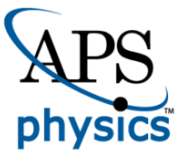

Scalable Randomized Benchmarking of Quantum Computers Using Mirror Circuits

Timothy Proctor, Stefan Seritan, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young

Phys. Rev. Lett. **129**, 150502 — Published 6 October 2022

DOI: 10.1103/PhysRevLett.129.150502

Scalable randomized benchmarking of quantum computers using mirror circuits

Timothy Proctor, Stefan Seritan, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young
Quantum Performance Laboratory, Sandia National Laboratories,
Albuquerque, NM 87185, USA and Livermore, CA 94550, USA
(Dated: September 9, 2022)

The performance of quantum gates is often assessed using some form of randomized benchmarking. However, the existing methods become infeasible for more than approximately five qubits. Here we show how to use a simple and customizable class of circuits—randomized mirror circuits—to perform scalable, robust, and flexible randomized benchmarking of Clifford gates. We show that this technique approximately estimates the infidelity of an average many-qubit logic layer, and we use simulations of up to 225 qubits with physically realistic error rates in the range 0.1-1% to demonstrate its scalability. We then use up to 16 physical qubits of a cloud quantum computing platform to demonstrate that our technique can reveal and quantify crosstalk errors in many-qubit circuits.

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 $n \sim 50$ [18–21]. In contrast, standard RB requires only efficient classical computations but it benchmarks composite gates from the n-qubit Clifford group. They require $O(n^2/\log n)$ 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 all 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

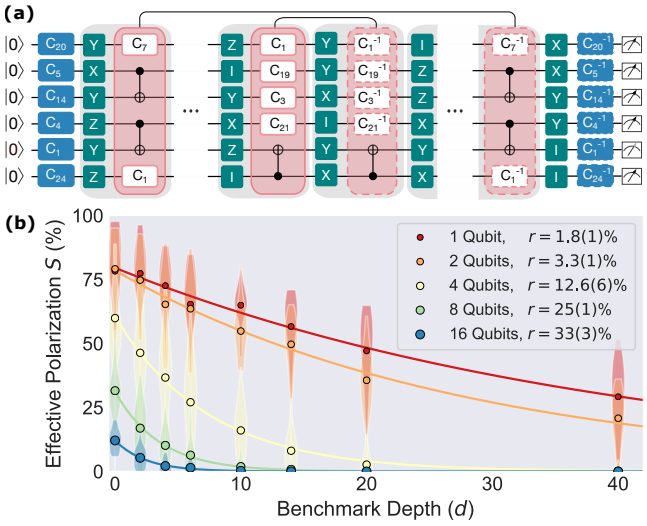

Figure 1. **Scalable RB with mirror circuits. (a)** Randomized mirror circuits over Clifford gates enable scalable RB. These circuits contain $d/2$ pairs of layers consisting of a layer and its inverse (pink boxes) sampled from some set of n-qubit Clifford layers, $d+1$ layers of uniformly random Pauli gates (green boxes), and a layer of uniformly random one-qubit Clifford gates and this layer's inverse (blue boxes). The number of "Pauli-dressed" layers (grey boxes) d is the circuit's *benchmark depth*. These circuits' "effective polarization" S, a quantity closely related to success probability, decays exponential with d. **(b)** Demonstrating our method on 1, 2, 4, 8 and 16 qubit subsets of IBM Q Rueschlikon. Points (violin plots) are the means (distributions) of S versus d, and the curves are fits to $S = Ap^d$. Each r is a rescaling of p that approximates the infidelity of an average Pauli-dressed n-qubit layer (uncertainties are 1σ here and thoughout).

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 $\mathbb{L} = \{L\}$ [34], a probability distribution Ω over $\mathbb{L}$, and a benchmark depth d that specifies the number of Pauli-dressed layers in the circuit. Both $\mathbb{L}$ and Ω are customizable, but we require that (1) each layer contains only Clifford gates, (2) each layer's inverse L^{-1} is also within $\mathbb{L}$, (3) $\Omega(L) = \Omega(L^{-1})$, and (4) Ω-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 to, e.g., CNOTs synthesized via SWAP chains, enabling comparisons between the errors in identical layer sets on different devices. All our distributions Ω 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 $\epsilon_\Omega := \sum_L \Omega(L)\epsilon(L)$, where Ω is a user-chosen distribution over $\mathbb{L}$, and $\epsilon(L)$ is the entanglement infidelity of the Pauli-dressed version of the n-qubit layer L (grey boxes, Fig. 1a). In all our demonstrations we do not compile the Paulis into the L layers, but this is permissible. MRB estimates ϵ_Ω using data from Ω-sampled randomized mirror circuit. For each circuit C that we run, we estimate its *effective polarization*

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

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 Ω, and run each one $N \geq 1$ times.

2. Estimate each circuit's effective polarization S.

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

Theory. We now show that MRB is reliable, i.e., $\overline{S}_d \approx Ap^d$ and $r_\Omega \approx \epsilon_\Omega$ 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 $\mathcal{E}(L)$ its error map, i.e., $\phi(L) = \mathcal{E}(L)U(L)$. Our theory starts from a single randomized mirror circuit C of benchmark depth d. So $C = F_0^{-1}P_dL_1^{-1}\cdots P_{1+d/2}L_{d/2}^{-1}P_{d/2}L_{d/2}\cdots P_1L_1P_0F_0$, 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 Ω-sampled layers and L_i^{-1} their inverses. The components (1), (2) and (3) are sampled independently. To compute $\overline{S}_d$, as a function of $\mathcal{E}(L)$, we can therefore average over (1-3) separately in turn.

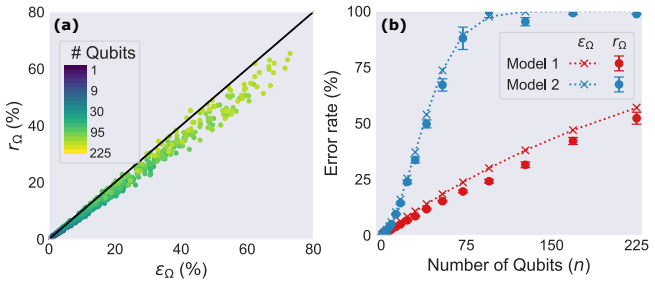

Figure 2. **Validating MRB with many-qubit simulations**. Simulations of MRB on up to 225 qubits show that it reliably approximates the infidelity of n-qubit layers, i.e., $r_\Omega \approx \epsilon_\Omega$. **(a)** r_Ω versus ϵ_Ω for randomly sampled error models. Each point was generated from an independent simulation (sampling an error model and circuits, simulating the circuits, and then applying the analysis to estimate r_Ω) for gates subject to stochastic Pauli errors. **(b)** r_Ω and ϵ_Ω versus n for two illustrative error models, with (model 2) and without (model 1) long-range crosstalk. This demonstrates the power of MRB to highlight crosstalk errors.

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 $\overline{S}_d \to 0$ as $d \to \infty$ 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 $C = F_0^{-1}L_1^{-1}\cdots L_d^{-1}L_d\cdots L_1F_0$ with each L_i's error map $\mathcal{E}(L_i)$ a stochastic Pauli channel. The composite superoperator for this circuit is $\phi(C) = \phi(F_0^{-1})\mathcal{E}_d\phi(F_0)$ where $\mathcal{E}_d \equiv \phi(L_1^{-1})\cdots\phi(L_d^{-1})\phi(L_d)\cdots\phi(L_1)$ is a stochastic Pauli channel, as each $U(L_i)$ is a Clifford operator.

The intial 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, $\overline{\mathcal{E}}_d \equiv \frac{1}{24^n}\sum_{F_0}[U(F_0^{-1})\mathcal{E}_dU(F_0)]$ 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 $\overline{\mathcal{E}}_d$ flips at least one output bit iff it applies X or Y to at least one qubit. So, if $\overline{\mathcal{E}}_d$ induces a weight k error (an error on k qubits) the circuit outputs s_C with a probability of $1/3^k$. Generally, a weight k error causes flips on j of the output bits with probability $M_{jk} = \binom{k}{j}\frac{2^j}{3^k}$. So $\vec{h} = M\vec{p}$ where h_k and p_k are the probabilities that k bits are flipped and that $\overline{\mathcal{E}}_d$ induces a weight k error, respectively, with $k = 0, \ldots, n$. By inverting M, we obtain $p_0 = \sum_{k=0}^{n}(-1/2)^k h_k \equiv H$. Because $p_0 = 1 - \epsilon(\mathcal{E}_d)$ where $\epsilon(\mathcal{E}_d)$ is $\mathcal{E}_d$'s entanglement infidelity, H

therefore equals $\mathcal{E}_d$'s entanglement fidelity, and S [Eq. (1)] its *polarization* $\gamma(\mathcal{E}_d) := 1 - 4^n\epsilon(\mathcal{E}_d)/(4^n - 1)$. 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 $S \approx A\gamma(\mathcal{E}_d)$ for some A.

We have related a randomized mirror circuit's S to the polarization of its superoperator [$\gamma(\mathcal{E}_d)$]. Now we relate $\gamma(\mathcal{E}_d)$ to the polarizations of the circuit's constituent layers [$\gamma(\mathcal{E}(L_i))$]. If every $\mathcal{E}(L_i)$ is an n-qubit depolarizing channel, with layer-dependent error rates, then $\gamma(\mathcal{E}_d) = \prod_{i=1}^{d}\gamma_{i^{-1}}\gamma_i$ where $\gamma_i \equiv \gamma(\mathcal{E}(L_i))$. More generally we argue that $\gamma(\mathcal{E}_d) \approx \prod_{i=1}^{d}\gamma_{i^{-1}}\gamma_i$. For two stochastic Pauli channels $\mathcal{E}_A$ and $\mathcal{E}_B$, $\gamma(\mathcal{E}_A\mathcal{E}_B) = \gamma(\mathcal{E}_A)\gamma(\mathcal{E}_B) + \eta$ where $\eta = \sum_j(\epsilon_{A,j} - \frac{\epsilon[\mathcal{E}_A]}{4^n-1})(\epsilon_{B,j} - \frac{\epsilon[\mathcal{E}_B]}{4^n-1}])$ and $\vec{\epsilon}_i$ is the vector of $4^n - 1$ Pauli error probabilities for $\mathcal{E}_i$. η quantifies the rate that errors cancel when composing the two channels, relative to the rate that they cancel when composing n-qubit depolarization channels. It is negligible unless $\vec{\epsilon}_A$ and $\vec{\epsilon}_B$ are sparse (e.g., if $\vec{\epsilon}_A = \vec{\epsilon}_B$ and the error probability is equally distributed over K errors, then $\eta = \epsilon(\mathcal{E}_A)^2\left[\frac{1}{K} - \frac{1}{4^n-1}\right]$). So, unless the Pauli error probability distributions of the L_i are sharply spiked, then $\gamma(\mathcal{E}_d) \approx \prod_{i=1}^{d}\gamma_{i^{-1}}\gamma_i$ for any randomized mirror circuit. Furthermore, because of the properties that we demand of Ω (see above), our circuits are "scrambling"—they locally randomize errors, and quickly spread them across many qubits. This suppresses error cancellation further [17]. So $\gamma(\mathcal{E}_d) \approx \prod_{i=1}^{d}\gamma_{i^{-1}}\gamma_i$ for a typical randomized mirror circuit.

Finally, we calculate the effect of averaging over the L_i layers (pink boxes, Fig. 1a). They are independently sampled from Ω, so $\overline{S}_d \approx A(\sum_L \Omega(L)\gamma_{L^{-1}}\gamma_L)^{d/2}$ where $\gamma_L \equiv \gamma(\mathcal{E}(L))$. That is, $\overline{S}_d \approx Ap^d$ where $p^2 \approx \sum_L \Omega(L)\gamma_{L^{-1}}\gamma_L$. Rewriting this in terms of ϵ_Ω and $\mathrm{Cov}_\Omega = [\sum_L \Omega(L)\epsilon(L^{-1})\epsilon(L)] - [\epsilon_\Omega]^2$ gives $p^2 \approx (1 - \frac{4^n}{4^n-1}\epsilon_\Omega)^2 + \frac{4^n}{4^n-1}\mathrm{Cov}_\Omega$. So if $\mathrm{Cov}_\Omega = 0$ then $r_\Omega \approx \epsilon_\Omega$. Cov_Ω quantifies the correlation between the error rate of a Ω-random layer L and its inverse L^{-1}, so $\mathrm{Cov}_\Omega \neq 0$ is likely. This covariance satisfies $\epsilon_\Omega(1 - \epsilon_\Omega) \geq \mathrm{Cov}_\Omega \geq -\epsilon_\Omega^2$, so $\epsilon_\Omega + O(\epsilon_\Omega^2) \gtrapprox r_\Omega \lessapprox \frac{\epsilon_\Omega}{2} + O(\epsilon_\Omega^2)$. Therefore r_Ω is never significantly large than ϵ_Ω, and it can be smaller by at most a factor of ≈ 2. The $\{\epsilon(L)\}$ distributions that get close to these bounds on Cov_Ω are not physically typical, e.g., the upper bound is saturated if $\epsilon(L) = \epsilon(L^{-1})$ and $\epsilon(L) = 0$ or $\epsilon(L) = 1$ for each L. We therefore conjecture that, for physically relevant $\{\epsilon(L)\}$, r_Ω typically only slightly underestimates ϵ_Ω. 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 15×15 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_Ω versus r_Ω. We observe that $r_\Omega \approx \epsilon_\Omega$, with r_Ω typically slightly less than ϵ_Ω, as expected from our theory. Quantifying estimation error by $\delta_{\mathrm{rel}} = \frac{r_\Omega - \epsilon_\Omega}{\epsilon_\Omega}$, we find that $\delta_{\mathrm{rel}} > -0.32$ in all 900 simulations and for each n its mean $\overline{\delta}_{\mathrm{rel}}$ satisfies $0.003 > \overline{\delta}_{\mathrm{rel}} > -0.16$. Although this systematic underestimation of ϵ_Ω is undesirable, it is arguably small enough to be insignificant (RB is typically used for rough estimates of gate performance rather than precision characterization).

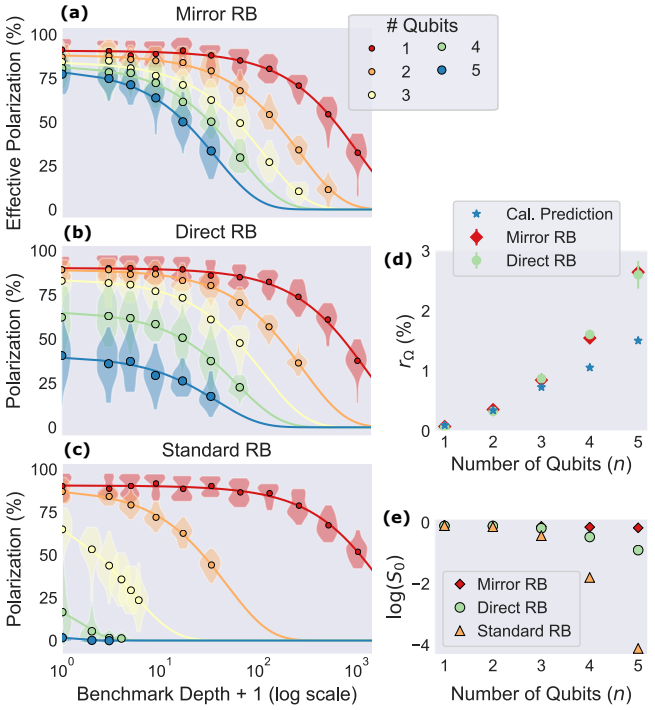

Figure 3. **Validating MRB using cloud access experiments.** MRB, DRB and standard RB on 1-5 qubits of IBM Q Quito. **(a-c)** The means (points) and distributions (violin plots) of the circuit polarizations versus benchmark depth (d), and fits to an exponential Ap^d (curves). **(d)** Error rates (r) 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.

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 $\epsilon(q)$, with $\epsilon(q)$ a slowly decreasing function of the distance (on the lattice) from q to the CNOT's location [40]. Fig. 2b shows r_Ω (points) and ϵ_Ω (dotted line) versus n for both models. We find that $r_\Omega \approx \epsilon_\Omega$ (averaged over n, $\overline{\delta}_{\mathrm{rel}} \approx -0.17$ and $\overline{\delta}_{\mathrm{rel}} \approx -0.08$ for the crosstalk-free and crosstalk models, respectively), and that r_Ω grows quadratically at low n under the crosstalk model—an effect that cannot be observed without running many-qubit circuits.

Validating MRB with cloud access experiments. To demon-

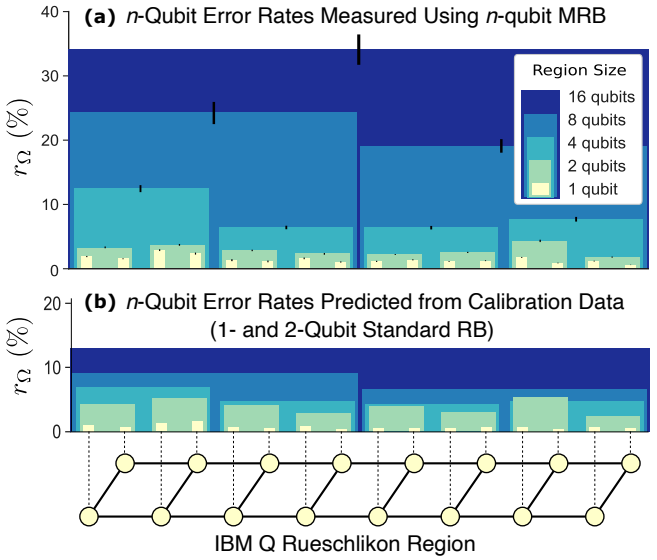

Figure 4. **Mapping out the performance of a 16-qubit processor.** MRB with was used to probe the performance of n-qubit regions of IBM Q Rueschlikon. **(a)** The measured error rate (r_Ω) for each qubit subset that was tested (black lines are 1σ uncertainties) and **(b)** the over-optimistic predictions from calibration data. The horizontal axis is a device schematic (nodes are qubits and edges the available CNOTs).

strate MRB and compare it to existing techniques, we ran MRB, DRB [17] and standard RB [3] on 1-5 qubits of IBM Q Quito [41]. DRB is designed to measure the same quantity as MRB (ϵ_Ω) and is known to be reliable but unscalable (because its circuits start by preparing a random n-qubit stabilizer state). For DRB and MRB we sampled layers with an expected CNOT density of $\xi = 1/8$ [40] (standard RB does not have flexible sampling and its error rate is incomparable). Fig. 3a-c shows that we observe exponential decays for all three methods and all n (for all methods $d = 0$ corresponds to the shortest allowed circuit, consisting of a random n-qubit Clifford and its inverse for standard RB and preparation in and measurement of a random stabilizer state for DRB). For standard RB and DRB we rescale the success probabilities P to polarizations $(P - 1/2^n)/(1 - 1/2^n)$ [this has no effect on the estimated r] for easier comparison with MRB.

Fig. 3a-c also highlights the fundamentally improved scaling of MRB. The $d = 0$ polarization (Fig. 3e) decays much more quickly with n for DRB and standard RB, because they use subroutines containing $O(n^2/\log n)$ gates, whereas $d = 0$ randomized mirror circuits use $O(n)$ gates. The error rates estimated by DRB and MRB are in close agreement for all n (Fig. 3d), validating MRB. We also predicted r_Ω from Quito's calibration data [40]. These predictions (stars, Fig. 3d) are consistent with our observations for $n = 1, 2$, but they are over-optimistic as n increases. This discrepancy indicates crosstalk errors caused by CNOTs. This is because IBM's one- and two-qubit calibration data are obtained from simultaneous one-qubit RB and isolated two-qubit RB (i.e., all other qubits are left idle) [39, 41], respectively. Therefore, the one-qubit error rates include contributions from any one-qubit gate crosstalk, whereas the two-qubit error rates do not include contributions from two-qubit gate crosstalk.

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 $16/n$ 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_Ω for all benchmarked regions and the predictions from the calibration data, respectively. The prediction underestimates r_Ω for $n > 2$, again signifying crosstalk 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 crosstalk. Similarly, running multiple MRB experiments with Ω 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 crosstalk 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.

ACKNOWLEDGMENTS

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 Życzkowski, "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**, 012307 (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 randomised 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 Rieffel, 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 OBrien, L Chen, BK Mitchell, RK Naik, DI 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 GSL Brandao, David A Buell, *et al.*, "Quantum supremacy using a programmable superconducting processor," Nature **574**, 505–510 (2019).

[20] Yunchao Liu, Matthew Otten, Roozbeh Bassirianjahromi, Liang Jiang, and Bill Fefferman, "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 expose the structure of the clifford group," IEEE Trans. Inf. Theory **67**, 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,"

Nature **434**, 39–44 (2005).
[31] Joel J Wallman and Joseph Emerson, "Noise tailoring for scalable quantum computation via randomized compiling," Phys. Rev. A **94**, 052325 (2016).
[32] Matthew Ware, Guilhem Ribeill, Diego Ristè, Colm A. Ryan, Blake Johnson, and Marcus P. da Silva, "Experimental pauli-frame randomization on a superconducting qubit," Phys. Rev. A **103**, 042604 (2021).
[33] Akel Hashim, Ravi K. Naik, Alexis Morvan, Jean-Loup Ville, Bradley Mitchell, John Mark Kreikebaum, Marc Davis, Ethan Smith, Costin Iancu, Kevin P. O'Brien, Ian Hincks, Joel J. Wallman, Joseph Emerson, and Irfan Siddiqi, "Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor," Phys. Rev. X **11**, 041039 (2021).
[34] n-qubit layers are also known as cycles or n-qubit gates.
[35] Akel Hashim, Stefan Seritan, Timothy Proctor, Kenneth Rudinger, Noah Goss, Ravi K Naik, John Mark Kreikebaum, David I Santiago, and Irfan Siddiqi, "Benchmarking verified logic operations for fault tolerance," arXiv:2207.08786 [quant-ph].
[36] Bryan H Fong and Seth T Merkel, "Randomized benchmarking, correlated noise, and ising models," arXiv preprint arXiv:1703.09747 .
[37] Jeffrey M Epstein, Andrew W Cross, Easwar Magesan, and Jay M Gambetta, "Investigating the limits of randomized benchmarking protocols," Phys. Rev. A **89**, 062321 (2014).
[38] Robin Harper, Ian Hincks, Chris Ferrie, Steven T Flammia, and Joel J Wallman, "Statistical analysis of randomized benchmarking," Phys. Rev. A **99**, 052350 (2019).
[39] Jay M Gambetta, AD Córcoles, Seth T Merkel, Blake R Johnson, John A Smolin, Jerry M Chow, Colm A Ryan, Chad Rigetti, S Poletto, Thomas A Ohki, *et al.*, "Characterization of addressability by simultaneous randomized benchmarking," Phys. Rev. Lett. **109**, 240504 (2012).
[40] See the Supplemental Material for details of the simulations and the demonstrations on IBM Q, which includes Refs. [60–63].
[41] IBM Quantum https://quantum-computing.ibm.com, (2021).
[42] Robin Harper and Steven T Flammia, "Estimating the fidelity of T gates using standard interleaved randomized benchmarking," Quantum Sci. Technol. **2**, 015008 (2017).
[43] Tobias Chasseur, Daniel M Reich, Christiane P Koch, and Frank K Wilhelm, "Hybrid benchmarking of arbitrary quantum gates," Phys. Rev. A **95**, 062335 (2017).
[44] Joel Wallman, Chris Granade, Robin Harper, and Steven T Flammia, "Estimating the coherence of noise," New J. Phys. **17**, 113020 (2015).
[45] Guanru Feng, Joel J Wallman, Brandon Buonacorsi, Franklin H Cho, Daniel K Park, Tao Xin, Dawei Lu, Jonathan Baugh, and Raymond Laflamme, "Estimating the coherence of noise in quantum control of a solid-state qubit," Phys. Rev. Lett. **117**, 260501 (2016).
[46] Sarah Sheldon, Lev S Bishop, Easwar Magesan, Stefan Filipp, Jerry M Chow, and Jay M Gambetta, "Characterizing errors on qubit operations via iterative randomized benchmarking," Phys. Rev. A **93**, 012301 (2016).
[47] Christopher J Wood and Jay M Gambetta, "Quantification and characterization of leakage errors," Phys. Rev. A **97**, 032306 (2018).
[48] T Chasseur and FK Wilhelm, "Complete randomized benchmarking protocol accounting for leakage errors," Phys. Rev. A **92**, 042333 (2015).
[49] Joel J Wallman, Marie Barnhill, and Joseph Emerson, "Robust characterization of loss rates," Phys. Rev. Lett. **115**, 060501 (2015).
[50] MA Rol, CC Bultink, TE O'Brien, SR de Jong, LS Theis, X Fu, F Luthi, RFL Vermeulen, JC de Sterke, A Bruno, *et al.*, "Restless tuneup of high-fidelity qubit gates," Phys. Rev. Appl. **7**, 041001 (2017).
[51] J Kelly, R Barends, B Campbell, Y Chen, Z Chen, B Chiaro, A Dunsworth, AG Fowler, I-C Hoi, E Jeffrey, *et al.*, "Optimal quantum control using randomized benchmarking," Phys. Rev. Lett. **112**, 240504 (2014).
[52] Alexander Erhard, Joel James Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban Adrian Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt, "Characterizing large-scale quantum computers via cycle benchmarking," Nat. Commun. **10**, 5347 (2019).
[53] Robin Harper, Steven T. Flammia, and Joel J. Wallman, "Efficient learning of quantum noise," Nat. Phys. **16**, 1–5 (2020).
[54] Steven T. Flammia and Joel J. Wallman, "Efficient estimation of Pauli channels," ACM Trans. Quant. Comp. **1**, 3 (2020).
[55] Steven T Flammia, "Averaged circuit eigenvalue sampling," arXiv:2108.05803 [quant-ph].
[56] 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).
[57] Karl Mayer, Alex Hall, Thomas Gatterman, Si Khadir 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].
[58] 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).
[59] 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), https://doi.org/10.5281/zenodo.6363115.
[60] Timothy Proctor, Kenneth Rudinger, Kevin Young, Mohan Sarovar, and Robin Blume-Kohout, "What randomized benchmarking actually measures," Phys. Rev. Lett. **119**, 130502 (2017).
[61] Joel J Wallman, "Randomized benchmarking with gate-dependent noise," Quantum **2**, 47 (2018).
[62] Seth T. Merkel, Emily J. Pritchett, and Bryan H. Fong, "Randomized Benchmarking as Convolution: Fourier Analysis of Gate Dependent Errors," Quantum **5**, 581 (2021).
[63] Arnaud Carignan-Dugas, Kristine Boone, Joel J Wallman, and Joseph Emerson, "From randomized benchmarking experiments to gate-set circuit fidelity: how to interpret randomized benchmarking decay parameters," New J. Phys. **20**, 092001 (2018).