Superconducting architecture demonstrating fast, tunable high-fidelity *CZ* gates with parametric control of *ZZ* coupling

X.Y. Jin,[1,2] Z. Parrott,[1,2] K. Cicak,[3] S. Kotler,[1,2,*] F. Lecocq,[3] J. Teufel,[3] J. Aumentado,[3] E. Kapit,[4] and R.W. Simmonds[3,†]

[1] *National Institute of Standards and Technology, Boulder, Colorado 80305, USA*

[2] *Department of Physics, University of Colorado, Boulder, Colorado 80309, USA*

[3] *National Institute of Standards and Technology, Boulder, Colorado 80305, USA*

[4] *Department of Physics, Colorado School of Mines, 1500 Illinois St., Golden, Colorado 80401, USA*

(Received 17 December 2024; revised 28 June 2025; accepted 21 October 2025; published 8 December 2025)

Future quantum information processors require tunable coupling architectures that can produce high-fidelity logical gates between two or more qubits. Parametric coupling is a powerful technique for generating tunable interactions between many qubits. Here, we present a highly flexible parametric coupling scheme with superconducting qubits that provides complete removal of residual *ZZ* coupling and the implementation of driven SWAP or SWAP-free controlled-*Z* (*CZ*) gates. Our fully integrated two-dimensional on-chip coupler design is only weakly flux tunable, cancels static linear coupling between the qubits, avoids internal coupler dynamics or excitations, and is extensible to multiqubit circuit-QED systems. Exploring gate fidelity versus gate duration allows us to maximize two-qubit gate fidelity, while providing insights into possible error sources for these gates. Randomized benchmarking over several hours reveals that the parametric SWAP *CZ* gate achieves an average fidelity of $99.44 \pm 0.09\%$ in a gate duration of 70 ns and that a dispersively driven parametric SWAP-free *CZ* gate attains an average fidelity of $99.47 \pm 0.07\%$ in only 30 ns. The fidelity remained above this value for over 8 h and peaked twice with a maximum of $99.67 \pm 0.14\%$. Overall, our parametric approach combines versatility, precision, speed, and high performance in one compact coupler design.

DOI: 10.1103/kmls-lgp5

I. INTRODUCTION

Quantum computers hold the potential to outperform classical computers by solving specific problems that are unattainable for classical systems [1–5]. To achieve quantum advantage in computation, two-qubit gates must be fast to maintain high fidelities within the entire architecture to allow for quantum error correction [6]. So far, it is not clear what the optimal coupling architecture should be, but it is clear that a larger-connectivity graph between elements combined with well-controlled quantum evolutions is advantageous for achieving both useful quantum computations and quantum simulations [3]. For superconducting architectures with finite coherence times, two of the most challenging problems that remain are (i) the ability to perform quantum gates over short timescales [7–10] with minimal uncontrolled or stray couplings to ensure high-fidelity operations, and (ii) to achieve this beyond just nearest-neighbor pairs and possibly between any two arbitrarily chosen qubits within an *N*-qubit architecture [3,11]. Thus far, there has been considerable progress in improving gate fidelities with tunable couplers, between nearest-neighbor qubits or even extended nearest-neighbor qubits [9,10,12–16]. Gate fidelities beyond 99.9% have been achieved with fluxonium qubits [15–17], with all coherence times beyond 100 μs and with one example [16] utilizing parametric gates.

As a way to move toward achieving all-to-all coupling and a more modular architecture, various approaches have been pursued to enable parametric interactions with superconducting circuits [18–23], analogous to some trapped-ion systems [3]. In fact, parametric coupling has emerged as an essential tool for producing efficient and high-fidelity quantum operations with various types of superconducting systems, including planar cavities [24], cavity QED [25], three-dimensional (3D) cavities [26], transmon qubits [27–33], and fluxonium qubits [16], and its application has

*Current Address: Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

†Contact author: raymond.simmonds@nist.gov

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

been expanded beyond two-qubit gates to include higher levels or multiqubit and cavity-QED systems [19,20,23,28,34–36]. With growing interest in parametric coupling for quantum computing, researchers continue to develop new techniques and explore innovative applications, driving new advancements in the field [16,22,37].

In this work, we show a two-dimensional (2D) architecture on chip that uses parametric drives to control various resonant and nonresonant coupling dynamics, providing a flexible coupling architecture [19,21,34,36,38–41] that can achieve fast high-fidelity two-qubit gates even with lower coherence transmons. Parametric techniques rely on modulating a controlled parameter at a specific frequency and phase with a variable amplitude, such as an electrically or mechanically variable capacitance or a current- or flux-dependent inductance [24,25,42–44]. A single parametric coupler is versatile, providing either single- or dual-mode amplification [42,44,45], frequency conversion and dispersive interactions [19,39], or nonreciprocal directional dynamics [45]. In this study, we present a parametric coupler design strategy that minimizes the number of on-chip control and readout ports and is capable of eliminating residual *ZZ* coupling between two transmon qubits, while performing various types of two-qubit gates: a parametric iSWAP gate, a SWAP-based controlled-*Z* (p-SWAP *CZ*) gate, and a dispersive parametric SWAP-free controlled-*Z* (p-SWAP-free *CZ*) gate. We find that the gate errors are a function of the gate duration, the shape of the pulse envelope, and the pulse amplitude. In particular, high-fidelity gates are vulnerable to very large-amplitude parametric drives that can rectify the qubit frequencies or induce single-qubit rotations from *n*-photon absorption at subharmonics of the individual qubit transition frequencies (for more details, see Appendix H). Identifying these issues provides us with a path to further improve parametric gates. Ultimately, this design can be extended to include more qubits and cavities [19] and such integrated multiqubit modules [20] could help facilitate necessary schemes for error correction [11,46,47]

II. CIRCUIT DESIGN

Our two-qubit system of study is displayed in Figs. 1(a) and 1(b) with two transmons, labeled "Left" (*L*) and "Right" (*R*), each having a main Josephson junction and a shared two-junction superconducting quantum interference device (SQUID) parametric coupler with total inductance L_s. Here, $\mathcal{L}_k = \Phi_o/2\pi I_k$ represents the (nonlinear)

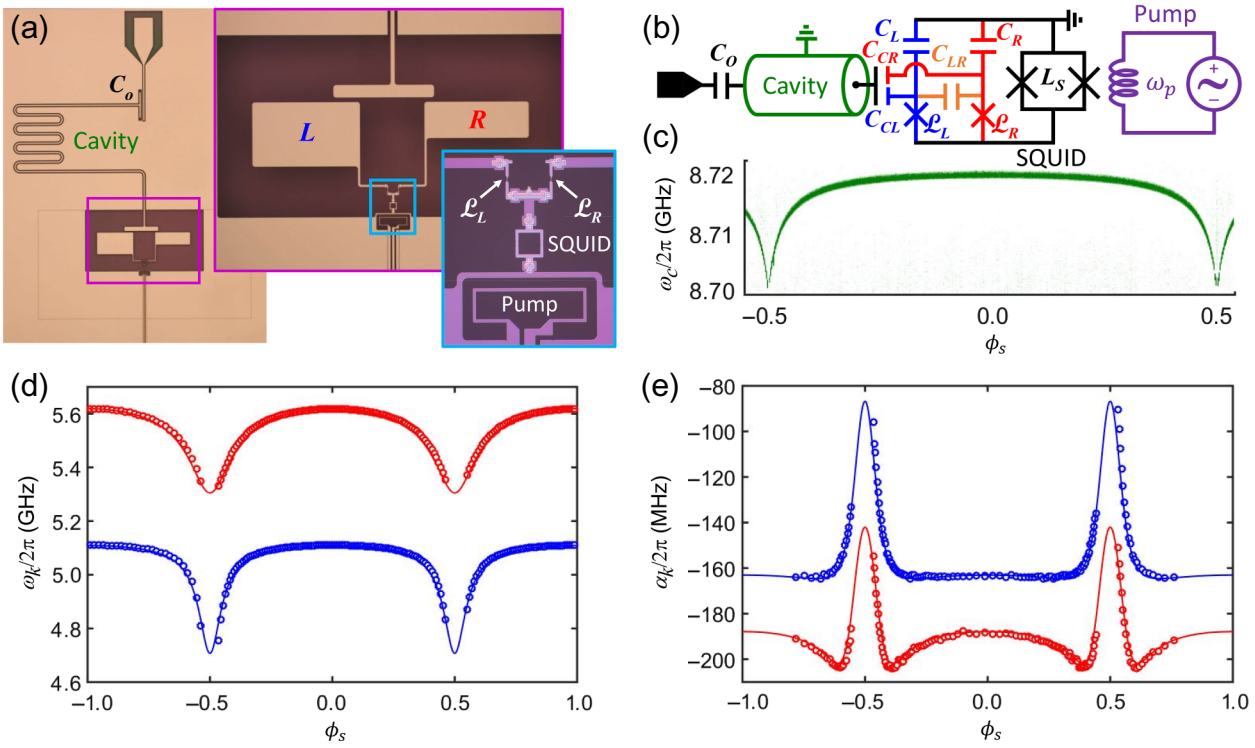


FIG. 1. A flux and parametrically tunable two-qubit system. (a) An optical micrograph of a two-qubit integrated symmetric superconducting quantum interference device (SQUID) system, with enlarged views to show more details of the niobium-on-sapphire-substrate device with aluminum shadow-evaporated junctions (lighter regions). (b) A schematic circuit diagram of the system, including the left transmon (*L*) and right transmon (*R*) plus the readout cavity. The parametric flux drive is applied through the shared SQUID coupler. (c) The cavity spectrum as a function of $\phi_s = \Phi_s/\Phi_0$. (d) A plot of the ϕ_s-periodic *L* and *R* transmon frequencies. (e) A plot of the ϕ_s-periodic *L* and *R* transmon anharmonicities.

Josephson inductance for $k \in L, R$ with nominally equal critical currents $I_o = I_L \approx I_R$ for the main qubit junctions, $E_J = I_o \Phi_o / 2\pi$ is the Josephson energy, and $\Phi_o = h/2e$ is the magnetic flux quantum. Each transmon has capacitance C_k (for $k \in L, R$), with mutual capacitance C_{LR}. The two qubits also share the same readout cavity, each capacitively coupled through C_{ck} (for $k \in L, R$), which also serves as the qubit drive line through the cavity coupling capacitor C_o to the feedline. This leads to a very weakly tunable cavity frequency that is periodic in the static flux bias $\phi_s = \Phi_s / \Phi_o$ of the coupler. Our compact layout only requires two on-chip ports for the whole system, one for the single-element excitations plus readout, and one for setting all static operating frequencies (static coupler bias) and for generating all coupled interactions (rf coupler "pump" drive). This type of architecture can be extended by adding more qubits or cavities to the coupler (see Sec. X), without increasing the number of on-chip control and readout ports [19,21,34].

In Fig. 1(d), we see that the two qubits are quite closely packed in frequency space and tune with ϕ_s in unison, so their frequencies will never collide. Here, we denote their frequencies as ω_L and ω_R, which represents the transition from the ground state $|gg\rangle$ to the joint two-qubit energy levels $|eg\rangle$ and $|ge\rangle$, respectively [for more details, see Fig. 2(c)]. The shared SQUID provides only a mild tunability of the two transmons, where the two nominally equal SQUID junctions are each about $20 \times E_J$, resulting in an inductive participation ratio of only a few percent at zero flux bias. The system is designed to operate at a fixed static flux, where the two qubit frequencies are fixed, as well as the static coupling between them. This mild tunability balances our ability to provide strong parametric interactions ($\lesssim$100 MHz), allowing us to perform two-qubit gates quickly, while still minimizing the susceptibility of either qubit ($\lesssim$1 GHz/Φ_o) to flux noise in the SQUID (for more details, see Appendix F) [19,48,49].

In Fig. 1(e), we see that the anharmonicity α_k (for $k \in L, R$) of the two transmons is roughly constant except for a sharp decrease as L_s increases rapidly near $\phi_s = 0.5$. Our device design differs considerably from most current coupler designs that usually incorporate a capacitively coupled SQUID that acts more like a tunable qubit or resonator between the qubits. In contrast, our SQUID coupler is fully integrated galvanically into the qubits themselves and its self-resonance frequency is >20 GHz for all ϕ_s, too high to be excited during coupled operations. This also means that the resonant frequency of the coupler will not collide with nearby qubits or cavities and that any additional degrees of freedom, internal coupler dynamics, or decoherence channels are unlikely to perturb its operation.

Our device design also takes advantage of the relatively large mutual capacitance C_{LR} between the two closely spaced transmons to largely cancel out the mutual inductive coupling due to the SQUID coupler. This is one of the major benefits of our architecture that allows us to pack the qubits close together and use the large mutual capacitance as a feature instead of as a bug. This helps to increase the packing density of qubits for scaling (see Sec. X). The L and R qubit electrodes are designed in size, orientation, and separation so that the negative capacitive coupling equals the positive inductive coupling of the SQUID at the static "cancellation-flux" bias, $\phi_c = \Phi_c / \Phi_o \approx 0.41$, where parametric coupling strengths will be strong.

III. STATIC *ZZ* COUPLING

At a flux bias far from ϕ_c, the residual static coupling g_s between the qubits is predominantly capacitive when $\phi_s < \phi_c$ or inductive when $\phi_s > \phi_c$. Because the SQUID coupler is fully integrated, when the two qubits (with $\alpha_k < 0$, for $k \in L, R$) are very strongly coupled ($g_s \gg 0$) with a static detuning of $\Delta_s = \omega_R - \omega_L \approx 0.5$ GHz, there is a very strong *static ZZ* coupling between them [7,50],

$$\zeta_s = 2g_s^2 \frac{(\alpha_L + \alpha_R)}{(\Delta_s - \alpha_L)(\Delta_s + \alpha_R)}. \tag{1}$$

In Fig. 2(a), we show an example of weak *ZZ* coupling near ϕ_c at $\phi_s \approx 0.386$, where $\zeta_s/2\pi \approx -6$ MHz, and in Fig. 2(b), we show ζ_s as a function of ϕ_s, where $|\zeta_s|/2\pi >$ 50 MHz at $\phi_s = 0$ or 0.5. Note that Eq. (1) only captures the induced *nonlinear* interactions from the static *linear* coupling between the two detuned qubits within the rotating-wave approximation (RWA), which goes to zero at ϕ_c. However, we see that at ϕ_c we have $|\min(\zeta_s)|/2\pi \approx 2$ MHz. This is due to the purely nonlinear coupling that results from a fraction of the current of the L transmon passing through the small qubit junction of the R transmon and vice versa. The size of this effect can be estimated by considering the counter-RWA contributions to ζ_s. Although the experimental devices discussed in this work have not achieved complete *static* cancellation of *ZZ* coupling, our design strategy can achieve this (see Appendix A). By placing ω_R within $|\alpha_R|$ of ω_L, the so-called "straddling regime" where $\Delta_s < |\alpha_R|$, these residual contributions can be completely eliminated to achieve $\zeta_s = 0$. This condition is utilized in numerous capacitively coupled transmon couplers to achieve a zero crossing for *ZZ* coupling [9,51]. Another strategy for cancellation of *ZZ* coupling relies on making $\alpha_L = -\alpha_R$ and can be achieved with flux-type qubits (or other qubits with positive anharmonicity) coupled to transmons [52,53]. Using our design layout, ANSYS Q3D, and the "scQubits" software package [54,55] to calculate the full interactions, we find very good agreement between our numerical predictions (solid lines) and the experimental data (circles) in Figs. 1(d), 1(e), and 2(b), including the residual $\min(\zeta_s)$ at ϕ_c (for more details, see Appendix B).

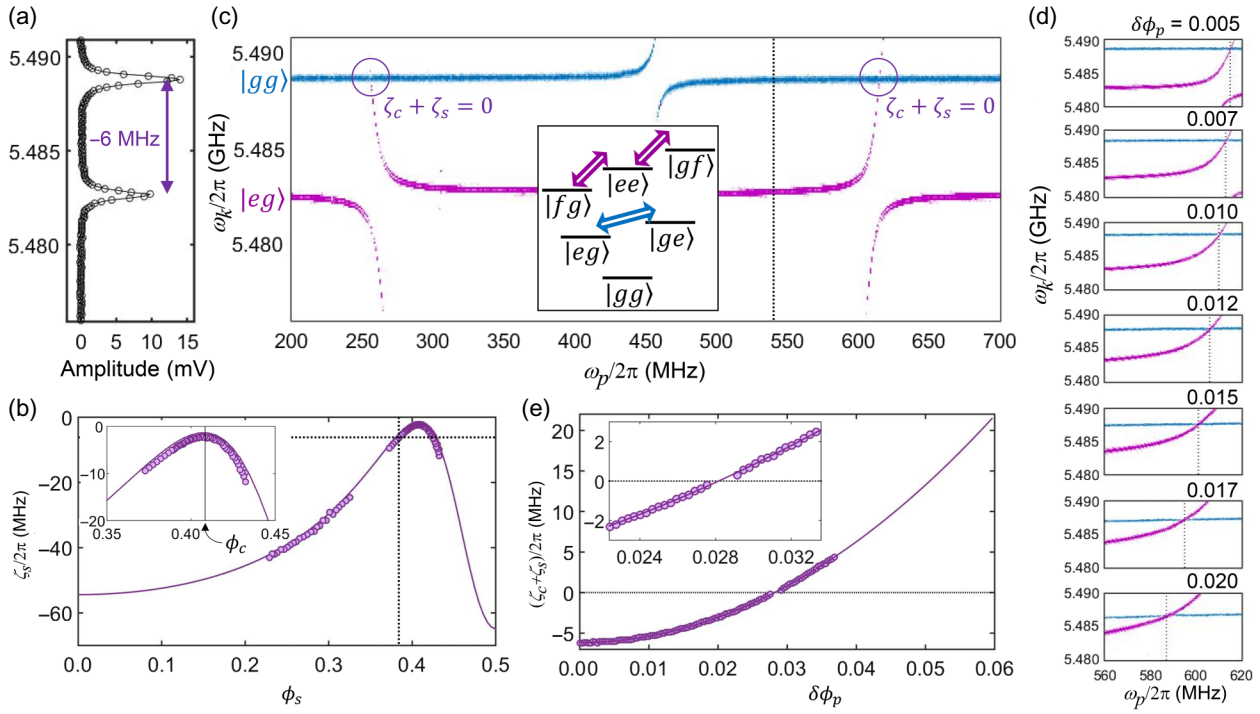


FIG. 2. The residual *ZZ* coupling and cancellation. (a) A spectroscopic scan showing *ZZ* coupling when $\phi_s \approx 0.386$. (b) The static *ZZ* coupling ζ_s as a function of ϕ_s: the dashed vertical line indicates when $\phi_s \approx 0.386$ and the horizontal line indicates when $\zeta_s \approx -6$ MHz. The inset shows an enlargement near ϕ_c (vertical line). (c) A spectroscopic scan in the presence of the parametric pump when $\phi_s \approx 0.386$. The circles indicate cancellation of the total *ZZ* coupling. The inset shows the two-qubit energy-level diagram with arrows indicating parametric transitions between the one- and two-excitation state manifolds. (d) A spectroscopic scan for increasing pump amplitude (see text for details). (e) $\zeta_s + \zeta_c$ as a function of the pump amplitude, showing the parametric cancellation of the *ZZ* coupling. The inset shows an enlargement near $\zeta_s + \zeta_c = 0$.

IV. PARAMETRIC INTERACTIONS: SPECTROSCOPIC

With the qubits detuned by Δ_s, energy exchange between two-qubit energy levels with an equal number of excitations is suppressed. The joint two-transmon levels are labeled $|LR\rangle$, denoting the individual L and R transmon levels as g, e, f, and so on [see Fig. 2(c)]. However, if in addition to the static flux bias ϕ_s, we drive the SQUID coupler with a cosinusoidal flux modulation, $\phi(t) = \delta\phi_p \cos(\omega_p t) + \phi_s$, then we can generate new coupling terms between any two of the two-qubit states $|b\rangle$ and $|t\rangle$ that reside within the same manifold of states with the same number of excitations written in an interaction frame with respect to the free Hamiltonian [25],

$$H_{Ip} = \hbar g_p \left(e^{i\omega_p t} + e^{-i\omega_p t}\right)\left(q_L q_R^\dagger e^{-i\Delta_{bt} t} + \text{h.c.}\right), \qquad (2)$$

where we will ignore sum frequency or counterrotating terms, $q_k^\dagger$ (q_k) is the creation (annihilation) operator of the kth transmon mode (for $k \in L, R$), and Δ_{bt} is the difference in frequency or detuning between the higher "top" level $|t\rangle$ and the lower "bottom" level $|b\rangle$. Here, we focus on $\omega_p = \Delta_{bt} + \Delta_p$, where Δ_p is the detuning from the difference frequency Δ_{bt}. The parametric coupling strength is related to the flux-dependent static coupling strength [25] through $2g_p = (dg_s(\phi_s)/d\phi)\delta\phi_p$ (for more details, see Appendix A). Of course, under ideal conditions $g_s(\phi_c) = 0$, while $dg_s(\phi_c)/d\phi \neq 0$. It is clear from Eq. (2) that when $\Delta_p = 0$, $\omega_p = \Delta_{bt}$ so that $H_{Ip} = \hbar g_p (q_L q_R^\dagger + q_L^\dagger q_R)$ provides a SWAP interaction between the qubits. These SWAP interactions are achieved by frequency conversion, where the parametric pump makes up for the energy difference between the states $|b\rangle$ and $|t\rangle$ within a given excitation manifold. Although not utilized in this work, it is also possible to connect between state manifolds that are different by a single excitation in each qubit (e.g., $|gg\rangle \leftrightarrow |ee\rangle$), where the parametric drive frequency occurs near the sum of the individual single-qubit excitation frequencies [16,56]. In general, the pump frequency can be detuned by Δ_p in order to sweep between parametrically induced resonant-type ($\Delta_p \ll g_p$) and dispersive-type ($\Delta_p \gg g_p$) interactions near the transition between the states $|b\rangle$ and $|t\rangle$ [56].

This behavior can been seen in the continuous-frequency domain [see Fig. 2(c)] by measuring the spectrum of the R qubit. Starting from $|gg\rangle$, when the spectroscopic tone is resonant with the R qubit transition, we move to state $|ge\rangle$. For a fixed pump amplitude of $\delta_p = 0.005$,

we see an avoided crossing in this spectrum with a coupling strength $2g_p/2\pi \approx 5$ MHz centered at $\omega_p/2\pi \approx 450$ MHz. This occurs when the two parametrically interacting states are $|b\rangle = |eg\rangle$ and $|t\rangle = |ge\rangle$ and $\omega_p = \Delta_{bt} = \omega_{ge} - \omega_{eg}$, exchanging the single excitation manifold of states, as depicted in the inset of Fig. 2(c). After π pulsing the L qubit, we start from $|eg\rangle$ and the R qubit spectroscopic tone moves the state to $|ee\rangle$ with a resonance frequency shifted by $\zeta_s/2\pi = -6$ MHz [see the line cut of both peaks shown in Fig. 2(a)]. As the pump tone increases above 200 MHz, an avoided crossing appears when $\omega_p = \omega_{gf} - \omega_{ee}$ and the two parametrically connected states are $|b\rangle = |ee\rangle$ and $|t\rangle = |gf\rangle$. And, lastly, when $\omega_p = \omega_{ee} - \omega_{fg}$, the interaction is between states $|b\rangle = |fg\rangle$ and $|t\rangle = |ee\rangle$. These two splittings occur for parametric interactions within the two-photon excitation manifold of states, as depicted in the inset of Fig. 2(c).

V. PARAMETRIC CANCELLATION OF *ZZ* COUPLING

The presence of the residual ZZ coupling near ϕ_c gives us an opportunity to show the versatility of our parametric coupling scheme by allowing us to use parametric dispersive shifts to cancel ζ_s [19,39]. Recall that the presence of ζ_s shifts the two-qubit energy-level spectrum [shown in the inset of Fig. 2(c)], so that the position of $|ee\rangle$ satisfies the relation $\omega_{ee} = \omega_{eg} + \omega_{ge} + \zeta_s$. Note, in Fig. 2(c), that there are two values for ω_p where the spectroscopic peaks for initial states $|gg\rangle$ and $|eg\rangle$ cross, as denoted by the circles. This occurs when the total ZZ coupling $\zeta_s + \zeta_c = 0$. The parametrically induced dispersive shifts ζ_c, analogous to the static case, follow from Eq. (1) with the reassignments $g_s \to g_p$ and $\Delta_s \to \Delta_p$. And, according to Eq. (1), as we increase the pump amplitude $\delta\phi_p$, g_p increases and the value for the detuning Δ_p that cancels ζ_s must also increase as well. This is observed in Fig. 2(d), as the cancellation point (dashed vertical line) moves increasingly to the left of Δ_{bt}. In addition, careful observation shows a clear decrease in the frequency of the spectroscopic line corresponding to the initial state $|gg\rangle$ as the pump amplitude increases due to a rectification effect from the curvature of the $\omega_k(\phi_s)$ curves (for more details, see Appendix E).

If we choose an ω_p that is maximally detuned from both parametric transitions from the single $(\omega_{ge} - \omega_{eg})$ and double $(\omega_{fg} - \omega_{ee})$ manifold of states, we can avoid unwanted SWAP interactions (as discussed below) and ensure purely dispersive shifts that can fully cancel ζ_s. An example is shown in Fig. 2(e), where we plot the total effective ZZ coupling $\zeta_s + \zeta_c$ as a function of pump amplitude for $\omega_p/2\pi = 545$ MHz, the vertical dashed line in Fig. 2(c). Starting with no parametric pumping and $\zeta_s + \zeta_c = -6$ MHz, as $g_p^2 \propto \delta\phi_p^2$ is increased, the spectroscopic peaks merge and we cross zero ZZ coupling. If we continue to increase the pump amplitude $\delta\phi_p$, ζ_c continues to increase as g_p^2 and dominates, leading to progressively stronger positive ZZ coupling.

Due to the scatter in the spectroscopic data, the quadratic fit allows us to identify zero ZZ coupling with an uncertainty of about ± 50 kHz. In order to improve this by an order of magnitude, we use cross-Ramsey measurements [51]. Recall that the total ZZ coupling between the two qubits, $\zeta_s + \zeta_c$, produces a frequency shift of the R (L) qubit when the state of the L (R) qubit changes from the ground state to the excited state. The cross-Ramsey experiment measures a Ramsey oscillation on one qubit while initializing the other qubit in either its ground state or excited state [the pulse sequence is shown in the top panel of Fig. 3(a) for a Ramsey on the R qubit]. Here, the delay time between the two $\pi/2$ pulses is fixed at δt and the rotation axis φ of the second $\pi/2$ pulse is swept by 2π, producing a Ramsey oscillation of the form $\cos(\varphi)$. In the presence of finite ZZ coupling, the Ramsey measurement acquires a phase $\delta\theta = (\zeta_s + \zeta_c)\delta t$ so that the oscillation becomes $\cos(\varphi + \delta\theta)$. We perform this experiment on the R (L) qubit with and without the π pulse on the L (R) qubit and measure the phase difference $\delta\theta$ near pump amplitudes where $\zeta_s + \zeta_c \approx 0$, as estimated from the spectroscopic data in Fig. 2(e).

As seen in Fig. 3(a), $\delta\theta$ shows a linear dependence on the pump amplitude and passes through zero. These more sensitive cross-Ramsey measurements have been performed on both qubits for several δt, showing all curves passing through a single pump amplitude that ensures $(\zeta_s + \zeta_c)/2\pi = 0 \pm 5.5$ kHz. As a further verification, we have performed randomized benchmarking (RB) [57,58] on both qubits simultaneously, as another sensitive test of residual ZZ coupling [51]. As shown in Fig. 3(b), we find that the error per gate for simultaneous RB, where we measure the probability that we have properly returned to $|gg\rangle$, is equal to the sum of the individual error rates found from single-qubit RB, which indicates that additional correlated errors from residual ZZ coupling are negligible.

VI. PARAMETRIC INTERACTIONS: TIME DOMAIN

Next, we show parametric interactions between the two qubits in the time domain, which leads us to performing specific gate operations. By applying a coherent pump signal in the vicinity of Δ_{bt}, parametric coupling induces transitions between the states $|b\rangle$ and $|t\rangle$ in a similar way to standard Rabi oscillations [24,25]. The probability of being in the initially empty state over time is given by

$$P(t) = \frac{4g_p^2}{\Omega^2}\sin^2\left(\frac{\Omega t}{2}\right), \tag{3}$$

where $\Omega = \sqrt{4g_p^2 + \Delta_p^2}$ is the parametric SWAP frequency when $\omega_p = \Delta_{bt} + \Delta_p$. When $\Delta_p = 0$, the on-resonance (in

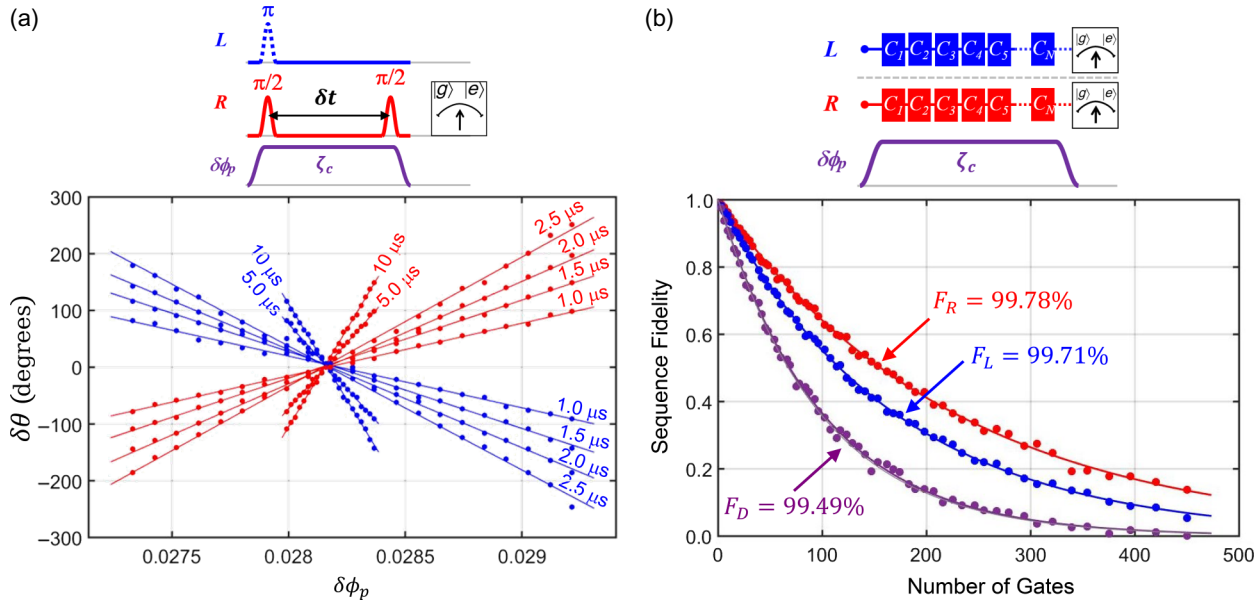


FIG. 3. The fine cancellation of *ZZ* coupling. (a) The pulse diagram of the cross-Ramsey measurement and data showing $\delta\theta$ as a function of pump amplitude $\delta\phi_p$ (see text for details). For each value of δt, all the red (blue) linear fit lines for the R (L) qubit data cross zero at the same pump amplitude, where the total *ZZ* coupling is eliminated. (b) The pulse diagram and data for single-qubit and simultaneous two-qubit randomized benchmarking (RB), showing that $\zeta_s + \zeta_c = 0$ with no additional correlated gate errors. The solid lines are fits to the data. The solid gray line is the expectation based on combining the single-qubit errors.

the rotating frame) SWAP frequency Ω is $2g_p$, which is equal to the total mode-splitting size seen in the spectroscopic data. Also, note that the probability amplitude of these oscillations reaches unity only when $\Delta_p = 0$. This will be important later when we discuss sources of gate errors.

An example of these parametric oscillations is shown in Figs. 4(a) and 4(b) between states $|b\rangle = |eg\rangle$ and $|t\rangle = |ge\rangle$

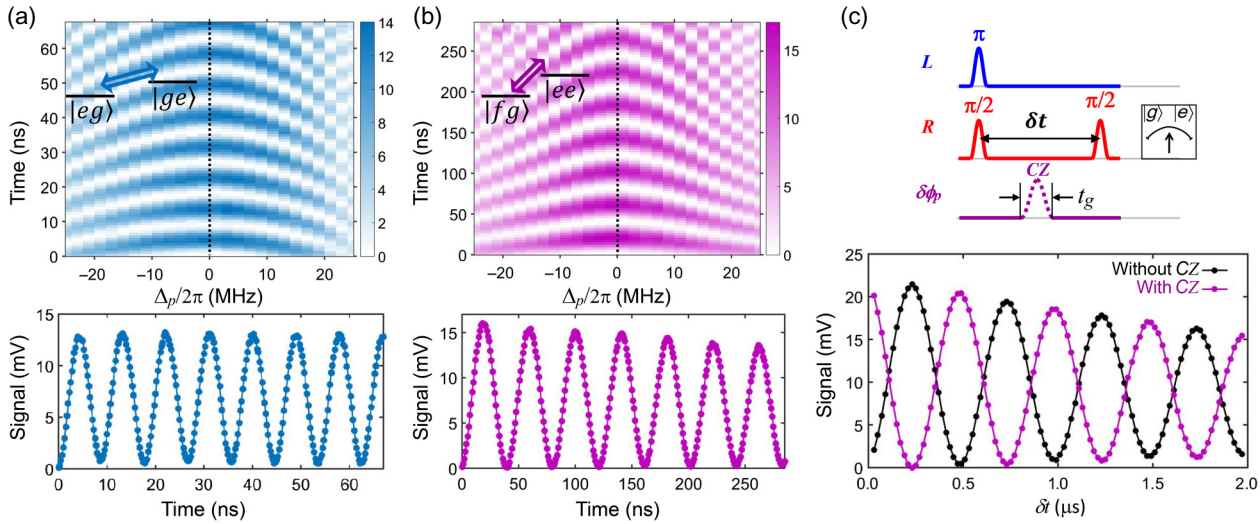


FIG. 4. Parametric SWAP oscillations and a demonstration of a SWAP-based controlled-*Z* (p-SWAP *CZ*) gate. (a) Top panel: parametric SWAPs between states $|eg\rangle$ and $|ge\rangle$ as a function of Δ_p. Bottom panel: Rabi oscillations with a maximum rate of about 100 MHz when $\Delta_p = 0$ (vertical dashed line in top panel). (b) Top panel: parametric SWAPs between states $|fg\rangle$ and $|ee\rangle$ as a function of Δ_p. Bottom panel: Rabi oscillations with a maximum rate of about 25 MHz when $\Delta_p = 0$ (vertical dashed line in top panel). (c) The pulse sequence for the cross-Ramsey experiment to verify operation of the p-SWAP *CZ* gate. Adding a *CZ* gate shifts the phase of the Ramsey oscillations by π as expected.

and $|b\rangle = |fg\rangle$ and $|t\rangle = |ee\rangle$, respectively. These data have been taken with a two-qubit device with a design similar to that shown in Fig. 1, but all made from aluminum, on a high-purity silicon substrate. Only Fig. 4 shows data from the all-aluminum device. The characteristic shape due to Eq. (3) is most visible in Fig. 4(b). Line cuts with a higher density of points have been taken through the center of these data when $\Delta_p = 0$. In Fig. 4(a), we show a maximum SWAP frequency of $\Omega/2\pi \approx 100$ MHz and Fig. 4(b) shows $\Omega/2\pi \approx 25$ MHz. This provides the speed necessary to perform fast two-qubit iSWAP or *CZ* gates in about 10–40 ns. Although we have performed iSWAP gates utilizing states $|b\rangle = |eg\rangle$ and $|t\rangle = |ge\rangle$ with an unoptimized fidelity of 95%, we focus our attention on the *CZ* gates described below.

VII. SWAP-INDUCED PARAMETRIC GATES

Now, let us consider a SWAP-induced parametric controlled-*Z* (p-SWAP *CZ*) gate [59], which can be realized via a complete rotation from the state $|ee\rangle$ to one of the other two-photon excited manifold of states $|fg\rangle$ or $|gf\rangle$ and back to $|ee\rangle$. It is important to note that this redundancy in the parametric transitions that can perform the two-qubit *CZ* gate provides an advantage when scaling the number of qubits (see Sec. X), providing more freedom to avoid parametric frequency collisions. Parametric two-qubit gates using similar principles have been realized with other systems [18,29,31], but with either limited parametric coupling strength [18,29] or reduced coherence times [31]. As we will see, our parametric coupling strategy can take advantage of high gate speeds to achieve a large ratio between coherence time and gate duration to minimize gate errors in the presence of qubit decoherence.

We show in Fig. 4(c) a simple demonstration of a p-SWAP *CZ* gate, again with the all-aluminum device. Here, we perform a cross-Ramsey experiment on the *R* qubit with a π pulse on the *L* qubit with and without a *CZ* gate applied halfway through the cross-Ramsey sequence. After accounting for rectification of the qubit frequencies, we observe a π phase shift between the two possible Ramsey oscillations. Cross-Ramsey experiments are helpful in the calibration procedures for optimizing *CZ* gates because the rectification of the qubit frequencies is dependent on ω_p and $\delta\phi_p$ (for more details, see Appendix E). In general, we have performed cross-Ramsey experiments to tune up all the *CZ* gates in this work (see Appendix I).

To quantify the p-SWAP *CZ* gate fidelity, we perform a standard interleaved RB procedure repeatedly over many hours [7,60]. To improve accuracy in estimating the two-qubit *CZ* gate fidelity, we have chosen to use a less costly form of interleaved RB that involves interleaving two-qubit *CZ* gates between simultaneously applied randomly chosen single-qubit Clifford ($C1$) gates (see Appendix I). This is a practical trade-off between constructing two-qubit Clifford ($C2$) gates from a relatively large number of single-qubit and *CZ* gates, which takes more time when compared to the relatively short energy-relaxation times of our two qubits (approximately 15 μs) [4,16,61–65]. In Fig. 5(a), we show the measured total error per gate ε_t as a function of the gate duration t_g (see the pulse diagram in Fig. 4(c), as t_g is defined by the edge-to-edge pulse width). First, we focus on pulse envelopes made from two fixed 20-ns Hann-function rise-and-fall pulse edges plus a variable-length plateau, where each RB data point has been averaged for about 9 h. For each gate duration, we have performed a suite of calibrations and we use RB [66] plus a covariance matrix adaptation evolution strategy (CMA ES) for numerical optimization [67] of the gate performance (for more details, see Appendix I).

In general, the total error per gate $\varepsilon_t = \varepsilon_{\text{dec}} + \varepsilon_{\text{con}}$ can be broken into decoherence-induced errors ε_{dec} and any control errors ε_{con}, which could include noise in the control electronics or coherent evolutions from unintended Hamiltonian dynamics. We observe that when the gate duration is long ($t_g > 150$ ns), the error rate is dominated by the individual qubit coherences, where $\varepsilon_{\text{dec}} = (2/5)(\Gamma_{1,\text{eff}} + 2\Gamma_{2,\text{eff}})t_g$, with $\Gamma_{1,\text{eff}} = (1/2)(1/T_{1,L} + 1/T_{1,R})$ and $\Gamma_{2,\text{eff}} = (1/2)(1/T_{2,L} + 1/T_{2,R})$ [14,68]. The dashed lines in Fig. 5(a) show the prediction when $T_{1,\text{eff}} = 1/\Gamma_{1,\text{eff}} = 16.3$ μs and $T_{2,\text{eff}} = 1/\Gamma_{2,\text{eff}} = 22.7$ μs, consistent with measured qubit coherences (for more details, see Appendix F). We also plot the individual contributions from $\Gamma_{1,\text{eff}}$ (dotted line) and $\Gamma_{2,\text{eff}}$ (dashed-dotted line), which shows the comparative roles that energy relaxation and phase coherence play in determining gate errors. This suggests that as t_g decreases with respect to the decoherence rates, the gate error should decrease steadily toward zero. However, our experimental results show that as t_g decreases, the observed ε_t deviates from the decoherence line when ε_{con} increases.

As a result, ε_t reaches a minimum value at $t_g = 70$ ns, where we achieve our best p-SWAP *CZ* gate fidelity of $99.44 \pm 0.09\%$, as shown in Fig. 5(b). If we compare this to the fidelity of an idle gate in place of the p-SWAP *CZ* gate, we find a much improved value of 99.88%, which suggests that $\varepsilon_{\text{con}} = 1.2 \times 10^{-3} < \varepsilon_{\text{dec}}$. This shows that simply idling during the gate in the presence of decoherence produces $\varepsilon_{\text{dec}} = 4.4 \times 10^{-3}$, close the theoretical decoherence-limit line when $t_g = 70$ ns ($\varepsilon_{\text{dec}} = 4.2 \times 10^{-3}$). As gate durations decrease further, $\varepsilon_{\text{con}} > \varepsilon_{\text{dec}}$ dominating the total error.

In order to investigate sources for ε_{con}, we have carried out a series of gate-fidelity measurements with variations in the shape of the pump-pulse envelope of the gate. Let us first consider that when $t_g = 80$ ns, the pulse envelope is made from equal durations of Hann-function pulse edges and a plateau. As the pulse width decreases further toward $t_g = 40$ ns, the spectrum of these pulses in frequency space

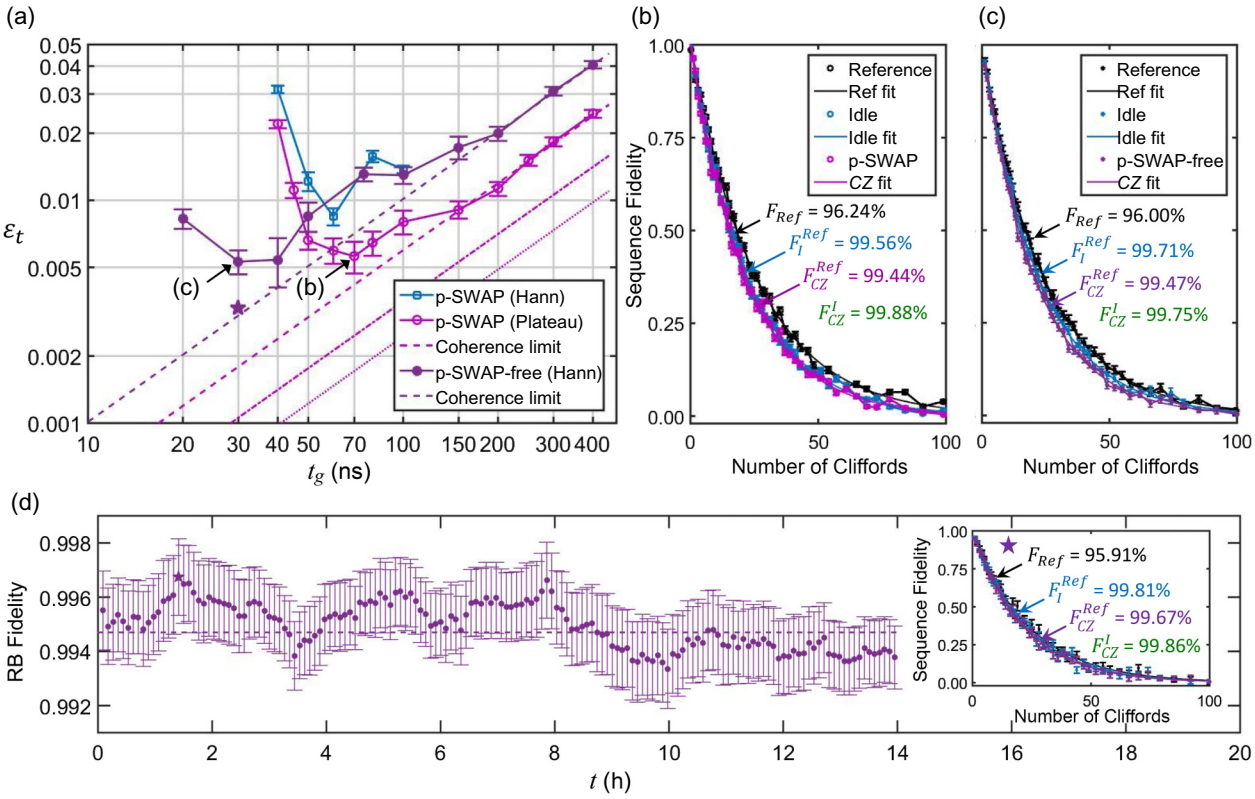


FIG. 5. Demonstration of parametric *CZ* gates. (a) RB for various parametric *CZ* gates as a function of the gate duration t_g. The dashed lines represent coherence limits. The individual contributions from decoherence are shown for energy relaxation ($\Gamma_{1,\text{eff}}$, dotted line) and phase coherence ($\Gamma_{2,\text{eff}}$, dashed-dotted line) for the plateau pulses. ⋆ shows the peak fidelity [see (d)]. (b) RB showing the best result for the p-SWAP *CZ* gate with fidelity $99.44 \pm 0.09\%$ for $t_g = 70$ ns when averaged over 9 h. (c) RB showing the result for the p-SWAP-free *CZ* gate with fidelity $99.47 \pm 0.07\%$ for $t_g = 30$ ns when averaged over about 18 h. (d) A running average of the p-SWAP-free *CZ* gate of duration 30 ns over 14 h with a peak fidelity of $99.67 \pm 0.14\%$ (⋆). The dashed horizontal line shows the average value [see (c)]. The inset shows RB results for the peak fidelity (⋆). All thin solid lines are RB fits to the data (see the text).

becomes wider, following more closely the Fourier transform of a pure Hann pulse, which is a broadened version of a rectangular window function. Thus, if a narrowing spectral width (decreasing t_g) leads to leakage errors as a limiting factor (increasing ε_{con}), then the pure Hann pulses would improve our gate fidelity. However, experiments show the opposite.

First, we have independently verified that leakage to higher transmon levels has not been a significant source of error [7]. Next, we have carried out *CZ*-gate experiments where the pump-pulse envelope has been a pure Hann function of different widths, as shown in Fig. 5(a). Note that this group of data have a total error per gate that is systematically higher than in the experiments with plateaued pulses for similar qubit coherence. We can understand the source of these errors by returning to the concept of how this gate functions.

Ideally, starting from the $|ee\rangle$ state, the parametric pump pulse operates at exactly $\omega_p = \Delta_{bt} = \omega_{ee} - \omega_{fg}$, so that the state evolution under the Hamiltonian [from Eq. (2)] occurs at the center of the Rabi curve [from Eq. (3), shown in Fig. 4(a)] returning all of the state population back to $|ee\rangle$ with a final phase accumulation of exactly π. In reality, because the frequencies of both qubits, their anharmonicities, and ζ_s all experience a slight adjustment due to the rectification effect that depends on pump amplitude $\delta\phi_p$ and frequency ω_p, it is difficult to maintain a perfect evolution of the states and cancellation of ζ_s (for more details, see Appendix E).

For our experiments, we have chosen a fixed cancellation tone for ζ_c to ensure proper single-qubit gates in the absence of any gate tones. For the gate tone, we have used a constant frequency $\omega_p = \tilde{\omega}_{ee} - \tilde{\omega}_{fg}$ set according to the rectified values of the two-qubit states at the maximum pump amplitude, as measured during parametric Rabi oscillations as we have swept Δ_p (see, e.g., Fig. 4 and Appendix E). Thus, during the rise and fall of the Hann-pulse envelope, the frequency ω_p is slightly detuned from optimal.

For low-amplitude pump pulses mostly comprised of a pulse plateau (larger t_g), the evolution of the gate matches closely with the idealized case and the residual

ZZ coupling is still negligible. On the contrary, for pure Hann pulses, the pump frequency only matches at the maximum amplitude of the pulse, modifying the Hamiltonian evolution of the gate [recall Eq. (3)]. In addition, as the pulse amplitude increases for shorter t_g, the static ζ_c tone is not completely cancelling ζ_s during the gate, although this effect is compensated for by the reduced gate duration over which it occurs. These problems could be properly addressed by varying the drive frequency $\omega_p(t)$ and cancellation tone $\zeta_c(t)$ over the duration of the gate, which we plan to explore in future experiments.

VIII. SWAP-FREE PARAMETRIC GATES

In order to remedy these issues, we next turn to a parametric *CZ* gate that is "SWAP-free." A nice feature of the p-SWAP-free *CZ* gate is that there is no requirement to optimize a single parametric drive frequency, making this gate more resilient to rectification-induced errors. Recall that in Fig. 2(e), we could cancel ζ_s or choose to drive harder so that $\zeta_s + \zeta_c > 0$, producing strong positive *ZZ* coupling. This is a much more natural way to realize a *CZ* gate [10], where a net frequency shift $\zeta_{cZ}(t)$ of the state $|ee\rangle$ over time t_g can accumulate the necessary conditional phase,

$$\pi = \int_0^{t_g} \zeta_{cZ}(t)dt. \tag{4}$$

To perform the gate, we add an additional tone $\zeta_g(t)$ along with the constant cancellation tone ζ_c, so that $\zeta_{cZ}(t) = \zeta_g(t) + \zeta_s + \zeta_c$. Any slight deviations over time in Δ_{bt} during the gate pulse due to rectification may lead to distortions so that $\zeta_{cZ}(t)$ is not optimal, but we only require the integral in Eq. (4) to be satisfied. As a first approximation, this can be compensated by a small increase or decrease in the gate duration. This means that we can use pure Hann pulses without any plateau and not suffer from additional errors from pump-induced rectification.

After performing a suite of calibrations and using the CMA ES (for more details, see Appendix I), we have measured ε_t for various t_g, in a similar way as for the p-SWAP *CZ* gate. We have found that for long gate durations, the p-SWAP-free *CZ* gate is limited by the combined decoherence of the two qubits, with $T_{1,\text{eff}} = 11.6\ \mu\text{s}$ and $T_{2,\text{eff}} = 12.0\ \mu\text{s}$ for this particular cooldown [see the dashed curve in Fig. 5(a) and see Appendix F for more details].

Again, as t_g is reduced, the error rate decreases following the decoherence prediction and then begins to deviate. The minimum gate error is observed to be $99.47 \pm 0.07\%$ with a pure Hann pulse with a total duration (rise and fall) of 30 ns, as is shown in Fig. 5(c). Again, we can also compare this to the fidelity of an idle gate in place of the p-SWAP-free *CZ* gate, which improves to 99.76%, showing that $\varepsilon_{\text{con}} = 2.4 \times 10^{-3} < \varepsilon_{\text{dec}}$. Here, we find that $\varepsilon_{\text{dec}} = 2.9 \times 10^{-3}$ is close to the theoretical decoherence-limit line when $t_g = 30$ ns ($\varepsilon_{\text{dec}} = 3.0 \times 10^{-3}$).

In Fig. 5(d), we plot the p-SWAP-free gate fidelity over time, with each point representing a 4-h moving average of a roughly 18-h-long repeating data set. We see that the maximum gate fidelity peaks at $99.67 \pm 0.14\%$ with respect to the reference (see $\star$ and inset). At this optimal point, with respect to the idle, the fidelity is 99.86% so that $\varepsilon_{\text{con}} = 1.4 \times 10^{-3}$ and $\varepsilon_{\text{dec}} = 1.9 \times 10^{-3}$, better than the theoretical decoherence-limit line when $t_g = 30$ ns ($\varepsilon_{\text{dec}} = 3.0 \times 10^{-3}$). Recall that this limit is assessed as an average over long periods of time, showing that coherence fluctuations also produce variations in gate fidelity over time, most likely due to two-level system (TLS) defects (for more details, see Appendix F). That said, the p-SWAP-free gate fidelity has peaked twice and has remained almost entirely above the mean value for over 8 h [see the dashed line and Fig. 5(c)].

IX. DISCUSSION OF DRIVE-INDUCED ERRORS

For both the p-SWAP and p-SWAP-free *CZ* gates, the gate error rises rapidly for the shortest gate durations. To investigate this, we have performed some high-amplitude pump-drive tests, where the initial two-qubit state has been $|gg\rangle$ (for more details, see Appendix C). We have found that it is possible for the parametric pump to drive multiphoton-induced single-qubit transitions when $\omega_p \approx \omega_k/n$, where n is the number of photons absorbed. For the highest pump amplitudes, if ω_p is in the vicinity of these frequencies, the pump can cause unwanted single-qubit Rabi rotations. In principle, these coherent single-qubit errors could be nulled out with additional deterministic single-qubit drives that could counteract these stray rotations [9,15]. In future designs, we plan to significantly decrease the capacitive coupling between the bias coil and the transmons in order to reduce any direct driving and resultant multiphoton absorption [40].

X. SCALING UP TO MULTIQUBIT ARCHITECTURES

In the way in which we have constructed the galvanic SQUID coupler, the cancellation properties between two qubits can be extended to three or more qubits to achieve all-to-all parametric couplings within a given cluster [34]. The induced parametric interactions are valid whether the coupled resonant circuits are completely linear (a cavity) or nonlinear (a qubit or qudit). As long as these elements have well-defined energy spectra that govern their excitations, then parametric couplings can help exchange energy between these elements. Thus, clusters can include mixtures of qubits and cavities. This has already been achieved in three-element systems that have included a lumped-element cavity as a thermal reservoir or for joint two-qubit

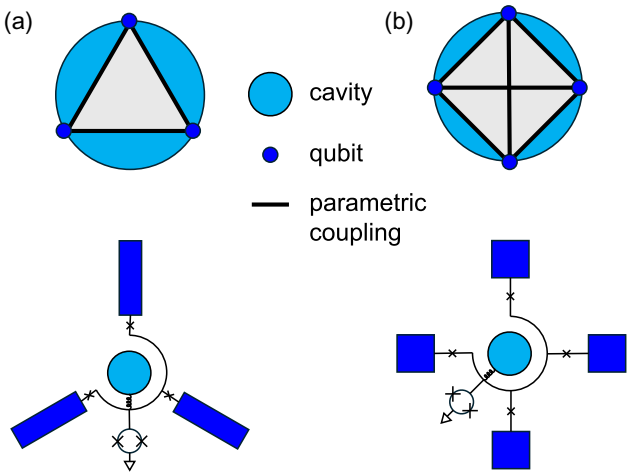


FIG. 6. Scaled-up clusters. Rough concept drawings for (a) three- and (b) four-qubit clusters with a parametrically coupled readout cavity.

readout [19,21]. To include additional transmons in our circuit topology, we simply tie together the current flowing through each additional transmon capacitor and its main Josephson junction into a single shared SQUID coupler. Two schematic examples are shown in Fig. 6 for three and four qubits, each including a cavity for readout. By sharing all the currents in this way and ensuring that the small Josephson junctions in the transmons have nominally identical critical currents, the inductive coupling from the galvanically connected SQUID coupler can then cancel all of the cross capacitances between the shared transmons (and the cavity). This is achieved by designing a lithographic layout that realizes the desired capacitance matrix that sets the capacitive coupling strengths. Achieving static cancellation helps to minimize both qubit crosstalk and unintended interactions, including coherence protection from the readout cavity [19].

Ideally, at the cancellation flux, all the transmons are effectively uncoupled with only minimal *ZZ* interactions. As discussed earlier, it is possible to eliminate static *ZZ* completely either with some transmons in the straddling regime or by combining flux-type qubits with positive anharmonicity with transmons (see Appendix A). Even without full cancellation, it is possible to cope with residual *ZZ* interactions by adding parametric dispersive shifts, as presented in the two-qubit experiments discussed already. In Ref. [19], it has been shown that these dispersive parametric techniques are additive, so that including additional relatively weak parametric microwave tones can completely cancel any residual *ZZ* couplings between the qubits (and cavities).

With parametric driving, we can generalize Eq. (2) for any two qubits (or cavities) a and b within a cluster. We can generate parametric coupling terms between any two of the n-qubit states $|b\rangle$ and $|t\rangle$ within a given excitation manifold by driving at the difference frequency between those states, following

$$H_{Ip} = \hbar g_p \left(e^{i\omega_p t} + e^{-i\omega_p t}\right)\left(q_a q_b^\dagger e^{-i\Delta_{bt} t} + \text{h.c.}\right) \quad (5)$$

Again, when $\Delta_p = 0$, $\omega_p = \Delta_{bt}$ and we have $H_{Ip} = \hbar g_p (q_a q_b^\dagger + q_a^\dagger q_b)$, which provides a SWAP interaction between only those two qubits (or cavities) a and b within the cluster. This then extends the two-qubit *CZ* gate strategies presented earlier within a larger cluster of qubits. In addition, n-qubit SWAP gates can also be achieved when they lie within the same excitation manifold of states (e.g., for $n = 3$, we could SWAP between $|eee\rangle \leftrightarrow |gef\rangle$), which is useful to extend *CZ* gates to larger n, such as the Toffoli gate. Simultaneous driving of single-qubit and parametric operations can also provide for a rich and predictable layering of multiqubit interactions.

In an effort to scale up this design concept, there are two major technical challenges in creating a single "cancellation" point or a relatively narrow region in flux for operating the device with the best performance. The first challenge is to carefully arrange the geometry of the capacitors and the second is to obtain Josephson junctions with the minimum amount of spread in their critical currents. At present, the creation of superconducting qubits owes most of its imprecision to the uncontrolled fabrication processes used for making small Josephson junctions, while the fabrication of planar capacitors is relatively accurate. As noted in Ref. [69], photolithographically defined capacitors have large dimensions (hundreds of micrometers) and small variations between qubits, while small Josephson junctions (100×100 nm^2) only have a precision of a few percent for obtaining targeted critical currents. However, it is still important to identify and incorporate parasitic inductive and capacitive couplings resulting from the lithographic layout.

Considering the known imperfections introduced by fabrication, the flexibility afforded by our design strategy matched with parametric operations has numerous benefits. Current parameter variations can cause each pair of elements in a multiqubit cluster to have unique cancellation conditions. This causes the minimum (including zero) cross-couplings for all pairs of elements to have some misalignment in flux. However, with a few percent variation expected, there will be a global minimum value of all cross-couplings at a particular flux where all the curves overlap. If we use this as the operation flux, the static couplings will have some overall increase compared to the designed values, but additional parametric tones can still remove these additional couplings. In addition, there is still flexibility in deciding what pair couplings to optimize for and adjust the operating flux as desired to maximize the cluster performance.

An important advantage of tunable parametric interactions is that the specific frequencies of the qubits and

cavities do not have to be extremely accurate. For larger clusters of qubits, we only require that the frequencies fall into a certain bandwidth of the order of a few gigahertz. The parametric approach provides more flexibility, as the pump frequencies required for operation reside near the difference (or sum) frequency of each qubit-qubit or qubit-cavity pair. The only requirement is for these transition frequencies to be separated from each other by about the strongest parametric coupling strength of approximately 100 MHz. As discussed above, a three-qubit gate can also be achieved by targeting the difference frequencies between any three-qubit states, which allows one to perform multiqubit gates such as the Toffoli gate [70] and other arbitrary n-qubit interactions that could be useful for quantum simulation or variational quantum eigensolvers [71,72].

At this time, it is not clear, given current fabrication restrictions that maintain low dielectric losses, exactly how far this strategy can be expanded beyond five qubits. For more elements coupled through a single SQUID, it may require fabricated crossovers, multilayers, or interposer layers, and computer-aided design software matched with optimization algorithms. However, these clusters could be coupled together to produce larger structures. In this way, there could be optimized interactions within each cluster with an increase in the qubit connectivity, while reducing static cross-couplings, as well as providing an *in situ* multiplexed or joint-qubit readout [19]. Optimizing the interaction between clusters and even moving beyond transmons is in our roadmap for future work.

In Fig. 7, we show a design concept for using tunable transmon couplers between clusters of three and four qubits, plus a cavity for multiplexed readout. The transmon couplers connect just two qubits within separate clusters with capacitive coupling [34]. It is important to utilize capacitive couplings in larger structures to avoid creating large superconducting loops when connecting clusters. If excess magnetic flux penetrates these parasitic loops, they can generate unwanted circulating currents.

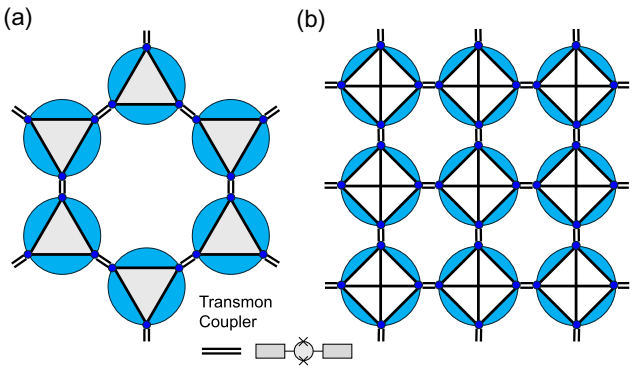


FIG. 7. Scaled-up clusters. (a) A three-qubit cluster with a readout cavity. (b) A four-qubit cluster with a readout cavity.

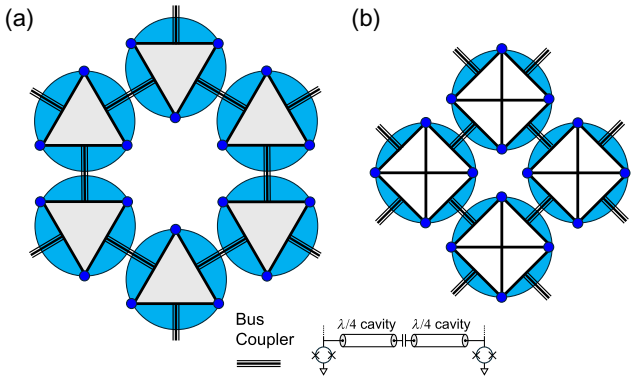


FIG. 8. Scaled-up clusters. (a) A three-qubit cluster with a readout cavity. (b) A four-qubit cluster with a readout cavity.

Not only can these currents introduce flux noise, but if they pass through Josephson junctions anywhere in the circuit topology, it will change their offset phase corrupting the designed circuit operation. Finally, large loops can also produce unintended cross-coupling of flux drive ports, which can severely complicate the operation of the whole architecture. Note that even with these small clusters, the number of programmable couplings is large. There are many possible lattice configurations that become software programmable. In the case of four qubits with a cavity for readout, one could perform parity readouts of all four qubits within a cluster. This could be a more efficient way to perform stabilizer checks for quantum error-correction protocols.

In order to wire up more connections, we propose a parametric bus coupler between different clusters (see Fig. 8). Here, the parametric bus coupler [24] is made from two $\lambda/4$ cavities capacitively coupled together. This leads to mode splitting of the fundamental mode into two possible quantum buses [73,74] for connecting the clusters together. The $\lambda/4$ cavities are galvanically grounded through the SQUIDs of their respective clusters. This provides parametric coupling between the qubits (and cavity) of the cluster and the bus. If the bus connects cluster A with cluster B, then parametric driving of the SQUIDs of both clusters at the proper difference frequencies can provide coupling $A_i \leftrightarrow B_j$, for $i,j \in \{1,2,\ldots,4\}$. These interactions with the bus could be performed sequentially [74] or simultaneously [73]. Here, the bus coupler provides even more connectivity between the clusters. This results in more possible lattice configurations with dynamic time-dependent couplings.

Finally, we discuss some expected trade-offs in performance versus the difficulties in extending this architecture to multiqubit clusters and lattices. Performance improvements would result from having small clusters of qubits with all-to-all coupling. This general increase in connectivity through parametric coupling means that multiqubit (or qubit-cavity) interactions are turned on or off

by microwave tones. These tones are programmable with software in a similar way to ion-trap or even neutral-atom approaches. This strategy provides an arbitrary multiqubit topology or Hamiltonian that is not "hard wired" by fixed capacitive connections. Here, the interactions are fully programmable by the user in time with carefully orchestrated parametric pulse streams.

The coupling of more qubits together with all-to-all coupling does come at a cost. First, achieving the correct capacitance matrix to provide cancellation of *ZZ* coupling becomes more involved as more qubits are added to each cluster and must be engineered carefully. Second, exponential scaling in the number of interaction frequencies can lead to parametric frequency collisions, where multiple transition frequencies lie near each other. This requires more care to avoid parametric transition overlap, but with the growing number of available gate transitions, ignoring a few problematic transitions or operations does not rule out a large amount of retained functionality. Third, the qubit frequencies of each cluster must be cycled properly throughout the architecture in order to reduce direct frequency collisions between qubits, cavities, or parametric transitions. This ability to recycle frequency zones has already been achieved with many of the industry processors already [4,10,46]. Finally, any failures in static *ZZ* cancellation would require additional low-amplitude detuned parametric drives to each cluster. As long as the these tones are low amplitude and crosstalk between similar clusters is low, we expect this technique to be successful.

XI. CONCLUSIONS

To conclude, we have demonstrated a versatile, hardware-efficient, fast, and high-performance parametric coupler between two mildly tunable transmon qubits. We have been able to eliminate stray *ZZ* coupling between the two qubits and we have performed three types of two-qubit gates: p-iSWAP, p-SWAP *CZ*, and p-SWAP-free *CZ*. We have measured an average gate error over many hours as a function of the gate duration and found maximum gate fidelities above 99% for the parametric *CZ* gates. The p-SWAP-free *CZ* gate has reached a peak fidelity of $99.67 \pm 0.14\%$ for a gate duration of 30 ns and remained above 99.47 for over 8 h. For longer gate durations above 150 ns, the gate error is almost entirely limited by qubit decoherence, while the shortest gates are limited mostly by erroneous Hamiltonian evolutions. Exploring different pulse shapes indicates that a significant portion of the control errors in the p-SWAP *CZ* gate can be ascribed to a rectification effect, which can be avoided when performing p-SWAP-free *CZ* gates.

Our results suggest that modest improvements in qubit coherence in next-generation devices should lead to parametric two-qubit gates with fidelities above 99.9%. In addition, future circuit designs with transmons should reduce residual static *ZZ* coupling to <300 kHz, whereas introducing flux-type qubits [52,53] provides exact cancellation of all residual static *ZZ* coupling. These architectural concepts can be scaled into integrated multiqubit cavity-QED systems to provide *in situ* parity readout useful for error correction [11,19] or stabilizing multiqubit entanglement [21].

ACKNOWLEDGMENTS

This work was partially performed under financial assistance from the U.S. Department of Commerce and the National Institute of Standards and Technology. The work of E.K. was supported by National Science Foundation (NSF) Grant No. PHY-1653820 and Army Research Office (ARO) Grant No. W911NF-18-1-0125. We thank Frederick W. Strauch for useful discussions and suggestions for improving gate fidelity in the experiments. We thank Manuel C. Castellanos-Beltran and Joshua Combes for commenting on the manuscript.

DATA AVAILABILITY

The data that support the findings of this paper are available from the authors upon reasonable request.

APPENDIX A: BASIC CIRCUIT MODEL

We can approximate each transmon frequency by

$$\omega_k \approx \frac{1}{\sqrt{(\mathcal{L}_k + L_s)C_k}} - E_{ck}/\hbar, \tag{A1}$$

where C_k represents the total capacitance (including any additional mutual capacitance) for each transmon, for $k \in \{L, R\}$, and $E_{ck} = (2\pi\hbar/\Phi_o)^2/(8C_k)$ is the charging energy for each transmon. At zero flux, where the SQUID inductance contributes about 4% of the total inductance of the transmon, we can neglect the contribution of the SQUID completely and extract the transmon junction inductance $\mathcal{L}_k$ and capacitance from measurements of the transmon transition frequency and anharmonicity. Here, we have

$$\mathcal{L}_k = \frac{1}{C_k}\frac{1}{(\omega_k + E_{ck}/\hbar)^2}, \tag{A2}$$

$$C_q = \frac{(2\pi\hbar/\Phi_o)^2}{8E_{ck}}. \tag{A3}$$

The $\mathcal{L}_k$ of each transmon has been made with nominally the same critical-current Josephson junction. The SQUID inductor has been made from two nominally identical SQUID junctions within a loop, each with a critical current that is η_s times larger than the main transmon junctions. As seen in the next section with a more detailed model, fits give $\eta_s = 16.3$, which is close to the as-drawn design value

of 15.2 for the area ratio of the SQUID and the transmon junctions.

Because transmons are nearly linear oscillators, we can use a simple model for linear coupled oscillators. In general, the two transmons share the dc-SQUID inductor, which leads to a sharing of current that provides an inductive coupling strength g_L. In addition, the large capacitors of the two transmons being relatively close together lead to a capacitive coupling with strength g_C. These coupling strengths are roughly given by

$$g_L = \frac{1}{2}\frac{L_s}{\sqrt{(\mathcal{L}_L + L_s)(\mathcal{L}_R + L_s)}}\sqrt{\omega_L \omega_R}, \tag{A4}$$

$$g_C = \frac{1}{2}\frac{C_{LR}}{\sqrt{(C_L + C_{LR})(C_R + C_{LR})}}\sqrt{\omega_L \omega_R}, \tag{A5}$$

where L_s and ω_k are flux dependent. The total static coupling is the difference of the inductive and capacitive contributions, $g_s = g_L - g_C$. As discussed in the main text, at a particular flux, there is a SQUID inductance L_s where $g_L = g_C$ and $g_s = 0$.

Equation (1) results from the assumption of only slow RWA terms in the combined inductive and capacitive couplings. A main source of the residual nonzero ZZ coupling can be attributed to counterrotating-wave (cRWA) terms in the coupled interactions.

Modeling the L and R transmon qubit system as Kerr oscillators with linear inductive (g_L) and capacitive (g_C) couplings leads to

$$\begin{aligned}\hat{H} &= \omega_L \hat{q}_L^\dagger \hat{q}_L + \frac{1}{2}\alpha_L \hat{q}_L^{\dagger 2}\hat{q}_L^2 + \omega_R \hat{q}_R^\dagger \hat{q}_R + \frac{1}{2}\alpha_R \hat{q}_R^{\dagger 2}\hat{q}_R^2 \\ &\quad + g_C(\hat{q}_L - \hat{q}_L^\dagger)(\hat{q}_R - \hat{q}_R^\dagger) + g_L(\hat{q}_L + \hat{q}_L^\dagger)(\hat{q}_R + \hat{q}_R^\dagger) \\ &= \omega_L \hat{q}_L^\dagger \hat{q}_L + \frac{1}{2}\alpha_L \hat{q}_L^{\dagger 2}\hat{q}_L^2 + \omega_R \hat{q}_R^\dagger \hat{q}_R + \frac{1}{2}\alpha_R \hat{q}_R^{\dagger 2}\hat{q}_R^2 \\ &\quad + (g_L - g_C)(\hat{q}_L^\dagger \hat{q}_R + \hat{q}_L \hat{q}_R^\dagger) + (g_L + g_C)(\hat{q}_L \hat{q}_R + \hat{q}_L^\dagger \hat{q}_R^\dagger).\end{aligned} \tag{A6}$$

Fully diagonalizing the Hamiltonian to second order in g_L and g_C gives [75] a cross-Kerr (or ZZ) term $\zeta_s \hat{q}_L^\dagger \hat{q}_L \hat{q}_R^\dagger \hat{q}_R$, with

$$\begin{aligned}\zeta_s/2 = &\frac{(g_L - g_C)^2}{\omega_R - \omega_L - \alpha_L} + \frac{(g_L - g_C)^2}{\omega_L - \omega_R - \alpha_R} \\ &- \frac{(g_C + g_L)^2}{\omega_L + \omega_R + \alpha_L + \alpha_R} + \frac{(g_C + g_L)^2}{\omega_L + \omega_R},\end{aligned} \tag{A7}$$

where $|\alpha_R| > |\alpha_L|$ and $\omega_R > \omega_L$. Defining $\Sigma = \omega_L + \omega_R$ and $\Delta_s = \omega_R - \omega_L$, one can write a simplified expression:

$$\zeta_s/2 = \frac{(g_L - g_C)^2(\alpha_L + \alpha_R)}{(\Delta_s - \alpha_L)(\Delta_s + \alpha_R)} + \frac{(g_L + g_C)^2(\alpha_L + \alpha_R)}{\Sigma(\Sigma + \alpha_L + \alpha_R)}. \tag{A8}$$

The first term of Eq. (A8) arises from the RWA of the coupling interaction and has been derived in other papers [50].

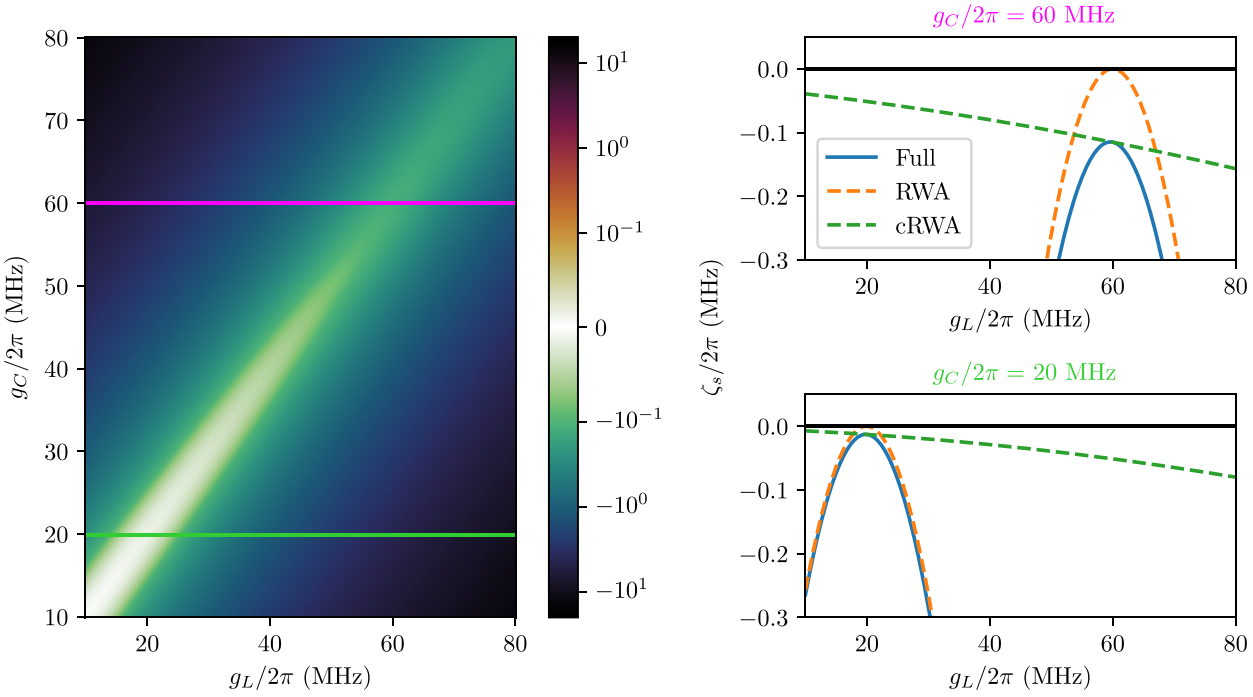


FIG. 9. The increased ζ_s minimum with increasing g_C for fixed qubit frequencies $\{\omega_L, \omega_R\}/2\pi = \{5.033, 5.627\}$ GHz and anharmonicities $\{\alpha_L, \alpha_R\}/2\pi = \{-194, -242\}$ MHz. The line cuts show the contributions of the rotating-wave (RWA) and counterrotating-wave (cRWA) terms to the total ζ_s.

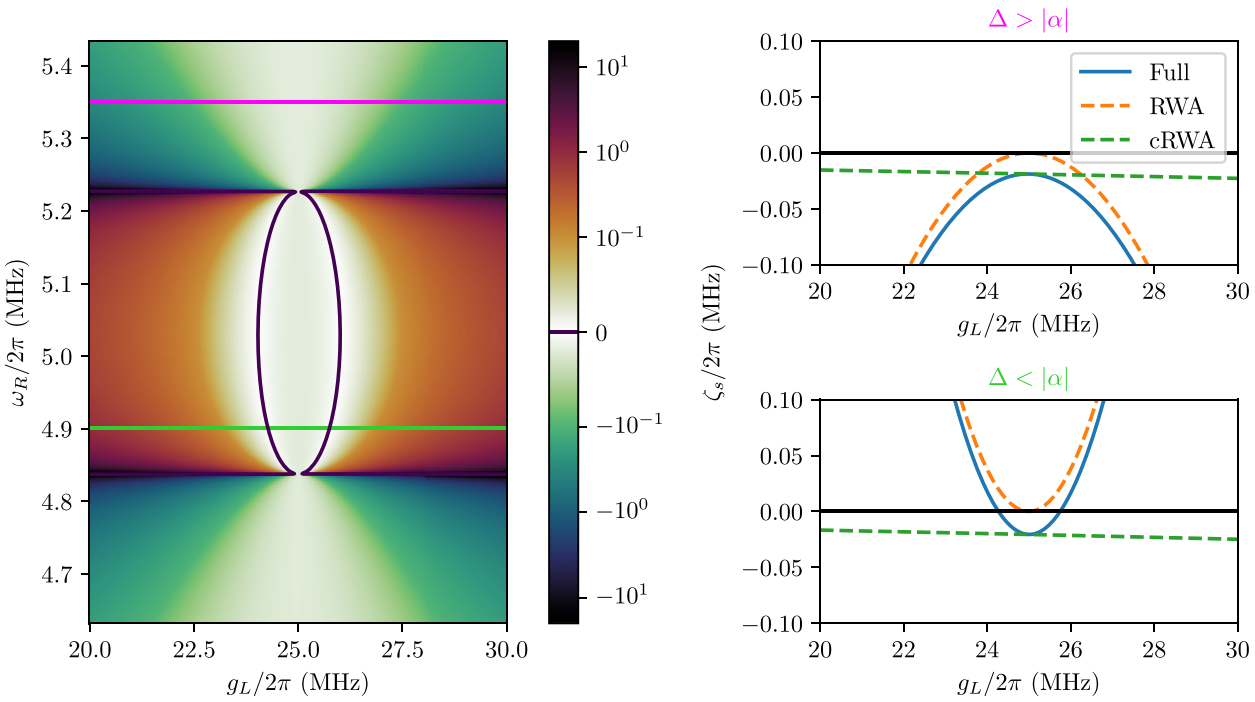


FIG. 10. The evaluation of Eq. (A8) for different values of the qubit frequency ω_R and inductive coupling g_L for fixed qubit frequency $\omega_L/2\pi = 5.033$ GHz, capacitive coupling $g_C/2\pi = 25$ MHz, and anharmonicities $\{\alpha_L, \alpha_R\}/2\pi = \{-194, -196\}$ MHz. The line cuts show the contributions of RWA and cRWA terms to the total ζ_s in and out of the straddling regime.

Often, the second term is ignored because it is generally quite small $\propto (g_L + g_C)^2/\Sigma^2 \ll 1$; but in our system, due to the two sources of coupling canceling in the first term at a specific flux bias ϕ_c, it is important to consider the size of the cRWA term proportional to $(g_L + g_C)^2$. The necessary inclusion of the cRWA terms has been noted with other transmon couplers when *ZZ* cancellation is important [9].

The residual nonzero value of ζ_s scales quadratically with the combination of inductive and capacitive couplings, as seen in Fig. 9. In the RWA-based design analysis, stronger parametric coupling comes with increased inductive coupling g_L, which needs to be balanced by an increase in capacitive coupling g_C in order to ensure a cancellation at ϕ_c. This necessarily increases the size of the cRWA terms.

Although the experimental devices discussed in this work have not achieved complete *static* cancellation of *ZZ* coupling, our design strategy can achieve this. In the so-called "straddling regime," where $\Delta_s < |\alpha_R|$, the first term of Eq. (A8) becomes positive (for transmons where $\alpha_k < 0$, for $k \in \{L, R\}$) while the second term stays negative, leading to two possible zero crossings (see Fig. 10). This condition is utilized in numerous capacitively coupled transmon couplers to achieve a zero crossing for *ZZ* coupling [9,51]. Although the model Hamiltonian of Eq. (A8) only approximates our galvanic SQUID coupling scheme, a full numerical circuit diagonalization of the more detailed circuit model in Appendix B also shows an exact cancellation of *ZZ* coupling with two locations where $\zeta_s = 0$, with appropriately chosen circuit parameters in the straddling regime. Finally, note that Eq. (A8) suggests another strategy for cancellation of *ZZ* coupling, namely, making $\alpha_L = -\alpha_R$. This can be achieved with flux qubits with positive anharmonicity coupled to transmons [52,53].

We can investigate the relative strength of the parametric coupling at each flux bias [24,25] by calculating the derivative of g_s with respect to ϕ, which contains both the inductive and capacitive contributions within the RWA,

$$\frac{\partial g_s}{\partial \phi} = \tag{A9}$$

$$= \left[\frac{1}{L_s}\frac{\partial L_s}{\partial \phi} + \frac{3}{2}\frac{1}{\omega_L}\frac{\partial \omega_L}{\partial \phi} + \frac{3}{2}\frac{1}{\omega_R}\frac{\partial \omega_R}{\partial \phi}\right] g_L - \left[\frac{1}{2}\frac{1}{\omega_L}\frac{\partial \omega_L}{\partial \phi} + \frac{1}{2}\frac{1}{\omega_R}\frac{\partial \omega_R}{\partial \phi}\right] g_C$$

$$\approx \left[\frac{1}{L_s}\frac{\partial L_s}{\partial \phi}\right] g_L, \tag{A10}$$

since $(1/\omega_k)(\partial \omega_k/\partial \phi) \ll (1/L_s)(\partial L_s/\partial \phi)$, so that

$$\frac{\partial g_s}{\partial \phi} \approx \frac{1}{2}\frac{\partial L_s}{\partial \phi}\frac{1}{\sqrt{(\mathcal{L}_L + L_s)(\mathcal{L}_R + L_s)}}\sqrt{\omega_L \omega_R}, \tag{A11}$$

and from Eq. (A1), $\partial\omega_k/\partial\phi \approx -(\omega_k/2)(\partial L_s/\partial\phi)/(\mathcal{L}_k + L_s)$ for $k \in \{L, R\}$, giving

$$\left|\frac{\partial g_s}{\partial\phi}\right| \approx \sqrt{\left|\frac{\partial\omega_L}{\partial\phi}\right|\left|\frac{\partial\omega_R}{\partial\phi}\right|}. \tag{A12}$$

Then, the parametric coupling strength can be estimated from [25]

$$g_p = \frac{1}{2}\left|\frac{\partial g_s}{\partial\phi}\right|\delta\phi_p \approx \frac{1}{2}\sqrt{\left|\frac{\partial\omega_L}{\partial\phi}\right|\left|\frac{\partial\omega_R}{\partial\phi}\right|}\delta\phi_p. \tag{A13}$$

This shows that the parametric coupling strength is directly related to the size of the frequency modulation of both qubits with flux. This is important when considering that the phase coherence of the qubits is inversely proportional to the same factors in the above expression for the parametric coupling strength. Thus, there is a tradeoff between achieving large parametric coupling strengths for relatively small flux modulations and maintaining high phase coherence in the presence of flux noise in the SQUID coupler (see Appendix F below).

APPENDIX B: DETERMINATION OF THE FULL CIRCUIT MODEL AND PARAMETERS

The main circuit parameters have been determined by considering the lumped-circuit model shown in Fig. 1(b) in the absence of the coplanar-waveguide readout cavity, apart from unavoidable capacitive loading to both the transmons and their mutual capacitive coupling. These effects are simply absorbed into the remaining capacitive parameter values used in the final circuit model. Here, the coplanar-waveguide cavity coupling capacitances (C_{CL} and C_{CR}) are explicitly accounted for in the two corresponding transmon shunting capacitances (C_L and C_R). To get a more accurate fitting of the qubit transition frequencies, it is important to consider the additional geometric inductance of the 8-μm-wide wires connecting the Josephson junction of each transmon to the roughly rectangular capacitor electrodes. A more detailed lumped-circuit schematic is shown in Fig. 11.

In this model, although not explicitly shown in Fig. 11, the Josephson-junction capacitances are initially assumed to be 1 fF for the qubit junctions and 18 fF for the SQUID junctions based on their relative areas. Their precise values do not significantly effect the fit; rather, they are set to reasonable values and included in the circuit model, which is necessary for treating the junctions in the charge basis. The remaining capacitance values are determined from the device layout. Using the ANSYS Q3D EXTRACTOR software, we can extract the capacitance matrix of the system from the design layout GDS file for the metal-wiring layer up to a scaling factor for the sapphire relative permittivity.

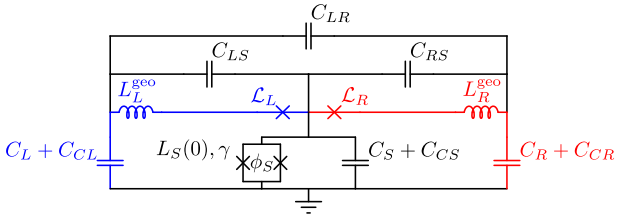


FIG. 11. The lumped-circuit model for parameter extraction. An extended circuit schematic diagram of Fig. 1(b) with the addition of parasitic geometric inductances $L_L^{\rm geo}$ and $L_R^{\rm geo}$. The final fitting circuit-element parameters are listed in Table I.

This restricts the number of free parameters in the circuit model and fixes the resultant circuit values to be consistent with the proportional geometries of the design. Thus, all the capacitances scale with this single factor corresponding to the sapphire substrate relative permittivity used in the Q3D simulation. While the ANSYS HFSS software can take into account an anisotropic relative permittivity reflective of sapphire, ANSYS Q3D only allows a simple isotropic permittivity. For both microstrip or coplanar-waveguide structures, analytic derivations show that an effective isotropic value can be taken as the geometric mean of the parallel and perpendicular components. Although we receive a room-temperature value for the relative permittivity of sapphire from the vendor supplying our source material, there are limited reports of low-temperature values in the current literature. The actual values can also be skewed by differences in the sapphire growth mechanisms and wafer-cutting and surface-polishing techniques, as well as differences in the low-temperature behavior of residual impurities. To fit the experimental data, we are left to determine the capacitive scaling factor and the inductances: the L and R transmon junctions, the SQUID junctions with potential asymmetry, and the geometric inductances. The geometric inductances are kept in a ratio comparable to that of their different lengths.

The lumped-circuit model shown in Fig. 11 is implemented in scQubits using the custom circuit class [54,55]. In order to diagonalize the system in a reasonable time with an exponentially large computational Hilbert space,

TABLE I. The fitted circuit parameter values.

Parameter	Value	Units	Parameter	Value	Units
$\mathcal{L}_L$	7.37	nH	$\mathcal{L}_R$	7.64	nH
$L_L^{\rm geo}$	0.16	nH	$L_R^{\rm geo}$	0.22	nH
C_L	83.48	fF	C_R	65.15	fF
C_{Lj}	1.12	fF	C_{Rj}	1.12	fF
$L_S(0)$	0.23	nH	γ	0.75	
C_S	4.30	fF	C_{CS}	0.14	fF
C_{CL}	18.34	fF	C_{CR}	12.44	fF
C_{LS}	1.68	fF	C_{RS}	1.55	fF
C_{Sj1}	23.04	fF	C_{Sj2}	17.28	fF
C_{LR}	7.34	fF			

hierarchical diagonalization is utilized [76]. This consists of a change of basis and then identification of subsystems within the circuit. The subsystems correspond to the SQUID coupler and the two transmons without the SQUID inductance. This returns the eigenenergies of the circuit evaluated at each value of external flux bias ϕ. The levels can be identified in increasing energy: E_{gg}, E_{eg}, E_{ge}, E_{fg}, E_{ee}, E_{gf}. From this, the qubit frequencies ω_k, anharmonicities α_k, and the ZZ coupling ζ_s can be calculated:

$$\hbar\omega_L = E_{eg} - E_{gg}, \tag{B1}$$

$$\hbar\omega_R = E_{ge} - E_{gg}, \tag{B2}$$

$$\hbar\alpha_L = (E_{fg} - E_{eg}) - (E_{eg} - E_{gg}), \tag{B3}$$

$$\hbar\alpha_R = (E_{gf} - E_{ge}) - (E_{ge} - E_{gg}), \tag{B4}$$

$$\hbar\zeta_s = (E_{ee} - E_{ge}) - (E_{eg} - E_{gg}). \tag{B5}$$

The extracted circuit parameters are determined by fitting the simulated qubit frequencies, anharmonicities, and static ZZ coupling to the experimentally measured data. The resulting fits are shown in Figs. 1(d), 1(e), and 2(b). Our fits correspond to a geometric mean relative permittivity of sapphire of $\epsilon_r = 11.49$. Additionally, the bare resonance of the SQUID coupler mode can be computed (not shown) and tunes from approximately 50 GHz to 20 GHz over the entire range of external flux ϕ. These values are far enough above any of the qubit transition frequencies of interest so that these modes can be completely neglected. This is an additional advantage of our circuit-design strategy.

APPENDIX C: SAMPLE FABRICATION, PACKAGE, AND SHIELDING

The samples have been fabricated on 76-mm sapphire wafers. The main wiring layer is formed by a fluorine-based reactive-ion dry-gas etch of a 100-nm-thick sputtered niobium film. Next, Josephson junctions are fabricated using the well-known "Dolan bridge technique" that relies on a double-angle ($\pm 14.3^\circ$) evaporation of aluminum using a bilayer-resist MMA(550 nm)/PMMA(240 nm) stack with an intermediate step for oxidization (100 mTorr for 300 s) of the first layer of aluminum (35 nm). The second deposition provides the top aluminum layer (75 nm) and forms a tunnel barrier across a narrow overlap region forming the Josephson junctions (the qubit-junction areas are approximately 100 nm $\times$ 120 nm, while the SQUID junctions are approximately 1.8 μm $\times$ 0.1 μm). This has been achieved in a custom deposition chamber with a load lock that allows computer-controlled oxidization for creating reproducible junction current densities. The final step requires connecting the Josephson junctions to the niobium-wiring layer through an aluminum patch layer that is electron-beam evaporated in the same chamber that was used for making junctions and then lifted off. To ensure superconducting contacts between all layers, we have used an ion-mill cleaning step (beam = 300 V, accelerator = 950 V, filament = 3.47 A) for 30 s before depositing aluminum.

Once diced, the chips are mounted into a sample box made of aluminum and bonded directly to nonmagnetic microwave SMA bulkhead feedthroughs for wiring connections, with additional bond wires between the ground plane on the chip and the inside of the box. This aluminum box is then placed inside another copper box of cylindrical shape. A niobium tube is then placed over this cylinder and is thermalized by a copper tube that covers it. A μ-metal tube is then placed over this cylinder, with another copper tube over that for thermalization. Finally, a copper can lined with black carbonized foam encloses all these layers. The coaxial cables are routed out of the ends of the cylinders and into nonmagnetic SMA feedthroughs that pass through the copper can. The entire shielded enclosure is mounted to the bottom of a dilution refrigerator and cooled to 10 mK.

APPENDIX D: EXPERIMENTAL SETUP

A detailed schematic of the instrumentation and cryogenic setup can be seen in Fig. 13. The device under test (DUT) is connected to the experimental setup through two SMA ports: a cavity port that applies both cavity and qubit drive tones, and a flux-bias port that provides both a dc flux bias and an rf flux bias for adding parametric drive tones. All control and readout microwave tones have been independently pulsed on two separate coaxial ports. The cavity and qubit drive signals have been combined at room temperature and sent to the cavity, while the parametric microwave tones have been sent through a bias tee to the pump port (see also Fig. 13). These separate tones have also required a pulse-timing calibration so that the tones could be properly aligned in time. The cavity pulses are constructed from one microwave source gated by a main arbitrary waveform generator (AWG), shown in Fig. 13. The cavity signal is attenuated and filtered as it is sent to the DUT and is then recovered after passing through cryogenic circulators at the millikelvin stage and then a number of amplification stages (an HEMT at 4 K and following amplifiers at room temperature). The original cavity tone, as well as the response from the cavity, are mixed with a detuned local oscillator tone so that the heterodyne signal can be amplified (HVA-200M-40B preamplifier) and digitized (Gage) for analysis. The qubit drive tones are generated by independent synthesized rf sources that are gated in time with the main AWG. The dc flux bias is provided by a low-noise Yokogawa 5760 and passes through multiple filter sections on the way to the DUT, passing through an Anritsu K250 bias tee at the millikelvin stage. The rf-flux-bias tone is generated in a similar fashion as the



FIG. 12. The shielding and packaging, plus the step-by-step shielding assembly.

qubit pulses, although at lower frequency. These parametric pulse tones are attenuated and pass through their own appropriate filter configuration on their way to the DUT, also passing through the Anritsu bias tee. Filters on the rf flux line are configured to create a pass band between approximately 0 and 1 GHz, which allows parametric modulation at the difference frequency between the qubits and the cavity, while fully isolating the cavity and the qubits at their resonance frequencies.

Our sample box is machined from high-purity aluminum in a 3D-cavity-like fashion, where the chip is secured by indium squished into the corners of a mounting ledge in the center of a cylindrical cavity with a lowest fundamental near 36 GHz, shown in Fig. 13. The wire-bond connections to the bias and cavity leads on chip pass through small cylindrical tunnels that lead directly to SMA connectors, without any additional circuit boards. This box is then mounted inside a cylindrical copper box that is then sequentially enclosed by alternating layers of magnetic shielding and copper tubes to increase thermalization, as shown in Fig. 12. The first shield tube is made from pure niobium and the second is cryoperm. The outermost copper shield can is lined with black carbonized foam. All of this has been mounted to the bottom of a dilution refrigerator that has a mixing-chamber shield that is tin plated and acts as an additional layer of thermal and magnetic shielding.

APPENDIX E: QUBIT RECTIFICATION AND CALIBRATION OF THE PARAMETRIC PUMP DRIVE

The data shown in Fig. 2(c) require measuring the qubit spectrum while sweeping the frequency ω_p of the parametric pump. During these measurements with a

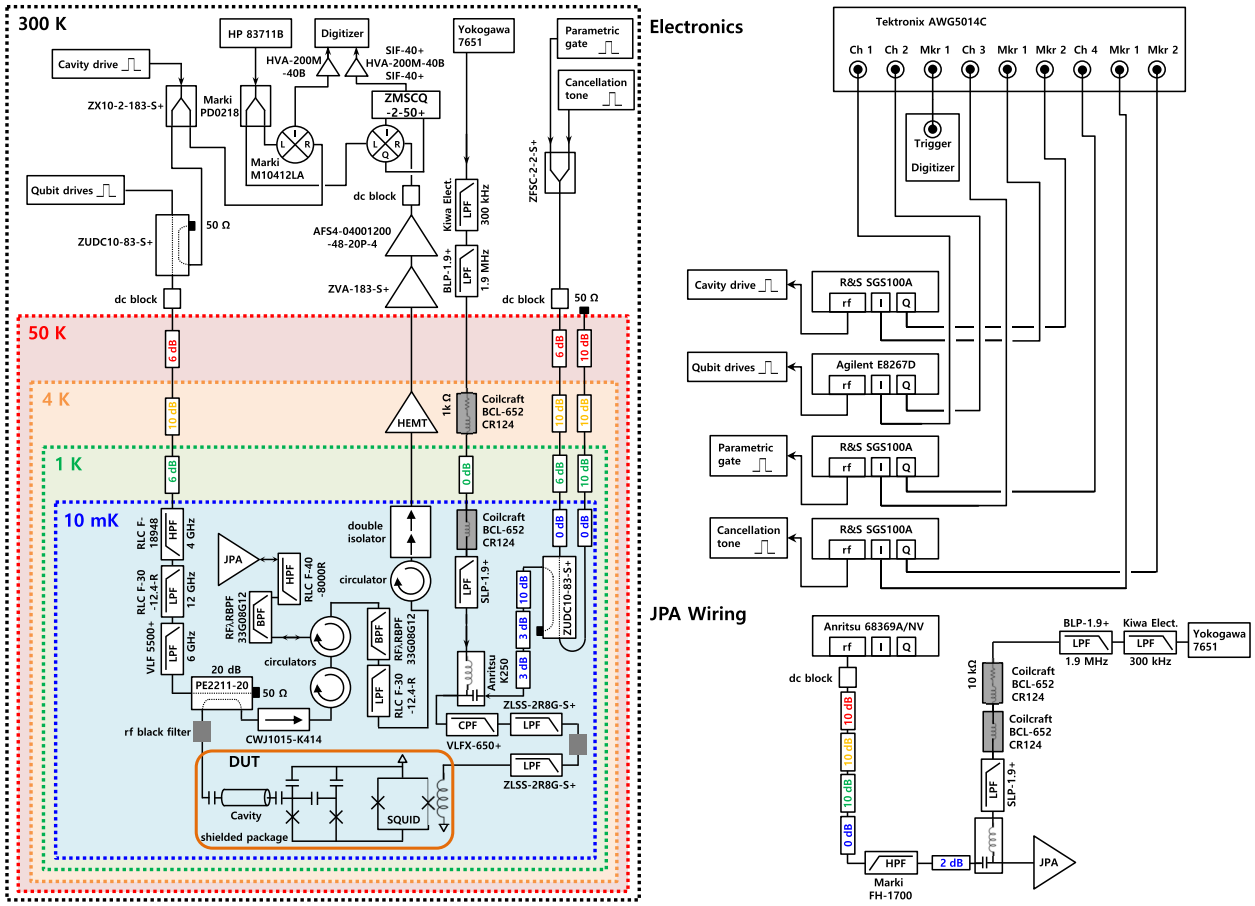


FIG. 13. A schematic of the experimental setup and wiring diagrams.

parametric drive, it is crucial to maintain a constant pump amplitude for all pump frequencies. In this appendix, we discuss a systematic calibration procedure to compensate for the standing waves in the pump line that produce an oscillatory dependence of the pump amplitude with frequency (see Fig. 15) and to extract the flux-modulation amplitude $\delta\phi_p$ of the pump at the device. In principle, the pump calibration can be performed at any static flux bias ϕ_s; however, it is most convenient to perform this calibration at $\phi_s = 0$, where rectification of the qubit frequencies is a maximum and first-order parametric interactions are suppressed.

In the presence of a parametric pump, the resonant frequency of the transmon $\omega_k(\phi_s)$ $(k \in \{L, R\})$ at a given static flux bias $\phi_s = \Phi_s/\Phi_0$ is shifted or rectified to a new average value $\overline{\omega_k(\phi_s)}$ due to any curvature in its flux dependence. If the flux applied to the SQUID loop is modulated such that $\phi(t) = \delta\phi_p \cos(\omega_p t) + \phi_s$, then $\omega_k(\phi(t))$ can be averaged over one period $\tau_p = 2\pi/\omega_p$ through a Taylor expansion, giving

$$\begin{aligned}\overline{\omega_k(\phi_s)} &= \frac{1}{\tau_p}\int_0^{\tau_p} \omega_k(\phi(t))dt \\ &= \frac{1}{\tau_p}\int_0^{\tau_p} [\omega_k(\phi_s) + \omega_k'(\phi_s)\delta\phi_p \cos(w_p t) \\ &\quad + \frac{1}{2}\omega_k''(\phi_s)\delta\phi_p^2 \cos^2(w_p t) + \cdots]dt \\ &= \omega_k(\phi_s) + \frac{1}{4}\omega_k''(\phi_s)\delta\phi_p^2 + O(\delta\phi_p^4)\cdots, \end{aligned} \tag{E1}$$

where $\omega_k''(\phi_s) = d^2\omega_k/d\phi^2$, evaluated at $\phi = \phi_s$. From this expression, it is clear that the rectified shift of the qubit frequency $\delta\omega_k = \overline{\omega_k(\phi_s)} - \omega_k(\phi_s) \approx \omega_k''(\phi_s)\delta\phi_p^2/4$ is dominated by contributions coming from the second derivative of $\omega_k(\phi)$ at ϕ_s or the curvature of the modulation of the qubit with flux. We can use the rectification of (either) qubit frequency $\delta\omega_k$ to quantify $\delta\phi_p$ as a function of the amplitude and frequency settings on the microwave signal generator for the pump,

$$\delta\phi_p \approx 2\sqrt{\frac{\delta\omega_k(\phi_s)}{\omega_k''(\phi_s)}}. \tag{E2}$$

A clear example of the rectification of the R transmon with the signal-generator pump-drive amplitude is shown

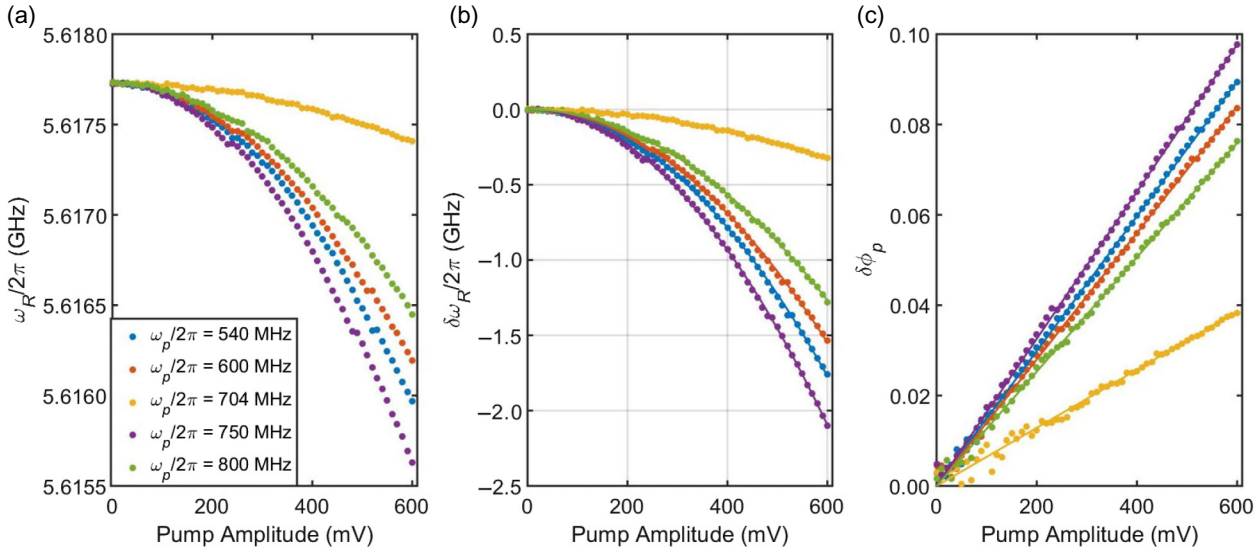


FIG. 14. Rectification of the resonant frequency of the R transmon. (a) Examples showing the rectification of ω_R (the R transmon) as a function of the signal-generator pump amplitude for five pump frequencies ($i = 1, 2, \ldots 5$ for $\omega_p/2\pi = 540$, 600, 704, 750, and 800 MHz). (b) The extracted $\delta\omega_R$ as a function of the signal-generator pump amplitude for the five pump frequencies. The solid lines represent a quadratic fit to the data. (c) The calibrated values for $\delta\phi_p$ as a function of the pump amplitude for each pump frequency. The solid lines are linear fits to the data with slopes S_i and intercept 0.

in Fig. 14 for five color-coded pump frequencies (540, 600, 704, 750, and 800 MHz) with $\phi_s = 0$. In Figs. 14(a) and 14(b), we show that the resonant frequency of the R transmon is quadratically dependent on the amplitude of the parametric pump, with differing strengths depending on the pump frequency. As shown in Fig. 14(c), using $\delta\omega_R$ and $\omega_R''(0)$ from fits to the flux-modulation curve of the qubit spectrum (see Fig. 1), we can use Eq. (E2) to calibrate $\delta\phi_p$ as a function of the pump amplitude for each pump frequency.

To perform the full calibration and compensation, we measure the resonant frequency of the R transmon while sweeping the pump frequency between 200 MHz and 800 MHz at a fixed amplitude setting for the microwave signal generator for the pump. We choose to perform this calibration at the symmetric point $\phi_s = 0$, where $\omega_R'(0) = 0$ (higher-order odd derivatives also vanish) in order to suppress strong parametric couplings between the transmons, which is known (to first order) to be proportional to $\sqrt{\omega_R'(\phi)}$ (see Appendix A above). In this way, any shifts $\delta\omega_R$ of the qubit frequency should be due primarily to rectification and proportional to $\delta\phi_p^2$ (as we have seen in Fig. 14). In order to maintain a constant $\delta\phi_p$ at the SQUID loop of the device for each ω_p, the amplitude setting on the microwave signal generator must vary to compensate for the frequency-dependent attenuation of the entire pump line. We use qubit rectification data across this full frequency range at constant pump-generator amplitude (not shown) to calculate a multiplication factor as a function of the pump frequency,

$$\lambda(\omega_p) = \frac{\lambda_0}{\sqrt{|\delta\omega_R(\omega_p)|}}, \tag{E3}$$

where the normalization constant λ_0 is chosen so that $\lambda = 1$ at a given pump frequency.

In Fig. 15, we show an example of this calibration result when measuring the spectrum of the R transmon in the presence of the parametric pump when $\phi_s = 0.386$ [as shown in Fig. 2(c) and Fig. 16]. Here, we have adjusted the amplitude setting on the microwave signal generator for the pump by multiplying our *nominal* amplitude $V_n = 33$ mV by $\lambda(\omega_p)$. The five pump frequencies ($i = 1, 2, \ldots 5$ for $\omega_p/2\pi = 540$, 600, 704, 750, and 800 MHz) where rectification has been measured directly in Fig. 14 are identified with corresponding colored dots at each $\lambda_i V_n$. The inset shows the ratio λ_i/λ_1 as a function of the inverse normalized slopes S_1/S_i, for the five pump frequencies. The solid line has a slope of 1 and intercept 0. This shows that λ can give the relative size of the inverse slopes $1/S$ for any ω_p. Therefore, after measuring $\lambda(\omega_p)$ and one single rectification curve measured at one pump frequency (see Fig. 14), we can find any $S(\omega_p)$ and calibrate the size of $\delta\phi_p$ at any ω_p.

This calibration and compensation procedure has been important for taking the data presented in this work. As an example, in Fig. 16 we show the R transmon spectrum as a function of the parametric pump frequency,

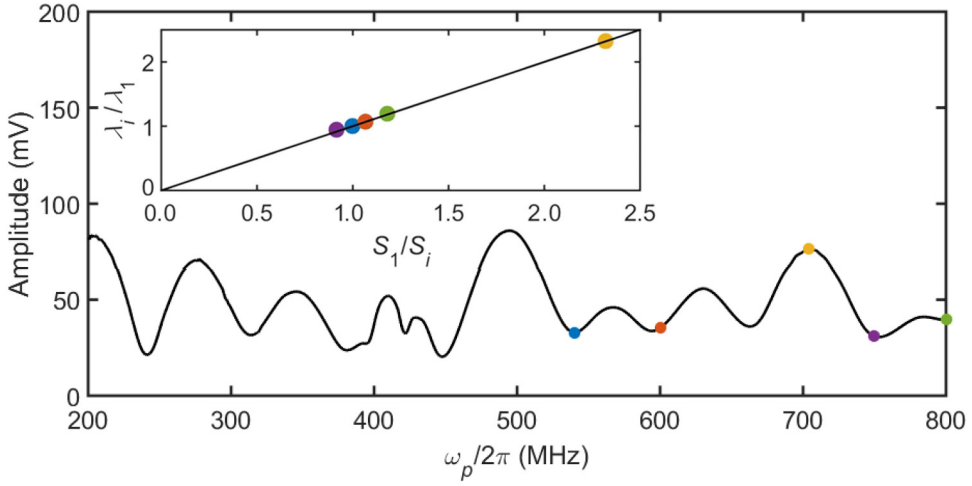


FIG. 15. The signal-generator amplitude as a function of the frequency after pump calibration. The amplitude of the signal generator is found by multiplying the target constant amplitude by the calibration factor $\lambda(\omega_p)$. The five pump frequencies ($i = 1, 2, \ldots 5$ for $\omega_p/2\pi = 540$, 600, 704, 750, and 800 MHz) where rectification has been measured directly in Fig. 14 are identified with corresponding colored dots at each $\lambda_i V_n$. The inset shows the ratio λ_i/λ_1 as a function of the inverse normalized slopes S_1/S_i, for the five pump frequencies. The solid line has a slope of 1 and intercept 0.

with and without the compensation adjustment of the pump signal-generator amplitude (shown in Fig. 15). Without the compensation, as shown in Fig. 16(a), multiple undulations appear as the pump frequency changes, making it difficult to clearly distinguish the avoided crossings due to the parametric mode couplings between the L and

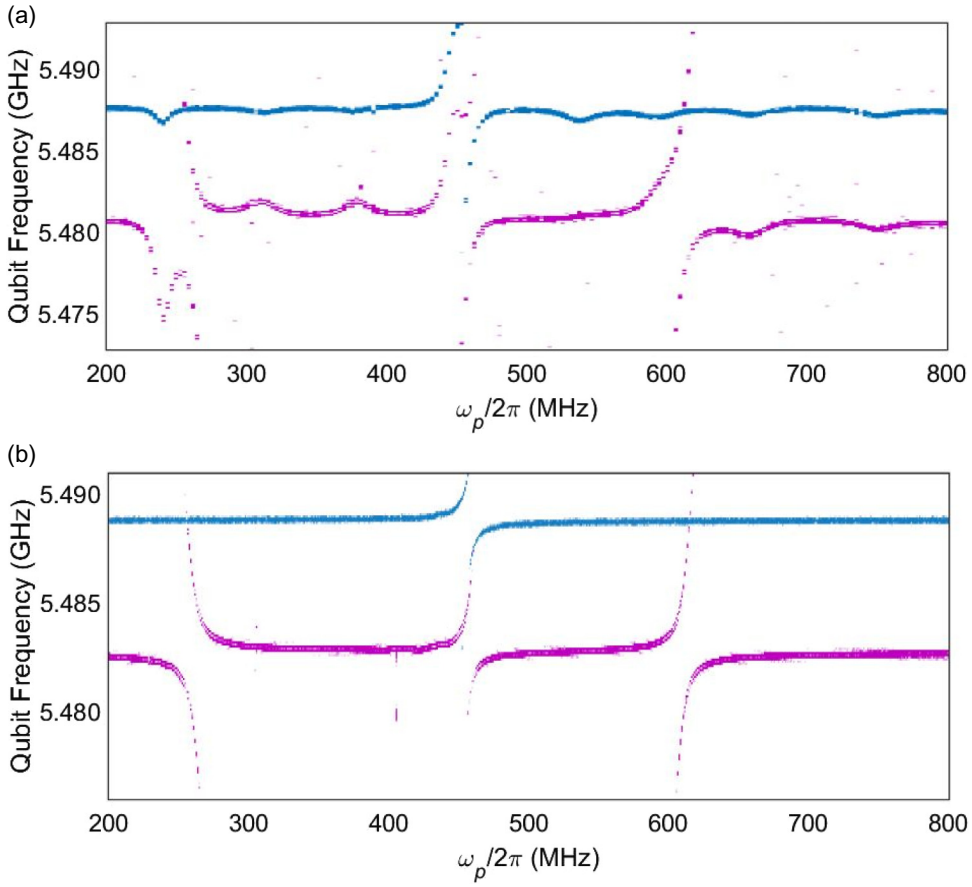


FIG. 16. Before and after pump calibration. (a) Without compensation. (b) With compensation [see Fig. 2(c)].

R transmons. Again, this indicates that the amplitude of the parametric pump is not uniform over the frequency range swept. In Fig. 16(b), we show the result of the same measurement as in Fig. 16(a), but with the compensating numerical factor $\lambda(\omega_p)$ taken into account. Note that the resonant frequency of the R transmon is now mostly constant over this range of pump frequencies, which clearly demonstrates a proper compensation for the unwanted frequency dependence of the standing waves in the pump line. Moreover, the undulations have disappeared, leaving a clear signature of the parametrically generated avoided crossings on top of a uniform background. Clear measurements of the two-qubit parametric interactions with pump detuning both in the frequency and time domain require a constant pump amplitude for all pump frequencies, ensuring the removal of these artifacts.

APPENDIX F: SAMPLE COHERENCE AND PERFORMANCE

1. Sample coherence versus flux and frequency

We have measured the coherence of our samples as a function of the static flux bias ϕ_s. This has included measuring the energy relaxation T_1 and the Ramsey dephasing time T_2^* of each qubit. The data for the main device used to produce the fast parametric gates are shown in Fig. 17. The coherence of the qubit can be plotted as a function of the bias flux, as shown in Figs. 17(a) and 17(b) (L, R), or as a function of the qubit frequency, as shown in Figs. 17(c) and 17(d) (L, R). Note that each qubit has a "dead zone" where the overall coherence is reduced significantly. When plotted against the qubit frequency, these regions look like a single-mode Purcell effect from a mode at 5 GHz and 5.59 GHz for the L and R transmons, respectively. These modes are *not* indicative of TLS defects because they survive multiple temperature cycles to room temperature; however, they do move and sometimes disappear and return. They are *not* consistent with the packaging (shown in Fig. 12) that encloses the chip in a cylindrical cavity with a lowest fundamental mode of approximately 36 GHz. Needless to say, these modes are still under investigation. Note that at $\phi_s = 0$, and the highest qubit frequencies, both qubits show Ramsey coherence times with $T_2^* \approx 2T_1$, where they are first-order insensitive to flux-bias noise. As discussed below, the solid and dashed-dotted lines show predictions accounting for flux-noise dephasing with $\sqrt{A_\Phi} = 4.1\ \mu\Phi_o$ using the average T_1 or the full T_1 data set, respectively.

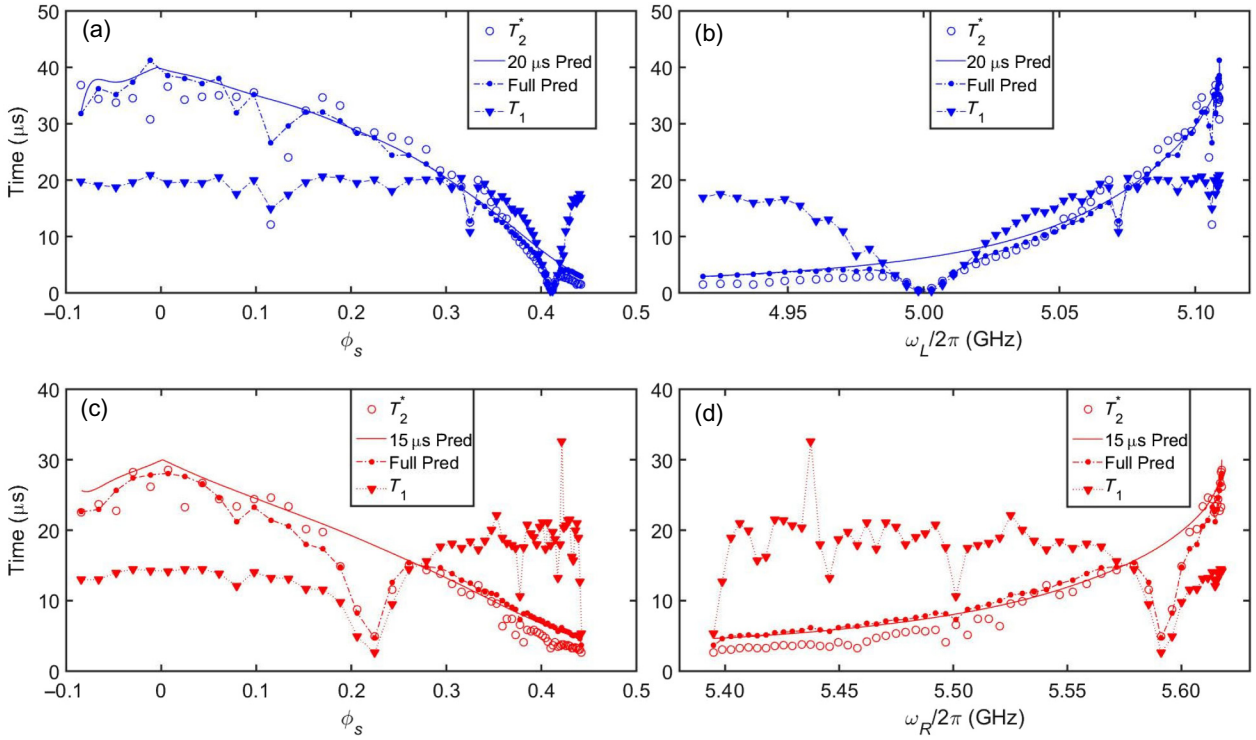


FIG. 17. The device coherence: the coherence for the L [R] qubit as a function of (a) [(b)] the bias flux and (c) [(d)] the qubit frequency. The solid and dashed-dotted lines show predictions for dephasing coming from flux noise with $\sqrt{A_\Phi} = 4.1\ \mu\Phi_o$ using an average T_1 or the full T_1 data set, respectively. Note that each qubit has a "dead zone" where the overall coherence is reduced significantly. When plotted against the qubit frequency, these regions look like a single-mode Purcell effect from a mode at 5 GHz and 5.59 GHz for the L and R qubits, respectively.

2. Coherence, flux sensitivity, and circuit design

Our design strategy balances strong tunable parametric interactions with significant protection from qubit dephasing from bias-flux noise. Much of this discussion follows that found in Noh *et al.* [19, Appendixes]. In our design, we reduce the slopes of the qubit frequencies with flux at the cancellation flux ϕ_c by relying on the weak tunability provided only by the shared SQUID coupler, which then reduces $\partial g_s/\partial\phi$ as well. However, the parametric coupling strength also depends on the size of the parametric pump amplitude $\delta\phi_p$. This means that it is possible to compensate a weak slope (lower flux-noise-induced dephasing rates) with a strong drive to obtain sizable parametric coupling strengths. We only need modulations of order 100 MHz in order to achieve about 100 MHz of coupling. As a benefit, our circuit-design strategy has far weaker sensitivity to flux noise than fully tunable transmons based on two-junction SQUIDs, allowing us to minimize dephasing across most of a flux quantum.

In Hutchings *et al.* [48], the authors describe optimizing a balance between flux tunability, a desirable feature for overcoming issues with critical current deviations from fabrication spread, and maintaining qubit phase coherence. In that work, they have shown that transmons with an asymmetry of 15:1 (the ratio of the left and right SQUID junction critical currents) with a sensitivity to flux ($|d\omega_q/d\Phi|/2\pi$) less than 1 GHz/Φ_o did not show significant additional dephasing when the flux noise was about 1.3 $\mu\Phi_o/\sqrt{\text{Hz}}$. Following their technique, we can calculate the flux sensitivity of our design from knowing the measured qubit frequency as a function of the flux. At the cancellation flux ϕ_c, the two qubits have nearly the same slope of approximately 1.5 GHz/Φ_o and $\lesssim$1.2 GHz/Φ_o where we have operated the two-qubit gates. These values are comparable to the levels found in Hutchings *et al.* [48] and can be reduced further as well, as found in Noh *et al.* [19].

We characterize the flux noise in the same way as in Hutchings *et al.* [48], where $\Gamma_\phi = \Delta\phi|\partial\omega_k/\partial\phi|$ ($k \in \{L, R\}$) represents the dephasing rate from low-frequency flux noise with a $1/f$ power spectrum $S_\Phi(f) = A_\Phi/|f|$ and $\Delta\phi = \sqrt{A_\Phi|\ln(\omega_{IR}\tau)|}/\Phi_o$, where $\omega_{IR}/2\pi = 1$ Hz is the infrared cutoff frequency and $\tau = 10$ μs, comparable to $1/\Gamma_\phi$. The lines in Fig. 17 are given by $1/T_2^* = 1/(2T_1) + \Gamma_\phi + \Gamma_{bg}$, where we use the average measured T_1, $\omega_k(\phi)$ as a function of the flux to calculate $|\partial\omega_k/\partial\phi|$ across the frequency range of the qubit, and $\Delta\phi$ is the best-fit value to the T_2^* data. Here, $\Gamma_{bg} = 1/T_2^* - 1/(2T_1)$ when $|\partial\omega_k/\partial\phi| = 0$ and represents the background dephasing when $\Gamma_\phi = 0$. In our case, $1/\Gamma_{bg} > 500$ μs for both transmons because $T_2^* \approx 2T_1$ at the maximum qubit frequencies. From our fit, we find one value, $\sqrt{A_\Phi} = 4.1$ $\mu\Phi_o$, slightly larger than more typical values of 1–3 $\mu\Phi_o$ found in other experiments [48]. In the future, the maximum T_1 and T_2^* values can be significantly improved with better materials, cleaner fabrication, and improved filtering,

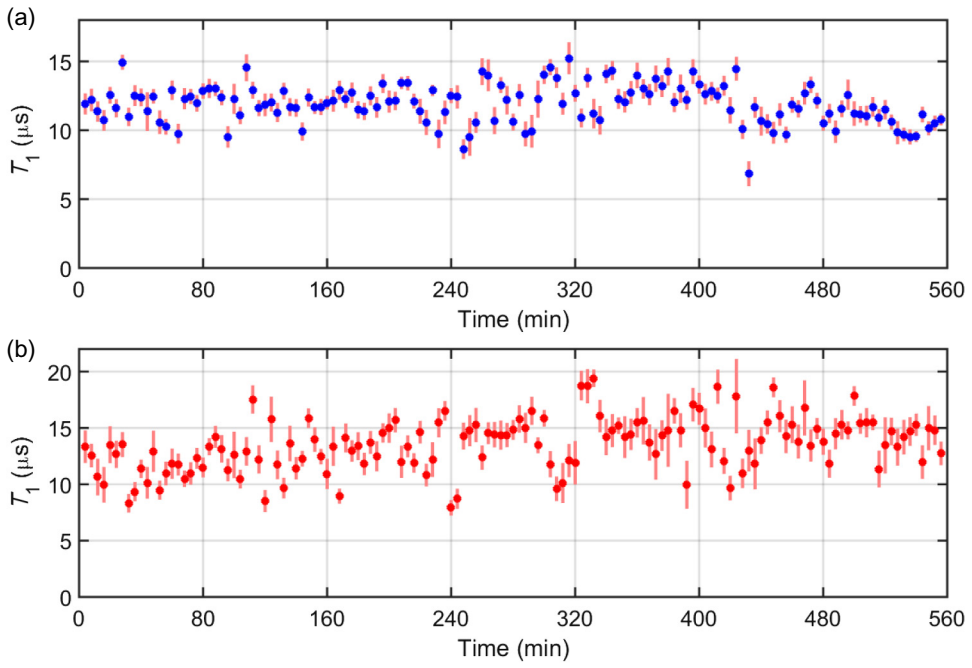


FIG. 18. Measuring relaxation times repeatedly. (a) T_1 as a function of the repetition time for the L qubit. (b) T_1 as a function of the repetition time for the R qubit. These data have been taken over the full timescale for each qubit at separate times, not interleaved. These data have come from a two-qubit device of similar design but fabricated with all-aluminum transmons on a silicon substrate.

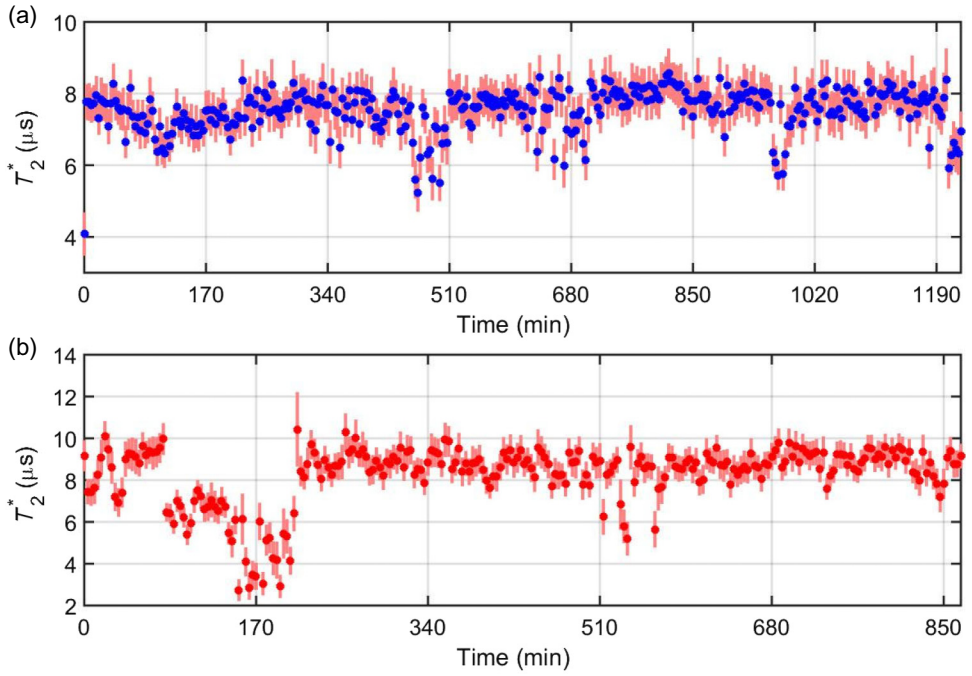


FIG. 19. Measuring Ramsey dephasing times repeatedly. (a) T_2^* as a function of the repetition time for the L qubit. (b) T_2^* as a function of the repetition time for the R qubit. These data have been taken over the full timescale for each qubit at separate times, not interleaved. These data have come from a two-qubit device of similar design but fabricated with all-aluminum transmons on a silicon substrate.

shielding, and packaging to further reduce the sample's exposure to flux noise and infrared radiation.

3. Sample coherence over time

Next, we plot the coherence times for both qubits over time by repeating measurements for long periods as shown in Figs. 18 and 19. Here, we have focused on separately measuring one quantity on one qubit for many minutes, then moving on to another quantity. These results show fluctuations over time with sometimes dramatic changes or jumps over relatively short periods of time. These data have been taken on a similar two-qubit design but made from all-aluminum wiring on high-purity silicon substrates. We have found that the coherence times are comparable to our niobium-on-sapphire samples.

Finally, in Figs. 20 and 21 we plot repeated *interleaved* coherence-time measurements for both qubits over time for the niobium-wiring device used for fast parametric gates. These data have been taken near the operational flux for performing the gates. They do not represent the best coherence times that we have found at this location or per cooldown, and they have not been taken during the time period of our final gate measurements. Here, we have focused on interleaving the measurements of each quantity on each qubit in order to look for correlations in the coherence fluctuations between the qubits. For these data, we have measured T_1 and T_2^* on the L qubit, then the same on the R qubit, and repeated this over time for approximately 28 h. Although we see fluctuations over time, there are no clear correlations between the two qubits.

4. Sample coherence limits on gate fidelity

First, we can check the limits on the single-qubit gate fidelities shown in Fig. 3 using the known individual qubit coherences. We can calculate the gate fidelity limit by comparing the single-qubit gate duration t_g to the energy-relaxation time T_1, or $F = \exp(-t_g/T_1)$. In our case, to avoid populating higher levels of the transmons, we have chosen a single-qubit gate pulse duration of $t_g = 40$ ns (with a 20-ns rise and a 20-ns fall). The T_1 values of the L and R qubits were about 14 μs and 20 μs, respectively, which gives energy-relaxation-limited single-qubit gate fidelities of $F_L \approx \exp(-40/14000) = 99.71\%$ and $F_R \approx \exp(-40/20000) = 99.80\%$, respectively. These values match those measured and shown in Fig. 3. The simultaneous result $F_D = 0.9971 \times 0.9978 = 99.49\%$, also shown in Fig. 3, shows no further reduction in gate fidelity from any extraneous correlations that could come from static coupling or classical or quantum crosstalk, or from additional parametric drives.

In order to calculate the limits on the two-qubit gate fidelity based on the individual qubit coherences in the main text, we have used the T_1 (energy decay) values to obtain $\Gamma_{1,\text{eff}} = (1/2)(1/T_{1,L} + 1/T_{1,R})$ and we

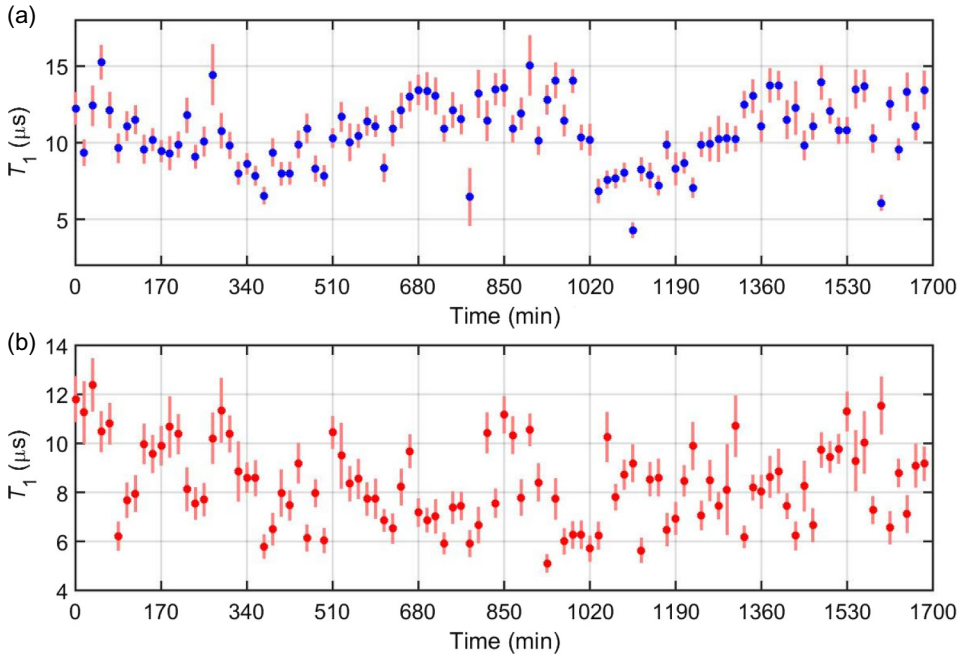


FIG. 20. Measuring relaxation repeatedly interleaved over time. (a) T_1 as a function of the repetition time for the L qubit. (b) T_1 as a function of the repetition time for the R qubit. These data have come from the main two-qubit device fabricated with niobium wiring and aluminum Josephson junctions.

have used the T_2^* (Ramsey decay) values to obtain an estimate of $\Gamma_{2,\mathrm{eff}}^* = (1/2)(1/T_{2,L}^* + 1/T_{2,R}^*)$. We have measured both T_1 and T_2^* before we have performed randomized benchmarking (RB). For the p-SWAP CZ data using plateau pulses, we have measured $T_{1,L} = 13.93$ μs and $T_{1,R} = 19.78$ μs and $T_{2,L}^* = 7.75$ μs and

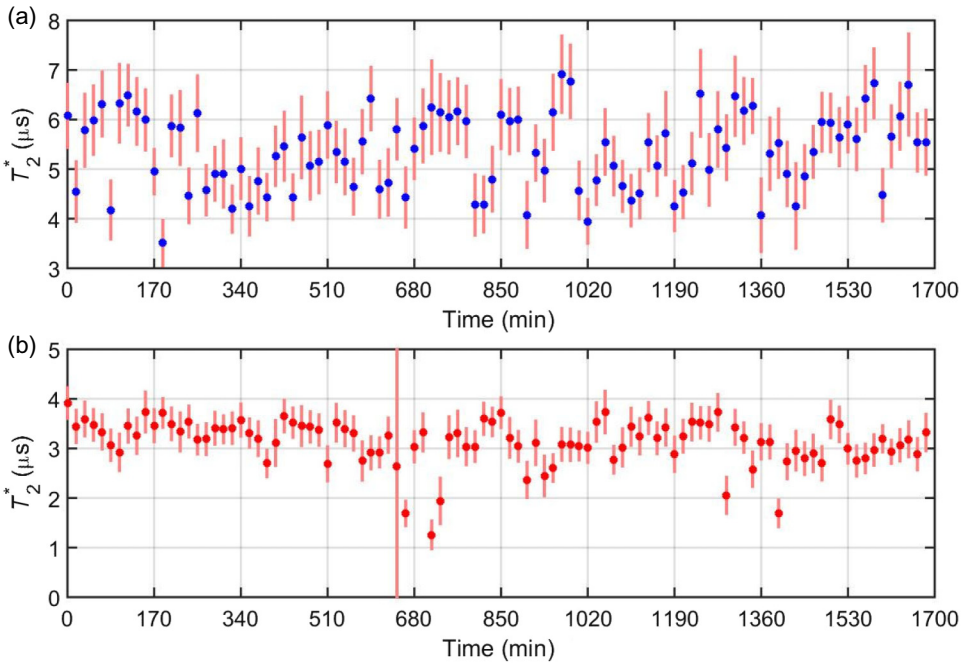


FIG. 21. Measuring Ramsey dephasing repeatedly interleaved over time. (a) T_2^* as a function of the repetition time for the L qubit. (b) T_2^* as a function of the repetition time for the R qubit. These data have come from the main two-qubit device fabricated with niobium wiring and aluminum Josephson junctions.

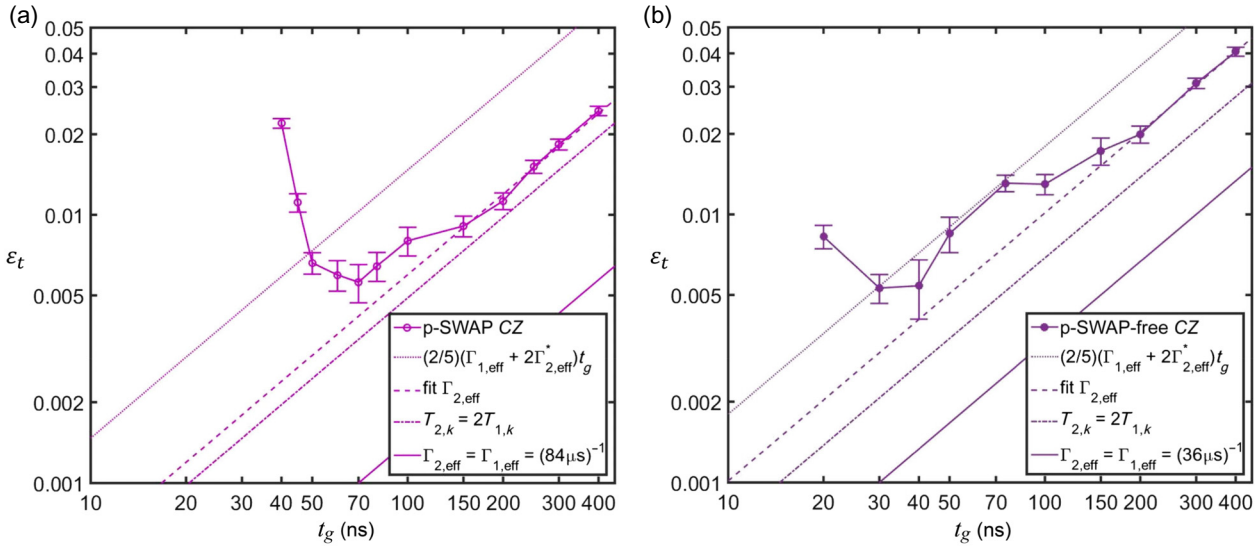


FIG. 22. The coherence limits on the gate fidelity. The total error per gate ε_t of gate duration t_g for (a) the p-SWAP CZ gate and (b) the p-SWAP-free CZ gate.

$T^*_{2,R} = 5.64$ μs. This then gives $1/\Gamma_{1,\mathrm{eff}} = 16.35$ μs and $1/\Gamma^*_{2,\mathrm{eff}} = 6.53$ μs. For the p-SWAP-free CZ data on a different cooldown, the measured values are a bit lower, with $T_{1,L} = 10.82$ μs and $T_{1,R} = 12.56$ μs and $T^*_{2,L} = 6.45$ μs and $T^*_{2,R} = 4.81$ μs. This then gives $1/\Gamma_{1,\mathrm{eff}} = 11.63$ μs and $1/\Gamma^*_{2,\mathrm{eff}} = 5.51$ μs. If we simply plot $\varepsilon^*_{\mathrm{dec}} = (2/5)(\Gamma_{1,\mathrm{eff}} + 2\Gamma^*_{2,\mathrm{eff}})t_g$ derived from the Ramsey data for these values, we find a much larger error than measured, as shown in Fig. 22 by the dotted lines. Using the Ramsey measurements overestimates the error per gate. This seems reasonable because RB tends to have a natural echoing effect that reduces low-frequency noise, which is why we require T_2 values that are related to the phase coherence of the qubits. If we assume that $\Gamma_{1,\mathrm{eff}}$ is an accurate effective mean value (recall above that T_1 values vary over time), then we can estimate the best possible performance when the phase coherence is a maximum or $T_{2,k} = 2T_{1,k}$ for $k \in L, R$. This then leads to $\varepsilon^{\mathrm{min}}_{\mathrm{dec}} = (4/5)\Gamma_{1,\mathrm{eff}}t_g$, which outperforms our measured data, as shown in Fig. 22 by the dashed-dotted lines. In order to find the effective phase-coherence decay rate $\Gamma_{2,\mathrm{eff}}$ that explains the data, we have chosen to use this as a fit parameter for data with $t_g \geq 200$ ns. The result is plotted as dashed lines in Fig. 22 and also in Fig. 5(a). The fit gives $T_{2,\mathrm{eff}} = 1/\Gamma_{2,\mathrm{eff}} = 22.7$ μs or $T_{2,\mathrm{eff}}/T_{1,\mathrm{eff}} \approx 1.4$ for the p-SWAP CZ data and $T_{2,\mathrm{eff}} = 1/\Gamma_{2,\mathrm{eff}} = 12.0$ μs or $T_{2,\mathrm{eff}}/T_{1,\mathrm{eff}} \approx 1.0$ for the p-SWAP-free CZ data. These are very reasonable effective values that are valid over the many hours during which the data are taken. if we continue to expect modest coherence from transmon qubits, where $T_{2,\mathrm{eff}} = T_{1,\mathrm{eff}}$ or $\Gamma_{2,\mathrm{eff}} = \Gamma_{1,\mathrm{eff}}$, then $\varepsilon^{2=1}_{\mathrm{dec}} = (6/5)\Gamma_{1,\mathrm{eff}}t_g$. We can then ask: What coherence values for $T_{1,\mathrm{eff}}$ limit the gate fidelity to 99.9% when $t_g = 70$ ns for the p-SWAP CZ or when $t_g = 30$ ns for the p-SWAP-free CZ gates? It turns out that we only require $T_{1,\mathrm{eff}} = 84$ μs for the p-SWAP CZ and $T_{1,\mathrm{eff}} = 36$ μs for the p-SWAP-free CZ gate. These results are shown as the solid curves in Fig. 22. This is at least a factor-of-2 improvement over the coherence of this sample. Coherence values beyond 100 μs have been achieved in the literature, which should lead to fast parametric gates with gate fidelities beyond 99.9%. This represents our goal for future devices.

There are many results in the community for quantum gates with a variety of different systems and coupling strategies. Unfortunately, providing an exhaustive chart as a fair comparison of all the different coupling methods and gate performance is difficult. In addition, for those couplers that are essentially qubits with their own energy-level structures, some gate errors may depend on the coherence of the coupler, including leakage to its internal states during operation. Any coupling strategy includes other metrics of interest, such as architecture build-out and scaling, apart from just gate fidelity and speed. For example, fluxoniums have the best results in the community, but scaling these systems is still in its infancy. In an attempt to show some quantitative comparison for the results cited throughout the paper, we have included Table II for parametrically, microwave, and adiabatic-flux-activated approaches to tunable coupling. We have taken the best result from each of these works for CZ gates when possible. Based on the T_1 and T_2 (T^*_2) values, we calculate $T_{1,\mathrm{eff}}$ and $T_{2,\mathrm{eff}}$ ($T^*_{2,\mathrm{eff}}$). Note that these values may not be the best metrics during gate operation, but have usually been listed as mean values over time at some idle point. For the table entries, the best performers have had $(T_{2,\mathrm{eff}}/T_{1,\mathrm{eff}}) \gtrsim 1$. Therefore, we have chosen to estimate the most probable modest

TABLE II. A summary of the key performance metrics: "Cubic T" refers to a cubic transmon coupled to a transmon, "X transmon" refers to an extended transmon coupler. Coupling is either "Capacitive" or "Galvanic" (inductive). A "Coupler" that is "Static" is not a separate structure.

Reference	Gate	F (%)	t_g (ns)	Qubits	Coupling	Coupler	$T_{1,\mathrm{eff}}$ (μs)	$T_{2,\mathrm{eff}}$ (μs)	$\epsilon_{\mathrm{dec}}^{2=1}$ (10^{-3})
Parametrically activated									
Reagor *et al.* [77]	*CZ*	95	278	Transmon	Capacitive	Transmon	34.1	6.95*	9.8
Hong *et al.* [30]	*CZ*	98.8	176	Transmon	Capacitive	Transmon	15.9	16.8*	13.3
Ganzhorn *et al.* [29]	*CZ*	99.1	188	Transmon	Capacitive	Transmon	70.0	39.4	3.2
Noguchi *et al.* [31]	*CZ*	97.1	50	Cubic T	Capacitive	Static	3.9	0.95*	15.4
Moskalenko *et al.* [13]	*CZ*	99.23	183.3	Fluxonium	Capacitive	Fluxonium	87.0	99.5	2.5
Zhang *et al.* [16]	$\sqrt{\mathrm{bSWAP}}$	99.91	102	Fluxonium	Galvanic	rf SQUID	300	273	0.4
Sete *et al.* [33]	*CZ*	99.34	64	Transmon	Capacitive	Transmon	22.0	26.1	3.5
Huber *et al.* [36]	*CZ*	98.30	339	Transmon	Capacitive	Transmon	23.4	14.9	17.4
This work	*CZ*	99.67	30	Transmon	Galvanic	dc SQUID	10.8	5.51*	3.3
Microwave activated									
Mitchell *et al.* [65]	*CZ*	99.43	201	Transmon	Capacitive	Static	65.0	81.3	3.7
Ding *et al.* [15]	*CZ*	99.922	50	Fluxonium	Capacitive	Transmon	560	195	0.1
Lin *et al.* [17]	controlled-NOT	99.94	60	Fluxonium	Galvanic	Static	260	171	0.3
Adiabatic flux activated									
Xu *et al.* [12]	*CZ*	99.5	30	Transmon	Capacitive	Transmon	6.54	0.86	5.5
Sung *et al.* [9]	*CZ*	99.76	60	Transmon	Capacitive	Transmon	60.0	27.7	1.2
Stehlik *et al.* [10]	*CZ*	99.85	46	Transmon	Capacitive	Transmon	76.0	105	0.7
Marxer *et al.* [14]	*CZ*	99.81	33	Transmon	Capacitive	X transmon	13.0	10.2	3.1

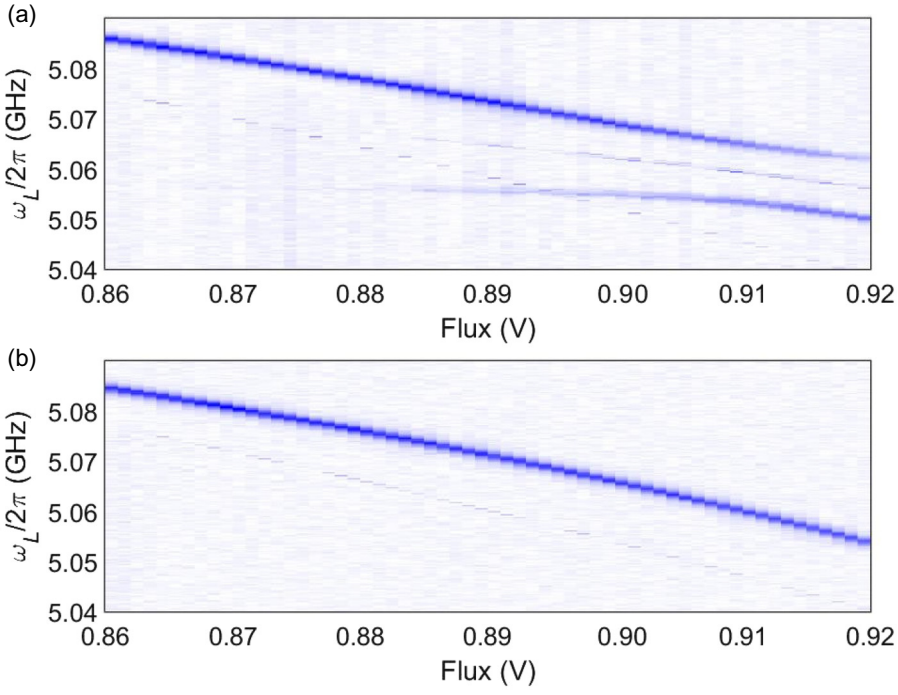


FIG. 23. The *L*-qubit spectroscopy. (a) This spectrum shows a clear approximately 10-MHz mode splitting, signaling the presence of a two-level system (TLS). (b) After this device was cycled in temperatures above 100 K and then recooled, the TLS disappeared, leaving a clear spectrum.

performance, $\varepsilon_{\rm dec}^{2=1} = (6/5)\Gamma_{1,\rm eff} t_g$ when the effective pair phase coherence equals the pair energy decay or $T_{2,\rm eff} = T_{1,\rm eff}$. Achieving $\varepsilon_{\rm dec}^{2=1} < 1 \times 10^{-3}$ is a near-term goal of all the strategies, with some fluxonium approaches already there. Note that qubit coherence, as apposed to coupler coherence (not listed), may *not* be the most important factor determining gate errors for some systems.

5. Sample coherence reductions due to TLS defects

One major difficulty we have had with our qubit samples is TLS defects. Although during a given cooldown the spectra of the qubits may look clean across an entire flux quantum, we have seen TLSs appear from nowhere. Usually, if our laboratory loses power and the devices warm up above 100 K, there is a chance that TLSs will appear in the usable spectra of the qubits. In general, we do not see them as strongly resonant TLSs within the spectra of the two qubits. Rarely, they may make one of the qubits have telegraphlike fluctuations in its frequency. More often, we just see the coherence drop or fluctuate more than usual. It is likely that the stability of our two-qubit gates [shown in Fig. 5(d)] is influenced by these types of fluctuations. More work is needed to carefully characterize and quantify these effects.

The strongest example of a TLS defect in our system is shown in Figs. 23 and 24. In Fig. 23(a), we clearly see a mode splitting in the spectrum of the L transmon, which indicates the presence of a strongly coupled TLS. Here, the coupling strength is about 10 MHz—quite significant. In this case, when taking measurements to tune up the two-qubit parametric gates, we have suddenly seen a drastic drop in the gate fidelity. Upon further inspection, it has been clear that the L qubit has experienced a significant coherence drop. This has been made clearer by measuring the Ramsey oscillations of both qubits, as shown in Fig. 24(a). Note that the Ramsey-oscillation amplitude is very low for the L transmon [and the frequency is too high—it was after this that we actually took the spectrum shown in Fig. 24(a)]. Note that the R qubit is not affected by the TLS. Our circuit design keeps the qubits in relatively close proximity, so it seems that the TLS is mostly likely within the small main Josephson junction of the L qubit.

We have tried many strategies to try to remove TLSs to improve coherence. We have moved the flux bias to extremely large positive and negative values, as well as modulating those fluxes with ramps over time. We have driven both ports with strong rf power in order to "rattle" the circuit. Some of these techniques worked, but not every time. We have also tried to warm up the device above the 1-K critical temperature of the aluminum Josephson junctions without much success. The most reliable technique is to increase the temperature of the device above about 100 K. In Figs. 23(b) and 24(b), we show that a temperature cycling of the device has removed the strongly coupled TLS completely. The fabrication of our main qubit Josephson junctions produces junction sizes that are relatively small, with junction areas near 130×130 nm^2. Although we may be avoiding most of the TLSs within the qubit junctions, it is still no guarantee. We believe that more materials research will be needed to produce nearly defect-free Josephson junctions that will ensure more stable devices and two-qubit gates.

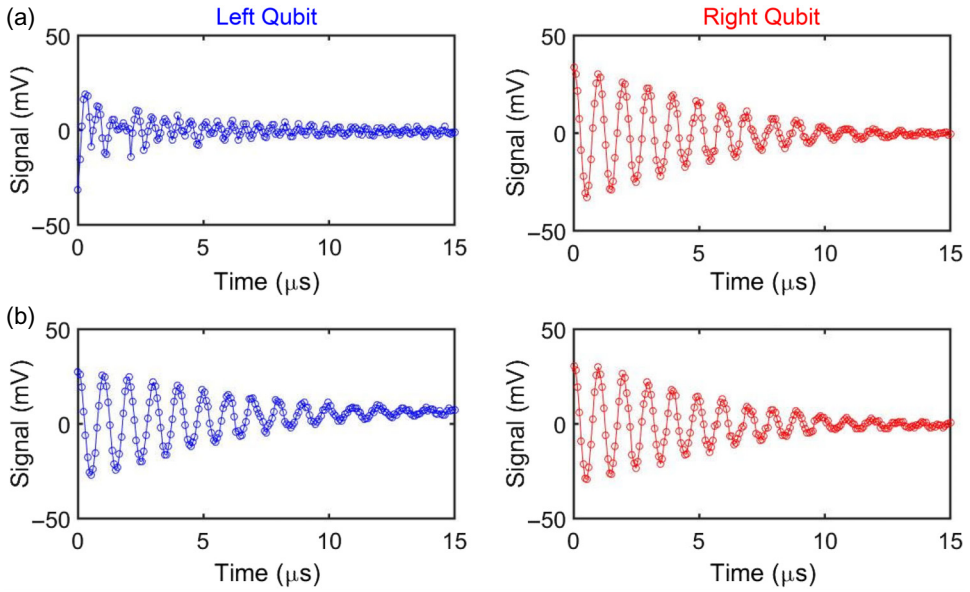


FIG. 24. The L- and R-qubit Ramsey oscillations. (a) In the presence of a strongly coupled TLS, the Ramsey oscillations at the operation flux are severely altered in the L qubit. (b) After this device has been cycled in temperature, the clear Ramsey oscillations at the operational flux are restored.

APPENDIX G: TWO-QUBIT READOUT AND FIDELITY

For our circuit design, we use a joint two-qubit readout with one coplanar-waveguide cavity coupled to both qubits. This simplifies the layout and requires only one cavity response to be monitored with a Josephson parametric amplifier (JPA) to achieve single-shot readout. The JPA that we use has been designed and fabricated by our group using a niobium-trilayer fabrication process [78]. This simplifies the measurement strategy to some degree, but also requires more effort to calibrate the readout properties in order to distinguish all four computational states of the two-qubit system. Because both qubits impart a dispersive shift onto the cavity, there are four cavity frequencies for the four states $|gg\rangle$, $|eg\rangle$, $|ge\rangle$, and $|ee\rangle$. For our system, the dispersive shifts from each qubit (for one excitation) are nearly equal, $\chi_L \approx \chi_R$, so that the states $|eg\rangle$ and $|ge\rangle$ are nearly indistinguishable [19]. We can see this by monitoring the readout cavity as a function of the drive frequency when the qubits are prepared in a mixture of all four two-qubit states. This leads to three possible phase curves with drive frequency as shown in Fig. 25. At the lowest and highest drive frequencies, the phase response is nearly the same for all states, but between 8.715 and 8.716 GHz there are three well-separated phase values that can distinguish three possibilities: $|gg\rangle$, $|ee\rangle$, and the overlapping response of $|eg\rangle$ and $|ge\rangle$. If we carefully measure the readout-cavity response when preparing either $|eg\rangle$ or $|ge\rangle$, we can measure the phase response for both qubits separately (see Fig. 26), revealing that $2\chi_L/2\pi \approx 2.8$ MHz and $2\chi_R/2\pi \approx 2.9$ MHz, with a difference of about 120 kHz.

We note that this type of joint two-qubit readout, where the χ shifts on the cavity are nearly equal for the two qubits, has been used before. The advantage of not identifying which qubit is excited is very useful for both entanglement-generation and parity-readout strategies [79–81]. This measurement strategy has other advantages. This same device could be used as a dual-rail logical qubit. Here, the two qubits can be used as a single logical qubit that spans the subspace of the states $|ge\rangle$ and $|eg\rangle$, similar in concept to recent device proposals [82]. Our measurement technique could provide us with the ability to perform midcircuit readout to determine if the system has left this subspace without significantly decohering the entangled states made from $|eg\rangle$ and $|ge\rangle$. "Erasure qubits" and this dual-rail approach are a promising new qubit architecture [83].

Also in Fig. 25, we show the single-shot histograms in the I-Q plane at several cavity drive frequencies. At the lowest (highest) drive frequency, we can distinguish only one (two) regions in I-Q space. This response splits into three regions at the central drive frequency, allowing us to clearly distinguish three distributions or "balls" for $|gg\rangle$, $|ee\rangle$, and the overlapping response of $|eg\rangle$ and

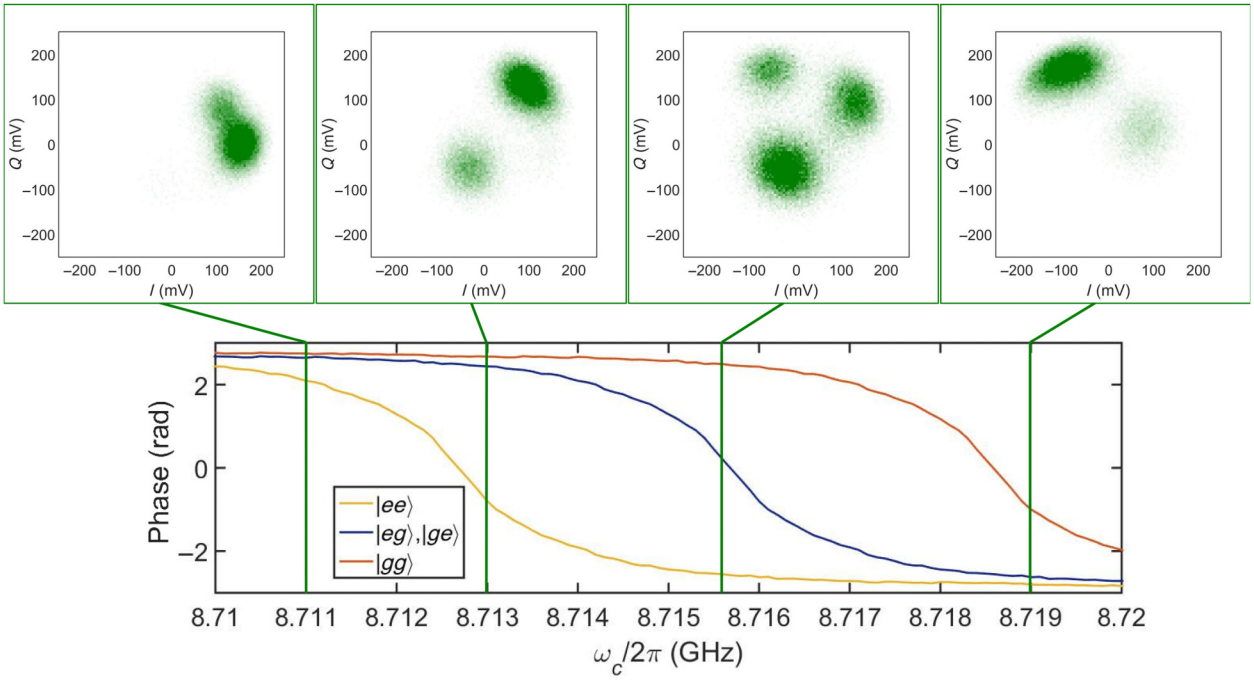


FIG. 25. The readout-cavity measurements as a function of the drive frequency. The main plot shows the average phase of the readout cavity as a function of the drive frequency when prepared in a mixture of two-qubit states. We also show I-Q phase portraits at several cavity frequencies. Note that between 8.715 and 8.716 GHz there are three well-separated regions that can distinguish three possibilities: $|gg\rangle$, $|ee\rangle$, and the overlapping response of $|eg\rangle$ and $|ge\rangle$.

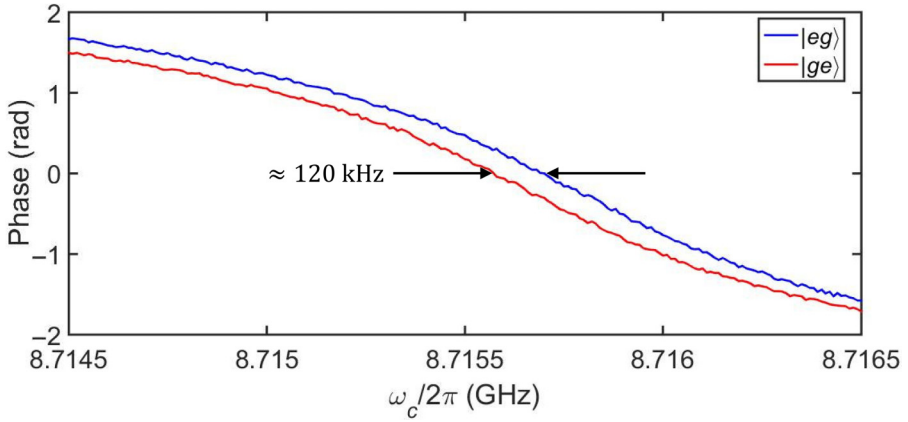


FIG. 26. Fine measurements of the readout cavity for states $|eg\rangle$ and $|ge\rangle$. By measuring the readout-cavity response when preparing either $|eg\rangle$ or $|ge\rangle$, we can see that both qubits produce nearly equal dispersive shifts on the cavity. The total shifts are $2\chi_L/2\pi \approx 2.8$ MHz and $2\chi_R/2\pi \approx 2.9$ MHz, with a difference of 120 kHz.

$|ge\rangle$. To maximize the two-qubit readout fidelity, we must calibrate the readout drive frequency and amplitude in order to maximally separate these balls in I-Q space. This can be quantified by measuring the three angles θ_i and three distances L_i that characterize the triangle formed by connecting the centroids of the three balls.

In Fig. 27, we show the three angles θ_i as a function of the drive frequency. In the I-Q portraits, we can see that above and below the central frequency the scalene triangles have acute angles where two of the balls have poor distinguishability. However, it is clear that near 8.7155 GHz the three balls are maximally separated, forming an equilateral triangle with all $\theta_i \approx 60°$. In Fig. 28, we show the three lengths L_i as a function of the cavity drive amplitude when driven near the optimal cavity drive frequency found in Fig. 27. Note that as the drive amplitude increases all three lengths increase, separating the three balls. We can see that near 1500 mV, the three distances are

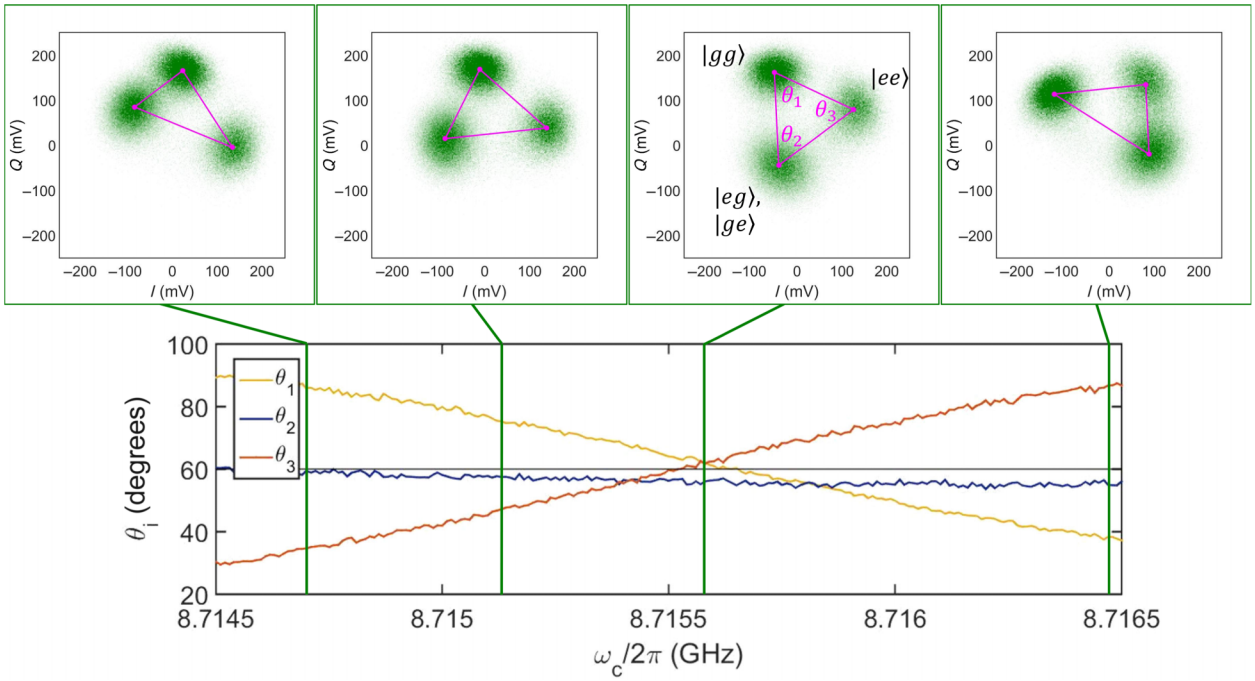


FIG. 27. The readout-cavity measurements as a function of the drive frequency at a drive amplitude of 1150 mV. The main plot shows the three angles θ_i that characterize the triangle formed by connecting the center of the three response regions in the I-Q plane as a function of the drive frequency when prepared in a mixture of two-qubit states. We can see that near 8.7155 GHz, the three regions are maximally separated mostly closely, forming an equilateral triangle.

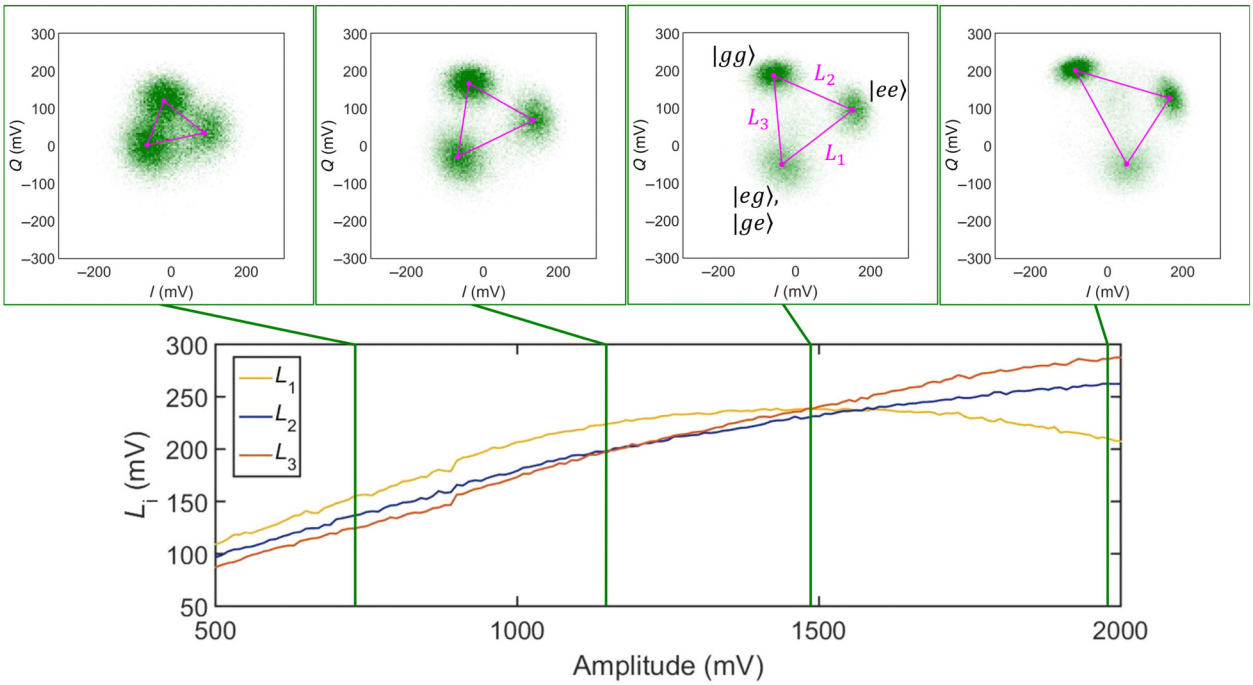


FIG. 28. The readout-cavity measurements as a function of the drive amplitude at drive frequency 8.7154 GHz. The main plot shows the three lengths L_i that characterize the distances connecting the centers of the three response regions in the I-Q plane as a function of the drive amplitude when prepared in four two-qubit states $|gg\rangle$, $|eg\rangle$, $|ge\rangle$, and $|ee\rangle$. We can see that near 1500 mV, the three distances are equally and maximally separated. Note that for large amplitudes, we can see major distortions and point-spreading of the three regions.

equally and maximally separated. In the I-Q portraits, we can see that for the largest amplitudes the triangle begins to distort, leading to scalene triangles. In addition, the density of points between balls increases, signifying that the readout is starting to cause unwanted transitions between the qubit states during the readout.

In order to avoid distortions and unwanted state transitions, we have chosen a moderate drive amplitude of 1150 mV at a cavity drive frequency of 8.71555 GHz, where the separations have been reasonably large and the balls have nearly formed an equilateral triangle. Under these conditions, we can estimate the separation fidelity for the measurement of $|gg\rangle$ by knowing the standard deviation σ and separation S of the three balls [84]. Here, we must sum the errors ($erf\ (S_i/2\sqrt{2}\sigma_i) = 3.0\%, 2.2\%, 2.7\%$) coming from the overlap of $|gg\rangle$ with the Gaussian balls of $|eg\rangle$, $|ge\rangle$, and $|ee\rangle$. This equates to a separation fidelity of 92.1%. Had we been able to operate at 1480 mV without adverse effects from increased qubit relaxation, this would have increased to 98.5%. Measurements of $|gg\rangle$ have also provided an estimate of the sample temperature of approximately 60 mK.

In order to fully distinguish the states $|eg\rangle$ and $|ge\rangle$ for monitoring the parametric vacuum Rabi oscillations like those shown in Fig. 4, we apply a π pulse to the L (R) qubit before driving the cavity readout tone in order to "shuffle" the population of $|eg\rangle$ and $|ge\rangle$ to $|gg\rangle$ and $|ee\rangle$ ($|ee\rangle$ and $|gg\rangle$), the two distinguishable outcomes. Shuffling either excited state to $|gg\rangle$ actually increases its fidelity by avoiding state decay during the readout, which was about $1 - \exp(-0.5/15) \approx 3\%$ for a 500-ns readout pulse with $T_1 = 15$ μs. The shuffle pulse performing the population swap is achieved with high fidelity since the π pulse is calibrated to have the proper rotation and have minimal leakage, as required for single-qubit RB.

For any set of logical operations, we can obtain the four probabilities P_{gg}, P_{eg}, P_{ge}, and P_{ee} by performing $2N$ iterations with readout. One set of N samples is taken without the shuffle pulse and then another set of N is taken with the shuffle pulse. Without the shuffle pulse (the native readout), we can calculate the probability of $P_{gg} = N_{gg}/N$ and $P_{ee} = N_{ee}/N$. With the (L qubit) shuffle pulse for the next N samples, we can assign $P_{eg} = N_{gg}/N$ and $P_{ge} = N_{ee}/N$. It is easy to verify that the statistics from the two sets of N measurements are in agreement. Specifically, $P_{(eg|ge)} = N_{(eg|ge)}/N$ from the first set of N samples without a shuffle pulse is equal to $P_{eg} + P_{ge}$ from the second set of N samples with a shuffle pulse. Likewise, $P_{(gg|ee)} = N_{(ge|eg)}/N = P_{gg} + P_{ee}$, with and without a shuffle pulse, respectively. Thus, we have all four probabilities, but with *half* the efficiency we would have had if we could have distinguished all four states with a single measurement.

The shuffle pulse can also distinguish the state $|ee\rangle$ from $|gf\rangle$ or $|fg\rangle$, all of which have dispersive shifts from $|gg\rangle$ of about -2χ and are nearly on top of each other in the I-Q plane. However, in this case, because the single qubit π pulses are tuned to the transition $|g\rangle \rightarrow |e\rangle$, they will not shuffle the $|f\rangle$ state. Thus, the shuffle pulse on the L qubit will swap $|ee\rangle \rightarrow |ge\rangle$, which is distinguishable from both $|gf\rangle \rightarrow |ef\rangle$ or $|fg\rangle \rightarrow |fg\rangle$. Here, the cavity frequency is shifted by approximately -3χ for $|ef\rangle$ and approximately -2χ for $|fg\rangle$. In the case shown in Fig. 4(b), we have actually performed another version of the shuffle by using a two-photon π pulse on the L qubit (driving hard on the transition $|g\rangle \rightarrow |f\rangle$ at $\omega_{gf}/2$) so that $|fg\rangle \rightarrow |gg\rangle$ and $|ee\rangle \rightarrow |ee\rangle$ (again, because the shuffle is off resonant). Now, the distinguishable states are $|gg\rangle$ and $|ee\rangle$ with a shift of approximately -2χ, on the cavity.

In terms of performing RB, all of our protocols return the system to the state $|gg\rangle$. Thus, we simply need to measure the probability of leaving $|gg\rangle$ in order to perform RB. Our native joint two-qubit readout achieves this without shuffle pulses. It is interesting to point out that early simultaneous RB experiments have been performed in this way, with joint two-qubit readout with a single cavity and two qubits [60,85].

APPENDIX H: TRANSMON EXCITATIONS FROM SUBHARMONIC PUMP DRIVES

For both kinds of CZ gates described in this work, the operation-induced error rises steeply as the parametric gate pulses get shorter than their optimal value, as seen in Fig. 5(a). This divergent error cannot be simply explained by the linearly expanding bandwidth in frequency space of shorter microwave pump pulses. To investigate these errors, we have performed a diagnostic experiment where we have measured the induced vacuum Rabi oscillations over a large bandwidth of microwave frequency for pulsed parametric drive amplitudes close to that used in CZ gates. Upon fine inspection of these data [shown in Fig. 29(a)], it is possible to see "streaks" at certain pump frequencies. If we retake these data, but start in $|gg\rangle$, then the parametric drives should, under ideal circumstances, *not* change the state whatsoever. However, as we can see in Fig. 29(b), we observe Rabi-like oscillations at multiple parametric drive frequencies indicative of two-qubit state transitions. We further interrogate these regions by enhancing the resolution in their locations and by using shuffle pulses (see Appendix G) in order to determine which qubit is being excited. In Fig. 29(c), we can see that one set of oscillations shows that the R transmon is excited, whereas the other oscillations are coming from the L transmon. By noting the pump frequencies at which this occurs, we find that each qubit is being uniquely driven at an integer fraction of its individual excitation frequency. In the case shown here, the R transmon absorbed 11 photons and the L transmon absorbed 10 photons. Once we have understood this problem, we have verified these excitations at many more pump locations, as highlighted in Fig. 30. Remarkably, we have been able to find evidence of multiphoton absorption with up to 16 photons. This is conclusive evidence that multiphoton transitions are possible with strong parametric drives and must be considered when performing strong parametric interactions.

The multiphoton excitation process of transmons described above has a highly nonlinear dependence on the drive amplitude. It remains extremely weak at lower to moderate powers, but then begins to grow rapidly at much larger drive powers and has been proposed as a way to maintain control of extremely Purcell-filtered transmons [86,87]. Thus, for slower gate speeds at lower power, we do not incur any state errors from this process. However, as we reduce the gate duration, the drive power must be increased to larger and larger amplitudes, eventually reducing the gate fidelity through these unwanted additional single-qubit bit-flip errors. Because this error mechanism is coherent, we could in principle apply additional single-qubit drives to compensate for the multiphoton operations [9,15]. It should be emphasized that this issue occurs at well-defined locations in frequency space and could be easier to avoid by further separating the two qubit frequencies. In addition, it is known that our bias coil and coupling design to the SQUID is not optimal for purely driving flux into the SQUID coil. Instead, capacitive coupling from the floating end of the coil to the transmon capacitors can significantly drive there transitions, as verified when we remove any bias line filter. In the future, we plan to significantly decrease this capacitive coupling by galvanically connecting the bias coil and SQUID with a better geometrical design to avoid multiphoton absorption and improve power transfer to flux in the SQUID [40].

APPENDIX I: AUTOMATED GATE OPTIMIZATION WITH CMA ES: COVARIANCE MATRIX ADAPTATION EVOLUTION STRATEGY

The covariance matrix adaptation evolution strategy (CMA ES) is a robust and proficient algorithm tailored for tackling intricate optimization challenges, especially those within the realm of continuous and nonlinear optimization. It works especially well when addressing optimization issues characterized by noisy discontinuous objective functions or those teeming with numerous local optima [88,89]. In the field of superconducting qubits, the CMA ES has been used to optimize pulse shapes in a closed-loop fashion to increase the fidelity of single-qubit gates. This process involves treating each output-point value of the AWG as a variable [67].

In this work, we have chosen to expand the application of the CMA-ES strategy to fine tune the parametric

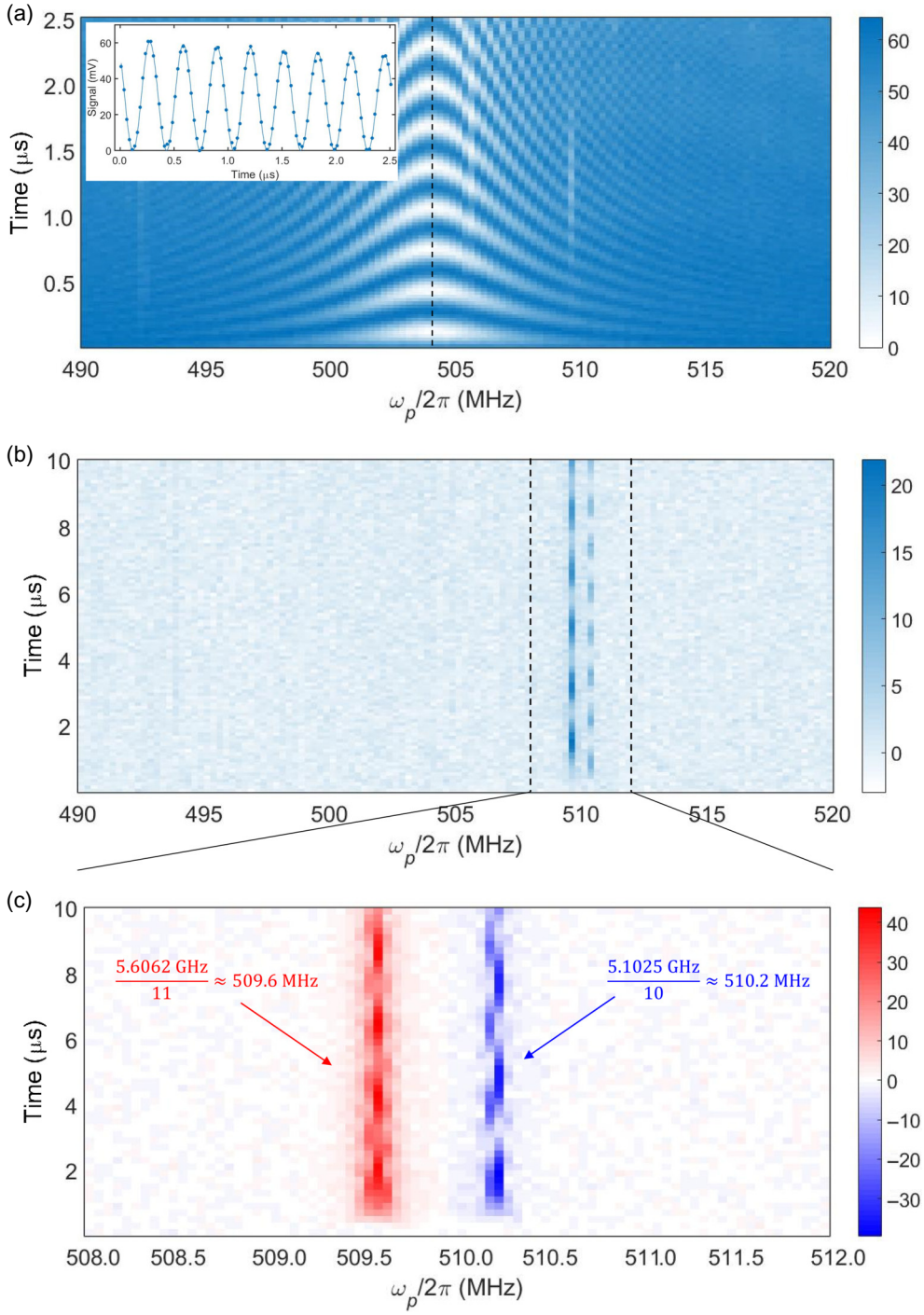


FIG. 29. Detecting multiphoton transitions in the transmons from the parametric pump drive. (a) Parametric oscillations between $|eg\rangle$ and $|ge\rangle$. The inset shows a line cut along the center of the oscillations. (b) The same measurement, but starting from the state $|gg\rangle$. We clearly can see two regions that show excitations in the measurements. (c) An enlargement of (b) showing that each qubit (L in blue and R in red) is being Rabi driven by the pump drive at the displayed subharmonic frequencies.

two-qubit gates in a timely manner. For this task, the physical parameters of the parametric gates are considered as adjustable variables, including the pump amplitude, the frequency, and the width of the pulses, plus the first- and second-order derivative removal by adiabatic gate (DRAG) pulse-shape control parameters. Although a full detailed description and mathematical discussion of the CMA ES is beyond the scope of this work, here we

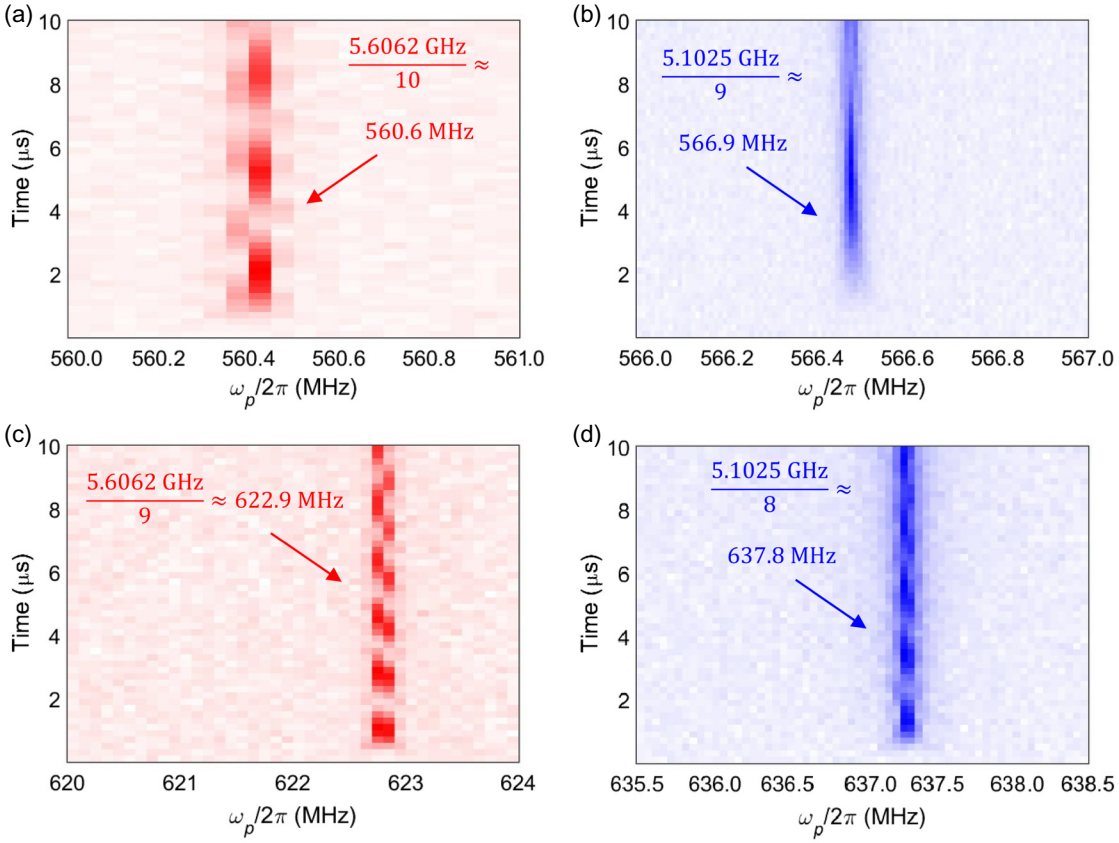


FIG. 30. More locations of multiphoton transitions. (a) A 10-photon subharmonic drive on the *R* qubit. (b) A 9-photon subharmonic drive on the *R* qubit. (c) A 9-photon subharmonic drive on the *L* qubit. (d) An 8-photon subharmonic drive on the *L* qubit.

describe the salient concepts, the general protocol, and our experimental realization.

1. Manual gate optimization and parameter space

Before running the CMA-ES protocol, we first optimize the parametric gate parameters "by hand" with simple parameter sweeps with basic metrics for performance. The initial tune-up of parametric gates involves either a SWAP oscillation (p-SWAP) as shown in Fig. 4(b) or a detuned parametric pulse to enhance *ZZ* coupling (p-SWAP-free) as shown in Fig. 2(e), both for the proper gate duration. This is first done roughly and then tested with the cross-Ramsey experiments, as shown in Fig. 4(c). Here, the cross-Ramsey oscillation changes phase by π if the phase of the *CZ* gate is π. In addition, we also quantify any leakage of the gate operation and adjust parameters to ensure that we do not leave the proper manifold of two-qubit states. For all gates, we calibrate the parametric *CZ* frequency, the *CZ* amplitude, and the *CZ* pulse width as well as the virtual *Z* gates on both qubits. To further increase accuracy, these parameters are swept while evaluating the gate decay number as a simple proxy for the gate fidelity. This value is calculated as the exponential decay rate (in units of the gate number) of the probability of finding the two qubits in state $|gg\rangle$ after a full sequence of gates. This is a measure of the success probability, or what we have labeled as the "Sequence Fidelity." These measurements require just enough points to characterize the probability decay, so that we can relatively quickly sweep the parameters of the parametric gate and find optimal values. As discussed above, and in Appendix E, we calibrate the virtual *Z* rotations as we did the other pump parameters. Examples for these initial calibrations are shown in Figs. 31 and 32.

2. The optimization protocol

In our work, we have repeated a simultaneous sequence of random single-qubit Clifford (*C*1) gates and interleaved *CZ* entangling gates (see below). For all sequences, we have added a final set of recovery gates to return to $|gg\rangle$ so that we could easily measure the probability of success or what we label "Sequence Fidelity." We have chosen two types of reference sequences, the results of which are shown in Figs. 5(b)–5(d, inset). The first reference, labeled "Reference" in the legend with fidelity "F_{Ref}", consists of

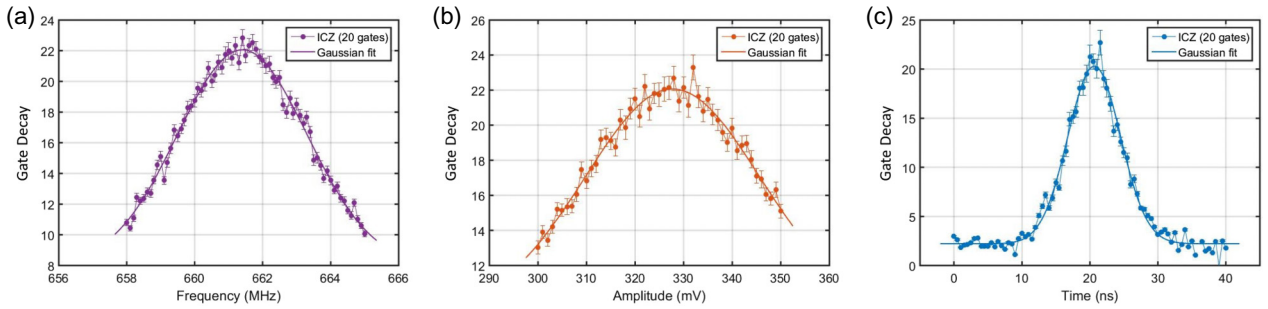


FIG. 31. Optimizing the parametric *CZ*-gate pump parameters by hand. (a) The gate decay number while adjusting the pump frequency. (b) The gate decay number while adjusting the pump amplitude. (c) The gate decay number while adjusting the pump-pulse width.

random single-qubit Clifford ($C1$) gates applied simultaneously to each qubit. The second reference, labeled "Idle" in the legend with fidelity "F_I^{Ref}", consists of random single-qubit Clifford ($C1$) gates applied simultaneously to each qubit interleaved with an "idle" or wait period with the same duration as the *CZ* gate. With these two reference curves, we have calculated F_{cZ}^{Ref} and F_{cZ}^{I}, respectively, according to Eq. (I1) after fitting the data to Eq. (I3), as described below. This has been done in order to experimentally assess the error rate due to the finite coherence times of the qubits as discussed in the main text (see also the theory discussed in Appendix F).

In general, the two-qubit gate fidelity can be deduced by the interleaved RB experiments following this equation:

$$F = 1 - \left(1 - \frac{P_{\text{int}}}{P_{\text{ref}}}\right)\left(\frac{2^n - 1}{2^n}\right), \tag{I1}$$

where $n = 2$ is the number of qubits, and the interleaved decay rate P_{int} and the reference decay rate P_{ref} per gate are determined experimentally. The renormalized data (where the decay starts from 1 and saturates at 0 with no offset) should follow

$$S = P^N, \tag{I2}$$

where S is the readout signal, N is the number of interleaved gates applied, and $0 < P < 1$ is a fitting parameter that represents the average probability of success. Example curves corresponding to different decay rates and gate fidelities are plotted in Fig. 33(a).

Assuming that the raw signal drift is negligible, it is possible to compare two sets of gate parameters and determine which one has higher fidelity, without actually having to measure the entire curve or calculate the actual fidelity of the gate. In Fig. 33(b), we show possible raw-data values, the final voltage readout of one quadrature of the cavity drive in millivolts. If we only measure this raw readout signal after M interleaved gates (the vertical dashed line in Fig. 33 with $M = 20$), then the value at that location will rise and fall with the fidelity. Thus, the gate-optimization problem is identical to a maximization problem of a multiparameter function with a large parameter space. This is

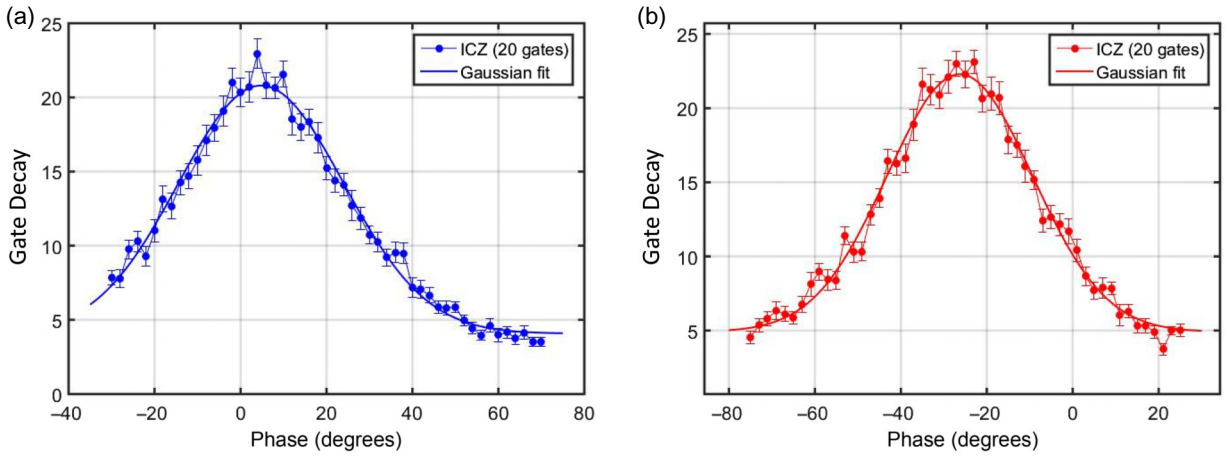


FIG. 32. Optimizing the virtual Z rotations on each qubit by hand during a parametric *CZ* gate. (a) The gate decay number while adjusting the virtual Z gate of the L qubit. (b) The gate decay number while adjusting the virtual Z gate of the R qubit.

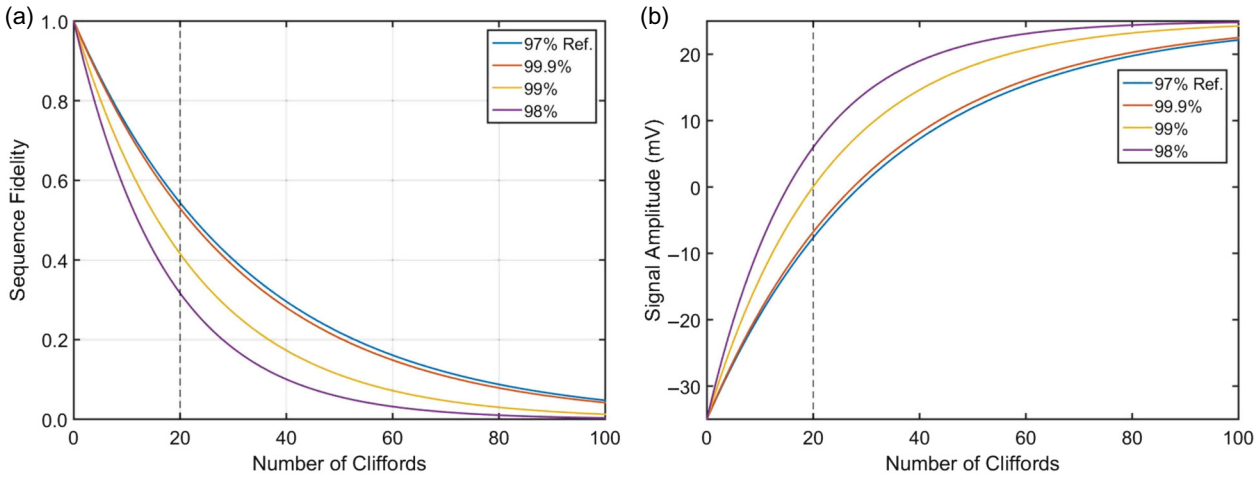


FIG. 33. Visualizing the sequence fidelity and the raw data. (a) Calculated sequence-fidelity curves F of 99.9%, 99.0%, and 98.0% based on a reference decay curve with $P_{\text{ref}} = 0.97$. (b) Possible raw-data curves corresponding to those shown in (a).

where the CMA-ES protocol is ideal—in order to investigate a large parameter space efficiently to find an optimized set of parameters.

During the experimental calibration, the raw data come in without renormalization; thus the fitting function in Eq. (I2) would become

$$S = AP^N + C, \tag{I3}$$

where A is the amplitude of decay and C is the offset constant. Depending on the experimental details, A could be positive or negative, meaning that the data could look like an exponential decay downward or an exponential growth upward. For the possible raw signal shown in Fig. 33(b), along the dashed line where $M = 20$, the signal amplitude decreases when the fidelity actually increases. Here, the optimization target of the CMA ES will become a minimization problem of the multiparameter function for this raw data.

In order to make the CMA-ES method sensitive, one should choose a M value that will reflect greater change in signal when fidelity improves. However, the distance between different fidelity curves is highly sensitive to the reference value. Clearly, a reference curve closer to 0.99 will allow more room to approach 0.99. In Fig. 33, we have chosen a reference decay rate of 0.97. Note that in order to achieve 0.999, one must get very close to the 0.97 curve and the difference between the reference and the other curves seems to be maximum when about 67% of the curve has decayed, or at the exponential gate decay number discussed above in Appendix I 1. Therefore, knowing the reference curve, it is possible to choose M to be close to the gate decay number. However, more gates also takes more time. Therefore, in practice we have chosen $M = 10$–20 in our calibrations, which saves time but also maintains a decent signal size to enable sufficient optimization.

3. Motivation and experimental realization of CMA ES

For simplicity, we assume that the two-qubit gate has only two parameters to optimize, namely, the gate frequency f and its amplitude a. Ideally, f and a are supposed to be orthogonal parameters, which means they can be calibrated independent of each other, and changing one parameter does not affect the optimization of the other. If this were true, the rough optimization of each individual parameter shown above in Appendix I 1 would have been sufficient. However, in real experiments, where the loss of the microwave signal itself is a function of f and the sample response is nonlinear to the applied gate amplitude a, the two parameters are no longer orthogonal. Instead of a one-dimensional scan for f and a, respectively, one needs multiple scans to find the optimized values and could end up finding only a local minimum instead of a global optimization point. A full 2D scan could solve the problem, but at the cost of a geometrically increasing search time. Considering more parameters, the optimization process would become an extremely long process, with no guarantee of locating the best existing parameter set. The application of the CMA ES could reduce the search time by more than a factor of 10, while maintaining if not exceeding the best optimization results.

Essentially, the CMA-ES optimization that we apply is a closed-loop iteration process in which each iteration will attempt to optimize the parameters based on the results of the previous iteration. We start with a set of manually input "original seed" parameters and several iteration-control parameters. They include the population of candidate (number of seeds) m, the selection pressure (survival rate) $0 < s < 1$, the scattering prefactor p, and the initial starting step size ϵ for the parameters, respectively. Using the (greatly simplified) CMA-ES protocol, we do the following:

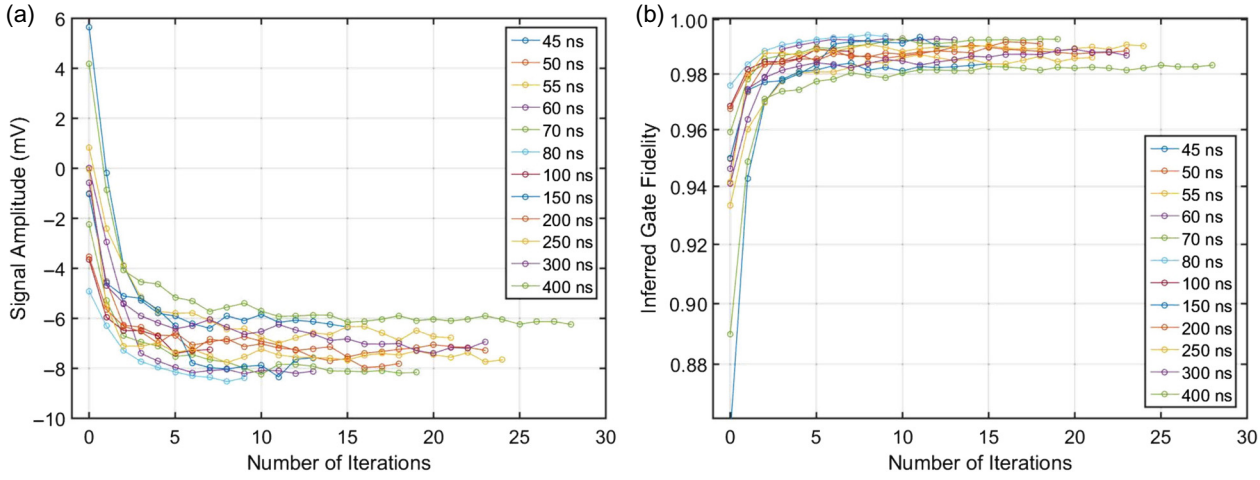


FIG. 34. The CMA-ES protocol to optimize the gate fidelity. (a) Raw data of the CMA-ES calibration up to 28 periods. (b) The CMA-ES calibration up to 28 periods converted to gate fidelities.

(1) We start with manually calibrated parameters, $f_0 = 661.5$ MHz and $a_0 = 328.5$ mV, and choose the initial step sizes of $\epsilon_f^0 = 5.0$ MHz and $\epsilon_a^0 = 5.0$ mV, based on the results from Appendix I 1.

(2) We then generate the first generation of seeds: ${f_1}^0, {f_2}^0, \ldots, {f_m}^0$, and $a_1^0, a_2^0, \ldots, a_m^0$, where $m = 50$ is the population mentioned above. The seeds are randomly distributed in the space of $f_0 \pm \epsilon_f^0$ and $a_0 \pm \epsilon_a^0$, respectively. We also decide to use a survival rate of $s = 20\%$.

(3) We pick one set of parameters $P_1^0 : \{{f_1}^0, a_1^0\}$ and build a two-qubit gate. We then generate an interleaved gate sequence with, e.g., $N + 1$ simultaneous randomized single-qubit Clifford ($C1$) gates with $N = 10$ two-qubit CZ gates interleaved. The last gates return the system to $|gg\rangle$. At the end of the sequence, we apply the readout pulse and record a signal value. We repeat this experiment for, e.g., $M = 50$ times, and get a mean value for Q_1^0.

(4) For each set of data P_i^0, we repeat (3) and get a mean value Q_i^0. At the end of the initial iteration, we get a series of Q^0 containing m numbers.

(5) Assuming that the optimal condition is to find the minimum. With a survival rate of s, only $m \times s = 10$ seeds should survive. We thus pick among Q^0 the ten smallest values (assuming that they are sorted) $Q^0_{1,2,\ldots,10}$, and trace down their corresponding parameters $P^0_{1,2,\ldots,10}$. Note that these numberings are different than the order in which experiments are carried out in (3).

(6) Based on ${f}^0_{1,2,\ldots,10}$ and $a^0_{1,2,\ldots,10}$, we calculate their mean values ${f_m}^0$ and a_m^0, and their standard deviations S_f^0 and S_a^0, respectively. Here, we adopt ${f_m}^0$ and a_m^0 to become the new parameters f_1 and a_1, and use the scattering prefactor p to calculate the new step sizes: $\epsilon_f^1 = p \times S_f^0$ and $\epsilon_a^1 = p \times S_a^0$.

(7) Now we can go back to step (2) and start the new iteration, where we generate the new sets of seeds in the new parameter space $f_1 \pm \epsilon_f^1$ and $a_1 \pm \epsilon_a^1$.

The change of raw signal versus the number of iterations is plotted in Fig. 34(a) for one gate calibration, where the target was to find the minimum, and $M = 10$ was chosen. The improvement of fidelity is clearly seen (although quantitatively unclear) in the first 15 iterations, after which the improvement saturates into the noise background. This was done for various gate durations, as indicated in the figure.

It is possible to convert the CMA-ES calibration results into a corresponding gate fidelity. To do this, we measure and fit the reference curve and obtain a value for P_{int} at the point at which the calibration was carried out. This has been done (curve not shown) and has yielded $P_{\text{ref}} = 0.964$. We can then reverse Eq. (I3) to calculate P_{int}:

$$P = [(S - C)/A]^{1/N}. \tag{I4}$$

Here, S is the CMA-ES raw data, whereas A and C are parameters obtained at the same time as fitting P_{ref}. We can then use Eq. (I1) to convert the CMA-ES data to fidelity values point by point. The results are shown in Fig. 34(b). Note that for the longest gate durations, the saturating gate fidelity is consistently lower, agreeing with the decoherence-limited predictions shown in Fig. 5 and described in Appendix F.

[1] P. W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, in *Proceedings 35th Annual Symposium on Foundations of Computer Science* (IEEE, Santa Fe, NM, USA, 1994), pp. 124–134.

[2] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, MA, USA, 2010).

[3] K. R. Brown, J. Kim, and C. Monroe, Co-designing a scalable quantum computer with trapped atomic ions, npj Quantum Inf. **2**, 16034 (2016).

[4] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[5] Y. Alexeev, D. Bacon, K. R. Brown, R. Calderbank, L. D. Carr, F. T. Chong, B. DeMarco, D. Englund, E. Farhi, B. Fefferman *et al.*, Quantum computer systems for scientific discovery, PRX Quantum **2**, 017001 (2021).

[6] J. P. Gaebler, T. R. Tan, Y. Lin, Y. Wan, R. Bowler, A. C. Keith, S. Glancy, K. Coakley, E. Knill, D. Leibfried, and D. J. Wineland, High-fidelity universal gate set for $^{9}Be^{+}$ ion qubits, Phys. Rev. Lett. **117**, 060505 (2016).

[7] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, Nature **508**, 500 (2014).

[8] J. Kelly, R. Barends, A. G. Fowler, A. Megrant, E. Jeffrey, T. C. White, D. Sank, J. Y. Mutus, B. Campbell, Y. Chen *et al.*, State preservation by repetitive error detection in a superconducting quantum circuit, Nature **519**, 66 (2015).

[9] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, Realization of high-fidelity CZ and *ZZ*-free iSWAP gates with a tunable coupler, Phys. Rev. X **11**, 021058 (2021).

[10] J. Stehlik, D. M. Zajac, D. L. Underwood, T. Phung, J. Blair, S. Carnevale, D. Klaus, G. A. Keefe, A. Carniol, M. Kumph, M. Steffen, and O. E. Dial, Tunable coupling architecture for fixed-frequency transmon superconducting qubits, Phys. Rev. Lett. **127**, 080505 (2021).

[11] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, High-threshold and low-overhead fault-tolerant quantum memory, Nature **627**, 778 (2024).

[12] Y. Xu, J. Chu, J. Yuan, J. Qiu, Y. Zhou, L. Zhang, X. Tan, Y. Yu, S. Liu, J. Li, F. Yan, and D. Yu, High-fidelity, high-scalability two-qubit gate scheme for superconducting qubits, Phys. Rev. Lett. **125**, 240503 (2020).

[13] I. N. Moskalenko, I. A. Simakov, N. N. Abramov, A. A. Grigorev, D. O. Moskalev, A. A. Pishchimova, N. S. Smirnov, E. V. Zikiy, I. A. Rodionov, and I. S. Besedin, High fidelity two-qubit gates on fluxoniums using a tunable coupler, npj Quantum Inf. **8**, 130 (2022).

[14] F. Marxer, A. Vepsäläinen, S. W. Jolin, J. Tuorila, A. Landra, C. Ockeloen-Korppi, W. Liu, O. Ahonen, A. Auer, L. Belzane *et al.*, Long-distance transmon coupler with CZ-gate fidelity above 99.8%, PRX Quantum **4**, 010314 (2023).

[15] L. Ding, M. Hays, Y. Sung, B. Kannan, J. An, A. Di Paolo, A. H. Karamlou, T. M. Hazard, K. Azar, D. K. Kim, B. M. Niedzielski, A. Melville, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, J. A. Grover, K. Serniak, and W. D. Oliver, High-fidelity, frequency-flexible two-qubit fluxonium gates with a transmon coupler, Phys. Rev. X **13**, 031035 (2023).

[16] H. Zhang, C. Ding, D. K. Weiss, Z. Huang, Y. Ma, C. Guinn, S. Sussman, S. P. Chitta, D. Chen, A. A. Houck, J. Koch, and D. I. Schuster, Tunable inductive coupler for high-fidelity gates between fluxonium qubits, PRX Quantum **5**, 020326 (2024).

[17] W.-J. Lin, H. Cho, Y. Chen, M. G. Vavilov, C. Wang, and V. E. Manucharyan, 24 days-stable CNOT gate on fluxonium qubits with over 99.9% fidelity, PRX Quantum **6**, 010349 (2025).

[18] M. Reagor, C. B. Osborn, N. Tezak, A. Staley, G. Prawiroatmodjo, M. Scheer, N. Alidoust, E. A. Sete, N. Didier, M. P. da Silva *et al.*, Demonstration of universal parametric entangling gates on a multi-qubit lattice, Sci. Adv. **4**, eaao3603 (2018).

[19] T. Noh, Z. Xiao, X. Y. Jin, K. Cicak, E. Doucet, J. Aumentado, L. C. G. Govia, L. Ranzani, A. Kamal, and R. W. Simmonds, Strong parametric dispersive shifts in a statically decoupled two-qubit cavity QED system, Nat. Phys. **19**, 1445 (2023).

[20] C. Zhou, P. Lu, M. Praquin, T.-C. Chien, R. Kaufman, X. Cao, M. Xia, R. S. K. Mong, W. Pfaff, D. Pekker, and M. Hatridge, Realizing all-to-all couplings among detachable quantum modules using a microwave quantum state router, npj Quantum Inf. **9**, 54 (2023).

[21] T. Brown, E. Doucet, D. Ristè, G. Ribeill, K. Cicak, J. Aumentado, R. Simmonds, L. Govia, A. Kamal, and L. Ranzani, Trade off-free entanglement stabilization in a superconducting qutrit-qubit system, Nat. Commun. **13**, 3994 (2022).

[22] A. D. Paolo, C. Leroux, T. M. Hazard, K. Serniak, S. Gustavsson, A. Blais, and W. D. Oliver, Extensible circuit-QED architecture via amplitude- and frequency-variable microwaves, arXiv:2204.08098.

[23] B. Marinelli, J. Luo, H. Ren, B. M. Niedzielski, D. K. Kim, R. Das, M. Schwartz, D. I. Santiago, and I. Siddiqi, Dynamically reconfigurable photon exchange in a superconducting quantum processor, arXiv:2303.03507.

[24] E. Zakka-Bajjani, F. Nguyen, M. Lee, L. R. Vale, R. W. Simmonds, and J. Aumentado, Quantum superposition of a single microwave photon in two different "colour" states, Nat. Phys. **7**, 599 (2011).

[25] M. Allman, J. D. Whittaker, M. Castellanos-Beltran, K. Cicak, F. da Silva, M. P. DeFeo, F. Lecocq, A. Sirois, J. D. Teufel, J. Aumentado, and R. W. Simmonds, Tunable resonant and nonresonant interactions between a phase qubit and *LC* resonator, Phys. Rev. Lett. **112**, 123601 (2014).

[26] Y. Lu, A. Maiti, J. W. O. Garmon, S. Ganjam, Y. Zhang, J. Claes, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, High-fidelity parametric beamsplitting with a parity-protected converter, Nat. Commun. **14**, 5767 (2023).

[27] N. Didier, E. A. Sete, M. P. da Silva, and C. Rigetti, Analytical modeling of parametrically modulated transmon qubits, Phys. Rev. A **97**, 022330 (2018).

[28] X. Li, Y. Ma, J. Han, T. Chen, Y. Xu, W. Cai, H. Wang, Y. P. Song, Z.-Y. Xue, Z.-Q. Yin, and L. Sun, Perfect quantum state transfer in a superconducting qubit chain with parametrically tunable couplings, Phys. Rev. Appl. **10**, 054009 (2018).

[29] M. Ganzhorn, G. Salis, D. J. Egger, A. Fuhrer, M. Mergenthaler, C. Müller, P. Müller, S. Paredes, M. Pechal, M. Werninghaus, and S. Filipp, Benchmarking the noise sensitivity of different parametric two-qubit gates in a single superconducting quantum computing platform, Phys. Rev. Res. **2**, 033447 (2020).

[30] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M.

P. da Silva, and B. R. Johnson, Demonstration of a parametrically activated entangling gate protected from flux noise, Phys. Rev. A **101**, 012302 (2020).

[31] A. Noguchi, A. Osada, S. Masuda, S. Kono, K. Heya, S. P. Wolski, H. Takahashi, T. Sugiyama, D. Lachance-Quirion, and Y. Nakamura, Fast parametric two-qubit gates with suppressed residual interaction using the second-order nonlinearity of a cubic transmon, Phys. Rev. A **102**, 062408 (2020).

[32] S. Li, D. Fan, M. Gong, Y. Ye, X. Chen, Y. Wu, H. Guan, H. Deng, H. Rong, H.-L. Huang *et al.*, Realization of fast all-microwave controlled-*Z* gates with a tunable coupler, Chin. Phys. Lett. **39**, 030302 (2022).

[33] E. A. Sete, V. Tripathi, J. A. Valery, D. Lidar, and J. Y. Mutus, Error budget of a parametric resonance entangling gate with a tunable coupler, Phys. Rev. Appl. **22**, 014059 (2024).

[34] T. Roy, Z. Li, E. Kapit, and D. Schuster, Two-qutrit quantum algorithms on a programmable superconducting processor, Phys. Rev. Appl. **19**, 064024 (2023).

[35] K. Nie, A. Bista, K. Chow, W. Pfaff, and A. Kou, Parametrically controlled microwave-photonic interface for the fluxonium, Phys. Rev. Appl. **22**, 054021 (2024).

[36] G. B. P. Huber, F. A. Roy, L. Koch, I. Tsitsilin, J. Schirk, N. J. Glaser, N. Bruckmoser, C. Schweizer, J. Romeiro, G. Krylov *et al.*, Parametric multielement coupling architecture for coherent and dissipative control of superconducting qubits, PRX Quantum **6**, 030313 (2025).

[37] A. Petrescu, C. Le Calonnec, C. Leroux, A. Di Paolo, P. Mundada, S. Sussman, A. Vrajitoarea, A. A. Houck, and A. Blais, Accurate methods for the analysis of strong-drive effects in parametric gates, Phys. Rev. Appl. **19**, 044003 (2023).

[38] Y. Lu, S. Chakram, N. Leung, N. Earnest, R. K. Naik, Z. Huang, P. Groszkowski, E. Kapit, J. Koch, and D. I. Schuster, Universal stabilization of a parametrically coupled qubit, Phys. Rev. Lett. **119**, 150502 (2017).

[39] D. R. Pérez, P. Varosy, Z. Li, T. Roy, E. Kapit, and D. Schuster, Error-divisible two-qubit gates, Phys. Rev. Appl. **19**, 024043 (2023).

[40] Z. Li, T. Roy, D. R. Pérez, K.-H. Lee, E. Kapit, and D. I. Schuster, Autonomous error correction of a single logical qubit using two transmons, Nat. Commun. **15**, 1681 (2024).

[41] Z. Li, T. Roy, D. R. Pérez, D. I. Schuster, and E. Kapit, Hardware efficient autonomous error correction with linear couplers in superconducting circuits, Phys. Rev. Res. **6**, 013171 (2024).

[42] B. Yurke, L. R. Corruccini, P. G. Kaminsky, L. W. Rupp, A. D. Smith, A. H. Silver, R. W. Simon, and E. A. Whittaker, Observation of parametric amplification and deamplification in a Josephson parametric amplifier, Phys. Rev. A **39**, 2519 (1989).

[43] L. Tian, M. S. Allman, and R. W. Simmonds, Parametric coupling between macroscopic quantum resonators, New J. Phys. **10**, 115001 (2008).

[44] J. D. Teufel, T. Donner, D. Li, J. W. Harlow, M. S. Allman, K. Cicak, A. J. Sirois, J. D. Whittaker, K. W. Lehnert, and R. W. Simmonds, Sideband cooling of micromechanical motion to the quantum ground state, Nature **475**, 359 (2011).

[45] L. Ranzani and J. Aumentado, Graph-based analysis of nonreciprocity in coupled mode system, New J. Phys. **17**, 023024 (2015).

[46] J. M. Gambetta, J. M. Chow, and M. Steffen, Building logical qubits in a superconducting quantum computing system, npj Quantum Inf. **3**, 2 (2017).

[47] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[48] M. D. Hutchings, J. B. Hertzberg, Y. Liu, N. T. Bronn, G. A. Keefe, M. Brink, J. M. Chow, and B. L. T. Plourde, Tunable superconducting qubits with flux-independent coherence, Phys. Rev. Appl. **8**, 044003 (2017).

[49] J. M. Chávez-Garcia, F. Solgun, J. B. Hertzberg, O. Jinka, M. Brink, and B. Abdo, Weakly flux-tunable superconducting qubit, Phys. Rev. Appl. **18**, 034057 (2022).

[50] J. Long, T. Zhao, M. Bal, R. Zhao, G. S. Barron, H.-S. Ku, J. A. Howard, X. Wu, C. R. H. McRae, X.-H. Deng, G. J. Ribeill, M. Singh, T. A. Ohki, E. Barnes, S. E. Economou, and D. P. Pappas, A universal quantum gate set for transmon qubits with strong *ZZ* interactions, arXiv:2103.12305.

[51] P. Mundada, G. Zhang, T. Hazard, and A. Houck, Suppression of qubit crosstalk in a tunable coupling superconducting circuit, Phys. Rev. Appl. **12**, 054023 (2019).

[52] P. Zhao, P. Xu, D. Lan, J. Chu, X. Tan, H. Yu, and Y. Yu, High-contrast *ZZ* interaction using superconducting qubits with opposite-sign anharmonicity, Phys. Rev. Lett. **125**, 200503 (2020).

[53] J. Ku, X. Xu, M. Brink, D. C. McKay, J. B. Hertzberg, M. H. Ansari, and B. L. T. Plourde, Suppression of unwanted *ZZ* interactions in a hybrid two-qubit system, Phys. Rev. Lett. **125**, 200504 (2020).

[54] P. Groszkowski and J. Koch, scQubits: A PYTHON package for superconducting qubits, Quantum **5**, 583 (2021).

[55] S. P. Chitta, T. Zhao, Z. Huang, I. Mondragon-Shem, and J. Koch, Computer-aided quantization and numerical analysis of superconducting circuits, New J. Phys. **24**, 103020 (2022).

[56] Z. Xiao, E. Doucet, T. Noh, L. Ranzani, R. W. Simmonds, L. C. G. Govia, and A. Kamal, Perturbative diagonalization for time-dependent strong interactions, Phys. Rev. Appl. **18**, 024009 (2022).

[57] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B: Quantum Semiclass. Opt. **7**, S347 (2005).

[58] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).

[59] F. W. Strauch, P. R. Johnson, A. J. Dragt, C. J. Lobb, J. R. Anderson, and F. C. Wellstood, Quantum logic gates for coupled superconducting phase qubits, Phys. Rev. Lett. **91**, 167005 (2003).

[60] A. D. Córcoles, J. M. Gambetta, J. M. Chow, J. A. Smolin, M. Ware, J. Strand, B. L. T. Plourde, and M. Steffen, Process verification of two-qubit quantum gates by randomized benchmarking, Phys. Rev. A **87**, 030301(R) (2013).

[61] Y. Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, B. Campbell, Z. Chen, B. Chiaro *et al.*,

Qubit architecture with high coherence and fast tunable coupling, Phys. Rev. Lett. **113**, 220502 (2014).

[62] L. Casparis, T. W. Larsen, M. S. Olsen, F. Kuemmeth, P. Krogstrup, J. Nygård, K. D. Petersson, and C. M. Marcus, Gatemon benchmarking and two-qubit operations, Phys. Rev. Lett. **116**, 150505 (2016).

[63] J. Helsen, X. Xue, L. M. K. Vandersypen, and S. Wehner, A new class of efficient randomized benchmarking protocols, npj Quantum Inf. **5**, 71 (2019).

[64] X. Xue, T. F. Watson, J. Helsen, D. R. Ward, D. E. Savage, M. G. Lagally, S. N. Coppersmith, M. A. Eriksson, S. Wehner, and L. M. K. Vandersypen, Benchmarking gate fidelities in a Si/SiGe two-qubit device, Phys. Rev. X **9**, 021011 (2019).

[65] B. K. Mitchell, R. K. Naik, A. Morvan, A. Hashim, J. M. Kreikebaum, B. Marinelli, W. Lavrijsen, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, Hardware-efficient microwave-activated tunable coupling between superconducting qubits, Phys. Rev. Lett. **127**, 200502 (2021).

[66] J. Kelly, R. Barends, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, I. C. Hoi, E. Jeffrey *et al.*, Optimal quantum control using randomized benchmarking, Phys. Rev. Lett. **112**, 240504 (2014).

[67] 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. **7**, 14 (2021).

[68] T. Abad, J. Fernández-Pendás, A. F. Kockum, and G. Johansson, Universal fidelity reduction of quantum operations from weak dissipation, Phys. Rev. Lett. **129**, 150504 (2022).

[69] J. B. Hertzberg, E. J. Zhang, S. Rosenblatt, E. Magesan, J. A. Smolin, J. -B. Yau, V. P. Adiga, M. Sandberg, M. Brink, J. M. Chow, and J. S. Orcutt, Laser-annealing Josephson junctions for yielding scaled-up superconducting quantum processors, npj Quantum Inf. **7**, 129 (2021).

[70] I. A. Simakov, G. S. Mazhorin, I. N. Moskalenko, S. S. Seidov, and I. S. Besedin, High-fidelity transmon-coupler-activated *CCZ* gate on fluxonium qubits, Phys. Rev. Appl. **21**, 044035 (2024).

[71] E. Altman, K. R. Brown, G. Carleo, L. D. Carr, E. Demler, C. Chin, B. DeMarco, S. E. Economou, M. A. Eriksson, K.-M. C. Fu *et al.*, Quantum simulators: Architectures and opportunities, PRX Quantum **2**, 017003 (2021).

[72] H. Singh, L. W. Bertels, D. Claudino, S. E. Economou, E. Barnes, N. P. Bauman, K. Kowalski, and N. J. Mayhall, Qubit-efficient quantum chemistry with the adapt variational quantum eigensolver and double unitary downfolding, J. Chem. Theory Comput. **21**, 8799 (2025).

[73] J. Majer, J. M. Chow, J. M. Gambetta, J. Koch, B. R. Johnson, J. A. Schreier, L. Frunzio, D. I. Schuster, A. A. Houck, A. Wallraff, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, Coupling superconducting qubits via a cavity bus, Nature **449**, 443 (2007).

[74] M. A. Sillanpää, J. I. Park, and R. W. Simmonds, Coherent quantum state storage and transfer between two phase qubits via a resonant cavity, Nature **449**, 438 (2007).

[75] G. F. Diotallevi, L. Reascos, and M. Benito, SymPT: A comprehensive tool for automating effective Hamiltonian derivations, arXiv:2412.10240.

[76] A. Kerman, Efficient numerical simulation of complex Josephson quantum circuits, arXiv:2010.14929.

[77] M. Reagor, W. Pfaff, C. Axline, R. W. Heeres, N. Ofek, K. Sliwa, E. Holland, C. Wang, J. Blumoff, K. Chou, M. J. Hatridge, L. Frunzio, M. H. Devoret, L. Jiang, and R. J. Schoelkopf, Quantum memory with millisecond coherence in circuit QED, Phys. Rev. B **94**, 014506 (2016).

[78] G. A. Peterson, Parametric coupling between microwaves and motion in quantum circuits: Fundamental limits and applications. ProQuest Dissertation and Theses, 27837930, University of Colorado at Boulder, 2020.

[79] N. Roch, M. E. Schwartz, F. Motzoi, C. Macklin, R. Vijay, A. W. Eddins, A. N. Korotkov, K. B. Whaley, M. Sarovar, and I. Siddiqi, Observation of measurement-induced entanglement and quantum trajectories of remote superconducting qubits, Phys. Rev. Lett. **112**, 170501 (2014).

[80] D. Ristè, M. Dukalski, C. A. Watson, G. de Lange, M. J. Tiggelman, Y. M. Blanter, K. W. Lehnert, R. N. Schouten, and L. DiCarlo, Deterministic entanglement of superconducting qubits by parity measurement and feedback, Nature **502**, 350 (2013).

[81] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, J. Heinsoo, J.-C. Besse, M. Gabureac, A. Wallraff, and C. Eichler, Entanglement stabilization using ancilla-based parity detection and real-time feedback in superconducting circuits, npj Quantum Inf. **5**, 69 (2019).

[82] A. Kubica, A. Haim, Y. Vaknin, H. Levine, F. Brandão, and A. Retzker, Erasure qubits: Overcoming the T_1 limit in superconducting circuits, Phys. Rev. X **13**, 041022 (2023).

[83] H. Levine, A. Haim, J. S. C. Hung, N. Alidoust, M. Kalaee, L. DeLorenzo, E. A. Wollack, P. Arrangoiz-Arriola, A. Khalajhedayati, R. Sanil *et al.*, Demonstrating a long-coherence dual-rail erasure qubit using tunable transmons, Phys. Rev. X **14**, 011051 (2024).

[84] B. T. Gard, Z. Parrott, K. Jacobs, J. Aumentado, and R. W. Simmonds, Fast high-fidelity quantum nondemolition readout of a superconducting qubit with tunable transverse couplings, Phys. Rev. Appl. **21**, 024008 (2024).

[85] J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, M. B. Ketchen, and M. Steffen, Characterization of addressability by simultaneous randomized benchmarking, Phys. Rev. Lett. **109**, 240504 (2012).

[86] M. Xia, C. Zhou, C. Liu, P. Patel, X. Cao, P. Lu, B. Mesits, M. Mucci, D. Gorski, D. Pekker, and M. Hatridge, Fast superconducting qubit control with sub-harmonic drives, arXiv:2306.10162.

[87] A. Sah, S. Kundu, H. Suominen, Q. Chen, and M. Möttönen, Decay-protected superconducting qubit with fast control enabled by integrated on-chip filters, Commun. Phys. **7**, 227 (2024).

[88] Wikipedia contributors, Covariance matrix adaptation evolution strategy—Wikipedia, the free encyclopedia (2023). https://en.wikipedia.org/wiki/CMA-ES. [Online; accessed June 15,2023].

[89] D. Nielsen, Deep Learning Cage Match: Gradient Descent vs CMA-ES, 2018, https://github.com/DuaneNielsen/cma-es.