

Implementation of a Quantum Addressable Router Using Superconducting Qubits

Connie Miao^{1,*}, Sébastien Léger^{1,*}, Ziqian Li¹, Gideon Lee², Liang Jiang² and David I. Schuster¹

¹*Departments of Physics and Applied Physics, Stanford University, Stanford, California 94305, USA*

²*Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA*

 (Received 7 April 2025; revised 14 August 2025; accepted 20 August 2025; published 13 November 2025)

The implementation of a quantum router capable of performing both quantum signal routing and quantum addressing (a Q^2 -router) represents a key step toward building quantum networks and quantum random access memories. We realize a Q^2 -router that uses fixed-frequency transmon qubits to implement a routing protocol based on two native controlled- i SWAP gates. These gates leverage a large ZZ interaction to selectively route information according to a quantum address. We find an estimated average routing fidelity of 95.3%, with errors arising primarily from decoherence or state preparation and measurement. We present a comprehensive calibration and characterization of both the c - i SWAP gates and the overall routing protocol through randomized benchmarking techniques and state tomography.

DOI: [10.1103/pq3x-cmw9](https://doi.org/10.1103/pq3x-cmw9)

I. INTRODUCTION

Routing information is crucial for classical processing and is even more essential in quantum systems due to the inherent constraint that quantum states cannot be duplicated [1]. The most general form of routing is achievable with a Q^2 -router, which allows both the data and address to be quantum mechanical [2]. Classical routing of quantum states plays a key role in quantum networking [3,4]. Similarly, the ability to direct information to superpositions of paths via quantum addresses enables efficient quantum random access memory (QRAM) [5–7], which can aid algorithms such as Grover’s search [8], quantum matrix inversion [9], quantum deep learning [10], and quantum simulation [11,12]. Additionally, quantum addressing itself offers applications in quantum communication [13–15], quantum calibration [16,17], data compression [18], and error filtration [19,20].

Numerous platforms have experimentally explored classical routing of quantum states, including routing between qubits [21–23] and the directional emission of microwave

[24–27] and optical [28–30] photons. However, the classical routers in these studies do not possess an internal addressing mechanism, as the routing direction must be encoded externally. This encoding is typically achieved, for instance, by applying different pulses corresponding to the routing direction or by altering the design of the sample. On the other hand, routing a state based on a quantum address has received various theoretical proposals based on photonics [31], Rydberg atoms [32], superconducting circuits [33–35], and hybrid architectures [34,36]. Alongside these proposals, several groups [37–39] have implemented three-qubit gates that could be used as building blocks for a quantum addressing operation. A Q^2 -router has been implemented on a photonic platform [40], but its dependence on non-deterministic post-selection hinders its applicability. As a result, a fully deterministic Q^2 -router has yet to be demonstrated.

In this work, we use superconducting circuits to realize a Q^2 -router able to handle both a quantum signal and a quantum address. For minimal overhead, we employ fixed-frequency transmon qubits. The routing protocol, inspired by previous theoretical proposals [34], consists of two controlled- i SWAP (c - i SWAP) gates. Each of these gates is natively implemented with an $|eg\rangle$ to $|gf\rangle$ two-qubit sideband gate and uses a large ZZ interaction to control the i SWAPs. By using two native c - i SWAPs, each taking the time of just a single two-qubit gate, we circumvent the need to decompose the quantum routing protocol into a much larger number of two-qubit gates. The fidelities of these c - i SWAPs are estimated to be $97.77 \pm 0.13\%$ and $98.62 \pm 0.02\%$ within the valid subspace. The routing

*These authors contributed equally to this work.

[†]Contact author: cmiao@stanford.edu

[‡]Contact author: sebleger@stanford.edu

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.

protocol has an average fidelity of 95.3% and we estimate that qubit decoherence and state preparation and measurement are the main sources of error. In Sec. II we present our router design and protocol. In Sec. III we explain how the c - i SWAPs used in the protocol are calibrated and we benchmark them individually using randomized benchmarking and error amplification. Finally, in Sec. IV we estimate the fidelity of the routing protocol using state tomography and error amplification.

II. ROUTER LAYOUT AND PROTOCOL

Our Q^2 -router is based on four fixed-frequency, grounded transmon qubits with fixed capacitive couplings. An optical image of the sample is shown in Fig. 1(a) with the connectivity labeled in Fig. 1(b). The circuit can be used in any of the four classical and quantum modes of information routing shown in Fig. 1(c).

To enable fast two-qubit gates, we have strong couplings on the order of $g/2\pi \sim 50$ MHz between the input and the switch as well as between the input and each of the two outputs. The key ingredient to enable native c - i SWAPs is a relatively strong ZZ interaction between the input and switch ($\zeta_{SI}/2\pi = 5.4$ MHz). Each qubit is also capacitively coupled to its own charge line for control and its own resonator, which enables four-qubit simultaneous readout. This sample was fabricated and packaged by the

MIT SQUILL foundry and the measured qubit parameters and lifetimes are reported in Table I in Appendix B.

In our convention, the input state [dark gray in Fig. 1(a)] is routed to output 1 (2) (red, blue) if the switch (purple) is in $|g\rangle$ ($|e\rangle$), or to a superposition of the outputs with the relative amplitudes given by the switch amplitudes in $|g\rangle$ or $|e\rangle$. This routing operation U_{route} can be written as

$$\begin{aligned} & |\psi\rangle_I \left(\alpha |0\rangle + \beta |1\rangle \right)_S |0\rangle_{O_1} |0\rangle_{O_2} \\ & \quad \downarrow U_{\text{route}} \\ & |0\rangle_I \left(\alpha |0\rangle_S |\psi\rangle_{O_1} |0\rangle_{O_2} + \beta |1\rangle_S |0\rangle_{O_1} |\psi\rangle_{O_2} \right), \end{aligned}$$

where I, S, O_1 , and O_2 represent the input, switch, output 1, and output 2 qubits respectively. To perform the routing protocol, we tile one c - i SWAP, which swaps the input and output 1 controlled on the switch in $|g\rangle$, with another c - i SWAP, which swaps the input and output 2 controlled on the switch in $|e\rangle$ [Fig. 1(d)]. Valid router states and operations are a subset of the full four-qubit state and control space. Most importantly, the router module should have at most one total photon shared between the input and outputs at all times, with the routing protocol swapping this photon between the input and the outputs. Thus, for simplicity, we always initialize the outputs in their ground states and so we expect the input to always be in its ground state at the end of the protocol. Because we only need the protocol to

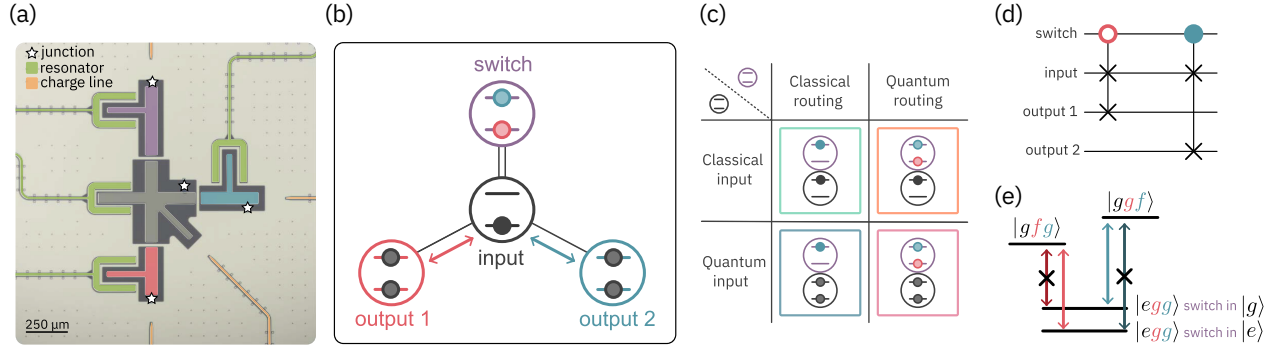


FIG. 1. Device and Q^2 -router concept. (a) False colored optical microscope image of the Q^2 -router device with four transmon qubits [switch (purple), input (dark gray), output 1 (red), and output 2 (blue)], their readout resonators (green) and charge drive lines (yellow), and their Josephson junctions (stars). (b) Schematic of our Q^2 -router concept in a possible three-qubit entangled final state after one routing operation. Shaded state circles indicate equal, entangled superposition states while solid circles indicate classical states. The colors of the switch states indicate to which output each switch state controls routing, and the shared gray coloring between input and output states indicate that excitation is routed from input to outputs according to the switch superposition. Lines between pairs of qubits indicate strong couplings; the switch and input are particularly strongly ZZ coupled (5.4 MHz). The red and blue arrows indicate the pairs of qubit between which we drive two-qubit swaps to implement the protocol. (c) Breakdown of the operational modes of a general quantum router depending on whether the input and routing are classical or quantum. The Q^2 -router is capable of performing the most general form of routing where both the input and routing are quantum. The colored boxes are used as a reference to the different operational modes in later figures of this work. (d) Quantum circuit describing the protocol: (1) i SWAP between input and output 1, controlled on switch in $|g\rangle$; (2) i SWAP between input and output 2, controlled on switch in $|e\rangle$. (e) Level diagram of the input and output states relevant to the native $|eg\rangle$ to $|gf\rangle$ swaps used in the routing protocol. We label levels as $|\text{in}, \text{out 1}, \text{out 2}\rangle$. The input-output 1 swap (light or dark red) is far-detuned from the input-output 2 swap (light or dark blue), allowing selective swapping. We control each swap on the state of the switch by using the large ZZ interaction between the switch and input, which blocks the swap when the switch starts in the wrong state due to a shift in the sideband transition frequency.

operate correctly within this restricted subspace, within the protocol any physical Z rotations can be compensated with a virtual one (Appendix G 2).

We realize our c - i SWAP gate by activating a four-wave mixing interaction between $|eg\rangle$ and $|gf\rangle$ of input and output 1 (2). We do so by driving a sideband pulse on output 1 (2) at the frequency difference between $|eg\rangle$ and $|gf\rangle$ [41] for a long enough time to perform $\sqrt{i}$ SWAP. To prevent phase errors due to off-resonant $\sqrt{i}$ SWAP and i SWAP gates in randomized benchmarking (Sec. III B), we decompose each i SWAP into two $\sqrt{i}$ SWAP pulses. To make the i SWAP conditional, we leverage the strong ZZ interaction between the input and switch. Because of this interaction, the input, and therefore the $|eg\rangle - |gf\rangle$ frequencies, depends on the state of the switch [Fig. 1(e)]. With an appropriate gate calibration, we can thus ensure that population is only transferred when the switch is in the desired state. We note that combined with the same ZZ -based principle to implement the control, other two-qubit sideband schemes could be used instead of $|eg\rangle - |gf\rangle$. We choose the $|eg\rangle - |gf\rangle$ interaction because it is a single photon transition, making it fast and simple to implement [41,42], especially given our relatively far-detuned qubit frequencies. With our protocol the outputs are encoded in the g - f subspace as opposed to the more standard g - e , which is a valid choice for either networking or QRAM applications (Appendix I).

III. C- i SWAP CALIBRATION AND CHARACTERIZATION

A. Calibration

To achieve high-fidelity c - i SWAPs, we must satisfy two goals. First, each sideband must implement a high-fidelity i SWAP. Second, the two pulses must be selective, performing the swap only when the switch is in the correct state $|S\rangle$ for the i SWAP to proceed, where $|S\rangle = |g\rangle$ for input to output 1 or $|S\rangle = |e\rangle$ for input to output 2.

To implement the i SWAP, we need to calibrate the drive frequency, amplitude, and gate time. Although the frequency of the transition is close to $f_{g-f}^{\text{out}} - f_{g-e}^{\text{in}}$, the exact values of these parameters are interdependent because the qubits are AC-stark shifted during the drive [41]. We first establish a relationship between these parameters through an initial calibration for different possible pulse lengths (Appendix C). We choose pulse parameters to maximize contrast given the finite $\zeta_{\text{SI}}/2\pi = 5.4$ MHz shift in the sideband, such that the desired transition is driven strongly to obtain a Rabi rate g_{eg-gf} while the undesired transition is weakly driven at Ω . It is possible to ensure no state transfer in the blocking state despite still having weak driving if we pick $\Omega = \sqrt{g_{eg-gf}^2 + \zeta_{\text{SI}}^2}$ to be $2n g_{eg-gf}$ where n is an integer [43] (see Appendix F for further discussion). We explain the calibration of the optimal point in Appendix C.

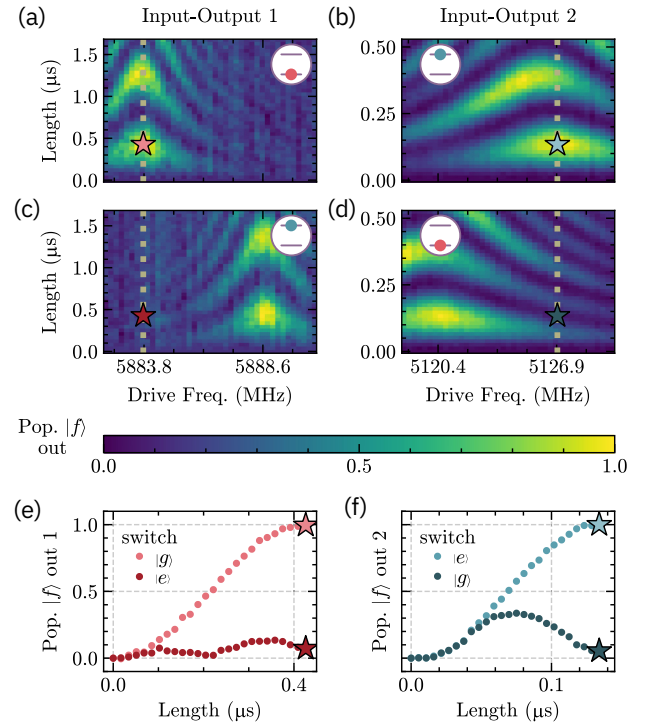


FIG. 2. **c - i SWAP calibration.** Rabi oscillations for each swap measuring the target output $|f\rangle$ population, sweeping pulse length and drive frequency. Because we would like to calibrate a $\sqrt{i}$ SWAP, we apply two stacked pulses whose total pulse time corresponds to the labeled axis. We initialize the input in $|e\rangle$ and omit the data for the input, which experiences a complementary oscillation. (a),(b) Sweep initializing the switch in $|S\rangle$, the state to allow the swap. The pulse time is marked with a light red or blue star. (c),(d) Sweep initializing the switch in $|\bar{S}\rangle$, the state to block the swap. With this initialization, there is a minimum of population transfer at the calibrated pulse time (dark red or blue star). (e),(f) Slices through the Rabi oscillations in previous panels at the calibrated swap frequency (dashed lines). At the calibrated length (stars), we observe an i SWAP (4π rotation) when the switch is in $|S\rangle$ ($|\bar{S}\rangle$).

In Figs. 2(a) and 2(b) we show the resulting Rabi oscillation between the input and output 1 (2) vs pulse length and drive frequency at the calibrated drive amplitude when the switch qubit is in $|S\rangle$. In Figs. 2(b) and 2(c), we prepare the switch in the blocking state $|\bar{S}\rangle = |e\rangle$ ($|\bar{S}\rangle = |g\rangle$) and again measure the Rabi oscillation between the input and output 1 (2) vs pulse length and drive frequency. We see that the resonant frequency is shifted and that the previously calibrated point now corresponds to a minimum of population in the output qubit. In Figs. 2(e) and 2(f), we show slices through the Rabi oscillations from the previous panels at the selected optimal frequency, initializing the switch in either $|S\rangle$ or $|\bar{S}\rangle$. The pulse times are respectively 426 and 134 ns, corresponding to $n = 2$ and $n = 1$ for output 1 and 2 (see Appendices C and E 3 for more details).

B. Benchmarking individual c-*i*SWAPs

Before benchmarking the full router protocol we first quantify the errors in the individual c-*i*SWAPs. We treat the subspace $|Seg\rangle - |Sgf\rangle$ as a single dual-rail logical qubit on which we perform reference and interleaved randomized benchmarking (RB) [44–47]. Within the logical subspace, the sideband is described by the single qubit rotation unitary

$$U^{\text{sub}}(\varphi) = \exp \left[-i\frac{\pi}{4} (e^{i\varphi} |Seg\rangle\langle Sgf| + e^{-i\varphi} |Sgf\rangle\langle Seg|) \right].$$

From U^{sub} , we define $X_{\pi/2}^{\text{sub}} \equiv U^{\text{sub}}(0)$ and $Y_{\pi/2}^{\text{sub}} \equiv U^{\text{sub}}(\pi/2)$. The Clifford set is constructed from $\{I, \pm X_{\pi/2}^{\text{sub}}, \pm Y_{\pi/2}^{\text{sub}}\}$. To perform the reference RB [Figs. 3(c) and 3(d)], we prepare the switch, input, and target output in $|Seg\rangle$ and apply a sequence of gates selected randomly from the Clifford group. For the interleaved RB [Figs. 3(e) and 3(f)], we add an extra gate that we wish to characterize after each randomly selected Clifford gate in the sequence [Fig. 3(b)] [47]. In both cases we then apply the inversion gate and measure the switch, input, and target output simultaneously to determine the probability of these qubits having returned successfully to their initial states. The full RB experimental details and data analysis procedures are described in Appendix E.

An advantage of treating the c-*i*SWAP as a conditional dual-rail single qubit is that we can distinguish errors that leave the subspace [Fig. 3(a)] by measuring the populations of all relevant qubit levels $\{|g/e\rangle \otimes |g/e\rangle \otimes |g/e/f\rangle\}$. We estimate the probability of remaining in the logical subspace $p_{\text{sub}} = p_{|Seg\rangle} + p_{|Sgf\rangle}$ at the end of the sequence, as shown in Figs. 3(c)–3(f). From fitting p_{sub} [Eq. (E2)] in the reference RB, we can estimate a photon loss per Clifford for each sideband of $L_{\text{IO}_1} = 1.88 \pm 0.16\%$ ($L_{\text{IO}_2} = 0.83 \pm 0.01\%$) for input-output 1 (2), where error bars represent 95% confidence from the fitting error. Interestingly, we see that the survival probability $p_{|Seg\rangle}$ closely matches the subspace survival p_{sub} , indicating that the gate error is dominated by photon loss. The remaining logical error is estimated from the survival probability post-selected on the probability to stay in the logical subspace, defined as $p_{\text{survival}} = p_{|Seg\rangle}/p_{\text{sub}}$, also shown in Figs. 3(c)–3(f). Fitting p_{survival} from the reference RB with an exponential [Eq. (E3)] gives the total error taking into account the photon loss error as $\epsilon_{\text{IO}_1} = 3.05 \pm 0.11\%$ ($\epsilon_{\text{IO}_2} = 1.36 \pm 0.01\%$) per Clifford. Notably, the reference RB for input-output 1 deviates from an exponential for circuits with over 150 gates, likely due to the small p_{sub} above this depth. We limit exponential fitting to depths < 150 to make this deviation apparent, though fitting up to 200 gates yields the same error estimate.

For the routing protocol, the only necessary gate is X_{π}^{sub} between the input and each output. To measure the fidelity

of this gate specifically we perform interleaved RB. Following the same experimental and fitting procedure as in the reference RB, we estimate both the photon loss contribution and the total error per gate, with the results shown in Figs. 3(e) and 3(f). For the *i*SWAP between input and output 1 (2), we estimate the average total error per X_{π}^{sub} to be $\epsilon_{\text{IO}_1}^X = 2.24 \pm 0.13\%$ ($\epsilon_{\text{IO}_2}^X = 1.38 \pm 0.02\%$), with photon loss rate per X_{π}^{sub} $L_{\text{IO}_1}^X = 1.32 \pm 0.08\%$ ($L_{\text{IO}_2}^X = 0.85 \pm 0.02\%$) (see Appendix E for more details). The gate error for input-output 2 is about half that of the gate with output 1. This difference is mainly due to the input-output 2 gate being shorter and therefore less sensitive to decoherence. These fidelities are comparable with other c-SWAP-like gates implemented with superconducting or bosonic qubits [37–39]. We note that because our protocol assumes there is at most one excitation shared between the input and output qubits, we do not measure the fidelity of our c-*i*SWAP on $|ef\rangle_{\text{IO}}$.

After benchmarking each *i*SWAP, we confirm proper control by the switch, using error amplification to highlight amplitude deviations from a perfect $2\pi n$ rotation when the switch is in the blocking state $|\bar{S}\rangle$. We prepare the switch in $|\bar{S}\rangle$ and measure the population in both $|\bar{S}gf\rangle$ and $|\bar{S}ge\rangle$ after applying $N = 15$ pulses [Figs. 3(g) and 3(h)]. We include both of these states to account for population that may have transferred to the output and decayed over the course of the error amplification. We use this method instead of RB because it coherently adds unintended small rotations, making it a more sensitive measure of the contrast. With this procedure, we find on average 0.3% (0.1%) undesired transfer per c-*i*SWAP between input and output 1 (2). To demonstrate the c-*i*SWAP on-off ratio we also initialize the switch in $|S\rangle$ and observe decaying oscillations as a function of N .

With their simplicity, high on-off ratio, and high fidelity in the non-blocking state, these gates meet all the key criteria for integration into the routing protocol.

IV. PROTOCOL BENCHMARKING

To quantify the performance of the router we apply the full protocol U_{route} and then perform three-qubit (3Q) state tomography on the switch, output 1, and output 2 [Fig. 4(a), experimental details in Appendix G 1]. In all of our 3Q tomography, we skip measurement of the input, as it is expected to always be in the ground state at the end of U_{route} regardless of the initialization. We select 16 cardinal states to initialize on the switch and input, constructed from the tensor product of classical ($|0\rangle, |1\rangle$) or superposition ($|+\rangle \equiv (|0\rangle + |1\rangle)/\sqrt{2}$, $|+i\rangle \equiv (|0\rangle + i|1\rangle)/\sqrt{2}$) states. We build the initial state preparation and tomography pulses using optimal control pulses to realize $g - e$ pulses that are robust against the strong ZZ coupling between the switch and input (Appendix D). Before

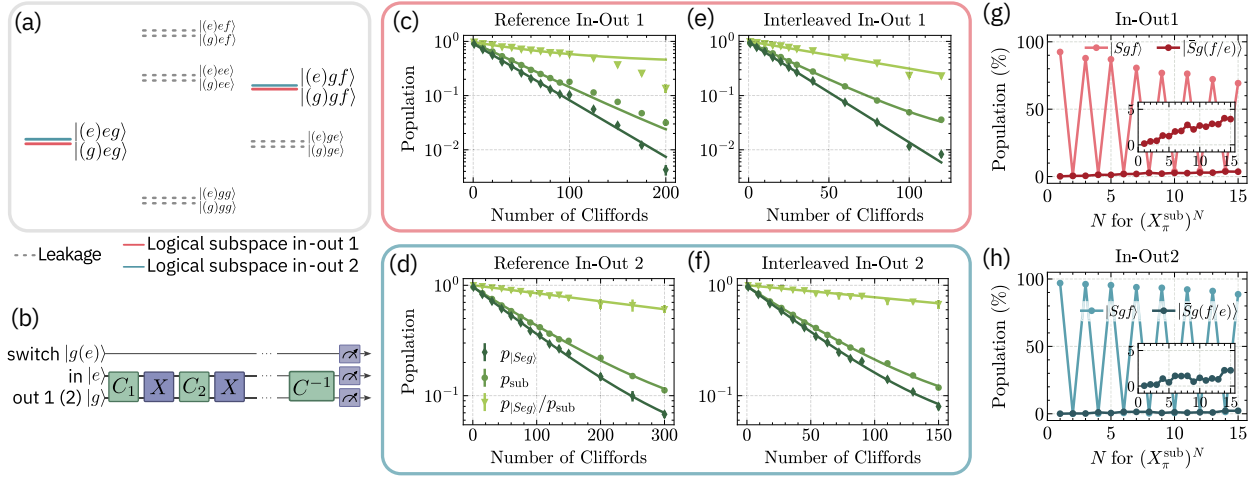


FIG. 3. **Subspace randomized benchmarking of c-iSWAPs.** (a) Level diagram of states probed for each randomized benchmarking experiment on the switch, input, and relevant output. We treat $|Seg\rangle - |Sgf\rangle$ as the logical two-level subspace and the Clifford group is defined relative to this subspace. Population measured in any other states is considered leakage. (b) Gate sequence used to implement interleaved RB. C_i is the i th randomly selected Clifford gate while C^{-1} is the inversion Clifford. For the reference RB, the interleaved X is removed. (c),(d) Reference RB measured population vs number of Clifford gates. $p_{|Seg\rangle}$ is the survival probability, $p_{\text{sub}} = p_{|Seg\rangle} + p_{|Sgf\rangle}$ is the probability to stay in the logical subspace and $p_{|Seg\rangle}/p_{\text{sub}}$ is the survival probability post-selected for the counts that are still in the subspace. (e),(f) Interleaved RB measured population vs number of Clifford gates. (g),(h) Error amplification of the c-iSWAP when the switch state is in either $|S\rangle$ (light red, light blue) or $|\bar{S}\rangle$ (dark red, dark blue), measuring the population in $|Sgf\rangle$ or in $|\bar{S}g(f/e)\rangle$ respectively. After 15 gates, we find 3.6% transfer (0.3% per c-iSWAP) into output 1 and 2.2% transfer (0.1% per c-iSWAP) into output 2 when the switch is in $|\bar{S}\rangle$.

applying the tomography pulses we bring the outputs back to the $g - e$ subspace using $\pi e - f$ pulses.

To characterize the performance of the protocol for each cardinal state initialization, we calculate the fidelity of the measured density matrix compared to the ideal final state. We show this comparison for the $|+\rangle_S |+\rangle_I$ initialization in Fig. 4(b). In Fig. 4(c) we show the average fidelity binned in each of the four configurations of the routing and input behavior (classical and quantum). The average fidelity over all 16 initializations is $\bar{\mathcal{F}} = 95.3\%$, where each fidelity is calculated as $\mathcal{F} = \langle \psi_{\text{ideal}} | \rho_{\text{meas}} | \psi_{\text{ideal}} \rangle$ [48], ρ_{meas} is the measured density matrix, and $|\psi_{\text{ideal}}\rangle$ is the ideal state. Next, we estimate the contributions to the measured error. Using a master equation simulation taking into account the coherence times of the $g - e$ transitions of all qubits and the $e - f$ transitions of the two outputs (Table I), we find an average estimated error of 2.7% due to decoherence. Another error source is the imperfect preparation of the initial states, with a large contribution due to ZZ. We estimate SPAM error by performing 2Q tomography on the initial states and sending the measured density matrices through the ideal routing unitary [Eq. (G3)] to use as the reference for the fidelity calculation. The difference in fidelity from this calculation compared to the previous calculation that references the ideal final state is about 1.7%. Both the decoherence and SPAM errors are compiled in Fig. 4(c) and match well with the averaged measured error. The remaining discrepancy between the

measured and total estimated errors likely arises from errors in reconstructing the density matrices in the 3Q and 2Q tomography (see Appendix H for a state-by-state comparison).

Typically, benchmarking techniques such as randomized benchmarking, cycle benchmarking, or cross entropy benchmarking [44,49,50] can be used to quantify gate fidelity while being insensitive to SPAM errors. However, these approaches are difficult to apply to the full protocol due to the restricted subspace that is allowed to have population. We instead estimate the routing fidelity by applying odd multiples of the protocol, where on every even iteration we apply the backward protocol U'_{route} , defined as reversing the order of the two c-iSWAPs [Fig. 4(a)]. The sequence is thus $(U_{\text{route}} U'_{\text{route}})^N U_{\text{route}}$ for variable N . This method of error amplification is not guaranteed to follow an exponential decay as it is sensitive to coherent error, but it is a useful metric as a comparison to the RB. We obtain the error per protocol ϵ by fitting the measured fidelity averaged for each configuration $\mathcal{F}(N)$ to

$$\mathcal{F}(N) = A(1 - \epsilon)^{2N+1}. \quad (1)$$

From this calculation, we find that the measured error rates are $1.6 \pm 0.9\%$, $2.6 \pm 1.3\%$, $2.4 \pm 0.9\%$, and $3.3 \pm 0.4\%$ respectively for the CrCi, CrQi, QrCi, and QrQi configurations of classical or quantum routing and input, where the errors bars represent 95% confidence from the fitting error,

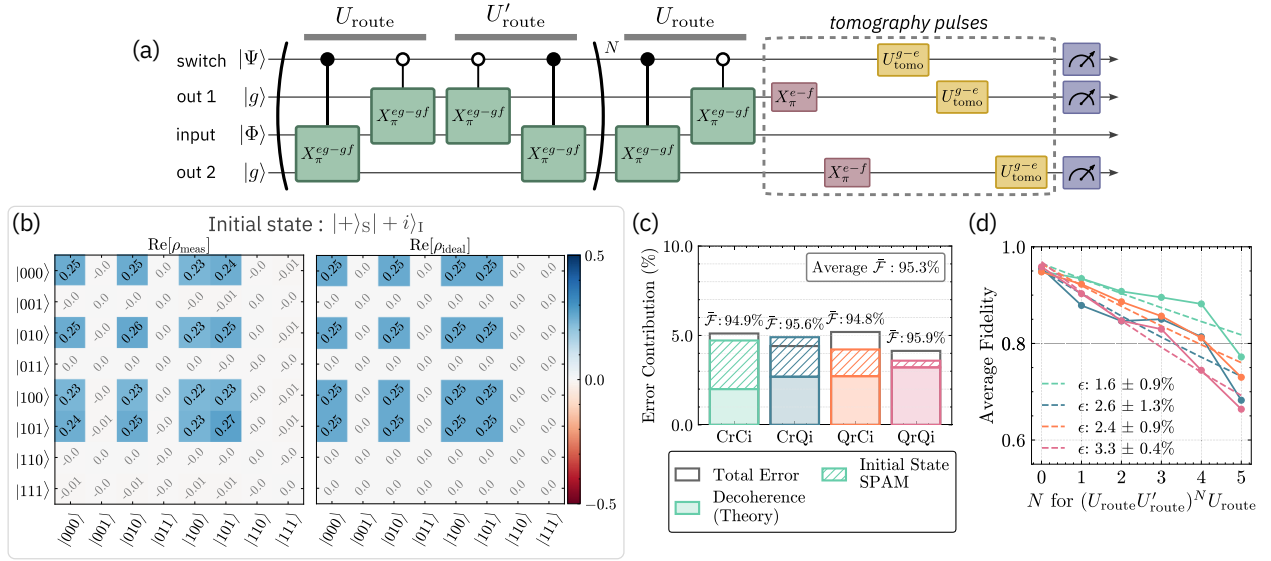


FIG. 4. Protocol fidelity from state tomography. (a) Pulse diagram for tomography experiments. Tomography pulses on the $g - e$ basis U_{tomo}^{g-e} are applied based on the measurement basis. A single protocol corresponds to the pulse diagram with $N = 0$. (b) Left (right) panel: real component of the measured (ideal) 3Q tomography on the switch and outputs 1 and 2 after one routing protocol U_{route} starting with the switch in $|+\rangle$ and the input in $|+i\rangle$. The measured density matrix has a fidelity of $\mathcal{F} = 96.4\%$ compared to the ideal final state. (c) Protocol fidelities and error contributions for the four switch and input configurations. C or Q denote classical or quantum, and r and i denote routing and input. Each configuration is averaged over the four corresponding cardinal states. Total error and fidelities (gray) are calculated compared to the ideal state as in (b). Decoherence error (shaded fill) is estimated from master equation simulations using the measured qubit coherence times (Table I). State preparation error (hatched fill) is estimated by calculating fidelity with respect to the measured initial state on the switch and input instead of the ideal one. (d) To amplify error per routing operation and disentangle gate error from SPAM, we measure the fidelity decay of the final state vs the number of applied forward (U_{route}) and backward (U'_{route}) protocols. The error is estimated using Eq. (1) and aligns with the decoherence limited fidelity.

for an overall fidelity of $97.5 \pm 0.5\%$. The error amplification results are shown in Fig. 4(d). These error rates closely match the decoherence-caused error rates estimated from simulation, 2.0%, 2.7%, 2.7%, and 3.2% respectively for the four configurations, suggesting that the routing protocol is likely decoherence-limited. These errors can be understood intuitively: for all configurations, the T_1 decay of any qubit impacts the fidelity. For QrCi, both the T_1 decay and the dephasing of the switch degrade the fidelity. For CrQi, because the quantum input state is transferred to one of the output states, the dephasing of either the input or outputs (but not the switch) contributes to the error. For QrQi, dephasing on any qubit affects the protocol. This analysis explains why the error-amplified and simulated decay-limited fidelities are both smallest for the CrCi configuration, increase for QrCi and CrQi, and are largest for QrQi.

Finally, we note that we cannot directly compare fidelities measured for individual c-*i*SWAPs via RB with those measured for the protocol via tomography. In the protocol, the two c-*i*SWAPs are applied sequentially, so the second c-*i*SWAP operates on an initial state that has undergone a wait time during which errors can occur, while no such wait time is relevant for the RB. In addition, the RB experiments always have one total excitation between the

input and the output while the tomography-based benchmarking averages over cardinal states with no excitations. These initializations have higher fidelities due to reduced sensitivity to decoherence. Finally, since the RB experiments are never initialized in a superposition state, only tomography-based benchmarking probes the capability for quantum routing.

V. CONCLUSION

We have implemented a Q^2 -router with minimal overhead, capable of deterministically routing classical or quantum signals via classical or quantum addresses. The routing protocol uses two native c-*i*SWAP gates, enabled by an $|eg\rangle - |gf\rangle$ drive for population transfer and a strong ZZ interaction between the switch and input qubits for selectivity. The measured routing fidelity is 95.3% from which about 1.7% error is estimated to come from SPAM. The remaining error is compatible with our estimate of a decoherence-limited protocol.

To enhance the future performance of Q^2 -routers, some straight forward solutions would be to either make the gate faster by increasing the ZZ interaction between the input and switch or improve our qubits' lifetimes. Also promising would be the implementation of an erasure scheme

such as that used in dual-rail encodings. Specifically, in our protocol, states where any output qubit ends in $|e\rangle$ due to photon loss are invalid. Detecting and removing shots where this occurs could significantly improve the fidelity of the routing protocol.

In this work, we opted for the simplest architecture to demonstrate the robustness of our procedure even when using only fixed-frequency elements. However, a compelling path for improvement is to enable control over ZZ interactions by introducing tunable couplers, particularly between the switch and input. Reducing always-on stray ZZ interactions would improve single-qubit gate robustness and state preparation. It would also generalize our c -iSWAP gate to operate in cases with multiple total excitations in the input and outputs. Given our measured fidelities, with additional ZZ control we expect to be able to use our router as a building block for a few-memory layer QRAM (Appendix I) or for a proof-of-concept quantum network.

ACKNOWLEDGMENTS

C.M. and S.L. acknowledge support from the Air Force Office of Scientific Research (AFOSR) Multidisciplinary Research Program of the University Research Initiative (MURI) Grant No. W911NF2010177. C.M. acknowledges additional support from NSF GRFP Grant No. DGE-2146755. G.L. and L.J. acknowledge support from the ARO (W911NF-23-1-0077), ARO MURI (W911NF-21-1-0325), AFOSR MURI (FA9550-19-1-0399, FA9550-21-1-0209, FA9550-23-1-0338), DARPA (HR0011-24-9-0359, HR0011-24-9-0361), NSF (OMA-1936118, ERC-1941583, OMA-2137642, OSI-2326767, CCF-2312755, OSI-2426975), NTT Research, Packard Foundation (2020-71479), and the Marshall and Arlene Bennett Family Research Program. The device used in this work was fabricated and packaged by the Superconducting Qubits at Lincoln Laboratory (SQUILL) Foundry at MIT Lincoln Laboratory, with funding from the Laboratory for Physical Sciences (LPS) Qubit Collaboratory. We also thank Lincoln Labs for providing a Josephson traveling-wave parametric amplifier. We gratefully thank the QICK team, especially Sho Uemura, for providing us with custom firmware and technical support. We acknowledge Aaron Trowbridge for useful discussions on optimal control theory, Arnold Mong for insights on efficient MLE for tomography with ZZ correction, and all the members of the Schuster Lab for stimulating discussions and technical support.

DATA AVAILABILITY

The data that support the findings of this article are openly available [51].

APPENDIX A: EXPERIMENTAL SETUP

The room temperature and cryogenic wiring setups are shown in Fig. 5. The sample is packaged for us by the

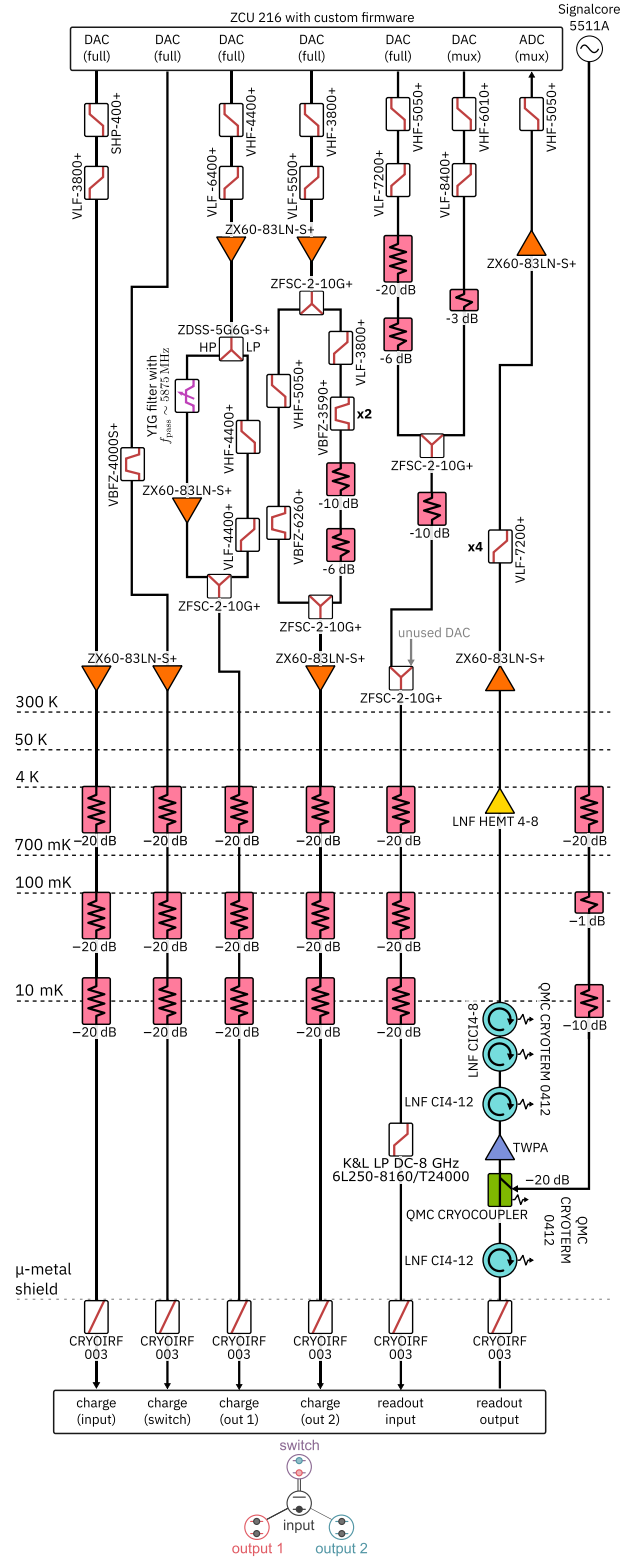


FIG. 5. Room temperature and cryogenic wiring.

MIT SQUILL foundry in a gold plated OFHC copper housing [52] and heat sunk via OFHC copper to the mixing chamber (MXC) plate of a BlueFors LD400 dilution refrigerator, which is at approximately 10 mK. The sample is entirely enclosed by an OFHC can that is coated on the inside with Berkeley black [53] to provide IR filtering then with two layers of μ -metal cans from MuShield to provide magnetic shielding. All input and output lines to the sample are filtered with CRYOIRF-003 IR filters placed inside the innermost can, and the readout input line is additionally filtered with a low pass filter from K&L with a cut off around 8 GHz. To perform all of our qubit control and readout, we use a Xilinx Zynq UltraScale+ RFSoc ZCU216 with custom firmware made for us by the QICK collaboration [54,55]. For qubit and sideband drives, we use four DAC generators (“full” type in QICK terminology) with sampling frequency $f_s = 6389.760$ MHz. We split and then recombine the output of the DACs for the two output qubits to enable separate filtering and amplification for the qubit and sideband drives. We use a YIG filter on the output 1 sideband drive to reduce populating the input qubit resonator, which is about 100 MHz away from the sideband frequency. All four DACs for qubit control are coherently phase reset at the start of each experiment iteration.

On the readout side, we use an additional full DAC to drive the resonator tone on the input qubit and a multiplexed (“mux” type) DAC to drive the resonator tones on the switch, output 1, and output 2. Both of these generators have $f_s = 6881.280$ MHz. To perform readout of all four resonator tones, we use a multiplexed ADC with $f_s = 2457.600$ MHz. Because we use a mixer-free setup, all readout tones are received in higher Nyquist zones of the mux ADC and are mapped back into the first Nyquist zone to perform the down-conversion. We use a TWPA fabricated by Lincoln Labs that provides about 20–30 dB of gain across the bandwidth of our resonator tones and we heat sink it via an OFHC copper finger to the MXC. We ensure that the environment surrounding the TWPA is impedance matched at both the readout frequencies (6–8 GHz) and idler frequencies (9–11 GHz) by using 4–12 GHz circulators on either side of the TWPA. Two more 4–8 GHz circulators provide additional isolation from the HEMT, which operates at 4–8 GHz and provides about 35 dB of gain. We use a SignalCore 5511A to provide a continuous pump tone for the TWPA. At room temperature, the readout signal is amplified by a ZX60-83LN-S+ amplifier from Minicircuits that is placed close to the fridge output flange. We use four VLF-7200+ low pass filters to attenuate the TWPA pump tone by approximately 20 dB before amplifying again with another ZX60-83LN-S+. An additional VHF-5050+ high pass filter reduces noise from the lower Nyquist zones before the final readout acquisition.

APPENDIX B: SYSTEM PARAMETERS AND SIMULATION

In this section, we present the system model and the measured parameters. The system is composed of four capacitively coupled transmon qubits. Its Hamiltonian can be written as

$$\begin{aligned} H &= H_q + H_c \\ H_q &= \sum_k \omega_k a_k^\dagger a_k + \frac{\alpha_k}{2} a_k^{\dagger 2} a_k^2 \\ H_c &= \sum_{k,l>k} g_{kl} (a_k^\dagger + a_k) (a_l^\dagger + a_l) \end{aligned} \quad (\text{B1})$$

where we consider an all-to-all coupling to account for parasitic couplings. ω_k and α_k are the qubits’ uncoupled frequencies and anharmonicities respectively and g_{kl} is the coupling between qubits k and l .

From this Hamiltonian we can perform both dispersive coupling and rotating wave approximations [56] to get

$$\begin{aligned} \tilde{H}_q &= \sum_k \tilde{\omega}_k \tilde{a}_k^\dagger \tilde{a}_k + \frac{\tilde{\alpha}_k}{2} \tilde{a}_k^{\dagger 2} \tilde{a}_k^2 \\ \tilde{H}_c &= \sum_{k,l>k} \zeta_{kl} \tilde{a}_k^\dagger \tilde{a}_k \tilde{a}_l^\dagger \tilde{a}_l \end{aligned} \quad (\text{B2})$$

where $\tilde{\omega}_k$ and $\tilde{\alpha}_k$ are respectively the dressed frequency and anharmonicity of qubit k and ζ_{kl} is the ZZ interaction between qubits k and l .

To estimate the dressed frequencies we use a Ramsey sequence for all individual qubits on the $|g\rangle - |e\rangle$ subspace. We estimate the dressed anharmonicity using the same sequence on the $|e\rangle - |f\rangle$ subspace and obtain $\tilde{\alpha} = \omega_{ef} - \omega_{ge}$. To estimate the ZZ interaction ζ_{kl} we apply a π pulse on the qubit k before performing a Ramsey sequence on qubit l . Finally, to estimate the bare parameters from the dressed ones, we perform a numerical diagonalization of Eq. (B1) where the bare parameters are optimized to minimize the difference between the numerically

TABLE I. **Qubit parameters.** ω_{res} is the readout resonator frequency, $\tilde{\omega}_{ge}$ is the qubit $g - e$ dressed transition frequency, $\tilde{\alpha}$ is the qubit dressed anharmonicity, $T_1^{ge/ef}$ is the relaxation time of the $g - e/e - f$ subspace, and $T_{2,E}^{ge/ef}$ is the T_2 echo coherence time of the $g - e/e - f$ subspace. Error bars are the standard errors from the fitted values.

Parameters	Switch	Input	Output 1	Output 2
$\omega_{\text{res}}/2\pi$ (GHz)	6.810	5.796	7.702	6.971
$\tilde{\omega}_{ge}/2\pi$ (GHz)	4.109	3.448	4.761	4.379
$\tilde{\alpha}/2\pi$ (GHz)	−0.226	−0.100	−0.188	−0.173
T_1^{ge} (μ s)	53 ± 5	92 ± 17	61 ± 3	59 ± 5
T_1^{ef} (μ s)	...	...	18 ± 2	51 ± 12
$T_{2,E}^{ge}$ (μ s)	58 ± 4	52 ± 5	54 ± 4	81 ± 6
$T_{2,E}^{ef}$ (μ s)	...	...	14 ± 2	35 ± 6

TABLE II. ZZ shifts $\zeta_{kl}/2\pi$ and corresponding couplings $g_{kl}/2\pi$ between pairs of qubits (MHz).

ZZ/coupling	Switch	Input	Output 1	Output 2
Switch	...	-5.39/52.4	-0.14/4.86	-0.56/2.67
Input	-5.39/52.4	...	-1.11/55.1	-1.31/43.6
Output 1	-0.14/4.86	-1.11/55.1	...	-0.58/6.76
Output 2	-0.56/2.67	-1.31/43.6	-0.58/6.76	...

computed and measured dressed parameters (reported in Tables I and II).

APPENDIX C: C-ISWAP GATE CALIBRATION

To implement the c-ISWAP gates between input and each of the two outputs, we use flat-topped pulses with Gaussian ramps to calibrate $\sqrt{i}$ SWAP gates. Two $\sqrt{i}$ SWAPs are tiled to build an i SWAP. There are four possible parameters for each $\sqrt{i}$ SWAP: amplitude A , frequency f , pulse length τ , and ramp time τ_r , all of which are interdependent. We begin by fixing τ_r to be 3 ns each, taking $2\sqrt{2}$ standard deviations for each ramp. We verify using simulations that this ramp time provides a good balance between speed and avoiding unwanted excitations. Importantly, when the ac-Stark shift effect is large, having a ramp time that is a sizable fraction of the total pulse length can cause large shifts in optimal frequency when changing the total length [46]. Next, to fix the (A, f, τ) triplet, we must satisfy two optimization goals. First, for a given τ , we want to find the unique (A, f) tuple that implements a high fidelity X_π^{sub} gate where X_π^{sub} refers to a π -pulse (i SWAP) in the $|eg\rangle - |gf\rangle$ subspace. Second, we want to pick the shortest τ that will not cause leakage between the two branches of c-ISWAPs. To satisfy these two goals, we follow a procedure in which we (1) use rough calibrations to obtain an initial (A, f, τ) at a range of τ s, (2) check each (A, f, τ) for leakage across branches of c-ISWAPs to pick an optimal τ , and (3) perform a final fine calibration of (A, f) at the selected τ .

- (1) **Rough calibration of A vs f relationship.** We would like to find a series of guess tuples of A_0 vs f_0 for a range of $\{\tau_0\}$ that we will sweep. To find the guess tuples (A_0, f_0) , we perform a Rabi chevron experiment sweeping A vs f at a fixed τ , find A_0 that maximizes the swap amplitude at each given f_0 , and fit the resulting points to a quadratic function. This experiment can be repeated at a few different τ s to ensure the fit is good across the full range of lengths [Figs. 6(a) and 6(b)]. We can then use this quadratic fit to calculate the expected frequency for any given amplitude, matching these (A_0, f_0) tuples to some τ_0 within the range of tested lengths.
- (2) **Looping over tuples: rough calibration of τ .** For each guess tuple (A_0, f_0, τ_0) where τ_0 is allowed to be off by a factor of 2 to 3, we know that the actual

best length τ_1 will fall within a rough range based on the input lengths for the amplitude vs frequency chevron above. Thus, for every (A_0, f_0) tuple, we perform a Rabi sweep across a reasonable range of lengths and fit the oscillation to a sinusoid to get an updated length parameter τ_1 . At this point, we fix the length parameter to τ_1 for all further calibrations.

- (3) **Looping over tuples: first refinement of f .** For each guess tuple (A_0, f_0, τ_1) , we perform a $(X_\pi^{\text{sub}}, X_{-\pi}^{\text{sub}})^N$ (“ π -minus- π ”) experiment, which finds the ac-Stark-shifted resonant frequency while being mostly insensitive to amplitude error. In this experiment, we sweep the pulse frequency and an integer number of cycles N [Fig. 7(a)]. As we approach the optimal frequency for the given length and amplitude, we observe slower and slower oscillations in N . This N vs frequency sweep can be fitted using a Gaussian [Fig. 7(c)] for the optimal frequency f_1 by multiplying $\prod_N (1 - P_{|1\rangle}(f))$ for the output qubit and $\prod_N (P_{|1\rangle}(f))$ for the input qubit, as in Appendix E3 of Ref. [57].
- (4) **Looping over tuples: first refinement of A .** For each guess tuple (A_0, f_1, τ_1) , we perform a $(X_\pi^{\text{sub}}, X_\pi^{\text{sub}})^{2N+1}$ [“ π -train”, Fig. 7(b)] experiment, which gives a result that is similar qualitatively to the previous π -minus- π experiment but sweeps the gain instead of the frequency. In the π -train experiment we choose to perform an odd number of X_π^{sub} so that a full swap has contrast with the case where no swap occurs at all. The optimal gain A_1 is fit to a Gaussian, multiplying $\prod_N (P_{|1\rangle}(f))$ for the output qubit and $\prod_N (1 - P_{|1\rangle}(f))$ for the input qubit [Fig. 7(d)]. At this point, we have a set of (A_1, f_1) tuples for each length τ_1 that are each within 1–2% of amplitude/a few 100 kHz from the best final tuple.
- (5) **Optimization for leakage.** We would now like to pick an optimal length τ^* from our set of calibrated tuples (A_0, f_1, τ_1) that corresponds to the fastest pulse that enables minimal transfer of population to the wrong output branch. As discussed in Sec. III A of the main text, to calibrate the swap between input and output 1 (2) conditioned on switch in $|g\rangle$ ($|e\rangle$), we prepare the switch in the wrong state $|e\rangle$ ($|g\rangle$) and observe the population transferred to output 1 (2). We pick the optimal length τ^* to be at the minima with the shortest gate length where the

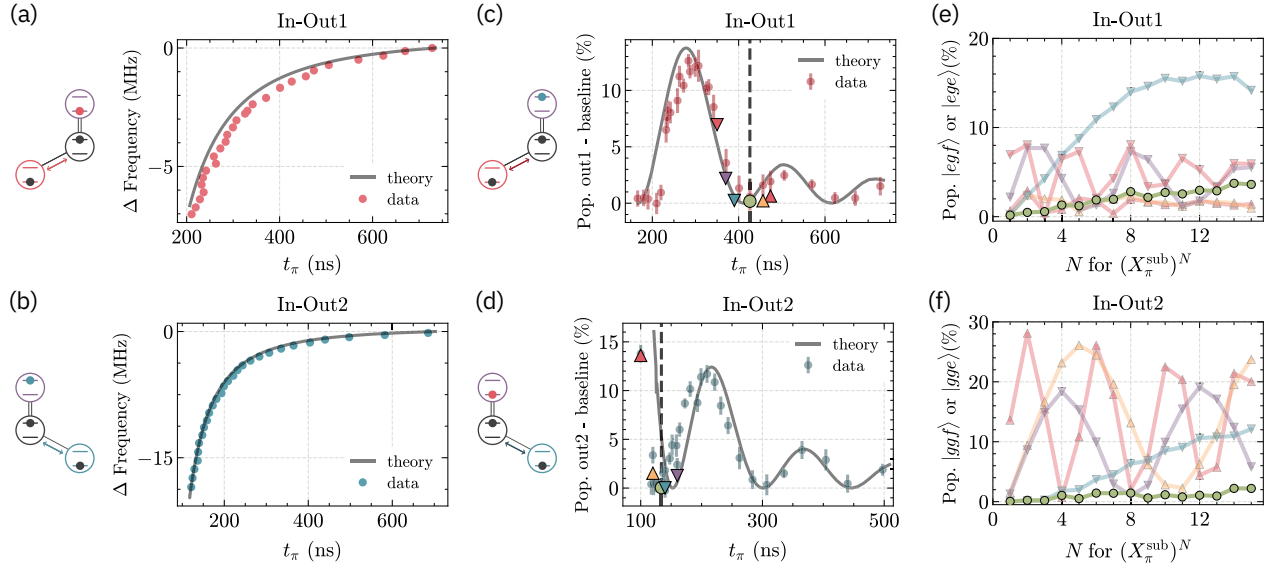


FIG. 6. **C-*i*SWAP length calibration.** (a),(b) To select the pulse lengths for each *i*SWAP, we calibrate a series of drive amplitude and frequency tuples corresponding to a selection of different pulse lengths τ for a $\sqrt{i}$ SWAP, preparing the switch in the state to allow the SWAP to proceed. We find a relationship between $t_\pi \equiv 2\tau$ and the calibrated frequency that matches well with the expected ac-Stark shift behavior from simulation of the system Hamiltonian without any fitting parameters. (c),(d) To select a pulse length that minimizes leakage across the two branches of the protocol, for each tuple of calibrated drive amplitude, frequency, and length, we prepare the switch in the forbidden state and measure how much population is still transferred to the target output. The observed oscillatory pattern in the transferred population vs t_π matches well with the simulation without fitting parameters (after subtracting off a baseline offset that we attribute to SPAM). We select a length τ^* for each *i*SWAP that sits at a minimum of this population transfer (black dashed line) before proceeding to the next layer of finer calibrations (Appendix C and Fig. 7). (e),(f) To quantify the amount of population that is erroneously transferred at τ^* when the switch is in the blocked state, we perform error amplification as in Figs. 3(g) and 3(f) for a few lengths near the minima [selected pulse lengths with their population at $N = 1$ highlighted with the corresponding colored marker on (c),(d)]. The green circles in (e),(f) correspond to the populations plotted in Figs. 3(g) and 3(f). We include both the $|\tilde{S}gf\rangle$ and $|\tilde{S}ge\rangle$ states in the population to ensure we capture all the undesired output population, including that which has decayed from $|f\rangle$ to $|e\rangle$ during the error amplification. The Rabi oscillations corresponding to the calibrated tuple of frequency, length, and amplitude are shown in Fig. 2.

experiments are well-behaved [Figs. 6(c) and 6(d)]. To verify that we have picked an optimal length and upper bound the undesired population transfer, we amplify the measurable population transfer error as in Figs. 3(g) and 3(h) for a few different lengths near τ^* (combined with corresponding calibrated amplitude and frequency) by applying $X_\pi^{\text{sub}} N$ times with the switch initialized in the wrong state and observe oscillations at lengths away from τ^* [Figs. 6(e) and 6(f)].

- (6) **Fine calibration of A, f .** Once τ^* is picked, we perform a fine calibration on the amplitude and frequency at τ^* . In our previous calibration, we have assumed that the amplitude obtained from the π -train experiment and the frequency obtained from the π -minus- π experiment are independent. However, this is not the case due to the ac-Stark shift induced by the drive [46]. We want to make sure that not only is the amplitude correct given the length and frequency, but also the frequency is correct given the length and amplitude. Thus, we perform

a fine calibration fixing the length to be τ^* where we first repeat the π -minus- π experiment at different amplitudes [Fig. 7(e)] and then repeat the π -train experiment at different frequencies [Fig. 7(f)]. We thus obtain two two-dimensional (2D) maps, each of which sweeps amplitude vs frequency. The final optimal tuple of (τ^*, A^*, f^*) is found where the optimal regions in the two 2D maps intersect. We find this tuple by finding the center of the blob that appears when multiplying the two maps together [Fig. 7(g)].

APPENDIX D: SINGLE QUBIT GATE CALIBRATION

Due to the strong, always-on, ZZ interaction between the qubits, implementing single qubit gates can be challenging. To circumvent this we use the commercial optimal control software Q-CTRL to find pulses that are less sensitive to frequency shifts. To do so, we use the dispersive Hamiltonian from Eq. (B2) in the frame rotating at $\tilde{\omega}_k$ for each

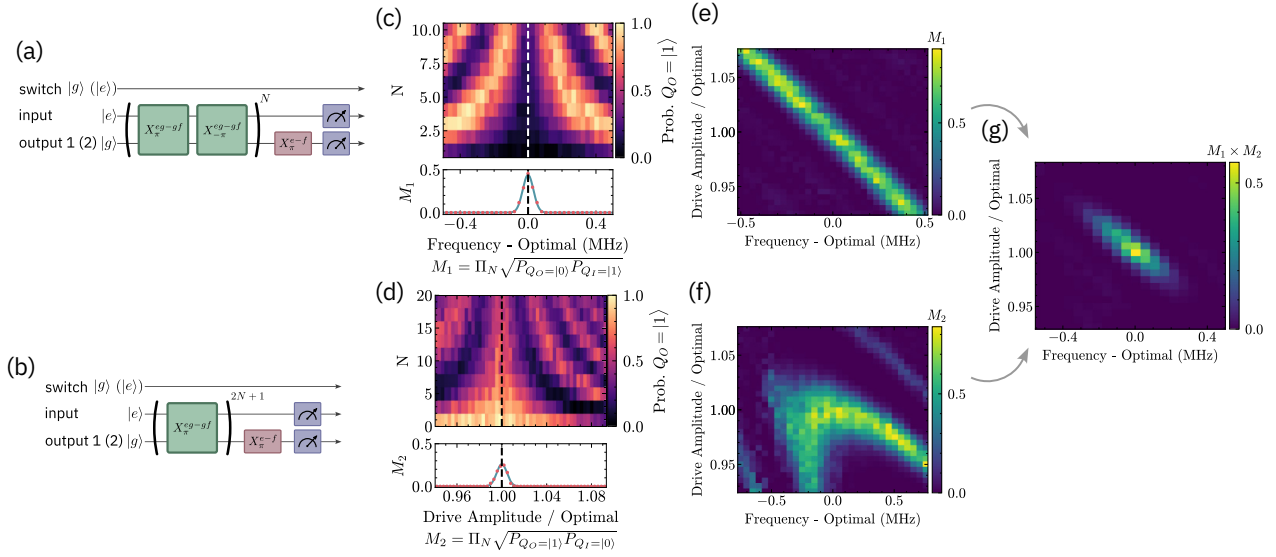


FIG. 7. **C-ISWAP fine calibration procedure.** (a),(b) Pulse sequences for (c),(e) and (d),(f) respectively. (a),(c),(e) represent a π -minus- π experiment, which amplifies frequency error while being robust against amplitude error. (b),(d),(f) represent a π -train experiment, which amplifies amplitude error. All data shown is scaled from 0 to 1. (c),(d) As the number of error amplifications N increases, we observe faster and faster oscillations in the frequency (amplitude) that is probed by the π -minus- π (π -train) experiments. By multiplying the population at each frequency (amplitude) over all N , we obtain an easy-to-fit Gaussian that is peaked at the optimal frequency (amplitude). As the number and magnitude of N 's sampled increases, the Gaussian peak narrows and the background reduces, allowing for continuous refinement of the optimal point. (e),(f) We perform the π -minus- π (π -train) experiments at a variety of different amplitudes (frequencies), always fixing the pulse length, so we that we obtain an amplitude vs frequency map for both the π -minus- π and π -train experiments. For each amplitude and frequency point, we perform the same multiplication over N as in (c),(d) and an additional multiplication over the two qubits measured (input and relevant output). For the π -minus- π experiment we obtain approximately straight lines following the Stark shift of the pulse. For the π -train experiment, we obtain quadratic curves that are symmetric about the Stark shift of the pulse. (g) To find the optimal amplitude, frequency tuple for the length we have chosen, we multiply together the results of (e),(f) to get their intersection, resulting in a blob whose center is the optimal tuple. This tuple is a “stable” point, where alternating updating the frequency from a π -minus- π experiment and updating amplitude from a π -train experiment will converge to the same amplitude and frequency.

individual qubit. Then for each qubit k we search for the best possible $X_{\pi/2}$ unitary given the control operator

$$H_{\text{drive},k} = (I_k(t) + iQ_k(t))\tilde{a}_k + \text{h.c.} \quad (\text{D1})$$

where the time-dependent coefficients I_k and Q_k are optimized by the solver for a fixed total pulse length that we input. We then vary the pulse length and pick the one that minimizes the infidelity. For the switch, input, output 1, and output 2, we find $X_{\pi/2}$ pulses of lengths 9, 20, 13, and 13 ns respectively. Once the optimal pulse shapes are found we then calibrate them experimentally using the following process:

- (1) **Ramsey experiment.** To find the pulse frequency we first perform a Ramsey experiment to estimate the dressed qubit frequency.
- (2) **Rabi oscillation vs amplitude.** Using the found qubit frequency and computed pulse shape, we measure the Rabi oscillations vs pulse strength, fit to a sinusoid, and take the drive strength at the maximum of the Rabi oscillation.

- (3) **Calibration of the virtual Z gate.** The pulses are designed to be optimal in the qubit frequency frame. However, we usually find experimentally that the rotation axis of the implemented pulse has a small tilt toward the Z-axis. We attribute this tilt to an ac-Stark shift that is not perfectly captured by the simulated evolution. To correct for this phase accumulation, we add a virtual Z phase at the end of each $X/2$ rotation. To calibrate the virtual Z rotation we perform a $(X_{\pi/2}, Z_\phi, X_{-\pi/2}, Z_\phi)^N$ sequence where the Z_ϕ gate is implemented virtually by updating the carrier phase of the next physical pulse [58]. We then measure the qubit population vs both N and ϕ . This experiment is similar to the π -minus- π experiment in Appendix C [Fig. 7(c)], updating the phase instead of sweeping the pulse frequency, and we can similarly fit the product over N to a Gaussian to find the optimal phase correction.
- (4) **Error amplification vs pulse amplitude.** Once the pulse frequency and virtual Z gate angle are known we perform a $(X_{\pi/2}, X_{\pi/2})^{2N}$ experiment and fit the product over N to a Gaussian to accurately estimate

the optimal pulse amplitude, similar to the π -train experiment in Appendix C [Fig. 7(d)].

We repeat these steps to get a unique frequency, drive amplitude, and virtual Z for each ZZ-shifted pulse. Doing so allows us to prepare higher fidelity product states when more than one qubit is excited.

All of the $e-f$ transitions are implemented with simple Gaussians where we take a total of $4\sqrt{2}$ standard deviations per $\pi/2$ pulse. Only the output qubits have $e-f$ transitions that are relevant to the protocol performance; these gates are 60 ns long each. We perform the calibrations for these pulses in a similar manner as for the optimal control-based $g-e$ pulses, except that we do not need calibrate the virtual Z corrections. We also ensure that the $e-f$ pulses are calibrated for the correct ZZ-shifted frequency whenever possible. Note that within the protocol, the $e-f$ pulse on output 1 will only ever expect to see the switch in $|g\rangle$, and the $e-f$ pulse on output 2 will only ever expect to see the switch in $|e\rangle$, so they do not need to be robust against frequency shifts.

Finally, we implement all $Y_{\pi/2}$ rotations by applying the calibrated $X_{\pi/2}$ pulse with the carrier phase shifted by $\pi/2$. The X_{π} and Y_{π} are constructed by repeating the respective $\pi/2$ pulse.

APPENDIX E: SUBSPACE RANDOMIZED BENCHMARKING OF C-ISWAP GATES

In this section we present the methods used to extract the average error per swap as well as the leakage out of the logical subspace of the c-ISWAPS.

1. Randomized benchmarking experiment

We perform two rounds of benchmarking, a reference standard randomized benchmarking experiment and an interleaved benchmarking experiment. We treat the states $|Seg\rangle - |Sgf\rangle$ as the two level logical subspace of a single dual-rail qubit, where $|S\rangle$ is the switch state that allows the SWAP to proceed. For each round of benchmarking, we initialize the system in $|Seg\rangle$ then apply a randomly selected sequence of gates from the 24 unique gates in the Clifford gate set. Each Clifford gate is constructed by tiling 0 (identity), 1, 2, or 3 $X_{\pi/2}^{\text{sub}}$ gates with phase offsets to implement rotations about different axes. At the end of each sequence, we append an extra inversion Clifford. To implement the inversion Clifford, we first calculate the product of all previous gates. If this product gate brings the $+Z$ -axis back to $+Z$, we simply take the inversion gate as the identity gate, since we only need to measure the counts along the Z -axis. Otherwise, we apply the calculated product gate with a negated phase, so that the ideal final state is $|Seg\rangle$. For each sequence depth we generate 20 variations to obtain statistics.

TABLE III. Readout fidelities (%) calculated as the number of correctly categorized shots over the total number of shots. We include errors for 95% confidence calculated from a typical window of a few-hour measurement.

Switch	Input	Output 1	Output 2
84.3 ± 0.5	76.1 ± 1.0	75.6 ± 0.4	80.4 ± 0.2

For each gate sequence, we would like to monitor the population of 12 total states, consisting of states within the subspace ($|Seg\rangle, |Sgf\rangle$) and leakage states outside of the subspace ($|\bar{S}eg\rangle, |\bar{S}gf\rangle, |*gg\rangle, |*ge\rangle, |*ee\rangle, |*ef\rangle$). Here, $|\bar{S}\rangle$ is the switch state that forbids the SWAP and $|*\rangle$ means both the $|g\rangle$ and $|e\rangle$ states on the switch. To distinguish these states, we perform two experiments with different readout parameters for every unique gate sequence that we construct. In the first experiment, we measure the switch, input, and target output qubits using readout parameters optimized to distinguish the $|g\rangle$ and $|e\rangle$ states of each qubit. In the second experiment, we re-run the gate sequence and again measure the switch, input, and target output but change the readout parameters for just the output to be optimized to distinguish between $|e\rangle$ and $|f\rangle$. In both cases, we bin each shot into $|g\rangle$ vs $|e/f\rangle$ or $|g/e\rangle$ vs $|f\rangle$. We require this two-step readout because we have a poor readout fidelity if we attempt to use a readout point that can simultaneously distinguish $|g\rangle$, $|e\rangle$, and $|f\rangle$. Our readout fidelities are reported in Table III.

To correct for readout error, we follow a modified version of the standard confusion matrix correction (i.e., in Appendix G 1 a). In particular, the number of bins for the measured counts is not equal to the number of possible final bins that we want to correct the counts into. To obtain the confusion matrix, we prepare each of the 12 possible states that we want to bin our final counts into and measure them twice using the two-step readout. From this procedure we obtain a confusion matrix M of size 12 rows $\times$ (8 + 8) columns, where the first and second set of eight columns correspond to the results of the first and second readout experiments respectively. Despite the confusion matrix being a rectangular matrix, we can still apply the confusion matrix to each measured set of 8 + 8 counts and use the methods described in Appendix G 1 c to find the most likely set of true measured counts binned into the 12 target final states.

2. Error extraction

To characterize the fidelity of a single SWAP, we consider the behavior of the probability $p_{|Seg\rangle}$ of returning to the target state, $|Seg\rangle$, after the sequence of gates. This probability should follow a double exponential as a function of the number of Clifford gates N [59], which is composed of the leakage out of the subspace measured by the parameter λ_1 , the logical error within the subspace measured by

the parameter λ_2 , and the seepage back into the system measured by the parameter A_0 :

$$p_{|Seg\rangle} = A_0 + B_0\lambda_1^N + C_0\lambda_2^N. \quad (E1)$$

In our case, the dominating errors are due to leakage triggered by a photon loss in one of the underlying qubits. Hence, we have $\lambda_1 \ll \lambda_2 < 1$ and this double exponential is difficult to fit directly. Instead, we restrict ourselves to the return probability $p_{|Seg\rangle}$ post-selected on the probability to stay within the logical subspace $|Seg\rangle - |Sgf\rangle$, given by $p_{\text{survival}} = p_{|Seg\rangle}/p_{\text{sub}}$. The probability to stay within the logical subspace follows a single exponential behavior:

$$p_{\text{sub}} = A + B\lambda_1^N, \quad (E2)$$

Since seepage is negligible p_{sub} and $p_{|Seg\rangle}$ will go to zero at large N and hence we have $A \sim A_0 \sim 0$. Then the post-selected survival probability is a single exponential that is easier to fit:

$$p_{\text{survival}} = C + D \left(\frac{\lambda_2}{\lambda_1} \right)^N; \quad (E3)$$

for a more rigorous derivation see Ref. [60]. Thus, for each benchmarking round, we first fit Eq. (E2) to obtain the leakage parameters A and λ_1 . We can estimate the average photon loss (or equivalently leakage) per Clifford or interleaved Clifford, respectively L_{ref} and L_{int} , as [59]:

$$L = (1 - A)(1 - \lambda_1). \quad (E4)$$

Second, we fit Eq. (E3) to obtain λ_2 since λ_1 is known from the previous fit. Using these two fits, we can estimate the average errors ϵ_{ref} or ϵ_{int} as in [59]

$$\epsilon = 1 - \frac{1}{d}[(d-1)\lambda_2 + 1 - L], \quad (E5)$$

where d is the subspace dimension ($d = 2$ in our case). Finally, using the error and photon loss rates from both the reference and interleaved benchmarking, we can estimate the photon loss and error per X_π swap as [61]

$$\begin{aligned} \epsilon_X &= 1 - \frac{1 - \epsilon_{\text{int}}}{1 - \epsilon_{\text{ref}}} \\ L_X &= 1 - \frac{1 - L_{\text{int}}}{1 - L_{\text{ref}}}. \end{aligned} \quad (E6)$$

We note that though we have benchmarked our system like a dual-rail qubit, unlike in a dual-rail qubit we do not treat leakage errors as erasures since they are not detected, so this error is included in the total errors.

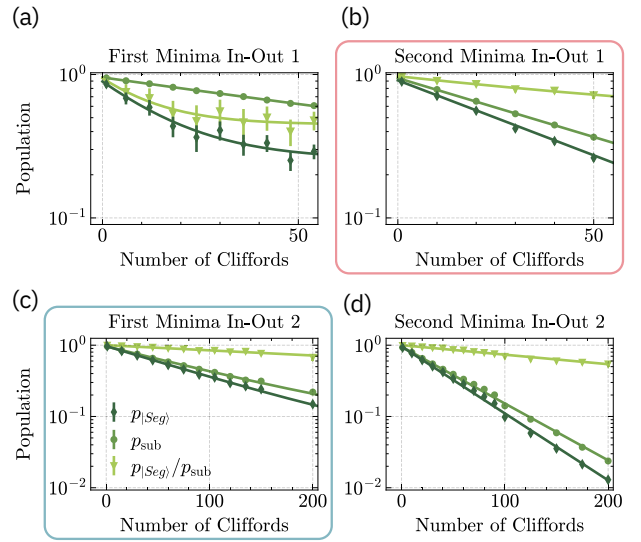


FIG. 8. (a),(b) RB for the i SWAP between input and output 1 for $n = 1$ and $n = 2$. (c),(d) RB for the i SWAP between input and output 2 for $n = 1$ and $n = 2$. The RB plots corresponding to the selected working points for input to output 1 (2) are boxed in red (blue) and are the same as those shown in Figs. 3(c) and 3(d).

3. Randomized benchmarking for alternative working points

In Sec. III A and Appendix C we saw that several pulse amplitudes can perform a good c- i SWAP, as long as the off-resonant drive rate respects $\Omega = 2n g_{eg-gf}$. These points correspond to a minimum of population transfer to the target output qubit when the switch is in the blocking state. If the gates are decoherence-limited, we expect the $n = 1$ point to have the best fidelity. Hence, we perform randomized benchmarking on both the $n = 1$ and $n = 2$ points for the i SWAP between output 1 and 2 (see Fig. 8) to pick the best parameter set.

For the i SWAP between input and output 1, the photon loss rate is $0.88 \pm 0.22\%$ and $1.88 \pm 0.16\%$ for respectively $n = 1$ and $n = 2$. However, the total errors are respectively $3.86 \pm 0.48\%$ and $3.05 \pm 0.11\%$. We hypothesize that the higher error rate for the shorter gate is due to a drop in the $|eg\rangle - |gf\rangle$ subspace coherence at larger drive strengths, as has been observed in several experiments [46,62]. As a consequence, the fastest gate has a lower overall fidelity despite a better photon loss rate, so we use $n = 2$ in the protocol.

On the other hand, for the i SWAP between input and output 2 we have a photon loss rate of $0.83 \pm 0.01\%$ and $1.82 \pm 0.02\%$ for respectively $n = 1$ and $n = 2$. In addition, the total error rates are respectively $1.36 \pm 0.01\%$ and $2.87 \pm 0.02\%$. For this gate, both the leakage and error rates improve for the shorter gate, indicating that this gate remains limited by photon loss error even at the larger drive strength. Thus, for the protocol we use the $n = 1$ working point.

APPENDIX F: ANALYSIS OF PULSE TIMING AND THE SIMULTANEOUS PROTOCOL

Because our protocol is limited by coherence, there is a possibility that performing the two c - i SWAPs simultaneously as opposed to sequentially could yield an improved fidelity by reducing the quantum circuit gate depth to 1. As shown in Figs. 2 and 6, because we are limited by our ZZ interaction strength, we achieve the fastest swaps possible by picking pulse parameters such that the allowed swap is swapped while the forbidden swap experiences an integer multiple of a 2π rotation. If we were to adjust the protocol to have simultaneous pulses, this behavior might still be achievable as long as there remains population in the input qubit during the entire course of each forbidden swap's $2\pi n$ rotation such that the rotation can be completed. Unfortunately, due to this condition, the simultaneous protocol cannot in fact yield a higher fidelity.

In our architecture, the swap between input and output 1 is slower than the swap between input and output 2. This means that it is no longer possible to satisfy our $2\pi n$ rotation condition. We can examine the resulting behavior in more detail by breaking down each of the swap cases separately. For simplicity, we initialize the system in either $|eggg\rangle$ ($|A\rangle$) or $|eegg\rangle$ ($|B\rangle$). When initializing in $|A\rangle$, the output 1 swap should bring us to $|ggfg\rangle$ ($|A'\rangle$), and the output 2 swap (if not properly blocked) would take us to $|gggf\rangle$ ($|C\rangle$). Similarly, when initializing in $|B\rangle$, the output 2 swap should bring us to $|eggf\rangle$ ($|B'\rangle$), and the output 1 swap (if not properly blocked) would take us to $|egfg\rangle$ ($|D\rangle$). In the simultaneous protocol, for the fast swap between input-output 2, the forbidden transition ($|A\rangle$ to $|C\rangle$) will have enough time to do the desired $2\pi n$ rotation before the input qubit population reduces to zero over the course of the allowed transition $|A\rangle$ to $|A'\rangle$ [Fig. 9(a)]. On the other hand, for the slow swap between input-output 1, the $2\pi n$ rotation on the forbidden transition ($|B\rangle$ to $|D\rangle$) is slower than the simultaneous allowed fast swap between $|B\rangle$ to $|B'\rangle$. Thus, some of the population will be stuck in $|D\rangle$ after the input has no remaining population and it is no longer possible to complete the rotation on the forbidden transition [Fig. 9(b)]. For these reasons, a simultaneous protocol is expected to perform worse than a sequential protocol on our architecture.

We note that on our architecture, even if we were to pick the same n for both swaps, the swaps would still be different lengths. This is because in the formula for calculating the Rabi rate g_{eg-gf} based on the condition stated in the main text, $\Omega = \sqrt{g_{eg-gf}^2 + \zeta_{SI}^2} = 2ng_{eg-gf}$, we have neglected the ZZ shift induced by the switch on the corresponding output $|f\rangle$ state. Taking into account $|f\rangle$, we have $\zeta_{tot} = \zeta_{SI}^{ee} - \zeta_{SO_i}^{ef}$ where ζ_{SI}^{ee} is the previously defined ZZ interaction between the switch and input $|e\rangle$ states and $\zeta_{SO_i}^{ef}$ is the same interaction but between the switch $|e\rangle$ and

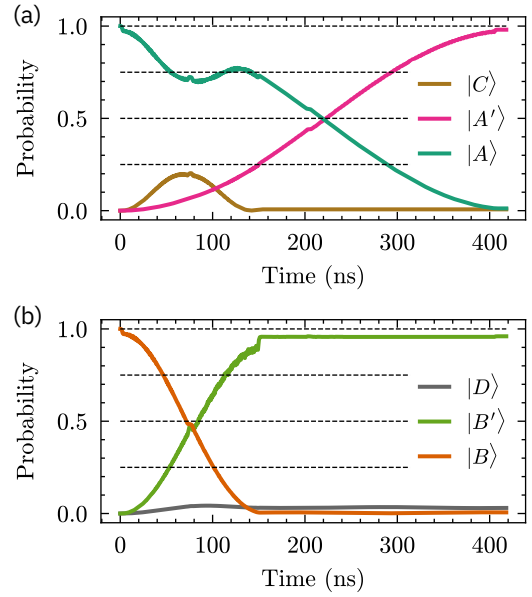


FIG. 9. **Simulated behavior of initial state populations under simultaneous protocol.** (a) Starting from $|eggg\rangle$ ($|A\rangle$), the forbidden transition driven by the input-output 2 swap has time to complete a $2\pi n$ rotation prior to the depletion of the input population by the allowed transition driven by the input-output 1 swap. (b) Starting from $|eegg\rangle$ ($|B\rangle$), the forbidden transition driven by the input-output 1 swap is too slow to complete a $2\pi n$ rotation prior to the depletion of the input population by the allowed transition driven by the input-output 2 swap. Thus, there is excess population leftover in an undesired state $|egfg\rangle$ ($|D\rangle$), reducing the fidelity compared to a sequential operation.

corresponding output $|f\rangle$ states. The adjusted Rabi rate on the forbidden transition is then $\Omega = \sqrt{g_{eg-gf}^2 + \zeta_{tot}^2}$, which ensures that it is not possible on our system for the two swaps to have the same length even if we were to pick the same n for each swap.

With a much larger ZZ interaction, we can expect we no longer need to satisfy the $2\pi n$ rotation condition on the forbidden swap, such that it would be possible to have a simultaneous protocol.

APPENDIX G: STATE TOMOGRAPHY

1. Tomography methods

In this work, we perform full quantum state tomography on either two or three qubits, which use the same procedure that can be adjusted for N qubits. The procedure can be broken down into five stages: (1) confusion matrix measurements, (2) Pauli basis target state measurements, (3) readout error correction, (4) guess density matrix construction from counts, (5) maximum likelihood estimation (MLE) to find the closest real density matrix from the unphysical guess.

a. Measurement of confusion matrix

To correct for readout error, we apply a standard confusion matrix correction. We define the measurement basis as the 2^N bins composed of all combinations of $|g\rangle$ and $|e\rangle$ on N qubits. All measurements in the tomography experiment are made via simultaneous readout tones on all N qubits. For each qubit, the raw demodulated I and Q values are rotated and binned via a fixed threshold into $|g\rangle$ or $|e\rangle$. The rotation angle and threshold are picked to maximize the fidelity between $|g\rangle$ and $|e\rangle$ for that qubit when all other qubits are in $|g\rangle$. We sort each shot into one of the 2^N bins by looking at thresholded states on all qubits. To obtain the $2^N \times 2^N$ confusion matrix M , we prepare each state in the measurement basis and sort shots into the measurement basis bins. The rows of M correspond to each prepared state and the columns correspond to the measured counts binned into each basis state.

b. Measurement of target state in Pauli basis

Assuming perfect measurements, standard state tomography would require $3^N - 1$ measurements [48]. We perform an overcomplete set of 3^N measurements corresponding to combinations of measurements along the X -, Y -, and Z -axes for each qubit. For each of these 3^N measurements, we measure all qubits simultaneously and bin individual qubit shots using the same angle and threshold as found during the calibration stage above, then bin collective shots into the measurement basis.

Since all readout tones measure qubits along the Z -axis, for the X (Y) measurement we perform a $Y_{-\pi/2}$ ($X_{\pi/2}$) “tomography pulse” to align the ± 1 eigenstates of X (Y) to the same eigenstate along Z . Note that in the presence of ZZ coupling, depending on the initial state and the measurement basis we are considering, the tomography pulses may be quite imperfect since the pulses must be agnostic to the initial state. This issue is addressed further in Sec. G 1 e.

c. Readout error correction

After completing all measurements, we move to the post-processing stages of the tomography. We start by correcting for readout error applying the measured confusion matrix on each Pauli basis measurement. Because a direct application of the inverse confusion matrix can sometimes result in negative counts, we use an optimization method to perform the correction as done in QISKIT [63]. We use the SLSQP optimization method implemented in the publicly available Python package `scipy` to minimize the cost function $f(\vec{p}) = \|\vec{M}'\vec{p} - \vec{p}_0\|$, where $\vec{p}_0$ is the column vector of measured counts for one Pauli basis scaled by the total number of counts to produce a probability, and $\vec{p}$ is the column vector of minimized probabilities. The minimization is performed with constraints for norm ($\|\vec{p}\| = 1$) and range (all $p_i \in [0, 1]$).

ALGORITHM 1. Efficient MLE for tomography (from Ref. [64]).

```

Define  $E = \{E_i\}$  as the set of eigenvalues of  $\rho_0$ , ordered
from largest to smallest
Define  $\{|E_i\rangle\}$  as the corresponding set of normalized eigen-
vectors of  $\rho_0$ 
Define  $d = 2^N$ 
Initialize  $E'$ , the eigenvalues for  $\rho_{\text{MLE}}$ , as  $E' = \{E'_i\} = 0 \forall i$ 
Initialize an accumulator  $a = 0$  and a stop counter  $i = d$ 
while  $E_{i-1} + a/i < 0$  do
     $a \leftarrow a + E_{i-1}$ 
     $i \leftarrow i - 1$ 
end while
for all  $j \in \{0, 1, \dots, i-1\}$  do
     $E'_j \leftarrow E_j + a/i$ 
end for
Initialize  $\rho_{\text{MLE}}$  in the shape of  $\rho_0$  with all zeros
for all  $j \in \{0, 1, \dots, i-1\}$  do
     $\rho_{\text{MLE}} \leftarrow \rho_{\text{MLE}} + E'_j |E_j\rangle\langle E_j|$ 
end for
return  $\rho_{\text{MLE}}$ 

```

d. Density matrix guess from corrected counts

Using only the readout error-corrected counts going forward, we construct an initial guess density matrix ρ_0 that will later be used as a starting point for the MLE. To do so we construct a tensor T where each $T_{i_1, i_2, \dots, i_N}$ is calculated by taking the expectation value of that Pauli matrix when the state is measured in the corresponding basis. Because all measurements are performed in the Z basis up to applying a rotation just before, T values corresponding to the Z and I basis use the same measurement results with an additional minus sign to correct for the difference in eigenvalues between Z and I . The guess ρ_0 is then formed by summing over tensor products over the four Pauli matrices weighted by the corresponding component of T :

$$\rho_0 = \frac{1}{2^N} \sum_{i_1, \dots, i_N \in \{X, Y, Z, I\}} T_{i_1, \dots, i_N} \sigma_{i_1} \otimes \dots \otimes \sigma_{i_N}. \quad (\text{G1})$$

e. Maximum likelihood estimation and ZZ correction

Perfectly matching the guess density matrix to the true density matrix would require an infinite number of measurements [64]. Thus, we need to use MLE to find the density matrix that is most likely to have produced the set of measurements that we have. We use an efficient MLE algorithm from [64], whose pseudocode we reproduce in Algorithm 1, to perform the MLE in on average less than 0.3 s for three qubit tomography. The algorithm requires that we have (1) an orthonormal Hermitian operator basis $\{\sigma_i\}_{i=1}^{d^2}$ where $d = 2^N$, (2) a set of values $\{m_i\}$ that are the average noisy measured expectation values of each σ_i , and (3) noise that is Gaussian.

Because of the presence of strong ZZ coupling between the qubits in the tomography, especially the two qubit

tomography on the switch and input, we modify the algorithm to correct for ZZ effects. As in Ref. [65], we do this by simulating the evolution of each measurement basis state under the tomography pulse waveforms. For each Pauli basis σ_i we thus build a corresponding evolution matrix. Instead of treating each of the measured counts as a measurement in the true σ_i basis, we treat them as counts in a modified basis $\{\tilde{\sigma}_i\}$ and perform the MLE relative to this basis.

To perform the simulation, we setup the Hamiltonian of the full qubit system according to Eq. (B2), using the measured dressed qubit frequencies (with all other qubits in $|g\rangle$) for $\tilde{\omega}_k$, measured dressed anharmonicities for $\tilde{\alpha}_k$, and measured average ZZ shift between qubits k and l ($\omega_{|11\rangle} - \omega_{|01\rangle} - (\omega_{|10\rangle} - \omega_{|00\rangle})$) for $\tilde{\zeta}_{kl}$. We then schedule the pulses corresponding to the Pauli basis we are considering, with the waveform amplitude and shape defined directly by the generated pulse shape from the optimal control. We note that it is important to match the timing of the simulated pulse with the timing of the true experimental pulse sequence (including idle times), as given our ZZ shift magnitudes, states can accumulate non-trivial phases due to ZZ effects with even a few ns of mismatch. We construct each initial state by looping over all the measurement basis states and defining the corresponding Fock basis state, then finding the eigenstate of the Hamiltonian that has the maximum overlap with the Fock basis state to build the dressed basis state. After sending each initial state through the pulse sequence, we take the partial trace over the result to reduce it to just the two lowest levels corresponding to the qubits that we are doing tomography on. Finally, from the set of all evolved basis states we define the evolution matrix R'_i corresponding to basis σ_i , with the evolved basis states along the columns.

Having obtained the simulated evolution matrices $\{R'_i\}$ that describe the performed rotations (as opposed to a perfect rotation $\{R_i\}$), we still need to “renormalize” the basis and find an updated set of counts that will be input into the MLE. We do this in two steps. First, we must determine the actual basis that the measured counts correspond to. Because we always perform measurements along the Z-axis, our experimental data consists of a set of n_0 counts marked as the +1 eigenstate along the Z-axis and n_1 counts marked as the -1 eigenstate along the Z-axis. We must find a map to some other deformed axis such that we can interpret the counts as a set of n_0 counts along the +1 eigenstate of this new axis and n_1 counts along the -1 eigenstate of the same axis. Second, because the rotations that were performed may not be orthogonal, we must orthonormalize the axes to satisfy the assumptions of the MLE algorithm.

We find that the true basis that corresponds to the measured counts is $\{\tilde{\sigma}_i\} = \{(R'_i)^\dagger \sigma_z R'_i\}$, which can be thought of geometrically as un-rotating the Z-axis by the simulated rotation. A diagram of this idea is shown in Fig. 10, where we have un-rotated the Z-axis by some R'_y , a deformed

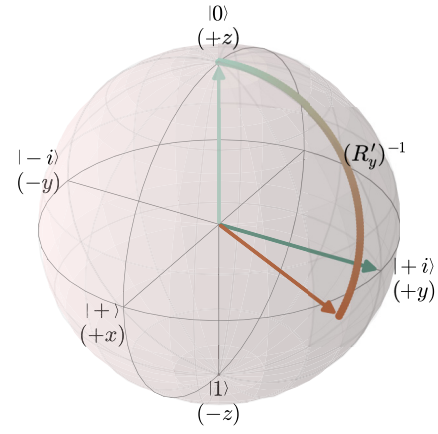


FIG. 10. During an ideal tomography pulse that corresponds to the target measurement basis σ_i , the Pauli basis should be rotated to the +Z-axis (i.e., dark green state should rotate to light green state; here we show a desired measurement in σ_y). Instead, under the influence of ZZ interactions, the rotation is deformed to R'_i (here, R'_y). In this case, the true axis along which the measured counts can be interpreted is represented by a rotation $(R'_y)^{-1}$ from the +Z-axis (orange state).

rotation about the X-axis that was originally intended as a tomography pulse for the Y-basis measurement.

Next, assuming the ZZ shifts are finite and $\{\tilde{\sigma}_i\}$ completely spans the Hilbert space (i.e., the performed rotations are good enough that no two rotations are exactly the same), we can use Gram-Schmidt decomposition on $\{\tilde{\sigma}_i\}$ and the measured counts to simultaneously orthonormalize the basis and obtain a corresponding set of counts. These renormalized counts can then finally be inputted to the MLE algorithm. For completeness, we provide the orthonormalization algorithm in pseudocode in Algorithm 2, where $\tilde{\sigma}_j$ is already initialized as $(R'_j)^\dagger \sigma_z R'_j$, and p_j is the T matrix element calculated by treating measurements as made in the Z basis if the target basis measured is X, Y, or Z; or in the I basis if the target basis is I.

Note that the final basis matrices $\{\tilde{\sigma}_i^\perp\}$ and counts $\{p_j^\perp\}$ still satisfy the assumption of [64]. The basis is orthonormal as required by the Gram Schmidt. Furthermore the basis remains Hermitian and the noise on the measurement data remains Gaussian (assuming the data started out Gaussian), both because the Gram-Schmidt procedure performs only linear transformations on the data.

2. Virtual Z

The routing protocol has the interesting property that it is possible to correct for arbitrary coherent phase errors simply by adding virtual Z gates on the switch and output qubits after each routing operation. In many multi-qubit gate operations, phase errors cannot always be arbitrarily corrected (e.g., a general *i*SWAP [66]), which at a high

ALGORITHM 2. Modified Gram-Schmidt for measurement basis orthonormalization.

```

Define  $d = 2^n$ 
for all  $j \in \{0, 1, \dots, d-1\}$  do
  Initialize  $\mathcal{N} = \sqrt{\text{tr}(\tilde{\sigma}_j, \tilde{\sigma}_j)}$  ▷ normalization
  Initialize  $\tilde{\sigma}_j^\perp = \tilde{\sigma}_j / \mathcal{N}$ 
  Initialize  $p_j^\perp = p_j / \mathcal{N}$ 
  for all  $k \in \{j+1, j+2, \dots, d-1\}$  do
     $\mathcal{M} = \text{tr}(\tilde{\sigma}_k, \tilde{\sigma}_j^\perp)$  ▷ calculate overlap
     $\tilde{\sigma}_k = \tilde{\sigma}_k - \mathcal{M}\tilde{\sigma}_j^\perp$  ▷ subtract parallel component
     $p_k = p_k - \mathcal{M}p_j^\perp$ 
  end for
   $\tilde{\sigma}_j^\perp = \sqrt{d}\tilde{\sigma}_j^\perp$ 
   $p_j^\perp = \sqrt{d}p_j^\perp$ 
end for
return  $\{p_j^\perp\}, \{\tilde{\sigma}_j^\perp\}$ 

```

level can be explained by there being an insufficient number of knobs to fully correct for the number of independent phases that can be accumulated during the operation. For a general two-qubit *i*SWAP, we can tune two phases, φ_0 and φ_1 , by applying a virtual Z gate on each of the two qubits. However, there are three possible phases that can arise during the gate: $\varphi_{|01\rangle}$, $\varphi_{|10\rangle}$, and $\varphi_{|11\rangle}$, corresponding to phase accumulations on each qubit individually and phase accumulation due to the ZZ interaction between the qubits [66]. As a result, one of these phases cannot be corrected.

In the routing case, we also work with two qubits worth of Hilbert space when we consider only the levels that are allowed to have population. Specifically, we can consider the input space basis as four possible $|ab\rangle$, a state on switch-input with output qubits in $|0\rangle$. Similarly, the output space basis consists of four $|abc\rangle$, a state on switch-output 1-output 2 with the input qubit in $|0\rangle$. Thus, within the relevant subspace, just like in the standard two qubit case there are three possible phases that can arise, φ_α , φ_β , and φ_γ , which correspond respectively to the phases of the $|010\rangle\langle 01|$, $|100\rangle\langle 10|$, and $|101\rangle\langle 11|$ matrix elements relative to $|000\rangle\langle 00|$. The unitary for the performed routing operation can be written as

$$\begin{aligned}
 & U_{|SO_1O_2\rangle\langle SI|}^{\text{route}} \\
 &= \begin{matrix} & \langle 00| & \langle 01| & \langle 10| & \langle 11| \\ \begin{matrix} |000\rangle \\ |010\rangle \\ |100\rangle \\ |101\rangle \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -ie^{-i\varphi_\alpha} & 0 & 0 \\ 0 & 0 & e^{-i\varphi_\beta} & 0 \\ 0 & 0 & 0 & -ie^{-i\varphi_\gamma} \end{pmatrix} \end{matrix}
 \end{aligned} \tag{G2}$$

while the desired unitary $U_{|SO_1O_2\rangle\langle SI|}^{\text{route, ideal}}$ is

$$U_{|SO_1O_2\rangle\langle SI|}^{\text{route, ideal}} = \begin{matrix} & \langle 00| & \langle 01| & \langle 10| & \langle 11| \\ \begin{matrix} |000\rangle \\ |010\rangle \\ |100\rangle \\ |101\rangle \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -i & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -i \end{pmatrix} \end{matrix} \tag{G3}$$

Unlike in the standard two qubit *i*SWAP case, however, we now have three possible phase knobs that we can adjust, on the switch, output 1, and output 2. Thus, we should be able to fully compensate for any coherent phase accumulated using just virtual Z operations on these three qubits. Each virtual Z operation on qubit q can be written as

$$Z_q = e^{-i\varphi_q/2} \begin{pmatrix} 1 & 0 \\ 0 & e^{i\varphi_q} \end{pmatrix}. \tag{G4}$$

Multiplying out the application of virtual Z gates on the switch, output 1, and output 2, we find

$$\begin{aligned}
 & Z_S Z_{O_1} Z_{O_2} U_{|SO_1O_2\rangle\langle SI|}^{\text{route}} = e^{-i\varphi_0} \\
 & \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -ie^{-i(\varphi_\alpha - \varphi_{O_1})} & 0 & 0 \\ 0 & 0 & e^{-i(\varphi_\beta - \varphi_S)} & 0 \\ 0 & 0 & 0 & -ie^{-i(\varphi_\gamma - \varphi_S - \varphi_{O_2})} \end{pmatrix},
 \end{aligned} \tag{G5}$$

where φ_0 is just a global phase $(\varphi_S + \varphi_{O_1} + \varphi_{O_2})/2$ that can be dropped from the analysis. Based on this expression, we are free to pick $\varphi_{O_1} = \varphi_\alpha$, $\varphi_S = \varphi_\beta$, and $\varphi_{O_2} = \varphi_\gamma - \varphi_\beta$, which will perfectly cancel the extra phases to get the desired routing unitary $U_{|SO_1O_2\rangle\langle SI|}^{\text{route, ideal}}$. We can also pick a different set of φ_S , φ_{O_1} , and φ_{O_2} to obtain any other arbitrary desired phase on this unitary.

Experimentally, different initial states are sensitive to different virtual Z rotations. For example, an ideally prepared $|0+\rangle_{SI}$ should route to $(|000\rangle + |010\rangle)_{SO_1O_2}$, which is only sensitive to the virtual Z on output 1, while an ideally prepared $|++\rangle_{SI}$ should route to $(|000\rangle - i|010\rangle + |100\rangle - i|101\rangle)_{SO_1O_2}$, which is sensitive to all three virtual Zs applied. To ensure that we have implemented a unitary, the set of three virtual Zs applied to each final state must be the same for all initial states.

To test the extent to which this is true, we sweep over the virtual Z applied to the switch, output 1, and output 2 on the measured final state 3Q tomography and find the tuple of phases that obtains the maximum fidelity for the final state (“individual optimization”). We then perform a second search where we look for the tuple of phases that obtains the maximum average fidelity when applied to all

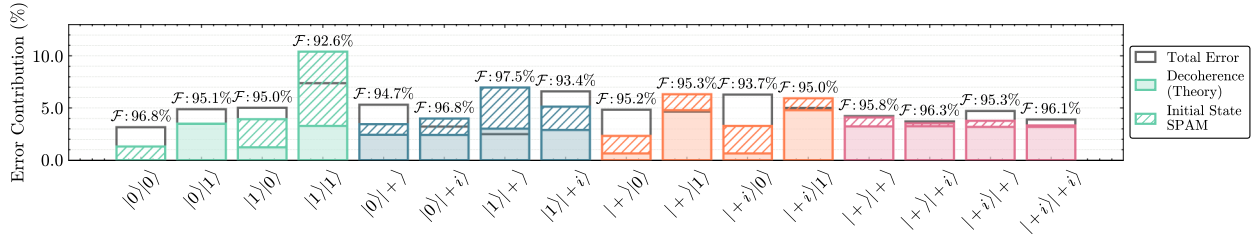


FIG. 11. **Error budget for all initial states.** State labels correspond to the initial state $|\text{switch}\rangle|\text{input}\rangle$ before applying the routing operation.

16 different initial states (“global optimization”). For a perfect unitary, the global and individual optimizations should return the same phase tuples. We ascribe any differences to timing errors, where a small discrepancy on the order of a few ns can cause noticeable accumulation of phases due to the large ZZ between the switch and input.

In particular, two relevant timescales are the delay between the initial state preparation and the tomography pulses to measure the initial state, and the possibly different delay between the preparation of the initial state and the start of the SWAP gates. The first timescale determines the phase accumulation on the state measured by the initial state tomography (e.g., Fig. 12) and is assumed to be 0 when we perform fidelity calculations referencing an ideal initial state. The second timescale determines the true state seen by the router when it begins the protocol. For different initial states, these delays can be slightly different depending on the exact pulse sequences used to prepare the initial state. The discrepancy between different initial state timings can result in the appearance of the unitary having a phase that is initial state dependent.

After optimizing the pulse timing in our setup, we are able to minimize the discrepancy such that we have a remaining fidelity drop of on average 0.4% when using global instead of individual phase optimization. We emphasize that the fidelities quoted in this work use global optimization since our goal is to implement a routing unitary.

APPENDIX H: ERROR CONTRIBUTION FOR ALL MEASURED INITIAL STATES

In this section, we elaborate on the error breakdown presented in Fig. 4(c). In Fig. 11 we show the error contribution from the theoretically estimated decoherence and the state preparation SPAM for all 16 cardinal states in which we initialize the router. For all states we use the convention $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$ and $|+i\rangle = (|0\rangle + i|1\rangle)/\sqrt{2}$. We confirm that the error estimation from decoherence and SPAM closely match the measured ones on average, although there is a state-by-state deviation that we postulate comes mostly from SPAM. Because of the large ZZ interaction between the switch and input, it is difficult to

properly prepare product states where one or both qubits have probability of being in $|e\rangle$. Furthermore, tomography pulses are also lower fidelity when measuring these states, and we find that even with ZZ correction (Appendix G 1 e), the reported density matrix does not perfectly recover states that we expect to be prepared with high fidelity. We rely on optimal control pulses (Appendix D) to improve the state preparation and tomography measurement, but we find that states such as $|1\rangle|1\rangle$ and $|1\rangle|+\rangle$ are still difficult to prepare and/or to measure with high fidelity.

We note, however, that for the protocol the preparation of the initial state is not relevant to the actual fidelity of the gate. Indeed, as long as the initial state of the system has no population in either of the output qubits, the protocol should be able to correctly route from the input to the outputs. Accordingly, we find that when we compare the measured 3Q density matrix to the measured initial state instead of to an ideal initial state, the measured fidelity improves by 1.7% on average. In Fig. 12(a), we show the real and imaginary components of the measured preparation of $|+\rangle|1\rangle$ (top row) and of the ideal preparation (bottom row). In Fig. 12(b), we compare the measured 3Q density matrix after sending the $|+\rangle|1\rangle$ state through the protocol to the 3Q density matrix generated by sending the measured initial state through an ideal protocol. The fidelity is then given by $\bar{F} = (\text{tr} \sqrt{\sqrt{\rho_{\text{meas}}} \rho_{\text{ideal}} \sqrt{\rho_{\text{meas}}}})^2$ [48], where ρ_{ideal} is the measured input state sent through ideal protocol and ρ_{meas} is the measured final density matrix.

APPENDIX I: SCALING THE ROUTER FOR A QRAM

In this section we propose how the Q^2 -router we have implemented could be integrated into a QRAM with a few address layers.

A QRAM can be implemented via either a gate-based architecture, which uses a standard QPU and compiles each query of the memory into other logical gates, or a router-based architecture, where queries are implemented via routing operations [7]. Compared to the gate-based approach, for the same memory size the router-based solution is exponentially more efficient in the number of

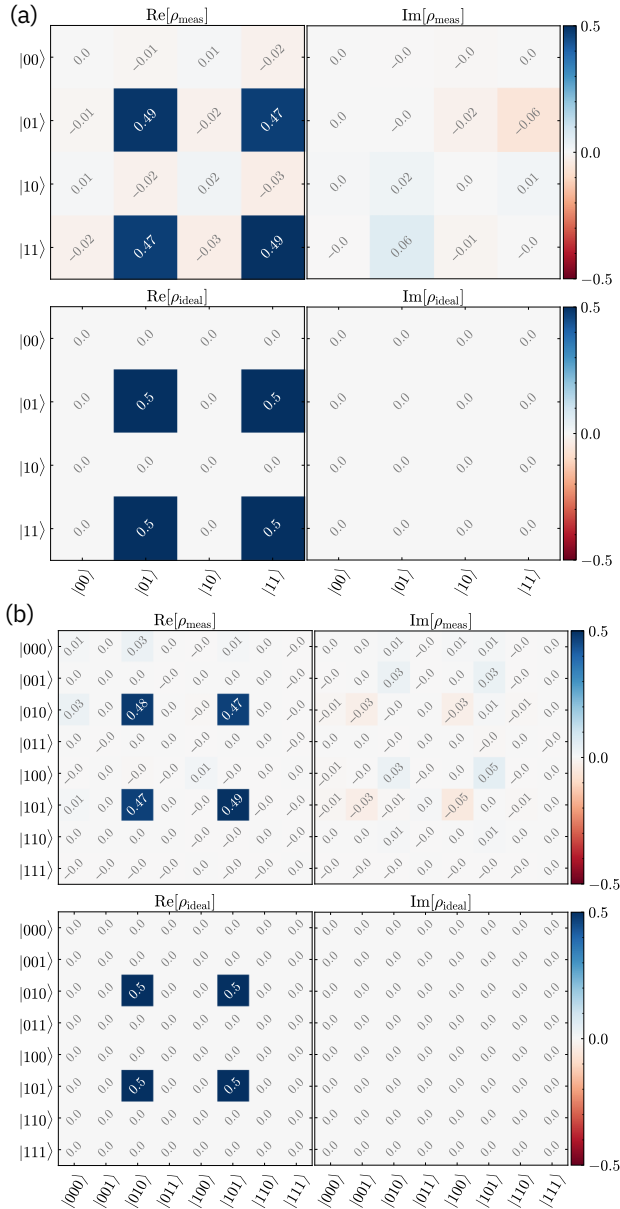


FIG. 12. **Tomography measurements for SPAM.** (a) 2Q tomography on $|+1\rangle_{SI}$. In the top (bottom) two panels we plot the measured (ideal) real and imaginary components of the density matrix. (b) 3Q tomography on $(|010\rangle + |101\rangle)_{SO1O2}$, the result of applying the protocol once on the initial state measured in (a). We plot the real and imaginary components of the measured density matrix (top row) and of the target density matrix obtained by applying a perfect protocol unitary on the measured initial state from (a) (bottom row).

sequential operations required for a single query [7]. Like a classical RAM fanout architecture, a router-based QRAM has a binary tree architecture where each level k of the tree corresponds to the k th bit of a base-2 memory address of length n . The binary tree can specify a total of 2^n memory locations. When information reaches a node (switch), it

can be routed to the left, right, or a superposition of the two depending on if the switch is in the 0, 1, or a superposition state. To query a single memory location or a superposition of memory locations, we can perform a routing operation from the root node of the tree all the way to the specified memory locations at the leaves of the tree, swap the information from the storage locations back into the tree, and perform the inverse routing operation to bring the retrieved information back to the root node.

To build a scalable QRAM, we propose to operate via the principles of the “bucket-brigade” algorithm [6]. Despite the QRAM architecture having a total number of qubits that is exponential in the total number of address bits n , this algorithm ensures that with a router-based QRAM architecture, each query only requires a number of $O(n)$ sequentially applied gates [5,6] and also improves noise resilience compared to a standard fanout architecture [5] by requiring a lower degree of entanglement within the tree to maintain the routing capability [33]. In the bucket-brigade architecture, an additional important step is a loading stage in which the switch state should be loaded with the address state prior to beginning the routing. This operation could be done on our router via a SWAP between the input qubit and the switch. In this paper we have simply initialized the switch directly to focus on the routing protocol only.

The routing protocol that we have implemented in this paper can be thought of as a routing operation in a single-address-bit QRAM. To perform a full query operation, in which we retrieve information from a superposition of address locations defined by the initial switch state and multiply the result by the initial input state, we can perform

$$U_{\text{query}} = Z_{\text{correction}}^{SO_1O_2} U_{\text{route}}^{-1} U_{\text{retrieve}} Z_{\text{correction}}^{SO_1O_2} U_{\text{route}}, \quad (11)$$

where $Z_{\text{correction}}^{SO_1O_2}$ is the tuple of virtual Z gates on the switch and outputs that we can calibrate via a single protocol operation (Appendix G 2) and U_{retrieve} is some operation that transfers information from a set of classical or quantum memory storage locations to the corresponding output qubit. For a classical database, for example, U_{retrieve} can be a set of classically-controlled Z gates on each final output-classical memory pair, where the control is the value of the classical memory at each address [67].

To tile our Q^2 -router into a QRAM with two or more address bits, we propose to use the same main operating principles with a few changes to ensure the bucket-brigade algorithm can be implemented. In Fig. 13 we show a schematic of a QRAM with two address bits. We first note that we do not need to use the $g - e$ subspace for all qubits in the binary tree. Instead, we can alternate between the $g - e$ and $g - f$ subspace for even and odd layers in the tree, which will allow the routing to be performed using $|eg\rangle - |gf\rangle$ sideband pulses as in our original Q^2 -router. In addition, we note that the output qubits of each layer become the input qubits of the following layer.

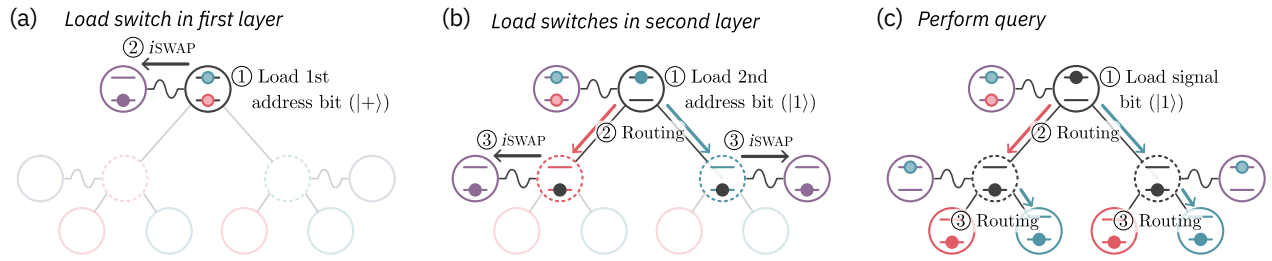


FIG. 13. Querying a two-address-bit QRAM. All diagrams show the state of the QRAM just after the first labeled step. We omit showing the states of qubits in the ground state when they do not participate in the labeled operations. Purple, black, red, and blue colors label the qubits that behave respectively as switches, inputs, and outputs 1 and 2 for the operation, following Fig. 1(b). (a) The first input qubit is in gray and the output qubits of the first layer, which will later act as the inputs of the second layer, are in red and blue. In this architecture, we route the input excitation in the tree using alternatively its $|e\rangle$ (solid circle) or $|f\rangle$ (dashed circle) states. Each non-switch qubit in the tree is coupled to its corresponding switch (purple) via a tunable coupler, (S-shaped line). Before performing a query, we load the switches one layer at a time as in the bucket-brigade algorithm for QRAM operation. To load the switches of both layers of the tree, we begin by loading the input with the first address state (here, $|+\rangle$) and transferring the state to the switch with an i SWAP (black arrow). (b) Next, we load the switches of the second layer. We initialize the input with the second address bit (here, $|1\rangle$, colored blue because it will act as a switch to output 2 once it is routed to the second layer switches) and use our Q^2 -router protocol to route this state to some superposition of the two outputs defined by the previously loaded first switch state ($|+\rangle$). We then perform an i SWAP between each second-layer switch and its corresponding input qubit. Since each Q^2 -router sub-module in the tree is essentially independent, these i SWAPs can be performed simultaneously to ensure that the gate time is constant as a function of the layer depth. (c) Once all the switches are initialized, we can run the query operation using some input state (here $|1\rangle$) by applying the protocol once on each switch-input-output 1-output 2 Q^2 -router sub-module, starting from the top of the tree. These routing protocols can also be applied simultaneously for each layer.

For the router-based bucket-brigade architecture [5], there are two main steps: (1) loading the switches with the address bits and (2) routing information between the input and output (and vice-versa). In Figs. 13(a) and 13(b) we show how the proposed architecture can be used to load the address bits in the switch. In addition to the native c - i SWAP presented in this article, we need an i SWAP between each input/output qubit and its respective switch. This i SWAP can be implemented from an $|eg\rangle - |gf\rangle$ sideband similar to the c - i SWAP that we have calibrated. The switches are loaded one layer at a time, where for the k -th address (qu)bit we load the bit in the first input, perform $k - 1$ layers of routing operations, and perform an i SWAP between each switch on the k th layer and its corresponding input qubit. By doing so, the address bit is loaded into all switches on the k th layer according to the superposition state defined by the previous address bit. After all address bits are loaded, the switches in the tree are in some complex superposition, and all other qubits are in the ground state. A full query operation can be accomplished simply by loading a final signal bit and applying each layer of routing operations sequentially from the top to bottom layer [Fig. 13(c)].

One caveat is that to scale up to multiple memory layers, we need to add a tunable ZZ interaction between each switch and its corresponding input qubit. If the ZZ interactions are always-on as in our Q^2 -router, they would block the c - i SWAPs of the following layer. This behavior

can be seen from Fig. 13(c) where once all the address bits have been loaded in the switches, without a tunable ZZ the switches of the second layer would ZZ -shift their corresponding attached qubits (the inputs of the second layer), such that during the routing procedure, the c - i SWAPs between the input of the first layer and the inputs of the second layer would no longer be resonant. A tunable coupler between each input qubit and its switch, implemented for example by a frequency-tunable transmon, would allow the ZZ interaction to be tuned close to zero in the idle position or activated only when we want to drive the c - i SWAPs of the corresponding layer.

-
- [1] V. Bužek and M. Hillery, Quantum copying: Beyond the no-cloning theorem, *Phys. Rev. A* **54**, 1844 (1996).
 - [2] K. Lemr, K. Bartkiewicz, A. Černoč, and J. Soubusta, Resource-efficient linear-optical quantum router, *Phys. Rev. A* **87**, 062333 (2013).
 - [3] J. I. Cirac, P. Zoller, H. J. Kimble, and H. Mabuchi, Quantum state transfer and entanglement distribution among distant nodes in a quantum network, *Phys. Rev. Lett.* **78**, 3221 (1997).
 - [4] H. J. Kimble, The quantum internet, *Nature* **453**, 1023 (2008).
 - [5] V. Giovannetti, S. Lloyd, and L. Maccone, Architectures for a quantum random access memory, *Phys. Rev. A* **78**, 052310 (2008).

- [6] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum random access memory, *Phys. Rev. Lett.* **100**, 160501 (2008).
- [7] S. Xu, C. T. Hann, B. Foxman, S. M. Girvin, and Y. Ding, in *56th Annual IEEE/ACM International Symposium on Microarchitecture* (2023), pp. 526–538.
- [8] L. K. Grover, in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96 (Association for Computing Machinery, New York, NY, United States, 1996), pp. 212–219.
- [9] A. W. Harrow, A. Hassidim, and S. Lloyd, Quantum algorithm for linear systems of equations, *Phys. Rev. Lett.* **103**, 150502 (2009).
- [10] N. Wiebe, A. Kapoor, and K. M. Svore, Quantum deep learning, [arXiv:1412.3489](https://arxiv.org/abs/1412.3489).
- [11] D. W. Berry, C. Gidney, M. Motta, J. R. McClean, and R. Babbush, Qubitization of arbitrary basis quantum chemistry leveraging sparsity and low rank factorization, *Quantum* **3**, 208 (2019).
- [12] R. Babbush, C. Gidney, D. W. Berry, N. Wiebe, J. McClean, A. Paler, A. Fowler, and H. Neven, Encoding electronic spectra in quantum circuits with linear T complexity, *Phys. Rev. X* **8**, 041015 (2018).
- [13] J. Miguel-Ramiro, A. Pirker, and W. Dür, Genuine quantum networks with superposed tasks and addressing, *npj Quantum Inf.* **7**, 1 (2021).
- [14] G. Chiribella and H. Kristjánsson, Quantum Shannon theory with superpositions of trajectories, *Proc. R. Soc. A: Math., Phys. Eng. Sci.* **475**, 20180903 (2019).
- [15] K. Bartkiewicz, A. Černoč, and K. Lemr, Using quantum routers to implement quantum message authentication and Bell-state manipulation, *Phys. Rev. A* **90**, 022335 (2014).
- [16] A. Barenco, A. Berthiaume, D. Deutsch, A. Ekert, R. Jozsa, and C. Macchiavello, Stabilization of quantum computations by symmetrization, *SIAM J. Comput.* **26**, 1541 (1997).
- [17] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf, Quantum fingerprinting, *Phys. Rev. Lett.* **87**, 167902 (2001).
- [18] J. Liu, C. T. Hann, and L. Jiang, Data centers with quantum random access memory and quantum networks, *Phys. Rev. A* **108**, 032610 (2023).
- [19] G. Lee, C. T. Hann, S. Puri, S. M. Girvin, and L. Jiang, Error suppression for arbitrary-size black box quantum operations, *Phys. Rev. Lett.* **131**, 190601 (2023).
- [20] J. Miguel-Ramiro, Z. Shi, L. Dellantonio, A. Chan, C. A. Muschik, and W. Dür, Superposed quantum error mitigation, *Phys. Rev. Lett.* **131**, 230601 (2023).
- [21] X. Wu, H. Yan, G. Andersson, A. Anferov, M.-H. Chou, C. R. Conner, J. Grebel, Y. J. Joshi, S. Li, J. M. Miller, R. G. Povey, H. Qiao, and A. N. Cleland, Modular quantum processor with an all-to-all reconfigurable router, *Phys. Rev. X* **14**, 041030 (2024).
- [22] 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**, 1 (2023).
- [23] R. K. Naik, N. Leung, S. Chakram, P. Groszkowski, Y. Lu, N. Earnest, D. C. McKay, J. Koch, and D. I. Schuster, Random access quantum information processors using multimode circuit quantum electrodynamics, *Nat. Commun.* **8**, 1904 (2017).
- [24] M. Pechal, J.-C. Besse, M. Mondal, M. Oppliger, S. Gasparinetti, and A. Wallraff, Superconducting switch for fast on-chip routing of quantum microwave fields, *Phys. Rev. Appl.* **6**, 024009 (2016).
- [25] B. Kannan, A. Almanakly, Y. Sung, A. Di Paolo, D. A. Rower, J. Braumüller, A. Melville, B. M. Niedzielski, A. Karamlou, K. Serniak, A. Vepsäläinen, M. E. Schwartz, J. L. Yoder, R. Winik, J. I.-J. Wang, T. P. Orlando, S. Gustavsson, J. A. Grover, and W. D. Oliver, On-demand directional microwave photon emission using waveguide quantum electrodynamics, *Nat. Phys.* **19**, 394 (2023).
- [26] A. Almanakly, B. Yankelevich, M. Hays, B. Kannan, R. Assouly, A. Greene, M. Gingras, B. M. Niedzielski, H. Stickler, M. E. Schwartz, K. Serniak, J. I.-J. Wang, T. P. Orlando, S. Gustavsson, J. A. Grover, and W. D. Oliver, Deterministic remote entanglement using a chiral quantum interconnect, [arXiv:2408.05164](https://arxiv.org/abs/2408.05164).
- [27] I.-C. Hoi, C. M. Wilson, G. Johansson, T. Palomaki, B. Peropadre, and P. Delsing, Demonstration of a single-photon router in the microwave regime, *Phys. Rev. Lett.* **107**, 073601 (2011).
- [28] I. Söllner, S. Mahmoodian, S. L. Hansen, L. Midolo, A. Javadi, G. Kiršanskė, T. Pregolato, H. El-Ella, E. H. Lee, J. D. Song, S. Stobbe, and P. Lodahl, Deterministic photon-emitter coupling in chiral photonic circuits, *Nat. Nanotechnol.* **10**, 775 (2015).
- [29] P. Lodahl, S. Mahmoodian, S. Stobbe, A. Rauschenbeutel, P. Schneeweiss, J. Volz, H. Pichler, and P. Zoller, Chiral quantum optics, *Nature* **541**, 473 (2017).
- [30] R. J. Coles, D. M. Price, J. E. Dixon, B. Royall, E. Clarke, P. Kok, M. S. Skolnick, A. M. Fox, and M. N. Makhonin, Chirality of nanophotonic waveguide with embedded quantum emitter for unidirectional spin transfer, *Nat. Commun.* **7**, 11183 (2016).
- [31] K. C. Chen, W. Dai, C. Errando-Herranz, S. Lloyd, and D. Englund, Scalable and high-fidelity quantum random access memory in spin-photon networks, *PRX Quantum* **2**, 030319 (2021).
- [32] F.-Y. Hong, Y. Xiang, Z.-Y. Zhu, L.-Z. Jiang, and L.-N. Wu, Robust quantum random access memory, *Phys. Rev. A* **86**, 010306 (2012).
- [33] D. Weiss, S. Puri, and S. Girvin, Quantum random access memory architectures using 3D superconducting cavities, *PRX Quantum* **5**, 020312 (2024).
- [34] C. T. Hann, C.-L. Zou, Y. Zhang, Y. Chu, R. J. Schoelkopf, S. M. Girvin, and L. Jiang, Hardware-efficient quantum random access memory with hybrid quantum acoustic systems, *Phys. Rev. Lett.* **123**, 250501 (2019).
- [35] Y.-J. Wang, S. Zhang, T.-P. Sun, Z.-A. Zhao, X.-F. Xu, X.-N. Zhuang, H.-Y. Liu, C. Xue, P. Duan, Y.-C. Wu, Z.-Y. Chen, and G.-P. Guo, Hardware-efficient quantum random access memory design with a native gate set on superconducting platforms, *Adv. Quantum Technol.* **8**, 2400519 (2025).
- [36] Z. Wang, H. Qiao, A. N. Cleland, and L. Jiang, Quantum random access memory with transmon-controlled phonon routing, *Phys. Rev. Lett.* **134**, 210601 (2025).

- [37] C.-K. Hu, J. Yuan, B. A. Veloso, J. Qiu, Y. Zhou, L. Zhang, J. Chu, O. Nurbolat, L. Hu, J. Li, Y. Xu, Y. Zhong, S. Liu, F. Yan, D. Tan, R. Bachelard, A. C. Santos, C. Villas-Boas, and D. Yu, Native conditional i swap operation with superconducting artificial atoms, *Phys. Rev. Appl.* **20**, 034072 (2023).
- [38] B. J. Chapman, S. J. De Graaf, S. H. Xue, Y. Zhang, J. Teoh, J. C. Curtis, T. Tsunoda, A. Eickbusch, A. P. Read, A. Koottandavida, S. O. Mundhada, L. Frunzio, M. Devoret, S. Girvin, and R. Schoelkopf, High-on-off-ratio beam-splitter interaction for gates on bosonically encoded qubits, *PRX Quantum* **4**, 020355 (2023).
- [39] C. W. Warren, J. Fernández-Pendás, S. Ahmed, T. Abad, A. Bengtsson, J. Biznárová, K. Debnath, X. Gu, C. Križan, A. Osman, A. Fadavi Roudsari, P. Delsing, G. Johansson, A. Frisk Kockum, G. Tancredi, and J. Bylander, Extensive characterization and implementation of a family of three-qubit gates at the coherence limit, *npj Quantum Inf.* **9**, 1 (2023).
- [40] X. X. Yuan, J.-J. Ma, P.-Y. Hou, X.-Y. Chang, C. Zu, and L.-M. Duan, Experimental demonstration of a quantum router, *Sci. Rep.* **5**, 12452 (2015).
- [41] S. Zeytinoglu, M. Pechal, S. Berger, A. A. Abdumalikov, A. Wallraff, and S. Filipp, Microwave-induced amplitude- and phase-tunable qubit-resonator coupling in circuit quantum electrodynamics, *Phys. Rev. A* **91**, 043846 (2015).
- [42] D. Egger, M. Ganzhorn, G. Salis, A. Fuhrer, P. Müller, P. K. Barkoutsos, N. Moll, I. Tavernelli, and S. Filipp, Entanglement generation in superconducting qubits using holonomic operations, *Phys. Rev. Appl.* **11**, 014017 (2019).
- [43] T. Tsunoda, J. D. Teoh, W. D. Kalfus, S. J. de Graaf, B. J. Chapman, J. C. Curtis, N. Thakur, S. M. Girvin, and R. J. Schoelkopf, Error-detectable bosonic entangling gates with a noisy ancilla, *PRX Quantum* **4**, 020354 (2023).
- [44] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, *Phys. Rev. A* **85**, 042311 (2012).
- [45] 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).
- [46] 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).
- [47] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, Efficient measurement of quantum gate error by interleaved randomized benchmarking, *Phys. Rev. Lett.* **109**, 080505 (2012).
- [48] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, 2010).
- [49] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Characterizing quantum supremacy in near-term devices, *Nat. Phys.* **14**, 595 (2018).
- [50] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, Characterizing large-scale quantum computers via cycle benchmarking, *Nat. Commun.* **10**, 5347 (2019).
- [51] S. Leger, and C. Miao, Supporting data for “Implementation of a quantum addressable router using superconducting qubits” (2025), <https://zenodo.org/records/16954111>.
- [52] S. Huang, B. Lienhard, G. Calusine, A. Vepsäläinen, J. Braumüller, D. K. Kim, A. J. Melville, B. M. Niedzielski, J. L. Yoder, B. Kannan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, Microwave package design for superconducting quantum processors, *PRX Quantum* **2**, 020306 (2021).
- [53] M. J. Persky, Review of black surfaces for space-borne infrared systems, *Rev. Sci. Instrum.* **70**, 2193 (1999).
- [54] C. Ding, M. Di Federico, M. Hatridge, A. Houck, S. Leger, J. Martinez, C. Miao, D. Schuster I, L. Stefanazzi, C. Stoughton, S. Sussman, K. Treptow, S. Uemura, N. Wilcer, H. Zhang, C. Zhou, and G. Cancelo, Experimental advances with the QICK (Quantum Instrumentation Control Kit) for superconducting quantum hardware, *Phys. Rev. Res.* **6**, 013305 (2024).
- [55] L. Stefanazzi, K. Treptow, N. Wilcer, C. Stoughton, S. Montella, C. Bradford, G. Cancelo, S. Saxena, H. Arnaldi, S. Sussman, A. Houck, A. Agrawal, H. Zhang, C. Ding, and D. I. Schuster, The QICK (Quantum Instrumentation Control Kit): Readout and control for qubits and detectors, *Rev. Sci. Instrum.* **93**, 044709 (2022).
- [56] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, Circuit quantum electrodynamics, *Rev. Mod. Phys.* **93**, 025005 (2021).
- [57] D. A. Rower, L. Ding, H. Zhang, M. Hays, J. An, P. M. Harrington, I. T. Rosen, J. M. Gertler, T. M. Hazard, B. M. Niedzielski, M. E. Schwartz, S. Gustavsson, K. Serniak, J. A. Grover, and W. D. Oliver, Suppressing counter-rotating errors for fast single-qubit gates with fluxonium, *PRX Quantum* **5**, 040342 (2024).
- [58] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, Efficient Z gates for quantum computing, *Phys. Rev. A* **96**, 022330 (2017).
- [59] C. J. Wood and J. M. Gambetta, Quantification and characterization of leakage errors, *Phys. Rev. A* **97**, 032306 (2018).
- [60] Y.-H. Chen and C. H. Baldwin, Randomized benchmarking with leakage errors, [arXiv:2502.00154](https://arxiv.org/abs/2502.00154).
- [61] M. A. Rol, F. Battistel, F. K. Malinowski, C. C. Bultink, B. M. Tarasinski, R. Vollmer, N. Haider, N. Muthusubramanian, A. Bruno, B. M. Terhal, and L. DiCarlo, Fast, high-fidelity conditional-phase gate exploiting leakage interference in weakly anharmonic superconducting qubits, *Phys. Rev. Lett.* **123**, 120502 (2019).
- [62] L. B. Nguyen, Y. Kim, A. Hashim, N. Goss, B. Marinelli, B. Bhandari, D. Das, R. K. Naik, J. M. Kreikebaum, A. N. Jordan, D. I. Santiago, and I. Siddiqi, Programmable Heisenberg interactions between Floquet qubits, *Nat. Phys.* **20**, 240 (2024).
- [63] G. Aleksandrowicz *et al.*, Qiskit: An open-source framework for quantum computing. Zenodo, 2019.
- [64] J. A. Smolin, J. M. Gambetta, and G. Smith, Efficient method for computing the maximum-likelihood quantum state from measurements with additive Gaussian noise, *Phys. Rev. Lett.* **108**, 070502 (2012).

- [65] T. Roy, Z. Li, E. Kapit, and D. I. Schuster, Tomography in the presence of stray inter-qubit coupling, [arXiv:2103.13611](#).
- [66] 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 \(2020\)](#).
- [67] C. T. Hann, G. Lee, S. Girvin, and L. Jiang, Resilience of quantum random access memory to generic noise, [PRX Quantum **2**, 020311 \(2021\)](#).