Three-Qubit Randomized Benchmarking

David C. McKay,* Sarah Sheldon, John A. Smolin, Jerry M. Chow, and Jay M. Gambetta
IBM T.J. Watson Research Center, Yorktown Heights, New York 10598, USA

(Received 20 December 2017; published 23 May 2019)

As quantum circuits increase in size, it is critical to establish scalable multiqubit fidelity metrics. Here we investigate, for the first time, three-qubit randomized benchmarking (RB) on a quantum device consisting of three fixed-frequency transmon qubits with pairwise microwave-activated interactions (cross-resonance). We measure a three-qubit error per Clifford of 0.106 for all-to-all gate connectivity and 0.207 for linear gate connectivity. Furthermore, by introducing mixed dimensionality simultaneous RB—simultaneous one- and two-qubit RB—we show that the three-qubit errors can be predicted from the one- and two-qubit errors. However, by introducing certain coherent errors to the gates, we can increase the three-qubit error to 0.302, an increase that is not predicted by a proportionate increase in the one- and two-qubit errors from simultaneous RB. This demonstrates the importance of multiqubit metrics, such as three-qubit RB, on evaluating overall device performance.

DOI: 10.1103/PhysRevLett.122.200502

As quantum circuits increase in size, the problem of characterization becomes more acute. Exponential growth of the state space with the number of qubits means that tomographic methods for reconstructing the system will require exponential resources. Indeed, the number of required measurements for quantum process tomography scales as 16^n [1], where n is the number of qubits. To avoid scaling issues, methods have focused on characterizing the primitive set of gates used to construct the universal gate set. At minimum, for n qubits, this set contains several one-qubit gates for all n qubits and $n-1$ two-qubit gates [2]. But how good is the assumption that multiqubit algorithmic fidelities will be predicted by these primitive gate fidelities? There are strong indications that this assumption fails due to cross talk and addressability errors. For example, surface code algorithms require constructing local five-qubit gates via sequential application of two-qubit CNOT gates in parallel across a multiqubit circuit. Surface codes are predicted to have a high threshold for correcting errors, but they are typically simulated with correlated noise only between qubits for which there is a direct gate [3]. In a recent five-qubit test of a logical qubit, the fidelity was greatly improved by compensating for ZZ terms to spectator (i.e., nonparticipating neighboring) qubits during the two-qubit gate [4]. In addition, several studies have observed that algorithmic and primitive gate fidelity do not always agree. For example, when four algorithms were run on two different five-qubit processors, there was no definitive agreement from primitive to algorithmic fidelity [5]. In a five-qubit device with measured two-qubit gate fidelities of 0.99, the state fidelity of a five-qubit Greenberger-Horne-Zeilinger state was 0.82 after applying four two-qubit gates [6]. Therefore, to predict the true algorithmic fidelity, we need to measure multiqubit fidelity metrics.

Fortunately, the issue of scaling can be circumvented if the goal is to characterize a process based on a few measures, e.g., average gate fidelity. Based on this idea, there have been several proposed techniques such as Monte Carlo sampling [7,8], compressed sensing [9], matrix product state tomography [10], and twirling protocols [11] which have been applied in a variety of multiqubit systems such as photons [12], NMR [13], and trapped ions [14]. Furthermore, the fidelity of certain multiqubit entangled states can be efficiently measured, as was demonstrated for 10- [15] and 12-qubit states [16]. However, a common drawback to these techniques is that the result is sensitive to preparation and measurement errors (sometimes exponentially so) and/or does not fully characterize the underlying gates. These problems are addressed by randomized benchmarking (RB) [17,18], where sequences of random Clifford gates equaling the identity operator are applied to a set of qubits. The decay of qubit polarization versus the sequence length measures the average fidelity of the Clifford set independent of preparation and measurement errors. RB is a method widely used to characterize gates in superconducting circuits [6,19–21], ion traps [17,22–24], neutral-atom traps [25], and NMR systems [26] and for solid-state spin qubits [27]. Extensions to RB have been proposed and implemented to measure specific gate errors via interleaving [28], purity [29,30], and leakage [31,32].

RB is designed to address fidelities in multiqubit systems in two ways. For one, RB can be performed by constructing sequences from the n-qubit Clifford group. Additionally, the n-qubit space can be subdivided into sets of qubits $\{n_i\}$ and n_i-qubit RB performed in each subset simultaneously [33]. Both methods give metrics of fidelity in the n-qubit space. Despite the availability of these two methods, there has been no demonstration of RB with $n > 2$, since it is viewed as sufficient to characterize only the primitive gate set. Here we show, for the first time, a variety of three-qubit RB combinations in a three-qubit fixed-frequency superconducting device. For all-to-all gate connectivity, we measure a three-qubit error per Clifford ($3Q$ EPC) of 0.106, which is well predicted by the primitive gate errors from simultaneous RB. However, we find a strong dependence on whether we perform gate calibrations collectively or individually; the error increases to 0.302 when gates are calibrated individually. Importantly, this increase in error is not predicted by a commensurate increase in the primitive gate errors as measured from simultaneous RB. The importance of collective gate calibrations was also highlighted by the recent 12-qubit cluster state work of Ref. [16]. We also show the importance of connectivity in devices as the $3Q$ EPC increases to 0.207 when we limit the device to have linear gate connectivity.

Before describing our experiment in detail, we first provide a brief summary of the RB method; a detailed discussion of RB can be found in Ref. [34]. The idea is to construct an m-length sequence of random n-qubit Clifford gates $\prod_i^{m-1}\{C_{n,i}\} = \tilde{C}_{n,m-1}$ which is appended by the inverse of the sequence $\tilde{C}_{n,m-1}^{-1}$. Such an inverse is efficiently calculated by the Gottesman-Knill theorem [35]. Starting in the state $|0\rangle^{\otimes n}$ and applying the full sequence of Clifford gates, we then measure the population in $|0\rangle$ of each qubit. This procedure is repeated l times for different random sequences, which, in the limit of large l, twirls the error map to a depolarizing error map $\Lambda[\rho] = \alpha\rho + (1-\alpha)\mathcal{I}/d$, where $p = 1-\alpha$ is the depolarizing probability. The population in $|0\rangle$ versus the sequence length fits to an exponential decay $A\alpha^m + B$ and the average error over the Clifford gates is

$$\mathrm{EPC} = \frac{2^n - 1}{2^n}(1-\alpha) \tag{1}$$

(for a wide variety of noise models [36–38]). State preparation and measurement errors do not affect the decay constant. The number of gates in the Clifford group grows superexponentially—there are 24 one-qubit gates, 11 520 two-qubit gates, and 92 897 280 three-qubit gates [39]. However, the method requires only fair sampling from this set. Each gate is constructed from a set of primitive gates, and the exact number of $1Q$ and $2Q$ gates required depends on the basis used. In this work, our $2Q$ gate is a controlled NOT (CNOT_{ij}), where i is the control and j is the target. We generate our $1Q$ and $2Q$ Clifford gates using the set of $1Q$ gates $\{I, X_{\pi/2}, X_{-\pi/2}, Y_{\pi/2}, Y_{-\pi/2}\}$ where $P_\theta = e^{-i\theta/2\hat{P}}$. With this gate set, there are 2.2083 $1Q$ primitive gates per $1Q$ Clifford and 1.5 CNOT gates and 12.2167 $1Q$ gates per $2Q$ Clifford. To generate the $3Q$ Cliffords, we use the set of $1Q$ gates $\{X_{\pi/2}, X_{-\pi/2}, Y_{-\pi/2}\}$ plus arbitrary Z rotations, which are software defined [19]; this is the set used by the Qiskit compiler [40]. For all-to-all connectivity, there are 3.5 CNOT gates and 11.6 $1Q$ gates (counting only X and Y). We use the Qiskit compiler to change the connectivity by removing one of the CNOT gates, which results in an average of 7.7 CNOT gates and 18.4 $1Q$ gates per $3Q$ Clifford. Sample $1Q$, $2Q$, and $3Q$ Cliffords are shown in Fig. 1.

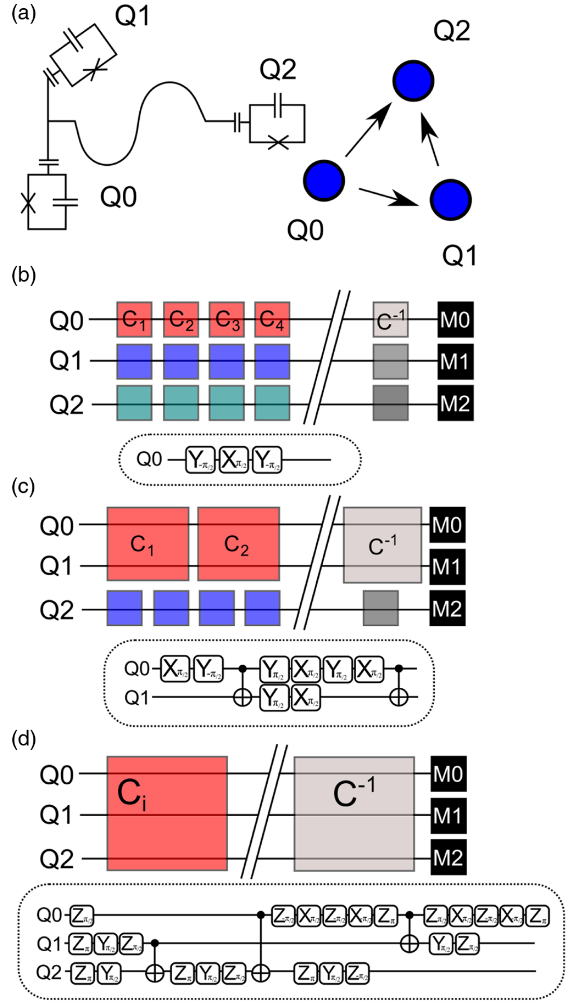

FIG. 1. (a) Schematic of the experimental setup and connectivity of the CNOT $2Q$ gates (control → target). (b) $1Q$ simultaneous RB $\{[0],[1],[2]\}$, (c) $2Q$-$1Q$ simultaneous RB $\{[0,1],[2]\}$, and (c) $3Q$ RB $\{[0,1,2]\}$. Under each is a sample (b) $1Q$, (c) $2Q$, and (d) $3Q$ Clifford gate.

In the case of multiqubit systems, RB may be performed on the full n-qubits (as detailed above) or on subsets of the system. For example, it is common to perform $2Q$ RB on the subset of two-qubits defining a CNOT gate while the other qubits are quiescent. As explained in Ref. [33], these RB data will not necessarily decay exponentially, because the other qubit subspaces are not twirled. Subsets are more rigorously characterized by simultaneous RB, which also measures some level of cross talk error since all qubits are active. Herein, we will use the notation $\{[i,j],\ldots,[k]\}$ to denote benchmarking where the mth set of n_m qubits is performing independent n_m-qubit RB. For example, $\{[0],[1,2]\}$ would indicate $1Q$ RB on qubit 0 and $2Q$

RB on qubits 1 and 2. The different combinations for three-qubits are shown in Fig. 1.

To test $3Q$ RB, we use a device comprised of three fixed-frequency superconducting transmon qubits ($Q0$, $Q1$, $Q2$) of frequencies (5.353,5.291,5.237) GHz coupled to a common 6.17 GHz bus resonator. Our $1Q$ gates are 44.8-ns-wide DRAG-shaped microwave pulses [41]. Our $2Q$ gates are Gaussian smoothed square microwave pulses applied to a qubit (the control) at the frequency of one of the other qubits (the target). This activates a cross-resonance interaction, which can be tuned to build a composite pulse CNOT gate of 240 ns; details are found in Ref. [42]. A schematic of the device and CNOT connectivity is shown in Fig. 1. More device details are given in Ref. [43].

For our three-qubit system, we consider eight possible RB combinations: simultaneous $1Q$ RB ($\{[0],[1],[2]\}$), separate $2Q$ RB ($\{[0,1]\}$,$\{[0,2]\}$,$\{[1,2]\}$), simultaneous $2Q$ RB and $1Q$ RB ($2Q-1Q$ RB) ($\{[0,1],[2]\}$, $\{[0,2],[1]\}$,$\{[1,2],[0]\}$), and, finally, $3Q$ RB ($\{[0,1,2]\}$). For each combination, we perform $l = 30$ averages (except for separate $2Q$ RB, where $l = 20$). For simultaneous RB, we attempt to match the sequence lengths on the different subsystems, so we use a ratio of 9∶1 $1Q$:$2Q$ Clifford gates for $2Q$-$1Q$ simultaneous RB. We perform these RB sequences under two different calibration procedures. In procedure A, we calibrate the $1Q$ gate parameters simultaneously, e.g., qubit frequency, pulse amplitude, and drag amplitude. In procedure B, we calibrate the $1Q$ gate parameters individually. In both cases, we calibrate the $2Q$ gates separately. To give a sense of the types of curves produced from $1Q$, $2Q$, and $3Q$ RB, a subset of the data from calibration A is shown in Fig. 2. The errors from the full RB set and for both calibrations are summarized in Table I.

The data from Table I demonstrate that $2Q$ gate errors from $2Q$-$1Q$ RB are worse, consistent with increased cross talk. There is one exception, CNOT_{12}, for calibration A, which decreases from 2.8×10^{-2} to 1.74×10^{-2}. This highlights the difference between the calibration procedures, mainly that they result in different calibrated values for the qubit frequency. The qubit frequencies in calibration A are shifted by the average ZZ interaction between pairs ($ZZ_{01} = 20$ kHz, $ZZ_{02} = 352$ kHz, and $ZZ_{12} = 114$ kHz). Since the ZZ_{02} shift is calibrated into the frequency of $Q2$ for calibration A, there is a Z error when benchmarking CNOT_{12} if $Q0$ is in the ground state; the opposite is true for calibration B and so the stand-alone CNOT_{12} RB error is very low (0.92×10^{-2}). Although there is only a subtle difference between the calibration procedures, there is a large difference between the $3Q$ RB errors, illustrating how $3Q$ RB can be a sensitive probe of such calibration procedures on algorithmic fidelity. Overall, calibrating the average ZZ into the qubit frequencies maximizes $3Q$ fidelity. The data in Table I also show the importance of connectivity, as omitting one of the CNOTs causes the algorithmic error to increase appreciably.

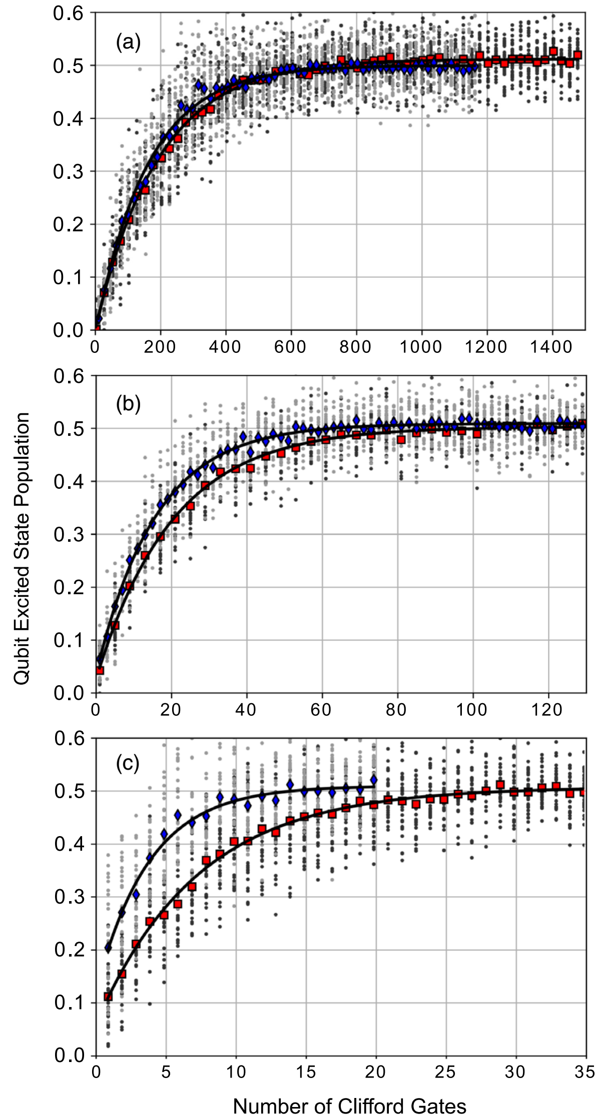

FIG. 2. Qubit 0 experimental data from different RB sequences for calibration A. Black lines are exponential fits to the data, and the gray points are from the individual trials. Red squares (blue diamonds) are the averages over these trials for the light gray (dark gray) points. (a) $1Q$ RB from simultaneous $1Q$ (red squares) and $2Q$-$1Q$ RB (blue diamonds). (b) $2Q$ RB for the 01 pair performed in isolation (red square) and simultaneously with $1Q$ RB on Q2 (blue diamonds). (c) $3Q$ RB for all-to-all connectivity (red squares) and for limited (no CNOT_{12}) connectivity (blue squares). The decay parameters from these fits are given in Ref. [43].

One of the main questions about $3Q$ RB is how much new information does it convey; i.e., can $3Q$ errors be predicted from the $1Q$ and $2Q$ errors (more specifically, the $1Q$ and $2Q$ depolarizing rates)? To answer this question, we calculate the predicted $3Q$ decay parameter α [converting to EPC using Eq. (1)]:

TABLE I. EPG (error per gate) and EPC (error per Clifford) from different RB experiments in $[Q0, Q1, Q2]$ order for $1Q$ (one-qubit) EPG and in order $[\mathrm{CNOT}_{01}, \mathrm{CNOT}_{02}, \mathrm{CNOT}_{12}]$ for the $2Q$ (two-qubit) EPG. $1Q$ EPG is the error per gate averaged over the set indicated in the main text. $2Q$ EPG is calculated from the $2Q$ EPC assuming the $1Q$ EPG from $\{[0],[1],[2]\}$ benchmarking (see [43] for details of this calculation). $3Q$ EPC omitting CNOT_{12} for calibration B was not measurable, because the error was too high to properly fit the data. The coherence-limited errors are calculated assuming only errors from T_1 and T_2. Variability in T_1 and T_2 between the calibrations is due to drift over the approximately 3 days between experiments. Errors reflect the uncertainty in the fit parameters.

	Calibration A	Calibration B
T_1	$[29, 50, 39]$ μs	$[42, 47, 35]$ μs
T_2	$[39, 75, 59]$ μs	$[61, 74, 46]$ μs
$1Q$ EPG coherence limit	$[6.5, 3.5, 4.4] \times 10^{-4}$	$[4.2, 3.6, 5.4] \times 10^{-4}$
$1Q$ EPG from $\{[0],[1],[2]\}$ RB	$[1.12(2), 0.86(1), 1.22(2)] \times 10^{-3}$	$[1.40(5), 0.81(1), 1.66(4)] \times 10^{-3}$
$1Q$ EPG from $\{[i],[j,k]\}$ RB	$[1.41(3), 0.95(2), 1.35(2)] \times 10^{-3}$	$[1.68(4), 0.95(2), 1.54(3)] \times 10^{-3}$
$2Q$ EPG coherence limit	$[6, 7, 5] \times 10^{-3}$	$[5, 6, 6] \times 10^{-3}$
$2Q$ EPG from $\{[i,j]\}$ RB	$[1.26(7), 1.15(8), 2.8(2)] \times 10^{-2}$	$[0.86(5), 2.8(1), 0.92(7)] \times 10^{-2}$
$2Q$ EPG from $\{[i,j],[k]\}$ RB	$[1.89(6), 1.62(6), 1.74(7)] \times 10^{-2}$	$[2.45(8), 4.2(2), 4.3(2)] \times 10^{-2}$
$3Q$ EPC from $\{[0,1,2]\}$ RB (all to all)	0.106(2)	0.302(6)
$3Q$ EPC from $\{[0,1,2]\}$ RB (omit CNOT_{12})	0.207(3)	

$$\alpha_{3Q} = \frac{\alpha_{1Q}^{N_1/3}\alpha_{2Q}^{2N_3/3}}{7}(1 + 3\alpha_{1Q}^{N_1/3}\alpha_{2Q}^{N_2/3} + 3\alpha_{1Q}^{2N_1/3}\alpha_{2Q}^{N_2/3}) \quad (2)$$

where N_2 (N_1) is the number of $2Q$ ($1Q$) gates per $3Q$ Clifford and $p_1 = 1 - \alpha_1 (p_2 = 1 - \alpha_2)$ are the $1Q$ ($2Q$) depolarizing probabilities. For simplicity, we assume that all $1Q$ gates and $2Q$ gates have the same depolarizing probability; see [43] for the general form of Eq. (2) and details of the derivation. The values discussed previously for N_1 and N_2 did not consider the finite duration of gates. In reality, there will be idle periods on some qubits, and characterizing idle periods as one-qubit gates, $N_1 = 34.7$ ($N_1 = 67.9$) for all-to-all (limited) connectivity. This is the number used for predicting the $3Q$ EPC.

For the $1Q$ and $2Q$ depolarizing probabilities in Eq. (2), we use two sets of numbers from Table I to calculate the predicted $3Q$ EPC shown in Table II. The first set are the coherence-limited EPGs, which unsurprisingly predict a much lower than measured $3Q$ EPC, indicating that the majority of errors are due to unwanted and uncompensated terms in the Hamiltonian such as cross talk. The second set of numbers are from $2Q$-$1Q$ simultaneous RB, which should be the most accurate measure of primitive gate errors. Indeed, for calibration A, the predicted $3Q$ EPC is accurate for both all-to-all and limited connectivity. However, in the case of calibration B, there is very little agreement between the predicted and measured $3Q$ EPC, demonstrating the utility of the $3Q$ RB fidelity as a novel multiqubit metric sensitive to subtle errors that are not fully revealed by benchmarking the primitive gates. In calibration B, the uncompensated ZZ errors are amplified by the specific structure of the $3Q$ Clifford gate, since there are idle periods on the spectator qubits while the other qubits perform the $2Q$ gate [this is schematically illustrated in Fig. 1(d)]. Simulations including the measured $1Q/2Q$ errors and ZZ predict well the observed $3Q$ RB data; see Ref. [43]. Since the implementation of the $3Q$ Clifford gate is not unique, certain constructions may amplify or attenuate different error terms; investigating such constructions in detail is left for future study.

In conclusion, we demonstrate, for the first time, $3Q$ RB and subset $2Q$-$1Q$ simultaneous RB. Although there is no true primitive three-qubit gate, $3Q$ RB measures a fidelity that is not captured by the one- and two-qubit gate metrics. As systems continue to increase in size and cross talk terms dominate the error, metrics such as $3Q$ RB will play an important role in benchmarking the true algorithmic fidelity of these large systems. Although $3Q$ RB does not indicate how to correct cross talk errors, it will play an important role in validating mitigation strategies. Software and hardware methods to suppress cross talk are an active area of

TABLE II. Predicted $3Q$ EPC from $1Q$ and $2Q$ EPG numbers listed in Table I by applying Eq. (2). See the main text for a detailed discussion of the calculation.

	Calibration A		Calibration B
	All to all	Omit CNOT_{12}	All to all
$3Q$ EPC from RB	0.106(2)	0.207(3)	0.302(6)
Coherence limit	0.044	0.094	0.041
$3Q$ EPC predicted from $\{[i],[j,k]\}$ RB	0.115(4)	0.226(6)	0.187(7)

research and may require the use of active elements such as tunable couplers [45,46].

We thank Firat Solgun, Markus Brink, Sami Rosenblatt, and George Keefe for modeling and fabricating the device. We thank Lev Bishop, Andrew Cross, Easwar Magesan, and Antonio Corcoles for discussions and manuscript comments. We thank Christopher Wood and Sergey Bravyi for help generating the Clifford gates. This work was supported by the Army Research Office under Contract No. W911NF-14-1-0124.

*dcmckay@us.ibm.com

[1] I. L. Chuang and M. A. Nielsen, Prescription for experimental determination of the dynamics of a quantum black box, J. Mod. Opt. **44**, 2455 (1997).

[2] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, England, 2000).

[3] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Surface codes: Towards practical large-scale quantum computation, Phys. Rev. A **86**, 032324 (2012).

[4] M. Takita, A. W. Cross, A. D. Córcoles, J. M. Chow, and J. M. Gambetta, Experimental Demonstration of Fault-Tolerant State Preparation with Superconducting Qubits, Phys. Rev. Lett. **119**, 180501 (2017).

[5] 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).

[6] R. Barends *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, Nature (London) **508**, 500 (2014).

[7] S. T. Flammia and Y.-K. Liu, Direct Fidelity Estimation from Few Pauli Measurements, Phys. Rev. Lett. **106**, 230501 (2011).

[8] M. P. da Silva, O. Landon-Cardinal, and D. Poulin, Practical Characterization of Quantum Devices without Tomography, Phys. Rev. Lett. **107**, 210404 (2011).

[9] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Quantum State Tomography via Compressed Sensing, Phys. Rev. Lett. **105**, 150401 (2010).

[10] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Efficient quantum state tomography, Nat. Commun. **1**, 149 (2010).

[11] O. Moussa, M. P. da Silva, C. A. Ryan, and R. Laflamme, Practical Experimental Certification of Computational Quantum Gates Using a Twirling Procedure, Phys. Rev. Lett. **109**, 070504 (2012).

[12] C. Schwemmer, G. Tóth, A. Niggebaum, T. Moroder, D. Gross, O. Gühne, and H. Weinfurter, Experimental Comparison of Efficient Tomography Schemes for a Six-Qubit State, Phys. Rev. Lett. **113**, 040503 (2014).

[13] D. Lu, H. Li, D.-A. Trottier, J. Li, A. Brodutch, A. P. Krismanich, A. Ghavami, G. I. Dmitrienko, G. Long, J. Baugh, and R. Laflamme, Experimental Estimation of Average Fidelity of a Clifford Gate on a 7-Qubit Quantum Processor, Phys. Rev. Lett. **114**, 140505 (2015).

[14] B. P. Lanyon, C. Maier, M. Holzäpfel, T. Baumgratz, C. Hempel, P. Jurcevic, I. Dhand, A. S. Buyskikh, A. J. Daley, M. Cramer, M. B. Plenio, R. Blatt, and C. F. Roos, Efficient tomography of a quantum many-body system, Nat. Phys. **13**, 1158 (2017).

[15] C. Song, K. Xu, W. Liu, C.-p. Yang, S.-B. Zheng, H. Deng, Q. Xie, K. Huang, Q. Guo, L. Zhang, P. Zhang, D. Xu, D. Zheng, X. Zhu, H. Wang, Y.-A. Chen, C.-Y. Lu, S. Han, and J.-W. Pan, 10-Qubit Entanglement and Parallel Logic Operations with a Superconducting Circuit, Phys. Rev. Lett. **119**, 180511 (2017).

[16] M. Gong *et al.*, Genuine 12-Qubit Entanglement on a Superconducting Quantum Processor, Phys. Rev. Lett. **122**, 110501 (2019).

[17] 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).

[18] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and Robust Randomized Benchmarking of Quantum Processes, Phys. Rev. Lett. **106**, 180504 (2011).

[19] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, Efficient z gates for quantum computing, Phys. Rev. A **96**, 022330 (2017).

[20] J. M. Chow, J. M. Gambetta, L. Tornberg, J. Koch, L. S. Bishop, A. A. Houck, B. R. Johnson, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, Randomized Benchmarking and Process Tomography for Gate Errors in a Solid-State Qubit, Phys. Rev. Lett. **102**, 090502 (2009).

[21] 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).

[22] J. P. Gaebler, T. R. Tan, Y. Lin, Y. Wan, R. Bowler, A. C. Keith, S. Glancy, K. Coakley, E. Knill, D. Leibfried, and D. J. Wineland, High-Fidelity Universal Gate Set for $^9Be^+$ Ion Qubits, Phys. Rev. Lett. **117**, 060505 (2016).

[23] C. J. Ballance, T. P. Harty, N. M. Linke, M. A. Sepiol, and D. M. Lucas, High-Fidelity Quantum Logic Gates Using Trapped-Ion Hyperfine Qubits, Phys. Rev. Lett. **117**, 060504 (2016).

[24] J. P. Gaebler, A. M. Meier, T. R. Tan, R. Bowler, Y. Lin, D. Hanneke, J. D. Jost, J. P. Home, E. Knill, D. Leibfried, and D. J. Wineland, Randomized Benchmarking of Multiqubit Gates, Phys. Rev. Lett. **108**, 260503 (2012).

[25] S. Olmschenk, R. Chicireanu, K. D. Nelson, and J. V. Porto, Randomized benchmarking of atomic qubits in an optical lattice, New J. Phys. **12**, 113007 (2010).

[26] 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).

[27] 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, and A. S. Dzurak, An addressable

quantum dot qubit with fault-tolerant control-fidelity, Nat. Nanotechnol. **9**, 981 (2014).
[28] 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, M. B. Ketchen, and M. Steffen, Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking, Phys. Rev. Lett. **109**, 080505 (2012).
[29] D. C. McKay, S. Filipp, A. Mezzacapo, E. Magesan, J. M. Chow, and J. M. Gambetta, Universal Gate for Fixed-Frequency Qubits via a Tunable Bus, Phys. Rev. Applied **6**, 064007 (2016).
[30] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, Estimating the coherence of noise, New J. Phys. **17**, 113020 (2015).
[31] C. J. Wood and J. M. Gambetta, Quantification and characterization of leakage errors, Phys. Rev. A **97**, 032306 (2018).
[32] J. J. Wallman, M. Barnhill, and J. Emerson, Robust characterization of leakage errors, New J. Phys. **18**, 043021 (2016).
[33] 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, M. B. Ketchen, and M. Steffen, Characterization of Addressability by Simultaneous Randomized Benchmarking, Phys. Rev. Lett. **109**, 240504 (2012).
[34] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A **85**, 042311 (2012).
[35] D. Gottesman, The Heisenberg representation of quantum computers, arXiv:quant-ph/9807006.
[36] 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).
[37] J. J. Wallman, Randomized benchmarking with gate-dependent noise, Quantum **2**, 47 (2018).
[38] T. Proctor, K. Rudinger, K. Young, M. Sarovar, and R. Blume-Kohout, What Randomized Benchmarking Actually Measures, Phys. Rev. Lett. **119**, 130502 (2017).
[39] M. Ozols, Clifford group (2008), http://home.lu.lv/~sd20008/papers/essays/Clifford%20group%20[paper].pdf.
[40] Qiskit SDK, online (2017), https://qiskit.org/.
[41] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, Simple Pulses for Elimination of Leakage in Weakly Nonlinear Qubits, Phys. Rev. Lett. **103**, 110501 (2009).
[42] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, Procedure for systematically tuning up cross-talk in the cross-resonance gate, Phys. Rev. A **93**, 060302(R) (2016).
[43] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.122.200502 for further device information, formula derivations, raw data and simulations, which includes Refs. [33,44].
[44] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).
[45] G. Zhang, P. S. Mundada, and A. A. Houck, Suppression of qubit crosstalk in a tunable coupling superconducting circuit, arXiv:1810.04182.
[46] R. C. Bialczak, M. Ansmann, M. Hofheinz, M. Lenander, E. Lucero, M. Neeley, A. D. O'Connell, D. Sank, H. Wang, M. Weides, J. Wenner, T. Yamamoto, A. N. Cleland, and J. M. Martinis, Fast Tunable Coupler for Superconducting Qubits, Phys. Rev. Lett. **106**, 060501 (2011).