SAND2019-11842C

Sandia National Laboratories

Emergent Errors in NISQ Devices

PRESENTED BY

Kevin Young

Robin Blume-Kohout, Erik Nielsen,
Tim Proctor, Kenny Rudinger, and Mohan Sarovar

Quantum Performance Laboratory
at Sandia National Laboratories
Livermore, CA and Albuquerque, NM
(and Gettysburg, PA and Los Angeles, CA)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

Emergent Errors

- Small scale characterizations (1- and 2- qubit gate fidelities) are not predictive of holistic device performance
- Errors responsible for this deviation are generically called **crosstalk**
- Crosstalk makes modeling NISQ systems even harder that you might think
  - More is *different*
    - Purpose of Anderson's paper was that increasingly complex systems have *new models* – eg., particle based vs continuum or statistical models
  - Models naively stay the same type, but grow exponentially in complexity
  - We can construct reduced models that accurately describe NISQ behavior, but remain learnable and tractable

Emergent Errors

- Small scale characterizations (1- and 2- qubit gate fidelities) are not predictive of holistic device performance
  - *I'll show some examples*
- Errors responsible for this deviation are generically called **crosstalk**
  - *I'll define crosstalk more rigorously and provide a statistical test to identify*
- Crosstalk makes modeling NISQ systems even harder that you might think
  - More is *different*
    - Purpose of Anderson's paper was that increasingly complex systems have *new models* – eg., particle based vs continuum or statistical models
  - Models naively stay the same type, but grow exponentially in complexity
  - We can construct reduced models that accurately describe NISQ behavior, but remain learnable and tractable
  - *I'll define some reduced models and a few protocols for fitting their parameters*
  - *I'll present a method for rigorously dealing with unmodeled error*

Characterizing the holistic performance of contemporary NISQ systems

Volumetric benchmarking with randomized mirror circuits

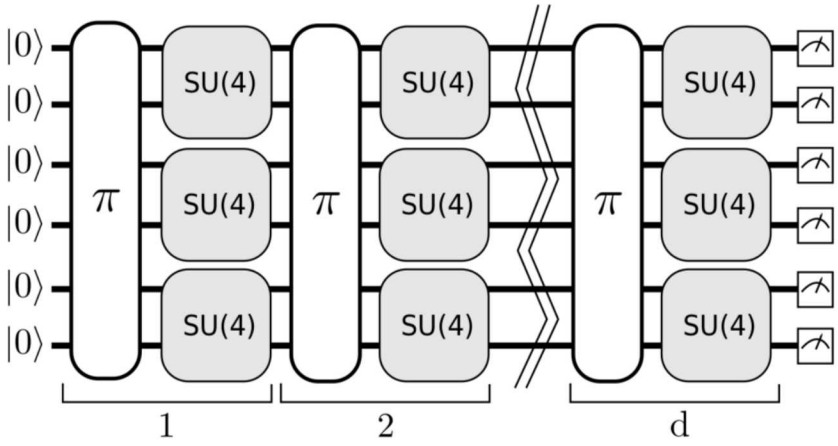
We already have benchmarks… why do we need more?

- We have lots of *low-level* benchmarks for characterizing few-qubit devices
  - Quantum characterization, verification and validation (QCVV) is an established field
  - A whole family of randomized benchmarking techniques
  - Gate set tomography
- But it is difficult to predict performance of large processors based on low-level benchmarks
  - Some failures only emerge at scale:
    - Crosstalk
    - Non-Markovianity
- Application-specific benchmarks would be great!
  - NISQ devices can't run any useful applications!
- We need high-level, holistic benchmarks able to assess the performance of as-built, NISQ processors that are sensitive to all the weird things that can go wrong as quantum devices scale up.

Inspiration: IBM's Quantum Volume

- It's relatively easy to print a bunch of superconducting qubits on a circuit board, but tuning them up and making them work together is much harder.
- IBM recognized:
  1. Adding more qubits doesn't increase the computational power of the device if they already decohere before the can all be coupled together.
  2. Decreasing the error rate per operation doesn't increase the computational power if you can already reliably access any state on all available qubits.
- What is the *effective* size of the device:
  - "What is the largest number of qubits on which the processor can reliably produce a random state?"

Quantum Volume



- A class of quantum circuits:
  - *d* qubits with *d* rounds of scrambling
- A measure of success for each *d*:
  - Is the *heavy-outcome probability* > *2/3?*
  - An outcome is *heavy* if it is predicted to have $p_{meas} > p_{med}$, where p_{med} is the expected median outcome probability.
- The quantum volume is defined as
  - Where *d** is the largest *d* for which the above experiment is successful

$$V_Q = 2^{d^*}$$

Generalizing Quantum Volume

- Knowing the largest square scrambling circuit that a device can run reliably is interesting, but does not give a complete picture of the device's performance
- Depending on compilation choices, some algorithms can trade qubit number (circuit width) for circuit depth, so we might be interested in the device performance at a range of widths and depths.

Framework for volumetric benchmarks

1. A map from integers *(w,d)* to ensembles of circuits *C(w,d)*
2. A measure of success for each circuit
3. A measure of overall success at each *(w,d)* pair
4. ***Optional*:** An experiment design specifying how the circuits are to be run (how many repetitions per circuit? interleaved? adaptively chosen circuits?)
5. Plot all the data (we'll get to this)

Framework for volumetric benchmarks

1. A map from integers *(w,d)* to ensembles of circuits *C(w,d)*

- Randomized benchmarking
  depth-*d* sequences of arbitrary *w*-qubit Cliffords (plus inversion)
- Direct randomized benchmarking
  depth-*d* sequences of arbitrary w-qubit, depth-1 circuit layers
- Rabi oscillations
  d consecutive repetitions of a single layer of local 1-qubit gates
- Idle tomography
  specific depth-*d* Rabi/Ramsey-type circuits, parallelized over *w*-qubits
- Grover iterations
  d iterations of a single *w*-qubit Grover step (alternating oracle marking and reflection)
- Trotterized Hamiltonian simulation
  d iterations of a single Trotter step for simulating a *w*-qubit, 2-local Hamiltonian

Framework for volumetric benchmarks

2. A measure of success for each circuit

- Randomized benchmarking
  Probability of unique correct outcome
- Direct randomized benchmarking
  Probability of unique correct outcome
- Rabi oscillations
  TVD between predicted and observed local outcome probabilities < threshold
- Idle tomography
  TVD between predicted and observed local outcome probabilities < threshold
- Grover iterations
  Heavy outcome probability, cross-entropy, or other metric
- Trotterized Hamiltonian simulation
  Heavy outcome probability, cross-entropy, or other metric

Framework for volumetric benchmarks

3. A measure of overall success at each *(w,d)* pair

- Randomized benchmarking
  Uniform average over all sampled circuits
- Direct randomized benchmarking
  Uniform average over all sampled circuits
- Rabi oscillations
  N/A (just one circuit at each w/d pair)
- Idle tomography
  All individual circuits must succeed
- Grover iterations
  Averaged single-circuit criterion
- Trotterized Hamiltonian simulation
  Require all circuits to succeed (based on single-circuit criterion)

Volumetric Benchmarking Plots

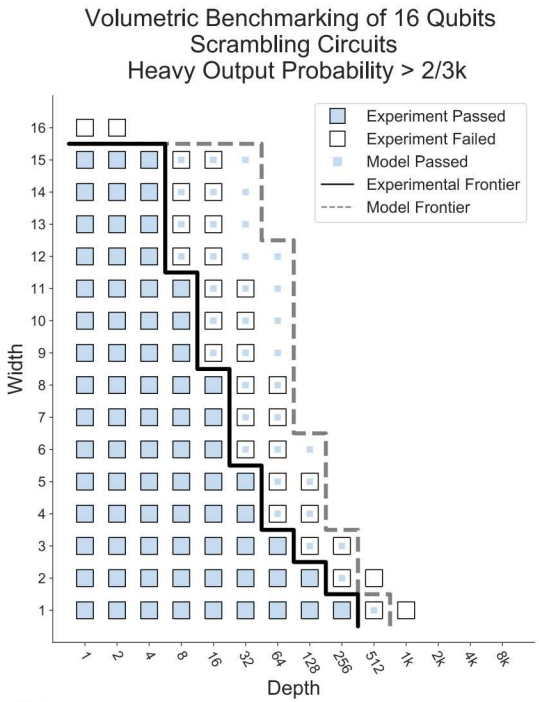


Figure 4(b). 16-qubit volumetric benchmarking using scrambling circuits. Pass/fail criterion: heavy-output probability $> 2/3$. The model was constructed using one- and two-qubit experiments. Qubit 16 was defective.

- Choose a set of widths and depths to measure and plot
- For each width/depth pair, run the circuits and indicate on the figure which ran successfully or successfully
- Caption should indicate what experiments were run

- In this figure we:
  - Use linear scaling on for the widths
  - Use logarithmic scaling for the depths
  - Show successful experiments in blue
  - Show unsuccessful experiments in white
  - Show successful model in small blue dots
  - Indicate Pareto front for model and experiment

Volumetric Benchmarking Plots

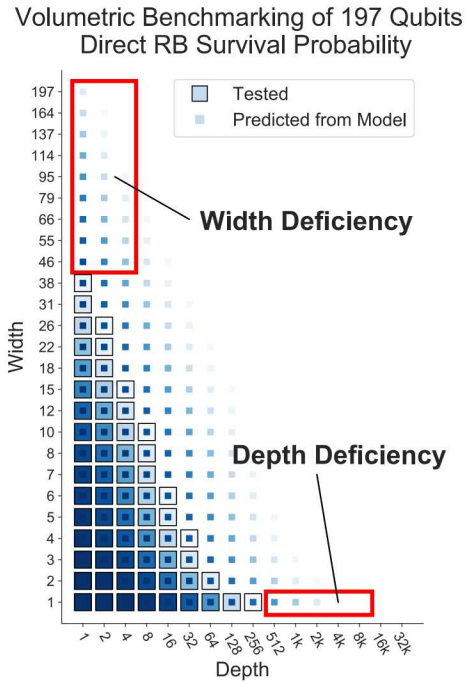


Figure 5(b). Performance of hypothetical 197-qubit device benchmarked with direct randomized benchmarking (DRB). Model predictions are based on a depolarizing noise model with parameters obtained using one- and two-qubit randomized benchmarking.

- Success metric might be continuous, rather than simply pass/fail
- Disagreement between models and experiment are immediately obvious when plotted together
- Width deficiencies can result from, eg., crosstalk
- Depth deficiencies can result from non-Markovianity

Volumetric Benchmarking Plots

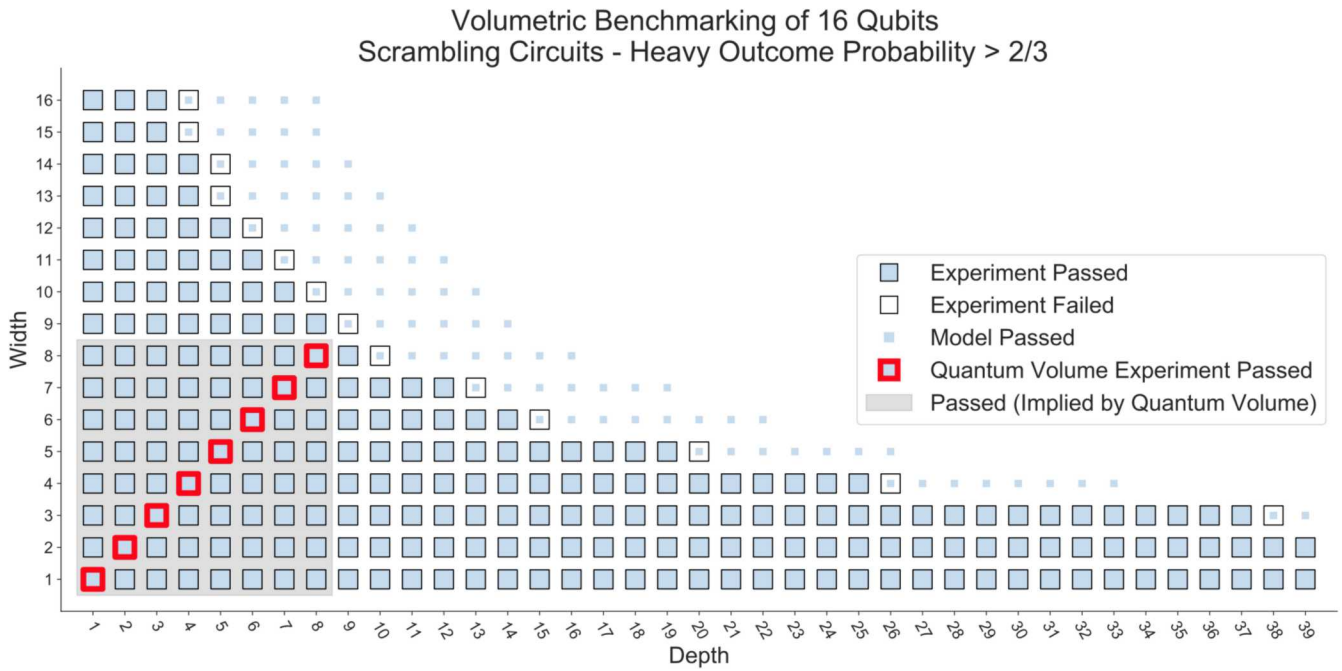


Figure 6(a). Volumetric benchmarking of a 16 qubit device using scrambling circuits. If at least 2/3 of the measurement results are heavy for a given width/depth pair, then the pair passes the test and is marked with a large, solid blue box. Using linear axes, the quantum volume experiments appear along the diagonal and are outlined with heavy, red lines. For this example, $\log_2(V_Q) = 8$. It is expected that scrambling circuits with both width and depth less than or equal to the quantum volume should succeed, and we highlight these with a grey background.

Volumetric Benchmarking Plots

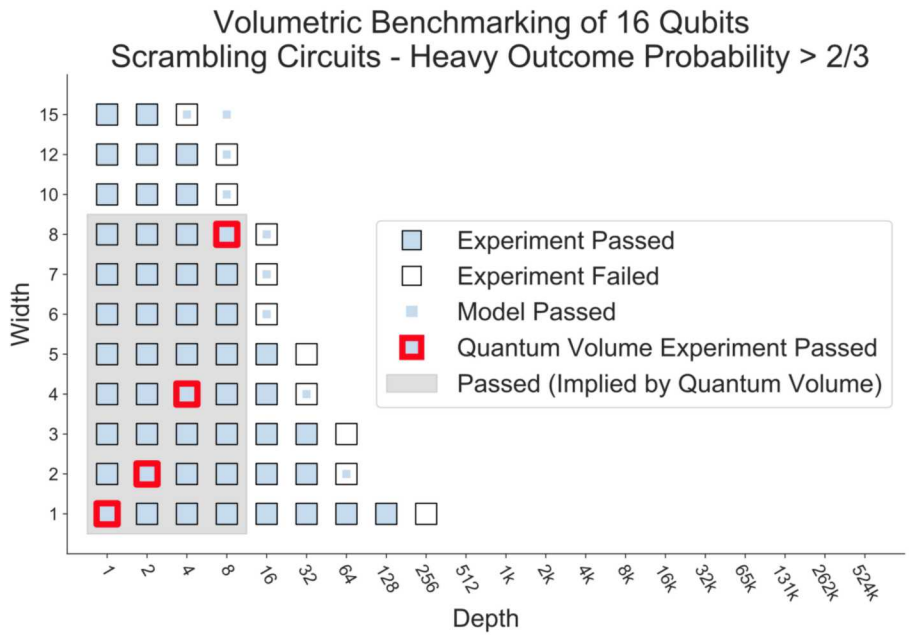
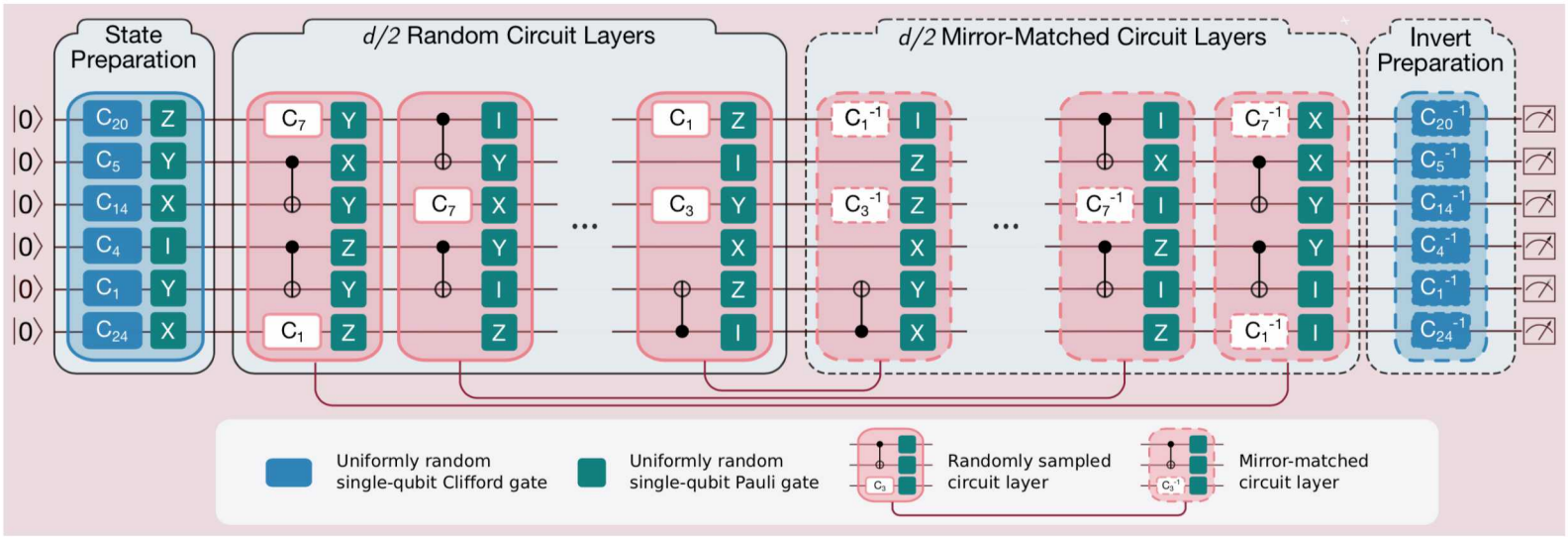


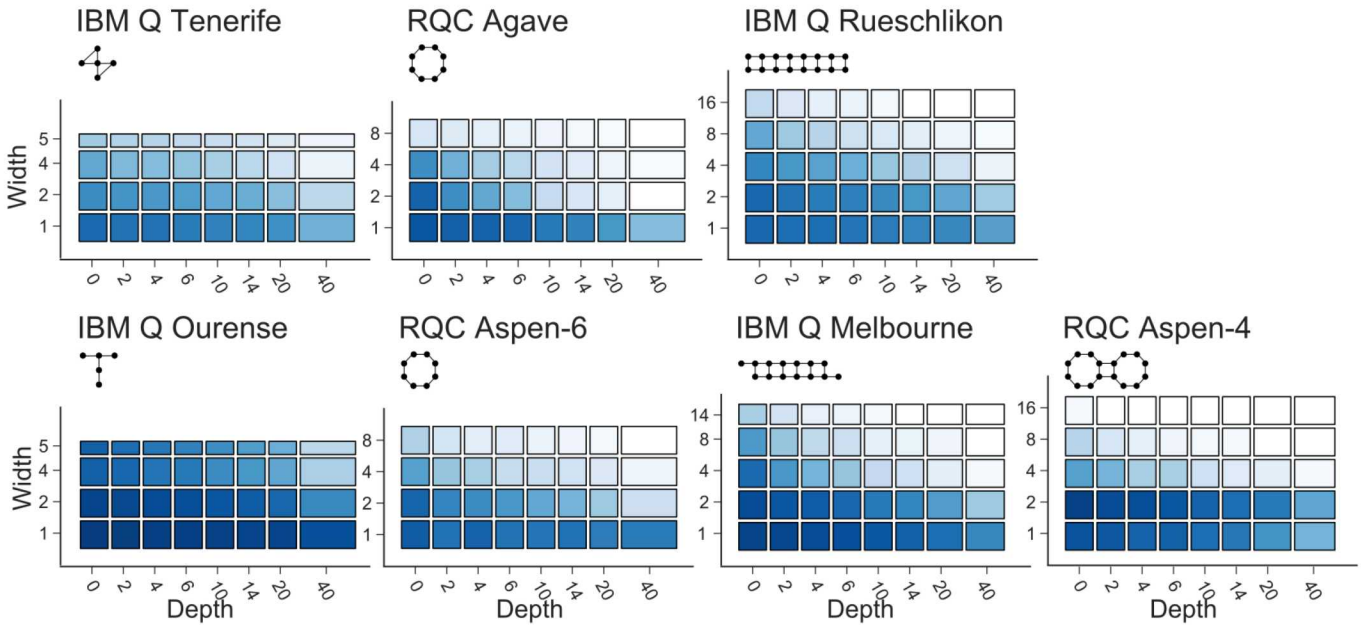
Figure 6(b). Volumetric benchmarking of a 16 qubit device using scrambling circuits. If at least 2/3 of the measurement results are heavy for a given width/depth pair, then the pair passes the test and is marked with a large, solid blue box. Using logarithmic axes, the quantum volume experiments appear along a curved line and are outlined with heavy, red lines. For this example, $\log_2(V_Q) = 8$. It is expected that scrambling circuits with both width and depth less than or equal to $\log_2(V_Q)$ should succeed, and we highlight these with a grey background.

Benchmarking NISQ devices with randomized mirror circuits



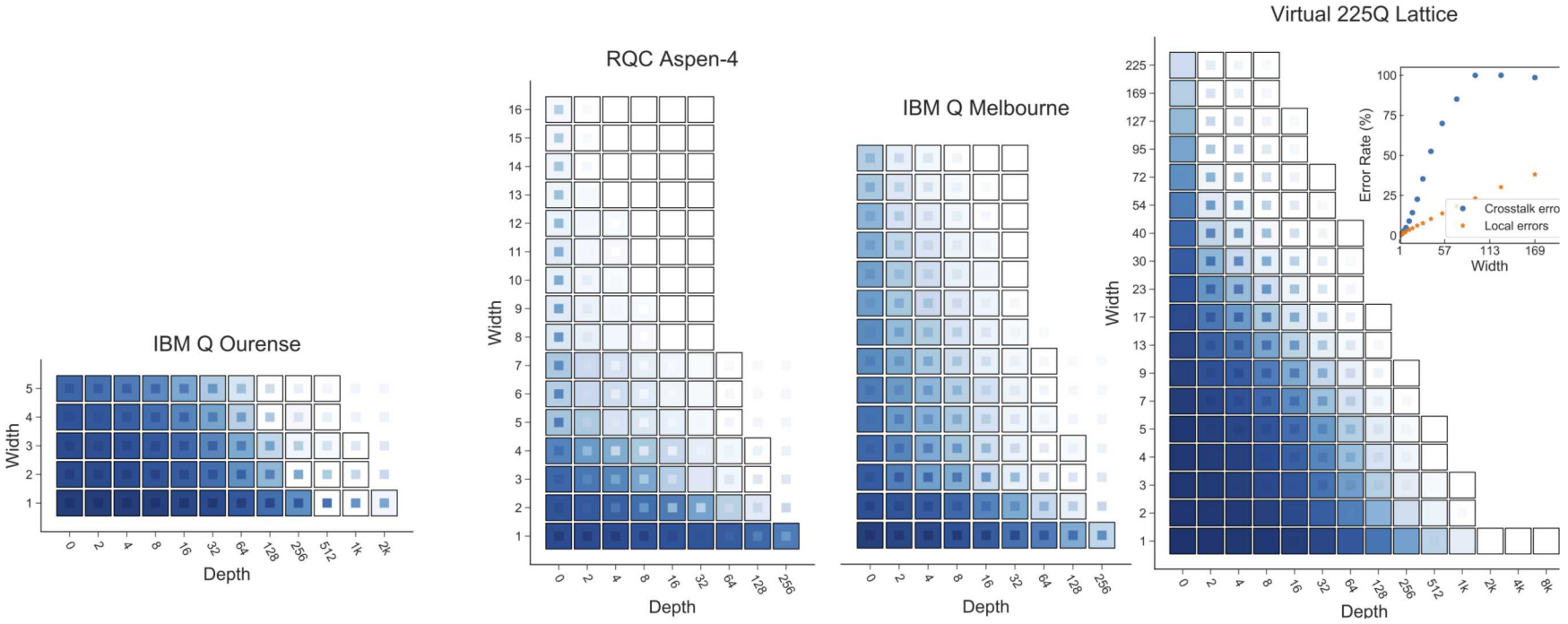
All circuits are Clifford, so predicting the outcome is easy (*eg.*, using Aaronson's CHP code)

Benchmarking NISQ devices with randomized mirror circuits

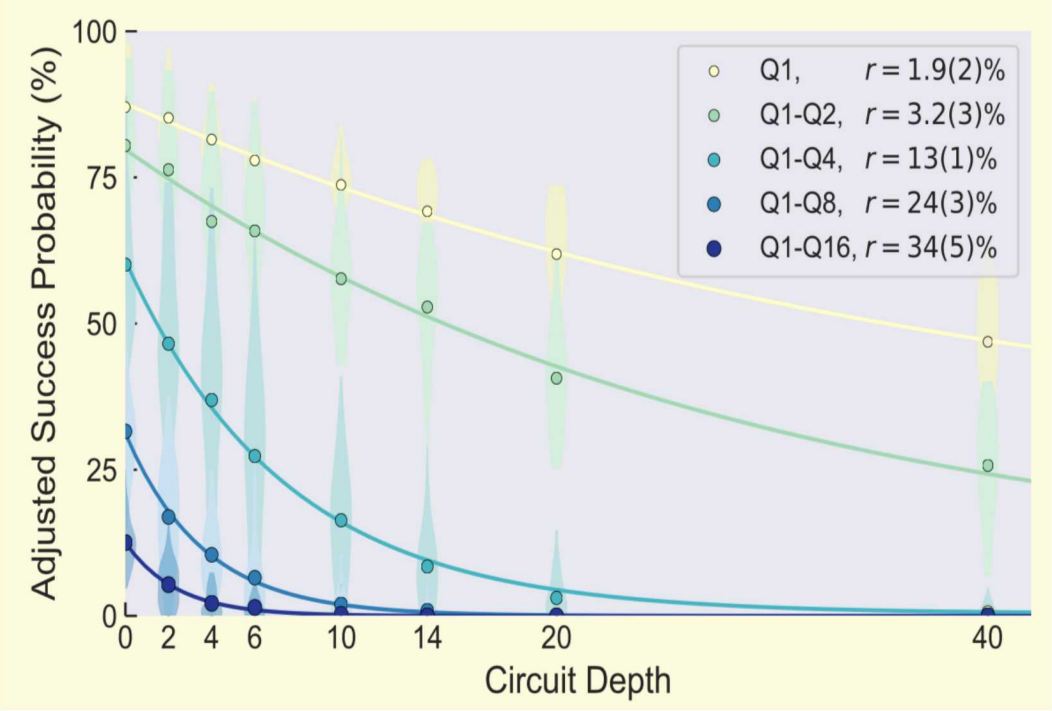


* Width-1 circuits are done simultaneously

Benchmarking NISQ devices with randomized mirror circuits



Benchmarking NISQ devices with randomized mirror circuits



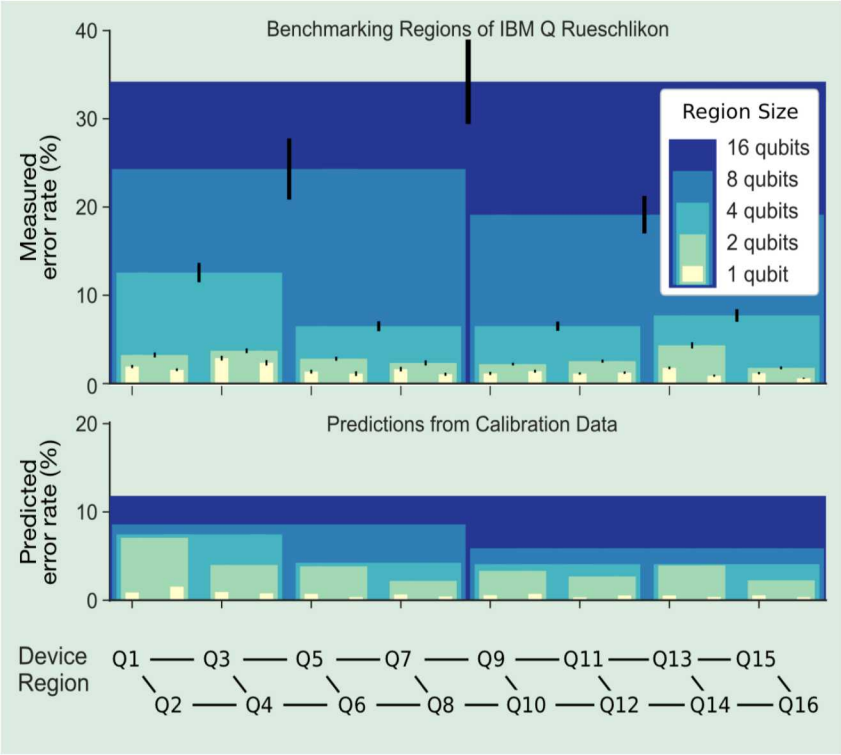
This is the adjusted success probability:

$$S = \sum_{k=0}^{w} \left(-\frac{1}{2}\right)^{k} h_k$$

Where h_k is the probability of an error with Hamming weight-k

Clear exponential decays for up to 16 qubits on real hardware. Each circuit family has a meaningful "error rate".

Benchmarking NISQ devices with randomized mirror circuits



Exceptionally rare for a device specification to capture crosstalk.

The measured error rate by region is **significantly** higher than the prediction *because of crosstalk*.

Defining crosstalk and identifying it in experiments

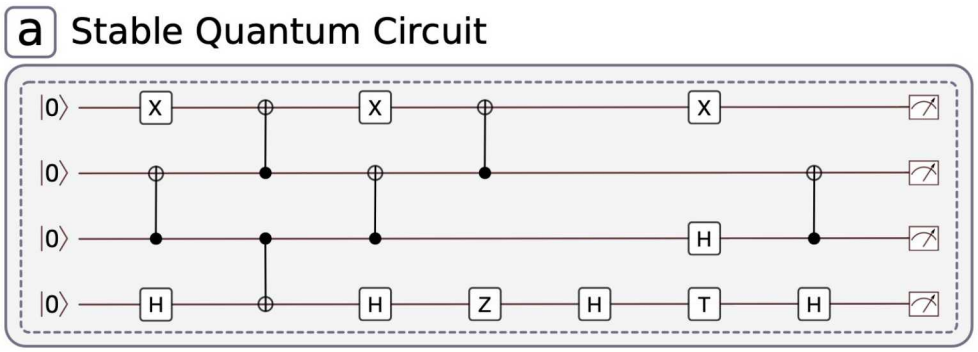
Conditional independence testing

What do we mean by *crosstalk*?

- **Crosstalk** – "Unwanted couplings between signal paths"
  - Couplings between transmons
  - Stray EM fields
  - Correlated 50Hz/60Hz line noise that influences all qubits
- **Crosstalk errors**
  - Any observable effect at the logical level (qubits/gates/circuits/outcome probabilities) that stems **uniquely** from some form of physical crosstalk
- We can define crosstalk by its absence

F. Mazda, *Telecommunications Engineer's Reference Book* (Elsevier, 1993).

Error model for a crosstalk-free QIP

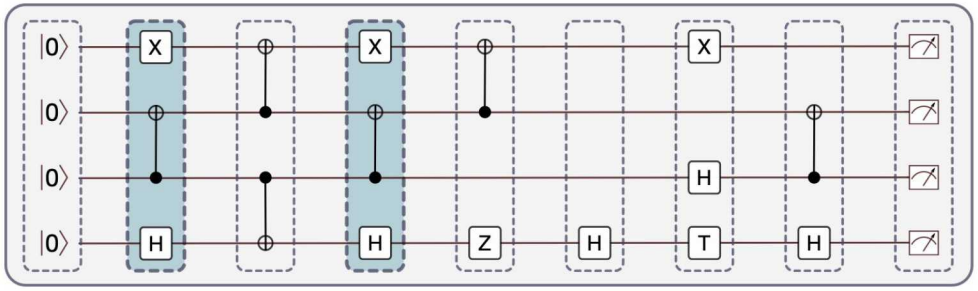


- The output probability of any given circuit is *stable* with respect to any external context, *eg*.:
  - Time of day
  - The circuit that was run previously
  - If Mercury is in retrograde

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)

Error model for a crosstalk-free QIP

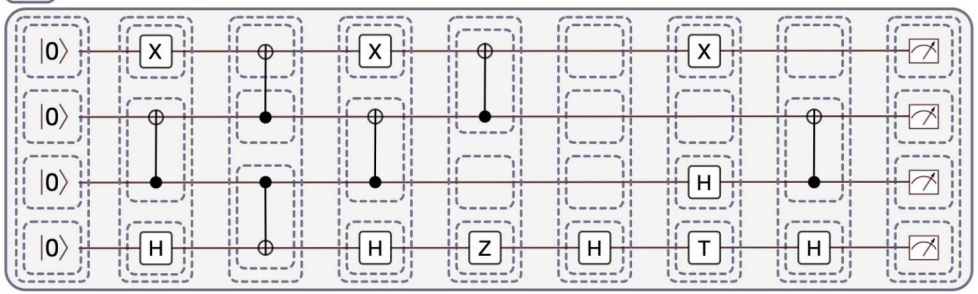
b Markovian Quantum Circuit



- We should be able, in principle, to describe and model each layer in the circuit with an n-qubit CPTP map.

Error model for a crosstalk-free QIP

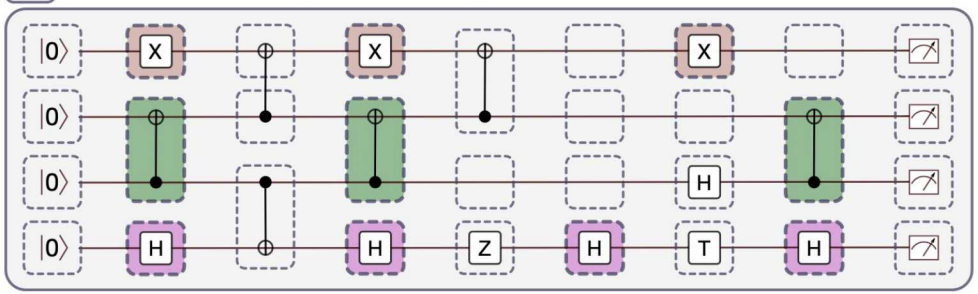
C Local Quantum Circuit



- We should, in principle, be able to describe and model each layer with an appropriate tensor product of CPTP maps.
- *Violations of this are* ***absolute crosstalk***

Error model for a crosstalk-free QIP

d Context-Independent Quantum Circuit

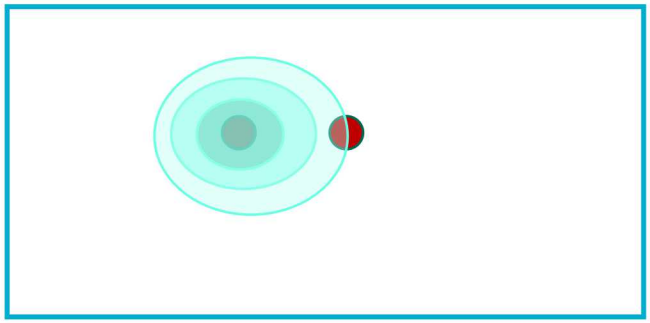


- There should be a unique CPTP map corresponding to each gate G on qubit(s) Q, regardless of what other gates are taking place on other qubits.
- *Violations of this constitute* **relative crosstalk**

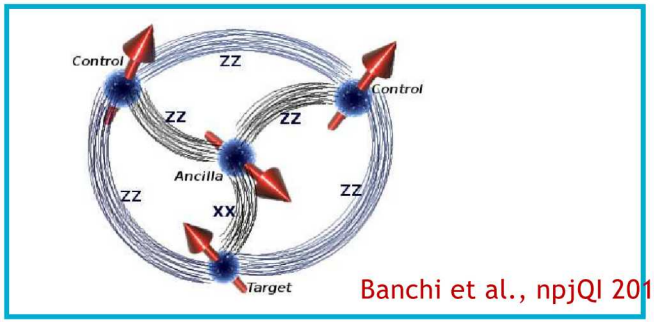
Defining crosstalk errors

- In summary, QIP is **crosstalk-free** if it obeys the following on arbitrary circuits:
- Locality of operations
  - Quantum circuit does not create correlation between qubits, or disjoint sets of qubits, unless that circuit contains multiqubit operations that explicitly couple them
- Independence of local operations
  - When an operation (gate, measurement, etc) appears in a quantum circuit acting on target qubits q at time t, the dynamical evolution of q at time t does not depend on what other operations (acting on disjoint qubits) appear in the circuit at the same time t.

Examples

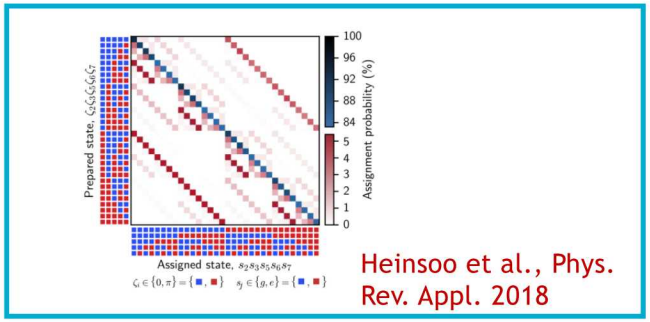


Pulse spillover, frequency overlap, poor addressability, etc.
Violates independence



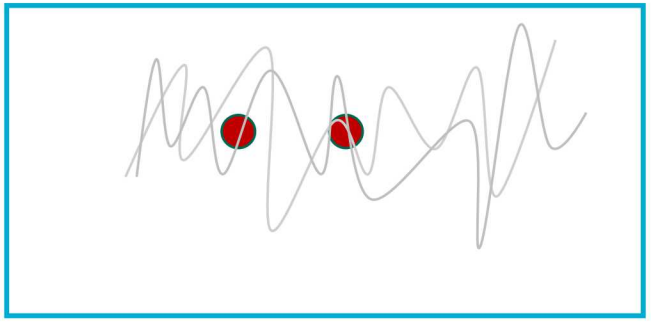
Banchi et al., npjQI 2015

Always-on Hamiltonian
Violates locality



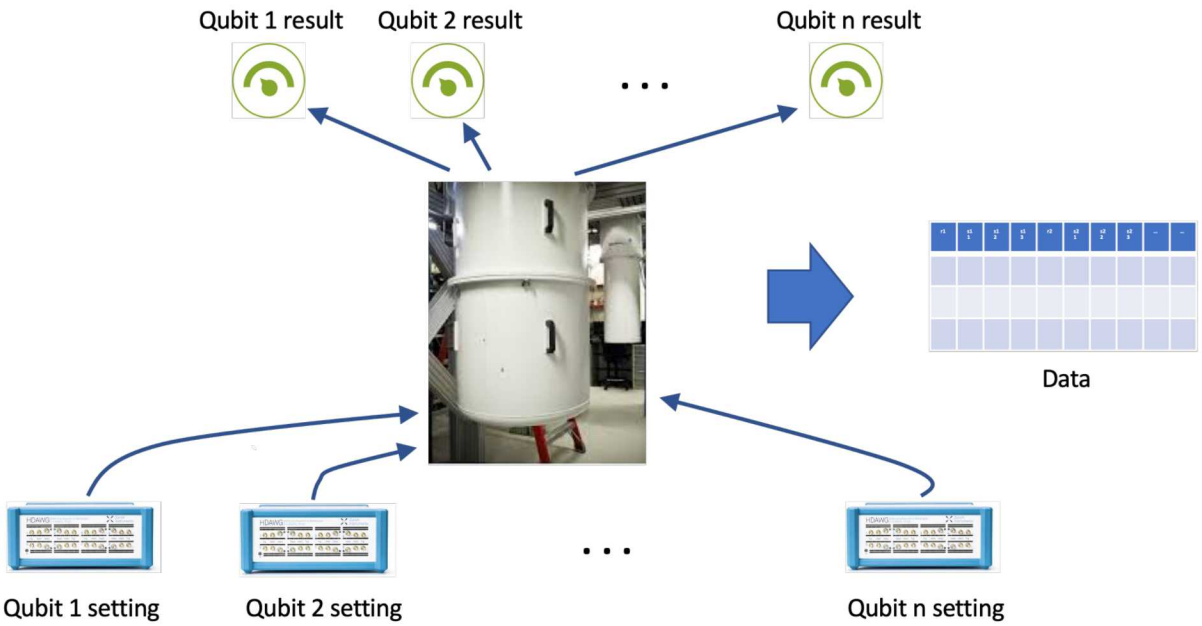
Heinsoo et al., Phys. Rev. Appl. 2018

Detection crosstalk
Violates locality

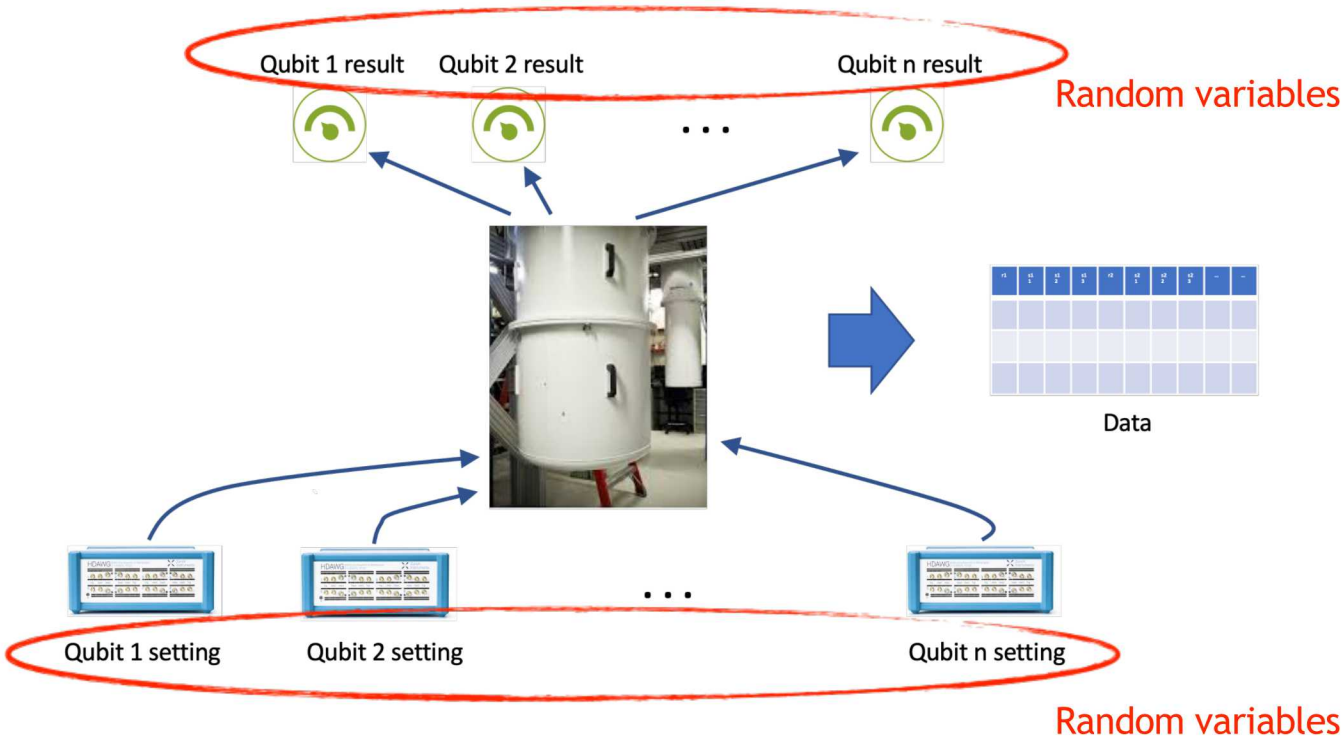


Correlated stochastic errors
Violates locality

The view from the lab



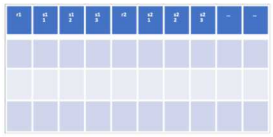
The view from the lab



Formalizing the view from the lab

Qubit 1 result Qubit 2 result ... Qubit *n* result

$X_{\text{result},1}$ $X_{\text{result},2}$ $X_{\text{result},n}$



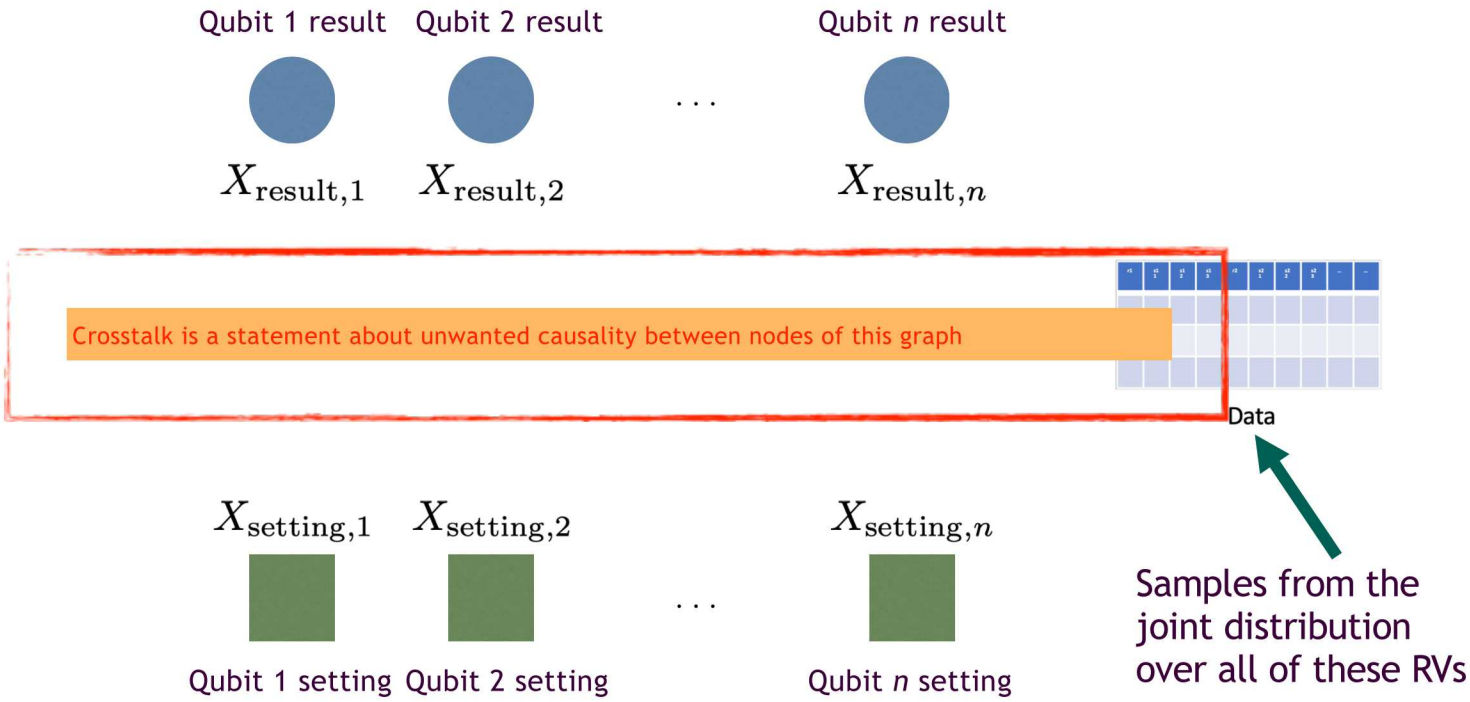
Data

$X_{\text{setting},1}$ $X_{\text{setting},2}$ ... $X_{\text{setting},n}$

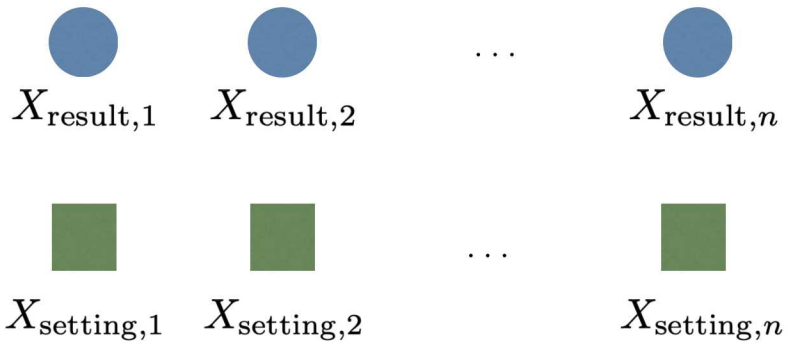
Qubit 1 setting Qubit 2 setting Qubit *n* setting

Samples from the joint distribution over all of these RVs

Formalizing the view from the lab



Formalizing the view from the lab



We define "no crosstalk" as

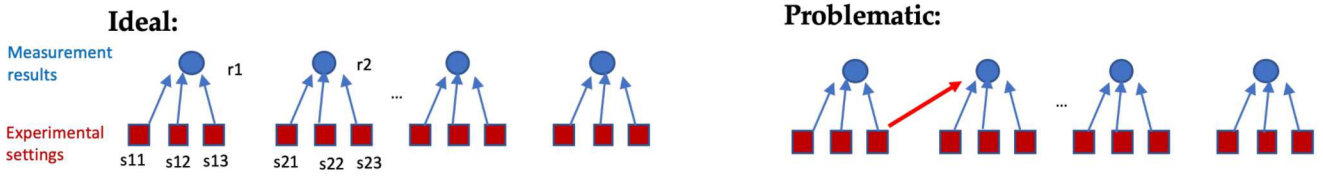
$$X_{\text{result},i} \perp (X_{\text{result},j}, X_{\text{setting},j}) \mid X_{\text{setting},i} \quad \forall j$$

i.e.,

$$P(X_{\text{result},i} \mid (X_{\text{result},j}, X_{\text{setting},j}), X_{\text{setting},i}) = P(X_{\text{result},i} \mid X_{\text{setting},i}) \quad \forall j$$

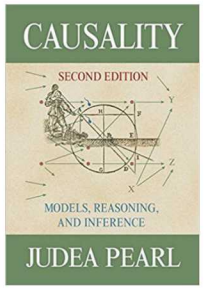
Implementation

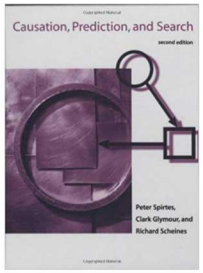
- Run a series of independent experiments on each of the qubits (or more generally, regions of multiple qubits) in parallel.
- The "settings" for each qubit/region can be:
  1. Gate sequence
  2. Initial state that is prepared
  3. Measurement basis
  4. Extrinsic information (*e.g.*, operating temperature)
- Collect measurements from these experiments.
- Reconstruct the causal graph from this data
  - Nodes are random variables representing settings and measurement results
  - Edges indicate a conditional dependence between nodes (a proxy for a causal relationship)



Causal graph reconstruction

We can build an implementation on a large body of literature dedicated to determining causal structure from sampled data.





Several algorithms exist for this task. We rely on *constraint-based* algorithms, that require two ingredients:

1. **A statistical test for conditional independence** $P(X_{\text{result},i} \mid (X_{\text{result},j}, X_{\text{setting},j}), X_{\text{setting},i}) = P(X_{\text{result},i} \mid X_{\text{setting},i})?$

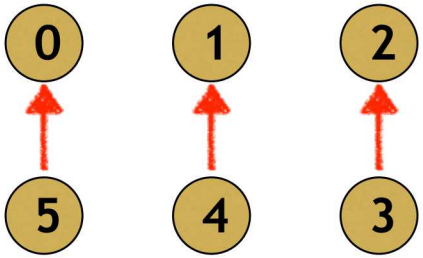
   e.g. G2 (log likelihood ratio) test

2. **A graph discovery algorithm that tests and prunes edges efficiently**

   e.g. the PC algorithm [Sprites & Glymour, 2000]

A number of statistical and algorithmic improvements are possible due to the structure in the crosstalk detection setting.

An example



6 qubits $X(\pi/2), Y(\pi/2), I$

Native gate set

Crosstalk error model

$$X_3(\pi/2) \rightarrow X_3(\pi/2) \otimes X_2(\epsilon)$$

$$X_4(\pi/2) \rightarrow X_4(\pi/2) \otimes X_1(\epsilon)$$

$$X_5(\pi/2) \rightarrow X_5(\pi/2) \otimes X_0(\epsilon)$$

Local depolarization at rate

$$r \sim \epsilon^2$$

Experiments:

1. Prepare each qubit in |0 >.
2. Measure in computational basis.
3. Setting for each qubit is the gate sequence applied to it.
4. Sequences are randomized benchmarking (RB)-like sequences.
5. Roughly ~390 total experiments.

An example

6 qubits $X(\pi/2), Y(\pi/2), I$

Native gate set

Crosstalk error model

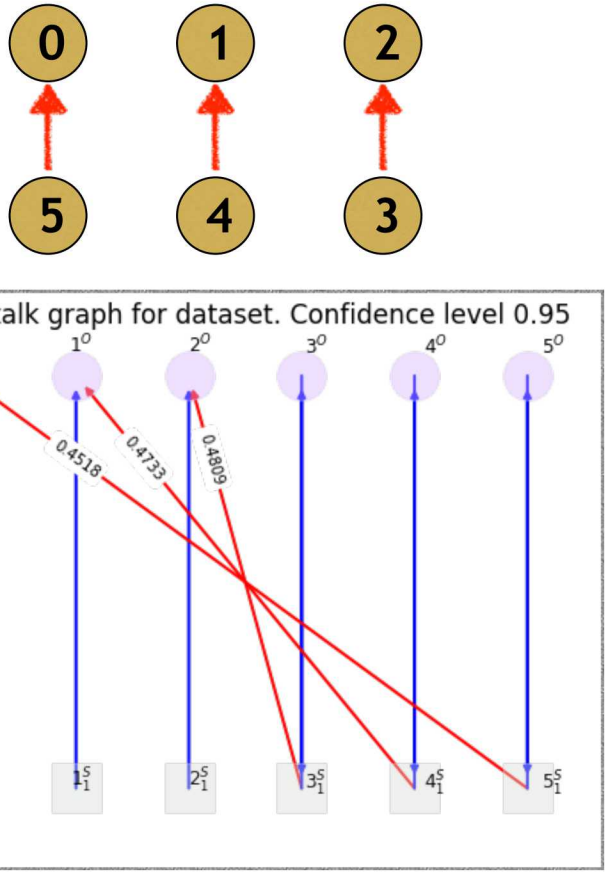
$$X_3(\pi/2) \rightarrow X_3(\pi/2) \otimes X_2(\epsilon)$$

$$X_4(\pi/2) \rightarrow X_4(\pi/2) \otimes X_1(\epsilon)$$

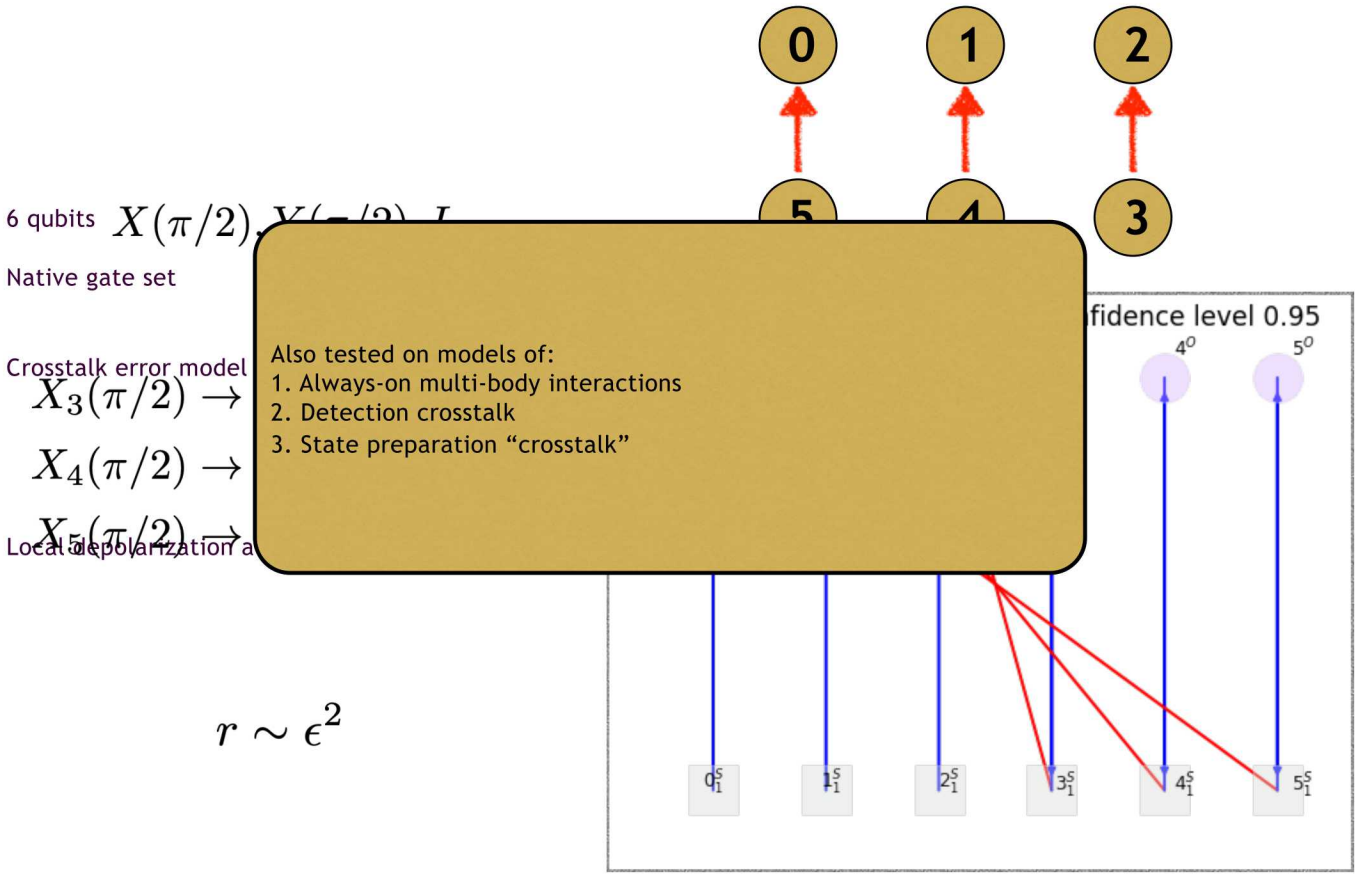
$$X_5(\pi/2) \rightarrow X_5(\pi/2) \otimes X_0(\epsilon)$$

Local depolarization at rate

$$r \sim \epsilon^2$$



An example



$$X(\pi/2)$$

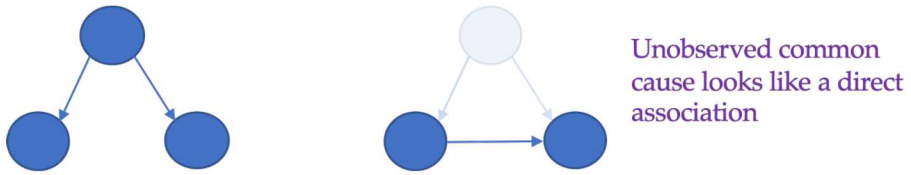
$$X_3(\pi/2) \rightarrow$$

$$X_4(\pi/2) \rightarrow$$

$$r \sim \epsilon^2$$

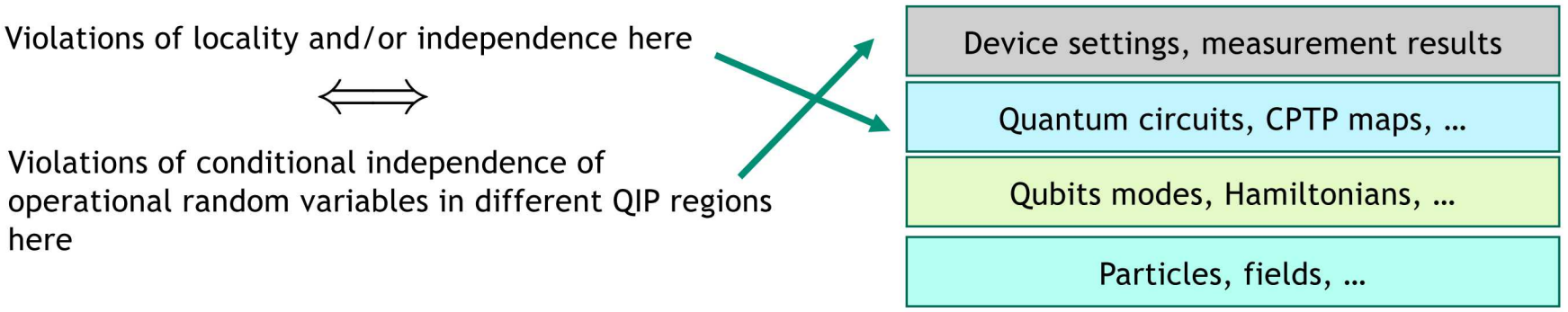
Summary

- General method to detect and quantify crosstalk:
  - Statistically motivated: can state results with degree of confidence and error bars.
  - Light-weight: requires number of experiments that scales linearly with number of qubits/regions.
  - Black-box: makes no assumptions about the form or physical origin of the crosstalk
- Pros and cons: the method captures all kinds of “crosstalk”, but as a consequence cannot distinguish between them.
- **Key issue**: crosstalk detection is confounded by drift. Must design experiment carefully to minimize drift confounding



Consistency

- We can show that the above definitions of crosstalk are equivalent to the definitions used in the crosstalk detection protocol (based on conditional dependency of operational random variables).



Defining and fitting reduced models for crosstalk

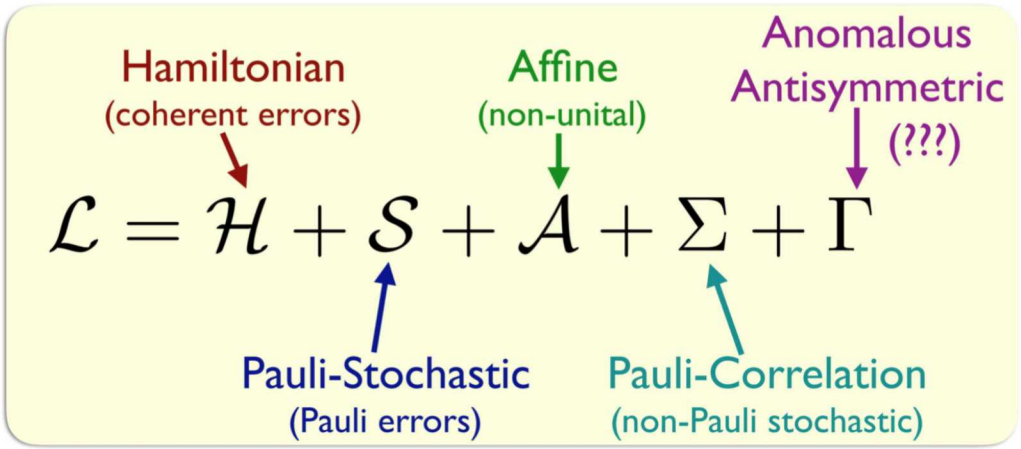
Idle tomography and reduced-model GST

Process matrix deconstruction

If $\boldsymbol{G}$ is the process matrix for a noisy version of unitary gate $\boldsymbol{G_0}$

$\boldsymbol{G}$: $\rho \rightarrow \boldsymbol{G}[\rho]$ $\qquad$ $\boldsymbol{G_0}$: $\rho \rightarrow U\rho U^\dagger$

then isolate the *error generator* by writing $\mathbf{G} = e^{\mathcal{L}}\mathbf{G}_0$.



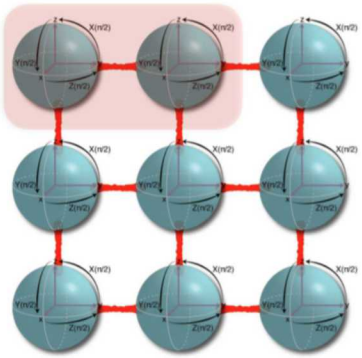
Get rid of the parts that don't matter

- A *generic* process matrix is described by $d^2(d^2\text{-}1)$ **rates** for all those distinct error generators. (240 rates for a 2-qubit process)
- Assuming all but $3(d^2\text{-}1)$ are negligible gives a simple reduced model:
  - 15 (d^2-1) **H**amiltonian error rates (a Pauli vector).
  - 15 (d^2-1) **S**tochastic error rates (a Pauli diagonal 2-tensor).
  - 15 (d^2-1) **A**ffine error rates (a Pauli vector).
- The Σ and Γ generators are less important and/or less likely to occur, so let's eliminate them from the model (=> assume they don't occur)

Restrict to low-weight generators

- $3(d^2\text{-}1)$ is less than d^4 , but still hopelessly big for d=2^N (N qubits).
- So let's focus on the qubits that are *idle* during a gate, and say:

It is reasonable to expect that idle qubits will only experience weight-1 and weight-2 errors.



- Weight-k generators act only on k qubits, e.g.

$$\dot{\rho}_{1\ldots N} = (X_1 \otimes X_2 \otimes \mathbb{1}_{3\ldots N})\,\rho\,(X_1 \otimes X_2 \otimes \mathbb{1}_{3\ldots N})$$

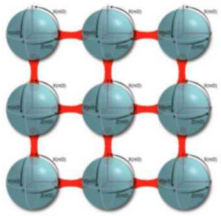
- Idle tomography measures the rates of all these errors efficiently. It will also *detect* many higher-weight errors, but not identify them.

Fitting the model

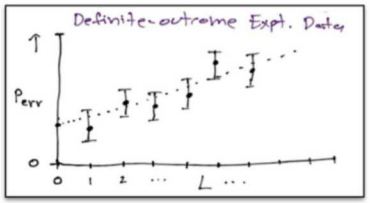
- We can fit the model with GST (maximum likelihood)
- Or we can use *idle tomography*

Idle tomography in 1 slide

- Define a small set of "experiments" in which each qubit is:
  1. Prepared in a Pauli eigenstate (e.g., $|\mathbf{0000}{+}{+}{+}{+}\rangle$).
  2. Idled for L clock cycles.
  3. Measured in a Pauli basis (e.g., **ZZZZXXXX**).



- We'll measure stochastic/affine errors with *definite-outcome experiments* (e.g. Z/Z or X/X as above), and Hamiltonian errors with *uniform-outcome experiments* (e.g. Z/Y or X/Z)



- To ensure SPAM-invariance, extract observable error rates from linear fits to data vs L.
- Each observable rate is a sum of $O(N)$ intrinsic rates ⇒ invert a matrix.
- Choose experiments cleverly to multiplex rates into few configurations.

Rigorously accounting for unmodeled error

Wildcard error budget

What if the models don't fit the data?

- Models will almost never *exactly* fit the data
  - If they do, you're probably over fitting
- We should be honest about how well our models do
- How can we convert "error in the predicted circuit outcomes" to "error in the underlying gates"?
- Proposed solution: **Wildcard TVD Budget**

Wildcard TVD Budget

- We assign to each operation a wildcard TVD budget (*W*)
- It quantifies the unmodeled error a gate *could* induce
- It is additive over all gates in a circuit
- The sum tells you how much we allow the observed outcome distribution of quantum circuits to differ from the model prediction.
- If a model assigns probability distribution P(i) to an event with possible outcomes {i}, and allows a wildcard TVD budget W for that event, then the model is consistent with observed frequencies F(i) iff there exists a probability distribution P'(i) such that
  - TVD(P,P') <= W
  - Data is consistent with P'

Interpreting wildcard TVD

- Wildcard error tells us how bad our reduced Markovian model is
- The *cause* of this badness could be:
  - Drift
  - Leakage
  - *Crosstalk*
- But it is difficult to quantitatively ascribe it to any of these error sources in particular
- Wildcard TVD can then be thought of as an additional error whose source is unknown *unless you've been exceptionally careful to rule out other sources*
- Wildcard error can be used to estimate how large a circuit you can perform before you can't trust your model

Acknowledgements

Lots of people to thank

Thanks

- IBM Quantum Experience and Rigetti Quantum Computing for generously providing time on their quantum computing platforms

Don’t blame the government!

- This work funded in part by Sandia’s Laboratory Directed Research and Development program.
- This material was funded in part by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research Quantum Testbed Program.
- This research was funded, in part, by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA).
- All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.
- 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.