

Experimental Characterization of Crosstalk Errors with Simultaneous Gate Set Tomography

Kenneth Rudinger,¹ Craig W. Hogle,² Ravi K. Naik[Ⓞ],³ Akel Hashim,³ Daniel Lobser,² David I. Santiago[Ⓞ],^{3,4} Matthew D. Grace[Ⓞ],⁵ Erik Nielsen,¹ Timothy Proctor,⁵ Stefan Seritan[Ⓞ],⁵ Susan M. Clark,² Robin Blume-Kohout,¹ Irfan Siddiqi,^{3,4,6} and Kevin C. Young^{5,*}

¹Quantum Performance Laboratory, Sandia National Laboratories, Albuquerque, New Mexico, 87185, USA

²Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

³Quantum Nanoelectronics Laboratory, Department of Physics, University of California at Berkeley, Berkeley, California 94720, USA

⁴Computational Research Division, Lawrence Berkeley National Lab, Berkeley, California 94720, USA

⁵Quantum Performance Laboratory, Sandia National Laboratories, Livermore, California 94550, USA

⁶Materials Sciences Division, Lawrence Berkeley National Lab, Berkeley, California 94720, USA



(Received 17 March 2021; accepted 8 October 2021; published 19 November 2021)

Crosstalk is a leading source of failure in multiqubit quantum information processors. It can arise from a wide range of disparate physical phenomena, and can introduce subtle correlations in the errors experienced by a device. Several hardware characterization protocols are able to detect the presence of crosstalk, but few provide sufficient information to distinguish various crosstalk errors from one another. In this article we describe how gate set tomography, a protocol for detailed characterization of quantum operations, can be used to identify and characterize crosstalk errors in quantum information processors. We demonstrate our methods on a two-qubit trapped-ion processor and a two-qubit subsystem of a superconducting transmon processor.

DOI: [10.1103/PRXQuantum.2.040338](https://doi.org/10.1103/PRXQuantum.2.040338)

I. INTRODUCTION

Quantum information processors have demonstrated one- and two-qubit quantum operations with error rates below the threshold required for fault-tolerant quantum computation [1–10]. One of the biggest obstacles to achieving similarly low error rates in large, integrated quantum processors is the appearance of a large class of errors known collectively as *crosstalk* [11–15]. Crosstalk can increase error rates on individual qubits, and can also cause errors on different qubits to become correlated with one another. These correlations are particularly damaging for error correction [16–18], and optimizing the power of quantum error correction requires understanding and strictly controlling crosstalk errors.

The underlying physical causes of crosstalk errors in quantum information processors are diverse. Perhaps the most familiar source is pulse spillover, wherein a

control pulse (i.e., laser, rf signal, etc.) on a target qubit unintentionally affects a neighboring qubit. But crosstalk errors can also occur due to, e.g., coherent coupling between qubits, shared quantum environments, or even shared classical environments that experience spatially correlated fluctuations. To reduce or mitigate crosstalk errors [19–35] and enable fault-tolerant quantum computation, experimentalists need characterization methods that provide detailed information about the specific crosstalk errors that occur in their processors.

A number of techniques to characterize [15,36–52] the impact of crosstalk have been developed and implemented. Randomized methods—such as simultaneous randomized benchmarking (RB) [41], correlated RB [42], cycle benchmarking [43], and Pauli noise learning [44]—are among the most popular, as they are generally simple to implement and analyze. However, these methods are typically sensitive to coherent errors at only second order [45–47], and rely on twirling techniques that obfuscate the underlying physical sources of observed errors. Model-free methods based on hypothesis testing of probability distributions [15,48] can identify the presence of crosstalk errors, but cannot characterize those errors. Methods based on quantum process tomography, such as selective process tomography [49] and direct characterization of quantum

*kyoung@sandia.gov

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

dynamics [50], can be adapted to provide insight into certain types of crosstalk (such as coherent two-qubit interactions) but are not designed to detect others (such as how gate errors depend on neighboring operations), and inherit many of the well-known problems of quantum process tomography [51]. Finally, specialized techniques, like cross-Rabi oscillations or direct capacitance measurements [52], are useful for learning specific physical parameters, but they are generally unable to detect other crosstalk errors that may be present, or even dominant, in a system.

In this article we demonstrate how to use gate set tomography (GST) [5,51,53] to perform a detailed investigation of crosstalk errors between two subsystems of a quantum information processor. GST is a protocol designed to provide detailed characterization of qubit dynamics by estimating a set of process matrices describing the various operations of a processor. So, in principle, we could simply perform GST on a multiqubit system, obtain multiqubit process matrices describing the gates, and look for the presence of crosstalk errors in these process matrices. But such process matrices are large and unwieldy. It is not clear what constitutes conclusive evidence for crosstalk errors, nor how to reliably distinguish real effects from statistical noise (finite-sample fluctuations).

Instead, we take a two-stage approach. First, we construct three parameterized models that, by design, allow for different degrees and kinds of crosstalk errors [see Table I and Figs. 1(b)–1(d)]. We fit them to data, and we evaluate how well they explain the data [54]. We can infer a surprising amount of information just from this evaluation, because we know exactly what kinds of crosstalk each model can describe, and for each model, we can quantify the amount of observed error that it failed to describe. We also use this analysis to select the simplest model that fits the data well, estimate its parameters to obtain “best fit” process matrices describing the gates and their errors, and analyze those process matrices in detail to understand the nature of the crosstalk errors. When the best-fit model only

TABLE I. Three nested families of process matrix models for quantum logic operations. Each successively larger model can capture richer forms of crosstalk between qubits than the previous ones. Also shown are the decompositions of the process matrices as tensor products of local operations when operations A and B are applied simultaneously to subsystems 1 and 2, respectively. Here N_p is the number of free parameters required to describe the model.

Model	Decomposition	N_p	Constraints	
			(i)	(ii)
Crosstalk-free	$\mathcal{G}_A^{(1)} \otimes \mathcal{G}_B^{(2)}$	86	✓	✓
Context dependent	$\mathcal{G}_{A,B}^{(1)} \otimes \mathcal{G}_{B,A}^{(2)}$	240	✓	
General	$\mathcal{G}_{A,B}^{(1,2)}$	1683		

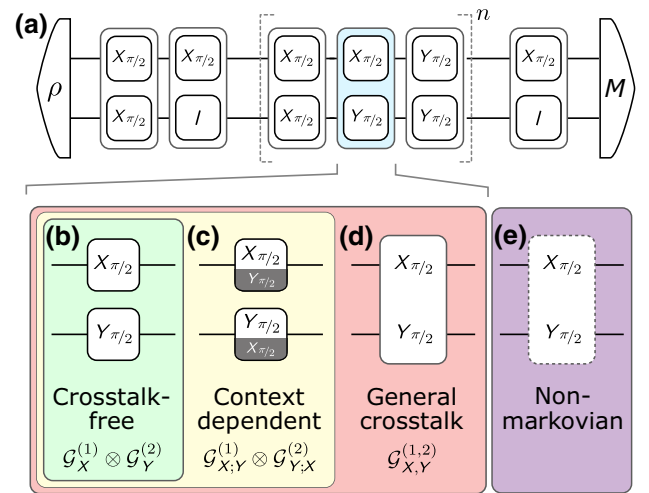


FIG. 1. Detecting and measuring crosstalk errors with gate set tomography. To probe crosstalk between two qubits, we execute a set of two-qubit quantum circuits (a) whose layers comprise parallel, single-qubit gates. These circuits consist of initialization into $\rho \approx |00\rangle\langle 00|$, a short state-preparation operation, an n -fold repeated *germ* operation, a short measurement-preparation operation, and finally measurement in the computational basis. The measured outcomes of these circuits are used to fit each of the three models in (b)–(d): (b) Crosstalk-free models assume that each elementary gate can be described by a single-qubit process matrix. (c) Context-dependent models assume that each gate can be described by a single-qubit process matrix conditioned on the neighboring qubit operation. (d) General crosstalk models assume that each two-qubit layer can be described by a full, two-qubit process matrix with no additional constraints. (e) Non-Markovian behavior cannot be described by two-qubit process matrices. We use tools from statistical model selection to decide which model offers the best balance between accuracy (describing the data) and simplicity (using the fewest parameters).

includes certain effects, we can apply customized analysis techniques that are tailored to those particular crosstalk errors. Finally, by comparing the measured errors and their magnitudes against candidate physical device models, we draw conclusions about the underlying physical causes of the observed crosstalk errors—and, potentially, how to mitigate them.

In this work we focus on identifying and analyzing crosstalk induced by parallel single-qubit gates on two single-qubit subsystems. However, all of our methods are straightforwardly extensible to larger systems, e.g., to crosstalk between two-qubit subsystems, induced by entangling gates. We demonstrate our methods on two U.S. Department of Energy (DOE) sponsored quantum computing testbed platforms—the transmon-based Advanced Quantum Testbed (AQT) [55] and a prototype of the trapped-ion-based Quantum Scientific Computing Open User Testbed (QSCOUT) [56–58]—and we discuss and compare the errors observed on these two devices.

II. MODELING CROSSTALK ERRORS

Crosstalk in quantum information processors can arise from a wide range of disparate physical mechanisms. Our goal in this work is to characterize the crosstalk errors in quantum logical operations induced by those mechanisms. The standard model for any Markovian error, including crosstalk errors [15], is a process matrix—a completely positive, trace-preserving map on density matrices. Multiqubit process matrices can describe a wide variety of crosstalk effects. To disaggregate different categories of crosstalk errors, we construct a hierarchy of process matrix models, each of which can only represent certain kinds of crosstalk errors. We fit these models to data, compare their ability to describe the observations, select the best model, and use it to draw quantitative conclusions about the crosstalk errors present in the device. In this section, we present (1) those models, (2) our methods for fitting them to data and evaluating fit quality, (3) how we analyze an estimated model, and (4) the analysis pipeline used in our experiments.

A. Families of crosstalk error models

Reference [15] defines two conditions that must be satisfied for a quantum processor to be free of crosstalk.

- (i) Locality of operations—the process matrices decompose as tensor products.
- (ii) Independence of local operations—each component in the tensor product depends only on what gate is acting on that subsystem.

Each condition defines a constraint on multiqubit process matrices, which can otherwise describe many forms of crosstalk. By systematically enforcing or relaxing constraints (i)–(ii), we can construct the three families of models shown in Table I. They describe increasingly complex crosstalk phenomena.

The crosstalk-free model cannot model any type of crosstalk in the system—each gate is required to act locally and independently. The process matrix representing each layer of gates is required to have a tensor-product form, and the local process matrix on each subsystem is independent of its context (i.e., which operations are applied to other subsystems).

The context-dependent model relaxes the independence constraint. Each gate still acts locally—a layer’s process matrix must have tensor-product form—but the local operations on a subsystem can vary from layer to layer (i.e., depend on context). This model can capture pulse-spillover effects, where the operations on one subsystem are perturbed by operations on the other. But, like the crosstalk-free model, the context-dependent model cannot model entangling Hamiltonians or correlated errors that create correlations between the two subsystems.

The general crosstalk model relaxes both constraints, and represents each parallel, single-qubit layer as a full two-qubit process matrix. This model can capture all Markovian crosstalk effects, including context dependence, entangling operations, and classical correlation.

In principle, a complete description of the system’s crosstalk errors could be extracted just by estimating the parameters of the general model. In practice, analyzing those process matrices and deciding which effects are statistically significant is difficult and time consuming. We let the model-fitting process do that work for us. If, e.g., the effects of crosstalk are not statistically significant then model selection criteria (see below) will indicate that the data are consistent with a crosstalk-free model. Conversely, if only the general model fits well then this is unambiguous evidence of nonlocal interactions, which we can track down by detailed analysis of process matrices.

Because our models utilize process matrices, they are all explicitly Markovian. But real devices are often *non-Markovian*—experiments on them yield data inconsistent with the predictions of any process matrix model. This has implications for the use of process matrix models to diagnose crosstalk. As discussed in Ref. [15], correlations induced by non-Markovianity can easily be mistaken for crosstalk errors, so finding some way to acknowledge and incorporate non-Markovian errors is extremely important. Later in this section we discuss techniques for quantifying non-Markovianity, and what to do when it (almost inevitably) appears.

To predict data, our models also need to describe errors in state preparation and measurement (SPAM). We only consider systems that initialize and measure all qubits simultaneously. In each of our models, the density matrices and positive operator-valued measure (POVM) effects that represent SPAM operations are required to respect the same constraints (e.g., tensor product decomposition) as the gate operations. Thus, both the crosstalk-free and context-dependent models incorporate initial states of the form $\rho = \rho^{(1)} \otimes \rho^{(2)}$ and POVM effects of the form $F_{ij} = F_i^{(1)} \otimes F_j^{(2)}$. SPAM operations in the general crosstalk model are unconstrained, and can describe entangled or correlated initial states, and arbitrarily correlated measurements.

B. Fitting models to data

Fitting process matrix models to data is typically done using some variant of process tomography [50]. While standard quantum process tomography can be used to characterize quantum gate operations, it is not *self-consistent*, i.e., it assumes access to input states and measurements that are already highly accurately characterized [51]. We use GST, a protocol introduced to solve this self-consistency problem [5,51,53]. GST experiments include all the circuits required for state, measurement, and

process tomography, plus a few additional circuits that establish a mutually consistent reference frame. This allows GST to fully characterize all state preparation, measurement, and gate operations on a processor, concurrently and self-consistently.

As illustrated in Fig. 1, GST relies on circuits of the form

$$\rho \text{---} \boxed{\mathbf{p}_i} \text{---} \boxed{\mathbf{g}_j^n} \text{---} \boxed{\mathbf{m}_k} \text{---} \mathbf{M}, \quad (1)$$

where \mathbf{p}_i is a fiducial state preparation subcircuit, \mathbf{m}_k is a fiducial measurement preparation subcircuit, and \mathbf{g}_j is an n -fold repeated “germ” subcircuit. Data from running these circuits can be used to estimate a process matrix for each distinct circuit layer [53]. In our experiments, we use and estimate the nine two-qubit layers formed by all possible parallel combinations of a three-element single-qubit gate set:

$$\mathbb{G} = \{a \otimes b : a, b \in \{X_{\pi/2}, Y_{\pi/2}, I\}\}. \quad (2)$$

Here $X_{\pi/2}$ and $Y_{\pi/2}$ are $\pi/2$ rotations around X and Y , respectively, and I is an idle gate. Our experiments use no entangling gates. As in standard GST [5,53], the fiducial operations are chosen to be informationally complete, and the germs are chosen to amplify all observable components of a full two-qubit process matrix model for each layer. See Table VI in Appendix A for the specific subcircuits used in our experiments.

Once data have been obtained, we use maximum likelihood estimation to find a best-fit estimate for each of the three crosstalk models, by varying its parameters θ to maximize the likelihood function $\mathcal{L}(\theta) = \Pr(\text{data}|\theta)$. We denote the maximum likelihood estimate of the i th model by $\hat{\theta}_i$. We denote the maximum value of the likelihood function for the i th model by $\mathcal{L}^{(i)} = \mathcal{L}(\hat{\theta}_i)$, and refer to it as the likelihood of model (i).

To construct, fit, and analyze these models, we use `pyGSTi` [59,60], a PYTHON implementation of GST that includes robust routines for fitting predictive models of quantum information processors to data, and analyzing the resulting estimates.

C. Comparing, selecting, and validating models

How well a candidate model (i) fits data is captured by its likelihood, $\mathcal{L}^{(i)}$. Extracting useful information requires some simple manipulations. We measure the quality of model (i)’s fit by the log-likelihood ratio between it and a “maximal model” that has no structure at all, and can assign an independent probability to each measurement outcome a of each circuit b ,

$$\lambda^{(i)} = -2 \ln(\mathcal{L}^{(i)}/\mathcal{L}_{\max}), \quad (3)$$

where the maximum likelihood of the maximal model,

$$\mathcal{L}_{\max} = \prod_{a,b} f_{a,b}^{N_b f_{a,b}}, \quad (4)$$

is achieved by predicting the observed frequency $f_{a,b}$ after N_b measurements of circuit b . The log-likelihood ratio is a standard hypothesis-testing statistic. Wilks’ theorem [61] states that when the data are actually generated by model (i) with some parameters θ , $\lambda^{(i)}$ is a χ_k^2 random variable, with k equal to the difference between the number of free parameters (N_p) for model (i) and the maximal model. Under this null hypothesis, $\langle \lambda^{(i)} \rangle = k$ and $\Delta \lambda^{(i)} = \sqrt{2k}$.

The data are inconsistent with model (i) if and only if $\lambda^{(i)}$ is inconsistent with a χ_k^2 distribution. In particular, when the data are not consistent with model (i), $\lambda^{(i)}$ will be larger than its expected value under the null hypothesis, k . How much model (i) is violated can be quantified by the number of standard deviations by which $\lambda^{(i)}$ exceeds its expected value under the null hypothesis,

$$N_\sigma^{(i)} = \frac{\lambda - k}{\sqrt{2k}}. \quad (5)$$

Ideally, we would simply choose the smallest model that fits the data—i.e., the smallest model for which N_σ is negligible. But in practice, most systems display enough non-Markovian behavior that no model—not even the general (process matrix) model—fits the data that well. In this case, we need a criterion for identifying how much better (or worse) one model is than another.

To derive such a criterion [54], we observe that Wilks’ theorem implies that removing exactly n “useless” parameters from a model increases $\langle \lambda \rangle$ by exactly n . So, if we consider two equally valid models (i) and (j), with (j) nested within (i) and having n fewer parameters than it, then we expect $\lambda^{(j)} - \lambda^{(i)} \approx n$. If we observe $\lambda^{(j)} - \lambda^{(i)} \gg n$, this suggests that the extra parameters in the larger model (i) are not useless—i.e., they describe real effects. But Akaike’s derivation of his eponymous Akaike information criterion (AIC) [62] demonstrates a scenario where using the larger model to fit those effects actually *decreases* predictive accuracy, unless $\lambda^{(j)} - \lambda^{(i)} \geq 2n$. We conclude that although there are multiple criteria for deciding which model is “better” in a given situation, they share a simple form: is $\lambda^{(j)} - \lambda^{(i)} \geq \alpha n$ for some α ?

To compare two nested models [63], we use a quantity that we call the *evidence ratio* of (i) against the smaller model (j) [54]:

$$\gamma^{(i,j)} = \left(\frac{\lambda^{(j)} - \lambda^{(i)}}{N_p^{(i)} - N_p^{(j)}} \right). \quad (6)$$

Here $N_p^{(i)}$ and $N_p^{(j)}$ indicate the numbers of parameters in models (i) and (j), respectively. If $\gamma^{(i,j)} \leq 1$ then there

is no evidence against the smaller model (j)—the larger model (i)’s extra parameters are functionally useless—and so we always choose (j). If $1 < \gamma^{(ij)} \leq 2$ then the data provide weak evidence against the smaller model, but the AIC suggests its predictions would still be more accurate. Even when $\gamma^{(ij)} > 2$, we may still choose the smaller model if we prioritize simplicity, but for any use case, there will be some threshold beyond which the smaller model must be rejected. In general, $\gamma^{(ij)}$ normalizes the weight of evidence against the smaller model, on a per-parameter basis, and provides a quantitative measure for comparing two models.

D. Quantifying unmodeled error

Statistical measures of model violation like $N_\sigma^{(i)}$ and $\gamma^{(ij)}$ quantify the amount of evidence for errors outside a given model. They do not quantify the magnitude of those errors. For example, they depend strongly on the amount of data taken. In many circumstances, we care more about the size of unmodeled errors than about the amount of evidence that they exist. In this work, we quantify the size of unmodeled errors using the *wildcard error* [64].

The wildcard error can quantify the per-gate deviation between a model’s predictions and observed data. To do this, we assign a *minimal wildcard model* to an estimate. A wildcard model assigns to each estimated gate g a number $w_g \geq 0$, and to each circuit C the total w for all the gates in it: $w_C = \sum_{g \in C} w_g$. Adding a wildcard model explicitly relaxes the estimate’s prediction for each circuit C : if the estimate originally predicted outcome distribution \vec{p}_C then the wildcard-augmented estimate predicts only that C ’s outcomes will be drawn from *some* \vec{p}_C whose total variation distance to the predicted distribution is bounded by w_C (that is, $\frac{1}{2} \|\vec{p}_C - \vec{p}_C\|_1 \leq w_C$). A minimal wildcard model is an assignment $\{w_g\}$ that just barely makes the estimate statistically consistent with the data. We only use single-parameter wildcard models that assign a single wildcard error rate ($W_{(i)}$) to all gates in the estimate of a model (i).

The minimal amount of W required to reconcile an estimate with data tells us whether unmodeled errors are dominant or negligible. In this work we use very simple wildcard models that assign a single wildcard error rate ($W_{(i)}$) to all gates in a model (i). We can compare W to the diamond error [ϵ_\diamond , Eq. (8)] [65–67] of the gates. The diamond error has the useful property that the total variation distance between the outcome distributions of ideal and noisy circuits is bounded above by the sum of the gate diamond errors (plus state preparation and measurement error). If the W assigned to the gates in a model’s estimate is significantly less than their average diamond error ($\bar{\epsilon}_\diamond$), that model explains most of the observed error. But if $W \geq \bar{\epsilon}_\diamond$, unmodeled errors may be dominant, and the model should probably be discarded or not taken seriously.

Unmodeled errors—heralded by significant W —can appear in our analysis from two distinct causes. If the general model cannot fit the data then its unmodeled errors constitute some sort of non-Markovian dynamics, since the general model (by construction) can model all Markovian errors on the gates. Any non-Markovian effect will also go unmodeled by the smaller models (crosstalk-free and context dependent). But certain crosstalk errors are also excluded by those models (again, by construction).

When data show evidence of non-Markovianity (as is often the case), none of the three models will fit the data well. But we can use wildcard error analysis to roughly estimate the magnitude of crosstalk errors even in the presence of non-Markovianity. To do so, we assign a wildcard error to each model. The general model serves as a baseline; only non-Markovian errors contribute to its W . A smaller model’s W accounts for both non-Markovian errors *and* the crosstalk errors excluded by that model. If the smaller model’s W is significantly higher, that indicates the presence of crosstalk errors that are not dominated by non-Markovianity. We will see examples of this scenario in the experimental data.

E. Metrics

The analysis in the preceding section can provide extensive high-level information about whether whole classes of crosstalk error are present or absent, and about their overall magnitude. But it also lets us select one of the three models as the best fit to the data—i.e., the one that best balances simplicity and explanatory power. Once this model has been selected, we examine the process matrices that it assigns to each gate. We can extract detailed performance metrics, identify dominant error channels, and/or use the process matrices to predict the processor’s performance on specific tasks and benchmarks.

Reductive gate error metrics like diamond distance or entanglement infidelity provide rough summaries of system performance, but to probe the details of estimated error models we transform process matrices to *error generators* [68]:

$$\mathcal{G} = e^{\mathcal{L}} \circ e^{\mathcal{H} + \Delta\mathcal{H}}. \quad (7)$$

Here \mathcal{G} is a gate’s process matrix, \mathcal{H} is a Hamiltonian superoperator that generates a perfect unitary implementation of the gate, $\Delta\mathcal{H}$ is a Hamiltonian error generator that generates the gate’s unitary errors [69], and \mathcal{L} is a nonunitary error generator that describes all nonunitary errors in the gate (see Ref. [68] for an extensive discussion). Using this notation, a gate’s diamond error is

$$\epsilon_\diamond = \frac{1}{2} \sup_{\rho} \|(\mathcal{G} \otimes \mathbb{I})(\rho) - (e^{\mathcal{H}} \otimes \mathbb{I})(\rho)\|_1, \quad (8)$$

where \mathbb{I} is the identity operator of dimension equal to that of \mathcal{G} and $e^{\mathcal{H}}$ is the target unitary gate. As discussed above, we use $\bar{\epsilon}_\diamond$ to denote the diamond error averaged over all gates in the gate set.

Gate sets like those we analyze here have a gauge freedom [53,70,71]; some of their parameters have no physical consequences and are unobservable. Gauge degrees of freedom appear in error generator representations as unobservable linear combinations of error generator coefficients. When we construct and examine estimates in this article, we manifest the gauge freedoms explicitly as unobservable constant offsets, and we measure crosstalk errors using strictly gauge-invariant properties constructed as differences between two coefficients with identical gauge freedoms. Gauge-invariant parameters of the Hamiltonian error generator include:

1. the coefficient (rate) of any error generator that commutes with the target gate, including
 - (a) the entire $\Delta\mathcal{H}$ for an idle operation,
 - (b) over- or under-rotation angles of any active gate,
2. the angle between the rotation axes of any two active gates,
3. the change in $\Delta\mathcal{H}$ between the same gate acting in two different contexts.

These will be sufficient for our analysis.

F. Testing quantum information processors for crosstalk

The previous sections each discussed an important element of a robust method for identifying and characterizing crosstalk errors in a quantum information processor. In this section we outline the steps taken for end-to-end characterization of crosstalk errors in the two experimental platforms discussed in the next section. For each processor, we collect all data in one contiguous experiment (details are given below), but here we present a step-by-step procedure for clarity.

First, to obtain a rough estimate of local and crosstalk error rates, we perform and analyze a form of simultaneous RB [41]. Simultaneous RB involves three distinct RB experiments: running RB on subsystem (1) while idling subsystem (2); idling (1) while running RB on (2); and running RB on (1) and (2) simultaneously. This yields two error rates (r_i and r_s , from the idle and driving contexts) for each subsystem. The change in each subsystem’s RB error rate ($r_s - r_i$), when the other subsystem is driven instead of idled, provides an estimate of how much error the gates on one subsystem induce on its neighbor. Simultaneous RB can be implemented with any variant of RB; we use *direct* RB (DRB), a variant of standard RB in which the Clifford

RB circuits [72] are replaced by uncompiled circuits over a system’s native gates [73].

Next, we analyze GST data. We fit our three models to this data (using `pyGSTi`), and evaluate their fit quality using N_σ , evidence ratios, and wildcard error. We use this information to deduce which forms of crosstalk are present, and to estimate their magnitude. In each use case, we then select the model that best explains the observed data, for further analysis using the techniques of Sec. E. The details of this analysis depend on the model.

- (i) Crosstalk-free model: if this model fits, there is no evidence for crosstalk. Each single-system gate is represented by a local process matrix, independent of context. It can be evaluated in the usual way.
- (ii) Context-dependent model: if this model is selected, the gates still act locally, but their action is context dependent. The model specifies the action of each single-system gate in several contexts, indexed by which operation is performed on the neighboring subsystem. The most relevant object of study is the *variation* in a gate’s action between contexts, which is gauge invariant and easily extracted from the local process matrices that describe it in different layers.
- (iii) General model: if this model is selected, at least one layer is inducing nonlocal (e.g., entangling or correlated) errors. Each process matrix must be analyzed to see if it produces correlated errors, and if so, what their kind and magnitude are.

III. EXPERIMENTAL DEMONSTRATION

We use our nested crosstalk models to investigate and characterize crosstalk errors on the two DOE quantum testbed platforms, AQT [55] and the QSCOUT prototype [56–58]. AQT is a transmon-based platform housed at Lawrence Berkeley National Laboratory and UC Berkeley. QSCOUT is a trapped-ion quantum computing platform housed at Sandia National Laboratories. By characterizing both devices, we are able to test the performance of our methods against vastly different physical error sources and experimental limitations.

We report experimentally measured values throughout this section. When possible, uncertainties are given, using concise notation, e.g., 1.234(5) indicates 1.234 ± 0.005 . All uncertainty intervals are 95% $\approx 2\sigma$ confidence intervals, obtained using either bootstrapping or likelihood ratio confidence interval methods.

A. The Advanced Quantum Testbed

Experiments on the AQT platform are performed using an eight-qubit superconducting transmon processor (AQT@LBNL Trailblazer8-v5.c2) [74]. The qubits are encoded as the $|0\rangle$ and $|1\rangle$ states of the transmons, and are coupled to their nearest neighbors in a ring

geometry. The demonstrations here focus on two next-nearest-neighbor transmons, labeled Q4 and Q6, whose fundamental transition frequencies range from 5.2 to 5.5 GHz, with anharmonicities around 270 MHz. Each qubit in the device has its own control line for applying $X_{\pi/2}$ and $Y_{\pi/2}$ gates, while any necessary Z gates are applied virtually through discrete phase shifts of subsequent $X_{\pi/2}$ and $Y_{\pi/2}$ gates [75].

1. Potential (and actual) sources of crosstalk in AQT

Crosstalk is a well-known problem in many transmon-based quantum processors. Two important crosstalk effects are coherent ZZ interactions, induced by shared microwave resonator modes, and pulse spillover. Microwave drive signals are often poorly localized on superconducting chips, so they rely on resonance mismatch to mitigate the impact of control spillover between qubits. However, this does not work perfectly, and the lingering interactions can manifest locally on spectator qubits as unwanted Rabi oscillations (if the drive is close to resonant with the spectator) or dispersive ac Stark shifts (if the drive is off-resonant). For neighboring qubits, such spillover crosstalk might even cause unwanted cross-resonance entangling interactions [76].

Using spectroscopic analysis, we find that the $|0\rangle\text{--}|1\rangle$ transition frequency of Q4 is nearly resonant with the $|1\rangle\text{--}|2\rangle$ transition of Q6. When microwave drive tones are applied to Q4, some of this power impinges on Q6 and this can therefore cause an ac Stark shift in the $|0\rangle\text{--}|1\rangle$ transition frequency of Q6. The size of this shift can be measured by driving Q4 on resonance while monitoring the frequency of Q6 via Ramsey spectroscopy.

We can correct for this drive-dependent Stark shift by adding an explicit *crosstalk compensation* pulse on Q6 that interferes destructively with the spillover pulse. This compensation pulse is optimized by first identifying the phase shift for which the Stark shift of Q6 is maximal, and then finding the relative amplitude that minimizes the error on Q6. This compensation tone is then built into each active operation that is applied to Q4.

2. Experiment design

To characterize crosstalk in the AQT platform, we run simultaneous DRB and GST experiments, both with and without crosstalk compensation. The GST circuit family we use is summarized in Table VI in Appendix A. We use circuits up to depth $\simeq 32$, resulting in 20,577 GST circuits in total. Our DRB circuits are constructed by sampling 30 two-qubit, simultaneous DRB circuits at each of eight exponentially spaced depths up to a maximum depth of 256. From each of these DRB circuits, we create two additional independent DRB circuits for which all of the gate operations on one or the other qubit are replaced with

idles, yielding 710 unique DRB circuits in total (if a circuit is duplicated, we simply gather twice as much data to avoid undersampling).

We combine the GST and DRB circuit sets into a single list and randomize its order (effectively interleaving the RB and GST circuits). Because neither GST nor RB are designed to be reliable when there are large drifts in a processor’s behavior over the duration of the experiment, we gather data in four batches. Crosstalk compensation is employed in alternate batches, and each batch consists of running every circuit in the combined circuit list 500 times. Using simple statistical tests described in Ref. [48], we confirm that the data in similar batches are statistically consistent. We then aggregate data from similar batches (with and without crosstalk compensation) into a single dataset containing 1000 shots per circuit. All data are taken over a period of approximately 10 h.

Our measurement is able to distinguish between $|0\rangle$, $|1\rangle$, and the leaked state $|2\rangle$. However, the GST and RB techniques used in this work are not designed to characterize leakage, so we discard any measurements for which either qubit is found in the $|2\rangle$ state. This requires discarding approximately 0.3% of all measurements when crosstalk compensation is used, and approximately 0.5% when it is not.

3. Randomized benchmarking

We use simultaneous DRB on AQT’s Q4 and Q6 to estimate (1) each qubit’s typical error per gate, and (2) how much that error rate changed when the other qubit is left idle or driven with random gate sequences. We call these “contexts.” Simultaneous DRB reveals each qubit’s error per gate in two contexts, the “spectator driven” context (r_s) and the “spectator idle” context (r_i). Its main result is the difference between these error rates ($r_s - r_i$), which we call the *context-to-context variation* of a qubit’s error per gate. Figure 2 summarizes the results, showing success probability decay curves for each qubit, in both contexts, and the estimated values of each r_s and r_i , with and without crosstalk compensation.

Without crosstalk compensation, we observe a striking asymmetry between the two qubits. Q4’s error per gate is approximately 0.07% regardless of context ($r_s \approx r_i$ and the estimated change of 0.006(6)% is within error bars of zero). But Q6’s error per gate shows significant context-to-context variation—it jumps from 0.082(5)% when Q4 is idled to 0.13(2)% when Q4 is driven. We conclude that there is significant crosstalk from Q4 to Q6, but no evidence for crosstalk in the opposite direction.

Enabling crosstalk compensation reduces Q6’s “spectator driven” error per gate to 0.100(9)%, without significantly changing other decay rates. So crosstalk compensation reduces the context-to-context variation of Q6’s error per gate from 0.05(2)% to 0.016(8)%.

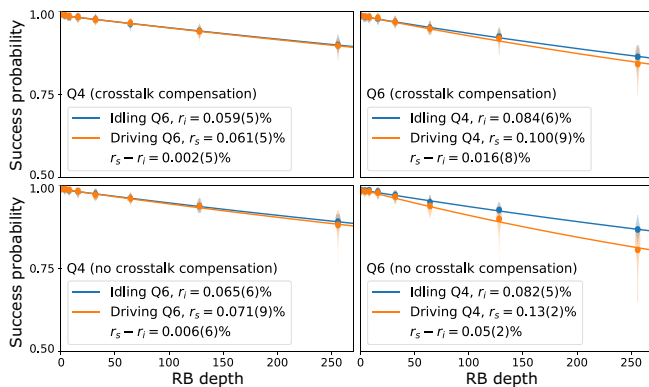


FIG. 2. Simultaneous DRB on the AQT platform. We run simultaneous DRB experiments (details in the main text) on the AQT device’s Q4–Q6 subsystem both with (top) and without (bottom) crosstalk compensation enabled. Estimates of each qubit’s average error-per-gate are extracted in two contexts: “idling spectator” (r_i) and “driving spectator” (r_s). The increase in each qubit’s r due to driving the spectator is a high-level measure of crosstalk-induced errors on that qubit. Uncertainties are 95% confidence intervals obtained from bootstrapping. Violin plot regions around each point indicate the distribution of the 30 individual DRB circuits whose average success probability is represented by the point. There is no statistically significant evidence of crosstalk-induced errors from Q6→Q4. However, there is significant evidence of crosstalk from Q4→Q6, and it is reduced (though not eliminated) by crosstalk compensation.

Simultaneous RB demonstrates that crosstalk is present, and that it is significant. It suggests that crosstalk effects are asymmetric (Q4→Q6), and that crosstalk compensation reduces them. But it is hard to draw further conclusions, because simultaneous RB does not reveal how each gate’s action depends on context, nor does it distinguish context-dependent errors from entangling or correlated ones. It is possible to extract some of this information from a more sophisticated analysis of RB data [42], but we can answer these questions definitively with GST data and analysis.

4. Comparing and selecting crosstalk models

We begin our analysis of the GST data by fitting all three models—crosstalk-free, context dependent, and general—to the two datasets (with and without crosstalk compensation), and evaluating their fit quality. Table II displays the results.

None of the three models fit either dataset perfectly. We evaluate each model’s fit quality by its log-likelihood ratio λ with respect to the maximal model [see Eq. (3)]. All six fits display $N_\sigma > 45$, i.e., at least 45σ of model violation. This constitutes strong statistical evidence of non-Markovian behavior that cannot be modeled by two-qubit process matrices. Under these circumstances, neither GST nor RB is guaranteed to be reliable, and caution is required when interpreting results.

However, large N_σ does not imply that non-Markovian errors are dominant. To compare the rates of modeled and unmodeled errors, we assign a wildcard error (see Sec. D) to each model’s best-fit estimate. The general model’s wildcard error rate is less than 0.2%, both with and without crosstalk compensation. This is much smaller than the GST model’s average diamond distance error rate [$\bar{\epsilon}_\diamond = 1.670(3)\%$ without crosstalk compensation, or $\bar{\epsilon}_\diamond = 1.097(3)\%$ with crosstalk compensation]. We conclude that Markovian errors dominate, and are captured by the largest GST model.

Since the general GST model explains most of the errors observed in the GST data, we investigate whether smaller models—context dependent and crosstalk-free—are equally consistent with the data. We can evaluate their fit relative to the general model using two criteria: evidence ratios or the change in wildcard error.

The context-dependent model is conclusively accepted by both criteria—it fits the data as well as the general model despite having more than 1400 fewer parameters. The evidence ratio between them is $\gamma = 0.9$ ($\gamma = 0.58$) without (with) crosstalk compensation, and the wildcard error increases by only 0.014%. We conclude there is *no* evidence for entangling or correlated crosstalk errors; the context-dependent model’s tensor product process matrices describe the observed data as well as possible by any (two-qubit) process matrices.

The crosstalk-free model, however, does not fit the data well. In the absence of crosstalk compensation, it is overwhelmingly rejected by both criteria—the evidence ratio between the context-dependent and crosstalk-free models is $\gamma = 495$, and the wildcard error required to reconcile it with the data is increased by more than $5\times$ ($W_{\text{crosstalk-free}} = 0.85\% \gg W_{\text{context dependent}} = 0.15\%$). This constitutes overwhelming evidence that crosstalk errors are present, and their magnitude is large (i.e., they make a substantial contribution to the total gate error rates).

When crosstalk compensation is enabled, the crosstalk-free model fits the data much better. Its wildcard error drops substantially ($W_{\text{crosstalk-free}} = 0.23\% \approx W_{\text{context dependent}} = 0.20\%$), and the evidence ratio in favor of the context-dependent model drops to $\gamma = 15.6$. This constitutes clear evidence that crosstalk is still present, but its magnitude and significance are greatly reduced by crosstalk compensation.

Our conclusions from this analysis are as follows.

- (i) Crosstalk errors are clearly present, but only local context-dependent errors.
- (ii) Non-Markovian errors are present at the 0.2%/gate level, but are dominated by Markovian errors (and by Markovian crosstalk errors).
- (iii) Crosstalk compensation reduces crosstalk errors significantly, but it does not eliminate them.

TABLE II. Fit quality metrics for three models on the AQT device. We tabulate metrics of fit quality and unmodeled error for three distinct models when fit to GST data from the AQT platform (both with and without crosstalk compensation). The performance of these models can be compared in detail using the log-likelihood ratio score λ_{LR} , the N_σ of model violation, and the residual wildcard error W . Each of these models further predicts an average diamond error $\bar{\epsilon}_\diamond$ for the gate set. Without crosstalk compensation, the context-dependent model is strongly preferred over the crosstalk-free model by the evidence ratio test γ , and the wildcard error is significantly reduced by moving to the larger model. The full Markovian model, however, requires nearly an order of magnitude more parameters, achieves only a small improvement in the likelihood ratio and the wildcard, and is rejected by the evidence ratio test. When crosstalk compensation is applied, the context-dependent model is again preferred by the evidence ratio test, but much more weakly, and the wildcard error is nearly constant across models. Crosstalk compensation further results in a $\simeq 35\%$ reduction in the average gate error.

Model	Without crosstalk compensation						With crosstalk compensation				
	N_p	N_σ	γ	$\lambda_{\text{LR}}[10^3]$	$W[10^{-3}]$	$\bar{\epsilon}_\diamond[10^{-3}]$	N_σ	γ	$\lambda_{\text{LR}}[10^3]$	$W[10^{-3}]$	$\bar{\epsilon}_\diamond[10^{-3}]$
General	1697	46.28	...	76.02	1.39	16.7(1)	49.92	...	77.21	1.89	11.0(1)
Context dependent	230	45.90	0.90	77.60	1.53	16.5(1)	47.70	0.58	78.23	2.03	10.6(1)
Crosstalk-free	86	248.27	494.60	148.82	8.54	15.94(5)	53.63	15.58	80.48	2.29	10.21(8)

- (iv) Both with and without crosstalk compensation, the best estimate (process matrices) to examine in detail is the context-dependent estimate.

5. Extracting detailed crosstalk error rates

We now examine the best GST estimates (with and without crosstalk compensation) in detail. A GST estimate of the context-dependent model specifies a two-qubit process matrix for each of the nine parallel-gate layers (e.g., $X_{\pi/2} \otimes I, Y_{\pi/2} \otimes X_{\pi/2}$, etc). Each two-qubit process matrix is the tensor product of two one-qubit process matrices. Each one-qubit process matrix describes one of three gates ($X_{\pi/2}, Y_{\pi/2}$, or I) acting on one of two qubits (Q4 or Q6), in one of three contexts ($X_{\pi/2}, Y_{\pi/2}$, or I applied to the other qubit). We represent each one-qubit process matrix using error generators (Sec. E), and examine the context-to-context variation of the error generators for each of the six single-qubit gates ($X_{\pi/2}, Y_{\pi/2}$, and I gates on Q4 and Q6).

The nonunitary parts of each error generator—e.g., rates of stochastic errors—have almost no statistically significant context-to-context variation. The very small number of variations that *are* statistically significant are comparable in magnitude to the non-Markovian errors (measured by the wildcard error at $W \approx 0.2\%$). In contrast, context-to-context variations in the unitary part of the gates' error generators [$\Delta\mathcal{H}$ in Eq. (7)] are both statistically significant and large.

Table III presents the coefficients of coherent X, Y , and Z Hamiltonians (in milliradians) for each gate in each context, both without (top) and with (bottom) crosstalk compensation enabled. The target rotation angle $\theta_0 = \pi/2 \times 10^3$ mrad is included where appropriate. This gate set has six gauge freedoms (corresponding to three-parameter unitary changes of basis on each qubit), which are reflected in this table by the unobservable constants c_1, \dots, c_6 . The key to interpreting these error rates is that constants (θ_0 and c_i)

appear identically in each column. So context-to-context variations in each gate—i.e., differences between entries in the same row—are gauge invariant.

When no crosstalk compensation is applied, every gate displays *some* statistically significant context-to-context variation. However—as suggested by simultaneous RB results—the variations are significantly larger for Q6. Its idle operation (I) experiences phase (Z) errors that change by 13.1(2) mrad when the spectator qubit is driven. Active gates ($X_{\pi/2}, Y_{\pi/2}$) on Q6 display similar variations in their rotation angles [e.g., Y errors on the $Y_{\pi/2}$ gate, which vary by 9.5(1) mrad] or axes [e.g., Y errors on the $X_{\pi/2}$ gate, which vary by 14.4(1.0) mrad]. In contrast, the largest variation observed in a Q4 gate is the rotation angle of the $X_{\pi/2}$ gate, which varies by 4.6(5) mrad. Figure 3 illustrates the variation in the effective Hamiltonians that generate $X_{\pi/2}$ and $Y_{\pi/2}$ gates on Q6, in the X - Y plane.

The context-to-context variations do not follow a simple pattern. For the idle gates, only phase (Z) errors are observed, and they depend significantly only on whether the spectator qubit is driven or not (rather than on which gate is performed on the spectator). For the active gates, however, both rotation angles and axes vary, and they depend not just on whether the spectator is driven, but on whether an $X_{\pi/2}$ or $Y_{\pi/2}$ gate is performed on the spectator. This detailed information about the nature of the crosstalk could in principle be compared to—or used to inform—physical models of gate context dependence, but we do not currently have such a model.

Enabling crosstalk compensation reduces the overall crosstalk significantly, but not uniformly. It has little effect on Q4's gates (which are already relatively good), but eliminates (a) essentially all context dependence for Q6's $Y_{\pi/2}$ and I gates, and (b) essentially all variation in the rotation angle of Q6's $X_{\pi/2}$ gate. Interestingly, variations in the $X_{\pi/2}$ gate's rotation axis are not eliminated (see Fig. 3).

TABLE III. Estimated rates (in milliradians) of all Hamiltonian (unitary) errors, on all gates, in all contexts, for the context-dependent model of the AQT platform. Each column of three values is the three components \vec{h} of the effective Hamiltonian that generates the unitary part of the estimated process (see the main text), represented as $H_{\text{eff}} = \vec{h} \cdot \vec{\sigma}$ [with $\vec{\sigma} = (X, Y, Z)$]. This gate set has six gauge freedoms, which are reflected in this table by the unobservable constants c_1, \dots, c_6 . Any value, or linear combination of values, containing *no* c 's is gauge invariant, and therefore physically meaningful. All values have been shifted relative to the target value $\theta_0 = \pi/4 \text{ rad} \simeq 785.4 \text{ mrad}$.

Gate on target qubit	Gate on spectator qubit (context)		
	I	$X_{\pi/2}$	$Y_{\pi/2}$
Without crosstalk compensation			
$I^{(4)}$	0.1(0.1)	0.2(0.1)	0.3(0.1)
	0.1(0.1)	0.1(0.1)	0.1(0.1)
	0.9(0.2)	1.2(0.2)	1.7(0.2)
$X_{\pi/2}^{(4)}$	$\theta_0 + 3.6(0.2)$	$\theta_0 + 3.7(0.1)$	$\theta_0 + 3.7(0.1)$
	$4.4(0.4) + c_1$	$0.6(0.4) + c_1$	$-0.2(0.4) + c_1$
	$1.0(4.1) + c_2$	$-1.9(4.1) + c_2$	$0.9(4.1) + c_2$
$Y_{\pi/2}^{(4)}$	$3.3(0.4) - c_1$	$0.3(0.4) - c_1$	$1.3(0.4) - c_1$
	$\theta_0 + 3.6(0.2)$	$\theta_0 + 3.5(0.1)$	$\theta_0 + 3.2(0.1)$
	$-2.8(5.6) + c_3$	$1.6(5.6) + c_3$	$1.2(5.6) + c_3$
$I^{(6)}$	$-0.1(0.1)$	0.2(0.1)	0.2(0.1)
	$-0.1(0.1)$	$-0.3(0.2)$	$-0.2(0.2)$
	3.0(0.2)	$-10.2(0.1)$	$-10.1(0.1)$
$X_{\pi/2}^{(6)}$	$\theta_0 + 6.6(0.1)$	$\theta_0 + 15.8(0.1)$	$\theta_0 + 15.8(0.1)$
	$-8.2(0.3) + c_4$	$-2.2(0.3) + c_4$	$-3.6(0.3) + c_4$
	$8.1(0.7) + c_5$	$-1.8(0.7) + c_5$	$-6.3(0.8) + c_5$
$Y_{\pi/2}^{(6)}$	$-3.6(0.3) - c_4$	$-4.6(0.3) - c_4$	$-5.9(0.3) - c_4$
	$\theta_0 + 6.9(0.1)$	$\theta_0 + 16.4(0.1)$	$\theta_0 + 16.1(0.1)$
	$7.3(12.1) + c_6$	$-3.5(12.2) + c_6$	$-3.8(12.2) + c_6$
With crosstalk compensation			
$I^{(4)}$	0.0(0.1)	0.1(0.1)	0.3(0.1)
	0.1(0.1)	0.1(0.1)	0.0(0.1)
	1.3(0.2)	1.8(0.2)	1.8(0.2)
$X_{\pi/2}^{(4)}$	$\theta_0 - 3.5(0.2)$	$\theta_0 - 3.6(0.1)$	$\theta_0 - 3.4(0.1)$
	$4.2(0.3) + c_7$	$0.1(0.3) + c_7$	$-0.8(0.3) + c_7$
	$0.9(1.6) + c_8$	$-1.5(1.6) + c_8$	$0.6(1.6) + c_8$
$Y_{\pi/2}^{(4)}$	$2.9(0.3) - c_7$	$0.1(0.3) - c_7$	$0.5(0.3) - c_7$
	$\theta_0 - 3.6(0.2)$	$\theta_0 - 3.6(0.1)$	$\theta_0 - 4.1(0.1)$
	$-2.6(3.3) + c_9$	$1.4(3.3) + c_9$	$1.3(3.3) + c_9$
$I^{(6)}$	0.1(0.1)	0.1(0.1)	$-0.1(0.1)$
	$-0.2(0.1)$	$-0.2(0.1)$	$-0.3(0.1)$
	2.4(0.2)	1.6(0.2)	1.3(0.2)
$X_{\pi/2}^{(6)}$	$\theta_0 + 6.9(0.3)$	$\theta_0 + 7.0(0.2)$	$\theta_0 + 7.3(0.2)$
	$-6.9(0.3) + c_{10}$	$-2.9(0.3) + c_{10}$	$-5.0(0.3) + c_{10}$
	$3.3(7.1) + c_{11}$	$0.6(7.1) + c_{11}$	$-3.9(7.1) + c_{11}$
$Y_{\pi/2}^{(6)}$	$-5.1(0.3) - c_{10}$	$-4.7(0.3) - c_{10}$	$-5.0(0.3) - c_{10}$
	$\theta_0 + 7.3(0.2)$	$\theta_0 + 7.6(0.2)$	$\theta_0 + 7.3(0.2)$
	$1.5(6.7) + c_{12}$	$-0.8(6.7) + c_{12}$	$-0.7(6.7) + c_{12}$

6. Discussion of crosstalk in AQT

The asymmetric crosstalk errors identified by our GST experiments are consistent with ac Stark shifts induced by the control fields. As discussed in Sec. 1, spectroscopic data predicted that driving Q6 should not influence Q4 because of the large discrepancy in relevant transition frequencies. Conversely, that same analysis predicted Q6 should experience phase shifts when Q4 is driven,

because the Q4 drive tone is near resonant with an excited state transition of Q6. The crosstalk errors we observe are consistent with these predictions.

B. The Quantum Scientific Computing Open User Testbed

Experiments on the QSCOUT platform are performed using a prototype system that is nearly identical to the

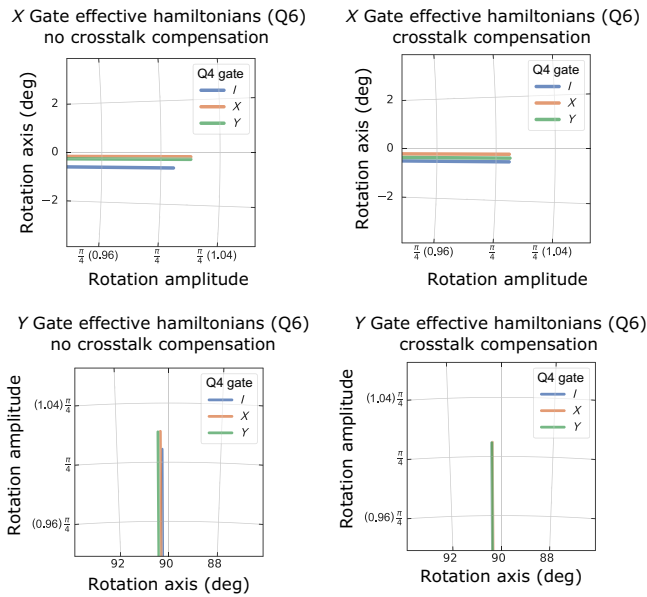


FIG. 3. Coherent errors in $X_{\pi/2}$ and $Y_{\pi/2}$ gates on AQT Q6 depend on context. We use GST to estimate one-qubit process matrices describing the $X_{\pi/2}$ (top) and $Y_{\pi/2}$ (bottom) gates on Q6 conditional on $\{I, X_{\pi/2}, \text{ and } Y_{\pi/2}\}$ gates performed at the same time on Q4, both without (left) and with (right) crosstalk compensation. We extract the effective Hamiltonian that generates the unitary part of each estimated process (see the main text), represent it as $H_{\text{eff}} = \hbar \cdot \vec{\sigma}$ [with $\vec{\sigma} = (X, Y, Z)$], and plot the projection of each one into the X - Y plane (all h_Z components are negligible). Each of the four panels shows a small region of the X - Y plane; each effective Hamiltonian is represented by a line from the origin to (h_X, h_Y) . Uncertainty regions are shown as ellipses, which are too small to be visible here. Both rotation angles and axes vary significantly from context to context; crosstalk compensation reduces this effect, and essentially eliminates it for the $Y_{\pi/2}$ gate.

deployed testbed [56,57]. It is configured to use two qubits encoded as the hyperfine clock states of a pair of $^{171}\text{Yb}^+$ ions [77] that are held in a Sandia-fabricated HOA 2.0 surface ion trap [78]. The ions are trapped together in a single pseudopotential and are spaced $4.5 \mu\text{m}$ apart. The trap frequencies are approximately 1 MHz axially and 2 MHz radially. The ions' hyperfine states are manipulated using a two-photon Raman transition via a pair of phase-locked copropagating frequency combs generated by a frequency-tripled Nd:YAG pulsed laser at 355 nm [79]. Each ion is individually addressed by first splitting a single laser beam into multiple beams, and sending each beam through a dedicated channel of a multichannel acousto-optic modulator (AOM) that allows for independent frequency, phase, and amplitude control. Each beam is then tightly focused using custom optics to a $0.8 \mu\text{m}$ axial waist radius, ideally impinging on only a single target ion. During the detection cycle, light from each ion is focused to a different core of a multicore fiber. Each core is then sent to its own

photomultiplier tube, allowing for distinguishable detection of the ions.

1. Potential (and actual) sources of crosstalk in QSCOUT

Several physical phenomena are expected to manifest as crosstalk errors in the QSCOUT hardware. The first is straightforward: control lasers targeted at one ion have a nonzero beam waist, allowing some light to spill over onto neighboring ions. Because all ions in the system are at the same resonant frequency, this light can cause slow, coherent Rabi oscillations on the neighbor ions. The laser waist is small relative to the ion spacing, so one might expect this effect to be relatively small. But its magnitude can be significant if the system is poorly aligned. A more subtle source of potential crosstalk is the internal dynamics of the multichannel AOM. The nonlinear optical crystals in the AOM are not perfectly isolated from one another, so acoustic drive tones applied to a target crystal can cause neighboring crystals to ring sympathetically. The neighboring channel will then be activated (or perturbed), and its target ion will experience an error. The signals that implement these drive tones can also electrically couple into neighboring channels. The wiring layout of the AOM predicts that next-nearest-neighbor correlations will be greatest. Correlated errors arising from common causes can also manifest as crosstalk. In the trapped-ion hardware we expect amplitude (or phase) fluctuations of the driving laser to result in correlated amplitude (or phase) errors at the qubits. Similarly, magnetic field fluctuations can result in correlated phase errors.

2. Experiment design

We use the same family of GST circuits for QSCOUT as for AQT (see Table VI in Appendix A). Because trapped-ion operations are slower, we take less data on the QSCOUT device. The number of counts per circuit is reduced from 1000 to 80, and the GST circuits are limited to depth $\simeq 8$ (rather than $\simeq 32$). We run simultaneous DRB circuits generated in exactly the same way as for AQT (maximum depth 256, 30 circuits at each of eight logarithmically spaced depths), and interleave them with the GST circuits in the same way. A total of 12,514 unique circuits are run in the QSCOUT experiment, including 11,813 GST circuits and 701 unique DRB circuits. Data are taken in eight batches, each comprising ten shots of every circuit. The device is recalibrated between each batch. Approximately one million individual circuit shots are performed over approximately 4 h.

The circuits we run on the QSCOUT platform are implemented using simple pulses, with no compensating pulse sequences of any kind. The QSCOUT hardware can implement gates using, e.g., BB1 composite pulse sequences

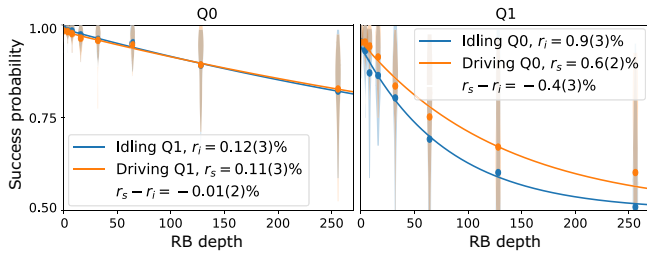


FIG. 4. Simultaneous direct randomized benchmarking on the QSCOUT platform. We run simultaneous DRB experiments (see the caption to Fig. 2) on a two-qubit QSCOUT processor. No crosstalk compensation is used in this experiment. Q0 performs well independent of context (i.e., whether Q1 is driven). In contrast, Q1 performs at least 5 times worse in both contexts, and its error per gate appears to depend on whether Q0 is driven. Oddly, however, driving Q0 actually reduces Q1’s error rate, suggesting a negative rate of crosstalk-induced errors. We identify the particular calibration protocol used as the probable cause (see the main text).

[80] that can yield significantly reduced error rates. However, they also require more time to implement, complicate the interpretation of unitary error generators, and can reduce the magnitude of certain crosstalk errors below detectable thresholds. So, for the purposes of this work, we restrict gate operations to simple, bare pulses.

3. Randomized benchmarking

We run simultaneous DRB on the QSCOUT platform to estimate the context-to-context variation of each qubit’s error per gate. Figure 4 shows the results for both Q0 and Q1. We observe an even more dramatic asymmetry than in the RB experiments on AQT; Q0’s error per gate is approximately 0.11(3)% regardless of whether Q1 is driven, but Q1’s error per gate is far higher [0.6(2)% – 0.9(3)%], and varies dramatically depending on whether Q0 is driven. Perhaps surprisingly, driving Q0 actually *improves* Q1’s performance. We hypothesize that this is a consequence of the calibration procedure. Only the $X_{\pi/2}^{(0)} X_{\pi/2}^{(1)}$ operation is calibrated. The $Y_{\pi/2}$ gates use the same calibrated pulses, but phase shifted by $\pi/2$, so effectively $Y_{\pi/2}^{(0)} Y_{\pi/2}^{(1)}$ is also calibrated. This calibration procedure optimizes the $X_{\pi/2}$ and $Y_{\pi/2}$ gates on each qubit for performance in the very specific context where the *same* gate is performed on the other qubit. Other circuit layers—e.g., $X_{\pi/2}^{(0)} Y_{\pi/2}^{(1)}$ or $X_{\pi/2}^{(0)} I^{(1)}$ are not necessarily well calibrated, and may experience different and/or larger crosstalk errors. Independent DRB probes the performance of layers that are strictly uncalibrated (e.g., $X_{\pi/2}^{(0)} I^{(1)}$, $Y_{\pi/2}^{(0)} I^{(1)}$, and $I^{(0)} I^{(1)}$ for Q0). Approximately one-third of the circuit layers used in simultaneous DRB have been calibrated, predicting that the simultaneous circuits should perform better—which is what we observe. We shall see further consequences of this calibration protocol in the GST data.

4. Comparing and selecting crosstalk models

We fit all three gate set models to the data (just one dataset in this case, since no crosstalk compensation is performed) and evaluate their fit quality. Table IV displays the results.

No model fits the data perfectly. The general model displays about 9σ of model violation, suggesting that the gates are somewhat non-Markovian. However, the wildcard error assigned to the general model is almost zero, suggesting that this non-Markovianity is barely visible. These results should not be compared directly to AQT—far less data are taken on QSCOUT, so the experiment is less able to detect and quantify non-Markovianity.

The evidence ratio between the general model and the (smaller) context-dependent model is extraordinarily small ($\gamma = 0.29$), indicating no evidence whatsoever of entangling or correlating crosstalk errors between the qubits [81]. The observed data can be described more or less perfectly by context-dependent local errors.

However, the evidence ratio between the context-dependent model and the crosstalk-free model clearly rejects the crosstalk-free model ($\gamma = 134$). This constitutes clear and overwhelming evidence of crosstalk (as expected, given the simultaneous RB results). We can also estimate the magnitude of the crosstalk errors from the wildcard error required to reconcile the best crosstalk-free estimate with the data ($W = 2.4\%$). Since the context-dependent model fits the data well, we can ascribe all crosstalk errors to context-to-context variation of local errors, which we proceed to examine in detail.

5. Extracting detailed crosstalk error rates

The GST estimate of the context-dependent model yields one-qubit process matrices describing each of the six gates in three different contexts. We analyze the error generators for these processes. As in the AQT analysis, the nonunitary error generators reveal almost no significant context-to-context variation, so we focus on unitary (coherent) errors.

Table V presents the coefficients of coherent X , Y , and Z Hamiltonians (in milliradians) for each gate in each context. The target rotation angle $\theta_0 = \pi/2 \times 10^3$ mrad is included where appropriate. This gate set has six gauge freedoms (corresponding to three-parameter unitary changes of basis on each qubit), which are reflected in this table by the constants c_1, \dots, c_6 , which are unobservable. Here θ_0 and c_i appear identically in each column, so context-to-context variations in each gate—i.e., differences between entries in the same row—are automatically gauge invariant.

We can immediately draw the following conclusions from Table V.

TABLE IV. Details of the three best-fit gate models on the QSCOUT device. The performance of these models can be compared in detail using the log-likelihood ratio score λ_{LR} , the N_σ of model violation, and the residual wildcard error W . Each of these models further predicts an average diamond error $\bar{\epsilon}_\diamond$ for the gate set. The context-dependent model is strongly preferred over the crosstalk-free model by the evidence ratio test γ , and the wildcard error is significantly reduced by moving to the larger model. The general crosstalk model, however, requires nearly an order of magnitude more parameters, achieves only a small improvement in the likelihood ratio and the wildcard, and is rejected by the evidence ratio test. Notably, W is nonzero for the general model despite being zero for the smaller context-dependent model. This can occur because the wildcard error calculation requires models with more parameters to fit better than those with fewer.

Model	N_p	N_σ	γ	$\lambda_{\text{LR}}[10^3]$	$W[10^{-3}]$	$\bar{\epsilon}_\diamond[10^{-3}]$
General	1697	9.33	...	36.17	0.02	59(1)
Context dependent	230	5.27	0.29	36.61	0	57(1)
Crosstalk-free	86	134.02	139.38	70.99	24.1	34.5(8)

- (i) Errors on both idle gates are small ($\leq 3 \times 10^{-3}$ radians) in all contexts, and almost exclusively indistinguishable from zero. The I gate on Q0 has no evident errors at all. The I gate on Q1 shows barely significant context-dependent rotations by about 3 mrad.
- (ii) The Z Hamiltonian error rates on each qubit's $X_{\pi/2}$ and $Y_{\pi/2}$ gates are also effectively negligible—they are all < 4 mrad in all contexts, and mostly indistinguishable from zero.

The remaining errors, with magnitudes of up to 0.1 radians (100 mrad), fall into two categories.

1. Over- or under-rotation errors, i.e., X Hamiltonian errors on $X_{\pi/2}$ gates and Y Hamiltonian errors on $Y_{\pi/2}$ gates.
2. “Tilt” errors that change a gate’s rotation axis, i.e., Y Hamiltonian errors on $X_{\pi/2}$ gates and X Hamiltonian errors on $Y_{\pi/2}$ gates.

All of these errors show significant context-to-context variation. Figure 5 illustrates this variation by depicting each gate’s angle and axis in the X - Y plane for each context. Gates on Q0 have rotation angles that vary by up to 33(7) mrad, and rotation axes that vary by up to 28(2) mrad. For gates on Q1, rotation angles vary by up to 170(40)

TABLE V. Estimated rates (in milliradians) of all Hamiltonian (unitary) errors, on all gates, in all contexts, for the context-dependent model of the QSCOUT platform. Each column of three values is the three components \vec{h} of the effective Hamiltonian that generates the unitary part of the estimated process (see the main text), represented as $H_{\text{eff}} = \vec{h} \cdot \vec{\sigma}$ [with $\vec{\sigma} = (X, Y, Z)$]. This gate set has six gauge freedoms, which are reflected in this table by the unobservable constants c_1, \dots, c_6 . Any value, or linear combination of values, that does *not* contain any c ’s is gauge invariant, and therefore physically meaningful. All values have been shifted relative to the target value, $\theta_0 = \pi/4$ rad $\simeq 785.4$ mrad, where appropriate.

Gate on target qubit	Gate on spectator qubit (context)		
	I	$X_{\pi/2}$	$Y_{\pi/2}$
$I^{(0)}$	0.8(1.8)	−0.5(1.3)	−0.6(1.5)
	1.0(1.8)	1.3(1.3)	−0.4(1.5)
	1.5(2.5)	0.4(2.0)	1.9(2.2)
$X_{\pi/2}^{(0)}$	$\theta_0 - 12.0(1.4)$	$\theta_0 - 4.6(1.2)$	$\theta_0 - 37.4(6.4)$
	$1.2(1.1) + c_1$	$1.5(1.1) + c_1$	$22.8(1.1) + c_1$
	$-0.3(1.1) + c_2$	$-1.3(1.1) + c_2$	$1.6(1.1) + c_2$
$Y_{\pi/2}^{(0)}$	$0.8(1.1) - c_1$	$26.6(1.1) - c_1$	$-1.8(1.1) - c_1$
	$\theta_0 - 12.5(1.7)$	$\theta_0 - 28.5(7.4)$	$\theta_0 - 5.6(1.2)$
	$-2.8(1.1) + c_3$	$1.2(1.1) + c_3$	$1.6(1.1) + c_3$
$I^{(1)}$	0.7(1.7)	−1.6(1.5)	−2.6(1.5)
	0.6(1.8)	3.2(1.3)	−1.4(1.7)
	0.9(2.5)	−0.5(2.1)	0.3(2.4)
$X_{\pi/2}^{(1)}$	$\theta_0 + 70.2(13.5)$	$\theta_0 - 14.3(5.8)$	$\theta_0 - 101.5(30.3)$
	$46.5(1.0) + c_4$	$-19.7(1.0) + c_4$	$-104.7(1.0) + c_4$
	$1.1(1.0) + c_5$	$0.5(1.0) + c_5$	$-1.7(1.0) + c_5$
$Y_{\pi/2}^{(1)}$	$-53.4(1.0) - c_4$	$-36.5(1.0) - c_4$	$13.9(1.0) - c_4$
	$\theta_0 + 70.7(15.4)$	$\theta_0 - 7.9(10.7)$	$\theta_0 - 9.1(4.2)$
	$-2.5(1.0) + c_6$	$-1.6(1.0) + c_6$	$3.9(1.0) + c_6$

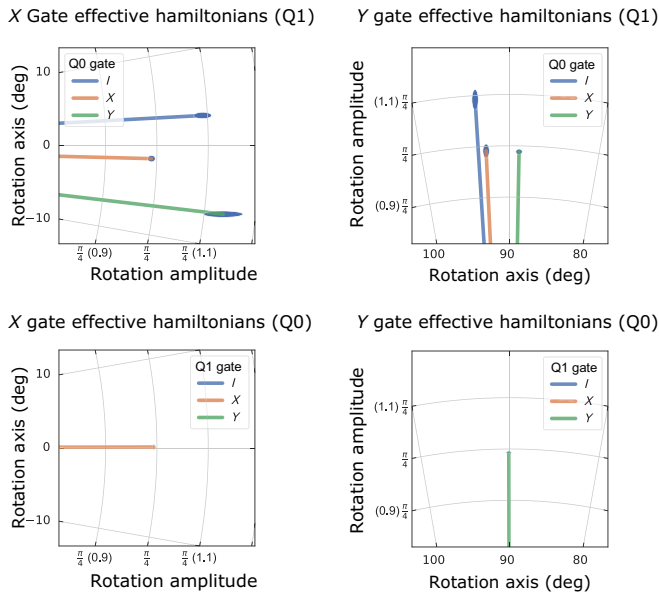


FIG. 5. Coherent errors in $X_{\pi/2}$ and $Y_{\pi/2}$ gates on QSCOUT qubits depend on context. We use GST to estimate one-qubit process matrices describing the $X_{\pi/2}$ (left) and $Y_{\pi/2}$ (right) gates on Q0 (bottom) and Q1 (top), conditional on $\{I, X_{\pi/2}, \text{ and } Y_{\pi/2}\}$ gates performed at the same time on the other qubit, in the QSCOUT system. Exactly as for Fig. 3, we extract the effective Hamiltonian that generates the unitary part of each estimated process (see the main text), represent it as $H_{\text{eff}} = \vec{h} \cdot \vec{\sigma}$ [with $\vec{\sigma} = (X, Y, Z)$], and plot the projection of each one into the X - Y plane (all h_Z components are negligible). Each of the four panels shows a small region of the X - Y plane; each effective Hamiltonian is represented by a line from the origin to (h_X, h_Y) . Uncertainty regions are shown as ellipses. Q0’s gates show essentially no context dependence. In contrast, the rotation angles and axes of the $X_{\pi/2}$ and $Y_{\pi/2}$ gates on Q1 show extremely strong dependence on the gate performed on the spectator. We also observe small context-dependent phase errors in the I gate (not shown here).

mrad, and rotation axes vary by up to 150(1) mrad. The context dependence of errors on Q1’s gates are approximately 3 – 5 times larger than Q0’s. Because these errors are coherent, the error per gate observed in RB scales as θ^2 , and so in RB experiments we expect to see 10–25 times more context-dependent error on Q1 than Q0. This is consistent with the observed results of simultaneous RB (Fig. 4).

The GST estimate confirms the conjecture we stated in the discussion of simultaneous RB results: errors are minimized when the same active gate ($X_{\pi/2}$ or $Y_{\pi/2}$) is performed on both qubits. These are the only layers that are explicitly calibrated. Active gates performed in other contexts show clear and significant calibration errors in both their rotation angles and their rotation axes (relative to the calibrated operation).

6. Discussion of QSCOUT results

Simultaneous RB experiments clearly demonstrate the existence of $Q0 \rightarrow Q1$ crosstalk, but also demonstrate the counterintuitive result that gates on Q1 perform better when Q0 is driven. GST experiments reveal what is actually happening: each active gate’s behavior depends not just on the “spectator driven” and “spectator idled” contexts, but on exactly what gate is performed on the spectator. Most of this effect is due to the calibration protocol, and the fact that only $X_{\pi/2}X_{\pi/2}$ and $Y_{\pi/2}Y_{\pi/2}$ gates are specifically calibrated. However, Table V shows additional variations between the other two contexts.

The crosstalk errors we observe in the QSCOUT system reflect the fact that the two qubits have identical energy splitting. Pulse spillover onto a spectator ion is therefore resonant with its qubit transition, leading to coherent Rabi oscillations around an axis on the equator of the Bloch sphere. In contrast, the AQT qubits have different frequencies, so spillover crosstalk acts very differently. Instead of coherent Rabi oscillations, it induces an ac Stark shift on the spectator qubit, which manifests as coherent rotation about the Z axis. The asymmetry in the QSCOUT error rates (particularly the errors in the idle gates) support a hypothesis that the crosstalk arises from beam pointing errors, rather than internal AOM phenomena, which are more likely to result in crosstalk errors that are symmetric between qubits.

Restricting device calibration to parallel operations has serious, observable impacts on the performance on non-parallel layers. This suggests that the active gates’ error rates could be reduced significantly *and* made less context dependent by explicitly calibrating all of the layers simultaneously. This would require a more complicated tune-up process.

IV. DISCUSSION

This paper presents both a novel device characterization method, and the results of its deployment on two experimental platforms. We separately discuss our conclusions about this method and our insights into the experimental results.

A. Protocols and methods

GST is best known as a replacement for process tomography, and for its use in constructing full-dimensional process matrices that represent gates’ action on their target space. But GST is inherently flexible [54], and we use that flexibility here to test *multiple* error models—ranging from full-dimensional process matrices down to a highly restricted crosstalk-free model—fits to data generated using parallel gates on two subsystems. This targeted adaptation of GST to probe crosstalk allows us to extract a lot of information at many levels of detail, ranging

from “Yes, there is crosstalk, but it is not entangling” down to the exact details of how much over-rotation each gate induces on the spectator.

Our results should be easy to reproduce, extend, and deploy in many experimental systems. They are explicitly platform agnostic, and they require only user-level access to a device and the ability to run simple circuits composed of device-native quantum operations. The data analysis routines are all built from free, widely available software tools, and can often be implemented with just a few lines of PYTHON code (e.g., `pyGSTi`). All of the experimental data and analysis code necessary to reproduce the results shown here are available upon request as supplementary material.

The investigation and results presented here highlight the complexity of crosstalk errors. If we had stopped after running simultaneous RB, we would have concluded simply that (1) the AQT qubits had a little bit of crosstalk and (2) the QSCOUT qubits had a small *negative* amount of crosstalk. The results of our detailed GST investigation do not contradict those findings—but they illustrate that crosstalk is not described by a single number. Crosstalk *changes* errors—e.g., depending on their context—which can cause unexpected harm irrespective of which context (e.g., idle neighbor or driven neighbor) induces the worse error. Eliminating crosstalk means removing all forms of context dependence entirely, not just ensuring that “idle” and “driven” randomized benchmarking experiments yield the same error per gate.

The approach presented here complements high-level crosstalk benchmarks, such as simultaneous RB, by constructing detailed models that can identify specific errors. This low-level diagnostic information can elucidate the physics of quantum information processors, enable better calibration, and inform design of next-generation systems. Detailed error models also enable more accurate estimates of device performance on real tasks, and noise-optimized decoders for quantum error correction.

These advantages do have a cost. Although GST experiments constitute the overwhelming majority of the circuits we run, the simultaneous DRB experiments are actually more sensitive to certain errors. GST and RB circuits are equally sensitive to the average stochastic error of a gate set, and they both amplify it proportional to circuit length L . We run GST circuits up to $L = 8$ or $L = 32$ (depending on platform), but RB circuits up to $L = 256$. As a result, the simultaneous RB analysis provides sharper information about the context dependence of average stochastic error than GST estimates do; GST-based predictions of the simultaneous DRB results are within error bars, but those error bars are larger than the estimated effect. The GST estimates provide far more information about coherent errors (and individual stochastic error rates), but simultaneous RB extracts its single summary benchmark with unmatched efficiency.

Extensions of our methods to many-qubit systems and two-qubit gates are possible. One obvious and immediately practical extension is to study the crosstalk induced by two-qubit gates (on a two-qubit subsystem) on neighboring one- and two-qubit subsystems. This requires probing at most four qubits at once, which is feasible with existing analytic machinery. Straightforwardly scaling our models to subsystems of $\gg 2$ qubits will quickly become impractical. Recent advances in many-qubit gate set tomography [82] suggest a path to reducing this overhead and enabling GST for efficient crosstalk characterization of large-scale quantum systems.

B. Experimental results

Our investigation reveals similar crosstalk errors in both platforms—the gate errors are significantly context dependent, but there is no evidence for entangling or correlated errors. And in both platforms, the dominant crosstalk errors are consistent with simple pulse spillover.

Our experiments find no statistically significant evidence for entangling or correlated stochastic errors in either experiment. We confirm using numerical simulations that our methodology is sensitive to entangling errors of this type (see the Appendix B), so our failure to detect them indicates their rate is below the detection threshold of this experiment. GST circuits’ sensitivity to coherent errors (of all types) increases proportional to their length, so in future work we intend to search for weak coherent ZZ errors in the AQT system using longer GST circuits. More sensitive GST experiments could also reveal context dependence in the nonunitary errors, and/or correlated stochastic noise.

The absence of entangling errors in the QSCOUT platform is unsurprising. This platform’s single-qubit gates are not expected to induce the ion-phonon couplings required to couple two ion qubits. Correlated stochastic errors are plausible and could be produced by a wide array of environment effects, but we see no evidence for such errors in the data. However, this experiment’s detection threshold is relatively high—the trapped-ion hardware has a relatively low data rate that limited both the circuit depth ($L = 8$) and the number of counts ($N = 80$). As a result, our estimates of stochastic error rates have low precision, and correlated errors could be hiding in the noise. More precise probing of these errors will require streamlined experiments.

One of the most impactful results of our analysis is that—in both platforms—the dominant crosstalk errors are restricted to (1) context-dependent local errors and (2) coherent unitary rotations. These restrictions single out a very small subset of all the possible crosstalk errors. It is always easier to characterize and track errors that lie in a constrained, low-dimensional set. Perhaps more importantly, context-dependent unitary errors are among the easiest to eliminate. Dynamical decoupling, optimal

control, or simple context-dependent calibration can all remove such errors. As demonstrated by AQT’s crosstalk compensation pulses, even simple techniques to cancel pulse spillover can improve device performance. Similarly, the dominant errors observed in QSCOUT experiments stem directly from how the gates are calibrated. Minor changes to that protocol—e.g., independent calibrations for “idle neighbor” and “driven neighbor” contexts—could reduce crosstalk errors below detectable thresholds.

Our analysis demonstrates the utility and performance of crosstalk compensation in the AQT system. Simultaneous GST can be used to enable continued, iterative reduction of crosstalk errors (via iterative calibration), but we do not think this is the right idea. Full simultaneous two-qubit GST requires too much overhead to be used in an active optimization loop. However, we believe it is possible to construct simpler, more targeted characterization protocols that focus on a particular type of crosstalk. These will run much faster, and be suitable for inclusion in active feedback loops. We propose that the role of “heavy” protocols like GST is to identify *which* crosstalk errors are dominant, so that specialized “lightweight” protocols can be deployed to tame them.

ACKNOWLEDGMENTS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. This material was funded in part by the by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research’s Quantum Testbeds for Science, Quantum Testbed Pathfinder, and Early Career Research Programs, by Sandia National Laboratories’ Laboratory Directed Research and Development Program, and 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, the U.S. Department of Energy, or the U.S. Government.

APPENDIX A: GATE SET TOMOGRAPHY CIRCUITS USED IN OUR EXPERIMENTS

Following Ref. [5] and as shown in Fig. 1 and Eq. (1), GST circuits consist of:

1. preparing the system in the all-zeros state,
2. applying a short preparation fiducial subcircuit, \mathbf{p}_i ,
3. applying a short germ subcircuit (repeated n times), \mathbf{g}_j^n ,

4. applying a short measurement fiducial subcircuit, \mathbf{m}_k ,
5. measuring the qubits in the computational basis.

In Table VI we list all $\{\mathbf{p}_i\}$, $\{\mathbf{g}_j\}$, and $\{\mathbf{m}_k\}$ subcircuits for the GST circuits used in our experiments. For each germ \mathbf{g}_j , n takes on the values $n = \lfloor L/\text{len}(\mathbf{g}_j) \rfloor$ for $L \in \{1, 2, 4, 8, \dots, L_{\max}\}$. For AQT, $L_{\max}^{\text{(AQT)}} = 32$, while for QSCOUT, $L_{\max}^{\text{(QSCOUT)}} = 8$. The GST circuit list includes all possible combinations of subcircuits of the form of Eq. (1).

The fiducials and germs are chosen numerically via pyGSTi [59] such that the fiducials generate an informationally complete set of states and measurements, and the germs are sensitive to all parameters in the general crosstalk model. Additionally, the germs and fiducials contain the circuits necessary to run isolated, single-qubit GST on each component qubit. This last requirement does not increase the number of fiducials required, but does necessitate the addition of two germs (the last two in Table VI).

APPENDIX B: SIMULATION

For all of the experiments presented in the main text, context-dependent models happen to provide the best balance of explanatory power and simplicity. This is evidence that correlated and entangling errors, which are specifically excluded from context-dependent models, are unnecessary to explain the observed data. Of course, not all experiments can be explained with these models. For instance, as a superconducting processor, the AQT platform might have been expected to experience weak ZZ-type coupling between the qubits [28]. In this section, we provide numerical evidence that our methods are capable of detecting a coherent entangling error and selecting the general crosstalk model, as long as the magnitude of the error is above a certain threshold, for a given amount of data gathered.

For our simulations, we use the same target gate set and GST circuit family as in our experiments. We set $L_{\max}^{\text{(sim)}} = 8$ and simulate 1000 measurements per circuit. Each circuit layer experiences an identical $Z^{(1)} \otimes Z^{(2)}$ entangling error of strength ϵ , described by a Hamiltonian error generator (which is applied concurrent with each gate operation):

$$\Delta\mathcal{H}(\rho) = -i \left[\frac{\epsilon}{2} Z^{(1)} \otimes Z^{(2)}, \rho \right]. \quad (\text{B1})$$

For each of ten distinct, exponentially spaced values of ϵ , we simulate data and fit our models. Results of these simulations are presented in Figs. 6–8.

1. Analysis of simulations

For the parameters used in our simulations, we identify a threshold value for the rate of the entangling error

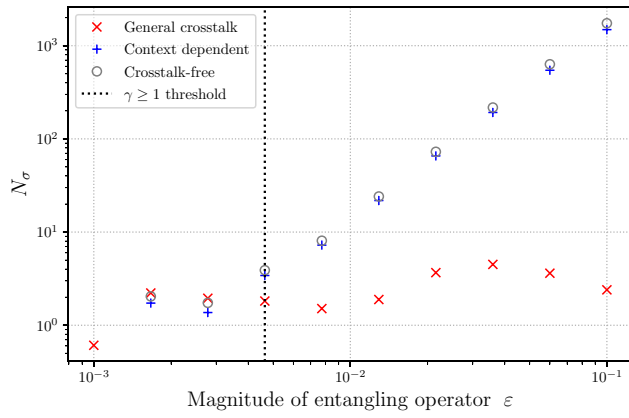


FIG. 6. Explanatory power of estimated crosstalk models versus strength of entangling error (simulation). Here, N_σ quantifies fit violation for each model. When $\epsilon < \epsilon^* = 4.6 \times 10^{-3}$, the predictions of all three estimated models are statistically consistent with the simulated data. As ϵ increases above this threshold, N_σ for the context-dependent and crosstalk-free models increases rapidly, while for the general crosstalk model, it remains relatively constant. In this regime, the general crosstalk model is preferred by the evidence ratio test.

$\epsilon^* \simeq 4.6 \times 10^{-3}$, above which we successfully detect the injected crosstalk *and* choose the crosstalk-containing model as the best fit. As shown in Fig. 6, for simulations with $\epsilon \geq \epsilon^*$, N_σ for the context-dependent and crosstalk-free models grows linearly with ϵ , but remains approximately constant for the general model. In these cases, the evidence ratio strongly (and correctly) selects the general model. Below the threshold, the context-dependent model is weakly preferred to the crosstalk-free model. The precision of GST scales with the number of measurements

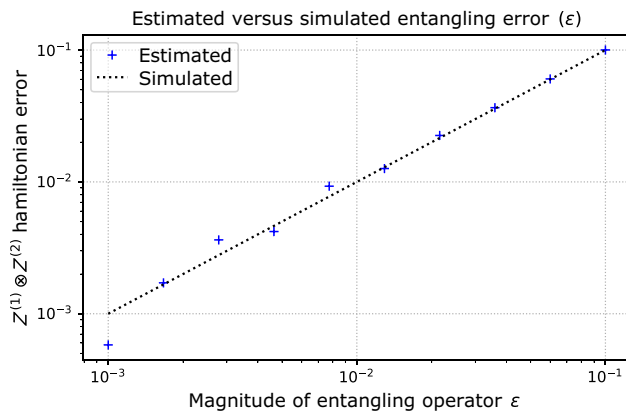


FIG. 7. Accuracy of the general crosstalk model's entangling error estimate (simulation). The general crosstalk model is able to accurately estimate the magnitude of a simulated entangling $Z \otimes Z$ crosstalk error. Shown here is the magnitude of the estimate of this error term (extracted from the crosstalk-free model) for the idle gate versus the true magnitude of the entangling error.

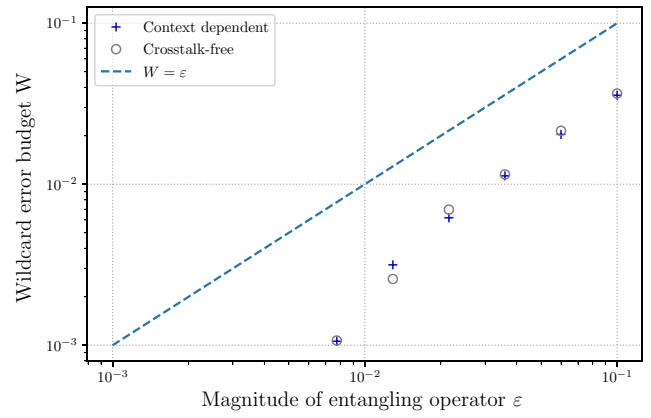


FIG. 8. Wildcard error for entangling-error-free models in the presence of entangling error (simulation). Here we show the wildcard error assigned to fits of the context-dependent and crosstalk-free models when an entangling error is present. For $\epsilon \leq \epsilon^* = 4.6 \times 10^{-3}$, the wildcard error budget for all three models is zero. However, for $\epsilon > 4.6 \times 10^{-3}$, the wildcard error for the context-dependent and crosstalk-free models increase approximately linearly with ϵ in the regime shown. The wildcard error for the general crosstalk model is zero for all values of ϵ , as this model always properly fits the data.

N and the maximum length L_{\max} as $O[1/(L\sqrt{N})]$, and we expect similar scaling in the threshold ϵ^* .

From each fit to the general crosstalk model we can extract an estimate of the strength of the entangling ZZ error present in any gate estimate. We do this for the general crosstalk model's idle gate estimate, and plot it against ϵ in Fig. 7. We see in all cases considered that the general crosstalk model is able to accurately reconstruct this entangling error, whether or not the general crosstalk model's evidence ratio indicates its selection over the other models.

In Fig. 8, we show the wildcard error W for the context-dependent and crosstalk-free models as a function of ϵ . For $\epsilon > \epsilon^*$, the wildcard error is proportional to, but non-trivially smaller than, the magnitude of the error. This indicates that the wildcard is capturing some of the unmodeled error, but not all of it. In fact, this is entirely expected behavior. The wildcard error is intended to capture the extra error *per gate* that is required for the model to be consistent with the data. For the wildcard to be approximately equal to the true value of the ZZ coherent error rate, there would need to exist a circuit of depth d for which the total variational distance between the observed and predicted probabilities is $\delta_{\text{TVD}}(p_o, p_p) = d\epsilon$. No such circuit exists, though some get close. For simulations with $\epsilon < \epsilon^*$, all models are consistent with the data, so W is zero.

-
- [1] Jerry M. Chow, Jay M. Gambetta, A. D. Córcoles, Seth T. Merkel, John A. Smolin, Chad Rigetti, S Poletto, George A. Keefe, Mary B. Rothwell, J. R. Rozen, Mark B. Ketchen,

- and M. Steffen, Universal Quantum Gate set Approaching Fault-Tolerant Thresholds with Superconducting Qubits, *Phys. Rev. Lett.* **109**, 060501 (2012).
- [2] R. Barends *et al.*, Superconducting quantum circuits at the surface code threshold for fault tolerance, *Nature* **508**, 500 (2014).
- [3] John P. Gaebler, Ting Rei Tan, Y. Lin, Y. Wan, R. Bowler, Adam C. Keith, S. Glancy, K. Coakley, E. Knill, D. Leibfried, and D. J. Wineland, High-Fidelity Universal Gate set for 9+ ion Qubits, *Phys. Rev. Lett.* **117**, 060505 (2016).
- [4] 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).
- [5] Robin Blume-Kohout, John King Gamble, Erik Nielsen, Kenneth Rudinger, Jonathan Mizrahi, Kevin Fortier, and Peter Maunz, Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography, *Nat. Commun.* **8**, 14485 (2017).
- [6] R Barends *et al.*, Diabatic Gates for Frequency-Tunable Superconducting Qubits, *Phys. Rev. Lett.* **123**, 210501 (2019).
- [7] V. Negîrneac, H. Ali, N. Muthusubramanian, F. Battistel, R. Sagastizabal, M. S. Moreira, J. F. Marques, W. Vlothuizen, M. Beekman, N. Haider, A. Bruno, and L. DiCarlo, High-fidelity controlled-z gate with maximal intermediate leakage operating at the speed limit in a superconducting quantum processor, [arXiv:2008.07411](https://arxiv.org/abs/2008.07411) [quant-ph] (2020).
- [8] R. Srinivas, S. C. Burd, H. M. Knaack, R. T. Sutherland, A. Kwiatkowski, S. Glancy, E. Knill, D. J. Wineland, D. Leibfried, A. C. Wilson, D. T. C Allcock, and D. H. Slichter, High-fidelity laser-free universal control of two trapped ion qubits, [arXiv:2102.12533](https://arxiv.org/abs/2102.12533) [quant-ph] (2021).
- [9] Shaowei Li, Anthony D. Castellano, Shiyu Wang, Yulin Wu, Ming Gong, Zhiguang Yan, Hao Rong, Hui Deng, Chen Zha, Cheng Guo, Lihua Sun, Chengzhi Peng, Xiaobo Zhu, and Jian-Wei Pan, Realisation of high-fidelity nonadiabatic CZ gates with superconducting qubits, *Npj Quantum Inf.* **5**, 84 (2019).
- [10] Youngkyu Sung, Leon Ding, Jochen Braumüller, Antti Vepsäläinen, Bharath Kannan, Morten Kjaergaard, Ami Greene, Gabriel O. Samach, Chris McNally, David Kim, Alexander Melville, Bethany M. Niedzielski, Mollie E. Schwartz, Jonilyn L. Yoder, Terry P. Orlando, Simon Gustavsson, and William D. Oliver, Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler, [arXiv:2011.01261](https://arxiv.org/abs/2011.01261) [quant-ph] (2020).
- [11] M. Brink, J. M. Chow, J. Hertzberg, E. Magesan, and S. Rosenblatt, in *2018 IEEE International Electron Devices Meeting (IEDM)* (2018), p. 6.1.1.
- [12] Dripto M. Debroy, Muyuan Li, and Shilin Huang, and Kenneth R. Brown, Logical performance of 9 qubit compass codes in ion traps with crosstalk errors, [arXiv:1910.08495](https://arxiv.org/abs/1910.08495) [quant-ph] (2019).
- [13] Manuel Gessner, Claude Fabre, and Nicolas Treps, Super-resolution limits from measurement crosstalk, [arXiv:2004.07228](https://arxiv.org/abs/2004.07228) [quant-ph] (2020).
- [14] Cupjin Huang, Xiaotong Ni, Fang Zhang, Michael Newman, Dawei Ding, Xun Gao, Tenghui Wang, Hui-Hai Zhao, Feng Wu, Gengyan Zhang, Chunqing Deng, Hsiang-Sheng Ku, and Jianxin Chen, and Yaoyun Shi, Alibaba cloud quantum development platform: Surface code simulations with crosstalk, [arXiv:2002.08918](https://arxiv.org/abs/2002.08918) [quant-ph] (2020).
- [15] Mohan Sarovar, Timothy Proctor, Kenneth Rudinger, Kevin Young, Erik Nielsen, and Robin Blume-Kohout, Detecting crosstalk errors in quantum information processors, *Quantum* **4**, 321 (2020).
- [16] Andrew J. Landahl, Jonas T. Anderson, and Patrick R. Rice, Fault-tolerant quantum computing with color codes, [arXiv:1108.5738](https://arxiv.org/abs/1108.5738) (2011).
- [17] Simon J. Devitt, William J. Munro, and Kae Nemoto, Quantum error correction for beginners, *Rep. Prog. Phys.* **76**, 076001 (2013).
- [18] Austin G. Fowler, Matteo Mariantoni, John M. Martinis, and Andrew N. Cleland, Surface codes: Towards practical large-scale quantum computation, *Phys. Rev. A* **86**, 032324 (2012).
- [19] Donovan Buterakos, Robert E. Throckmorton, and S. Das Sarma, Error correction for gate operations in systems of exchange-coupled singlet-triplet qubits in double quantum dots, *Phys. Rev. B Condens. Matter* **98**, 035406 (2018).
- [20] Andre R. R. Carvalho, Harrison Ball, Michael J. Biercuk, Michael R. Hush, and Felix Thomsen, Error-robust quantum logic optimization using a cloud quantum computer interface, [arXiv:2010.08057](https://arxiv.org/abs/2010.08057) [quant-ph] (2020).
- [21] Yanzhu Chen, Maziar Farahzad, Shinjae Yoo, and Tzu-Chieh Wei, Detector tomography on IBM quantum computers and mitigation of an imperfect measurement, *Phys. Rev. A* **100**, 052315 (2019).
- [22] Stephen Crain, Clinton Cahall, Geert Vrijsen, Emma E. Wollman, Matthew D. Shaw, Varun B. Verma, Sae Woo Nam, and Jungsang Kim, High-speed low-crosstalk detection of a 171yb+ qubit using superconducting nanowire single photon detectors, *Commun. Phys.* **2**, 97 (2019).
- [23] Dripto M. Debroy and Kenneth R. Brown, Extended flag gadgets for low-overhead circuit verification, [arXiv:2009.07752](https://arxiv.org/abs/2009.07752) [quant-ph] (2020).
- [24] Yongshan Ding, Pranav Gokhale, Sophia Fuhui Lin, Richard Rines, Thomas Propson, and Frederic T. Chong, Systematic crosstalk mitigation for superconducting qubits via Frequency-Aware compilation, [arXiv:2008.09503](https://arxiv.org/abs/2008.09503) [quant-ph] (2020).
- [25] Akel Hashim, Ravi K. Naik, Alexis Morvan, Jean-Loup Ville, Bradley Mitchell, John Mark Kreikebaum, Marc Davis, Ethan Smith, Costin Iancu, Kevin P. O'Brien, Ian Hincks, Joel J. Wallman, Joseph Emerson, and Irfan Siddiqi, Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor, [arXiv:2010.00215](https://arxiv.org/abs/2010.00215) [quant-ph] (2020).
- [26] Jaseung Ku, Xuexin Xu, Markus Brink, David C. McKay, Jared B. Hertzberg, Mohammad H. Ansari, and B. L. T. Plourde, Suppression of unwanted ZZ interactions in a hybrid Two-Qubit system, [arXiv:2003.02775](https://arxiv.org/abs/2003.02775) [quant-ph] (2020).
- [27] Swarnadeep Majumder, Leonardo Andreta de Castro, and Kenneth R. Brown, Real-time calibration with spectator qubits, *Npj Quantum Inf.* **6**, 1 (2020).

- [28] Pranav S. Mundada, Gengyan Zhang, Thomas Hazard, and Andrew A. Houck, Suppression of Qubit Crosstalk in a Tunable Coupling Superconducting Circuit, *Phys. Rev. Appl.* **12**, 054023 (2019), [arXiv:1810.04182](https://arxiv.org/abs/1810.04182) [quant-ph].
- [29] Prakash Murali, David C. McKay, Margaret Martonosi, and Ali Javadi-Abhari, in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '20* (Association for Computing Machinery, New York, NY, USA, 2020), p. 1001, [arXiv:2001.02826](https://arxiv.org/abs/2001.02826) [quant-ph].
- [30] Alireza Seif, Kevin A. Landsman, Norbert M. Linke, Caroline Figgatt, C. Monroe, and Mohammad Hafezi, Machine learning assisted readout of trapped-ion qubits, *J. Phys. B At. Mol. Opt. Phys.* **51**, 174006 (2018).
- [31] Sarah Sheldon, Easwar Magesan, Jerry M. Chow, and Jay M. Gambetta, Procedure for systematically tuning up cross-talk in the cross-resonance gate, *Phys. Rev. A* **93**, 060302(R) (2016).
- [32] Neereja Sundaresan, Isaac Lauer, Emily Pritchett, Easwar Magesan, Petar Jurcevic, and Jay M. Gambetta, Reducing unitary and spectator errors in cross resonance with optimized rotary echoes, [arXiv:2007.02925](https://arxiv.org/abs/2007.02925) [quant-ph] (2020).
- [33] T. F. Watson, S. G. J. Philips, E. Kawakami, D. R. Ward, P. Scarlino, M. Veldhorst, D. E. Savage, M. G. Lagally, Mark Friesen, S. N. Coppersmith, M. A. Eriksson, and L. M. K. Vandersypen, A programmable two-qubit quantum processor in silicon, *Nature* **555**, 633 (2018).
- [34] Adam Winick, Joel J. Wallman, and Joseph Emerson, Simulating and mitigating crosstalk, [arXiv:2006.09596](https://arxiv.org/abs/2006.09596) [quant-ph] (2020).
- [35] Yuan Xu, Ji Chu, Jiahao Yuan, Jiawei Qiu, Yuxuan Zhou, Libo Zhang, Xinsheng Tan, Yang Yu, Song Liu, Jian Li, Fei Yan, and Dapeng Yu, High-fidelity, high-scalability two-qubit gate scheme for superconducting qubits, [arXiv:2006.11860](https://arxiv.org/abs/2006.11860) [quant-ph] (2020).
- [36] Deanna M. Abrams, Nicolas Didier, Shane A. Caldwell, Blake R. Johnson, and Colm A. Ryan, Methods for Measuring Magnetic Flux Crosstalk between Tunable Transmons, *Phys. Rev. Appl.* **12**, 064022 (2019), [arXiv:1908.11856](https://arxiv.org/abs/1908.11856) [quant-ph].
- [37] Stefania Balasiu, Characterization of Multi-Qubit algorithms with randomized benchmarking, [https://qudev.phys.ethz.ch/static/content/science/Documents/master/Stefania Balasiu-MasterThesis.pdf](https://qudev.phys.ethz.ch/static/content/science/Documents/master/Stefania%20Balasiu-MasterThesis.pdf) (2018).
- [38] J. Kelly *et al.*, Optimal Quantum Control Using Randomized Benchmarking, *Phys. Rev. Lett.* **112**, 240504 (2014).
- [39] S. Krinner, S. Lazar, A. Remm, C. K. Andersen, N. Lacroix, G. J. Norris, C. Hellings, M. Gabureac, C. Eichler, and A. Wallraff, Benchmarking Coherent Errors in Controlled-Phase Gates due to Spectator Qubits, *Phys. Rev. Appl.* **14**, 024042 (2020).
- [40] C. Piltz, T. Sriarunothai, A. F. Varón, and C. Wunderlich, A trapped-ion-based quantum byte with 10(-5) next-neighbour cross-talk, *Nat. Commun.* **5**, 4679 (2014).
- [41] Jay M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, John A. Smolin, Jerry M. Chow, Colm A. Ryan, Chad Rigetti, S. Poletto, Thomas A. Ohki, Mark B. Ketchen, and M. Steffen, Characterization of Addressability by Simultaneous Randomized Benchmarking, *Phys. Rev. Lett.* **109**, 240504 (2012).
- [42] David C. McKay, Andrew W. Cross, Christopher J. Wood, and Jay M. Gambetta, Correlated randomized benchmarking, [arXiv:2003.02354](https://arxiv.org/abs/2003.02354) [quant-ph] (2020).
- [43] Alexander Erhard, Joel J. Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban A. Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt, Characterizing large-scale quantum computers via cycle benchmarking, *Nat. Commun.* **10**, 5347 (2019).
- [44] Robin Harper, Steven T. Flammia, and Joel J. Wallman, Efficient learning of quantum noise, [arXiv:1907.13022](https://arxiv.org/abs/1907.13022) [quant-ph] (2019).
- [45] Timothy Proctor, Kenneth Rudinger, Kevin Young, Mohan Sarovar, and Robin Blume-Kohout, What Randomized Benchmarking Actually Measures, *Phys. Rev. Lett.* **119**, 130502 (2017).
- [46] Richard Kueng, David M. Long, Andrew C. Doherty, and Steven T. Flammia, Comparing Experiments to the Fault-Tolerance Threshold, *Phys. Rev. Lett.* **117**, 170502 (2016).
- [47] Eric Huang, Andrew C. Doherty, and Steven Flammia, Performance of quantum error correction with coherent errors, *Phys. Rev. A* **99**, 022313 (2019).
- [48] Kenneth Rudinger, Timothy Proctor, Dylan Langharst, Mohan Sarovar, Kevin Young, and Robin Blume-Kohout, Probing Context-Dependent Errors in Quantum Processors, *Phys. Rev. X* **9**, 021045 (2019).
- [49] Eugene Dumitrescu and Travis S. Humble, Discrimination of correlated and entangling quantum channels with selective process tomography, *Phys. Rev. A* **94**, 042107 (2016), [arXiv:1607.05292](https://arxiv.org/abs/1607.05292) [quant-ph].
- [50] Masoud Mohseni, A. T. Rezakhani, and D. A. Lidar, Quantum-process tomography: Resource analysis of different strategies, *Phys. Rev. A* **77**, 032322 (2008).
- [51] Seth T. Merkel, Jay M. Gambetta, John A. Smolin, Stefano Poletto, Antonio D. Córcoles, Blake R. Johnson, Colm A. Ryan, and Matthias Steffen, Self-consistent quantum process tomography, *Phys. Rev. A* **87**, 062119 (2013).
- [52] Firat Solgun, David P. DiVincenzo, and Jay M. Gambetta, Simple impedance response formulas for the dispersive interaction rates in the effective hamiltonians of low anharmonicity superconducting qubits, *IEEE Trans. Microw. Theory Tech.* **67**, 928 (2019).
- [53] Erik Nielsen, John King Gamble, Kenneth Rudinger, Travis Scholten, Kevin Young, and Robin Blume-Kohout, Gate set tomography, [arXiv:2009.07301](https://arxiv.org/abs/2009.07301) [quant-ph] (2020).
- [54] Erik Nielsen, Kenneth Rudinger, Timothy Proctor, Kevin Young, and Robin Blume-Kohout, Efficient flexible characterization of quantum processors with nested error models, [arXiv:2103.02188](https://arxiv.org/abs/2103.02188) [quant-ph] (2021).
- [55] AQT@LBL - SC Qubit Testbed, <https://aqt.lbl.gov/> (), accessed: 2020-03-01.
- [56] Susan M. Clark, Craig William Hogle, Kevin Young, and Daniel Lynn Stick, Demonstrating robustness of analogue quantum simulators, (2020).
- [57] Susan M. Clark, Daniel Lobser, Melissa Revelle, Christopher G. Yale, David Bossert, Ashlyn D. Burch, Matthew

- N. Chow, Craig W. Hogle, Megan Ivory, Jessica Pehr, Bradley Salzbrener, Daniel Stick, William Sweatt, Joshua M. Wilson, Edward Winrow, and Peter Maunz, Engineering the quantum scientific computing open user testbed (QSCOUT): Design details and user guide, [arXiv:2104.00759](https://arxiv.org/abs/2104.00759) [quant-ph] (2021).
- [58] Quantum Scientific Computing Open User Testbed (QSCOUT), <https://www.sandia.gov/quantum/Projects/QSCOUT.html>, accessed: 2020-03-01.
- [59] Erik Nielsen, Robin J. Blume-Kohout, Kenneth M. Rudinger, Timothy J. Proctor, and Lucas Saldyt *et al.*, *Python GST Implementation (PyGSTi) v. 0.9*, Tech. Rep. (Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2019).
- [60] Erik Nielsen, Kenneth Rudinger, Timothy Proctor, Antonio Russo, Kevin Young, and Robin Blume-Kohout, Probing quantum processor performance with pygsti, *Quantum Sci. Technol.* **5**, 044002 (2020).
- [61] Samuel S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Ann. Math. Stat.* **9**, 60 (1938).
- [62] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Automat. Contr.* **19**, 716 (1974).
- [63] Our crosstalk models form a hierarchy, so they are strictly nested—for any pair of them, one contains the other as a subset.
- [64] Robin Blume-Kohout, Kenneth Rudinger, Erik Nielsen, Timothy Proctor, and Kevin Young, Wildcard error: Quantifying unmodeled errors in quantum processors, [arXiv:2012.12231](https://arxiv.org/abs/2012.12231) (2020).
- [65] Aleksei Yur'evich Kitaev, Quantum computations: Algorithms and error correction, *Uspekhi Matematicheskikh Nauk* **52**, 53 (1997).
- [66] John Watrous, Semidefinite programs for completely bounded norms, [arXiv:0901.4709](https://arxiv.org/abs/0901.4709) (2009).
- [67] Yuval R. Sanders, Joel J. Wallman, and Barry C. Sanders, Bounding quantum gate error rate based on reported average fidelity, *New J. Phys.* **18**, 012002 (2015).
- [68] Robin Blume-Kohout, Marcus P. da Silva, Erik Nielsen, Timothy Proctor, Kenneth Rudinger, Mohan Sarovar, and Kevin Young, A taxonomy of small Markovian errors, [arXiv:2103.01928](https://arxiv.org/abs/2103.01928) [quant-ph] (2021).
- [69] This is a slightly different error generator representation than that presented in Ref. [68]; we use *during-gate* generators for (only) the Hamiltonian sector, because it is more convenient.
- [70] Łukasz Rudnicki, Zbigniew Puchała, and Karol Życzkowski, Gauge invariant information concerning quantum channels, *Quantum* **2**, 60 (2018).
- [71] Junan Lin, Brandon Buonacorsi, Raymond Laflamme, and Joel J. Wallman, On the freedom in representing quantum operations, *New J. Phys.* **21**, 023006 (2019).
- [72] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson, Scalable and Robust Randomized Benchmarking of Quantum Processes, *Phys. Rev. Lett.* **106**, 180504 (2011).
- [73] Timothy J. Proctor, Arnaud Carignan-Dugas, Kenneth Rudinger, Erik Nielsen, Robin Blume-Kohout, and Kevin Young, Direct Randomized Benchmarking for Multiqubit Devices, *Phys. Rev. Lett.* **123**, 030503 (2019).
- [74] AQT@LBL - AQT Capabilities, ' [https://aqt.lbl.gov/wp-content/uploads/sites/4/2021/06/AQT Capabilities 20210-601rev.pdf](https://aqt.lbl.gov/wp-content/uploads/sites/4/2021/06/AQT%20Capabilities%20210-601rev.pdf) (), accessed: 2021-07-28.
- [75] David C. McKay, Christopher J. Wood, Sarah Sheldon, Jerry M. Chow, and Jay M. Gambetta, Efficient Z gates for quantum computing, *Phys. Rev. A* **96**, 022330 (2017).
- [76] A. D. Patterson, J. Rahamim, T. Tsunoda, P. A. Spring, S. Jebari, K. Ratter, M. Mergenthaler, G. Tancredi, B. Vlastakis, M. Esposito, and P. J. Leek, Calibration of a Cross-Resonance Two-Qubit Gate between Directly Coupled Transmons, *Phys. Rev. Appl.* **12**, 064013 (2019).
- [77] S. Olmschenk, K. C. Younge, D. L. Moehring, D. N. Matsukevich, P. Maunz, and C. Monroe, Manipulation and detection of a trapped Yb^+ hyperfine qubit, *Phys. Rev. A* **76**, 052314 (2007).
- [78] Peter Maunz, *High Optical Access Trap 2.0*, Tech. Rep. SAND2016-0796R (Sandia National Laboratories, Albuquerque, NM, 2016).
- [79] R. Islam *et al.*, Beat note stabilization of mode-locked lasers for quantum information processing, *Opt. Lett.* **39**, 3238 (2014).
- [80] S Wimperis, Broadband, narrowband, and passband composite pulses for use in advanced NMR experiments, *J. Magn. Reson. A* **109**, 221 (1994).
- [81] Evidence ratios less than 1 are typically caused by extremely small sample counts, when many events appear 0 or 1 times, and stem from breakdown of the Gaussian ansatz used in deriving Wilks' theorem. Since each circuit is repeated $N = 80$ times, it is unsurprising to see this for the QSCOUT data.
- [82] Erik Nielsen, *Efficient Scalable Tomography of Many-Qubit Quantum Processors*, Tech. Rep. (Sandia National Laboratories, Albuquerque, NM, 2020).