Hardware-Efficient Leakage-Reduction Scheme for Quantum Error Correction with Superconducting Transmon Qubits

F. Battistel,[1,*] B.M. Varbanov,[1] and B.M. Terhal[1,2]

[1] *QuTech, Delft University of Technology, P.O. Box 5046, Delft 2600 GA, Netherlands*
[2] *JARA Institute for Quantum Information, Forschungszentrum Juelich, Juelich D-52425, Germany*

(Received 2 March 2021; revised 16 June 2021; accepted 2 July 2021; published 26 July 2021)

Leakage outside of the qubit computational subspace poses a threatening challenge to quantum error correction (QEC). We propose a scheme using two leakage-reduction units (LRUs) that mitigate these issues for a transmon-based surface code, without requiring an overhead in terms of hardware or QEC-cycle time as in previous proposals. For data qubits, we consider a microwave drive to transfer leakage to the readout resonator, where it quickly decays, ensuring that this negligibly disturbs the computational states for realistic system parameters. For ancilla qubits, we apply a $|1\rangle \leftrightarrow |2\rangle$ π pulse conditioned on the measurement outcome. Using density-matrix simulations of the distance-3 surface code, we show that the average leakage lifetime is reduced to almost one QEC cycle, even when the LRUs are implemented with limited fidelity. Furthermore, we show that this leads to a significant reduction of the logical error rate. This LRU scheme opens the prospect for near-term scalable QEC demonstrations.

DOI: 10.1103/PRXQuantum.2.030314

Quantum computing has recently reached the milestone of quantum supremacy [1] thanks to a series of improvements in qubit count [2,3], gate fidelities [4–15], and measurement fidelities [16–18]. The next major milestones include showing a quantum advantage [19–22] and demonstrating quantum error correction (QEC) [3,23–31]. Errors accumulate over time in a quantum computer, leading to an entropy increase that severely hinders the accuracy of its output. Thus, QEC is necessary to correct errors and remove entropy from the computing system. If the overall physical error rate is below a certain noise threshold for a given QEC-code family, the logical error rate decreases exponentially with the code distance d at the price of a poly(d) overhead, thus allowing one to extend the computational time. Recently, small-size instances of error-detecting [29,30] and error-correcting [3] codes have been experimentally realized. To further demonstrate fault tolerance, it is crucial to scale up these systems and show that larger distance codes consistently lead to lower logical error rates than smaller distance codes [31].

Leakage outside of the computational subspace [8–10,12,32–37], present in leading quantum-computing platforms such as superconducting qubits and trapped ions, poses a particularly threatening challenge to fault tolerance [23,38–48]. Leakage can increase entropy by making measurement outcomes no longer point to the underlying errors and can effectively reduce the code distance [46]. Furthermore, leakage can last for many QEC cycles [40], making operations on a leaked qubit fail and possibly spread correlated errors through the code [31,39,46]. In particular, leakage falls outside the stabilizer formalism of QEC as it cannot be decomposed in terms of Pauli errors. Stabilizer codes [49,50] and their decoders are thus typically ill-suited to deal with leakage, leading to a significant increase of the logical error rate [42,45,46]. If the average leakage lifetime $l^{\mathcal{L}}_{\text{avg}}$, that is, the average number of QEC cycles that a qubit stays leaked (after leaking in the first place), fulfills $l^{\mathcal{L}}_{\text{avg}} = \mathcal{O}(1)$ QEC cycles and $l^{\mathcal{L}}_{\text{avg}} \ll d$ then, for low-enough error rates, a threshold is likely to exist [39] as leakage would have a relatively local effect in space and time. Because of a finite energy-relaxation time, leakage does indeed last for $l^{\mathcal{L}}_{\text{avg}} = \mathcal{O}(1)$ QEC cycles. However, in practice, it is important how large $l^{\mathcal{L}}_{\text{avg}}$ is, since, if it is low, the noise threshold is expected to be higher. Shortening the relaxation time to reduce $l^{\mathcal{L}}_{\text{avg}}$ is not effective as this increases the physical error rate as well.

A leakage-reduction unit (LRU) [38,39,41,42,47,48,51,52] is an operation introducing a seepage mechanism besides that of the relaxation channel. A LRU converts leakage into regular (Pauli) errors and shortens the average leakage lifetime, ideally to one QEC cycle. As discussed above, this is expected to lead to a higher noise threshold,

[*] battistel.fra@protonmail.com

but not as high as for the case without leakage, since the leakage rate effectively adds to the regular error rate thanks to the LRU. As an alternative to the use of LRUs, postselection based on leakage detection has been adopted [46] as a near-term method to reduce the logical error rate. While leakage detection could also be used to apply LRUs in a targeted way, postselection is not scalable. By shortening the lifetime to $l_{\text{avg}}^{\mathcal{L}} = \mathcal{O}(1) \ll d$, the use of LRUs is instead a scalable approach.

In its imperfect experimental implementation a LRU can either introduce extra Pauli errors or mistakenly induce leakage on a nonleaked qubit. Furthermore, in the context of the surface code the LRUs investigated so far [41,42,47] introduce an overhead in terms of hardware and QEC-cycle time. Specifically, these LRUs are variants of the swap-LRU, in which the qubits are swapped at the end of each QEC cycle, taking alternatively the role of data and ancilla qubits. In this way every qubit is measured every two QEC cycles. The core of the swap-LRU is the fact that the measured qubits are reset to the computational subspace after the measurement. This can be accomplished by a scheme that unconditionally maps $|1\rangle$ and $|2\rangle$ (and possibly $|3\rangle$ [48]) to $|0\rangle$ [53–55], or conditionally using real-time feedback [28,56]. Under the standard assumption that the SWAP gates swap the states of two qubits only if none of them is leaked (which does not necessarily hold in experiment [48]), $l_{\text{avg}}^{\mathcal{L}}$ is ideally shortened to two QEC cycles. On the downside, for the pipelined surface-code scheme in Ref. [57], the pipeline is disrupted as qubits cannot be swapped until the measurement and reset operations are completed, leading overall to an increase up to 50% of the QEC-cycle time depending on the reset time. The extra controlled-Z (CZ) gates, needed to implement the SWAP gates, can cause additional errors or leakage as the CZ gate is the major source of leakage in transmons [8–10,12,32–35]. Moreover, in the surface code an extra row of qubits is needed to perform all the SWAP gates [41], which is a non-negligible overhead in the near term. All these issues increase the physical error rate by a considerable amount, thus requiring an increase in the system size to compensate for that (assuming that the error rates are still below threshold).

In this work we propose two separate LRUs for data and ancilla qubits that use resources already available on chip, namely the readout resonator for data qubits (res-LRU) and a $|1\rangle \leftrightarrow |2\rangle$ π pulse conditioned on the measurement outcome for ancilla qubits (π-LRU). In particular, the use of the res-LRU avoids the necessity to swap data and ancilla qubits to be able to reset the data qubits. The res-LRU is a modification of the two-drive scheme in Refs. [53–55] to a single drive to deplete only the population in $|2\rangle$ but not $|1\rangle$, making it a LRU rather than a reset scheme. We additionally show that this negligibly affects the coherence within the computational subspace in an experimentally accessible regime, with a low probability of mistakenly inducing leakage as long as the thermal population in the readout resonator is relatively small. This allows us to unconditionally use res-LRU in the surface code in every QEC cycle. In the pipelined scheme [57] the res-LRU easily fits within the time in which the data qubits are idling while the ancilla qubits are finishing being measured. As the π-LRU can be executed as a short pulse at the end of the measurement time with real-time feedback, our LRU scheme overall does not require any QEC-cycle time overhead. Using density-matrix simulations [46,50,58] of the distance-3 surface code (Surface-17), we show that the average leakage lifetime is reduced to almost one QEC cycle when res-LRU and π-LRU with realistic performance are employed. Furthermore, compared to the case without LRUs, the logical error rate is reduced by up to 30%. The proposed res-LRU and π-LRU can be straightforwardly adapted to QEC-code schemes other than that of Ref. [57] and the res-LRU is potentially applicable to superconducting qubits with higher anharmonicity than transmons. The demonstrated reduction serves as evidence of scalability for our LRU scheme, even though we cannot estimate a noise threshold as we have simulated only one size of the surface code. To explore larger codes, it is necessary to use less computationally expensive simulations [23,39,42] that use a simplified version of our error model at the cost of losing some information contained in the density matrix. Furthermore, to optimize the noise threshold, the LRUs can be supplied with a leakage-aware decoder [23,39,42,59–61] that uses measurement information about leakage to better correct leakage-induced correlated errors.

I. READOUT-RESONATOR LRU

The readout resonator has been used [53–55] to reset a transmon qubit to the $|0\rangle$ state, depleting the populations in $|1\rangle$ and $|2\rangle$. Targeting the $|20\rangle \leftrightarrow |01\rangle$ transition, with the notation $|\text{transmon}, \text{resonator}\rangle$, those populations are swapped onto the readout resonator, where they quickly decay due to the strong coupling to the transmission-line environment. Magnard *et al.* [53] used two drives simultaneously, while Zeytinoğlu *et al.* [54] and Egger *et al.* [55] used these drives in a three-step process. Here we adapt these techniques to use a single drive in a single step to deplete the population in $|2\rangle$ only.

A LRU is defined [38] as an operation such that (1) the incoming leakage population is reduced after the application of the LRU, (2) the induced leakage when applied to a nonleaked state is ideally 0. We thus ensure below that not only leakage is reduced but also that the effect that the drive has on a nonleaked transmon is as small as possible.

A. Transmon-resonator Hamiltonian

We consider a transmon capacitively coupled to a resonator and to a dedicated microwave drive line. The resonator possibly employs a Purcell filter that we do not include explicitly. In a frame rotating at the transmon-drive frequency ω_d for both the resonator and the transmon, the Hamiltonian is time independent and is given by

$$H = H_0 + H_c + H_d, \tag{1}$$

$$H_0 = \delta^r a^\dagger a + \delta^q b^\dagger b + \frac{\alpha}{2}(b^\dagger)^2 b^2, \tag{2}$$

$$H_c = g(ab^\dagger + a^\dagger b), \tag{3}$$

$$H_d = \frac{\Omega}{2}(e^{i\phi} b + e^{-i\phi} b^\dagger), \tag{4}$$

where a and b are the creation operators for the resonator and the transmon, respectively; $\delta^r = \omega_r - \omega_d$ and $\delta^q = \omega_q - \omega_d$ with ω_r and ω_q the resonator and transmon frequencies, respectively; $\alpha < 0$ is the transmon anharmonicity; g corresponds to the capacitive coupling; Ω and ϕ are the transmon-drive amplitude and phase, respectively. The phase is not relevant for the results in this work and we fix it to $\phi = 0$.

We can qualitatively understand [see Fig. 1(a)] that H contains an effective coupling $\tilde{g}$ between $|20\rangle$ and $|01\rangle$. If ω_d matches the transition frequency between the “bare” $|20\rangle$ and $|01\rangle$, these two states are degenerate in the rotating frame and they are connected by two paths (at lowest order) via either $|11\rangle$ or $|10\rangle$. If $\Delta{:=}\omega_q - \omega_r \gg g$ and $\delta^q \gg \Omega$ then $|11\rangle$ and $|10\rangle$ are occupied only “virtually” and one gets purely an effective $|20\rangle \leftrightarrow |01\rangle$ coupling. Modulo a constant term, in the two-dimensional subspace $\mathcal{S} = \text{span}\{|20\rangle, |01\rangle\}$ we can write H in Eq. (1) as $H|_{\mathcal{S}} \equiv -\eta(\omega_d)Z/2 + \tilde{g}(\omega_d)X$ for an appropriate function η [an approximation can be extracted from Eq. (A31)]. As a function of ω_d, this Hamiltonian gives rise to a $|20\rangle \leftrightarrow |01\rangle$ avoided crossing centered at a frequency ω_d^* [see Fig. 1(b)] where $\eta(\omega_d^*) = 0$. The energy separation at the center of the avoided crossing is then $2\tilde{g}(\omega_d^*)$.

In order to quantitatively study the action of H, we unitarily transform it using a Schrieffer-Wolff transformation e^S [62–65]. Let $\{|ij\rangle_D\}$ be the basis of eigenvectors of $H_0 + H_c$ (the transmon-resonator “dressed” basis). In the dispersive regime ($g \ll \Delta$), with respect to a first-order Schrieffer-Wolff transformation S_1 in the perturbation parameter g/Δ, such that $e^{-S_1}|ml\rangle \approx |ml\rangle_D$, we get (see Appendix A)

$$H^D{:=}e^S H e^{-S} \approx e^{S_1} H e^{-S_1} \tag{5}$$

$$= H_0^D + H_{d1}^D + H_{d2}^D, \tag{6}$$

with

$$H_0^D = \left(\delta^r - \sum_{m=0}^{\infty} \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m|\right) a^\dagger a + \sum_{m=1}^{\infty} \left(m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}}\right)|m\rangle\langle m|, \tag{7}$$

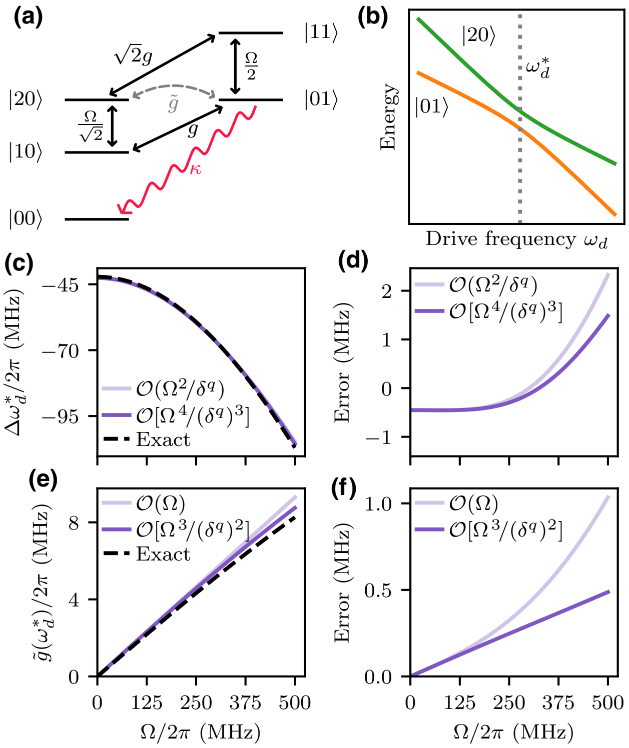

FIG. 1. Concept of the readout-resonator LRU. (a) The state $|20\rangle$ (with the notation $|$transmon,resonator$\rangle$) is connected to $|01\rangle$ by two main paths via either $|11\rangle$ or $|10\rangle$, due to the capacitive coupling g or the transmon-drive amplitude Ω, respectively. This generates an effective coupling $\tilde{g}$ that can be used to swap $|20\rangle \leftrightarrow |01\rangle$. The latter quickly decays to $|00\rangle$ due to the typically high coupling κ of the readout resonator to the transmission-line environment, overall removing leakage from a leaked transmon. (b) In the rotating frame of the drive, $|20\rangle$ and $|01\rangle$ show an avoided crossing as a function of the drive frequency ω_d, centered at ω_d^*. The effective coupling $\tilde{g}(\omega_d^*)$ is equal to half the energy separation at that point. (c),(e) The values $\Delta\omega_d^*{:=}\omega_d^* - (2\omega_q + \alpha - \omega_r)$ and $\tilde{g}(\omega_d^*)$ are respectively evaluated either exactly by full numerical diagonalization of H in Eq. (1), or by approximate analytical formulas (see Sec. I A and Appendix A) for the parameters in Table I. The absolute errors with respect to the exact curves are shown in (d) and (f), respectively.

$$H_{d1}^D = \frac{\Omega e^{i\phi}}{2} b + \text{H.c.}, \tag{8}$$

$$H_{d2}^D = \frac{\Omega e^{i\phi}}{2}\left(a \sum_{m=0}^{\infty} \frac{g\Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| + a^\dagger \sum_{m=0}^{\infty} \frac{g\alpha\sqrt{m+1}\sqrt{m+2}}{\Delta_m \Delta_{m+1}} |m\rangle\langle m+2|\right) + \text{H.c.}, \tag{9}$$

where $\Delta_m{:=}\Delta + \alpha m$ and $\{|m\rangle\}$ are transmon states. Here H_0^D is diagonal and contains the dispersive shifts, H_{d1}^D is the transmon drive now in the unitarily transformed frame, and

H_{d2}^D contains an indirect resonator drive and couplings of the kind $a^\dagger|m\rangle\langle m+2|$. In particular, for $m=0$ in Eq. (9), we get a lowest-order approximation of $\tilde{g}$:

$$\tilde{g} \approx \frac{\Omega g \alpha}{\sqrt{2}\Delta(\Delta+\alpha)}. \qquad (10)$$

Note that at this order there is no dependence on ω_d. Furthermore, $\tilde{g}$ would vanish for $\alpha = 0$, since the two paths in Fig. 1(a) fully destructively interfere in that case. Since α is low for transmons, one can expect that Ω needs to be relatively large for $\tilde{g}$ to be substantial.

For the drive to be most effective, it is important that ω_d matches ω_d^*. If $g = 0 = \Omega$, there is no avoided crossing but $|20\rangle$ and $|01\rangle$ simply cross at $\omega_{d,0}^* \equiv 2\omega_q + \alpha - \omega_r$, as can be straightforwardly computed from H_0 in Eq. (2). This value is shifted due to the capacitive coupling [as can be seen from Eq. (7)], as well as due to the possibly strong drive. For $g \neq 0$ and $\Omega \neq 0$, one can either compute ω_d^* by full numerical diagonalization of H and find the avoided crossing as a function of ω_d, or one can find an (approximate) analytical expression. For the latter, we use another Schrieffer-Wolff transformation (rather than the resolvent method in Ref. [54], which does not give the full Hamiltonian) to account for the effect of the transmon drive H_{d1}^D and to compute ω_d^* up to order $\Omega^4/(\delta^q)^3$; see Appendix A. We also use this transformation to compute $\tilde{g}$ up to order $\Omega^3/(\delta^q)^2$. Figures 1(c) and 1(e) compare the analytical approach with the exact numerical results for $\Delta\omega_d^* = \omega_d^* - \omega_{d,0}^*$ and $\tilde{g}(\omega_d^*)$, respectively, given the parameters in Table I. We consider six energy levels for the transmon and three for the resonator as we see that the exact curves converge for such choice. In Figs. 1(c) and 1(d) we see that the two approximations are both pretty good, while in Figs. 1(e) and 1(f) we see that Eq. (10) deviates by up to 1 MHz from the exact value at high Ω and that the absolute error with respect to the exact $\tilde{g}(\omega_d^*)$ scales in a seemingly quadratic way. Instead, the higher-order approximation stays closer to the exact curve and the error scales linearly. We expect that the remaining gap would be mostly filled by also considering higher orders in g/Δ in the first Schrieffer-Wolff transformation, since increasing only the order of approximation in Ω/δ^q does not provide a significant improvement in Fig. 1(d).

TABLE I. Parameters used both in the analysis and Lindblad simulations of the readout-resonator LRU, similar to the experimental ones in Ref. [27]. The transmon parameters correspond to the target parameters of a high-frequency data qubit in Sec. II.

Parameter	Transmon	Readout resonator
Frequency $\omega/2\pi$	6.7 GHz	7.8 GHz
Anharmonicity $\alpha/2\pi$	−300 MHz	Not applicable
Coupling $g/2\pi$	135 MHz	
Average photon number $\bar{n}$	Not applicable	0.005
Relaxation time T_1	30 μs	16 ns ($\kappa/2\pi$ = 10 MHz)
Dephasing time T_2	30 μs (flux noise)	32 ns

B. Performance of the readout-resonator LRU

Given the theoretical understanding of the transmon-resonator system, we devise a pulse to minimize the population in $|2\rangle$ on a leaked transmon. We consider the pulse shape

$$\Omega(t) = \begin{cases} \Omega \sin^2\left(\pi \dfrac{t}{2t_{\text{rise}}}\right) & \text{for } 0 \le t \le t_{\text{rise}}, \\ \Omega & \text{for } t_{\text{rise}} \le t \le t_p - t_{\text{rise}}, \\ \Omega \sin^2\left(\pi \dfrac{t_p - t}{2t_{\text{rise}}}\right) & \text{for } t_p - t_{\text{rise}} \le t \le t_p, \end{cases} \qquad (11)$$

similarly to Ref. [54], where t_p is the total pulse duration, at a fixed frequency $\omega_d(t) = \omega_d$. Hence, there are four parameters to optimize over, i.e., Ω, ω_d, t_p, and t_{rise}. We fix $t_{\text{rise}} = 30$ ns since we observe that this strongly suppresses nonadiabatic transitions out of the manifold of interest: for example, $|20\rangle$ is coupled to $|10\rangle$ by the drive but they are quite off-resonant, so only a fast pulse can cause "nonvirtual" transitions between them. Indeed, for $t_{\text{rise}} \lesssim 10$ ns, there appear ripples (for an example, see Ref. [54]) in, e.g., the $|20\rangle$ and $|10\rangle$ populations when the drive is turned on and off, leading to a reduction in performance. We expect that an improved pulse shape can shorten t_{rise}. However, we do not explore this given the long maximum t_p allowed in our surface-code scheme ($t_p \le T_{\text{slot}} = 440$ ns; see Sec. II A).

We use Lindblad simulations of the transmon-resonator system to optimize over Ω, ω_d, and t_p. The Lindblad equation is given by

$$\dot{\rho} = -i[H^D, \rho] + \sum_j \left(K_j \rho K_j^\dagger - \frac{1}{2}\{K_j^\dagger K_j, \rho\}\right) \qquad (12)$$

with $\{K_j\}$ the quantum jump operators. We express (and solve) this equation in the exact unitarily transformed frame. That is, while in Sec. I A we have used a first-order Schrieffer-Wolff transformation e^{S_1} [see Eq. (5)], in the numerics we compute the full transformation e^S [see also Eq. (5)]. In this way we find the basis that exactly diagonalizes $H_0 + H_c$ and express H_d in this basis as well, without any further Schrieffer-Wolff transformation like in Sec. I A. In other words, the simulations reproduce the dynamics under the Hamiltonian in Eqs. (1) to (4) without any approximation.

The Hamiltonian parameters are the same as in Sec. I A and are reported in Table I, including the noise parameters. In particular, while we neglect the transmon thermal population, we include it for the resonator since it determines the leakage that the pulse induces when the transmon is not leaked, as we discuss below. The resonator thermal state is given by [67]

$$\sigma_{\mathrm{th}} \approx \left(1 - \frac{\bar{n}}{1+2\bar{n}}\right)|0\rangle\langle 0| + \frac{\bar{n}}{1+2\bar{n}}|1\rangle\langle 1| \tag{13}$$

for low average photon number $\bar{n}$. We consider dressed relaxation and dephasing, as given below, assuming that this is a good model in the dispersive regime. In the unitarily rotated frame, the employed jump operators $\{K_j\}$ are explicitly given by

$$\frac{1}{\sqrt{T_1^r}}a = \sqrt{\kappa}a, \qquad \sqrt{\frac{\bar{n}}{1+\bar{n}}}\sqrt{\kappa}a^\dagger, \qquad \sqrt{\frac{2}{T_\phi^r}}a^\dagger a, \tag{14}$$

$$\frac{1}{\sqrt{T_1^q}}b, \qquad \sqrt{\frac{2}{T_\phi^q}}b^\dagger b, \tag{15}$$

where $T_\phi = (1/T_2 - 1/2T_1)^{-1}$ and where we consider six energy levels for the transmon and three for the resonator. Note that, e.g., for a, going back to the original frame it holds that $e^{-S}ae^{S} = \sum_{l=0}^{1}\sqrt{l+1}|l\rangle_D\langle l+1|_D = a_D$ by the definition of e^S, corresponding indeed to relaxation in the dressed basis. By considering dressed relaxation and dephasing, the effective relaxation time T_1^q of the transmon is not shortened by the fact that it is coupled to a lossy resonator (Purcell effect). We assume that this is also a good approximation during driving as the drive couples eigenstates that mostly have the same number of excitations in the resonator (except for $|20\rangle$ and $|01\rangle$ when the drive is near resonant with this transition and causes a strong mixing of these states). We thus mimic the use of a Purcell filter but without including it in the simulations since that would increase the Hilbert-space dimension in a computationally expensive way.

For each choice of (Ω, ω_d), we optimize t_p such that, given the initial state $|2\rangle\langle 2| \otimes \sigma_{\mathrm{th}}$, the leakage population $p^{|2\rangle} = \langle 2|\,\mathrm{Tr}_r[\rho(T_{\mathrm{slot}})]|2\rangle$ at the end of the available time slot is minimized [see Fig. 2(a)]. The states $|20\rangle$ and $|01\rangle$ approximately form a two-level system with additional damping from $|01\rangle$ to $|00\rangle$; thus, the drive effectively induces damped Rabi oscillations [68] between them. Oscillations occur only for $\tilde{g} > \kappa/4$ [68] (underdamped regime), while for $\tilde{g} = \kappa/4$ (critical regime) or $\tilde{g} < \kappa/4$ (overdamped regime), the populations in $|20\rangle$ and $|01\rangle$ simply decay in an exponential fashion without forming any minimum. For the parameters in Table I, the critical drive amplitude that gives $\tilde{g} = \kappa/4$ is $\Omega_{\mathrm{cr}}/2\pi \approx 143$ MHz. Thus, for $\Omega \leq \Omega_{\mathrm{cr}}$, the best strategy is to drive until $p^{|2\rangle}$ reaches a (low) practically stable value (which is in general not 0 when the full system is taken into account). Here, with the given κ, we find that this occurs in a time comparable to T_{slot} only from about $\Omega = \Omega_{\mathrm{cr}}$, so for $\Omega \leq \Omega_{\mathrm{cr}}$, we drive for the entire T_{slot}. For $\Omega > \Omega_{\mathrm{cr}}$, the optimization has many local minima as a function of t_p, corresponding to the minima of the $|20\rangle \leftrightarrow |01\rangle$ oscillations induced by the drive. Here we choose to target the first minimum as in Refs. [54,55] since it is the fastest approach. For a sudden pulse, this minimum would occur around $\pi/2\tilde{g}$ for sufficiently small κ, whereas we find heuristically that a good initial guess for the optimization is $\pi/2\tilde{g}_{\mathrm{damp}}$ with $\tilde{g}_{\mathrm{damp}}{:=}\sqrt{\tilde{g}^2 - (\kappa/4)^2}e^{-\kappa/7\tilde{g}}$ for larger κ. Then, for the optimization over t_p, we use the bounds $t_p - 2t_{\mathrm{rise}} \in [0, 1.1 \times \pi/2\tilde{g}_{\mathrm{damp}}]$ (using the bounded Brent method in *scipy*; see [69] for the code). While using a longer t_p in the

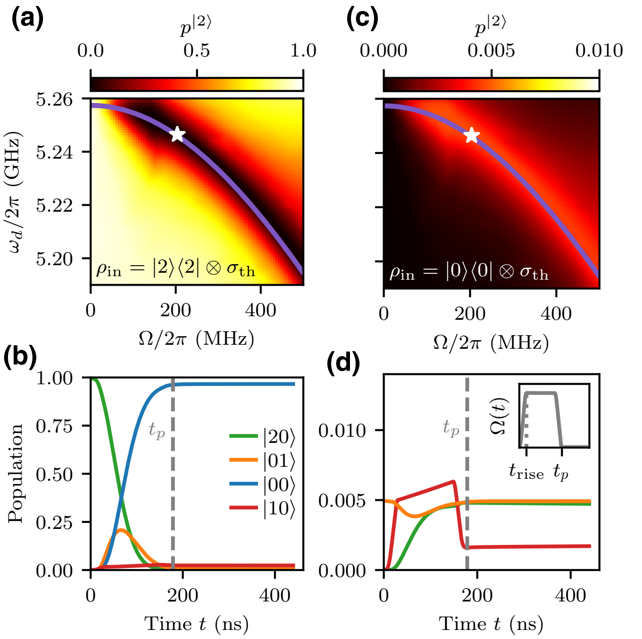

FIG. 2. Lindblad simulations of the transmon-resonator system for the readout-resonator LRU. In (a),(b) the initial state is $|2\rangle\langle 2| \otimes \sigma_{\mathrm{th}}$, while in (c),(d) it is $|0\rangle\langle 0| \otimes \sigma_{\mathrm{th}}$, where σ_{th} is the resonator thermal state. (a),(c) Transmon leakage population $p^{|2\rangle} = \langle 2|\,\mathrm{Tr}_r[\rho(T_{\mathrm{slot}})]|2\rangle$ at the end of the time slot of $T_{\mathrm{slot}} = 440$ ns. For each choice of (Ω, ω_d), we optimize the total pulse duration $t_p \leq T_{\mathrm{slot}}$ to minimize $p^{|2\rangle}$ given the initial state $|2\rangle\langle 2| \otimes \sigma_{\mathrm{th}}$ for fixed $t_{\mathrm{rise}} = 30$ ns. The white star indicates the chosen operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_p = 178.6$ ns) with $p^{|2\rangle}_{\mathrm{op.}} \approx 0.5\%$ in (a). The induced leakage in (c) is $p^{|2\rangle} \approx 0.48\%$ at the operating point. The purple line corresponds to the higher-order estimate of the optimal drive frequency ω_d^* as a function of Ω [see Fig. 1(c)]. The heatmaps are sampled using the ADAPTIVE package [66]. (b),(d) Time evolution of the populations in a few selected states for the operating point. The vertical dashed line indicates the used t_p. The inset in (d) shows a schematic of the pulse $\Omega(t)$.

underdamped regime (possibly even greater than the allotted T_{slot}) would eventually lead to an even lower leakage population [53], it is not necessarily desirable as a longer t_p may mean that the disturbance to a nonleaked transmon is greater as well (see Appendix B 2).

While the procedure above optimizes t_p given a certain pair (Ω, ω_d), we use the package ADAPTIVE [66] to choose the next pair to sample and we iterate this process. This package searches a given parameter space (here $\Omega/2\pi \in [0, 500 \text{ MHz}]$, $\omega_d/2\pi \in [5.19, 5.26 \text{ GHz}]$) in a finer way where the given cost function changes faster. Here we use $(\log p^{|2\rangle})^2$ as the cost function since it changes faster where $p^{|2\rangle}$ is small, allowing us to get both a high-resolution heatmap (see Fig. 2) and a good first estimation of the $p^{|2\rangle}$ minima in a single run. Then we run a local optimization with tight bounds around some of these candidate points for fine tuning.

In Fig. 2(a) one can observe a band with low $p^{|2\rangle}$, as desired. This band occurs at drive frequencies slightly above $\omega_d^*(\Omega)$, which one would expect to be optimal based on Sec. I A. We attribute this to the fact that a significant share of the time is taken by the rise and fall of the pulse, where $\Omega(t)$ is smaller than the maximum. We find that one can choose a broad range of Ω values to achieve a $p^{|2\rangle} \gtrsim 0.5\%$, from 130 MHz (slightly below the critical point) to deep in the underdamped regime. However, other considerations apply, namely, on the high end using a very high Ω poses strong experimental requirements on the drive, while on the low end the pulse takes much longer and it is not *a priori* given that driving at the critical point would be best. Actually, note that driving at the critical point with good performance is possible only due to the relatively high T_{slot} for the given κ. In the following we choose the point marked by a star in Fig. 2 as the operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_p = 178.6$ ns). This point reaches $p^{|2\rangle}_{\text{op.}} \approx 0.5\%$ while least affecting the coherence within the computational subspace (see Appendix B 1). We attribute the fact that this minimum does not reach 0 to reheating from $|00\rangle$ to $|01\rangle$, as well as transmon decoherence (resonator pure dephasing would contribute as well but here $T^r_\phi = \infty$) and interactions with higher-energy levels. We note that in Fig. 2(a) we find good $p^{|2\rangle} \lesssim 5\%$ up to $\Omega/2\pi \gtrsim 100$ MHz, which could be used to further ease the requirements on the drive (see Sec. II).

The time evolution for a few selected states is shown in Fig. 2(b) for the operating point, given the initial state $|2\rangle\langle 2| \otimes \sigma_{\text{th}}$. The first few nanoseconds make $|20\rangle$ rotate into $|01\rangle$, while the latter decays relatively fast to $|00\rangle$ due to the large relaxation rate κ of the readout resonator. Already after approximately 220 ns the remaining $|01\rangle$ population has practically returned to the thermal state. The repetition of the pulse, such as in the surface code (see Sec. II) at every QEC cycle, does not thus lead to heating of the resonator with these system parameters (see Sec. III for a discussion about other parameter regimes).

We now evaluate the effect of the pulse on a nonleaked transmon [see Figs. 2(c) and 2(d)]. There should ideally be no effect, except for an acquired single-qubit phase that can be easily determined and corrected by either a real or virtual Z rotation. First, if the transmon is in $|0\rangle$ and there is some thermal population in the resonator, part of the state is supported on $|01\rangle$, which rotates into $|20\rangle$ in the same way as the opposite process by unitarity. Figure 2(c) shows that indeed the induced leakage is greater where $p^{|2\rangle}$ is lower in Fig. 2(a). However, due to the low $\bar{n} = 0.005$, the induced leakage is also overall low [$p^{|2\rangle} \approx 0.48\%$ in Fig. 2(c) at the operating point, which is comparable to state-of-the-art CZ leakage rates; see Sec. II B] and can be made even lower by engineering colder resonators. If the initial state is $|1\rangle\langle 1| \otimes \sigma_{\text{th}}$, there is little induced leakage ($p^{|2\rangle} \approx 0.02\%$ at the operating point and $p^{|2\rangle} \lesssim 0.04\%$ across the whole landscape) as the drive is off-resonant with transitions from this state. Second, the pulse might affect the coherence times of the transmon by driving transitions within or outside the computational subspace (and back), as the small but non-negligible transitory population in $|10\rangle$ in Figs. 2(b) and 2(d) seems to suggest. However, we find that both the effective T_1^q and T_2^q are only marginally affected as a function of Ω (see Appendix B 1). This is because stronger pulses cause a somewhat stronger disturbance to the qubit, but they are shorter so that in total the effect is small.

II. SURFACE CODE WITH LRUS

A. Layout and operation scheduling

We study the distance-3 rotated surface code [see Fig. 3(a)], nicknamed Surface-17, in the presence of leakage and LRUs. We follow the frequency and pipelined scheme in Ref. [57], in which the nine data qubits are subdivided into three high- and six low-frequency ones. The four X and the four Z ancilla qubits have an intermediate frequency. We consider the flux-pulse implementation of the CZ gate [9,10,32–34] for tunable-frequency transmons, in which the transmon with the greater frequency is lowered towards the other one with a flux pulse. With this technique, fluxed transmons are prone to leakage. This means that the high-frequency data qubits and all the ancilla qubits can leak. As shown in Ref. [46], leakage can last for many QEC cycles and can be quite detrimental to the logical performance of the code. Here we address these issues with the res-LRU for high-frequency data qubits and with the π-LRU for ancilla qubits, as described below. If, due to a different implementation of the CZ gates (or due to leakage mobility [46,48]), the low-frequency data qubits can also leak, one can apply the res-LRU to them as well, but we do not explore this here.

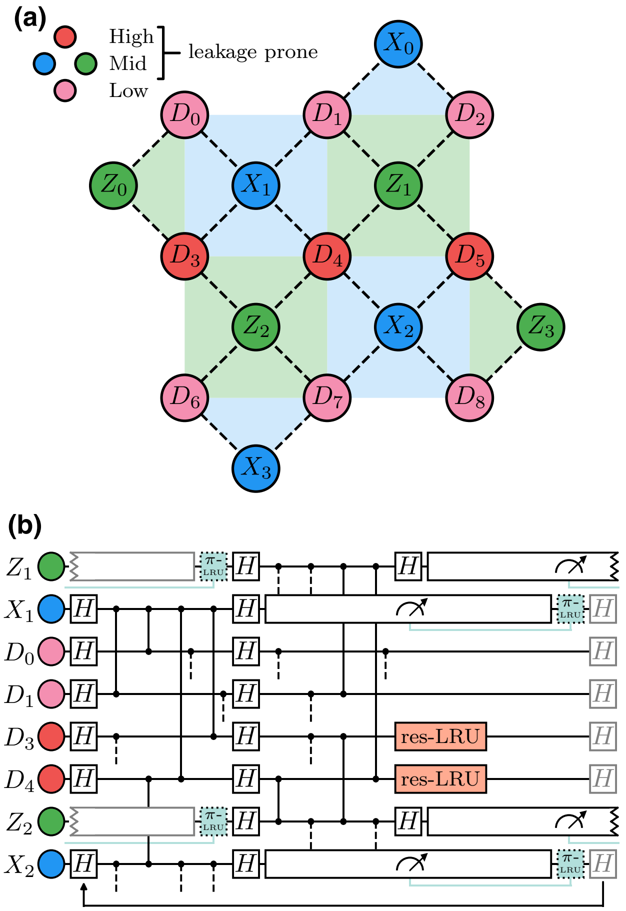

FIG. 3. (a) Schematic overview of the Surface-17 layout [46,57]. Pink (red) circles with D labels represent low- (high-)frequency data qubits, while blue (green) circles with X (Z) labels represent ancilla qubits, which have an intermediate frequency. Ancilla qubits and high-frequency data qubits are prone to leakage during the CZ gates. (b) The quantum circuit for a single QEC cycle employed in simulation for the unit-cell scheduling defined in Ref. [57], in which we insert the LRUs. The res-LRUs (orange) are applied unconditionally on the high-frequency data qubits after the CZ gates, while the π-LRUs (teal) are applied on the ancilla qubits depending on the measurement outcome. Gray elements correspond to operations belonging to the previous or the following QEC cycle. The duration of each operation is given in Appendix C 1. The arrow at the bottom indicates the repetition of QEC cycles.

The circuit executed for each QEC cycle is shown in Fig. 3(b). The X-type and Z-type parity-check units are implemented in an interleaved way, with the CZ gates for one unit being applied while the other ancilla-qubit type is measured. The duration of each operation is summarized in Appendix C 1, with a total QEC-cycle duration of 800 ns. The data qubits are idling for a considerable amount of time, namely $T_{\text{slot}} = 440$ ns, while the ancilla qubits are measured. We choose this time slot as the ideal place to apply the res-LRUs, introduced in Sec. I, to the high-frequency data qubits. Note that the optimal pulse selected in Sec. I B, which was simulated for the target parameters of the high-frequency data qubits, takes about $t_p = 180$ ns and thus easily fits within this time slot (see Sec. III for a discussion about other parameter regimes).

For the ancilla qubits, there is no available time slot to apply the res-LRU. A possibility would be to make the QEC-cycle time longer by inserting these LRUs when the measurement is completed. However, this approach would lower the logical error rate of the code by a non-negligible amount. On the other hand, ancilla qubits are measured and the (analog) measurement outcome contains information about leakage [46]. We choose to use a different type of LRU altogether that uses this information. Specifically, we consider a $|1\rangle \leftrightarrow |2\rangle$ π pulse, conditioned on the measurement outcome reporting a $|2\rangle$. Below we discuss further details of the implementation of this π-LRU.

B. Implementation of the LRUs in the density-matrix simulations

We use density-matrix simulations [50] using the open-source package QUANTUMSIM [58] to study Surface-17 with res-LRUs and π-LRUs. We include relaxation and dephasing (T_1 and T_2), as well as flux-dependent T_2 and leakage rate L_1 during the CZ gates, following the same error model as in Ref. [46]. We define L_1 as the average leakage from the computational to the leakage subspace [70]. The state of the art is $L_1 \approx 0.1\%$ [9,10], although the actual L_1 is expected to be higher when operating a multitransmon processor [30,71]; thus, here we consider up to $L_1 = 0.5\%$. We assume that single-qubit gates do not induce any leakage as their leakage rates are typically negligible compared to the CZ gates [5,36,37]. The noise parameters used are reported in Appendix C 1. Furthermore, during a CZ gate with a leaked transmon, the nonleaked transmon acquires a phase called the leakage conditional phase [46]. We select these phases uniformly at random (see Appendix C 3) and, in contrast to Ref. [46], we then keep them fixed for every Surface-17 simulation in this work. This makes it easier to recognize trends as a function of the LRU parameters. In Appendix C 3 we discuss the variability of the logical error rate depending on the leakage conditional phases. We do not consider further leakage from $|2\rangle$ to $|3\rangle$ in subsequent CZ gates [46] as we expect it to be negligible when LRUs make $|2\rangle$ short lived.

1. res-LRU for data qubits

In the simulations, leakage-prone transmons are modeled as three-level systems and nonleakage-prone transmons as two-level systems, leading to an already computationally expensive size for the density matrix. As a consequence, we do not include the readout resonator explicitly in these simulations. The resonator is initially in the ground state and is returned to it at the end of the time slot, approximately. We can thus trace the resonator out and model the res-LRU on the transmon qubit as an incoherent $|2\rangle \mapsto |0\rangle$ relaxation (see Appendix C 1 a for details). Furthermore, in Sec. I B we have observed that the res-LRU can also cause a nonleaked transmon to partially leak, so we include that as an incoherent $|0\rangle \mapsto |2\rangle$ excitation.

Calling $p_i^{|j\rangle}, p_f^{|j\rangle}$ the populations before and after the res-LRU, we define the leakage-reduction rate $0 \leq R \leq 1$ as $R = 1 - p_f^{|2\rangle}$ conditioned on an initially fully leaked transmon, i.e., for $p_i^{|2\rangle} = 1$. Furthermore, we define the average res-LRU leakage rate L_1^{LRU} as the average of the induced leakage starting from either $|0\rangle$ or $|1\rangle$ (consistently with the definition for CZ gate [70]), with probability 1/2 each. Since almost all induced leakage comes from $|0\rangle$ (see Sec. I B), this means that $p_f^{|2\rangle} \approx 0$ for $p_i^{|1\rangle} = 1$ and that $p_f^{|2\rangle} \approx 2L_1^{\mathrm{LRU}}$ for $p_i^{|0\rangle} = 1$ (neglecting relaxation effects as the used $T_1 = 30\ \mu$s is relatively long). Combining these two definitions one gets the expression

$$p_f^{|2\rangle} \approx (1 - R)p_i^{|2\rangle} + 2L_1^{\mathrm{LRU}} p_i^{|0\rangle} \tag{16}$$

for an arbitrary incoming state. Note that, given these definitions, Figs. 2(a) and 2(c) respectively show a heatmap of $1 - R$ and $2L_1^{\mathrm{LRU}}$ for the considered transmon-resonator parameters. In particular, the operating point achieves $R \approx 99.5\%$ and $L_1^{\mathrm{LRU}} \approx 0.25\%$. The achieved leakage reduction can be compared with that given purely by relaxation during T_{slot}, namely $R_{T_1} = 1 - e^{-T_{\mathrm{slot}}/(T_1/2)} = 2.9\%$, which shows that the LRU provides a much stronger additional seepage channel.

2. *π-LRU for ancilla qubits*

The dispersive readout of a transmon qubit is in general performed by sending a pulse to the readout resonator, integrating the reflected signal to obtain a point in the in-phase and quadrature (IQ) plane and depleting the photons in the resonator (either passively by relaxation or actively with another pulse) [16–18]. The measured point is compared to one or more thresholds to declare the measurement outcome. These thresholds are determined as to optimally separate the distributions for the different outcomes, which have a Gaussian(-like) form. Here we assume that the distribution for $|2\rangle$ is sufficiently separated from $|0\rangle$ and $|1\rangle$ [16]. This is generally expected to be possible thanks to the different dispersive shift. Then one uses three thresholds in the IQ plane to distinguish between $|0\rangle$, $|1\rangle$, and $|2\rangle$ (or two if $|2\rangle$ is well separated from, e.g., $|0\rangle$). We also assume that an outcome can be declared during photon depletion, thus enabling real-time conditional feedback. This is challenging to perform in 200–300 ns in experiment due to the classical-postprocessing requirements, but it has been previously achieved [28,56]. We can then apply the π-LRU immediately at the end of the depletion time. The $|1\rangle \leftrightarrow |2\rangle$ π pulse is expected to be implementable as a simple pulse in the same way and time as single-qubit gates (20 ns) and with comparable, coherence-limited fidelity.

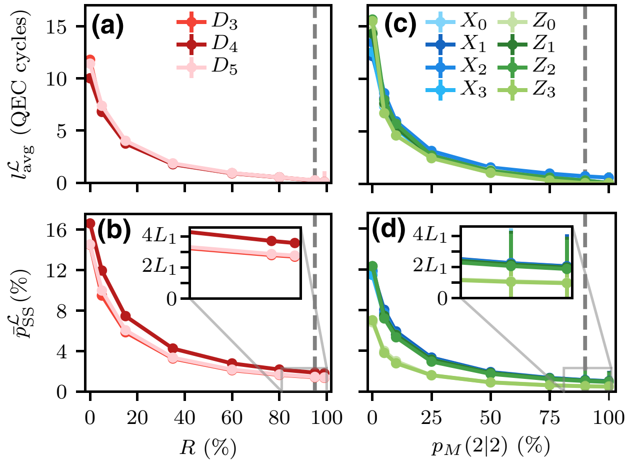

FIG. 4. Average leakage lifetime $l_{\mathrm{avg}}^{\mathcal{L}}$ [(a),(c)] and leakage steady state $\bar{p}_{\mathrm{SS}}^{\mathcal{L}}$ [(b),(d)] as a function of the leakage-reduction rate R for data qubits [(a),(b)] and as a function of the readout probability $p_M(2|2)$ for ancilla qubits [(c),(d)]. Here we fix the CZ leakage rate to $L_1 = 0.5\%$. The insets in (b) and (d) show that $\bar{p}_{\mathrm{SS}}^{\mathcal{L}}$ tends to approximately $N_{\mathrm{flux}}L_1$ ($N_{\mathrm{flux}} = 4$ for D_4, 3 for D_3, D_5, 1 for Z_0, Z_3, and 2 for the remaining ancilla qubits). The vertical dashed lines correspond to the values used in Sec. II D. These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are mostly smaller than the symbol size.

If conditional feedback is not possible in the allotted time, one can either increase the QEC-cycle duration (at the cost of extra decoherence for all qubits, scaling as $1 - e^{-t_{\mathrm{extra}}/T_2}$ per qubit per QEC cycle) or postpone the conditional gate to the next QEC cycle. In the latter case, one source of error corresponds to the ancilla qubit already seeping before the application of the π-LRU, which then causes it to leak instead. The probability of this error is already low and is expected to become even lower with longer T_1 times and lower-leakage CZ gates. The other errors are the Z rotations (depending on the leakage conditional phases) that the leaked ancilla qubit spreads for at least one extra QEC cycle, as well as the fact that the parity-check stays disabled. We do not simulate these variants and we expect a relatively low logical-performance loss, corresponding to an average leakage lifetime of about two QEC cycles (see Figs. 4 and 9).

Readout-declaration errors are expected to affect the performance of the π-LRU. On the one hand, an incorrect declaration of $|1\rangle$ as a $|2\rangle$ makes the π pulse induce leakage. On the other hand, declaring a $|2\rangle$ as a $|1\rangle$ would lead to leakage not being corrected and lasting for at least one extra QEC cycle. We define the readout matrix M with entries $M_{ij} =: p_M(i|j)$ being the probability that the actual state $|j\rangle$ resulting from the projective measurement is declared as an $|i\rangle$. In the simulations we use

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & p_M(1|1) & 1 - p_M(1|1) \\ 0 & 1 - p_M(2|2) & p_M(2|2) \end{pmatrix}. \tag{17}$$

In particular, this means that we do not consider declaration errors within the computational subspace. While that would change the value of the logical error rate since the error syndrome gets corrupted, it is not relevant for evaluating the performance of the π-LRU since a $|0\rangle$ mistaken for a $|1\rangle$ or vice versa does not trigger the π-LRU anyway. Furthermore, we assume that a $|0\rangle$ cannot be mistaken as a $|2\rangle$ since their readout signals are often much more separated than the signals of $|1\rangle$ and $|2\rangle$. Note that if a $|0\rangle$ (rather than a $|1\rangle$, as we assume in this work) could be mistakenly declared as a $|2\rangle$ then a $|1\rangle \leftrightarrow |2\rangle$ π pulse does not induce leakage, so here we consider the worst-case scenario for the π-LRU.

C. Average leakage lifetime and leakage steady state

Once a qubit leaks, it tends to remain leaked for a significant amount of time, up to 10–15 QEC cycles on average [46]. Starting from an initial state with no leakage, the probability that a qubit is in the leaked state tends towards a steady state within a few QEC cycles. It was shown in Ref. [46] that this evolution is well captured by a classical Markov process with leakage (seepage) rate $\Gamma_{\mathcal{C}\to\mathcal{L}}$ ($\Gamma_{\mathcal{L}\to\mathcal{C}}$) per QEC cycle, where $\mathcal{C}$ ($\mathcal{L}$) is the computational (leakage) subspace. Note that here $\mathcal{L}$ is one dimensional, corresponding to $|2\rangle$. In our error model, without accounting for LRUs, these rates are approximately given by

$$\Gamma_{\mathcal{C}\to\mathcal{L}} \approx N_{\text{flux}} L_1, \tag{18}$$

$$\Gamma_{\mathcal{L}\to\mathcal{C}} \approx N_{\text{flux}} L_2 + (1 - e^{-t_c/(T_1/2)}), \tag{19}$$

where N_{flux} is in how many CZ gates the transmon is fluxed during a QEC cycle, t_c is the duration of a QEC cycle, and L_1 (L_2) is the average leakage (seepage) probability of a CZ gate [70]. Thus, the two native mechanisms that generate seepage are the CZ gates themselves and relaxation.

The major effect of a LRU is to effectively increase $\Gamma_{\mathcal{L}\to\mathcal{C}}$ in Eq. (19) by introducing an extra seepage mechanism. Hence, we expect that $\Gamma^{\text{LRU}}_{\mathcal{L}\to\mathcal{C}} \sim \Gamma_{\mathcal{L}\to\mathcal{C}} + R$ for data qubits and $\Gamma^{\text{LRU}}_{\mathcal{L}\to\mathcal{C}} \sim \Gamma_{\mathcal{L}\to\mathcal{C}} + p_M(2|2)$ for ancilla qubits, preventing leakage from accumulating and lasting long for large R or $p_M(2|2)$.

The average leakage lifetime $l^{\mathcal{L}}_{\text{avg}}$ is the average duration of leakage and, for a Markov process, it is calculated as

$$\begin{aligned} l^{\mathcal{L}}_{\text{avg}} &= \sum_{n=1}^{\infty} n\mathbb{P}(\text{stay in } \mathcal{L} \text{ for } n \text{ QEC cycles}) \\ &= \sum_{n=1}^{\infty} n(1 - \Gamma_{\mathcal{L}\to\mathcal{C}})^{n-1}\Gamma_{\mathcal{L}\to\mathcal{C}} \\ &= \frac{1}{\Gamma_{\mathcal{L}\to\mathcal{C}}}, \end{aligned} \tag{20}$$

thus assuming that the qubit starts in $\mathcal{L}$. The evolution of the leakage probability $\bar{p}^{\mathcal{L}}(n)$, averaged over surface-code runs, as a function of the QEC-cycle number n is well approximated by [46]

$$\bar{p}^{\mathcal{L}}(n) = \frac{\Gamma_{\mathcal{C}\to\mathcal{L}}}{\Gamma_{\mathcal{C}\to\mathcal{L}} + \Gamma_{\mathcal{L}\to\mathcal{C}}}(1 - e^{-(\Gamma_{\mathcal{C}\to\mathcal{L}}+\Gamma_{\mathcal{L}\to\mathcal{C}})n}). \tag{21}$$

The steady state is the long-time limit and is given by

$$\bar{p}^{\mathcal{L}}_{\text{SS}} = \lim_{n\to\infty} \bar{p}^{\mathcal{L}}(n) = \frac{\Gamma_{\mathcal{C}\to\mathcal{L}}}{\Gamma_{\mathcal{C}\to\mathcal{L}} + \Gamma_{\mathcal{L}\to\mathcal{C}}}. \tag{22}$$

For ancilla qubits, $\bar{p}^{\mathcal{L}}(n)$ can be computed directly from the "true" measurement outcomes (i.e., without declaration errors on top). For data qubits, it can be computed from the density matrix. Specifically, for data qubits, we evaluate $\bar{p}^{\mathcal{L}}(n)$ immediately after the CZ gates.

Figure 4 shows $l^{\mathcal{L}}_{\text{avg}}$ and $\bar{p}^{\mathcal{L}}_{\text{SS}}$ extracted from the Surface-17 simulations by fitting $\bar{p}^{\mathcal{L}}(n)$ to Eq. (21) for each qubit. We can indeed observe that these quantities drop substantially for both data and ancilla qubits. The decays follow an inverse proportionality as, e.g., for data qubits

$$l^{\mathcal{L}}_{\text{avg}} = \frac{1}{\Gamma^{\text{LRU}}_{\mathcal{L}\to\mathcal{C}}} \sim \frac{1}{\Gamma_{\mathcal{L}\to\mathcal{C}} + R} \sim \frac{1}{R}, \tag{23}$$

$$\bar{p}^{\mathcal{L}}_{\text{SS}} = \frac{\Gamma^{\text{LRU}}_{\mathcal{C}\to\mathcal{L}}}{\Gamma^{\text{LRU}}_{\mathcal{C}\to\mathcal{L}} + \Gamma^{\text{LRU}}_{\mathcal{L}\to\mathcal{C}}} \sim \frac{\Gamma^{\text{LRU}}_{\mathcal{C}\to\mathcal{L}}}{\Gamma^{\text{LRU}}_{\mathcal{L}\to\mathcal{C}}} \sim \frac{\Gamma^{\text{LRU}}_{\mathcal{C}\to\mathcal{L}}}{R}, \tag{24}$$

for sufficiently large R and small $\Gamma^{\text{LRU}}_{\mathcal{C}\to\mathcal{L}}$. For ancilla qubits, we expect, similarly, a $1/p_M(2|2)$ dependence. The lifetime drops from values $\gtrsim 10$ to approximately 1, which is the minimum value it can achieve (some points drop below 1 within error bars as it is difficult for the fit to estimate such a short lifetime). As of course the LRUs do not prevent leakage from occurring during the CZ gates in the first place, one cannot expect the steady state to reach 0 even for a perfect LRU ($R = 1$), but rather $\bar{p}^{\mathcal{L}}_{\text{SS}} \sim \Gamma^{\text{LRU}}_{\mathcal{C}\to\mathcal{L}} \approx N_{\text{flux}}L_1$ ($+L_1^{\text{LRU}}$ if the LRU can mistakenly induce leakage). Figures 4(b) and 4(d) show that this is indeed the case.

Figure 4 also demonstrates that both $l^{\mathcal{L}}_{\text{avg}}$ and $\bar{p}^{\mathcal{L}}_{\text{SS}}$ get close to their minimum values already for

$R, p_M(2|2) \gtrsim 80\%$. This suggests that res-LRU and π-LRU may not necessarily need to be perfect to provide a good logical performance in Surface-17. This means that one could use, e.g., a weaker pulse to implement the res-LRU or that the readout of $|2\rangle$ may not need to be particularly optimized in practice.

D. Logical performance

In the simulations the logical qubit is initialized in $|0\rangle_L$ and the logical fidelity $\mathcal{F}_L(n)$ is computed at the end of each QEC cycle as the probability that the decoder correctly determines whether a logical error has occurred or not. We do not perform a similar analysis with initial state $|+\rangle_L$ or other states as the density-matrix simulations are computationally expensive and we expect a similar performance. The logical error rate ε_L per QEC cycle can be extracted by fitting $\mathcal{F}_L(n) = [1 + (1 - 2\varepsilon_L)^{n-n_0}]/2$, where n_0 is a fitting parameter (usually close to 0) [50]. We evaluate ε_L for the upper bound (UB) decoder that uses the complete density-matrix information to infer a logical error, and for the minimum-weight perfect-matching (MWPM) decoder. Detailed information about these decoders can be found in Refs. [50,72] and an overview is given in Appendix C 1 b.

By mapping a leaked qubit back to the computational subspace, a LRU does not fully remove a leakage error but can at most convert it into a regular (Pauli) error. Hence, it is not to be expected that ε_L in the presence of leakage can be restored to the value at $L_1 = 0$. We consider realistic parameters for the LRUs. Specifically, we use $R = 95\%$, $L_1^{\text{LRU}} = 0.25\%$, $p_M(2|2) = 90\%$, and $p_M(1|1) = 99.5\%$. We have shown in Sec. I B that the first two parameters can be attained with realistic parameters for the transmon-readout system, while the last two are close to achievable in experiment [14,53]. In particular, while the operating point has $R = 99.5\%$, we conservatively choose $R = 95\%$ here. Note that $p_M(1|1) = 99.5\%$ is quite high. We argue that the state of the art can be squeezed as the threshold to distinguish between $|1\rangle$ and $|2\rangle$ in the IQ plane could be moved towards $|2\rangle$, rather than placing it in the middle as is common practice. In this way one would slightly reduce $p_M(2|2)$ in favor of $p_M(1|1)$ if $p_M(1|1)$ is not high enough. A broader study of the logical performance as a function of the LRU parameters can be found in Appendix C 2.

Figure 5 shows the reduction in ε_L as a function of the CZ leakage rate L_1 when LRUs with the given parameters are employed. Using only the res-LRU or the π-LRU lowers $\varepsilon_L^{\text{MWPM}}$ by basically the same amount, while $\varepsilon_L^{\text{UB}}$ is lower for the π-LRU than for the res-LRU. We attribute this to the fact that the UB decoder directly uses the information in the density matrix, while the MWPM decoder relies on the measured syndrome, thus being more susceptible to ancilla-qubit leakage. When both LRUs are used, we see that ε_L is reduced by an amount that is close to the sum of the reductions when only one kind of LRU is used. As expected, ε_L is not restored to the value at $L_1 = 0$, but the reduction is overall significant and can reach up to 30% for both MWPM and UB decoders compared to the case without LRUs.

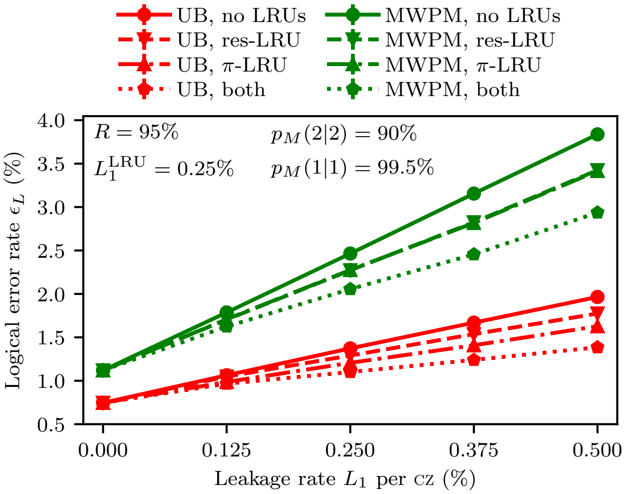

FIG. 5. Logical error rate ε_L per QEC cycle for the upper bound (red) and minimum-weight perfect-matching (green) decoders versus the CZ leakage rate L_1, in the cases with no LRUs, only res-LRU, only π-LRU, and both LRUs (the point without leakage at $L_1 = 0$ is always without LRUs as well). These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are smaller than the symbol size.

III. DISCUSSION

In this work we have introduced a leakage-reduction scheme using res-LRUs and π-LRUs that does not require any additional hardware or a longer QEC cycle. Furthermore, while the scheme in Ref. [48] is applicable only to ancilla qubits, our combination of res-LRU for data qubits and π-LRU for ancilla qubits enables us to significantly reduce leakage in the whole transmon processor. We have shown with detailed simulations using realistic parameters that the reset scheme in Refs. [53–55] can be adapted to be a LRU without significantly affecting the states in the computational subspace, allowing us to unconditionally apply the res-LRU in the surface code. The use of the res-LRU for data qubits, as well as the use of the π-LRU for ancilla qubits, leads to a substantial reduction of the average leakage lifetime and leakage steady state, preventing leakage from lasting more than approximately one QEC cycle on average, even when the LRUs are imperfect and can introduce leakage themselves. Using full density-matrix simulations of Surface-17, we have demonstrated that this leads to a significant reduction of the logical error rate for both the UB and MWPM decoders.

Regarding the practical implementation of the res-LRU, the required drive amplitude is relatively strong, similarly to that used in the experiments in Refs. [53–55]. It is thus important that the microwave crosstalk is minimized by careful engineering of the drive lines. Furthermore, in a multitransmon processor it is relevant that the drive frequency does not accidentally match any two-qubit or neighboring single-qubit transitions. For example, in the original scheme in Ref. [57] that we followed, the target frequencies are 6.7, 6.0, and 4.9 GHz for high-, mid-, and low-frequency qubits, respectively, and 7.8 GHz for the readout resonator [27]. In particular, the mid-frequency qubits (the ancilla qubits) are parked around 5.4–5.5 GHz during measurement, with their $|1\rangle \leftrightarrow |2\rangle$ transition around 5.1–5.2 GHz. This is close to the optimal drive frequency found in Sec. I B (approximately 5.25 GHz), which can lead to an indirect ancilla-qubit drive mediated by the bus resonator, albeit weaker. The difficulty of precise frequency targeting in fabrication can further lead to undesired frequency collisions. These issues can be alleviated by choosing slightly different transmon and resonator frequencies and anharmonicities to make the drive more off-resonant with that transition (combined with better frequency targeting [73]), or they can be mitigated altogether by using tunable couplers [1,11,14]. The res-LRU is compatible with tunable-coupler schemes and their possibly different operation scheduling than in Ref. [57], as well as potentially applicable to superconducting qubits that use a resonator for dispersive readout other than the transmon. Tunable couplers would also be advantageous to fully protect the res-LRU performance from residual *ZZ* crosstalk, even though we find that a cumulative *ZZ* interaction up to about 2 MHz can be tolerated with fixed couplers (see Appendix B 3). Beside this, if the low-frequency data qubits can leak depending on the implementation of the CZ gate, the res-LRU can be applied to them in the same time slot as the high-frequency ones. If the thermal population in the readout resonator is relatively high in a given experiment, the effect of a correspondingly high L_1^{LRU} can potentially be mitigated by applying res-LRU conditionally on the detection of leakage by a set of hidden Markov models [46].

Regarding the viability of inserting the res-LRU in the surface-code time scheduling, the necessary condition is that $t_p \leq T_{\mathrm{slot}}$. We can express T_{slot} as $T_{\mathrm{slot}} = t_m - 4t_{\mathrm{CZ}}$, where t_m is the measurement time for the ancilla qubits. Slower CZ gates might make T_{slot} too short, although CZ gates even faster than 40 ns (as assumed here) have been realized in 15 ns [12]. The measurement time can be further broken down into readout-pulse time and photon-depletion time, $t_m = t_{\mathrm{read}} + t_{\mathrm{depl}}$. Both of these would be reduced by a larger κ; however, assuming that the κ of ancilla- and data-qubit resonators are comparable, t_p would be reduced as well. Even if we keep t_p and t_{CZ} fixed to the values in this work, we get $t_m \geq 340$ ns, which is significantly lower than $t_m = 580$ ns, as considered here. A desirable, additional condition to the necessary one is that $T_{\mathrm{slot}} - t_p \geq 4/\kappa$, i.e., that there is enough leftover time in T_{slot} to allow for the data-qubit resonator to return the thermal state, where we estimate that four decay constants would suffice (together with the fact that the resonator is already relaxing during t_p). Assuming similar depletion time for data- and ancilla-qubit resonators, this roughly means that the res-LRU is easily applicable if t_p is smaller or similar to t_{read}. Note that in this work we have $T_{\mathrm{slot}} - t_p \sim 16/\kappa$ and $t_p < t_{\mathrm{read}}$. If the additional condition above is not satisfied, one could demand that at least the resonator has returned to the thermal state before the res-LRU in the following QEC cycle, i.e., $T_{\mathrm{slot}} - t_p + 8t_{\mathrm{CZ}} + 2t_H \geq 4/\kappa$. In this case the disadvantage would be that the presence of a fraction of a photon in the resonator would cause additional data-qubit dephasing, especially during the first few CZ gates. As the extra photon is present only when the qubit was previously leaked, we expect this disadvantage to be small as long as the overall leakage rate is small. If even the relaxed additional condition is violated, on top of the additional dephasing the resonator would also heat up, effectively leading to a higher L_1^{LRU} in the QEC cycle(s) following that in which the qubit leaked. As this effect also scales with L_1, we expect that it would not be an issue as long as κ is not very low (allowing for at most one extra QEC cycle to thermalize we get $\kappa/2\pi \geq 1$ MHz). Otherwise, leakage would not really be removed from the system but would be largely moved back and forth from the transmon to the resonator.

The demonstrated reduction in the average leakage lifetime and in the logical error rate is expected to lead to a higher noise threshold for the surface code in the presence of leakage, compared to the case without LRUs. Furthermore, for error rates below threshold (both regular and leakage), we believe that the logical error rate would be exponentially suppressed with increasing code distance when employing LRUs. Without LRUs, this might hold only when the code distance is sufficiently larger than the average leakage lifetime ($d \gg l_{\mathrm{avg}}^{\mathcal{L}}$). For smaller distances, the relatively long correlated error chains induced by leakage might lead to a subexponential scaling. To study the noise threshold and subthreshold behavior, it is necessary to implement simulations of large code sizes that use a simplified error model, such as a stochastic error model for leakage and Pauli errors [23,39,42]. We expect that the demonstrated MWPM logical error rate can be further lowered by the use of decoders [23,39,42,59–61] that use information about leakage extracted directly or indirectly (e.g., with hidden Markov models [46]) from the measurement outcomes.

The data underlying this work, as well as the code to analyze it, are available online [74]. The code used to generate the data is available upon request to the corresponding author.

ACKNOWLEDGMENTS

We thank L. DiCarlo for his comments on the manuscript. F.B., B.M.V., and B.M.T. are supported by ERC Grant EQEC No. 682726. Most simulations were performed with computing resources granted by RWTH Aachen University under projects rwth0566 and rwth0669.

APPENDIX A: APPROXIMATE TRANSMON-RESONATOR HAMILTONIAN

1. Schrieffer-Wolff transformation

In this appendix we explain the concept of the Schrieffer-Wolff transformation (SWT) [62–64] and derive the equations that we use in the following sections.

Consider a Hamiltonian

$$H = H_0 + \epsilon V, \tag{A1}$$

expressed in a certain basis $\{|\psi_n\rangle\}$, where H_0 is block diagonal with respect to this basis and the perturbation V can be taken as block off-diagonal without loss of generality (block-diagonal terms can be included in the definition of H_0). Furthermore, we assume that $||V|| = \mathcal{O}(1)$ and $\epsilon \ll \Delta_{ij}$, where we set Δ_{ij} as the minimum energy separation between blocks i and j.

The SWT corresponds to finding an anti-Hermitian matrix S such that

$$H' := e^S H e^{-S} \tag{A2}$$

is block diagonal. In other words, calling $\{|\bar{\psi}_n\rangle\}$ the basis of eigenstates of H, $e^S = \sum_n |\psi_n\rangle\langle\bar{\psi}_n|$. The matrix S can be expanded in a series

$$S = \sum_{k=1}^{\infty} \epsilon^k S_k, \tag{A3}$$

where each S_k is block off-diagonal. If $\epsilon \ll \Delta_{ij}$, one can expect the first order (S_1) to provide a good approximation; otherwise, one needs to consider higher orders depending on ϵ (although the series does not always converge for extensive systems [63]). Using the Baker-Campbell-Hausdorff formula, one gets

$$H' = e^S H e^{-S} = \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{[S, [S, \ldots, [S}_{k \text{ times}}, H], \ldots]]. \tag{A4}$$

The procedure for the SWT is to group terms of the same order in ϵ in this formula and set the block off-diagonal part of H' to 0, thus getting equations for $\{S_k\}$, in the usual case with *two* blocks [63]. One uses the relationships

$$[\text{diagonal}, \text{diagonal}] = \text{diagonal}, \tag{A5}$$

$$[\text{diagonal}, \text{off-diagonal}] = \text{off-diagonal}, \tag{A6}$$

$$[\text{off-diagonal}, \text{off-diagonal}] = \text{diagonal}. \tag{A7}$$

However, the last line only holds for the case with two blocks. In the following we consider the generalization of the SWT to the case with an arbitrary number of blocks [64]. We use the notation $O_{\rm D}$ and $O_{\rm OD}$ for the block diagonal and off-diagonal parts of an operator $O = O_{\rm D} + O_{\rm OD}$, respectively.

Here we expand H and S up to $k = 3$ in Eq. (A4), assuming that the fourth-order block off-diagonal term is negligible. We get the following pieces:

$$H_0 \quad \text{at the zeroth order,} \tag{A8}$$

$$V + [S_1, H_0] \quad \text{at the first order,} \tag{A9}$$

$$[S_1, V] + \tfrac{1}{2}[S_1, S_1, H_0] + [S_2, H_0] \quad \text{at the second order,} \tag{A10}$$

$$\begin{aligned}&[S_2, V] + \tfrac{1}{2}([S_2, S_1, H_0] + [S_1, S_1, V] + [S_1, S_2, H_0]) \\ &\quad + \tfrac{1}{6}[S_1, S_1, S_1, H_0] + [S_3, H_0] \quad \text{at the third order,}\end{aligned} \tag{A11}$$

$$\begin{aligned}&[S_3, V] + \tfrac{1}{2}([S_1, S_3, H_0] + [S_2, S_2, H_0] + [S_3, S_1, H_0] \\ &\quad + [S_1, S_2, V] + [S_2, S_1, V]) + \tfrac{1}{6}([S_1, S_1, S_1, V] \\ &\quad + [S_2, S_1, S_1, H_0] + [S_1, S_2, S_1, H_0] + [S_1, S_1, S_2, H_0]) \\ &\quad + \tfrac{1}{24}[S_1, S_1, S_1, S_1, H_0] \quad \text{at the fourth order.}\end{aligned} \tag{A12}$$

Setting the block off-diagonal parts at first, second, and third orders to 0 we get

$$[H_0, S_1] = V, \tag{A13}$$

$$[H_0, S_2] = \tfrac{1}{2}[S_1, V]_{\rm OD}, \tag{A14}$$

$$\begin{aligned}[H_0, S_3] = {}& \tfrac{1}{2}[S_2, V]_{\rm OD} + \tfrac{1}{3}[S_1, S_1, V_{\rm D}]_{\rm OD} \\ & + \tfrac{1}{12}[S_1, S_1, V_{\rm OD}]_{\rm OD},\end{aligned} \tag{A15}$$

where we have used the first equation to simplify the following ones. These equations can be solved iteratively for S_k (given knowledge of the eigenstates of H_0). The Hamiltonian H' is then block diagonal up to fourth order and is explicitly given by

$$\begin{aligned}H' = {}& H_0 + \frac{\epsilon^2}{2}[S_1, V]_{\rm D} + \epsilon^3 \left(\tfrac{1}{2}[S_2, V]_{\rm D} + \tfrac{1}{12}[S_1, S_1, V_{\rm OD}]_{\rm D}\right) \\ & + \epsilon^4 \left(\tfrac{1}{2}[S_3, V]_{\rm D} - \tfrac{1}{24}[S_1, S_1, S_1, V_{\rm DOD}]_{\rm D} \right. \\ & \left. - \tfrac{1}{6}[S_2, S_1, V_{\rm OD}]_{\rm D} + \tfrac{1}{12}[S_1, S_2, V_{\rm OD}]_{\rm D}\right).\end{aligned} \tag{A16}$$

This expression has been simplified using Eqs. (A13) to (A15), together with the fact that, e.g., $[S_k, \ldots, \ldots_{\mathrm{D}}]_{\mathrm{D}} = 0$ since S_k is block off-diagonal.

2. SWT of the capacitive coupling

We consider the Hamiltonian $H = H_0 + H_c + H_d$ of a driven transmon capacitively coupled to a resonator, as given in Eqs. (1) to (4).

The SWT of H_c up to first order in the perturbation parameter $\epsilon = g/\Delta$, where $\Delta = \omega_q - \omega_r$, is implemented using the matrix [65]

$$S_1 = g \sum_{m=1}^{\infty} \frac{\sqrt{m}}{\Delta + \alpha(m-1)} (a|m\rangle\langle m-1| - \mathrm{H.c.}), \quad \text{(A17)}$$

where $\{|m\rangle\}$ are transmon states and where we have absorbed ϵ in the definition of S_1. The Hamiltonian in the unitarily transformed frame as defined in Sec. I A is then given by

$$H^D \approx e^{S_1} H e^{-S_1} = e^{S_1}(H_0 + H_c)e^{-S_1} + e^{S_1} H_d e^{-S_1} \quad \text{(A18)}$$

with

$$\begin{aligned} e^{S_1}(H_0 + H_c)e^{-S_1} &= H_0 + \tfrac{1}{2}[S_1, H_c] \\ &\approx \delta^r a^\dagger a + \sum_{m=1}^{\infty} \left(m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}} \right) |m\rangle\langle m| \\ &\quad - a^\dagger a \sum_{m=0}^{\infty} \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| \qquad \text{(A19)} \\ &:= H_0^D, \end{aligned}$$

where we define $\Delta_m = \Delta + \alpha m = \Delta - |\alpha| m$ as $\alpha < 0$ for transmons. The second term above contains a Stark shift of the transmon frequency and the last term is the state-dependent dispersive shift. The approximation in Eq. (A19) is due to the fact that we have ignored a double-excitation exchange term coming from $[S_1, H_c]$, since it is proportional to $g\alpha/(\Delta_m \Delta_{m-1})$. This is negligible for low anharmonicity and also for $\omega_r > \omega_q$ as then $\Delta < 0$ and $|\Delta_m|$ increases with m. If instead $\omega_r < \omega_q$, $\Delta > 0$ and $|\Delta_m|$ decreases with m, so even if the approximation is good for the two lowest levels, there can be some higher level that does not sit well within the dispersive regime. However, in this work we consider a system with $\omega_r > \omega_q$; hence, we do not need to take this into account.

The drive Hamiltonian in the unitarily transformed frame takes the form

$$e^{S_1} H_d e^{-S_1} = \underbrace{\frac{\Omega e^{i\phi}}{2} b + \mathrm{H.c.}}_{:=H_{d1}^D} + \underbrace{\frac{\Omega e^{i\phi}}{2} \left(a \sum_{m=0}^{\infty} \frac{g\Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| + a^\dagger \sum_{m=0}^{\infty} \frac{g\alpha\sqrt{m+1}\sqrt{m+2}}{\Delta_m \Delta_{m+1}} |m\rangle\langle m+2| \right) + \mathrm{H.c.}}_{:=H_{d2}^D} \quad \text{(A20)}$$

The last term contains a first-order approximation in g/Δ of the $|20\rangle \leftrightarrow |01\rangle$ effective coupling $\tilde{g}$, which is linear in Ω. However, the "pure" drive term H_{d1}^D can be quite strong, so we need to evaluate how it affects $\tilde{g}$ and the rest of the Hamiltonian.

3. SWT of the pure drive Hamiltonian

Summarizing, in the unitarily transformed frame the original Hamiltonian H takes (approximately) the form

$$H^D \approx H_0^D + H_{d1}^D + H_{d2}^D, \quad \text{(A21)}$$

where H_0^D is given in Eq. (A19) and H_{d1}^D, H_{d2}^D are given in Eq. (A20).

We now want to find an additional SWT transformation $S' = S_1' + S_2' + S_3'$, with H_{d1}^D taking the role of V in Appendix A 1, defining a "double-dressed" Hamiltonian

$$\begin{aligned} H^{\mathrm{DD}} &:= e^{S'} H^D e^{-S'} \\ &= \underbrace{e^{S'}(H_0^D + H_{d1}^D)e^{-S'}}_{=:H_0^{\mathrm{DD}}} + \underbrace{e^{S'} H_{d2}^D e^{-S'}}_{=:H_d^{\mathrm{DD}}}, \end{aligned} \quad \text{(A22)}$$

such that H_0^{DD} is fully diagonal up to third order in the perturbation parameter $\epsilon = \Omega/\delta^q$. Then H_d^{DD} gives the couplings within the manifold of interest ($|20\rangle, |01\rangle$) and outside of it. We absorb ϵ^k in the definition of S_k' so it does not explicitly appear below.

Following Appendix A 1, to find S_1', we need to solve Eq. (A13), i.e.,

$$[H_0^D, S_1'] = H_{d1}^D, \quad \text{(A23)}$$

in this specific case. Bracketing it with the eigenstates $\{|ml\rangle\}$ of H_0^D, with the notation $|\text{transmon, resonator}\rangle$, we get the matrix elements of S_1' as

$$\langle ml|S_1'|nk\rangle = \frac{\langle ml|H_{d1}^D|nk\rangle}{E_{ml}^D - E_{nk}^D}, \tag{A24}$$

where $\{E_{ml}^D\}$ are the eigenenergies of H_0^D, which can be easily inferred from Eq. (A19). We neglect the dispersive shift since it is proportional to α/Δ. Then

$$\langle ml|S_1'|nk\rangle = \frac{\Omega}{2}\left(-\frac{\sqrt{m+1}\delta_{m,n-1}\delta_{l,k}}{\delta^q + \alpha m + g^2\Delta_{-1}/(\Delta_{m-1}\Delta_m)}e^{i\phi} + \frac{\sqrt{m}\delta_{m,n+1}\delta_{l,k}}{\delta^q + \alpha(m-1) + g^2\Delta_{-1}/(\Delta_{m-2}\Delta_{m-1})}e^{-i\phi}\right), \tag{A25}$$

where $\delta_{i,j}$ is the Kronecker delta. From this equation one can infer that

$$S_1' = -\frac{\Omega}{2}e^{i\phi}\sum_{m=0}^{\infty}\frac{\sqrt{m+1}}{\delta_m^q}|m\rangle\langle m+1| - \text{H.c.}, \tag{A26}$$

where we have defined $\delta_m^q = \delta^q + \alpha m + g^2\Delta_{-1}/(\Delta_{m-1}\Delta_m)$.

Having derived S_1', we can compute S_2' from Eq. (A14), i.e.,

$$[H_0^D, S_2'] = \tfrac{1}{2}[S_1', H_{d1}^D]_{\text{OD}} \tag{A27}$$

with

$$[S_1', H_{d1}^D] = -\frac{\Omega^2}{2}\sum_{m=0}^{\infty}\frac{\tilde{\delta}_m^q}{\delta_m^q\delta_{m-1}^q}|m\rangle\langle m| - \frac{\Omega^2}{4}\sum_{m=0}^{\infty}\sqrt{m+1}\sqrt{m+2}\left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q}\right)(e^{2i\phi}|m\rangle\langle m+2| + \text{H.c.}), \tag{A28}$$

where $\tilde{\delta}_m^q = \delta^q - \alpha + g^2\Delta_{-1}\Delta_{3m}/(\Delta_m\Delta_{m-1}\Delta_{m-2})$. Clearly, the first term is the diagonal part while the second term is the off-diagonal part. With a similar procedure as that used for S_1', it follows that

$$S_2' = \frac{\Omega^2}{8}e^{2i\phi}\sum_{m=0}^{\infty}\frac{\sqrt{m+1}\sqrt{m+2}}{\delta_m^q + \delta_{m+1}^q}\left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q}\right)|m\rangle\langle m+2| - \text{H.c.} \tag{A29}$$

We can then compute S_3' from Eq. (A15), i.e.,

$$[H_0^D, S_3'] = \tfrac{1}{2}[S_2', H_{d1}^D]_{\text{OD}} + \tfrac{1}{3}[S_1', S_1', H_{d1\text{D}}^D]_{\text{OD}} + \tfrac{1}{12}[S_1', S_1', H_{d1\text{OD}}^D]_{\text{OD}}. \tag{A30}$$

The result is

$$\begin{aligned} S_3' = \Omega^3 e^{i\phi}\sum_{m=0}^{\infty}|m\rangle\langle m+1|\Bigg\{\frac{1}{12}\frac{\sqrt{m+1}}{(\delta_m^q)^3}\left(\frac{\tilde{\delta}_{m+1}^q}{\delta_{m+1}^q} - \frac{\tilde{\delta}_m^q}{\delta_{m-1}^q}\right) + \frac{1}{96\delta_m^q}\Bigg[(m+2)\sqrt{m+1}\frac{\delta_m^q + 4\delta_{m+1}^q}{\delta_{m+1}^q(\delta_m^q + \delta_{m+1}^q)}\left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q}\right) \\ - \sqrt{m+1}m\frac{4\delta_{m-1}^q + \delta_m^q}{\delta_{m-1}^q(\delta_{m-1}^q + \delta_m^q)}\left(\frac{1}{\delta_{m-1}^q} - \frac{1}{\delta_m^q}\right)\Bigg]\Bigg\} - \text{H.c.} + \frac{\Omega^3}{96}e^{3i\phi}\sum_{m=0}^{\infty}|m\rangle\langle m+3|\frac{\sqrt{m+1}\sqrt{m+2}\sqrt{m+3}}{\delta_m^q + \delta_{m+1}^q + \delta_{m+2}^q} \\ \times\left[\frac{3\delta_{m+2}^q - \delta_{m+1}^q - \delta_m^q}{\delta_{m+2}^q(\delta_m^q + \delta_{m+1}^q)}\left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q}\right) - \frac{3\delta_m^q - \delta_{m+1}^q - \delta_{m+2}^q}{\delta_m^q(\delta_{m+1}^q + \delta_{m+2}^q)}\left(\frac{1}{\delta_{m+1}^q} - \frac{1}{\delta_{m+2}^q}\right)\right] - \text{H.c.} \end{aligned} \tag{A31}$$

We can eventually use Eqs. (A26), (A29), and (A31) together with Eq. (A16) to obtain H_0^{DD} [defined in Eq. (A22)]:

$$\begin{aligned}
H_0^{\mathrm{DD}} = {} & \delta^r a^\dagger a + \sum_{m=0}^{\infty} |m\rangle\langle m| \Bigg(m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}} - \frac{\Omega^2 \tilde{\delta}_m^q}{4\delta_m^q \delta_{m-1}^q} - \frac{\Omega^4}{32} \Bigg[\frac{m+1}{(\delta_m^q)^3} \left(\frac{\tilde{\delta}_{m+1}^q}{\delta_{m+1}^q} - \frac{\tilde{\delta}_m^q}{\delta_{m-1}^q} \right) \\
& - \frac{m}{(\delta_{m-1}^q)^3} \left(\frac{\tilde{\delta}_m^q}{\delta_m^q} - \frac{\tilde{\delta}_{m-1}^q}{\delta_{m-2}^q} \right) \Bigg] - \frac{\Omega^4}{192} \Bigg\{ \frac{1}{\delta_m^q} \Bigg[(m+2)(m+1) \frac{\delta_m^q + 5\delta_{m+1}^q}{\delta_{m+1}^q (\delta_m^q + \delta_{m+1}^q)} \left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q} \right) \\
& - (m+1)m \frac{5\delta_{m-1}^q + \delta_m^q}{\delta_{m-1}^q (\delta_{m-1}^q + \delta_m^q)} \left(\frac{1}{\delta_{m-1}^q} - \frac{1}{\delta_m^q} \right) \Bigg] - \frac{1}{\delta_{m-1}^q} \Bigg[(m+1)m \frac{\delta_{m-1}^q + 5\delta_m^q}{\delta_m^q (\delta_{m-1}^q + \delta_m^q)} \left(\frac{1}{\delta_{m-1}^q} - \frac{1}{\delta_m^q} \right) \\
& - m(m-1) \frac{5\delta_{m-2}^q + \delta_{m-1}^q}{\delta_{m-2}^q (\delta_{m-2}^q + \delta_{m-1}^q)} \left(\frac{1}{\delta_{m-2}^q} - \frac{1}{\delta_{m-1}^q} \right) \Bigg] \Bigg\} + \frac{\Omega^4}{96} \Bigg[\frac{(m+2)(m+1)}{\delta_m^q + \delta_{m+1}^q} \left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q} \right)^2 \\
& - \frac{m(m-1)}{\delta_{m-2}^q + \delta_{m-1}^q} \left(\frac{1}{\delta_{m-2}^q} - \frac{1}{\delta_{m-1}^q} \right)^2 \Bigg] \Bigg) - a^\dagger a \sum_m \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m|. \qquad \text{(A32)}
\end{aligned}$$

We note that this expression implicitly contains all cross terms between the perturbative parameters g/Δ and Ω/δ^q up to the chosen orders. The approximate coupling Hamiltonian H_d^{DD} [defined in Eq. (A22)] up to second order in Ω/δ^q is instead given by

$$\begin{aligned}
H_d^{\mathrm{DD}} &= H_{d2}^D + [S_1', H_{d2}^D] + [S_2', H_{d2}^D] + \tfrac{1}{2}(S_1', [S_1', H_{d2}^D]) \\
&=: H_{\mathrm{eff.\,coupl.}}^{\mathrm{DD}} + H_{\mathrm{resid.}}^{\mathrm{DD}}, \qquad \text{(A33)}
\end{aligned}$$

where

$$\begin{aligned}
H_{\mathrm{eff.\,coupl.}}^{\mathrm{DD}} = {} & e^{i\phi} a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+2| \Bigg\{ \tilde{g}_m \Bigg[1 - \frac{\Omega^2}{8} \left(\frac{m+3}{(\delta_{m+2}^q)^2} + \frac{m+2}{(\delta_{m+1}^q)^2} + \frac{m+1}{(\delta_m^q)^2} + \frac{m}{(\delta_{m-1}^q)^2} \right) \Bigg] \\
& + \frac{\Omega^2}{4} \left(\frac{\sqrt{m+1}\sqrt{m+3}}{\delta_m^q \delta_{m+2}^q} \tilde{g}_{m+1} + \frac{\sqrt{m}\sqrt{m+2}}{\delta_{m-1}^q \delta_{m+1}^q} \tilde{g}_{m-1} \right) + \frac{\Omega^2}{4} \sqrt{m+1}\sqrt{m+2} \\
& \times \left(\frac{g_{m+2}'}{\delta_m^q (\delta_m^q + \delta_{m+1}^q)} - \frac{g_{m+1}'}{\delta_m^q \delta_{m+1}^q} + \frac{g_m'}{\delta_{m+1}^q (\delta_m^q + \delta_{m+1}^q)} \right) \Bigg\} + \mathrm{H.c.} \qquad \text{(A34)}
\end{aligned}$$

with

$$\tilde{g}_m := \frac{g\alpha\Omega\sqrt{m+1}\sqrt{m+2}}{2\Delta_m \Delta_{m+1}}, \qquad \text{(A35)}$$

$$g_m' := \frac{g\Omega\Delta_{-1}}{2\Delta_m \Delta_{m-1}}, \qquad \text{(A36)}$$

and

$$\begin{aligned}
H_{\mathrm{resid.}}^{\mathrm{DD}} = {} & (e^{i\phi} a + \mathrm{H.c.}) \sum_{m=0}^{\infty} |m\rangle\langle m| \Bigg\{ g_m' \Bigg[1 - \frac{\Omega^2}{4} \left(\frac{m+1}{(\delta_m^q)^2} + \frac{m}{(\delta_{m-1}^q)^2} \right) \Bigg] + \frac{\Omega^2}{4} \left(\frac{m+1}{(\delta_m^q)^2} g_{m+1}' + \frac{m}{(\delta_{m-1}^q)^2} g_{m-1}' \right) \\
& + \frac{\Omega^2}{4} \left(\frac{\sqrt{m+1}\sqrt{m+2}\tilde{g}_m}{\delta_m^q (\delta_m^q + \delta_{m+1}^q)} + \frac{\sqrt{m}\sqrt{m+1}\tilde{g}_{m-1}}{\delta_m^q \delta_{m-1}^q} + \frac{\sqrt{m-1}\sqrt{m}\tilde{g}_{m-2}}{\delta_{m-1}^q (\delta_{m-2}^q + \delta_{m-1}^q)} \right) \Bigg\} \\
& - \frac{\Omega}{2} e^{2i\phi} a \sum_{m=0}^{\infty} |m\rangle\langle m+1| \frac{\sqrt{m+1}}{\delta_m^q} (g_{m+1}' - g_m') + \mathrm{H.c.}
\end{aligned}$$

$$
\begin{aligned}
&-\frac{\Omega}{2}a^{\dagger}\sum_{m=0}^{\infty}|m\rangle\langle m+1|\left(\frac{\sqrt{m+1}}{\delta^q_m}(g'_{m+1}-g'_m)+\frac{\sqrt{m+2}}{\delta^q_{m+1}}\tilde{g}_m-\frac{\sqrt{m}}{\delta^q_{m-1}}\tilde{g}_{m-1}\right)+\text{H.c.}\\
&+\frac{\Omega^2}{4}e^{3i\phi}a\sum_{m=0}^{\infty}|m\rangle\langle m+2|\sqrt{m+1}\sqrt{m+2}\left(\frac{g'_{m+2}}{\delta^q_m(\delta^q_m+\delta^q_{m+1})}-\frac{g'_{m+1}}{\delta^q_m\delta^q_{m+1}}+\frac{g'_m}{\delta^q_{m+1}(\delta^q_m+\delta^q_{m+1})}\right)+\text{H.c.}\\
&-\frac{\Omega}{2}e^{2i\phi}a^{\dagger}\sum_{m=0}^{\infty}|m\rangle\langle m+3|\left(\frac{\sqrt{m+1}}{\delta^q_m}\tilde{g}_{m+1}-\frac{\sqrt{m+3}}{\delta^q_{m+2}}\tilde{g}_m\right)+\text{H.c.}\\
&+\frac{\Omega^2}{4}e^{3i\phi}a^{\dagger}\sum_{m=0}^{\infty}|m\rangle\langle m+4|\left(\frac{\sqrt{m+1}\sqrt{m+2}\tilde{g}_{m+2}}{\delta^q_m(\delta^q_m+\delta^q_{m+1})}-\frac{\sqrt{m+4}\sqrt{m+1}\tilde{g}_{m+1}}{\delta^q_m\delta^q_{m+3}}+\frac{\sqrt{m+3}\sqrt{m+4}\tilde{g}_m}{\delta^q_{m+3}(\delta^q_{m+3}+\delta^q_{m+2})}\right)+\text{H.c.}\qquad\text{(A37)}
\end{aligned}
$$

All terms in $H^{\text{DD}}_{\text{resid.}}$ are relatively small and off-resonant with the $|20\rangle \leftrightarrow |01\rangle$ transition so we expect them to have a small effect and we do not proceed with higher orders of SWT s.

4. Analysis of the $|20\rangle \leftrightarrow |01\rangle$ avoided crossing

In this appendix we give the methods used to calculate the curves in Figs. 1(c) and 1(e).

We define ω_d^* as the drive frequency corresponding to the center of the $|20\rangle \leftrightarrow |01\rangle$ avoided crossing of the full Hamiltonian H as given in Eq. (1). Then the exact value of the effective $|20\rangle \leftrightarrow |01\rangle$ coupling $\tilde{g}$ is given by half the energy separation at that point. The avoided crossing can be found numerically by exact diagonalization as a function of ω_d.

In the subspace $\mathcal{S} = \text{span}\{|20\rangle, |01\rangle\}$ we can write H as $H|_{\mathcal{S}} \equiv -\eta(\omega_d)Z/2 + \tilde{g}(\omega_d)[\cos(\phi)X + \sin(\phi)Y] = -\eta(\omega_d)Z/2 + \tilde{g}(\omega_d)X$ for $\phi = 0$ as in Sec. I A. As we want to implement a $|20\rangle \leftrightarrow |01\rangle$ π rotation, we note that the choice of ϕ, i.e., the choice of rotation axis in the equator of the Bloch sphere, is irrelevant. We have also ignored a term proportional to the identity I, which gives a phase difference with respect to states outside of $\mathcal{S}$, in particular between the computational and leakage subspaces of the transmon. However, this phase is irrelevant if $|20\rangle$ is swapped entirely onto $|01\rangle$ since the latter decays and dephases fast, thus suppressing any phase coherence. As demonstrated in Sec. I B, the res-LRU can reach a very high R, for which the effect of this phase is then minimal. Assuming that $H^{\text{DD}}_{\text{resid.}}$ in Eq. (A37) is negligible, an analytical approximation of η is given by

$$\eta(\omega_d) \approx \langle 20|H_0^{\text{DD}}(\omega_d)|20\rangle - \langle 01|H_0^{\text{DD}}(\omega_d)|01\rangle, \qquad \text{(A38)}$$

where we have made the dependence of H_0^{DD} in Eq. (A32) on ω_d explicit. This holds since then H_0^{DD} accounts for all the Stark shifts of $|20\rangle$ and $|01\rangle$ due to the capacitive coupling and the drive (up to the given orders). The center of the avoided crossing is found by imposing the condition $\eta(\omega_d) = 0$. As the explicit expression that can be extracted from Eq. (A32) is not analytically solvable, we use the secant method available in *scipy* to find ω_d^* that fulfills this condition in Eq. (A38). It is then straightforward to compute the (approximate) analytical estimate for the effective coupling as $\tilde{g}(\omega_d^*) = |\langle 01|H^{\text{DD}}_{\text{eff. coupl.}}(\omega_d^*)|20\rangle|$ from Eq. (A34), which is plotted in Fig. 1(e).

APPENDIX B: FURTHER CHARACTERIZATION OF THE READOUT-RESONATOR LRU

1. Effective T_1 and T_2 due to the drive

In this appendix we discuss the effects of the readout-resonator LRU within the computational subspace when applied to a nonleaked transmon. As pulses at different (ω_d, Ω) points have a different duration t_p, it would not be fair to report an effective T_1 and T_2 during t_p. That is, stronger pulses potentially produce lower T_1 and T_2, but they also take less time to implement the LRU. However, the overall disturbance to the qubit is a combination of these two factors. We thus report an effective T_1 and T_2 during the whole time slot of $T_{\text{slot}} = 440$ ns, leading to a uniform metric for the whole (ω_d, Ω) landscape. Specifically, to estimate T_1, we prepare the state $|1\rangle\langle 1| \otimes \sigma_{\text{th}}$, we simulate the Lindblad equation in Eq. (12), and we evaluate the remaining population $p^{|1\rangle}$ in $|1\rangle$ at the end of the time slot after tracing out the resonator. Assuming that $p^{|1\rangle} = e^{-T_{\text{slot}}/T_1}$ we then compute T_1 by inverting this formula. To estimate T_2, we prepare $|+\rangle\langle +| \otimes \sigma_{\text{th}}$ and we evaluate the decay of the off-diagonal transmon matrix element $|0\rangle\langle 1|$ as this is directly available in simulation (rather than simulating a full Ramsey experiment). We then invert $|\langle 0|\,\text{Tr}_r[\rho(T_{\text{slot}})]|1\rangle| = e^{-T_{\text{slot}}/T_2}/2$ to get T_2.

Figure 6 shows the resulting effective T_1 and T_2. In Fig. 6(a) one can see that T_1 decreases by at most 15% as a function of Ω, showing that a short t_p mostly counterbalances the effect of a strong Ω. In particular, $T_1 \approx 27.1$ μs at the operating point. On the other hand, one can note that T_1 dips around $\Omega_{\text{cr}}/2\pi = 143$ MHz, where the pulses are

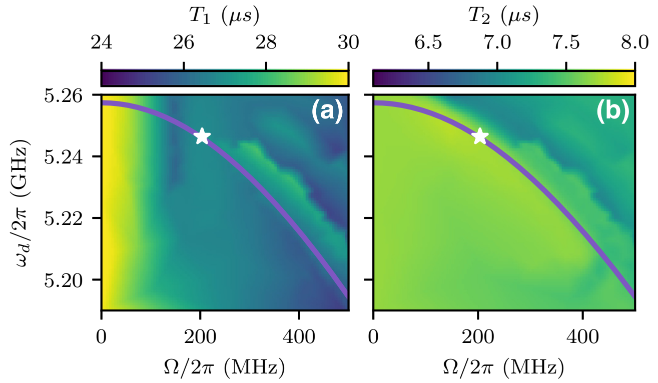

FIG. 6. Effective T_1 (a) and T_2 (b) that account for the extra decoherence caused by the drive during the time slot $T_{\text{slot}} =$ 440 ns. We can see that the variation is small as a function of the drive amplitude compared to the values at $\Omega = 0$. The white star indicates the chosen operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_p = 178.6$ ns; see Sec. I B). The purple line corresponds to the higher-order estimate of the optimal drive frequency ω_d^* as a function of Ω [see Fig. 1(c)]. The heatmaps are sampled using the ADAPTIVE package [66].

very long, suggesting that driving slightly into the underdamped regime is favorable. In Fig. 6(b) one can see that the value of T_2 is about 7.7 μs at $\Omega = 0$, i.e., when no pulse is applied. This has to be contrasted with the input T_2 parameter of 30 μs inserted in the Lindblad equation (see Table I). We assume that that implicitly accounts for dephasing caused by flux noise only. Photon-shot noise from the resonator is a further dephasing source that is explicitly included in these simulations. The combination of flux and photon-shot noise leads to the actual effective T_2 reported in Fig. 6(b). We note that if $\bar{n} = 0$ then the effective T_2 at $\Omega = 0$ would exactly match the input of 30 μs. While the effective T_2 can be restored from 7.7 to 30 μs with colder resonators or by engineering different system parameters altogether, the important information from Fig. 6(b) is that T_2 barely changes as a function of Ω. Combined with the similar result for T_1, this means that the drive causes only a marginal effect within the computational subspace. Note that in the region where the readout-resonator LRU is most effective [just above the purple line in Fig. 6(b)], T_2 is even slightly higher than at $\Omega = 0$ (7.9 versus 7.7 μs). We attribute this to the fact that the pulse temporarily reduces the excited-state population in the resonator [see Fig. 2(d)]. In this way photon-shot noise is reduced until the resonator rethermalizes, but at the cost of some leakage of the transmon.

In Fig. 2(d) one can note that a non-negligible amount of population ends up in $|10\rangle$ from the initial state $|0\rangle\langle 0| \otimes \sigma_{\text{th}}$. This corresponds to an excitation rate $T_1^{\uparrow} \approx 256$ μs at the operating point. We backtrack this source of error to a combination of the drive and the jump operator $a^{\dagger}$, corresponding to the drive inducing a transmon excitation rate based on the resonator excitation rate. However, as here $T_1^{\uparrow} \gg \max\{T_1, T_2\}$, it is not a limiting factor and we have not included it in the Surface-17 simulations.

2. Long-drive limit in the underdamped regime and its drawbacks as a LRU

In this appendix we compare the reset schemes in Refs. [54,55] versus Ref. [53] in terms of their performance as a LRU in the underdamped regime. The approach in Refs. [54,55], which we have adopted in Sec. I B, aims at swapping $|20\rangle$ and $|01\rangle$ by targeting the first minimum of the oscillations induced by the drive (switching the drive off afterwards). As shown in Sec. I B, this approach allows for a residual leakage population $p_{\text{op.}}^{|2\rangle} \approx 0.5\%$ at the operating point [see Fig. 2(a)], given our parameters (see Table I). While this already reaches thermal-state levels (here $\bar{n} = 0.5\%$) with the considered system parameters, the approach in Ref. [53] could be used in general to achieve an even lower or similar $p^{|2\rangle}$ (in particular for lower κ).

The approach in Ref. [53] keeps the drive on for a much longer period of time (at least one more oscillation), allowing both the populations in $|20\rangle$ and $|01\rangle$ to decay to almost 0, modulo thermal excitations. Figure 7 shows that it is indeed possible to suppress these populations to thermal-state levels, where we use the same (Ω, ω_d) as at the operating point (see Sec. I B). However, we see that, for the operating point, there is almost no gain by using this approach. Furthermore, this approach costs much more time and could exceed $T_{\text{slot}} = 440$ ns if κ is not as high as assumed here. In particular, in that case the first few minima after the first one could be slightly higher, due to transmon decoherence, and one would need to wait even longer to overcome this effect.

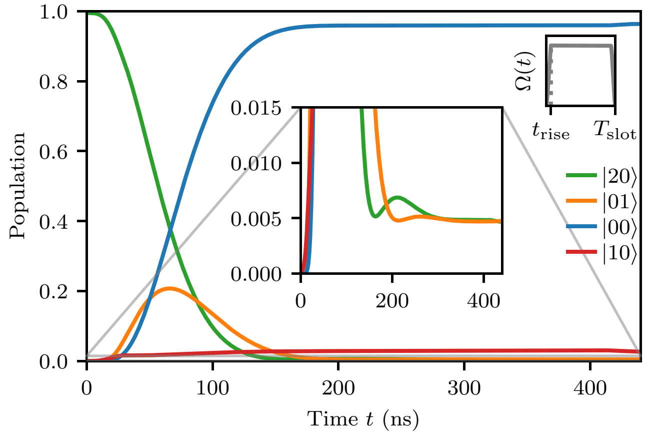

FIG. 7. Time evolution from the initial state $|2\rangle\langle 2| \otimes \sigma_{\text{th}}$ for $t_{\text{rise}} = 30$ ns and for an otherwise always-on drive during T_{slot}. This is simulated with the same $\Omega/2\pi \approx 204$ MHz and $\omega_d/2\pi \approx$ 5.2464 GHz as at the operating point in Fig. 2.

Another disadvantage of the approach in Ref. [53] is that the disturbance to the qubit is stronger as the drive is kept on for a longer period of time. For example, in Fig. 7 one can see that $|00\rangle$ and $|10\rangle$ reach an equilibrium thanks to the drive (even in the presence of relaxation), where the population in $|10\rangle$ is higher than in Fig. 2(b). By evaluating T_1 we find that $T_1 \approx 23$ μs instead of 27 μs (see Appendix B 1). Furthermore, if one would have to use a $t_p > T_{\text{slot}}$ when κ is lower than here, then the QEC cycle would get longer, affecting the coherence of all qubits, not only of the high-frequency data qubits to which the res-LRU is applied.

3. Sensitivity to residual *ZZ* crosstalk

In a multitransmon chip, each transmon is coupled to one or more neighbors. In general, if the coupling is not tunable, there can be some residual *ZZ* crosstalk, i.e., a shift of the transmon frequency by an amount ζ based on whether each neighboring transmon is in $|1\rangle$ instead of $|0\rangle$. In this appendix we study the effect of this *ZZ* coupling on the readout-resonator LRU, which we assume being tuned up when all neighbors are in $|0\rangle$. We do not include neighboring transmons in our simulations, so we mimic it by shifting the transmon frequency (while keeping the drive parameters fixed).

In Fig. 8 we perform the analysis for the operating point (see Sec. I B), which resides in the underdamped regime, and for the critical point. In both cases the leakage-reduction rate R scales seemingly quadratically. In the underdamped regime the pulse targets the first minimum of the damped Rabi oscillations, so it is more sensitive to a variation in frequency than in the critical regime. However, we can observe that, for $|\zeta|/2\pi \lesssim 2$ MHz (note that this is the cumulative *ZZ* coupling over all neighbors), R stays above 95%, which is the conservative value we used in Sec. II D and for which the logical error rate is already close to optimal in Surface-17 (see Appendix C 2). Regarding other performance parameters of the LRU, we find that L_1^{LRU} scales in the same relative way as R by unitarity, whereas T_1, T_2, and $T_1^{\uparrow}$ vary by $\lesssim 1\%$.

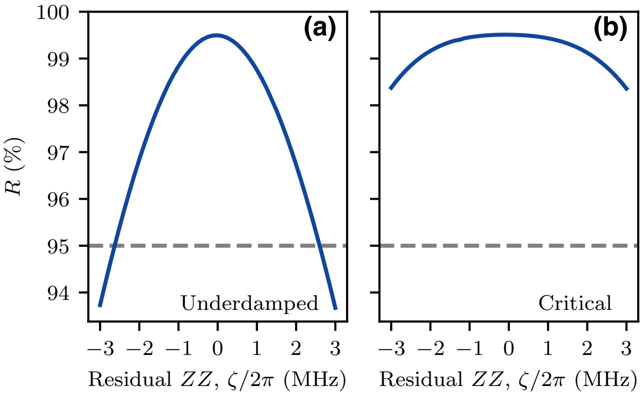

FIG. 8. Sensitivity of the leakage-reduction rate R of the readout-resonator LRU as a function of the overall residual *ZZ* coupling ζ. (a) Underdamped regime, specifically at the operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_p = 178.6$ ns; see Sec. I B). (b) Critical regime ($\Omega/2\pi \approx 143$ MHz, $\omega_d/2\pi \approx 5.252$ GHz, $t_p = 440$ ns).

APPENDIX C: FURTHER SURFACE-17 CHARACTERIZATION

1. Details about the density-matrix simulations

The parameters used in this work are reported in Table II.

a. res-LRU in QUANTUMSIM

A comprehensive review of the density-matrix simulations and the use of the QUANTUMSIM package [58] is available in Refs. [46,50]. In this appendix we explain the specific implementation of the newly introduced res-LRU, expressed in the Pauli transfer matrix formalism.

We construct a “phenomenological” Lindblad model with input parameters R, L_1^{LRU}, and $t_{\text{res-LRU}}$. We use the Pauli transfer matrix $S_{\text{res-LRU}} = S_{\uparrow} S_{\downarrow}$, where $S_{\downarrow}$ is the Pauli transfer matrix of the superoperator $\mathcal{S}_{\downarrow} = e^{t_{\text{res-LRU}}\mathcal{L}_{\downarrow}}$ and the Lindbladian $\mathcal{L}_{\downarrow}$ has the quantum jump operator

$$K_{\downarrow} = \frac{1}{\sqrt{t_{\text{res-LRU}}/[-\log(1-R_{\text{sim}})]}} |0\rangle\langle 2| \qquad \text{(C1)}$$

TABLE II. The parameters for the qubit coherence times and for the gate, LRU, measurement, and QEC-cycle durations used in the density-matrix simulations. The interaction point corresponds to the frequency to which a transmon is fluxed to implement a CZ gate, whereas the parking point corresponds to the frequency at which the ancilla qubits are parked during measurement [57].

Parameter	Value
Relaxation time T_1	30 μs
Sweetspot pure-dephasing time $T_{\phi,\text{max}}$	60 μs
High-frequency pure-dephasing time at interaction point $T_{\phi,\text{int}}$	8 μs
Mid-frequency pure-dephasing time at interaction point $T_{\phi,\text{int}}$	6 μs
Mid-frequency pure-dephasing time at parking point $T_{\phi,\text{park}}$	8 μs
Low-frequency pure-dephasing time at parking point $T_{\phi,\text{park}}$	9 μs
Single-qubit gate time t_{gate}	20 ns
Two-qubit interaction time t_{int}	30 ns
Single-qubit phase-correction time t_{PC}	10 ns
Readout-resonator LRU time $t_{\text{res-LRU}}$	100 ns
$|1\rangle \leftrightarrow |2\rangle$ π-pulse time $t_{\pi\text{-LRU}}$	20 ns
Measurement time t_m	580 ns
QEC-cycle time t_c	800 ns

with R_{sim} to be determined. Besides this, $\mathcal{L}_\downarrow$ has the standard qutrit jump operators for relaxation and dephasing [46]. On the other hand, $S_\uparrow$ is the Pauli transfer matrix of the superoperator $\mathcal{S}_\uparrow = e^{\mathcal{L}_\uparrow}$ and the Lindbladian $\mathcal{L}_\uparrow$ has a single jump operator

$$K_\uparrow = \frac{1}{\sqrt{1/[-\log(1-2L_1^{\mathrm{LRU}})]}}|2\rangle\langle 0|, \qquad \text{(C2)}$$

since relaxation and dephasing during $t_{\text{res-LRU}}$ are already accounted for by $S_\downarrow$. In this way, calling $p_i^{|j\rangle}, p_f^{|j\rangle}$ the populations before and after the res-LRU, if we apply $S_{\text{res-LRU}}$ on a nonleaked transmon, we get $p_f^{|2\rangle} = 2L_1^{\mathrm{LRU}} p_i^{|0\rangle}$, consistently with Sec. II B 1. Instead, if we apply $S_{\text{res-LRU}}$ to a leaked transmon ($p_i^{|2\rangle} = 1$), we get $p_f^{|2\rangle} \approx 1 - R_{\mathrm{sim}} + 2L_1^{\mathrm{LRU}}$. By fixing $R_{\mathrm{sim}} = R + 2L_1^{\mathrm{LRU}}$ we match the definition of R in Sec. II B 1 as well. The approximation is very good for large R and low L_1^{LRU}, which is precisely the interesting regime for res-LRU that we have explored.

b. Decoding

In this appendix we provide additional information on the UB and MWPM decoders [50,72].

The UB decoder considers the 32 computational states that differ by a purely X error on top of $|0\rangle_L$ and that are independent (i.e., they cannot be obtained from each other by multiplication with an X-type stabilizer). At the end of each QEC cycle n, each possible final Z syndrome is compatible with a pair of these states, where one can be associated with $|0\rangle_L$ and the other with $|1\rangle_L$ as they differ by the application of any representation of X_L. The largest overlap of these two states with the diagonal of the density matrix at QEC cycle n corresponds to the maximum probability of correctly guessing whether a X_L error has occurred or not upon performing a logical measurement of Z_L. The latter is assumed to be performed by measuring all data qubits in the $\{|0\rangle, |1\rangle, |2\rangle\}$ basis and computing the overall parity. To compute the parity, we assume that a $|2\rangle$ is declared as a $|1\rangle$ since decoders usually do not use information about leakage (and since measurements often declare $|2\rangle$ as a $|1\rangle$ rather than as a $|0\rangle$). Then the UB decoder computes $\mathcal{F}_L(n)$ by weighing this probability with the chance of measuring the given final Z syndrome (conditioned on the density matrix) and by summing over all possible syndromes. In other words, the UB decoder always finds the correction that maximizes the likelihood of the logical measurement returning the initial state, here $|0\rangle_L$. As the UB decoder uses information generally hidden in the density matrix, it gives an upper bound to the performance of any realistic decoder, which can at most use the syndrome information extracted via the ancilla qubits.

The MWPM decoder tries to approximate the most likely correction by finding the lowest weight correction, which is a good approximation when physical error rates are relatively low. As the ancilla qubits can be faulty, the decoding graph is three dimensional. In particular, we allow for spacelike edges corresponding to data-qubit errors, timelike edges corresponding to ancilla-qubit errors, and spacetimelike edges corresponding to data-qubit errors occurring in the middle of the

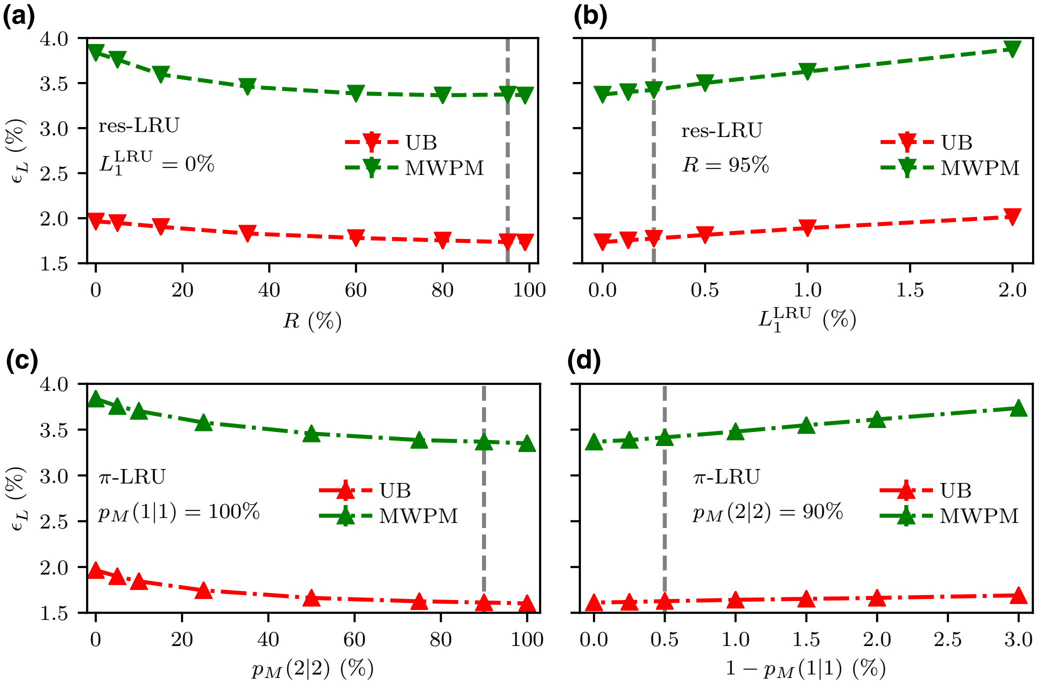

FIG. 9. Logical error rate ϵ_L per QEC cycle as a function of various LRU parameters. In (a) and (b) we use only the res-LRU, while in (c) and (d) we use the π-LRU. We fix $L_1 = 0.5\%$ for all. Vertical dashed lines indicate the values considered in Sec. II D. These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are smaller than the symbol size.

parity-check circuit. The weights are extracted with the adaptive algorithm in Ref. [75] from a simulation (10^5 runs of 20 QEC cycles each) without leakage and an otherwise identical error model. Similarly to the UB decoder, for decoding, we assume that a $|2\rangle$ is declared as a $|1\rangle$ since the standard MWPM decoder does not account for leakage.

2. Logical error rate as a function of the LRU parameters

We study the variation in the logical error rate ε_L per QEC cycle as a function of the performance parameters of the LRUs. Here we fix $L_1 = 0.5\%$ as it is easier to visualize variations in ε_L with a relatively large L_1. The leakage-reduction rate R and the readout probability $p_M(2|2)$ play similar roles for the res-LRU and π-LRU, respectively. In Figs. 9(a) and 9(c) one can see that this is the case and that the values of ε_L at the parameters used in Sec. II D [$R = 95\%$ and $p_M(2|2) = 90\%$] are very close to their best values (at least for this system size). This shows that the advantages of a larger R or $p_M(2|2)$ are marginal. We attribute this to the fact that leakage is exponentially suppressed with an already quite large exponent. Furthermore, the parameters L_1^{LRU} and $1 - p_M(1|1) = p_M(2|1)$, regulating the induced leakage, play similar roles as well, as Figs. 9(b) and 9(d) show. We see that ε_L is more sensitive to $L_1{}^{\mathrm{LRU}}$ and $1 - p_M(1|1)$ compared to R and $p_M(2|2)$. In particular, we see that ε_L is slightly larger at the parameters used in Sec. II D [$L_1^{\mathrm{LRU}} = 0.25\%$ and $1 - p_M(1|1) = 0.5\%$] rather than at 0, although the difference is small.

3. Effect of the leakage conditional phases on the logical error rate

As defined in the main text, the leakage conditional phases are the phases that a nonleaked transmon acquires when interacting with a leaked one during a CZ gate. Here we denote them as $\phi^{\mathcal{L}}_{\mathrm{flux}}$ and $\phi^{\mathcal{L}}_{\mathrm{stat}}$ depending on whether the lower- or the higher-frequency transmon of the pair is leaked, respectively, and we use $\phi^{\mathcal{L}}$ to indicate either of them. Furthermore, in this appendix we use the notation $|\text{low-frequency transmon, high-frequency transmon}\rangle$. Note that, for a CZ gate between two qutrits, in principle, there are nine phases ($\phi_{00}, \phi_{01}, \phi_{10}, \phi_{11}, \phi_{02}, \phi_{20}, \phi_{21}, \phi_{12}, \phi_{22}$), where the first four are fixed to $0, 0, 0, \pi$, respectively. Of the five phases containing a $|2\rangle$ we consider only two of them here, i.e., $\phi^{\mathcal{L}}_{\mathrm{stat}} = \phi_{02} - \phi_{12}$ and $\phi^{\mathcal{L}}_{\mathrm{flux}} = \phi_{20} - \phi_{21}$ as defined above. This is because in our leakage model [46] we set to 0 the coherence between the computational and leakage subspaces of each qutrit, motivated by the fact that leakage is projected relatively fast and that the stabilizer measurements ideally prevent any interference effect. This means that the individual phases are global phases, whereas their difference cannot be gauged away when the nonleaked qubit is in a superposition of $|0\rangle$ and $|1\rangle$.

For a flux-based CZ gate with conditional phase π for $|11\rangle$, ideally one should have $\phi^{\mathcal{L}}_{\mathrm{flux}} = 0$ and $\phi^{\mathcal{L}}_{\mathrm{stat}} = \pi$ [46] as $|02\rangle$ acquires a conditional phase equal and opposite to $|11\rangle$. If only $|12\rangle$ and $|21\rangle$ are coupled in the three-excitation manifold, it holds that $\phi^{\mathcal{L}}_{\mathrm{stat}} = \pi - \phi^{\mathcal{L}}_{\mathrm{flux}}$. The strength of the repulsion times the CZ duration gives, e.g.,

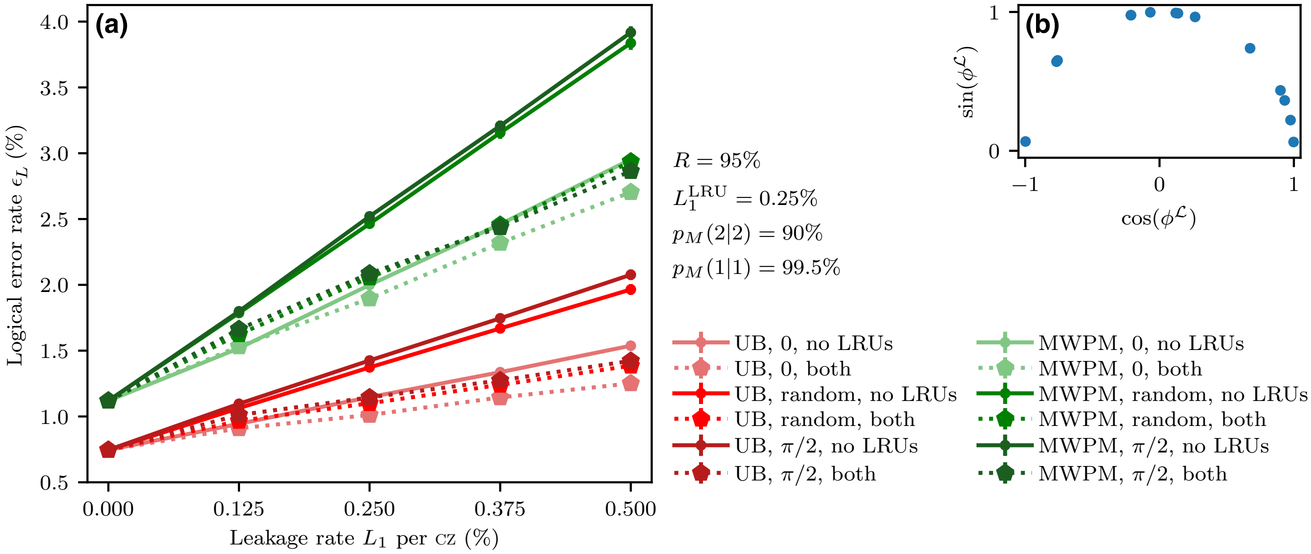

FIG. 10. Variation of the logical error rate ε_L for different choices of leakage conditional phases $\phi^{\mathcal{L}}$. (a) The logical error rate ε_L per QEC cycle for the UB (shades of red) and MWPM (shades of green) decoders versus L_1, in the cases with no LRUs and both LRUs, each for all $\phi^{\mathcal{L}}$ set to $0, \pi/2$, or uniformly random in $[0, \pi]$. These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are mostly smaller than the symbol size. (b) The random values for $\phi^{\mathcal{L}}$ used across this work. These values are extracted from a uniform distribution in $[0, \pi]$. We have excluded negative values as $\pm\phi^{\mathcal{L}}$ corresponds to the same chance of spreading a Z error under the twirling action of the parity-check measurements.

$\phi^{\mathcal{L}}_{\text{flux}} \sim \pi/4$ for the parameters in Ref. [46]. However, $|03\rangle$ interacts with $|12\rangle$ and $|21\rangle$ and breaks the relationship above, for which we can consider $\phi^{\mathcal{L}}_{\text{flux}}$ and $\phi^{\mathcal{L}}_{\text{stat}}$ as effectively unconstrained. The randomized values used across the main text are reported in Fig. 10(b). We use 14 values, of which three for $\phi^{\mathcal{L}}_{\text{stat}}$ and three for $\phi^{\mathcal{L}}_{\text{flux}}$ when each high-frequency data qubit is leaked or interacts with a leaked ancilla qubit, respectively, and eight only for $\phi^{\mathcal{L}}_{\text{stat}}$ when each ancilla qubit is leaked and interacts with a low-frequency data qubit (as low-frequency data qubits cannot leak themselves).

In this appendix we study the dependence of the logical error rate ε_L on the leakage conditional phases, without discussing how one would engineer the system to tune them to certain values. The best-case scenario to minimize ε_L is to set all $\phi^{\mathcal{L}} = 0$, since no Z rotations are spread then. Instead, the worst-case scenario corresponds to all $\phi^{\mathcal{L}} = \pi/2$, since, under the twirling effect of the parity-check measurements, this corresponds to spreading a Z error with 50% chance. Note that, if all $\phi^{\mathcal{L}} = \pi$, overall the spread errors amount to a stabilizer (except in the QEC cycle in which leakage occurs), so it is close to the best-case scenario.

Figure 10(a) compares the logical performance for both UB and MWPM decoders in the cases where $\phi^{\mathcal{L}} = 0$, $\phi^{\mathcal{L}} = \pi/2$ and when they are random as in Fig. 5 and in the rest of this work. First, one can note that the performance of random $\phi^{\mathcal{L}}$ is very close to the worst-case scenario ($\phi^{\mathcal{L}} = \pi/2$). This is due to the fact that it is not necessary to spread an error on every qubit with 50% chance each to cause a logical error with high probability. One can also note that just tuning all $\phi^{\mathcal{L}} = 0$ without implementing LRUs is almost as good (or even better) as using the LRUs when the $\phi^{\mathcal{L}}$ are random. We attribute this to the fact that one of the major effects of the LRUs is to prevent correlated errors being spread by a leaked qubit for many QEC cycles. Tuning $\phi^{\mathcal{L}} = 0$ achieves this as well, but it still does not address the fact that the code distance is effectively reduced if a data qubit stays leaked and that the full stabilizer information is not accessible as long as an ancilla qubit is leaked. Indeed, using LRUs even when $\phi^{\mathcal{L}} = 0$ always allows for a lower logical error rate [see Fig. 10(a)]. Furthermore, the reduction in distance and the corruption of the stabilizer information suggest that a threshold would still likely be low without using LRUs.

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

[2] P. Jurcevic, A. Javadi-Abhari, L. S. Bishop, I. Lauer, D. F. Bogorin, M. Brink, L. Capelluto, O. Günlük, T. Itoko, N. Kanazawa, *et al.*, Demonstration of quantum volume 64 on a superconducting quantum computing system, Quantum Sci. Technol. **6**, 025020 (2021).

[3] L. Egan, D. M. Debroy, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Newman, M. Li, K. R. Brown, M. Cetina, and C. Monroe, Fault-Tolerant Operation of a Quantum Error-Correction Code, ArXiv:2009.11482 (2020).

[4] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo, Restless Tuneup of High-Fidelity Qubit Gates, Phys. Rev. Appl. **7**, 041001 (2017).

[5] Z. Chen, J. Kelly, C. Quintana, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Lucero, *et al.*, Measuring and Suppressing Quantum State Leakage in a Superconducting Qubit, Phys. Rev. Lett. **116**, 020501 (2016).

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

[7] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, Procedure for systematically tuning up cross-talk in the cross-resonance gate, Phys. Rev. A **93**, eid 060302 (2016).

[8] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M. P. da Silva, and B. R. Johnson, Demonstration of a parametrically activated entangling gate protected from flux noise, Phys. Rev. A **101**, 012302 (2020).

[9] 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).

[10] 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).

[11] 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**, eid 054062 (2018).

[12] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Z. Chen, K. Satzinger, R. Barends, *et al.*, Demonstrating a Continuous Set of Two-Qubit Gates for Near-Term Quantum Algorithms, Phys. Rev. Lett. **125**, 120504 (2020).

[13] M. Kjaergaard, M. E. Schwartz, A. Greene, G. O. Samach, A. Bengtsson, M. O'Keeffe, C. M. McNally, J. Braumüller, D. K. Kim, P. Krantz, *et al.*, Programming a quantum computer with quantum instructions, ArXiv:2001.08838 (2020).

[14] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, *et al.*, Realization of High-Fidelity CZ and *ZZ*-Free iSWAP Gates with a Tunable Coupler, Phys. Rev. X **11**, 021058 (2021).

[15] T. P. Harty, D. T. C. Allcock, C. J. Ballance, L. Guidoni, H. A. Janacek, N. M. Linke, D. N. Stacey, and D. M. Lucas, High-Fidelity Preparation, Gates, Memory, and Readout of a Trapped-Ion Quantum Bit, Phys. Rev. Lett. **113**, 220501 (2014).

[16] E. Jeffrey, D. Sank, J. Y. Mutus, T. C. White, J. Kelly, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, *et al.*, Fast Accurate State Measurement with Superconducting Qubits, Phys. Rev. Lett. **112**, 190504 (2014).

[17] C. C. Bultink, M. A. Rol, T. E. O'Brien, X. Fu, B. C. S. Dikken, C. Dickel, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, R. N. Schouten, and L. DiCarlo, Active Resonator Reset in the Nonlinear Dispersive Regime of Circuit QED, Phys. Rev. Appl. **6**, 034008 (2016).

[18] 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).

[19] S. Bravyi, D. Gosset, and R. König, Quantum advantage with shallow circuits, Science **362**, 308 (2018).

[20] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

[21] R. Babbush, J. R. McClean, M. Newman, C. Gidney, S. Boixo, and H. Neven, Focus beyond quadratic speedups for error-corrected quantum Advantage, PRX Quantum **2**, 010103 (2021).

[22] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, S. Boixo, M. Broughton, B. B. Buckley, D. A. Buell, *et al.*, Hartree-fock on a superconducting qubit quantum computer, Science **369**, 1084 (2020).

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

[24] D. Ristè, S. Poletto, M. Z. Huang, A. Bruno, V. Vesterinen, O. P. Saira, and L. DiCarlo, Detecting bit-flip errors in a logical qubit using stabilizer measurements, Nat. Commun. **6**, 6983 (2015).

[25] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, J. M. Chow, and J. M. Gambetta, Demonstration of Weight-Four Parity Measurements in the Surface Code Architecture, Phys. Rev. Lett. **117**, 210505 (2016).

[26] V. Negnevitsky, M. Marinelli, K. K. Mehta, H.-Y. Lo, C. Flühmann, and J. P. Home, Repeated multi-qubit readout and feedback with a mixed-species trapped-ion register, Nature **563**, 527 (2018).

[27] 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).

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

[29] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, N. Lacroix, G. J. Norris, M. Gabureac, C. Eichler, and A. Wallraff, Repeated quantum error detection in a surface code, Nat. Phys. **16**, 875 (2020).

[30] 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, ArXiv:2102.13071 (2021).

[31] Z. Chen, K. J. Satzinger, J. Atalaya, A. N. Korotkov, A. Dunsworth, D. Sank, C. Quintana, M. McEwen, R. Barends, P. V. Klimov, *et al.*, Exponential suppression of bit or phase flip errors with repetitive error correction, ArXiv:2102.06132 (2021).

[32] F. W. Strauch, P. R. Johnson, A. J. Dragt, C. J. Lobb, J. R. Anderson, and F. C. Wellstood, Quantum Logic Gates for Coupled Superconducting Phase Qubits, Phys. Rev. Lett. **91**, 167005 (2003).

[33] L. DiCarlo, J. M. Chow, J. M. Gambetta, L. S. Bishop, B. R. Johnson, D. I. Schuster, J. Majer, A. Blais, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, Demonstration of two-qubit algorithms with a superconducting quantum processor, Nature **460**, 240 (2009).

[34] J. M. Martinis and M. R. Geller, Fast adiabatic qubit gates using only σ_z control, Phys. Rev. A **90**, 022307 (2014).

[35] V. Tripathi, M. Khezri, and A. N. Korotkov, Operation and intrinsic error budget of a two-qubit cross-resonance gate, Phys. Rev. A **100**, 012301 (2019).

[36] A. P. Babu, J. Tuorila, and T. Ala-Nissila, State leakage during fast decay and control of a superconducting transmon qubit, npj Quantum Inf. **7**, 30 (2021).

[37] 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, ArXiv:2003.05952 (2020).

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

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

[40] J. Ghosh, A. G. Fowler, J. M. Martinis, and M. R. Geller, Understanding the effects of leakage in superconducting quantum-error-detection circuits, Phys. Rev. A **88**, 062329 (2013).

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

[42] M. Suchara, A. W. Cross, and J. M. Gambetta, Leakage suppression in the toric code, Quantum Inf. Comput. **15**, 997 (2015).

[43] N. C. Brown and K. R. Brown, Comparing Zeeman qubits to hyperfine qubits in the context of the surface code: $^{174}\mathrm{Yb}^{+}$ and $^{171}\mathrm{Yb}^{+}$, Phys. Rev. A **97**, 052301 (2018).

[44] N. C. Brown, M. Newman, and K. R. Brown, Handling leakage with subsystem codes, New J. Phys. **21**, 073055 (2019).

[45] N. C. Brown and K. R. Brown, Leakage mitigation for quantum error correction using a mixed qubit scheme, Phys. Rev. A **100**, 032325 (2019).

[46] 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).

[47] N. C. Brown, A. W. Cross, and K. R. Brown, *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)* (2020), p. 286.

[48] M. McEwen, D. Kafri, Z. Chen, J. Atalaya, K. J. Satzinger, C. Quintana, P. V. Klimov, D. Sank, C. Gidney, A. G. Fowler, *et al.*, Removing leakage-induced correlated errors in superconducting quantum error correction, Nat. Commun. **12**, 1761 (2021).

[49] 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).

[50] T. E. O'Brien, B. Tarasinski, and L. DiCarlo, Density-matrix simulation of small surface codes under current and projected experimental noise, npj Quantum Inf. **3**, 39 (2017).

[51] D. Hayes, D. Stack, B. Bjork, A. Potter, C. Baldwin, and R. Stutz, Eliminating Leakage Errors in Hyperfine Qubits, Phys. Rev. Lett. **124**, 170501 (2020).

[52] V. Langrock and D. P. DiVincenzo, A reset-if-leaked procedure for encoded spin qubits, ArXiv:2012.09517 (2020).

[53] 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).

[54] S. Zeytinoğlu, M. Pechal, S. Berger, A. A. Abdumalikov, A. Wallraff, and S. Filipp, Microwave-induced amplitude- and phase-tunable qubit-resonator coupling in circuit quantum electrodynamics, Phys. Rev. A **91**, 043846 (2015).

[55] 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).

[56] D. Ristè, C. C. Bultink, K. W. Lehnert, and L. DiCarlo, Feedback Control of a Solid-State Qubit Using High-Fidelity Projective Measurement, Phys. Rev. Lett. **109**, 240502 (2012).

[57] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, Scalable Quantum Circuit and Control for a Superconducting Surface Code, Phys. Rev. Appl. **8**, 034021 (2017).

[58] The *quantumsim* package can be found at https://quantumsim.gitlab.io/.

[59] T. M. Stace and S. D. Barrett, Error correction and degeneracy in surface codes suffering loss, Phys. Rev. A **81**, 022317 (2010).

[60] S. Nagayama, A. G. Fowler, D. Horsman, S. J. Devitt, and R. V. Meter, Surface code error correction on a defective lattice, New J. Phys. **19**, 023050 (2017).

[61] J. M. Auger, H. Anwar, M. Gimeno-Segovia, T. M. Stace, and D. E. Browne, Fault-tolerance thresholds for the surface code with fabrication errors, Phys. Rev. A **96**, 042316 (2017).

[62] J. R. Schrieffer and P. A. Wolff, Relation between the Anderson and Kondo Hamiltonians, Phys. Rev. **149**, 491 (1966).

[63] S. Bravyi, D. P. DiVincenzo, and D. Loss, Schrieffer-Wolff transformation for quantum many-body systems, Ann. Phys. **326**, 2793 (2011).

[64] E. Magesan and J. M. Gambetta, Effective hamiltonian models of the cross-resonance gate, Phys. Rev. A **101**, 052308 (2020).

[65] M. Boissonneault, J. M. Gambetta, and A. Blais, Dispersive regime of circuit QED: Photon-dependent qubit dephasing and relaxation rates, Phys. Rev. A **79**, 013819 (2009).

[66] B. Nijholt, J. Weston, J. Hoofwijk, and A. Akhmerov, *Adaptive:* Parallel active learning of mathematical functions (Zenodo, 2019).

[67] H. P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press, Oxford, 2002).

[68] S. Haroche and J. Raimond, *Exploring the Quantum: Atoms, Cavities, and Photons*, Oxford Graduate Texts (Oxford University Press, Oxford, 2006).

[69] https://doi.org/10.4121/14762052.

[70] C. J. Wood and J. M. Gambetta, Quantification and characterization of leakage errors, Phys. Rev. A **97**, 032306 (2018).

[71] S. Krinner, S. Lazar, A. Remm, C. Andersen, N. Lacroix, G. Norris, C. Hellings, M. Gabureac, C. Eichler, and A. Wallraff, Benchmarking Coherent Errors in Controlled-Phase Gates due to Spectator Qubits, Phys. Rev. Appl. **14**, 024042 (2020).

[72] T. E. O'Brien, B. M. Varbanov, and S. T. Spitz, *Qgarden* (Github, 2019), https://github.com/obriente/qgarden.

[73] J. B. Hertzberg, E. J. Zhang, S. Rosenblatt, E. Magesan, J. A. Smolin, J.-B. Yau, V. P. Adiga, M. Sandberg, M. Brink, J. M. Chow, and J. S. Orcutt, Laser-annealing Josephson junctions for yielding scaled-up superconducting quantum processors, ArXiv:2009.00781 (2020).

[74] F. Battistel, B. M. Varbanov, and B. M. Terhal, Data for: "Hardware-efficient leakage-reduction scheme for quantum error correction with superconducting transmon qubits," https://doi.org/10.4121/c.5320331 (2021).

[75] 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).