

Demonstrating Coherent Quantum Routers for Bucket-Brigade Quantum Random Access Memory on a Superconducting Processor

Sheng Zhang,^{1,2,3} Yun-Jie Wang,⁴ Peng Wang^{1,2,3}, Ren-Ze Zhao^{1,2}, Xiao-Yan Yang^{1,2}, Ze-An Zhao,^{1,2} Tian-Le Wang^{1,2}, Hai-Feng Zhang^{1,2}, Zhi-Fei Li^{1,2}, Yuan Wu,^{1,2} Hao-Ran Tao,⁵ Liang-Liang Guo,⁵ Lei Du,⁵ Chi Zhang,⁵ Zhi-Long Jia,⁵ Wei-Cheng Kong,⁵ Zhuo-Zhi Zhang^{1,2,3}, Xiang-Xiang Song,^{1,2,3} Yu-Chun Wu^{1,2,6}, Zhao-Yun Chen,^{6,*} Peng Duan^{1,2,†} and Guo-Ping Guo^{1,2,5,‡}

¹Laboratory of Quantum Information, *University of Science and Technology of China*, Hefei, Anhui, 230026, China

²CAS Center For Excellence in Quantum Information and Quantum Physics, *University of Science and Technology of China*, Hefei, Anhui, 230026, China

³Suzhou Institute for Advanced Research, *University of Science and Technology of China*, Suzhou, Jiangsu, 215123, China

⁴Institute of the Advanced Technology, *University of Science and Technology of China*, Hefei, Anhui, 230088, China

⁵Origin Quantum, Hefei, Anhui, 230088, China

⁶Institute of Artificial Intelligence, *Hefei Comprehensive National Science Center*, Hefei, Anhui, 230088, China

 (Received 27 May 2025; revised 16 April 2026; accepted 24 July 2026; published 25 August 2026)

Quantum routers (QRouters) are essential components of bucket-brigade quantum random access memory (QRAM), enabling quantum applications such as Grover's search and quantum machine learning. Despite significant theoretical advances, achieving scalable and coherent QRouters experimentally remains challenging. Here, we demonstrate coherent quantum routers using a superconducting quantum processor, laying a practical foundation for scalable QRAM systems. The quantum router at the core of our implementation utilizes the transition composite gate scheme, wherein auxiliary energy levels temporarily mediate conditional interactions. By leveraging two-qubit primitives in its construction, the quantum router preserves straightforward initialization and calibration while significantly reducing circuit depth and duration relative to traditional gate-based decompositions. Moreover, by encoding routing addresses in the nonadjacent qutrit states $|0\rangle$ and $|2\rangle$, our design inherently enables erasure-detection capability, providing efficient postselection to mitigate routing errors. Experimentally, we achieve an average fidelity of 94.8% across three individual QRouters, and we validate scalability through a two-layer quantum routing network achieving a fidelity of 82.4%. Our results represent a significant advancement in quantum routing technology, providing enhanced fidelity, built-in error resilience, and practical scalability crucial for the development of future QRAM and large-scale quantum computing architectures.

DOI: [10.1103/h5m3-qm9](https://doi.org/10.1103/h5m3-qm9)

Subject Areas: Quantum Physics,
Quantum Information

I. INTRODUCTION

Quantum random access memory (QRAM) stands as a cornerstone of quantum computing [1–8], enabling superposition-based access to classical data—a crucial capability

for algorithms such as Grover's search [9], quantum machine learning [10–13], and quantum simulation [14–19].

Among various QRAM designs, the bucket-brigade architecture [1,20] stands out for its favorable scaling and noise resilience [21], which organizes QRAM access as a conditional routing process along a binary-tree network. The key enabling element in this scheme is the quantum router (QRouter)—a conditional switch that coherently routes a quantum input to one of two output paths based on the superposed state. Through recursive composition of such routers, a complete binary-tree routing network can be constructed to support superposition-based queries for QRAM.

*Contact author: chenzhaoyun@iai.ustc.edu.cn

†Contact author: pengduan@ustc.edu.cn

‡Contact author: gpguo@ustc.edu.cn

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

While the concept of quantum routers has been widely explored in both theory and experiment [22–37], most of these studies focus on quantum communication scenarios, where routing typically refers to controlling the propagation of flying qubits. While preserving coherence in transit, they are predominantly routed via nondeterministic [34,35], incomplete-transmission methods [29,31], rendering them unsuitable for the reversible bucket-brigade QRAM. In contrast, stationary qubit systems, such as superconducting circuits [38–40], are particularly well suited to support fully coherent, deterministic, and cascable routing [37,41–47]. Recent notable progress includes the first experimental demonstration of deterministic quantum routing on a superconducting processor via native three-qubit operations as well as proposals for three-layer QRAM implementations based on subspace-equivalent digital decompositions [37,48]. As with many superconducting implementations, engineering considerations remain: Native multiqubit routing requires complex initialization and calibration, while traditional digital decomposition based on standard one- and two-qubit gates inherently deepens circuits [49,50], exacerbating decoherence and cumulative gate errors compared to the native implementations. These challenges highlight the need for new architectures that bridge the gap between router-level coherence and QRAM-level scalability through compact, hardware-efficient designs.

In this work, we address these challenges by implementing QRouters based on the recently proposed transition composite gate (TCG) scheme [51]. The TCG scheme uses auxiliary energy levels as intermediate states to mediate conditional interactions, allowing for shallower gate sequences than conventional Clifford decompositions. This results in improved coherence and reduced circuit overhead—an essential feature for routing networks. By relying on two-qubit primitives rather than native multi-qubit gates, the approach preserves simple initialization and calibration, thereby enabling scalable circuit synthesis and implementation. Furthermore, by encoding the address qubit in the nonadjacent qutrit states $|0\rangle$ and $|2\rangle$, we enable an erasure-detection mechanism [48,52–57], through which enormous routing errors can be identified and discarded by postselection. This provides lightweight error mitigation without the need for ancillary hardware.

Leveraging both shallow circuit depth and erasure-detection-based error suppression, we present the first experimental demonstration of a two-layer coherent quantum routing system for bucket-brigade QRAM on a superconducting quantum processor. Specifically, we experimentally construct and characterize three individual QRouters with an average fidelity of 94.8% and further demonstrate a two-layer quantum routing network achieving a fidelity of 82.4%. These results establish a scalable and error-resilient routing architecture with minimal overhead, representing a significant step toward practical QRAM systems.

II. QRROUTER ARCHITECTURE AND IMPLEMENTATION

A QRouter implements a conditional SWAP operation between an input qubit Q_I and one of two path qubits (Q_L or Q_R), depending on the quantum state of an address qutrit Q_C . The logical function is defined as

$$U_{\text{QRouter}} = |L\rangle\langle L| \otimes \text{SWAP}_{Q_L, Q_I} + |R\rangle\langle R| \otimes \text{SWAP}_{Q_R, Q_I}, \quad (1)$$

where $|L\rangle$ and $|R\rangle$ are two orthogonal address states (e.g., $|0\rangle$, $|1\rangle$, or $|2\rangle$) encoded in Q_C . The QRouter, thus, conditionally transfers the quantum state of Q_I to either Q_L or Q_R , realizing address-dependent data routing.

Figure 1(a) illustrates the logical structure of a QRouter and how multiple such routers are recursively composed into a binary-tree network to support superposition-based queries. The physical implementation of the QRouter, shown in Fig. 1(b), consists of two controlled-SWAP (CSWAP) gates and two address-flipping operations. Each CSWAP conditionally exchanges Q_I with Q_L or Q_R , based on the state of Q_C .

Direct realization of three-qubit CSWAP gates is experimentally demanding. To overcome this, we employ the transition composite gate (TCG) scheme [51], which

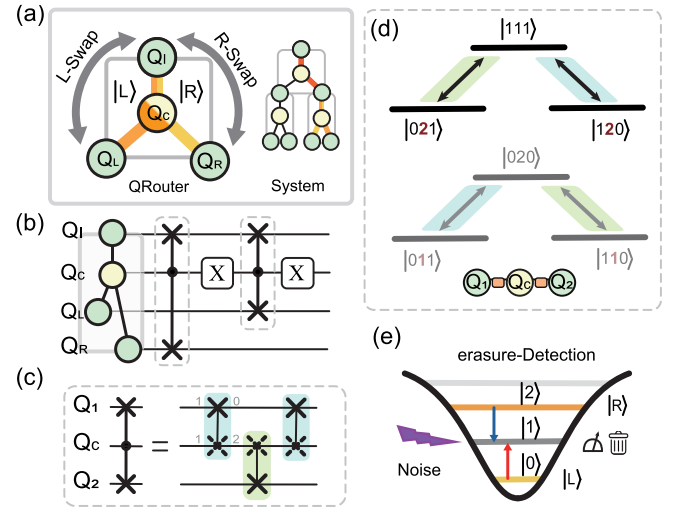


FIG. 1. Design and operation of the QRouter and routing network. (a) Logical structure of a QRouter and its recursive integration into a binary-tree routing network. Each QRouter routes an input qubit (Q_I) to a path qubit (Q_L or Q_R) based on the address state of Q_C . (b) Circuit decomposition of the QRouter using two CSWAP gates and two address-flipping operations. (c) Gate-level realization of CSWAP via the TCG scheme, using three $\sqrt{\text{CZ}}$ gates that induce transitions between $|11\rangle$ and $|02\rangle$. (d) Transition pathways for the TCG-based CSWAP under address states $|1\rangle$ (gray) and $|2\rangle$ (black). (e) Illustration of routing errors caused by leakage to $|1\rangle$ and the erasure-detection protocol for error suppression through postselection.

TABLE I. Gate counts and circuit depth for different QRouter implementations. Comparison between standard Clifford decomposition and the TCG-based QRouter under two encoding schemes: nonerasure (address states $|0\rangle$ and $|1\rangle$) and erasure scheme (address states $|0\rangle$ and $|2\rangle$). TCG significantly reduces the number of single- and two-qubit gates, as well as circuit depth, compared to the Clifford-based implementation. The erasure scheme incurs slightly more single-qubit gates due to address-state flipping but retains overall efficiency.

QRouter	Clifford	TCG ^{nonerasure}	TCG ^{erasure}
N_{1q}	20	2	6
N_{2q}	16	6	6
Depth	30	8	12

decomposes the CSWAP into a sequence of three two-qubit entangling gates ($\sqrt{CZ}$), as shown in Fig. 1(c). The key idea is to temporarily access the noncomputational $|2\rangle$ level of transmons to enable state transitions such as $|11\rangle \leftrightarrow |20\rangle$ or $|02\rangle$ [58,59], thus implementing effective conditional operations with reduced gate depth.

The transition pathways for the TCG-based CSWAP are shown in Fig. 1(d). A conditional SWAP can be achieved through sequences like $|011\rangle \rightarrow |020\rangle \rightarrow |110\rangle$ or $|021\rangle \leftrightarrow |111\rangle \leftrightarrow |120\rangle$, depending on whether the control qutrit Q_C is in $|1\rangle$ or $|2\rangle$ (gray and black pathways, respectively). When Q_C is in the $|0\rangle$ state, the SWAP operation is inhibited.

This TCG-based approach represents a key advantage of our design: By engineering intermediate-level transitions, we achieve a shallower gate decomposition compared to traditional Clifford-based methods, reducing both circuit depth and accumulated error. Table I quantitatively compares gate counts and circuit depths for different decomposition strategies, with further details available in Appendix B 1.

The address qutrit Q_C can be encoded in two configurations: $|L\rangle = |0\rangle$, $|R\rangle = |1\rangle$ (nonerasure scheme) or $|L\rangle = |0\rangle$, $|R\rangle = |2\rangle$ (erasure scheme). The latter enables an additional mechanism for error mitigation where the Q_C functions as an erasure qubit [52,53]. As shown in Fig. 1(e), gate imperfections, decoherence, or thermal population can cause Q_C to transition into the intermediate $|1\rangle$ state, which leads to routing errors. In the erasure scheme, these erroneous events can be postselected by measuring the final state of Q_C . If Q_C is found in $|1\rangle$, the corresponding data are discarded. This built-in erasure-detection mechanism allows us to suppress a dominant class of routing errors without introducing additional hardware or qubits—another core advantage of our QRouter design.

We implemented the QRouter using ten transmon qubits on the superconducting processor *Wukong*, with additional details provided in Appendix A 1. The $\sqrt{CZ}$ gates are realized by tuning frequencies of qubits and coupler close

to the $|02\rangle \leftrightarrow |11\rangle$ interaction point wherein these two states accomplish a SWAP operation. The gates are further optimized through Floquet sequence experiments [60–62] (see Appendix A 2). Address preparation and flipping (X) were carried out using single-qubit gates designed with the derivative reduction by adiabatic gate scheme [63]. The state $|0\rangle \leftrightarrow |1\rangle$ flip uses an X_π^{01} gate, while the state $|1\rangle \leftrightarrow |2\rangle$ flip uses an X_π^{12} gate.

All address qubits were measured at the end of each experiment. Unless otherwise specified, we used the erasure scheme encoding.

III. CHARACTERIZATION OF QUANTUM ROUTER

We begin by characterizing the conditional routing and the coherence properties of the QRouter, where Fig. 2 illustrates the experimental setups. First, we

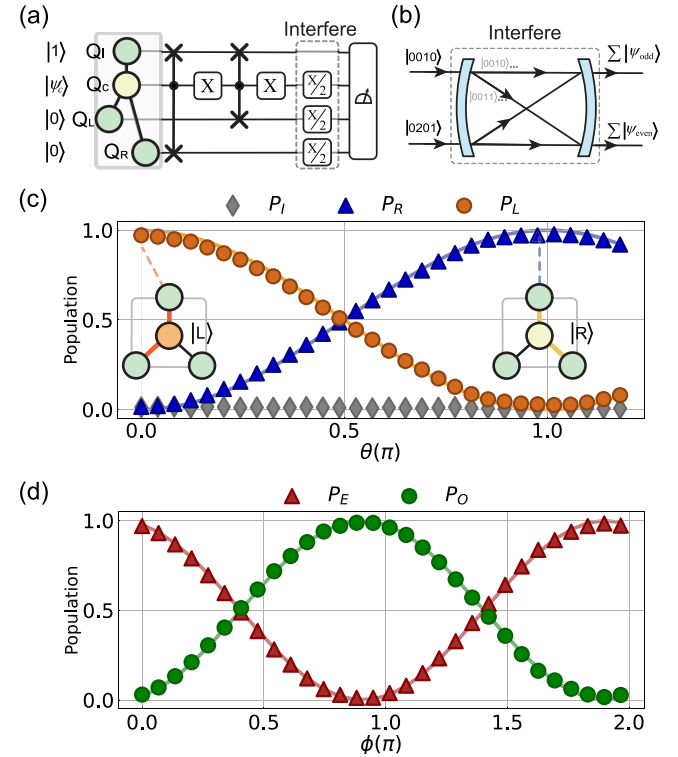


FIG. 2. Verification of conditional and coherent routing properties of the QRouter. (a) Experimental setup for testing both transmission (via θ variation) and coherence (via ϕ variation) of the QRouter with address qubit Q_C initialized in the state $|\psi_c\rangle = \cos(\theta/2)|0\rangle + e^{i\phi}\sin(\theta/2)|2\rangle$. Q_I is prepared in state $|1\rangle$, and the path qubits are prepared in state $|0\rangle$. (b) Interference scheme using three $X_{\pi/2}$ gates to probe phase-sensitive entanglement. (c) Routing population versus θ , showing expected $\cos^2(\theta/2)$ and $\sin^2(\theta/2)$ behavior for path qubits Q_L and Q_R . (d) Phase-coherent interference oscillations in the population of odd- and even-parity states as a function of ϕ , indicating address-path coherence. Dots represent experimental data (5000 shots), and solid lines are theoretical predictions.

prepared the address qutrit in a superposition state $|\psi_c\rangle = \cos(\theta/2)|0\rangle_c + e^{i\phi}\sin(\theta/2)|2\rangle_c$ and observed how information flows from the input qubit Q_I to path qubits Q_L and Q_R . As shown in Fig. 2(c), we varied θ and found that the population distributions of P_L (P_R) changed continuously with adjusted θ , following the theoretical $\cos^2(\theta/2)$ [$\sin^2(\theta/2)$] behavior (lines). The residual population of the input qubit (P_I) stayed below a certain threshold (2%) after routing, indicating high transmission efficiency of approximately 98%. Second, to verify the phase coherence during routing, we scanned ϕ and applied three $X_{\pi/2}$ operations in parallel to interfere with the generated entangled states of Q_I , Q_C , Q_L , and Q_R , as illustrated in Fig. 2(b). The phase of the address changes the phase difference between the entangled states generated under this coherent control. This leads to oscillations in the population of odd-parity states ($P_O = \sum_k P_{|k\rangle}$, $k \in \{0001, 0010, 0200, 0211\}$). Here, to evaluate the parity of the routing network, the address state $|2\rangle$ is treated as a single excitation. By compensating for a single-qubit phase $\delta\phi$, P_O is aligned to oscillate as $(1 - \sin\phi)/2$. Even-parity states ($P_E = \sum_k P_{|k\rangle}$, $k \in \{0000, 0011, 0201, 0210\}$) exhibited complementary behaviors, indicating a ϕ -driven coherence process, as shown in Fig. 2(d). The high-contrast routing in Fig. 2(c) and clear interference patterns in Fig. 2(d) demonstrate that our QRouter effectively achieves deterministic routing with address-path entanglement. More theoretical and experimental details are provided in Appendix C.

To benchmark the fidelity of our QRouter, we employed a standard quantum state tomography (QST), with the setup shown in Fig. 3(a). After routing under various addresses, QST reconstructed the experimental state density matrix (ρ) of Q_C , Q_L , and Q_R , with fidelities as $F_{\text{QST}} = \langle \Psi | \rho | \Psi \rangle$, where $|\Psi\rangle$ is the target state. Figures 3(c) and 3(d) display QST results under addresses $|0\rangle$ and $|+\rangle = (|0\rangle + |2\rangle)/\sqrt{2}$, respectively, where corresponding fidelities are 95.66% and 93.36%. We additionally provide numerical simulations covering more representative input and address configurations, with methods and complete results in Appendix D.

We further developed a random access test (RAT) to evaluate the system's average fidelity [see Fig. 3(b)]. Here, random address sequences $\{(a_1, a_1), \dots, (a_n, a_n), (a_{n+1})\}$ ($a_i \in \{|0\rangle, |2\rangle, |+\rangle, |-\rangle\}$), where $|\pm\rangle = (|0\rangle \pm |2\rangle)/\sqrt{2}$, were loaded into $2N + 1$ QRouters, with each sequence repeated 30 times. Then, we measured Q_I , Q_L , and Q_R to determine the population $P_{|k'\rangle}$ ($k' \in \{000, 001, 010, 011\}$). We regarded the consistency between ideal ($P_{|t\rangle}^0$) and experimental ($P_{|t\rangle}^e$) populations, $M_{\text{RAT}} = 1 - \sum_t |P_{|t\rangle}^0 - P_{|t\rangle}^e|$, as the quantitative measure, where $|t\rangle$ is the routing target state. We further fit $M_{\text{RAT}}(N) = l_1 + l_2 F_{\text{RAT}}^{2N+1}$ to extract the fidelity F_{RAT} , where l_1 and l_2 are fitting parameters. The RAT experiments are detailed in Appendix E. Ultimately, as

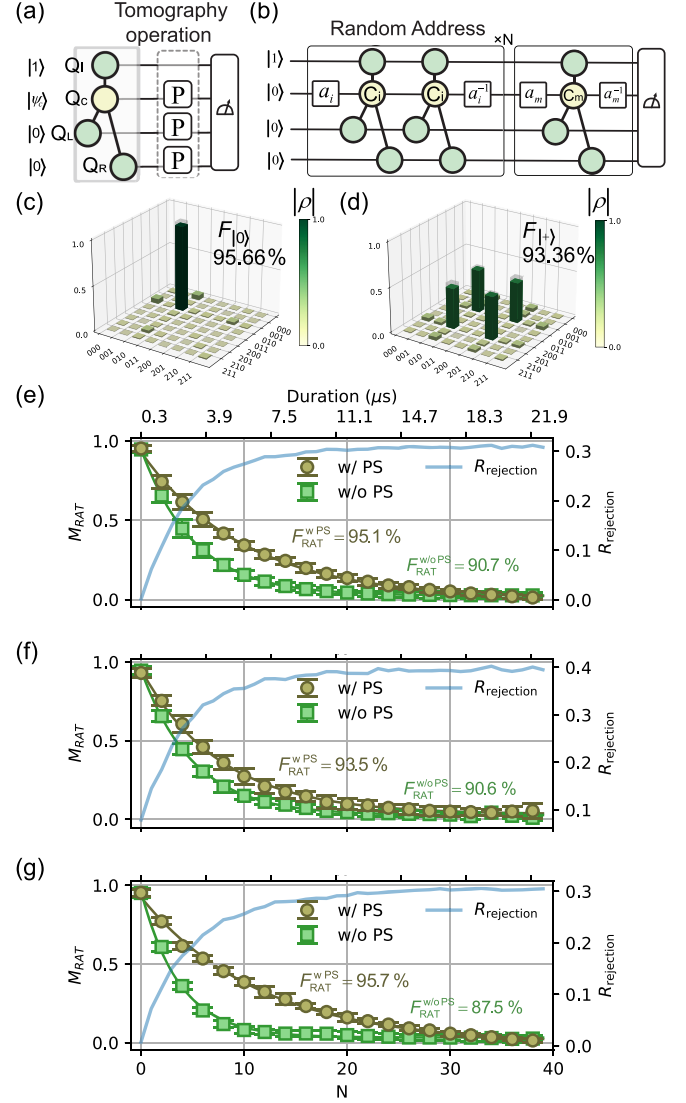


FIG. 3. QST and RAT of the QRouter. (a) Experimental circuit used for QST under different addresses. The tomography includes single-qubit basis rotations (P) and simultaneous readout. (b) Circuit structure of the RAT, consisting of $N + 1$ blocks, with the input qubit initially prepared in the $|1\rangle$ state. The first N blocks have random addresses a_i , with two QRouters sharing the same address to implement round-trip information transfer. The final block contains a single QRouter with a random address, transferring the information to data qubits. (c),(d) Reconstructed density matrices ρ from QST under addresses $|0\rangle$ and $|+\rangle = (|0\rangle + |2\rangle)/\sqrt{2}$, with fidelities 95.66% and 93.36%, respectively. (e)–(g) Extracted M_{RAT} values as a function of depth N (and corresponding duration) for three QRouters under erasure and nonerasure schemes. The blue curve represents the rejection rate ($R_{\text{rejection}}$) during the postselection. Each data point is the mean of 30 independent experiments. Within each experiment, 30 random trials are performed, with 5000 measurement shots per trial.

shown in Figs. 3(e)–3(g), the three QRouters exhibit fidelities of $95.1\% \pm 0.11\%$, $93.5\% \pm 0.15\%$, and $95.7\% \pm 0.04\%$ under the erasure scheme ($F_{\text{RAT}}^{\text{w/PS}}$, brown) with their respective

rejection rates $R_{\text{rejection}}$ (blue), compared to only 90.7% $\pm 0.28\%$, 90.6% $\pm 0.14\%$, and 87.5% $\pm 0.11\%$ under the nonerasure scheme ($F_{\text{RAT}}^{\text{w/o PS}}$, green). Within our experimental duration, $R_{\text{rejection}}$ for the single-layer router increases and then saturates as erasure-detectable errors accumulate. This steady-state plateau reflects a dynamic balance within the error-erasure QRouter sequences consisting of CSWAP and X gates. The continuous pumping of population into the $|2\rangle$ state by the alternating X gates, and its subsequent cascade into $|1\rangle$, is ultimately offset by the natural relaxation from $|1\rangle$ back to $|0\rangle$. Overall, the erasure protocol improves QRouter performance while keeping $R_{\text{rejection}}$ within an acceptable range. More details are provided in Appendix F.

The dominant error channels in our device are decoherence and leakage. These contributions are discussed in detail in Appendixes A 1, A 2, and G. Experimentally, these errors arise primarily from three technical factors: the absence of Purcell filtering, low-frequency noise, and short-timescale distortions of the $\sqrt{\text{CZ}}$ waveforms [61,64–67]. We expect that combining the erasure protocol with targeted hardware improvements and physically intuitive pulse optimization strategies, such as sudden net-zero (SNZ) shaping, will materially improve the QRouter's fidelity while preserving straightforward calibration [68,69]. The erasure protocol also shows

robustness to thermal errors and to unwanted SWAP-type processes in the qubit-qubit and qubit-coupler subspaces, as detailed in Appendix G. Therefore, this high-fidelity QRouter, endowed with independent routing capabilities, shows strong potential for constructing scalable routing networks.

IV. TWO-LAYER QUANTUM ROUTING NETWORK

Cascaded quantum routers form the foundational architecture of bucket-brigade QRAM, where conditional data routing is achieved through a binary-tree network of QRouter units [see Fig. 4(a)]. While we have demonstrated the performance of individual QRouters, scaling the system to multilayer networks introduces additional practical challenges. In particular, query efficiency (i.e., the circuit depth and timing required to resolve a path) and layout efficiency (the physical feasibility of embedding the network onto a quantum processor) must be addressed for scalable implementation [70].

To overcome these challenges, we adopt several strategies tailored to the constraints of planar superconducting hardware. For layout optimization, we employ a triangular tiling topology that preserves binary-tree connectivity while minimizing interroutier wiring and avoiding loops,

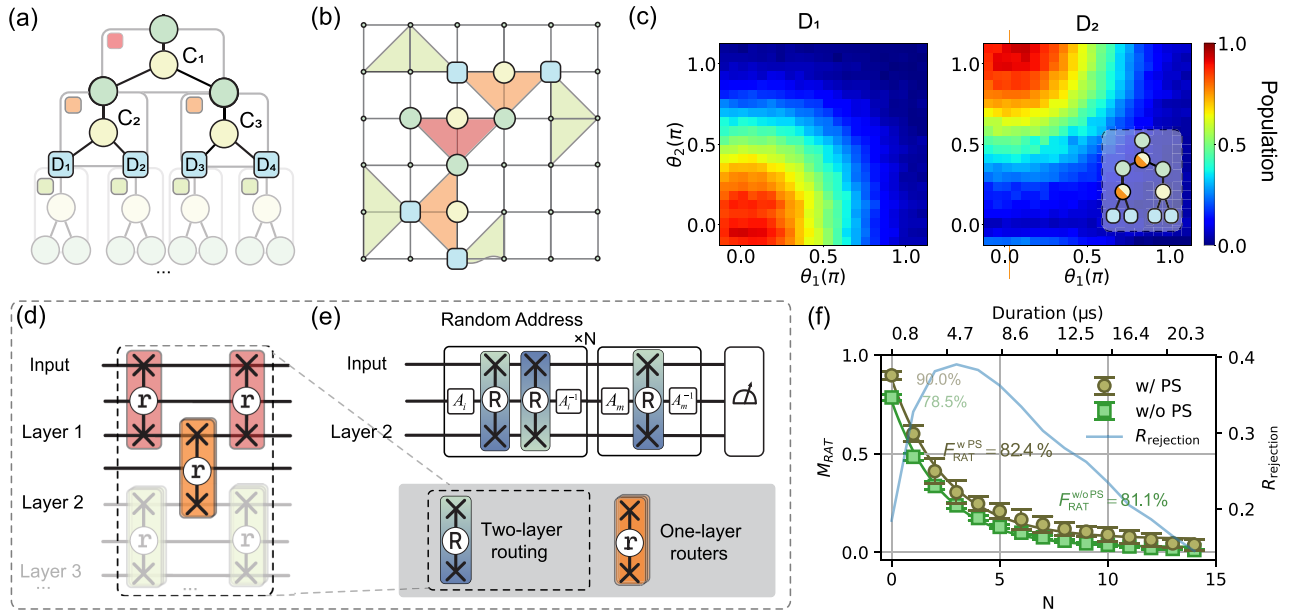


FIG. 4. Design and characterization of a two-layer quantum routing network. (a) Logical diagram of the binary-tree routing structure used in bucket-brigade QRAM. (b) Qubit layout on the superconducting processor, based on a triangular tiling scheme for optimized planar connectivity. (c) Measured populations of data qubits D_1 and D_2 as functions of address parameters (θ_1, θ_2) encoded in address qubits C_1 and C_2 , demonstrating controlled excitation transfer. (d) Simplified circuit structure enabling symmetric top-down and bottom-up routing, supporting layerwise parallel activation of QRouters. (e) RAT protocol for the two-layer network using randomized address sequences across blocks. (f) Extracted M_{RAT} values from RAT measurements as a function of routing depth N (and corresponding duration), for both erasure and nonerasure schemes. The blue curve represents the rejection rate ($R_{\text{rejection}}$) during the postselection. Each data point is the mean of 30 independent experiments. Within each experiment, 30 random trials are performed, with 5000 measurement shots per trial.

as shown in Fig. 4(b) and further discussed in Appendix B 3. To reduce query latency, we adopt a similar circuit structure inspired by Ref. [71], as illustrated in Fig. 4(d), with full circuit details provided in Appendix B 4. This circuit follows a symmetric “ $\vee$ ” structure to support both forward (top-down) and backward (bottom-up) routing—an essential feature for full QRAM operations. Under this architecture, layerwise parallel activation of QRouters becomes possible in both directions, allowing routing information to propagate through the network with minimal latency and maintaining temporal symmetry. By overlapping noninterfering gate executions across adjacent layers, this parallelized scheduling further compresses the overall circuit depth, substantially reducing decoherence overhead. We also present a simplified unidirectional variant that retains routing functionality while additionally compressing the circuit depth. Details are provided in Appendixes B 4 and B 5.

As a proof-of-concept demonstration, we implemented a two-layer routing network using the above strategies and evaluated its routing performance and fidelity. This setup represents the minimal nontrivial configuration for a cascaded QRAM-compatible system and serves as an essential validation of the network’s scalability.

First, we investigated address-path entanglement within the system. Using the setup illustrated in the inset in Fig. 4(c), we excited the top qubit and continuously varied the addresses (A_1^c, A_2^c) of C_1 and C_2 through θ adjustments in multiple routing tests, while monitoring the data qubits D_1 and D_2 . The experimental results in Fig. 4(c) depict the population distributions of D_1 and D_2 after routing. When $\theta_1 = 0$ and $\theta_2 = 0$, the population of D_1 reaches its maximum ($P_1^{\max} = 93.96\%$), consistent with the expected $\cos^2(\theta_1/2)\cos^2(\theta_2/2)$. A similar trend is observed for D_2 ($P_2^{\max} = 92.51\%$) when $\theta_1 = 0$ and $\theta_2 = \pi$, demonstrating the system’s effective routing capability for information transfer.

We further conducted random access tests to evaluate the system’s average fidelity. A randomized address sequence $\{(A_1, A_1), (A_2, A_2), \dots, (A_n, A_n), (A_{n+1})\}$ was loaded, where each A_k comprised three random addresses $\{a_1, a_2, a_3\}$. Each address sequence was initialized within the corresponding block and reset after routing, as shown in Fig. 4(e). The fidelity was determined by comparing the experimentally measured populations of the input qubits and four data qubits with the theoretically expected results, as illustrated in Fig. 4(f). The measured M_{RAT} values at $N = 0$ were 90.0% with the erasure scheme and 78.5% without. The performance gradually declined with increasing routing depth due to factors such as addresses’ two-photon transitions, imperfect operations, and data information decay. Notably, the system employing the erasure scheme consistently outperformed its nonerasure counterpart, with fitted average fidelities of $82.4\% \pm 0.91\%$ and $81.1\% \pm 0.30\%$, respectively. For the two-layer routing

network, $R_{\text{rejection}}$ exhibits a pronounced peak followed by a decrease. We suspect that coherent errors on the address qubit (perhaps due to crosstalk [72,73], known to be present in the scaled-up system) may be responsible for the sharper turnaround in the rejection rate in the two-layer case. These additional disturbances can break the original dynamic pumping balance, serving as a possible mechanism for this phenomenon. Concurrently, while decoherence and leakage remain the dominant error channels, these crosstalk mechanisms degrade the effectiveness of the postselection, resulting in the reduced incremental fidelity gain observed in the two-layer device. Mitigating such crosstalk is a central goal of our ongoing work. As detailed in Appendix G, our theoretical analysis and simulations confirm that the erasure advantage persists, underscoring its potential for scalable QRAM implementations.

V. ERROR BUDGET

To pinpoint the physical mechanisms limiting routing fidelity, we construct quantitative error budgets by integrating our experimental data with numerical simulations of the RAT. We evaluate the system performance across two distinct scales: the isolated component level and the cascaded two-layer network.

At the component level, the error decomposition in Fig. 5(a) reveals that environmental decoherence and coherent leakage errors dominate the baseline routing infidelities. Notably, in the standard nonerasure baseline, QRouter 3 exhibits a coherent error penalty that substantially exceeds the theoretical upper bound for combined decoherence and $\sqrt{\text{CZ}}$ gate leakage (“Max Est.”). We attribute this excess error to the destructive impact of

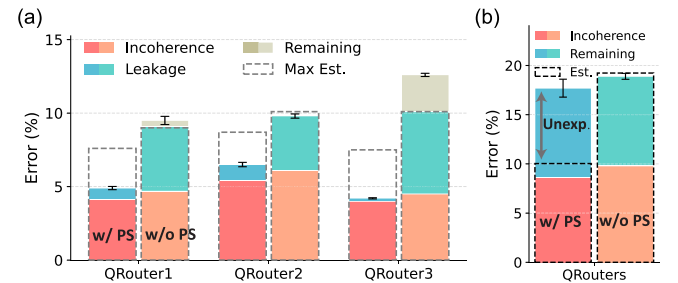


FIG. 5. Quantitative error budgets across component and system scales. (a) Component-level error decomposition for the three individual QRouters. The dashed boxes (“Max Est.”) represent the theoretical upper bound of errors contributed by environmental decoherence (coral) and the maximum calibrated $\sqrt{\text{CZ}}$ leakage (cyan). (b) System-level error budget for the integrated two-layer network. The dashed boxes (“Est.”) denote the theoretical error expectations derived from the independent multiplication of single-layer component fidelities. While the nonerasure baseline (“w/o PS”) strictly follows this independent accumulation model, the error-erasure protocol (“w/ PS”) exhibits an “unexpected” coherent error gap, quantifying the operational penalty introduced at scale (“Unexp.”).

uncharacterized parasitic SWAP-type interactions, which corrupt the computational address state in standard schemes. By contrast, the nonadjacent state encoding of our error-erasure protocol affords inherent structural robustness against this class of coherent errors, ensuring the total error remains within or near the anticipated boundaries, detailed in Appendix G.

Figure 5(b) presents the estimated error contributions for the expanded two-layer routing network. Given the complex interplay of coherent errors in a scaled system, we aggregate all participating coherent channels into a single “remaining” error metric.

By calculating the expected two-layer fidelity through the independent multiplication of single-layer component fidelities (dashed “Est.”), we explicitly expose the inherent operational overhead of active error mitigation. While the baseline nonerasure scheme aligns with its independent theoretical projection, the erasure protocol exhibits an “unexpected” performance gap. We evaluate the expected two-layer fidelity by multiplying the independent single-layer fidelities. The nonerasure scheme closely aligns with the resulting theoretical expectation (80.8%). Conversely, the erasure protocol exhibits a performance gap from its expected value (90.0%), revealing its scaling penalty. In this scaled configuration, the supplementary pulses required for intermediate address manipulation for error erasure may exacerbate classical and quantum crosstalk across the network, subsequently narrowing the practical gains offered by the erasure scheme. Consequently, while the protocol effectively shields against localized coherent SWAPs, fully unlocking the potential of scalable routing networks necessitates the hardware-level suppression of these underlying control deficiencies.

VI. DISCUSSION AND CONCLUSION

In this study, we demonstrated the implementation of hardware-efficient, deterministic, and complete-transmission QRouter on a superconducting quantum processor. Our work constitutes one of the first experimental realizations of a two-layer quantum routing network, enabling address-dependent, reversible routing with minimal hardware overhead. While previous efforts have largely remained at the theoretical level or focused on single-router demonstrations, our cascaded routing network showcases scalable, coherent routing compatible with grid-based superconducting hardware, marking a critical step toward practical QRAM architectures.

The coherent QRouter achieved reversible routing with a high transmission efficiency of 98%, and the two-layer network maintained an overall efficiency of 93%. In random access tests—a benchmark we introduce for scalable quantum routing systems—the one-layer and two-layer configurations achieved average fidelities of 94.8% and 82.4%, respectively. These results validate theoretical observations that shallow circuits and

postselection can effectively suppress errors and improve fidelity in noisy intermediate-scale quantum devices [74–76].

In terms of single-layer router performance, both our scheme and the implementations reported recently [37,48] utilized QST and achieved comparable fidelities. For more complex routing systems, Ref. [48] reports the query fidelity after executing a standard QRAM protocol. Given the scalability challenges and increasing measurement overhead associated with QST, our evaluation focuses instead on the average path performance of the routing network. Architecturally, Ref. [37] minimizes depth through two native C-ISWAP gates enabled by a strong state-dependent ZZ interaction, whereas Ref. [48] combines subspace-optimized CZ-based decompositions with teleportation-assisted connectivity to realize QRAM query. Our TCG approach constructs modular routers from two-qubit transition gates, reducing depth relative to conventional digital decompositions while providing intrinsic qutrit-based erasure detection without additional ancilla qubits. The demonstrated planar two-layer integration and scalable RAT characterization, thus, complement the native-gate and end-to-end QRAM approaches of Refs. [37,48].

Through detailed characterization and simulation, we identified decoherence and control-induced leakage as dominant sources of infidelity. These observations provide concrete targets for system-level optimization. Techniques such as postselection, high-coherence fabrication, and waveform optimization offer practical pathways to mitigate such effects without introducing additional quantum overhead. Furthermore, in the two-layer system, potential crosstalk issues additionally degrade the effectiveness of the postselection, thereby reducing the net gain of the error-erasure scheme. This highlights the need for further research into parallelized circuit design and corresponding crosstalk suppression techniques. Our recent research [77] quantifies the error-rate thresholds required for scalable QRAM and offers actionable optimization targets for implementing larger, practical systems.

As the routing network scales further, new challenges emerge, including QRouter layout overlap, increased crosstalk [72,78,79], and error accumulation across multiple routing paths. Addressing these issues will require advances in hardware layout, control sequencing, and short-range quantum interconnects [80–82]. We also anticipate that the current circuit design offers significant potential for further optimization to reduce the overall gate-depth overhead. We anticipate that synthesizing circuits at a higher structural level within the TCG framework, specifically through direct implementation using QRouter primitives, will yield more compact and higher-performance routing architectures [51]. Building upon these optimized router networks, future implementations of the full QRAM data-fetch pipeline can be directly constructed

by adapting highly efficient memory access protocols [71]. Nevertheless, our results establish both the architectural viability and the experimental feasibility of scalable QRAM-compatible routing and provide a practical blueprint for extending coherent quantum data access on near-term superconducting platforms.

ACKNOWLEDGMENTS

This work has been supported by the National Key Research and Development Program of China (Grant No. 2023YFB4502500) and the National Natural Science Foundation of China (Grant No. 12404564). This work is partially carried out at the University of Science and Technology of China Center for Micro and Nanoscale Research and Fabrication.

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX A: EXPERIMENTAL SETUP AND GATE CALIBRATIONS

1. Superconducting processor and qubit layout

We conducted experiments on a superconducting quantum processor named Wukong, comprising 72 transmon superconducting qubits (Q) and 126 transmon-type couplers (C) arranged in a gridlike topology. The topology of the chip is shown in Fig. 6.

We selected ten specific qubits as our experimental platform, because their spatial arrangement forms three connected triangular modules, satisfying the connectivity requirements for constructing a two-layer quantum routing network. Their basic parameters and decoherence performance are listed in Table II. As shown, the typical energy relaxation (T_1) and dephasing (T_2) times are on the order of ten to a few tens of microseconds. Because only $Q2$, $Q6$, and $Q8$ are utilized as control qutrits (address nodes) in our routing architecture, calibrations for the second excited state $|2\rangle$ (such as T_1^{21} and $T_2^{21,E}$) are exclusively performed for these three qubits. The corresponding entries for the remaining data and path qubits are, therefore, omitted.

The relatively short overall coherence times observed on this device arise from two primary factors. First, the absence of Purcell filtering increases the qubits' coupling to the environment, which accelerates energy relaxation and degrades coherence [65,66]. Second, we have observed that low-frequency noise in the control hardware further exacerbates decoherence.

2. Optimization of $\sqrt{CZ}$ gates via Floquet sequences

Serving as the primary component, the $\sqrt{CZ}$ gate enables the precise interference pathways essential for the quantum router demonstrated in this work. This gate relies on the full exchange between the $|11\rangle$ and $|02\rangle$ states, where the higher-frequency qubit occupies the latter. However, quantifying microscopic coherent errors, such as leakage into the noncomputational $|2\rangle$ state, is challenging with standard single-shot measurements. To

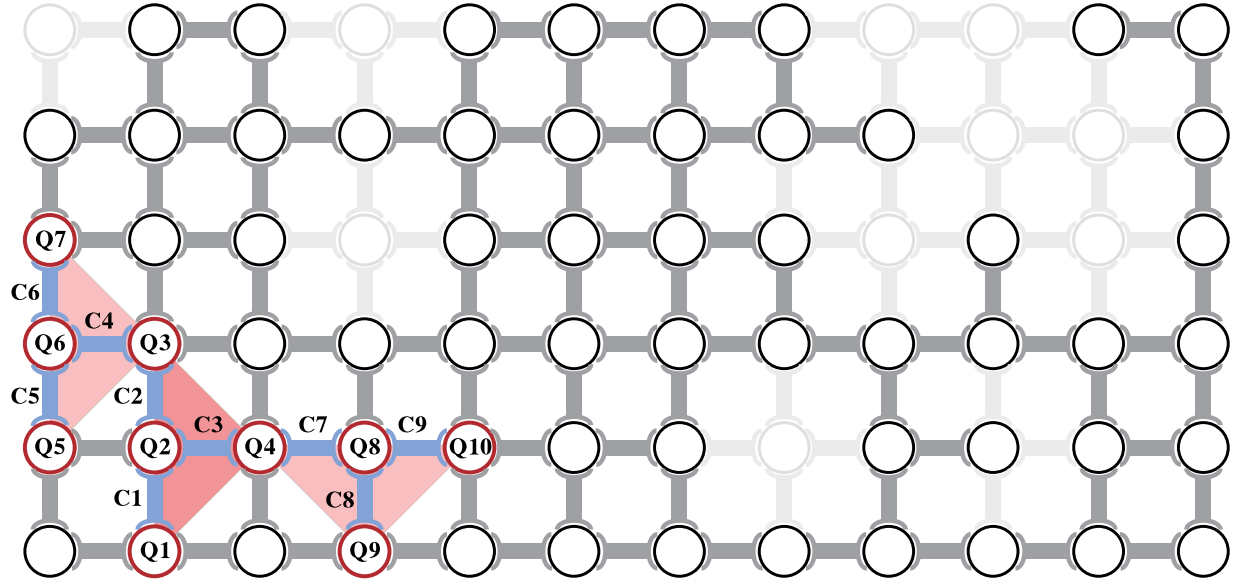


FIG. 6. The topology of the superconducting processor Wukong and the specific qubits used in the experiment. Solid dark circles and connections indicate well-performing qubits and couplers across the chip, while faded light ones correspond to malfunctioning or uncalibrated components. Ten qubits (highlighted with dark red outlines, $Q1$ – $Q10$) and their corresponding couplers (highlighted in blue, $C1$ – $C9$) were selected as our experimental platform. These are arranged into three connected triangles (shaded in light red), where each triangle functions as a modular quantum router.

TABLE II. Relevant qubit parameters and performances for the device under test. We list qubit transition frequencies from states $|i\rangle$ to $|j\rangle$ (f_{ij}) and anharmonicities α , while $f_{12} = f_{01} + \alpha$. We also give qubit coherence parameters (T_1^{10} , $T_2^{10,E}$, $T_2^{21,E}$, T_1^{21} , and Γ_{10}/Γ_{21}), single-qubit readout fidelities (F_0/F_1), and single-qubit gate fidelities (F_{sq}). T_1^{ij} is the relaxation time from state $|i\rangle$ to $|j\rangle$, and $T_2^{ij,E}$ is the T_2 time obtained by the spin-echo experiment between state $|i\rangle$ and $|j\rangle$. The decoherence ratio Γ_{10}/Γ_{21} is T_1^{21}/T_1^{10} . Note that parameters related to the $|2\rangle$ state are marked with $\dots$ for data and path qubits, as they are calibrated only for the designated control qubits ($Q2$, $Q6$, and $Q8$).

Parameters	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$	$Q9$	$Q10$
f_{01} (GHz)	4.269	4.622	4.163	4.238	4.308	4.587	4.107	4.401	4.315	4.130
α (MHz)	-246.6	-188.4	-242.4	-244.1	-236.7	-233.9	-243.5	-238.7	-248.3	-245.4
T_1^{10} (μ s)	22	19	20	22	17	24	8	17	16	23
$T_2^{10,E}$ (μ s)	9	11	8	9	8	10	7	8	7	1
T_1^{21} (μ s)	$\dots$	15	$\dots$	$\dots$	$\dots$	17	$\dots$	14	$\dots$	$\dots$
$T_2^{21,E}$ (μ s)	$\dots$	7	$\dots$	$\dots$	$\dots$	8	$\dots$	7	$\dots$	$\dots$
Γ_{10}/Γ_{21}	$\dots$	0.79	$\dots$	$\dots$	$\dots$	0.70	$\dots$	0.82	$\dots$	$\dots$
F_0 (%)	95.6	96.2	97.5	93.4	96.0	96.7	96.0	96.7	95.2	96.8
F_1 (%)	89.0	91.1	90.7	88.5	91.4	88.8	89.4	89.7	91.8	90.3
F_{sq} (%)	99.8	99.7	99.7	99.3	99.7	99.8	99.7	99.6	99.8	99.7

amplify these small errors into measurable population oscillations, we employ a repeating gate circuit known as a Floquet sequence. This section presents the theoretical analysis and experimental optimization of the $\sqrt{CZ}$ gate, along with relevant statistical data. We began by systematically characterizing this operation to determine suitable Floquet sequences and optimization targets [60–62]. The mathematical representation of the $i\text{SWAP}_{02\leftrightarrow 11}$ operation in the $\{|11\rangle, |02\rangle\}$ basis is given by

$$i\text{SWAP}_{02\leftrightarrow 11}(\vartheta, \eta) = \begin{pmatrix} e^{i\eta} \cos(\vartheta/2) & -i \sin(\vartheta/2) \\ -i \sin(\vartheta/2) & e^{-i\eta} \cos(\vartheta/2) \end{pmatrix}, \quad (\text{A1})$$

$$\sqrt{CZ} = i\text{SWAP}_{02\leftrightarrow 11}(\pi, \eta) = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}, \quad (\text{A2})$$

where ϑ and η denote the Rabi angle and the controlled phase acquired during the operation [68]. Setting $\vartheta = \pi$ realizes the ideal $\sqrt{CZ}$, which fully exchanges the populations of states $|11\rangle$ and $|02\rangle$.

Floquet theory can be used to analyze the dynamics of repeated $\sqrt{CZ}$ gates. To construct the Floquet sequence, we interleave the two-qubit gates with single-qubit phase rotations. The evolution under N applications of this composite gate $(Z_{qq} \cdot \sqrt{CZ} \cdot Z_{qq})^N$ is described by

$$(Z_{qq} \cdot \sqrt{CZ} \cdot Z_{qq})^N = e^{-iN\Omega\sigma_n} = I \cos N\Omega - i\sigma_n \sin N\Omega, \quad (\text{A3})$$

where

$$\Omega = \arccos[\cos(\vartheta/2) \cos(\zeta/2 - \eta)], \quad (\text{A4})$$

$$Z_{qq} = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\zeta/2} \end{pmatrix}, \quad (\text{A5})$$

$$\sigma_n = \begin{pmatrix} \cos \alpha & \sin \alpha \\ \sin \alpha & -\cos \alpha \end{pmatrix}, \quad (\text{A6})$$

$$\alpha = \arctan\left(\frac{\tan(\vartheta/2)}{\sin(\eta/2 - \zeta)}\right). \quad (\text{A7})$$

Here, Z_{qq} represents a Z-gate operation applied between the $\sqrt{CZ}$ gates. Interleaving these single-qubit phase rotations (ζ) allows us to tune the interference conditions within the sequence, isolating the specific dynamics of the $\{|11\rangle, |02\rangle\}$ subspace.

For the ideal case where $\vartheta = \pi$, we have

$$\Omega = \pi/2, \quad \alpha = \pi/2, \quad (\text{A8})$$

$$(Z_{qq} \cdot \sqrt{CZ} \cdot Z_{qq})^N = \begin{pmatrix} e^{i(\eta+\zeta/2)} \cos(N\pi/2) & -i \sin(N\pi/2) \\ -i \sin(N\pi/2) & e^{-i(\eta+\zeta/2)} \cos(N\pi/2) \end{pmatrix}. \quad (\text{A9})$$

As a result, the population of state $|11\rangle$ (P_{2i}^{11} , $i = 0, 1, 2, \dots$) reaches 1 after an even number of gates, while the population of state $|02\rangle$ (P_{2i+1}^{02}) reaches 1 after an odd number of gates. Under the application of $2m$ successive $\sqrt{CZ}$ gates, we define the average population of these two states as the cost function $C_I^m = \sum_{i=0}^{m-1} (P_{2i+1}^{02} + P_{2i}^{11})/2m$, which characterizes both leakage and decoherence. For ideal operations without relaxation or leakage, C_I^m evaluates to 100%.

In our experiment, we first adjusted the basic waveform parameters of the $\sqrt{CZ}$ gate using a peak-finding algorithm

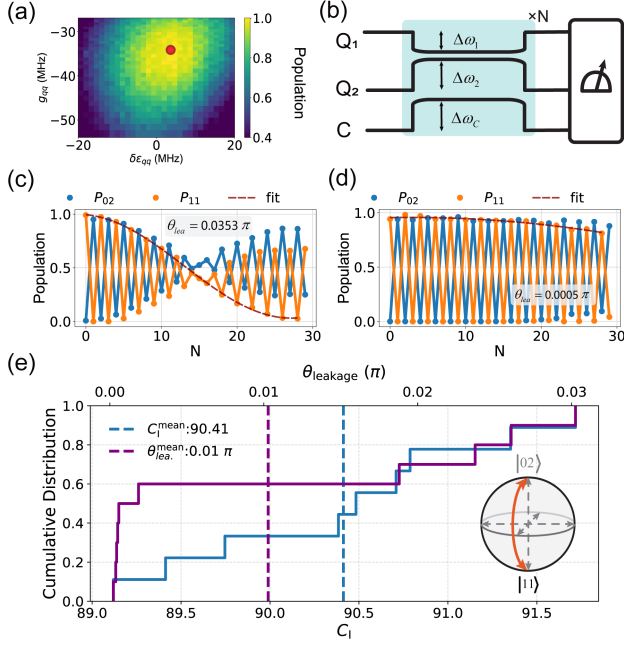


FIG. 7. Optimization of the $\sqrt{\text{CZ}}$ gate. (a) Landscape of the $|02\rangle$ state population (P_{02}) as a function of the energy detuning ($\delta\epsilon_{qq} = \epsilon_{|11\rangle} - \epsilon_{|02\rangle}$) and the effective coupling strength (g_{qq}) for a fixed operation time of 25 ns. (b) Schematic illustration of the N -repeat Floquet sequence setup used for optimization. (c),(d) State population evolution as a function of sequence depth N before (c) and after (d) Nelder-Mead optimization. The slow envelope modulation corresponding to the accumulated leakage error is visibly suppressed in (d). (e) Distribution of the optimization cost C_l^{15} (blue) and the extracted leakage angle θ_{lea} (purple) for the nine $\sqrt{\text{CZ}}$ gates calibrated in our experiment (5000 shots each). The ideal cost function value is 100%.

to achieve full population exchange between states $|11\rangle$ and $|02\rangle$ [see Fig. 7(a)]. Next, we employed an N -repeat gate circuit along with a Nelder-Mead (NM) algorithm to optimize the detailed waveform parameters in parallel [Fig. 7(b)]. Given the susceptibility of experimental environments to parameter fluctuations, the gradient-free NM method provides a robust approach for this multidimensional optimization. Specifically, we adjusted the frequency modulation amplitudes of the two interacting qubits and their shared coupler ($\Delta\omega_1, \Delta\omega_2, \Delta\omega_C$) under a fixed gate duration of 25 ns. The optimization objective was to maximize $C_l^{m=15} = \sum_{i=0}^{14} (P_{2i+1}^{(02)} + P_{2i}^{(11)})/30$, minimizing unwanted rotations and the influence of spurious two-level systems.

Figures 7(c) and 7(d) compare the gate performance before and after optimization. Since an ideal $\sqrt{\text{CZ}}$ gate corresponds to a π rotation in this subspace, any nonideal deviation (defined as the leakage angle θ_{lea}) accumulates with the sequence depth N . This phase accumulation manifests as a slow envelope modulation in the population

oscillations. We extract this leakage angle θ_{lea} by fitting the population data at odd or even depths to the function $a \sin[N(\pi - \theta_{\text{lea}}) + b] + c$, where a , b , and c are fitting parameters. After optimization, this slow oscillation envelope was suppressed, indicating a reduction in leakage. Similar optimization procedures were performed for all relevant qubit pairs, with the results summarized in Fig. 7(e). The average leakage angle was controlled to approximately 0.01π .

We observed that, even when using optimized parameters and a flattop Gaussian waveform, some $\sqrt{\text{CZ}}$ gates exhibited minor residual leakage. A technical source of this issue is waveform distortion, which prevents ideal non-adiabatic population transfer between states $|11\rangle$ and $|02\rangle$ within the 25 ns. In our system, such distortion originates from the arbitrary waveform generators, cryogenic wiring, and sample packaging. While digital predistortion techniques, such as finite impulse response or infinite impulse response filters [64,67,83–85], can partially mitigate these effects, they are often insufficient for complete suppression. Optimal control methods like gradient ascent pulse engineering [86–88] provide a theoretical pathway to minimize this leakage. However, their computational complexity and demanding calibration requirements pose practical challenges. Machine-learning-based offline physical models can reduce this computational overhead [89], yet a more hardware-efficient alternative is the SNZ waveform scheme [68,69]. By leveraging precise interference control, SNZ pulses ensure complete population exchange between the $|11\rangle$ and $|02\rangle$ energy levels, effectively mitigating the impact of imperfect evolutions caused by short-timescale distortions. Furthermore, SNZ schemes exhibit inherent robustness to long-timescale drifts and generalized waveform distortions. Integrating this physically driven waveform engineering strategy is expected to enable higher-fidelity two-qubit gates for future scalable routing architectures without sacrificing the simplicity of the calibration work flow.

APPENDIX B: CONSTRUCTION OF QUANTUM ROUTING SYSTEM

1. Controlled-SWAP based on the TCG scheme

Implementing the controlled operations necessary to construct a quantum router via standard Clifford decomposition typically incurs prohibitive circuit depths, making the system highly vulnerable to decoherence in near-term superconducting platforms. To address this overhead, this section outlines a hardware-efficient implementation scheme for the CSWAP operation [50,90]. In previous work, we proposed the TCG scheme, demonstrating that controlled operations can be implemented with reduced circuit depths by utilizing higher-energy levels as auxiliary states through transition pathway engineering [51]. Under this framework, the TCG-based CSWAP gate requires only three

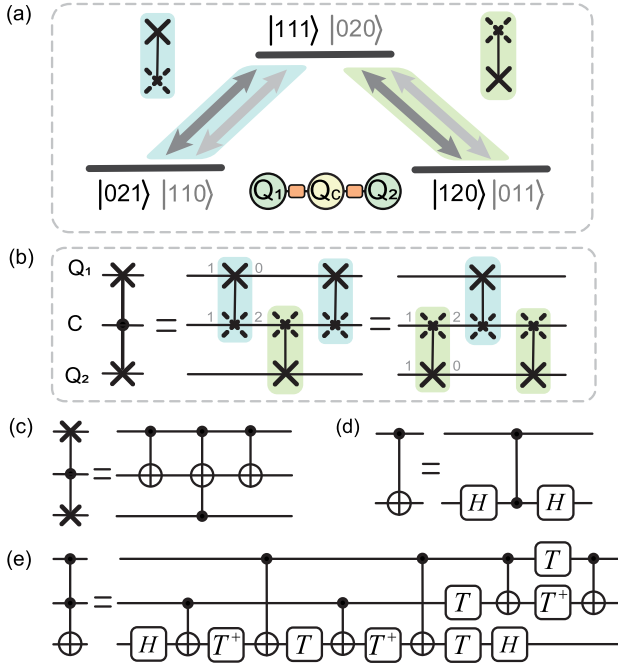


FIG. 8. CSWAP operation based on the TCG scheme. (a) Transition pathways of our CSWAP operation. Two parallel transition paths ($|110\rangle \leftrightarrow |020\rangle \leftrightarrow |011\rangle$, $|120\rangle \leftrightarrow |111\rangle \leftrightarrow |021\rangle$) enable CSWAP conditional on the control qutrit states ($|1\rangle, |2\rangle$). (b) Two equivalent composite implementations of the CSWAP operation, each requiring only three two-qubit gates. (c)–(e) Standard CSWAP construction based on Clifford decompositions, requiring nine single-qubit gates and eight two-qubit CNOT gates.

$\sqrt{CZ}$ gates, with the corresponding transition pathways shown in Fig. 8(a). Analysis of these transition pathways reveals that this operation exhibits multiple-flag switching capabilities. Functioning as a controlled switch, it remains in the off state (no SWAP operation) when the control unit is in state $|0\rangle$. Conversely, it conducts (activates the SWAP operation) when the control unit is in states $|1\rangle$ or $|2\rangle$. The addressing $\{|0\rangle, |1\rangle\}$ is a nonerasure scheme, while the $\{|0\rangle, |2\rangle\}$ one is an error-erasure scheme, as detailed in Appendix F.

This effective CSWAP operation can be described in the Hilbert subspace spanned by the basis states $\{|011\rangle, |020\rangle, |110\rangle, |111\rangle, |021\rangle, |120\rangle\}$:

$$\begin{aligned} \text{CSWAP}_{Q_1 Q_C Q_2} &= \sqrt{CZ}_{Q_1 Q_C} \otimes I_{Q_2} \cdot I_{Q_1} \otimes \sqrt{CZ}_{Q_C Q_2} \cdot \sqrt{CZ}_{Q_1 Q_C} \otimes I_{Q_2} \\ &\quad \quad \quad (B1) \end{aligned}$$

$$\begin{aligned} &= I_{Q_1} \otimes \sqrt{CZ}_{Q_C Q_2} \cdot \sqrt{CZ}_{Q_1 Q_C} \otimes I_{Q_2} \cdot I_{Q_1} \otimes \sqrt{CZ}_{Q_C Q_2} \\ &\quad \quad \quad (B2) \end{aligned}$$

$$= (-1) \cdot \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad (B3)$$

where

$$\sqrt{CZ}_{Q_1 Q_C} \otimes I_{Q_2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 & 0 & 0 \\ 0 & -i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -i & 0 \\ 0 & 0 & 0 & -i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad (B4)$$

$$I_{Q_1} \otimes \sqrt{CZ}_{Q_C Q_2} = \begin{pmatrix} 0 & -i & 0 & 0 & 0 & 0 \\ -i & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -i \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -i & 0 & 0 \end{pmatrix}, \quad (B5)$$

and I denotes the identity operation. Here, Q_1 and Q_2 are data qubits, and Q_C is the control qutrit.

The effective CSWAP operation can be implemented in two equivalent ways, each corresponding to a distinct yet symmetric transition path, as shown in Fig. 8(b). In both cases, the operation matrix is block diagonal, and the SWAP occurs within two decoupled subspaces of the Hilbert space. Compared to the conventional Clifford-based decomposition, which requires circuit depths up to 14 [Figs. 8(c)–8(e)], this scheme provides a significant reduction in operational complexity.

We acknowledge that similar work has been conducted recently in this field [91]. However, our analysis offers a distinct perspective through the TCG framework by proposing the utilization of higher energy levels as active switches. Furthermore, while the complete CSWAP described above is bidirectional, many quantum routing tasks, such as QRAM address loading and unloading, inherently require only unidirectional information transfer. To accommodate this, the short-path TCG (SPTCG) utilizes truncated transition pathways specifically designed for unidirectional transmission [51]. For instance, transferring information unidirectionally from Q_1 to Q_2 requires only two $\sqrt{CZ}$ gates, as detailed in Appendix B 5. This SPTCG variant retains the benefit of rapid initialization while minimizing circuit depth, thereby providing a

foundational building block for the scalable QRAM architecture.

2. Quantum router construction

The core function of the QRouter is to distribute quantum information from an input node to distinct spatial paths conditional on a quantum control state. This section details the structural design and construction of the QRouter module.

For simplicity, we analyze the nonerasure scheme ($|L\rangle = |0\rangle, |R\rangle = |1\rangle$) as a representative case. Structurally, the QRouter module is constructed by concatenating two CSWAP gates, interleaved with specific single-qubit operations, as shown in Fig. 13(a). The primary design objective is to ensure that the input qubit (Q_I) is exclusively routed to either the left path qubit (Q_L) or the right path qubit (Q_R) depending on the logical state of the control qubit (Q_C).

To enforce this mutually exclusive routing, a bit-flip operation (X gate) is applied to the control qubit between the two constituent CSWAP operations. The full logical routing operation, denoted as $U_{Q_I Q_C Q_L Q_R}$, is constructed as follows:

$$\begin{aligned} U_{Q_I Q_C Q_L Q_R} &= (I_{Q_I} \otimes X_{Q_C} \otimes I_{Q_L} \otimes I_{Q_R}) \cdot (\text{CSWAP}_{Q_I Q_C Q_L} \otimes I_{Q_R}) \\ &\quad \cdot (I_{Q_I} \otimes X_{Q_C} \otimes I_{Q_L} \otimes I_{Q_R}) \cdot (\text{CSWAP}_{Q_I Q_C Q_R} \otimes I_{Q_L}). \end{aligned} \quad (\text{B6})$$

The operational mechanism of this constructed circuit is deterministic: If Q_C is initially in the state $|0\rangle$, the first CSWAP gate remains inactive. The subsequent X gate flips Q_C to $|1\rangle$, thereby activating the second CSWAP gate and routing the input data to Q_L . Conversely, if Q_C is initially in the state $|1\rangle$, the first CSWAP gate activates, routing the data to Q_R . The following X gate then flips Q_C to $|0\rangle$, preventing the second CSWAP gate from executing. The final X gate in Eq. (B6) serves to restore the control qubit to its original logical state, preserving the control information for subsequent operations or uncomputation.

This hardware logic can be mathematically represented in the restricted Hilbert subspace spanned by the ordered computational basis $\{|0010\rangle, |1000\rangle, |0101\rangle, |1100\rangle\}$, where the basis corresponds to the state vector $|Q_I Q_C Q_L Q_R\rangle$. Within this basis, the constructed QRouter operation evaluates to the following unitary matrix:

$$U_{Q_I Q_C Q_L Q_R} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \quad (\text{B7})$$

This block-off-diagonal matrix confirms that the constructed sequence of gates effectively realizes the

conditional operations. Having established the functional logic and structural composition of a single QRouter module, scaling this architecture into a network necessitates a systematic strategy for physical implementation on a hardware processor.

3. Topological mapping of routing trees on 2D grids

With the construction of an individual QRouter established, scaling this architecture into a functional QRAM prototype requires mapping a cascaded binary-tree network onto a physical processor. As the system scales, the exponential growth in required quantum components presents a hardware-engineering challenge. Specifically, embedding the prototype's binary-tree architecture into a standard planar 2D grid topology imposes severe spatial constraints, particularly when accounting for unavoidable fabrication defects in specific qubits or couplers [70].

To address this topological constraint, we propose a densely packed triangular scheme. In this model, each fundamental QRouter is represented as an isosceles triangle comprising four physically adjacent qubits on the 2D grid. The overall QRAM network is then formed by a cascade of such triangles, effectively emulating the logical binary tree. To optimize qubit utilization while mitigating the impact of defective hardware components, the constituent triangles are packed in accordance with two geometric design principles: (i) Triangles must be connected at their vertices in pairs to form the routing tree structure, and (ii) the resulting network must not enclose any isolated regions (i.e., avoiding closed loops) to ensure that the QRouters do not physically overlap.

To implement this scheme, we developed a heuristic layout algorithm that initiates from a designated seed triangle [labeled S in Fig. 9(a)] and iteratively expands triangular links into topologically favorable regions. By computing all possible seed locations across the processor, we identified the most compact and defect-tolerant configurations. The specific layout utilized in our experiment, shown in Fig. 9(a), was selected based on an evaluation of regional qubit coherence, operational crosstalk, and prevailing hardware constraints.

Furthermore, we simulated the optimal packing geometries for an ideal Wukong processor with a complete 12×6 topology. The optimal configurations consistently place the seed triangle near the geometric center of the processor grid, indicating a radial expansion strategy for maximizing packing efficiency.

Structural analysis reveals that this densely packed scheme can accommodate up to a five-layer routing system on this grid without topological conflict. Scaling beyond this five-layer limit on a purely planar architecture introduces physical overlaps between separate QRouters, necessitating short-range quantum interconnects or dynamic SWAP networks to cross intermediate routing lines. This observation highlights the limits of planar embedding and

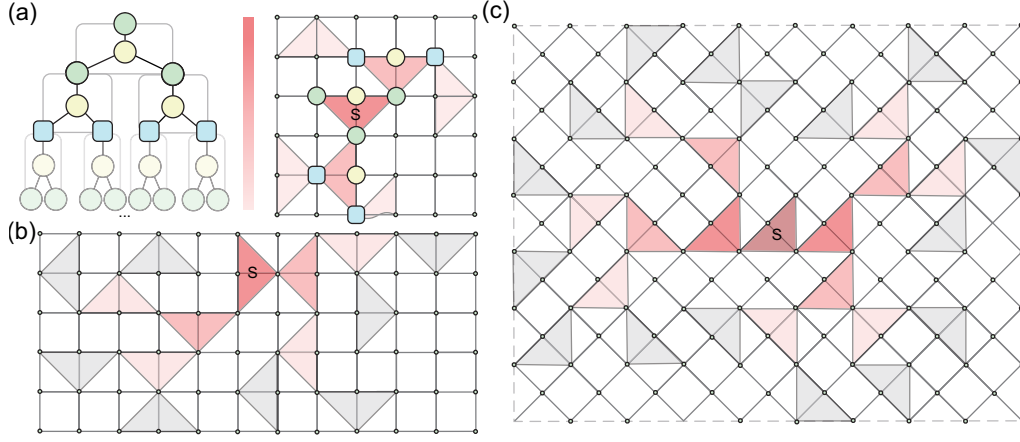


FIG. 9. Layout of the routing network for QRAM on a grid-based quantum processor. (a) QRouters are arranged in an isosceles triangular pattern within a regular 2D grid topology. By applying optimization constraints and our layout algorithm, we achieved an optimal arrangement on the physical processor. This illustrates the layout for a two-layer quantum routing network, where the color gradient from dark red to light red indicates ascending network layers. The label S denotes the “seed” triangle, which serves as the foundational starting point for our network layout algorithm. Within each QRouter, the central qubit (yellow) serves as the address node, while the peripheral qubits (green and blue) function as data nodes. To accommodate hardware defects and practical decoherence limitations, a portion of the ideal layout is truncated for the experimental implementation of the two-layer system. (b) An optimal simulated layout for a four-layer routing network on a defect-free 12×6 processor grid. (c) An example of the maximally dense five-layer system layout achievable on a defect-free processor.

the necessity of hardware-aware layout designs. Nonetheless, the proposed triangular packing scheme establishes a scalable topological framework for arranging QRAM hardware.

4. Circuit design and parallelization for QRAM

Translating the static physical layout of the cascaded network into a functional memory requires a synchronized circuit execution protocol. While QRAM represents the comprehensive memory application, its operational core is fundamentally driven by the quantum routing system. Because the complex memory access protocols are entirely orchestrated by the dynamic signal distribution of individual QRouter modules, our work deliberately emphasizes the design and optimization of this routing system. By defining the routing system as a network of independent modular components, the logical operations of a single QRouter naturally extend to the entire architecture, enabling consistent and parallel execution across the array.

The overall QRAM structure, depicted in Fig. 10(a), adopts a binary-tree architecture. The physical qubits are hierarchically divided into an input bus, address layers, and data layers. At the top of the tree resides the bus, which orchestrates both address and data transmission. Each gray box denotes a QRouter module, whose intermediate qubits serve as control nodes, thereby forming the address layer. The remaining qubits constitute the data layer, responsible for routing and temporarily storing quantum information. In practical architectures, the lowest-level data qubits are typically connected to classical memory resources to facilitate quantum-classical data transduction [3,5–7,50,92].

To address the query latency characteristic of deep cascaded networks, we designed the quantum circuits illustrated in Fig. 10(b). According to the circuit definitions, the underlying operations are classified into SWAP gates, unidirectional QRouters, and bidirectional QRouters (blue-green). The memory access protocol proceeds in three parallelized phases: (i) address setting (A_i) to establish coherent traversal routes; (ii) data fetch, where data inputs (D_i) propagate along the defined paths to interact with the memory via data copy operations (M_i); and (iii) uncomputing, where the sequence is coherently reversed through the same route to reset the routing network.

To maintain uniform query latency across all possible address paths, the SWAP and QRouter operations are applied in a fixed sequential order regardless of the queried address. Across the cascaded routing system, operations targeting nodes within the same architectural layer operate on disjoint sets of qubits and are, thus, executed simultaneously as unified layered operations. Furthermore, as successive address or data bits propagate through the network, operations across different layers overlap in execution time. This controlled combination of intralayer concurrency and pipelined parallelism is enabled by the localized structure of the digital QRouter components, thereby compressing the total circuit duration.

We also note that specific memory operations inherently involve unidirectional data flow. During the initial and final stages of the QRAM process (e.g., the address setting and uncomputing sequences), information propagates strictly forward or backward, invoking the unidirectional routing modules (green blocks in Fig. 10). As established in

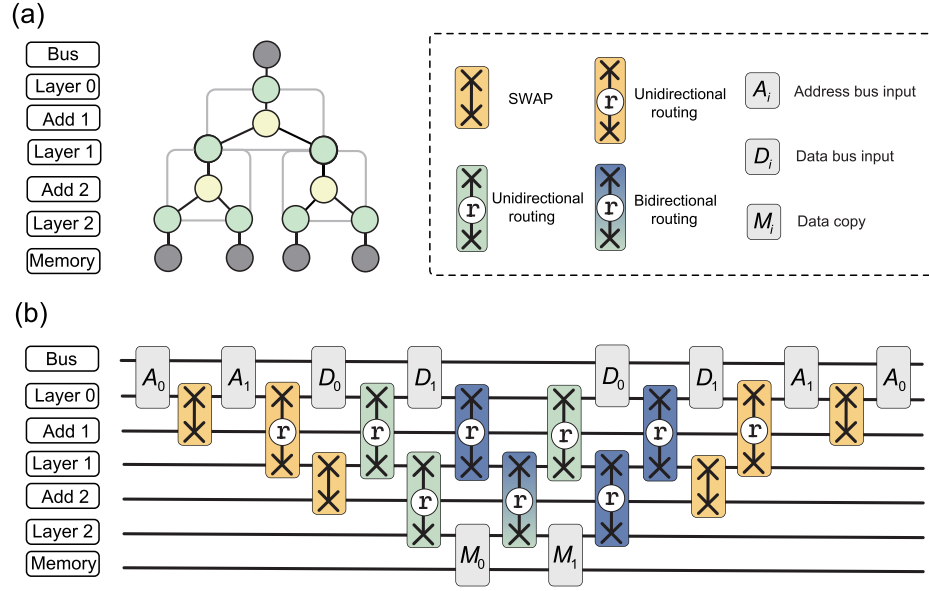


FIG. 10. Hierarchical structure and quantum circuits for bucket-brigade QRAM operations. (a) Schematic of the binary-tree architecture. Qubits are functionally categorized: green circles represent data qubits, yellow circles indicate control (address) qubits, and gray boxes define individual QRouter modules. The dashed legend box defines the specific circuit components: SWAP operations (orange), unidirectional routing modules (green), and bidirectional routing modules (blue-green). It also introduces the operational notations for address bus input (A_i), data bus input (D_i), and data copy (M_i). (b) Generalized quantum circuit for QRAM data access. The execution flow is orchestrated into three parallelized phases: address setting, data fetch (interacting with memory via M_i), and uncomputing via inverse operations to reset the network.

Appendix B 1, this unidirectional requirement allows the substitution of the full QRouter with the optimized SPTCG module. Furthermore, under the TCG framework, the constituent SWAP gates can be decomposed into a combination of a single $\sqrt{CZ}$ gate and local single-qubit rotations, replacing the standard three-CNOT implementation. This reduction in operational overhead directly mitigates cumulative gate errors, increasing the viability of executing complex routing algorithms on current generation processors [51].

To physically implement the routing network with minimal decoherence overhead, we employ a highly parallelized gate scheduling strategy. Figure 11 illustrates this optimization for a two-layer routing system. While the abstract logical circuit [Fig. 11(b)] implies a sequential cascade of routing operations, the underlying gate-level executions can be significantly overlapped. As depicted in the topological layout [Fig. 11(a)] and the corresponding gate sequence [Fig. 11(c)], the operations of the second-layer QRouters (green and yellow blocks) do not strictly require the complete termination of the first-layer QRouter (red block). Instead, once the initial excitation transfer to the intermediate nodes (Q_3 and Q_4) is achieved, the subsequent routers can immediately initiate their local entangling sequences in a pipelined fashion. A naive, fully serialized execution of this three-node network would necessitate a total two-qubit gate depth of 24 (comprising eight layers per individual QRouter). By actively overlapping noninterfering gate layers, our parallel-optimized circuit compresses the total depth to 18. This reduction in

circuit depth is highly advantageous for near-term superconducting processors, as it restricts the temporal window for environmental decoherence and directly mitigates the accumulation of coherent gate errors across the expanded QRAM architecture.

In summary, the integration of synchronized circuit parallelism, TCG-optimized operations, and hardware-aware triangular layouts establishes a comprehensive foundation for the quantum routing system. This holistic codesign approach significantly reduces the physical depth and operational latency of the routing circuits. Ultimately, by providing a complete bottom-up blueprint that spans from fundamental gate-level optimizations to system-level execution, this work demonstrates a practical pathway for scaling complex architectures such as QRAM on near-term superconducting platforms.

5. Unidirectional routing via short-path TCG

While the preceding sections focused on the implementation of fully bidirectional quantum routers, practical QRAM architectures often feature distinct data-loading (writing) and data-retrieval (reading) phases. Because these specific processes strictly require only unidirectional information transfer, employing a full, symmetric routing protocol introduces unnecessary operational overhead. To optimize for these task-specific deployments, this section presents the design of unidirectional quantum routers implemented via the SPTCG scheme [51]. The SPTCG

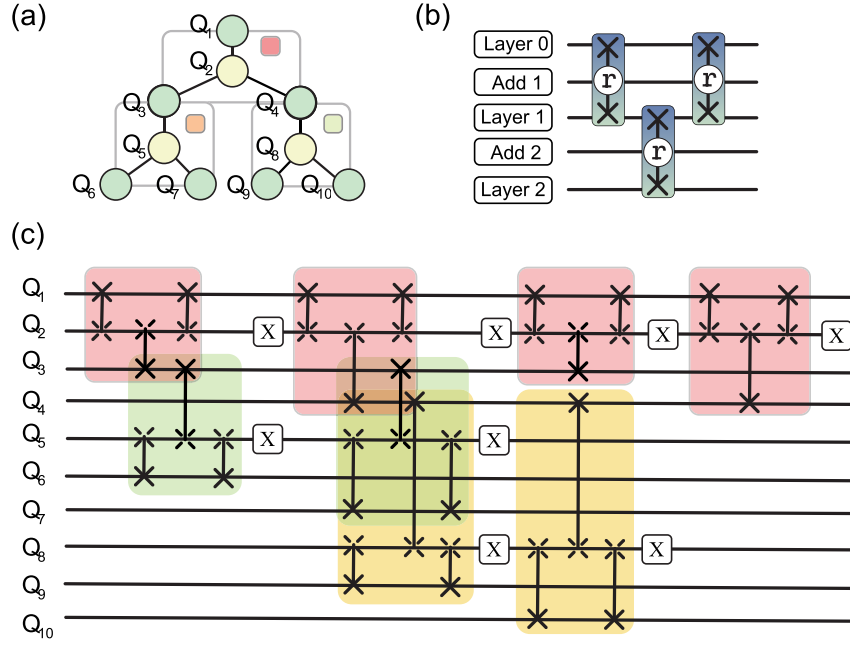


FIG. 11. Gate-level parallelization of the two-layer quantum routing circuit. (a) Topological layout of the two-layer routing system, with individual QRouter modules highlighted by distinct color blocks. (b) Logical circuit representation of the generalized two-layer routing network. (c) Parallel-optimized gate sequence for the circuit depicted in (b). By scheduling the execution of the second-layer QRouters (green and yellow blocks) to overlap with the latter stages of the first-layer QRouter (red block), the total gate depth is compressed from 24 to 18.

scheme deliberately employs a truncated, asymmetric evolution process to achieve directed state transitions. By utilizing only half of the available transition pathways compared to the full TCG framework, this approach significantly reduces the required quantum resources and simplifies the underlying circuit implementation.

To physicalize this concept, we introduce the short-path controlled-SWAP (SP-CSWAP) operation, as illustrated in Fig. 12. This operation breaks the symmetry of the full CSWAP gate to dictate a unidirectional excitation transfer (unidirectional arrow), as shown in Fig. 12(a). For instance, the abbreviated sequence shown in Fig. 12(b) utilizing the $\{|0\rangle, |1\rangle\}$ control basis enforces the cascaded evolution $|110\rangle \rightarrow |020\rangle \rightarrow |011\rangle$, routing quantum data from Q_1 to Q_2 . For data retrieval (e.g., from Q_2 to Q_1), one simply applies the spatially reversed pulse sequence in Fig. 12(c). Note that, while this shorter-depth scheme implements unidirectional routing, it trades off compatibility with our qutrit-based error-erasure protocol (which relies on the $|2\rangle$ address encoding).

This inherent directional duality naturally serves as the foundational building block for constructing two specialized memory variants: quantum write-only memory (QWOM) and quantum read-only memory (QROM) [3]. As depicted in Figs. 12(c) and 12(d), QWOM relies on one operational basis to direct data strictly from the bus into the memory register, whereas QROM utilizes the inverse basis to extract data from the memory back to the bus. By replacing the generalized bidirectional routers with these targeted SP-CSWAP modules, the required circuit depth is

reduced to approximately 67% of the original implementation. This substantial compression in circuit depth mitigates cumulative gate errors and decoherence, making the streamlined SPTCG architecture a highly scalable and compelling candidate for near-term specialized quantum memory applications.

APPENDIX C: EXPERIMENTAL VERIFICATION OF ROUTING FUNCTIONALITY IN THE QUANTUM ROUTING SYSTEM

1. Deterministic routing and address-path entanglement

The primary objective of a quantum router is to direct incoming quantum information to designated spatial paths dictated by a control state. Establishing a fully functional QRouter requires demonstrating two core functionalities. First, the device must act as a deterministic switch, accurately routing data to a specific path when provided with a definite, classical-like address state. Second, and crucially, it must preserve quantum coherence when the address is prepared in a superposition. Under such conditions, the router does not collapse the input onto a single trajectory. Rather, it coherently distributes the data across multiple paths, generating entanglement between the address and the physical spatial paths. This address-path entanglement serves as the definitive signature of a genuine quantum routing process.

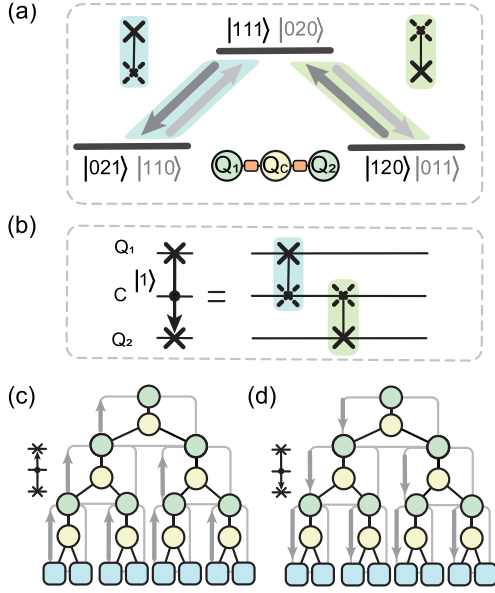


FIG. 12. Unidirectional quantum routing based on the SPTCG scheme. (a) Transition pathways of the SPTCG scheme for the short-path CSWAP (SP-CSWAP) operation. Unlike the full CSWAP gate, the state evolution in the SP-CSWAP operation is asymmetric and unidirectional (e.g., $|110\rangle \rightarrow |020\rangle \rightarrow |011\rangle$). (b) Circuit implementation of the SP-CSWAP using the control basis $\{|0\rangle, |1\rangle\}$. (c),(d) Schematic representations of quantum write-only memory (QWOM) and quantum read-only memory (QROM), respectively. By employing the complementary operations detailed in (b) and (c), the QRouter facilitates targeted, unidirectional data transfer for these specific memory applications.

We evaluate these dual capabilities in our hardware. Within this framework, we decouple the validation process into two distinct experimental phases. This decoupling is achieved by parametrizing the control qutrit's initial state with two degrees of freedom: $|\psi_c\rangle = \cos(\theta/2)|L\rangle + e^{i\phi} \sin(\theta/2)|R\rangle$. Here, $|L\rangle = |0\rangle$ denotes the inactive control state, while $|R\rangle$ represents the active control flag, which can be flexibly chosen as either $|1\rangle$ or $|2\rangle$ due to the multiflag switching capability of our TCG scheme. Specifically, varying the amplitude parameter θ allows us to assess the deterministic routing distribution, while tuning the phase parameter ϕ enables us to verify the preservation of quantum coherence and the resulting entanglement.

First, we verify the routing governed by θ . We prepare the input qubit (Q_I) in the excited state $|1\rangle$, forming the initial combined state $|\Psi_{\text{in}}\rangle = |1\rangle_{Q_I} \otimes |\psi_c\rangle_{Q_C} \otimes |00\rangle_{Q_L Q_R}$. Applying the routing operator $U_{Q_I Q_C Q_L Q_R}$ (derived in Appendix B) to this state induces the spatial distribution. Expressed as a column vector in the relevant operational basis $\{|0L10\rangle, |1L00\rangle, |0R01\rangle, |1R00\rangle\}$, the initial state S and the final routed state S' are, respectively,

$$S = e^{i\tau}(0, \cos(\theta/2), 0, e^{i\phi} \sin(\theta/2))^T, \quad (\text{C1})$$

$$U_{Q_I Q_C Q_L Q_R} S = -e^{i\tau}(\cos(\theta/2), 0, e^{i\phi} \sin(\theta/2), 0)^T, \quad (\text{C2})$$

where τ is an adjustable global phase. From $U_{Q_I Q_C Q_L Q_R} S$, we can extract the expected theoretical populations for the respective routing paths, denoted as P_L and P_R :

$$P_L = P_{|0L10\rangle} = \cos^2(\theta/2), \quad (\text{C3})$$

$$P_R = P_{|0R01\rangle} = \sin^2(\theta/2). \quad (\text{C4})$$

As shown in Fig. 13(b), the measured populations P_L and P_R closely follow this expected θ dependence. Although this agreement validates the device's capability to act as a tunable population distributor, identical statistics could be produced by a classical probabilistic switch.

To definitively confirm the genuinely quantum nature of this routing process, the second step is to verify the phase relationship governed by ϕ . We initialize the control qutrit in an equal superposition ($\theta = \pi/2$), resulting in $|\Psi_{\text{in}}\rangle = (|1L00\rangle_{Q_I Q_C Q_L Q_R} + e^{i\phi}|1R00\rangle_{Q_I Q_C Q_L Q_R})/\sqrt{2}$. After the routing operation, the input qubit Q_I deterministically returns to the $|0\rangle$ state and entirely disentangles from the rest of the system. The remaining three qubits (Q_C , Q_L , and Q_R) are left in a tripartite entangled Greenberger-Horne-Zeilinger-like cat state: $(|L10\rangle + e^{i\phi}|R01\rangle)/\sqrt{2}$. In this state, the relative quantum phase ϕ is nonlocally stored across the three-qubit system. Represented in the reduced three-qubit basis $\{|L00\rangle, |L01\rangle, \dots, |R11\rangle\}_{Q_C Q_L Q_R}$, this intermediate state S'_+ is

$$S'_+ = \frac{1}{\sqrt{2}}(0 \ 0 \ 1 \ 0 \ 0 \ e^{i\phi} \ 0 \ 0)^T. \quad (\text{C5})$$

Because the relative phase ϕ cannot be extracted through direct population measurements, we map this phase information onto population observables via a parity interference measurement. By applying global $X_{\pi/2}$ rotations to the three remaining qubits ($X_{\pi/2}^{\otimes 3}$), we force the separate components of the cat state to interfere. This yields the final state S''_+ :

$$S''_+ = X_{\pi/2}^{\otimes 3} \cdot S'_+ = \frac{1}{4} \begin{pmatrix} -i - e^{i\phi} \\ -1 - ie^{i\phi} \\ 1 + ie^{i\phi} \\ -i - e^{i\phi} \\ -1 - ie^{i\phi} \\ i + e^{i\phi} \\ -i - e^{i\phi} \\ -1 - ie^{i\phi} \end{pmatrix}. \quad (\text{C6})$$

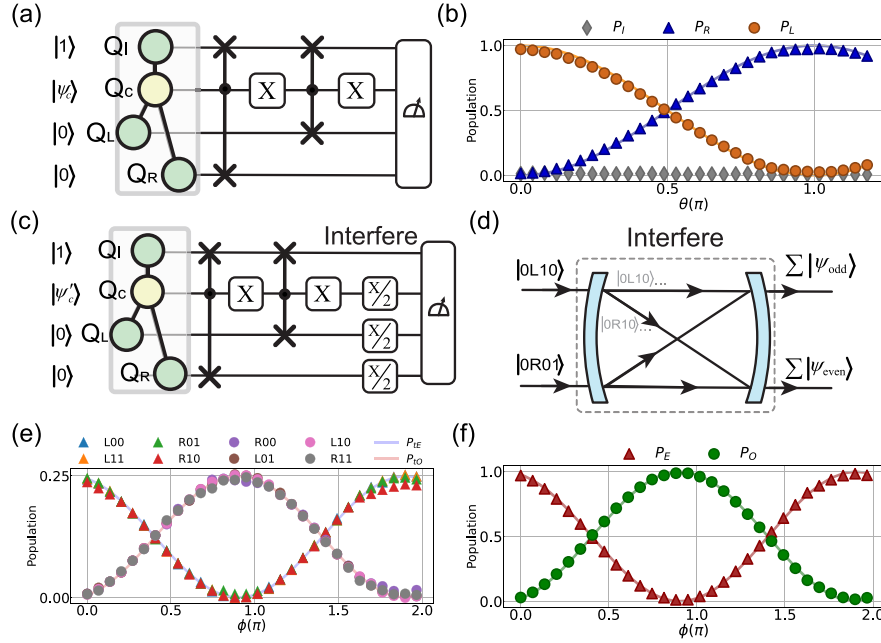


FIG. 13. Experimental verification of address-path entanglement in the QRouter. (a),(c) Experimental pulse sequences for characterizing the amplitude (θ) and phase (ϕ) dependence of the routing operations, respectively. (b) Measured routing populations as a function of θ . The deterministic routing ratio follows the theoretical probability governed by θ while remaining independent of the relative phase ϕ . (d) Schematic of the phase-interference protocol. Following the routing operation, the input qubit disentangles (returning to the $|0\rangle$ state), leaving the remaining subsystems in a three-qubit entangled cat state $(|L10\rangle + e^{i\phi}|R01\rangle)/\sqrt{2}$. Applying global $X_{\pi/2}$ rotations maps the relative phase ϕ onto the excitation parity of the final state. (e) Interference patterns obtained by sweeping the address phase ϕ . The measured populations of individual logically odd- and even-parity states oscillate in agreement with the theoretical $[1 \mp \sin(\phi + \delta\phi)]/8$ distributions, incorporating the phase offset $\delta\phi$. (f) Aggregated populations for the odd-parity (P_O) and even-parity (P_E) subspaces extracted from (e), demonstrating complementary oscillations that confirm coherent address-path entanglement.

The resulting probability $P = |S'_+|^2$ becomes

$$P = \frac{1}{8} \begin{pmatrix} 1 + \sin \phi \\ 1 - \sin \phi \\ 1 - \sin \phi \\ 1 + \sin \phi \\ 1 - \sin \phi \\ 1 + \sin \phi \\ 1 + \sin \phi \\ 1 - \sin \phi \end{pmatrix}. \quad (C7)$$

We categorize these basis states by their logical excitation parity. By mapping the inactive state $|L\rangle$ to a logical 0 and the active flag $|R\rangle$ to a logical 1, the total population of logically odd-parity states (i.e., $|L01\rangle$, $|L10\rangle$, $|R00\rangle$, and $|R11\rangle$) is denoted as P_O , which evaluates to $[1 - \sin(\phi + \delta\phi)]/2$. Conversely, the sum of logically even-parity populations (P_E) evaluates to $[1 + \sin(\phi + \delta\phi)]/2$. Here, $\delta\phi$ represents a deterministic single-qubit phase offset that is compensated using virtual

Z gates [93]. As depicted in Figs. 13(e) and 13(f), the measured odd- and even-parity populations exhibit the expected high-contrast complementary oscillations. This interference confirms the successful generation of coherent address-path entanglement, establishing the module as a genuine quantum router.

2. Cascaded routing and system-level verification

While the preceding section validates the generation of address-path entanglement within a single QRouter, practical QRAM architectures require quantum information to cascade through multiple routing nodes. To demonstrate the scalability and addressing accuracy of the physical layout, we performed system-level routing verification on our two-layer quantum routing network. Figure 14 illustrates the experimental configurations and the corresponding results.

In the first set of experiments, we initialized the root input qubit to the excited state $|1\rangle$ and validated the sequential addressability of the network by continuously varying the control addresses (A_1^c and A_2^c) of nodes C_1 and C_2 . This is achieved by adjusting the amplitude parameters θ_1 and θ_2 and subsequently measuring the excitation transfer to the target data qubits D_1 and D_2 . The resulting

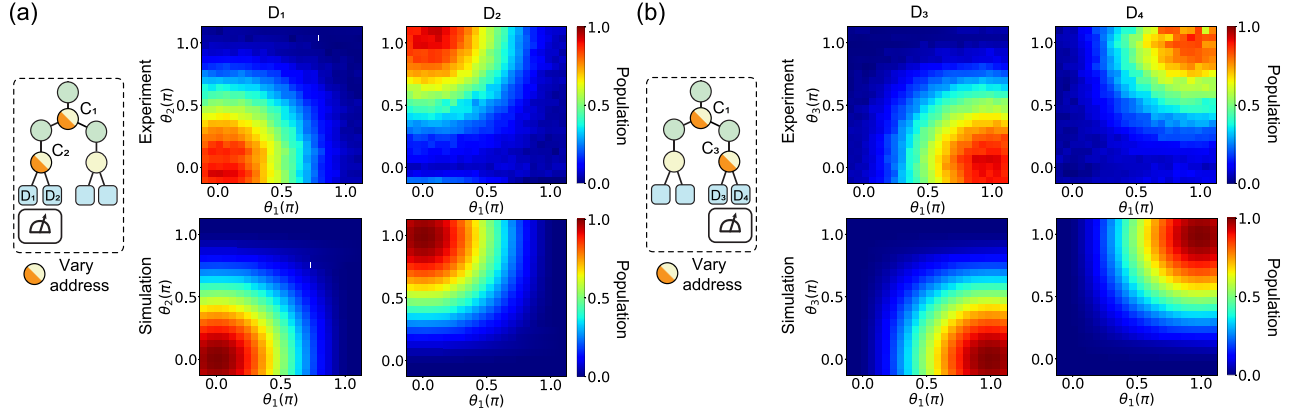


FIG. 14. Experimental verification of two-layer system routing performance. (a),(b) Characterization of the routing behavior for two independent network branches, targeting the data qubit pairs $\{D_1, D_2\}$ and $\{D_3, D_4\}$, respectively. The color maps compare the experimentally measured target-qubit populations (top rows) with numerical simulations (bottom rows) as a function of the cascaded address-control parameters. Specifically, (a) maps the population against θ_1 and θ_2 , while (b) maps it against θ_1 and θ_3 . The experimental data demonstrate strong agreement with the theoretical transmission amplitudes across both hardware branches.

population landscapes, shown in Fig. 14(a), reveal that, when $\theta_1 = 0$ and $\theta_2 = 0$, the population of D_1 reaches a maximum of $P_1^{\max} = 93.96\%$. This observation aligns with the theoretical $\cos^2(\theta_1/2)\cos^2(\theta_2/2)$ dependence. Similarly, for $\theta_1 = 0$ and $\theta_2 = \pi$, we observe a maximum population of $P_2^{\max} = 92.51\%$ in D_2 , corresponding to the $\cos^2(\theta_1/2)\sin^2(\theta_2/2)$ transmission amplitude. These results confirm reliable and selective quantum routing behavior across successive network layers, a conclusion further corroborated by matching numerical simulations.

A second set of experiments assessed an independent routing branch targeting path qubits D_3 and D_4 , as shown in Fig. 14(b). Here, the initial excitation is directed to the right network branch by setting $\theta_1 = \pi$. By subsequently varying θ_3 of node C_3 , we observe that the population of D_3 reaches a maximum of $P_3^{\max} = 92.23\%$ at $\theta_3 = 0$, which is consistent with the theoretical $\sin^2(\theta_1/2)\cos^2(\theta_3/2)$ distribution. Correspondingly, for $\theta_3 = \pi$, the population of D_4 maximizes at $P_4^{\max} = 88.38\%$, aligning with the $\sin^2(\theta_1/2)\sin^2(\theta_3/2)$. The slightly reduced excitation transfer observed for D_4 is a direct consequence of the limited coherence properties of its designated physical carrier (qubit Q_7), which exhibits a comparatively short relaxation time of $T_1 = 8 \mu\text{s}$.

In summary, the successful execution of these independent routing branches validates the structural viability of the cascaded architecture. Despite the inevitable decoherence inherent in near-term superconducting processors, the cascaded network maintains high routing fidelity and deterministic addressing accuracy. These system-level results confirm the functional integrity of the architecture under a densely packed triangular layout, thereby establishing the experimental foundation required to deploy the complex, highly parallelized QRAM circuits.

APPENDIX D: TOMOGRAPHIC CHARACTERIZATION

To evaluate the performance of our quantum routing architecture, we begin in this section by employing traditional QST to reconstruct the exact output states of a single QRouter. By combining experimental QST with extended numerical simulations, we uncover a state- and path-dependent decoherence mechanism inherent to the QRouter. Because characterizing every possible routing path via full QST requires an exponentially growing number of measurements and becomes practically unscalable for larger networks, these findings directly motivate the need for a holistic, scalable benchmarking protocol. This unified metric, which we introduce as the RAT, is detailed in the subsequent section.

1. QST for QRouters

To verify the coherent generation of the routed states at the fundamental hardware level, we first performed QST on a single QRouter under various address configurations. However, because raw tomographic data in superconducting processors are highly susceptible to readout classification errors, extracting an accurate physical density matrix requires a rigorous data processing pipeline.

Before evaluating the experimental fidelities, we adopted the iterative Bayesian unfolding (IBU) method [94] to correct these measurement errors. This approach overcomes the inherent limitations of conventional matrix-inversion techniques, such as the generation of unphysical negative probability entries and the amplification of statistical uncertainties.

For our n -qubit system, we define a $2^n \times 2^n$ response matrix R that maps the ideal probability distribution of the prepared states to the experimentally observed distribution.

Here, $R_{ij} = \Pr(\text{meas} = i | \text{prep} = j)$ represents the conditional probability of obtaining outcome i given that state j was prepared. The matrix R is constructed through exhaustive single-shot calibration experiments on the target qubits, under the assumption of negligible readout crosstalk. Given the experimentally measured noisy probability distribution w , the IBU algorithm iteratively estimates the corrected distribution v as

$$v_j^{(t+1)} = \sum_{i=0}^{2^n-1} w_i \frac{R_{ij} v_j^{(t)}}{\sum_{k=0}^{2^n-1} R_{ik} v_k^{(t)}}, \quad (\text{D1})$$

where t is the iteration index and $v^{(0)}$ is the initial guess, typically chosen as a uniform distribution. In our implementation, the number of IBU iterations is capped at a maximum of 50.

Following the IBU readout correction, we utilized a maximum-likelihood estimation procedure to extract the closest positive semidefinite, trace-one physical density matrix [95]. The final QST fidelity is then computed as $F_{\text{QST}} = \langle \Psi | \rho | \Psi \rangle$, where $|\Psi\rangle$ denotes the ideal target pure state and ρ represents the experimentally reconstructed density matrix.

With this rigorous pipeline established, we characterized the QRouter performance both with and without the application of the erasure-detection scheme, as illustrated in Fig. 15. We performed three-qubit QST on the subsystem Q_C, Q_L, Q_R , and experimental results indicate that erasure-detection improves the overall QST fidelity by approximately 3%–5%, though further hardware-level enhancement remains possible. The residual errors primarily stem from unavoidable environmental decoherence, imperfect initial state preparation, and residual leakage during the $\sqrt{CZ}$ operations.

2. QST simulations and state-dependent performance

While experimental QST provides an exact snapshot of a single routing operation, evaluating the architecture's scalability requires understanding how errors accumulate over deep cascaded circuits across diverse input states. To investigate this, we performed full QST-based numerical simulations to assess the routing performance across a representative set of classical and quantum input states.

As detailed in Table III, these simulations focus on the nonerasure protocol and rigidly follow the experimental QST procedure. Setting the baseline decoherence parameters to $T_1 = 20 \mu\text{s}$ and $T_2 = T_1/2$, the resulting output density matrices following a single-layer QRouter operation are presented in Figs. 16(a)–16(f). These single-layer simulations demonstrate coherent routing while revealing an inherent state-dependent sensitivity to decoherence: Input states with fewer physical excitations or less superposition content consistently yield higher QST fidelities.

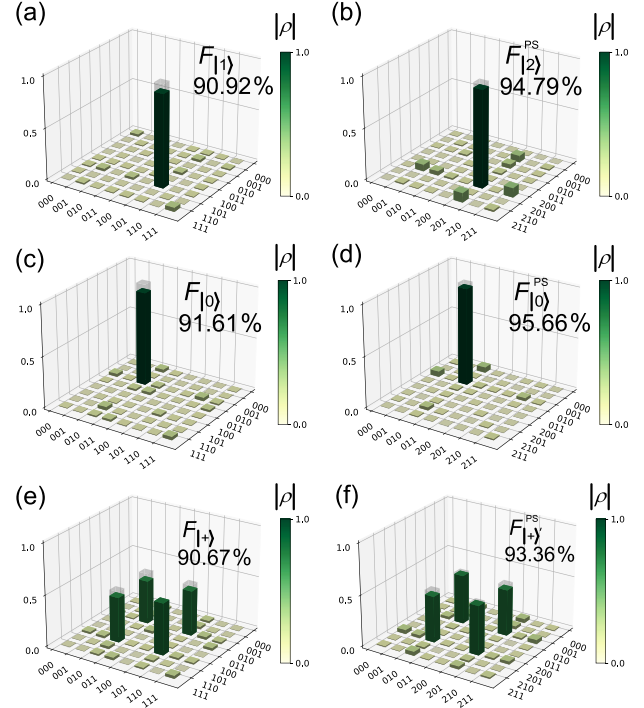


FIG. 15. QST results for QRouters. Left column: density matrices obtained under the nonerasure scheme for addresses $|1\rangle$, $|0\rangle$, and $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, shown in (a), (c), and (e), respectively. Right column: density matrices obtained after postselection under the erasure scheme for addresses $|2\rangle$, $|0\rangle$, and $|+\rangle = (|0\rangle + |2\rangle)/\sqrt{2}$, shown in (b), (d), and (f), respectively. The reconstructed matrices are averaged over ten independent QST experiments, each utilizing 10 000 measurement shots.

To further quantify how these state-dependent errors accumulate over deep circuits, we extended our simulations to sequences of $2N + 1$ consecutive QRouter operations, motivated by the tomographic error analysis methodology in Ref. [37]. We extracted an effective decay factor by fitting the resulting sequence fidelities to the empirical model:

$$f_{\text{QST}}(N) = a + bp^{2N+1}, \quad (\text{D2})$$

where $1 - p$ is the decay factor and a and b are fitting parameters. Under the baseline decoherence condition, Figs. 17(b) and 17(c) summarize the fidelity decay trajectories and the extracted decay factors. Notably, when both the address and information registers are prepared in a quantum superposition (labeled as QQ), the routing fidelity exhibits the steepest decay. This is consistent with the combined detrimental impact of longitudinal relaxation and rapid dephasing acting simultaneously on both registers. Conversely, when only the information qubit is excited (CC0), the accumulated error is minimized, as the

TABLE III. Input-output state types for the QRouter simulations. The state-type nomenclature follows the format {information state type, address state type, address value}. The letters Q and C denote whether the respective register is prepared in a quantum superposition or a classical basis state. The numerical suffix “1” (or “0”) indicates that the selected target address is the state $|1\rangle$ (or $|0\rangle$).

State type	Input state	Target state
CC0	$ 1000\rangle$	$ 0010\rangle$
CC1	$ 1100\rangle$	$ 0101\rangle$
QC0	$\frac{1}{\sqrt{2}}(1000\rangle + 0000\rangle)$	$\frac{1}{\sqrt{2}}(0010\rangle + 0000\rangle)$
QC1	$\frac{1}{\sqrt{2}}(1100\rangle + 0100\rangle)$	$\frac{1}{\sqrt{2}}(0101\rangle + 0100\rangle)$
CQ	$\frac{1}{\sqrt{2}}(1000\rangle + 1100\rangle)$	$\frac{1}{\sqrt{2}}(0010\rangle + 0101\rangle)$
QQ	$\frac{1}{2}(0000\rangle + 0100\rangle + 1000\rangle + 1100\rangle)$	$\frac{1}{2}(0000\rangle + 0100\rangle + 0010\rangle + 0101\rangle)$

population relaxation of the information qubit becomes the dominant error channel.

A necessary caveat when interpreting these fitted decay factors is the presence of distinct asymptotic fidelities across different state types. For example, when the information register is in a superposition, the fitted decay factor obtained for an excited address (QC1) can appear smaller than that for a ground-state address (QC0), despite excited-address paths experiencing strictly stronger longitudinal decay. This mathematical artifact arises because QC0 final states have a greater intrinsic overlap with the global ground state, which artificially raises the asymptotic

convergence level and confounds direct comparisons of the raw fitted parameter p . Because our empirical model assumes a common idealized asymptote, p should not be directly compared across configurations with different physical convergence limits. We acknowledge this caveat and plan to investigate generalized, asymptote-aware fitting procedures in future work.

Finally, to isolate the impact of pure dephasing, Figs. 17(d) and 17(e) present analogous simulation results for a dephasing-suppressed scenario ($T_1 = 20 \mu\text{s}$, $T_2 = 2T_1$). As expected, reducing transverse dephasing uniformly increases fidelities and lowers the effective decay

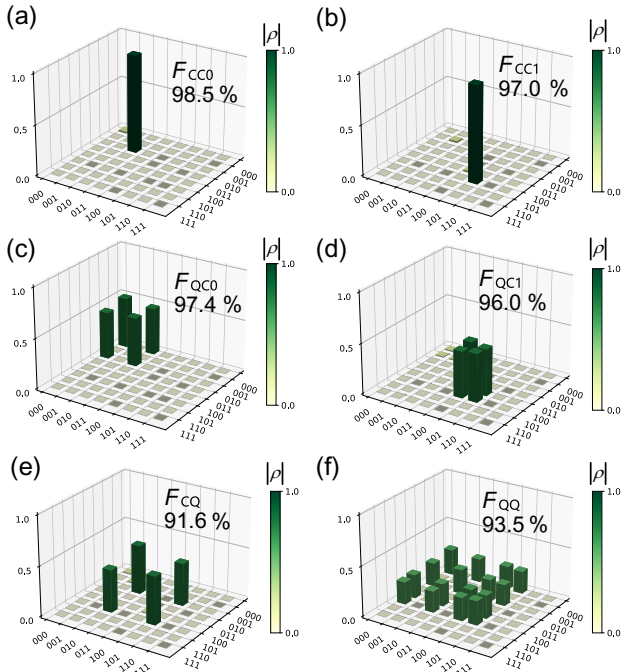


FIG. 16. Quantum state tomography simulation results under different initial states. (a)–(f) display the simulated output density matrices for a representative set of classical and quantum input states (as defined in Table III) following a single-layer QRouter operation.

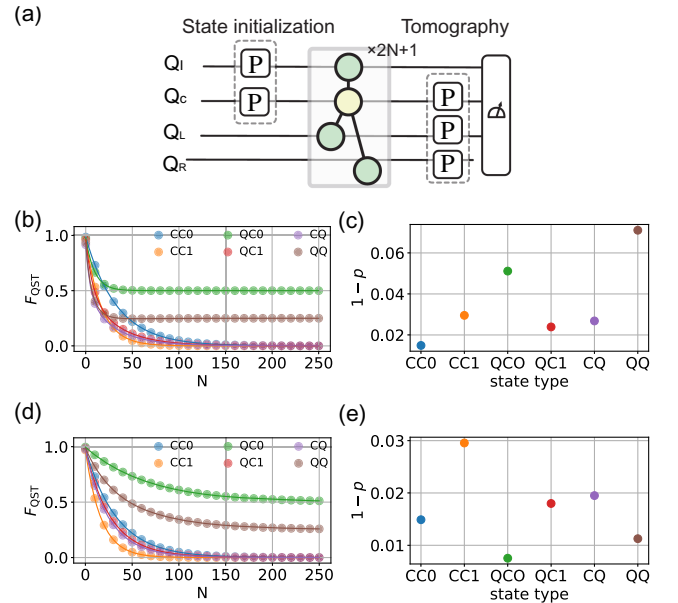


FIG. 17. Simulated QST fidelity decay under repeated QRouter operations. (a) Experimental setup. We evaluate the QST fidelity after a sequence of $2N + 1$ consecutive QRouter operations for various input state types. (b),(c) QST fidelity (F_{QST}) decay curves and the corresponding extracted decay rates under the decoherence condition $T_2 = T_1/2$. (d),(e) QST fidelity decay curves and the corresponding extracted decay rates under the improved decoherence condition $T_2 = 2T_1$.

rates. The most significant improvements are observed for state types involving superpositions in either register.

Crucially, these state-dependent sensitivities underscore a fundamental challenge in QRAM architectures: Different paths within a bucket-brigade network incur vastly different error budgets depending on their specific excitation profiles. Because full QST requires an exponentially growing number of measurements and cannot efficiently capture this holistic, path-dependent variation in a large-scale system, evaluating the entire routing network path by path is practically unscalable. This intrinsic nonuniformity and scaling bottleneck serve as the primary physical motivation for introducing a unified, scalable benchmarking metric, the RAT, which we detail in Appendix E.

APPENDIX E: THE RAT PROTOCOL

1. Physical motivation

The primary motivation for developing the RAT protocol stems from the path-dependent nature of address decoherence inherent in binary-tree routing architectures. In the bucket-brigade (BB)-QRAM framework, the number of independently addressable paths scales exponentially with the layer count L , as illustrated in Fig. 18(a). Crucially, each routing path exhibits distinct performance characteristics due to the varying decoherence vulnerabilities of its specific address configuration. This path-dependent decoherence process consequently induces highly nonuniform error profiles across the routing network.

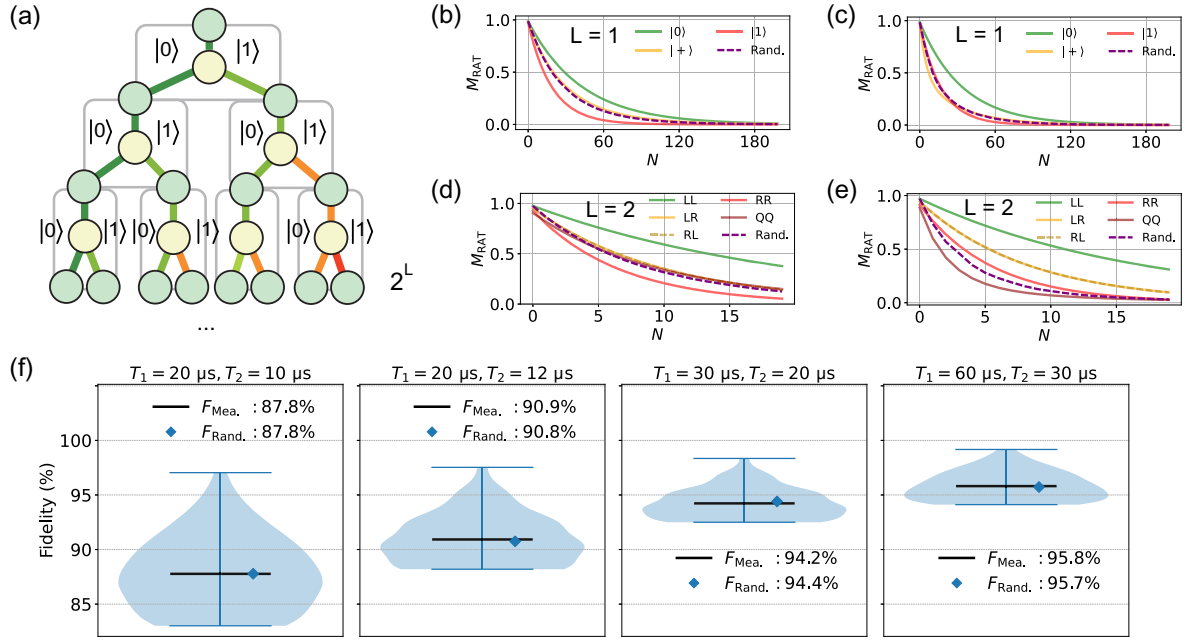


FIG. 18. Path-dependent decoherence and its impact on quantum routing performance. (a) Schematic illustration of path-dependent decoherence in the quantum routing network. Paths dictated by an address register in the global ground state suffer the least coherence loss (marked in green), while paths routing through all-excited address qubits experience the most severe decoherence (marked in red). This disparity in decoherence across different paths becomes increasingly pronounced as the router depth L increases. (b) One-layer ($L = 1$) comparison of transmission error rates for fixed addresses ($|0\rangle$, $|1\rangle$, $|+\rangle$) and for randomly sampled addresses (labeled “Rand.”), with $T_2 = 2T_1$. The fixed-address error rates correlate strongly with the excited-state population of the corresponding address register. (c) The same comparison as in (b) but for a stronger dephasing regime ($T_2 = T_1/2$). The reduced T_2 increases error rates across all address types but impacts superposition addresses the most severely. (d) Two-layer ($L = 2$) comparison of fixed and random addresses. The label C_1C_2 denotes the first- and second-layer address choices, where L (left) indicates the ground-state branch and R (right) indicates the excited-state branch. “QQ” denotes the set of fully superposed addresses. Results are shown for $T_2 = 2T_1$ and mirror the trends observed in (b). (e) The same as (d) but for $T_2 = T_1/2$. The trends are analogous to (c), with superposition-addressed paths exhibiting pronounced fidelity loss under stronger dephasing. (f) Comparison between fixed-address transmission simulations and random-address test fidelities across various decoherence parameter sets. Each column displays a violin plot summarizing the probability distribution of simulated transmission fidelities across all 64 possible address paths for that specific parameter set. Specifically, the horizontal width of the shaded region at any given vertical axis value represents the probability density (or relative frequency) of routing paths achieving that specific fidelity. The solid black line within each violin indicates the sample mean value of all paths, while the diamond marker positioned to the right of each violin denotes the corresponding random address transmission fidelity (RAT) for the same parameter set.

Specifically, we consider two distinct physical error mechanisms: longitudinal relaxation (amplitude damping) and transverse relaxation (dephasing). Under pure longitudinal relaxation, the error rates correlate directly with the total number of excited states present in the address configuration. For instance, the routing path dictated by an all-excited address register [marked in red in Fig. 18(a)] suffers maximal fidelity degradation, whereas the path corresponding to the global ground-state address (marked in green) demonstrates significantly reduced error susceptibility. The introduction of transverse relaxation further compounds this disparity, as paths relying heavily on superposition amplitudes experience enhanced dephasing penalties. Device-level nonuniformities, particularly qubit-specific variations in coherence times, further amplify these performance discrepancies across individual network paths.

Because the number of paths grows exponentially, traditional benchmarking methods that require exhaustive per-path tomographic analysis quickly become intractable. To overcome this scalability bottleneck, we propose the RAT protocol—a characterization scheme specifically tailored for large-scale quantum routing systems. By employing stochastic address sampling, the RAT efficiently averages out these path-dependent errors, providing a scalable, single-value metric for comprehensive system performance evaluation.

2. Theoretical analysis

To formally understand the physical implications of the RAT in a quantum router network, we model the system as a binary tree of depth L , representing a QRAM architecture. There are $N = 2^L$ possible pathways from the root to the leaves, indexed by a binary string $\mathbf{x} \in \{0, 1\}^L$.

In the RAT protocol, an address sequence $\mathbf{a} = (a_1, a_2, \dots, a_L)$ is generated, where each address state a_i is uniformly sampled from the set $\mathcal{A} = \{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$. We assume the data qubit experiences a path-dependent depolarizing channel with an intrinsic fidelity $F_{\mathbf{x}}$ when traversing the specific path $\mathbf{x}$, while the address qubits are considered ideal to isolate the routing transmission errors.

a. Path averaging

To evaluate the overall performance of a spatially heterogeneous QRAM, it is crucial that the benchmarking protocol does not bias toward any specific high-performing or defective hardware branch. The RAT protocol achieves this by utilizing a randomized address sequence that inherently balances the population distribution across the entire network. Furthermore, the inclusion of coherent states simulates the complex, simultaneous multibranch queries typical of quantum algorithms. We formalize this path-averaging mechanism and query simulation in the following theorems.

Theorem 1 (uniform path sampling). Under the uniform sampling of address states $a_i \in \mathcal{A}$, the expected population distribution (probability of data routing) across all 2^L pathways is perfectly uniform, despite the absence of the $|\pm i\rangle$ states.

Proof. For a single quantum router at layer i , let $p(L|a_i)$ and $p(R|a_i)$ denote the population fraction of the data qubit routed to the left and right branches, respectively. Based on the controlled-SWAP dynamics,

- (i) if $a_i \in \{|0\rangle, |1\rangle\}$, the routing is deterministic: $p(L|0) = 1, p(R|0) = 0$, and $p(L|1) = 0, p(R|1) = 1$;
- (ii) if $a_i \in \{|+\rangle, |-\rangle\}$, the data qubit is routed into an equal superposition. Tracing out the relative phase yields a balanced population: $p(L|\pm) = 1/2$ and $p(R|\pm) = 1/2$.

Since each state is chosen with probability $P(a_i) = 1/4$, the marginal probability of routing to the left branch is

$$\mathbb{E}[p(L)] = \frac{1}{4} \left(1 + 0 + \frac{1}{2} + \frac{1}{2} \right) = \frac{1}{2}. \quad (\text{E1})$$

By symmetry, $\mathbb{E}[p(R)] = 1/2$. Because the address state at each layer is chosen independently, the expected probability of reaching any specific leaf $\mathbf{x}$ is $\prod_{i=1}^L (1/2) = 1/2^L$. Thus, the sampling covers all branches uniformly. ■

Theorem 2 (arithmetic mean of network fidelity). Assuming ideal addressing operations, the overall survival fidelity $\bar{F}_{\text{RAT}}$ measured by the RAT protocol is exactly the arithmetic mean of the fidelities of all individual physical paths in the router network.

Proof. For a specific address sequence $\mathbf{a}$, let $p(\mathbf{x}|\mathbf{a})$ be the deterministic population fraction reaching path $\mathbf{x}$. The fidelity of the network for this specific query is $F(\mathbf{a}) = \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{a}) F_{\mathbf{x}}$. The RAT protocol repeats this process over a uniformly sampled ensemble of sequences $\mathbf{a}$. The measured fidelity is the expectation value over this ensemble:

$$\bar{F}_{\text{RAT}} = \mathbb{E}_{\mathbf{a}} \left[\sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{a}) F_{\mathbf{x}} \right] = \sum_{\mathbf{x}} \mathbb{E}_{\mathbf{a}} [p(\mathbf{x}|\mathbf{a})] F_{\mathbf{x}}. \quad (\text{E2})$$

According to Theorem 1, $\mathbb{E}_{\mathbf{a}} [p(\mathbf{x}|\mathbf{a})] = 1/2^L$ for all $\mathbf{x}$. Therefore,

$$\bar{F}_{\text{RAT}} = \frac{1}{2^L} \sum_{\mathbf{x}} F_{\mathbf{x}}. \quad (\text{E3})$$

This demonstrates that the $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$ basis selection perfectly hedges the spatial heterogeneity of the network, yielding the true average fidelity of the entire QRAM structure. ■

Beyond spatial averaging, the inclusion of the superposition states intrinsically simulates the complex query

workloads of a QRAM, which is mathematically described by the distribution of activated route subsets.

Theorem 3 (distribution of activated route subsets). In a single RAT shot on a QRAM of depth L , the number of simultaneously activated branches (leaves), denoted as M , follows a log-binomial distribution over the support $\{1, 2, 4, \dots, 2^L\}$.

Proof. Let k be the number of coherent address states ($|+\rangle$ or $|-\rangle$) within a randomly generated sequence of length L . Because the address line at layer i globally controls all routers at that depth, a classical state ($|0\rangle$ and $|1\rangle$) maintains the current number of active paths, whereas a coherent state ($|+\rangle$ and $|-\rangle$) coherently splits every active path, multiplying the total number of branches by 2.

Thus, the number of activated leaves is given by $M = 2^k$. Since the probability of selecting a coherent state at any layer is $P_{\text{coh}} = P(|+\rangle) + P(|-\rangle) = 1/2$, the number of coherent splittings k follows a binomial distribution $k \sim \mathcal{B}(L, 1/2)$. The probability of activating exactly $M = 2^k$ branches is, therefore,

$$P(M = 2^k) = \binom{L}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{L-k} = \binom{L}{k} \frac{1}{2^L}. \quad (\text{E4})$$

■

Theorem 3 highlights the superiority of the RAT protocol over fixed-path characterizations. By randomly drawing from $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$, the RAT automatically transitions between distinct operational regimes: single-point queries ($k = 0, M = 1$), full-tree broadcasts ($k = L, M = 2^L$), and random subset queries ($0 < k < L$). Consequently, the extracted exponential decay parameter F_{RAT} does not merely represent a local gate error; rather, it serves as a holistic metric indicating the average probability of successful data retrieval under arbitrary coherent superposition workloads in QRAM applications.

b. Initial state selection of RAT

Beyond the routing sequence, the choice of the data qubit's initial state is equally critical for error diagnostic sensitivity. In our experiment, the initial state is selected to be a single excitation injected into a global vacuum background, namely, $|D_{\text{in}}\rangle = |1\rangle$ and $|\Omega_P\rangle = |00\dots 0\rangle$. Rather than tracking mere population transfer, this specific state preparation tracks a unique excitation token, maximizing the observability of nonunitary dissipation and computational bit flips throughout the network.

Theorem 4 (optimal sensitivity to dissipative and bit-flip errors). Let the initial state of the data qubit be $|D_{\text{in}}\rangle$ and the initial state of all N spatial routing paths be the global vacuum state $|\Omega_P\rangle = \bigotimes_{k=1}^N |0\rangle_{P_k}$. Selecting $|D_{\text{in}}\rangle = |1\rangle$ provides the theoretically maximal observable contrast for characterizing both amplitude damping (energy

relaxation) and bit-flip errors during the routing process, compared to the $|0\rangle$ state initialization.

Proof. Consider the physical action of a quantum router network as an operator $\hat{U}_{\text{route}}$ that maps the state of the data qubit into the spatial modes of the paths. We evaluate the sensitivity of the population measurement at the target leaf to two dominant hardware error channels: the amplitude damping (AD) channel $\mathcal{E}_{\text{AD}}$ parametrized by decay rate γ and the bit-flip channel $\mathcal{E}_X$ parametrized by probability p_x .

Case 1: Vacuum initialization ($|D_{\text{in}}\rangle = |0\rangle$)—The total initial state is $|\Psi_0\rangle = |0\rangle_D \otimes |\Omega_P\rangle$. Since the controlled-SWAP operations of the router are excitation preserving, the ideal output state remains in the zero-excitation manifold, $|\Psi_{\text{id}}\rangle = |00\dots 0\rangle$. The AD channel is defined by Kraus operators $M_0 = |0\rangle\langle 0| + \sqrt{1-\gamma}|1\rangle\langle 1|$ and $M_1 = \sqrt{\gamma}|0\rangle\langle 1|$. Applying $\mathcal{E}_{\text{AD}}$ to the ideal output yields

$$\mathcal{E}_{\text{AD}}(|00\dots 0\rangle\langle 00\dots 0|) = |00\dots 0\rangle\langle 00\dots 0|. \quad (\text{E5})$$

The measured population error $\Delta P = P_{\text{ideal}} - P_{\text{measured}} = 1 - 1 = 0$. The sensitivity to amplitude damping is strictly zero. The system is entirely blind to T_1 relaxation.

Case 2: Single-excitation injection ($|D_{\text{in}}\rangle = |1\rangle$)—The initial state is $|\Psi_0\rangle = |1\rangle_D \otimes |\Omega_P\rangle$. Ideally, the network routes this single quantum of energy to a target path k (or a superposition of paths), meaning the ideal output ρ_{id} lies entirely within the single-excitation subspace $\mathcal{H}^{(1)} = \text{span}\{|0\dots 1_k\dots 0\rangle\}$. Applying the AD channel to any state $\rho \in \mathcal{H}^{(1)}$ results in

$$\mathcal{E}_{\text{AD}}(\rho) = (1-\gamma)\rho + \gamma|00\dots 0\rangle\langle 00\dots 0|. \quad (\text{E6})$$

When measuring the successful routing population $P_{\text{success}} = \text{Tr}(\hat{\Pi}^{(1)}\rho_{\text{out}})$ (where $\hat{\Pi}^{(1)}$ is the projector onto the target excitation paths), the measured population strictly decays by γ . The signal contrast $\Delta P_{\text{AD}} = \gamma$ is maximized.

Furthermore, consider a localized bit-flip error X occurring with probability p_x on the active path. For the $|1\rangle$ excitation, $X|1\rangle \rightarrow |0\rangle$, projecting the state back into the vacuum manifold and decreasing the success population by p_x . If the X error occurs on an idle (vacuum) path, $X|0\rangle \rightarrow |1\rangle$, it injects a spurious excitation, pushing the system into the two-excitation manifold $\mathcal{H}^{(2)}$, which can be distinctly identified as leakage or crosstalk background.

Therefore, initializing the system with exactly one excitation against a vacuum background ($|1\rangle \otimes |00\dots 0\rangle$) establishes an “excitation token tracking” paradigm. Any measured deviation from $\sum P_{\text{leaves}} = 1$ linearly and maximally quantifies the combined effect of nonunitary damping and computational bit flips.

Theorem 4 justifies the specific state preparation in our experiment. In a realistic QRAM application, data are encoded as arbitrary superpositions $\alpha|0\rangle + \beta|1\rangle$. Since the

$|0\rangle$ component is inherently immune to amplitude damping during transit, the $|1\rangle$ component represents the “worst-case scenario” for state survival. By characterizing the routing fidelity using $|D_{\text{in}}\rangle = |1\rangle$, the extracted RAT decay rate serves as a rigorous lower bound (the most conservative estimate) for the coherence preservation capacity of the QRAM network.

3. Numerical verification

To establish the RAT as a representative and scalable metric, we conducted comprehensive comparative simulations between fixed-path and random-access routing behaviors. For the fixed-path configuration, the simulation involves averaging the fidelities over $2N$ deterministic address sequences followed by a final randomized routing step. In contrast, for the random-path protocol, every routing operation is executed with randomized address states. The simulation results are presented in Figs. 18(b)–18(e), where the curves labeled “Rand” correspond to the random-address-based RAT characterization, while the others represent fixed-address routing performance.

The analysis begins with RAT simulations for one-layer ($L = 1$) and two-layer ($L = 2$) QRouters under experimentally relevant parameters ($T_1 = 20 \mu\text{s}$, $r = 0.8$). The complete RAT sequence is evaluated numerically using discrete matrix multiplications of the corresponding evolutionary operations, and the open-system dynamics model is obtained from Ref. [96]. Figures 18(b) and 18(d) depict the dephasing-suppressed scenario ($T_2 = 2T_1$). Here, ground-state paths consistently demonstrate maximum routing fidelity, all-excited-state paths exhibit minimal fidelity, and superposition-rich addresses occupy an intermediate performance regime. Conversely, in the $T_2 = T_1/2$ case shown in Figs. 18(c) and 18(e), the reduced transverse coherence time heavily amplifies dephasing effects. This causes superposition-addressed paths to experience significantly greater fidelity degradation compared to the $T_2 = 2T_1$ configuration. These simulations reveal the distinct, path-dependent degradation mechanisms driven by longitudinal and transverse decoherence, highlighting the urgent need for a unified metric capable of capturing the average system behavior.

Crucially, the RAT protocol fulfills this requirement by inherently quantifying the arithmetic average of the routing performance across all individual paths. Because uniform sampling over the address subspace ensures that the routing signal traverses every path with equal probability, the RAT fidelity theoretically converges to the global mean fidelity. To validate this mathematical equivalence, we compared the arithmetic mean of the per-path transmission fidelities ($F_{\text{Mea}} = \sum_{i=1}^m F_{A_i}/m$, where A_i denotes the i th address configuration. Specifically, in our two-layer simulation where $L = 2$, the routing network comprises three address qubits, each evaluated across four distinct address states, yielding a total of $m = 4^3 = 64$ comprehensive address

configurations) against the stochastically acquired RAT fidelity F_{Rand} under various decoherence regimes.

As explicitly shown in Fig. 18(f), the random access fidelity F_{Rand} demonstrates near-identical alignment with the network-wide average fidelity F_{Mea} . This convergence numerically confirms the RAT’s mathematical validity as a statistically representative metric for system-wide routing performance. The minor discrepancies occasionally observed between these two quantities arise solely from finite-sampling errors inherent in stochastic address preparation. However, these minor deviations scale inversely with the sample size and remain well within the statistical tolerance thresholds required for scalable quantum routing characterization.

APPENDIX F: ERROR-ERASURE SCHEME

1. Mechanism of the error-erasure protocol

Before quantifying the specific physical error channels, it is essential to analyze the underlying mechanisms by which these errors propagate. In conventional quantum routing, the address register relies strictly on the two-level subspace $\{|0\rangle, |1\rangle\}$. Consequently, when amplitude damping corrupts the address state, the resulting errors remain confined within this computational subspace. This mechanism induces routing failures that are indistinguishable and uncorrectable. One effective strategy to overcome this limitation is to introduce an observable error flag. Our error-erasure architecture deliberately redirects these physical errors into a designated, noncomputational flag state. By transforming address errors into identifiable erasures, this mechanism allows for error mitigation via postselection, bypassing the massive hardware overhead associated with active quantum error correction or traditional multi-qubit erasure schemes [7].

To realize a scalable, high-performance quantum router without introducing ancillary qubits, we construct the router using the TCG scheme and harness the higher-energy degrees of freedom inherent in superconducting transmons, as described in Appendix B. Specifically, we encode the valid routing address in the nonadjacent subspace $\{|0\rangle, |2\rangle\}$ for the error-erasure scheme while deliberately reserving the intermediate energy level $|1\rangle$ as an error flag state, independent of the computational logic. The core advantage of this encoding strategy lies in the fact that a dominant fraction of first-order physical errors experienced by the quantum system during operation will deterministically drive the system into this flag state. First, during decoherence dominated by amplitude damping, an address qubit in the excited state $|2\rangle$ preferentially undergoes single-photon decay, dropping into the $|1\rangle$ state. Second, under thermal excitation, a qubit in the ground state $|0\rangle$ absorbs environmental energy and similarly transitions to the $|1\rangle$ state. Furthermore, during the execution of two-qubit entangling gates (such as an imperfect

$\sqrt{\text{CZ}}$ gate), control pulse distortions or parasitic interactions can induce coherent leakage, irreversibly transferring population from the computational subspace to parasitic levels containing the $|1\rangle$ state. Consequently, owing to the adjacency of the energy-level structure, the $|1\rangle$ state naturally acts as a physical convergence point to intercept these first-order hardware errors.

Leveraging this error-mapping mechanism, we execute a postselection measurement at the conclusion of the entire quantum routing operation. Specifically, when reading out the address qubit, the detection of the $|1\rangle$ state serves as an unambiguous indicator that a physical-level error has occurred during the routing process. Subsequently, during the classical data processing phase, we directly discard this portion of the data. Considering the energy relaxation of the address qubits, typically only lower-probability, second-order transition processes (e.g., the cascaded decay $|2\rangle \rightarrow |1\rangle \rightarrow |0\rangle$) will result in undetectable errors that ultimately degrade the final fidelity.

2. Analytical model and numerical validation

To understand the decoherence dynamics within the periodic quantum routing protocol, we develop a continuous-time model in this section. We first analyze the ideal operational sequence, demonstrating how the strictly alternating gate operations lead to a monotonic accumulation of the error flag state ($|1\rangle$). This idealized analytical framework is then validated against exact discrete numerical simulations. Subsequently, to bridge the gap between theoretical predictions and empirical observations in scaled multilayer architectures, we extend our model to incorporate parasitic leakage channels. This generalized framework captures the nonmonotonic anomalous dynamics observed in experiments, providing possible insights into the error mechanisms of highly parallelized quantum circuits.

In practical periodic quantum routing circuits, the fundamental operational sequence along the primary decoherence pathway consists of interleaved multiqubit entangling gates and X_{02} pulses (as illustrated in Appendix B 2). Under continuous cascaded amplitude damping ($|2\rangle \rightarrow |1\rangle \rightarrow |0\rangle$), the X_{02} pulses act as instantaneous exchange operations, periodically swapping the populations of the $|0\rangle$ and $|2\rangle$ states. Consequently, over any two consecutive cycles, the system spends an equal effective duration exposed to the specific dissipation dynamics of both $|0\rangle$ and $|2\rangle$. This temporal symmetry effectively averages out the state-dependent dissipation across the asymmetric gate intervals. This cyclic sequence establishes an effective pumping mechanism, driving a monotonic accumulation of population within the error flag state ($|1\rangle$) in an ideal isolated system. However, it is crucial to emphasize that this dynamical symmetry is strictly confined to the internal execution phase of the individual ideal QRouter. This local characteristic must be carefully distinguished from the perspective of the overall RAT

architecture. At the system level, the impact of address errors across different global routing paths remains inhomogeneous. This global asymmetry arises from various factors, including the variations in specific hardware performances across different physical pathways, address state preparation and loading and the fact that not all routers within the network are consistently in an active working state during the routing process.

To analytically model this process, we define a duty cycle ratio $k = t_{\text{Mq}}/t_X$, where t_{Mq} and t_X represent the duration of the multiqubit gate and the X_{02} gate, respectively. Let $\epsilon = \gamma t_X$ be the dimensionless error accumulated during a single X_{02} gate, where $\gamma = \Gamma_{10}$ is the base relaxation rate. The single-qubit gate relaxation propagators for the X_{02} and multiqubit gate intervals can be expanded to the first order in ϵ :

$$D_X = \exp(L \cdot t_X) \approx I + \epsilon L, \quad (\text{F1})$$

$$D_{\text{Mq}} = \exp(L \cdot t_{\text{Mq}}) \approx I + k\epsilon L, \quad (\text{F2})$$

where L is the dimensionless continuous-time relaxation generator governing the natural cascaded decay.

The discrete evolution operator for a single cycle M_{single} , applied from right to left, consists of an X_{02} gate followed by the combined decoherence of both operations:

$$M_{\text{single}} = D_{\text{Mq}} \cdot D_X \cdot X \approx (I + k\epsilon L)(I + \epsilon L)X \quad (\text{F3})$$

$$\approx [I + (k+1)\epsilon L]X. \quad (\text{F4})$$

Because the X_{02} gate flips the relevant subspace, a continuous generator must be evaluated over two consecutive cycles to form a complete identity-mapped envelope. The double-cycle operator is

$$\begin{aligned} M_{\text{Router}} &= M_{\text{single}}^2 \approx [I + (k+1)\epsilon L]X[I + (k+1)\epsilon L]X \\ &= [I + (k+1)\epsilon L][I + (k+1)\epsilon LXL]X^2 \\ &\approx I + (k+1)\epsilon(L + XLX). \end{aligned} \quad (\text{F5})$$

Since the double cycle spans a duration equivalent to 2ϵ in base units, we define the generalized effective continuous-time generator L_{eff} such that $M_{\text{Router}} = I + 2\epsilon L_{\text{eff}}$:

$$L_{\text{eff}} = \frac{k+1}{2}(L + XLX). \quad (\text{F6})$$

To construct the effective equations, we first define the state space and the amplitude damping generator L . We represent the macrostate of the transmon as a probability column vector $\vec{P} = (P_0, P_1, P_2)^T$. This vector is spanned by the standard canonical basis, where each basis element directly corresponds to the normalized population residing in the energy levels $|0\rangle$, $|1\rangle$, and $|2\rangle$, respectively.

Within this basis, the continuous-time Markovian dynamics of the cascaded amplitude damping channel are governed by the classical rate equations (i.e., the diagonal projection of the Lindblad master equation):

$$\frac{d}{dt} \begin{pmatrix} P_0 \\ P_1 \\ P_2 \end{pmatrix} = \gamma \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & \eta \\ 0 & 0 & -\eta \end{pmatrix} \begin{pmatrix} P_0 \\ P_1 \\ P_2 \end{pmatrix}. \quad (\text{F7})$$

By scaling the time with respect to the base relaxation rate $\gamma = \Gamma_{10}$, we introduce the decay enhancement ratio $\eta = \Gamma_{21}/\Gamma_{10}$ (noting that $\eta = 1/r$, where r is the lifetime ratio, $r = \Gamma_{10}/\Gamma_{21}$) to extract the dimensionless continuous-time relaxation generator L . Physically, its off-diagonal elements L_{ij} (where $i \neq j$) denote the transition rate *from* state j *to* state i , while the diagonal elements L_{jj} represent the total decay rate *out* of state j . Thus, the generator L and the corresponding pumped term XLX are evaluated as

$$L = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & \eta \\ 0 & 0 & -\eta \end{pmatrix}, \quad XLX = \begin{pmatrix} -\eta & 0 & 0 \\ \eta & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \quad (\text{F8})$$

Substituting these matrices into the generalized effective generator, we obtain a matrix:

$$\begin{aligned} L_{\text{eff}} &= \frac{k+1}{2} \left[\begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & \eta \\ 0 & 0 & -\eta \end{pmatrix} + \begin{pmatrix} -\eta & 0 & 0 \\ \eta & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \right] \\ &= \frac{k+1}{2} \begin{pmatrix} -\eta & 1 & 0 \\ \eta & -2 & \eta \\ 0 & 1 & -\eta \end{pmatrix}. \end{aligned} \quad (\text{F9})$$

Transitioning to a continuous -time variable $t = n\epsilon$, the evolution is governed by $[(d\vec{P})/dt] = L_{\text{eff}}\vec{P}$. Extracting the explicit components, we obtain

$$\frac{dP_1}{dt} = \frac{k+1}{2} [\eta P_0 - 2P_1 + \eta P_2], \quad (\text{F10})$$

$$\frac{dP_2}{dt} = \frac{k+1}{2} [P_1 - \eta P_2]. \quad (\text{F11})$$

Remarkably, by invoking the probability conservation $P_0 + P_2 = 1 - P_1$, the differential equation for the flag state P_1 decouples into a first-order linear ordinary differential equation:

$$\frac{dP_1}{dt} = \frac{k+1}{2} [\eta - (\eta + 2)P_1]. \quad (\text{F12})$$

Setting $[(dP_1)/dt] = 0$, we find that the theoretical limit for the steady-state population is fundamentally

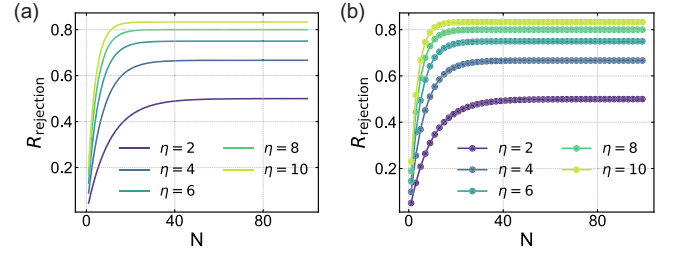


FIG. 19. Comparison of theoretical predictions and numerical simulations for the postselection rejection rate $R_{\text{rejection}}$. (a) The analytical continuous-time model showing $R_{\text{rejection}}$ for various decay enhancement ratios η , evaluated at a fixed duty cycle ratio of $k = 5/2$. (b) Discrete numerical simulation against the operational cycle index N . Both panels demonstrate identical monotonic transient dynamics and steady-state limits. The color map indicates the transition from a moderate population accumulation ($\eta = 2$, dark purple) to a severe dissipation bottleneck ($\eta = 10$, light yellow).

independent of the duty cycle ratio k , relying strictly on the intrinsic decay enhancement ratio η :

$$P_1^* = \frac{\eta}{\eta + 2}. \quad (\text{F13})$$

Similarly, the steady state for the second excited state is derived as $P_2^* = [1/(\eta + 2)]$.

Given the initial condition $P_1(0) = 0$, the time evolution of $P_1(t)$ is a strict, monotonic exponential approach to the asymptote, entirely devoid of any transient overshoot:

$$P_1(t) = P_1^*(1 - e^{-\lambda t}), \quad \text{where } \lambda = \frac{k+1}{2}(\eta + 2). \quad (\text{F14})$$

This rigorous analytical framework proves that the strictly alternating sequence balances the dynamic dissipation, reducing the temporal asymmetry introduced by the multi-qubit gates.

To validate our theoretical framework, we compare the predictions of the analytical model with numerical simulations of the discrete operational sequence. Figure 19(a) illustrates the theoretically derived postselection rejection rate $R_{\text{rejection}}$ as a function of the normalized effective time for varying decay enhancement ratios η , with the duty cycle ratio fixed at $k = 5/2$ ($t_{\text{Mq}} = 75$ ns, $t_X = 30$ ns here). In parallel, Fig. 19(b) presents the corresponding numerical results. The simulation data are plotted against the discrete operational cycle index N . To mitigate the classical computational overhead of the exact numerical simulations, we assigned a relatively short baseline relaxation time of $T_1 = 5$ μ s.

The comparison reveals a strong qualitative consistency between the two methodologies. Both approaches capture the defining monotonic dynamics of the system: a smooth transient accumulation followed by a gradual relaxation toward a stable asymptote. Furthermore, the dynamic dependence on η is consistently reflected in both panels. Increasing η

elevates the final steady-state limit, corroborating the analytical predictions derived in the previous section.

We also observe subtle quantitative discrepancies between the datasets, most notably a slight divergence in the exact asymptotic convergence levels for extreme parameters, such as $\eta = 10$. These deviations stem fundamentally from the first-order approximation utilized to construct the effective continuous-time generator, which smooths over the higher-order intracycle discrete decoherence jumps. Despite these minor deviations, the continuous model serves as a highly robust and computationally efficient analytical tool. It maps the complex discrete quantum routing circuit into a tractable mathematical framework, providing a reliable approximation for predicting the scaling behavior of the postselection overhead.

In the fundamental analysis of the quantum routing circuit, the idealized continuous-time model predicts a monotonic accumulation of the flag state population P_1 , smoothly reaching a plateau in a strictly isolated one-layer system. However, in more complex experimental implementations, particularly in the scaled two-layer architecture, empirical data reveals an anomalous signature: The evolution of $R_{\text{rejection}}$ exhibits non-monotonic behavior, characterized by a pronounced transient peaking followed by degradation over time.

To account for this phenomenon, we hypothesize that, in more complex architectures, the high degree of parallelization in quantum circuits and factors such as microwave crosstalk may invalidate the initially stable pumping model. Specific contributing factors include the leakage of the transmon qubit to higher energy levels (e.g., $|3\rangle$ and $|4\rangle$) driven by unintended excitations and spurious interactions, as well as parasitic energy transfer to adjacent couplers. These effects cause the energy originally intended for the computational subspace to be erroneously pumped into uncontrolled trap states. To theoretically encapsulate these complex dynamics, we abstract these additional leakage channels into a single generalized leakage state, denoted as $|\mathcal{L}\rangle$.

Assuming that the generalized leakage state $|\mathcal{L}\rangle$ is coupled to the auxiliary state $|2\rangle$ via the aforementioned parasitic mechanisms, we augment the local symmetric generator by introducing a forward parasitic transition rate, γ_{leak} (from $|2\rangle$ to $|\mathcal{L}\rangle$), and a generalized return rate, γ_{back} (representing the relaxation from $|\mathcal{L}\rangle$ back to $|2\rangle$ or the active computational subspace). Consequently, the state of the system is extended to a four-dimensional probability vector $\vec{P} = (P_0, P_1, P_2, P_{\mathcal{L}})^T$. The asymmetric single-step dissipation generator L_{leak} is defined as

$$L_{\text{leak}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & \eta & 0 \\ 0 & 0 & -(\eta + \gamma_{\text{leak}}) & \gamma_{\text{back}} \\ 0 & 0 & \gamma_{\text{leak}} & -\gamma_{\text{back}} \end{pmatrix}. \quad (\text{F15})$$

By incorporating this cross-block coupling into the strictly alternating double-cycle sequence, we apply the

symmetrization operator over two adjacent cycles. The resulting effective continuous-time generator $L_{\text{eff}} = [(k+1)/2](L_{\text{leak}} + XL_{\text{leak}}X)$ assumes a strictly symmetric form:

$$L_{\text{eff}} = \frac{k+1}{2} \begin{pmatrix} -(\eta + \gamma_{\text{leak}}) & 1 & 0 & \gamma_{\text{back}} \\ \eta & -2 & \eta & 0 \\ 0 & 1 & -(\eta + \gamma_{\text{leak}}) & \gamma_{\text{back}} \\ \gamma_{\text{leak}} & 0 & \gamma_{\text{leak}} & -2\gamma_{\text{back}} \end{pmatrix}. \quad (\text{F16})$$

The corresponding evolution is governed by the master equation $[(d\vec{P})/d\tau] = [2/(k+1)]L_{\text{eff}}\vec{P}$, where τ is the scaled effective time ($\tau = [(k+1)/2]t$). In this context, the additional generalized leakage breaks the dynamic equilibrium of the system, disabling the originally stable pumping source. To obtain a rigorous analytical understanding of this nonmonotonic dynamical process, we further derive the time-dependent evolution formula for the flag state $P_1(\tau)$ based on this equation.

The system of differential equations for the primary computational subspace can be extracted from the matrix L_{eff} :

$$\frac{dP_0}{d\tau} = -(\eta + \gamma_{\text{leak}})P_0 + P_1 + \gamma_{\text{back}}P_{\mathcal{L}}, \quad (\text{F17})$$

$$\frac{dP_1}{d\tau} = \eta P_0 - 2P_1 + \eta P_2, \quad (\text{F18})$$

$$\frac{dP_2}{d\tau} = P_1 - (\eta + \gamma_{\text{leak}})P_2 + \gamma_{\text{back}}P_{\mathcal{L}}. \quad (\text{F19})$$

Exploiting the inherent dynamic symmetry of the system, we define the joint survival probability within the driving subspace as $P_+ = P_0 + P_2$. Summing the differential equations for P_0 and P_2 yields

$$\frac{dP_+}{d\tau} = -(\eta + \gamma_{\text{leak}})P_+ + 2P_1 + 2\gamma_{\text{back}}P_{\mathcal{L}}. \quad (\text{F20})$$

Crucially, the global probability conservation of the system requires $P_0 + P_1 + P_2 + P_{\mathcal{L}} = 1$, which inherently constrains the population of the generalized leakage state to $P_{\mathcal{L}} = 1 - P_+ - P_1$. Substituting this conservation constraint into the aforementioned equations allows us to eliminate $P_{\mathcal{L}}$, reducing the original four-dimensional homogeneous system into a two-dimensional nonhomogeneous linear differential equation system:

$$\frac{d}{d\tau} \begin{pmatrix} P_+ \\ P_1 \end{pmatrix} = \begin{pmatrix} -(\eta + \gamma_{\text{leak}} + 2\gamma_{\text{back}}) & 2(1 - \gamma_{\text{back}}) \\ \eta & -2 \end{pmatrix} \begin{pmatrix} P_+ \\ P_1 \end{pmatrix} + \begin{pmatrix} 2\gamma_{\text{back}} \\ 0 \end{pmatrix}. \quad (\text{F21})$$

The nonhomogeneous nature of this system physically indicates that, unlike a perfect sink model, the system

population does not asymptotically collapse to zero but rather reaches a nonzero steady-state thermal equilibrium. By setting the derivatives to zero ($(dP_+/d\tau) = 0$ and $(dP_1/d\tau) = 0$), we can solve for the asymptotic limit of the flag state population, $P_1^* \equiv P_1(\infty)$, which mathematically defines the height of the final degraded plateau:

$$P_1^* = \frac{\eta\gamma_{\text{back}}}{\gamma_{\text{leak}} + (\eta + 2)\gamma_{\text{back}}}. \quad (\text{F22})$$

This steady-state solution is consistent under physical limits: In the absolute trap limit ($\gamma_{\text{back}} \rightarrow 0$), the flag state eventually reaches zero, i.e., $P_1^* \rightarrow 0$; whereas in the ideal closed-system limit ($\gamma_{\text{leak}} \rightarrow 0$), the formula seamlessly converges to the classical monotonic plateau limit of the one-layer system, $P_1^* \rightarrow [\eta/(\eta + 2)]$.

The transient dynamical behavior of the system is determined by the eigenvalues of the corresponding two-dimensional homogeneous matrix. Solving the characteristic equation yields two real roots, $s_{1,2}$:

$$s_{1,2} = \frac{-(\eta + \gamma_{\text{leak}} + 2\gamma_{\text{back}} + 2) \pm W'}{2}, \quad (\text{F23})$$

where the generalized scaling factor W' incorporates the influence of the reversible relaxation dynamics:

$$W' = \sqrt{(\eta + \gamma_{\text{leak}} + 2\gamma_{\text{back}} + 2)^2 - 8[\gamma_{\text{leak}} + (\eta + 2)\gamma_{\text{back}}]}. \quad (\text{F24})$$

To map these effective parameters back to the real physical time domain t , we scale the eigenvalues by the duty cycle, such that $\lambda_{1,2} = [(k + 1)/2]s_{1,2}$. Assuming the system is perfectly initialized in the address state $|2\rangle$ [with initial conditions $P_+(0) = 1$ and $P_1(0) = 0$], solving the aforementioned nonhomogeneous differential equation system yields the exact analytical time-dependent formula for the evolution of the flag state $P_1(t)$:

$$P_1(t) = P_1^* + \frac{\eta + s_2 P_1^*}{W'} e^{\lambda_1 t} - \frac{\eta + s_1 P_1^*}{W'} e^{\lambda_2 t}. \quad (\text{F25})$$

Equation (F25) provides a mathematical description of the complex anomalous dynamics observed in the experiment. The first term, P_1^* , establishes the steady-state thermalized noise floor determined by the interplay between the drive-induced leakage γ_{leak} and the intrinsic relaxation γ_{back} . The subsequent biexponential terms mathematically capture the transient peaking process: Since both decay constants λ_1 and λ_2 are strictly negative, the flag state population experiences a rapid initial surge (forming a peak) before inevitably decaying under the influence of cross-block coupling, ultimately settling smoothly at the asymptotic steady-state limit P_1^* .

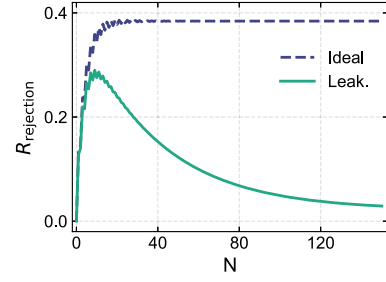


FIG. 20. Evolution of the rejection rate ($R_{\text{rejection}}$) as a function of cycle index N . The dashed line (Ideal) demonstrates the continuous accumulation plateau expected in an isolated system. The solid line (Leak.) illustrates the transient peaking and subsequent gradual degradation, captured by our reversible leakage model with parameters $\gamma_{\text{leak}} = 0.55$, $\gamma_{\text{back}} = 0.01$, $k = 5/2$, and $\eta = 1/0.8$.

As corroborated by Fig. 20, the deviation from the ideal plateau reflects the underlying hardware limitations of the scaled architecture. Based on our theoretical framework, we estimate that a severe disparity between the forward leakage and the intrinsic return rate ($\gamma_{\text{leak}} \gg \gamma_{\text{back}}$) drives the gradual degradation of the flag state population following its initial transient peak. This estimation suggests that, in highly parallelized multilayer routing circuits, collective parasitic interactions and unintended leakages into $|L\rangle$ eventually outpace the population accumulation within the target computational subspace. Consequently, while our postselection mechanism effectively mitigates errors in the short-depth regime, this continuous parasitic drain inherently bounds its asymptotic performance in deep circuits.

To mitigate these secondary effects and further enhance the scalability of the routing protocol, future architectural optimizations should focus on suppressing these parasitic channels. Potential solutions include implementing more stringent hardware and pulse-level isolation techniques to minimize crosstalk-induced leakage in parallel operations. Furthermore, exploring architectures with an expanded redundant Hilbert space could provide the system with enhanced capacity to absorb and dissipate these parasitic excitations, thereby extending the operational depth and maintaining the integrity of the primary computational subspace.

3. The scalability of error-erasure scheme

To assess how the erasure protocol scales in larger routing networks, we extrapolate the performance to varying numbers of layers L . We first consider a baseline limited primarily by environmental decoherence, focusing on the worst-case path where all address registers are in excited states [Figs. 21(a) and 21(b)]. Figure 21(c) plots the resulting error rates for two coherence scenarios ($T_1 = 20$ and $60 \mu\text{s}$, with $r = 0.8$).

As L increases, the nonerasure scheme exhibits a rapidly rising error that grows faster than linearly, driven by the

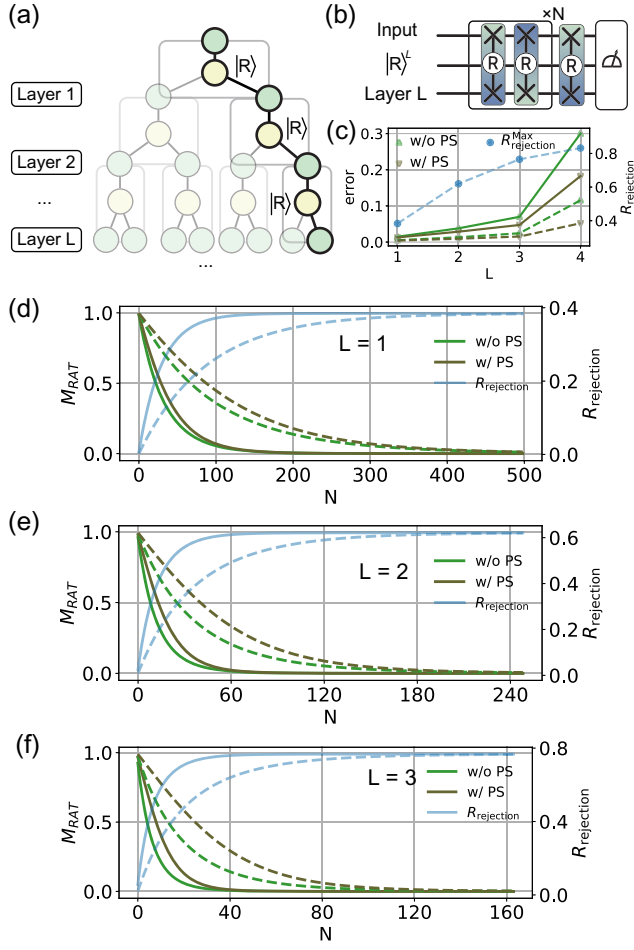


FIG. 21. Simulated validation of the error-erasure scheme in an L -layer quantum routing system. (a) Structural diagram of the single-chain quantum routing system. We probe the worst-case single path (all-excited address), which suffers the strongest address-induced decoherence and provides a stringent test of the erasure protocol. (b) Simplified circuit diagram. The top input qubit is initialized in the excited state, all address registers are initialized in $|R\rangle^i$ ($|1\rangle$ or $|2\rangle$), and $2N + 1$ consecutive routing operations are applied. (c) Comparison of error rates versus L under pure decoherence. Two scenarios are shown: $T_1 = 20 \mu\text{s}$ (solid lines) and $T_1 = 60 \mu\text{s}$ (dashed lines) with $T_2 = 2T_1$. The maximum rejection rates ($R_{\text{rejection}}^{\text{Max}}$) are marked in blue. (d)–(f) Detailed RAT results for $L = 1$, $L = 2$, and $L = 3$. The brown and green curves indicate the error-erasure and nonerasure schemes, respectively. The blue curve tracks the rejection rate.

combined decoherence of the address and data registers. Address errors are particularly damaging, because they inherently corrupt the routing control. By suppressing the impact of address decoherence, the erasure protocol achieves uniformly lower errors across all tested sizes. However, this advantage requires a practical trade-off: The postselection overhead increases with system size. The peak rejection rate $R_{\text{rejection}}^{\text{Max}}$ rises from approximately 0.4 at $L = 1$ to about 0.8 at $L = 3$. Figure 21(c) reports the maximum $R_{\text{rejection}}$ as a function of L :

$$R_{\text{rejection}}^{\text{Max}}(L) = 1 - [1 - R_{\text{rejection}}^{\text{Max}}(1)]^L, \quad (\text{F26})$$

where $R_{\text{rejection}}^{\text{Max}}(1)$ is the maximum $R_{\text{rejection}}$ of the one-layer system. This scaling relationship allows us to estimate the postselection overhead for larger-scale architectures, providing a practical evaluation for the resource costs of error erasure.

We suspect that extra errors on the address qubit, perhaps due to crosstalk, which is known to be present in the system, may be responsible for the sharper turnaround in the rejection rate observed in the two-layer experimental case, as discussed in Appendix F 2. However, actual experimental deviations are multifaceted, such that other explanations remain possible. For instance, imperfect a_i and a_i^{-1} routing pulses, which are not explicitly included in this specific simulation, could also contribute to the observed dynamics. These simulations serve primarily to illustrate how the accumulation of coherent errors can lead to such phenomenological effects. Ultimately, deploying scalable routing architectures requires managing these practical trade-offs. Notably, even in the absence of the erasure scheme, the underlying TCG-based implementation retains a substantial circuit-depth advantage relative to traditional CZ-based decompositions, providing a favorable baseline. Furthermore, because the erasure scheme can be applied locally, selective deployment on highly susceptible nodes offers a flexible path to maximizing end-to-end fidelity in near-term distributed architectures.

Looking forward, bridging the gap between current component-level demonstrations and practical large-scale deployment will require coordinated advances. On the control front, optimizing waveforms to minimize $\sqrt{\text{CZ}}$ -induced leakage and employing machine-learning-assisted crosstalk compensation are critical priorities [68,69,72,86,87]. Simultaneously, device-level upgrades, such as enhanced Purcell filtering to suppress radiative loss [61,64], will extend intrinsic coherence limits. As established in our recent architectural study [77], operating a deep, five-layer BB-QRAM necessitates pushing the equivalent depolarization rate to a stringent threshold of $p \lesssim 0.01$. We believe that unifying the structural resilience of the error-erasure protocol with these ongoing hardware and control optimizations establishes a viable trajectory toward that ambitious operational milestone.

APPENDIX G: COMPREHENSIVE ERROR ANALYSIS AND ERROR BUDGETS

To understand the physical mechanisms degrading our quantum routing architecture, the primary objective of this section is to construct a quantitative error budget that analyzes the specific contributions of dominant error channels. Our experimental baseline is established using the RAT protocol. By extracting the average operational fidelity through depth fitting, RAT reduces state preparation

and measurement errors. This allows us to cleanly isolate the intrinsic error rate ($1 - F_{\text{RAT}}$) of the routing network, which serves as the total error variance to be subsequently decomposed into distinct physical sources.

To accurately attribute this total intrinsic error to specific physical mechanisms within the superconducting processor, we must first establish theoretical models for each individual noise channel. Consequently, this comprehensive analysis is structurally organized to progress from fundamental noise modeling to system-level synthesis. Specifically, Appendixes G 1 a and G 1 b first analyze the incoherent environment errors, namely, environmental decoherence and thermal excitations. Following this, Appendixes G 2 a–G 2 c evaluate the coherent control errors, isolating the impacts of coherent leakage and parasitic SWAP interactions induced by imperfect $\sqrt{\text{CZ}}$ operations. Next, in Appendix G 3, we integrate these isolated theoretical models with our experimental RAT measurements. This synthesis yields a comprehensive quantitative error budget, explicitly breaking down the primary error contributions to the routing infidelities at both the component and system levels. Finally, building upon this comprehensive error understanding, Appendix G 4 performs a targeted parameter sweep to explicitly verify the advantages of the error-erasure scheme and establish the specific hardware requirements necessary for realizing high-performance quantum routers in the future.

1. Incoherent environment error

a. Impact of decoherence

Among all physical error sources in superconducting processors, decoherence, comprising amplitude damping (T_1) and pure dephasing (T_ϕ), establishes the fundamental fidelity limit of any quantum operation. Specifically, our error-erasure protocol helps mitigate the impact of amplitude damping on the address register by identifying aberrant populations in the $|1\rangle$ flag state. While this targeted approach yields neutral benefits regarding pure address dephasing and data-qubit relaxation, it nonetheless provides a advantage against the dominant amplitude damping channels. With this defined scope, we now solve the open-system dynamics to derive the theoretical fidelity penalty imposed by amplitude damping on both the standard nonerasure baseline and our postselected scheme.

To model the nonerasure baseline, we first consider the standard two-level qubit system. The relaxation rate from the excited state $|1\rangle$ to the ground state $|0\rangle$ is denoted by $\Gamma_{10} = 1/T_{10}$, where T_{10} is the relaxation time. The corresponding population dynamics are given by

$$\dot{A}_{|1\rangle} = -\Gamma_{10}A_{|1\rangle}, \quad (\text{G1})$$

$$\dot{A}_{|0\rangle} = \Gamma_{10}A_{|1\rangle}, \quad (\text{G2})$$

with the initial condition $A_{|1\rangle}(t=0) = 1$, yielding the standard exponential decay:

$$A_{|1\rangle}(t) = e^{-\Gamma_{10}t}, \quad (\text{G3})$$

$$A_{|0\rangle}(t) = 1 - e^{-\Gamma_{10}t}, \quad (\text{G4})$$

where $A_{|i\rangle}$ represents the probability of state $|i\rangle$. Under the nonerasure scheme, both the routing data and the address state subjected to this first-order amplitude damping process are irreversibly corrupted within the computational basis.

By contrast, the erasure scheme operates within a three-level transmon configuration. The presence of the higher excited state $|2\rangle$ introduces an additional damping pathway characterized by $\Gamma_{21} = 1/T_{21}$, representing the relaxation from $|2\rangle$ to $|1\rangle$. Because direct transitions from $|2\rangle$ to $|0\rangle$ are second-order processes with a significantly lower probability, the full open-system damping dynamics of this three-level configuration are cascaded:

$$\dot{A}_{|2\rangle} = -\Gamma_{21}A_{|2\rangle}, \quad (\text{G5})$$

$$\dot{A}_{|1\rangle} = \Gamma_{21}A_{|2\rangle} - \Gamma_{10}A_{|1\rangle}, \quad (\text{G6})$$

$$\dot{A}_{|0\rangle} = \Gamma_{10}A_{|1\rangle}. \quad (\text{G7})$$

Given the initial condition $A_{|2\rangle}(t=0) = 1$, the state populations evolve as

$$A_{|2\rangle}(t) = e^{-\Gamma_{21}t}, \quad (\text{G8})$$

$$A_{|1\rangle}(t) = \frac{\Gamma_{21}(e^{-\Gamma_{10}t} - e^{-\Gamma_{21}t})}{\Gamma_{21} - \Gamma_{10}}, \quad (\text{G9})$$

$$A_{|0\rangle}(t) = \frac{\Gamma_{10}e^{-\Gamma_{21}t} - \Gamma_{21}e^{-\Gamma_{10}t}}{\Gamma_{21} - \Gamma_{10}} + 1. \quad (\text{G10})$$

As established, the erasure protocol utilizes postselection to detect and discard trials that fall into the intermediate $|1\rangle$ state. Consequently, the effective amplitude damping in the postselected system manifests only through the second-order decay from $|2\rangle$ directly to $|0\rangle$. The renormalized populations of the target states, conditioned on the absence of the $|1\rangle$ flag, are therefore given by $A'_{|0\rangle}(t) = A_{|0\rangle}/(A_{|0\rangle} + A_{|2\rangle})$ and $A'_{|2\rangle}(t) = A_{|2\rangle}/(A_{|0\rangle} + A_{|2\rangle})$.

To analytically map this to our QRouter, we model amplitude damping on the fundamental CSWAP unit. This minimal module is composed of two data qubits ($q_{1,2}$) and one control qutrit (q_c), assuming identical Γ_{10} and Γ_{21} decay rates for all transmon qubits. Figures 22(a) and 22(b) illustrate the corresponding energy-level diagrams, focusing on the typical initial states: $|120\rangle$ for the erasure scheme and $|110\rangle$ for the nonerasure baseline.

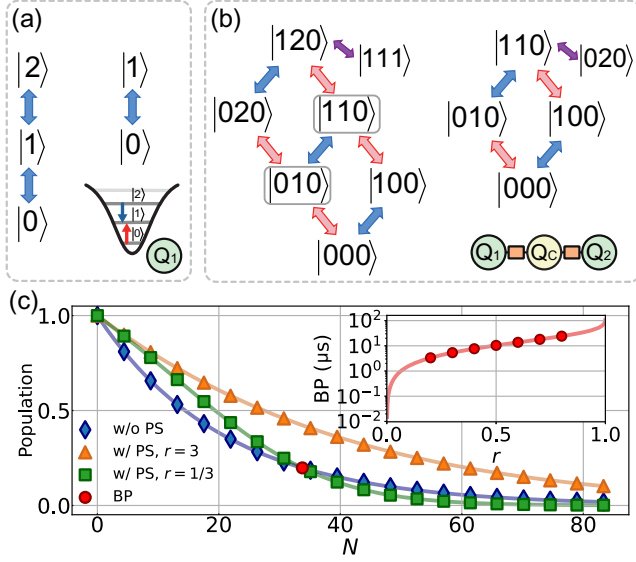


FIG. 22. Energy-level diagrams and analytical amplitude damping evaluation. (a) Energy-level diagram for a single qutrit. The erasure scheme utilizes postselection to detect state evolution into the $|1\rangle$ state caused by amplitude damping and thermal excitation. (b) Energy-level diagram for the three-qubit system. Blue arrows indicate data mutation, red arrows denote address mutation, and purple arrows represent coherent leakage. (c) Comparison of the erasure and nonerasure schemes under varying ratios of Γ_{10} to Γ_{21} ($r = \Gamma_{10}/\Gamma_{21}$). We set $\Gamma_{10} = 1/15 \mu s^{-1}$ and a CSWAP duration of 90 ns. Data points represent QuTiP simulations for CSWAP^{2N} [97], while solid lines denote the theoretical predictions. The balance point (BP, marked in red) reflects the temporal equilibrium between the two schemes when $r < 1$. The inset illustrates the relationship between the BP and r . As r increases, the BP occurs at progressively later times.

The system dynamics for the initial state $|120\rangle$ under amplitude damping are

$$\dot{A}_{|120\rangle} = -(\Gamma_{10} + \Gamma_{21})A_{|120\rangle}, \quad (G11)$$

$$\dot{A}_{|020\rangle} = \Gamma_{10}A_{|120\rangle} - \Gamma_{21}A_{|020\rangle}, \quad (G12)$$

$$\dot{A}_{|110\rangle} = \Gamma_{21}A_{|120\rangle} - 2\Gamma_{10}A_{|110\rangle}, \quad (G13)$$

$$\dot{A}_{|010\rangle} = \Gamma_{21}A_{|020\rangle} + \Gamma_{10}A_{|110\rangle} - \Gamma_{10}A_{|010\rangle}, \quad (G14)$$

$$\dot{A}_{|100\rangle} = \Gamma_{10}A_{|110\rangle} - \Gamma_{10}A_{|100\rangle}, \quad (G15)$$

$$\dot{A}_{|000\rangle} = \Gamma_{10}A_{|010\rangle} + \Gamma_{10}A_{|100\rangle}. \quad (G16)$$

Solving the coupled differential equations for the initial state $|120\rangle$ under this continuous evolution with initial

condition $A_{|120\rangle}(t=0) = 1$, the postselected amplitude of the target state evaluates to

$$A'_{|120\rangle} = \frac{A_{|120\rangle}}{A_{|120\rangle} + A_{|020\rangle} + A_{|100\rangle} + A_{|000\rangle}} = \frac{1}{e^{(\Gamma_{10} + \Gamma_{21})t} + \lambda}, \quad (G17)$$

$$\lambda = \frac{\Gamma_{21}(e^{\Gamma_{10}t} - e^{\Gamma_{21}t})}{\Gamma_{21} - \Gamma_{10}}, \quad (G18)$$

where λ accounts for the statistical effect of the postselection normalization. Conversely, the target state evolution for the nonerasure baseline experiences unmitigated amplitude damping on both the data and control registers:

$$A'_{|110\rangle} = e^{-2\Gamma_{10}t}. \quad (G19)$$

By taking the difference between Eqs. (G17) and (G19), we identify three intersection points where the two schemes yield identical populations: $t_1 = 0$, $t_2 = \infty$, and a critical balance point (BP) at t_3 :

$$t_3 = -\frac{\log(1 - \Gamma_{10}/\Gamma_{21})}{\Gamma_{10}}, \quad \text{valid when } \Gamma_{10} < \Gamma_{21}. \quad (G20)$$

This BP, illustrated in Fig. 22(c), serves as a fundamental metric for evaluating practical system setups. It dictates that the erasure technique yields a net fidelity advantage over the nonerasure baseline prior to time t_3 . Conversely, if $\Gamma_{10} > \Gamma_{21}$ (defining the ratio $r = \Gamma_{10}/\Gamma_{21} > 1$), the postselection technique strictly outperforms the baseline across all timescales. To validate these dynamics, Fig. 22(c) plots two representative scenarios ($r = 1/3$ and $r = 3$) extracted by simulating discrete CSWAP^{2N} sequences initialized in $|120\rangle$ (erasure) and $|110\rangle$ (baseline). Furthermore, the inset confirms the reliability of the analytical formula for t_3 over the parameter space $r \in (0, 1)$.

To assess the combined effects of environmental noise on system-level QRouter performance, we expand our scope beyond pure amplitude damping and conduct full numerical simulations of the RAT utilizing a Lindblad master equation approach. In this comprehensive model, full decoherence, encompassing both amplitude damping and pure dephasing, is modeled using a complete qutrit, specific formalism adapted from Ref. [96]. The vectorized density matrix evolves under the 9×9 decoherence matrix:

$$\begin{pmatrix} \rho_{00}(t) \\ \rho_{01}(t) \\ \rho_{02}(t) \\ \rho_{10}(t) \\ \rho_{11}(t) \\ \rho_{12}(t) \\ \rho_{20}(t) \\ \rho_{21}(t) \\ \rho_{22}(t) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 - e^{-\Gamma_{10}t} & 0 & 0 & 0 & V_1 \\ 0 & e^{-\Gamma_2t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & e^{-\Gamma_3t} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & e^{-\Gamma_2t} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e^{-\Gamma_{10}t} & 0 & 0 & 0 & V_2 \\ 0 & 0 & 0 & 0 & 0 & e^{-\Gamma_4t} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & e^{-\Gamma_3t} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & e^{-\Gamma_4t} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & e^{-\Gamma_{21}t} \end{pmatrix} \begin{pmatrix} \rho_{00}(0) \\ \rho_{01}(0) \\ \rho_{02}(0) \\ \rho_{10}(0) \\ \rho_{11}(0) \\ \rho_{12}(0) \\ \rho_{20}(0) \\ \rho_{21}(0) \\ \rho_{22}(0) \end{pmatrix}, \quad (\text{G21})$$

where

$$V_1 = 1 - \frac{\Gamma_{10}e^{-\Gamma_{21}t} - \Gamma_{21}e^{-\Gamma_{10}t}}{\Gamma_{10} - \Gamma_{21}}, \quad (\text{G22})$$

$$V_2 = \frac{\Gamma_{21}(e^{-\Gamma_{21}t} - e^{-\Gamma_{10}t})}{\Gamma_{10} - \Gamma_{21}}. \quad (\text{G23})$$

Here, the amplitude damping rates are $\Gamma_{21} = 1/T_1^{21}$ and $\Gamma_{10} = 1/T_1^{10}$, while the corresponding dephasing rates governing the off-diagonal coherences are denoted by $\Gamma_2 = 1/T_2^{10}$, $\Gamma_3 = 1/T_2^{21}$, and $\Gamma_4 = 1/T_2^{20}$.

To evaluate the error-erasure protocol in routing system, we must account for the intrinsic penalty of encoding the active address in the shorter-lived $|2\rangle$ state ($r = \Gamma_{10}/\Gamma_{21} < 1$). Building upon our analytical model, we anticipate that the BP will similarly emerge in the quantum router, determining the operational depth at which postselection yields a net advantage. We simulate this process using our numerical framework to pinpoint when the benefits of error filtration outweigh the penalty of accelerated decay.

We conduct RAT simulations (Appendix E) utilizing a coherence of $T_1 = 20 \mu\text{s}$ and $T_2 = 2T_1$, following the same decoherence dynamics specified in Eq. (G21). As depicted in Fig. 23(a), comparing the depth-dependent performance across representative values ($r \in \{1/3, 0.77, 3\}$) confirms the emergence of the BP when $r < 1$. Figure 23(b) further illustrates the routing error ($1 - F_{\text{RAT}}$), extracted from the depth-fitted decay curves, as a function of r . As r increases, the probability of the detrimental cascade $|2\rangle \rightarrow |1\rangle \rightarrow |0\rangle$ occurring within a single routing window diminishes (purple curve). The erasure and nonerasure curves intersect at the break-even point $r' = 0.38$ (red triangle). For any $r > r'$, the erasure protocol proves profitable. Crucially, our hardware typically operates near $r'' = 0.77$ (orange triangle), firmly positioning the device in the regime where the error-erasure architecture provides a definitive fidelity advantage. We also evaluated the erasure protocol with postselection

disabled (dark red curve). In this scenario, parity with the nonerasure baseline is reached only around $r = 1$,

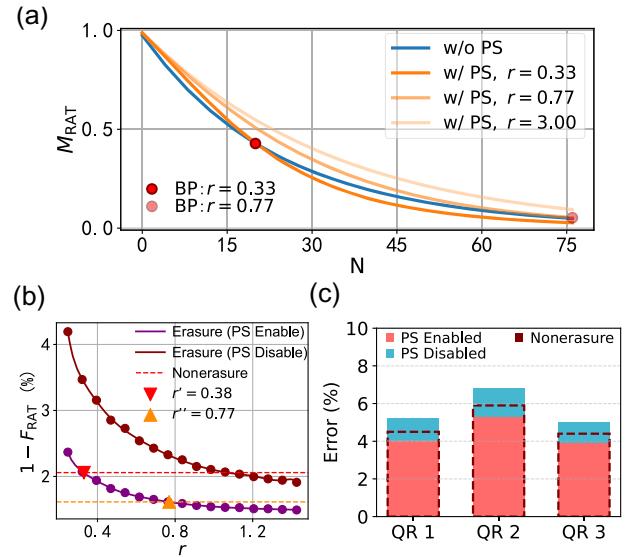


FIG. 23. Impact of the decay ratio $r = \Gamma_{10}/\Gamma_{21}$ on the error-erasure protocol. (a) RAT comparison between the error-erasure protocol (orange) and the nonerasure baseline (blue) for several representative values of r . For $r < 1$, a BP emerges that determines, for a given depth, whether enabling erasure is advantageous. (b) RAT error rate ($1 - F_{\text{RAT}}$) as a function of r . The plot compares the erasure scheme with postselection (PS) enabled (purple solid line), the erasure scheme with PS disabled (dark red solid line), and the traditional nonerasure baseline (red dashed line). The performance of the erasure scheme with PS surpasses the nonerasure baseline when $r > r' = 0.38$ (red triangle). Our system's typical operating point lies near $r'' = 0.77$ (orange triangle), safely within the regime of net fidelity improvement. Notably, if PS is disabled, the erasure scheme suffers a severe error penalty due to the fast decay of $|2\rangle$, matching the baseline only at $r = 1$. (c) Estimated decoherence errors for the three experimental QRouters (QR 1–3). The stacked bars represent the error-erasure protocol with postselection enabled (salmon) and the additional penalty incurred if postselection is disabled (cyan). The dashed dark-red boxes outline the error for the nonerasure baseline.

reflecting the inherent penalty of briefly populating the shorter-lived $|2\rangle$ level if errors cannot be subsequently discarded. Figure 23(c) quantifies the decoherence impact on the three QRouters utilizing measured experimental parameters. Consistent with theoretical predictions, the results confirm that, while the native higher-energy encoding incurs a larger raw penalty (represented by the higher total error when PS is disabled), active PS overcomes this overhead. Consequently, the fully enabled error-erasure scheme yields a net reduction in energy relaxation errors across all QRouters compared to the nonerasure baseline.

In summary, our analytical models quantify the performance impact induced by decoherence. The results clearly demonstrate that the error-erasure scheme provides a substantial net gain against amplitude damping errors, provided that the ratio r exceeds the break-even threshold. However, because this operational advantage strictly relies on the postselection technique, it imposes a penalty on the effective data yield. Therefore, realizing the protocol's full potential in scalable implementations requires a rigorous architectural trade-off: Future system designs must optimally balance the hardware decoherence profile against the inherent data-rejection costs and overhead associated with active postselection.

b. Impact of thermal excitations

While the preceding section quantified the fidelity limits imposed by downward amplitude damping, finite-temperature environments also induce upward thermal excitations, populating unwanted higher energy levels in the address register. It is worth noting that the raw readout assignment fidelity of the ground state (e.g., the F_0 values reported in Table II) can, to some extent, serve as a conservative indicator of these thermal effects. Compared to recent state-of-the-art reports [47,98–100], these initial observations indicate a non-negligible thermal load, which we attribute primarily to excess thermal radiation coupled through multiple microwave control lines, alongside imperfect cryogenic filtering and attenuation.

However, raw assignment fidelity invariably conflates the true physical populations with measurement-induced errors, such as amplifier noise and state overlap in the I - Q plane. To accurately isolate the pure physical thermal occupancy from measurement infidelity, we extracted the true equilibrium ground-state populations (P_0) by performing a physically constrained Gaussian mixture model fitting on the undriven dispersive readout I - Q histograms [101,102]. Through this deconvolution method, we determined the actual P_0 in our device to be in the range of 95.4%–98.8%. By extending our Lindblad master equation to incorporate upward thermal jump operators, this section isolates and quantifies the exact fidelity penalty imposed by this specific thermal environment, providing the second component of our incoherent error analysis.

To quantitatively assess this impact at the system level, we extended our RAT Lindblad simulations by incorporating upward-jump thermal channels across the transmon manifold (i.e., $|0\rangle \rightarrow |1\rangle$ and $|1\rangle \rightarrow |2\rangle$ transitions), sweeping the effective thermal occupancy via the measured P_0 baseline. For this analysis, we maintained $T_1 = 20 \mu\text{s}$, $T_2 = 2T_1$, and $r = 0.8$. Following the thermal model outlined in Ref. [103], the thermalized background state of a single transmon mode is approximated as

$$\rho_q = (1 - k)|0\rangle\langle 0| + k|1\rangle\langle 1|, \quad (\text{G24})$$

where k denotes the effective excited-state fraction, leaving $1 - k$ as the measurable ground-state population (P_0) following an identity operation. Within the Lindblad master equation framework, these thermal excitations are driven by the jump operator $L_\uparrow$:

$$L_\uparrow = l_\uparrow \cdot a^\dagger = \sqrt{\frac{k}{1 - k} \cdot \frac{1}{T_1}} \cdot a^\dagger, \quad (\text{G25})$$

where $a^\dagger$ serves as the appropriate creation operator.

The simulation results, presented in Fig. 24(a), confirm that, as P_0 decreases (i.e., the thermal load increases), the RAT error $1 - F_{\text{RAT}}$ rises monotonically. However, across the entire explored thermal range, the error-erasure protocol consistently outperforms the nonerasure baseline. As

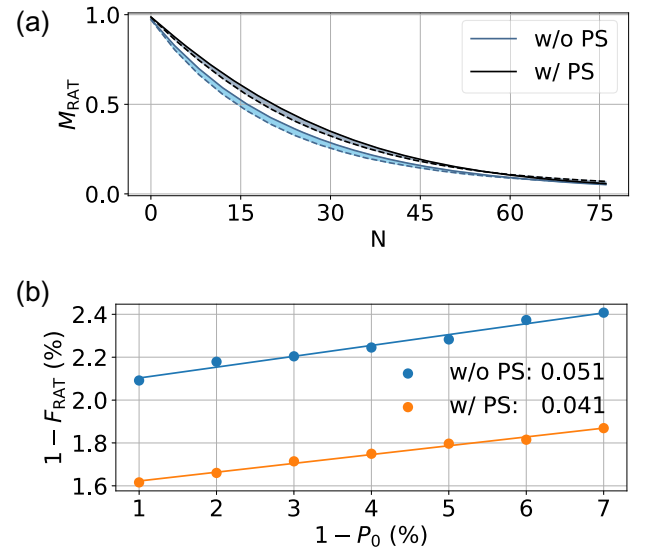


FIG. 24. Impact of thermal excitations on routing fidelity. (a) Simulated RAT curves for two thermal-population settings: solid lines correspond to a low-thermal-load environment ($P_0 = 0.99$), while dashed lines represent an elevated thermal load ($P_0 = 0.93$). (b) Comparison of the F_{RAT} decay under increasing thermal excitation (decreasing P_0). Linear fits yield degradation slopes of 0.041 for the error-erasure protocol and 0.051 for the nonerasure baseline, quantitatively demonstrating the erasure scheme's enhanced resilience to thermal fluctuations.

extracted from the linear-fit guides in Fig. 24(b), the erasure curve exhibits a shallower degradation slope (0.041) compared to the nonerasure curve (0.051).

This enhanced thermal robustness is a direct consequence of the physical mechanism described in Appendix F 1. In the standard nonerasure scheme, a thermal excitation on a $|0\rangle$ -addressed router flips the address to $|1\rangle$, causing a completely valid but erroneous reversal of the routing path. By contrast, in our erasure architecture, the identical $|0\rangle \rightarrow |1\rangle$ thermal jump merely pushes the system into the designated error flag state. Consequently, a significant fraction of these thermally induced routing faults are intercepted and filtered out via postselection. Nevertheless, at excessively high thermal excitation levels, thermally induced errors occurring directly on the data qubits, which lack a dedicated flag state, will still accumulate and eventually bottleneck the overall QRouter performance. Therefore, proactive physical mitigation remains essential. Practical measures to suppress the effective thermal occupancy seen by the chip include implementing stronger cryogenic attenuation, deploying specialized infrared shielding (e.g., Eccosorb filters) on control lines, and exploring advanced optical-control or active cooling techniques [104].

2. Coherent control errors

Having fully accounted for the incoherent errors arising from system-environment coupling, we now transition to evaluating the coherent errors induced by imperfect gate operations. Realizing a flawless $\sqrt{\text{CZ}}$ evolution in practical superconducting transmon architectures remains a formidable challenge. During gate operations, near-resonance conditions combined with waveform distortions, spectral crowding, and parasitic couplings inevitably drive unintended coherent population transfers [105]. Although both the error-erasure and nonerasure protocols rely on the same underlying $\sqrt{\text{CZ}}$ gate, they exhibit different resiliencies to these coherent errors.

To complete the theoretical foundation for our experimental error budget, this section models the three primary coherent error channels stemming from such non-ideal gate dynamics. We first analyze coherent leakage

out of the qubits' computational subspace, which is a mechanism we directly calibrate and quantify experimentally. Subsequently, we construct analytical models for parasitic qubit-qubit (QQ) SWAP interactions and unintended qubit-coupler (QC) energy exchanges. Although the exact parameters for these specific parasitic SWAPs are not individually isolated in our current hardware characterization, we perform a detailed quantitative simulation analysis of their behavior. By establishing these analytical models, we provide a mathematical framework that can accurately quantify their specific contributions to the overall error budget once these parameters are explicitly measured in future experiments. In the current context, this analysis allows us to estimate and explain the possible origins of the residual infidelities observed in the routing network.

a. Leakage from imperfect $\sqrt{\text{CZ}}$ operations

The most prominent coherent error channel during the QRouter operation is leakage from the designated computational subspace. In practical superconducting implementations, realizing a perfect population transfer between the $|11\rangle$ and $|02\rangle$ states during the $\sqrt{\text{CZ}}$ gate is inherently limited by spectral crowding and waveform distortions. Consequently, population inevitably leaks out of the target computational subspace. Unlike environmental decoherence, which acts as a unidirectional incoherent loss, this gate-induced leakage is fundamentally a coherent process. If unmitigated, these coherent errors accumulate through complex interference pathways across deep routing circuits. To accurately extract leakage contributions for our final error budget, we model this process using coherent evolution matrices.

At the single-gate level, an imperfect $\sqrt{\text{CZ}}$ interaction effectively implements a detuned or miscalibrated $i\text{SWAP}_{02 \leftrightarrow 11}(\vartheta \neq \pi, \eta)$ (A1). Assuming a negligible phase error ($\eta = 0$) for clarity, we characterize the strength of this parasitic interaction via the effective angle ϑ . To capture the full coherent dynamics necessary to calculate the resulting routing penalty, the distorted routing operation, denoted as CSWAP' , must be modeled within an extended relevant Hilbert subspace spanned by $\{|011\rangle, |020\rangle, |110\rangle, |111\rangle, |021\rangle, |120\rangle\}$. The coherent evolution matrix is explicitly given by

$$\text{CSWAP}' = \begin{pmatrix} \cos(\vartheta/2) & -i \sin(\vartheta)/2 & -\sin^2(\vartheta/2) & 0 & 0 & 0 \\ -i \sin(\vartheta)/2 & K_1 & K_2 & 0 & 0 & 0 \\ -\sin^2(\vartheta/2) & K_2 & K_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & K_1 & K_2 & -i \sin(\vartheta)/2 \\ 0 & 0 & 0 & K_2 & K_3 & -\sin^2(\vartheta/2) \\ 0 & 0 & 0 & -i \sin(\vartheta)/2 & -\sin^2(\vartheta/2) & \cos(\vartheta/2) \end{pmatrix}, \quad (\text{G26})$$

where the matrix elements are defined as $K_1 = \cos^3(\vartheta/2) - \sin^2(\vartheta/2)$, $K_2 = -i/2 \sin(\vartheta) - i/2 \cos(\vartheta/2) \sin(\vartheta)$, and $K_3 = \cos^2(\vartheta/2) - \cos(\vartheta/2) \sin^2(\vartheta)$.

Because this leakage mechanism is coherent, its accumulation over multiple routing cycles is sensitive to the relative phases dynamically acquired by the participating states. Specifically, the phase of the address qubit's auxiliary states ($|1\rangle$ and $|2\rangle$) modulates the interference between the intended computational subspace and the leakage subspace. In our experimental setup, one source of such phase shifts is the ac Stark effect induced by nonresonant microwave drives during address-flipping operations (e.g., X_π^{01} and X_π^{12}).

We numerically probe this coherent interference mechanism by injecting a controllable relative phase, denoted as $\delta\phi$, into the leakage component of the address register. As illustrated in Fig. 25(a), this parameter is tuned via a virtual Z rotation $Z_{\delta\phi}^{(lea)}$. By scanning this relative phase, we can determine whether the intended computational routing

pathway and the parasitic leakage pathway undergo constructive or destructive interference.

As demonstrated in the decoherence-free RAT simulation [Fig. 25(b)], when $\delta\phi = 0$, the interference is destructive, yielding a lower error response for both schemes (solid lines). However, when $\delta\phi = \pi$, constructive interference amplifies the coherent leakage, increasing the routing error in the nonerasure baseline (dashed blue line). In contrast, the error-erasure protocol (dashed black line) suppresses the impact of this leakage-induced constructive interference. This is because leakage events lead to an accumulation of population in the $|1\rangle$ state of the address qubit, which can subsequently be detected and flagged. These coherent errors are intercepted and discarded during the final postselection. This resilience is maintained when realistic environmental decoherence is included in the simulations, as evidenced by Fig. 25(c).

This analytical and experimental evidence indicates that while active phase engineering (e.g., $\delta\phi$ tuning via virtual Z frame updates) can mitigate leakage for standard architectures, the error-erasure protocol offers a more robust and scalable approach against leakage-induced errors. Furthermore, we acknowledge that leakage introduces coherent interference that cannot be perfectly averaged out by the RAT protocol within an isolated QRouter. Strictly speaking, the accompanying coherent oscillations deviate from a pure exponential decay. Nevertheless, for small leakage errors and in the presence of natural decoherence, these long-period oscillations are heavily damped, allowing us to accurately approximate them as a relaxation-type error. Moreover, because the delicate phase conditions required for such coherent interference are overwhelmingly lost in larger, scaled routing systems, treating leakage as a Markovian channel is physically justified. Consequently, we employ a standard exponential model to extract the effective routing error rate in our integrated analyses.

b. Qubit-qubit SWAP interactions

While the preceding section addressed population loss to noncomputational states, imperfect $\sqrt{CZ}$ operations also induce parasitic population exchanges entirely within the active multiqubit set. Specifically, near-resonance conditions and waveform distortions can trigger unintended coherent SWAPS between adjacent data and address qubits. To assess their impact, we should quantitatively model this specific error mechanism. We first consider coherent SWAPS between the qubit states $|10\rangle_{QQ} \leftrightarrow |01\rangle_{QQ}$ and $|12\rangle_{QQ} \leftrightarrow |21\rangle_{QQ}$, which are well documented in transmon architectures [105,106]. By representing these coherent SWAP-type perturbations as a unitary swapping operation $U_{SWAP}^{QQ}(\vartheta_s)$ acting in the two-qubit subspace spanned by $\{|01\rangle, |10\rangle, |12\rangle, |21\rangle\}$, we can model this error channel:

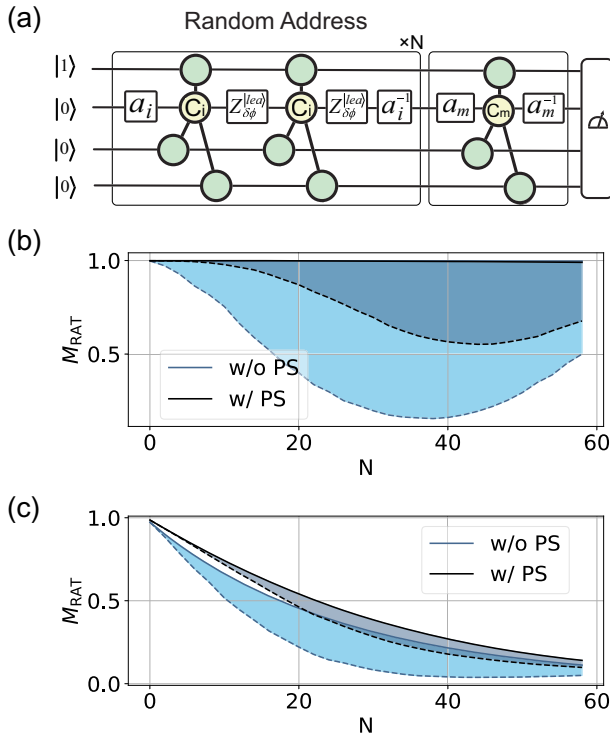


FIG. 25. RAT experiments probing the coherent interference of leakage errors. (a) Experimental setup used to control the leakage phase. The phase of the leakage component on the address qubit is deterministically varied via a virtual Z gate ($Z_{\delta\phi}^{(lea)}$). (b) RAT comparison between the error-erasure and nonerasure protocols in an idealized, decoherence-free simulation ($\vartheta = 0.98\pi$). (c) RAT comparison including realistic decoherence ($T_1 = 20 \mu s$, $T_2 = 2T_1$, $r = 0.8$). The erasure scheme robustly suppresses leakage-induced constructive interference across all phase regimes.

$$U_{\text{SWAP}}^{QQ}(\vartheta_s) = \begin{pmatrix} \cos(\vartheta_s/2) & -i \sin(\vartheta_s/2) & 0 & 0 \\ -i \sin(\vartheta_s/2) & \cos(\vartheta_s/2) & 0 & 0 \\ 0 & 0 & \cos(\vartheta_s/2) & -i \sin(\vartheta_s/2) \\ 0 & 0 & -i \sin(\vartheta_s/2) & \cos(\vartheta_s/2) \end{pmatrix}. \quad (\text{G27})$$

Here, ϑ_s is the effective SWAP angle. We assume comparable SWAP rates for both transitions, consistent with near-resonance conditions. The imperfect operation is then described by $U'_{\sqrt{\text{CZ}}} = U_{\sqrt{\text{CZ}}} \cdot U_{\text{SWAP}}^{QQ}$, where $U_{\sqrt{\text{CZ}}}$ is the operation of the ideal $\sqrt{\text{CZ}}$ gate.

To quantify the impact of these coherent SWAPs, we evaluate the routing performance at both the component level (CSWAP operations) and the system level (RAT). First, to isolate the error accumulation within the basic routing component, we simulate CSWAP^{2N} sequences. For the nonerasure baseline, we separately prepare the initial states $|100\rangle$ and $|110\rangle$ and monitor the corresponding output populations $P_{|100\rangle}$ and $P_{|110\rangle}$. For the erasure scheme, we separately prepare $|100\rangle$ and $|120\rangle$ and track $P_{|100\rangle}$ and $P_{|120\rangle}$. Performance is summarized by the average target population: $\bar{P}_{\text{ne}} = (P_{|100\rangle} + P_{|110\rangle})/2$ without erasure, and $\bar{P}_{\text{e}} = (P_{|100\rangle} + P_{|120\rangle})/2$ with erasure. In the absence of SWAP perturbations, the target population remains unity. When a representative perturbation of $\vartheta_s = 0.01\pi$ is introduced, coherent error accumulation drives distinct population dynamics, as shown in Fig. 26(c). The non-erasure baseline (dashed red curve) undergoes continuous degradation, with the target population dropping to zero at $N \approx 48$. In contrast, the erasure scheme with postselection disabled (solid purple curve) decays at a lower rate. When postselection is enabled (solid black curve), the target population loss is further mitigated.

Moving to the system level, RAT simulations performed without decoherence [Fig. 26(d)] reveal the scaling of these errors across the QRouter. As the perturbation strength increases ($\vartheta_s = 0.005\pi$ and 0.01π), the nonerasure M_{RAT} metric (dashed blue lines) decays monotonically toward zero. Conversely, the erasure protocol (solid green lines) maintains higher M_{RAT} values across the tested depth. The coherent nature of the SWAP error manifests as oscillations, distinctly visible in the $\vartheta_s = 0.01\pi$ erasure curve, which reaches a local minimum near $N = 40$ before rebounding. Across these coherent cycles, the erasure protocol limits the maximum routing failure relative to the baseline.

Mechanistically, the SWAP channel U_{SWAP}^{QQ} degrades routing by transferring population between the information qubit and the address register. In the nonerasure protocol, this transfer simultaneously scrambles the information qubit state and directly flips the computational address, inducing systematic routing faults. In the erasure protocol,

the nonadjacent address encoding ($|0\rangle$ and $|2\rangle$) provides structural isolation, making the address less susceptible to total inversion via a single exchange. At the same time,

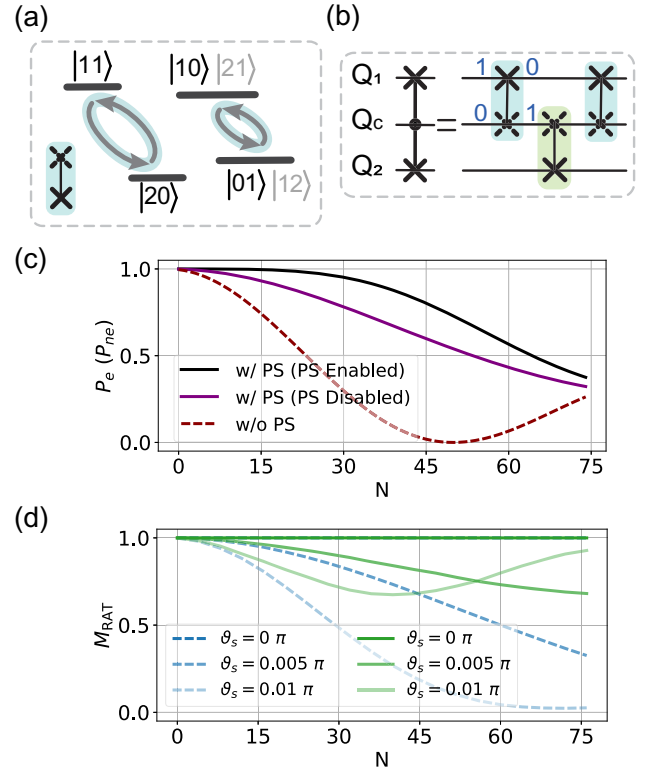


FIG. 26. Simulation study of parasitic qubit-qubit SWAP errors. (a) Energy-level schematic illustrating the intended transition alongside the parasitic $|10\rangle_{QQ} \leftrightarrow |01\rangle_{QQ}$ and $|12\rangle_{QQ} \leftrightarrow |21\rangle_{QQ}$ SWAP processes induced during an imperfect $\sqrt{\text{CZ}}$ operation. (b) Circuit-level representation demonstrating how a $|10\rangle_{QQ} \leftrightarrow |01\rangle_{QQ}$ SWAP exchanges population between the address qutrit (Q_c) and a data qubit (Q_1, Q_2). This transfers information out of the data path and corrupts the address state for subsequent routing cycles. (c) Target state population dynamics under repeated CSWAP^{2N} sequences with a fixed effective SWAP angle $\vartheta_s = 0.01\pi$. The nonerasure baseline is denoted by the dashed red curve, while the erasure protocol is shown with postselection disabled (solid purple curve) and enabled (solid black curve). (d) Simulated RAT metric (M_{RAT}) as a function of depth N for varying perturbation strengths ($\vartheta_s = 0, 0.005\pi, 0.01\pi$). Dashed blue lines and solid green lines correspond to the nonerasure and erasure protocols, respectively.

parasitic exchanges with an information qubit in the $|1\rangle$ state inherently push the address register into the intermediate $|1\rangle$ state, specifically, $|10\rangle_{QQ} \rightarrow |01\rangle_{QQ}$ (when the address is $|0\rangle$) or $|12\rangle_{QQ} \rightarrow |21\rangle_{QQ}$ (when the address is $|2\rangle$). Furthermore, postselection actively identifies and discards trials where the parasitic SWAP populates the intermediate $|1\rangle$ state. Consequently, corrupted addresses are prevented from propagating through the network, preserving a higher overall routing fidelity.

c. Qubit-coupler SWAP interactions

Beyond interqubit interactions, parasitic SWAP errors can also occur between a transmon qubit and its adjacent tunable coupler. During the $\sqrt{CZ}$ operation, the high-frequency qubit can become nearly resonant with the coupler, enabling unintended energy transfer [105]. Waveform distortion can amplify this effect, corrupting the qubit state and degrading routing performance. We model this process by an effective QC SWAP $U_{\text{SWAP}}^{\text{QC}}(\vartheta_c)$ acting in the subspace spanned by $\{|10\rangle_{\text{QC}}, |01\rangle_{\text{QC}}\}$:

$$U_{\text{SWAP}}^{\text{QC}}(\vartheta_c) = \begin{pmatrix} \cos(\vartheta_c/2) & -i \sin(\vartheta_c/2) \\ -i \sin(\vartheta_c/2) & \cos(\vartheta_c/2) \end{pmatrix}, \quad (\text{G28})$$

where ϑ_c is the effective SWAP angle. In a three-qubit chain, such interactions occur between the target qubit and its neighboring couplers. For this analysis, we assume a uniform ϑ_c across these pairs.

Applying the sequential evaluation framework established in the preceding QQ SWAP analysis, we quantified the impact of these QC interactions. Component-level simulations of the CSWAP^{2N} sequences [Fig. 27(c)] reveal distinct population dynamics under SWAP perturbations. The nonerasure baseline (dashed red curve) undergoes rapid, coherent degradation, dropping to a local minimum near 0.2 at $N = 50$. In contrast, the erasure scheme systematically maintains a higher target population, remaining above 0.7 at $N = 75$. Notably, the erasure trajectories with postselection enabled (solid black) and disabled (dashed purple) are perfectly overlapping.

Moving to the system level, RAT simulations [Fig. 27(d)] confirm these topological trends across the full routing network. As the perturbation strength increases to $\vartheta_c = 0.01\pi$ and 0.02π , the nonerasure M_{RAT} metric (dashed blue lines) decays sharply. Conversely, the erasure protocol (solid green lines) exhibits superior resilience, consistently bounding the routing failure rate below the baseline across the simulated depth.

The overlapping curves in Fig. 27(c) numerically illustrate the underlying mechanism of this resilience, which differs from the QQ SWAP case. In a QC SWAP, population is diverted entirely out of the computational qubit subspace and into the active coupler mode. Consequently, this specific error channel avoids generating a detectable event

on the auxiliary $|1\rangle$ flag state, rendering the active postselection mechanism ineffective. However, the erasure scheme inherently retains a robust structural advantage: Because the nonadjacent address encoding ($|0\rangle$ and $|2\rangle$) dictates a distinct participation profile during the intermediate routing primitive, the parasitic QC interaction corrupts a localized transition rather than directly inverting the entire address state.

This inherent structural isolation strictly limits the propagation of routing faults across the binary tree, a feature that manifests in the system-level RAT simulations [Fig. 27(d)]. As the effective QC perturbation strength increases, the nonerasure baseline undergoes a rapid,

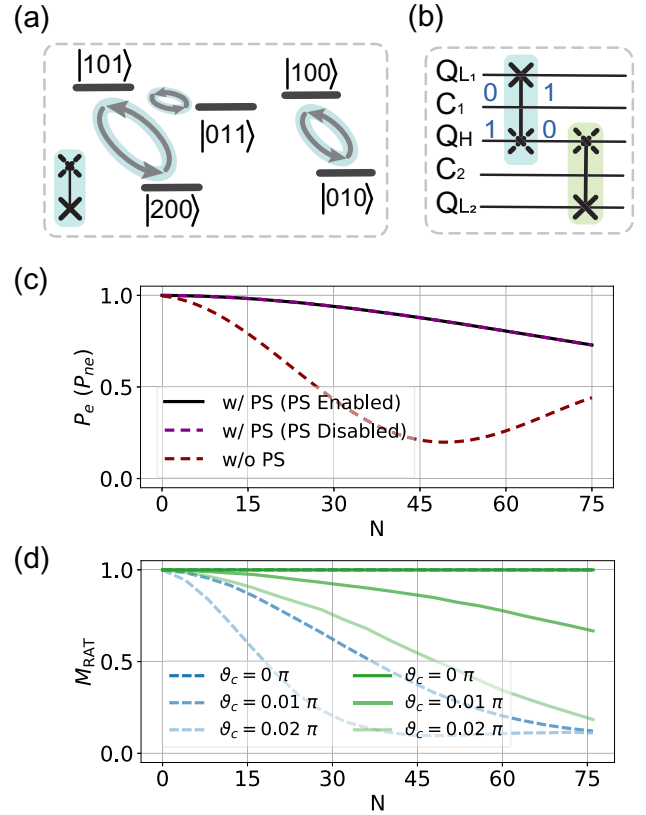


FIG. 27. Simulation study of parasitic qubit-coupler SWAP errors. (a) Energy-level schematic illustrating the parasitic qubit-coupler SWAP process induced during an imperfect $\sqrt{CZ}$ operation. This process induces $|101\rangle_{\text{QCQ}} \leftrightarrow |011\rangle_{\text{QCQ}}$ and $|100\rangle_{\text{QCQ}} \leftrightarrow |010\rangle_{\text{QCQ}}$ transitions, where the first index denotes the high-frequency qubit. (b) Circuit-level representation demonstrating how a QC swap diverts population into the coupler mode, corrupting the address state. (c) Target state population dynamics under CSWAP^{2N} sequences. The nonerasure baseline is denoted by the dashed red curve. The erasure protocol is shown with postselection disabled (dashed purple curve) and enabled (solid black curve), which perfectly overlap in this error regime. (d) Simulated RAT metric (M_{RAT}) as a function of depth N under varying QC SWAP perturbation strengths ($\vartheta_c = 0, 0.01\pi, 0.02\pi$). Dashed blue lines and solid green lines correspond to the nonerasure and erasure protocols, respectively.

monotonic collapse in the routing metric M_{RAT} . In stark contrast, the structural isolation of the erasure protocol significantly delays this fidelity decay. For instance, under a perturbation of $\vartheta_c = 0.01\pi$, the erasure scheme preserves an M_{RAT} above 0.6 at a depth of $N = 75$, whereas the nonerasure baseline drops below 0.15. This pronounced divergence confirms that even when the active postselection filter is bypassed, the topological resilience of the erasure encoding bottlenecks the proliferation of coherent errors across deep cascaded circuits.

We recognize that these distinct scaling behaviors may account for the deviations observed in our final system integration, where the erasure scheme suppresses the overall impact of unspecified coherent SWAP channels. As emphasized previously, while direct experimental characterization of these specific parasitic couplings via Floquet calibration is planned for future investigations, our current theoretical models can bound their contribution to the residual routing infidelity.

3. Error budgets for the experimental system

Having established robust analytical models for both incoherent environmental noise and coherent control imperfections, we now transition to evaluating the physical implementation. In this section, we construct error budgets to estimate the primary contributions to the infidelities observed in our experimental quantum routing system. The parameters utilized in these numerical simulations are directly extracted from the independent device calibrations detailed in Appendix A. We analyze the hardware performance across two distinct scales: first, by assessing the localized errors at the component level (individual QRouters) and, second, by evaluating the topological error accumulation within the integrated two-layer network. Through this quantitative evaluation, we isolate the underlying error channels, identifying environmental decoherence and $\sqrt{\text{CZ}}$ gate leakage as the dominant sources of routing infidelity while attributing the residual discrepancies to the aforementioned parasitic SWAP interactions.

a. Component-level error decomposition

To establish the baseline hardware performance, we first evaluate the individual QRouters at the component level. Figure 28(c) presents the RAT experimental results alongside the depth-fitted fidelities for the three QRouters. Employing the error-erasure protocol yields component-level fidelities of $95.1\% \pm 0.11\%$, $93.5\% \pm 0.15\%$, and $95.7\% \pm 0.04\%$. In comparison, the standard nonerasure baseline exhibits lower fidelities of $90.7\% \pm 0.28\%$, $90.6\% \pm 0.14\%$, and $87.5\% \pm 0.11\%$, respectively.

By incorporating the independently measured decoherence and gate parameters (Appendix A) into our numerical simulations, we decompose the total error into contributions from individual physical mechanisms, as

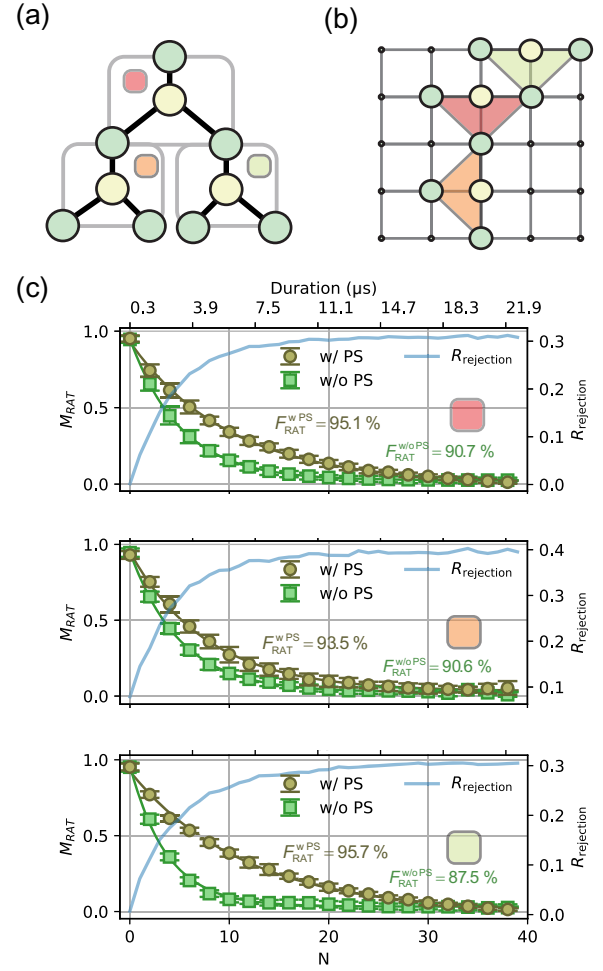


FIG. 28. QRouter distribution and RAT experimental data. (a) Layout of the two-layer quantum routing system used in the experiment. (b) Physical placement of the corresponding QRouters on the quantum processing unit layout. (c) RAT experimental results for each QRouter at the component level, evaluated under both the error-erasure and nonerasure protocols.

summarized in Fig. 29(d). The estimated thermal contributions are negligible ($< 0.2\%$) compared to decoherence and control errors, leaving environmental decoherence and coherent gate errors as the dominant factors in the fundamental error budget. We establish a theoretical upper bound for the combined error from decoherence and maximum calibrated gate leakage, indicated by the dashed boxes in Fig. 29(d). A discrepancy is observed in QRouter 3, where the error of the nonerasure baseline substantially exceeds this theoretical limit. We suspect that the parasitic SWAP-type channels account for this excess residual error as modeled in Appendixes G 2 b and G 2 c. In standard routing schemes, these coherent interactions blindly corrupt the address state and induce severe routing failures. By contrast, the nonadjacent state encoding of the error-erasure protocol affords inherent structural robustness against this specific class of coherent errors, ensuring a consistently lower net error across all three devices. As emphasized

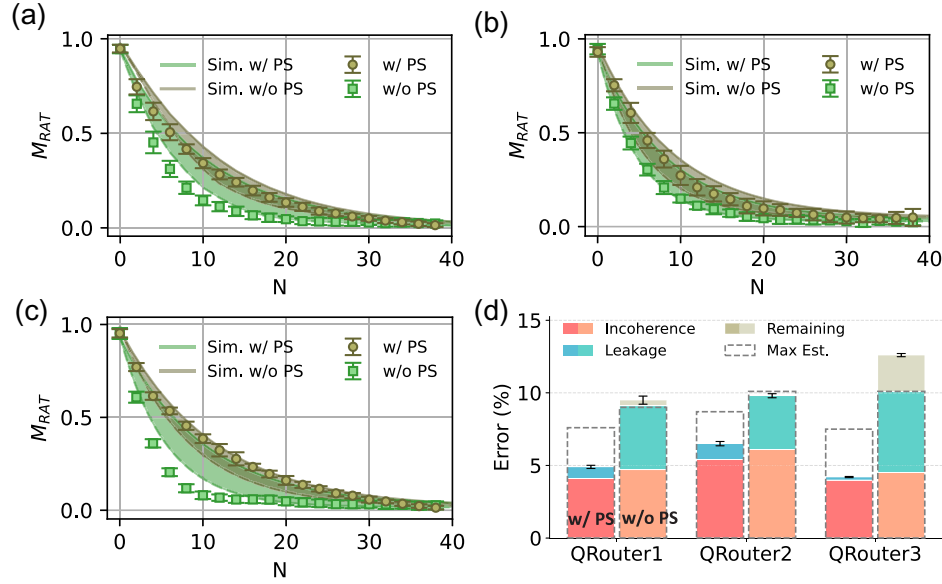


FIG. 29. The comparison between simulations and experiments and inferred error budgets for the QRouters. (a)–(c) Each QRRouter’s comparison between experimental data (markers) and numerical simulations (shaded regions). The shaded regions indicate the simulated response bounded by the minimal and maximal leakage scenarios. (d) Inferred error budgets for each QRRouter derived from the simulations. Coral bars show the incoherence contributions, while cyan bars indicate leakage errors. The dashed block gives the simulated maximum estimated error to each QRRouter, while the khaki bars show the remaining errors in the experiments.

previously, direct experimental characterization of these parasitic SWAP channels is deferred to future work, where we intend to perform fSim-type gate calibrations within a Floquet framework [62].

b. System-level error accumulation

Figure 30 presents the estimated error budget for the full two-layer routing network. The depth-fitted fidelities for the integrated two-layer system evaluate to $82.4\% \pm 0.91\%$ utilizing the error-erasure protocol and $81.1\% \pm 0.30\%$ for the nonerasure baseline.

To quantitatively analyze how errors accumulate across this integrated network, we first establish the theoretical fidelity expectation for a two-layer routing cycle [Fig. 11(b)]. We define the noisy evolution superoperator for the i th layer router as $\tilde{U}_i = \mathcal{E}_d(t) \circ \mathcal{E}_{gi} \circ \mathcal{U}_{\text{ideal},i}$, where $\mathcal{U}_{\text{ideal},i}$ is the ideal unitary transformation, $\mathcal{E}_{gi}$ represents the localized control error channel, and $\mathcal{E}_d(t)$ denotes the time-dependent decoherence channel. Within each round-trip block of our RAT circuit, the quantum state is subjected to the operations $\tilde{U}_1$, $\tilde{U}_2$, and $\tilde{U}_1$ in sequence, that is, $\tilde{U}_{\text{cycle}} = \tilde{U}_1 \tilde{U}_2 \tilde{U}_1$.

However, the physical error depth does not scale linearly with the number of operations due to two key architectural mechanisms. First, once the state is routed downwards (via $\tilde{U}_2 \tilde{U}_1$), the quantum information is localized at the target data qubits. The subsequent operation $\tilde{U}_1$ executed by the upstream router acts on an empty subspace for the information carrier, closely approximating the identity

operation. Second, through circuit parallelization, the execution of the layered routers overlaps in time, compressing the total decoherence duration to approximately

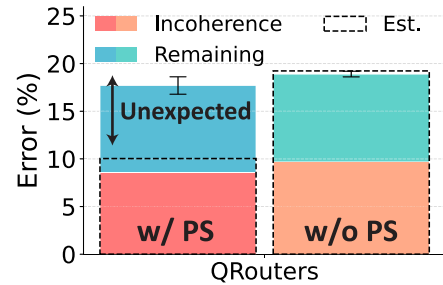


FIG. 30. System-level error budget and the coherent overhead penalty. The bar plot compares the total experimental routing error ($1 - F_{\text{RAT}}$) of the error-erasure protocol (w/ PS) and the standard nonerasure baseline (w/o PS) in the integrated two-layer network. The total error is decomposed into fundamental incoherent environmental contributions (coral) and the remaining error (cyan). The dashed outlines (Est.) denote the theoretical error expectations derived from the product of single-layer component fidelities. Notably, the standard baseline’s experimental error closely aligns with its theoretical estimate. In stark contrast, the erasure scheme exhibits a significant “unexpected” error gap between the component-level independent projection (approximately 10% error) and the integrated system-level reality (approximately 17.6% error). This explicitly labeled gap visually quantifies the system-level overhead introduced by complex multiqubit crosstalk and frequency crowding during scaled operations.

two sequential operations ($2t$). Consequently, the physical impact of a complete round-trip cycle on the initial state ρ_0 can be simplified to a single directional transmission: $\tilde{U}_{\text{cycle}}(\rho_0) \approx \tilde{U}_2 \tilde{U}_1(\rho_0)$.

Given this effective error depth of two, the theoretical average fidelity of the two-layer network can be explicitly formulated by compounding the individual component fidelities along the two symmetric routing paths. This overall two-layer fidelity is estimated as $F_{\text{two-layer}} = F_{\text{QR1}} \times (F_{\text{QR2}} + F_{\text{QR3}})/2$, where F_{QR1} represents the fidelity of the first-layer router and F_{QR2} and F_{QR3} correspond to the fidelities of the two possible target routers in the second layer. For the standard nonerasure baseline, this expected fidelity evaluates to $0.907 \times (0.906 + 0.875)/2 \approx 80.8\%$. For the error-erasure protocol, the expected fidelity evaluates to $0.951 \times (0.935 + 0.957)/2 \approx 90.0\%$.

Comparing these theoretical projections with our integrated RAT measurements reveals the operational overhead of the respective schemes. The experimental nonerasure fidelity (81.1%) is consistent with its theoretical projection (80.8%), indicating that errors in the standard baseline accumulate primarily independently. In contrast, the experimental erasure fidelity (82.4%) falls notably below its theoretical independent-component projection (90.0%).

This numerical discrepancy explicitly quantifies the system-level penalty introduced by the erasure scheme. The supplementary microwave pulses required for address manipulation into the $|2\rangle$ encoding exacerbate frequency collisions, classical microwave crosstalk, and residual quantum ZZ couplings across the expanded multiqubit network. As reflected in the error budget in Fig. 30, while the erasure protocol reduces the fundamental decoherence contribution from 9.8% to 8.6%, these unmitigated crosstalk phenomena cause the “remaining error” (the total experimental error excluding decoherence) to remain comparable at 9.0% (compared to 9.1% for the nonerasure scheme). Therefore, while the nonadjacent encoding provides robust immunity against localized SWAP errors, mitigating these global coherent interactions remains an essential requirement for scaling high-fidelity quantum routing networks.

4. Hardware requirements for high-fidelity QRouters

Having decomposed the underlying infidelities of our current physical implementation, we now pivot to a forward-looking perspective. By leveraging the comprehensive error models validated in the preceding sections, we project the specific hardware specifications required to cross the threshold of high-fidelity quantum routing. This analysis not only quantifies the practical scalability of our proposed error-erasure scheme, but also establishes a quantitative road map for next-generation hardware development.

To isolate the fundamental performance limits imposed by environmental decoherence, we assume ideal coherent control for these projections. Under this assumption, we

perform a two-dimensional parameter sweep over the primary relaxation and dephasing channels. Specifically, we simulate the QRouter’s RAT fidelity by varying the longitudinal relaxation time T_1 and the ratio of transverse to longitudinal coherence times, T_2/T_1 . The relaxation time of the second excited state is fixed at a constant ratio $r = 0.8$.

Figures 31(a) and 31(b) present the simulated fidelity landscapes for the standard nonerasure baseline and our

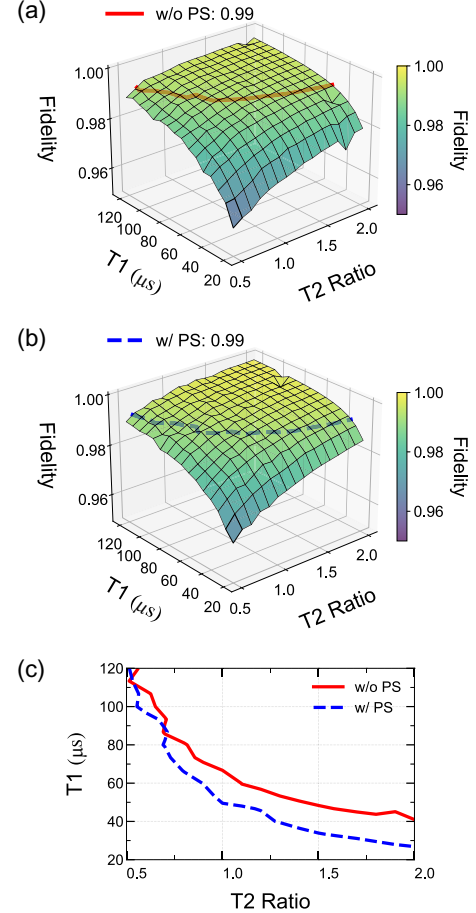


FIG. 31. Decoherence requirements for high-fidelity QRouter performance. Simulated RAT fidelity landscapes as a function of the longitudinal relaxation time T_1 and the coherence ratio T_2/T_1 for (a) the standard nonerasure baseline and (b) the proposed error-erasure protocol. The simulations assume ideal coherent control to isolate the fundamental impact of decoherence, with the higher-level relaxation ratio fixed at $r = \Gamma_{10}/\Gamma_{21} = 0.8$. The 99% fidelity thresholds are highlighted by the red solid line and the blue dashed line on the respective surfaces. (c) A 2D projection comparing the 99% fidelity boundaries. The region above each curve indicates the parameter space satisfying the $\mathcal{F} \geq 0.99$ criterion. The error-erasure scheme (blue dashed line) consistently demands less stringent T_1 and T_2 hardware requirements than the nonerasure approach (red solid line), demonstrating its capability to significantly lower the technological threshold for high-fidelity quantum routing. The minor fluctuations in the data arise from finite sampling.

proposed error-erasure protocol, respectively. As anticipated, the overall fidelity scales monotonically with concurrent improvements in both T_1 and the T_2/T_1 ratio (T_2 ratio). To establish concrete hardware targets for scalable architectures, we extract the parameter contours corresponding to a critical 99% operational fidelity threshold ($\mathcal{F} \geq 0.99$).

A direct comparison of these threshold requirements is explicitly delineated in the projected 2D phase diagram in Fig. 31(c). The parameter space above the respective boundary curves represents the regime where 99% fidelity is physically achievable. The data demonstrate that the error-erasure scheme (blue dashed line) relaxes the decoherence constraints demanded by the nonerasure baseline (red solid line) across the entire parameter space. For instance, operating at a standard coherence limit of $T_2 = T_1$, achieving 99% routing fidelity without erasure necessitates a robust relaxation time of $T_1 \approx 60 \mu\text{s}$. Notably, as T_2 decreases, the relaxation of the T_1 constraint afforded by the erasure scheme diminishes. This behavior is attributed to the accumulation of dephasing errors, which the erasure protocol is unable to mitigate. Consequently, this imposes a requirement on the T_2 coherence time. In stark contrast, our structural erasure protocol substantially lowers this hardware barrier to approximately $50 \mu\text{s}$. Ultimately, these results furnish specific materials and fabrication targets for future high-performance QRouter implementations.

-
- [1] V. Giovannetti, S. Lloyd, and L. Maccone, *Quantum random access memory*, *Phys. Rev. Lett.* **100**, 160501 (2008).
 - [2] 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).
 - [3] A. Paler, O. Oumarou, and R. Basmadjian, *Parallelizing the queries in a bucket-brigade quantum random access memory*, *Phys. Rev. A* **102**, 032608 (2020).
 - [4] R. Asaka, K. Sakai, and R. Yahagi, *Quantum random access memory via quantum walk*, *Quantum Sci. Technol.* **6**, 035004 (2021).
 - [5] 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).
 - [6] S. Jaques and A. G. Rattew, *Qram: A survey and critique*, [arXiv:2305.10310](https://arxiv.org/abs/2305.10310).
 - [7] D. K. Weiss, S. Puri, and S. M. Girvin, *Quantum random access memory architectures using 3d superconducting cavities*, *PRX Quantum* **5**, 020312 (2024).
 - [8] 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(R) (2012).
 - [9] L. K. Grover, *Quantum mechanics helps in searching for a needle in a haystack*, *Phys. Rev. Lett.* **79**, 325 (1997).
 - [10] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, *Quantum machine learning*, *Nature (London)* **549**, 195 (2017).
 - [11] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio, and P. J. Coles, *Challenges and opportunities in quantum machine learning*, *Nat. Comput. Sci.* **2**, 567 (2022).
 - [12] B. Duan, J. Yuan, C.-H. Yu, J. Huang, and C.-Y. Hsieh, *A survey on hhl algorithm: From theory to application in quantum machine learning*, *Phys. Lett. A* **384**, 126595 (2020).
 - [13] R. Steijl, *Quantum algorithms for nonlinear equations in fluid mechanics*, in *Quantum Computing and Communications*, edited by Y. Zhao (IntechOpen, Rijeka, 2020) Chap. 2.
 - [14] I. M. Georgescu, S. Ashhab, and F. Nori, *Quantum simulation*, *Rev. Mod. Phys.* **86**, 153 (2014).
 - [15] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, *Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets*, *Nature (London)* **549**, 242 (2017).
 - [16] Z.-Y. Chen *et al.*, *Enabling large-scale and high-precision fluid simulations on near-term quantum computers*, *Comput. Methods Appl. Mech. Eng.* **432**, 117428 (2024).
 - [17] Y. Kotukh, *Cryptanalysis of the system based on word problems using logarithmic signatures*, *Radiotekhnika*, <https://api.semanticscholar.org/CorpusID:245324463> (2021).
 - [18] 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).
 - [19] Z.-Y. Chen, C. Xue, S.-M. Chen, B.-H. Lu, Y.-C. Wu, J.-C. Ding, S.-H. Huang, and G.-P. Guo, *Quantum approach to accelerate finite volume method on steady computational fluid dynamics problems*, *Quantum Inf. Process.* **21**, 137 (2022).
 - [20] V. Giovannetti, S. Lloyd, and L. Maccone, *Architectures for a quantum random access memory*, *Phys. Rev. A* **78**, 052310 (2008).
 - [21] C. T. Hann, G. Lee, S. M. Girvin, and L. Jiang, *Resilience of quantum random access memory to generic noise*, *PRX Quantum* **2**, 020311 (2021).
 - [22] L. Zhou, S. Yang, Y. X. Liu, C. P. Sun, and F. Nori, *Quantum zeno switch for single-photon coherent transport*, *Phys. Rev. A* **80**, 062109 (2009).
 - [23] C. Gonzalez-Ballester, E. Moreno, F. J. Garcia-Vidal, and A. Gonzalez-Tudela, *Nonreciprocal few-photon routing schemes based on chiral waveguide-emitter couplings*, *Phys. Rev. A* **94**, 063817 (2016).
 - [24] M.-T. Cheng, X.-S. Ma, J.-Y. Zhang, and B. Wang, *Single photon transport in two waveguides chirally coupled by a quantum emitter*, *Opt. Express* **24**, 19988 (2016).
 - [25] L. Zhou, L.-P. Yang, Y. Li, and C. P. Sun, *Quantum routing of single photons with a cyclic three-level system*, *Phys. Rev. Lett.* **111**, 103604 (2013).
 - [26] Y. T. Zhu and W. Z. Jia, *Single-photon quantum router in the microwave regime utilizing double superconducting*

- resonators with tunable coupling, *Phys. Rev. A* **99**, 063815 (2019).
- [27] J. Lu, Z. H. Wang, and L. Zhou, *T-shaped single-photon router*, *Opt. Express* **23**, 22955 (2015).
- [28] J. Lu, L. Zhou, L.-M. Kuang, and F. Nori, *Single-photon router: Coherent control of multichannel scattering for single photons with quantum interferences*, *Phys. Rev. A* **89**, 013805 (2014).
- [29] Z. Wang, Y. Wu, Z. Bao, Y. Li, C. Ma, H. Wang, Y. Song, H. Zhang, and L. Duan, *Experimental realization of a deterministic quantum router with superconducting quantum circuits*, *Phys. Rev. Appl.* **15**, 014049 (2021).
- [30] Y.-l. Ren, S.-l. Ma, J.-k. Xie, X.-k. Li, M.-t. Cao, and F.-l. Li, *Nonreciprocal single-photon quantum router*, *Phys. Rev. A* **105**, 013711 (2022).
- [31] Y. Li, Z. Bao, Z. Wang, Y. Wu, J. Wang, J. Yang, H. Xiong, Y. Song, H. Zhang, and L. Duan, *Quantum switch for itinerant microwave single photons with superconducting quantum circuits*, *Phys. Rev. Appl.* **21**, 044030 (2024).
- [32] L. Du, Y.-T. Chen, J.-H. Wu, and Y. Li, *Nonreciprocal interference and coherent photon routing in a three-port optomechanical system*, *Opt. Express* **28**, 3647 (2020).
- [33] G. S. Agarwal and S. Huang, *Optomechanical systems as single-photon routers*, *Phys. Rev. A* **85**, 021801(R) (2012).
- [34] 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).
- [35] K. Bartkiewicz, A. Černoč, and K. Lemr, *Implementation of an efficient linear-optical quantum router*, *Sci. Rep.* **8**, 13480 (2018).
- [36] N. E. Palaiodimopoulos, S. Ohler, M. Fleischhauer, and D. Petrosyan, *Chiral quantum router with Rydberg atoms*, *Phys. Rev. A* **109**, 032622 (2024).
- [37] C. Miao, S. Léger, Z. Li, G. Lee, L. Jiang, and D. I. Schuster, *Implementation of a quantum addressable router using superconducting qubits*, *PRX Quantum* **6**, 040335 (2025).
- [38] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, J. Majer, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Charge-insensitive qubit design derived from the cooper pair box*, *Phys. Rev. A* **76**, 042319 (2007).
- [39] R. Barends, J. Kelly, A. Megrant, D. Sank, E. Jeffrey, Y. Chen, Y. Yin, B. Chiaro, J. Mutus, C. Neill, P. O'Malley, P. Roushan, J. Wenner, T. C. White, A. N. Cleland, and J. M. Martinis, *Coherent Josephson qubit suitable for scalable quantum integrated circuits*, *Phys. Rev. Lett.* **111**, 080502 (2013).
- [40] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Tunable coupling scheme for implementing high-fidelity two-qubit gates*, *Phys. Rev. Appl.* **10**, 054062 (2018).
- [41] S. E. Rasmussen, K. S. Christensen, and N. T. Zinner, *Controllable two-qubit swapping gate using superconducting circuits*, *Phys. Rev. B* **99**, 134508 (2019).
- [42] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, *Quantum supremacy using a programmable superconducting processor*, *Nature (London)* **574**, 505 (2019).
- [43] W. Feng and D. W. Wang, *Quantum fredkin gate based on synthetic three-body interactions in superconducting circuits*, *Phys. Rev. A* **101**, 062312 (2020).
- [44] T. Roy, S. Hazra, S. Kundu, M. Chand, M. P. Patankar, and R. Vijay, *Programmable superconducting processor with native three-qubit gates*, *Phys. Rev. Appl.* **14**, 014072 (2020).
- [45] X. Gu, J. Fernández-Pendás, P. Vikstål, T. Abad, C. Warren, A. Bengtsson, G. Tancredi, V. Shumeiko, J. Bylander, G. Johansson, and A. F. Kockum, *Fast multi-qubit gates through simultaneous two-qubit gates*, *PRX Quantum* **2**, 040348 (2021).
- [46] C. W. Warren, J. Fernández-Pendás, S. Ahmed, T. Abad, A. Bengtsson, J. Biznárová, K. Debnath, X. Gu, C. Križan, A. Osman *et al.*, *Extensive characterization and implementation of a family of three-qubit gates at the coherence limit*, *npj Quantum Inf.* **9**, 44 (2023).
- [47] D. Gao *et al.*, *Establishing a new benchmark in quantum computational advantage with 105-qubit zuhongzhi 3.0 processor*, *Phys. Rev. Lett.* **134**, 090601 (2025).
- [48] F. Shen *et al.*, *A bucket-brigade quantum random access memory*, *Nat. Phys.* **22**, 745 (2026).
- [49] A. Barenco, C. H. Bennett, R. Cleve, D. P. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. A. Smolin, and H. Weinfurter, *Elementary gates for quantum computation*, *Phys. Rev. A* **52**, 3457 (1995).
- [50] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, England, 2010).
- [51] S. Zhang *et al.*, *Realizing scalable conditional operations through auxiliary energy levels*, *Sci. China Inf. Sci.* (to be published).
- [52] A. Kubica, A. Haim, Y. Vaknin, H. Levine, F. Brandão, and A. Retzker, *Erasure qubits: Overcoming the T_1 limit in superconducting circuits*, *Phys. Rev. X* **13**, 041022 (2023).
- [53] S. Vittal, P. Das, and M. Qureshi, *Eraser: Towards adaptive leakage suppression for fault-tolerant quantum computing*, [10.1145/3613424.3614251](https://arxiv.org/abs/10.1145/3613424.3614251) (2023).
- [54] K. S. Chou, T. Shemma, H. McCarrick, T.-C. Chien, J. D. Teoh, P. Winkel, A. Anderson, J. Chen, J. C. Curtis, S. J. de Graaf *et al.*, *A superconducting dual-rail cavity qubit with erasure-detected logical measurements*, *Nat. Phys.* **20**, 1454 (2024).
- [55] S. J. de Graaf, S. H. Xue, B. J. Chapman, J. D. Teoh, T. Tsunoda, P. Winkel, J. W. Garmon, K. M. Chang, L. Frunzio, S. Puri *et al.*, *A mid-circuit erasure check on a dual-rail cavity qubit using the joint-photon number-splitting regime of circuit QED*, *npj Quantum Inf.* **11**, 1 (2025).
- [56] H. Levine *et al.*, *Demonstrating a long-coherence dual-rail erasure qubit using tunable transmons*, *Phys. Rev. X* **14**, 011051 (2024).
- [57] D. K. Weiss, S. Puri, and S. M. Girvin, *Quantum random access memory architectures using superconducting cavities*, [arXiv:2310.08288](https://arxiv.org/abs/2310.08288).
- [58] B. Foxen *et al.* (Google AI Quantum), *Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms*, *Phys. Rev. Lett.* **125**, 120504 (2020).
- [59] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C.

- McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Realization of high-fidelity cz and zz-free iswap gates with a tunable coupler*, *Phys. Rev. X* **11**, 021058 (2021).
- [60] F. Arute *et al.*, *Observation of separated dynamics of charge and spin in the Fermi-Hubbard model*, [arXiv:2010.07965](https://arxiv.org/abs/2010.07965).
- [61] C. Neill, T. McCourt, X. Mi, Z. Jiang, M. Niu, W. Mruczkiewicz, I. Aleiner, F. Arute, K. Arya, J. Atalaya *et al.*, *Accurately computing the electronic properties of a quantum ring*, *Nature (London)* **594**, 508 (2021).
- [62] J. A. Gross, É. Genois, D. M. Debroy, Y. Zhang, W. Mruczkiewicz, Z.-P. Cian, and Z. Jiang, *Characterizing coherent errors using matrix-element amplification*, *npj Quantum Inf.* **10**, 123 (2024).
- [63] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, *Simple pulses for elimination of leakage in weakly nonlinear qubits*, *Phys. Rev. Lett.* **103**, 110501 (2009).
- [64] J. Butscher, S. Krinner, and A. Wallraff, *Shaping of fast flux pulses for two-qubit gates: Inverse filtering* (2018).
- [65] E. A. Sete, J. M. Martinis, and A. N. Korotkov, *Quantum theory of a bandpass purcell filter for qubit readout*, *Phys. Rev. A* **92**, 012325 (2015).
- [66] Y. Sunada, S. Kono, J. Ilves, S. Tamate, T. Sugiyama, Y. Tabuchi, and Y. Nakamura, *Fast readout and reset of a superconducting qubit coupled to a resonator with an intrinsic purcell filter*, *Phys. Rev. Appl.* **17**, 044016 (2022).
- [67] L.-L. Guo, P. Duan, S. Zhang, X.-X. Yang, C. Zhang, L. Du, H.-F. Zhang, H.-R. Tao, T.-L. Wang, Z.-L. Jia, Z.-Y. Chen, and G.-P. Guo, *Universal scalable characterization and correction of pulse distortions in controlled quantum systems*, *Phys. Rev. Appl.* **21**, 064060 (2024).
- [68] 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).
- [69] V. Negîrneac, H. Ali, N. Muthusubramanian, F. Battistel, R. Sagastizabal, M. S. Moreira, J. F. Marques, W. J. Vlothuizen, M. Beekman, C. Zachariadis, N. Haider, A. Bruno, and L. DiCarlo, *High-fidelity controlled-z gate with maximal intermediate leakage operating at the speed limit in a superconducting quantum processor*, *Phys. Rev. Lett.* **126**, 220502 (2021).
- [70] Y. Wang, S. Zhang, T. Sun, Z. Zhao, X. Xu, X. Zhuang, H. Liu, C. Xue, P. Duan, Y. Wu, Z. Chen, and G. Guo, *Hardware-efficient quantum random access memory design with a native gate set on superconducting platforms*, *Adv. Quantum Technol.* **8**, 2400519 (2025).
- [71] Z.-Y. Chen, C. Xue, Y.-J. Wang, T.-P. Sun, H.-Y. Liu, X.-N. Zhuang, M.-H. Dou, T.-R. Zou, Y. Fang, Y.-C. Wu, and G.-P. Guo, *Efficient and error-resilient data access protocols for a limited-sized quantum random access memory*, [arXiv:2303.05207](https://arxiv.org/abs/2303.05207).
- [72] X.-Y. Yang, H.-F. Zhang, L. Du, H.-R. Tao, L.-L. Guo, T.-L. Wang, Z.-L. Jia, W.-C. Kong, Z.-Y. Chen, P. Duan *et al.*, *Fast, universal scheme for calibrating microwave crosstalk in superconducting circuits*, *Appl. Phys. Lett.* **125** (2024).
- [73] P. Wang, B.-H. Lu, T.-L. Wang, S. Zhang, Z.-Y. Chen, H.-F. Zhang, R.-Z. Zhao, X.-Y. Yang, Z.-A. Zhao, Z.-Z. Zhang, X.-X. Song, Y.-C. Wu, P. Duan, and G.-P. Guo, *Spectator leakage elimination in cz gates via tunable coupler interference on a superconducting quantum processor*, [arXiv:2507.14531](https://arxiv.org/abs/2507.14531).
- [74] J. Preskill, *Quantum Computing in the NISQ era and beyond*, *Quantum* **2**, 79 (2018).
- [75] J. W. Z. Lau, K. H. Lim, H. Shrotriya, and L. C. Kwek, *Nisq computing: Where are we and where do we go?*, *AAPPS Bull.* **32**, 27 (2022).
- [76] U. Khalid, J. U. Rehman, S. N. Paing, H. Jung, T. Q. Duong, and H. Shin, *Quantum network engineering in the nisq age: Principles, missions, and challenges*, *IEEE Network* **38**, 112 (2024).
- [77] Y.-J. Wang, T.-P. Sun, X.-N. Zhuang, X.-F. Xu, H.-Y. Liu, C. Xue, Y.-C. Wu, Z.-Y. Chen, and G.-P. Guo, *Refined criteria for gram error suppression via efficient large-scale gram simulator*, *Phys. Rev. Appl.* **25**, 044069 (2026).
- [78] P. Mundada, G. Zhang, T. Hazard, and A. Houck, *Suppression of qubit crosstalk in a tunable coupling superconducting circuit*, *Phys. Rev. Appl.* **12**, 054023 (2019).
- [79] P. Zhao, K. Linghu, Z. Li, P. Xu, R. Wang, G. Xue, Y. Jin, and H. Yu, *Quantum crosstalk analysis for simultaneous gate operations on superconducting qubits*, *PRX Quantum* **3**, 020301 (2022).
- [80] S. Rodrigo, S. Abadal, C. G. Almudéver, and E. Alarcón, *Modelling short-range quantum teleportation for scalable multi-core quantum computing architectures*, [10.1145/3477206.3477461](https://arxiv.org/abs/10.1145/3477206.3477461) (2021).
- [81] X. Li, Y. Ma, J. Han, T. Chen, Y. Xu, W. Cai, H. Wang, Y. P. Song, Z.-Y. Xue, Z. Q. Yin, and L. Sun, *Perfect quantum state transfer in a superconducting qubit chain with parametrically tunable couplings*, *Phys. Rev. Appl.* **10**, 054009 (2018).
- [82] M. Christandl, N. Datta, T. C. Dorlas, A. Ekert, A. Kay, and A. J. Landahl, *Perfect transfer of arbitrary states in quantum spin networks*, *Phys. Rev. A* **71**, 032312 (2005).
- [83] B. R. Johnson, *Controlling photons in superconducting electrical circuits*, ProQuest Dissertations and Theses, 2011, <https://www.proquest.com/dissertations-theses/controlling-photons-superconducting-electrical/docview/884221402/se-2>.
- [84] Z. Chen, *Metrology of Quantum Control and Measurement in Superconducting Qubits* (University of California, Santa Barbara, 2018), <https://api.semanticscholar.org/CorpusID:196194358>.
- [85] Z. Yan, Y.-R. Zhang, M. Gong, Y. Wu, Y. Zheng, S. Li, C. Wang, F. Liang, J. Lin, Y. Xu, C. Guo, L. Sun, C.-Z. Peng, K. Xia, H. Deng, H. Rong, J. Q. You, F. Nori, H. Fan, X. Zhu, and J.-W. Pan, *Strongly correlated quantum walks with a 12-qubit superconducting processor*, *Science* **364**, 753 (2019).
- [86] N. J. Glaser, F. A. Roy, I. Tsitsilin, L. Koch, N. Bruckmoser, J. Schirck, J. H. Romeiro, G. B. P. Huber, F.

- Wallner, M. Singh, G. Krylov, A. Marx, L. Södergren, C. M. F. Schneider, M. Werninghaus, and S. Filipp, *Sensitivity-adapted closed-loop optimization for high-fidelity controlled-z gates in superconducting qubits*, [arXiv:2412.17454](#).
- [87] B. Chiaro and Y. Zhang, *Active leakage cancellation in single qubit gates*, *Phys. Rev. Lett.* **135**, 130601 (2025).
- [88] E. Genois, N. J. Stevenson, N. Goss, I. Siddiqi, and A. Blais, *Quantum optimal control of superconducting qubits based on machine-learning characterization*, [arXiv:2410.22603](#).
- [89] E. Genois, N. J. Stevenson, N. Goss, I. Siddiqi, and A. Blais, *Quantum optimal control of superconducting qubits based on machine-learning characterization*, *Phys. Rev. Appl.* **24**, 034073 (2025).
- [90] 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. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *High-on-off-ratio beam-splitter interaction for gates on bosonically encoded qubits*, *PRX Quantum* **4**, 020355 (2023).
- [91] X.-L. Li, Z. Tao, K. Yi, K. Luo, L. Zhang, Y. Zhou, S. Liu, T. Yan, Y. Chen, and D. Yu, *Hardware-efficient and fast three-qubit gate in superconducting quantum circuits*, *Front. Phys.* **19**, 51205 (2024).
- [92] W. Shi, N. K. Kundu, M. R. McKay, and R. Malaney, *Error-mitigated quantum random access memory*, [arXiv:2403.06340](#).
- [93] 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).
- [94] B. Nachman, M. Urbanek, W. A. de Jong, and C. W. Bauer, *Unfolding quantum computer readout noise*, *npj Quantum Inf.* **6**, 84 (2020).
- [95] F. S. Richards, *A method of maximum-likelihood estimation*, *J. R. Stat. Soc. Ser. B* **23**, 469 (1961).
- [96] A. Morvan, V. V. Ramasesh, M. S. Blok, J. M. Kreikebaum, K. O'Brien, L. Chen, B. K. Mitchell, R. K. Naik, D. I. Santiago, and I. Siddiqi, *Qutrit randomized benchmarking*, *Phys. Rev. Lett.* **126**, 210504 (2021).
- [97] J. Johansson, P. Nation, and F. Nori, *Qutip 2: A Python framework for the dynamics of open quantum systems*, *Comput. Phys. Commun.* **184**, 1234 (2013).
- [98] *Suppressing quantum errors by scaling a surface code logical qubit*, *Nature (London)* **614**, 676 (2023).
- [99] Z. Meng *et al.*, *Simulating unsteady flows on a superconducting quantum processor*, *Commun. Phys.* **7**, 349 (2024).
- [100] *Quantum error correction below the surface code threshold*, *Nature (London)* **638**, 920 (2025).
- [101] T. Uusnäkki, T. Mörstedt, W. Teixeira, M. Rasola, and M. Möttönen, *Experimental realization of a quantum heat engine based on dissipation-engineered superconducting circuits*, *Nat. Commun.* **17**, 6054 (2025).
- [102] W. Teixeira, T. Mörstedt, A. Viitanen, H. Kivijärvi, A. Gunyhó, M. Tiiri, S. Kundu, A. Sah, V. Vadimov, and M. Möttönen, *Many-excitation removal of a transmon qubit using a single-junction quantum-circuit refrigerator and a two-tone microwave drive*, *Sci. Rep.* **14**, 13755 (2024).
- [103] F. Battistel, B. M. Varbanov, and B. M. Terhal, *Hardware-efficient leakage-reduction scheme for quantum error correction with superconducting transmon qubits*, *PRX Quantum* **2**, 030314 (2021).
- [104] N. Li, Y.-H. Li, D.-J. Fan, L.-C. Han, Y. Xu, J. Lin, C. Guo, D.-D. Li, M. Gong, S.-K. Liao *et al.*, *Optical transmission of microwave control signal towards large-scale superconducting quantum computing*, *Opt. Express* **32**, 3989 (2024).
- [105] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Realization of high-fidelity cz and zz-free iswap gates with a tunable coupler*, *Phys. Rev. X* **11**, 021058 (2021).
- [106] B. Foxen *et al.* (Google AI Quantum), *Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms*, *Phys. Rev. Lett.* **125**, 120504 (2020).