Experimental Pauli-frame randomization on a superconducting qubit

Matthew Ware, Guilhem Ribeill, Diego Ristè,* Colm A. Ryan,† Blake Johnson,‡ and Marcus P. da Silva§
Raytheon BBN Technologies, 10 Moulton Street, Cambridge, Massachusetts 02138, USA

(Received 17 December 2020; accepted 22 March 2021; published 8 April 2021)

The promise of quantum computing with imperfect qubits relies on the ability of a quantum computing system to scale cheaply through error correction and fault tolerance. While fault tolerance requires relatively mild assumptions about the nature of qubit errors, the overhead associated with coherent and non-Markovian errors can be orders of magnitude larger than the overhead associated with purely stochastic Markovian errors. One proposal to address this challenge is to randomize the circuits of interest, shaping the errors to be stochastic Pauli errors but leaving the aggregate computation unaffected. The randomization technique can also suppress couplings to slow degrees of freedom associated with non-Markovian evolution. Here, we demonstrate the implementation of *Pauli-frame randomization* in a superconducting circuit system, exploiting a flexible programming and control infrastructure to achieve this with low effort. We use high-accuracy gate-set tomography to characterize in detail the properties of the circuit error, with and without the randomization procedure, which allows us to make rigorous statements about Markovianity as well as the nature of the observed errors. We demonstrate that randomization suppresses signatures of non-Markovian evolution to statistically insignificant levels, from a Markovian model violation ranging from 43σ to 1987σ, down to violations between 0.3σ and 2.7σ under randomization. Moreover, we demonstrate that, under randomization, the experimental errors are well described by a Pauli error model, with model violations that are similarly insignificant (between 0.8σ and 2.7σ). Importantly, all these improvements in the model accuracy were obtained without degradation to fidelity, and with some improvements to error rates as quantified by the diamond norm. This demonstrates the ability of Pauli-frame randomization to shape noise into forms that are more benign for quantum error correction and fault tolerance.

DOI: 10.1103/PhysRevA.103.042604

I. INTRODUCTION

Large-scale quantum computation poses a number of design and control challenges. Significant efforts are in progress [1–3] to meet and overcome challenges associated with initial-state preparation, maintaining coherence, implementing universal gates, and measuring qubits reliably—all key criteria for building scalable quantum computers [4]. As the system coherence times continue to grow, coherent errors can become the dominant source of error. These errors can originate from miscalibration of qubit rotations, unintentional control frequency detunings, or interactions between systems that are otherwise assumed to be decoupled—all ubiquitous problems for experimental quantum computers. These errors are also particularly difficult to simulate in multiqubit systems, as they can interfere constructively and destructively, making predictions about the performance of quantum error correction codes and fault-tolerant computation quite difficult [5–7]. Moreover, theoretical lower bounds on the tolerable rates for coherent errors indicate they may be much more damaging than stochastic errors [8–11]. One way to address this problem is to transform coherent errors into incoherent, stochastic errors, such as random bit and phase flips. Here, we use a superconducting qubit system to implement *Pauli-frame randomization* (PFR) [12–14] and show that coherent errors can be reshaped into stochastic Pauli errors. We also discuss some additional benefits of the randomization process, such as decoupling of slow non-Markovian noise [15].

One significant challenge in determining whether PFR has indeed made coherent errors stochastic is their small magnitude. Owing to the community's progress towards fault tolerance, the magnitudes of these errors are on the order of 10^{-3} or less in state-of-the-art devices. Measuring such small errors reliably runs into limitations of various characterization approaches: Standard tomography is sensitive to preparation and measurement imperfections and has very low accuracy, while randomized benchmarking estimates a quantity (closely associated with the average fidelity [16–18]) that does not differentiate between coherent and stochastic errors, and cannot test if errors corresponds to Pauli error models or not. In this demonstration we use *gate-set tomography* (GST) [19–23], a tomographic reconstruction technique that provides (1) insensitivity to state preparation and measurement (SPAM) errors, (2) nearly quantum-limited accuracy, and (3) an open

*Present address: Quantum Engineering Solutions, Keysight Technologies, One Broadway, Cambridge, MA 02412.
†Present address: AWS Center for Quantum Computing, Pasadena, CA 91125.
‡Present address: IBM T. J. Watson Research Center, Yorktown Heights, NY 10598; blake.johnson@ibm.com
§Present address: Microsoft Quantum, One Microsoft Way, Redmond, WA 98052; marcus.silva@microsoft.com

source library for experiment design and data analysis [24]. Critically, GST also allows us to accurately quantify not only the behavior of the diamond norm error [25,26] and average infidelity [16] under randomization, but also detailed features of individual gate errors and the degree to which the evolution is well described by a Markovian, time-invariant model [27]—all of which help confirm the predicted Pauli error model behavior, despite the presence of general imperfections in the randomization operations.

The remainder of the paper is organized as follows. In Sec. II we describe PFR, and discuss how to test its implementation, in a statistically rigorous manner, in Sec. III. Section IV describes the experiments as well as the infrastructure required to create and process randomized sequences. Finally, in Sec. V we discuss the experimental results.

II. PAULI-FRAME RANDOMIZATION

Pauli-frame randomization (PFR) is a noise-shaping technique that reduces general noise to effective random Pauli errors between computational gates [12–14]. If the computational gates consist of Clifford group operations [28] (a set of operations sufficient for the most promising approaches to error correction and fault tolerance), the effect of these random Pauli operations can be easily tracked [29,30] so that the computation can be unrandomized by simply reinterpreting the measurement results. While this randomization is designed to have no impact on the ideal computation, it effectively *symmetrizes* the error, much as *twirling* [31–36] and randomized decoupling [15], leading to an effective error operation that corresponds to a mixture of Pauli group operations known as a *Pauli channel*, or a *Pauli error model*.

These results can be derived in the limit of perfect randomization operations and gate-independent errors as follows. Consider a set of ideal (noisy) [37] Clifford group quantum operations $\mathcal{C}_i$ ($\tilde{\mathcal{C}}_i$). Any sequence of ideal Clifford operations can be randomized by inserting uniformly random Pauli group operations between the Clifford group operations. Since Clifford group operations transform the Pauli group operation to other Pauli group operations, the overall effect of these random Pauli group operations can be canceled out by applying a final single Pauli group operation at the end of the sequence of gates. Moreover, since the Pauli group is a subgroup of the Clifford group, one may simply combine the ith random Pauli operation $\mathcal{P}_i$ with the ith Clifford group operation $\mathcal{C}_i$, to obtain a random Clifford group operations $\mathcal{D}_i$ [38]. In other words, a given sequence of Clifford group operations $\mathcal{C}_L\mathcal{C}_{L-1}\cdots\mathcal{C}_2\mathcal{C}_1$ becomes

$$\underbrace{\mathcal{P}_{L+1}\mathcal{C}_L\mathcal{P}_L}_{\mathcal{D}_L}\underbrace{\mathcal{C}_{L-1}\mathcal{P}_{L-1}}_{\mathcal{D}_{L-1}}\cdots\underbrace{\mathcal{C}_2\mathcal{P}_2}_{\mathcal{D}_2}\underbrace{\mathcal{C}_1\mathcal{P}_1}_{\mathcal{D}_1}, \tag{1}$$

which results in the randomized sequence of Clifford group operations $\mathcal{D}_L\mathcal{D}_{L-1}\cdots\mathcal{D}_2\mathcal{D}_1$. In essence, under PFR, a single realization of a randomized sequence of Clifford group operations simply corresponds to a different sequence of Clifford group operations.

It is possible to choose all $\mathcal{P}_i$ independently at random and compensate for their action by flipping observed measurement outcomes in postprocessing (as, by construction, we only measure in the computational basis). In order to simplify postprocessing, we instead choose $\mathcal{P}_{L+1}$ to cancel the effect that all other random Pauli group operations would have on measurement results (i.e., $\mathcal{P}_{L+1}$ is a Pauli frame correction before measurement). In this way the measurement outcome of the randomized and unrandomized experiments can be treated exactly the same, with no additional postprocessing for the randomized experiments.

We can analyze the sequences above with the simplifying assumption of gate-independent errors by replacing each operation with its noisy counterpart. We write the noisy operations $\tilde{\mathcal{D}}_i = \mathcal{E}\mathcal{D}_i$ [where $\mathcal{E}$ is an arbitrary but fixed completely-positive trace-preserving (CPTP) map] to obtain

$$\tilde{\mathcal{D}}_L\tilde{\mathcal{D}}_{L-1}\cdots\tilde{\mathcal{D}}_2\tilde{\mathcal{D}}_1 \tag{2}$$

$$= \mathcal{E}\mathcal{D}_L\mathcal{E}\mathcal{D}_{L-1}\cdots\mathcal{E}\mathcal{D}_2\mathcal{E}\mathcal{D}_1 \tag{3}$$

$$= \mathcal{E}\mathcal{P}_{L+1}\mathcal{C}_L\mathcal{P}_L\mathcal{E}\mathcal{C}_{L-1}\mathcal{P}_{L-1}\cdots\mathcal{E}\mathcal{C}_2\mathcal{P}_2\mathcal{E}\mathcal{C}_1\mathcal{P}_1. \tag{4}$$

Defining $\mathcal{P}^{\mathcal{C}} = \mathcal{C}\mathcal{P}\mathcal{C}^\dagger$, we can write $\mathcal{C}\mathcal{P} = \mathcal{P}^{\mathcal{C}}\mathcal{C}$. Similarly, we define $\mathcal{P}_{n:1} = \mathcal{P}_n\mathcal{P}^{\mathcal{C}_{n-1}}_{n-1:1}$ (with the base case $\mathcal{P}_{1:1} = \mathcal{P}_1$). With these definitions, the entire sequence can then be rewritten as $\mathcal{E}\,\mathcal{P}_{L+1:1}\,\mathcal{C}_L\,\mathcal{P}^{\mathcal{C}_{L-1}}_{L-1:1}\,\mathcal{E}\,\mathcal{P}^{\mathcal{C}_{L-1}}_{L-1:1}\,\mathcal{C}_{L-1}\cdots\mathcal{P}^{\mathcal{C}_1}_{1:1}\,\mathcal{E}\,\mathcal{P}^{\mathcal{C}_1}_{1:1}\mathcal{C}_1$, where, in the experiments described here, we have chosen $\mathcal{P}_{L+1:1}$ to be the identity. In other words, we choose $\mathcal{P}_i$ uniformly at random for $1 \leqslant i \leqslant L$, and choose $\mathcal{P}_{L+1}$ to get a trivial $\mathcal{P}_{L+1:1}$. Averaging over many uniformly random choices of Pauli operations in Eq. (4), we transform each $\mathcal{E}$ in the sequence into $\overline{\mathcal{E}} = \frac{1}{d^2}\sum_i \mathcal{P}_i\mathcal{E}\mathcal{P}_i$, which correspond to twirling $\mathcal{E}$ over the Pauli group. This, in turn, ensures that the effective error $\overline{\mathcal{E}}$ associated with each gate in the sequence corresponds to a statistical mixture of Pauli operations [34], as desired [39].

The calculation outlined above does require rather strong assumptions about the properties of the noise (i.e., that it is gate independent and Markovian), but due to similarities to randomized benchmarking (RB) [40–44], which has been shown to require weaker assumptions [18], we expect that these strong assumptions are not strictly necessary. In the remainder of this paper we focus on how to test such a hypothesis, and implement these tests on the natural imperfections of a superconducting qubit experiment.

III. HYPOTHESIS TESTING

The task of checking whether the result of applying PFR to an experiment does indeed result in a Pauli channel is subtle. Modern experiments have very high fidelity to ideal operations so checking that the unrandomized errors are not well described by the Pauli channel—i.e., determining that PFR is necessary—is already challenging, since error rates can be on the order of 10^{-3} or less. In both cases, it is natural to consider long sequences of operations to amplify sensitivity to these small errors.

We choose to use long-sequence gate-set tomography (GST) [21–23] to observe these small effects, and use a readily available open-source package for experiment design and data analysis [24], with minor modifications. At heart, GST is a sophisticated refinement of a quantum process tomography [45,46], providing a complete reconstruction of the action of quantum operations. In particular, GST is an iterative procedure that refines the tomographic reconstruction of a set of

gates by comparing predictions about long gate sequences to experimental observations, and adjusting the reconstruction for better agreement. Since long sequences allow for small perturbations to accumulate, this technique yields unparalleled accuracy [21–23,47].

Even with a reconstruction in hand, another subtle question is how to quantify the distance between reconstructed errors and a Pauli error model—i.e., the degree of "non-Pauliness" of the noise. We use the likelihood ratio test for this purpose [48,49], which requires a hierarchy of nested models. The null hypothesis H_0 is taken to be that the statistics for each sequence in the GST experiments leads to a separate binomial distribution of outcomes. More explicitly, for the null hypothesis we only assume that the sequences correspond to reproducible experiments with well-defined measurement statistics, and ignore the gate structure of the sequences. This corresponds to not making any assumption about Markovianity or time independence of the system evolution. We then consider two hypotheses nested within H_0: that each gate in the sequence corresponds to a fixed linear operation acting on the system (we call this first hypothesis H_1), and that each gate in the sequence corresponds to a fixed Clifford group operation followed by a fixed Pauli stochastic error operation (we call this second hypothesis H_2). The consistency of H_0 with H_2, for some reconstructed Pauli error model, will be taken as our measure of Pauliness.

As we indicated, these hypotheses are nested: H_2 is a special case of H_1, and H_1 is a special case of H_0, meaning that if the statistical tests indicated H_2 is consistent with H_0, the same will be true of H_1. The statistical tests will only be able to test if the proposed hypotheses are consistent with H_0, in the sense that a hypothesis cannot have a higher likelihood than another hypothesis it is nested into.

We fit data to a model under H_0 by the maximum-likelihood estimation of the binomial distribution parameter p associated with each GST sequence. We fit data to a model under H_1 using a progressive refinement of the maximum-likelihood estimation, a heuristic developed for GST [24]. We fit data to a model under H_2 by projecting the fit of H_1 into a generalized monomial matrix (described below), determined by the corresponding noiseless Clifford group operation. The first two fits are part of the standard routines within GST, while the last fit is a small extension to the existing GST routines.

The fitting of data to a model under H_2 proceeds as follows. In the Pauli-Liouville representation [50–56], a Clifford group operation is a monomial matrix—each row or column has a single nonzero matrix element, and this matrix element is ± 1. In the presence of a Pauli error model, a noisy Clifford group operation will be a generalized monomial matrix, where the ± 1 elements of the noiseless matrix are replaced by numbers in the interval $[-1, 1]$ (but the 0 matrix elements remain unchanged). Collectively, these matrix elements must live in a simplex equivalent to the probability simplex for the Pauli channel [35]. Thus, the projection of an H_1 model onto an H_2 model simply corresponds to identifying which matrix elements should be set to zero (i.e., which matrix elements are zero in the ideal gate), and then adjusting the remaining nonzero matrix elements so that the resulting matrix lies in the appropriate simplex.

A. Badness of fit

We quantify how well the data are explained with each of the hypotheses discussed above by computing a metric for the quality of the fits obtained. The basis for this calculation is $\mathcal{L}(H_i)$, the likelihood of the observed data given the model fitted under a particular hypothesis H_i.

Following Wilk's theorem [49], we know the log-likelihood ratio $-2\ln\frac{\mathcal{L}(H_i)}{\mathcal{L}(H_0)}$ has a distribution that asymptotically (in the same size) approaches a χ^2 distribution with degrees of freedom given by the difference in the dimensionality of the two nested hypotheses, under the assumption that the null hypothesis is true. The mean and variance of the asymptotic distribution for the log-likelihood ratio are determined by the number of degrees of freedom. Given the fitted models, the likelihood of the observations under the various hypotheses are computed, and we follow the convention of reporting the difference between the observed statistic and the mean predicted by Wilk's theorem, in units of the standard deviation of the appropriate χ^2 distribution, and call this quantity N_σ. Intuitively, if this "badness-of-fit" number is large, we favor the null hypothesis (i.e., the hypothesis fit is bad), but if this number is small, both the simpler hypothesis and the null hypotheses are valid, and the simpler hypothesis is favored as a parsimonious model. We emphasize this "badness-of-fit" parameter cannot be obtained by characterization techniques such as RB that do not have an implicit model built in.

The log-likelihood ratios allow us to quantify whether (a) the observations are consistent with a Markovian error model (i.e., whether H_1 is plausible), and (b) whether the observations are consistent with a Clifford group operation with a Pauli error model (i.e., whether H_2 is plausible). In particular, we are interested in testing whether the answer to these questions changes when we apply PFR to our experiments. For this, it is necessary to look at the likelihood of hypotheses in different data sets (i.e., the unrandomized and the randomized GST experiments). Likelihoods cannot be meaningfully compared across different data sets. Instead, we simply consider the plausibility of the different hypothesis for the different data sets, while taking great care to ensure that the data sets are representative of the same noise and error environment, as we now describe.

IV. EXPERIMENTS

A. Device parameters

To test the hypotheses of Sec. III, we implement the PFR procedure on a superconducting qubit device (see Fig. 1). The device consists of four fixed-frequency, transmon qubits, designed to be similar to those described in Ref. [3]. The qubits are uncoupled but read out through a common Purcell filter. For the PFR experiment in this paper, only one qubit (Q_1) is measured. Q_1 is dispersively coupled to a readout resonator with a center frequency of $\omega_r/2\pi = 7.112$ GHz, $\kappa/2\pi = 3.4$ MHz, which is in turn capacitively coupled to a quarter-wave Purcell filter with external $Q = 22$ and a center frequency of $\omega_f = 7.27$ GHz [57] enabling fast qubit readout. Q_1 has a fixed 0-1 transition frequency of $\omega_q/2\pi = 4.432$ GHz with an anharmonicity $\alpha/2\pi = 308$ MHz. Coherence times measured for Q_1 are $T_1 = 10\ \mu$s, $T_2 = 13\ \mu$s, and

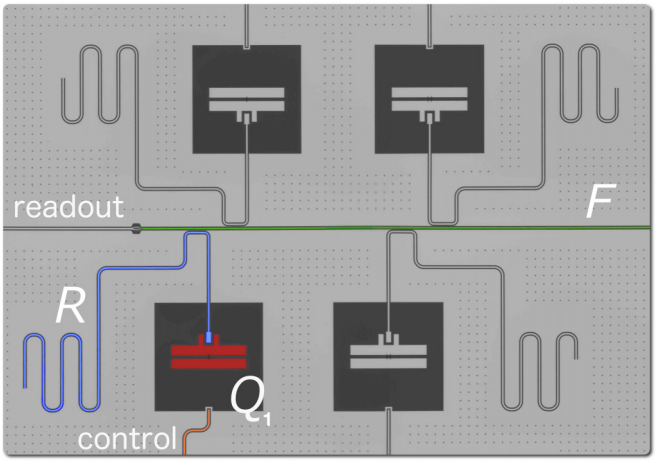

FIG. 1. False-color micrograph showing qubit Q_1 (red), resonator *R* (blue), and Purcell filter *F* (green). The qubit is dispersively coupled to a $\lambda/4$ readout resonator which is capacitively coupled to a Purcell filter with a $Q = 22$. Qubit control is done through a dedicated drive line (orange) and all qubit readout is done via a central feed line incorporating a Purcell filter.

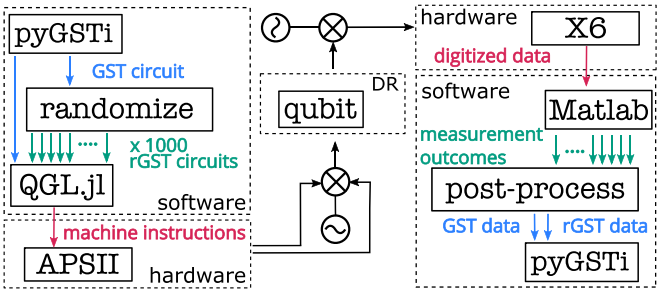

FIG. 2. Experimental data flow. Basic GST sequences are created using pyGSTi [24]. This basic experiment is then randomized 1000 times as described in section II. Each randomization and the original experiment get compiled by the QGL.jl compiler that translates sequence instruction into instructions implemented by APSII pulse sequencer [65]. The qubit response is digitized by an Innovative Integration X6 digitizer card and organized with a *Matlab* experimental framework. The single-shot data from each experiment are then postprocessed into counts that pyGSTi uses to reconstruct the gate set process matrices and the goodness-of-fit metrics.

a Hahn echo time of $T_{\text{echo}} = 16\ \mu$s (the other qubits in the device were not characterized). The pulses used were 50 ns long, leading to an expected average gate infidelity (diamond norm distance) of at least $\sim$0.2% per pulse ($\sim$1% per pulse). Since we use two pulses in the implementation of the gates discussed here, we expect an infidelity of no less than $\sim$0.4% and diamond norm distance of no less than $\sim$2% (ignoring all other sources of error and ignoring the effects of PFR).

B. Electronics and software stack

Making the PFR process experimentally tractable requires leveraging a complex software and hardware control infrastructure. The first hurdle is the sheer number of experiments needed. Long-sequence GST (ℓGST) experiments require a large set ($\sim$3500) of long circuits, each with up to 6155 gates. To ensure high accuracy, we produce a large number of measurement shots, 1000 per sequence. Under PFR, we take a single measurement shot per randomized sequence, resulting in 1000 unique GST circuits for each of the $\sim$3500 ℓGST circuits originally specified. Thus, in total, we measure over 3.5 million unique sequences to obtain high tomographic reconstructions for the gates in the unrandomized and the randomized experiments [58]. The second challenge is running the experiments in a way that allows the most direct comparison between the randomized and unrandomized cases—doing so allows us to minimize the impact of drift when comparing how the hypothesis tests from Sec. III fare on the different data sets. To achieve this, the unrandomized and the randomized sequences should be run in an interleaved fashion to ensure they experience the same noise environment (to the extent possible). These requirements necessitate hardware that can execute a large number of very long circuits, and to quickly alternate between them.

To address these issues we use a custom sequence compiler written in JULIA [59] called QGL.jl [60], providing a 4$\times$ compilation speed-up per circuit over an earlier Python version through a combination of parallel computation and other efficiency improvements—this ensured we were able to compile the 3.5 million unique sequences into pulse sequences in a reasonable amount of time. To minimize the runtime overhead we leverage a custom arbitrary pulse sequencer with gateware dedicated to implementing quantum circuits [61].

A rough outline of the process is as follows. Standard, one-qubit ℓGST sequences are created using the GST experiment and analysis software pyGSTi [24]. For the data presented here, we choose the maximal sequence length in GST to be 6150 gates, to ensure the experiment will have high accuracy, and be sensitive to non-Markovianity over timescales long compared to qubit coherence times. The ℓGST sequences are then randomized as described in Sec. II. It is worth emphasizing the lengths of the randomized circuits are unchanged as the Pauli group operations are combined with a neighboring Clifford group operation [62], much as the randomized compiling proposal of Ref. [14].

The collection of uniquely randomized GST (rGST) circuits and the original unrandomized circuits are then passed to the QGL.jl compiler which translates qubit gate instructions into machine instructions. This involves not only mapping high-level instructions to control pulses but also time-ordering and synchronizing instruction playback between all qubit control and readout channels. We note, this process takes significantly more time than the generation or randomization of the GST circuits. The compiled instructions are then passed to the pulse sequencer which is used to control the qubit. Due to the nature of the randomization process, the rGST experiments lack any kind of repetition or subroutine structure which rules out any efficient storage in hardware memory of a complete set of circuits. To address this, the set of randomized experiments are broken into groups of ten in order to fit in the control hardware memory. These ten single-shot runs of rGST were then interleaved with ten shots of unrandomized ℓGST. The process is repeated 100 times for each data point in Fig. 3. The complete process flow for a single round of experiment generation is illustrated in Fig. 2.

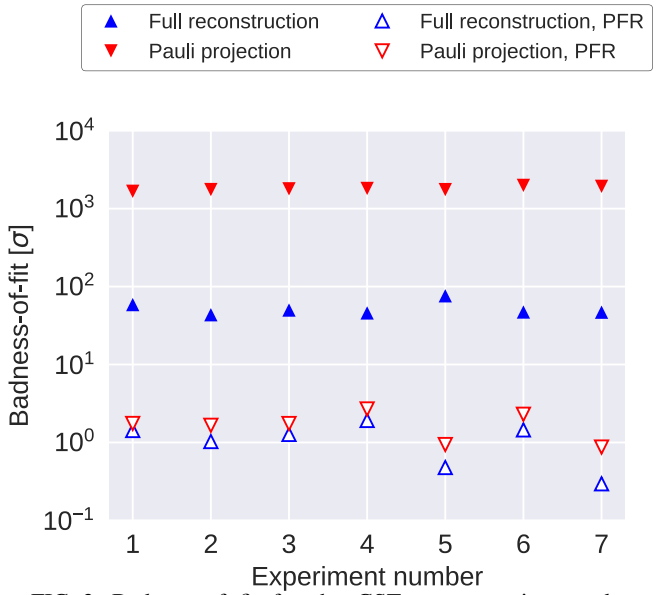

FIG. 3. Badness of fit for the GST reconstructions under a Markovian error model (H_1, upward blue triangles), and a Markovian stochastic Pauli error model (H_2, downward red triangles), as quantified by the log-likelihood-ratio statistic. This statistic is presented as the difference from the predicted mean of the χ^2 distribution (from Wilk's theorem), in units of the standard deviation of that same distribution, under the assumption that H_0 is true. Both experiments without randomization (solid triangles) and with randomization (open triangles) are considered. As discussed in the text, the 3.5 million unique sequences that comprise the tomography experiments for randomized and unrandomized experiments are measured in an interleaved fashion, so that both reconstructions should experience the same physical noise conditions. The entire collection of tomography experiments is repeated seven times to illustrate the behavior observed persists, and thus unlikely to be the result of statistical fluctuations, and is robust to drift in our system (each of these seven experiments lasted roughly 1 h).

In the canonical construction of the Clifford group, elements are composed of multiple native π and $\frac{\pi}{2}$ pulses, which leads to Clifford group elements being implemented by different numbers of native gates and pulses of nonuniform length. To account for this, we use a "diatomic" implementation of the group where each Clifford group operation is performed with two $X_{\pi/2}$ pulses of fixed length (50 ns) and three possible Z-frame updates [56,63]. This diatomic approach ensures all Clifford operations have equal duration. The room-temperature measurement signals were processed with an autodyne technique described in Ref. [64] using the BBN-QDSP digitization architecture [65] for the *Innovative Integrations X6-1000M* digitizer card. The final state assignment is then fed into the pyGSTi package for gate-set reconstruction. pyGSTi also provides the likelihood of H_0 and H_1, while custom code generates the likelihood of H_2—from these likelihoods, we obtain the likelihood ratio statistic and compare it to the predictions from Wilk's theorem.

V. RESULTS

The experiment outlined in Sec. III was performed to test the effectiveness of PFR. This process was repeated seven times, each taking roughly 1 h to complete. The repetitions allow us to observe how drift affects the results over an operationally meaningful amount of time.

One of the critical questions of this work is the validity of H_2 (the hypothesis that gates are well described by Clifford group operations followed by stochastic Pauli noise) and the Markovian behavior of qubit evolution under PFR. Data addressing this question can be seen in Fig. 3 where the GST model violation is plotted in terms of N_σ both with and without randomization. Several features are immediately apparent: (1) The Markovian fits (H_1) to the unrandomized experiments (solid triangles) are orders of magnitude worse than the randomized experiments (open triangles); (2) the data projected to a Pauli error model (H_2) in the unrandomized cases (red open triangles pointing down) are roughly three orders of magnitude worse than the randomized experiments (red open triangles pointing down); and (3) there is little difference between the quality of the fits under all the hypotheses for the randomized experiments (open triangles). In terms of the hypotheses outlined previously, for unrandomized experiments there is a large likelihood discrepancy between H_0 and the simpler hypotheses, greatly favoring the non-Pauli, non-Markovian H_0 model (H_1 is 43σ–76σ away from the predictions from Wilk's theorem, and H_2 is 1754σ–1987σ away), while for the randomized experiments all hypotheses have a comparable likelihood (within 0.3σ–2.7σ of the predictions from Wilk's theorem), so it is reasonable to take the simplest hypothesis (the Markovian, stochastic Pauli error model H_2) as the best explanation for those observations.

We should note that, despite the base level of H_1 model violation measured in the unrandomized data (a signature of non-Markovianity) appearing large at 1987σ, it is largely consistent with observations in other systems under similar circumstances (see, e.g., ion trap experiments without drift control or decoupling pulses [22]).

These features strongly indicate the noise in the absence of randomization is not well described by a Markovian error model, which follows from comment (1) above. Also apparent from comment (2) is that even the best Markovian error model is not well approximated by a Pauli model in the absence of randomization. Conversely, these features indicate the noise under PFR is very well described by a Markovian Pauli error model. In much simpler terms, the features of non-Pauli error models (i.e., nontrivial off-diagonal matrix elements in the Pauli-Liouville representation [55]) are insignificant in the reconstructions of the randomized experiments, as Fig. 4 illustrates. These separations are persistent over many repeats of the experiment, and the separation of many orders of magnitude indicates that PFR worked in these experiments not only quantitatively, but qualitatively, i.e., unrandomized experiments have strong non-Markovian features, while randomized experiments were well explained by Markovian Pauli error models.

Behavior of error metrics under randomization

The badness-of-fit results illustrate that the models derived under randomization are much more useful in explaining the observations than the models derived without randomization. A natural question that arises is whether this comes at the cost

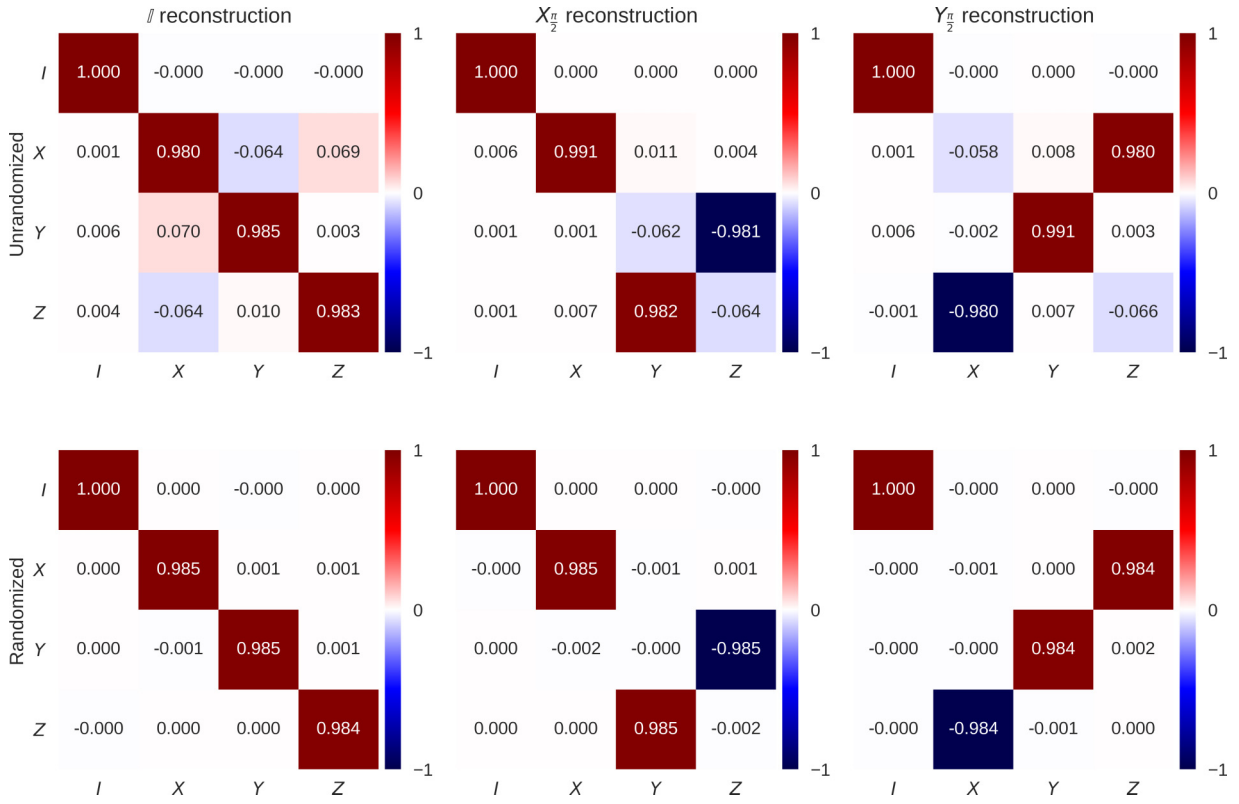

FIG. 4. Matrix representations of the reconstructed processes corresponding to the $\mathbb{I}$, $X_{\frac{\pi}{2}}$, and $Y_{\frac{\pi}{2}}$ operations in the first of the seven experiments performed (details of the other six experiments and confidence interval computation can be found in Ref. [66]). Without randomization (top row) there are significant off-diagonal contributions, corresponding to non-Pauli errors. With randomization (bottom row) there are no statistically significant off-diagonal contributions, indicating the errors correspond to a Pauli error model, as expected. Error bars (95% confidence) for this experiment are smaller than ± 0.0028 (unrandomized) and ± 0.0023 (randomized).

of degrading the performance of the gates. Here, we demonstrate that this is not the case, and that in fact the performance of the gates improves under PFR.

We computed the average gate infidelity [67] and the diamond norm distance [68] for the reconstructed gates under normal operation and under PFR, as depicted in Fig. 5. The observed average gate infidelities (diamond norm distances) are roughly double (quadruple) the expected coherence limits of the device, which may be explained by dynamical effects in the gate implementation which may be addressed by more careful pulse shaping [69] (and which are not accounted for in the coherence limit calculation mentioned earlier). We observe no appreciable difference between the infidelity of randomized and unrandomized experiments, while the diamond norm distance is reduced by a factor of 3–5 under PFR. This is consistent with the well-known behavior of the infidelity and the diamond norm under small coherent errors—namely, the infidelity is only sensitive to coherent errors to second order, while the diamond distance is sensitive to first order [70].

The diamond norm distance of unrandomized experiments monotonically increases over the course of the seven experiments, a behavior consistent with drift in the qubit and control parameters with respect to calibrations. The qubit control parameters are calibrated only once, at the beginning of the first experiment. This drift may at least partially explain the violation of the time-independent Markovian model represented by H_1, since these parameters appear to be continuously and systematically drifting throughout the seven experimental runs, but this is a small effect since H_1 still makes accurate predictions within each of the runs.

It should be noted that the drift is not apparent in the randomized experiments (even in the diamond norm distance), despite these experiments being run under the same conditions as the unrandomized experiments. This indicates that the drift was averaged away under PFR, a behavior consistent with coherent errors.

VI. SUMMARY

We have demonstrated that Pauli-frame randomization reduces both the non-Markovian features and the non-Pauli model features of errors in single-qubit experiments. This demonstration relies on long-sequence gate-set tomography, which yields high-accuracy reconstructions of all operations used in the experiments. This in turn required a high degree of automation to capture and process the $\sim$7 million measurement shots per hour. In the absence of randomization, the experiments were shown to have strong non-Markovian features, and the best Markovian model in that case was also shown to have strong features inconsistent with Pauli error models. In the presence of Pauli-frame randomization, the experiments were shown to be highly consistent with a Markovian Pauli error model, as predicted. As quantified by log-likelihood-

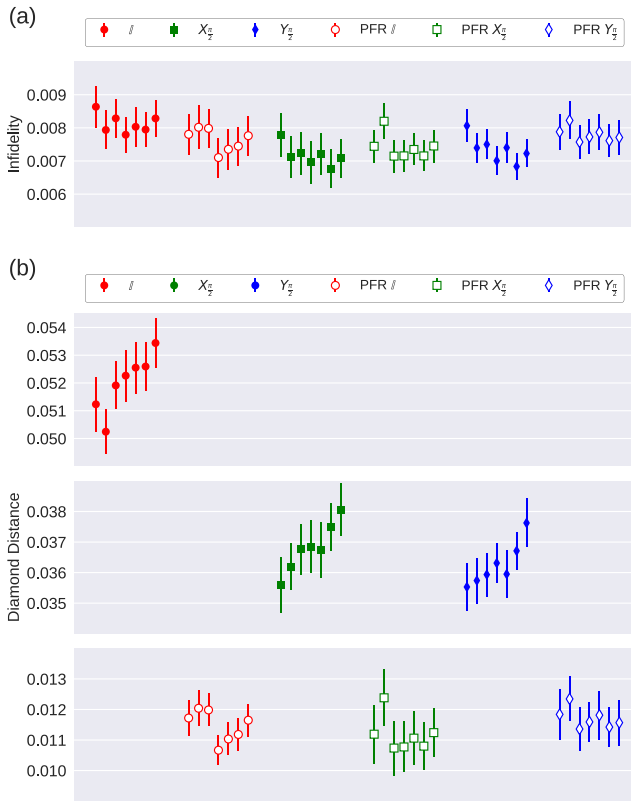

FIG. 5. (a) Average gate infidelity and (b) diamond norm distance estimates for all three gates (colors) on a sequence of seven separate experiments over several hours. Data for different gates are horizontally offset for clarity, but we emphasize each of the experiments leads to a reconstruction of all three gates, for both randomized (solid symbols) and unrandomized (open symbols) gate-sets under identical conditions—the only distinction between the seven experiments is the passage of time. All quantities are computed from the reconstructed process matrices. For the randomized case (open symbols), the infidelity and the diamond norm are comparable, at ≈1%. For the unrandomized experiments, there is significant deviation between the diamond norm error rate and the infidelity, suggesting the presence of coherent errors that affect the infidelity metric only weakly (and which are suppressed in the randomized experiments). A monotonic upward trend in the diamond norm distance of unrandomized experiments (solid symbols) implies the presence of systematic drift in the control pulses, which is also suppressed by randomization (open symbols). Error bars are 95% confidence intervals [analytically for (a), and with the Hessian provided by pyGSTi for (b)].

ratio statistic, the violation of Markovian and Pauli error models in the unrandomized experiments is highly significant, as high as 1987σ, while the violations of Markovian Pauli error models in the randomized experiments are statistically insignificant, less than 2.7σ in most of the experiments. This several orders-of-magnitude separation between randomized and unrandomized experiments was persistent across seven repeats of the experiment, indicating the noise-shaping effect of Pauli-frame randomization is robust to drift in the control parameters and fluctuations in the noise environment.

Areas for future work include speeding up the experiments using techniques such as active reset [71,72], and pushing randomization process onto the hardware field-programmable gate array (FPGA), which would allow for data acquisition of randomized Clifford group circuits without the user having to manually precompile random circuits.

Note added. Recently, similar results were independently reported by Hashim *et al.* [73].

The experimental data, along with scripts used to perform the analysis and plot the results, can be found in Ref. [66]. These include the full tomographic reconstruction of gate sets for all seven experiments, along with the raw counts needed for these reconstructions.

ACKNOWLEDGMENTS

We thank Robin Blume-Kohout, Erik Nielsen, Joel Wallman, and Joseph Emerson for fruitful discussions, and the anonymous referees for comments that helped improved the presentation of the paper. We also thank Alan Howsare, Adam Moldawer, and Ram Chelakara at Raytheon Integrated Defense Systems for sample fabrication. The qubit design and fabrication were funded by Internal Research and Development at Raytheon BBN Technologies and directed by Thomas Ohki. This work was funded by LPS/ARO Grant No. W911NF-14-C-0048. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

M.W., G.R., and M.P.S. wrote the manuscript. M.W., G.R., and D.R. performed the experiment, while C.R. and B.J. built the control software and electronics infrastructure. G.R. and M.W. designed the device and assisted in fabrication. M.P.S. proposed the experiment, and wrote the experiment generation scripts. M.W. and M.P.S. performed the data analysis.

[1] J. Kelly, R. Barends, A. G. Fowler, A. Megrant, E. Jeffrey, T. C. White, D. Sank, J. Y. Mutus, B. Campbell, Yu. Chen, Z. Chen, B. Chiaro, A. Dunsworth, E. Lucero, M. Neeley, C. Neill, P. J. J. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, J. Wenner, and J. M. Martinis, Scalable *in situ* qubit calibration during repetitive error detection, Phys. Rev. A **94**, 032321 (2016).

[2] 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).

[3] J. M. Chow, J. M. Gambetta, E. Magesan, D. W. Abraham, A. W. Cross, B. R. Johnson, N. A. Masluk, C. A. Ryan, J. A. Smolin, S. J. Srinivasan, and M. Steffen, Implementing a strand of a scalable fault-tolerant quantum computing fabric, Nat. Commun. **5**, 4015 (2014).

[4] D. P. DiVincenzo, The physical implementation of quantum computation, Fortschr. Phys. **48**, 771 (2000).

[5] P. Iyer, M. P da Silva, and D. Poulin, Critical parameters of a noise model that affect fault tolerant quantum computation on a single qubit, in *APS March Meeting 2016* (Bulletin of the American Physical Society, 2016), abstract: P44.00003.

[6] A. Darmawan, N. Delfosse, P. Iyer, and D. Poulin, Surprising facts about quantum error correction, Sydney Quantum Information Theory Workshop, Coogee, Australia, 2017, https://www.physique.usherbrooke.ca/~dpoulin/utilisateur/files/seminaires/2017_Coogee.pdf.

[7] P. Iyer and D. Poulin, A small quantum computer is needed to optimize fault-tolerant protocols, Quantum Sci. Technol. **3**, 030504 (2018).

[8] B. M. Terhal and G. Burkard, Fault-tolerant quantum computation for local non-Markovian noise, Phys. Rev. A **71**, 012336 (2005).

[9] P. Aliferis, D. Gottesman, and J. Preskill, Quantum accuracy threshold for concatenated distance-3 codes, Quantum Inf. Comput. **6**, 97 (2006).

[10] D. Aharonov and M. Ben-Or, Fault-tolerant quantum computation with constant error rate, SIAM J. Comput. **38**, 1207 (2008).

[11] H. K. Ng and J. Preskill, Fault-tolerant quantum computation versus Gaussian noise, Phys. Rev. A **79**, 032318 (2009).

[12] O. Kern, G. Alber, and D. L. Shepelyansky, Quantum error correction of coherent errors by randomization, Eur. Phys. J. D **32**, 153 (2005).

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

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

[15] L. Viola and E. Knill, Random Decoupling Schemes for Quantum Dynamical Control and Error Suppression, Phys. Rev. Lett. **94**, 060502 (2005).

[16] M. A. Nielsen, A simple formula for the average gate fidelity of a quantum dynamical operation, Phys. Lett. A **303**, 249 (2002).

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

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

[19] C. Stark, Self-consistent tomography of the state-measurement Gram matrix, Phys. Rev. A **89**, 052109 (2014).

[20] 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).

[21] R. J. Blume-Kohout, J. K. Gamble, E. Nielsen, P. L. W. Maunz, T. L. Scholten, and K. M. Rudinger, Turbocharging quantum tomography, Technical Report No. SAND2015-0224, Sandia National Laboratories, 2015 (unpublished).

[22] 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).

[23] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout, Gate set tomography, arXiv:2009.07301.

[24] E. Nielsen, L. Saldyt, J. Gross, T. Scholten, K. Rudinger, T. J. Proctor, and D. Nadlinger, pyGSTi, https://github.com/pyGSTio/pyGSTi (2017).

[25] A. Y. Kitaev, A. Shen, and M. N. Vyalyi, *Classical and Quantum Computation*, Vol. 47 (American Mathematical Society, Providence, RI, 2002).

[26] J. Watrous, Semidefinite programs for completely bounded norms, Theory Comput. **5**, 217 (2009).

[27] We define the operations to be Markovian if they can be well approximated by linear transformations of the state (density operator)—we consider the coarse-grained evolution at the timescale of the primitive gates, and do not consider issues such as divisibility of operations, or their infinitesimal generators.

[28] D. Gottesman and I. L. Chuang, Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations, Nature (London) **402**, 390 (1999).

[29] D. Gottesman, Heisenberg representation of quantum computers, in *Group 22: Proceedings of the XXII International Colloquium on Group Theoretical Methods in Physics*, edited by S. P. Corney, R. Delbourgo, and P. D. Jarvis, International Press Lectures and Conference Proceedings in Physics (International Press, Cambridge, MA, 1999), pp. 32–43.

[30] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).

[31] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, Mixed-state entanglement and quantum error correction, Phys. Rev. A **54**, 3824 (1996).

[32] D. P. DiVincenzo, D. W. Leung, and B. M. Terhal, Quantum data hiding, IEEE Trans. Inf. Theory **48**, 580 (2002).

[33] 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).

[34] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).

[35] M. Silva, E. Magesan, D. W. Kribs, and J. Emerson, Scalable protocol for identification of correctable codes, Phys. Rev. A **78**, 012347 (2008).

[36] 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).

[37] Throughout this discussion, we represent quantum operations by the corresponding superoperators, denoted with calligraphic uppercase letters ($\mathcal{A}$, $\mathcal{B}$, etc.), while the noisy implementations are denoted in the same way but with a tilde.

[38] This argument carries through independently of how the Clifford group operation is physically implemented (one pulse or multiple pulses, as long as $\mathcal{D}_i$ is applied consistently regardless of where may appear in a sequence).

[39] Although the error in the last gate of the sequence is not randomized due to our choice to have the last randomization operation cancel all others, we can treat this as a measurement error. Alternatively, we could choose all randomization operations uniformly at random and changed the measurement outcome to undo the randomization, so that effectively the error in the last gate would also be twirled over the Pauli group.

[40] J. Emerson, R. Alicki, and K. Zyczkowski, Scalable noise estimation with random unitary operators, J. Opt. B **7**, S347 (2005).

[41] 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).

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

[43] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A **85**, 042311 (2012).

[44] 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).

[45] J. F. Poyatos, J. I. Cirac, and P. Zoller, Complete Characterization of a Quantum Process: The Two-Bit Quantum Gate, Phys. Rev. Lett. **78**, 390 (1997).

[46] 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).

[47] D. Greenbaum, Introduction to quantum gate set tomography, arXiv:1509.02921.

[48] J. Neyman and E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Philos. Trans. R. Soc. A **231**, 289 (1933).

[49] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. **9**, 60 (1938).

[50] K. Blum, *Density Matrix Theory and Applications* (Plenum, New York, 1981).

[51] D. W. Leung, Towards robust quantum computation, Ph.D. thesis, Stanford University, 2000.

[52] B. Rahn, A. C. Doherty, and H. Mabuchi, Exact performance of concatenated quantum codes, Phys. Rev. A **66**, 032304 (2002).

[53] J. Fern, J. Kempe, S. N. Simic, and S. Sastry, Generalized performance of concatenated quantum codes—a dynamical systems approach, IEEE Trans. Autom. Control **51**, 448 (2006).

[54] J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, M. B. Ketchen, and M. Steffen, Universal Quantum Gate Set Approaching Fault-Tolerant Thresholds with Superconducting Qubits, Phys. Rev. Lett. **109**, 060501 (2012).

[55] S. Kimmel, M. P. da Silva, C. A. Ryan, B. R. Johnson, and T. Ohki, Robust Extraction of Tomographic Information via Randomized Benchmarking, Phys. Rev. X **4**, 011050 (2014).

[56] B. R. Johnson, M. P. da Silva, C. A. Ryan, S. Kimmel, J. M. Chow, and T. A. Ohki, Demonstration of robust quantum gate tomography via randomized benchmarking, New J. Phys. **17**, 113019 (2015).

[57] E. Jeffrey, D. Sank, J. Y. Mutus, T. C. White, J. Kelly, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. Megrant, P. J. J. O'Malley, C. Neill, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, Fast Accurate State Measurement with Superconducting Qubits, Phys. Rev. Lett. **112**, 190504 (2014).

[58] The number of randomizations could be much lower [14], or, equivalently, more measurement shots could be collected per randomized sequences, but we chose to take the extreme limit of one shot per randomized sequence to avoid any subtle questions about how many shots would be safe to take per randomized sequence before correlations became significant.

[59] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, Julia: A fresh approach to numerical computing, SIAM Rev. **59**, 65 (2017).

[60] C. Ryan, D. Ristè, and B. Johnson, QGL.jl: master branch (August 2016), https://github.com/BBN-Q/QGL.jl/.

[61] BBN APS II, 2017, http://www.raytheon.com/capabilities/rtnwcm/groups/public/documents/content/aps-datasheet.pdf.

[62] The time necessary to generate these random sequences is minimal, on the order of a handful of minutes for all 3.5 million unique sequences, since it requires only the simulation of Pauli errors propagating in Clifford circuits. If we were not taking the additional step of compensating for the randomization in the final measurement, the generation of these sequences would be even faster, since it would not require the propagation of the randomizing Pauli operations.

[63] 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).

[64] C. A. Ryan, B. R. Johnson, J. M. Gambetta, J. M. Chow, M. P. da Silva, O. E. Dial, and T. A. Ohki, Tomography via correlation of noisy measurement records, Phys. Rev. A **91**, 022118 (2015).

[65] C. A. Ryan, B. R. Johnson, D. Ristè, B. Donovan, and T. A. Ohki, Hardware for dynamic quantum computing, Rev. Sci. Instrum. **88**, 104703 (2017).

[66] M. Ware, G. Ribeill, D. Ristè, C. A. Ryan, B. Johnson, and M. P. da Silva, Data set and analysis for "Experimental Pauli-frame randomization on a superconducting qubit", doi: 10.5281/zenodo.4279722, Zenodo (2020).

[67] In this work we use the standard definitions of average gate infidelity as $1 - \mathcal{F}(\mathcal{U}, \tilde{\mathcal{U}})$, with average gate fidelity $\mathcal{F}$ defined as [16]

$$\mathcal{F}(\mathcal{U}, \tilde{\mathcal{U}}) = \frac{\mathrm{tr}\tilde{\mathcal{U}}\mathcal{U}^\dagger + d}{d^2 + d}$$

.

[68] The diamond distance between two processes [25,26] is defined as

$$\mathcal{D}_\diamond = \frac{1}{2}||\tilde{\mathcal{U}} - \mathcal{U}||_\diamond = \sup_\rho \frac{1}{2}||(\tilde{\mathcal{U}} \otimes \mathcal{I}_d - \mathcal{U} \otimes \mathcal{I}_d)(\rho)||_1,$$

with d being the total system dimension.

[69] J. M. Chow, L. DiCarlo, J. M. Gambetta, F. Motzoi, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, Optimized driving of superconducting artificial atoms for improved single-qubit gates, Phys. Rev. A **82**, 040305(R) (2010).

[70] R. Kueng, D. M. Long, A. C. Doherty, and S. T. Flammia, Comparing Experiments to the Fault-Tolerance Threshold, Phys. Rev. Lett. **117**, 170502 (2016).

[71] 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, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo, Restless Tuneup of High-Fidelity Qubit Gates, Phys. Rev. Appl. **7**, 041001 (2017).

[72] D. Ristè, C. C. Bultink, K. W. Lehnert, and L. DiCarlo, Feedback Control of a Solid-State Qubit Using High-Fidelity Projective Measurement, Phys. Rev. Lett. **109**, 240502 (2012).

[73] A. Hashim, R. K. Naik, A. Morvan, J.-L. Ville, B. Mitchell, J. M. Kreikebaum, M. Davis, E. Smith, C. Iancu, K. P. O'Brien, I. Hincks, J. J. Wallman, J. Emerson, and I. Siddiqi, Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor, arXiv:2010.00215.