Certified Speed-Limit Diagnostics for Deployed Superconducting Quantum Processors

Carles Marín

***Abstract*—How far are deployed quantum gates from the fastest evolution quantum mechanics allows? We answer for pre-Heron IBM processors, mainly Falcon and Eagle, using public, version-pinned calibration snapshots and inequalities whose proofs are machine-checked in Lean 4. Heron snapshots lack a Hamiltonian. For single-qubit gates, with the drive-strength convention fixed by IBM's calibrated π pulses, the X gate lasts a median 11.3 times the two-level speed limit at full drive amplitude (1646 qubits, 41 backends). Most of this gap is a design choice: IBM drives at about a tenth of the anharmonicity, and at that drive its pulses are about 2 times the square-pulse limit. Within each device's full amplitude, optimal control on IBM's model gives gates about 10 times faster than the median. For two-qubit gates (1758 coupled pairs, 43 backends), the known qubit-subspace minimal time of controlled-NOT-class gates under exchange coupling (literature) gives a median ratio 4.22; across the snapshots it rose, not monotonically, from 2.78 (2021) to 4.26 (2025) as gates got longer, coinciding with the change from Falcon to Eagle. With the second level and fast arbitrary local control, the minimal time of a controlled-Z gate with ideal local control lies between 0.295 and 0.331 of the qubit-only exchange time (a two-copy bound and a certified explicit protocol, version 3); bounded local control reaches 0.35. We also formalize an entangling-time bound in arbitrary local dimension, known for qubits and new only as a machine-checked assembly. Applied to IBM's chip-to-chip state transfer (version 2), the same diagnostic places the transfers at 5.8–7.9 times the single-mode minimal time, whose lower bound is machine-checked in Isabelle, and the coherence error of the fastest protocol four to six times below the measured errors. The contribution is a reproducible diagnostic built on the machine-checked inequalities, with explicit guarantees.**

***Index Terms*—Quantum speed limit, superconducting qubits, transmon, gate time, cross-resonance, Mandelstam–Tamm, Margolus–Levitin, entangling time, formal verification, Lean, reproducible computation.**

Companion site. https://karlesmarin.github.io/blueprints/qsl-diagnostics/: every claim of this paper with its Lean theorem or script, the programs and data, and laboratories in the browser that redraw Figs. 1–4 and diagnose the reader's own gates (Section X). This preprint: doi:10.5281/zenodo.23783382.

I. Introduction

GATE duration enters every figure of merit of a quantum processor: it sets how many operations fit inside the coherence time, and it trades against leakage and control error. Quantum speed limits (QSLs)—Mandelstam–Tamm (MT) [1], Margolus–Levitin (ML) [2] and their fractional-moment family due to Zieliński and Zych [3]—bound how fast a state can evolve to an orthogonal (or partially distinguishable) one given an energy scale. For two-qubit gates the analogous object is the minimal interaction time under fast local control [4], [5], [6].

C. Marín is an independent researcher (e-mail: karlesmarin@gmail.com).

A natural diagnostic is then the *gap*: the calibrated gate duration divided by the speed limit implied by the device's own Hamiltonian parameters. Two things make this diagnostic delicate. First, the energy scale must be the right one: with the laboratory Hamiltonian of a transmon the MT/ML bound measures the carrier frequency and is vacuous ($\sim 0.1\,\mathrm{ns}$ against gate times of tens to hundreds of ns); the meaningful scales are the drive amplitude in the rotating frame (one qubit) and the coupling (two qubits). Second, the conventions in vendor data (factors of 2 and 2π, which parameter is measured and which is derived) must be fixed against independent data, or the diagnostic reports the convention instead of the device.

This paper contributes five things.

(1) Certified bounds. The inequalities listed in Table I are proved in Lean 4 in the public development [25] (the two-qubit gate bound $\pi/(2j_q)$ behind the headline 4.22 is taken from the literature and is not among them): Margolus–Levitin, an L^1 form of Mandelstam–Tamm, the Zieliński–Zych family for every exponent $\alpha > 0$ (whose key auxiliary inequality the original paper states without proof), a time-dependent single-qubit bound with its equality case for fixed-axis pulses and its finite-tolerance version, an entangling-time bound in arbitrary local dimension with its shifted, locally equivalent (ECR) and finite-tolerance versions (Section II), and, for the CZ in d-level transmons, the norm bounds behind its asymptotic tightness and the key lemma of the refined bound (Section VII).

(2) A convention fixed by independent data. The drive-strength parameter is pinned by IBM's calibrated π pulses through an exact pulse-area identity that holds on all 777 qubits with pulse definitions, up to which of two pulse normalizations reaches the qubit, which also reveals that it is derived from calibration rather than measured independently (Section III).

(3) A sweep of deployed hardware: single-qubit gates on 1646 qubits, two-qubit gates on 1758 pairs, and a 2021–2025 time series (Sections IV–VI); for the single-qubit sweep, a falsifier and three rival hypotheses were written before the analysis, and we report their fate.

(4) Feasible times found by optimal control on the vendor's own model, bracketed by lower bounds (certified, or proved in the text) and, for the CZ, by an explicit construction, which in d-level transmons approaches the entangling-time bound as $d \to \infty$ at fixed interaction norm, with ideal local control (Section VII).

(5) An application beyond gates (version 2): IBM's chip-to-chip state transfer [21] runs at 5.8–7.9 times the single-mode minimal time, whose lower bound we machine-check

in Isabelle; the master equation of the optimal protocol puts its coherence error four to six times below the measured errors (Section VIII).

Closest prior work. Ishida and Hasegawa [12] apply speed limits in the inverse direction on the Heron processor `ibm_torino`: they measure gate durations from job wall-time by gate-time amplification (X: 32 ns, CZ: 70 ns) and infer model-free lower bounds on energy scales—through MT, on the time-averaged energy spread, and through ML, on an effective time-independent Hamiltonian—finding them of the order of typical drive energies and concluding that the gates operate near the speed limit. For a fixed-axis π rotation $H(t) = f(t)\sigma_x/2$ with $f \geq 0$ from $|0\rangle$ one has $\langle H\rangle = 0$ and $\Delta E(t) = f(t)/2$, and the MT estimator equals the time average of ΔE exactly, whatever the pulse shape (Section II): the ratio of duration to that limit is 1 by construction, so the estimator fixes an energy scale rather than a gap. (Here energy is the spread of the drive Hamiltonian in the rotating frame, not energy delivered by the control electronics.) We instead take the scale from the calibrated Hamiltonian (maximum drive amplitude, exchange coupling) and report duration/limit ratios. The two views are numerically compatible once the denominators are stated: at the drive IBM actually uses, the X gate is within about a factor 2 of the square-pulse limit (pulse shape and the pulse normalization), and about 11 times above the limit at full drive amplitude (Section IV-C). Their measurement needs no model and covers a current Heron device; our diagnostic needs the vendor Hamiltonian, so it covers pre-Heron (mainly Falcon and Eagle) snapshots only, but spans 41 backends and 2021–2025 and rests on machine-checked inequalities.

We claim no new physics: the speed limits are classical, the two-qubit constants are known for qubits, and the transmon model is IBM's. The exceptions are mathematical: the d-level quadrature code for the CZ and its asymptotic tightness were not found in the sources we read (Section VII). To our knowledge no prior work applies a proof-assistant-certified QSL to public calibration data with a per-gate gap; a search that finds nothing does not prove novelty. Section IX states what the certificate covers.

II. Certified Bounds

Units have $\hbar = 1$. The statements in the list items, the Theorem and the corollaries below are theorems of the Lean 4 development [25]; the remarks and numerical checks are not (module `Formalizations/Physics/QuantumMechanics/SpeedLimit`, commit `42f307b`); Table I lists the theorem names and the status of each statement in the literature.

A. Single-system speed limits

Let a state have populations $p_n \geq 0$, $\sum_n p_n = 1$, on energies E_n, and autocorrelation $S(\tau) = \sum_n p_n e^{-iE_n\tau}$. If $S(\tau) = 0$ (orthogonality) at $\tau \geq 0$:

- *Margolus–Levitin* (`QSL.margolus_levitin`): with $E_n \geq 0$, $\pi \leq 2\langle E\rangle\tau$.
- *Mandelstam–Tamm, L^1 form* (`QSL.mandelstam_tamm_L1`): $1 \leq D_1\tau$ with $D_1 = \sum_n p_n|E_n - \bar{E}|$; with `QSL.mandelstam_tamm_weak`, $1 \leq \Delta E\,\tau$. The full MT constant $\tau \geq \pi/(2\Delta E)$ is *not* proved in the development.
- *Zieliński–Zych* (`QSL.zielinski_zych_all`, `QSL.zielinski_zych_time_all`): with $E_n \geq E_0$ on populated levels and $M_\alpha = \sum_n p_n(E_n - E_0)^\alpha$, $\pi^\alpha \leq 2M_\alpha\tau^\alpha$ for every $\alpha > 0$. The auxiliary inequality $x^\alpha - \pi^\alpha/2 + (\pi^\alpha/2)\cos x + \alpha\pi^{\alpha-1}\sin x \geq 0$ ($x \geq 0$) is stated without proof in [3] for all $\alpha > 0$; Lean proves it for $0 < \alpha \leq 1$ (`QSL.zz_inequality`) and for $\alpha \geq 1$ (`QSL.zz_inequality_of_one_le`).

For a two-level system driven at constant Rabi frequency Ω, ML gives $\tau \geq \pi/\Omega$ for $|0\rangle \to |1\rangle$, saturated by a square π pulse. This is the single-qubit reference time used in Section IV. For a *time-dependent* drive, which is what a shaped pulse needs, two further statements are certified. (i) Any piecewise-constant evolution whose Hamiltonians satisfy $\|H_k - c_k\mathbb{1}\| \leq \Omega_{\max}/2$ for some real c_k (spectral width at most $\Omega_{\max}$; for $H = (\Omega/2)\,\hat{n}\cdot\vec{\sigma}$ this is $|\Omega| \leq \Omega_{\max}$) moves any state by a Fubini–Study angle at most $(\Omega_{\max}/2)\,T$ (`QSL.Gate.fsAngle_le_time_halfWidth`); hence $|0\rangle \to |1\rangle$ needs $T \geq \pi/\Omega_{\max}$ for every pulse shape. (ii) For a fixed-axis drive $H(t) = f(t)\,\sigma_x/2$ with $f \geq 0$ and $\int f \leq \pi$, starting from $|0\rangle$ (a state orthogonal to the axis on the Bloch sphere), the generalized MT bound $\int \Delta E\,dt \geq \arccos|\langle\psi_0|\psi_T\rangle|$ holds with *equality*, the trajectory being proved to solve the Schrödinger equation (`QSL.FixedAxis.mandelstam_tamm_integral_eq_solution`): every fixed-axis π pulse has speed-limit ratio exactly 1, whatever its shape. This is the identity behind the near-unit ratios reported for π pulses in [12]. It does not extend to DRAG pulses, whose quadrature makes the axis time dependent (try it on the Bloch sphere with any envelope, a DRAG term or a detuning).

B. Entangling time in arbitrary local dimension

Two systems with finite index types (arbitrary local dimensions, ancillas included) evolve under $\prod_k \exp[-i\tau_k(A_k \otimes 1 + 1 \otimes B_k + H)]$, with A_k, B_k Hermitian and unbounded (piecewise-constant local control), $\tau_k \geq 0$, and a constant Hermitian interaction H with $\|H\| \leq J$ (operator norm). In this section J is always a bound on that norm; in Sections V and VII the exchange *coefficient* is written j_q or J and the corresponding norm is stated each time (it is J on two levels and $2J$ on three).

Theorem (`QSL.Qudit.arccos_le_entangling_time`). If the evolution maps a unit product vector to Φ, and every unit product vector has overlap at most λ with Φ, then $\arccos\lambda \leq J\sum_k \tau_k$.

The overlap of a state with product vectors is bounded by its largest Schmidt coefficient. For a two-level Bell vector $(e_{00} + e_{11})/\sqrt{2}$ inside qudits of any dimension $d \geq 2$, $\lambda = 1/\sqrt{2}$, hence $J\sum_k \tau_k \geq \pi/4$ (`QSL.Qudit.pi_div_four_le_entangling_time_fin`). The bound is attained: $H = J\,Z \otimes X$ reaches the Bell vector from a product vector in one step of length $\pi/(4J)$ (`QSL.Qudit.entangling_time_tight`). The proof is a Duhamel estimate in the Fubini–Study angle chained over control steps. An earlier chord-distance version in the same development (`QSL.TwoQubit.`

`entangling_time_lower_bound`) gives the weaker constant $\sqrt{2-\sqrt{2}} \approx 0.7654$, 2.6% below the exchange-coupling time $\pi/(4J) = 0.7854/J$.

Consequence for CNOT. CNOT maps $|+\rangle|0\rangle$ to the Bell state, so any implementation of CNOT on the qubit subspace of two d-level systems needs $JT \geq \pi/4$, with J the norm of the interaction on the *full* d-level space. An independent numerical check (`cnot_m3.py`) of the relevant overlap constant gives $m_d = 0.707107$ ($d = 2$) and 0.707105 ($d = 3$) for the largest overlap of a CNOT output with product vectors, over 45 starts, consistent with $1/\sqrt{2}$; a maximization from finitely many starts is a check, not a certificate. The same holds for every gate only *locally equivalent* to CNOT, such as IBM's echoed cross-resonance (ECR): such a gate maps some product vector to $(P \otimes Q)$ applied to a Bell vector, local unitaries do not change the overlap with product vectors, and so $JT \geq \pi/4$ (`QSL.Qudit.pi_div_four_le_entangling_time_local`; with a shifted coupling, `..._shift_local`).

Two further corollaries. (a) A multiple of the identity in H can be absorbed into the local control, so $\|H\|$ may be replaced by $\|H - c\mathbb{1}\|$ for any real c, i.e. by the spectral half-width (`QSL.Qudit.arccos_le_entangling_time_shift`); for the Heisenberg coupling $XX+YY+ZZ$ this gives 2 instead of 3 and the Bell time $\pi/8$ (`QSL.Qudit.pi_div_eight_le_heisenberg_time`). (b) If the final state only has fidelity $1-\varepsilon$ with the Bell vector, the triangle inequality of the Fubini–Study angle gives $J\sum_k \tau_k \geq \pi/4 - \arcsin\sqrt{\varepsilon}$ (`QSL.Qudit.pi_div_four_sub_le_entangling_time_approx`); likewise a two-level transfer to fidelity $1-\varepsilon$ needs $\Omega_{\max} T \geq 2\arccos\sqrt{\varepsilon}$ (`QSL.Gate.two_arccos_sqrt_le_time_approx`). Here ε is the infidelity of a specific output state, not an average gate infidelity.

What is known. For two qubits without ancillas the result is known with equality: Dür *et al.* [4] solve the Schmidt-coefficient dynamics, $P(t) = \sin^2[h_{\max}t + \varphi_0]$, which gives a Bell state at $\pi/(4h_{\max})$; Vidal, Hammerer and Cirac [5] give the interaction cost $C_H(\mathrm{CNOT}) = (\pi/4)(1/h_1)$. The interaction-picture machinery for qudits with $\|H_c\| \leq J$ and unbounded local control is in Basyildiz, Jameson and Gong [7] (Sec. II, Eqs. (1)–(7)), applied there to iSWAP with an orthogonality criterion; the step "overlap with a product $\leq$ largest Schmidt coefficient" combined with the arccos form of MT is in Rudnicki [10]. The statement for arbitrary d and ancillas with the Schmidt criterion was not found as a written theorem in the sources we read; it is a short corollary of [7], [10], and a reader of [7] will see it as a natural extension—the authors write that "a similar but possibly more complicated analysis can be applied to obtain the speed limit for other two-qubit gates." What we add is the machine-checked statement. The initial state must be a product (ancillas included): with pre-entangled ancillas the argument does not apply [5], [11].

III. Data and Conventions

A. Sources

All data are public snapshots shipped with Qiskit under the Apache 2.0 licence: the fake backends of `qiskit-ibm-runtime` 0.51.0 [19] (`props_*.json`: per-qubit T_1, T_2, frequency, anharmonicity, gate errors and durations; `conf_*.json`: the Duffing-model Hamiltonian variables w_q, δ, `omegad` and j_q in rad/ns), and the calibrated pulse definitions (`defs_*.json`, 32 backends) of `qiskit-terra` 0.24.1 [20]. No computation mixes the two packages: the analyses that use pulse shapes (convention, peak drive, cross-resonance, Figs. 1–2, per-qubit durations) read only the `defs` and `conf` of terra 0.24.1, and the sweeps read only the `conf` and `props` of runtime 0.51.0. For the time series (Section VI) we pulled 13 Qiskit releases (terra 0.16–0.25 except 0.18.3, which did not download; runtime 0.20–0.51), giving 101 unique (backend, date) snapshots, of which 56 from 41 backends are usable. Snapshots are calibrations of a given date, not live data; IBM has been withdrawing pulse-level access and frequency/anharmonicity data [12], which makes version-pinned snapshots the citable, freezable source. Our own data show the same trend: the snapshots of the current Heron processors ship an empty Hamiltonian (in `conf_torino.json`, `hamiltonian = {'h_latex': ''}`, with no `omegad` and no j_q), so `torino`, `fez`, `marrakesh` and `kingston` are absent from our sweeps. The diagnostic therefore covers pre-Heron processors only, mainly of the Falcon and Eagle generations; the single-qubit sweep also contains the older `armonk`, `yorktown`, `manhattan` and `brooklyn` snapshots (`barrido.csv`).

B. Model

IBM's own model is a Duffing oscillator per transmon, in the frame rotating at the qubit frequency with the rotating-wave approximation (RWA): $H = (\delta/2)\,n(n-1) + \frac{1}{2}\big(u(t)\,b^\dagger + u(t)^*\,b\big)$, with $u(t) = c \cdot \mathtt{omegad} \cdot D(t)$ for a drive of complex amplitude $D(t)$ (both quadratures; $\Omega = |u|$ is the Rabi frequency), and an exchange coupling $j_q(b_i^\dagger b_j + b_i b_j^\dagger)$ between coupled transmons. The speed limit of an X gate at full amplitude $|D| = 1$ is then $\tau_{\mathrm{QSL}} = \pi/(c \cdot \mathtt{omegad})$ (and half of that for $\sqrt{X}$).

C. Fixing the drive convention

The factor c is not documented in the snapshots. We fixed it before computing any gap, with the pre-specified rival H3 ("a different convention, factor 2 or 2π"; `ficha_ideas.toml`). The independent datum is the calibrated X pulse itself: its rotation angle must be π. Across the 32 backends with pulse definitions, $\mathtt{omegad} \cdot |\mathrm{amp}| \cdot \mathrm{area} = 1.148\,\pi$ with $c = 1$ and $2.296\,\pi$ with $c = 2$ (median, with a simple discretization of the area) (`pulsos_x.py`). We adopt $c = 1$: the conventions with a factor 2 or 2π are excluded. The residual ≈ 1.15 has an exact explanation in the pulse normalization. With $g(x) = \exp[-(x - N/2)^2/(2\sigma^2)]$, $N = 4\sigma$ and $S = \sum_{n=0}^{N-1} g(n + \frac{1}{2})$, let $A_{\mathrm{rec}} = dt\,[S - N\,g(-\frac{1}{2})]$ be the area of the sampled Gaussian with the edge value subtracted but *without* renormalizing the peak, and $A_Q = dt\,[S - N\,g(-1)]/[1 - g(-1)]$ the area of Qiskit's lifted Gaussian. Then $\mathtt{omegad} \cdot |\mathrm{amp}| \cdot A_{\mathrm{rec}} = \pi$ on all 777 qubits with pulse definitions, to $1.2 \cdot 10^{-7}$ (`cruces_area_cz.py`); the residual is A_Q/A_{rec}, which tends to $1/(1 - e^{-2}) =$

TABLE I
CERTIFIED STATEMENTS USED IN THIS PAPER (LEAN 4, COMMIT `42F307B`) AND THEIR STATUS IN THE LITERATURE.

Lean theorem (prefix `QSL.`)	statement	status
`margolus_levitin`	$\pi \le 2\langle E\rangle\tau$	classical [2]
`mandelstam_tamm_L1, _weak`	$1 \le D_1\tau \le \Delta E\,\tau$	weak form of [1]
`zielinski_zych_all`	$\pi^\alpha \le 2M_\alpha\tau^\alpha$, all $\alpha > 0$	stated in [3]; key inequality unproved there
`Qudit.arccos_le_entangling_time`	$\arccos\lambda \le J\sum\tau_k$	corollary of [7], [10]
`Qudit.pi_div_four_le_entangling_time_fin`	Bell: $J\sum\tau_k \ge \pi/4$, any d	$d = 2$: [4], [5]
`Qudit.entangling_time_tight`	attained by $J\,Z\otimes X$	$d = 2$: [4], [5]
`Qudit.pi_div_four_le_entangling_time_local, _shift_local`	same for $(P\otimes Q)$ Bell (ECR)	corollary
`TwoQubit.entangling_time_lower_bound`	chord form, constant $\sqrt{2-\sqrt{2}}$	weaker earlier version
`Qudit.arccos_le_entangling_time_shift`	$\|H\| \to \|H - c\mathbb{1}\|$	corollary; $d = 2$: [4]
`Qudit.pi_div_eight_le_heisenberg_time`	Heisenberg Bell: $T \ge \pi/8$	$d = 2$, no ancilla: [4]
`Gate.fsAngle_le_time_halfWidth`	angle $\le (\Omega_{\max}/2)\,T$, piecewise $H(t)$	standard (shifted MT)
`FixedAxis.mandelstam_tamm_integral_eq_solution`	$\int\Delta E = \arccos\|\langle\psi_0\|\psi_T\rangle\|$, fixed axis	equality case of [1]
`Qudit.pi_div_four_sub_le_entangling_time_approx`	Bell to fidelity $1-\varepsilon$: $J\sum\tau_k \ge \pi/4 - \arcsin\sqrt{\varepsilon}$	robust corollary; formalized here
`Gate.two_arccos_sqrt_le_time_approx`	transfer to fidelity $1-\varepsilon$: $\Omega_{\max}T \ge 2\arccos\sqrt{\varepsilon}$	two-level; tight
`Exchange.exchange_norm_ge, _le, ratio_tendsto_one`	$\lambda_d^2/2 \le \kappa_d \le 2(d-1)$; $2\kappa_d/\lambda_d^2 \to 1$	not found in the sources read
`Exchange.refined_cz_lemma`	$\frac{1}{4}\mathrm{Tr}[(P_A\otimes P_B)H_c^2] \le \frac{97}{32}J^2$, rank-2 codes, $d = 3$	lemma of the refined CZ bound

1.15652 and reproduces, without fitting, both the width dependence of a simple discretization (1.1428, 1.1482, 1.1523 for $\sigma = 24, 40, 80$) and the values 1.1552–1.1565 of Qiskit's convention (`identidad_calibracion.py`). The check covers all qubits. So `omegad` is consistent with a drive whose peak is $(1 - e^{-2})|\mathrm{amp}|$ in units of `omegad`; under Qiskit's normalization the model would over-rotate by 1.1565. The snapshots do not say which waveform reaches the qubit. The full-amplitude ratio below uses `omegad` only and does not depend on this; the peak-drive figures do, and we give both readings.

The same check exposes a property of the data: for equal σ, the product `omegad`·amp is identical to 9 digits across 27 backends. Hence `omegad` is *not* an independent measurement; IBM derives it from the π-pulse calibration. Any correlation between gate duration and `omegad` is therefore partly circular (Section IV).

D. Exclusions and falsifier

The pre-specified falsifier was a ratio $\tau_X/\tau_{\mathrm{QSL}} < 1$ on any qubit, impossible for an exact X under a correct convention (an approximate gate with trace error e allows down to $(2/\pi)\arcsin(1-e)$, i.e. 0.991 for $e = 10^{-4}$). It fired: 254 qubits came out faster than their own limit. They belong to the snapshots of `brussels`, `prague` and `strasbourg`, whose sample time $dt \ge 1$ ns differs from that of every other snapshot (0.22 ns) and suggests placeholder values; we have no documentation of their provenance. These backends are excluded, as are qubits reporting zero anharmonicity; the exclusion was decided after the falsifier fired, and we report it as such. Controls: a hand computation on `armonk` reproduces the script (ratio 2.631), and changing the unit of `omegad` moves the median by exactly 2π (`barrido.py`). One time-series snapshot (`toronto`, 2021-03-15, median two-qubit duration 6315 ns, ratio ~ 40) is an outlier, probably a degraded calibration; it is kept in Table IV, and removing it changes the 2021 two-qubit median from 2.78 to 2.74 and the single-qubit one not at all (`toronto_sensibilidad.py`).

TABLE II
SINGLE-QUBIT GAP, IBM SNAPSHOTS (`QISKIT-IBM-RUNTIME` 0.51.0; 41 BACKENDS, 1646 QUBITS). $\tau_{\mathrm{QSL}} = \pi/$`OMEGAD` FOR X (FULL DRIVE AMPLITUDE, $|D| = 1$).

quantity	median	percentiles
$\tau_X/\tau_{\mathrm{QSL}}$ (5th–95th)	11.30	[7.19, 15.54]
$\tau_{\sqrt{X}}/\tau_{\mathrm{QSL}}(\sqrt{X})$	22.61	—
$\tau_X/(2\pi/\|\delta\|)$ (5th–95th)	15.45	[10.90, 18.70]
$\Omega_{\mathrm{peak}}/\|\delta\|$ (750 qubits, 10th–90th)	0.0907	[0.0607, 0.0992]

IV. SINGLE-QUBIT GATES

A. The gap

Over 41 backends and 1646 qubits (`barrido.py`, output `barrido.csv`), the X gate lasts a median $\tau_X/\tau_{\mathrm{QSL}} = 11.30$ times its two-level speed limit at full drive amplitude ($|D| = 1$, $\tau_{\mathrm{QSL}} = \pi/$`omegad`), 5th–95th percentile $[7.19, 15.54]$, minimum 2.37 (Table II). Qubits of one backend share a duration, so they are not independent: the median of the 41 per-backend medians is 11.37, and a bootstrap that resamples whole backends gives a 95% interval $[11.02, 11.65]$ for the median (`mediana_por_backend.py`). IBM does not drive at that amplitude; Section IV-C separates the part of the gap set by the drive actually used from the part set by amplitude headroom. The $\sqrt{X}$ gate has the same duration and half the rotation, so its median ratio is 22.61. Against the leakage time scale $2\pi/|\delta|$, the X gate lasts a median 15.45 periods $[10.90, 18.70]$.

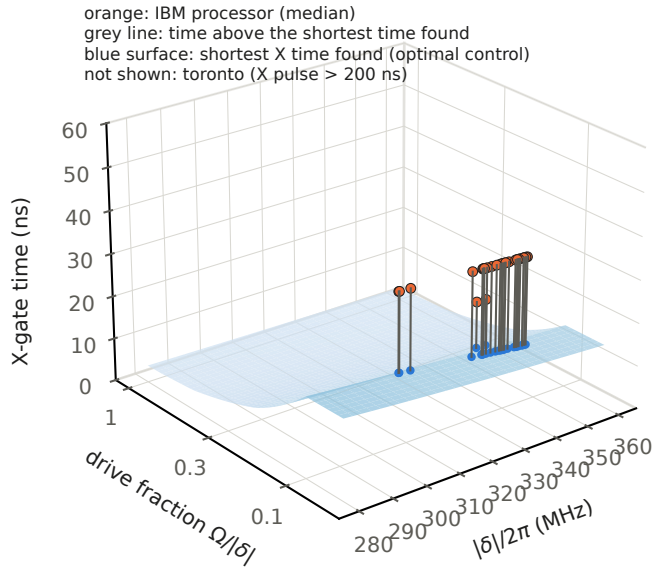


Fig. 1. Where IBM single-qubit gates sit relative to the model. Blue surface: shortest X-gate time found by optimal control on a 4-level transmon (Table V) as a function of the anharmonicity and of the drive fraction $\Omega/|\delta|$. Orange: one point per IBM processor with calibrated pulses (median anharmonicity, median drive fraction $\Omega_{\mathrm{peak}}/|\delta|$, X duration); the grey line is the time above the shortest time found at the same drive fraction. `toronto` (2560-sample X pulse) is not shown but is included in the medians. The surface for $r \geq 0.8$ uses non-converged 4-level values. Section VII. Interactive version.

B. Hypotheses written before the sweep

Before computing we wrote three rivals with predicted ranges (`ficha_ideas.toml`): H1, amplitude-limited gates at 1–2× the limit; H2, leakage-limited gates, $\tau_X \propto 1/|\delta|$, ratio 2–6, uncorrelated with `omegad`; H3, a convention artefact. Outcome: H1 is rejected against the full-amplitude limit (the gates are at $\approx 11\times$); against the peak drive actually used the ratio is ≈ 2.16 (Section IV-C), just above H1's range, which did not specify the drive; with the peak $(1-e^{-2})|\mathrm{amp}|$ consistent with `omegad` (Section III) it is 1.87, inside that range, so this comparison depends on the reading, while the full-amplitude one does not; H2 has no support (R^2 of $\log \tau_X$ on $\log(1/|\delta|)$ is 0.087, and the gates last $\approx$ 15 leakage periods); the factor-2 and factor-2π conventions of H3 are excluded by the π-pulse check, and the remaining 15% is the lifted-Gaussian normalization (Section III). The R^2 on $\log(1/\mathtt{omegad})$ is 0.692 (shuffled control 0.001), but because `omegad` is derived from the same pulse's amplitude and width (Section III) this is partly circular and is not independent physical evidence.

C. What sets the duration

Writing $\tau_X/\tau_{\mathrm{QSL}} = \mathrm{duration} \cdot dt \cdot \mathtt{omegad}/\pi$, the gap factors into ≈ 2.16—the duration over A_{rec}, i.e. the pulse-shape factor 1.87 (a 4σ Gaussian with its edge removed against a square pulse of the same peak) times the normalization $1/(1-e^{-2})$ of Section III—times $1/\mathrm{amp} \approx 5$. Since $\mathtt{omegad} \cdot |\mathrm{amp}| \cdot A_{\mathrm{rec}} = \pi$, the ratio is exactly $\tau_X/\tau_{\mathrm{QSL}} = \mathrm{duration} \cdot dt/(|\mathrm{amp}|A_{\mathrm{rec}}) \approx 2.16/|\mathrm{amp}|$: with a nearly fixed envelope, the spread of the ratio is the spread of the calibrated amplitudes (IBM calibrates the π pulse at $\mathrm{amp} \approx 0.2$: an amplitude reserve). Neither the speed limit nor the anharmonicity fixes the duration; engineering choices do: a duration fixed in samples per backend and a low amplitude.

The calibrated peak Rabi frequency sits at a nearly fixed fraction of the anharmonicity: over 750 qubits, $\Omega_{\mathrm{peak}}/|\delta|$ has median 0.0907 and 10th–90th percentile $[0.0607, 0.0992]$ (`cruce_formulas.py`), i.e. a cap near 0.1, taking the peak as $\mathtt{omegad}\,|\mathrm{amp}|$; with the peak $(1-e^{-2})\,\mathtt{omegad}\,|\mathrm{amp}|$ consistent with `omegad` (Section III) the median is 0.078. We read this as IBM's design rule, not a physical limit.

The headline with its denominators. The median 11.3 is measured against $\pi/\Omega_{\max}$ with $\Omega_{\max} = \mathtt{omegad}$, the drive at full amplitude $|D| = 1$. The calibrated peak drive is $\Omega_{\mathrm{peak}} \approx 0.09|\delta|$; the factor ≈ 5 between that drive and full amplitude is headroom chosen by design, and at the drive actually used the gap is the shape factor 1.87 if the peak is $(1-e^{-2})\,\mathtt{omegad}\,|\mathrm{amp}|$, or ≈ 2.16 if it is $\mathtt{omegad}\,|\mathrm{amp}|$ (Section III). Optimal control on IBM's own model (Section VII, Table V) bounds the gain at the same peak drive: since $T \geq 2\arcsin(1-e)/\Omega_{\mathrm{peak}}$, a gate with $T\Omega_{\mathrm{peak}}/\pi = R_{\mathrm{peak}}$ can be shortened by at most $R_{\mathrm{peak}}/0.991$, i.e. 2.18 with the nominal peak or 1.89 with the edge-subtracted one (at the slightly higher cap $r = 0.1$ the fastest pulse found, ≈ 16 ns against 35.6 ns, gives 2.2); and, within the full amplitude, a median gate of $11.3\,\tau_{\mathrm{QSL}}$ against a minimum between $0.99\,\tau_{\mathrm{QSL}}$ (bound) and $1.12\,\tau_{\mathrm{QSL}}$ (found at the median $r_{\mathrm{full}} = 0.351$, Section VII), i.e. about 10–11 times; for the Falcon example (35.6 ns) at that amplitude, 7.4–8.3 times. A gate of ≈ 3.4 ns needs about three times that amplitude. In short: IBM runs the drive at about a tenth of the anharmonicity; at that drive its pulses are about 2 times the square-pulse limit at that drive and can be shortened by at most 2.18 (nominal peak) or 1.89 (edge-subtracted peak) at the same peak drive; within the full amplitude the model admits gates about 10 times faster than the median. The figure 11.3 should not be read as "gates are 11 times slower than possible"; those $\approx 10\times$ are a model prediction (Section IX).

Per-qubit durations. Since all qubits of a backend share one duration, one may ask how much would be gained by giving each qubit the shortest duration compatible with that cap (drive fraction 0.099, the observed 90th percentile), keeping the pulse shape (`mejora_por_cubit.py`, 750 qubits on 31 backends). The median per-qubit gain is only $\times 1.09$ ($\times 1.07$ if a layer waits for its slowest qubit): the cap binds almost every qubit alike. The backend `cairo` is a natural experiment: its common X duration is 21.3 ns where the cap would require a median of 31.6 ns, i.e. it runs at a drive fraction near 0.15, and works.

V. Two-Qubit Gates

A. Pairs

From the snapshots we extract every coupled pair with a coupling j_q in the Hamiltonian variables and a calibrated two-qubit gate (`pares.py`): 1758 pairs in 43 backends (1019 ECR, 739 CX). The median coupling is $j_q = 0.01268$ rad/ns

TABLE III
TWO-QUBIT GAP (1758 PAIRS, 43 BACKENDS; 1019 ECR, 739 CX). QUBIT SUBSPACE, $\|H_{\rm int}\| = j_q$.

bound	source	median ratio
CNOT-class gate, $\pi/(2j_q)$	[9], [5] (not in Lean)	4.22
Bell preparation, $\sqrt{2-\sqrt{2}}/j_q$	Lean, chord form	8.7
Bell preparation, $\pi/(4j_q)$	Lean, sharp	2× row 1

($g/2\pi \approx 2.0$ MHz), the median detuning 115 MHz ($\approx 60\, j_q$), and the median gate duration 548 ns (5th–95th percentile $[277, 698]$ ns).

B. Which bound

In the qubit subspace the IBM coupling is $j_q(\sigma_i^+\sigma_j^- + \text{h.c.}) = (j_q/2)(XX+YY)$, of operator norm j_q. Two bounds apply, and they answer different questions.

Literature bound for the gate. With unbounded local control, no leakage, and an XY coupling, the minimal time of a CNOT-class gate (CNOT, CZ, ECR) is $\pi/(2j_q)$. This follows from Howard *et al.* [9], Appendix C ("$T_{\min} = \pi/(4g)$ holds for H_0, H_0^{XY} and H_0^{XXZ}," with $g_H = j_q/2$ in their normalization) and from the majorization theory of Vidal–Hammerer–Cirac [5]. The detuning is a local term absorbed by the unbounded local control, and bounded drives are a subset of unbounded ones, so within the qubit subspace the real duration cannot beat it [our inference, not stated in that form in the sources].

Our certified bound for state preparation. The Lean theorem of Section II bounds the time to *create* a Bell state from a product state: $\pi/(4J)$, attainable. A full CNOT on an XY coupling needs twice that, which is why the literature bound is the stronger one for gates. Ours is the one that is machine-checked, and it survives extra transmon levels: with $|2\rangle$ included the norm of the exchange term grows ($\|a^\dagger b + ab^\dagger\| = 1, 2, 3.4641$ for $d = 2, 3, 4$), and the CNOT bound becomes $\pi/(4J)$ with that larger J. With leakage levels allowed, the only rigorous constant in the literature is for iSWAP [7]; for CNOT/ECR with $|2\rangle$ no constant was found in the sources read. Ours, $\pi/(8j_q)$ for $d = 3$, is not tight; Section VII sharpens it to $0.4511/j_q$ and, in version 3, to $0.4632/j_q$ for the CZ.

C. Gaps

Table III gives both ratios. Against the literature bound $\pi/(2j_q)$ the median of the per-pair ratios is 4.22 (the ratio of the median duration to the median bound would be 4.4). Against our certified chord constant $\sqrt{2-\sqrt{2}}/j_q$ the median ratio is 8.7; the sharp Lean constant $\pi/4$ lowers it by 2.6% to 8.4, and since $\pi/(4j_q)$ is exactly half of $\pi/(2j_q)$ the certified ratio is exactly twice the literature ratio. Both are correct; they bound different tasks. Both assume $\|H_{\rm int}\| = j_q$, i.e. the qubit subspace. With the second transmon level the norm grows to $2j_q$, the certified time bound halves to $\pi/(8j_q)$ and the certified ratio doubles to ≈ 16.9: the bound becomes weaker, which says nothing about how much of that room is reachable.

Numerical controls of the certified bound (`control_dos_cubits.py`): the Duhamel inequality has 0 violations in 20 000 random pairs, with maximum ratio 0.99998 (the constant 1 is sharp), and the bound has 0 violations in 3000 evolutions with random local control.

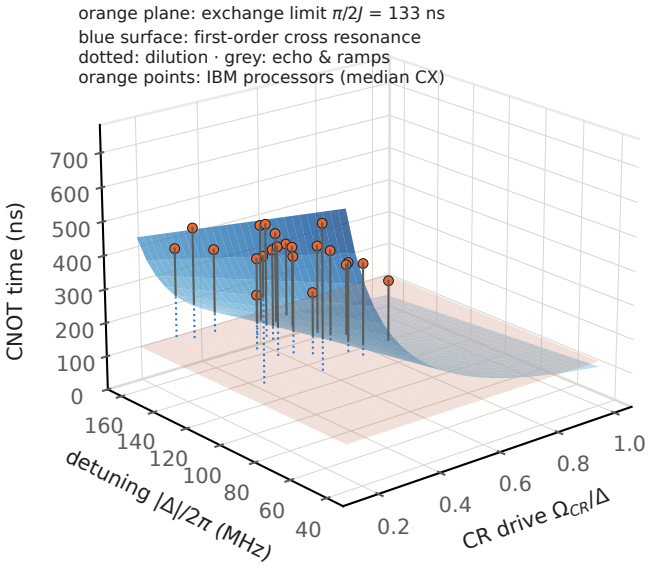


Fig. 2. Two gaps in the two-qubit gate. Orange plane: exchange speed limit $\pi/(2J) = 133$ ns at the median coupling of the 29 processors with pulse data ($J = 0.0118$ rad/ns; the median over all 1758 pairs of Section V is 0.01268). Blue surface: CNOT time allowed by first-order cross resonance, $t = (\pi/2)/\nu_{ZX}$, as a function of the drive fraction Ω_{CR}/Δ and the detuning. Orange points: one per IBM processor with calibrated CX pulses (median drive fraction, detuning and CX duration). Dotted: dilution by the cross-resonance mechanism (plane to surface); grey: the remainder (surface to point), from echo pulses and ramps and from the error of the first-order model. Median J and anharmonicity. Interactive version.

D. Cross-resonance against its first-order formula

IBM's CX is built from echoed cross-resonance (CR) pulses. To first order the CR drive of amplitude Ω on the control, at detuning Δ, produces a ZX rate $\nu_{ZX} = -(J\Omega/\Delta)\,\delta_c/(\delta_c+\Delta)$ (the coefficient of $ZX/2$ to first order in J and Ω, [18, Eq. (4.26)], with $\Delta = \omega_c - \omega_t$ and $H = (\nu_{ZX}/2)\,ZX$), and a CNOT needs $\theta_{ZX} = \nu_{ZX} t_{\rm CR} = \pi/2$. On 1212 calibrated CX gates in 29 backends (`cruce_formulas.py`, pulse definitions of `qiskit-terra` 0.24.1) the median of $\theta/(\pi/2)$ is 1.109, 10th–90th percentile $[0.870, 1.472]$, with 59% in $[0.8, 1.2]$. The median lies 11% above the target, with a wide spread. The calibrated drive sits at a nearly fixed fraction of the detuning: $|\Omega_{\rm CR}/\Delta|$ has median 0.451.

The two gaps share one pattern: the control runs at a fixed fraction of an energy splitting it must not excite—$\Omega_{\rm peak} \approx 0.09|\delta|$ for one qubit, $\Omega_{\rm CR} \approx 0.45\Delta$ for two. Caveats: the formula is first order (no ZZ, IX or classical crosstalk), ramp areas use Qiskit's "lifted" Gaussian-square convention, and the angles have not been checked against a full simulation. The calibrated drives lie outside the stated range of the perturbative expansion ($\Omega/\Delta \ll 1$; 53% of the gates have $\Omega/2\pi > 50$ MHz). With the third-order term of [18, Eq. (4.25)], integrated over the actual pulse shape, the median falls to 0.898 ($[0.659, 1.105]$); a numerical block

TABLE IV
MEDIAN GAPS BY YEAR (56 USABLE SNAPSHOTS, 41 BACKENDS). TWO-QUBIT RATIO AGAINST $\pi/(2j_q)$.

	2021	2022	2023	2024	2025
$\tau_X/\tau_{\mathrm{QSL}}$	11.28	11.32	11.21	11.10	11.87
$\tau_{2q}/(\pi/2j_q)$	2.78	2.64	4.32	3.89	4.26
snapshots	27	4	2	16	7

diagonalization (the least-action method of [18, App. A1], four levels per transmon) gives 0.967 ($[0.795, 1.159]$, 81% in $[0.8, 1.2]$) (`fuentes/cruce_mg2020.py`). Flipping the sign of Δ moves the median away from 1 (0.664), as a control should. One convention is unresolved: in the Qiskit backend Hamiltonians the control channel of the pair (c, t) is weighted by the drive strength of the *target* ($\mathtt{omegad}_t$), whereas we use that of the control; with the former the first-order median is 1.089.

VI. TIME SERIES 2021–2025

Snapshots across Qiskit releases (Section III; `serie_temporal.py`, `serie_porque.py`, `duracion_error.py`) give the evolution of both gaps (Table IV).

Single-qubit: frozen. The ratio to the full-amplitude limit stays at ≈ 11 from 2021 to 2025: a design constant, consistent with the engineering reading of Section IV-C.

Two-qubit: duration, not the archived coupling. The ratio grew from 2.78 (2021) to 4.26 (2025). The archived coupling did not change: $g/2\pi \approx 1.9$–2.1 MHz in every processor family, and within a processor j_q is *identical* at every date—a model value carried over, not re-measured. The change is the duration: Eagle ECR gates (median 562 ns) against Falcon CX gates (≈ 377 ns), and recalibrations lengthen gates and never shorten them in the examples examined (`bogota` 366 → 544 ns, `perth` 380 → 491 ns, `lagos` 299 → 452 ns). The 2022 and 2023 medians rest on 4 and 2 snapshots respectively (Table IV; source: `serie_temporal.py`). The trend describes the snapshots in the series, which are all pre-Heron: it reflects the transition from Falcon CX to Eagle ECR gates, not the whole fleet. The current Heron generation, with CZ gates of $\approx$ 68–70 ns (68 ns median in our `torino` snapshot; 70 ns measured on `ibm_torino` in [12]), is not in the series, because its snapshots carry no Hamiltonian (Section III).

No sign that lengthening lowers the error. Across the 80 snapshots with two-qubit durations and errors, Spearman's correlation between median two-qubit duration and median gate error is $+0.13$ ($p = 0.23$), which does not exclude a benefit. For the same pair lengthened between two dates (71 cases) the error rose (median change in log error $+0.21$; it fell in only 31%); for pairs shortened (40 cases) it fell. These comparisons are observational: other calibration changes and drift happen at the same dates.

A methodological lesson. We then asked whether lengthened pairs had worse qubits. Pooled across dates, T_1 of lengthened pairs fell more than that of unchanged pairs (Mann–Whitney $p = 0.0017$; `degradacion.py`). Under a stratified control (same backend, same dates; 18 strata, 84 lengthened pairs) the effect vanished (Wilcoxon $p = 0.90$ for T_1, $p = 0.99$ for T_2); a pooled frequency-shift effect ($p = 4 \cdot 10^{-12}$) vanished likewise (`frecuencia_degradados.py`). The pooled effects are consistent with confounding by epoch: lengthenings concentrate in backends and periods in which the whole chip's T_1 fell. We retracted the reading that lengthened pairs had degraded qubits: after controlling for backend and date the data do not support pair-specific degradation. This does not show what does drive lengthening. Any cross-date effect in calibration snapshots should be tested stratified by backend and date before it is read.

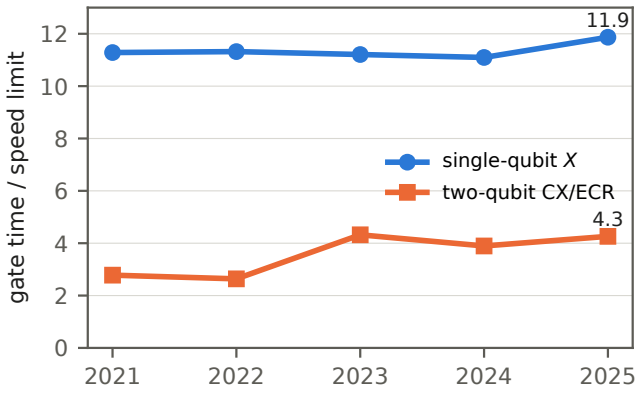


Fig. 3. Per-year median of the snapshot medians: single-qubit ratio frozen near 11, two-qubit ratio growing with the gate duration. Interactive version, split by gate or processor size and followed processor by processor.

VII. HOW FAR THE MODEL ALLOWS

Metrics and controls. In all three problems the error is $1 - F$ with $F = |\mathrm{Tr}(V^\dagger PUP)|/m$, V the target on the m-dimensional qubit subspace, P the projector on it and U the full propagator; F ignores the global phase, and leakage lowers it because PUP is then sub-unitary. The leakage reported is $1 - \|PUP\|_F^2/m$, the average over computational inputs. The control sets differ: for X, one complex dipole drive $|u| \le r|\delta|$ plus free virtual Z rotations before and after; for the CNOT, a bounded cross-resonance drive on the control, a weak drive on the target and free single-qubit rotations before and after; for the CZ, arbitrary local Hermitian generators on each transmon, bounded by $a_{\max}$ during the gate, plus arbitrary instantaneous local unitaries before and after. The last set is much wider than a microwave drive on a Duffing transmon.

A. Single qubit

We ran optimal control (`grape_1q.py`) on a 4-level Duffing transmon in units $|\delta| = 1$, with a piecewise-constant complex drive bounded by $|\Omega| \le r|\delta|$, and searched, on a grid of times, for the shortest time at which an X gate with error $< 10^{-4}$ including leakage is found. The two-level control returns $T = \pi/\Omega_{\max}$ exactly. The results (Table V) follow

$$T_{\mathrm{found}} \approx \max\big(\pi/\Omega_{\max},\ \sim 1.1 \cdot 2\pi/|\delta|\big), \qquad (1)$$

TABLE V
SHORTEST X-GATE TIME FOUND BY OPTIMAL CONTROL (AN UPPER BOUND ON THE MINIMUM), 4-LEVEL TRANSMON, ERROR $< 10^{-4}$ WITH LEAKAGE, UNITS $|\delta| = 1$ (`GRAPE_1Q.PY`).

$r = \Omega_{\max}/\lvert\delta\rvert$	T_{found}	$T_{\text{found}}/(\pi/\Omega_{\max})$
0.1	33.0	1.05
0.2	16.5	1.05
0.3	12.0	1.15
0.5	8.2	1.30
0.8	7.1	—
1.0	6.9	—
1.5	7.3	—

a speed-limit branch for $r \lesssim 0.3$ and a leakage-dominated branch for $r \gtrsim 0.8$. Equation (1) is an empirical fit to a coarse grid, not a theorem; it misses the values found at $r = 0.3$–0.5 by 13–16%, where neither branch dominates. The entries are feasible times found, i.e. upper bounds on the true minimum: a larger cap admits every pulse of a smaller one, so the true minimum cannot increase with r, and the value at $r = 1.5$ above that at $r = 1.0$ shows that the search is local. A lower bound holds for any number of levels. With $H(t) = H_{\text{diag}}(t) + \frac{1}{2}(u\,b^\dagger + u^* b)$, $|u| \le \Omega_{\max}$, and $p_0 = |a_0|^2$ the population of $|0\rangle$, the only way out of $|0\rangle$ is through $|1\rangle$, so $|\dot{p}_0| \le |u|\,|a_0||a_1| \le |u|\sqrt{p_0(1-p_0)}$ and the angle $\theta = \arccos\sqrt{p_0}$ obeys $|\dot{\theta}| \le |u|/2$. An X gate takes θ from 0 to $\pi/2$, hence $T \ge \pi/\Omega_{\max}$ for any truncation; with a transfer fidelity $|\langle 1|\psi_T\rangle|^2 \ge 1 - \varepsilon$, $T \ge 2\arccos\sqrt{\varepsilon}/\Omega_{\max}$. (This short proof is ours and not formalized; the two-level case is `QSL.Gate.two_arccos_sqrt_le_time_approx`.) In the trace metric of Table V the same argument bounds $|U_{10}|, |U_{01}| \le \sin(\Omega_{\max}T/2)$ for $T \le \pi/\Omega_{\max}$, so an error e needs $T \ge 2\arcsin(1-e)/\Omega_{\max}$; for $r \le 0.2$ the times found are 6% above it ($e = 10^{-4}$), so there the minimum is known to 6%; at the cap $\Omega_{\max} = 0.1|\delta|$ ($|\delta| = 2.07$ rad/ns), slightly above the calibrated peak drive and therefore not the same-peak comparison of Section IV-C, the archived 35.6 ns lies between 2.2 and 2.4 times the true minimum (bound 15.0 ns for that tolerance, pulse found 15.9 ns).

For IBM's typical $|\delta| \approx 2.07$ rad/ns, the fastest pulse found at IBM's own cap $r = 0.1$ takes ≈ 16 ns, against 35.6 ns for Falcon single-qubit gates (e.g. `jakarta`): a factor ≈ 2.2 at that cap, which is slightly above the calibrated peak drive (median 0.091). At $r = 1$ the 4-level search finds $T = 6.9/|\delta|$ (3.3 ns), but the truncation matters at this drive: the 4-level pulse at $T = 7.0/|\delta|$ keeps $1 - F = 7 \cdot 10^{-9}$ in 4 levels but degrades to 5.5–$7.2 \cdot 10^{-4}$ with 5 to 12, and the Duhamel bound on the distance to a larger model, $\eta_4 = \int |u|\,\|\langle 3|UP\|\,dt = 0.55$, does not certify it. A pulse optimized with 6 levels at the same time reaches $1 - F = 4.0 \cdot 10^{-6}$ there and $3.6 \cdot 10^{-6}$ with 8 and 12 levels (`truncacion_1q.py`), so a gate of ≈ 3.4 ns ($T = 7.0/|\delta|$), more than ten times shorter than IBM's, survives in the converged model; it is a numerical witness in an idealized model, not a physical floor. The table entries for $r \ge 0.8$ are 4-level values and are not converged in the number of levels. At IBM's cap the 4-level pulse is unaffected by more levels ($\eta_4 = 0.01$). The drive fraction $r = 1$ is not available in the analysed snapshots: the full amplitude $|D| = 1$ corresponds

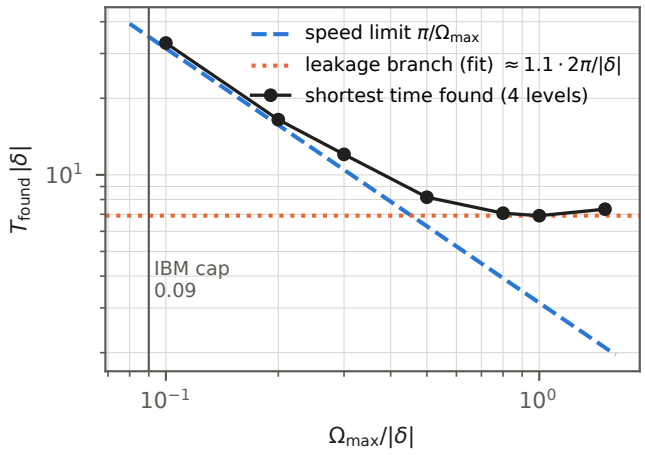


Fig. 4. Shortest X-gate time found by optimal control on a 4-level transmon (error $< 10^{-4}$, leakage included) against the drive bound: speed-limit branch $\pi/\Omega_{\max}$ and leakage-dominated branch $\approx 1.1 \cdot 2\pi/|\delta|$ (empirical fit). The vertical line marks the cap observed in IBM calibrations. Interactive version.

to $r_{\text{full}} =$ `omegad`$/|\delta|$, median 0.351 (5th–95th percentile $[0.24, 0.61]$; `tercera_revision.py`). At $r = 0.351$ a 6-level search finds an X at $T = 10.0/|\delta| = 4.8$ ns with $1 - F = 2.8 \cdot 10^{-8}$, unchanged with 12 levels, while the bound above gives $T \ge 8.87/|\delta| = 4.3$ ns; so within the amplitude budget of the median device a 35.6 ns Falcon gate is 7.4 times the fastest gate found and at most 8.3 times the bound, i.e. 7.4–8.3 times the minimum of the model, and the median gate ($11.3\,\tau_{\text{QSL}}$, with $\tau_{\text{QSL}} = 8.95/|\delta|$ at $r = 0.351$) is 10–11 times. As an independent point of comparison, a 4.16 ns single-qubit pulse on a superconducting transmon, with 99.76% fidelity and 0.044% leakage, has been demonstrated by optimal control [15]; its elementary pulses are $\pm X/2$ and $\pm Y/2$ characterized by Clifford-based protocols, a different task from our simulated X gate, so it gives context, not validation. Limits: the time grid is coarse, the search is local, and there is no noise or line distortion in the model.

B. Two qubits

Cross-resonance. We optimized a CNOT on two 3-level transmons with a cross-resonance drive bounded by $|\Omega_{\text{CR}}| \le r\Delta$ at IBM median parameters ($J/\Delta = 0.0176$, $\delta/\Delta = -2.7$, 10 segments; `grape_cr.py`). In units of the exchange time $\pi/(2J)$ the shortest feasible times found (error $< 10^{-3}$, upper bounds on the minimum) are 3.27 ($r = 0.25$), 1.82 ($r = 0.45$, IBM's calibrated fraction) and 1.50 ($r = 0.7$). Since a larger cap admits every pulse of a smaller one, 1.50 is also feasible for every $r \ge 0.7$; independent searches at $r = 1.0$ and 1.5 returned longer times, a failure of the local search, which we confirmed by re-expressing the $r = 0.7$ pulse under the larger caps (identical error) and continuing from it to shorter times, which did not reach 10^{-3} down to 1.2 (`grape_cr_continuacion.py`). The searches use coarse time grids and few random starts; a repeated run at $r = 0.7$, $T = 1.50$ reached only $1.2 \cdot 10^{-3}$. What the data support is that at IBM's fraction the fastest CNOT found is ≈ 1.2 times

TABLE VI
CZ ON THE QUBIT SUBSPACE OF TWO EXCHANGE-COUPLED 3-LEVEL TRANSMONS: BEST ERROR $1-F$ AGAINST GATE TIME T AND THE BOUND ON EACH LOCAL CONTROL $\|A_k\|, \|B_k\| \le a_{\max}$ (GRAPE_CZ_TORCH.PY, VERIFICAR_CZ.PY). QUBIT-ONLY LIMIT $\pi/(2J) = 1.571/J$; LOWER BOUNDS $\pi/(8J) = 0.393/J$ (CERTIFIED) AND $0.4511/J$ (REFINED, SECTION VII), $0.4632/J$ (TWO COPIES, VERSION 3); EXPLICIT CONSTRUCTION $\pi/(6J) = 0.524/J$, CERTIFIED PROTOCOL $0.5200/J$ (VERSION 3). EACH CELL IS THE BEST OF UP TO 12 (NO CAP) OR 8 (CAPPED) RANDOM STARTS; DIFFERENCES BELOW 10^{-8} BETWEEN COLUMNS ARE OPTIMIZER NOISE.

$T\,J$	none	$40J$	$20J$	$10J$
0.40	$1.6 \cdot 10^{-2}$			
0.45	$5.0 \cdot 10^{-3}$	$5.3 \cdot 10^{-3}$	$5.5 \cdot 10^{-3}$	
0.50	$3.7 \cdot 10^{-4}$	$4.2 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$1.3 \cdot 10^{-3}$
0.55	$7 \cdot 10^{-12}$	$8 \cdot 10^{-13}$	$2 \cdot 10^{-11}$	$2 \cdot 10^{-8}$
0.60	$3 \cdot 10^{-10}$			$1 \cdot 10^{-9}$

slower than at $r = 0.7$; they do not establish an optimum in r.

CZ with the second level. Two 3-level transmons with a fixed exchange coupling $J(b_1^\dagger b_2 + \text{h.c.})$ and fast local control (any Hermitian A_k, B_k on each transmon, 8 piecewise-constant segments, exact gradients; `grape_cz_torch.py`). The target is CZ on the qubit subspace with no leakage; the action outside it is free, as in [7]. Every solution is recomputed independently in float64 (`verificar_cz.py`). As a control, the qubit-only search reproduces the known time $\pi/(2J)$ [9] (a gate at $\pi/(2J)$, none at $0.89\,\pi/(2J)$). With the second level: *Bounds.* Write T_* for the minimal time of an exact CZ with ideal (unbounded, instantaneous) local control, and $T_*(e)$ for the minimal time of a CZ with error at most e in the trace metric used throughout. CZ maps $|{+}{+}\rangle$ to a maximally entangled state, so the entangling-time theorem of Section II with $\|H_c\| = 2J$ ($d = 3$) gives $T_* \ge \pi/(8J) = 0.393/J$. For a gate with error e, add a two-level reference to each transmon and start from the product (across $AA'{:}BB'$) of two local maximally entangled states: the overlap of the resulting state with the one produced by CZ is exactly $|\mathrm{Tr}(CZ^\dagger PUP)|/4 = 1-e$, leakage included, and the CZ state has Schmidt coefficients $(1/\sqrt{2}, 1/\sqrt{2})$ across that cut, while the references leave $\|H_c\| = 2J$ unchanged. The tolerance version of the theorem (`QSL.Qudit.pi_div_four_sub_le_entangling_time_approx`, with $\varepsilon = 1-(1-e)^2$) then gives $T_*(e) \ge [\pi/4 - \arccos(1-e)]/(2J)$ (this instantiation with references is ours, not formalized; `tercera_revision.py`). Preparing the single output $CZ|{+}{+}\rangle$ is a much easier task: local operations map $|{+}{+}\rangle$ to $|12\rangle$, on $\{|12\rangle, |21\rangle\}$ the coupling acts as $2J\sigma_x$, and local operations map the result back, so the state-preparation bound is attained exactly: it is a valid but weak bound for the gate. An explicit construction gives an upper bound. Writing $H_c = \frac{J}{2}(Q\otimes Q + \Pi\otimes\Pi)$ with $Q = b + b^\dagger$ and $\Pi = i(b^\dagger - b)$, fast conjugations of *one* transmon by $W = V_Q\,\mathrm{diag}(1, e^{2\pi i/3}, e^{4\pi i/3})V_Q^\dagger$ (V_Q diagonalizing Q) average $\Pi\otimes\Pi$ away and keep $\frac{J}{2}Q\otimes Q$; encoding each qubit in the two extreme eigenvectors of Q (eigenvalues $\pm\sqrt{3}$) gives $\frac{3J}{2}Z\otimes Z$, hence a CZ (up to local phases) in interaction time $\pi/(6J) = 0.5236/J$, in the limit of arbitrarily fast local operations (`cruces_area_cz.py`). The gate bound itself can be sharpened, because the gate must act coherently on all logical inputs (the lemma is certified in Lean as `QSL.Exchange.refined_cz_lemma`, with its equality case, and checked numerically in `cota_cz_refinada.py`; the time argument built on it is not formalized). Let $L(t)$ apply the same local controls as $U(t)$ but without H_c, and $z(t) = U(t)^\dagger L(t)|\Omega_0\rangle$ with $|\Omega_0\rangle$ the reference-extended product state above. Then $\dot z = iU^\dagger H_c L|\Omega_0\rangle$, whose Fubini–Study speed is the spread of H_c in $L(t)|\Omega_0\rangle$; that state has physical reduced state $(P_A\otimes P_B)/4$ with P_A, P_B rank-2 projectors, and for any such projectors $\frac{1}{4}\mathrm{Tr}[(P_A\otimes P_B)H_c^2] \le \frac{97}{32}J^2$ (equality at $P = \mathbb{1} - |v\rangle\langle v|$, $|v\rangle = \sqrt{7/8}|0\rangle + \sqrt{1/8}|2\rangle$). Since $L(T)|\Omega_0\rangle$ is a product across $AA'{:}BB'$ and the CZ state is at Fubini–Study distance at least $\pi/4$ from every product, $T_*(e) \ge [\pi/4 - \arccos(1-e)]/(J\sqrt{97/32})$, i.e. $T_* \ge 0.4511/J$ and $T_*(2\cdot 10^{-8}) \ge 0.4510/J$, 15% above the norm bound. With ideal local control, $0.4511/J \le T_* \le \pi/(6J) = 0.5236/J$ (the upper value as a limit of finite protocols); the construction is within 16% of the bound. Preparing one Bell state and implementing the whole CZ therefore have different minimal times in this model. *Version 3: both ends improved.* Apply two identical copies of a protocol to $\chi_0 = \psi^-_{A_1A_2}\otimes\psi^-_{B_1B_2}$, the singlets of the qubit subspace. Local controls preserve the symmetric and antisymmetric subspaces of each pair, and the copies are exchanged by a symmetry, so the state stays in $\wedge^2\otimes\wedge^2 \oplus \mathrm{Sym}^2\otimes\mathrm{Sym}^2$; an exact CZ must carry it from the first sector to the second, at a rate at most $\|K\|$, with K the block of $H_c^{(1)} + H_c^{(2)}$ between them. In exact arithmetic $\|K\|^2 = \frac{23}{2}J^2$ (`cz_v3_norma_K.py`), so $T_* \ge \pi/(\sqrt{46}\,J) = 0.4632/J$. In the other direction, the average of 18 local conjugates of H_c keeps, besides $\frac{J}{2}Q\otimes Q$, a three-state block $\{|\ell\ell\rangle, |mm\rangle, |hh\rangle\}$ in the eigenbasis of $Q = b + b^\dagger$; with a local detuning of $|m\rangle$ the block returns exactly at $T_2 = 3\pi/\big(2(1+\sqrt{65})J\big) = 0.5200016/J$ and the code undergoes a CZ up to local phases. A second-order product of 64 steps followed by an exact correction is certified by a Krawczyk test in interval arithmetic, $T_* \le 0.520023/J$ (`cz_v3_certificado_salida.txt`; two negative controls are rejected). Hence $0.4632/J \le T_* \le 0.5200/J$, between 0.295 and 0.331 of the qubit-only time.

Larger local dimension. The same construction works for d-level transmons: encoding each qubit in the two extreme eigenvectors of $Q_d = b + b^\dagger$ (eigenvalues $\pm\lambda_d$, with $\lambda_d/\sqrt{2}$ the largest zero of the Hermite polynomial H_d) gives a CZ in interaction time $\pi/(2J\lambda_d^2)$, while the entangling-time theorem gives $\pi/(4J\kappa_d)$ with $\kappa_d = \|b^\dagger\otimes b + b\otimes b^\dagger\|$. The ratio of the two bounds is $2\kappa_d/\lambda_d^2$: 2, 1.33, 1.27, 1.20, 1.09, 1.05 for $d = 2, 3, 4, 5, 10, 20$, and it tends to 1. This is certified in Lean: $\lambda_d^2/2 \le \kappa_d \le 2(d-1)$ (`QSL.Exchange.exchange_norm_ge`, `QSL.Exchange.exchange_norm_le`), $\lambda_d \ge 2(1 - 1/m)\sqrt{d-m}$ for the largest eigenvalue (`QSL.Exchange.lamMax_ge`), and $2\kappa_d/\lambda_d^2 \to 1$ (`QSL.Exchange.ratio_tendsto_one`); the rate is $O(d^{-2/3})$ by the asymptotics of the largest Hermite zero (DLMF 18.16.17, not formalized). At a fixed interaction norm $\|H_c\| = J\kappa_d$, the CZ time of the construction therefore approaches the Bell-preparation bound:

the entangling-time bound becomes tight for the whole gate as $d \to \infty$, with ideal local control. The averaging step is an instance of the no-overhead simulation framework of [17]; the quadrature code and the asymptotic tightness were not found in the sources we read. Both bounds can be computed for any d and J in the blueprint.

Numerical solutions. With local controls of at most $10J$ during the gate (plus instantaneous local unitaries before and after), a CZ is realized to error $\le 2\cdot10^{-8}$ (leakage $\le 4\cdot10^{-8}$) at $T = 0.55/J$, i.e. at 0.35 of the qubit-only time and 5% above the construction. At $T = 0.45/J$ and $0.50/J$, relaxing the cap to $20J$, $40J$ or none did not improve the solutions found ($\approx 5\cdot10^{-3}$ and $\approx 4\cdot10^{-4}$); this does not identify what limits them, and none was found at $0.50/J$. For the same task and metric, $0.4510/J \le T_*(2\cdot10^{-8}) \le 0.55/J$; the bound also shows that no protocol reaches $e \le 2\cdot10^{-8}$ at $T = 0.45/J$ (it needs $e \ge 1.9\cdot10^{-6}$ there), consistent with the $5\cdot10^{-3}$ found. The solutions populate the second level: up to 60% of the population of a computational input leaves the qubit subspace during the gate (`poblacion_cz.py`). Each code state of the construction has population $1/3$ in $|2\rangle$, so each encoded basis input has $5/9 \approx 56\%$ outside the qubit subspace (superpositions range from 0 to $8/9$); together with the $\sqrt{2}$-enhanced elements $\langle 20|H_c|11\rangle$, $\langle 02|H_c|11\rangle$ identified for CNOT [16] and iSWAP [7], this suggests a mechanism for our solutions that we have not isolated. For the qutrit CZ, $CZ_3|k,l\rangle = e^{2\pi ikl/3}|k,l\rangle$, with a capacitive coupling and drive cap $40g$, the optimal-control time reported in [8] is close to the qubit limit. That is a different gate: on the qubit subspace it is $\mathrm{diag}(1,1,1,e^{2\pi i/3})$, not $\mathrm{diag}(1,1,1,-1)$, and coupling and controls differ, so the two times are not comparable.

Heisenberg coupling and ancillas: an identity. For $H = XX + YY + ZZ$ and fast local control, the smallest largest-Schmidt-coefficient found from a product state at time T is $\lambda(T) = \cos(2T)$ to 10^{-13} without ancillas and to 10^{-8} with one ancilla qubit per side (`grape_bell_torch.py`), up to the Bell time $\pi/8$ [4]. The same identity $\lambda(T) = \cos(wT)$, with w the spectral half-width, holds for Ising ZZ and XY couplings, normalized as $H_{\mathrm{XY}} = (XX + YY)/2$ ($w = 1$ for both), with and without ancillas (errors $\le 3\cdot10^{-7}$). The norm bound of Section II gives rate $\|H\| = 3$; the identity says the right rate is the spectral half-width $(E_{\max} - E_{\min})/2 = 2$. This is a certified corollary (`QuditEntanglingShift`): since $H - c\mathbb{1}$ differs from H by $c\mathbb{1} = (c\mathbb{1}) \otimes \mathbb{1}$, which can be absorbed in the local control, $\arccos\lambda \le T\,\|H - c\mathbb{1}\|$ for every real c; for Heisenberg $\|H + \mathbb{1}\| = 2$ (because $H + \mathbb{1} = 2\,\mathrm{SWAP}$), whence $T \ge \pi/8$, which [4] attains. The certified bound gives $\lambda(T) \ge \cos(2T)$ for every control, and an explicit trajectory reaches it: since $H = 2\,\mathrm{SWAP} - \mathbb{1}$, $e^{-iHT}|01\rangle = e^{iT}[\cos(2T)|01\rangle - i\sin(2T)|10\rangle]$, whose largest Schmidt coefficient is $\cos(2T)$ for $T \le \pi/8$. Up to the Bell time the minimum is therefore exactly $\cos(2T)$; the numerics are checks of this. Local ancillas do not change the spectrum of the coupling and so cannot lower this bound; beyond the Bell time they do allow more entanglement than two qubits can hold. For two qubits without ancillas the rate coincides with the $h_{\max} = \mu_1 + \mu_2$ of [4].

TABLE VII
IBM CHIP-TO-CHIP TRANSFERS [21] AGAINST THE SINGLE-MODE OPTIMUM $\pi/(\sqrt{2}\,g)$, FOR $g/2\pi = 8$ AND 9 MHZ (44.2 AND 39.3 NS). ERROR $= 1 - F$ AS REPORTED. COHERENCE AT T_g: IBM'S ERROR MODEL (THEIR EQS. (1)–(2), OCCUPATIONS $\frac{1}{2}$ QUBITS, $\frac{1}{2}$ CABLE; OUR REPRODUCTION). COHERENCE AT $T_{\min}$: LINDBLAD MASTER EQUATION OF THE OPTIMAL PROTOCOL WITH THE SAME COHERENCE TIMES. INTERACTIVE VERSION FOR YOUR OWN LINK.

Link	T_g (ns)	$T_g/T_{\min}$	error	coh. at T_g	coh. at $T_{\min}$
Q240–Q247	312	7.06–7.94	0.71%	0.63%	0.13–0.15%
Q241–Q246	296	6.70–7.53	1.13%	0.87%	0.23–0.26%
Q242–Q245	256	5.79–6.52	1.01%	0.61%	0.16–0.18%
Q243–Q244	312	7.06–7.94	1.14%	0.94%	0.22–0.24%

VIII. APPLICATION: A CHIP-TO-CHIP STATE TRANSFER

The same diagnostic applies to operations other than gates. Sundaresan *et al.* [21] transfer a qubit state between two 120-qubit processors through an 80 cm aluminium coaxial link (an "L-coupler"), relaying the excitation through one standing mode of the cable. They report the maximal qubit–mode coupling, $g/2\pi = 8$–9 MHz (their Fig. 1f), the transfer duration T_g and the fidelity of each of four links (their Extended Data Table 2), and the coherence times of qubits and cable modes (Extended Data Tables 1 and 2). The coupling waveforms are not published.

Model. In the single-excitation subspace, with the rotating-wave approximation and one resonant mode, the relay is a three-level problem on $|A\rangle$ (first qubit), $|C\rangle$ (cable mode), $|B\rangle$ (second qubit): $H(t) = g_1(t)\,(|A\rangle\langle C| + \mathrm{h.c.}) + g_2(t)\,(|C\rangle\langle B| + \mathrm{h.c.})$ with complex $g_i(t)$, $|g_i(t)| \le g$ and $g = 2\pi \times (g/2\pi)$.

Certified floor. Since $\|H(t)\| = (|g_1|^2 + |g_2|^2)^{1/2} \le \sqrt{2}\,g$ and the transfer $|A\rangle \to |B\rangle$ needs a Fubini–Study angle $\pi/2$, the bound `QSL.Gate.fsAngle_le_time_halfWidth` of Section II (with shift $c = 0$; it holds in any dimension) gives $T \ge \pi/(2\sqrt{2}\,g)$ for piecewise-constant couplings of any phase. It is certified and not tight.

Exact optimum. For this model the minimal time is $\pi/(\sqrt{2}\,g)$, reached with both couplings constant and at their maximum: the isotropic case of Theorem 1 of Boscain, Chambrion and Charlot [22]. The lower half, $gT \ge \pi/\sqrt{2}$, is machine-checked in Isabelle/HOL (`Relay_Minimal_Time.relay_minimal_time`, entry `Quantum_Speed_Limits` [23]) for the amplitudes $x = |a_A|$, $z = |a_B|$, which satisfy $|\dot{x}|, |\dot{z}| \le g\sqrt{1 - x^2 - z^2}$ away from a countable set of times; the proof uses an explicit calibration function, and the reduction from the Schrödinger equation to these inequalities (one line, valid also with several cable modes and detunings when g is the norm of the coupling vector of each end) is in the text, not in Isabelle. We also reproduced the optimum numerically with piecewise-constant controls, real and complex (`tiempo_minimo.py`): no control found reaches full transfer for $gT < \pi/\sqrt{2}$. With these couplings the excitation spends a quarter of the time in the cable ($\int_0^T |c(t)|^2 dt = T/4$). The sequential protocol (two swaps) takes π/g.

Gaps. The ratios of Table VII are ratios to the optimum of the single-mode model with the published maximal couplings; they do not show that the device admits a speed-up by that factor at the same fidelity. The transfers last 5.8–7.9

times the single-mode optimum (Table VII), 4.1–5.6 times the sequential time and 11.6–15.9 times the certified floor. The 126 ns that the authors give as a target would still be 2.9–3.2 times the optimum (diagnose your own link: the same numbers, a master-equation curve over the duration and a reading of which coherence time dominates). By IBM's own error model most of the measured error is coherence at these durations. At the optimum, with the same coherence times, the coherence error of the optimal protocol is 0.13–0.26% (master equation, `lindblad_relevo.py`, checked against an independent integrator `lindblad_rk4.py`), four to six times below the measured errors. The additive model of Eq. (2) evaluated with the occupations of the optimal protocol gives less, 0.09–0.17%: it is exact for relaxation but not for pure dephasing while the excitation is shared between elements (`comparar_modelos.py`). This is the coherence error of one admissible protocol, not a floor: a slower protocol can lose less (with lossless cable and lossy qubits the sequential protocol, at twice the duration, has a smaller error, `falsar_v198.py`), and a floor would need an inequality valid for every control.

Caveats. (i) The cable has many modes (free spectral range ≈ 150 MHz, $g/\mathrm{FSR} \approx 0.05$–$0.06$ by the authors' figures); protocols that use several modes are not covered by the single-mode optimum, and for them the certified floor and the Isabelle bound $\pi/(\sqrt{2}\,G)$ hold with the collective coupling norm G of each end, which is not reported. (ii) g is the maximal coupling; the ratios measure the distance to what that maximum allows, not to the waveforms used. (iii) The reported fidelity and the error model are IBM's; the occupations $\frac{1}{2}$–$\frac{1}{2}$ are IBM's assumption, and for the qubits we use, as IBM does, the shorter coherence time of each pair. (iv) The master equation takes dephasing as Markovian at the rate set by the Ramsey T_2; low-frequency noise would change the dephasing terms. The computation is in `relevo/cocientes.py`, `relevo/diag_relevo.py` and `relevo/suelo_lindblad.py`, with the data in `relevo/datos/ibm_2610.29986.toml`.

IX. What the Certificate Does and Does Not Guarantee

The certified objects are the inequalities, not the gaps. The theorems of Table I are machine-checked for all inputs satisfying their hypotheses. Each reported gap is a computation on vendor data; its trust rests on the controls of Section III, not on a proof.

The model is IBM's, not a measurement. Every speed limit is evaluated in IBM's Duffing model with the RWA. If the model misdescribes a device, the gap inherits the error.

`omegad` is derived. It comes from the π-pulse calibration (Section III), and a factor $1/(1-e^{-2})$ separates the area with which `omegad` was derived from Qiskit's normalized pulse; which of the two waveforms reaches the qubit is not in the snapshots. The single-qubit gaps carry this uncertainty, and duration–`omegad` correlations are partly circular.

Hypotheses of the two-qubit theorem. The Lean theorem assumes a constant interaction of bounded norm and piecewise-constant local controls, with a product initial state. Applying it to IBM gates assumes that the drives act locally and that the static coupling is the only interaction (no classical crosstalk). The single-qubit time-dependent bound and the ECR corollary are stated in Lean for piecewise-constant controls; a continuous pulse is covered only as a limit of such controls, which is not formalized (Section II). The norm $\|H_{\mathrm{int}}\| = j_q$ holds only in the qubit subspace, and j_q itself is a model value repeated across dates.

Bounds that are not in Lean. The literature gate bound $\pi/(2j_q)$ behind the headline 4.22, its application to real durations (our inference, Section V), the multilevel single-qubit bound $T \geq \pi/\Omega_{\max}$ and its trace-metric version (proved in the text, Section VII), the instantiation with references, the time argument of the refined CZ bound $0.4511/J$ (its $97/32$ lemma is in Lean), the two-copy bound and the certified protocol of version 3 (exact and interval computations, not proofs in Lean), the CZ construction in $d = 3$ and in dimension d, and the rate $O(d^{-2/3})$ are not formalized (the limit of the ratio of bounds is) (Lean certifies $\pi/(8J)$).

Truncation. The single-qubit times found with 4 levels at $r \geq 0.8$ are not converged in the number of levels (Section VII); the converged fastest gate, ≈ 3.4 ns, comes from a 6-level search checked up to 12 levels.

Effective cross-resonance models. The CR check of Section V uses effective Hamiltonians (first order, third order, and a static numerical block diagonalization); none is a time-dependent simulation of the calibrated pulse sequence with echoes and cancellation tones.

Coarse optimal control. Tables V and VI come from coarse grids (time steps, 8–10 piecewise-constant segments, finitely many random starts), without noise; the times reported are feasible times found, i.e. upper bounds on the minimum, and a time at which no gate was found proves nothing.

Snapshots, not live data; no hardware validation. All data are dated snapshots. We have run nothing on hardware: IBM has been withdrawing pulse-level access [12], so the feasible times found in Section VII are model predictions. Black-box timing on live hardware is nevertheless feasible: gate-time amplification [12], which reads gate durations from job wall-time, could validate the snapshot durations used here. We have not done so.

Relation to prior work. The closest work, Ishida and Hasegawa [12], runs the inverse direction: from gate times measured by amplification on `ibm_torino` they infer energy scales through ML and MT, and conclude that current gates are close to the QSL. We take the energy scale from the calibration (maximum drive amplitude, coupling) and ask how far the gate is from it. The two answer different questions, and the conclusions must be read with their energy scales attached. For a fixed-axis single-qubit π rotation their MT estimator equals the time-averaged energy spread of the pulse, so the gate sits at that limit by construction; their limit uses the drive actually applied, our headline 11.3 uses the full-amplitude drive, which IBM does not use. At the realized drive our own numbers give a gap of about 2 (Section IV-C), compatible with their reading. Their measurement is model-free and on a Heron device; ours requires the vendor Hamiltonian and covers only pre-Heron snapshots. Basilewitsch, Dlaska and Lechner [13] compute

optimal-control QSLs with Sycamore parameters; Howard *et al.* [9] run two-qubit gates at the speed limit on a custom chip; Basyildiz, Gong and Ashhab [8] compare a theoretical qubit-CZ limit of $\approx$ 46 ns with a 580 ns experiment. That experiment is a *qutrit* CZ, so the pair mixes a qubit limit with a qutrit gate. Wu *et al.* [14] test the Zieliński–Zych bounds on a transmon–resonator experiment. None of these is machine-checked, and none that we found sweeps a vendor's fleet. Our work complements them: a fleet-wide, version-pinned diagnostic with certified inequalities.

X. Reproducibility

All scripts and data are in the companion package of the blueprint, https://karlesmarin.github.io/blueprints/qsl-diagnostics/companion/ (Apache 2.0): `convencion.py` (three-level check of the drive convention), `pulsos_x.py` and `identidad_omegad.py`, `identidad_calibracion.py` (convention from calibrated π pulses), `barrido.py` (single-qubit sweep, `barrido.csv`), `pares.py` (two-qubit pairs, `pares.csv`), `cruce_formulas.py` (CR and peak-Rabi checks), `serie_temporal.py`, `serie_porque.py`, `duracion_error.py` (time series, `serie.csv`), `degradacion.py` and `frecuencia_degradados.py` (stratified controls), `mejora_por_cubit.py` (per-qubit durations), `mediana_por_backend.py` (backend-weighted median, clustered bootstrap), `toronto_sensibilidad.py` (outlier sensitivity), `control_dos_cubits.py` and `cnot_m3.py` (numerical controls of the certified bounds), `grape_1q.py`, `grape_cr.py` and `grape_cr_continuacion.py` (monotonicity check), `grape_cz_torch.py` with its independent check `verificar_cz.py`, `grape_bell_torch.py`, `grape_anchura.py`, `poblacion_cz.py` (optimal control), `fuentes/cruce_mg2020.py` (higher-order CR check), `cruces_area_cz.py` (pulse-area identity, CZ construction) and `truncacion_1q.py` (truncation check), `tercera_revision.py` (gate-metric bounds, r_{full}), `cota_cz_refinada.py` (refined CZ bound). Data: `qiskit-ibm-runtime` 0.51.0 and `qiskit-terra` 0.24.1 fake-backend snapshots (Apache 2.0), copied in `datos/` and `datos_serie/`. The Lean sources are at https://github.com/karlesmarin/formalizations, commit `42f307b`. The calibrated pulses (`qiskit-terra` 0.24.1) are copied in `datos_pulsos/`; a self-check rebuilds `barrido.csv` inside the package and compares it with the archived file. A companion blueprint, https://karlesmarin.github.io/blueprints/qsl-diagnostics/, maps each claim of this paper to its Lean theorem or script. Its laboratories run in the reader's browser on the reader's own values: Figs. 1–4 redrawn live; the fixed-axis equality on the Bloch sphere for any envelope; the certified floors of Section II for a list of the reader's own gates (diagnose your device), which fires the falsifier of Section III on any gate below its floor; and the d-level CZ bounds of Section VII. Each result can be downloaded as CSV or shared as a link with its inputs. For Section VIII (version 2): `relevo/datos/ibm_2610.29986.toml` (the published values, each with its table), `relevo/cocientes.py` (ratios), `relevo/diag_relevo.py` (IBM's error model), `relevo/lindblad_relevo.py` with its independent check `relevo/lindblad_rk4.py` (master equation), `relevo/suelo_lindblad.py`, `relevo/comparar_modelos.py`, `relevo/falsar_v198.py` (additive model and the slower protocol), `relevo/tiempo_minimo.py` (numerical optimum) and `relevo/heya_satd.py`, which reproduces the published Lindblad simulation of an IBM SATD transfer [24] ($2.0 \cdot 10^{-3}$ and $6.0 \cdot 10^{-3}$ against $2 \cdot 10^{-3}$ and $6.2 \cdot 10^{-3}$) as a validation of the simulator. The Isabelle sources of `Quantum_Speed_Limits` are in the same package.

XI. Conclusion

On public calibration snapshots of pre-Heron IBM processors (mainly Falcon and Eagle), single-qubit gates run at a median $11.3\times$ their two-level speed limit at full drive amplitude, a ratio within about 5% of 11.3 in every year from 2021 to 2025. That denominator is an amplitude IBM does not use: it runs the drive at about a tenth of the anharmonicity, and at that drive its pulses are about 2 times the square-pulse limit (the shape factor 1.87, or 2.16 with Qiskit's peak normalization) and can be shortened by at most 2.18 (nominal peak) or 1.89 (edge-subtracted peak) at the same peak drive; the remaining ≈ 5 is amplitude headroom, a design choice. Optimal control on IBM's model finds ≈ 16 ns X gates at IBM's own amplitude cap and 4.8 ns at its full amplitude (where the minimum is at least 4.3 ns): within the same amplitude budget the model admits gates about 10 times faster than the median gate (7.4–8.3 times for a 35.6 ns Falcon gate); ≈ 3.4 ns needs three times that amplitude. These are feasible times in an idealized model, not physical floors. Two-qubit gates run at a median $4.22\times$ the known qubit-subspace CNOT-class bound $\pi/(2j_q)$; in the snapshots that ratio rose from 2.78 to 4.26 between 2021 and 2025 as gates got longer, not monotonically and coinciding with the change from Falcon CX to Eagle ECR, while the archived couplings did not change. The current Heron generation (CZ $\approx$ 68–70 ns) is not covered. The inequalities of Table I are machine-checked, including an any-dimension entangling-time bound that is new only as a certified assembly of known pieces. On the model, optimal control finds a cross-resonance CNOT ≈ 1.2 times faster at drive fraction 0.7 than at IBM's 0.45, and, with the second transmon level, a CZ (error $2 \cdot 10^{-8}$) in 0.35 of the qubit-only exchange time with local controls of at most $10J$; both are feasible times found, not proven minima. The open questions are the true minimum between the two-copy bound $0.4632/J$ and the certified protocol $0.5200/J$ (version 3) (in d-level transmons the gap between the entangling bound and the construction closes as $d \to \infty$, a limit certified in Lean), and a certified bound that joins the speed-limit and leakage branches of (1), for which we have no result. The same diagnostic applied to IBM's chip-to-chip state transfer (version 2) places it at 5.8–7.9 times the single-mode minimal time $\pi/(\sqrt{2}\,g)$,

whose lower bound is machine-checked in Isabelle, with a coherence error at that time four to six times below the measured errors; this is the error of one protocol, not a floor.

Conflict of Interest

The author declares no conflict of interest. This work received no funding.

Use of AI Tools

This work was carried out with an AI assistant. The author checked all statements, proofs and computations and takes full responsibility for the content.

References

[1] L. Mandelstam and I. Tamm, "The uncertainty relation between energy and time in non-relativistic quantum mechanics," *J. Phys. USSR*, vol. 9, pp. 249–254, 1945.

[2] N. Margolus and L. B. Levitin, "The maximum speed of dynamical evolution," *Physica D*, vol. 120, pp. 188–195, 1998.

[3] B. Zieliński and M. Zych, "A generalization of Margolus–Levitin bound," *Phys. Rev. A*, vol. 74, 034301, 2006. arXiv:quant-ph/0603076.

[4] W. Dür, G. Vidal, J. I. Cirac, N. Linden, and S. Popescu, "Entanglement capabilities of nonlocal Hamiltonians," *Phys. Rev. Lett.*, vol. 87, 137901, 2001. arXiv:quant-ph/0006034.

[5] G. Vidal, K. Hammerer, and J. I. Cirac, "Interaction cost of nonlocal gates," *Phys. Rev. Lett.*, vol. 88, 237902, 2002. arXiv:quant-ph/0112168.

[6] K. Hammerer, G. Vidal, and J. I. Cirac, "Characterization of non-local gates," *Phys. Rev. A*, vol. 66, 062321, 2002. arXiv:quant-ph/0205100.

[7] B. Basyildiz, C. Jameson, and Z. Gong, "Speed limits of two-qubit gates with qudits," arXiv:2312.09218, 2023.

[8] B. Basyildiz, Z. Gong, and S. Ashhab, "Speed limits of two-qutrit gates," arXiv:2510.07742, 2025.

[9] J. Howard, A. Lidiak, C. Jameson, B. Basyildiz, K. Clark, T. Zhao, M. Bal, J. Long, D. P. Pappas, M. Singh, and Z. Gong, "Implementing two-qubit gates at the quantum speed limit," *Phys. Rev. Research*, vol. 5, 043194, 2023. arXiv:2206.07716.

[10] Ł. Rudnicki, "Quantum speed limit helps interpret geometric measure of entanglement," *Phys. Rev. A*, vol. 104, 032417, 2021. arXiv:2107.11877.

[11] S. Bravyi, "Upper bounds on entangling rates of bipartite Hamiltonians," *Phys. Rev. A*, vol. 76, 052319, 2007. arXiv:0704.0964.

[12] N. Ishida and Y. Hasegawa, "Energy inference of black-box quantum computers using quantum speed limit," arXiv:2512.15472, 2025.

[13] D. Basilewitsch, C. Dlaska, and W. Lechner, "Comparing planar quantum computing platforms at the quantum speed limit," *Phys. Rev. Research*, vol. 6, 023026, 2024. arXiv:2304.01756.

[14] Y. Wu, J. Yuan, C. Zhang, Z. Zhu, J. Deng, X. Zhang, P. Zhang, Q. Guo, Z. Wang, J. Huang, C. Song, H. Li, D.-W. Wang, H. Wang, and G. S. Agarwal, "Testing the unified bounds of quantum speed limit," *Phys. Rev. A*, vol. 110, 042215, 2024. arXiv:2403.03579.

[15] M. Werninghaus, D. J. Egger, F. Roy, S. Machnes, F. K. Wilhelm, and S. Filipp, "Leakage reduction in fast superconducting qubit gates via optimal control," *npj Quantum Inf.*, vol. 7, 14, 2021. arXiv:2003.05952.

[16] S. Ashhab, F. Yoshihara, T. Fuse, N. Yamamoto, A. Lupascu, and K. Semba, "Speed limits for two-qubit gates with weakly anharmonic qubits," *Phys. Rev. A*, vol. 105, 042614, 2022.

[17] P. Wocjan, M. Rötteler, D. Janzing, and T. Beth, "Simulating Hamiltonians in quantum networks: Efficient schemes and complexity bounds," *Phys. Rev. A*, vol. 65, 042309, 2002. arXiv:quant-ph/0109088.

[18] E. Magesan and J. M. Gambetta, "Effective Hamiltonian models of the cross-resonance gate," *Phys. Rev. A*, vol. 101, 052308, 2020. arXiv:1804.04073.

[19] Qiskit IBM Runtime, version 0.51.0, `fake_provider` backends, Apache 2.0. https://github.com/Qiskit/qiskit-ibm-runtime.

[20] Qiskit Terra, version 0.24.1, `fake_provider` backends, Apache 2.0. https://github.com/Qiskit/qiskit (package `qiskit-terra` on PyPI).

[21] N. M. Sundaresan *et al.*, "Meter-scale quantum interconnects between large quantum processors," arXiv:2610.29986 (2026).

[22] U. Boscain, T. Chambrion, and G. Charlot, "Nonisotropic 3-level quantum systems: complete solutions for minimum time and minimum energy," *Discrete Contin. Dyn. Syst. Ser. B*, vol. 5, pp. 957–990, 2005; arXiv:quant-ph/0409022.

[23] C. Marín, "Quantum Speed Limits: the Zieliński–Zych Bound and Time–Occupation Bounds for Qubit–Bus–Qubit State Transfer," Isabelle/HOL entry (October 2026); sources in the companion package.

[24] K. Heya *et al.*, "Randomized benchmarking of a remote CNOT gate via a meter-scale microwave link," *Phys. Rev. Lett.*, vol. 135, art. 200801, 2025; arXiv:2502.15034.

[25] C. Marín, "Formalizations" (Lean 4), https://github.com/karlesmarin/formalizations, commit `42f307b`, 2026. Companion text: C. Marín, "The clock that never ticks: a machine-checked path from the dilogarithm to quantum speed limits in Lean 4," Zenodo, 2026, doi:10.5281/zenodo.20675271.