Direct Randomized Benchmarking for Multiqubit Devices

Timothy J. Proctor,[1] Arnaud Carignan-Dugas,[2] Kenneth Rudinger,[3] Erik Nielsen,[3]
Robin Blume-Kohout,[3] and Kevin Young[1]
[1]*Quantum Performance Laboratory, Sandia National Laboratories, Livermore, California 94550, USA*
[2]*Institute for Quantum Computing and the Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*
[3]*Quantum Performance Laboratory, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA*

(Received 14 December 2018; published 19 July 2019)

Benchmarking methods that can be adapted to multiqubit systems are essential for assessing the overall or "holistic" performance of nascent quantum processors. The current industry standard is Clifford randomized benchmarking (RB), which measures a single error rate that quantifies overall performance. But, scaling Clifford RB to many qubits is surprisingly hard. It has only been performed on one, two, and three qubits as of this writing. This reflects a fundamental inefficiency in Clifford RB: the n-qubit Clifford gates at its core have to be compiled into large circuits over the one- and two-qubit gates native to a device. As n grows, the quality of these Clifford gates quickly degrades, making Clifford RB impractical at relatively low n. In this Letter, we propose a direct RB protocol that mostly avoids compiling. Instead, it uses random circuits over the native gates in a device, which are seeded by an initial layer of Clifford-like randomization. We demonstrate this protocol experimentally on two to five qubits using the publicly available ibmqx5. We believe this to be the greatest number of qubits holistically benchmarked, and this was achieved on a freely available device without any special tuning up. Our protocol retains the simplicity and convenient properties of Clifford RB: it estimates an error rate from an exponential decay. But, it can be extended to processors with more qubits—we present simulations on 10+ qubits—and it reports a more directly informative and flexible error rate than the one reported by Clifford RB. We show how to use this flexibility to measure separate error rates for distinct sets of gates, and we use this method to estimate the average error rate of a set of CNOT gates.

DOI: 10.1103/PhysRevLett.123.030503

With quantum processors incorporating 5 to 20 qubits now commonplace [1–12], and 50+ qubits expected soon [13–15], efficient, holistic benchmarks are becoming increasingly important. Isolated qubits or coupled pairs can be studied in detail with tomographic methods [16–20], but the required resources scale exponentially with the qubit number n, making these techniques infeasible for $n \gg 2$ qubits. And, although an entire device could be characterized two qubits at a time, this often results in overoptimistic estimates of device performance that ignore crosstalk and collective dephasing effects. What is needed instead is a family of holistic benchmarks that quantify the performance of a device as a whole. Randomized benchmarking (RB) methods [21–29] avoid the specific scaling problems that afflict tomography—in RB, both the number of experiments [30] and the complexity of the data analysis [25] are independent of n—but introduce a new scaling problem in the form of gate compilation.

Although a quantum processor's native gates typically include only a few one- and two-qubit operations, the "gates" benchmarked by RB are elements of an exponentially large n-qubit group 2-design (e.g., the Clifford group). These gates must be compiled into the native gate set [31,32]. As the number of qubits increases, the circuit depth and infidelity of these compiled group elements grow rapidly, rendering current RB protocols impractical for relatively small n, even with state-of-the-art gates. The industry-standard protocol laid out by Magesan *et al.* [24,25]—which we will refer to as Clifford randomized benchmarking (CRB)—has been widely used to benchmark [33–44] and calibrate [45,46] both individual qubits and pairs of qubits, but we are aware of just one reported application to three qubits [47] and none to four or more.

Another consequence of compilation is that, instead of quantifying native gate performance, CRB measures the error per compiled group element. Although this is sometimes translated into a native gate error rate (e.g., by dividing it by the average circuit size of a compiled Clifford gate [42–44]), this is *ad hoc* and not always reliable [48]. Moreover, error rates obtained this way are hard to interpret for $n \gg 1$ CRB, where error rates can vary widely between native gates.

In this Letter, we propose and demonstrate direct randomized benchmarking (DRB), which is an RB protocol that directly benchmarks the native gates of a device.

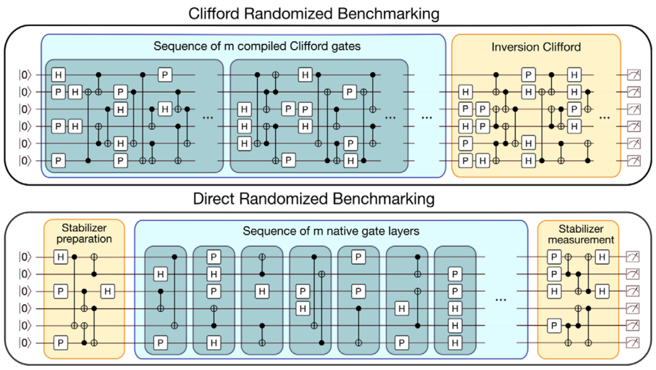


FIG. 1. Illustration of circuits used in Clifford RB and the streamlined direct RB protocol that we propose.

Like CRB, our DRB protocol utilizes random circuits of variable lengths, but these circuits directly contain the native gates of the device rather than compiled Clifford operations (see Fig. 1). Our protocol is not infinitely scalable, but the simplified structure enables DRB to be successfully implemented on significantly more qubits than CRB. Moreover, DRB preserves the core simplicity of CRB: it estimates an error rate from an exponential decay.

We anticipate that DRB will be an important tool for characterizing current multiqubit devices. For this reason, this Letter focuses on the practical applications of DRB. We present experiments on two to five qubits and simulations on up to 14 qubits. These examples show that DRB works, they demonstrate how our protocol improves on current methods, and they show that DRB can be implemented on significantly more than two qubits on current devices. We follow these demonstrations with arguments for why DRB is broadly reliable, but this Letter does not contain a comprehensive theory for DRB—that will be presented in a series of future papers.

Direct randomized benchmarking.—DRB is a protocol to directly benchmark the native gates in a device. There is flexibility in defining a device's "native gates." For DRB, we only require that they generate the n-qubit Clifford group $\mathbb{C}_n$ [49]. Normally, they will be all the n-qubit Clifford operations that can be implemented by depth-one circuits, e.g., by parallel one- and two-qubit gates (see Fig. 1). So we refer to these operations as either circuit layers or (n-qubit) native gates.

Just as CRB uses sequences of random Clifford gates, DRB uses sequences of random circuit layers. But, whereas the Clifford gates in CRB are supposed to be uniformly random, DRB allows the circuit layers to be sampled according to a user-specified probability distribution Ω. Many distributions are permissible but, to ensure reliability, Ω must have support on a subset of the gates that generates $\mathbb{C}_n$, and Ω random circuits must quickly spread errors (see later in this Letter).

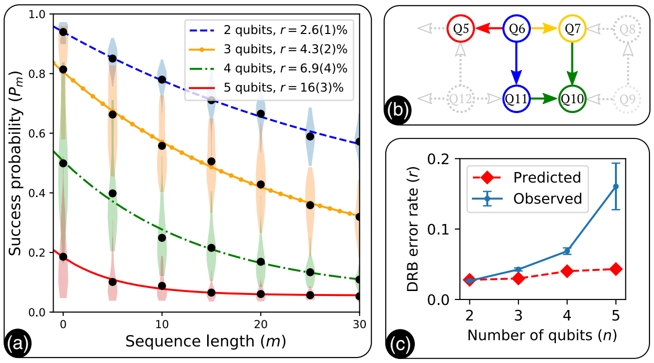
The n-qubit DRB protocol is defined as follows (note that all operations are assumed to be imperfect): (1) For a range of lengths $m \geq 0$, repeat the following $k_m \gg 1$ times: (1.1) Sample a uniformly random n-qubit stabilizer state $|\psi\rangle$. (1.2) Sample an m-layer circuit $\mathcal{U}_m$, where each layer is drawn independently from some user-specified distribution Ω over all n-qubit native gates. (1.3) Repeat the following $N \geq 1$ times: (1.3.1) Initialize the qubits in $|0\rangle^{\otimes n}$. (1.3.2) Implement a circuit to map $|0\rangle^{\otimes n} \to |\psi\rangle$. (1.3.3) Implement the sampled circuit $\mathcal{U}_m$. (1.3.4) Implement a circuit that maps $\mathcal{U}_m|\psi\rangle$ to a known computational basis state $|s\rangle$. (1.3.5) Measure all n qubits and record whether the outcome is s (success) or not (failure). (2) Calculate the average probability of success P_m at each length m, averaged over the k_m randomly sampled circuits and the N trials for each circuit. (3) Fit P_m to $P_m = A + Bp^m$, where A, B, and p are fit parameters. (4) The Ω-averaged DRB error rate of the native gates is $r = (4^n - 1)(1 - p)/4^n$. The n-dependent rescaling used above is different from that in common usage [23–25]. Using our convention, r corresponds to the probability of an error when the errors are stochastic (see later in this Letter). This is particularly convenient when varying n.

DRB is similar to the earliest implementations of RB. Both the one-qubit RB experiments of Knill *et al.* [23] and the three-qubit experiments of Ryan *et al.* [50] utilized random sequences of group generators, and so were specific examples of DRB without the stabilizer state preparation step and flexible sampling. These additional features, however, are essential to DRB: they make DRB provably reliable under broad conditions, and they allow us to separate the error rate into contributions from distinct sets of gates.

What DRB measures.—To interpret DRB results, it is important to understand what DRB measures. Assume that the gate errors are stochastic, which can be enforced to a good approximation by, e.g., Pauli-frame randomization [51–53] or by following each layer in DRB with a random n-qubit Pauli gate. Then, whenever Ω-random circuits quickly increase the weight of errors, r is a good estimate of the probability that an error happens on an Ω-average native gate. That is, $r \approx \epsilon_\Omega \equiv \sum_i \Omega(\mathcal{G}_i)\epsilon_i$, where ϵ_i is the probability of an error on the n-qubit native gate $\mathcal{G}_i$. Later, we will derive this relationship.

Because r depends on the sampling distribution, they should be reported together. A similar but hidden variability also exists in CRB—the CRB r depends on the Clifford gate compiler. This compiler dependence in CRB is inconvenient because the properties of multiqubit Clifford gate compilers are difficult to control. In contrast, because we directly choose Ω, we can control how often each gate appears in the random circuits to estimate the error rates of particular interest.

Experiments on two to five qubits.—To demonstrate that DRB is useful and behaves correctly on current multiqubit devices, we used it to benchmark two- to five-qubit subsets of the publicly accessible ibmqx5 [1,2]. The ibmqx5 native gates comprise CNOTs and arbitrary one-qubit gates [2,54]; we benchmarked a set of n-qubit gates consisting of parallel



Number of qubits in circuits	2	3	4	5
Qubits in circuits	Q6+Q11	+Q7	+Q10	+Q5
Mean CNOT ϵ estimated via n-qubit DRB (%)	3.4(1)	5.4(4)	8.2(9)	20(6)
Mean CNOT ϵ estimated via 2-qubit CRB (%)	3.6	3.4	4.4	4.5

FIG. 2. Experimental two- to five-qubit DRB on ibmqx5. (a) Success probability decays. The points are average success probabilities P_m, and the violin plots show the distributions of the success probabilities at each length over circuits (there are 28 circuits per length). The curves are obtained from fitting to $P_m = A + Bp^m$ and $r = (4^n - 1)(1 - p)/4^n$. (b) Schematic of ibmqx5. Colors match those in Fig. 2(a) and correspond to the additional qubits (and CNOTs) added from $n \to n+1$-qubit DRB. (c) Observed r vs n, and predictions from one- and two-qubit CRB calibration data. (d) Estimates of the average error rate of the CNOT gates in n-qubit circuits, obtained by comparing data in Fig. 2(a) with additional DRB data that used circuits with fewer CNOTs per layer.

applications of all directly available CNOTs and all one-qubit Clifford gates.

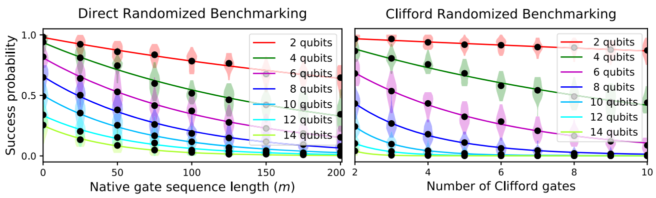
Figure 2 summarizes our results. Figure 2(a) demonstrates that DRB was successful on two to five qubits: an exponential decay is observed, and r is estimated with reasonable precision (bootstrapped 2σ uncertainties are shown). To our knowledge, this is the largest number of qubits holistically benchmarked to date, which was made possible by the streamlined nature of DRB (see Fig. 1). To interpret these results, it is necessary to specify the circuit sampling. Each layer was sampled as follows: with probability p_{CNOT}, we uniformly choose one of the CNOTs and add it to the sampled layer; for all n or $n-2$ remaining qubits, we independently and uniformly sample a one-qubit gate and add it to the layer. For the data in Fig. 2(a), $p_{\mathrm{CNOT}} = 0.75$. We also implemented experiments with $p_{\mathrm{CNOT}} = 0.25$; see the Supplemental Material [55] for these data and further experimental details.

Using this sampling, the average number of CNOTs per layer is p_{CNOT}, which is independent of n. Therefore, r will vary little with n if CNOT errors dominate, the error rates are reasonably uniform over the CNOTs, and n-qubit benchmarks are predictive of benchmarks on more than n qubits. Instead, the observed r increases quickly with n. This is quantified in Fig. 2(c), where we compare each observed r to a prediction r_{cal} obtained from the ibmqx5 CRB calibration data (one-qubit error rates from simultaneous one-qubit CRB [2,56,60] and CNOT error rates from CRB on isolated pairs [56]). These predictions are calculated both by using $r \approx \epsilon_\Omega$ and via a DRB simulation using a crosstalk-free error model that is consistent with the calibration data. Both methods agree, confirming that the increase in r with n is not due to a failure of DRB. For $n = 2$, r_{cal} and r are similar, demonstrating that n-qubit DRB and CRB are consistent. But, as n increases, r diverges from r_{cal}. This shows that the effective error rates of the one-qubit and/or two-qubit gates in the device change as we implement circuits over more qubits, demonstrating that DRB on more than two qubits can detect errors that are not predicted by one- and two-qubit CRB (calibration data) or two-qubit DRB (our data). This highlights the value of holistic benchmarking for multiqubit devices.

Using the data from Fig. 2(a) ($p_{\mathrm{CNOT}} = 0.75$) alongside additional data with $p_{\mathrm{CNOT}} = 0.25$ sampling [61], we can estimate the average error rate of the CNOT gates in n-qubit circuits. For each n, and using $r \approx \sum_i \Omega(\mathcal{G}_i)\epsilon_i$, we have $\vec{r} \approx M\vec{\epsilon}$, where $\vec{r} = (r_{0.75}, r_{0.25})$ with $r_{0.75}$ (respectively, $r_{0.25}$) denoting the r obtained with $p_{\mathrm{CNOT}} = 0.75$ (respectively, $p_{\mathrm{CNOT}} = 0.25$) sampling; $\vec{\epsilon} = (\epsilon_A, \epsilon_B)$ with ϵ_A (respectively, ϵ_B) denoting the average error rate of those n-qubit gates containing one CNOT in parallel with one-qubit gates on the other qubits (respectively, n parallel one-qubit gates); and $M = \frac{1}{4}\left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right)$. Therefore, ϵ_A and ϵ_B can be estimated using $\vec{\epsilon} = M^{-1}\vec{r}$, and so—by estimating the average one-qubit gate error rate from ϵ_B and removing this contribution from ϵ_A—we can estimate the mean CNOT error rate vs n. Estimates are given in Fig. 2(d). For two qubits, our estimate of the CNOT error rate is similar to the prediction from the calibration data, and so our methodology seems consistent with CRB techniques. In contrast, our results show that CNOTs perform substantially worse in circuits on more than two qubits than they do in two-qubit circuits. This is likely due to CNOT crosstalk; i.e., CNOTs affect "spectator" qubits.

DRB simulations.—We have shown that DRB works on current multiqubit devices, and so we now demonstrate with simulations that $r \approx \epsilon_\Omega \equiv \sum_i \Omega(\mathcal{G}_i)\epsilon_i$. Assume n qubits with native gates consisting of parallel CNOT, idle I, Hadamard H, phase P gates ($P|x\rangle = i^x|x\rangle$), and all-to-all connectivity. We model gate errors by assuming that, after each CNOT (respectively, one-qubit gate), the qubits involved in the gate are independently, with a probability of 0.25% (respectively, 0.05%), subject to a random σ_x, σ_y, or σ_z error. So, the CNOT error rate is $\approx$ 0.5%. We simulated DRB with a sampling distribution defined by randomly pairing up the qubits, applying a CNOT to a pair with probability $\tilde{p}_{\mathrm{CNOT}} = 0.5$, and applying uniformly random one-qubit gates (H, P, or I) to all qubits that do not have a CNOT acting on them. Figure 3 shows simulated 2- to 14-qubit DRB and CRB data. DRB has succeeded: the decay is exponential, and

$$
\begin{aligned}
r \approx \epsilon_\Omega &= 1 - (0.5 \times 0.9975^2 + 0.5 \times 0.9995^2)^{n/2} \\
&\approx n \times 0.15\%.
\end{aligned}
$$



Number of qubits	2	4	8	14
Ω-averaged error rate ϵ_Ω (%)	0.300	0.599	1.19	2.08
DRB r (%)	0.29(1)	0.59(2)	1.17(4)	2.2(2)
CRB r (%)	1.8(2)	9.9(4)	38(2)	Failed
CNOT-rescaled CRB r (%)	0.83(9)	1.02(4)	0.97(5)	Failed
Depth-rescaled CRB r (%)	0.16(2)	0.39(2)	0.65(3)	Failed

FIG. 3. Simulation of DRB and CRB for 2 to 14 qubits with a simple error model. The n-qubit DRB error rate is $r \approx n \times 0.15\%$, which is consistent with the simulated sampling-averaged native gate error rate ϵ_Ω.

In contrast, the CRB r grows rapidly with n—for only four-qubit CRB $r \approx 10\%$—and CRB fails for $n > 12$, demonstrating that DRB can be implemented on more qubits than CRB. Moreover, the CRB error rates r rescaled as $r_{\mathrm{RCRB}} = 1 - (1-r)^{1/\alpha}$ [42–44,48], with α as the average compiled Clifford circuit depth or CNOT count, are not simple functions of the native gate error rates (see Fig. 3).

This example is illustrative but simplistic. So, in the Supplemental Material [55], we present additional simulations with large, nonuniform CNOT error rates, as well as limited qubit connectivity. We also simulate the CNOT error rate estimation method used on the ibmqx5 data, validating the technique.

DRB theory.—We now provide a theory for DRB of gates with Pauli-stochastic errors. DRB circuits consist of preparing a uniformly random n-qubit stabilizer state ψ, a circuit $\mathcal{U}_m = \mathcal{G}_{s_m} \cdots \mathcal{G}_{s_1}$ with m layers $\mathcal{G}_i$ sampled according to Ω, and a stabilizer measurement projecting onto $\mathcal{U}_m|\psi\rangle$. For now, assume that the stabilizer state preparation and measurement (SSPAM) are perfect. In the stochastic error model, each time $\mathcal{U}_m$ is applied, there is some faulty implementation in which $\tilde{\mathcal{U}}_m = \mathcal{P}_{s_m}\mathcal{G}_{s_m} \cdots \mathcal{P}_{s_1}\mathcal{G}_{s_1}$, with $\mathcal{P}_{s_i}$ either some Pauli error or the identity. DRB aims to capture the rate that these $\mathcal{P}_{s_i}$ deviate from the identity. Because ψ is a stabilizer state, the measurement will register success if, and only if, one of the following holds: (S1) No errors occur in $\tilde{\mathcal{U}}_m$, i.e., all $\mathcal{P}_i = \mathbb{1}$; (S2) 2+ errors occur in $\tilde{\mathcal{U}}_m$ but, when they propagate through the circuit, they cancel, i.e., multiple $\mathcal{P}_i \neq \mathbb{1}$ but $\tilde{\mathcal{U}}_m = i^k \mathcal{U}_m$ (for some $k = 0, 1, 2, 3$); or (S3) 1+ errors occur in $\tilde{\mathcal{U}}_m$ that do not cancel, but they are nonetheless unobserved by the stabilizer measurement, i.e., $\tilde{\mathcal{U}}_m \neq i^k \mathcal{U}_m$ but $\tilde{\mathcal{U}}_m|\psi\rangle = i^l \mathcal{U}_m|\psi\rangle$.

The DRB average success probability P_m is obtained by averaging $P(\mathcal{U}_m, \psi) = |\langle\psi|\mathcal{U}_m^\dagger \tilde{\mathcal{U}}_m|\psi\rangle|^2$ over ψ, $\mathcal{U}_m$ and the possible Pauli errors. We may then write $P_m = s_1 + (1-s_1)(s_2 + (1-s_2)s_3)$, where s_1 is the probability of S1 (i.e., no errors), s_2 is the probability of S2 conditioned on 1+ errors occurring, and s_3 is the probability of S3 conditioned on 1+ errors occurring and the errors not canceling. Because ϵ_Ω is the Ω-averaged error rate per layer, $s_1 = (1-\epsilon_\Omega)^m$. A uniformly random stabilizer state ψ is an eigenstate of any Pauli error with a probability of $(2^n - 1)/(4^n - 1)$, and so $s_3 = (2^n - 1)/(4^n - 1) \approx 2^{-n}$. This is one of the motivations for the state preparation step in DRB.

In order to understand the effect of s_2 on P_m, we consider two regimes: small n ($\lesssim 3$), and not-so-small n ($\gtrsim 3$). In both regimes, we expect Pauli errors to occur, at most, once every several layers and to be low weight, with support on only a few qubits. In the not-so-small n regime, errors propagating through a sequence of one- and two-qubit gates are likely to quickly increase in weight [62–64] (due to the demands we made of Ω earlier). Subsequent errors are therefore very unlikely to cause error cancellation. If each layer is a uniformly random Clifford gate (as in uncompiled CRB), any Pauli error is randomized to one of the $4^n - 1$ possible n-qubit Pauli errors at each step. So, the probability that another error cancels with an earlier error is $\simeq 1/4^n$, implying that $s_2 \lesssim 1/4^n$. In DRB, we expect error cancellation at a rate only slightly above this. Therefore, s_2 contributes negligibly to P_m, and so $P_m \approx 2^n + (1-2^n)(1-\epsilon_\Omega)^m$. This is an exponential with a decay rate of ϵ_Ω. Verifying this error scrambling process for a given sampling distribution Ω is computationally efficient in the number of qubits. Distributions that do not scramble the errors quickly (e.g., if two-qubit gates are rare) can yield decays that are not simple exponentials. These should be avoided.

For small n, the probability of cancellation (s_2) is not negligible for any distribution. But, because n is small, we only need a few random circuit layers of Clifford-group generators to implement approximate Clifford twirling, and so P_m may be computed using the resulting effective depolarizing channel. Such channels are well known to lead to exponential decays [25]. However, s_2 (a function of m) now contributes significantly to the DRB decay constant p, and so $p \not\approx 1 - \epsilon_\Omega$. This motivates $r = (4^n - 1)(1-p)/4^n$, which removes the unwanted s_2 contribution in $1 - p$. Let each layer be followed by a depolarizing map $\mathcal{D}_\lambda$, where $\mathcal{D}_\lambda[\rho] = \lambda\rho + (1-\lambda)\mathbb{1}/2^n$. Then, $P_m = (1 - 2^{-n})\lambda^m + 2^{-n}$ but the error rate of $\mathcal{D}_\lambda$ is $\epsilon = (4^n - 1)(1-\lambda)/4^n$. Of course, in the large-n limit, $\epsilon \to 1 - \lambda$.

Above, we assumed perfect SSPAM, which is unrealistic. The SSPAM operations are almost m independent, and so errors in the SSPAM are almost entirely absorbed into A and B in $P_m = A + Bp^m$ as normal in RB [24,25]. The only m-dependent impact is from correlations between the stabilizer states that are prepared and measured—they are perfectly correlated (respectively, uncorrelated) at $m = 0$ (respectively, $m \to \infty$). This causes an inconsequentially small tendency to overestimate the gate error rate—because SSPAM contributes an error of $1 - \mathrm{avg}_i[(1-\epsilon_{i,\mathrm{SSPAM}})^2]$ at $m = 0$ but a smaller error of $1 - (\mathrm{avg}_i[1-\epsilon_{i,\mathrm{SSPAM}}])^2$ at $m \to \infty$, where $\epsilon_{i,\mathrm{SSPAM}}$ is the error in creating or measuring the ith stabilizer state.

DRB remains effective with coherent errors—with any one-qubit gates that generate the one-qubit Clifford group, independently random one-qubit gates on each qubit are sufficient to quickly twirl coherent errors to Pauli-stochastic errors, implying that errors can only coherently combine between a few layers in a DRB circuit (in contrast to the uncontrolled coherent addition within a compiled Clifford gate in CRB). But linking r to a formal notion of the gate error rate is more subtle with coherent errors, which is in direct analogy with CRB [65–67], as will be discussed in future work.

Conclusions.—Benchmarking methods for multiqubit systems are essential for assessing the performance of current and near-term quantum processors. But, currently, there are no reliable methods that can be easily and routinely applied to more than two qubits with current device performance. In this Letter, we have introduced and demonstrated direct randomized benchmarking, which is a method that streamlines the industry-standard Clifford randomized benchmarking technique [24,25] so that it can be applied to more qubits. DRB retains the core simplicity of CRB, our protocol directly measures the quantities of most interest—the error rates of the native gates in a device—and it is user configurable, allowing a variety of important error rates to be estimated. Our experimental demonstrations were on two to five qubits and, using a publicly accessible device [1,2], we set a record for the number of qubits holistically benchmarked. The tools we used are available as open-source code [68], and they support any device connectivity. So, we anticipate that 5 to 10+ qubits will soon be benchmarked with our protocol, providing important insights into state-of-the-art device performance. Finally, the techniques of DRB can also be applied to extend and improve the full suite of RB methods [26,27,60,69–77], and varied-sampling DRB provides an alternative to both interleaved CRB [77] and "interleaved DRB" for estimating individual error rates, demonstrating the broad applicability and impact of DRB.

T. J. P. thanks Scott Aaronson for helpful correspondence on how to efficiently generate random stabilizer states. K. R. thanks Jay Gambetta, Diego Moreda, and Ali Javadi for QISKit support. We acknowledge use of the IBM Q Experience for this work. This Letter describes objective technical results and analyses. Any subjective views or opinions that might be expressed in the Letter do not necessarily represent the views of the U.S. Department of Energy or the U.S. Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, which is a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under Contract No. DE-NA0003525. This research was funded, in part, by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). 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 IARPA, the ODNI, or the U.S. Government.

[1] IBM, IBM Q Experience, https://quantumexperience.ng.bluemix.net/qx/devices (retrieved 2018).

[2] B. Abdo *et al.*, IBM Q Experience backend information, https://github.com/QISKit/ibmqx-backend-information (retrieved 2018).

[3] N. M. Linke, D. Maslov, M. Roetteler, S. Debnath, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe, Experimental comparison of two quantum computing architectures, Proc. Natl. Acad. Sci. U.S.A. **114**, 3305 (2017).

[4] C. Figgatt, D. Maslov, K. A. Landsman, N. M. Linke, S. Debnath, and C. Monroe, Complete 3-qubit grover search on a programmable quantum computer, Nat. Commun. **8**, 1918 (2017).

[5] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. Schuyler Fried, S. Hong *et al.*, Unsupervised machine learning on a hybrid quantum computer, arXiv:1712.05771.

[6] N. Friis, O. Marty, C. Maier, C. Hempel, M. Holzäpfel, P. Jurcevic, M. B. Plenio, M. Huber, C. Roos, R. Blatt *et al.*, Observation of Entangled States of a Fully Controlled 20-Qubit System, Phys. Rev. X **8**, 021012 (2018).

[7] C. A. Riofrío, D. Gross, S. T. Flammia, T. Monz, D. Nigg, R. Blatt, and J. Eisert, Experimental quantum compressed sensing for a seven-qubit system, Nat. Commun. **8**, 15305 (2017).

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

[9] C. Neill, P. Roushan, K. Kechedzhi, S. Boixo, S. V. Isakov, V. Smelyanskiy, A. Megrant, B. Chiaro, A. Dunsworth, K. Arya *et al.*, A blueprint for demonstrating quantum supremacy with superconducting qubits, Science **360**, 195 (2018).

[10] C. Song, K. Xu, W. Liu, C.-p. Yang, S.-B. Zheng, H. Deng, Q. Xie, K. Huang, Q. Guo, L. Zhang *et al.*, 10-Qubit Entanglement and Parallel Logic Operations with a Superconducting Circuit, Phys. Rev. Lett. **119**, 180511 (2017).

[11] X. Fu, M. A. Rol, C. C. Bultink, J. van Someren, N. Khammassi, I. Ashraf, R. F. L. Vermeulen, J. C. De Sterke, W. J. Vlothuizen, R. N. Schouten *et al.*, An experimental microarchitecture for a superconducting quantum processor, in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture* (ACM, Cambridge, 2017), pp. 813–825.

[12] Intel, Intel delivers 17-qubit superconducting chip with advanced packaging to qutech, https://newsroom.intel.com/news/intel-advances-quantum-neuromorphic-computing-research/ (retrieved 2018).

[13] J. Kelly, Engineering superconducting qubit arrays for quantum supremacy, Bull. Am. Phys. Soc. (2018).

[14] IBM, IBM announces advances to IBM quantum systems & ecosystem, 2017, https://www-03.ibm.com/press/us/en/pressrelease/53374.wss.
[15] Intel, 2018 CES: Intel advances quantum and neuromorphic computing research, 2018, https://newsroom.intel.com/news/intel-advances-quantum-neuromorphic-computing-research/.
[16] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Self-consistent quantum process tomography, Phys. Rev. A **87**, 062119 (2013).
[17] R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz, Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography, Nat. Commun. **8**, 14485 (2017).
[18] D. Greenbaum, Introduction to quantum gate set tomography, arXiv:1509.02921.
[19] S. Kimmel, G. H. Low, and T. J. Yoder, Robust calibration of a universal single-qubit gate set via robust phase estimation, Phys. Rev. A **92**, 062315 (2015).
[20] K. Rudinger, S. Kimmel, D. Lobser, and P. Maunz, Experimental Demonstration of Cheap and Accurate Phase Estimation, Phys. Rev. Lett. **118**, 190502 (2017).
[21] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B Quantum Semiclass. Opt. **7**, S347 (2005).
[22] J. Emerson, M. Silva, O. Moussa, C. Ryan, M. Laforest, J. Baugh, D. G. Cory, and R. Laflamme, Symmetrized characterization of noisy quantum processes, Science **317**, 1893 (2007).
[23] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).
[24] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and Robust Randomized Benchmarking of Quantum Processes, Phys. Rev. Lett. **106**, 180504 (2011).
[25] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A **85**, 042311 (2012).
[26] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, Characterizing universal gate sets via dihedral benchmarking, Phys. Rev. A **92**, 060302(R) (2015).
[27] A. W. Cross, E. Magesan, L. S. Bishop, J. A. Smolin, and J. M. Gambetta, Scalable randomised benchmarking of non-clifford gates, npj Quantum Inf. **2**, 16012 (2016).
[28] W. G. Brown and B. Eastin, Randomized benchmarking with restricted gate sets, Phys. Rev. A **97**, 062323 (2018).
[29] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, Real randomized benchmarking, Quantum **2**, 85 (2018).
[30] J. Helsen, J. J. Wallman, S. T. Flammia, and S. Wehner, Multi-qubit randomized benchmarking using few samples, arXiv:1701.04299.
[31] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).
[32] K. N. Patel, I. L. Markov, and J. P. Hayes, Efficient synthesis of linear reversible circuits, Quantum Inf. Comput. **8**, 282 (2008).
[33] J. Yoneda, K. Takeda, T. Otsuka, T. Nakajima, M. R. Delbecq, G. Allison, T. Honda, T. Kodera, S. Oda, Y. Hoshi *et al.*, A quantum-dot spin qubit with coherence limited by charge noise and fidelity higher than 99.9%, Nat. Nanotechnol. **13**, 102 (2018).
[34] D. M. Zajac, A. J. Sigillito, M. Russ, F. Borjans, J. M. Taylor, G. Burkard, and J. R. Petta, Resonantly driven CNOT gate for electron spins, Science **359**, 439 (2018).
[35] T. F. Watson, S. G. J. Philips, E. Kawakami, D. R. Ward, P. Scarlino, M. Veldhorst, D. E. Savage, M. G. Lagally, M. Friesen, S. N. Coppersmith *et al.*, A programmable two-qubit quantum processor in silicon, Nature (London) **555**, 633 (2018).
[36] J. M. Nichol, L. A. Orona, S. P. Harvey, S. Fallahi, G. C. Gardner, M. J. Manfra, and A. Yacoby, High-fidelity entangling gate for double-quantum-dot spin qubits, npj Quantum Inf. **3**, 3 (2017).
[37] M. Veldhorst, J. C. C. Hwang, C. H. Yang, A. W. Leenstra, B. De Ronde, J. P. Dehollain, J. T. Muhonen, F. E. Hudson, K. M. Itoh, A. Morello *et al.*, An addressable quantum dot qubit with fault-tolerant control-fidelity, Nat. Nanotechnol. **9**, 981 (2014).
[38] A. D. Córcoles, J. M. Gambetta, J. M. Chow, J. A. Smolin, M. Ware, J. Strand, B. L. T. Plourde, and M. Steffen, Process verification of two-qubit quantum gates by randomized benchmarking, Phys. Rev. A **87**, 030301(R) (2013).
[39] T. Xia, M. Lichtman, K. Maller, A. W. Carr, M. J. Piotrowicz, L. Isenhower, and M. Saffman, Randomized Benchmarking of Single-Qubit Gates in a 2D Array of Neutral-Atom Qubits, Phys. Rev. Lett. **114**, 100503 (2015).
[40] A. D. Córcoles, E. Magesan, S. J. Srinivasan, A. W. Cross, M. Steffen, J. M. Gambetta, and J. M. Chow, Demonstration of a quantum error detection code using a square lattice of four superconducting qubits, Nat. Commun. **6**, 6979 (2015).
[41] Z. Chen, J. Kelly, C. Quintana, R. Barends, B. Campbell, Yu Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Lucero *et al.*, Measuring and Suppressing Quantum State Leakage in a Superconducting Qubit, Phys. Rev. Lett. **116**, 020501 (2016).
[42] J. T. Muhonen, A. Laucht, S. Simmons, J. P. Dehollain, R. Kalra, F. E. Hudson, S. Freer, K. M. Itoh, D. N. Jamieson, J. C. McCallum *et al.*, Quantifying the quantum gate fidelity of single-atom spin qubits in silicon by randomized benchmarking, J. Phys. Condens. Matter **27**, 154205 (2015).
[43] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, Nature (London) **508**, 500 (2014).
[44] J. Raftery, A. Vrajitoarea, G. Zhang, Z. Leng, S. J. Srinivasan, and A. A. Houck, Direct digital synthesis of microwave waveforms for quantum computing, arXiv:1703.00942.
[45] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno *et al.*, Restless tuneup of high-fidelity qubit gates, Phys. Rev. Applied **7**, 041001 (2017).
[46] J. Kelly, R. Barends, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, I.-C. Hoi, E. Jeffrey *et al.*, Optimal Quantum Control Using Randomized Benchmarking, Phys. Rev. Lett. **112**, 240504 (2014).

[47] D. C. McKay, S. Sheldon, J. A. Smolin, J. M. Chow, and J. M. Gambetta, Three Qubit Randomized Benchmarking, Phys. Rev. Lett. **122**, 200502 (2019).

[48] J. M. Epstein, A. W. Cross, E. Magesan, and J. M. Gambetta, Investigating the limits of randomized benchmarking protocols, Phys. Rev. A **89**, 062321 (2014).

[49] Extensions to other group 2-designs are possible.

[50] C. A. Ryan, M. Laforest, and R. Laflamme, Randomized benchmarking of single- and multi-qubit control in liquid-state NMR quantum information processing, New J. Phys. **11**, 013034 (2009).

[51] E. Knill, Quantum computing with realistically noisy devices, Nature (London) **434**, 39 (2005).

[52] M. Ware, G. Ribeill, D. Riste, C. A. Ryan, B. Johnson, and M. P. da Silva, Experimental demonstration of pauli-frame randomization on a superconducting qubit, arXiv:1803.01818.

[53] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, Phys. Rev. A **94**, 052325 (2016).

[54] Qiskit, https://qisqkit.org.

[55] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.123.030503 for additional data and details, which include Refs. [56–59].

[56] IBM Q Experience (private communication).

[57] J. Dehaene and B. De Moor, Clifford group, stabilizer states, and linear and quadratic operations over GF(2), Phys. Rev. A **68**, 042318 (2003).

[58] E. Hostens, J. Dehaene, and B. De Moor, Stabilizer states and clifford operations for systems of arbitrary dimensions and modular arithmetic, Phys. Rev. A **71**, 042315 (2005).

[59] R. Koenig and J. A. Smolin, How to efficiently select an arbitrary Clifford group element, J. Math. Phys. (N.Y.) **55**, 122202 (2014).

[60] J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki *et al.*, Characterization of Addressability by Simultaneous Randomized Benchmarking, Phys. Rev. Lett. **109**, 240504 (2012).

[61] These data were obtained after the device had been recalibrated. This is not ideal, but the error rates did not change substantially; details are in the Supplemental Material [55].

[62] W. Brown and O. Fawzi, Decoupling with random quantum circuits, Commun. Math. Phys. **340**, 867 (2015).

[63] Y. Sekino and L. Susskind, Fast scramblers, J. High Energy Phys. 10 (2008) 065.

[64] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, Commun. Math. Phys. **346**, 397 (2016).

[65] T. Proctor, K. Rudinger, K. Young, M. Sarovar, and R. Blume-Kohout, What Randomized Benchmarking Actually Measures, Phys. Rev. Lett. **119**, 130502 (2017).

[66] J. J. Wallman, Randomized benchmarking with gate-dependent noise, Quantum **2**, 47 (2018).

[67] A. Carignan-Dugas, K. Boone, J. J. Wallman, and J. Emerson, From randomized benchmarking experiments to gate-set circuit fidelity: How to interpret randomized benchmarking decay parameters, New J. Phys. **20**, 092001 (2018).

[68] E. Nielsen, R. Blume-Kohout, L. Saldyt, J. Gross, T. Scholten, K. Rudinger, T. Proctor, and J. K. Gamble, PyGSTi version 0.9.7.8, (2019), https://zenodo.org/record/3262737#.XSTJbC2ZMUF, http://www.pygsti.info, and https://github.com/pyGSTio/pyGSTi.

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

[70] T. Chasseur and F. K. Wilhelm, Complete randomized benchmarking protocol accounting for leakage errors, Phys. Rev. A **92**, 042333 (2015).

[71] J. J. Wallman, M. Barnhill, and J. Emerson, Robust Characterization of Loss Rates, Phys. Rev. Lett. **115**, 060501 (2015).

[72] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, Estimating the coherence of noise, New J. Phys. **17**, 113020 (2015).

[73] G. Feng, J. J. Wallman, B. Buonacorsi, F. H. Cho, D. K. Park, T. Xin, D. Lu, J. Baugh, and R. Laflamme, Estimating the Coherence of Noise in Quantum Control of a Solid-State Qubit, Phys. Rev. Lett. **117**, 260501 (2016).

[74] S. Sheldon, L. S. Bishop, E. Magesan, S. Filipp, J. M. Chow, and J. M. Gambetta, Characterizing errors on qubit operations via iterative randomized benchmarking, Phys. Rev. A **93**, 012301 (2016).

[75] R. Harper and S. T. Flammia, Estimating the fidelity of T gates using standard interleaved randomized benchmarking, Quantum Sci. Technol. **2**, 015008 (2017).

[76] T. Chasseur, D. M. Reich, C. P. Koch, and F. K. Wilhelm, Hybrid benchmarking of arbitrary quantum gates, Phys. Rev. A **95**, 062335 (2017).

[77] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki *et al.*, Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking, Phys. Rev. Lett. **109**, 080505 (2012).