

Leakage as the Price of Binary Truncation: A Scaling Law for Transmon Leakage, Removal Capacity, and Its Ontological Reading

Leakage as the Price of Binary Truncation: A Scaling Law for Transmon Leakage, Removal Capacity, and Its Ontological Reading

Abstract

Superconducting transmon processors lose population from the computational $\{|0\rangle, |1\rangle\}$ subspace into higher levels of an anharmonic oscillator, and this leakage is normally treated as an engineering defect to be suppressed pulse by pulse. The gap addressed here is conceptual and quantitative: no existing framework states how much leakage a gate must produce as a consequence of encoding a bosonic ladder into two levels, nor what removal capacity that production rate demands of a processor architecture. We adopt the perturbative and exact three-level leakage formulas of the Planck-Fidelity analysis of binary truncation of bosonic ladders, insert illustrative transmon parameters ($f_\Omega = 20$ MHz, $f_\alpha = 200$ MHz, $\Omega/|\alpha| = 0.05$), and derive a per-gate time-averaged leakage $\overline{L}_2 = 0.01$ with quadratic scaling in $\Omega/|\alpha|$. Combining this production rate with an assumed 50 ns gate cycle gives a leakage production rate of $2 \times 10^5 \text{ s}^{-1}$, so holding steady-state leakage at 10^{-4} requires a removal capacity $\Gamma_{\text{rem}} = 2 \times 10^9 \text{ s}^{-1}$, which exceeds the fastest leakage hopping rate reported for tunable-coupler transmon arrays (~ 10 MHz) by a factor of 200. For a state-of-the-art $d = 12$ transmon the binary map discards a fraction $f_{\text{disc}} = 0.721057$ of the Hilbert-space information content, and a labeled projection with an assumed proportionality constant gives a surface-code overhead factor $\chi \approx 3$. The residual leakage is interpreted as the empirical cost of projecting a continuous bosonic degree of freedom onto a discrete bit, a falsifiable platform-ordering prediction is stated, and the removal-capacity deficit is shown to justify dedicated leakage-reduction hardware.

1. Introduction

Builders of superconducting quantum processors face a persistent error channel that their error-correcting codes were not designed to absorb: population escaping the two-level computational subspace of each qubit. Leakage excitations traveling between qubits generate correlated errors, induce leakage multiplication, decrease gate fidelity, and increase systemic randomness, and typical error-correction codes protect only against qubit errors [4]. Every gate operated on a transmon therefore carries a hidden tax, paid in population that must be pumped back or it accumulates until the code fails. The cost is concrete: leakage-removal units, DRAG pulse shaping, and frequency-crowding mitigation all consume calibration time, gate cycles, and design margin that would otherwise serve computation. The scale of the stakes is documented by the state-of-the-art review [9]:

35billioninglobalinvestmenthasgoneintosuperconductingarchitectureswhosetwo – levelbehaviorismaintainedbysuppressiontechniques \ 1.9%\$.

What changes if the result of this paper holds? If per-gate leakage is not an avoidable defect but a fixed function of the drive-to-anharmonicity ratio, then processor architects can budget removal capacity from first principles — before building the chip — by reading a single scaling law off the device Hamiltonian. It would also reframe an ontological question in operational terms. The "it from bit" conjecture holds that discrete binary information is fundamental; under that reading, a transmon should natively confine to two levels, and leakage should be a contingent engineering artifact tending toward zero. The alternative reading, advanced by the QNFO research program, is that the transmon is a bosonic mode coerced into two-level behavior, so that "the qubit is the correction term, not the identity" [9]. Under the alternative, leakage measures the cost of the coercion and should scale in a lawful, platform-dependent way: large for bosonic carriers (transmons, fluxonium, cat qubits) and structurally small for fermionic carriers (spin qubits), whose two-level map matches the physical spectrum rather than truncating it.

This paper does not settle the ontology. It derives the scaling law, evaluates it numerically with fully shown arithmetic, and computes the removal-capacity deficit that the law implies for a representative transmon architecture, anchoring the removal side in the measured leakage hopping rates of tunable-coupler transmon arrays [4]. The contributions are:

- A restatement of the binary-truncation leakage bound $L_2 = 8(\Omega/|\alpha|)^2 \sin^2(\alpha T_g/2)$ and the exact three-level result $P_{2,\text{max}} = 2r^2/(1 + 2r^2)$ as the operative scaling law for transmon leakage, with numerical evaluation at illustrative parameters (Sections 3 and 4.1, from the derivations of [8]).
- A computation of the leakage production rate per qubit, $2 \times 10^5 \text{ s}^{-1}$ under stated assumptions, and of the removal capacity $\Gamma_{\text{rem}} = 2 \times 10^9 \text{ s}^{-1}$ needed to hold steady-state leakage at 10^{-4} (Section 4.3).
- A quantified deficit: the fastest leakage hopping rate in tunable-coupler transmon arrays, ~ 10 MHz [4], falls short of the required removal capacity by a factor of 200 (Section 4.4).
- A computation of the discarded-information fraction of a $d = 12$ transmon, $f_{\text{disc}} = 0.721057$, quantifying how much of the physical Hilbert space the binary map discards (Section 4.5, following [8]).
- A labeled projection of the surface-code overhead factor $\chi \approx 3$ under an assumed proportionality constant (Section 4.6).
- A falsifiable platform-ordering prediction — bosonic leakage burden exceeds fermionic — with explicit falsification conditions (Sections 5 and 6).

2. Background and Related Work

The transmon Hamiltonian is $H = 4E_C n^2 - E_J \cos(\phi)$, where $n = Q/(2e)$ is the Cooper-pair number operator, ϕ the superconducting phase difference, $E_C = e^2/(2C)$ the charging energy, and $E_J = I_c \Phi_0/(2\pi)$ the Josephson energy; the $\cos(\phi)$ potential is periodic and supports an infinite ladder of eigenstates, so the system "is not, and has never been, two-level" [9]. In the transmon regime $E_J \gg E_C$, expanding the cosine to fourth order yields an approximately harmonic oscillator with a small quartic correction, and it is this correction — the anharmonicity — that is the only feature distinguishing the device from a pure oscillator [9]. At $E_J/E_C = 325$, the highest ratio reported for a functional transmon, the device supports $d = 12$ resolvable energy levels and its relative anharmonicity $\alpha_r \equiv |\alpha|/\omega_{01}$ is 1.9%; the same source documents the approximately \ 35billioninglobalinvestmentandproposesafalsifiablecalorimetryprotocol : compareheatdissipatedinfreeevolutionofatransmonpre, |1rangle for100nsagainstheatdissipatedbyaDRAG – calibratedgatesequenceimplementingthesameidentityrotation, isolatingthetranslator d=12,\alpha_{r}=1.9%\$ figures as the state-of-the-art anchor.

The quantitative core of our scaling law comes from the Planck-Fidelity analysis of binary truncation of bosonic ladders [8]. That work derives a first-order leakage bound $L_2 = 8(\Omega/|\alpha|)^2 \sin^2(\alpha T_g/2)$ for finite-time gates driven through the raising operator, an exact three-level peak leakage $P_{2,\text{max}} = 2r^2/(1 + 2r^2)$ with time average $\overline{P}_2 = r^2/(1 + 2r^2)$ where $r = f_\Omega/f_\alpha$, and a base-truncation entropy $S_{\text{Base}} = \ln 6 = 1.791759$ nats for $d = 12$, i.e. a discarded-information fraction $f_{\text{disc}} = 0.721057$ of the $\ln 12$ -nat Hilbert-space content; it further notes that one $d = 12$ oscillator would require $\lceil \log_2 12 \rceil = 4$ qubits for lossless binary encoding,

against $\log_2 12 = 3.584963$ qubits of fractional capacity [8]. Its own limitations section states that the perturbative bound is first order in Ω , neglects levels $n \geq 3$, and holds for finite-time finite-amplitude drives of the stated form under the rotating-wave approximation, with shaped envelopes able to null first-order leakage only at isolated, measure-zero parameter points [8]. We reuse its formulas and arithmetic unchanged and build the removal-capacity analysis on top of them.

On the removal side, the leakage-mobility study of transmons with tunable couplers [4] establishes three points we use directly. First, leakage beyond the qubit subspace is a considerable error source and a challenge even for quantum error correction, since typical codes are protected only against qubit errors [4]. Second, leakage is mobile: excitations hop among qubits, causing correlated errors, leakage multiplication, and gate-fidelity degradation, and mobility depends on lattice architecture, couplers, and operating parameters [4]. Third, quantitatively, coupler operating points that cancel single-excitation exchange or ZZ interactions do not simultaneously suppress leakage hopping; in the resonant case leakage hopping rates fall in the range of 0.8–10 MHz, and next-nearest-neighbor hopping reaches approximately 0.1 MHz when the nearest site is far-detuned [4]. These hopping rates are the only removal-transport speeds we treat as measured, and Section 4.4 compares them against the production-derived removal requirement.

The calibration-loop analysis [10] supplies a complementary constraint: under subspace-restricted measurements, the total-variation distance between a strict two-level ontology and a three-level anharmonic carrier with leakage amplitude $|c| = 0.1$ is exactly zero, so population tomography cannot distinguish the ontologies; the degeneracy degrades as $(\Omega/2\Delta)^2$ at larger drive, and breaking it once a leakage-resolving observable exists requires shot budgets of $N \geq 2500$ (one sigma) and $N \geq 22500$ (three sigma) against a per-shot readout error scale $\varepsilon_{\text{ro}} = 9 \times 10^{-5}$ [10]. This matters here because it explains why the leakage channel this paper quantifies is invisible to the certification apparatus actually used on processors: the standard measurement chain is confined to the computational subspace. The platform pedagogy [11] completes the picture of what control and readout physically do — control pulses couple to the electric dipole of the plasmon mode, readout measures a macroscopic observable correlated with the qubit state — and states that a qubit is a constrained two-level subspace of a physical degree of freedom, which working engineers have never doubted, while flagging the fragile view that each abstraction layer from physical qubit to fault-tolerant computation compounds gauge freedoms [11]. Our scaling law gives the fragile view a number.

Three further works anchor the methodological posture. The European strategy analysis [1] argues that progress on the fundamental puzzles unsolved by the Standard Model requires a judicious combination of abstract methods in theoretical physics and precise experimental scrutiny, with precision measurements of rare processes as the low-energy alternative to direct high-energy searches, and notes that future colliders will produce 10^{12} Z bosons to enable such determinations [1]; we borrow that two-sided discipline: an abstract claim (ontological priority of the continuous) must be paired with a precision observable (leakage rates). The linear-collider report [2] demonstrates the pattern of deriving a facility's feasibility from its scaling laws before construction, presenting the design and technology basis of a 0.5–1.5 TeV linear collider and the accelerator-physics lessons of the SLC as the foundation for the next machine's design choices [2]; our removal-capacity budget plays the same pre-construction role for leakage architecture. The MHD design study [6] shows scaling-law extrapolation in practice: tokamak ignition is projected from empirical scaling laws, and stabilization requirements — e.g. the ECCD current for neoclassical-tearing-mode stabilization computed with a Lundquist number $S = 5.56 \times 10^6$ and bootstrap fraction $j_b/j_p = 0.3$ — are derived before engineering designs are frozen [6].

Finally, two works from outside physics supply a structural analogy we use once each. The API-documentation study [5] found that 66.5% of analyzed API elements carry at least one usage directive, i.e. that a reusable abstraction requires an extensive constraint-documentation layer; the transmon's pulse-shaping and leakage-removal apparatus is exactly such a directive layer for the qubit abstraction. The digital-archive study [3] shows that a classification scheme built for one purpose (its FashionStyle14 categories) "does not cover all contemporary fashion styles and their substyles in a mutually exclusive and collectively exhaustive manner" [3]; the $\{|0\rangle, |1\rangle\}$ radix is a classification of the same kind — purpose-built, and provably non-exhaustive over the underlying state space. The MPEG-21 digital-item declaration work [7] demonstrates that a formal representation language can capture complex hierarchical assets with declared structure and rights; we use it once as evidence that hierarchical digital representations are conventionally engineered artifacts whose choice of radix is a design decision, as the choice of a two-level computational subspace is a design decision over a bosonic ladder [7].

3. Methods

We proceed in five steps, using only formulas already derived in [8] and measured rates from [4], plus explicitly stated assumptions where the sources are silent.

Step 1 (leakage per gate). The operative scaling law is the first-order bound of [8],

$$L_2 = 8 \left(\frac{\Omega}{|\alpha|} \right)^2 \sin^2 \left(\frac{\alpha T_g}{2} \right),$$

where Ω is the drive Rabi frequency, $|\alpha|$ the anharmonicity magnitude, and T_g the gate duration. Its gate-averaged form, for gate times uniform over the anharmonicity beat period (mean of $\sin^2$ equal to $1/2$), is

$$\overline{L}_2 = 4 \left(\frac{\Omega}{|\alpha|} \right)^2.$$

The exact three-level companion, with $r = f_\Omega/f_\alpha$ the ratio of drive frequency to anharmonicity frequency scales, is

$$P_{2,\text{max}} = \frac{2r^2}{1 + 2r^2}, \quad \overline{P}_2 = \frac{r^2}{1 + 2r^2}.$$

Step 2 (production rate). We convert per-gate leakage into a volumetric production rate $\Lambda = \overline{L}_2 \cdot f_g$, where f_g is the gate cycle rate. The gate cycle time is an assumption of this paper: we take a cycle of twice the resonant π -pulse duration $t_\pi = 25$ ns computed for the illustrative parameter set, i.e. $T_{\text{cyc}} = 50$ ns, $f_g = 2 \times 10^7$ s $^{-1}$.

Step 3 (removal requirement). We model per-qubit leakage population P_L as

$$\frac{dP_L}{dt} = \Lambda - \Gamma_{\text{rem}} P_L,$$

with Γ_{rem} an effective removal rate. Steady state gives $P_{L,\text{ss}} = \Lambda/\Gamma_{\text{rem}}$, so holding leakage at a target P^* requires $\Gamma_{\text{rem}} \geq \Lambda/P^*$. The target $P^* = 10^{-4}$ is our assumption, chosen as a plausible leakage budget per qubit for a code that tolerates only qubit errors [4]; we treat it as a free parameter and report the scaling.

Step 4 (entropy of truncation). Following [8], for a ladder of d levels the full Hilbert-space information content is $\ln d$ nats and the base-truncation entropy retained by the two-level map's classical shadow is

$$S_{\text{Base}} = \ln \left(\frac{d}{2} \right),$$

with discarded-information fraction

$$f_{\text{disc}} = \frac{S_{\text{Base}}}{\ln d}.$$

Step 5 (overhead projection). As a labeled projection, we assume that each leaked excitation must be corrected by an additional physical qubit and that the required overhead factor scales linearly in the peak leakage,

$$\chi = 1 + \kappa P_{2,\text{max}},$$

with the proportionality constant set to $\kappa = 100$ as an assumption of this paper, of the order of typical surface-code qubit budgets. This projection is single-purpose: it translates the leakage number into a resource-estimate multiplier and carries no ontological weight.

Step 6 (comparison and interpretation). We compare Γ_{rem} against the leakage hopping rates of [4] (0.8–10 MHz resonant; ≈ 0.1 MHz next-nearest-neighbor), compute the discarded-information fraction of the $d = 12$ ladder following [8], and state the ontological reading and its falsification conditions. No experimental data are reported in this paper; every number is computed below with shown arithmetic or labeled a projection.

4. Analysis

4.1 Leakage per gate at the illustrative operating point

Inputs. From the illustrative parameter sets of [8] (explicitly assumptions, not measurements): $f_{\Omega} = 20$ MHz, $f_{\alpha} = 200$ MHz, and $\Omega/|\alpha| = 0.05$.

Compute the ratio:

$$r = \frac{f_{\Omega}}{f_{\alpha}} = \frac{20}{200} = 0.1, \quad r^2 = 0.01.$$

Exact three-level peak leakage:

$$P_{2,\text{max}} = \frac{2r^2}{1 + 2r^2} = \frac{2 \times 0.01}{1 + 2 \times 0.01} = \frac{0.02}{1.02} = 0.0196078 \approx 1.96 \times 10^{-2}.$$

Time-averaged three-level leakage:

$$\bar{P}_2 = \frac{r^2}{1 + 2r^2} = \frac{0.01}{1.02} = 0.00980392 \approx 9.80 \times 10^{-3}.$$

Perturbative bound at worst-case gate timing ($\alpha T_g/2 = \pi/2$, so $\sin^2 = 1$):

$$L_2 = 8 \left(\frac{\Omega}{|\alpha|} \right)^2 \sin^2 \left(\frac{\alpha T_g}{2} \right) = 8 \times (0.05)^2 \times 1 = 8 \times 0.0025 = 0.02.$$

Gate-averaged:

$$\bar{L}_2 = 8 \times 0.0025 \times \frac{1}{2} = 0.01.$$

Scaling check. Doubling $\Omega/|\alpha|$ from 0.05 to 0.10 gives, through the peak form, $L_2(0.10)/L_2(0.05) = (0.10/0.05)^2 = 4$, and indeed $8 \times (0.10)^2 = 0.08 = 4 \times 0.02$. The quadratic scaling in $\Omega/|\alpha|$ is confirmed.

4.2 The scaling law as an architectural constraint

The law $\bar{L}_2 = 4(\Omega/|\alpha|)^2$ has one consequence that is independent of the illustrative numbers: per-gate leakage can be driven down only by increasing anharmonicity relative to drive strength. But the transmon's anharmonicity is bounded above by its design: at the state of the art, $\alpha_r = 1.9\%$ [9], and increasing E_J/E_C to suppress charge noise simultaneously suppresses $|\alpha|$. The transmon therefore sits in a design corner where the truncation cost is minimized in decoherence terms and maximized in leakage terms. This is the quantitative content of the statement that the qubit is the correction term, not the identity [9]: the correction term's magnitude, 1.9%, is exactly the denominator of the leakage law.

4.3 Leakage production rate and removal requirement

Input (assumption). Gate cycle time $T_{\text{cyc}} = 50$ ns, chosen as twice the π -pulse duration computed from the illustrative parameters:

$$t_{\pi} = \frac{1}{2f_{\Omega}} = \frac{1}{2 \times 20 \times 10^6 \text{ s}^{-1}} = 2.5 \times 10^{-8} \text{ s} = 25 \text{ ns}.$$

Then

$$f_g = \frac{1}{T_{\text{cyc}}} = \frac{1}{50 \times 10^{-9} \text{ s}} = 2 \times 10^7 \text{ s}^{-1}.$$

Production rate using $\bar{L}_2 = 0.01$ from Section 4.1:

$$\Lambda = \bar{L}_2 \cdot f_g = 0.01 \times 2 \times 10^7 \text{ s}^{-1} = 2 \times 10^5 \text{ s}^{-1}.$$

That is, each qubit injects population into the leakage manifold at 2×10^5 events per second while gates run continuously.

Removal requirement for target $P^* = 10^{-4}$ (assumption):

$$\Gamma_{\text{rem}} \geq \frac{\Lambda}{P^*} = \frac{2 \times 10^5 \text{ s}^{-1}}{10^{-4}} = 2 \times 10^9 \text{ s}^{-1}.$$

Both Λ and Γ_{rem} are population rate constants in s^{-1} , not angular frequencies, so the comparison with the ordinary-frequency hopping rates of [4] in Section 4.4 is direct.

Accumulation time without removal. Time for leakage population to reach $P = 10^{-2}$ at production Λ :

$$t_{\text{fill}} = \frac{10^{-2}}{2 \times 10^5 \text{ s}^{-1}} = 5 \times 10^{-8} \text{ s} = 50 \text{ ns}.$$

One gate cycle. Without active removal, a continuously gated transmon reaches percent-level leakage within a single cycle at this operating point — which is the operational reason leakage-reduction units exist as dedicated hardware [9].

4.4 Removal-capacity deficit against measured hopping rates

The measured leakage hopping rates in tunable-coupler transmon arrays are 0.8–10 MHz in the resonant case and approximately 0.1 MHz for next-nearest-neighbor hopping [4]. Converting to s^{-1} :

$$10 \text{ MHz} = 1.0 \times 10^7 \text{ s}^{-1}, \quad 0.8 \text{ MHz} = 8 \times 10^5 \text{ s}^{-1}, \quad 0.1 \text{ MHz} = 1 \times 10^5 \text{ s}^{-1}.$$

Deficit factors against the requirement $\Gamma_{\text{rem}} = 2 \times 10^9 \text{ s}^{-1}$:

$$\frac{2 \times 10^9}{1.0 \times 10^7} = 200 \quad (\text{fastest resonant hopping}),$$

$$\frac{2 \times 10^9}{8 \times 10^5} = 2500 \quad (\text{slowest resonant hopping}),$$

$$\frac{2 \times 10^9}{1 \times 10^5} = 2 \times 10^4 \quad (\text{next-nearest-neighbor hopping}).$$

Even the fastest measured leakage transport channel is short of the required removal capacity by a factor of 200, under the stated assumptions ($\bar{L}_2 = 0.01$, $f_g = 2 \times 10^7 \text{ s}^{-1}$, $P^* = 10^{-4}$). The deficit scales linearly in each assumption: halving the per-gate leakage to 0.005 halves the requirement to $1 \times 10^9 \text{ s}^{-1}$ and the deficit to 100; relaxing the target to $P^* = 10^{-3}$ divides the requirement by 10, leaving a deficit factor of 20 at the fastest hopping rate. Passive mobility alone therefore cannot close the gap at any operating point in this family; dedicated, faster removal mechanisms (resonator-mediated swap-and-reset, driven removal units) are structurally required. This is our own assessment from the deficit arithmetic: the measured hopping rates of [4] bound passive transport, and no passive rate in that measured range reaches the required Γ_{rem} .

4.5 Discarded information of the binary map

Following [8], for a $d = 12$ ladder the full Hilbert-space information content is

$$\ln 12 = 2.484906 \text{ nats},$$

and the base-truncation entropy retained by the two-level map's classical shadow is

$$S_{\text{Base}} = \ln 6 = 1.791759 \text{ nats} = \frac{1.791759}{\ln 2} = 2.584963 \text{ bits}.$$

The discarded fraction is

$$f_{\text{disc}} = \frac{\ln 6}{\ln 12} = \frac{1.791759}{2.484906} = 0.721057,$$

i.e. 72.1% of the oscillator's information content is outside what the binary radix efficiently represents; the radix efficiency is $\eta = 1 - 0.721057 = 0.278943 \approx 27.9\%$, and one $d = 12$ oscillator would require $\lceil \log_2 12 \rceil = 4$ qubits for lossless binary encoding, against $\log_2 12 = 3.584963$ qubits of fractional capacity [8]. The leakage computed in Sections 4.1–4.4 is the dynamical face of this static 72.1% deficit: what the map discards, the drive partially re-injects as population outside the computational subspace.

4.6 Observability of the leakage channel

The calibration-loop result [10] bounds what the standard certification apparatus can see. Under subspace-restricted measurements, a two-level ontology and a three-level carrier with leakage amplitude $|c| = 0.1$ are observationally identical with total-variation distance $D_{\text{TV}} = 0$ exactly; the degeneracy breaks only as $(\Omega/2\Delta)^2$ at larger drive, and resolving it against the readout error floor $\varepsilon_{\text{ro}} = 9 \times 10^{-5}$ requires $N \geq 2500$ shots (one sigma) or $N \geq 22500$ (three sigma) once a leakage-resolving observable exists [10]. The signal-to-floor margin for $|c|^2 = 0.01$ is

$$\frac{|c|^2}{\varepsilon_{\text{ro}}} = \frac{0.01}{9 \times 10^{-5}} = 111.1 \approx 111,$$

as reported in [10]. The leakage this paper quantifies is therefore real, large relative to the readout floor, and yet invisible to the tomography actually run on processors — a measurement-architecture gap, not a smallness of the effect.

4.7 Surface-code overhead projection (labeled projection)

Under the Step 5 assumptions ($\kappa = 100$, linear scaling in peak leakage), the overhead factor at the illustrative operating point is

$$\chi = 1 + \kappa P_{2,\text{max}} = 1 + 100 \times 0.0196078 = 1 + 1.96078 = 2.96078 \approx 3.0.$$

This is a projection, not a derivation from error-correction theory: the constant $\kappa = 100$ is an assumption of this paper, and the linear scaling form is untested against full decoder simulations. It is reported because it converts the leakage number into the resource-estimate language that architects use, and its uncertainty is dominated by κ .

5. Results

All numbers below are computed in Section 4 with shown arithmetic, or labeled projections under stated assumptions. No experimental or simulated data are reported in this paper.

R1. Per-gate leakage at the illustrative operating point (Section 4.1; inputs $f_\Omega = 20$ MHz, $f_\alpha = 200$ MHz, $\Omega/|\alpha| = 0.05$, all assumptions following [8]). The exact three-level peak leakage is $P_{2,\max} = 0.02/1.02 = 1.96 \times 10^{-2}$ and its time average is $\bar{P}_2 = 9.80 \times 10^{-3}$; the perturbative bound gives $L_2 = 0.02$ at worst-case gate timing and a gate-averaged $\bar{L}_2 = 0.01$. The scaling is quadratic in $\Omega/|\alpha|$: doubling the ratio to 0.10 multiplies L_2 by 4 ($0.02 \rightarrow 0.08$). The quadratic exponent is the robust feature; the absolute coefficient is first order in Ω and valid only for the stated drive class [8].

R2. Leakage production rate (Section 4.3; assumption $T_{\text{cyc}} = 50$ ns, so $f_g = 2 \times 10^7$ s $^{-1}$). With $\bar{L}_2 = 0.01$, the per-qubit production rate is $\Lambda = \bar{L}_2 f_g = 2 \times 10^5$ s $^{-1}$. Without removal, leakage reaches 10^{-2} population in $t_{\text{fill}} = 50$ ns, one gate cycle. Λ is linear in both $\bar{L}_2$ and f_g .

R3. Removal-capacity requirement (Section 4.3; assumption $P^* = 10^{-4}$). Holding steady-state leakage at P^* requires $\Gamma_{\text{rem}} \geq \Lambda/P^* = 2 \times 10^9$ s $^{-1}$ per qubit. The requirement scales as $1/P^*$: at $P^* = 10^{-3}$ it falls to 2×10^8 s $^{-1}$.

R4. Removal-capacity deficit against measured hopping rates (Section 4.4; measured rates from [4]: 0.8–10 MHz resonant, ≈ 0.1 MHz next-nearest-neighbor). Deficit factors against $\Gamma_{\text{rem}} = 2 \times 10^9$ s $^{-1}$ are 200 at the fastest resonant hopping (1.0×10^7 s $^{-1}$), 2500 at the slowest resonant hopping (8×10^5 s $^{-1}$), and 2×10^4 at next-nearest-neighbor hopping (1×10^5 s $^{-1}$). Each deficit is linear in $\bar{L}_2$, in f_g , and in $1/P^*$.

R5. Discarded-information fraction of the binary map (Section 4.5, following [8]). For the $d = 12$ state-of-the-art ladder [9], $\ln 12 = 2.484906$ nats, $S_{\text{Base}} = \ln 6 = 1.791759$ nats = 2.584963 bits, and $f_{\text{disc}} = 1.791759/2.484906 = 0.721057$; the radix efficiency is $\eta = 0.278943$, and lossless binary encoding of one $d = 12$ oscillator requires $\lceil \log_2 12 \rceil = 4$ qubits against $\log_2 12 = 3.584963$ fractional. This figure depends only on d and is exact given the entropy definition of [8].

R6. Observability margin (Section 4.6, from [10]). Against the readout error floor $\varepsilon_{\text{ro}} = 9 \times 10^{-5}$, a leakage amplitude $|c|^2 = 0.01$ stands at a signal-to-floor margin $|c|^2/\varepsilon_{\text{ro}} = 111.1 \approx 111$, while subspace-restricted tomography has total-variation distance $D_{\text{TV}} = 0$ from the two-level ontology; resolving the degeneracy requires $N \geq 2500$ shots (one sigma) or $N \geq 22500$ (three sigma) once a leakage-resolving observable exists [10].

R7. Surface-code overhead projection (Section 4.7; labeled projection, assumptions $\kappa = 100$ and linear scaling in peak leakage). $\chi = 1 + \kappa P_{2,\max} = 1 + 100 \times 0.0196078 = 2.96078 \approx 3.0$. This is a resource-estimate translation, not a derivation from error-correction theory; its uncertainty is dominated by the assumed κ and the untested linear form.

6. Discussion

What the result establishes and what it does not. The paper establishes a conditional chain: *if* per-gate leakage is governed by the binary-truncation bound of [8], *and* the illustrative parameters are taken as representative, *then* the removal capacity required to hold leakage at 10^{-4} exceeds the fastest measured passive leakage transport in tunable-coupler arrays by a factor of 200 [4]. Each link carries its own assumptions, and the chain is only as strong as them. The scaling law itself is derived, not measured; the parameter set is illustrative; the target $P^* = 10^{-4}$ is a budgeting choice; and the hopping rates of [4] characterize passive mobility between coupled transmons, not the throughput of engineered removal units, which can in principle operate faster than natural hopping by driving the swap explicitly.

The ontological reading. The results support a specific, restricted reading of the "it from bit" conjecture. If binary information were ontologically fundamental in the operational sense — if the transmon's native dynamics preferred the two-level subspace — leakage would be a contingent defect suppressible toward zero without structural cost. What is observed instead, and quantified here, is a leakage production that is a fixed function of the truncation: the drive re-injects exactly the population the radix discards, at a rate $\bar{L}_2 = 4(\Omega/|\alpha|)^2$ set by the 1.9% anharmonicity that is the transmon's only non-oscillator feature [9]. Under this reading, leakage is the empirical price of the digital map, and the 72.1% discarded-information fraction of Section 4.5 is its static counterpart. The reading remains interpretive: the same numbers are consistent with the mundane statement that transmons are weakly anharmonic oscillators and must be engineered accordingly. The paper's claim is that the two statements are the same statement, and that the scaling law makes the ontological version falsifiable.

Falsification conditions. The main result — the removal-capacity deficit and its reading as a truncation cost — would be falsified by any of the following observations. (i) A transmon architecture achieving per-gate leakage far below $4(\Omega/|\alpha|)^2$ at practical drive strengths, with shaped pulses nulling leakage generically rather than at isolated parameter points; [8] itself states that shaped envelopes can null first-order leakage only at measure-zero parameter points, but a demonstration to the contrary at a dense set of operating points would break the law. (ii) A measurement showing leakage production in operating transmons orders of magnitude below 2×10^5 s $^{-1}$ at comparable gate rates, which would mean the illustrative parameters misrepresent practice by more than the assumed linear sensitivity tolerates. (iii) A passive or engineered removal mechanism sustaining $\Gamma_{\text{rem}} \sim 10^9$ s $^{-1}$ per qubit without dedicated fast hardware, collapsing the deficit. (iv) The platform-ordering prediction failing: spin-qubit platforms showing leakage burdens comparable to or exceeding bosonic platforms at matched gate quality would sever the link between truncation and leakage, and with it the ontological reading.

Limitations list. The result rests on the following assumptions, with sensitivity stated for each.

- *Perturbative, three-level model* [8]: first order in Ω , neglects levels $n \geq 3$, holds for finite-time drives of the stated form under the rotating-wave approximation. Leakage at larger drive or with strong pulse shaping could deviate; the quadratic scaling is the most robust feature, the absolute coefficient the least.
- *Illustrative parameters* ($f_\Omega = 20$ MHz, $f_\alpha = 200$ MHz, $\Omega/|\alpha| = 0.05$): assumptions, not measurements [8]. The production rate Λ is linear in $\bar{L}_2$ and in f_g , so a factor-2 error in either shifts the deficit by the same factor.
- *Gate cycle time* $T_{\text{cyc}} = 50$ ns: an assumption of this paper. Faster cycles raise Λ proportionally; slower cycles lower it.
- *Target* $P^* = 10^{-4}$: a budgeting choice. The requirement scales as $1/P^*$; at $P^* = 10^{-3}$ the deficit at the fastest hopping rate falls to 20, which passive mobility still cannot close but which dedicated hardware plausibly could.
- *Hopping rates as removal ceiling* [4]: the 0.8–10 MHz rates characterize passive mobility, not driven removal units; treating them as the ceiling is conservative for the deficit claim but not a measurement of removal-unit throughput.
- *Overhead projection* ($\kappa = 100$, linear in peak leakage): untested against decoder simulations; reported only as a resource-estimate translation.
- *Single-qubit treatment*: the analysis is per qubit and ignores leakage multiplication and correlated errors, which [4] identifies as mobility's principal harms; a multi-qubit treatment could raise the effective removal demand above the single-qubit figure.

Open questions. What is the actual throughput of a driven swap-and-reset removal unit, and does it approach 10^9 s $^{-1}$? How does the leakage law modify under realistic shaped envelopes and counter-rotating terms, which [10] lists as neglected effects? Does the platform ordering hold when gate quality, connectivity, and readout differences between platforms are controlled for? These questions define the experimental program the scaling law proposes.

7. Practical Implications and Engineering Pathways

(a) Who would use the result and what changes. The users are superconducting processor architects and leakage-removal-unit designers. If the scaling law holds, the removal-capacity requirement $\Gamma_{\text{rem}} \geq \bar{L}_2 f_g / P^*$ becomes a design-input computed from the device Hamiltonian and the target gate rate before tape-out, replacing the current iterate-and-measure calibration of leakage budgets. The design parameters that change are: the specified throughput of leakage-removal units (the deficit factor of 200 under the stated assumptions indicates passive mobility cannot be the removal mechanism, so removal units must be driven and resonator-mediated); the allocation of gate-cycle time to removal operations; and the choice of operating point, since the law makes explicit that raising $|\alpha|/\Omega$ is the only first-order lever on leakage. For error-correction architects, the discarded-information fraction and the overhead projection supply a leakage line-item to add to qubit budgets. No commercial or market figure is claimed; the work is foundational in the sense that its immediate output is a budget equation, not a device.

(b) Readiness level. This is purely analytical work that has not been peer reviewed: under the NASA Technology Readiness Level definitions it sits at TRL 1–2 (basic principles observed and reported; concept formulated). The scaling law is the first-order binary-truncation leakage bound derived in [8], together with its exact three-level companion and entropy figures, and it is anchored to one set of measured hopping rates [4]; no hardware demonstration informed by the law exists. Analytical work of this kind does not exceed TRL 3, and no manufacturing readiness is claimed.

(c) Smallest next experiment. The cheapest de-risking measurement is a single-qubit leakage production test: drive one transmon at a calibrated set of $\Omega/|\alpha|$ values spanning 0.02–0.10 with fixed gate rate, measure steady-state leakage population with a leakage-resolving observable (e.g., $|2\rangle$ -selective spectroscopy), and fit $\bar{L}_2$ against $(\Omega/|\alpha|)^2$. The shot budgets required are those computed in Section 4.6: $N \geq 2500$ per point at one sigma against the readout floor $\varepsilon_{\text{ro}} = 9 \times 10^{-5}$ [10]. The outcome that refutes the law is a fitted exponent significantly below 2, or a measured coefficient inconsistent with 4 at the stated confidence, across the drive range. A second, larger test is the platform comparison test of the bosonic-over-fermionic ordering prediction of Section 6 (falsification condition (iv)): matched gate-quality leakage budgets for one bosonic and one fermionic platform.

(d) Binding constraints. The constraints that bind first, in order: (i) *measurement*, not fabrication — the leakage-resolving observable required to test the law does not exist in the standard subspace-restricted readout chain, and building it is the gating step (Section 4.6); (ii) *calibration data* — the law's coefficient must be validated against measured leakage on real devices, since the illustrative parameters are assumptions; (iii) *removal hardware* — if the deficit holds, removal units with effective rates far above measured hopping are needed, and their speed is bounded by resonator swap times and reset fidelity, quantities this paper does not measure; (iv) *frequency crowding* — raising $|\alpha|/\Omega$ by increasing drive competes with the crowding constraints that already shape transmon frequency allocations, an interaction this paper does not model. Cost drivers, scale-up behavior, and regulatory dimensions are open questions; no figure for them is derivable from the present analysis.

(e) Order-of-magnitude bounds (labeled projections). Under the assumptions already stated ($\bar{L}_2 = 0.01$, $f_g = 2 \times 10^7 \text{ s}^{-1}$, $P^* = 10^{-4}$), the removal capacity per qubit is $\Gamma_{\text{rem}} = 2 \times 10^9 \text{ s}^{-1}$; for a code cycle of order $1 \mu\text{s}$ (an assumption of this projection), the removal operations required per qubit per cycle number $\Gamma_{\text{rem}} \times 10^{-6} = 2 \times 10^3$, i.e. removal must be a continuous, resonantly mediated process rather than a periodic reset. These are projections; their validity is conditional on every assumption in the chain, and the sensitivity analysis of Section 4.4 gives their linear scaling.

(f) Risks and misuse paths. The principal misuse is rhetorical: the ontological framing ("the transmon is not a qubit") can be quoted to imply that transmon processors are fundamentally unviable, which the analysis does not show — the deficit quantifies a removal requirement, not an impossibility, and the transmon remains the platform against which the law is most testable. A second risk is overreading the overhead projection $\chi \approx 3$ as an error-correction result; it is a resource-estimate translation under an assumed constant and should not enter design reviews as a derived overhead.

8. Conclusion

This paper derives, from the binary-truncation leakage analysis of bosonic ladders [8], an operative scaling law for transmon leakage, $\bar{L}_2 = 4(\Omega/|\alpha|)^2$, evaluates it at an illustrative operating point with fully shown arithmetic, and converts it into an architectural requirement: holding steady-state leakage at 10^{-4} under a 50 ns gate cycle demands a removal capacity of $2 \times 10^9 \text{ s}^{-1}$ per qubit, exceeding the fastest measured passive leakage hopping rate in tunable-coupler transmon arrays by a factor of 200 [4]. The same binary map discards 72.1% of the information content of the $d = 12$ state-of-the-art ladder [9, 8], and the leakage channel it produces, while standing a factor of ≈ 111 above the readout error floor, is exactly invisible to the subspace-restricted tomography actually run on processors [10].

The interpretation offered is operational rather than metaphysical: leakage is the dynamical price of projecting a bosonic ladder onto a two-level radix, paid at a rate set by the anharmonicity that is the device's only non-oscillator feature. Whether this price is a contingent engineering defect or the signature of the continuous being prior to the discrete is exactly the question the falsification conditions of Section 6 are designed to settle: a generic pulse-shaping null of the leakage, a passive removal mechanism at 10^9 s^{-1} , or a failure of the bosonic-over-fermionic platform ordering would each dissolve the reading. The paper's durable contribution, independent of the ontology, is the budget equation $\Gamma_{\text{rem}} \geq \bar{L}_2 f_g / P^*$ and the demonstration that, at any operating point in the stated family, passive mobility cannot satisfy it — so leakage removal must be designed as dedicated, driven hardware, with its throughput specified from the scaling law before the chip is built.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.18.6-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 17 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] Physics Briefing Book. arXiv:1910.11775v2. <https://arxiv.org/abs/1910.11775v2> [2] Physics and Technology of the Next Linear Collider: A Report Submitted to Snowmass '96. arXiv:hep-ex/9605011v1. <https://arxiv.org/abs/hep-ex/9605011v1> [3] A Novel Approach to Analyze Fashion Digital Archive from Humanities. arXiv:2107.08351v2. <https://arxiv.org/abs/2107.08351v2> [4] Leakage Mobility and Passive Leakage Removal in Transmons with Tunable Couplers. arXiv:2607.00688v1. <https://arxiv.org/abs/2607.00688v1> [5] What Should Developers Be Aware Of? An Empirical Study on the Directives of API Documentation. arXiv:1205.6363v1. <https://arxiv.org/abs/1205.6363v1> [6] MHD analysis on the physical designs of CFETR and HFRC. arXiv:2107.11742v1. <https://arxiv.org/abs/2107.11742v1> [7] Representing Digital Assets using MPEG-21 Digital Item Declaration. arXiv:cs/0508065v1. <https://arxiv.org/abs/cs/0508065v1> [8] QNFO: Planck Fidelity: Fundamental Leakage Cost of Binary Truncation of Bosonic Ladders [9] QNFO: The Two-Level Lie: The Transmon Is Not a Qubit — And the Entire Field Knows It [10] QNFO: The Calibration Loop as Fixed Point: Ontological Underdetermination of the Qubit by Its Own Metrology [11] QNFO: No Thing There: Control, Readout, and Self-Referential Metrology in Engineered Quantum Systems