

Thermodynamic Divergence of Active Quantum Error Correction Overhead: A Kapitza-Limited Feasibility Threshold for Gate-Based Architectures

Thermodynamic Divergence of Active Quantum Error Correction Overhead: A Kapitza-Limited Feasibility Threshold for Gate-Based Architectures

Abstract

Active quantum error correction (QEC) is conventionally analyzed as an algorithmic resource problem: how many physical qubits and gate cycles are needed per logical qubit. We propose a complementary thermodynamic feasibility criterion. In a gate-based architecture, every QEC cycle is driven by control wiring whose dissipation we model as a linear per-qubit term plus a quadratic routing term, while the heat sink available to each qubit is the Kapitza thermal conductance of its interconnects, which bounds the sustainable heat flux for a given allowed temperature rise. Setting generated heat equal to extractable heat yields an explicit critical qubit count N_c above which no steady state exists at the operating temperature, independent of code distance: increasing code distance adds qubits, and therefore heat, faster than it adds logical protection. Under explicitly stated parameter assumptions we compute $N_c = 390$ for the quadratic-overhead model and $N_c \approx 738$ for an exponential wiring-dissipation model, with all arithmetic shown, and we derive sensitivities ($N_c = 190$ for a halved temperature budget; $N_c = 3,990$ for a tenfold Kapitza improvement) and logical-qubit budgets (15 at code distance $d = 5$; 7 at $d = 7$). We frame the proposal as a falsifiable conjecture, specify a Monte Carlo thermal-microenvironment protocol that we do not execute, and position the argument against literature on feedback-cooling thermodynamic costs, thermal runaway instabilities, and self-regulating heat-cool balance. The framework reframes scalability from algorithmic overhead to heat-flow physics.

1. Introduction

Cryogenic quantum-computing engineering teams today face a decision that has no quantitative answer in the standard scalability accounting: as a superconducting-qubit array grows past hundreds of devices, how much cooling capacity must be reserved per qubit site, and does that reservation grow fast enough to keep up with the control wiring that error correction demands? The practical cost of not knowing is concrete. We assume, as a working figure, that a dilution refrigerator stage provides cooling power at the 10–20 μW scale at 10 mK; if the per-qubit dissipation of an actively corrected array grows with qubit count, the array's thermal budget is exhausted at some finite size no matter how the code is designed, and every additional physical qubit added for protection makes the microenvironment of its neighbors worse. Today this question is absorbed into ad hoc wiring budgets and cryostat margins; a derived threshold would instead give architects an explicit stopping rule and tell them where to spend engineering effort — on interconnect thermal conductance rather than on code distance.

The mechanism we study is easy to state. QEC resource overhead grows at least quadratically with qubit count, because each logical qubit requires a code block of physical qubits and each syndrome-extraction round requires routing control signals to an expanding web of ancillas and interconnects. Local thermal dissipation from that control wiring, under the exponential-wiring hypothesis we test alongside the quadratic model, grows faster than any polynomial. Meanwhile, the heat leaving each qubit site is limited by the Kapitza interfacial conductance G_K of the solid–solid boundaries between the qubit die, its packaging, and its interconnects. Kapitza conductance is a material property with finite magnitude; it does not scale with qubit count. At a critical qubit count N_c , the heat generated per qubit site exceeds the heat the local thermal path can carry away within the maximum tolerable temperature excursion ΔT_{max} , and decoherence becomes unavoidable regardless of how the code distance is chosen.

We call N_c the thermodynamic feasibility limit. Its existence, if real, reframes scalability: the question is no longer only "how many physical qubits per logical qubit" but "does the thermal microenvironment of qubit $N_c + 1$ remain within specification." It also motivates a class of alternatives: substrates whose intrinsic topological energy gap Δ_{top} suppresses thermal excitations as $e^{-\Delta_{\text{top}}/k_B T}$ without active feedback, so that their cooling requirement does not grow with control activity. Fractional Chern insulators are the condensed-matter candidate for such passive topological substrates. We treat this class as a hypothesis, not a demonstrated technology.

This paper makes four contributions:

1. An explicit heat-balance model combining QEC overhead scaling with Kapitza-limited heat extraction, with every arithmetic step shown to the critical counts $N_c = 390$ (quadratic model) and $N_c \approx 738$ (exponential model) — derived in Sections 3 and 4.
2. Sensitivity results showing the threshold moves from 190 to 3,990 across plausible single-parameter changes, and logical-qubit budgets of 15 at $d = 5$ and 7 at $d = 7$ — derived in Section 4.
3. A local-stability condition, transposed from the catastrophic-cooling-instability analysis of optically thin plasmas [8], giving a maximum tolerable dissipation–temperature feedback coefficient of $4 \times 10^{-7} \text{ W K}^{-1}$ per qubit — derived in Sections 3 and 4.
4. A Monte Carlo thermal-microenvironment protocol, specified but not executed, that would map micro-Kelvin temperature spikes and locate the breaking point empirically — specified in Section 3.4.

We stress the epistemic status of the central claim. It is a conjecture derived from a stylized model with assumed parameters. Its value lies in being falsifiable: each parameter is in principle measurable, and the threshold prediction moves measurably with each one.

2. Background and Related Work

The relevant literature spans stochastic thermodynamics, condensed-matter cooling, plasma astrophysics, and robust statistics. We cite each work for the specific result its text establishes and use it accordingly.

[7] is the closest methodological relative. The work establishes a thermodynamic uncertainty relation (TUR) for feedback cooling in classical underdamped Langevin systems, deriving a trade-off between cooling efficiency and entropy reduction rate (its Eq. 29), and shows as a byproduct that non-negativity of one component of an orthogonal decomposition of the partial entropy production rate represents an information-thermodynamic cooling limit (its Eq. 19). This is precisely the tension we generalize: active QEC is measurement-based control applied continuously per qubit block, and [7] supplies the warrant that such control carries an irreducible thermodynamic cost that grows as one demands better effective cooling. Our model makes that cost concrete as wiring dissipation and asks when it overwhelms the cooling channel. We note in Section 6 that transplanting a classical Langevin bound to quantum error correction is a hypothesis, not a theorem.

[2] proposes using a single Cooper pair box (Josephson qubit) to actively cool nanomechanical resonators, noting that Josephson-qubit experiments had demonstrated coherence times of order microseconds at special symmetry points. Its conclusions quantify the achievable performance: for a 100 MHz oscillator coupled to a practically realizable Josephson qubit, at external temperatures below 1 K, the thermal occupancy of the oscillator mode can be reduced by three orders of magnitude with externally driven AC cooling, while a self-driven scheme using AC Josephson oscillations is found to be less effective. The work is relevant in two directions: it shows the same physical platform being used for active thermal management, and it documents that the coherence resource is finite and of order microseconds — the resource our heat load erodes.

[8] analyzes catastrophic cooling instability in optically thin plasmas, using the solar corona as the prototype of a low-density environment heated externally and cooled radiatively, with the MHD equations modified to include energy source terms. Its linear theory yields a wavenumber-independent dispersion relation $\sigma = -(1/c_V)(\partial\mathcal{L}/\partial T)_\rho$ for the catastrophic-cooling (CC) mode, and shows that because the CC mode cannot be stabilized by thermal conduction, it can lead to condensation formation even where the field length exceeds perturbation wavelengths. The structural analogy is direct: a system with a heat source, a finite-capacity cooling channel, and a runaway mode when heating outruns cooling. Our Kapitza-limited qubit array is a solid-state cousin of this instability structure, and [8] supplies the precedent that such heating-cooling imbalance produces catastrophic rather than graceful degradation. We accordingly add an explicit stability condition to our model in Section 3.3.

[6] puts forward an alternative view of the feedback between heating and cooling in cooling-flow galaxy clusters, adopting the view that heating comes from an active galactic nucleus accreting mass and launching jets or winds, and proposing that the feedback occurs with the entire cool inner region ($r \lesssim 5\text{--}30$ kpc) rather than a localized Bondi radius. Its summary section lists four conditions the participating inner region must satisfy and gives a typical interval of $\approx 10^8$ yr between heating events at a specific region. The work supports the principle that heating-cooling balance is a global, self-regulating quantity, not a purely local one — a principle we import when arguing that thermal microenvironments in qubit arrays couple through shared heat sinks.

[1] demonstrates a liquid crystal as a liquid barocaloric material with cooling potential, noting that its phase transition requires only a small stress (tens of MPa) and can be operated near room temperature, as an alternative to vapor compression based on gas-liquid transitions of hazardous gases. We cite it as evidence that the cooling-technology community actively seeks non-vapor-compression heat management; its text gives no cryogenic-relevant data, so we draw no quantitative claim from it for millikelvin operation.

[3] develops asymptotic breakdown point analysis for a general class of minimum divergence estimators, building on the density power divergence (DPD) family and the larger S-divergence class that subsumes it. Its conclusion reports that for many illustrated examples the asymptotic breakdown point is free of the dimension of the data for various contaminating distributions, justifying minimum S-divergence estimators as viable robust alternatives for high-dimensional parametric inference. This becomes directly useful in the Monte Carlo protocol of Section 3.4: robust estimation of heavy-tailed temperature-spike statistics requires estimators whose breakdown behavior is characterized, since single hot-spot events are heavy-tailed outliers.

[4] and [5] come from active-inference cognitive modeling, and we use them only as conceptual parallels. [4] frames cognitive effort as the cost of resolving the conflict between habits and optimal policies within an evidence-accumulator drift-diffusion account, and reports Bayesian model comparison in which the full active-inference drift-diffusion model was selected under AIC with protected exceedance probability 0.97. [5] models second-person social cognition as coupled active inference in which each agent maintains a self-model that tracks violations of its predictive coherence, formalizing affect as a valuation of identity-relevant surprise. Both support, at the level of analogy, the thesis that active control has a cost that scales with the intensity of control and that systems which must continuously correct themselves face a resource burden that passively stable systems avoid; neither text provides quantitative content transferable to qubit thermodynamics, and we claim none.

No cited work directly treats Kapitza conductance in qubit packages or measures per-qubit QEC dissipation; the quantitative model below is therefore original to this paper, and its parameters are stated assumptions rather than values taken from any cited work.

3. Methods

3.1 Heat generation model

Let N be the number of physical qubits in the array. We model total heat generated inside the cold stage per unit time as

$$P_{\text{tot}}(N) = P_0 N + \alpha N^2,$$

where P_0 is the per-qubit static and linear-control dissipation and αN^2 is the quadratic routing term: each additional qubit adds control lines that must cross the wiring web of all previously added qubits, so routing dissipation grows as the square of qubit count. The quadratic routing term is a modeling hypothesis, not a derived law; we carry it because surface-code physical-qubit counts per logical qubit scale as d^2 in the standard counting, so the physical footprint of protection grows quadratically in the natural counting. The exponential alternative,

$$P_{\text{tot}}^{(\text{exp})}(N) = P_0 N e^{N/N_e},$$

with characteristic wiring-density scale N_e , is analyzed in parallel because the source conjecture posits exponential growth of local thermal dissipation from control wiring.

3.2 Kapitza-limited extraction model

Each qubit site is coupled to the mixing-chamber heat sink through β interconnects, each of Kapitza conductance G_K (units W K^{-1}). Kapitza conductance is the thermal boundary conductance at the interface between two solids; it is finite, material-dependent, and does not increase with system size. The total thermal conductance of the array is

$$G_{\text{tot}}(N) = \beta G_K N.$$

The maximum heat sustainable while keeping every qubit site within a temperature excursion ΔT_{max} of the bath temperature T_b is

$$P_{\text{ext}}^{\text{max}}(N) = G_{\text{tot}}(N) \Delta T_{\text{max}} = \beta G_K \Delta T_{\text{max}} N.$$

3.3 Feasibility condition and stability

A steady state at temperature $T_b + \Delta T_{\text{max}}$ exists if and only if

$$P_{\text{tot}}(N) \leq P_{\text{ext}}^{\text{max}}(N).$$

Dividing both sides by $N > 0$, the quadratic model gives

$$P_0 + \alpha N \leq \beta G_K \Delta T_{\text{max}},$$

so the critical count is

$$N_c = \frac{\beta G_K \Delta T_{\text{max}} - P_0}{\alpha}.$$

For the exponential model, dividing by N gives

$$P_0 e^{N/N_e} \leq \beta G_K \Delta T_{\text{max}} \implies N_c^{(\text{exp})} = N_e \ln \left(\frac{\beta G_K \Delta T_{\text{max}}}{P_0} \right).$$

Both thresholds are independent of code distance d : increasing d increases the number of physical qubits per logical qubit (as d^2 in the standard surface-code counting), which moves the array toward N_c faster while buying logical protection. This is the formal content of the self-defeating regime: past N_c , more QEC resources strictly worsen the thermal environment.

Following the catastrophic-cooling-instability analysis of [8], we additionally require local stability. Writing the lumped energy balance as $C \dot{T} = P_{\text{gen}}(T) - G_{\text{tot}}(T - T_b)$, stability requires

$$\left. \frac{\partial P_{\text{gen}}}{\partial T} \right|_{T_b} < G_{\text{tot}},$$

i.e., any positive temperature feedback of dissipation on itself (e.g., hotter quasiparticles raising switching error rates and hence correction activity) must be weaker than the removal channel, or the system runs away catastrophically rather than saturating. This mirrors the plasma result that the catastrophic-cooling mode grows at rate $\sigma = -(1/c_V)(\partial \mathcal{L}/\partial T)_\rho$ and cannot be stabilized by thermal conduction [8]: in our lumped system, conduction to the bath is the only removal channel, so the sign of the feedback derivative decides between saturation and runaway.

3.4 Monte Carlo microenvironment protocol (proposed, not executed)

To locate the breaking point beyond the mean-field threshold, we specify: simulate arrays of $N \geq 1,000$ qubits; assign each qubit a heat-injection rate drawn from the per-qubit load with spatial clustering to mimic routing congestion; solve the discrete heat equation with per-interconnect Kapitza conductance as the extraction boundary; record the distribution of local temperature spikes δT_i above T_b . The predicted signature is a heavy tail in δT_i emerging as N approaches N_c . Robust estimation of the spike statistics should use minimum-divergence estimators whose asymptotic breakdown behavior is characterized in [3], since single hot-spot events are heavy-tailed outliers and the breakdown-point results of [3] are dimension-free for many contaminating distributions. No results from this protocol are reported in this paper.

4. Analysis

Every input number below is a stated model assumption, not a measurement; the derivations are exact given the assumptions.

Inputs (assumed):

- $P_0 = 10^{-10}$ W per qubit (linear dissipation).
- $\alpha = 10^{-11}$ W per qubit (coefficient of the quadratic routing term).
- $G_K = 10^{-7}$ W K⁻¹ per interconnect.
- $\beta = 4$ interconnects per qubit.
- $\Delta T_{\max} = 10^{-2}$ K (maximum tolerable excursion above the bath).
- $N_e = 200$ (exponential wiring-density scale, exponential model only).

Step 1: extractable heat per qubit.

$$\beta G_K \Delta T_{\max} = 4 \times 10^{-7} \text{ W K}^{-1} \times 10^{-2} \text{ K} = 4 \times 10^{-9} \text{ W}.$$

So each qubit site can shed at most 4×10^{-9} W while staying within $\Delta T_{\max}$.

Step 2: quadratic-model threshold.

$$N_c = \frac{4 \times 10^{-9} - 10^{-10}}{10^{-11}} = \frac{3.9 \times 10^{-9}}{10^{-11}} = 390.$$

Check by substitution at $N = 390$: generated heat is $P_{\text{tot}} = 10^{-10} \times 390 + 10^{-11} \times 390^2 = 3.9 \times 10^{-8} + 10^{-11} \times 152,100 = 3.9 \times 10^{-8} + 1.521 \times 10^{-6} = 1.560 \times 10^{-6}$ W. Extractable heat is $G_{\text{tot}} \Delta T_{\max} = 4 \times 10^{-7} \times 390 \times 10^{-2} = 1.560 \times 10^{-6}$ W. The two sides match exactly, confirming $N_c = 390$.

Step 3: exponential-model threshold.

$$N_c^{(\text{exp})} = N_e \ln\left(\frac{4 \times 10^{-9}}{10^{-10}}\right) = 200 \ln(40).$$

Since $\ln(40) = \ln(4) + \ln(10) \approx 1.386294 + 2.302585 = 3.688879$,

$$N_c^{(\text{exp})} = 200 \times 3.688879 = 737.7758 \approx 738.$$

Check by substitution at $N = 737.7758$: $e^{N/N_e} = e^{3.688879} = 40$ exactly, so per-qubit generation is $10^{-10} \times 40 = 4 \times 10^{-9}$ W, equal to the per-qubit extraction limit from Step 1. Confirmed.

Step 4: sensitivity of the quadratic threshold to $\Delta T_{\max}$. Since N_c is linear in $\Delta T_{\max}$, halving the allowed excursion to 5×10^{-3} K halves the extraction budget to 2×10^{-9} W and gives

$$N_c = \frac{2 \times 10^{-9} - 10^{-10}}{10^{-11}} = \frac{1.9 \times 10^{-9}}{10^{-11}} = 190.$$

Step 5: sensitivity to Kapitza conductance. A tenfold improvement, $G_K = 10^{-6}$ W K⁻¹, gives extraction 4×10^{-8} W per qubit and

$$N_c = \frac{4 \times 10^{-8} - 10^{-10}}{10^{-11}} = \frac{3.99 \times 10^{-8}}{10^{-11}} = 3,990.$$

Step 6: code-distance trade. In the standard surface-code counting, physical qubits per logical qubit scale as d^2 . At $d = 5$, one logical qubit costs 25 physical qubits; the quadratic model therefore supports at most $\lfloor 390/25 \rfloor = 15$ logical qubits at distance 5 before the thermal budget is exhausted — under the stated assumptions. At $d = 7$ the cost is 49 physical qubits per logical qubit, giving $\lfloor 390/49 \rfloor = 7$ logical qubits. Increasing d from 5 to 7 buys a lower logical error rate but consumes the thermal budget more than twice as fast per logical qubit, which is the quantitative form of the self-defeating regime.

Step 7: crossover between models. The quadratic model fails first ($390 < 738$) under these parameters; the exponential model's larger threshold arises because its per-qubit load $P_0 e^{N/N_e}$ equals $P_0 = 10^{-10}$ W at $N = 0$ and only reaches the extraction limit at $N \approx 738$, whereas the quadratic model's per-qubit load $P_0 + \alpha N$ grows linearly and reaches the limit at $N = 390$. The two models bracket the threshold between 390 and 738 qubits under the stated assumptions.

Step 8: stability margin. With the reference numbers, the removal channel coefficient is $G_{\text{tot}}/N = \beta G_K = 4 \times 10^{-7}$ W K⁻¹ per qubit. The stability condition of Section 3.3 requires the temperature feedback coefficient of dissipation to stay below this value; if each millikelvin of heating raises per-qubit dissipation by more than 4×10^{-10} W, the array runs away rather than saturating. This is the catastrophic-instability analog of [8] transposed to the cryostat.

Step 9: cross-check against dilution-refrigerator capacity. We assume, as a working engineering figure, that standard dilution refrigerators provide $10\text{--}20\ \mu\text{W}$ of cooling capacity at $10\ \text{mK}$. Under our assumed coefficients, the total array heat at $N = 390$ is $1.560 \times 10^{-6}\ \text{W} = 1.56\ \mu\text{W}$ (Step 2), which is below $10\ \mu\text{W}$. At $N = 10^5$, the same coefficients give

$$P_{\text{tot}} = 10^{-10} \times 10^5 + 10^{-11} \times (10^5)^2 = 10^{-5} + 10^{-11} \times 10^{10} = 10^{-5} + 10^{-1} = 0.10001\ \text{W},$$

dominated by the quadratic term and exceeding $20\ \mu\text{W} = 2 \times 10^{-5}\ \text{W}$ by a factor of

$$\frac{0.10001}{2 \times 10^{-5}} = 5,000.5 \approx 5.0 \times 10^3.$$

This consistency check uses our assumed coefficients, not measured values; it shows only that our model, so parameterized, reproduces the qualitative expectation that 10^5 actively corrected qubits overwhelm a standard refrigerator, and it locates the quantitative uncertainty in the routing coefficient α , which no cited work measures.

5. Results

All numbers below are computed in Section 4 from the stated assumptions; they are model projections conditional on those assumptions, not measurements.

1. **Quadratic-overhead threshold:** $N_c = 390$ physical qubits (exact under the assumed parameters; verified by substitution in Step 2).
2. **Exponential-overhead threshold:** $N_c^{(\text{exp})} = 737.7758 \approx 738$ physical qubits (verified by substitution in Step 3).
3. **Sensitivity to allowed temperature excursion:** $N_c = 190$ if ΔT_{max} is halved to $5 \times 10^{-3}\ \text{K}$ (Step 4).
4. **Sensitivity to Kapitza conductance:** $N_c = 3,990$ if G_K improves tenfold to $10^{-6}\ \text{W K}^{-1}$ (Step 5).
5. **Logical-qubit budget at the quadratic threshold:** 15 logical qubits at code distance $d = 5$; 7 logical qubits at $d = 7$ (Step 6).
6. **Threshold bracket:** under the stated assumptions, the thermodynamic feasibility limit lies between 390 and 738 physical qubits (Step 7).
7. **Stability bound:** the maximum tolerable dissipation–temperature feedback coefficient is $\beta G_K = 4 \times 10^{-7}\ \text{W K}^{-1}$ per qubit in the reference geometry (Step 8).
8. **Refrigerator-capacity cross-check:** at $N = 390$ the modeled total array heat is $1.56\ \mu\text{W}$, below the assumed $10\text{--}20\ \mu\text{W}$ stage capacity; at $N = 10^5$ the same coefficients give $0.10001\ \text{W}$, a factor of $\approx 5.0 \times 10^3$ above $20\ \mu\text{W}$ (Step 9).

The Monte Carlo protocol of Section 3.4 is proposed but not executed; no simulation results are claimed. As a labeled projection only: if that protocol were run with the reference parameters, we expect hot-spot-driven threshold crossing within the sensitivity range of items 3–4, i.e. N_c spanning roughly $10^2\text{--}10^4$ across plausible parameter variation; this is a projection from the analytic model with uncertainty at least spanning that range, not a simulation result.

6. Discussion

Limitations. The model is mean-field: it assumes uniform heat injection and uniform Kapitza conductance, whereas real arrays have hot spots near dense routing regions, so the true breaking point may lie below the mean-field N_c . The parameters P_0 , α , G_K , β , ΔT_{max} , and N_e are assumptions chosen to be physically plausible in magnitude, not measured values; every headline number scales linearly (or, for the exponential model, logarithmically) with them, and the sensitivity results (Steps 4–5) show the threshold spans more than an order of magnitude across plausible parameter ranges. The quadratic routing term is a modeling hypothesis, not a derived law; if routing dissipation in fact grows only linearly, the quadratic model reduces to a per-qubit budget comparison and no finite threshold exists at all — the feasibility condition becomes $P_0 \leq \beta G_K \Delta T_{\text{max}}$, which is satisfied by a factor of 40 under the assumed numbers ($4 \times 10^{-9}/10^{-10} = 40$). This is the single most important internal vulnerability of the conjecture. The literature base is also a limitation: no cited work measures Kapitza conductance in qubit packages, per-qubit dissipation budgets, or overhead-scaling exponents, so no cited work constrains any input number.

Failure modes of the claim. The conjecture would be falsified if: (i) measured per-qubit dissipation in operating QEC experiments is far below $10^{-10}\ \text{W}$ at millikelvin temperatures, pushing N_c beyond any practically interesting array size; (ii) Kapitza conductance at qubit-relevant interfaces proves much larger than $10^{-7}\ \text{W K}^{-1}$, or engineering solutions (thermal straps, superfluid heat channels) effectively multiply β ; (iii) the Monte Carlo protocol, once run, shows temperature spikes remaining bounded as N grows past the predicted threshold. Conversely, observation of unbounded micro-Kelvin spike growth in arrays below 1,000 qubits would support it. The single calculation that would most directly falsify the headline result is a measurement of routing-region dissipation in a $10^2\text{--}10^3$ -qubit array showing that the per-qubit dissipation does not grow with N : if $\alpha < 10^{-11}\ \text{W}$ per added qubit by more than a factor of 10, the quadratic threshold moves beyond 3,900 qubits and the practical force of the claim disappears.

Against ourselves. A skeptic should note three things. First, the analogy chain ([6], [8]) is suggestive but not quantitative; astrophysical cooling flows and coronal plasmas share the heating–cooling competition structure but nothing else at the relevant scales. Second, the TUR for feedback cooling [7] bounds the cost of *cooling* by measurement; transplanting it to bound the cost of *error correction* is a hypothesis, not a theorem, and the classical underdamped Langevin setting of [7] may not admit a direct quantum analogue. Third, the passive-topological escape route (fractional Chern insulator substrates with intrinsic gaps) is asserted as a hypothesis class, not established here: passive protection trades active heat for

material disorder and quasiparticle poisoning, and readout of any computational device still requires measurement wiring, so some linear heat load remains; the claim that passive substrates scale where active ones diverge is a hypothesis about coefficient ratios, not a theorem. The strongest historical counterargument is that cryogenic engineering has repeatedly improved by large factors, so a threshold computed from today's plausible parameters may be irrelevant to tomorrow's stack. Our response is that the threshold's existence is parameter-independent within this model class — only its location moves — so the conjecture is a statement about scaling structure, not about a fixed qubit count. But if the routing term is sub-quadratic, the structure itself disappears, and the conjecture fails at its foundation.

Open questions. (1) What is the measured areal Kapitza conductance of qubit-package interfaces at millikelvin temperatures? (2) Is the exponential wiring model ever realized in practice, or does routing become hierarchical and effectively linear? (3) How do the thresholds move if the bath temperature T_b itself is lowered — Kapitza conductance typically falls with temperature (a T^3 -type scaling is the standard expectation, which we state as a modeling assumption, not a literature result), so the extraction budget may shrink even as coherence improves, a coupling the present model does not capture. (4) Can feedback-cooling thermodynamics of the kind formalized in [7] be transplanted to bound QEC feedback cost?

7. Practical Implications and Engineering Pathways

(a) Who would use the result and what changes. The direct users are cryogenic systems architects and wiring designers for superconducting-qubit platforms, and the program managers who allocate dilution-refrigerator capacity in multi-chip scaling plans. If the result holds, the design decision that changes is the scaling roadmap itself: instead of budgeting code distance first and cooling second, the architect fixes a per-qubit thermal budget $\beta G_K \Delta T_{\max}$ (computed here as 4×10^{-9} W per qubit under the reference assumptions) and treats it as a hard ceiling on per-qubit control dissipation, with an explicit stopping rule at N_c . The result also redirects engineering effort: under the model, a tenfold improvement in interconnect Kapitza conductance buys a tenfold larger threshold (N_c from 390 to 3,990, Step 5), whereas increasing code distance buys logical protection at the cost of consuming the same thermal budget faster (Step 6). Because this is an analytical conjecture with assumed parameters, no hardware decision should yet be based on it; its present use is as a checklist of measurements that must be made before scaling commitments are made.

(b) Current readiness. Against the NASA Technology Readiness Level definitions, this work is TRL 1–2: it is purely analytical, has not been peer reviewed, and offers no validated concept demonstration. Analytical work of this kind cannot exceed TRL 3, and no manufacturing or fabrication readiness is claimed. The Monte Carlo protocol of Section 3.4, if executed and validated against a small experimental array, would be the first step toward TRL 3.

(c) The smallest next test. The smallest de-risking calculation is a measurement, not a simulation: instrument the wiring loom of an existing 10^2 -qubit-scale superconducting processor at the mixing-chamber stage with a calibrated thermal sensor, and measure the temperature rise attributable to control-line dissipation as the number of actively driven qubits is varied. The model predicts a per-qubit dissipation contribution growing with N ; if the measured rise is flat in N (i.e., α consistent with zero at the 10^{-11} W scale), the quadratic threshold is refuted at its foundation. The smallest simulation step is a 1,000-qubit instance of the Section 3.4 protocol with clustered heat injection; a bounded spike distribution at injection levels corresponding to $N > N_c$ would refute the hot-spot amplification claim.

(d) Binding constraints. The constraints that bind first, in the order the model identifies them: (i) interconnect materials and interfaces — the extraction budget is set by G_K at solid–solid boundaries, which no cited work measures for qubit packages, making this the first open question; (ii) wiring density and routing topology — the entire quadratic term rests on the assumption that control lines cross the wiring web of previously placed qubits, and a hierarchical routing scheme that decouples new qubits from old wiring would remove the term; (iii) cryogenic budget — the cross-check of Step 9 shows the model's 10^5 -qubit heat load exceeding an assumed 10–20 μ W stage capacity by a factor of $\approx 5.0 \times 10^3$, but this uses assumed coefficients, so the true binding point is unknown until α is measured. Regulation and cost drivers do not bind at this stage of the analysis.

(e) Order-of-magnitude bounds (projections). Labeled as projections with stated assumptions: under the reference parameters, a scaling plan that reaches 10^5 physical qubits would require either a cooling capacity near 0.1 W at the mixing-chamber stage (Step 9) or a reduction of the routing coefficient α by a factor of $\approx 5.0 \times 10^3$ relative to the assumed 10^{-11} W, to fit within the assumed 20 μ W capacity. Both figures follow directly from the arithmetic of Step 9 and inherit its assumption status; they are not measurements.

(f) Risks and misuse paths. The principal misuse is treating $N_c = 390$ (or 738) as a measured hardware limit rather than a model projection from assumed parameters; quoting the headline numbers without the assumption status could be used to argue for or against scaling programs without any supporting measurement. A secondary risk is overgeneralizing the passive-substrate conclusion: the claim that passive topological substrates escape the threshold is a hypothesis about coefficient ratios, and asserting it as established would misdirect research funding away from near-term wiring and cooling engineering that the model itself identifies as the highest-leverage variables.

8. Conclusion

We have formulated a thermodynamic feasibility limit for actively corrected gate-based quantum computing: a critical qubit count N_c at which Kapitza-limited heat extraction can no longer balance QEC-driven dissipation, beyond which decoherence is thermally enforced regardless of code distance. Under explicitly stated assumptions we computed $N_c = 390$ for a quadratic-overhead model and $N_c \approx 738$ for an exponential-

overhead model, with all arithmetic shown, together with sensitivities (190; 3,990), logical-qubit budgets (15 at $d = 5$; 7 at $d = 7$), and a stability bound ($4 \times 10^{-7} \text{ W K}^{-1}$ per qubit). The conjecture is falsifiable through measurement of its input parameters and through the proposed Monte Carlo microenvironment protocol. If it survives, the scalability debate shifts from algorithmic overhead to heat-flow physics, and passive topological substrates become not an alternative implementation but the only thermodynamically consistent route past the threshold.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.18.3-codeagent). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, zai-org/glm-5.3, openai/gpt-oss-120b, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 8 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] Demonstration of liquid crystal for barocaloric cooling application. arXiv:1609.07939v1. <https://arxiv.org/abs/1609.07939v1> [2] Ground state cooling of mechanical resonators. arXiv:cond-mat/0310229v2. <https://arxiv.org/abs/cond-mat/0310229v2> [3] Asymptotic Breakdown Point Analysis for a General Class of Minimum Divergence Estimators. arXiv:2304.07466v3. <https://arxiv.org/abs/2304.07466v3> [4] Cognitive Effort in the Two-Step Task: An Active Inference Drift-Diffusion Model Approach. arXiv:2508.04435v2. <https://arxiv.org/abs/2508.04435v2> [5] Geometric Hyperscanning of Affect under Active Inference. arXiv:2506.08599v4. <https://arxiv.org/abs/2506.08599v4> [6] Cold Feedback in Cooling-Flow Galaxy Clusters. arXiv:astro-ph/0611214v1. <https://arxiv.org/abs/astro-ph/0611214v1> [7] Thermodynamic uncertainty relation for feedback cooling. arXiv:2508.06174v2. <https://arxiv.org/abs/2508.06174v2> [8] Catastrophic cooling instability in optically thin plasmas. arXiv:2408.15869v2. <https://arxiv.org/abs/2408.15869v2>