Fast Flux-Activated Leakage Reduction for Superconducting Quantum Circuits

Nathan Lacroix,[1,2] Luca Hofele,[1,2] Ants Remm,[1,*] Othmane Benhayoune-Khadraoui,[3] Alexander McDonald,[3] Ross Shillito,[3] Stefania Lazar,[1,†] Christoph Hellings,[1,2] François Swiadek,[1,2] Dante Colao-Zanuz,[1,2] Alexander Flasby,[1,2,4] Mohsen Bahrami Panah,[1,2,4] Michael Kerschbaum,[1,2,4] Graham J. Norris,[1,2] Alexandre Blais,[3,5] Andreas Wallraff,[1,2,4] and Sebastian Krinner[1,†]

[1]*Department of Physics, ETH Zurich, CH-8093 Zurich, Switzerland*
[2]*Quantum Center, ETH Zurich, 8093 Zurich, Switzerland*
[3]*Institut Quantique and Département de Physique, Université de Sherbrooke, Sherbrooke J1K2R1, Quebec, Canada*
[4]*ETH Zurich—PSI Quantum Computing Hub, Paul Scherrer Institute, CH-5232 Villigen, Switzerland*
[5]*Canadian Institute for Advanced Research, Toronto, Ontario, Canada*

(Received 26 September 2023; accepted 12 February 2025; published 25 March 2025)

Quantum computers will require quantum error correction to reach the low error rates necessary for solving problems that surpass the capabilities of conventional computers. One of the dominant errors limiting the performance of quantum error correction codes across multiple technology platforms is leakage out of the computational subspace arising from the multilevel structure of qubit implementations. Here, we present a resource-efficient universal leakage reduction unit for superconducting qubits using parametric flux modulation. This operation removes leakage down to our measurement inaccuracy of 7×10^{-4} in approximately 50 ns with a low error of $2.5(1) \times 10^{-3}$ on the computational subspace, thereby reaching durations and fidelities comparable to those of single-qubit gates. We demonstrate that using the leakage reduction unit in repeated weight-two stabilizer measurements reduces the total number of detected errors in a scalable fashion to close to what can be achieved using leakage-rejection methods that do not scale. Our approach does not require additional control electronics or on-chip components and is applicable to both auxiliary and data qubits. These benefits make our method particularly attractive for mitigating leakage in large-scale quantum error correction circuits, a crucial requirement for the practical implementation of fault-tolerant quantum computation.

DOI: 10.1103/PhysRevLett.134.120601

Quantum error correction (QEC) protocols [1,2] offer a promising path to close the gap between physical error rates achievable on quantum computing devices and the low logical error rates necessary to solve computational problems that are intractable for conventional computers [3]. However, the efficient suppression of logical errors typically relies on the assumption that physical errors occur independently in space and time, and that physical systems used as qubits have no more than two levels [4,5]. Yet leakage, a phenomenon in which an excitation leaves the two-level computational subspace used to perform quantum operations, is a source of highly correlated errors, likely due to its long-lived character over many quantum error correction cycles [6,7]. Consequently, leakage poses a significant challenge to achieve fault-tolerant quantum computation [8–14]. Leakage occurs across a wide range of technology platforms, including trapped-ion systems [15,16], semiconductor quantum dots [17], neutral atoms [18], and superconducting circuits. For superconducting circuits, leakage arises predominantly from control inaccuracies in single-qubit gate operations [19–22], two-qubit gate operations [23–26], and readout [27–29].

To mitigate the effect of leakage, so-called leakage reduction units (LRUs) have been proposed to convert leakage errors into Pauli-like errors in the computational subspace at regular intervals during the computation [8]. Most proposals for LRUs consist of involved teleportation circuits [8,9], of auxiliary qubit resets in combination with periodic swaps between auxiliary and data qubits [10], or of dedicated filter circuits that allow for the dissipation of only the leakage state [30,31], all of which add overhead to quantum error correction protocols or to the device architecture. Therefore, initial leakage-mitigation schemes for superconducting qubits [6,32] focused on removing leakage using a multilevel reset operation [6,33–35]. However, such an operation also resets the states of the computational subspace [36] and can therefore only be applied to auxiliary qubits at the end of an error correction cycle. Such a scheme was recently extended to remove leakage of data qubits using an additional leakage-swap gate followed by a second auxiliary-qubit reset operation [7]. It is only very

*Present address: Atlantic Quantum, Cambridge, Massachusetts 02139, USA.
†Present address: Zurich Instruments, CH-8005 Zurich, Switzerland.

recently that a universal LRU, a single operation that can be applied to data and auxiliary qubits, has been demonstrated based on the proposal of Ref. [14] using a second-order microwave-activated coupling (previously used in Ref. [33] for qutrit reset) between the leakage state and the readout resonator [37].

In this Letter, we present a resource-efficient, fast, universal flux-activated LRU that couples the leakage state of a flux-tunable transmon qubit [38] to its readout resonator. The engineered coupling, resulting from a parametric qubit frequency modulation, is a first-order transition and therefore the LRU can be fast [39], reaching durations and fidelities comparable to those of single-qubit gates. Additionally, unlike the method used in Ref. [7], it can also be performed when the readout resonator frequency is higher than the qubit frequency, a common architectural choice [34,40–42] to avoid complications from readout-induced transitions that arise when the readout resonator frequency is lower than the qubit frequency [29,43].

We realize the LRU by coupling the first leakage state of a flux-tunable transmon qubit, $|f\rangle$, to a readout resonator-Purcell filter system that is strongly coupled to a feedline acting as a dissipative environment, as illustrated with a simplified energy-level diagram in Fig. 1(a) and a full circuit diagram of the system in Fig. 1(b). For clarity, we consider a single readout mode in the energy-level diagram although we have two hybridized readout resonator-Purcell filter modes [44]; see Appendix A of the Supplemental Material [45] for all relevant device parameters. We realize the coupling by applying a flux pulse $\phi(t)$ with amplitude ϕ_a and modulation frequency ω_m to the flux line of the qubit [35,51,52], as depicted in Fig. 1(c) and detailed in Appendix B.

Because the qubit is operated at its upper flux-noise-insensitive bias point (i.e., with a dc flux bias of $\phi_{dc} = 0$), the flux modulation results in a qubit frequency modulation with leading-order sidebands at $\pm 2\omega_m$ [39,51,52]. This modulation lowers the average qubit frequency, with the difference from its value at the upper flux-noise-insensitive bias point defined as the modulation amplitude ω_a, which we infer in the experiments as described in Appendix C. When the high-frequency sideband [top dashed blue line in Fig. 1(a)] of $|f0\rangle$ is resonant with $|e1\rangle$, population is transferred from $|f0\rangle$ to $|e1\rangle$. Here, the first state label corresponds to the state of the transmon qubit and the second to the Fock state of the resonator mode. This resonance condition is fulfilled when

$$2\omega_m = |\bar{\omega}_{ef} - \omega_r| \approx |\bar{\omega}_{ge} + \alpha - \omega_r| = |\bar{\Delta} + \alpha|, \quad (1)$$

where ω_{ge} (ω_{ef}) is the transition frequency from $|g\rangle$ to $|e\rangle$ ($|e\rangle$ to $|f\rangle$) at the bias point, $\alpha = \omega_{ef} - \omega_{ge}$ is the transmon anharmonicity, and $\Delta = \omega_{ge} - \omega_r$ is the detuning between the qubit frequency and the transition frequency ω_r of the readout resonator mode. The overline symbol $\bar{\omega}_{kl}$ indicates the transition frequency from $|k\rangle$ to $|l\rangle$ time-averaged over the duration of the modulation pulse, i.e., $\bar{\omega}_{ge} \approx \omega_{ge} - \omega_a$ and $\bar{\omega}_{ef} \approx \omega_{ef} - \omega_a$.

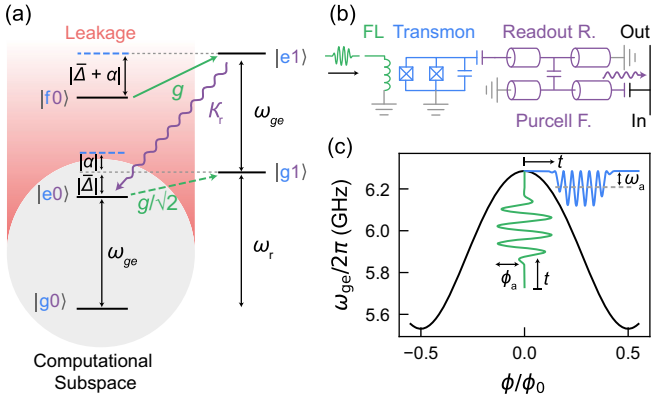

FIG. 1. Concept of the leakage reduction unit. (a) Energy-level diagram of a transmon qubit with transition frequency ω_{ge} and anharmonicity α coupled to a resonator mode with transition frequency ω_r. Sidebands generated by a modulation of the qubit frequency are indicated with dashed blue lines. These sidebands enable the coupling of the leaked state $|f0\rangle$ to the state $|e1\rangle$, which decays back to the computational subspace state $|e0\rangle$. See main text for details. (b) Circuit diagram of the elements required for the implementation of an LRU for a flux-tunable transmon qubit (blue), a flux line (green), and a readout resonator-Purcell filter system (purple) coupled to a feedline (black). (c) Modulating the magnetic flux in the SQUID loop of the qubit (Appendix A) using a Gaussian-filtered modulation pulse $\Phi(t)$ (green line) results in a modulation of the qubit transition frequency ω_{ge} (blue line), leading to a parametric coupling to the resonator mode.

The resulting coupling strength g depends on the modulation amplitude [39,52]

$$g = \sqrt{2} g_{qr} J_1(\omega_a/2\omega_m), \quad (2)$$

with g_{qr} the coupling strength between the qubit and the readout resonator, and $J_1(\cdot)$ the first Bessel function of the first kind [53]. After the leaked population has been transferred to the readout resonator, the coupled system decays back to the computational state $|e0\rangle$ on the timescale of the effective decay rate of the resonator mode $\kappa_r/2\pi = 16.4$ MHz (Appendix A).

When the flux modulation pulse is tuned to satisfy the resonance condition of the LRU [Eq. (1)], an analogous parametric transition from $|e0\rangle$ to $|g1\rangle$ with a coupling strength $g/\sqrt{2}$ is detuned by only $|\alpha|$, as illustrated in Fig. 1(a). It is essential to suppress residual driving of this transition because it affects the computational subspace. To achieve this, we ensure that the bandwidth of the $|e0\rangle$ sideband is much smaller than $|\alpha|$ by filtering the rising and falling edges of the flux pulse using a Gaussian kernel of width $\sigma = 5$ ns; see Appendix B. Furthermore, to suppress Purcell decay of the high-frequency $|e0\rangle$ sideband, we use a device architecture with an individual Purcell filter for each qubit and readout circuit parameters that ensure that the transmission through the readout resonator-Purcell filter system is suppressed at a detuning $|\alpha|$ from resonance [54,55].

We first identify suitable operating points for the LRU, i.e., pairs of (ω_m, ω_a) satisfying the resonance condition. Specifically, we prepare the qubit in $|f\rangle$, apply a flux modulation pulse with a fixed duration of $t = 100\,\text{ns} > 1/\kappa_r$, and subsequently measure the transmon qubit with three-state readout [33]. We sweep the modulation frequency and the modulation amplitude and identify four resonances yielding low $|f\rangle$ population after 100 ns; see the four slanted spectral lines in Fig. 2(a). The high-modulation-frequency resonance doublet corresponds to the parametric transition from $|f0\rangle$ to $|e1\rangle$ from the qubit into either one of the two resonator-Purcell filter modes. We use the highest frequency resonance of this doublet to implement the LRU. The two lower-frequency resonances are induced by a second harmonic process; see Appendix D for details. For each of the spectral lines, the modulation frequency required to reach resonance increases linearly as a function of the modulation amplitude with a slope of approximately $1/2$ as the mean qubit frequency is shifted by ω_a during flux modulation; see also Eq. (1).

In a second calibration step, we fix the modulation amplitude and frequency, and vary the duration of the pulse τ to extract the minimal duration τ_{LRU} of the pulse yielding the lowest population of $|f\rangle$. For the operating point $O = (\omega_m/2\pi = 564, \omega_a/2\pi = 128)$ MHz [purple circle in Fig. 2(a)], the achieved parametric coupling g is large with respect to $\kappa_r/4$, which results in underdamped oscillations [35] of the $|f\rangle$ population with a first minimum of $6(1) \times 10^{-4}$ after a pulse duration of only 34.5 ns (54.5 ns when accounting for the rising- and falling-edge buffers as detailed in Appendix B); see Fig. 2(b). The exhaustive depletion of the population in $|f\rangle$, down to the single-shot readout inaccuracy of approximately 7×10^{-4} (Appendix A), demonstrates the high effectiveness of the LRU. Ultimately, we expect the residual $|f\rangle$ population to be limited by the thermal occupation of the readout resonator ($\sim 2 \times 10^{-4}$), in which case the LRU can drive the transition in the opposite direction, i.e., from $|e1\rangle$ to $|f0\rangle$. The population dynamics of all three transmon eigenstates are in good agreement with master-equation simulations [solid lines in Fig. 2(b)]; see Appendix E for details.

To gain further insight into the relationship between the modulation amplitude and the duration of the LRU, we measure the time evolution of the transmon population for four modulation amplitudes [purple markers in Fig. 2(a)] and extract the corresponding τ_{LRU}. As expected from simulations, τ_{LRU} decreases approximately as $1/\omega_a$ in our parameter regime; see purple markers in Fig. 2(c).

Although leakage errors can significantly impede the performance of QEC protocols, they are infrequent, typically occurring at a rate of 0.1%–1% per qubit per QEC cycle [6,55,56]. Consequently, in practice the LRU acts on a state within the computational subspace most of the time, and it is therefore imperative to minimize its effect on this subspace. To this end, we extract the qubit lifetime T_1, Ramsey coherence time T_2^*, and pure dephasing time

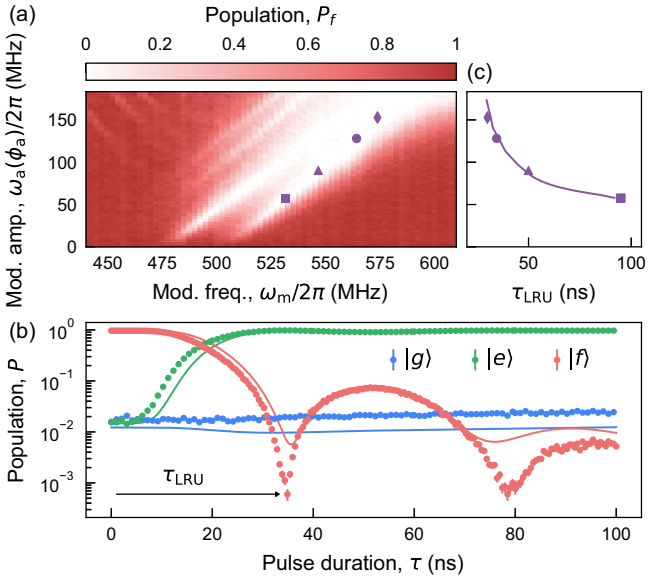

FIG. 2. Calibration of the leakage reduction unit. (a) $|f\rangle$ population after a 100-ns-long LRU as a function of the modulation frequency ω_m and the modulation amplitude ω_a. Four operating points for an LRU are indicated with purple symbols. (b) Experimentally measured (dots) and simulated (lines) time evolution of the population of $|g\rangle$ (blue), $|e\rangle$ (green), $|f\rangle$ (red) of the transmon qubit initially prepared in $|f\rangle$ when applying the modulation pulse with the parameters indicated by a purple circle in (a). Error bars correspond to the standard error from 25 000 single-shot measurements. (c) Measured (markers) and simulated (line) duration of the leakage reduction unit, τ_{LRU}, as a function of the modulation amplitude ω_a.

$T_\phi = 2T_1T_2^*/(2T_1 - T_2^*)$ when applying the modulation pulse as a function of the modulation amplitude; see Fig. 3(a). We observe a reduction in lifetime and pure dephasing time, which we mostly attribute to population loss due to the repeated crossing of two-level defect modes [57,58] and an increased sensitivity to flux noise, respectively; see Appendix F for details. Ongoing efforts to reduce the number of strongly coupled two-level defect modes on superconducting devices [59] are expected to help mitigate these losses. In future work, a detailed investigation of T_1 as a function of modulation frequency could help distinguish losses due to interactions with two-level defect modes and from losses due to the off-resonantly driven $|e0\rangle - |g1\rangle$ transition. For all operating points, we observe that $T_1 < T_\phi$, which indicates that errors on the computational subspace are mostly T_1-limited.

In addition, to extract the average error of the LRU on the computational subspace, we perform interleaved randomized benchmarking [60] in which the LRU is benchmarked against a perfect identity operation. For the operating point O [the coherence times of which are indicated by the blue arrow in Fig. 3(a)], we obtain an average gate error of 0.25(1)%; see Fig. 3(b). In comparison, the error for an idle operation of the same duration as the LRU is about 0.1%, showing that performing the LRU causes errors on the

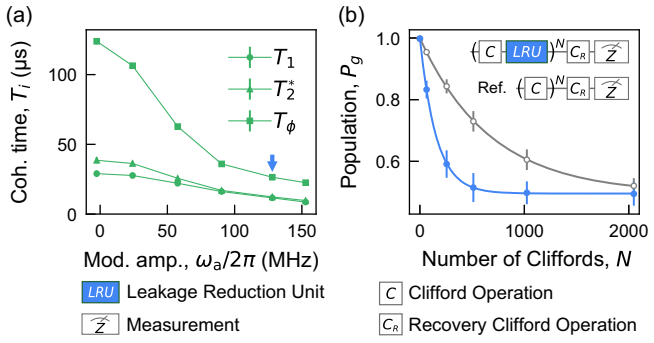

FIG. 3. Effect of the LRU on the computational subspace. (a) Effective lifetime T_1 (circles), dephasing time T_2^* (triangles), and pure dephasing time T_ϕ (squares) as a function of the modulation amplitude ω_a. The modulation amplitude $\omega_a = 128$ MHz used for (b) is indicated with a blue arrow. (b) Interleaved randomized benchmarking (blue) performed as shown in the upper quantum circuit diagram and reference randomized benchmarking (gray) performed as shown in the lower quantum circuit diagram.

computational subspace of the same order of magnitude as coherence-limited single-qubit gates. The measured error is in good agreement with a calculated coherence limit [61] of 0.24%, which takes into account the reduction in T_1 and T_2^* during the LRU. We choose the operating point O for all further experiments because it provides a good compromise between LRU duration and errors on the computational subspace.

To demonstrate the benefits of using an LRU in QEC experiments despite the small errors it causes on the computational subspace, we perform repeated cycles of a weight-two Z-type stabilizer measurement [55] with and without LRU; see Fig. 4(a) for the full quantum circuit diagram. The two data qubits (red dots) are initialized in one of the four Z-basis eigenstates and the parity of the state is mapped onto the auxiliary qubit (green dot) as shown in Fig. 4(a). An LRU can be applied to the auxiliary qubit, which is subsequently measured using single-shot three-level readout [55]. Scheduling the LRU just before the midcircuit auxiliary-qubit readout, rather than after, prevents potential dephasing and Stark shift of the auxiliary qubit that could otherwise be induced by the residual photon population in the readout resonator at the end of the LRU ($n_{\mathrm{res}} \lesssim 0.2$). The entire stabilizer cycle of a fixed duration of 0.7 μs is repeated m times.

We find that when the LRU is applied, the accumulation of population in $|f\rangle$ of the auxiliary qubit after 50 cycles, averaged over the four data-qubit input states, is reduced by approximately a factor of 10 to $\sim 3.5 \times 10^{-3}$ [green dots in Fig. 4(b)], compared to $\sim 3.4 \times 10^{-2}$ when the LRU is not applied (gray dots). We observe a background residual leakage of about 2×10^{-3} on average after a single cycle, higher than the minimum $|f\rangle$ population reported in Fig. 2(b). We attribute this residual leakage to a frequency collision leading to state-dependent readout-induced leakage; see Appendix G. This frequency collision can be

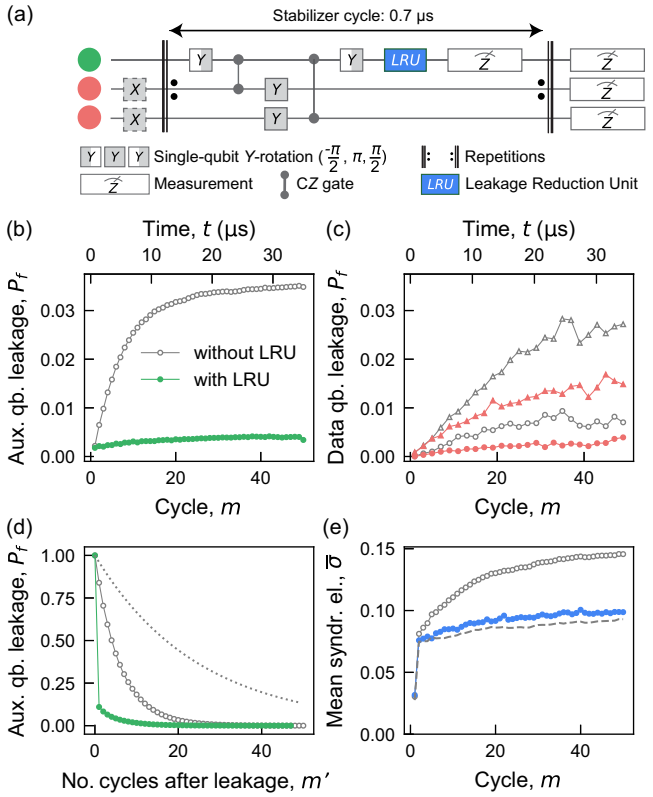

FIG. 4. Integration of the LRU in a weight-two Z-type stabilizer measurement. (a) Stabilizer circuit with one auxiliary qubit (green dot) and two data qubits (red). (b) Leakage of the auxiliary qubit with (blue) and without (gray) LRU in each stabilizer cycle as a function of the number of executed cycles m. (c) Leakage of data qubits D1 (circles) and D2 (triangles) with (red) and without (gray) the LRU. (d) Leakage lifetime in the stabilizer circuit with (green) and without (gray) the LRU, and in a separate characterization measurement (gray dotted line). (e) Mean syndrome element $\bar{\sigma}$ with the LRU (blue dots), with neither the LRU nor leakage rejection (gray dots), and with leakage rejection instead of the LRU (dashed gray line).

avoided in the future by adapting the design frequencies of the qubits and readout resonators. When considering solely the accumulation of leakage in addition to this background value, we calculate that the LRU leads to a twentyfold reduction in leakage accumulation. Moreover, we find that the application of the LRU to the auxiliary qubit also reduces leakage accumulation on data qubits, as shown in Fig. 4(c). We attribute this effect to a decrease in leakage transport [6,7] that arises only when the auxiliary qubit is in $|f\rangle$. The differences in leakage between the two data qubits are currently not understood. We provide a proof-of-principle experiment to further suppress leakage accumulation on data qubits; see Appendix H. We intend to integrate such data-qubit LRUs in quantum error correction circuits in future work, which has the potential to further reduce leakage-induced errors.

Furthermore, we extract the effective lifetime of a leakage event in the stabilizer circuit by postselecting on

runs in which leakage is detected on the auxiliary qubit and counting the average number of cycles in which the auxiliary qubit is consecutively read out in $|f\rangle$ after the initial leakage event. We find that the LRU achieves the goal of reducing the leakage lifetime on the auxiliary qubit to close to a single cycle of the repeated stabilizer measurements, while the lifetime is on the order of six cycles when no LRU is used; see Fig. 4(d). In comparison, the $|f\rangle$-state lifetime of the auxiliary qubit in an independent T_1 measurement (dashed gray line) is much longer, approximately 24.6 cycles, which provides further evidence for leakage transport away from the auxiliary qubit during the repeated stabilizer measurement. From the reduction of the leakage lifetime, we infer that both space- and time-correlated errors caused by leakage are suppressed [7].

To further investigate the impact of the LRU on the total number of detected errors by the stabilizer, we construct the error syndrome in each cycle $\sigma_m = (1 - s_m \times s_{m-1})/2$ from the current (m) and the previous $(m-1)$ measured stabilizer values s, with $\sigma = 1$ indicating an error and $\sigma = 0$ indicating no error, respectively [55,62]. When averaging over all circuit runs and possible data-qubit input states, we find that applying the LRU reduces the mean error syndrome value $\bar{\sigma}$ from $\sim$0.15 to $\sim$0.1 after 50 cycles [Fig. 4(e)]. Moreover, applying the LRU significantly decreases the probability of observing correlated syndrome elements separated by more than one cycle; see Appendix I for details. These results suggest that the LRU suppresses leakage-induced time-correlated errors and consequently reduces the total number of errors by approximately 33%. To further assess the performance of our approach, we compare the use of the LRU to a leakage-rejection method [dashed gray line in Fig. 4(e)] that discards experimental runs in which a leakage event on the auxiliary qubit is detected using three-level readout [55,56]. Note that this method is not suited for large-scale QEC experiments because the amount of experimental runs left after leakage-rejection decreases exponentially with the number of QEC cycles and qubits. By contrast, employing the LRU results in nearly the same performance as the leakage-rejection method, with the key benefit of scalability.

In summary, we have demonstrated a fast leakage reduction unit based on parametric flux modulation taking only $\sim$50 ns, which effectively removes leakage down to our qubit readout error of 7×10^{-4}. Moreover, it is high-fidelity, causing only an error of $2.5(1) \times 10^{-3}$ on the computational subspace. Our LRU thus approaches durations and fidelities comparable to those of single-qubit gates. We successfully integrated the LRU in a weight-two stabilizer measurement, thereby significantly improving its performance. Simulations show that the ability to suppress leakage will become even more relevant when executing large-scale quantum error correction circuits [7]. Furthermore, we show how the LRU protocol can be extended to mitigate leakage from even higher-excited transmon states using a sequence of modulated pulses with the appropriate modulation frequencies. As demonstrated in Appendix H, such a multilevel LRU can remove population from both the state $|f\rangle$ and the third-excited transmon state $|h\rangle$ in approximately 120 ns. In the future, the presented LRU can also be applied to data qubits (Appendix H) and thereby further reduce the total number of errors.

The LRU introduced in this work offers several advantages compared to other recent developments in leakage suppression [7,37]. First, our LRU is 4 times faster than the one presented in Ref. [37], resulting in a reduction of idling errors on all qubits, which often constitute a substantial fraction of the total error budget [32,63]. Second, the modulation pulses are generated by the same electronics that also generate pulses for the two-qubit gates, avoiding additional cost and complexity of the experimental setup. Finally, employing parametric coupling for realizing the LRU enables its use in a wide range of qubit-frequency configurations. Hence, this work showcases that the flux-activated parametric LRU is a promising approach to effectively suppress leakage in large-scale error correction circuits, which is an essential requirement for the practical implementation of fault-tolerant quantum computation.

Acknowledgments—The authors are grateful for valuable discussions with Markus Müller and for feedback on the manuscript by Ilya Besedin. The team in Zurich acknowledges financial support by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the U.S. Army Research Office Grant W911NF-16-1-0071, by the EU Flagship on Quantum Technology H2020-FETFLAG-2018-03 project 820363 OpenSuperQ, by the National Centre of Competence in Research Quantum Science and Technology (NCCR QSIT), a research instrument of the Swiss National Science Foundation (SNSF), by the SNFS R'equip Grant 206021-170731, by the EU program H2020-FETOPEN project 828826 Quromorphic, and by ETH Zurich. S. K. acknowledges financial support from Fondation Jean-Jacques and Felicia Lopez-Loreta and the ETH Zurich Foundation. The work in Sherbrooke was undertaken thanks in part to funding from NSERC, Canada First Research Excellence Fund and ARO W911NF-18-1-0411, the Ministère de l'Économie et de l'Innovation du Québec, and the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator. M. M. acknowledges support by the U.S. Army Research Office Grant W911NF-16-1-0070.

The authors declare no competing interests.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government.

Data availability—All data is available from the corresponding authors upon reasonable request.

[1] D. Gottesman, An introduction to quantum error correction and fault-tolerant quantum computation, in Proceedings of Symposia in Applied Mathematics, edited by S. J. Lomonaco, Jr., (2010) [arXiv:0904.2557].

[2] B. M. Terhal, Quantum error correction for quantum memories, Rev. Mod. Phys. **87**, 307 (2015).

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

[4] E. Knill, R. Laflamme, and W. H. Zurek, Resilient quantum computation: Error models and thresholds, Proc. R. Soc. A **454**, 365 (1998).

[5] D. Aharonov and M. Ben-Or, Fault-tolerant quantum computation with constant error rate, arXiv:quant-ph/9906129.

[6] M. McEwen *et al.*, Removing leakage-induced correlated errors in superconducting quantum error correction, Nat. Commun. **12**, 1761 (2021).

[7] K. C. Miao *et al.*, Overcoming leakage in quantum error correction, Nat. Phys. **19**, 1780 (2023).

[8] P. Aliferis and B. M. Terhal, Fault-tolerant quantum computation for local leakage faults, Quantum Inf. Comput. **7**, 139 (2007).

[9] A. G. Fowler, Coping with qubit leakage in topological codes, Phys. Rev. A **88**, 042308 (2013).

[10] J. Ghosh and A. G. Fowler, Leakage-resilient approach to fault-tolerant quantum computing with superconducting elements, Phys. Rev. A **91**, 020302(R) (2015).

[11] M. Suchara, A. W. Cross, and J. M. Gambetta, Quantum Inf. Comput. **15**, 997 (2015).

[12] C. C. Bultink, T. E. O'Brien, R. Vollmer, N. Muthusubramanian, M. W. Beekman, M. A. Rol, X. Fu, B. Tarasinski, V. Ostroukh, B. Varbanov, A. Bruno, and L. DiCarlo, Protecting quantum entanglement from leakage and qubit errors via repetitive parity measurements, Sci. Adv. **6**, eaay3050 (2020).

[13] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. DiCarlo, and B. M. Terhal, Leakage detection for a transmon-based surface code, npj Quantum Inf. **6**, 102 (2020).

[14] 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).

[15] R. Stricker, D. Vodola, A. Erhard, L. Postler, M. Meth, M. Ringbauer, P. Schindler, T. Monz, M. Müller, and R. Blatt, Experimental deterministic correction of qubit loss, Nature (London) **585**, 207 (2020).

[16] D. Hayes, D. Stack, B. Bjork, A. C. Potter, C. H. Baldwin, and R. P. Stutz, Eliminating leakage errors in hyperfine qubits, Phys. Rev. Lett. **124**, 170501 (2020).

[17] R. W. Andrews, C. Jones, M. D. Reed, A. M. Jones, S. D. Ha, M. P. Jura, J. Kerckhoff, M. Levendorf, S. Meenehan, S. T. Merkel, A. Smith, B. Sun, A. J. Weinstein, M. T. Rakher, T. D. Ladd, and M. G. Borselli, Quantifying error and leakage in an encoded Si/SiGe triple-dot qubit, Nat. Nanotechnol. **14**, 747 (2019).

[18] C. Wu, S. Kumar, Y. Kan, D. Komisar, Z. Wang, S. I. Bozhevolnyi, and F. Ding, Room-temperature on-chip orbital angular momentum single-photon sources, Sci. Adv. **8**, eabk3075 (2022).

[19] 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).

[20] Z. Chen *et al.*, Measuring and suppressing quantum state leakage in a superconducting qubit, Phys. Rev. Lett. **116**, 020501 (2016).

[21] M. Werninghaus, D. J. Egger, F. Roy, S. Machnes, F. K. Wilhelm, and S. Filipp, Leakage reduction in fast superconducting qubit gates via optimal control, npj Quantum Inf. **7**, 14 (2021).

[22] S. Lazăr, Q. Ficheux, J. Herrmann, A. Remm, N. Lacroix, C. Hellings, F. Swiadek, D. C. Zanuz, G. J. Norris, M. B. Panah, A. Flasby, M. Kerschbaum, J. Besse, C. Eichler, and A. Wallraff, Calibration of drive nonlinearity for arbitrary-angle single-qubit gates using error amplification, Phys. Rev. Appl. **20**, 024036 (2023).

[23] R. Barends *et al.*, Digitized adiabatic quantum computing with a superconducting circuit, Nature (London) **534**, 222 (2016).

[24] 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).

[25] M. C. Collodo, J. Herrmann, N. Lacroix, C. K. Andersen, A. Remm, S. Lazar, J.-C. Besse, T. Walter, A. Wallraff, and C. Eichler, Implementation of conditional phase gates based on tunable ZZ interactions, Phys. Rev. Lett. **125**, 240502 (2020).

[26] V. Negirneac, 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).

[27] D. Sank, Z. Chen, M. Khezri, J. Kelly, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. Fowler *et al.*, Measurement-induced state transitions in a superconducting qubit: Beyond the rotating wave approximation, Phys. Rev. Lett. **117**, 190503 (2016).

[28] R. Shillito, A. Petrescu, J. Cohen, J. Beall, M. Hauru, M. Ganahl, A. G. Lewis, G. Vidal, and A. Blais, Dynamics of transmon ionization, Phys. Rev. Appl. **18**, 034031 (2022).

[29] M. Khezri, A. Opremcak, Z. Chen, K. C. Miao, M. McEwen, A. Bengtsson, T. White, O. Naaman, D. Sank, A. N. Korotkov, Y. Chen, and V. Smelyanskiy, Measurement-induced state transitions in a superconducting qubit: Within the rotating-wave approximation, Phys. Rev. Appl. **20**, 054008 (2023).

[30] T. Thorbeck and B. Abdo, Electrical circuits for leakage reduction units, U.S. patent No. 11183989B1, 2021, Holder: International Business Machines.

[31] O. Ahonen, J. Heinsoo, P. Lähteenmäki, M. Möttönen, J. Rönkkö, J. Salo, J. Santos, and J. Tuorila, Qubit leakage

error reductions, U.S. patent No. 2022006458A1, 2022, Holder: IQM Finland Oy.
[32] Z. Chen *et al.*, Exponential suppression of bit or phase errors with cyclic error correction, Nature (London) **595**, 383 (2021).
[33] P. Magnard, P. Kurpiers, B. Royer, T. Walter, J.-C. Besse, S. Gasparinetti, M. Pechal, J. Heinsoo, S. Storz, A. Blais, and A. Wallraff, Fast and unconditional all-microwave reset of a superconducting qubit, Phys. Rev. Lett. **121**, 060502 (2018).
[34] D. Egger, M. Werninghaus, M. Ganzhorn, G. Salis, A. Fuhrer, P. Müller, and S. Filipp, Pulsed reset protocol for fixed-frequency superconducting qubits, Phys. Rev. Appl. **10**, 044030 (2018).
[35] Y. Zhou, Z. Zhang, Z. Yin, S. Huai, X. Gu, X. Xu, J. Allcock, F. Liu, G. Xi, Q. Yu, H. Zhang, M. Zhang, H. Li, X. Song, Z. Wang, D. Zheng, S. An, Y. Zheng, and S. Zhang, Rapid and unconditional parametric reset protocol for tunable superconducting qubits, Nat. Commun. **12**, 5924 (2021).
[36] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Surface codes: Towards practical large-scale quantum computation, Phys. Rev. A **86**, 032324 (2012).
[37] J. F. Marques, H. Ali, B. M. Varbanov, M. Finkel, H. M. Veen, S. L. M. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider, B. M. Terhal, and L. DiCarlo, All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits, Phys. Rev. Lett. **130**, 250602 (2023).
[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] F. Beaudoin, M. P. da Silva, Z. Dutton, and A. Blais, First-order sidebands in circuit QED using qubit frequency modulation, Phys. Rev. A **86**, 022305 (2012).
[40] C. K. Andersen, A. Remm, S. Lazar *et al.*, Repeated quantum error detection in a surface code, Nat. Phys. **16**, 875 (2020).
[41] J. F. Marques, B. M. Varbanov, M. S. Moreira, H. Ali, N. Muthusubramanian, C. Zachariadis, F. Battistel, M. Beekman, N. Haider, W. Vlothuizen, A. Bruno, B. M. Terhal, and L. DiCarlo, Logical-qubit operations in an error-detecting surface code, Nat. Phys. **18**, 80 (2022).
[42] Y. Wu *et al.*, Strong quantum computational advantage using a superconducting quantum processor, Phys. Rev. Lett. **127**, 180501 (2021).
[43] J. Cohen, A. Petrescu, R. Shillito, and A. Blais, Reminiscence of classical chaos in driven transmons, PRX Quantum **4**, 020312 (2023).
[44] F. Swiadek, R. Shillito, P. Magnard, A. Remm, C. Hellings, N. Lacroix, Q. Ficheux, D. C. Zanuz, G. J. Norris, A. Blais, S. Krinner, and A. Wallraff, Enhancing dispersive readout of superconducting qubits through dynamic control of the dispersive shift: Experiment and theory, PRX Quantum **5**, 040326 (2024).
[45] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.134.120601, which includes Refs. [46–50], for additional information and experimental details.
[46] S. Krinner, S. Storz, P. Kurpiers, P. Magnard, J. Heinsoo, R. Keller, J. Lütolf, C. Eichler, and A. Wallraff, Engineering cryogenic setups for 100-qubit scale superconducting circuit systems, Eur. Phys. J. Quantum Technol. **6**, 2 (2019).
[47] C. Macklin, K. O'Brien, D. Hover, M. E. Schwartz, V. Bolkhovsky, X. Zhang, W. D. Oliver, and I. Siddiqi, A near-quantum-limited Josephson traveling-wave parametric amplifier, Science **350**, 307 (2015).
[48] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, Circuit quantum electrodynamics, Rev. Mod. Phys. **93**, 025005 (2021).
[49] S. Krinner, P. Kurpiers, B. Royer, P. Magnard, I. Tsitsilin, J.-C. Besse, A. Remm, A. Blais, and A. Wallraff, Demonstration of an all-microwave controlled-phase gate between far-detuned qubits, Phys. Rev. Appl. **14**, 044039 (2020).
[50] S. T. Spitz, B. Tarasinski, C. W. J. Beenakker, and T. E. O'Brien, Adaptive weight estimator for quantum error correction in a time-dependent environment, Adv. Quantum Technol. **1**, 1800012 (2018).
[51] J. D. Strand, M. Ware, F. Beaudoin, T. A. Ohki, B. R. Johnson, A. Blais, and B. L. T. Plourde, First-order sideband transitions with flux-driven asymmetric transmon qubits, Phys. Rev. B **87**, 220505 (2013).
[52] S. A. Caldwell *et al.*, Parametrically activated entangling gates using transmon qubits, Phys. Rev. Appl. **10**, 034050 (2018).
[53] G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge University Press, Cambridge, England, 1995).
[54] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, and C. Eichler, Rapid high-fidelity multiplexed readout of superconducting qubits, Phys. Rev. Appl. **10**, 034040 (2018).
[55] S. Krinner, N. Lacroix, A. Remm, A. D. Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, Realizing repeated quantum error correction in a distance-three surface code, Nature (London) **605**, 669 (2022).
[56] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, Demonstrating multi-round subsystem quantum error correction using matching and maximum likelihood decoders, Nat. Commun. **14**, 2852 (2023).
[57] P. V. Klimov *et al.*, Fluctuations of energy-relaxation times in superconducting qubits, Phys. Rev. Lett. **121**, 090502 (2018).
[58] J. Lisenfeld, A. Bilmes, A. Megrant, R. Barends, J. Kelly, P. Klimov, G. Weiss, J. M. Martinis, and A. V. Ustinov, Electric field spectroscopy of material defects in transmon qubits, npj Quantum Inf. **5**, 105 (2019).
[59] D. C. Zanuz, Q. Ficheux, L. Michaud, A. Orekhov, K. Hanke, A. Flasby, M. B. Panah, G. J. Norris, M. Kerschbaum, A. Remm, F. Swiadek, C. Hellings, S. Lazăr, C. Scarato, N. Lacroix, S. Krinner, C. Eichler, A. Wallraff, and J.-C. Besse, Mitigating losses of superconducting qubits strongly coupled to defect modes, arXiv:2407.18746.

[60] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, Efficient measurement of quantum gate error by interleaved randomized benchmarking, Phys. Rev. Lett. **109**, 080505 (2012).

[61] S. M. Assad, O. Thearle, and P. K. Lam, Maximizing device-independent randomness from a Bell experiment by optimizing the measurement settings, Phys. Rev. A **94**, 012304 (2016).

[62] J. Kelly *et al.*, State preservation by repetitive error detection in a superconducting quantum circuit, Nature (London) **519**, 66 (2015).

[63] Google Quantum AI, Suppressing quantum errors by scaling a surface code logical qubit, Nature (London) **614**, 676 (2023).