Collaborative Partitioning of Hybrid Quantum–Classical Workloads

Cost-Model-Driven Planning for Tensor-Network/QPU Execution

Salvatore Zammuto
Technical University of Munich
Chair of Computer Architecture and Parallel Systems
Garching, Germany
salvatore.zammuto@tum.de

Martin Schulz (Advisor)
Technical University of Munich
Chair of Computer Architecture and Parallel Systems
Garching, Germany
schulzm@in.tum.de

Abstract

Quantum processors sit beside CPUs and GPUs in HPC centers, yet nothing decides, before execution, which part of a workload belongs where, and on a noisy device that decision needs two currencies at once, wall-clock time and error. We supply the missing layer for matrix-product-state simulation. A computation is partitioned at one bond: the weakly entangled side is contracted exactly, the strongly entangled fragment becomes a single ancilla-free staircase circuit driven by a seam register of fewer qubits, whose sampling cost is bounded by χ^2/ε^2 and adapts to the Schmidt spectrum. A planner prices every candidate: classical time and quantum time from transpiled gate durations, against a four-term error envelope, always retaining a classical fallback, and emits a per-run receipt comparing prediction with measurement. On Ising chains of up to 64 sites the envelope covers the realized error on every executed plan, and the planner correctly declines the QPU, naming the device quality that would make it viable.

CCS Concepts

• **Computer systems organization** → **Quantum computing**; • **Computing methodologies** → *Massively parallel and high-performance simulations*; • **Software and its engineering** → *Scheduling*.

Keywords

quantum–classical integration, tensor networks, cost models, performance prediction, HPCQC

1 Problem and Motivation

HPC centers are attaching quantum processing units (QPUs) to classical systems [2], yet the two are managed by different logic. For any CPU or GPU a scheduler can ask, before dispatch, how long work will take and whether offloading it is worthwhile. For a QPU that question has no operational answer, and it is strictly harder: quantum devices return noisy results, so "worth it" must be judged in two currencies at once, wall-clock seconds and accuracy. Until quantum resources are planned, priced and error-bounded like classical ones, they remain a silo rather than one more device in the machine room.

We address this for one flagship workload: computing observables of quantum spin systems held as tensor networks. Tensor networks are the dominant classical method for such simulations, and their cost is governed by a single measurable quantity, the entanglement across a cut, which is computable classically in advance.

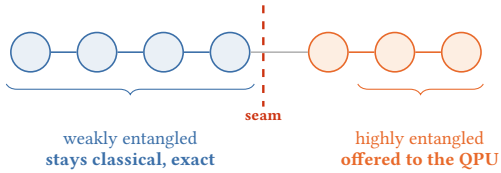


Figure 1: Collaborative partitioning at one bond. The classical side contracts its half exactly *and* tells the quantum side which seam states to sample and how many shots each deserves, so a low-entanglement seam makes the quantum share cheaper automatically.

2 Background and Related Work

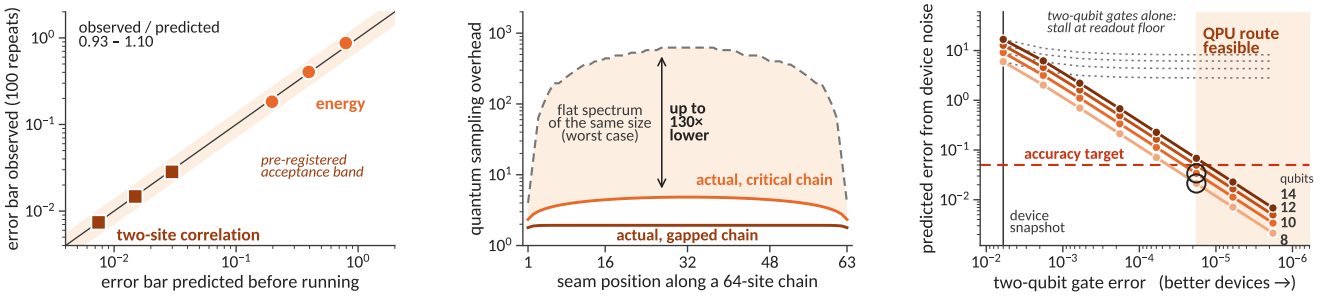
Circuit knitting partitions a quantum circuit into subcircuits that fit smaller devices. Brenner et al. [1] characterize the optimal sampling overhead of cutting n wires as $O(16^n)$ without and $O(4^n)$ with classical communication between fragments. Such methods ignore the state being cut and pay a fixed price for that generality.

A second family exploits structure. Entanglement forging [3] splits a state into two halves and recovers the cross terms classically, doubling the simulable size; it is the closest prior scheme to ours. Hybrid tensor networks [7] place classical and quantum tensors in one network, and Ran [4] encodes a matrix-product state (MPS) into a circuit of one- and two-qubit gates. TensorQC [5] applies a hybrid cost model at the circuit level to schedule cut fragments, and qTPU [6] compiles hybrid tensor-network programs, trading classical work against a gate-error score.

The physics primitives largely exist; what is missing is the systems layer that prices both routes in seconds and in error on the answer, keeps an explicit classical fallback, and audits every prediction against thresholds fixed in advance.

3 Approach and Uniqueness

We partition one computation at a single bond of an MPS (Figure 1). The large, weakly entangled part stays classical and is contracted exactly. The small, strongly entangled fragment goes to the QPU: its right-canonical tensors compose into an isometry realized as one fixed staircase circuit whose only input is a seam register of $r = \log_2 \chi_c$ qubits. Preparing basis states or equal superpositions there yields every fragment quantity required, and because the outer boundary bond is trivial the register provably decouples—no ancilla, no controlled state preparation, no Hadamard test.



(a) Predicted vs. observed shot-noise error bars over 100 repetitions per point: ratio 0.93–1.10.

(b) Quantum sampling overhead: 1.8–4.8× at every seam, up to 130× below a flat spectrum.

(c) Quantum route meets the target at two-qubit error $\approx 2 \times 10^{-5}$ (all channels improving).

Figure 2: Evidence from one frozen setup: Ising chains up to 64 sites, a synthetic device-noise snapshot, accuracy target $\varepsilon = 0.05$.

The collaboration is two-way: the classically known Schmidt spectrum fixes which seam states are sampled and how shots are allocated, so an orderly seam automatically makes the quantum share cheaper—sampling cost is bounded by χ^2/ε^2 in the worst case and far lower for decaying spectra. Against the best-known cutting bound the worst-case order matches rather than improves: the gain is adaptivity, not an exponential separation.

An error envelope $E = B_1 + B_2 + B_3 + 2B_4$ (truncation, compilation, device-noise bias, shot noise) and a cost model, classical $t = a \cdot n \cdot \chi^3 + b$ fitted on the target machine, quantum time from transpiled gate durations under a committed noise snapshot, price every plan in a grid over cut, seam size, backend and shots, always including the fully classical plan. The planner picks the fastest plan with $E \le \varepsilon$; every execution emits a receipt storing predictions beside measurements. Six calibration targets with numeric thresholds were registered before any experiment ran.

4 Results and Contributions

On transverse-field Ising chains up to 64 sites the predictions hold. Split and unsplit contractions agree to 10^{-10}. Predicted shot-noise error bars are calibrated: over 100 repetitions per grid point the empirical-to-predicted ratio lies in $[0.93, 1.10]$ against a pre-registered band of $[0.7, 1.4]$ (Figure 2a). Under the synthetic noise snapshot the envelope covered the realized error on all 10 executed plans, and the chosen plan matched the best measured feasible plan (zero regret, with one feasible plan). Wall-clock quotes are the reported failure (median error 60% vs. a 30% target). Quantum sampling overhead follows the physics: 1.8–4.8× at every seam (Figure 2b).

The headline is a quantitative answer to when the QPU helps. Today the planner correctly keeps every configuration classical: one H100 GPU contracts our largest seam ($\chi = 256$) in 0.45 s where the CPU needs 26.6 s. Scaling the snapshot, the quantum route first meets the accuracy target once two-qubit error falls to 1.8×10^{-5} with all error channels improving together, and wins on time only for seams $\chi \gtrsim 250$ (extrapolated); better gates alone never suffice, since readout error keeps the region closed (Figure 2c). Device noise is 80–96% of the predicted error, so hardware progress feeds straight into the quote. We contribute an ancilla-free, spectrum-adaptive boundary-contraction estimator; a planner that prices both routes in seconds and error and can decline the QPU; and calibrated, pre-registered evidence, including a reported failure, yielding a device-quality threshold HPC operators can act on.

Acknowledgments

This research is part of the Quantum Development Environment, System Software & Integration (Q-DESSI) consortium of the Munich Quantum Valley (MQV), supported by the Bavarian state government with funds from the Hightech Agenda Bayern Plus. It is further 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.

AI-generated text disclosure. Claude Fable 5 was used as a general-purpose writing assistant for drafting and copy-editing this abstract and for scaffolding plotting and analysis code. All results, analyses and claims were produced and verified by the authors, who accept complete accountability for the contents of this text.

References

[1] Lukas Brenner, Christophe Piveteau, and David Sutter. 2025. Optimal wire cutting with classical communication. *IEEE Transactions on Information Theory* 71, 10 (2025), 7742–7752. doi:10.1109/TIT.2025.3601047

[2] Lukas Burgholzer, Jorge Echavarria, Patrick Hopf, Yannick Stade, Damian Rovara, Ludwig Schmid, Ercüment Kaya, Burak Mete, Muhammad Nufail Farooqi, Minh Thanh Chung, Marco De Pascale, Laura Schulz, Martin Schulz, and Robert Wille. 2026. The Munich Quantum Software Stack: Connecting end users, integrating diverse quantum technologies, accelerating HPC. In *Proceedings of the Supercomputing Asia and International Conference on High Performance Computing in Asia Pacific Region (SCA/HPCAsia)*. ACM, 55–67. doi:10.1145/3773656.3773669

[3] 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. doi:10.1103/PRXQuantum.3.010309

[4] Shi-Ju Ran. 2020. Encoding of matrix product states into quantum circuits of one- and two-qubit gates. *Physical Review A* 101, 3 (2020), 032310. doi: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. In *20th USENIX Symposium on Operating Systems Design and Implementation (OSDI 26)*. USENIX Association, 2367–2388.

[7] Xiao Yuan, Jinzhao Sun, Junyu Liu, Qi Zhao, and You Zhou. 2021. Quantum simulation with hybrid tensor networks. *Physical Review Letters* 127, 4 (2021), 040501. doi:10.1103/PhysRevLett.127.040501