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

Yusuke Teranishi[1], Hayato Yunoki[2], Daniel Chung[1], Nils Wittemeier[1], Toru Shibamiya[1]

[1]QunaSys Inc., Tokyo, Japan [2]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan Contact: teranishi@qunasys.com

Introduction

- **Quantum Phase Estimation (QPE)** on **fault-tolerant quantum computers (FTQC)** computes molecular ground-state energies with controllable precision, with direct impact on catalyst and materials design.
- Before FTQC hardware exists, we must confirm by execution that **analytically estimated quantum resources actually deliver the target accuracy**: build the circuit, run it, and check the sampled energies.
- Qubitization QPE circuits, however, carry many helper (ancilla) qubits and **billions of gates**.

Contributions

Established reduction and communication techniques, combined into one end-to-end workflow:

1. Qubitization QPE for hydrogen chains **H_2–H_{10} (up to 42 qubits) on 1,024 NVIDIA H100 GPUs**, the largest size that fits in memory on this system.
2. **Stage-by-stage** resources for H_4 at chemical accuracy (error ≤ 1.6 milli-Hartree, mHa): **99 → 28 qubits, 1.78 B → 0.32 B gates**.
3. **Executed, not only estimated**: H_2 and H_4 reach chemical accuracy against exact energies from full configuration interaction (FCI), and H_4 yields the full measurement distribution.

Background: Why This Needs a Supercomputer

- QPE reads out the energy as the phase of a **walk operator**, a unitary that **block-encodes** (embeds) the Hamiltonian (energy operator) via **PREPARE** (loads term weights) and **SELECT** (applies terms) circuits.
- Qubitization gives **controllable precision and fewer gates** than Trotterization (step-by-step time evolution), at the cost of extra ancilla qubits; it underpins FTQC resource estimates such as FeMoco [Lee 2021].
- Simulating n qubits stores 2^n double-precision complex amplitudes (2^{n+4} bytes): **every added qubit doubles memory**.

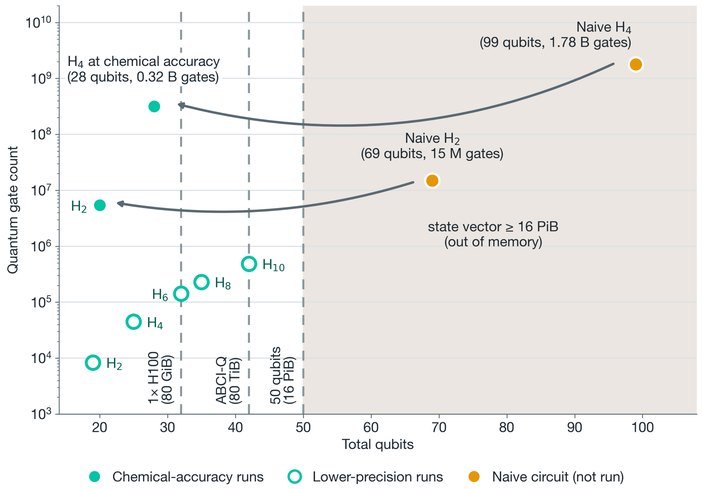


Figure 1: Where our circuits fall. Memory caps the qubit count (dashed lines); the gate count sets the runtime. Naive chemical-accuracy circuits lie beyond 50 qubits; arrows show the reduction stack (Fig. 2).

Method: Making Qubitization QPE Simulatable

End-to-end workflow

1. Molecular Hamiltonian for hydrogen chains H_2–H_{10} (STO-3G minimal basis)
2. **Double Factorization (DF) block encoding** (PREPARE / SELECT) → walk operator
3. QPE ancilla register + **sin-state initialization**
4. Compilation to the simulation gate set
5. **Distributed state-vector simulation** on up to 1,024 GPUs (NVIDIA cuStateVec backend)
6. Sampled energy distribution → ground-state energy

Reduction stack

Four established techniques, starting from the plain **linear combination of unitaries (LCU)** block encoding:

- **Simulation gate set**: the same circuit expressed in gates the simulator applies natively (multi-controlled gates, rotations). Exact; fewer gates, but not FTQC gate costs.
- **Double Factorization (DF)**: low-rank factorization of the Hamiltonian. Its truncation is an approximation, kept within the accuracy budget.
- **Möttönen PREPARE**: loads DF coefficients exactly via continuous rotations, with no data-loading ancillas.
- **Sin-state initialization**: a sine-shaped QPE register sharpens the energy peak, so one fewer QPE ancilla meets the target, halving the gates. It changes the output distribution, not an exact rewrite.

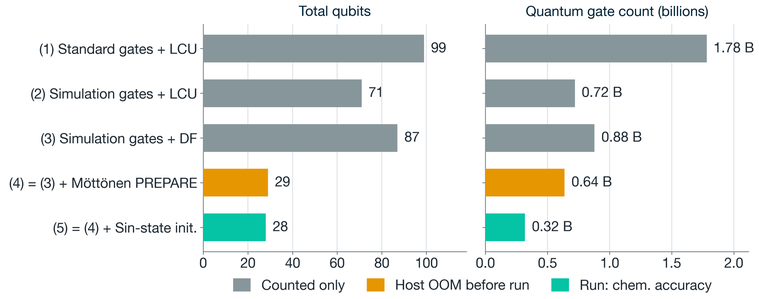


Figure 2: Resources of the H_4 chemical-accuracy circuit at each stage, counted in that stage's gate set (not hardware runtime). The stack turns a 99-qubit, 1.78 B-gate circuit into an executed 28-qubit, 0.32 B-gate one. DF alone adds qubits at this size ((2) → (3)); later stages more than compensate.

Platform & Operating at Scale

- **QURI SDK Enterprise** (QunaSys): its multi-node multi-GPU backend extends **quri-parts-cuquantum** on NVIDIA cuQuantum (cuStateVec).
- The 2^n amplitudes are split across GPUs, so some gates need inter-GPU exchange. **Lazy qubit reordering** and **communication hierarchization** cut this traffic [Teranishi 2025].
- **ABCI-Q** (AIST, Japan): **1,024 H100 GPUs** (256 nodes × 4, 80 GiB each); peak use **74 TiB device + 92 TiB host memory**.

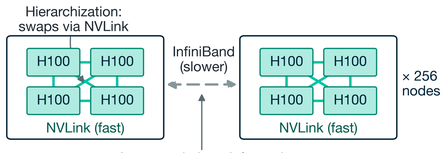


Figure 3: Two-tier network: exchanges favor fast intra-node links.

- **Targeted runs, not a grid search**: CPU-side error analysis and resource counting chose the hyperparameters (PREPARE precision, DF truncation, QPE ancillas), so GPU time went only to runs expected to succeed.

Results & Discussion

- **Scale**: qubitization QPE ran for **H_2–H_{10} (up to 42 qubits, 64 TiB)** on 1,024 H100 GPUs. **H_2 and H_4 reached chemical accuracy**; H_6–H_{10} ran at lower QPE precision to test scale.
- **Accuracy**: the H_4 run sampled **1 M shots in ≈ 1.5 h**; the distribution peak is within **0.53 mHa** of FCI, the exact energy of the same molecular model (Fig. 4).
- **Strong scaling**: one 32-qubit H_2 circuit on the same backend runs **173× faster on 512 GPUs** than on one (4,155 s → 24 s, Fig. 5). This measures parallel scaling, not the gain of individual optimizations.

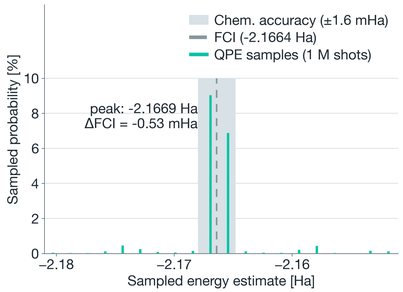


Figure 4: H_4 QPE energy samples (1 M shots) vs. FCI.

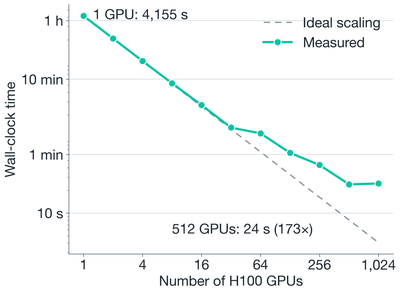


Figure 5: Strong scaling of a 32-qubit H_2 QPE circuit.

Discussion

- **Speed**: beyond 512 GPUs each GPU holds fewer qubits than the walk operator acts on, so qubit reordering, and thus communication, rises sharply. Larger circuits should keep scaling further.
- **Capacity**: GPU memory caps the qubit count, but **host memory at compile time** caps the gate count: beyond ≈ 0.6 B gates (H_4 stage (4), H_6+) it is exhausted.

Distance to industrial application

Same algorithmic framework (qubitization QPE) as the FeMoco nitrogen-fixation benchmark [Lee 2021]:

Axis	This work (executed)	FeMoco [Lee 2021]
System	H_2–H_{10}, 4–20 spin orbitals	FeMoco, 108–152 spin orbitals
Factorization	DF	Tensor hypercontraction (THC)
Logical qubits	≈ 42	≈ 2,100
Toffoli count	10^7–10^8	≈ 5.3×10^9
Runtime	1.5–9 h (1,024 GPUs)	≈ 4 days (FTQC)

The gap is **scale-up within the same qubitization-QPE framework** (DF → THC), not a change of principle; our runs check, at small scale, that estimated resources deliver the target accuracy.

Conclusion & Future Work

- Chemically accurate, end-to-end qubitization QPE executed for H_2 and H_4, with runs up to 42 qubits on 1,024 GPUs and the resources of each reduction stage measured.
- **Future work**: HPC-oriented quantum-circuit compilers (fast, memory-lean, distributed) to overcome the billion-gate compilation wall beyond H_6, and DF → THC factorization toward industrial molecules.

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.

References: [Lee 2021] J. Lee et al., PRX Quantum 2, 030305 (2021).
[Teranishi 2025] Y. Teranishi et al., ACM Trans. Quantum Computing 6(4), 27 (2025).