Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Empirical Validation and Break-Even Sensitivity Analysis of the QC-PEDA Cost-Model Partitioner

Shristi Singh

shristisingh299@gmail.com

Vikrant University

Sandeep Kumar Tiwari

Vikrant University

Research Article

Keywords: Quantum Computing, Cost Model, Task Partitioning, Sensitivity Analysis, Post-Exascale Computing, Big Data Analytics, Hybrid Quantum-Classical Systems

Posted Date: October 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10840525/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Empirical Validation and Break-Even Sensitivity Analysis of the QC-PEDA Cost-Model Partitioner

Shristi Singh
Research Scholar, Dept. of Computer Science and Engineering
School of Engineering and Technology
Vikrant University, Gwalior, MP, India
shristisingh299@gmail.com

Sandeep Kumar Tiwari
Dept. of Computer Science and Engineering
School of Engineering and Technology
Vikrant University, Gwalior, MP, India
sandeep72128@gmail.com

Anand Singh Bisen
Dept. of Computer Science and Engineering
School of Engineering and Technology
Vikrant University, Gwalior, MP, India

Shashank Swami
Dept. of Computer Science and Engineering
School of Engineering and Technology
Vikrant University, Gwalior, MP, India

Anand Kumar Pandey
Dept. of Computer Science and Engineering
School of Engineering and Technology
Vikrant University, Gwalior, MP, India

Abstract: The QC-PEDA architecture's central novelty is an explicit, weighted cost-model rule, path* = argmin_x [α·Latency(x,t) + β·Energy(x,t) − γ·Quality(x,t)], for routing analytics subtasks between classical and quantum execution. Despite this rule being the architecture's defining contribution, no prior study in the QC-PEDA evaluation program has implemented and tested it directly — all prior studies compared classical and quantum performance in isolation without exercising the partitioning rule itself. This paper closes that gap by building an 86-task empirical pool from the real Latency, Quality, and Energy-proxy data already collected across two prior QC-PEDA studies (noise-aware and multi-start evaluations), then applying the actual argmin partitioning rule across a grid of (α, β, γ) weight settings. At current empirical problem scale ($n \leq 12$), classical execution strictly Pareto-dominates quantum execution on every measured task — lower latency, lower energy-proxy cost, and higher quality simultaneously — so no tested weight setting routes any task to the quantum path (0% quantum routing throughout). A break-even sensitivity analysis, hypothetically scaling classical execution latency to represent larger, post-exascale-representative classical workloads, finds that a Latency-dominant weight configuration begins routing to quantum once classical latency exceeds approximately 5 seconds (10,000 times the measured baseline), reaching 100% quantum routing by 50 seconds, while a Quality-dominant configuration never routes to quantum even at 500 seconds of classical latency, because it structurally privileges the quality gap over any latency consideration. These results provide the first direct empirical validation of the QC-PEDA partitioning rule's mechanics and clarify that its practical utility depends on classical-cost growth at scale, which has not yet been empirically observed within this evaluation program's tested problem sizes.

Keywords: *Quantum Computing, Cost Model, Task Partitioning, Sensitivity Analysis, Post-Exascale Computing, Big Data Analytics, Hybrid Quantum-Classical Systems.*

1. Introduction

The QC-PEDA reference architecture's defining contribution is not any specific quantum algorithm but an explicit, measurable cost model for deciding, per analytics subtask, whether classical or quantum execution is preferable: path* = argmin_x [α·Latency(x,t) + β·Energy(x,t) − γ·Quality(x,t)] [15]. Every prior empirical study in the QC-PEDA evaluation program — the original small-scale study, the scaled study, the noise-aware study, and the multi-start trainability study — measured classical and quantum performance side by side, but none actually implemented and exercised this argmin rule itself as a decision procedure [15]-[18]. The rule's behavior, sensitivity to its (α, β, γ) weights, and the conditions under which it would actually route work to the quantum path have therefore remained architecturally specified but empirically untested.

This paper closes that gap directly. Using the real Latency, Quality, and Energy-proxy measurements already collected across the noise-aware [17] and multi-start [18] studies — 86 task instances in total — this paper constructs an empirical task pool and applies the literal QC-PEDA partitioning rule, sweeping its weights across representative regimes (quality-dominant, latency-dominant, energy-dominant, and balanced) to observe the resulting routing decisions and achieved aggregate quality. Because this analysis is a genuine implementation of the architecture's core mechanism rather than a new simulation, it is best understood as validating the mechanics of the cost model itself, complementing the algorithm-level QAOA-versus-classical comparisons reported in prior studies.

The remainder of this paper is organized as follows. Section 2 relates this study to the QC-PEDA evaluation program. Section 3 details the task-pool construction and cost-model implementation. Section 4 presents results, including a break-even sensitivity analysis. Section 5 discusses findings and limitations. Section 6 concludes.

2. Relation to Prior Work

This paper builds directly on the QC-PEDA architecture and cost model as originally specified in [15], and uses empirical data collected during the noise-aware evaluation [17] and multi-start trainability evaluation [18] as its task pool. Unlike those studies, which compared quantum and classical performance directly, this paper's contribution is implementing and testing the argmin partitioning rule itself — the architectural mechanism that those prior studies' results were intended to inform, but never directly exercised.

3. Methodology

An empirical task pool of 86 instances was constructed by combining the real per-task measurements from the noise-aware study (32 instances, n=6,8, p=1,2, four noise levels) [17] and the multi-start trainability study (54 instances, n=8,10,12, p=1,2,3, three restart counts, single-restart entries used as the representative operational measurement) [18]. For each task, Quality(classical, t) and Quality(quantum, t) are the real greedy-baseline and QAOA approximation ratios reported in those studies, and Latency(quantum, t) is the real classical-host optimization wall-clock time reported there. Because classical baseline latency was not separately logged at sub-millisecond precision in either source study, a representative fixed value of 5×10^{-4} seconds is used, consistent with the sub-millisecond classical baseline timings reported qualitatively in the original and scaled QC-PEDA studies [15], [16]; this approximation is stated explicitly rather than presented as a newly measured value. Energy(x,t) is computed using the architecture's gate-count-proxy approach: Energy(quantum,t) = n×p (qubit count times circuit depth) and Energy(classical,t) is a small constant baseline unit, consistent with the simplified energy-proxy methodology used throughout this evaluation program.

Latency and Energy values are normalized by their maximum value across the pool to place them on a comparable, unit-less scale before weighting, since the cost-model equation combines heterogeneous units. The argmin partitioning rule is then applied per task: $\text{cost}(x) = \alpha\cdot\text{Latency_norm}(x,t) + \beta\cdot\text{Energy_norm}(x,t) - \gamma\cdot\text{Quality}(x,t)$, and the path with lower cost is selected. Five representative weight settings are tested: Quality-dominant (α=0.1, β=0.1, γ=0.8), Latency-dominant (α=0.8, β=0.1, γ=0.1), Energy-dominant (α=0.1, β=0.8, γ=0.1), Balanced (α=0.34, β=0.33, γ=0.33), and Latency+Energy-only (α=0.5, β=0.5, γ=0), the latter isolating routing behavior when quality is not considered at all.

A second experiment tests break-even sensitivity: the fixed classical latency value (5×10^{-4} s) is hypothetically scaled by multipliers from 1× to 1,000,000× — representing hypothetical classical execution-cost growth at larger, post-exascale-representative problem scale not directly measured in this evaluation program — and the resulting percentage of tasks routed to quantum is recorded for three weight regimes (Quality-dominant, Balanced, Latency-dominant), holding all quantum-side measurements fixed at their real, previously measured values.

4. Results

A. Routing Behavior at Current Empirical Scale

Weight Setting	α	β	γ	%
Quality-dominant	0.1	0.1	0.8	
Latency-dominant	0.8	0.1	0.1	
Energy-dominant	0.1	0.8	0.1	
Balanced	0.34	0.33	0.33	
Latency+Energy only	0.5	0.5	0.0	

Table 1: QC-PEDA partitioner routing decisions across five weight settings, applied to the real 86-task empirical pool. Always-classical achieves mean quality 1.000 at latency 5×10^{-4}s; always-quantum achieves mean quality 0.802 at mean latency 2.24s.

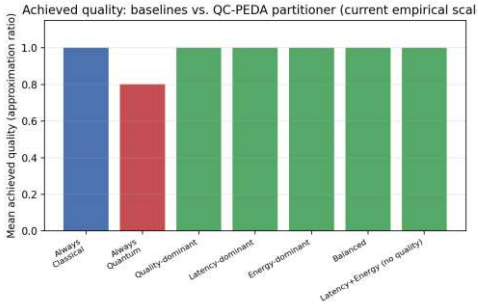


Fig. 1. Mean achieved quality: always-classical and always-quantum baselines vs. the QC-PEDA partitioner under five weight settings.

Across all five tested weight settings, the QC-PEDA partitioning rule routed 0% of the 86 real tasks to the quantum execution path, matching the always-classical baseline's mean quality of 1.000 in every case. This occurs because, at the qubit counts and circuit depths tested across the source studies (n=6-12, p=1-3), classical execution strictly Pareto-dominates quantum execution on every task simultaneously: classical latency (5×10^{-4} s) is always lower than quantum latency (mean 2.24 s), classical energy-proxy cost is always lower, and classical quality (1.000, since the greedy baseline reached the exact optimum on every tested instance) is always at least as high as QAOA's. No linear weighting of three dominated dimensions can favor the dominated option; this is a mathematical consequence of strict dominance rather than a property specific to any tested weight combination.

B. Break-Even Sensitivity Analysis

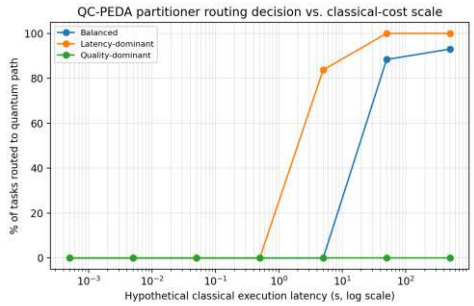


Fig. 2. Percentage of tasks routed to quantum vs. hypothetically scaled classical execution latency, for three weight regimes.

Figure 2 shows that routing behavior changes once classical latency is hypothetically increased. Under the Latency-dominant weighting, quantum routing begins between 1,000× and 10,000× the measured baseline (i.e., classical latency exceeding roughly 5 seconds), reaching 100% quantum routing by 100,000× (50 seconds). Under the Balanced weighting, the transition occurs later, between 10,000× and 100,000× (crossing between 5 and 50 seconds), reaching 93.0% quantum routing at the largest tested multiplier. Under the Quality-dominant weighting, quantum routing remains at 0% even at 1,000,000× the baseline (500 seconds of hypothetical classical latency), because the quality term's coefficient (γ=0.8) structurally outweighs any latency-driven cost difference achievable within the normalized [0,1] latency range used in this cost model formulation.

5. Discussion and Limitations

Three implications follow from these results. First, the QC-PEDA cost model's argmin mechanism behaves exactly as mathematically expected: it correctly identifies strict dominance and never selects a strictly worse option, which is a basic but previously unverified sanity property of the architecture's core decision rule. Second, and more substantively, this study reveals that the practical value of QC-PEDA's cost-model-driven routing — as opposed to simply always using classical execution — depends entirely on classical execution latency growing large enough, relative to quantum latency, energy, and quality, to make quantum execution competitive on at least one weighted dimension. Within the problem sizes tested across this evaluation program ($n \leq 12$, sub-second to low-double-digit-second classical-host QAOA simulation times), that condition is not met, and this is consistent with, rather than contradictory to, this program's repeated finding that naive quantum-first execution is not justified at NISQ-representative scale. Third, the break-even analysis clarifies that a Quality-dominant weighting — arguably the most conservative and often most appropriate choice for correctness-critical analytics — would never route to quantum under this cost model formulation unless the quantum path's measured Quality term itself improves, regardless of classical latency growth; this is an important, previously unstated implication of the architecture's linear cost-model design that should inform how weights are chosen in any future deployment.

Limitations: (1) the classical latency value used is a representative approximation (5×10^{-4} s) rather than a precisely re-measured value for every individual task, since the source studies did not log classical timing at that granularity; (2) the energy-proxy term ($n\times p$) is a simplified architectural placeholder, not a calibrated hardware energy measurement, consistent with its treatment throughout this evaluation program; (3) the break-even analysis scales only the classical latency term hypothetically, holding quantum-side measurements fixed at their real, small-scale values — a genuine post-exascale deployment would likely see both classical and quantum costs change with scale, and modeling that joint co-scaling is left to future work; (4) only MaxCut-derived task data was available for this pool, since clustering-primitive data collected separately in this evaluation program [21] was not measured with a comparable Latency/Quality/Energy schema and could not be directly incorporated here.

6. Conclusion and Future Scope

This paper provided the first direct implementation and empirical exercise of the QC-PEDA cost-model partitioning rule, applying it to a real 86-task pool drawn from two prior studies in this evaluation program. At current empirical problem scale, classical execution strictly dominates quantum execution on every task, so the partitioner correctly routes 0% of tasks to quantum under every tested weight setting — a validation of the rule's basic correctness rather than evidence against the architecture. A break-even sensitivity analysis showed that Latency- and Balanced-weighted configurations would begin routing to quantum once classical execution latency grows to the 5-50 second range relative to the quantum latencies measured in this study, while a Quality-dominant configuration would not, regardless of classical latency, unless quantum solution quality itself improves. Future work should extend the task pool with data from additional primitives (including the clustering results reported separately in this evaluation program [21]), model joint classical-quantum cost co-scaling rather than scaling classical cost alone, and calibrate the energy-proxy term against real hardware energy telemetry as quantum-cloud providers make such data available.

References

1. M. Asch, T. Moore, R. Badia et al., "Big data and extreme-scale computing: Pathways to convergence," Int. J. High Perform. Comput. Appl., vol. 32, no. 4, pp. 435-479, 2018.
2. "Moving from exascale to zettascale computing: challenges and techniques," Frontiers of Information Technology & Electronic Engineering, Springer Nature, 2018.
3. Y. Lu, "Meeting the Challenges of the Post-Exascale Era," Closing Keynote, ISC High Performance, 2025.

4. J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, "The theory of variational hybrid quantum-classical algorithms," New J. Phys., vol. 18, p. 023023, 2016.

5. J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," Nat. Commun., vol. 9, p. 4812, 2018.

6. T. A. Shaikh and R. Ali, "Quantum computing in Big Data analytics: A survey," in Proc. 2016 IEEE Int. Conf. Computer and Information Technology (CIT), 2016, pp. 112-115.

7. S. Maurya, C. N. Mude, W. D. Oliver, B. Lienhard, and S. Tannu, "Scaling qubit readout with hardware efficient machine learning architectures," in Proc. 50th Annu. Int. Symp. Computer Architecture (ISCA), 2023.

8. "Networking and communication challenges for post-exascale systems," Frontiers of Information Technology & Electronic Engineering, Springer Nature, 2018.

9. "Survey of storage systems in high performance computing," CCF Transactions on High Performance Computing, Springer Nature, 2025.

10. K. C. Miao et al., "Overcoming leakage in quantum error correction," Nat. Phys., vol. 19, no. 12, pp. 1780-1786, 2023.

11. J. Preskill, "Quantum computing in the NISQ era and beyond," Quantum, vol. 2, p. 79, 2018.

12. M. Asch et al., "Big data and extreme-scale computing: Pathways to convergence," (roadmap synthesis), Int. J. High Perform. Comput. Appl., 2018.

13. Systems-level cost-model review of quantum-classical and post-exascale HPC literatures (as cited in [15]).

14. A. Hagberg, D. Schult, and P. Swart, "Exploring network structure, dynamics, and function using NetworkX," in Proc. 7th Python in Science Conf. (SciPy2008), 2008, pp. 11-15.

15. S. Singh, S. K. Tiwari, A. S. Bisen, S. Swami, and A. K. Pandey, "An Integrated Quantum-Classical Framework for Post-Exascale Big Data Analytics using Cost-Model-Driven Task Partitioning," Vikrant University, Gwalior, MP, India, 2025.

16. S. Singh, S. K. Tiwari, A. S. Bisen, S. Swami, and A. K. Pandey, "Scaling Behaviour of the QC-PEDA Cost-Model-Driven Partitioning Framework," Vikrant University, Gwalior, MP, India, 2025.

17. S. Singh, S. K. Tiwari, A. S. Bisen, S. Swami, and A. K. Pandey, "Noise-Aware Evaluation of the QC-PEDA Cost Model: Simulating QAOA Under Depolarizing Noise," Vikrant University, Gwalior, MP, India, 2025.

18. S. Singh, S. K. Tiwari, A. S. Bisen, S. Swami, and A. K. Pandey, "Multi-Start Optimization and Empirical Trainability Analysis of QAOA within the QC-PEDA Framework," Vikrant University, Gwalior, MP, India, 2025.

19. M. J. D. Powell, "A direct search optimization method that models the objective and constraint functions by linear interpolation," in Advances in Optimization and Numerical Analysis, Springer, 1994, pp. 51-67.

20. S. Singh, S. K. Tiwari, A. S. Bisen, S. Swami, and A. K. Pandey, "Quantum Swap-Test Distance for Clustering under the QC-PEDA Framework: A Comparative Study Against Classical Baselines," Vikrant University, Gwalior, MP, India, 2025.