Chemically Accurate End-to-End Classical Simulation of Qubitization-Based Quantum Phase Estimation on 1,024 GPUs

Yusuke Teranishi
QunaSys Inc.
Tokyo, Japan
teranishi@qunasys.com

Hayato Yunoki
Graduate School of Information Science and Technology,
The University of Tokyo
Tokyo, Japan

Daniel Chung
QunaSys Inc.
Tokyo, Japan

Nils Wittemeier
QunaSys Inc.
Tokyo, Japan

Toru Shibamiya
QunaSys Inc.
Tokyo, Japan

***Abstract*—Future fault-tolerant quantum computers (FTQC) are expected to compute molecular energies precisely, for example to design catalysts, with an algorithm called quantum phase estimation (QPE). Resource estimates predict how many qubits and gates such a calculation needs, but only running the circuit shows whether it returns an accurate energy. Because such hardware does not exist yet, we simulated the circuits on a GPU supercomputer. We shrank a chemical-accuracy circuit for the hydrogen chain H_4 from 99 qubits and 1.78 billion gates to 28 qubits and 0.32 billion gates, and simulated it on ABCI-Q's 1,024 NVIDIA H100 GPUs in 1.5 hours. The most frequent of its one million sampled energies lies 0.53 milli-Hartree (mHa) from the exact value, within chemical accuracy.**

***Index Terms*—quantum phase estimation, qubitization, distributed state-vector simulation, GPU supercomputing**

I. Introduction

Fault-tolerant quantum computers (FTQC), which correct their own errors, are expected to compute molecular energies precisely, helping to design catalysts and materials. The main algorithm for this task is quantum phase estimation (QPE), and its qubitization-based variant underlies leading resource estimates for industrial molecules.

Such estimates tell us how many qubits and gates a calculation needs, but not whether a circuit built with them returns the right energy; only running the whole pipeline shows this. No FTQC exists yet, so we simulate the circuits exactly on classical hardware.

We combine established circuit-reduction and communication techniques into one end-to-end workflow and make three contributions. First, qubitization QPE ran for the hydrogen chains H_2–H_{10} (up to 42 qubits) on 1,024 NVIDIA H100 GPUs. Second, we counted the qubits and gates after each reduction stage. Third, the H_2 and H_4 runs reached chemical accuracy.

II. Why This Needs a Supercomputer

A classical simulator stores one complex number for every possible state of the qubits: 2^n numbers, or 2^{n+4} bytes, for n qubits. Every extra qubit therefore doubles the memory. A single 80 GiB GPU holds up to 32 qubits, 42 qubits need 64 TiB, and 50 qubits would need 16 PiB.

Qubitization circuits are large because they add helper (ancilla) qubits to encode the molecule's energy operator, the Hamiltonian, into a walk operator: a unitary whose phase QPE reads out as the energy. At chemical accuracy (an error of 1.6 mHa, the usual tolerance in chemistry), a straightforward circuit needs 69 qubits for H_2, and 99 qubits with 1.78 billion gates for H_4. Both exceed 50 qubits, so their state would need over 16 PiB; the circuits must first be made smaller.

III. Making the Circuits Simulatable

Our workflow builds the Hamiltonian of a hydrogen chain (STO-3G minimal basis), turns it into a QPE circuit, compiles the circuit, and simulates it on up to 1,024 GPUs. The sampled energies are compared with full configuration interaction (FCI), the exact energy of the same molecular model.

Starting from the plain linear combination of unitaries (LCU) encoding, four established techniques shrink the circuit in stages (Fig. 1). The simulation gate set writes the same circuit in gates the simulator applies directly; this is exact and cuts gates, but these counts are not FTQC costs. Double Factorization (DF) compresses the Hamiltonian; its truncation is an approximation kept within the accuracy budget. Möttönen PREPARE, the circuit loading the DF coefficients, does so exactly with rotations, removing most helper qubits (87 to 29). Sin-state initialization prepares the QPE register in a sine shape, which sharpens the energy peak so that one fewer QPE qubit meets the same target, halving the gates; it changes the output distribution rather than rewriting the circuit exactly. Together, the circuit shrinks from 99 to 28 qubits and from 1.78 to 0.32 billion gates. Stages (1) to (3) were only counted,

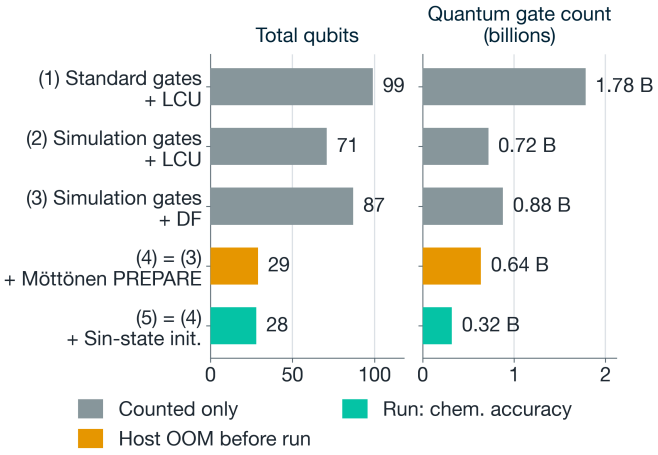


Fig. 1. Qubits and gates of the H_4 chemical-accuracy circuit after each reduction stage, counted in that stage's gate set.

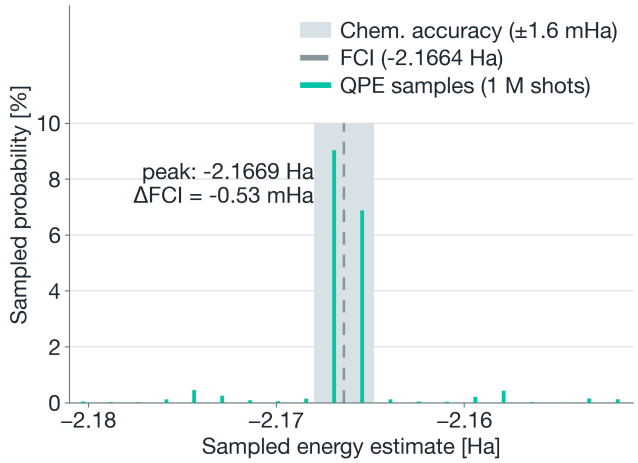


Fig. 2. H_4 QPE energy distribution (1 M shots), peaking within 0.53 mHa of FCI.

stage (4) exceeded host memory before running, and stage (5) was run.

IV. Platform

We use QURI SDK Enterprise, whose multi-GPU backend builds on NVIDIA cuQuantum (cuStateVec), on AIST's ABCI-Q (256 nodes with four 80 GiB H100 GPUs each). The state is split across GPUs, so some gates require data exchange. Lazy qubit reordering and communication hierarchization reduce this traffic, preferring fast links within a node. Error analysis on CPUs selected the circuit parameters, so GPU time went only to runs expected to succeed.

V. Results and Discussion

Qubitization QPE ran for H_2 through H_{10}, up to 42 qubits. H_2 and H_4 reached chemical accuracy, while H_6–H_{10} ran at lower precision to test scale. The H_4 run sampled one million energies in 1.5 hours, and the most frequent value lies 0.53 mHa from FCI (Fig. 2).

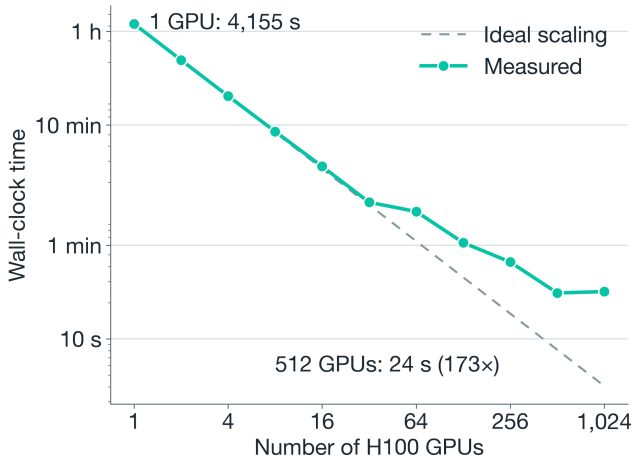


Fig. 3. Strong scaling of the largest QPE circuit that fits on a single GPU (32 qubits).

To measure parallel speedup, we ran one 32-qubit H_2 circuit on the same backend with 1 to 1,024 GPUs. The runtime fell from 4,155 seconds on one GPU to 24 seconds on 512 GPUs, a 173-fold speedup (Fig. 3). This measures scaling, not the gain from individual optimizations. Beyond 512 GPUs, each GPU holds fewer qubits than the walk operator acts on, so reordering and communication rise sharply; larger circuits should keep scaling further.

Two resources limit what we can run. GPU memory limits the qubit count, while host memory during compilation limits the gate count: beyond about 0.6 billion gates (H_4 stage (4), H_6 and larger) it runs out.

VI. Outlook

Estimates for industrial molecules, such as the nitrogen-fixation cofactor FeMoco, use the same framework at larger scale: about 2,100 logical qubits, with tensor hypercontraction (THC) instead of DF. Reaching them requires fast, memory-lean, distributed circuit compilers. Our runs check, at small scale, that estimated resources deliver the target accuracy, and the toolchain supports verifying future FTQC algorithms.

Acknowledgment

This work was supported by computational resources of the Quantum-AI Hybrid Computing Infrastructure (ABCI-Q), awarded under the "ABCI-Q Grand Challenge" Program of the National Institute of Advanced Industrial Science and Technology (AIST). The multi-node, multi-GPU state-vector backend used in this work builds on quri-parts-cuquantum, developed with AIST under the Cross-ministerial Strategic Innovation Promotion Program (SIP). Portions of the text were drafted with AI assistance (Anthropic Claude, OpenAI Codex); the authors verified all content.