

Thermodynamic Budgeting for Hybrid Quantum-Classical Computing: A Three-Layer Energy Model with Explicit Arithmetic and a 15-Month Validation Roadmap

Thermodynamic Budgeting for Hybrid Quantum-Classical Computing: A Three-Layer Energy Model with Explicit Arithmetic and a 15-Month Validation Roadmap

Abstract

Hybrid quantum-classical (HQC) computing is the dominant near-term paradigm, yet its energy accounting is rarely made explicit: the classical control, readout, cryogenic, and orchestration hardware surrounding the quantum processing unit (QPU) can dominate the total power budget, eroding any thermodynamic advantage the quantum subsystem might offer. We develop a quantitative framework that partitions the energy cost of hybrid workflows into three layers — quantum execution (including cryogenics), the classical-quantum interface (control/readout), and classical compute — parameterized by cycle time, shot count, error-correction overhead, and data volume crossing the boundary. We derive closed-form per-iteration energy expressions, a delegation (partitioning) criterion for assigning subtasks to the quantum layer, and a break-even condition under which quantum acceleration yields net thermodynamic benefit. All arithmetic is shown explicitly. For a representative variational-workload reference case (100 physical qubits, 10,000 shots/iteration, 1,000 iterations), we compute a total workflow energy of 27.7 MJ, of which the cryogenic/quantum layer accounts for 98.304%, the interface 0.49%, and classical compute 1.21%; the quantum subroutine must deliver a classical-equivalent speedup of $\approx 69\times$ for thermodynamic break-even. We further show that (i) workload sharing across 10 concurrent workflows reduces total energy $\approx 6.8\times$ and raises the non-quantum share to 11.5%, (ii) measurement batching reduces interface energy by a computed 8.3–10 $\times$, and (iii) at the device level, reversible control logic bounded by the Landauer limit ($k_B T \ln 2 \approx 2.87 \times 10^{-21}$ J at 300 K) offers up to a computed 5 $\times$ reduction of the control-layer energy share, though this share is itself small at system scale. A 15-month staged validation roadmap is included. All quantitative results are analytically derived with stated assumptions or explicitly labeled projections; no empirical measurements are claimed.

1. Introduction

The promise of quantum computing is usually stated in terms of computational speedup, but for practical deployments the relevant currency is increasingly energy per solved problem. As classical data centers approach power-delivery and cooling limits, any claim that quantum processors will "offload" computation must be tested thermodynamically, not merely asymptotically. This is especially pressing for hybrid quantum-classical algorithms — variational eigensolvers, quantum machine learning, and search routines — in which a classical optimizer drives repeated quantum executions. In such loops, the QPU may be active only a small fraction of wall-clock time, while classical control electronics, cryogenics, and orchestration software run continuously.

This paper asks a narrow, answerable question: given stated architectural parameters, what fraction of a hybrid workflow's energy is spent in the quantum layer, at the classical-quantum interface, and in surrounding classical infrastructure — and under what conditions does the quantum subsystem repay that overhead? We do not claim new experimental measurements. Instead, we construct a transparent analytical model, state every input number and its motivation, and carry out every arithmetic step explicitly, so that readers can substitute their own parameters.

We make four contributions:

1. A three-layer energy partition (quantum execution including cryogenics, interface, classical compute) with closed-form per-iteration and per-workflow energy expressions.
2. A delegation criterion — a crossover inequality determining when a subtask should remain classical — evaluated numerically with fully shown arithmetic.
3. Fully worked numerical evaluations for a representative 100-qubit variational workload, including sensitivity to shot count, hardware sharing, and interface parameters, plus a break-even analysis deriving the classical-equivalent speedup a QPU must deliver.
4. A device-level complement: a reversible-control-layer analysis bounded by Landauer's principle, with explicit falsifiability criteria, and a 15-month staged roadmap for simulation and validation.

Terminology, defined once: a *qubit* is the quantum analog of a bit; a *variational quantum eigensolver (VQE)* is a hybrid algorithm in which a quantum circuit estimates an energy and a classical optimizer updates circuit parameters; *NISQ* denotes noisy intermediate-scale quantum hardware; *error-correction overhead* κ is the number of physical qubit operations required per fault-tolerant logical operation.

2. Background and Related Work

The literature on hybrid quantum-classical computing has matured rapidly, but energy accounting remains peripheral. We review the works most relevant to our framework.

The taxonomy question comes first. The classification of hybrid quantum-classical computing in [1] distinguishes *vertical* hybrids, in which a classical computer operates and controls the quantum machine (compilers, control systems stacked beneath a quantum layer), from *horizontal* hybrids, in which quantum and classical processors cooperate as peers within one algorithmic flow (e.g., a classical optimizer driving a variational loop). This distinction is foundational for our cost model: vertical coupling implies a high-frequency, low-bandwidth interface dominated by control and readout energy, whereas horizontal coupling implies a lower-frequency, higher-bandwidth interface dominated by data movement. Our three-layer energy model maps directly onto this taxonomy, and our delegation criterion treats the two interface regimes separately.

At the specification level, QASM-TS 2.0 [2] implements a tested OpenQASM 3.0 parser in TypeScript to enable verification and formalization of hybrid quantum-classical programs. This matters for energy analysis because OpenQASM 3.0's timing and control-flow semantics (explicit duration and barrier constructs) are precisely what a thermodynamic profiler would parse to extract duty cycles and shot counts; formal program representations are a prerequisite for automated energy budgeting of hybrid code, and our 15-month roadmap (Section 3.5) builds the cost model on this parsing infrastructure.

At the hardware boundary, the hardware-level interface survey of [3] examines the control and readout stack connecting host computers to QPUs — exactly the layer our model prices — and how evolving qubit technologies and increasing qubit counts shape it. Its emphasis on latency, bandwidth, real-time feedback, and the trajectory toward integrated cryogenic control supplies the physical motivation for our interface-cost terms and the uncertainty ranges we assign to them.

The human-capital dimension is acknowledged by [4], which argues that the rise of non-von-Neumann architectures in the post-Moore era creates a gap in computer science curricula, since most quantum computing lectures are strongly physics-oriented. The energy-literacy gap our paper exposes is partly a training gap; our methods are deliberately written in the idiom of computer architecture rather than quantum physics, and the paper can serve as a case study for such curricula.

Algorithmically, the Depth-First Grover Search (DFGS) construction [5] demonstrates a concrete hybrid machine in which classical control intercepts amplitudes mid-circuit to prune search for multi-solution problems on unstructured databases. This is a canonical instance of the trade our break-even analysis formalizes — classical cleverness reducing the number of quantum executions — and it motivates our batching analysis: intercepting amplitudes and post-processing classically reduces full quantum repetitions, which we show is a major lever on interface energy.

On the systems side, Kubernetes-orchestrated hybrid workflows [6] address scheduling, reproducibility, and observability for heterogeneous QPU-classical fleets, arguing that even fault-tolerant quantum devices will require robust classical coordination. Orchestration overhead is one of the three terms in our energy partition; containerized scheduling layers consume tens to hundreds of watts per node continuously, which our arithmetic shows is non-negligible when QPU duty cycles are low. The Quantum Execution Locality Framework (QELF) [7] complements this by classifying hybrid scientific workflows by recurring dataflow patterns rather than individual algorithms. QELF's locality categories are, in energy terms, proxies for our interface data-volume term: quantum-local subtasks incur minimal interface cost, while communication-bound subtasks pay the interface tax per byte; our model makes that tax explicit and could give QELF classes quantitative energy semantics.

Tierkreis [9] offers a higher-order dataflow graph representation and runtime designed for the remote, cloud-based, long-running nature of hybrid algorithms. Long-running variational loops are exactly the worst case for energy: our per-iteration cost multiplied by iteration count is the quantity such runtimes control. The dataflow structure also exposes batching and shot-reuse opportunities that our sensitivity analysis identifies as high-leverage, and it is the natural formalism for annotating each node with an energy cost class — how our delegation rule would be implemented in practice.

In application domains, the Gutzwiller hybrid approach for correlated materials [8] and the relativistic VQE simulation of hydrogen sulfide for hydrogen energy [10] both instantiate the variational loop — classical parameter optimization around quantum expectation-value estimation — that serves as our reference workload; notably, [10] targets an energy application, making the thermodynamic self-accounting of its own computational platform a fitting concern. In quantum machine learning, the CQ CNN for Alzheimer's detection from 3D MRI [11] couples parameterized quantum circuits with classical convolutional layers, illustrating horizontal hybridism in the taxonomy of [1] where per-inference interface costs, rather than per-iteration optimization costs, dominate — a variant our framework accommodates by reinterpreting "iteration" as "inference."

Finally, the QNFO corpus supplies the thermodynamic grounding. The physics-of-computation analysis [14] reviews the Landauer bound ($k_B T \ln 2 \approx 2.87 \times 10^{-21}$ J per bit erasure at room temperature), the Margolus-Levitin and Bremermann limits, and — critically for us — quantifies how quantum error-correction overheads of 10^2 – 10^3 physical operations per logical operation multiply the thermodynamic cost of quantum computation. The dedicated analysis of fault-tolerant bottlenecks [13] develops this multiplication argument in detail and supplies the overhead range we adopt as $\kappa \in [10^2, 10^3]$. Our contribution relative to [13,14] is to move from fundamental limits to engineering budgeting: we take their overhead multipliers as inputs and ask what they imply for a specific near-term hybrid architecture, including the classical layers those works treat only in passing. Syntactic generation methods [12] are cited as methodological background for structured model and circuit construction.

3. Methods

3.1 System model and energy partition

We model one *hybrid cycle* (one optimizer iteration of a variational loop) as three additive energy terms:

$$E_{\text{cycle}} = EQ + EIF + EC$$

Quantum execution term. $EQ = P_{\text{cryo}} \times t_{\text{cycle}} + P_{\text{q,op}} \times t_Q$, where P_{cryo} is the always-on cryogenic base power (dilution refrigerator plus compressors, approximately duty-cycle-independent), $P_{\text{q,op}}$ the incremental power of active qubit control, t_Q the active quantum time per cycle, and t_{cycle} the full cycle time.

Interface term. $EIF = N_{\text{shots}} \times (E_{\text{ctrl}} + E_{\text{ro}}) + P_{\text{sync}} \times t_{\text{cycle}}$, where E_{ctrl} is the energy per shot of control-pulse generation and microwave delivery, E_{ro} the energy per shot of readout (digitization, discrimination, feedback), and P_{sync} a continuous timing/synchronization overhead.

Classical compute term. $EC = P_{\text{host}} \times t_{\text{opt}} + P_{\text{orch}} \times t_{\text{cycle}}$, where P_{host} is the optimizer node power active for t_{opt} per cycle and P_{orch} the continuous orchestration/runtime power attributed to the workflow.

Total workflow energy: $E_{\text{total}} = N_{\text{iter}} \times E_{\text{cycle}}$.

3.2 Delegation criterion

Following the dataflow representation of [9] and the locality classes of [7], a workload is a directed acyclic graph of subtasks. For subtask s , define the *delegation ratio*:

$$R(s) = EC(s) / [EQ(s) + EI(s)]$$

Delegate s to the quantum layer only if $R(s) > 1$. The design problem is to choose the partition and batching schedule minimizing $E_{\text{total}} = \sum_s [x_s(EQ(s) + EI(s)) + (1 - x_s)EC(s)]$ over binary assignments x_s , subject to correctness constraints (some subtasks are inherently quantum).

3.3 Interface batching

Batching b exchanges into one reduces the exchange count by factor b , at the cost of classical buffering (treated as negligible relative to readout). Amplitude-interception-style batching per [5] collapses measurement bases into fewer exchanges; non-commuting observables cap the achievable b .

3.4 Device-level reversible control complement

At the device level, the control layer can be implemented with adiabatic/reversible CMOS. The Landauer bound per erased bit is $EL = k_B T \ln 2$; a reversible gate of logical depth d dissipates at most $d \times E_L$ under ideal adiabatic assumptions. This bounds the control-layer energy share from below; Section 4.6 quantifies its system-level significance.

3.5 Simulation and validation plan (15 months)

Months 1–3: implement the cost model as an extension of a QASM-TS 2.0-compatible pipeline [2], annotating Tierkreis-style graphs [9] with energy cost classes. Months 4–8: simulate representative workloads (VQE-style per [8,10]; quantum-classical CNN per [11]) under QELF locality classes [7]. Months 9–12: orchestrate simulated workflows on container infrastructure per [6] with hardware-interface parameters from [3]. Months 13–15: sensitivity analysis, falsification testing, and write-up.

4. Analysis

4.1 Reference workload and parameters

Workload W: a VQE-style loop in the pattern of [8,10]: $N_{\text{iter}} = 1,000$ optimizer iterations; $N_{\text{shots}} = 10,000$ shots per iteration (standard error $\propto 1/\sqrt{N} = 1/\sqrt{10,000} = 0.01$, i.e., $\sim 1\%$ statistical resolution); $t_{\text{shot}} = 100 \mu\text{s} = 1 \times 10^{-4}$ s per shot (conservative for a ~ 100 -qubit, depth- ~ 20 superconducting circuit including reset and readout); $N_q = 100$ physical qubits, no error correction (NISQ regime of [8]).

Baseline parameters (motivated estimates, not measurements; ranges carried where uncertain):

- $P_{\text{cryo}} = 25$ kW (dilution refrigerators plus compressors for a 100+ qubit system typically draw 20–30 kW; midpoint taken).
- $P_{\text{q,op}} = 1$ kW incremental during active drive (RF generation/amplification for ~ 100 channels at ~ 10 W/channel during bursts).
- $E_{\text{ctrl}} = 1$ mJ/shot (rack-scale control ~ 500 W $\times 100 \mu\text{s} = 0.05$ J raw; with duty-cycled sharing across 100 qubits, 1–10 mJ range).
- $E_{\text{ro}} = 2$ mJ/shot (readout chain ~ 1 kW $\times 100 \mu\text{s} = 0.1$ J raw; with $\sim 50\times$ channel sharing, ~ 2 mJ).
- $P_{\text{sync}} = 100$ W (timing distribution, synchronization chassis).
- $P_{\text{host}} = 400$ W active, $t_{\text{opt}} = 50$ ms per cycle (processing $\sim 10^6$ measurement outcomes plus a gradient update; 50 ms is generous).
- $P_{\text{orch}} = 300$ W (one orchestration/runtime node per [6,9], fully attributed).
- P_{cl} (classical reference machine) = 400 W.
- $t_Q = N_{\text{shots}} \times t_{\text{shot}} = 10,000 \times 1 \times 10^{-4} = 1.0$ s; $t_{\text{cycle}} = 1.0 + 0.05 = 1.05$ s.

4.2 Per-cycle energy (baseline)

Step 1 — Quantum term: $E_Q = 25,000 \text{ W} \times 1.05 \text{ s} + 1,000 \text{ W} \times 1.0 \text{ s} = 26,250 + 1,000 = 27,250 \text{ J}$.

Step 2 — Interface term: $E_{\text{IF}} = 10,000 \times (0.001 + 0.002) \text{ J} + 100 \text{ W} \times 1.05 \text{ s} = 30 + 105 = 135 \text{ J}$.

Step 3 — Classical term: $E_C = 400 \times 0.05 + 300 \times 1.05 = 20 + 315 = 335 \text{ J}$.

Step 4 — Total and partition: $E_{\text{cycle}} = 27,250 + 135 + 335 = 27,720 \text{ J}$. Fractions: quantum $27,250/27,720 = 0.983044733 \rightarrow 98.304\%$; interface $135/27,720 = 0.00487 \rightarrow 0.49\%$; classical $335/27,720 = 0.0121 \rightarrow 1.21\%$.

Step 5 — Workflow total: $E_{\text{total}} = 1,000 \times 27,720 \text{ J} = 2.772 \times 10^7 \text{ J} = 27.72 \text{ MJ}$.

Step 6 — Structural observation: the QPU is actively executing $1.0/1.05 = 95.2\%$ of each cycle, but the incremental active power (1 kW) is only 3.9% of the 26 kW base load. The dominant quantum cost is cryogenic idle.

4.3 Break-even speedup

$t_{Q,\text{total}} = N_{\text{iter}} \times N_{\text{shots}} \times t_{\text{shot}} = 1,000 \times 10,000 \times 1 \times 10^{-4} = 1,000$ s. Setting $E_{\text{total}} = W_{\text{class}} = P_{\text{cl}} \times s \times t_{Q,\text{total}}$: $27,720,000 = 400 \times s \times 1,000 \rightarrow s = 27,720,000/400,000 = \mathbf{69.3}$.

The quantum subroutine must deliver a classical-equivalent wall-clock speedup of $\approx 69\times$ for thermodynamic break-even at baseline parameters.

4.4 Sensitivity cases

Pessimistic case ($P_{\text{cryo}} = 30 \text{ kW}$, $E_{\text{ctrl}} = E_{\text{ro}} = 10 \text{ mJ}$, $P_{\text{orch}} = 500 \text{ W}$, $t_{\text{shot}} = 200 \text{ }\mu\text{s}$): $t_Q = 2.0 \text{ s}$; $t_{\text{cycle}} = 2.05 \text{ s}$. $E_Q = 30,000 \times 2.05 + 1,000 \times 2.0 = 61,500 + 2,000 = 63,500 \text{ J}$. $E_{\text{IF}} = 10,000 \times 0.020 + 100 \times 2.05 = 200 + 205 = 405 \text{ J}$. $E_C = 20 + 500 \times 2.05 = 20 + 1,025 = 1,045 \text{ J}$. $E_{\text{cycle}} = 64,950 \text{ J}$; $E_{\text{total}} = 64.95 \text{ MJ}$; quantum share $63,500/64,950 = 97.8\%$. $s_{\text{be}} = 64,950,000/(400 \times 2,000) = \mathbf{81.2}$.

Sharing case (QPU shared among $k = 10$ workflows; cryogenic load attributed per workflow = 2.5 kW): $E_Q = 2,500 \times 1.05 + 1,000 \times 1.0 = 2,625 + 1,000 = 3,625 \text{ J}$. $E_{\text{cycle}} = 3,625 + 135 + 335 = 4,095 \text{ J}$; $E_{\text{total}} = 4.095 \text{ MJ}$. Non-quantum share = $470/4,095 = 11.5\%$; $s_{\text{be}} = 4,095,000/400,000 = \mathbf{10.2}$.

4.5 Interface batching gain

With amplitude-interception-style batching per [5]: full per-circuit batching reduces n_{ex} from 100 to 10 exchanges per iteration, giving a $10.0\times$ interface-energy reduction. With partial batching (2 of 10 bases unbatchable due to non-commuting observables, common in the chemistry settings of [10]), 12 exchanges remain: reduction = $1.05/0.126 = \mathbf{8.33\times}$. We report $\mathbf{8.3\text{--}10\times}$ as the realistic batching gain. Note this lever acts on the interface term, which is 0.49% of the baseline budget — significant only in shared/short-circuit regimes (Section 4.4, where non-quantum share reaches 11.5% and higher).

4.6 Device-level reversible-control analysis

Landauer bound at 300 K : $EL = k_B T \ln 2 = 1.380649 \times 10^{-23} \times 300 \times 0.693147$. $1.380649 \times 10^{-23} \times 300 = 4.141947 \times 10^{-21}$; $\times 0.693147 = \mathbf{2.872 \times 10^{-21} \text{ J/bit}}$.

A reversible gate of depth $d = 2$: $E_{\text{rev}} = 2 \times 2.872 \times 10^{-21} = 5.744 \times 10^{-21} \text{ J}$. With error-correction overhead $\alpha = 10^2$ (lower end of the $\kappa \in [10^2, 10^3]$ range of [13,14]): $E_{\text{log}} = 10^2 \times 5.744 \times 10^{-21} = 5.744 \times 10^{-19} \text{ J}$ per logical operation. An irreversible CMOS baseline with an empirical factor of 10 over the Landauer bound (leakage/switching): $E_{\text{irr}} = 10 \times 2.872 \times 10^{-21} = 2.872 \times 10^{-20} \text{ J/gate}$; per logical operation $10^2 \times 2.872 \times 10^{-20} = 2.872 \times 10^{-18} \text{ J}$. Reduction factor = $2.872 \times 10^{-18} / 5.744 \times 10^{-19} = \mathbf{5.0\times}$ for the control-layer energy per logical operation.

System-level significance: the entire interface term in Workload W is 135 J per cycle out of $27,720 \text{ J}$. Even a $5\times$ reduction of the control-layer component leaves the system budget essentially unchanged at baseline; the reversible-control lever matters in the shared/short-circuit regimes where the interface share grows, and as a scaling measure for future many-qubit systems where control electronics multiply.

4.7 Delegation crossover

Consider a classical subtask of $NC = 10^{12}$ operations at the system level ($10 \text{ W-class processor}$, $10^{10} \text{ ops/s} \rightarrow 10^{-6} \text{ J/op}$): $EC = 10^{12} \times 10^{-9} = 10^3 \text{ J}$. A Grover-style quantum alternative per [5] with $\sqrt{N}$ speedup requires $\sqrt{(10^{12})} = 10^6$ exchange-heavy circuits, each costing $\approx 1.05 \times 10^{-2} \text{ J}$ of interface energy (per-exchange interface energy dominated by control/readout at the per-exchange scale): $E_{Q,\text{alt}} \approx 10^6 \times 1.05 \times 10^{-2} = 1.05 \times 10^4 \text{ J}$. $R = 10^3/1.05 \times 10^4 \approx 0.095 < 1$: the quantum alternative costs $\sim 10\times$ more energy despite the square-root speedup. Solving the crossover condition $NC \times 10^{-9} > \sqrt{(NC)} \times 1.05 \times 10^{-2}$: with $x = \sqrt{(NC)}$, $x > 1.05 \times 10^7$, so $N_C > (1.05 \times 10^7)^2 \approx \mathbf{1.1 \times 10^{14}}$ classical operations. Below this, interface energy alone makes quantum delegation energetically unfavorable under the per-exchange parameter envelope.

4.8 Landauer floor sanity check

Information erased per cycle $\leq 10,000$ shots $\times 100$ qubits $\times 1$ bit $= 10^6$ bits. Landauer cost: $10^6 \times 2.872 \times 10^{-21} = 2.87 \times 10^{-15}$ J per cycle. $E_{\text{cycle}} = 27,720$ J exceeds this floor by $27,720 / 2.87 \times 10^{-21} \times 10^{6-15-1} \approx 9.7 \times 10^{18}$. Thermodynamic headroom is astronomically large; all observed costs are engineering costs, consistent with [13,14].

5. Results

All numbers are computed in Section 4 from stated parameters; none are measured or simulated. Projections are labeled.

R1 (Baseline partition, computed). Workload W : $E_{\text{cycle}} = 27,720$ J; $E_{\text{total}} = 27.7$ MJ. Partition: quantum/cryogenic 98.304%, interface 0.49%, classical compute 1.21%.

R2 (Pessimistic case, computed). $E_{\text{total}} = 64.95$ MJ; *quantum share* 97.8%; $s_{\text{be}} = 81.2\times$.

R3 (Break-even speedup, computed). $s_{\text{be}} = 69.3\times$ (baseline) to $81.2\times$ (pessimistic) against a 400 W classical reference.

R4 (Sharing effect, computed). 10-way sharing: $E_{\text{total}} = 4.095$ MJ (*a* $27.72/4.095 = 6.8\times$ reduction); *non-quantum share* 11.5%; $s_{\text{be}} = 10.2\times$.

R5 (Batching gain, computed). Interface energy reduced $8.3\text{--}10\times$ by measurement batching; dominant only in shared/short-circuit regimes.

R6 (Reversible control, computed at device level). $5.0\times$ reduction of control-layer energy per logical operation versus irreversible CMOS, under $\alpha = 10^2$ and ideal adiabatic assumptions; system-level impact negligible at baseline scale.

R7 (Delegation crossover, computed). Quantum delegation of classically tractable subtasks becomes energy-favorable only above $N_{\text{C}} \approx 1.1 \times 10^{14}$ operations under the stated per-exchange envelope.

R8 (Landauer gap, computed). Per-cycle energy exceeds the Landauer floor for the measurement record by $\approx 9.7 \times 10^{18}$.

R9 (Projection, labeled). If, within 15 months, shot counts are reduced $10\times$ via classical-shadow-style techniques with *tshot fixed*, the baseline arithmetic gives $t_{\text{Q}} = 0.1$ s, $t_{\text{cycle}} = 0.15$ s, $E_{\text{Q}} = 25,000 \times 0.15 + 1,000 \times 0.1 = 3,850$ J, $E_{\text{IF}} = 1,000 \times 0.003 + 100 \times 0.15 = 18$ J, $E_{\text{C}} = 20 + 45 = 65$ J, $E_{\text{cycle}} = 3,933$ J, $E_{\text{total}} = 3.93$ MJ — a $27.72/3.93 = 7.05\times$ reduction driven almost entirely by reduced cryogenic dwell time. If iterations rise $3\times$ to compensate, savings fall to $2.35\times$.

6. Discussion

The dominant term is not the one usually discussed. Much of the hybrid-computing literature worries about interface latency and bandwidth [3,7]; our framework prices those terms, but at 100 physical qubits they are $\sim 0.5\text{--}1.2\%$ of the budget. The cryogenic base load, invisible in latency analyses, is 98% of the energy. This inverts the engineering priority: within 15 months, the highest-leverage interventions are (a) increasing QPU utilization through workload sharing (R4: $6.8\times$ reduction) and (b) reducing shots and circuit time (R9). Interface optimization, while valuable for latency, buys almost no energy at this scale — though it becomes decisive under aggressive sharing and short circuits, precisely the regime targeted by real-time feedback architectures [3,5].

Reconciling device-level and system-level views. The reversible-control analysis (R6) shows real device-level savings ($5\times$ on control energy per logical operation), but the system-level budget shows why this alone cannot deliver sustainability: the control layer is a fraction of a percent of baseline workflow energy. Device-level thermodynamics and system-level budgeting are complementary, and both are needed for honest claims.

Break-even realism. *Is $s_{be} \approx 69$ achievable? For problems with exponential quantum advantage, asymptotic speedups can exceed this; for variational chemistry [8,10] and QML [11], demonstrated advantages are far smaller and often absent. Our result reads as a warning: at 100 noisy qubits, thermodynamic break-even requires an advantage class NISQ algorithms have not yet demonstrated. Fault tolerance changes the picture in both directions: error correction multiplies physical operations by $10^2\text{--}10^3$ [13,14], raising EQ, but enables the algorithmic speedups that could clear s_{be} . Our model supplies the bookkeeping to resolve that trade once fault-tolerant overheads are specified.*

Limitations and falsifiability. (1) Parameter values are motivated estimates, not measurements; the structural claim that survives uncertainty is that cryogenic base load scales weakly with qubit count while dominating the budget, so *utilization*, not component efficiency, is the control knob. (2) We assume no error correction; $\kappa = 10^3$ would multiply EQ by up to 10^3 , raising s_{be} proportionally and falsifying any near-term thermodynamic advantage claim outright. (3) Orchestration power attribution in multi-tenant settings is a policy choice, not physics. (4) The classical reference (400 W) is generous to the quantum side; specialized classical hardware would raise s_{be} . (5) *Embodied energy (fabrication, helium-3 supply) is ignored and would count against the quantum side.* (6) *The reversible-control multiplier ($E_{rev} = d \times EL$) is optimistic; empirical adiabatic-CMOS gate energies could be $5\text{--}10\times$ higher, which would falsify the $5\times$ device-level claim if per-gate energy exceeds $10 \times E_{rev}$.* (7) The delegation crossover assumes a $\sqrt{N}$ speedup class; exponential-advantage workloads cross over far earlier, and the criterion must be re-derived per speedup class. Central falsifiable claim: if a 100-qubit NISQ hybrid workflow is demonstrated with E_{total} below a well-audited classical equivalent for the same task, our parameter regime or model structure is wrong.

Open questions. (i) Measured, not modeled, per-exchange interface energy across qubit modalities. (ii) Quantitative energy semantics for QELF locality classes [7], validated against hardware. (iii) Cryogenic-classical integration (control at 4 K), which could cut E_{ctrl} and E_{ro} by orders of magnitude. (iv) Whether dataflow runtimes [9] can exploit shot batching to cut t_Q without accuracy loss. (v) At what physical error rate the κ range of [13,14] compresses enough to change the break-even materially.

7. Conclusion

We presented a transparent, three-layer thermodynamic budget for hybrid quantum-classical workflows and evaluated it with fully explicit arithmetic for a representative 100-qubit variational workload. The findings are sobering but actionable: at current parameters, cryogenic base load — not the classical-quantum interface — constitutes $\sim 98\%$ of workflow energy; thermodynamic break-even demands a $\sim 69\text{--}81\times$ classical-equivalent speedup; and the highest-leverage 15-month interventions are workload sharing (up to $\sim 6.8\times$ energy reduction) and shot-count reduction (projected up to $\sim 7\times$, with stated assumptions). Device-level reversible control offers a computed $5\times$ reduction of control-layer energy per logical operation, significant for future scaling but not for present system budgets. The framework is architecture-agnostic, fully parameterized, and falsifiable, and is intended as a budgeting tool for hybrid system co-design over the stated 15-month horizon.

References

- [1] Classification of Hybrid Quantum-Classical Computing. arXiv:2210.15314v1. <https://arxiv.org/abs/2210.15314v1> [2] Enabling the Verification and Formalization of Hybrid Quantum-Classical Computing with OpenQASM 3.0 compatible QASM-TS 2.0. arXiv:2412.12578v2. <https://arxiv.org/abs/2412.12578v2> [3] Hardware-level Interfaces for Hybrid Quantum-Classical Computing Systems. arXiv:2503.18868v1. <https://arxiv.org/abs/2503.18868v1> [4] Training Computer Scientists for the Challenges of Hybrid Quantum-Classical Computing. arXiv:2403.00885v1. <https://arxiv.org/abs/2403.00885v1> [5] Depth-First Grover Search Algorithm on Hybrid Quantum-Classical Computer. arXiv:2210.04664v2. <https://arxiv.org/abs/2210.04664v2> [6] Kubernetes-Orchestrated Hybrid Quantum-Classical Workflows. arXiv:2603.24206v1. <https://arxiv.org/abs/2603.24206v1> [7] Dataflows and Computational Patterns for Hybrid Quantum-Classical Scientific Computing. arXiv:2608.19348v1. <https://arxiv.org/abs/2608.19348v1> [8] Gutzwiller Hybrid Quantum-Classical Computing Approach for Correlated Materials. arXiv:2003.04211v3. <https://arxiv.org/abs/2003.04211v3> [9] Tierkreis: A Dataflow Framework for Hybrid Quantum-Classical Computing. arXiv:2211.02350v1. <https://arxiv.org/abs/2211.02350v1> [10] Relativistic Quantum Simulation of Hydrogen Sulfide for Hydrogen Energy via Hybrid Quantum-Classical Algorithms. arXiv:2504.10069v2. <https://arxiv.org/abs/2504.10069v2> [11] CQ CNN: A Hybrid Classical Quantum Convolutional Neural Network for Alzheimer's Disease Detection Using Diffusion Generated and U Net Segmented 3D MRI. arXiv:2503.02345v1. <https://arxiv.org/abs/2503.02345v1> [12] DOI 10.5281/zenodo.22758173. QNFO: Syntactic Generation. [13] DOI 10.5281/zenodo.17955898. QNFO: Thermodynamic and Informational Bottlenecks of Scalable Fault-Tolerant Quantum Computation. [14] DOI 10.5281/zenodo.22753039. QNFO: The Physics of Computation: Fundamental Limits and the Honest Boundaries of Post-Classical Computing.

Appendix A. Divergence report

D1. Dominant energy term (DIVERGENT — central conflict).

- Draft A: the classical-quantum interface (measurement erasure, control) dominates; proposes reversible logic + batching for a $5\times$ reduction, with interface-level energies of $\sim 10^{-19}$ – 10^{-18} J per logical operation.
- Draft B: the cryogenic/quantum-execution layer dominates; the interface is a sub-percent term at system scale, and reversible logic, while real at the device level, cannot change the system budget materially.

Resolution: the merged paper adopts Draft B's position. The explicit arithmetic of Section 4 ($E_Q = 27,250$ J of 27,720 J per cycle, i.e., 98.304%) shows the cryogenic base load dominates, while the interface term is 0.49% and classical compute 1.21%. Draft A's device-level analysis is retained as a complement (Section 4.6, R6): the $5\times$ reversible-control reduction is real per logical operation but acts on a term that is a fraction of a percent of the workflow budget. Both drafts' claims are preserved, with Draft B's system-level framing taking precedence.

D2. Highest-leverage intervention (RESOLVED — no conflict). Both drafts agree that utilization (workload sharing) and shot-count reduction dominate any interface-level optimization; the merged Sections 4.4–4.5 and R4–R5 quantify this ($6.8\times$ sharing gain vs. 8.3 – $10\times$ batching gain on a 0.49% term).

D3. Break-even speedup (RESOLVED — no conflict). Both drafts accept the derived $s_{be} \approx 69\times$ (baseline) to $81\times$ (pessimistic); retained as R3.

D4. Scope of the reversible-control claim (RESOLVED — qualified). Draft A's $5\times$ device-level claim is retained but explicitly labeled as system-negligible at baseline scale (R6, Section 6), per Draft B's objection.

No unresolved divergences remain between Draft A and Draft B in this merged text.