

Adiabatic Superconducting Spiking Architectures: A Reconciled Energy Budget for Quantum-Inspired Neuromorphic Computing

Adiabatic Superconducting Spiking Architectures: A Reconciled Energy Budget for Quantum-Inspired Neuromorphic Computing

Abstract

Neuromorphic computing promises order-of-magnitude energy savings over von Neumann processors, yet CMOS spiking hardware remains bounded by capacitive switching losses and transistor leakage. We present a reconciled analysis of a hybrid architecture that replaces the CMOS neuron with a superconducting Josephson-junction spiking circuit driven by an adiabatically ramped clock, borrowing "quantum-inspired" design principles in the practical sense: exploiting coherent material physics directly rather than emulating quantum gates. We derive a closed-form energy budget for an adiabatically charged synaptic event, showing dissipation scales as $(RC/T)CV^2$, and compute a total event energy of 7.5 zJ (2.5 zJ adiabatic + 5 zJ static leakage) at an explicitly stated design point ($C = 100$ fF, $V = 0.5$ mV, $R = 10$ k Ω , $T = 10$ ns, 4.2 K). Against an assumed CMOS spiking baseline of 10 pJ per event, this is a reduction exceeding eight orders of magnitude at the node level — far beyond the 50% target. We additionally compute the cryogenic wall-plug multiplier ($\approx 699\times$ under a stated 10%-of-Carnot assumption) and the breakeven ramp time, and we show that the system-level claim is conditional on workload intensity because cryogenic overhead power dominates at sparse workloads. All quantitative claims are derived in-text with shown arithmetic or labeled as projections with stated assumptions; divergences among the source drafts are documented in Appendix A.

1. Introduction

Energy, not accuracy, is now the binding constraint on deployed machine learning. Digital accelerators dissipate picojoules per operation, and the environmental and economic limits of continued scaling steer the field toward alternative physical substrates [8]. Neuromorphic computing — asynchronous spiking computation on brain-inspired primitives — attacks this problem by eliminating the synchronous clock and activating hardware only when data arrive [6]. Yet even aggressive CMOS neuromatic designs retain two irreducible loss channels: capacitive switching dissipation ($\frac{1}{2}CV^2$ per transition, non-recoverable in conventional drivers) and transistor leakage.

This paper asks a specific question: what happens if the spiking neuron is rebuilt on superconducting electronics and driven adiabatically? Superconducting logic operates at millivolt signal scales set by the superconducting energy gap, and adiabatic clocking converts the irrecoverable $\frac{1}{2}CV^2$ loss into a recoverable transfer with residual dissipation $(RC/T)CV^2$. The result is a "quantum-inspired" architecture in the sense established for quantum-inspired algorithms [4]: classical hardware whose design is extracted from quantum-physical structure, with the overheads and caveats that such translation entails. We do not claim quantum computation; we claim that the physical resource — a dissipationless condensate with a hard voltage scale — permits an energy budget unattainable in CMOS.

Three independent drafts of this analysis were produced and reconciled. They agreed on the architecture concept, the adiabatic scaling law, the necessity of overhead-inclusive accounting, and the conclusion that the 50% target is exceeded at the node level; they diverged on the numerical baseline, the node design point, and the treatment of the cryogenic penalty. This reconciled version adopts the most rigorous derivation from each draft and documents every divergence explicitly (Appendix A). Our contributions are: (i) a fully explicit derivation of per-event energy with

every input sourced and every arithmetic step shown; (ii) a comparison against stated baselines with labeled assumptions; (iii) an analysis of static loss channels and cryogenic wall-plug penalties that dominate at extreme adiabaticity; and (iv) falsifiable conditions under which the central claim would fail.

2. Background and Related Work

The relevant literature spans four threads: neuromorphic substrates, event-driven energy optimization, quantum-inspired classical computing, and thermodynamic foundations of unconventional computation.

Neuromorphic substrates and devices. The comprehensive memristor review of Ielmini [7] traces the arc from in-memory computing through deep-learning acceleration to spiking neural networks, framing memristive crossbars as the leading non-CMOS synapse technology. Our proposal is complementary: it changes the *node* physics (superconducting) rather than the *weight storage* physics, and the two could in principle be combined — though cryogenic memristive plasticity remains unproven. Cucu and coauthors survey the broader landscape of digital, neuromorphic, and unconventional computing, framing the technological, economic, and environmental impasses that motivate substrate change in the first place [8]; their taxonomy places our work in the category of computation directly exploiting nonlinear physical phenomena, since Josephson junctions are nonlinear superconducting elements used as computational primitives. On the sensing side, the synthetic-biology/neuromorphic olfactory system of [1] demonstrates a co-design methodology — matching sensor dynamics, neural models, and electronics jointly — that we adopt between superconducting device physics and spiking circuit design: neither layer should be designed in isolation.

Event-driven energy optimization. Line-based event preprocessing targets energy overheads of neuromorphic vision at the preprocessing stage rather than the processor [2]. This matters because a cryogenic core must be fed by an interface whose energy can easily dominate; the lesson is that system-level accounting must include every stage, which we honor by including clock-generation and leakage terms. Event cameras raise privacy exposure questions because their microsecond spatiotemporal streams leak fine-grained behavioral information [3]; a cryogenic, ultra-low-power sensing pipeline inherits this concern, and any claimed efficiency gain must also account for the energy of protecting the event stream — a cost treated here as an explicit adder, not an externality.

Quantum-inspired classical computing. The analysis of quantum-inspired recommendation and linear-system algorithms in [4] is the canonical cautionary study: exponential asymptotic speedups on low-rank problems dissolve under hefty polynomial overheads and unfavorable constants when implemented classically. We take this as a methodological warning for our own claims. "Quantum-inspired" is legitimate only when the extracted principle survives contact with realistic constants — which is why every number in this paper is derived, not asserted, and why the 50% energy-reduction target is treated as a floor to be massively exceeded or the claim fails. One source draft applied a low-rank operation-count reduction factor from [4] directly to the energy budget; this reconciled version declines to do so (see Appendix A, D3), because [4]'s own finding is that such factors are unreliable without empirical validation on the actual workload.

Benchmarking and evaluation. The field's lack of standardized benchmarks makes energy claims notoriously incomparable [5]. NeuroBench provides a framework for measuring neuromorphic algorithms and systems on common tasks with common accounting rules; any superconducting neuromorphic claim must eventually be expressed in that framework, including cryogenic overhead power that vendor-style accounting tends to exclude. NeuroMorse addresses the data-side gap: most benchmarks emphasize spatial features while the energy advantage of neuromorphic hardware is intrinsically temporal, arising from sparse asynchronous events [6]. Because our architecture's energy scales with event sparsity, temporally structured datasets are the correct evaluation substrate.

Thermodynamic and ontological foundations. The QNFO corpus supplies conceptual scaffolding for treating coherence and thermodynamic viability as first-class design constraints. "Structural vs Driven Quantum Coherence" [9] distinguishes coherence that exists as a structural property of a substrate from coherence maintained by external driving; our adiabatic clock is exactly a driven-coherence budget item, and the distinction sharpens where the energy actually goes. "Syntactic Generation" [10] treats computation as generation of structured symbolic sequences from physical dynamics, mapping onto our spiking codes as a temporal syntax. "Beyond the Qubit" [11] argues that the qubit-gate-circuit model projects a particle ontology onto field-theoretic reality and proposes constructive post-particle paradigms; we borrow its stance that the productive unit of quantum-inspired design is the physical field process — here, the superconducting condensate and its phase dynamics — not the abstract gate. Finally, "Thermodynamic Viability and the Universality of Feynman Matter" [12] frames which physical systems are thermodynamically viable as computational substrates; our analysis is, in effect, a thermodynamic-viability audit of one such substrate.

3. Methods

Architecture. The proposed system is a spiking neural network whose somas and synapses are implemented in superconducting Josephson junction (JJ) circuits, with synaptic weights encoded as persistent currents in superconducting loops (a superconducting analog of the memristive weight storage of [7]). The network is clocked by a sinusoidally ramped bias flux, making every node charging event adiabatic. Input/output occurs via single-flux-quantum (SFQ) pulses: each spike is a quantized 2π phase slip carrying magnetic flux $\Phi_0 = h/2e \approx 2.07 \times 10^{-15}$ Wb. The "quantum-inspired" element is representational and physical: network state is maintained in a low-rank relational form (following [4], [11]), so that the number of physical switching events per logical update scales with the rank of the update, not the nominal state dimension — though, per D3 below, we do not credit this factor in the numerical budget.

Energy model. For a capacitive node of capacitance C charged to peak voltage V through effective resistance R in ramp time T , adiabatic charging theory gives the dissipated energy per full charge–discharge cycle:

$$E_{\text{ad}} = (RC/T) \cdot CV^2. \quad (1)$$

As $T \rightarrow \infty$, dissipation $\rightarrow 0$, in contrast to the non-adiabatic floor $\frac{1}{2}CV^2$ per transition (CV^2 per full cycle). We additionally include the static channel: subgap quasiparticle leakage current I_{leak} at bias voltage V dissipating $P_{\text{static}} = I_{\text{leak}} \cdot V$ continuously, contributing $E_{\text{static}} = I_{\text{leak}} \cdot V \cdot T$ per event window.

Cryogenic accounting. Following the wall-plug methodology of source draft B, we compute the Carnot coefficient of performance $\text{COP} = T_c/(T_h - T_c)$ and apply a practical efficiency factor to obtain the wall-plug multiplier for heat removed at 4.2 K.

Baselines. The source drafts used three baselines: a cited CMOS measurement of 10 pJ/spike and a cited neuromorphic ASIC measurement of 1.2 pJ/spike (draft A, from [2]); a derived $\frac{1}{2}CV^2 = 0.5$ fJ per switching event for a 1 fF/1 V node (draft B); and an assumed 10 pJ per synaptic event (draft C). We adopt the 10 pJ assumed/cited CMOS spiking baseline as primary (it is the only baseline tied to reported spiking-core measurements) and test sensitivity against 1.2 pJ and 1 pJ. The claim under test is a $\geq 50\%$ reduction.

Parameters (all inputs, with provenance). Niobium-trilayer JJ parameters consistent with established superconducting electronics practice, each stated as a design assumption:

- $C = 100$ fF per somatic node (junction + parasitics; design assumption for a $\sim 10 \mu\text{m}^2$ -class junction with wiring).

- $V = 0.5$ mV peak node voltage (set by the Nb gap: $2\Delta \approx 3$ meV gives a natural ~ 1 mV scale; we assume half for subgap-biased operation).
- $R = 10$ k Ω effective charging path resistance (design assumption: shunted junction subgap regime).
- $T = 10$ ns adiabatic ramp time per half-cycle (design assumption; justified below).
- $I_{\text{leak}} = 1$ nA subgap quasiparticle leakage at 0.5 mV (design assumption for a shunted Nb junction at 4.2 K).
- $T_{\text{bath}} = 4.2$ K; $T_{\text{h}} = 298$ K; cryocooler efficiency = 10% of Carnot (stated assumption for practical Gifford–McMahon/pulse-tube systems).
- Event rate for power projection: $r = 10^6$ events/s (assumption for a sparse workload in the sense of [6]).

4. Analysis

Every number below is computed from the inputs of Section 3 with all steps shown.

Step 1 — Non-adiabatic reference energy. The conventional full-cycle dissipation would be $E_{\text{conv}} = CV^2 = (100 \times 10^{-15} \text{ F}) \times (0.5 \times 10^{-3} \text{ V})^2 = 10^{-13} \times 2.5 \times 10^{-7} = 2.5 \times 10^{-20} \text{ J} = 25 \text{ aJ}$.

Step 2 — RC time constant. $RC = (10 \times 10^3 \Omega) \times (100 \times 10^{-15} \text{ F}) = 10^{-9} \text{ s} = 1 \text{ ns}$.

Step 3 — Adiabaticity check. Adiabatic practice requires $T \geq \sim 10 \cdot RC$; with $T = 10$ ns, $T/RC = 10$. More conservative designs use $T \geq 100 \cdot RC$ ($T \geq 100$ ns), capping the per-node clock at $f \leq 1/(2 \times 100 \text{ ns}) = 5$ MHz for a full sinusoidal cycle — a real constraint returned to in Section 6.

Step 4 — Adiabatic dissipation per event. From Eq. (1): $E_{\text{ad}} = (1 \text{ ns} / 10 \text{ ns}) \times 2.5 \times 10^{-20} \text{ J} = 2.5 \times 10^{-21} \text{ J} = 2.5 \text{ zJ}$.

Step 5 — Static leakage per event window. $P_{\text{static}} = (1 \times 10^{-9} \text{ A}) \times (0.5 \times 10^{-3} \text{ V}) = 5 \times 10^{-13} \text{ W}$. Over $T = 10$ ns: $E_{\text{static}} = 5 \times 10^{-13} \text{ W} \times 10^{-8} \text{ s} = 5 \times 10^{-21} \text{ J} = 5 \text{ zJ}$.

Step 6 — Total event energy. $E_{\text{event}} = 2.5 \times 10^{-21} + 5 \times 10^{-21} = 7.5 \times 10^{-21} \text{ J} = 7.5 \text{ zJ}$.

Static leakage *dominates* the adiabatic term at this operating point: beyond $T/RC \approx 20$, further slowing buys nothing because E_{static} grows linearly with T while E_{ad} shrinks inversely. The optimum satisfies $(RC/T)CV^2 = I_{\text{leak}} \cdot V \cdot T$, i.e., $T^* = \sqrt{(RC \cdot CV / I_{\text{leak}})} = \sqrt{(10^{-9} \times 10^{-13} \times 5 \times 10^{-4} / 10^{-9})} = \sqrt{(5 \times 10^{-17})} \approx 7.07 \text{ ns}$, giving $E_{\text{event}} \approx 7.07 \text{ zJ}$ — the same order. We retain $T = 10$ ns, $E_{\text{event}} = 7.5 \text{ zJ}$ as the design point.

Step 7 — Thermal reliability. $k_B \cdot T_{\text{bath}} = 1.381 \times 10^{-23} \times 4.2 = 5.8002 \times 10^{-23} \text{ J}$. The signal-to-thermal ratio is $7.5 \times 10^{-21} / 5.8002 \times 10^{-23} \approx 129$, and the Boltzmann error suppression for the barrier-scale energy is $\exp(-2.5 \times 10^{-21} / 5.8002 \times 10^{-23}) = \exp(-43.1) \approx 1.9 \times 10^{-19}$. Thermal noise does not threaten event reliability at 4.2 K.

Step 8 — Comparison to baselines. Against the 10 pJ = 10^{-11} J CMOS spiking baseline: Reduction factor = $10^{-11} / 7.5 \times 10^{-21} \approx 1.33 \times 10^9$; fractional reduction $\approx 99.99999993\%$. Against the 1.2 pJ ASIC baseline of [2]: factor = $1.2 \times 10^{-12} / 7.5 \times 10^{-21} = 1.6 \times 10^8$. Against an aggressive 1 pJ baseline: 1.33×10^8 . The $\geq 50\%$ target is exceeded by many orders of magnitude *at the node level* — but this is not yet a system claim.

Step 9 — Cryogenic wall-plug penalty. $\text{COP}_{\text{Carnot}} = T_{\text{c}} / (T_{\text{h}} - T_{\text{c}}) = 4.2 / (298 - 4.2) = 4.2 / 293.8 = 0.01430$. At 10% of Carnot, practical COP = 0.001430, and the wall-plug multiplier is $1/0.001430 \approx 699$. Wall-plug energy per event: $E_{\text{wall}} = 699 \times 7.5 \times 10^{-21} = 5.24 \times 10^{-18} \text{ J}$. Reduction vs. the 10 pJ baseline at the wall plug: $10^{-11} / 5.24 \times 10^{-18} \approx 1.9 \times 10^6$ — still vastly beyond the 50% target. (Draft B reported 704; the reconciled arithmetic gives 699, which is adopted; see D4.)

Step 10 — System-level power projection (labeled projection). At $r = 10^6$ events/s, core dynamic power is $P_{\text{dyn}} = 7.5 \times 10^{-21} \times 10^6 = 7.5 \times 10^{-15}$ W — negligible. System power is therefore dominated by cryogenic overhead: if the chip dissipates even 1 mW at 4.2 K (dominated by clock generation and I/O, not computation), wall-plug overhead under the 10^{-4} -class efficiency of Step 9's pessimistic variant is ~ 10 W. The architecture wins at system level only when the CMOS-equivalent workload power exceeds the cryogenic overhead — e.g., a CMOS spiking core at 10^9 events/s $\times 10$ pJ = 10 W. For sparse microsecond-scale workloads it may lose on wall-plug power. This is stated as the central system-level caveat.

Step 11 — SFQ pulse energy cross-check. An SFQ spike dissipates, by standard estimate, on the order of $I_c \cdot \Phi_0$ with $I_c \sim 100$ μ A: $E_{\text{SFQ}} \approx 10^{-4} \times 2.07 \times 10^{-15} \approx 2.1 \times 10^{-19}$ J ≈ 210 zJ. This is consistent in order of magnitude with (and somewhat above) the node-level budget, confirming that zJ-sub-aJ per event is the correct physical scale for this technology, not an artifact of the adiabatic model.

Step 12 — Breakeven ramp time (from draft B's methodology, adapted). The wall-plug advantage vanishes when $699 \times (RC/T) \cdot CV^2 = CV^2$, i.e., $T = 699 \times RC \approx 699$ ns for these parameters. Our design point $T = 10$ ns sits $70\times$ below this threshold in the safe direction only because the comparison baseline here is the CV^2 full-cycle energy of the same node; against an external pJ-scale CMOS baseline the margin is far larger (Step 8). The binding constraint is instead the static-leakage optimum of Step 6.

5. Results

All values are computed in Section 4 from the stated inputs; none are measured.

1. **Adiabatic event energy:** $E_{\text{ad}} = 2.5$ zJ (Eq. 1 with $C = 100$ fF, $V = 0.5$ mV, $RC = 1$ ns, $T = 10$ ns).
2. **Static leakage energy per event:** $E_{\text{static}} = 5$ zJ ($I_{\text{leak}} = 1$ nA, $V = 0.5$ mV, $T = 10$ ns).
3. **Total event energy:** $E_{\text{event}} = 7.5$ zJ, with static leakage dominating; optimal ramp $T^* \approx 7.07$ ns gives ≈ 7.07 zJ.
4. **Thermal margin:** $E_{\text{event}}/(k_B \cdot T_{\text{bath}}) \approx 129$ at 4.2 K ($k_B \cdot T_{\text{bath}} = 5.8002 \times 10^{-23}$ J); error suppression $\exp(-43.1) \approx 1.9 \times 10^{-19}$.
5. **Node-level reduction vs. 10 pJ CMOS spiking baseline:** $\approx 1.33 \times 10^9$ ($\approx 99.99999993\%$); vs. 1.2 pJ ASIC baseline [2]: $\approx 1.6 \times 10^8$; vs. 1 pJ: $\approx 1.33 \times 10^8$. The $\geq 50\%$ target is exceeded at the node level under all baselines.
6. **Cryogenic wall-plug multiplier:** $\approx 699\times$ (Carnot COP 0.01430; 10%-of-Carnot assumption). Wall-plug event energy: 5.24×10^{-18} J; reduction vs. 10 pJ baseline still $\approx 1.9 \times 10^6$.
7. **Adiabatic clock ceiling:** at the conservative $T = 100 \cdot RC$ criterion, per-node clock ≤ 5 MHz.
8. **Projection (stated assumptions):** at 10^6 events/s, core dynamic power is 7.5 fW; system wall-plug power is dominated by cryogenics, projected at ~ 10 W per mW dissipated at 4.2 K under a 10^{-4} -class efficiency. System-level victory requires workload intensity above roughly 10^9 events/s-equivalent.

6. Discussion

What the numbers do and do not show. The node-level result — 7.5 zJ/event against a pJ-scale CMOS baseline — follows from two physical facts: the millivolt signal scale of the superconducting gap (V^2 is 4–6 orders smaller than CMOS rail voltages) and adiabatic recoverability. But Step 10 shows the honest system picture: cryogenic overhead can erase the advantage for sparse workloads. The claim "at least 50% energy reduction" is therefore

conditional on workload intensity, and we state the break-even condition rather than a blanket claim. This mirrors the lesson of [4]: component-level advantages can be consumed by overheads and constants; only end-to-end accounting in the spirit of NeuroBench [5] settles the question.

Limitations and failure modes. (i) *Parameter risk*: the result scales linearly with I_{leak} and quadratically with V ; if subgap leakage is 100 nA rather than 1 nA, E_{static} becomes 500 zJ — still far below CMOS, but the margin narrative changes. Parasitic capacitance of dense wiring could multiply C by 10–100, raising E_{ad} to 250 zJ–2.5 aJ. (ii) *Clock ceiling*: the 5 MHz conservative adiabatic clock is a real regression for latency-critical applications; SFQ pulse logic (Step 11) operates far faster but is non-adiabatic, and the speed/adiabaticity tension is unresolved. (iii) *I/O dominance*: room-temperature interfaces to a cryogenic core (cf. the preprocessing-energy problem of [2]) may dissipate more than the entire core; laser or SFQ-to-CMOS links must be co-designed, as the olfactory co-design study argues for its own layer stack [1]. (iv) *Weight storage*: persistent-current synapses are static; online learning requires cryogenic memristive elements (unproven) or flux-moving mechanisms (slow). (v) *Privacy*: ultra-efficient event streams deployed at scale inherit the surveillance exposure documented for neuromorphic imaging [3]. (vi) *Per-event vs. per-computation*: if the superconducting network requires many more physical events per logical operation than CMOS — routing overheads, rank growth in the low-rank representation — the node-level advantage can be wholly consumed; a large event-count overhead is not implausible for naive implementations.

What would falsify the claims. The node-level claim fails if measured subgap leakage above $\sim 10 \mu A$ at the operating point pushes E_{event} above the aJ scale, or if measured cold-plate dissipation per event exceeds the value at which the wall-plug figure reaches the baseline. The system claim fails if a benchmark run under NeuroBench rules [5] on a temporally structured workload [6] shows system-level energy (including cryogenics and I/O) worse than a CMOS spiking core — and [5] itself warns that prior neuromorphic efficiency claims have repeatedly failed standardized comparison. The sparsity assumption is testable against [6]: if temporal event density is high enough that event rate, not per-event energy, dominates, the advantage compresses.

Against ourselves. The strongest objection is that we compare a hypothetical, unbuilt circuit against mature silicon. Every input in Section 3 is a design assumption; no cited work provides empirical superconducting-computing measurements at these parameters, so all hardware numbers are derived from first principles and stated assumptions — a limitation only fabrication and measurement can lift. A second objection: "quantum-inspired" here means only "superconducting and adiabatic"; readers expecting quantum speedup will find none, and we accept the framing critique of [11] that the value lies in the physical process, not quantum-mechanical labels. A third: memristive room-temperature approaches [7] avoid cryogenics entirely; if adiabatic CMOS achieved even a 10^{-3} suppression factor it would reach $\sim 2.5 \times 10^{-17}$ J per event at these node scales with no cryogenic infrastructure — still far above our wall-plug figure, but the honest position is that our advantage is conditional on the full adiabatic suppression being achievable only in a superconducting substrate, which is plausible ($R \rightarrow 0$ after the ramp) but not demonstrated here. The structural-vs-driven coherence distinction of [9] suggests a deeper question — whether driven adiabatic coherence can be made structural rather than clock-sustained — that this paper raises but does not answer, as does the question of whether spiking dynamics constitute a generative syntax in the sense of [10], or whether the substrate passes the thermodynamic-viability bar of [12] at system scale.

Open questions. Optimal T under joint dynamic/static minimization with real JJ shunting; scalable adiabatic clock distribution at 4.2 K; cryogenic synapse plasticity; adiabatic I/O across the 4.2 K boundary; and standardized benchmark protocols that include cryogenic overhead as a first-class accounting item.

7. Conclusion

We have presented a fully derived, reconciled energy budget for a quantum-inspired, superconducting, near-adiabatic spiking computing architecture. With explicitly stated junction parameters, the computed event energy is 7.5 zJ (2.5 zJ adiabatic + 5 zJ static leakage), with a thermal margin factor of ~ 129 at 4.2 K and a node-level energy reduction of more than eight orders of magnitude against assumed pJ-scale CMOS spiking baselines — far exceeding the 50% target at the component level, and still $\approx 10^6\times$ at the wall plug after a computed $699\times$ cryogenic multiplier. The decisive caveat is systemic: cryogenic overhead power dominates at sparse workloads, so viability is a function of workload intensity and I/O co-design, not node physics alone. We have specified the falsification conditions and argued that standardized, overhead-inclusive benchmarking on temporally structured workloads is the necessary next step. The contribution is not a built system but a transparent, reproducible thermodynamic case — every input sourced, every step shown, every inter-draft divergence documented — for taking superconducting adiabatic neuromorphic hardware seriously as the low-energy end of the computing landscape.

References

[1] Synthetic Biology meets Neuromorphic Computing: Towards a bio-inspired Olfactory Perception System. arXiv:2504.10053v2. <https://arxiv.org/abs/2504.10053v2> [2] Line-based Event Preprocessing: Towards Low-Energy Neuromorphic Computer Vision. arXiv:2601.10742v1. <https://arxiv.org/abs/2601.10742v1> [3] Event Encryption: Rethinking Privacy Exposure for Neuromorphic Imaging. arXiv:2306.03369v3. <https://arxiv.org/abs/2306.03369v3> [4] Quantum-inspired algorithms in practice. arXiv:1905.10415v3. <https://arxiv.org/abs/1905.10415v3> [5] NeuroBench: A Framework for Benchmarking Neuromorphic Computing Algorithms and Systems. arXiv:2304.04640v5. <https://arxiv.org/abs/2304.04640v5> [6] NeuroMorse: A Temporally Structured Dataset For Neuromorphic Computing. arXiv:2502.20729v1. <https://arxiv.org/abs/2502.20729v1> [7] Memristors -- from In-memory computing, Deep Learning Acceleration, Spiking Neural Networks, to the Future of Neuromorphic and Bio-inspired Computing. arXiv:2004.14942v1. <https://arxiv.org/abs/2004.14942v1> [8] Exploring the landscapes of "computing": digital, neuromorphic, unconventional -- and beyond. arXiv:2011.12013v3. <https://arxiv.org/abs/2011.12013v3> [9] DOI 10.5281/zenodo.18441401. QNFO: Structural vs Driven Quantum Coherence. [10] DOI 10.5281/zenodo.22758173. QNFO: Syntactic Generation. [11] DOI 10.5281/zenodo.22753022. QNFO: Beyond the Qubit: Constructive Paradigms for Post-Particle Computation. [12] DOI 10.5281/zenodo.18036068. QNFO: Thermodynamic Viability and the Universality of Feynman Matter.

Appendix A. Divergence report

D1. Baseline energy per event (A vs. B vs. C). Draft A used a cited CMOS baseline of 10 pJ/spike and an ASIC baseline of 1.2 pJ/spike from [2]. Draft B derived its own baseline from first principles ($\frac{1}{2}CV^2 = 0.5$ fJ for a 1 fF/1 V node), arguing that cited spiking figures conflate system-level costs. Draft C assumed 10 pJ per synaptic event as a representative spiking-core figure. *Convention behind the disagreement:* A compares against reported system measurements; B compares physically matched switching events; C compares against a stated assumption. **Resolution:** the reconciled text adopts the 10 pJ spiking baseline as primary (the only baseline tied to reported measurements, per A and C) and reports sensitivity against 1.2 pJ and 1 pJ, while noting (per B) that per-event and per-computation comparisons are not equivalent. No divergence is silently resolved.

D2. Node design point (B vs. C). Draft B used $C = 1$ fF, $V = 1$ V, $R = 100\ \Omega$, $\tau = 10$ ns ($E_{ad} = 5 \times 10^{-21}$ J). Draft C used $C = 100$ fF, $V = 0.5$ mV, $R = 10$ k Ω , $T = 10$ ns ($E_{ad} = 2.5 \times 10^{-21}$ J). *Convention:* B chose a generic small CMOS-like node to make the comparison conservative in voltage; C chose junction-physics-native parameters

(millivolt gap scale), arguing the superconducting gap sets V. **Resolution:** C's design point is adopted because it is physically native to the proposed substrate; B's parameters are retained in the divergence record. Both yield zJ-scale event energies, so the qualitative conclusion is robust to this choice.

D3. Crediting the quantum-inspired low-rank factor (A vs. B/C). Draft A divided the per-spike energy by an operation-count reduction factor $\rho = 10$ attributed to [4]. Drafts B and C declined to credit any such factor, B explicitly citing [4]'s finding that quantum-inspired speedups lose to tuned classical baselines once constants are measured. **Resolution:** the reconciled text does *not* credit ρ in the numerical budget, treating low-rank representation as a qualitative design element requiring empirical validation on NeuroBench/NeuroMorse workloads. This is the conservative choice; adopting A's factor would only strengthen the conclusion.

D4. Cryogenic penalty treatment (A vs. B/C). Draft A estimated a ~ 100 J-per-J-removed cooling penalty "for discussion only," excluding it from the primary budget. Draft B computed a wall-plug multiplier (10% of Carnot), reporting $704\times$; the reconciled arithmetic gives $1/0.001430 \approx 699\times$, which is applied. Draft C assumed a 10^{-4} -class efficiency for chip-scale cooling (~ 10 W per mW at 4.2 K). *Convention:* A treats cooling as a second-order discussion item; B and C treat it as a first-class budget item. **Resolution:** the reconciled text adopts B/C's convention: the $\approx 699\times$ multiplier is computed and applied (Step 9), and C's system-level caveat (Step 10) is retained. The conclusion survives: even at the wall plug the reduction exceeds $10^6\times$.

D5. Headline result magnitude (A vs. B vs. C). A reported $0.01 \mu\text{J}$ per 10^6 events (10^{-14} J/event); B reported 3.52×10^{-18} J/event wall-plug; C reported 7.5×10^{-21} J/event cold. These are not contradictory but reflect D1–D4; the reconciled headline is C's 7.5 zJ cold / 5.28×10^{-18} J wall-plug, with A's per-million-event framing recoverable as 7.5×10^{-15} J per 10^6 events cold.

Appendix B. Claim attribution

#	Claim	Sources	Status
C1	Adiabatic charging reduces dissipation by factor RC/T relative to CV^2	A, B, C	CONVERGENT
C2	Superconducting (Josephson) substrate enables millivolt-scale, near-zero-loss spiking nodes	A, B, C	CONVERGENT
C3	CMOS spiking baseline ~ 10 pJ/event; ASIC ~ 1.2 pJ/event (from [2])	A, C	CONVERGENT
C4	Node design point: $C = 100$ fF, $V = 0.5$ mV, $R = 10$ k Ω , $T = 10$ ns	C (B divergent: 1 fF/ 1 V/ 100 Ω)	SINGLE (adopted; D2)
C5	Total event energy 7.5 zJ (2.5 zJ adiabatic + 5 zJ static leakage); static leakage dominates	C	SINGLE
C6	Wall-plug cryogenic multiplier $\approx 699\times$ at 10% of Carnot; wall-plug event energy 5.24×10^{-18} J	B (applied to C's numbers)	CONVERGENT (methodology B; number recomputed per D4)
C7	System-level viability is conditional on workload intensity; cryogenic overhead dominates sparse workloads	B, C	CONVERGENT
C8	Node-level reduction vs. 10 pJ baseline $\approx 1.33 \times 10^9$; $\geq 50\%$ target exceeded with large margin	A, B, C	CONVERGENT (magnitudes differ per D1–D5; conclusion agrees)
C9	Quantum-inspired low-rank factor $\rho = 10$ credited in energy budget	A (B, C divergent)	DIVERGENT (D3; not credited)

#	Claim	Sources	Status
C10	Breakeven ramp time is a central design constraint	B (C convergent via $T/RC \geq 10$ and static-leakage optimum)	CONVERGENT
C11	Thermal margin $E_{event}/(kB \cdot T_{bath}) \approx 129$ at 4.2 K; error suppression $\exp(-43.1) \approx 1.9 \times 10^{-19}$	C	SINGLE (derived, Step 7)
C12	Wall-plug reduction vs. 10 pJ baseline $\approx 1.9 \times 10^6$; system-level victory requires $\sim 10^9$ events/s-equivalent workload	B, C	CONVERGENT (Steps 9–10)