Model-Based Optimization of Superconducting Qubit Readout

Andreas Bengtsson,[1] Alex Opremcak,[1] Mostafa Khezri,[1] Daniel Sank,[1] Alexandre Bourassa,[1]
Kevin J. Satzinger,[1] Sabrina Hong,[1] Catherine Erickson,[1] Brian J. Lester,[1] Kevin C. Miao,[1]
Alexander N. Korotkov,[1,2] Julian Kelly,[1] Zijun Chen,[1] and Paul V. Klimov[1]
[1]*Google Quantum AI, Santa Barbara, 93111 California, USA*
[2]*Department of Electrical and Computer Engineering, University of California, Riverside, 92521 California, USA*

(Received 8 August 2023; accepted 2 February 2024; published 8 March 2024)

Measurement is an essential component of quantum algorithms, and for superconducting qubits it is often the most error prone. Here, we demonstrate model-based readout optimization achieving low measurement errors while avoiding detrimental side effects. For simultaneous and midcircuit measurements across 17 qubits, we observe 1.5% error per qubit with a 500 ns end-to-end duration and minimal excess reset error from residual resonator photons. We also suppress measurement-induced state transitions achieving a leakage rate limited by natural heating. This technique can scale to hundreds of qubits and be used to enhance the performance of error-correcting codes and near-term applications.

DOI: 10.1103/PhysRevLett.132.100603

Superconducting qubits have achieved measurement errors below 1% for single qubits [1–4] thanks to advancements including dispersive readout [5] and quantum-limited parametric amplifiers [6]. However, the increasing scale and complexity of algorithms bring a new set of challenges beyond simple single-qubit measurements. For example, quantum error correction requires measurements to be performed simultaneously, in the middle of circuits, and in a repetitive fashion. Midcircuit measurements must be fast to avoid decohering qubits not being measured, but faster readout can increase measurement and leakage errors. In recent error-correction experiments with superconducting qubits, readout-induced leakage severely limited performance [7,8], and in another, 20% of the total error (twice that of the measurement itself) was due to qubit idling during measurement [9].

Typically, readout is calibrated *in situ*, meaning that control parameters like pulse amplitude and frequency are varied until the observed measurement error is minimized. While effective at minimizing measurement error for isolated qubits, it can fail to capture other destructive processes like leakage or residual resonator photons. Additionally, effects such as qubit-qubit coupling impose nonlocality in that the optimal values for one qubit depend on the values of neighboring qubits. In turn, qubits optimized in isolation tend to perform poorly when measured simultaneously. Thus, optimizing multiqubit measurement requires searching a parameter space where the dimension scales linearly with the number of qubits, while attempting to minimize many metrics at once. This task rapidly becomes intractable by *in situ* parameter sweeps as the number of qubits grows.

In this Letter, we present an *ex situ* (model based) optimization technique for readout parameters. *Ex situ* optimization which allows us to explore a larger parameter space and minimize errors that are difficult or costly to measure, compared to *in situ* optimization where the speed is limited by the data rate of the quantum processor. In designing our model-based approach, we tackle three challenges which are often seen in quantum optimal control [10]. First, if the models do not accurately capture the present error channels, *ex situ* optimization is likely to perform worse than *in situ*. Second, the models must be evaluated quickly to be able to actually explore a larger space. These two challenges are conflicting in that more accurate models typically result in slower evaluation; for instance, a quantum simulation of the system dynamics would be too slow. Third, we must use an optimization algorithm that can find a reasonably good minimum without requiring a prohibitively long runtime.

We begin by describing representative models for readout error channels relevant to a Sycamore processor [11], which consists of superconducting frequency-tunable transmon qubits, each coupled to their own readout resonator. The models are quick to evaluate and we demonstrate that they accurately predict a variety of metrics over a wide range of parameters. We then use them together with the snake optimizer [12] to minimize the errors for 17 qubits in a distance-3 surface code. We achieve 1.5% measurement error per qubit in 500 ns (from the start of the readout until the system is ready for the next operation), while also reducing any additional errors like reset and leakage.

The models fall into two categories: predictive or heuristic. Ideally, we would only have predictive models (models that accurately predict error rates), but this is not always feasible as the computation might be too inaccurate or take too long. In those cases we use heuristic models to steer the optimizer away from parameter regions where

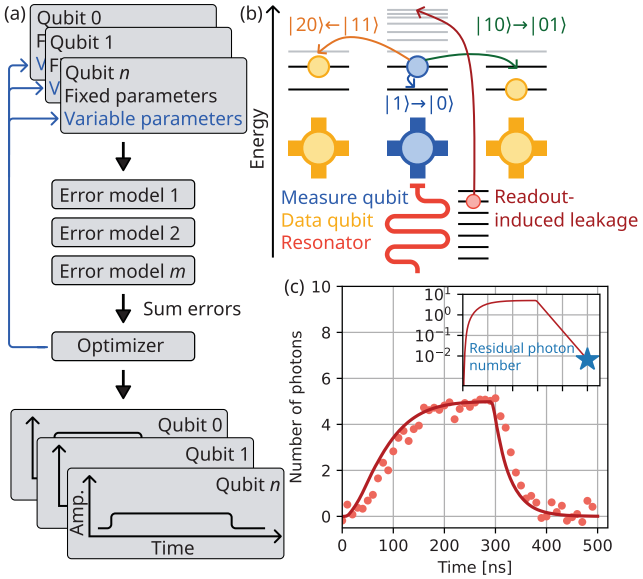

FIG. 1. (a) The optimization workflow. We build error models from fixed parameters like resonator frequency and linewidth, which we then run an optimizer on to find a set of variable parameters (e.g., pulse amplitude and length) that gives low errors. The output of the optimizer is a unique pulse shape for each qubit, as well as qubit frequency (not shown). (b) Examples of readout error mechanisms. The middle qubit, which is in $|1\rangle$, is coupled to two other qubits. During readout of the middle qubit its excitation can relax to $|0\rangle$; swap into $|1\rangle$ of the right qubit; or combine with the excitation in the left qubit to $|2\rangle$. Additionally, photons in the readout resonator can excite the qubit up to a high state. (c) Number of photons in the readout resonator as a function of time, both simulated (solid line) and measured (circles). The inset shows the simulated values on a logarithmic scale. Residual photons can cause reset and dephasing errors.

errors are large, but difficult to quantify. We use predictive models for the signal-to-noise ratio (SNR), qubit relaxation during readout, and residual resonator photons. Heuristic models include measurement-induced state transitions [13,14], and coupling to neighboring qubits. We sum all models into a single optimization cost function. The process is illustrated in Fig. 1(a).

Each model takes one or several input parameters describing the properties of the qubits, their readout resonators, or the measurement system itself. In total there are seven such parameters: (i) qubit anharmonicity $\alpha < 0$; (ii) resonator-qubit coupling $g(\omega_q)$; (iii) bare resonator frequency ω_r; (iv) measurement efficiency η [5]; (v) resonator linewidth κ; (vi) qubit relaxation rate as a function of frequency $\Gamma_1(\omega_q)$; and (vii) a calibrated reference for the readout pulse power at the processor.

We characterize these using a suite of metrology experiments [15]. Most of them are static and characterized just once. However, the qubit relaxation time is known to fluctuate [16] and is therefore remeasured just before optimization.

Additionally, there are four parameters which we can tune and optimize: (i) qubit frequency during readout ω_q; (ii) readout pulse amplitude B_0; (iii) readout pulse length t_p; and (iv) readout ringdown length t_r.

The ringdown length is needed for midcircuit measurements to allow the resonator to decay back to its ground state before other operations can resume. We choose to use a fixed total readout time ($t_p + t_r = 500$ ns), allowing us to synchronize gates, as well as reduce the number of optimization parameters to three.

A key parameter derived from the model inputs is the separation between resonator frequencies for the states $|0\rangle$ and $|1\rangle$, i.e., the dispersive shift $2\chi(\omega_q)$, given by [17],

$$\chi(\omega_q) = \frac{g(\omega_q)^2 \alpha}{(\omega_q - \omega_r)^2 [1 + \alpha/(\omega_q - \omega_r)]} \left(1 - \frac{\omega_q - \omega_r}{\omega_q}\right). \quad (1)$$

The shift is tunable since it depends on the qubit frequency; absent other constraints, the SNR per measurement photon is maximized when $2\chi = \kappa$.

A second derived parameter is the field in the resonator as a function of time $\beta(t)$, which is found by solving

$$\frac{d\beta}{dt} = \sqrt{\kappa} B(t) + (i\Delta - \kappa/2)\beta(t), \quad (2)$$

where $B(t)$ is the readout drive amplitude and has the dimension $\sqrt{\text{photons per time}}$, and Δ is the frequency difference between the drive and the dressed resonator. In the rest of this Letter we restrict the drive to be in the center of the resonator frequencies corresponding to $|0\rangle$ and $|1\rangle$, i.e., $\Delta = \pm\chi$, since that yields the highest SNR in the parameter regime we are interested in. For both states, we find the corresponding $\beta_{|0\rangle}(t)$ and $\beta_{|1\rangle}(t)$ by numerically solving Eq. (2).

The applied readout pulse, together with the noise (assumed to be Gaussian) in the readout chain, leads to a certain SNR, from which we derive the corresponding probability to misidentify the state. Given $\delta\beta(t) = \beta_{|0\rangle}(t) - \beta_{|1\rangle}(t)$, the SNR is calculated as

$$\mathrm{SNR} = 2\eta\kappa \frac{|\int_0^{t_p+t_r} \delta\beta(t) w(t) dt|^2}{\int_0^{t_p+t_r} |w(t)|^2}, \quad (3)$$

where $w(t)$ is an integration window function, which we set to $\delta\beta(t)^*$ to maximize SNR. From SNR we derive the corresponding error,

$$\epsilon_{\text{separation}} = \frac{1}{2} \mathrm{erfc}\left(\frac{\sqrt{\mathrm{SNR}}}{2}\right). \quad (4)$$

During readout the qubit might decay and potentially cause a measurement error. To calculate the error rate we need the qubit frequency during readout, which is changing throughout the process due to the ac-Stark effect, and the corresponding relaxation rates at those frequencies. The latter is measured by a standard relaxation experiment versus qubit frequency; while the former can be found via Eqs. (1) and (2),

$$\omega_q(t) = \omega_q(0) + 2|\beta_{|1\rangle}(t)|^2 \chi[\omega_q(0)], \tag{5}$$

where we have assumed that the ac-Stark shift is strictly linear. Given $\Gamma_1(\omega_q)$, we calculate the relaxation error as

$$\epsilon_{\mathrm{relaxation}} = \int_0^{T_0} \Gamma_1[\omega_q(t)] \mathrm{d}t. \tag{6}$$

We approximate T_0 to be the point where SNR is half of its maximum, since a relaxation event beyond that point should not change the measurement outcome. We also choose to ignore the upward transition rate $|0\rangle \rightarrow |1\rangle$, since the timescale for that process is much longer than for relaxation and the readout.

The resonator must be mostly depleted of photons before the next operation can begin, since any remaining photons cause qubit dephasing. Additionally, in our architecture such photons directly translate into reset errors since it is based on the swap interaction between the qubit and its resonator [18], and a photon in the resonator could be swapped into either $|1\rangle$ or $|2\rangle$, depending on the state. We use the mean photon number in the resonator at the end of the readout,

$$\epsilon_{\mathrm{photon}} = \frac{|\beta_{|0\rangle}(T)|^2 + |\beta_{|1\rangle}(T)|^2}{2}, \tag{7}$$

as the error model to minimize both the reset error and qubit dephasing.

Shown in Fig. 1(c) are the expected photon number $|\beta_{|0\rangle}(t)|^2$ and the corresponding measured values (via spectroscopic measurements of the ac-Stark shift [14]). The measurement technique is not sensitive to the small frequency shifts occurring at the low photon numbers toward the end of readout; however, since the agreement is good during the pulse itself we can use the model to infer what the final photon number is, which in this example is 0.005.

In Fig. 2, we show predicted and measured values for $\epsilon_{\mathrm{separation}}$ and $\epsilon_{\mathrm{relaxation}}$ as a function of qubit frequency, pulse amplitude, and pulse length. We additionally show the predicted residual photon number, though we do not have a sensitive enough technique to reliably measure this quantity at the modeled levels. Overall, we see good agreement between measured and simulated values, with the exception for $\epsilon_{\mathrm{relaxation}}$ at low amplitudes and lengths where $\epsilon_{\mathrm{separation}}$ is large. That parameter regime should be avoided and accurately predicting $\epsilon_{\mathrm{relaxation}}$ there is less important. The cause of the discrepancy could be due to the approximations in Eq. (6) or from the inaccuracy in trying to extract a small error on top of a large error.

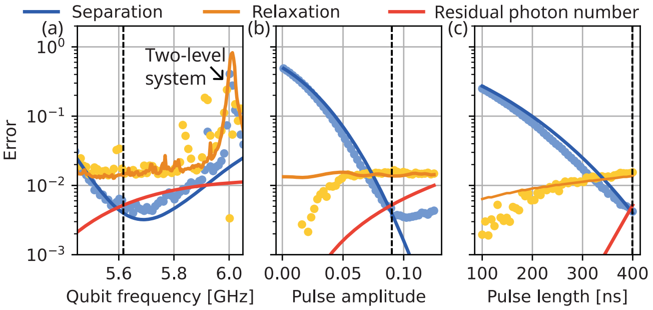

FIG. 2. Error models and their dependence on readout parameters. (a) Separation and relaxation errors, and residual photon number, versus the qubit frequency with amplitude and length kept fixed. Circles are measured data, and lines are simulated values. As the qubit frequency changes, we track the readout pulse frequency to be centered between the two dressed resonator states. The peak at 6 GHz is due to a two-level system defect. (b) and (c) show the same models, but versus pulse amplitude and length, respectively. The total readout time is kept constant, such that when the pulse length increases, the ringdown time decreases. The nonswept parameters are kept fixed at the values indicated by the dashed vertical lines in the respective panels.

We can understand the trade-offs in readout optimization by studying the predictive models. If we only consider these three models, the ideal pulse would be short and with high amplitude, since the SNR approximately scales quadratically with the amplitude [Fig. 2(b)] and linearly with the pulse length [Fig. 2(c)], while $\epsilon_{\mathrm{photon}}$ also scales quadratically with the amplitude, but exponentially with the pulse length (for a fixed total time). Additionally, a short readout pulse minimizes $\epsilon_{\mathrm{relaxation}}$.

However, a high pulse amplitude can be problematic for several reasons. For example, it leads to measurement-induced state transitions [13,14], which occurs when resonator photons are transferred to the qubit and excite it far beyond the computational subspace, as illustrated in Fig. 1(b). While this high state may lead to a measurement error, it is more importantly immune to our reset protocol [18], making it particularly destructive for mid-circuit measurements and quantum error correction [7,8]. Using the model in Ref. [14], valid only for $\omega_q > \omega_r$, we define a heuristic that constrains the maximum photon number in the resonator,

$$\max(|\beta(t)|^2) < a e^{b(\omega_q - \omega_r)} - \sqrt{a e^{b(\omega_q - \omega_r)}}, \tag{8}$$

where a and b are extracted from numerical simulations and only dependent on α and g [14].

Finally, we introduce a model related to the coupling between qubits. Our qubits are laid out on a square grid where a pair of qubits have four relevant coupling channels, $|01\rangle \leftrightarrow |10\rangle$, $|11\rangle \leftrightarrow |20\rangle$, $|11\rangle \leftrightarrow |02\rangle$, and $|12\rangle \leftrightarrow |21\rangle$.

We heuristically model the errors associated with these channels using a sum of Lorentizians,

$$\epsilon_{\text{coupling}} = \sum_i c_i \frac{\gamma_i}{2} \frac{\pi}{(\omega_q - \omega_i)^2 + \gamma_i^2/4}, \qquad (9)$$

where c_i, γ_i, and ω_i are the amplitude, width, and center frequency of each transition. We use a heuristic to avoid having to model the time dependence of the qubit frequency and its effect on the measurement errors. That dependence is complicated by the ac-Stark effect, which imposes both a frequency shift due to mean number of photons in the resonator [Eq. (5)], as well as frequency broadening due to photon number fluctuations. By using wide and large enough Lorentzians the optimizer avoids any qubit-qubit interactions. Assuming couplings between nearest and next-nearest neighbors there are up to 32 frequency collisions for each qubit.

We now continue to the actual optimization. While any global optimizer can be used, we choose to employ the snake optimizer [12], which has successfully optimized single and two-qubit gate parameters for a variety of quantum algorithms [9,11,19]. More optimization details are found in Supplemental Material, Ref. [20]. As our experimental platform we use 17 qubits in a distance-3 surface code layout, illustrated in Fig. 3(a). The optimization takes 1 min and includes 1.7×10^6 evaluations of the cost function. Afterward, the resulting parameters are uploaded to the control system, and the only remaining calibration is to find the discrimination line to distinguish between $|0\rangle$ and $|1\rangle$ for each qubit. We choose to not model this since we can efficiently measure it simultaneously across all qubits, and it does not conflict with the other parameter choices.

We compare three different optimization strategies to evaluate the performance of our model-based approach. The first strategy is *in situ* optimization where we choose a fixed pulse length (300 ns) and perform a sequence of 1D sweeps to find the optimal pulse frequency, amplitude, qubit frequency, and integration window. The second strategy is *ex situ* optimization using a partial cost function consisting of only the predictive models, i.e., no qubit-qubit coupling or measurement-induced state transitions models. The third strategy is *ex situ* optimization using a complete cost function consisting of all available models. For each strategy, we quantify three important aspects: measurement errors, reset errors, and leakage. Note that we do not benchmark the performance of the optimizer itself, e.g., how well it finds the actual global minimum. The performance aspects of the snake have been recently studied in Ref. [22].

We benchmark measurement errors by preparing 200 random initial states over all qubits and then sampling 2000 measurement outcomes for each initial state. We then repeat the procedure, but this time using only the measure qubits to mimic the surface code midcircuit measurements. We compare all outcomes with the known initial states and extract the errors, seen in Fig. 3(b), and calculate the measurement error as $[P(1|0) + P(0|1)]/2$, where $P(f|i)$ is the probability of preparing $|i\rangle$ and measuring $|f\rangle$. Note that state preparation errors will show up as measurement errors in this protocol. The complete *ex situ* optimizer achieves an average measurement error of 1.5% per qubit, while *in situ* and partial *ex situ* optimization achieve 1.9% and 4.7%, respectively. Overall, *in situ* and complete *ex situ* optimization have similar performance with the exception of a few high-error outliers for the *in situ* optimizer. For instance, the largest outlier is caused by $|11\rangle \leftrightarrow |02\rangle$ swapping between two neighboring qubits, which the *ex situ* optimizer is able to avoid [20]. Out of the 1.5% error per qubit, we are able to account for 1.2% when we include the contributions from state preparation, separation error, and relaxation error [20]. We estimate that the state preparation error is 0.4%, which, if accurate, should be subtracted from the values above.

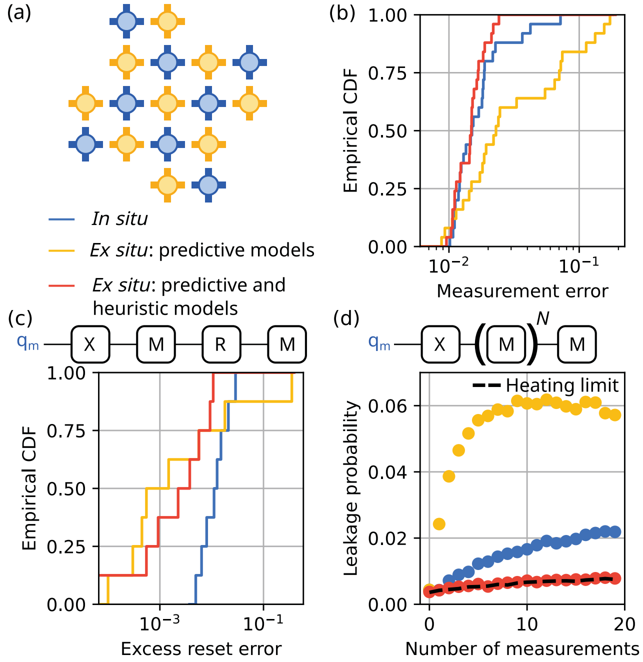

FIG. 3. Benchmarking of the optimized readout performance. We compare three optimization strategies: *in situ*; *ex situ* with only predictive models; *ex situ* with predictive and heuristic models. (a) The distance-3 surface code with 9 data qubits (yellow) and 8 measure qubits (blue), used for the benchmarking. (b) Simultaneous measurement errors for two cases: all qubits, only measure qubits. We prepare a set of random states across the qubits and perform simultaneous measurements. The data show the combination of the two cases. (c) Reset error added by a preceding measurement, benchmarked on the measure qubits. The excess reset error is caused by residual photons in the readout resonator. (d) Average leakage probability in the measure qubits after preparing $|1\rangle$ and performing N measurements. The dashed line shows the heating limit where the measurements are replaced by an equivalent amount of waiting time.

However, since state preparation errors should affect all three strategies the same we have chosen to be conservative and not do the subtraction.

Next, we benchmark reset errors added by readout for the measure qubits only (data qubits do not need reset). We prepare $|1\rangle$, perform measurements immediately followed by reset and another round of measurements. In the case of no errors we expect the second measurement to yield $|0\rangle$. We also perform the same sequence but without the first round of measurements and subtract that to remove the intrinsic reset and measurement errors. The results are shown in Fig. 3(c). Complete *ex situ* optimization adds on average an additional 0.4% of reset errors, compared to 4.7% and 1.4% for partial and *in situ*, respectively.

For the final benchmark, we quantify qubit leakage. Again, we focus on the measure qubits and prepare $|1\rangle$ as that makes the qubits more likely to leak, and then perform a variable number of measurements. We append a final and different measurement that is able to discriminate if the qubit has left the computational subspace. Figure 3(d) shows the probability of leakage as a function of the number of measurements. Complete *ex situ* optimization suppresses leakage down to an average of 0.8% after 20 measurement rounds, comparable to the heating limit as measured by repeating the experiment with no readout pulses but an equivalent amount of waiting time. After the same number of rounds, the partial and *in situ* strategies have leakage populations of 5.7% and 2.2%, respectively.

Comparing the optimization strategies, we see that complete *ex situ* using both predictive and heuristic models outperforms the others in all three benchmarks. It is able to achieve lower measurement errors, while also adding less reset and leakage errors for midcircuit measurements. The partial *ex situ* optimizer generally performs worse than *in situ* optimization. This is likely due to the lack of an amplitude limiting model, which tends to drive the optimizer toward short and high-amplitude pulses, which in turn leads to state transitions. This emphasizes that for model-based optimization to work well, the models have to account for all dominant error mechanisms, even if only as heuristics.

In conclusion, we demonstrated model-based optimization for superconducting qubit readout achieving low measurement errors (1.5%) for both midcircuit and terminal measurements. For midcircuit measurements, we also observed suppressed reset errors (0.4%) and no increase in leakage due to readout. We accomplished this by overcoming the challenges stated in the introduction: the presented models accurately capture the relevant error channels, and they can be evaluated 10 000 times faster (1 min vs 1 week for the parameter space used here) than measuring errors directly in hardware, which unlocks the ability to use a global optimizer. Based on recent work in Ref. [22] we believe the snake optimizer and these models will scale to at least 1000 qubits.

Our model-based readout optimization strategy has already been employed in several large experiments, such as the demonstration of a distance-5 surface code [9] with a measurement error of 1.9% per qubit, and a 70 qubit random-circuit sampling experiment [19] with an error of 1.3% per qubit. While the performance is among the best observed for repetitive and simultaneous measurements in superconducting qubits, even better performance will be needed to be well below the error-correcting threshold. In particular, the readout time has to be shorter to avoid data-qubit idling errors.

We believe that the error rates achieved in this Letter are close to optimal for the given processor, and that the path to more performant readout is through longer relaxation times, higher measurement efficiencies, and more optimized circuit parameters. While we treated the circuit parameters as fixed, we could include them as optimization parameters to inform the design of future processors. However, more research is needed to find the optimal readout circuit for superconducting qubits.

We thank the broader Google Quantum AI team for fabricating the processor, building and maintaining the cryogenic system, and general hardware and software infrastructure that enabled this experiment. We also thank V. Sivak and W. Livingston for providing comments on the manuscript, and L. Martin for helping with the crosstalk analysis.

[1] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Pechal, M. Mondal, M. Oppliger, C. Eichler *et al.*, Rapid high-fidelity single-shot dispersive readout of superconducting qubits, Phys. Rev. Appl. **7**, 054020 (2017).

[2] Y. Sunada, S. Kono, J. Ilves, S. Tamate, T. Sugiyama, Y. Tabuchi, and Y. Nakamura, Fast readout and reset of a superconducting qubit coupled to a resonator with an intrinsic Purcell filter, Phys. Rev. Appl. **17**, 044016 (2022).

[3] L. Chen, H.-X. Li, Y. Lu, C. W. Warren, C. J. Križan, S. Kosen, M. Rommel, S. Ahmed, A. Osman, J. Biznárová *et al.*, Transmon qubit readout fidelity at the threshold for quantum error correction without a quantum-limited amplifier, npj Quantum Inf. **9**, 26 (2023).

[4] F. Swiadek, R. Shillito, P. Magnard, A. Remm, C. Hellings, N. Lacroix, Q. Ficheux, D. C. Zanuz, G. J. Norris, A. Blais *et al.*, Enhancing dispersive readout of superconducting qubits through dynamic control of the dispersive shift: Experiment and theory, arXiv:2307.07765.

[5] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, Circuit quantum electrodynamics, Rev. Mod. Phys. **93**, 025005 (2021).

[6] J. Aumentado, Superconducting parametric amplifiers: The state of the art in Josephson parametric amplifiers, IEEE Microw. Mag. **21**, 45 (2020).

[7] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and

M. Takita, Demonstrating multi-round subsystem quantum error correction using matching and maximum likelihood decoders, Nat. Commun. **14**, 2852 (2023).

[8] J. Marques, H. Ali, B. Varbanov, M. Finkel, H. Veen, S. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider *et al.*, All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits, Phys. Rev. Lett. **130**, 250602 (2023).

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

[10] C. P. Koch, U. Boscain, T. Calarco, G. Dirr, S. Filipp, S. J. Glaser, R. Kosloff, S. Montangero, T. Schulte-Herbrüggen, D. Sugny *et al.*, Quantum optimal control in quantum technologies. Strategic report on current status, visions and goals for research in Europe, Eur. Phys. J. Quantum Technol. **9**, 19 (2022).

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

[12] P. V. Klimov, J. Kelly, J. M. Martinis, and H. Neven, The snake optimizer for learning quantum processor control parameters, arXiv:2006.04594.

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

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

[15] D. Sank, A. Opremcak, A. Bengtsson, M. Khezri, Z. Chen, O. Naaman, and A. Korotkov, System characterization of dispersive readout in superconducting qubits, arXiv:2402.00413.

[16] J. J. Burnett, A. Bengtsson, M. Scigliuzzo, D. Niepce, M. Kudra, P. Delsing, and J. Bylander, Decoherence benchmarking of superconducting qubits, npj Quantum Inf. **5**, 54 (2019).

[17] M. Khezri, *Dispersive Measurement of Superconducting Qubits* (University of California, Riverside, 2018).

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

[19] A. Morvan, B. Villalonga, X. Mi, S. Mandrà, A. Bengtsson, P. Klimov, Z. Chen, S. Hong, C. Erickson, I. Drozdov *et al.*, Phase transition in random circuit sampling, arXiv:2304.11119.

[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.132.100603, which includes Ref. [21], for more details and benchmarking results on a per qubit basis.

[21] 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).

[22] P. V. Klimov, A. Bengtsson, C. Quintana, A. Bourassa, S. Hong, A. Dunsworth, K. J. Satzinger, W. P. Livingston, V. Sivak, M. Y. Niu *et al.*, Optimizing quantum gates towards the scale of logical qubits, arXiv:2308.02321.