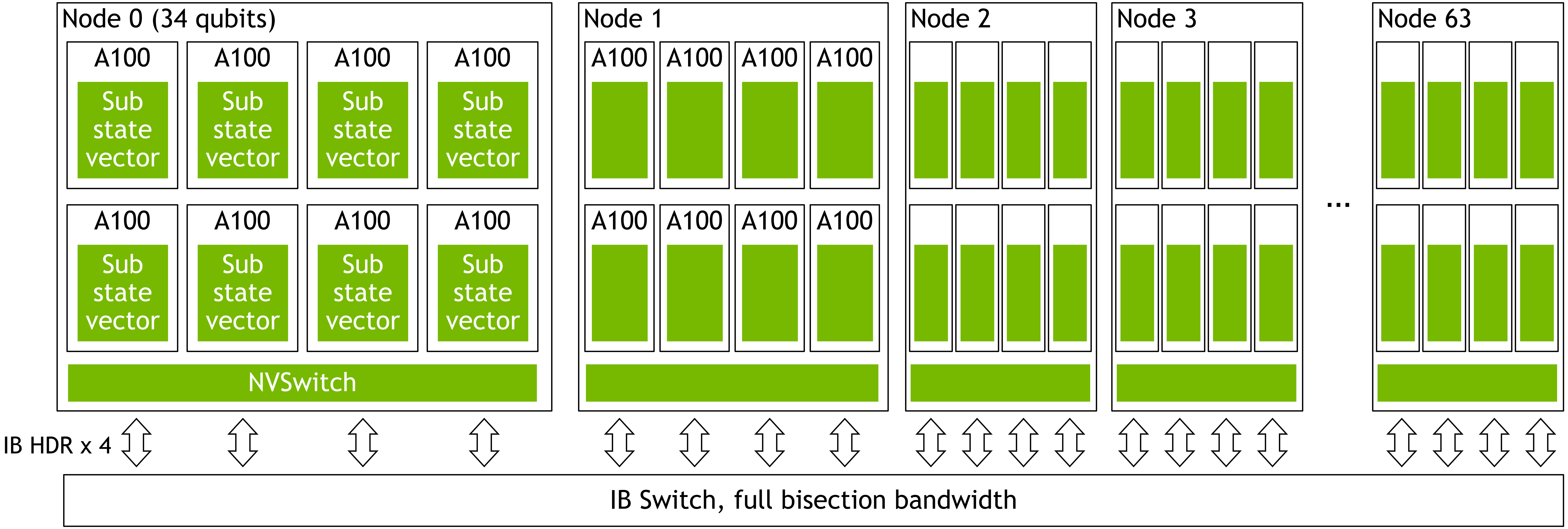
40 qubit state vector distributed to 64 nodes

Single GPU

- 31 qubits (c128)
- 32 GiB = 16 bytes x 2^{31}

- Equally slice the state vector
  - Allocate one slice on one GPU
- +1 qubits by doubling # GPUs

- Up-to 64 nodes
  - 31 qubits, 32 GiB in device
  - 3 qubits, 8 GPUs / node
  - 6 qubits, 64 nodes



Gate application for distributed state vector

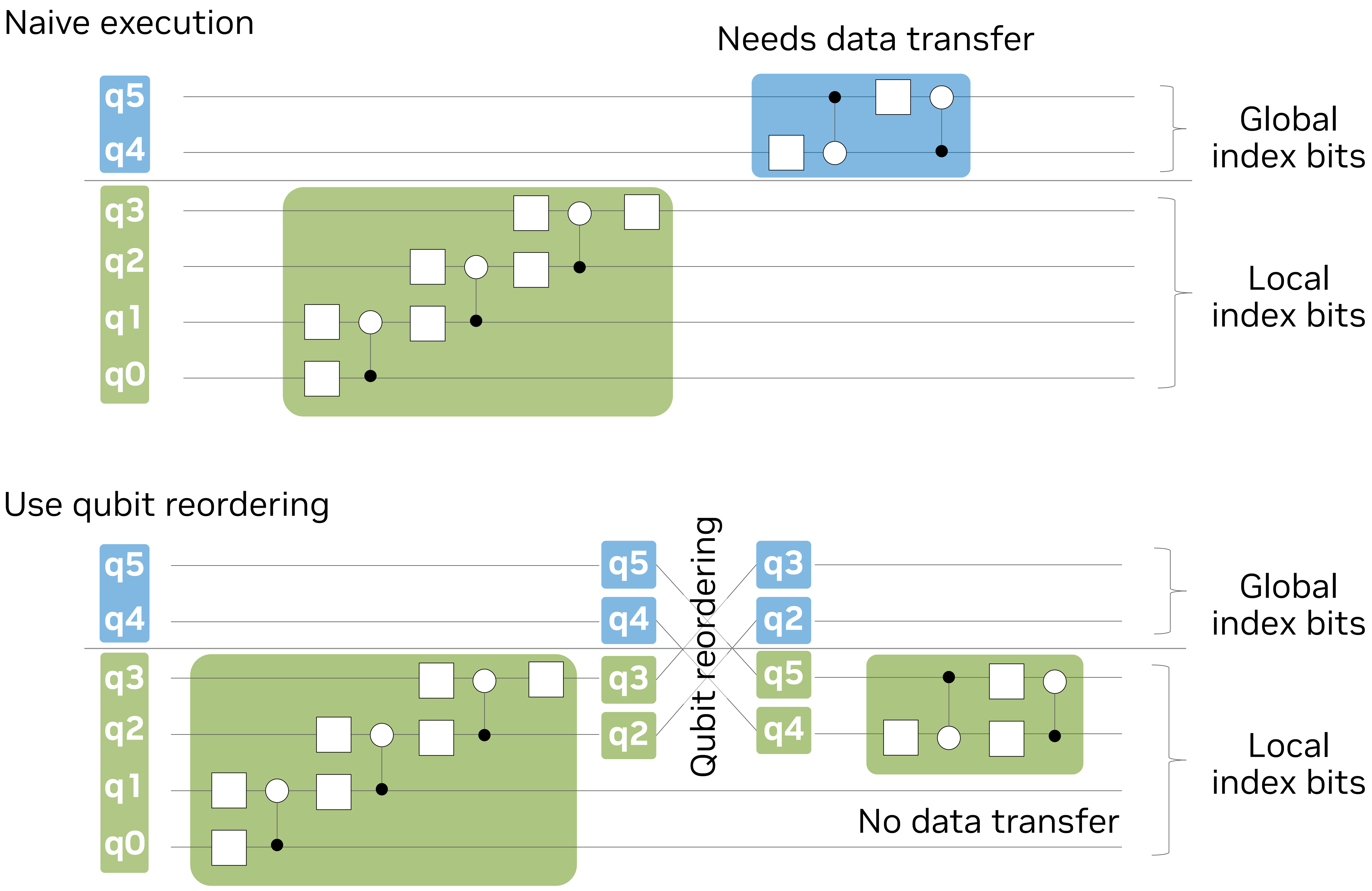
Example of single qubit gate application

Qubits map to the index bits of state vector

Gate is applied in each GPU

Gate is applied on two GPUs
Access to two GPUs
Data transfer happens

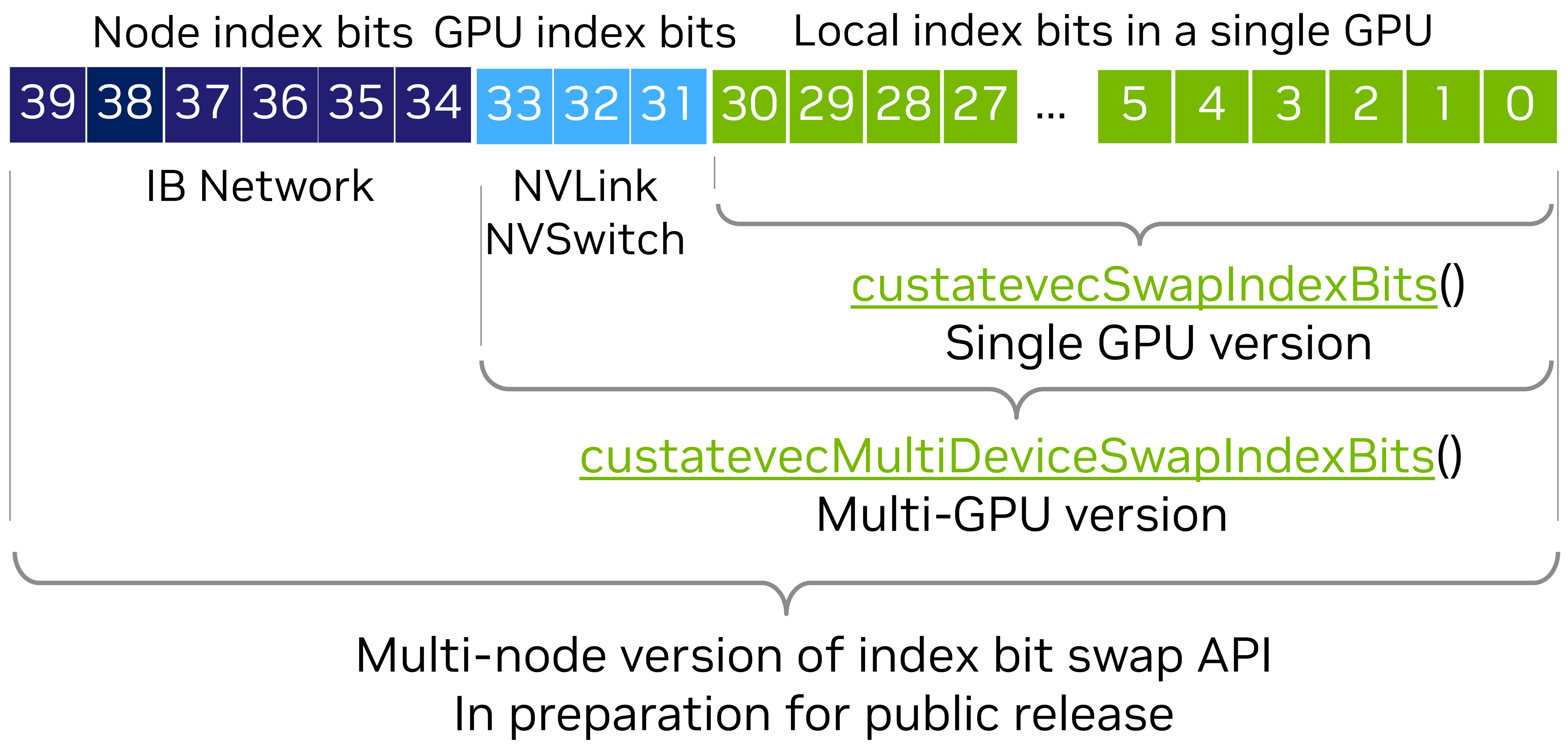
Qubit Reordering



- Swap and update mapping
  - Qubits
  - State vector index bits
- Data transfer for gates acting on global index bits
- Swapping qubit position
  - moves gates to act on local index bits

Index bit swap API

"SwapIndexBits" API in libcustatevec

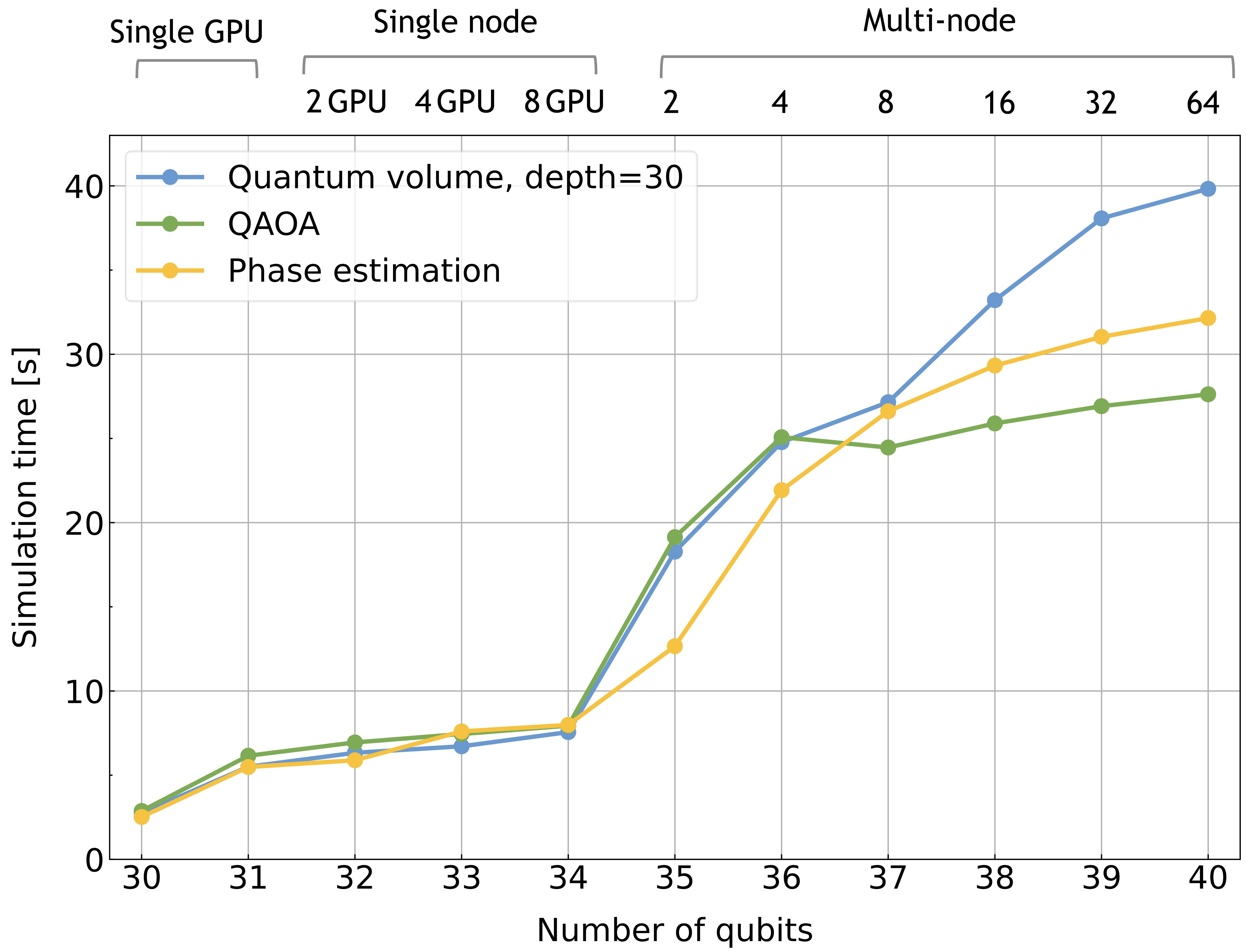


Qubit reordering:

- Proactively utilize faster NVLink/NVSwitch
  - 300 GB/[sec•GPU] (unidirectional)
- Reduce usage of IB network
  - 12.5 GB/[sec•GPU] (unidirectional)

Simulation performance

Weak scaling, c128



Circuits

- Quantum volume, depth=30
- QAOA
- Quantum phase estimation

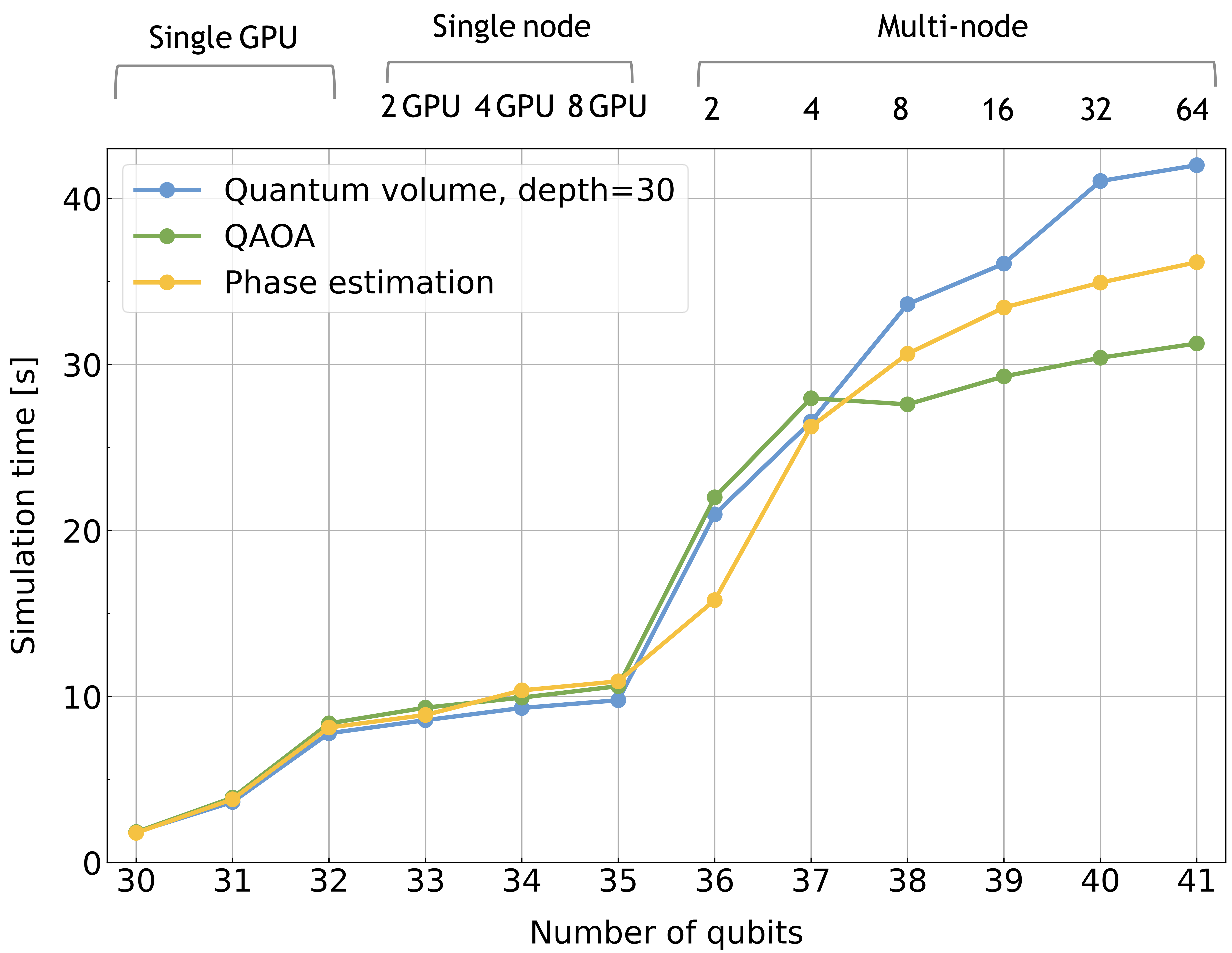
Simulator options

- Complex 128
- Gate fusion: 5 qubits

Ref: Quantum volume, QAOA, QPE

Simulation performance

Weak scaling, c64



Simulator options

- Complex 64
- Gate fusion: 4 qubits

Reached 41 qubits

- sizeof(Complex64) = 8
- sizeof(Complex128) = 16

Very similar performance to c128 results

Summary

Multi-node state vector simulation

- Executed on ABCI Computing nodes(A)
- 40 qubit (c128) and 41 qubit (c128) simulations with 64 nodes, 512 GPUs
- Best-in-class performance

Circuits

- Quantum volume, QAOA, Quantum phase estimation

Accuracy (not discussed in the presentation)

- Quantum phase estimation gives results with almost the same accuracy for c128 and c64 simulations

Public release in cuQuantum Appliance 22.11

- Available to all public users

Links

NVIDIA Technical Blog

- Achieving Supercomputing-Scale Quantum Circuit Simulation with the NVIDIA cuQuantum Appliance | NVIDIA Technical Blog
- Best-in-Class Quantum Circuit Simulation at Scale with NVIDIA cuQuantum Appliance | NVIDIA Technical Blog

cuQuantum

- cuQuantum SDK | NVIDIA
- Documentation: cuQuantum SDK
- Public github repository: github.com, benchmark

NGC

- NVIDIA cuQuantum Appliance | NVIDIA NGC

AIST

- 産総研ABCIを活用し、世界最速の量子回路シミュレーションに成功 - 成果公開 | デジタルアーキテクチャ研究センター | 産総研 (aist.go.jp)

NVIDIA