Quantum information processors suffer from a wide variety of errors that must be quantified if their performance is to be understood and improved. A processor’s errors are commonly probed using randomized benchmarks that involve running random circuits [1–21]—e.g., standard randomized benchmarking (RB) [3, 4] or one of its many variants [3–17] cross-entropy benchmarking [18], or the quantum volume benchmark [21]. Randomized benchmarks are appealing because they aggregate many kinds of error into one number that quantifies average performance over a large circuit ensemble. Unlike tomographic techniques [22] that estimate a set of parameters that may be exponentially large in the number of qubits (n), randomized benchmarks hold the potential for scalable performance assessment.

Yet current randomized benchmarks have one of two scaling problems. Quantum volume and cross-entropy benchmarking require classical computations that are exponentially expensive in n, becoming infeasible beyond n ≈ 50 [18–21]. In contrast, standard RB requires only efficient classical computations but it benchmarks composite gates from the n-qubit Clifford group. They require $O(\pi/\log n)$ two-qubit gates to implement [23–25], so the fidelity of a typical n-qubit Clifford decreases quickly with n. Lower compilation overheads [e.g., $O(\log n)$ are possible with access to many-qubit gates [26], but in realistic architectures the circuit depth required to implement a typical Clifford will increase with the number of qubits. As a result, standard RB has only been implemented on up to three qubits [27], and even its streamlined variant “direct RB” (DRB) has only been implemented on up to 5 qubits [17].

In this Letter we introduce a simple, flexible, and robust RB method that removes the Clifford compilation bottleneck that limits current methods. We show how randomized mirror circuits (Fig. 1a) enable scalable RB of Clifford gates. This work advances circuit mirroring [28], a recently introduced method for scalable benchmarking of quantum computers. Ref. [28] shows how mirror circuits can be used to map out how a quantum computer’s performance on circuits depends on their widths and depths (volumetric benchmarking [29]), but it doesn’t show how to quantify gate fidelity. Here, we show how to use randomized mirror circuits to estimate the infidelity of an average Pauli-dressed [30–33] n-qubit circuit layer (Fig. 1a, grey boxes), and we present a theory that proves that this method—mirror RB (MRB)—is reliable. MRB can be applied whenever a typical n-qubit circuit layer has significantly non-zero fidelity, enabling RB of hundreds or even thousands of qubits with physically realistic error rates [$O(10^{-2})$–$O(10^{-3})$]. We demonstrate and validate MRB on up to 225 qubits using simulations (Fig. 2), and on up to 16 physical qubits using IBM Q’s cloud quantum computing platform (Figs. 1, 3 and 4).

Randomized mirror circuits. MRB uses randomized mirror circuits [28], shown in Fig. 1a. By design, each random-
ized mirror circuit $C$ should ideally always produce a single bit string $s_C$ that is efficient to compute. Distributions over these circuits are parameterized by an $n$-qubit layer set $L = \{L\}$, a probability distribution $\Omega$ over $L$, and a benchmark depth $d$ that specifies the number of Pauli-dressed layers in the circuit. Both $L$ and $\Omega$ are customizable, but we require that (1) each layer contains only Clifford gates, (2) each layer’s inverse $L^{-1}$ is also within $L$, (3) $\Omega(L) = \Omega(L^{-1})$, and (4) $\Omega$-random layers quickly locally randomize an error (local “twirling”) and spread it across multiple qubits. Condition (4) is also required for reliable DRB, and the circumstances under which it is satisfied have been studied in detail [17].

For all demonstrations herein, the layer set consists of parallel applications of CNOTs between connected qubits and all 24 single-qubit Clifford gates. This enables transparent quantification of the errors caused by native two-qubit gates, including crosstalk. Note, however, that our method can be applied, e.g., CNOTs synthesized via SWAP chains, enabling comparisons between the errors in identical layer sets on different devices. All our distributions $\Omega$ have a similar form whereby sampling a layer consists of: (1) sampling some CNOTs, and (2) sampling uniformly random single-qubit Clifford gates for all qubits not acted on by those CNOTs.

**Mirror RB.** MRB aims to measure $\epsilon_{MRB} = \sum_i \epsilon_i(L)\epsilon(L_i)$, where $\Omega$ is a user-chosen distribution over $L$, and $\epsilon(L)$ is the entanglement infidelity of the Pauli-dressed version of the $n$-qubit layer $L$ (gray boxes, Fig. 1a). In all our demonstrations we do not compile the Paulis into the $L$ layers, but this is permissible. MRB estimates $\epsilon_{MRB}$ using data from $\Omega$-sampled randomized mirror circuits. For each circuit $C$ that we run, we estimate its effective polarization

$$S = \frac{4^n}{4^n - 1} \left[ \sum_{k=0}^{n} \left( \frac{1}{2} \right)^k h_k \right] - \frac{1}{4^n - 1},$$

where $h_k$ is the probability that the circuit outputs a bit string that is a Hamming distance of $k$ from its target bit string ($s_C$). As our theory (below) shows, the simple additional analysis in computing $S$ mitigates the limited “twirling” enacted by our circuits.

MRB is the following protocol:

1. For a range of integers $d \geq 0$, sample $K$ randomized mirror circuits of benchmark depth $d$ where $d$ is even (see Fig. 1a), using the distribution $\Omega$, and run each one $N \geq 1$ times.

2. Estimate each circuit’s effective polarization $S$.

3. Fit $S_d$, the mean of $S$ at benchmark depth $d$, to $S_d = Ap^d$, where $A$ and $p$ are fit parameters, and then compute $r_{MRB} = (4^n - 1)^{1/p}/4^n$ as an estimate of $\epsilon_{MRB}$.

**Theory.** We now show that MRB is reliable, i.e., $S_d \approx Ap^d$ and $r_{MRB} \approx \epsilon_{MRB}$ under broad conditions. We assume that errors are Markovian [22] but not necessarily gate-independent (many, but not all, non-Markovian errors appear Markovian within random circuits [35–37]). We use $U(L)$ and $\phi(L)$ to denote the $n$-qubit superoperators that represent a layer $L$’s perfect and imperfect implementations, respectively, and $\epsilon(L)$ its error map, i.e., $\phi(L) = \epsilon(L) U(L)$. Our theory starts from a single randomized mirror circuit $C$ of benchmark depth $d$. So $C = F_0^{-1}P_d L_1^{-1} \cdots P_{1+\delta} L_1^{-1} P_{1+\delta_1}L_2^{-1} \cdots \cdots P_{1+\delta} L_n^{-1} P_{1+\delta_1}$, where $(1) P_i$ are Pauli layers, $(2) F_0$ and $F_0^{-1}$ consist of one-qubit Clifford gates, and $(3) L_i$ are $\Omega$-sampled layers and $L_i^{-1}$ their inverses. The components (1), (2) and (3) are sampled independently.

To compute $S_d$, as a function of $\epsilon(L)$, we can therefore average over (1-3) separately in turn.

The Pauli layers (green boxes, Fig. 1a) are independent, uniformly random, and interleaved between every other layer. They therefore have two effects: they randomize the target bit string ($s$), which guarantees that $S_d \rightarrow 0$ as $d \rightarrow \infty$ to a good approximation [38], and they twirl the errors on the $L_i$ layers into stochastic Pauli errors [30–33]. So we can analyze the “residual” circuit $C = F_0^{-1}L_1^{-1} \cdots L_n^{-1}L_d^{-1} \cdots L_1 F_0$ with each $L_i$’s error map $\epsilon(L)$ a stochastic Pauli channel. The composite superoperator for this circuit is $\phi(C) = \phi(F_0^{-1})\epsilon_{\Omega}(\phi(F_0))$ where $\epsilon_{\Omega} = \epsilon(L_1^{-1}) \cdots \epsilon(L_d^{-1})\epsilon(L_d) \cdots \epsilon(L_1)$ is a stochastic Pauli channel, as each $U(L_i)$ is a Clifford operator.

The initial layer (blue boxes, Fig. 1a) $F_0$ contains independent, uniformly random single-qubit Clifford gates. Averaging over this implements local 2-design twirling on each qubit [39]. That is, $\bar{\epsilon}_{\Omega} = \frac{1}{F_0} \sum_p U(F_0)^{-1}\epsilon_{\Omega}(U(F_0))$ is a stochastic Pauli channel with equal marginal probabilities to induce an $X$, $Y$ or $Z$ error on any fixed qubit. An error induced by $\bar{\epsilon}_{\Omega}$ flips at least one output bit iff it applies $X$ or $Y$ to at least one qubit. So, if $\bar{\epsilon}_{\Omega}$ induces a weight $k$ error (an error on $k$ qubits) the circuit outputs $s_C$ with a probability of $1/2^k$. Generally, a weight $k$ error causes flips on $j$ of the output bits with probability $M_{jk} = \binom{n}{j} \frac{1}{2^j}$. So $\bar{H} = \bar{M} \bar{p}$ where $\bar{h}_k$ and $p_k$ are the probabilities that $k$ bits are flipped and that $\bar{\epsilon}_{\Omega}$ induces a weight $k$ error, respectively, with $k = 0, \ldots, n$. By inverting $\bar{M}$, we obtain $p_0 = \sum_{k=0}^{n} (-\frac{1}{2})^k \bar{h}_k \equiv H$. Because $p_0 = 1 - \epsilon(\bar{\epsilon}_{\Omega})$ where $\epsilon(\bar{\epsilon}_{\Omega})$ is $\bar{\epsilon}_{\Omega}$’s entanglement infidelity, $H$...
therefore equals $E_d$’s entanglement fidelity, and $S$ [Eq. (1)] its polarization $γ(E_d) := 1 - 4^d η(χ(E_d)) / (4^d - 1)$. State preparation and measurement (SPAM) errors also contribute to $S$ (and $H$), as do errors in $F_0$ and $F_0^{-1}$. But their effect is approximately $d$-independent, so $S \approx Aγ(E_d)$ for some $A$.

We have related a randomized mirror circuit’s $S$ to the polarization of its superoperator $[γ(E_d)]$. Now we relate $γ(E_d)$ to the polarizations of the circuit’s constituent layers $[γ(E(L_i))]$. If every $E(L_i)$ is an $n$-qubit depolarizing channel, with layer-dependent error rates, then $γ(E_d) = \prod_{i=1}^d γ_{\gamma_i}$ where $γ_i \equiv γ(E(L_i))$. More generally we argue that $γ(E_d) \approx \prod_{i=1}^d γ_{\gamma_i}$. For two stochastic Pauli channels $E_A$ and $E_B$, $γ(E_AE_B) = γ(E_A)γ(E_B) + η$ where $η = \sum_j (ε(A,j) - 1/4^d)(ε(B,j) - 1/4^d)$ and $ε_i$ is the vector of $4^d - 1$ Pauli error probabilities for $E_i$. $η$ quantifies the rate that errors cancel when composing the two channels, relative to the rate that they cancel when composing $n$-qubit depolarizing channels. It is negligible unless $ε_A$ and $ε_B$ are sparse (e.g., if $ε_A \equiv ε_0$ and the error probability is equally distributed over $K$ errors, then $η = ε(ε_0)^2 / (K - 1/4^d)$).

We simulated MRB on 1-225 qubits with randomly sampled stochastic Pauli error models. The qubits were arranged on a $15 \times 15$ lattice (the layer set is described above). We independently sampled a total of 900 MRB circuit sets with a range of $n \in [1,225]$. We used a distribution $Ω$ whereby a layer sampled from $Ω$ has an expected CNOT density of $1/8$. For each MRB circuit set we used a different randomly sampled error model, consisting of biased and correlated Pauli errors with one- and two-qubit gates having an expected infidelity of 0.1% and 1%, respectively [40]. Fig. 2a shows $ε_0$ versus $r_0$. We observe that $r_0 \approx ε_0$, with $r_0$ typically only slightly underestimates $ε_0$. This is supported by our simulations and our demonstrations on physical qubits.

Simulations. We simulated MRB on 1-225 qubits with randomly sampled stochastic Pauli error models. The qubits were arranged on a $15 \times 15$ lattice (the layer set is described above). We independently sampled a total of 900 MRB circuit sets with a range of $n \in [1,225]$. We used a distribution $Ω$ whereby a layer sampled from $Ω$ has an expected CNOT density of $1/8$. For each MRB circuit set we used a different randomly sampled error model, consisting of biased and correlated Pauli errors with one- and two-qubit gates having an expected infidelity of 0.1% and 1%, respectively [40]. Fig. 2a shows $ε_0$ versus $r_0$. We observe that $r_0 \approx ε_0$, with $r_0$ typically only slightly underestimates $ε_0$. This is supported by our simulations and our demonstrations on physical qubits.

Validating MRB with cloud access experiments. To dem-

Figure 3. Validating MRB using cloud access experiments. MRB, DRB and standard RB on 1-5 qubits of IBM Q Quito. (a-c) The means (points) and distributions (violin plots) of the circuit polarizations versus benchmark depth ($d$), and fits to an exponential $Ap^d$ (curves). (d) Error rates ($r$) obtained from the fit’s decay rate for DRB and MRB versus the number of qubits ($n$), and the values predicted from calibration data. The DRB and MRB error rates are in close agreement, validating MRB against the reliable but unscalable DRB protocol. The measured $r$ diverges from the predictions of Quito’s calibration data as $n$ increases, indicating crosstalk. (e) The mean polarizations at $d = 0$ ($S_0$) decrease rapidly with $n$ for DRB and standard RB [at best log($S_0$) $= 1 - O(n^2 / \log n)$] making them infeasible beyond a few qubits, whereas log($S_0$) $= 1 - O(n)$ for MRB. $S_0$ satisfies 0.003 $> \overline{δ_{rel}} > -0.16$. Although this systematic underestimation of $ε_0$ is undesirable, it is arguably small enough to be insignificant (RB is typically used for rough estimates of gate performance rather than precision characterization).

To show how MRB can be used to reveal crosstalk errors, we simulated it on our hypothetical 225-qubit processor with two illustrative models, one with and one without crosstalk. The crosstalk-free model consisted of 0.5% readout error on each qubit, and depolarization on the one- and two-qubit gates, with 0.1% and 1% error rates, respectively. In the crosstalk model, each CNOT also caused the error probability for qubit $q$ to increase by $ε(q)$, with $ε(q)$ a slowly decreasing function of the distance (on the lattice) from $q$ to the CNOT’s location [40]. Fig. 2b shows $r_0$ (points) and $ε_0$ (dotted line) versus $n$ for both models. We find that $r_0 \approx ε_0$ (averaged over $n$, $δ_{rel} \approx -0.17$ and $δ_{rel} \approx -0.08$ for the crosstalk-free and crosstalk models, respectively), and that $r_0$ grows quadratically at low $n$ under the crosstalk model—an effect that cannot be observed without running many-qubit circuits.
Mapping out a processor’s performance. MRB can be used to map out performance of a processor’s $n$-qubit layers when varying both $n$ and the embedding of those qubits, as we demonstrate on IBM Q Rueschlikon (16 qubits) [41]. For $n \in \{1, 2, 4, 8, 16\}$ we divided Rueschlikon into $16/e$ regions, and ran randomized mirror circuits on each region (the one-qubit circuits were performed simultaneously to match IBM’s calibration experiments) [40]. In this demonstration, we fixed the expected number of CNOTs in a layer to $1/2$. Fig. 1b shows exponential decays for one region of each size (the leftmost regions in Fig. 4), and Figs. 4a and 4b show $r_\Omega$ for all benchmarked regions and the predictions from the calibration data, respectively. The prediction underestimates $r_\Omega$ for $n > 2$, again signifying crossstalk induced by CNOTs (see discussion above).

Discussion. In this Letter we have introduced a technique that enables holistic RB of hundreds or thousands of qubits, while retaining the core simplicity of standard RB—fitting data from random circuits to an exponential. We anticipate that techniques based on standard RB [10, 14, 39, 42–51] can be enhanced using ideas introduced here. For example, MRB does not require compilation of subroutines so it removes the circuit scheduling complexities that plague simultaneous standard RB [14, 39], suggesting that MRB will be more powerful for probing crossstalk. Similarly, running multiple MRB experiments with $\Omega$ varied could be used to isolate the error rates of different subsets of layers [17]. This would enable reliable predictions of the performance of many-qubit, randomly-compiled circuits [30–33] (randomized compiling guarantees that layer fidelities are sufficient to predict overall circuit performance [35], which is not true otherwise [28]).

Our demonstrations on a cloud quantum computing platform revealed and quantified crossstalk errors that are invisible to one- and two-qubit RB, highlighting the need for scalable methods like ours. Outside the paradigm of RB there are a variety of methods for testing $n$-qubit circuit layers, and our technique complements them. For example, cycle benchmarking [33, 52] and Pauli noise estimation [53, 54] can characterize a Pauli-dressed $n$-qubit layer. These techniques extract more information about a layer’s errors, but, unlike MRB, they test only one (or a few) of a processor’s many possible $n$-qubit layers. Methods for extracting more information from mirror circuit data, e.g., by using the techniques of Refs. [52–54], are an intriguing possibility [55, 56].

Our method is built on a particular type of randomized mirror circuits, but circuit mirroring [28] is a flexible tool that could be used to construct a range of randomized benchmarks with complementary properties to ours. For example, mirror circuits can contain non-Clifford gates [28], which suggests a route to scalable RB of universal gate sets, and scalable “full stack” benchmarks.

Since the completion of this manuscript, Mayer et al. [57] presented a complementary theory for MRB that assumes gate-independent errors and a 2-design gate set.
This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories and the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research through the Quantum Testbed Program. Sandia National Laboratories is a multi-program laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA-0003525. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the U.S. Department of Energy, or the U.S. Government, or the views of IBM. We thank the IBM Q team for technical support.

All data and analysis code are available at 10.5281/zenodo.5197714. Our circuit sampling code is available in pyGSTi [58, 59].

REFERENCES

[1] Joseph Emerson, Robert Alicki, and Karol Zyczkowski, “Scalable noise estimation with random unitary operators,” J. Opt. B Quantum Semiclass. Opt. 7, S347 (2005).
[2] Joseph Emerson, Marcus Silva, Osama Moussa, Colm Ryan, Martin Laforest, Jonathan Baugh, David G Cory, and Raymond Laflamme, “Symmetrized characterization of noisy quantum processes,” Science 317, 1893–1896 (2007).
[3] Easwar Magesan, Jay M Gambetta, and Joseph Emerson, “Scalable and robust randomized benchmarking of quantum processes,” Phys. Rev. Lett. 106, 180504 (2011).
[4] Easwar Magesan, Jay M Gambetta, and Joseph Emerson, “Characterizing quantum gates via randomized benchmarking,” Phys. Rev. A 85, 042311 (2012).
[5] Emanuel Knill, D Leibfried, R Reichle, J Britton, RB Blakestad, JD Jost, C Langer, R Ozeri, S Seidelin, and DJ Wineland, “Randomized benchmarking of quantum gates,” Phys. Rev. A 77, 021203 (2008).
[6] Arnaud Carignan-Dugas, Joel J Wallman, and Joseph Emerson, “Characterizing universal gate sets via dihedral benchmarking,” Phys. Rev. A 92, 060302 (2015).
[7] Andrew W Cross, Easwar Magesan, Lev S Bishop, John A Smolin, and Jay M Gambetta, “Scalable randomized benchmarking of non-clifford gates,” npj Quantum Inf. 2, 16012 (2016).
[8] Winton G. Brown and Bryan Eastin, “Randomized benchmarking with restricted gate sets,” Phys. Rev. A 97, 062323 (2018).
[9] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, “Real randomized benchmarking,” Quantum 2, 85 (2018).
[10] Easwar Magesan, Jay M Gambetta, Blake R Johnson, Colm A Ryan, Jerry M Chow, Seth T Merkel, Marcus P da Silva, George A Keefe, Mary B Rothwell, Thomas A Ohki, et al., “Efficient measurement of quantum gate error by interleaved randomized benchmarking,” Phys. Rev. Lett. 109, 080505 (2012).
[11] Jonas Helsen, Xiao Xue, Lieven MK Vandersypen, and Stephanie Wehner, “A new class of efficient randomized benchmarking protocols,” npj Quantum Inf 5, 71 (2019).
[12] Jonas Helsen, Sepehr Nezami, Matthew Reagor, and Michael Walter, “Matchgate benchmarking: Scalable benchmarking of a continuous family of many-qubit gates,” Quantum 6, 657 (2022).
[13] Jahan Claes, Eleanor Kiefel, and Zhihui Wang, “Character randomized benchmarking for non-multiplicity-free groups with applications to subspace, leakage, and matchgate randomized benchmarking,” PRX Quantum 2, 010351 (2021).
[14] David C McKay, Andrew W Cross, Christopher J Wood, and Jay M Gambetta, “Correlated randomized benchmarking,” arXiv:2003.02354 [quant-ph].
[15] Jonas Helsen, Ingo Roth, Emilio Onorati, Albert H Werner, and Jens Eisert, “A general framework for randomized benchmarking,” PRX Quantum 3, 020357 (2022).
[16] Alexis Morvan, VV Ramasesh, MS Blok, JM Kreikebaum, K O’Brien, L Chen, BK Mitchell, RK Naik, DJ Santiago, and I Siddiqi, “Qutrit randomized benchmarking,” Phys. Rev. Lett. 126, 210504 (2021).
[17] Timothy J Proctor, Arnaud Carignan-Dugas, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young, “Direct randomized benchmarking for multiqubit devices,” Phys. Rev. Lett. 123, 030503 (2019).
[18] Sergio Boixo, Sergei V Isakov, Vadim N Smelyanskiy, Ryan Babbush, Nan Ding, Zhang Jiang, Michael J Bremner, John M Martinis, and Hartmut Neven, “Characterizing quantum supremacy in near-term devices,” Nat. Phys. 14, 595 (2018).
[19] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G S Brandao, David A Buell, et al., “Quantum supremacy using a programmable superconducting processor,” Nature 574, 505–510 (2019).
[20] Yunchao Liu, Matthew Otten, Roozbeh Bassiri-Jahromi, Liang Jiang, and Bill Jeffersm, “Benchmarking near-term quantum computers via random circuit sampling,” arXiv:2105.05232 [quant-ph].
[21] Andrew W Cross, Lev S Bishop, Sarah Sheldon, Paul D Nation, and Jay M Gambetta, “Validating quantum computers using randomized model circuits,” Phys. Rev. A 100, 032328 (2019).
[22] Erik Nielsen, John King Gamble, Kenneth Rudinger, Travis Scholten, Kevin Young, and Robin Blume-Kohout, “Gate set tomography,” Quantum 5, 557 (2021).
[23] Scott Aaronson and Daniel Gottesman, “Improved simulation of stabilizer circuits,” Phys. Rev. A 70, 052328 (2004).
[24] Ketan N Patel, Igor L Markov, and John P Hayes, “Efficient synthesis of linear reversible circuits,” Quantum Inf. Comput. 8, 282–294 (2008).
[25] Sergey Bravyi and Dmitri Maslov, “Hadamard-free circuits exist,” Quantum Inf. Process. 7, 4546–4563 (2021).
[26] Nikodem Grzesiak, Andrii Maksymov, Pradeep Niroula, and Yunseong Nam, “Efficient quantum programming using EASE gates on a trapped-ion quantum computer,” Quantum 6, 634 (2022).
[27] David C McKay, Sarah Sheldon, John A Smolin, Jerry M Chow, and Jay M Gambetta, “Three qubit randomized benchmarking,” Phys. Rev. Lett. 122, 200502 (2019).
[28] Timothy Proctor, Kenneth Rudinger, Kevin Young, Erik Nielsen, and Robin Blume-Kohout, “Measuring the capabilities of quantum computers,” Nat. Phys. 18, 75–79 (2022).
[29] Robin Blume-Kohout and Kevin C Young, “A volumetric framework for quantum computer benchmarks,” Quantum 4, 362 (2020).
[30] E Knill, “Quantum computing with realistically noisy devices,”
See the Supplemental Material for details of the simulations and experimental procedures.

Jay M Gambetta, AD Corcoles, Seth T Merkel, Blake R Johnson, Joseph Emerson, and Irfan Siddiqi, "Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor," Phys. Rev. X 11, 041039 (2021).

$n$-qubit layers are also known as cycles or $n$-qubit gates.

Averaged circuit eigenvalue sampling,

arXiv:2108.05303 [quant-ph].

Robin Harper, Steven T. Flammia, and Joel J. Wallman, "Efficient learning of quantum noise," Nat. Phys. 16, 1-5 (2020).

Steven T. Flammia and Joel J. Wallman, "Efficient estimation of Pauli channels," ACM Trans. Quant. Comp. 1, 3 (2020).

Steven T. Flammia, "Averaged circuit eigenvalue sampling," arXiv:2108.05303 [quant-ph].

Andrii Maksymov, Jason Nguyen, Vandiver Chaplin, Yunseong Nam, and Igor L Markov, "Detecting qubit-coupling faults in ion-trap quantum computers," 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA), 387-399 (2022).

Karl Mayer, Alex Hall, Thomas Gartner, Si Khadr Halit, Kenny Lee, Justin Bohnet, Dan Gresh, Aaron Hankin, Kevin Gilmore, and John Gaebler, "Theory of mirror benchmarking and demonstration on a quantum computer," arXiv [quant-ph] arXiv:2108.10431 [quant-ph].

Erik Nielsen, Kenneth Rudinger, Timothy Proctor, Antonio Russo, Kevin Young, and Robin Blume-Kohout, "Probing quantum processor performance with pyGSTi," Quantum Sci. Technol. 5, 044002 (2020).

Erik Nielsen, Stefan Seritan, Timothy Proctor, Kenneth Rudinger, Kevin Young, Antonio Russo, Robin Blume-Kohout, Robert Payton Kelly, John King Gamble, and Lucas Saldyt, "PyGSTi version 0.9.10.1." (2022), doi.org/10.5281/zenodo.6363115.

SUPPLEMENTAL MATERIAL

Simulations

Here we provide details of the simulations of mirror RB presented in the main text (see Fig. 2). All the error models that we simulated consisted of only stochastic Pauli errors. This enabled simulations that are efficient (polynomial) in the
number of qubits \( n \). We implemented “weak” simulation using a standard stochastic unravelling, i.e., for error model \( M \) and circuit \( C \), we sampled from \( P(x \mid C, M) \), denoting \( C \)'s output distribution over bit strings under error model \( M \). Under our error models, for each layer in \( C, M \) implies a probability distribution over the \( n \)-qubit Pauli operators (encoding a stochastic Pauli channel) that specifies the probability of each Pauli occurring after that layer (so the probability of the identity is the probability of no error). To obtain a sample from \( P(x \mid C, M) \), we sampled a Pauli operator instance after each layer in \( C \) (from \( M \)'s error distributions) to obtain a sequence of Clifford operators. We then simulated that sampled sequence using CH{P [23] (using an interface built into pyGSTi [58, 59]), obtaining a single bit string. All of our simulations consisted of generating \( N = 100 \) samples for each distinct circuit.

In the main text we presented the results of two distinct sets of simulations, shown in Figs. 2a and 2b. In both cases, the simulations are of \( n \)-qubit mirror RB for 18 different values of \( n \), ranging from \( n = 1 \) up to \( n = 225 \) qubits with exponential spacing. These simulations are of an \( n \) qubit subset of 225 qubits that are arranged on a \( 15 \times 15 \) square lattice with CNOT gates available between nearest-neighbor qubits. We independently sampled 50 distinct mirror RB experiment designs (a set of mirror RB circuits) at each \( n \), with \( K = 30 \) circuits per benchmark depth. The randomized mirror circuits depend on two user-specified components: the \( n \)-qubit layer set \( L \) and the sampling distribution \( \Omega \) over \( L \). The layer set that we used was constructed in the same way as in all of our demonstrations: \( L \) consisted of all parallel applications of CNOTs between connected qubits and all 24 single-qubit Clifford gates. The sampling distribution was the “edge grab” sampler, with a two-qubit gate density of \( 1/8 \), that was introduced and described in detail in the supplemental material of Ref. [28].

Fig. 2a shows the results of simulating each of these \( 50 \times 18 = 900 \) mirror RB experiment designs once, with an independent, randomly sampled error model for each mirror RB experiment design. For each RB experiment design, an error model was sampled as follows. The readout for each qubit is assigned a bit flip error probability, sampled uniformly and independently for each qubit from the interval \([0\%, 1\%]\). For each qubit, each of the 24 single-qubit gates (from the set of all 24 single-qubit Clifford gates) is independently assigned an \( n \)-qubit error map that acts after the gate. This error map has \( 3 + 3K \) error rate parameters that we sample, where \( K \) is number of neighbors for this qubit (\( K = 4 \) unless the qubit is on the edge of the lattice). The error map contains the 3 weight-1 Pauli errors on the target qubit, with error rates of \( \{ l_i \}_{i=1}^{15} \), and the 3K weight-1 Pauli errors that can occur on neighbouring qubits, with rates \( \{ c_i \}_{i=1}^{3K} \). These rates are sampled as follows. We sample \( \gamma \) uniformly from the interval \([0\%, 0.2\%]\), \( \kappa \) uniformly from the interval \([0.5, 1]\), \( l_i \) uniformly from the interval \([0, 1]\) for \( i = 1, 2, 3 \), and \( c_i \) uniformly from the interval \([0, 1]\) for \( i = 1, \ldots, 3K \). If the gate is a \( \sigma_z \)-basis rotation we set \( c_i = 0 \) for all \( i \). We then set \( l_i = \gamma k i + c_i (1 - \gamma) k i \), for \( i = 1, \ldots, 3K \). Therefore, the total rate of errors that this gate induces on its target qubit is \( \epsilon_{\text{target}} = \kappa \gamma \), and the total rate of errors that it induces on its neighbours is \( \epsilon_{\text{neighbours}} = \gamma (1 - \kappa) \) — except for \( \sigma_z \)-basis rotation where \( \epsilon_{\text{target}} = 0 \) (we make this choice as \( \sigma_z \)-basis rotation do not typically involve a physical pulse). The error map for each CNOT gate (corresponding to each edge on the lattice) is constructed in the same way, except that (1) \( \gamma \) is sampled uniformly from the \([0\%, 2\%]\), and (2) there are 15 possible Pauli errors on the target qubit, and so \( \{ l_i \}_{i=1}^{15} \). In each of the 900 independent simulations, we extract \( \epsilon_2 \) from the error model and we estimate \( \epsilon_3 \) from the simulated data. Fig. 2a shows a scatter plot of \( \epsilon_2 \) versus \( \epsilon_3 \).

The simulations in Fig. 2b consisted of sampling one mirror RB experiment design for each \( n \) under two different error models — with and without long-range crosstalk errors. In the first model (model 1)

- The readout on each qubit is subject to a bit flip error with a probability of 0.5%.
- Each one-qubit gate (any one of the 24 single-qubit Clifford gates) is followed by a one-qubit depolarizing channel with an entanglement infidelity of 1%.
- Each CNOT gate is followed by a two-qubit depolarizing channel with an entanglement infidelity of 1%.

The second model (model 2) is the same as model 1 except that the CNOT gates also cause long-range crosstalk errors. In particular, each CNOT is still followed by a two-qubit depolarizing channel with an entanglement infidelity of 1% on its two target qubits, \( q_1 \) and \( q_2 \), but it also causes local one-qubit depolarization on all other qubits. It applies a one-qubit depolarizing channel on qubit \( q \) with an entanglement infidelity of \( \delta(q) \), where \( \delta(q) \) is a decreasing function of the distance (on the lattice) from \( q \) to the CNOT location. The specific function used was \( \delta(q) = 0.0035 \times 0.999^{d(q-q_l-q_r)} \) where \( d(q-q_l-q_r) \) is the minimum of the distance between \( q \) and \( q_1 \) and the distance between \( q \) and \( q_2 \). For each of the two models, at each \( n \) we extract \( \epsilon_2 \) from the error model and we estimate \( \epsilon_3 \) from the simulated data. Fig. 2b shows \( \epsilon_2 \) and \( \epsilon_3 \) versus \( n \) for both error models.

**IBM Q Quito demonstrations**

Our demonstration on IBM Q Quito (ibmq_quito version 1.1.3, a 5-qubit Falcon r4T processor) was performed in June 2021. It consisted of running mirror, direct and standard RB on 1-5 qubits. Randomized mirror circuits are constructed using two user-specified components: the \( n \)-qubit layer set \( L \) and the sampling distribution \( \Omega \) over \( L \). The layer set that we used in this demonstration was constructed in the same way as in all of our demonstrations: \( L \) consisted of all parallel applications of CNOTs between connected qubits and all 24 single-qubit Clifford gates. The sampling distribution was the “edge grab” sampler, with a two-qubit gate density of \( 1/8 \), that was introduced and described in detail in the supplemental material of Ref. [28]. For each \( n \), we used exponentially spaced benchmark depths \( d \) (see Fig. 3). For each \( n \) and \( d \), we sampled and ran \( K = 30 \) randomized mirror circuits.
Direct RB [17] contains the same flexible sampling as mirror RB: the user specifies $L$ and $\Omega$. Direct RB is designed to measure the same average layer infidelity ($\epsilon_0$), but because direct RB and mirror RB are different protocols there is no a priori guarantee that their empirical error rates $r_\Omega$ will be equal. One of the motivations for running direct RB alongside mirror RB is to show that we obtain approximately the same result. Use the same $L$ and $\Omega$ as in our concurrent mirror RB demonstration.

For our purposes, an $n$-qubit direct RB circuit of benchmark depth $d$ consists of (1) a circuit that generates a uniformly random $n$-qubit stabilizer state, (2) a sequence of $d$ Pauli-dressed layers, consisting of a uniformly random Pauli layer and then a layer $L$ sampled from $\Omega$, and (3) a circuit that returns the qubits to a uniformly random computational basis state. To implement direct RB, it is necessary to compile the sub-circuits of steps (1) and (3) into the available gates (and it is the large circuits generated by this compilation when $n \gg 1$ that makes direct RB unscalable). We do so using the open-source compilers in pyGSTi. To analyze the direct RB data, we fit the mean success probability versus depth to $A p^d + \frac{1}{2^r}$, and then $r_\Omega = \frac{4^d - 1}{4^d} (1 - p)$ as with mirror RB. We are able to fix the asymptote to $\frac{1}{2^r}$ because the target bit string is randomized [38]. In Fig. 3 we present polarization decays, instead of success probability decays, but note that this is only a rescaling.

Our standard RB demonstrations follow the standard procedure for RB over the Clifford group [3, 4], except that we randomized the ideal output bit string by only inverting each sequence of random Clifford gates up to a randomly sampled Pauli operator. Although this is not current standard practice, it has been advocated for elsewhere — because it means that the success probability decay will asymptote to $\frac{1}{2^r}$ [38]. In particular, and as with direct RB, we fit the mean success probability versus depth to $A p^d + \frac{1}{2^r}$ (for standard RB, we use the number of Clifford subroutines minus two to define the benchmark depth $d$, as then the shortest circuits allowed by all three methods correspond to $d = 0$). Standard RB does not contain any flexible sampling, but it is necessary to compile the $n$-qubit Clifford subroutines in the circuits into the available gates. As with direct RB, we do so using the open-source compilers in pyGSTi. These compilations are unlikely to be optimal (e.g., in the number of CNOTs in the circuits), and so a slower decrease in the $d = 0$ success probability with $n$ (or, equivalently, a slower decrease in the $d = 0$ mean polarization $\overline{S}_0$, which is what is shown in Fig. 3e) than observed in our demonstration might have been achievable. However, at least $O(n^2/\log(n))$ two-qubit gates are required to implement a typical $n$-qubit Clifford operator [23–25], so it is not possible to improve upon the scaling seen in our results.

In the main text we do not present the error rates $r$ obtained from standard RB, as they are not directly comparable to the direct and mirror RB error rates. Specifically, standard RB measures (up to some subtleties [60–63]) the infidelity of an average $n$-qubit Clifford gate, whereas direct and mirror RB approximately measure $\epsilon_0$ — the error rate of an average layer. The observed $n$-qubit Clifford RB error rate $r_n$ is $r_1 = 0.041(1)\%$, $r_2 = 1.97(5)\%$, $r_3 = 17.9(7)\%$, $r_4 = 67(6)\%$ and $r_5 = 95(7)\%$, where we have defined $r = \frac{4^d - 1}{4^d} (1 - p)/4^d$ rather than using the more common definition of $r = (2^d - 1)(1 - p)/2^d$ for consistency with the convention that we have used for mirror and direct RB (which we choose due to the arguably preferable properties of entanglement infidelity compared to average gate infidelity).

IBM Q Rueschlikon demonstrations

Our demonstration on IBM Q Rueschlikon (ibmqx5 version 1.1.0, a 16-qubit processor) was performed in July 2018. It consisted of running mirror RB on multiple $n$-qubit subsets of the processor, for $n \in \{1, 2, 4, 8, 16\}$. For each $n$, we divided Rueschlikon into $19/4$ regions (as shown in Fig. 4), and we ran randomized mirror circuits on each region. The one-qubit circuits were performed simultaneously to match IBM’s calibration experiments. These circuits were constructed using a sampling distribution $\Omega$ for which sampling a layer consisted of (1) selecting a pair of connected qubits uniformly at random, and adding a CNOT between those qubits to the layer with probability $50\%$, and (2) sampling independent and uniformly random single-qubit Clifford gates to apply to all qubits that do not have a CNOT acting on them. A layer sampled from this $\Omega$ therefore contains no CNOT gates with $50\%$ probability and otherwise it contains one CNOT gate. Note that this sampling distribution is not suitable for much larger processors (unlike the edge grab sampler used in our simulations and the demonstrations on IBM Q Quito). This is because the CNOT density $\xi \to 0$ as $n \to \infty$ and circuits containing very few two-qubit gates cannot be scrambling [17]. The data from these demonstrations was also presented in Ref. [28] (see Fig. 1 therein), where it was used to construct volumetric benchmarking plots [29] rather than estimate layer error rates.