Collaborative Partitioning of Hybrid Quantum-Classical Workloads

Cost-Model-Driven Planning for Hybrid Tensor Network – QPU Execution

Salvatore Zammuto, Martin Schulz (Advisor)
Chair of Computer Architecture and Parallel Systems
Technical University of Munich, Germany

0. Overview

Background: HPC centers co-host CPUs, GPUs and QPUs, but plan them apart. A GPU offload is priced in time; **a QPU offload must also be priced in accuracy**.

RQ: Should the Quantum Computer help *at all*? If yes, when?
Approach: Price it like any accelerator offload.

1. Split the computation at a seam.
2. Quote time AND error for every option.
3. Pick the fastest that meets the accuracy target.
4. Audit every quote with a receipt.

	Accuracy	Speed	Known before running?
Classical route only — never uses the quantum device	↑ exact — the reference	↑ today · measured; ↓ at scale · projected — fast now; cost explodes as the seam grows	◐ its own cost only — the quantum option is never priced
Always offload — fixed policy: send it to the QPU	↓ device noise dominates today	↓ today · measured; ↑ at scale · projected — better scaling for some selected problems	↓ error depends on the device; not quoted in advance
This work — plan → price → choose → receipt	= matches the better route, by construction	= zero measured regret (one feasible plan)	✓ both routes priced in advance

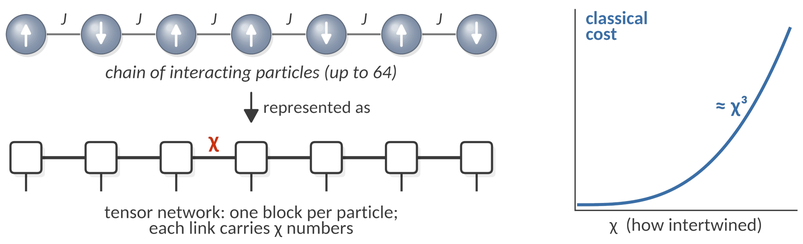
The goal **is not to win a column**, but to **find the best compromise**.

1. Problem Statement

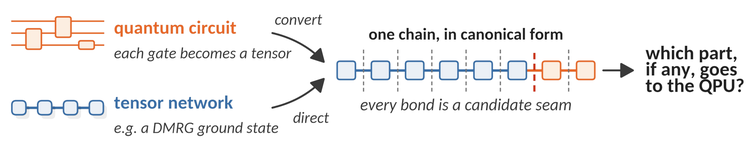
HPC centers now host quantum processors (**QPUs**) next to CPUs and GPUs. Before offloading work to a GPU, we can predict whether it pays off. A QPU needs **two numbers: how long and how accurate**, because its answers carry device noise.

Tensor networks are the natural meeting point: the workhorse for simulating quantum systems on classical machines, with a cost set by one measurable quantity, **how intertwined the two sides of a split are (χ)**.

a The workload

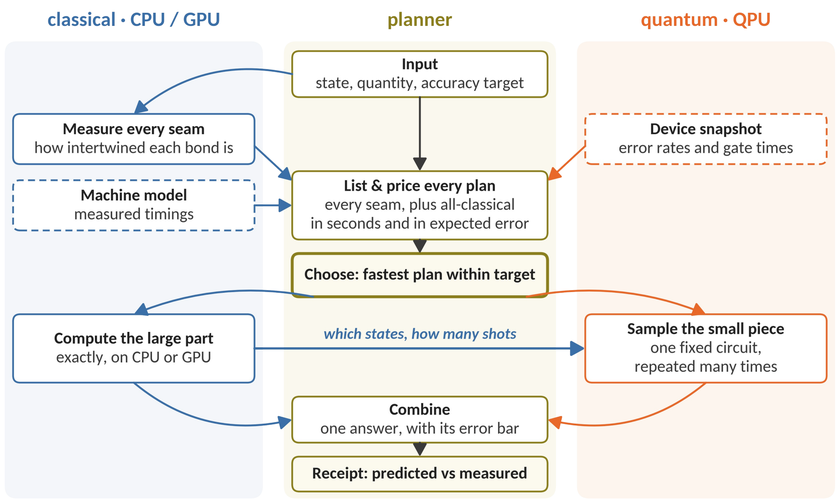


b Two ways in, one representation



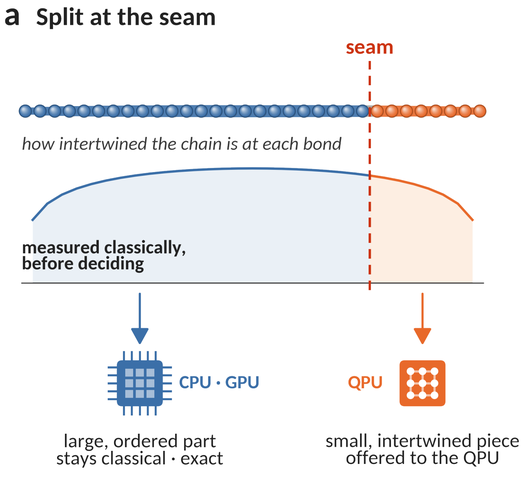
Why it matters. Until a QPU can be quoted like any accelerator, in seconds ***and*** in accuracy, schedulers, users and centers cannot plan with it: it stays a silo.

2. End-to-End Workflow

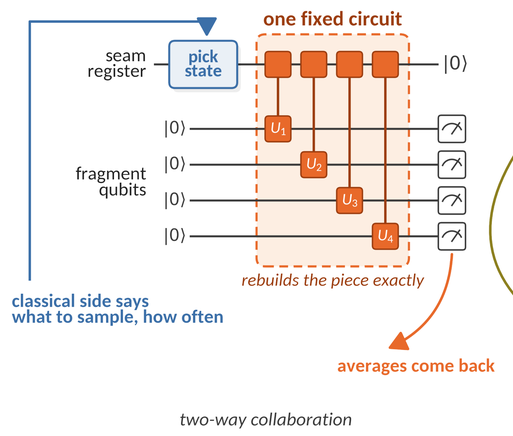


3. Collaborative Partitioning

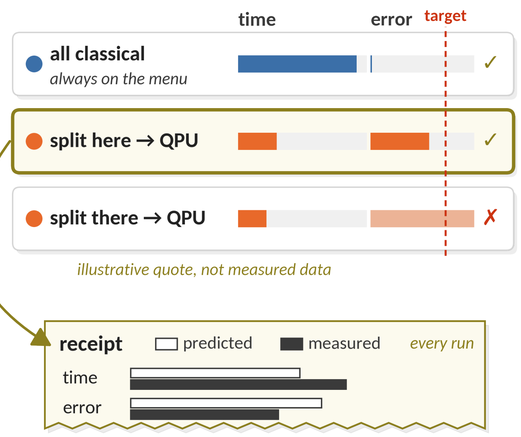
a Split at the seam



b Hand off with one fixed circuit



c Quote · choose · receipt



Time, in seconds

Measured models of this machine. Classical: **a·n·χ³ + b**, fitted per machine. Quantum: gate durations × shots + per-call overhead.

Error on the answer

Truncation + compilation + device noise + 2 × shot noise. **Every term is computed before the run.**

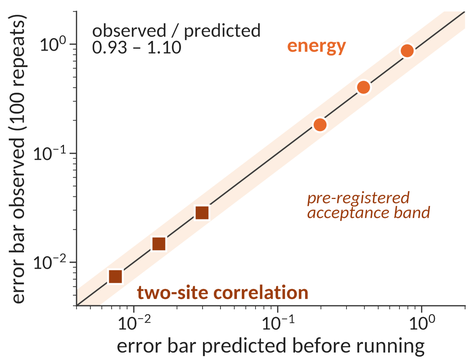
The rule

Pick the **fastest plan whose predicted error meets the target**. The all-classical plan is always a candidate, so the planner can say no.

4. Evaluations

Setup. Transverse-field Ising chains, 32–64 sites, gapped and critical; ground states by DMRG. Quantum route run on Qiskit Aer under a frozen **synthetic** device snapshot (CZ error 0.6%, readout 1.5%). Classical route timed on CPU and on one NVIDIA H100, which holds ≈0.45 s up to χ = 256 where the CPU needs 27 s. Accuracy target ε = 0.05.

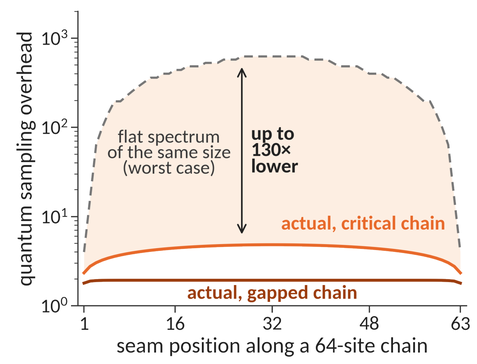
a Are the predicted error bars right?



0.93–1.10 observed ÷ predicted

For shot noise, yes: 100 repetitions per point, inside the pre-registered band (noiseless simulator).

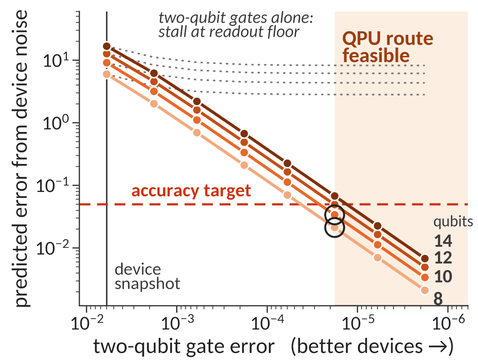
b What does the quantum side pay?



≤ 4.8× sampling overhead, any seam

Set by the physics, not by seam size: a flat spectrum of the same size would cost up to 625×.

c When does the QPU route become viable?



≈ 2×10⁻⁵ two-qubit error

All error channels improving together, the QPU route first meets the target here; winning on time also needs a larger seam than we measured. Our noise bound is conservative (6–8× above observed).

Conclusions

Today this workload stays classical, but the QPU is now priced on equal terms: we can state the answer, the evidence behind it, and what the QPU still needs.

What we found. Shot-noise error bars matched 100 repetitions per point; the predicted envelope covered the realized error on all 10 simulated plans. Device noise is 80–96% of the quantum route's predicted error, so hardware progress feeds straight into the quote. Wall-clock quotes still miss their pre-registered target (median error 60% vs 30%).

What comes next. Real-device calibration snapshots; a tighter noise bound with a coverage guarantee; partitions beyond one seam and beyond chains; quotes served to the runtimes HPC centers already use.

References

[1] Andrew Eddins, Mario Motta, Tanvi P. Gujarati, Sergey Bravyi, Antonio Mezzacapo, Charles Hadfield, and Sarah Sheldon. 2022. Doubling the size of quantum simulators by entanglement forging. PRX Quantum 3, 1 (2022), 010309. https://doi.org/10.1103/PRXQuantum.3.010309
[2] Xiao Yuan, Jinzhao Sun, Junyu Liu, Qi Zhao, and You Zhou. 2021. Quantum simulation with hybrid tensor networks. Phys. Rev. Lett. 127, 4 (2021), 040501. https://doi.org/10.1103/PhysRevLett.127.040501
[3] Lukas Brenner, Christophe Piveteau, and David Sutter. 2025. Optimal wire cutting with classical communication. IEEE Trans. Inform. Theory 71, 10 (2025), 7742–7752. arXiv:2302.03366
[4] Shi-Ju Ran. 2020. Encoding of matrix product states into quantum circuits of one- and two-qubit gates. Phys. Rev. A 101, 3 (2020), 032310. https://doi.org/10.1103/PhysRevA.101.032310
[5] Wei Tang and Margaret Martonosi. 2025. TensorQC: Towards scalable distributed quantum computing via tensor networks. arXiv:2502.03445 [cs.ET].
[6] Nathaniel Tornow, Emmanouil Giortamis, Dennis Sprokholt, Christian Mendl, and Pramod Bhatotia. 2026. qTPU: Hybrid Tensor Networks for Quantum-Classical Acceleration. OSDI 26.

Acknowledgments

The research is part of the Quantum Development Environment, System Software & Integration (Q-DESSI) consortium of the Munich Quantum Valley (MQV), which is supported by the Bavarian state government with funds from the Hightech Agenda Bayern Plus. Furthermore, this research is funded by BMIMI, BMWET, the State of Upper Austria, and the State of Tyrol within the COMET module Quantum Algorithm Engineering (FFG Grant no. 923923) managed by the Austrian Research Promotion Agency FFG.

Code · data · receipts