SAND20XX-XXXXR
LDRD PROJECT NUMBER: 201879
LDRD PROJECT TITLE: Efficient, Predictive Tomography of Multi-Qubit Quantum Processors
PROJECT TEAM MEMBERS: Robin Blume-Kohout, Erik Nielsen, Kenneth Rudinger, Mohan Sarovar, Kevin Young

ABSTRACT:

After decades of R&D, quantum computers comprising more than 2 qubits are appearing. If this progress is to continue, the research community requires a capability for precise characterization ("tomography") of these enlarged devices, which will enable benchmarking, improvement, and finally certification as mission-ready. As world leaders in characterization -- our gate set tomography (GST) method is the current state of the art – the project team is keenly aware that *every* existing protocol is either (1) catastrophically inefficient for more than 2 qubits, or (2) not rich enough to predict device behavior. GST scales poorly, while the popular randomized benchmarking technique only measures a single aggregated error probability. This project explored a new insight: that the combinatorial explosion plaguing standard GST could be avoided by using an ansatz of few-qubit interactions to build a complete, efficient model for multi-qubit errors. We developed this approach, prototyped it, and tested it on a cutting-edge quantum processor developed by Rigetti Quantum Computing (RQC), a US-based startup. We implemented our new models within Sandia's PyGSTi open-source code, and tested them experimentally on the RQC device by probing crosstalk. We found two major results: first, our schema worked and is viable for further development; second, while the Rigetti device is indeed a "real" 8-qubit quantum processor, its behavior fluctuated significantly over time while we were experimenting with it and this drift made it difficult to fit our models of crosstalk to the data.

INTRODUCTION:

Quantum computing – building systems of high-fidelity *qubits* (quantum bits) and implementing quantum algorithms on them to solve hard problems – is one of the rapidly expanding frontiers of computation and technology, with the promise to revolutionize a wide if poorly understood range of national security, commercial, and scientific problems. But as a field of engineering it remains in its infancy, and only a few years ago the largest "quantum computers" comprised just 2 connected qubits. In the past half-decade, this status quo has exploded, and today there are several fully connected quantum processors with 5, 8, or even 16 qubits working together. This number is expected to grow rapidly.

But despite this growth, there is a long road ahead to *useful* quantum computers, which will require thousands or millions of qubits to do anything truly revolutionary. Although as few as 50-100 perfect qubits might demonstrate a useful speedup on some problem, real qubits are noisy and error prone, and require *quantum error correction*, which demands significant overhead

Sandia National Laboratories is a multi-mission 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.

(perhaps 100x in the total number of qubits). Error correction also places a tremendous premium on understanding and being able to predict the noise properties of the physical qubits in the processor, so that it can be mitigated (by tailored error correction) or debugged (by smart physicists). This motivates *quantum characterization, verification, and validation* (QCVV), the field of R&D that seeks to probe noise and errors in qubits and report their strength and type.

But rich, predictive QCVV methods – "tomography" – are arduous, complicated, and computationally intensive. Worse, their difficulty and cost scales exponentially with the number of qubits, meaning that although they can be applied to 2 qubits at once today, scaling beyond 3 seems impossible. Our goal in this project was to explode *this* status quo, by finding ways to efficiently but accurately and predictively characterize multi-qubit systems comprising more than 2 qubits. We conjectured that the main obstacle to doing so was the proliferation of model parameters in the n-qubit quantum process matrices used to model noise, and that by devising "reduced models" based on physical intuitions about the nature of qubit systems, we could capture the behavior of n-qubit systems effectively with many fewer parameters and avoid the computational demands of those larger models.

In short, our purpose was to rapidly explore a new, revolutionary approach to qubit characterization. This project's success would depend critically on an open science question: *What is the form of noise in as-built quantum processors*? For the first time, we directly addressed this question. Our preliminary results indicate that our hypothesis is correct, and that it will enable practical tools critical for developing future quantum computers.

We set out to develop and implement, within Sandia's existing PyGSTi software framework, new algorithms for experimental characterization of multi-qubit quantum processors. These algorithms are based on efficient, physically motivated, statistical models for quantum noise that have many fewer parameters than standard "tomographic" models, which grow exponentially with processor size. We also arranged to deploy these prototype algorithms and test them on Rigetti Quantum Computing's (RQC's) 8-qubit device, in close cooperation with RQC, to measure the RQC device's performance. We tested whether our software implementation could reliably find an optimal fit to simulated noise (it could), and sought to determine whether the models accurately predicted device behavior. (This was not possible because the RQC device drifted too much over time.) Overall, this project *did* validate the new approach, and we accomplished a remarkably broad (though not complete) characterization of noise characteristics in a cutting-edge 8-qubit quantum processor.

DETAILED DESCRIPTION OF EXPERIMENT/METHOD:

Motivation for Generator GST

One of the primary achievements of this EE LDRD was our initial success testing an idea for extending tomography to more than two qubits. The idea was to reduce the number of free parameters specifying each gate *G*, and perform tomography on the resulting "reduced" or "sparse" gate-set model. The sparse model is defined by first demanding that $G = G_0 e^{L}$, where G_0 is some known ideal operation and *L* is an "error generator" that must take a particular form (the *Lindbladian* form) if it is to generate valid physical dynamics:

$$L(\rho) = \sum_i h_i [H_i, \rho] + \sum_{ij} o_{ij} \left(B_i \rho B_j^\dagger - \frac{1}{2} B_i B_j^\dagger \rho - \frac{1}{2} \rho B_i^\dagger B_j \right)$$

The H_i and the B_i range over an orthogonal basis of operators that span the Hilbert-Schmidt space in which quantum states (density matrices) live. For multi-qubit (*n*-qubit) systems, we can take each to be a tensor product of *n* Pauli matrices (plus the identity **I**), where *n* is the number of qubits. For instance, for two qubits a *complete* basis would be the 4^2=16 matrices {**II**, **IX**, **IY**, **IZ**, **XI**, … **ZZ**}. If a complete basis were used for the H_i and B_i, and no restrictions placed on the O(4^n) h_i and O(4^{2n}) o_{ij} coefficients, the resulting *G* would be unconstrained. That is, any type of qubit error could be represented. But this requires a large number of parameters, and makes even 3-qubit analysis computationally intractable.

Our central hypothesis was that not all of these terms are needed to adequately capture the behavior of a physical gate -- in essence, that the *logarithm* of *G* (or, equivalently, of GG_0^{-1}) is *sparse*. The H_i terms correspond to coherent (Hamiltonian-generated) errors on the qubits, whereas the B_i terms describe non-unitary (e.g. stochastic) errors. The diagonal (i=j) stochastic terms, when the B_i are Pauli operators, describe *Pauli-stochastic* errors. These are used very commonly in quantum error correction, and are expected to dominate the error budget in many physical systems. Our hypothesis, then, was that we only needed to keep the Hamiltonian and stochastic terms and could simply drop the rest. Perhaps even more critically, we also hypothesized that we could limit the *weight* of the error operators -- i.e., the number of non-identity Pauli matrices in the H_i or B_i product -- to be below some threshold. And, finally, we could limit the positions in which the non-identity operators occur. For instance, in a 5-qubit chain we could require that the weight of the H_i and B_i be at most 2, and we could require that the two non-identity operators must correspond to physically adjacent qubits.

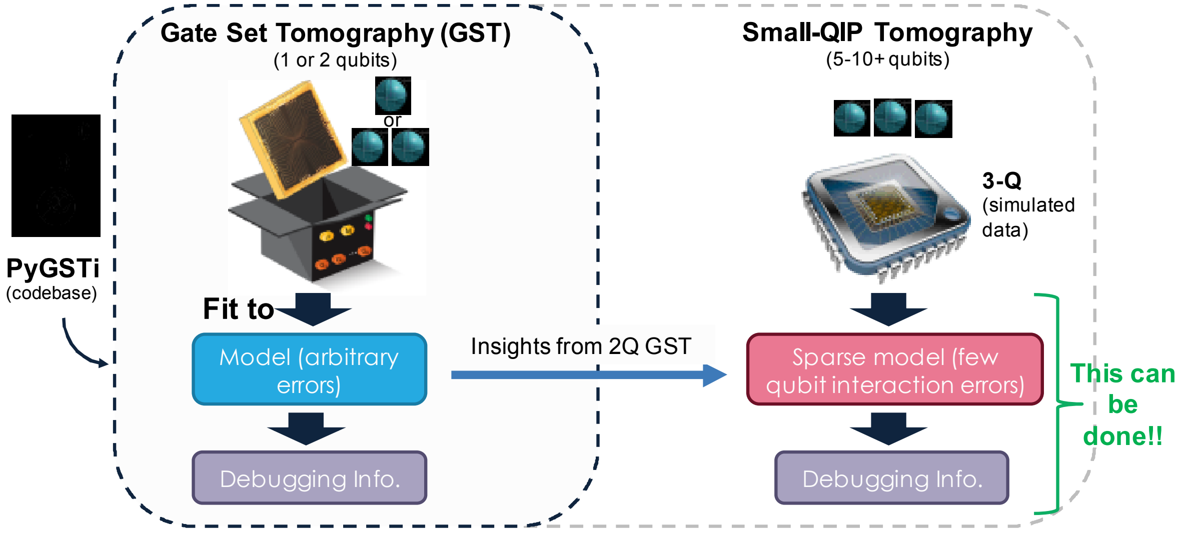
We set out to test these *sparse gate models* in this LDRD. The error-generator-based gate parameterizations outlined above were implemented within our pyGSTi open-source code, and tested for correctness using single and 2-qubit analyses. Because experimental data was unavailable during most of the LDRD, we simulated a linear chain of 3 qubits with standard Pauli-stochastic error and performed tomography on the data using a sparse model that included only Hamiltonian and Pauli-stochastic errors.

We found that our tomography procedure (detailed below) was able to reconstruct the processes that generated the data. While this leaves open the question of whether typical experimental data can also be reconstructed using the technique (something we hoped at the outset to also address), it affirmatively answers an important even more fundamental question: can a GST-like tomography procedure be used with a well-motivated sparse model to increase the system size? We have shown that if an experimental system did behave according to a simple Pauli-stochastic noise model, then this sparse-model tomography would be able to reconstruct that model from the data, and this is an essential first step toward characterizing larger systems of qubits.

Sparse Generator GST

The tomographic procedure we prototyped here is similar to gate set tomography (GST), in that it was a model-based inference method that maximizes a likelihood function over a class of models. However, there are several key differences. Most significantly, the model is “sparse”: using the construction described above, we model 3 qubit operations with just 1520 degrees of freedom, in contrast to the 33,280 of the “full” model that would be used by standard “dense” GST. A second important difference from standard GST is how the gate sequences were selected. GST gate sequences are comprised of two “fiducial” gate sequences with a repeated “germ” sequence between them. The fiducial sequences ensure that an informationally complete set of effective preparation and measurement “states” can be produced, and the germ sequences are chosen to ensure that the likelihood function is sensitive to every element of every gate matrix, i.e., *all* 33K parameters of the full model. The many fewer parameters of our sparse model in principle requires many fewer gate sequences (and thus experimental data), but determining *which* sequences are required is a nontrivial task that is not performed by standard GST - at least not entirely. The way standard GST selects “germ” sequences that ensure sensitivity to all parameters in the model can be -- and was -- generalized to select *only* sequences that amplify to the sparse model's parameters. This generalized procedure, however, still assumed informationally complete fiducial sequences (as GST does), and so our procedure began with a complete set and then removed unnecessary (preparation, measurement) pairs of fiducials - an algorithm we term “fiducial pair reduction”. The remaining sequences are sufficient (ideally minimally-sufficient) to give sensitivity to the sparse model's parameters. Our full tomography protocol can thus be outlined by the following steps:

1. Sparse model creation
2. Gate sequence selection
    (a) Select complete sets of fiducial sequences
    (b) Select sparse-model germ sequences
    (c) Perform fiducial pair reduction to eliminate unnecessary fiducials.
3. Maximum likelihood optimization (using a local-search Levenberg-Marquardt algorithm)



The figure above provides a schematic of the tomography we performed on a simulated system of 3-qubits, and diagrammatically indicates the parallels between GST and the tomography procedure we used. Successfully making such a dramatic reduction of the degrees of freedom is equivalent to taking the GST optimization, which we expect would converge well if it were tractable (because GST typically does), and placing 33,280-1520 = 31,760 constraints on the optimization space. The fact that imposing so many constraints resulted in an optimization landscape which was still amenable to standard local-search methods is an important and non-trivial statement, and greatly increases our confidence that tomography using such sparse models may be possible far beyond the 3-qubit case studied here.

RESULTS:

In addition to developing new techniques, protocols, and analysis methods (described in “Method” above), we deployed both the new techniques *and* variants of existing “standard” GST to probe the 8-qubit superconducting qubit (transmon) devices developed by RQC and accessed under the terms of a CRADA with RQC.

Here, we summarize what we learned about the *performance* of the RQC devices, without attempting to discuss or analyze the device architecture. However, a few basic details are necessary. The system is an 8-qubit ring (see figure in Section 3), in which the even-numbered qubits (Q0, Q2, Q4, Q6) are tunable transmons, and the odd-numbered ones (Q1, Q3, Q5, Q7) are fixed-frequency transmons. Each of the 8 pairs of nearest neighbors (0-1, 1-2, …, 7-0) is capacitively coupled. In normal operation, the coupling (roughly 3 MHz) is overwhelmed by a frequency separation of O(100 MHz) between the qubits, which effectively decouples them. To perform entangling gates, the decoupling is turned off by tuning the qubits into resonance. All 8 transmons are treated as qubits; there is not a dramatic difference in performance or addressability between them. An authoritative discussion of this architecture can be found in *arXiv:1706.06570*.

The goal of our research at SNL was, and is, to perform *holistic characterization* on an RQC device. This means building a successful, rich, predictive model of more than 2 qubits at once. This is an ambitious goal relative to the state of the art, which can be described roughly as the convex hull of three different approaches:

1. **Randomized benchmarking** (RB), which has been extensively performed on 1- and 2-qubit subsystems and could, in principle, be extended to 8 or more. RB is descriptive rather than predictive, and seeks to measure only O(1) descriptive properties of the system studied.
2. **Gate-set tomography** (GST), which SNL has developed for several years, and has been performed extensively on 1-qubit subsystems and somewhat less frequently on 2-qubit subsystems. GST is predictive and holistic *on these small subsystems*, but as currently formulated does not scale to more qubits well at all.
3. **Entangled-state preparation**, which is an ad-hoc technique developed and used by a wide variety of groups to demonstrate *something* about a many-qubit QIP. This usually consists of (1) preparing one or more Bell/GHZ/W states using a quantum circuit, and then (2) demonstrating that it worked by either (a) state tomography, (b) direct measurement of stabilizers, or (c) measuring an entanglement witness.

Our goal was to do what GST does for 1-2 qubits, but efficiently. We believe this is possible because we believe that although the number of distinct errors that *could* happen to *N* qubits scales exponentially with *N*, the number that actually *do* occur (and thus need to be modeled) scales polynomially, and more tractably. This is an ambitious goal and goes well beyond the scope of the project reported here. *That* scope is strictly to prototype the techniques, to determine whether they have potential, and to perform pathfinder experiments that reveal obstacles. Put more informally, we sought to jump into the pool and splash around vigorously – i.e., to start inventing solutions to the problems posed by our research goals, and testing them to see (a) whether they work, and (b) how they need to be modified in the face of the real world.

It's important to recognize that, at this time, there's no standard or satisfactory way to fully characterize an 8-qubit device! The problem is not yet well-posed. So this report will not: (1) answer all the questions, (2) even suppose that we know what the right questions are, (3) claim (in most cases) to have authoritative results, or (4) try to hide the fact that a lot of the things that we tried have failed. Most of those failures were productive and inspiring; we can usually figure out why our idea didn't work, and come up with a better idea for the next round.

Summary of findings

"The RQC device" is not a single device. RQC has a relatively rapid design cycle, and replaces their device-under-test fairly frequently with new and improved chips. Furthermore, since we focus on the performance of *operations* rather than on intrinsic physical properties of the qubit chip, the control hardware and calibrations (which are also being constantly tuned and improved) are at least as essential to device performance as the chip. The results reported here are snapshots at a range of times spanning roughly 1 month (mid-July to mid-August, 2017), and are not expected to even be consistent with each other, much less representative of future results.

Single-qubit GST, performed on each of the 8 qubits individually over <1 day, showed that all the qubits can be operated at an error rate of 1% or better. Individual gate infidelities for Idle, $X_{\pi/2}$, and $Y_{\pi/2}$ gates ranged from 0.19% to 1.04% (with a single outlier: the Idle on qubit 1 displayed 1.83% infidelity). Diamond norm error rates for the same gates ranged from 0.9% to 5.1%, apparently due to coherent errors stemming from imperfect calibration. Five of the eight qubits displayed very low rates of non-Markovian behavior (e.g. drift); Qubit 6 appears to be wildly unstable, and Qubits 1-2 display significant non-Markovianity. The Achilles heel of these qubits appears to be SPAM (state preparation and measurement) error, with SPAM errors between 10.2% and 23.4% observed for the well-behaved qubits.

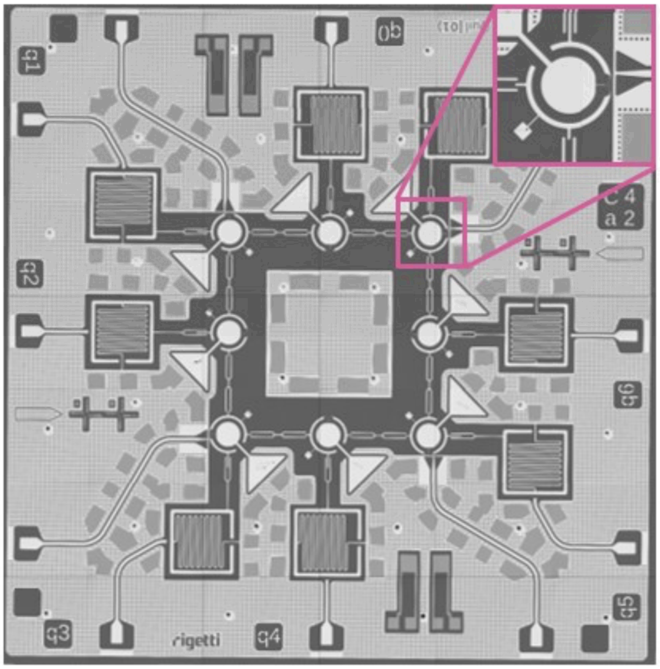
Two-qubit GST was performed on one tuned-up CPHASE gate between a pair of adjacent qubits, [Q0/Q7]. As of August 15, RQC did not feel it was feasible to get all 8 C-PHASE gates tuned up on the current device under test, although tuning up more than one of them may be possible. This experiment behaved well, giving us reasonably high confidence in the results. The Q0-Q7 CPHASE gate displayed a process infidelity of 9.5%, and a diamond norm error of 18.2%. Most of the process infidelity (about 9% out of 9.5%) appears to be due to stochastic errors, but there is a significant coherent error that contributes up to 0.5% of the process infidelity and dominates the diamond norm. Both stochastic and coherent errors are dominated by weight-1 (local) Z errors (dephasing), especially on Q0 (which suffers 3x as much error as Q7). The gate has five eigenvalues >0.95, four of which probably correspond to the four computational basis states. All of this suggests that RQC can probably achieve a fairly uniform error rate of ~4-5% for this CPHASE gate if they can achieve better and more stable control of the local qubit energy splitting during the CPHASE gate. Getting below 4-5% is likely to require other techniques.

Several different experimental tests and protocols were deployed to look for two kinds of crosstalk: *idle crosstalk* that manifests as an unwanted coupling between two qubits, and *operation crosstalk* in which operations on one qubit influence neighbors. These tests suggested the presence of crosstalk (at least idle crosstalk, which is expected in this architecture), but were inconclusive because of overwhelming levels of *drift*. So far, our crosstalk characterization experiments have to be considered a productive failure because the crosstalk is overwhelmed by drift. However, these techniques *did* work in simulations. We conclude that crosstalk can be characterized using the low-weight generator model, but that we need to devote more work to making it robust to drift.

Several techniques were used to detect or probe drift. However, almost none of our experiments were intended to detect drift – instead, we detected it in the course of routine sanity checking on the crosstalk experiments – so we do not currently have a full and formal picture of what's drifting and how. Our impression at this time is that: (1) the gates themselves seem to be very stable; (2) the SPAM operations seem prone to drift (sometimes wildly); (3) our best guess is that the measurement/readout procedure is drifting, possibly due to the classification procedure for the I-Q readouts; and (4) drift varies drastically across different sorts of experiments and from qubit to qubit, from unimportant (though detectable) to catastrophic.

System survey, control/access limitations

The RQC 8-qubit chip looks like this:



RQC's "system" is really multiple chips, installed in dilution refrigerators at various times. All the devices we have tested are variations on the same design: 8 qubits arranged in a ring geometry (see figure above). Even-numbered qubits (Q0, Q2, Q4, Q6) are fixed-frequency transmons, while odd-numbered ones (Q1, Q3, Q5, Q7) are tunable transmons. Each of the 8 pairs of nearest neighbors (0-1, 1-2, …, 7-0) is capacitively coupled, and the coupling can be turned on and off to isolate the qubits or perform entangling gates between them. All 8 transmons are treated as qubits, not resonators, and from our QCVV perspective there is not a dramatic difference in performance or addressability between them.

The original research plan anticipated remote access and control over a device (via the internet) beginning in January 2017. SNL would have queued and run experiments on the device in a similar fashion to how the IBM Quantum Experience (QX) is controlled. As of August 2017, this sort of control access is not available. In lieu of direct control, SNL generated experimental designs, and conveyed them to RQC to be run on the device. The resulting datasets are shared (usually as Python *pickle* files containing pyGSTi *dataset* objects), and either or both of SNL and RQC analyze them using pyGSTi functionality.

This mode of operation was generally successful, but its drawbacks include occasional communication failures and latency. On the positive side, it encourages collaboration and forces both parties to understand more about the other's perspective and constraints. However, from the SNL perspective, it limits creativity and spontaneity in a measurable way compared, for example, with the IBM QX where we can (1) devise a new experimental design and run it immediately, (2) know in advance exactly what control is possible and what is not, and (3) know exactly *how* our experimental design will be implemented. As QIPs start to behave more like classical computers, the concept of an API or well-defined interface becomes more and more important. However, APIs are inevitably difficult for hardware that is under development.

RQC's stated goal is to deploy a flexible, powerful control system that can stream arbitrary gates to the 8-qubit chip more or less in real time. However, this control system did not become fully (or even effectively) operational during the timeframe of this project. The control system used by RQC in FY17 behaves functionally much more like a typical arbitrary waveform generator (AWG). It can run the same pulse sequence (typically 1 circuit, although small batches seem possible too) N times in rapid succession, but if many sequences need to be run, then there is a significant "reloading" delay between them. The repetition rate for individual sequences is about 2000 – 6000 per second (at least for circuits of length up to L=128 gates), but the reload time is between 0.1 and 0.6 seconds. Performing N=2000 repetitions of K=2000 distinct sequences could take anywhere from 0.9 to 670 hours, depending on how many repetitions are done in each pass. This obviously motivates doing a single pass through the circuits, which is unfortunate because GST and its variants become *much* more reliable when the data is "rastered" by doing N passes through the K circuits and taking one count on each pass. Single-pass experiments are completely vulnerable to drift – if the system changes over the minutes or hours required to step through all the circuits, then different circuits effectively run on totally different qubits. Trying to reconstruct "the gates" from this is like reconstructing a dinosaur from the mixed-together bones of a triceratops and a tyrannosaurus – the results are generally meaningless.

In practice, we have compromised by doing 2 or 3 passes on most experiments. This doesn't provide enough time resolution to track drift, to factor it out of the reconstruction, or to reliably detect drift in *all* circuits whose probabilities drift. But it does enable us to reliably detect whether *something* about the system is drifting, to quantify roughly how much it's drifting, and (generally) to distinguish SPAM drift from gate drift.

1-qubit gate performance

1-qubit GST is well-established, but it played a critical role in this project for two goals: First, to establish the "baseline" performance of the device; and second, to test whether we could successfully characterize an 8-qubit device using standard protocols.

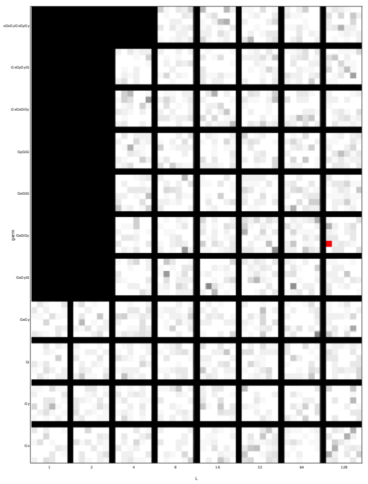
RQC performed "straight out of the box" 1-qubit GST on all 8 qubits in succession. Both RQC and SNL analyzed the data; the SNL analysis uses newer algorithms that yield almost the same statistical results, but provides a more powerful and flexible reporting format. The experimental design used SNL's standard { Idle, $X_{\pi/2}$, and $Y_{\pi/2}$} gateset for each qubit, 6x6=36 fiducial pairs for LGST, and 11 germs with sequence lengths up to L=128 to amplify noise, for a total of K=2353 circuits on each qubit, each repeated N=600 times in a single pass for a total of NK=1.4x10^{6} clicks on each of 8 qubits.

The results are fairly "normal" — no big surprises. The high-level results of probing all 8 qubits individually are gathered into the following table

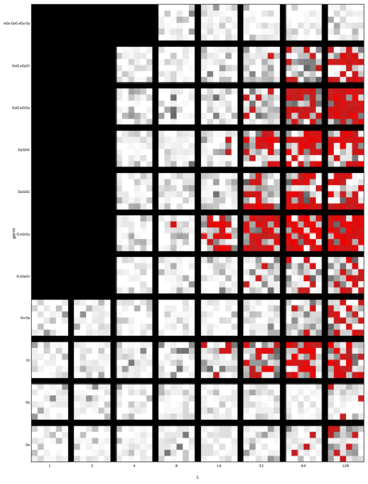
	Model violation	1-F(***Id***)	1-F(***X***)	1-F(***Y***)	◆/2(***Id***)	◆/2(***X***)	◆/2(***Y***)	SPAM
Qubit 0	11.5 σ	0.35%	0.27%	0.28%	0.6%	1.9%	1.9%	10.4%
Qubit 1	157 σ	1.83%	0.41%	0.47%	5.1%	0.9%	1.0%	17.5%
Qubit 2	390 σ	1.04%	0.28%	0.21%	2.1%	0.8%	0.8%	12.7%
Qubit 3	12.2 σ	0.65%	0.38%	0.36%	3.9%	1.9%	1.9%	23.4%
Qubit 4	11.4 σ	0.63%	0.62%	0.56%	2.9%	1.0%	0.9%	10.2%
Qubit 5	4 σ	0.29%	0.19%	0.18%	2.5%	1.0%	0.9%	21.1%
Qubit 6	1050 σ	0.31%	0.09%	0.04%	1.1%	0.3%	0.3%	42.4%
Qubit 7	5.8 σ	0.84%	0.64%	0.55%	3.9%	1.5%	1.3%	17.2%
Uncertainty		0.03%	0.03%	0.03%	0.3%	0.3%	0.3%	0.2%

A. ***Model violation***: The first thing we look at is non-Markovianity: how well could we fit the data to a Markovian CPTP map model? Here, Qubit 6 is a disaster (roughly 1000 σ violation of the model), and Qubit #2 is pretty bad (400 σ). Qubit #1 is shaky (160 σ), but the rest (Q0,Q3,Q4,Q5,Q7) are looking pretty good (<=12 σ). Careful examination of the per-sequence model violation (see figure below showing qubits Q7, Q2, and Q6 as examples), most of what we see is consistent with what we've seen in other places. It's distributed over the germs, it gets worse with L, and it's worst for the idle gate. We conjecture that this is coming from qubit frequency drift (manifesting mostly as an idle gate that shifts between under- and over-rotation). **However, Qubit 6 is doing something else**. There's a very weird and distinct signature in the data, suggesting that as RQC worked its way through the sequences, something shifted very abruptly about 50% of the way through (in the middle of the L=32 sequences), and then shifted again near the end (in the middle of the L=128 sequences).

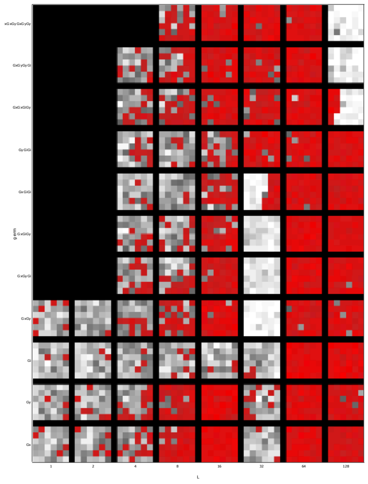
Model violation: Qubit #7



Model violation: Qubit #2



Model violation: Qubit #6



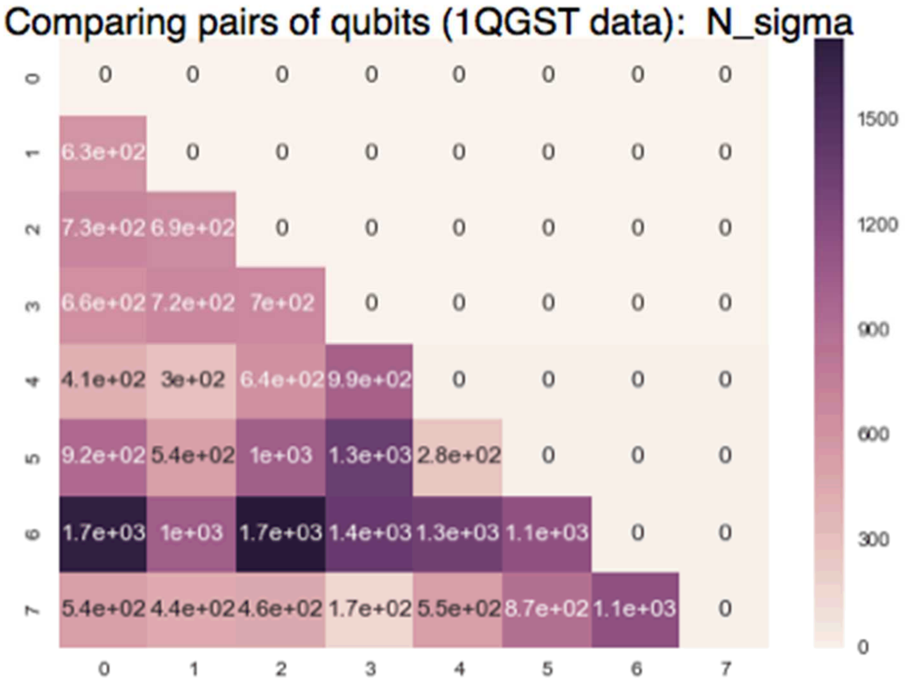
We will generally ignore Qubit 6 from here on out. Although it appears to perform pretty well (maybe even better than the others) when it's working, the degree to which it jumps around makes it impossible to say much with confidence, so any blanket statements should not be taken to apply to Qubit 6.

B. ***Process fidelity and stochastic errors***: On each of the 7 remaining qubits, all three gates display process infidelities between 0.2% and 0.8% (uncertainty here is about 0.03%) with two exceptions: the Idle gates on Qubits #1 and #2 are at 1.8% and 1.0%, respectively. All the X and Y rotations are between 0.2% (Qubit #5) and 0.6% (Qubits #4 and #7). For all qubits, X and Y gates are very similar (as they should be, since they differ only by a phase shift). For all qubits except #1 and #2, the Idle is between 1.1x and 2x worse than the X/Y gates. This is probably because the X and Y gates are echoing out some of the noise that afflicts the Idle. All the process infidelity is coming from stochastic errors (coherent errors are not a significant source of process infidelity for these qubits). In general, we see both "phase-flip" stochastic errors that commute with the gate, and "bit-flip" stochastic errors that flip the eigenstates of the gate. (Note that this is slightly different terminology from the usual in quantum computing — usually, "bit flip" means an X or Y error, while "phase flip" means a Z error. Our terminology agrees with that if the gate is a Z or Idle gate, but if the gate is an X gate then a "phase flip" error is an X error, and a "bit flip" error is a Z or Y error.) Phase flips seem to be occurring at a higher rate, by a factor of 1.1x – 2x. In other words, there is both T1 (bit flip) and T2 (phase flip) errors, but the noise is somewhat T2 dominated, indicating that fluctuations in the strength of the control pulse (or the qubit frequency) are contributing about 2/3 of the infidelity, while thermal decay is contributing about 1/3.

C. ***Diamond norm and coherent errors***: The 1/2- diamond norm error rates range from 0.8% (X and Y gates on Qubit #2) to 5% (Idle gate on Qubit #1). The X and Y gates range from 0.8% to 1.9%, while the Idles range from 0.6% (Qubit #0) to 5% (Qubit #2). Uncertainties here are around 0.3%. All of the extra error above infidelity appears to be coming from coherent errors, and could probably be tuned away. The Idles generally seem to be small Z rotations, while the X and Y gates seem to be dominated by over/under-rotation errors. Qubits #1, #2, and maybe #7 have some tilt error (angle between X and Y axes of rotation off by more than 0.1%).

D. ***SPAM error***: This is pretty high and pretty variable, ranging from 10% - 23% (except for Qubit 6). There's not much more to say here; we can't reliably tell from these experiments whether it's due to state preparation or measurement, but other factors suggest that the measurement is the weakest aspect of these qubits.

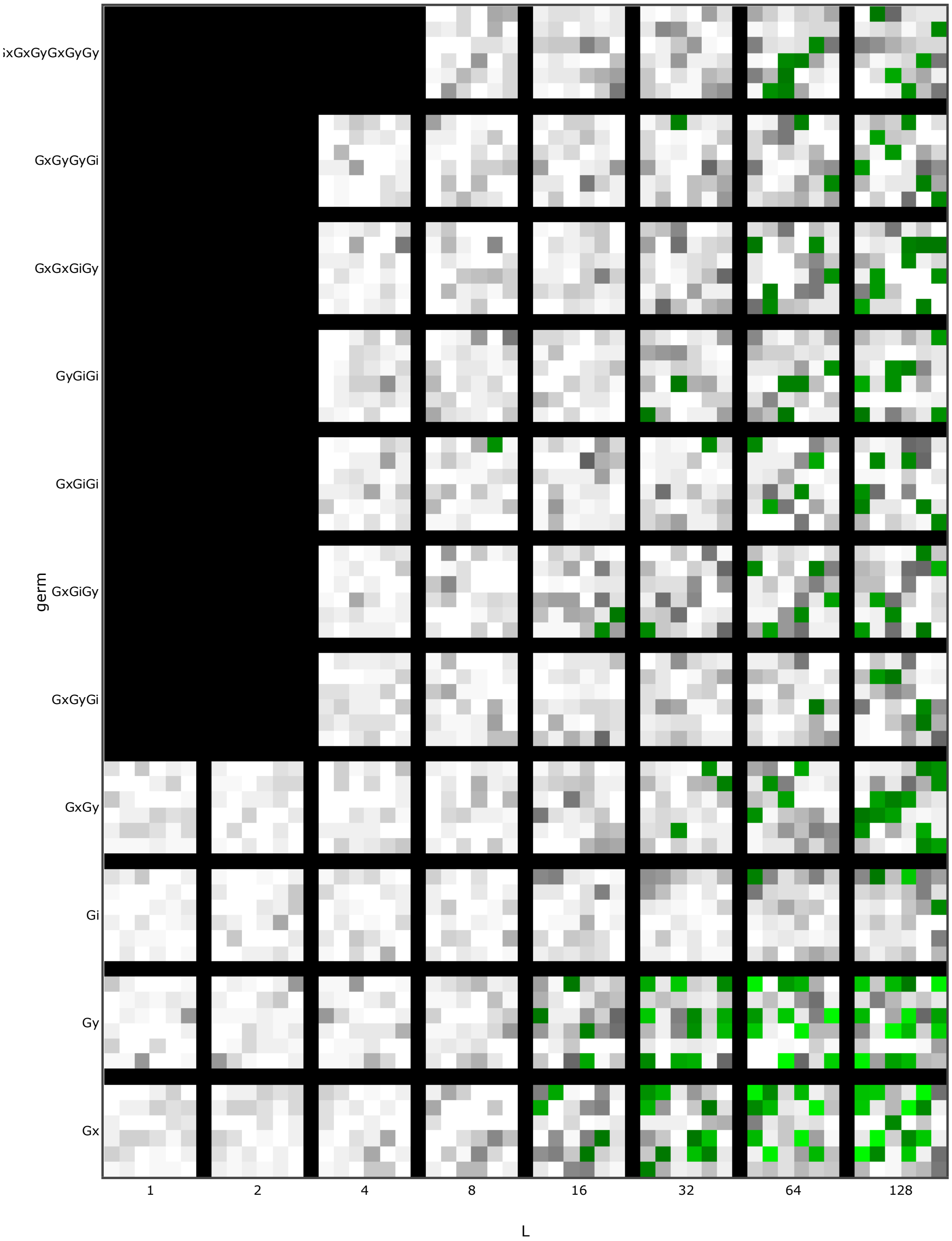
E. ***Similarity of qubits***: Since most of the qubits seemed to be well-behaved, we tried something new: we asked how *similar* each pair of qubits appeared to be, judging just from the 1Q-GST data. In other words, how plausible is it that (say) the data from Qubit #1 and Qubit #3 actually came from identical qubits? We know they're not identical, but how close are they? This is measured in # of σ of model violation, and this is roughly comparable to the model violation numbers given above. So 12 σ would indicate that two qubits are roughly as consistent with *each other* as each of them is with *itself*. On the other hand, 1000 σ would indicate that the difference between those qubits is roughly as big as the instability of Qubit #6. This analysis is shown in the figure below.


Comparing pairs of qubits (1QGST data): N_sigma

0	0	0	0	0	0	0	0	0
1	6.3e+02	0	0	0	0	0	0	0
2	7.3e+02	6.9e+02	0	0	0	0	0	0
3	6.6e+02	7.2e+02	7e+02	0	0	0	0	0
4	4.1e+02	3e+02	6.4e+02	9.9e+02	0	0	0	0
5	9.2e+02	5.4e+02	1e+03	1.3e+03	2.8e+02	0	0	0
6	1.7e+03	1e+03	1.7e+03	1.4e+03	1.3e+03	1.1e+03	0	0
7	5.4e+02	4.4e+02	4.6e+02	1.7e+02	5.5e+02	8.7e+02	1.1e+03	0
	0	1	2	3	4	5	6	7

The most similar pair of qubits is [#3,#7], with about 170 σ difference. Other pairs with <300 σ difference are [#4,#5] and [#1,#4]. #6 is clearly off in its own world.

To put these numbers in context, we examined the most similar pair ([3,7]) in detail to identify <u>what</u> makes them different. The datasets are compared directly and visualized in terms of "per sequence discrepancy" in the figure below; green squares indicate circuits for which the observed frequencies differed by a statistically significant amount (between the two 1Q-GST datasets).



It turns out that the dominant discrepancy between Qubit #3 and Qubit #7 is quite simply the rotation angle of the X and Y gates. (The fact that the "Gi" germ row is almost all grey confirms this, and the pattern of green boxes in the Gx and Gy germ rows tells us that it's the rotation angle. The decoherence rates are also different by a factor of about 2, but the rotation angle is a bigger effect). So how different is this dominant difference? Well, Qubit #3 has a rotation angle of 0.5034*Pi, while Qubit #7 has a rotation angle of 0.5100*Pi. So the biggest difference here is an 0.0066*Pi rotation angle. (The two qubits' Gx gates have decoherence rates of 0.5% and 0.9%, respectively). So, they're different — but not as much as one might expect from a 170 σ violation. Basically, this is a small difference in the lab, but GST is really good at detecting small differences.

2-qubit gate performance

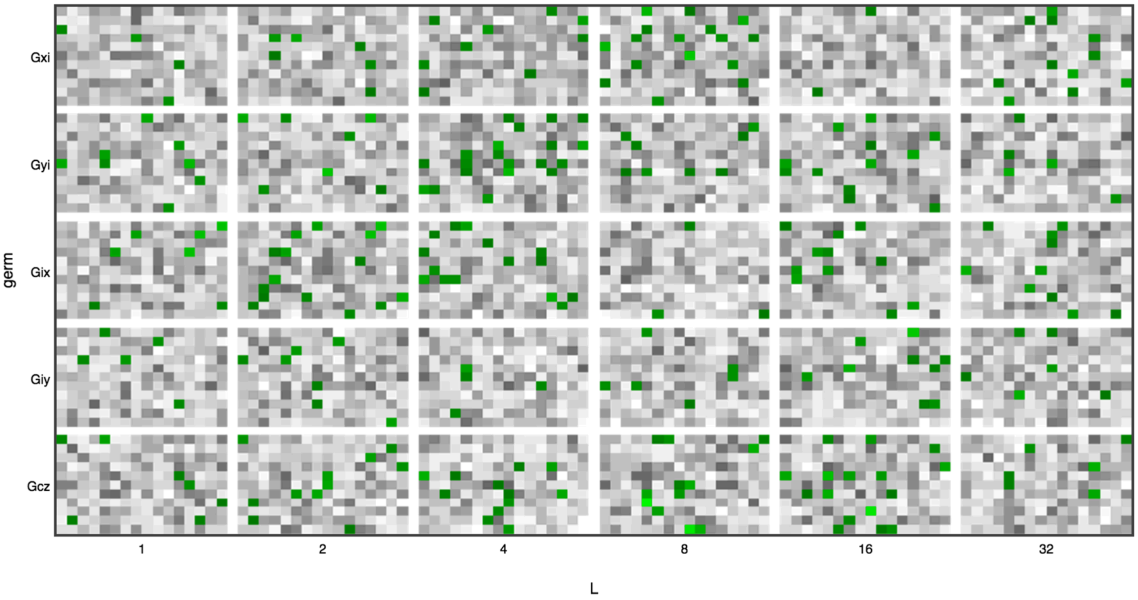
The current "standard practice" 2-qubit GST assumes a set of 5 gates (X and Y rotations on each qubit, plus an entangling gate that for RQC is a C-PHASE), which yields 1088 gauge-invariant parameters to be estimated. Amplifying all these parameters requires 63 distinct germs repeated to yield sequences of length L=1,2,4,…, and each of these germ-powers is bracketed between 16x11 = 176 fiducial pairs.

This means that a "full" 2-qubit GST experiment with L_{max} = 32 contains almost 50,000 distinct circuits. Although we have done such an experiment in another collaboration, the RQC device's control limitations make it infeasible. We therefore deployed a technique called *fiducial pair reduction* (FPR), which just means identifying circuits that provide redundant information and throwing them out. We developed an aggressive form of FPR that discards all but about 1/16 of the circuits. This yielded an experimental design for L_{max} = 32 with just 2,754 circuits, and one for L_{max} = 128 with just 3,889 circuits.

When RQC ran the L_{max} = 128 version of this experiment and we analyzed the data, we found that although the analysis generally made sense, we got radically different results from (a) a CP-constrained fit to the data, and (b) an unconstrained fit. Furthermore, these very different estimates appeared to fit the data roughly equally well (the CP-constrained fit was worse, but not shockingly so). The unconstrained fit indicated a gate fidelity around 90%, while the CP-constrained fit indicated about 30%. This discrepancy persisted in the eigenvalues of the gate, which are purely gauge-invariant (and therefore ruled out the possibility that one of the fits was being badly gauge-fixed). So, two very different theories about the C-PHASE gate – that it had process fidelity 0.9, and that it had process fidelity 0.3 – were both roughly as consistent with the data. This isn't supposed to happen, because the germs are chosen to amplify all parameters and ensure that any theory other than the correct one is inconsistent with the data. We tentatively attribute this behavior to a combination of (a) aggressive FPR that reduced the experiment's statistical robustness, and (b) significant non-Markovianity in the data, possibly due to drift (no drift detection was incorporated into the experimental design for this experiment).

To deal with this problem, we sent RQC a revised experiment in which FPR was *only* applied to the 58 germs of length >1. All five of the germs corresponding to single gates were bracketed between all 176 fiducial pairs. This was intended to guarantee that (1) the gates' eigenvalues were very reliably probed, and (2) any non-Markovianity would show up with a recognizable signature. The new experiment contained 8,010 circuits, and two passes of N=500 counts each were performed.

The first analysis we did was to examine and compare the two passes, to see which (if any) circuits showed statistically inconsistent results between Pass 0 and Pass 1. This is shown in the figure below for the 5,280 sequences corresponding to L repetitions of a single gate. Green squares represent statistically significant detections of change between the two passes; if the experiment was completely drift-free, then with 95% probability *every* square would be a shade of grey.



This confirmed the following points:

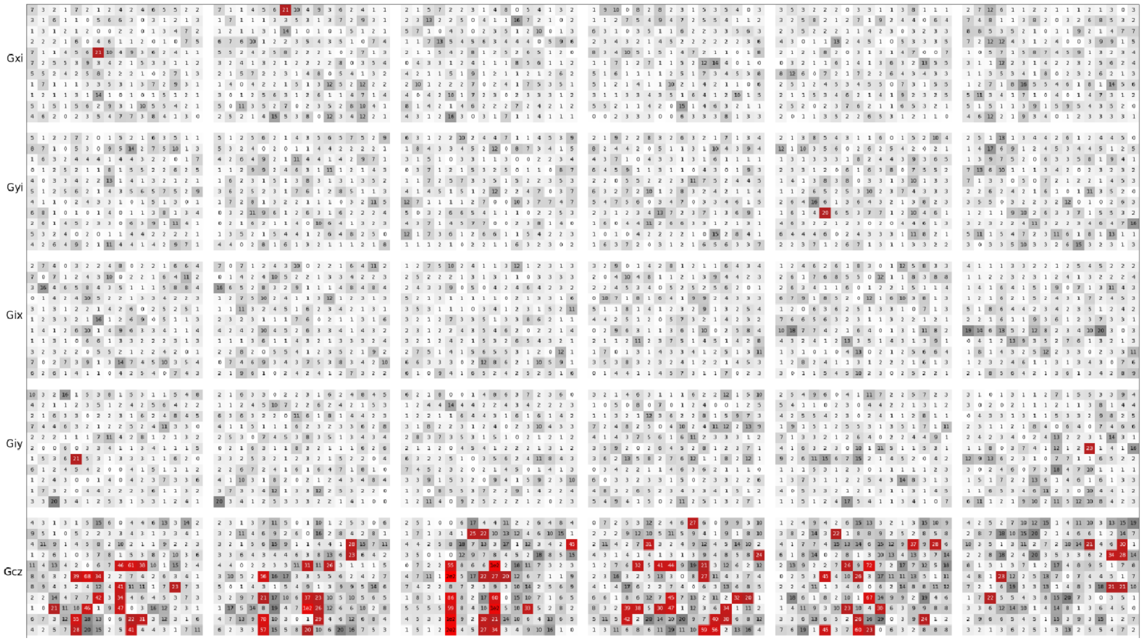
1. Drift isn't too bad in this dataset. Most squares are grey, indicating relatively little drift for that circuit. The green ones are statistically significant, but by inspection they don't correspond to dramatic shifts in the observed probabilities.
2. Detected drift is not strongly correlated with L. This strongly suggests that it's not the gates that are drifting, but more probably the SPAM. We suspect the measurement classifier.
3. Detected drift is not strongly correlated with which gate is being performed. This tends to support our suspicion that the dominant mechanism of drift has to do with SPAM rather than gates.
4. No consistent pattern is apparent. This tends to support the theory that what's going on is primarily random fluctuations in the performance of the measurement classifier, rather than anything correlated with a specific aspect of the gates or SPAM.

We also checked (although this is getting a bit ahead of ourselves) that the GST fits for each pass individually (N=500) are very similar and consistent.

Next, we looked for model violation. While the data *are* significantly non-Markovian – the fit to the combined (N=1000 counts per circuit) dataset violates the model by about 200 σ. However, this is common in our experience. Moreover, the model violation rises steadily and predictably with the length and number of circuits (see table below, especially N_σ vs L).

L	$2\Delta\log(\mathcal{L})$	k	$2\Delta\log(\mathcal{L})-k$	$\sqrt{2k}$	N_σ	N_s	N_p	Rating
0	1227.4	1105	122.42	47.011	2.6	2193	1088	★★★★
1	1227.4	1105	122.42	47.011	2.6	2193	1088	★★★★
2	3946.9	3058	888.88	78.205	11.4	4146	1088	★★★★
4	1×10^4	6550	3495.2	114.46	30.5	7638	1088	★★★★
8	2×10^4	12004	1×10^4	154.95	72	13092	1088	★★★★
16	4×10^4	17473	2×10^4	186.94	1×10^2	18561	1088	★★★★
32	6×10^4	22942	3×10^4	214.21	2×10^2	24030	1088	★★★★

We also examined the per-sequence model violation to see which circuits were noticeably non-Markovian. This is shown in the figure below for the 5,280 circuits corresponding to L repetitions of single gates.

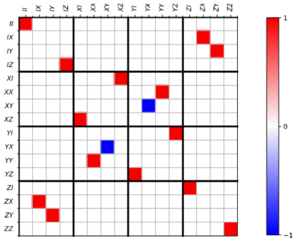
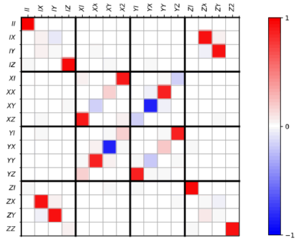


This is actually *remarkably* well-behaved 2-qubit data (especially given the relatively high number of counts, N=1000). Non-Markovianity is almost entirely restricted to the C-PHASE gate, and this pattern persists in the other 58 germs (not shown); only germs involving a C-PHASE gate are non-Markovian. We strongly suspect that this is "true" non-Markovianity stemming from a persistent environment formed by the other 6 qubits, and that what looks like non-Markovianity in this 2-qubit reduced subsystem is actually entangling crosstalk.

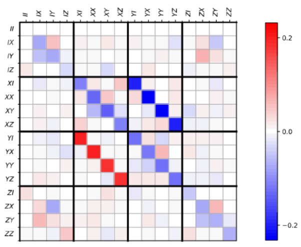
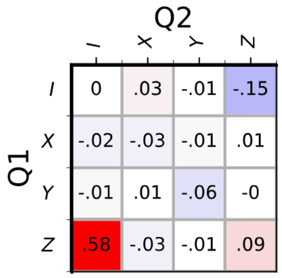
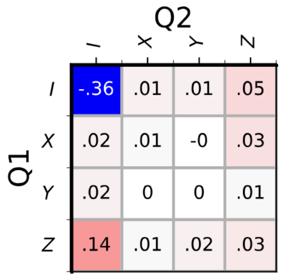
All of this suggests that the estimated gates should be meaningful, and describe the operations being performed at the time of this experiment (at least). The estimates of the single-qubit gates (as 2-qubit process matrices) are not especially interesting; they're generally consistent with the results of 1-qubit GST. We therefore focus on the C-PHASE gate. However, it's worth noting that we don't believe that our current analysis has selected the best gauge in which to describe the gates. This analysis chooses a gauge that makes *all* the gates as close as possible to their targets. Because the C-PHASE gate is noisier than the 1-qubit gates, the 1-qubit gates should define the reference frame (gauge). When we re-set the gauge this way, we expect to get a slightly different detailed noise model (neither the process fidelity nor the eigenvalues of the C-PHASE gate will change meaningfully, and we expect the diamond norm error to increase slightly).

The process infidelity of the estimated C-PHASE gate is 9.5%, and its diamond norm error is 18.2%. (For what it's worth, the Jamiolkowski trace distance is 17.2%, indicating that applying this gate to one half of a maximally entangled state is *almost* the best way to distinguish it from an ideal C-PHASE gate with a single use).

The estimated process matrix for the C-PHASE gate is compared to its target in the figure below:

Gate	Target Superoperator (Pauli-prod basis)	Estimated Superoperator (Pauli-prod basis)
Gcz		

And here is the error *generator* and its Hamiltonian and Pauli-stochastic components:

Gate	Error Generator	Hamiltonian Projections	Stochastic Projections
Gcz			

Hamiltonian Projections

	I	X	Y	Z
I	0	.03	-.01	-.15
X	-.02	-.03	-.01	.01
Y	-.01	.01	-.06	-0
Z	.58	-.03	-.01	.09

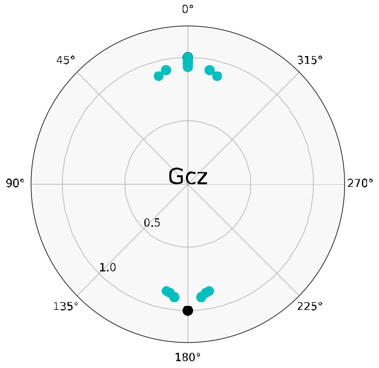
Stochastic Projections

	I	X	Y	Z
I	-.36	.01	.01	.05
X	.02	.01	-0	.03
Y	.02	0	0	.01
Z	.14	.01	.02	.03

The single largest error for this gate is a coherent ZI error (a Z rotation on the first qubit, Q0), followed by IZ (-Z rotation on the second qubit) and ZZ (over rotation of the gate). It's interesting that the Pauli-stochastic errors are dominated by the same terms: ZI, IZ, and ZZ, in almost the same proportions. This suggests a fairly straightforward physical mechanism: The Hamiltonian terms that commute with the gate itself (ZI, IZ, and ZZ all commute with the C-

PHASE) are insufficiently controlled: their average value is not quite right (producing a systematic coherent error) and they are also fluctuating around that average value (producing stochastic errors that commute with the gate. Lending some additional support to this theory is the fact that the next most significant *coherent* errors are weight-2 XX and YY terms, both of which commute with the ZZ interaction. However, this could also be a gauge-fixing artifact.

We've focused so far on gauge-dependent quantities. An important sanity check on these is to check gauge-*invariant* quantities – the gate eigenvalues, mostly – and see if they are consistent. The eigenvalues of the estimated C-PHASE gate are shown in the table below:

Gate	Eigenvalues	Polar Plot
Gcz	$0.9012e^{i3.022}$ $0.9012e^{-i3.022}$ $0.8626e^{i2.943}$ $0.8626e^{-i2.943}$ $0.8724e^{i2.978}$ $0.8724e^{-i2.978}$ $0.8831e^{i0.267}$ $0.8831e^{-i0.267}$ $0.9159e^{i0.192}$ $0.9159e^{-i0.192}$ 0.9222 0.9997 0.9813 0.9507 $0.958e^{i0.000}$ $0.958e^{-i0.000}$	

The absolute values of these eigenvalues describe how the quantum state space shrinks along certain principal axes as the C-PHASE is applied over and over. Their phases describe how it rotates (and can be roughly interpreted as rotation angles). For a perfect unitary process, these would all have absolute value 1, and would come in conjugate pairs whose phases are differences between eigenvalues of the unitary rotation operator. Real eigenvalues (e.g. 1) represent fixed points of the map – they define the generalized rotation axis (which remains fixed as other states rotate about it). *Every* CPTP map has at least one eigenvalue that is exactly 1 (a fixed point).

Here, all the magnitudes are between 0.86 and 0.98 (excluding the fixed point). Their average is 0.914, which is consistent with the process fidelity of 0.905 if we blame the remaining 0.9% infidelity on coherent errors. There are 5 large (>0.95) eigenvalues, 4 of which probably correspond to the 4 Pauli operators that commute with C-PHASE (II, IZ, ZI, ZZ). The ideal C-PHASE has two eigenvalues: 1 with multiplicity 10, and -1 with multiplicity 6. The 6 eigenvalues that should have phase $\pm\pi$ are reasonably close, under-rotating by 0.12 – 0.16 radians, but the biggest coherent errors are the spurious phases of 0.267 and 0.192 that correspond to the ZI Hamiltonian.

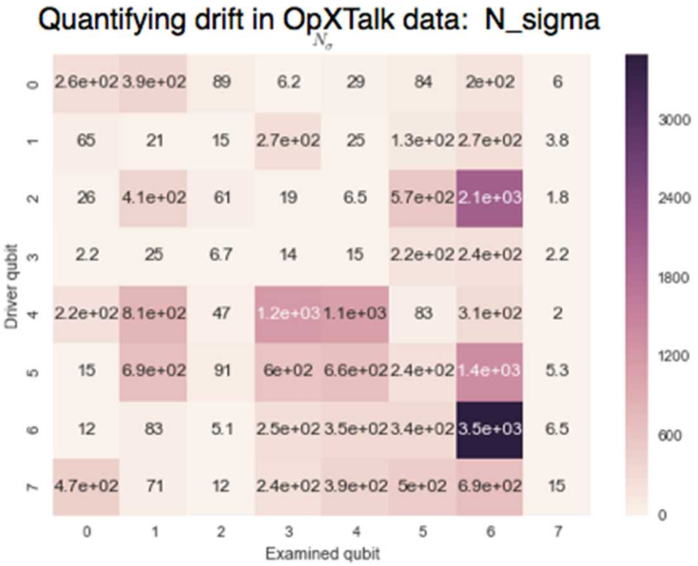
Drift and crosstalk

Using the fruits of the theoretical/conceptual development in the first part of this project, we developed and performed (in collaboration with RQC) a variety of experiments to detect and characterize crosstalk. However, these were preliminary and represented our first experimental attempt to detect crosstalk in a systematic way. The main result – so far – is that we've determined most of the qubits have so much <u>drift</u> that we can't apply our crosstalk tests in a straightforward way.

We asked RQC to run two passes on all the crosstalk experiments. In the absence of drift, Pass 0 and Pass 1 of each circuit should be statistically identical. Checking this won't catch all forms of drift (e.g., probably not 60hz drift or 1/f noise, which are high frequency dominated), and it doesn't really let us <u>characterize</u> the drift, but at least we can do a sanity check.

Unfortunately, a lot of the experiments failed the sanity check. We divided the data into a lot of subsets — generally corresponding to either idle or operation crosstalk tests on specific pairs of qubits — and then "traced out" all but 1 or 2 qubits from the measurement outcomes, by ignoring the measurement outcomes on 6 or 7 of the qubits. This yields a "coarse-grained" dataset describing either a 2-outcome measurement on 1 qubit, or a 4-outcome measurement on 2 qubits. Then, for each subdataset, we go through the sequences and compare the observed frequencies from Pass 0 to those from Pass 1. Typically, we'll compute a statistical measure of how implausible it is that both frequencies came from the same underlying probability, and we'll add that up over all the sequences in the subdataset. Then we ask whether the sum is consistent with normal statistical fluctuations or not — and, if not, how many σ it is away from consistency.

Here is one of those plots:



For this plot, we took operation crosstalk data (where one qubit is driven while the other 7 are monitored (using LGST) in parallel. For each of the 7 driven qubits (rows) we look at each of the 8 spectator qubits in turn (columns), ignoring measurement outcomes on other qubits, and look for drift in its behavior. Drift is quantified here by # of σ model violation, but these numbers aren't directly comparable to the ones discussed previously (these are different experiments from

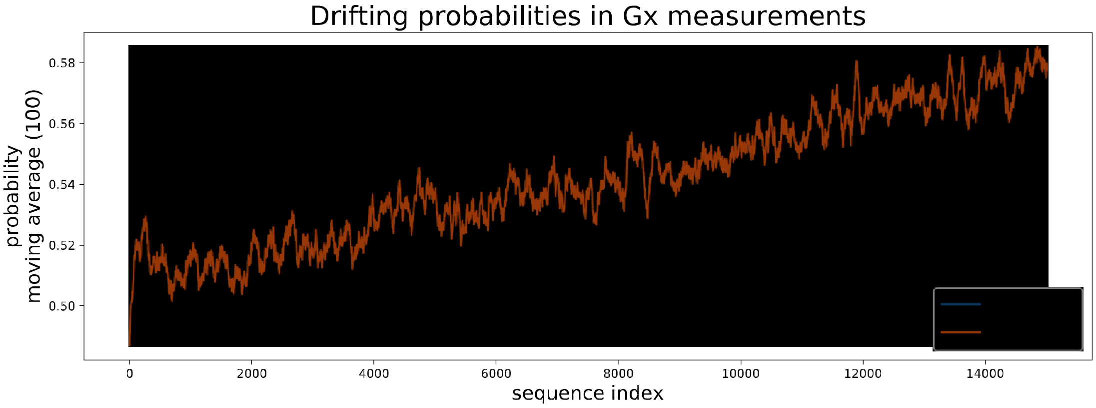
1-qubit GST, with a different number of component circuits).

Basically, what we see here is that all qubits except Q7 and Q2 display a lot of drift. And Q2 is questionable — it's better than the others, but it's still got 90 σ of drift in two of the experiments (when Q5 was driven and when #1 was driven). Qubit Q6 is, again, uniquely bad. This is probably what caused its bad 1QGST results.

If we set aside Q7 and Q2 as being pretty good, and Q6 as clearly a mess, then the remainder are harder to describe. Each of them appears to have drifted significantly ($N_\sigma > 200$) in at least one experiment. There's no clear pattern as to which experiments each qubit *did* drift in — which is to say that whether Qubit X drifts doesn't seem to depend on whether we're driving Qubit Y. Which is as expected. But the sporadic nature of it (Q1 has 810 σ of drift in one experiment and just 21 σ in another) suggests that we're not seeing *homogenous* drift (e.g. Brownian motion), but something more like occasional jumps from one regime to another.

We also dug into the data to see how big the changes were from one pass to another. And, importantly, whether the typical shift in sequence probability increased with the *length* of the sequence (L). We found that:

1. There are a fairly large number of sequences for which the observed frequency shifts by 20-40%. This is not a small effect, and it's not even close to the statistical significance threshold. There are individual sequences for which the jump from Pass 0 to Pass 1 is statistically significant at the 100 σ level.
2. **The size of the typical jump does *not* seem to correlate strongly with L.** This is very important, because it suggests that it's not the gates that are drifting — it's probably the SPAM. Effects that are independent of L are SPAM effects. We suspect this is actually a *measurement* effect — perhaps due to recalibrating the classifier.
3. In a separate experiment specifically intended to study drift (see figure below), we saw drift that was clearly SPAM drift rather than gate drift. The probabilities for two Ramsey experiments, one with a single X gate and one with 445 consecutive X gates, drifted more or less in unison. However, that drift was relatively slow and mild. We conjecture that the drift seen here is related to the measurement classifier, and that it gets much worse when the measurement has more (e.g. 256) outcomes.



So far, we have not been able to do any really solid crosstalk characterization because the drift has gotten in the way. We *did* start running analyses on both datasets (operation and idle). Summary of results so far:

Idle: Here, the goal is basically to do GST on the 2-qubit idle (Gii) gate between each pair. Unfortunately, so far when we've run this, the optimizer hangs and doesn't converge. This might be for a variety of reasons, but usually it happens when the data is confusing enough that GST's usual assumptions about things are wrong. (For example, we assume that when we reconstruct Gii^4, it's going to be pretty close to the 4th power of Gii. But if the qubit changed drastically between when we studied Gii^4 and when we studied Gii, that may be false, and the optimizer gets lost in local minima). The fact that we *know* drift was a major problem here suggests that this is likely to be the culprit.

Operation: Here, we want to see what happens to each of the 7 spectator qubits when we drive a target qubit. The first step in doing so is to compare two datasets — one where we drove the target, and one where we didn't — and see if they're different. In most cases they are – but in almost all cases, this difference is smaller than the difference between Pass 0 and Pass 1. When it's not smaller than the drift for a particular circuit, we see many other circuits whose "drift" difference is bigger, which makes it hard to have confidence that when we see a difference between "drive" and "no drive", we're not getting faked out by simple drift. Nonetheless, a scan of the data suggests that there *is* probably cross-driving between pairs [3,4] and [1,2]. It's tempting to guess that there might be some between [5,6] and maybe [1,6] as well, except that it's almost impossible to say anything about Q6 because it drifts so wildly. Some of our experiments may have caught it within a period of stability, but it's hard to be sure when we know that it jumps around.

We can probably extract more information from the crosstalk data with effort, but the amount of drift (whatever it is) is going to seriously hamper what we can get out. A better approach would be to eliminate/reduce the drift and redo the experiments. Alternatively, if the drift cannot be eliminated, we'll seek to do some experiments to try and figure out how things are drifting, so that we know what we're up against, and then redo the crosstalk experiments in a different way that is designed to mitigate the effects of drift, by arranging the experimental design so that sequences that get directly compared are performed as close to each other in time as possible.

DISCUSSION:

Theory: This project achieved *all* of its theoretical and computational goals (which were of specific limited scope, given the small amount of time and money available). We developed the initial idea of "efficient, low-weight generators" into a concrete model. We implemented that model in pyGSTi. And then we designed algorithms to fit data to that model, formulated experiments to provide that data, simulated the experiments and ran the fitting algorithms on the resulting data, and confirmed that the modified pyGSTi software was indeed able to reconstruct the gate-set that generated the data.

In the process of all this success, we *also* discovered a lot of good challenges for future research. (These might be called "failures" in contexts other than Exploratory Express). Having successfully tamed the explosive increase in parameter count with the number of qubits, we discovered that the computing power required to compute circuit probabilities (which has to be done millions of times to compute the likelihood function in the data-fitting optimization loop) still scales exponentially with the number of qubits, and this rapidly becomes the bottleneck. This is currently a *bottleneck*, not a *showstopper*, and we are developing creative ways to work around it (either by speeding up simulation, or by limiting experimental design to circuits that can be simulated faster).

We also found, in the process of deploying our nascent protocols to characterize the RQC system, that *drift* is a much bigger obstacle to characterizing crosstalk than it is for characterizing local gates with GST, and than we had expected. After devising models for crosstalk based on low-weight generators, we had then proceeded to develop *tests* for crosstalk that are based on statistical variation between different experiments. However, these experiments need to be done simultaneously (with their counts interleaved) in order to average out the effects of drift. We discovered that RQC did not have the capability to do this rapid interleaving in FY17, and this forced us to resort to "chunked" experiments that are much more drift-sensitive. We have reacted by shifting our focus away from *testing* to *characterization*, by making our experiments more over-complete (as standard GST is), and by focusing on experimental design to minimize the effects of drift.

Experiment: The RQC 8-qubit device appears to be a legitimate 8-qubit QIP. All eight qubits (with the exception, right now, of Q6) are functional and controllable with stochastic error rates between 0.25% and 1%. Diamond norm errors are somewhat higher, but this is due to coherent errors that can probably be calibrated away. These are some of the most stable 1-qubit gates we have seen, and in the right situation they are extremely Markovian. This supports the conjecture that coherent errors can be eliminated by better calibration. It is worth noting, however, that while 1% error rates are consistent with the design spec, and respectable, they are not competitive with the best in the world, which are around 100 times lower.

SPAM – and probably measurement – is currently a weakness of the device. All the qubits have around 10% SPAM error, which we suspect is due to measurement failures. This is something of a big deal for both VQE (variational quantum eigensolvers) and quantum error correction. However, we suspect that some of this error is due to classifier failure, so it's not clear how much of the measurement error is intrinsic to the system and its control rather than to suboptimal post-processing.

2-qubit gates are the glaring weakness. A full set of 8 nearest-neighbor C-PHASE gates is essential for this to be a *useful* QIP. RQC has consistently shown the ability to tune up a genuine C-PHASE gate between two good qubits – but this will need to be done for *all* pairs, and this is clearly a big part of their current agenda. But it's a bit worse than that, because 10% error rates are not only uncompetitive, but also a complete showstopper for all applications. Our analysis of the error generators suggests that 5% error rates might be relatively low-hanging fruit (better control of the Z-basis Hamiltonian, and/or dynamical decoupling), but it's not clear how they will go beyond this. This gate does not appear to be anywhere close to T1 limited.

Crosstalk is clearly present, but the dominant form of crosstalk appears at this time to be idle crosstalk (always-on ZZ interactions) rather than operation crosstalk. However, our conclusions at this time are worse than preliminary; we need to deal with the SPAM drift before we can draw any useful conclusions about crosstalk, and we haven't looked for crosstalk caused by 2Q gates.

ANTICIPATED OUTCOMES AND IMPACTS:

This project produced a wealth of follow-on impacts. Most notably, we used the preliminary results established here, and the proof of principle that was demonstrated with pyGSTi, to propose a full 3-year LDRD project that received extremely positive reviews and was funded. The Exploratory Express program functioned exactly as (we believe) intended: we were able to rapidly develop an idea, prove it out, confirm that it had legs, and then develop it into a more solid proposal. This work is underway now, and we are extremely positive on its eventual outcomes – previous LDRD projects with similar goals in the past led to enthusiastic external sponsorship, and we have good reason to believe that this R&D (if successful) will continue this tradition.

Specific support for this goal comes from another concrete outcome: the success of this exploratory investigation enabled us to present, with high confidence, a roadmap for the future of quantum device characterization for one of our SPP sponsors. This was *not* a use of LDRD funds to support SPP work; instead, our successful LDRD research, conducted independently and at the same time as more focused SPP contract work, allowed us to outline a future roadmap for where that sponsor might expect future work by all of its researchers to go. Our vision was well received and paves the way for future continued interactions with that sponsor, as well as continued leadership in the QCVV field.

More concretely, both this project and the follow-on LDRD work are expected to produce *research* outcomes (both theoretical constructions and specific software implementations of protocols) that will provide critical support for Sandia, external, and US government R&D in quantum computing. We have successfully shown that it is feasible to turn realistic amounts of data into precise, predictive models of quantum processors. We've also identified some of the obstacles that stand in the way of concrete tools to do this. In a few years, we will have overcome these obstacles, and we expect to be routinely probing and identifying the noise and error modes in systems of 3-10 qubits. This will enable experimentalists to debug their qubits; it will enable theorists and architects to foresee errors and mitigate them; and it will enable sponsors and program managers to make wise decisions about allocation of resources and forecast the capabilities of processors before they are fully deployed.

CONCLUSION:

The premise of this project was simple: *Quantum tomography is important but hard. We might be able to make it much easier by using reduced models of noise.*

This premise is now a promise: *We can make tomography work faster, better, and cheaper by using low-weight generator models.*

We demonstrated that this works, and in the process of doing so, we took the first steps toward useful tools for characterizing multi-qubit systems. Then, we deployed those proto-tools on a real (if noisy) 8-qubit processor. We turned our failures in this exploratory process into valuable lessons for future development, including the importance of fast forward simulation, and the frustrating ability of drift to interfere with otherwise robust characterization protocols. We expect that our continued R&D in this area, funded by Sandia's LDRD program, will produce concrete tools that will be useful (and essential) for debugging qubits and supporting the continued growth of quantum computing for national security and economic competitiveness.