SAND2020-2678C

Sandia National Laboratories

Demonstrating Scalable Benchmarking of Quantum Computers

PRESENTED BY

Timothy Proctor

Kenneth Rudinger, Kevin Young, Erik Nielsen, and Robin Blume-Kohout

Quantum Performance Laboratory
@ Sandia National Laboratories
Livermore, CA and Albuquerque, NM

Quantum Performance Laboratory

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.

Quantifying a quantum computer's capabilities

You have a noisy quantum computer. What do *you* want to know about it?

- The lab physicist:

What do I change in my lab to make it work better?

(or, how do I get a Nature/Science paper out of it?)

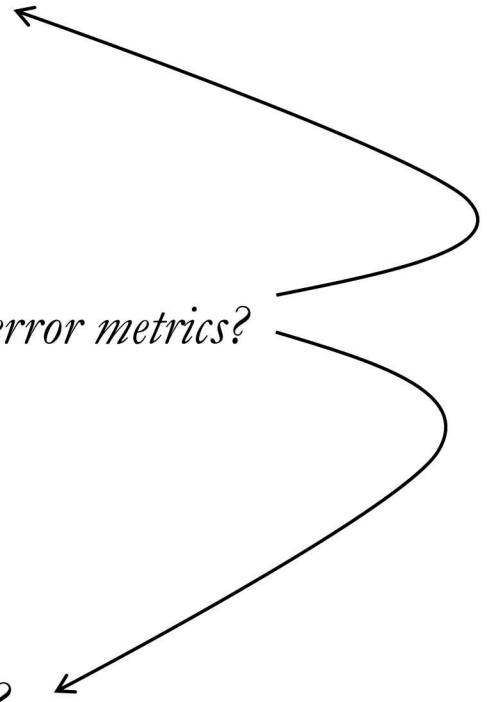
- The quantum computer characterization ("QCVV") specialist:

How does it perform according to my favorite characterization protocol and error metrics?

(or, can I run my latest characterization protocol on it and get into PRL?)

- A potential user:

What quantum programs can I successfully run on it?



Quantifying a quantum computer's capabilities

- A quantum computer's **capability set**:

  The set of *all* quantum programs it can successfully run.

- *Every* quantum program is categorized as successfully implementable on that hardware or not.

  $Circuit_1$: ***Yes***, $Circuit_2$: ***Yes***, $Circuit_3$: ***No***, $Circuit_4$: ***Yes***, $Circuit_5$: ***No***, ….

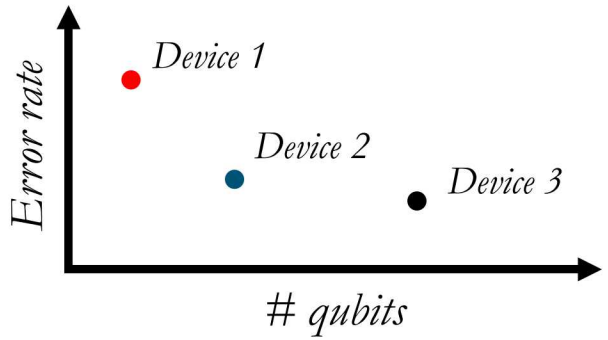
- We need to specify what running a program "successfully" means. E.g., TVD from ideal output distribution is below some threshold.

- In complete generality, the capability set of a quantum computer isn't efficiently representable.

- But we don't expect real hardware to have pathological errors. *We expect to be able to approximately predict how well a circuit will run from some summary properties of that circuit.*

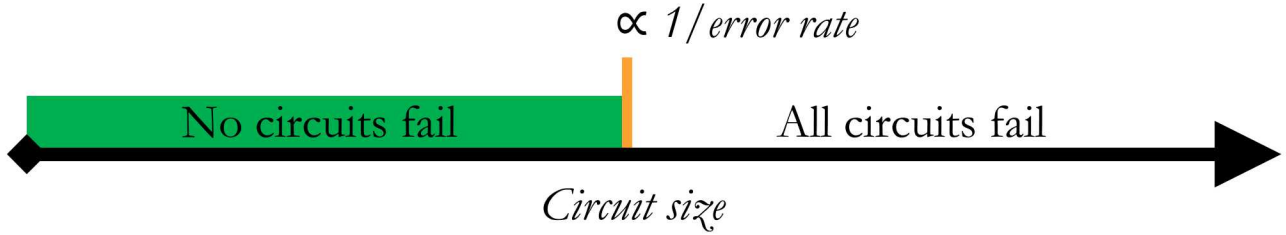
Quantifying a quantum computer's capabilities

- Consider the widely-used {# qubits, error rate} error model.



- The only meaningful interpretation of a {# qubits, error rate} pair is that it is implying that a circuit's success probability *approximately* depends only on it's size.

- This is a *very* efficiently representable capability set.

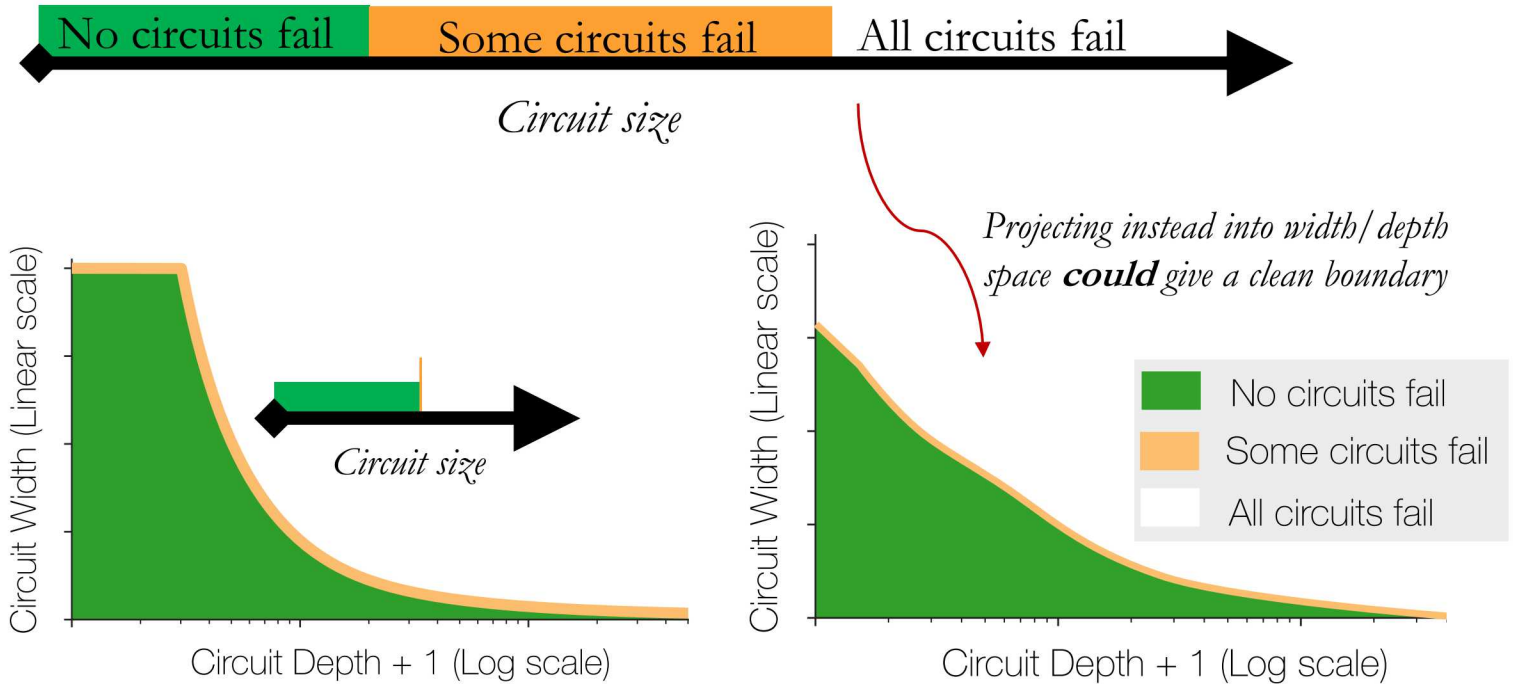


Projecting a processor's capability set onto width/depth space

- The circuit-size error model is naïve.

Example: It is not satisfied by a device subject only to local depolarization but with a depolarization rate that varies between qubits.

- We can still use the circuit-size "projection", but it will often have a fuzzy success boundary
- A simple generalization is to project circuits onto width/depth space.[1]
- A broader class of errors result in sharp success/fail boundary
- And we can still learn a lot using this representation even when the boundary isn't sharp.



[1]Blume-Kohout and Young, arXiv:1904.05546.

Estimating a processor's capability set

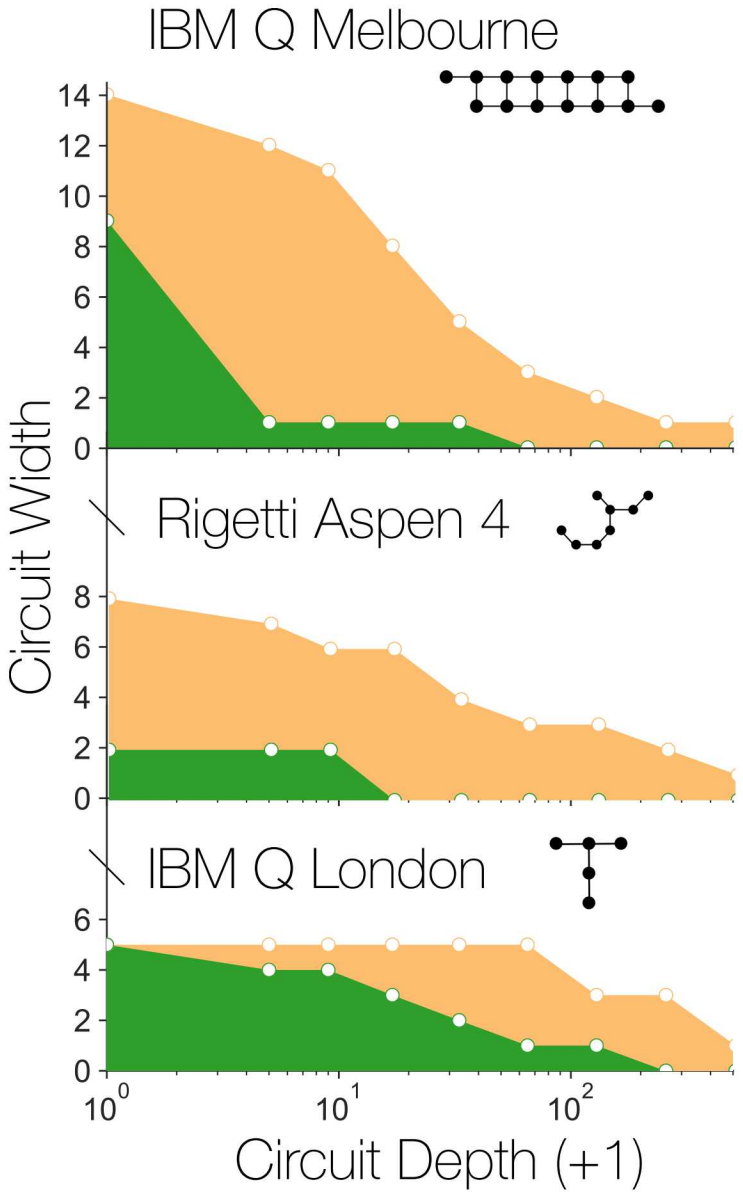
We're going to present:

- Methods for probing a device's capability set.

These methods scale to 1000s of qubits!

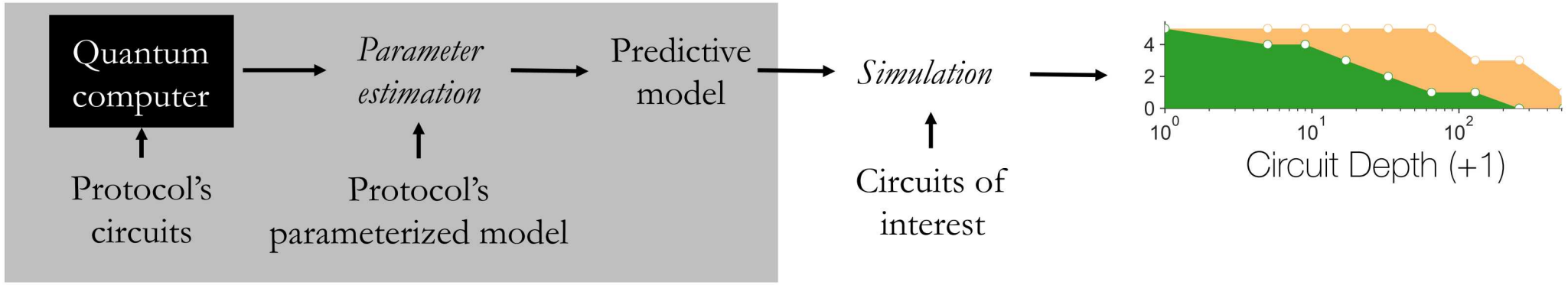
- Experiments on current hardware from IBM and Rigetti.

The results reveal that current devices display structured errors that cannot be adequately quantified by randomized benchmarks.

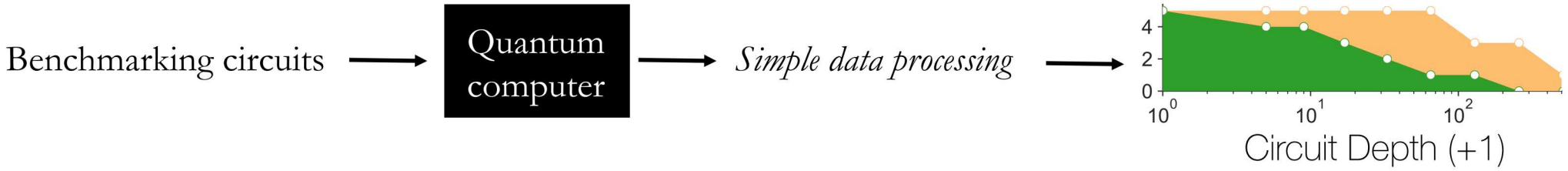


Mapping out a quantum computer's capability set

- We'd like to estimate a quantum computer's capability set.
- There's two distinct schema for doing this:
  - Build a predictive, mathematical model for the device (using device-specific physical modelling or a characterization protocol such as Gate set tomography[1], Cycle Benchmarking[2], Pauli noise estimation[3], or some flavor of RB[4]).



  - **Benchmarks:** run a set of test circuits. ***This is the route we take.***



[1]Sandia's QPL, Blume-Kohout *et al* , Nat. Commun. 8 144586 (2017), [2]Erhard *et al*, arXiv:1902.08543 (2019), [4]Emerson et al, J. Opt. B 7 S347 (2005), etc.

What properties do we want of a test circuit class?

- We expect the width/depth projection to be a good starting point.

1. The test circuits should have independently variable width and depth.

- We want to be able to test any near-term hardware using these circuits (up to 1000s of qubits).

2. The circuit class should contain very shallow circuits at each width.

- We don't trust that hardware performance depends only a circuit's width and depth.

3. The test circuit class should contain a wide range of circuit "types".

4. The circuits should be sensitive to all important types of error.

- We need to be able to quantify how well any circuit from the class ran.

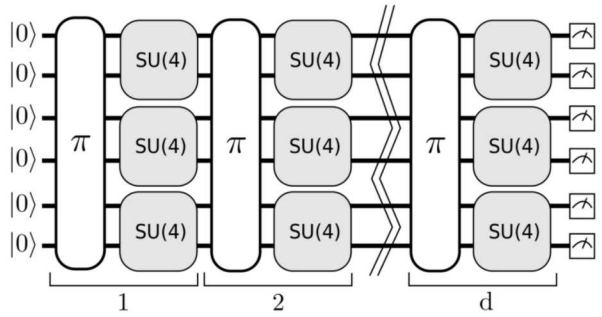
5. Every circuit should have an efficiently estimable success metric.

Scalable benchmarking circuits

1. Independently variable width *w* and depth *d*.
2. Some very shallow circuits at each width.
3. A wide range of circuit types.
4. Sensitive to all important types of error.
5. Efficiently estimable success metric.

*We're **not** say that current benchmarking methods aren't useful!*

No current benchmarking circuits satisfy all these criteria! So we need to make our own.

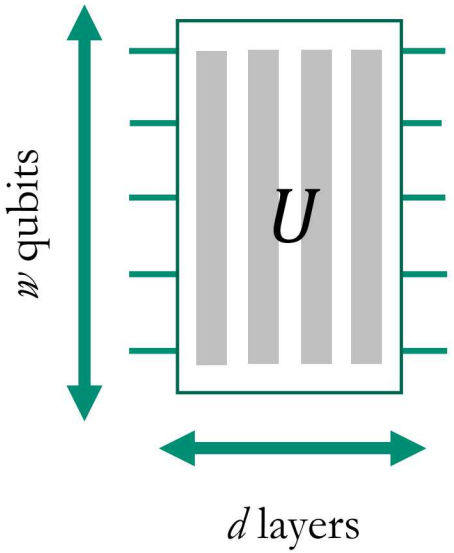


- Quantum volume[1] circuits? Fail (1-3) and (5).
- Cross-entropy benchmarking[2] circuits? Fail (3) and (5).
- Circuits from a current Randomized Benchmarking[3] method? Fail (2) and (3).
- Benchmarking with exemplar algorithm circuits[4]? Depends, but can fail all 5 criteria.

[1]Cross *et al.*, PRA 100, 032328 (2019). [2]Boixo *et al*, Nat. Phys. 14 595 (2018). [3]Emerson *et al*, JOB 7 S347 (2005). [4]Linke *et al*, PNAS 114 2205 (2017).

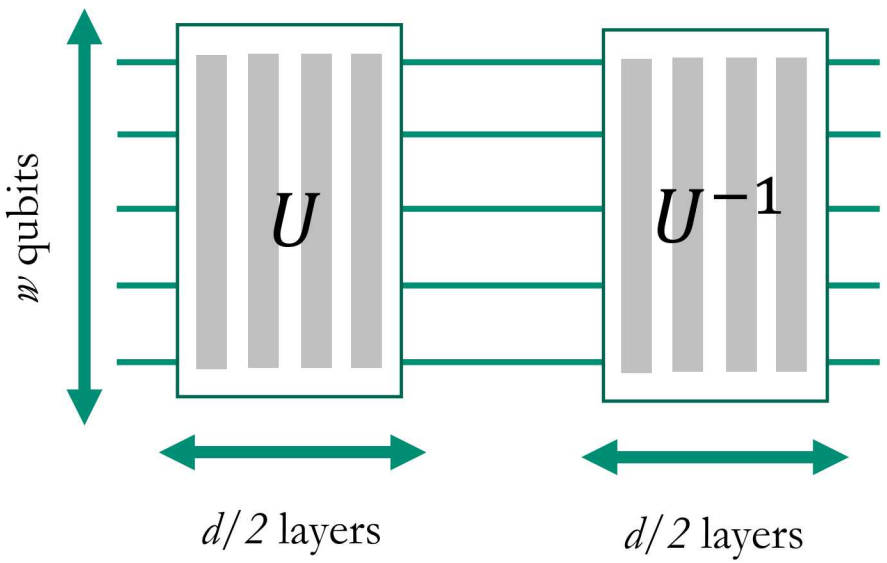
Scalable benchmarking circuits

1. Independently variable width w and depth d.
2. Some very shallow circuits at each width.
3. A wide range of circuit types.
4. Sensitive to all important types of error.
5. ~~Efficiently estimable success metric.~~



Scalable benchmarking circuits

1. Independently variable width w and depth d.
2. Some very shallow circuits at each width.
3. A wide range of circuit types.
4. ~~Sensitive to all important types of error.~~
5. Efficiently estimable success metric.



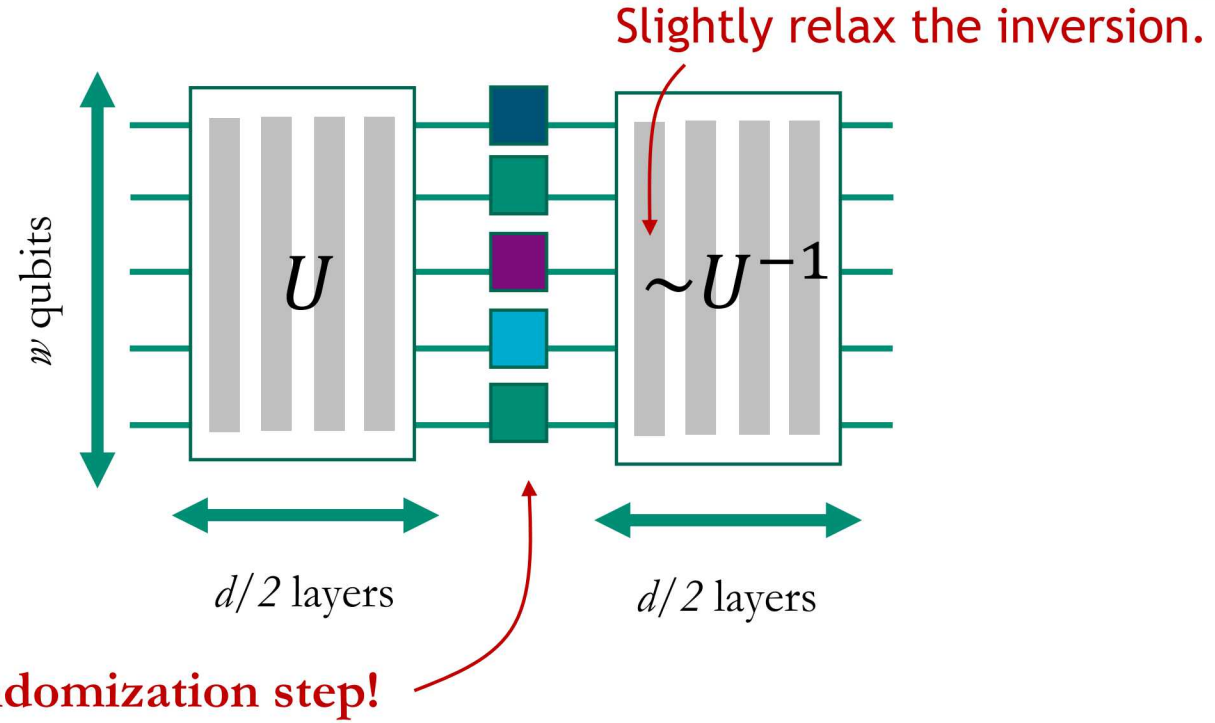
Insensitive to errors whereby $G * G^{-1} = I$ (matching over/under rotations).

The idea of motion-reversal isn't new! (see, e.g., the Loschmidt echo[1] and Emerson's original RB paper[2])

[1]Loschmidt 1876. [2]Emerson *et al*, J. Opt. B 7 S347 (2005).

Scalable benchmarking circuits: *Mirror circuits*

1. Independently variable width w and depth d.
2. Some very shallow circuits at each width.
3. A wide range of circuit types.
4. Sensitive to all important types of error.
5. Efficiently estimable success metric.



We still have an efficiently estimable success metric for arbitrary width circuits under certain conditions.
Sufficient (but not necessary) conditions:

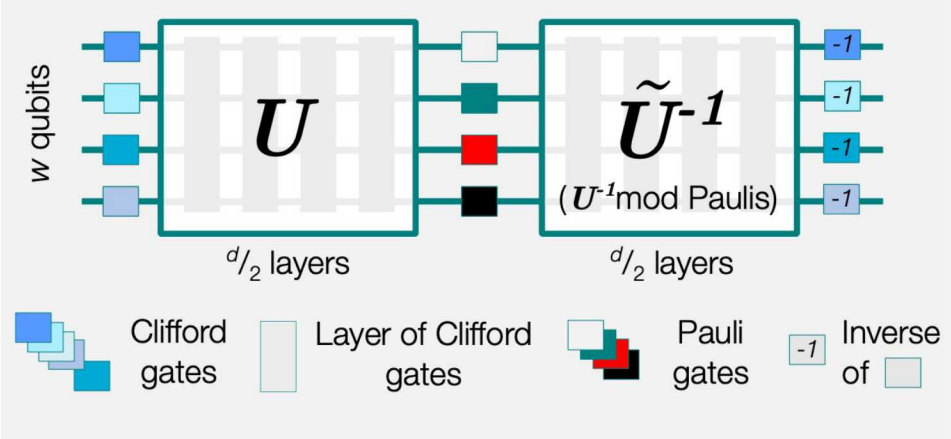
1. *All gates are Clifford operators.*
2. *The randomization gates are Pauli operators.*

Then the error-free output is a random, efficiently estimable[1] bit-string.

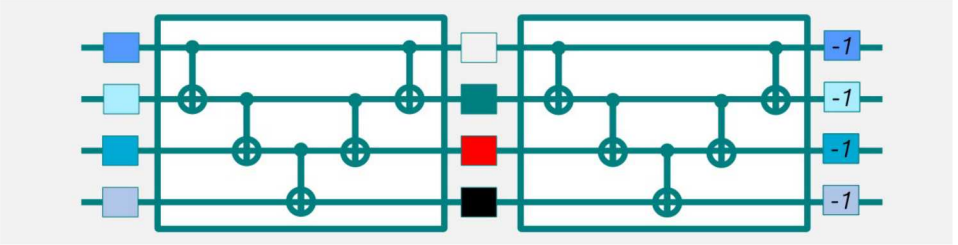
[1]Aaronson and Gottesman (2004).

Scalable benchmarking circuits: *Mirror circuits*

Mirror circuits are *very* general. What types of mirror circuits might make good benchmarks?

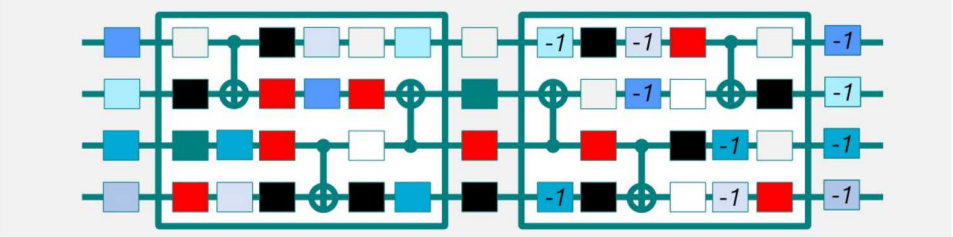


Algorithm-like circuits

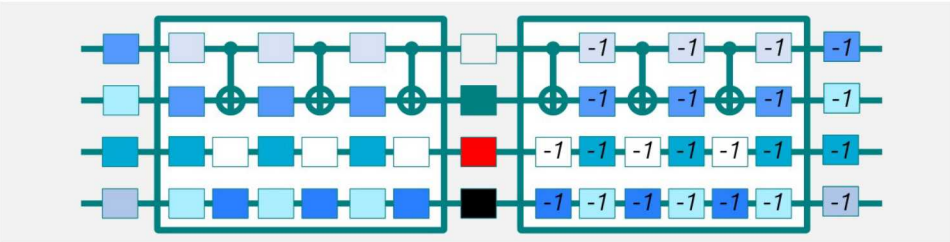


The focus in this talk

Randomized, unstructured circuits

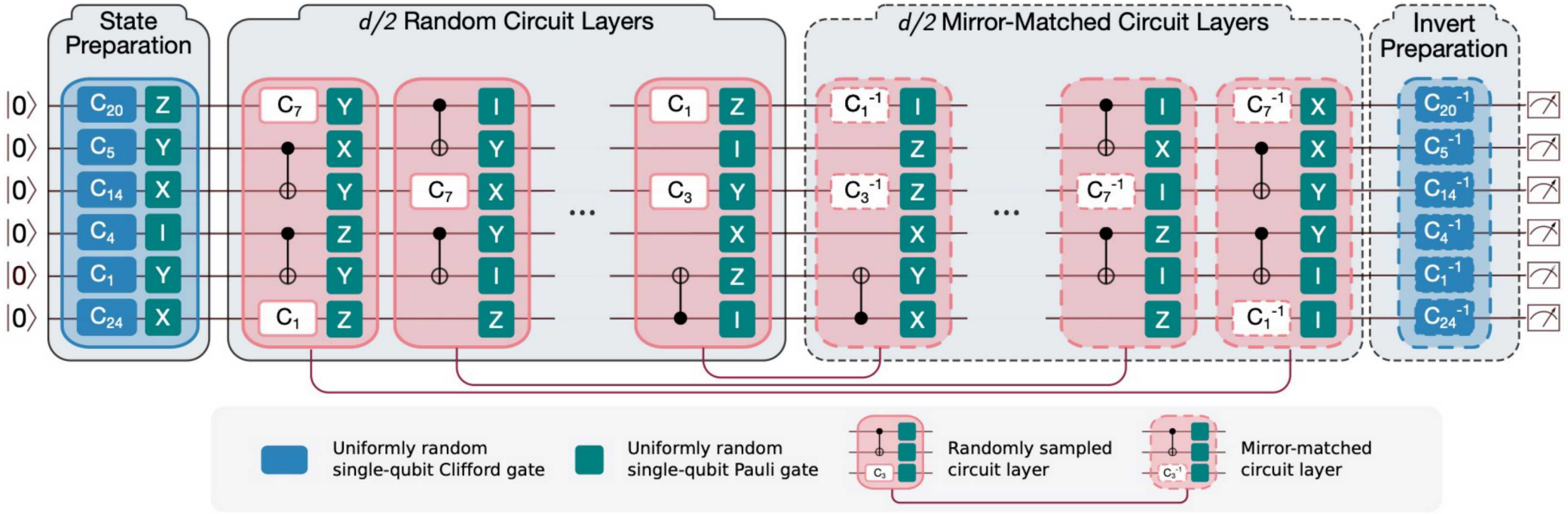


Repetitive error-amplifying circuits



A slide for the Randomized Benchmarking super-fans

The *randomized* mirror circuits that we use:



The dual-layers (pink boxes) of the "compute" circuit consist of:

- A layer of gates, from any set of Clifford group generators, sampled according to a user-specified distribution that is a fast scrambler.[1]
- A layer of uniformly random Pauli gates.

The "uncompute" circuit is the reverse computation but with *independent* random Paulis.

[1]Proctor *et al,* PRL (2019)

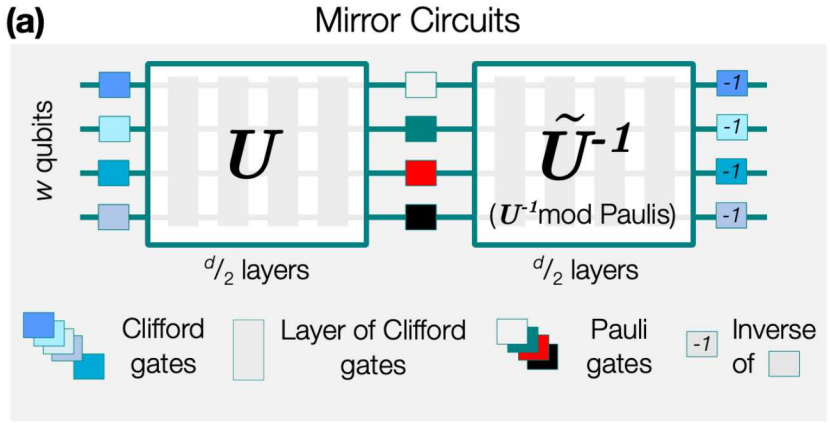
Mapping capability regions with randomized mirror circuits

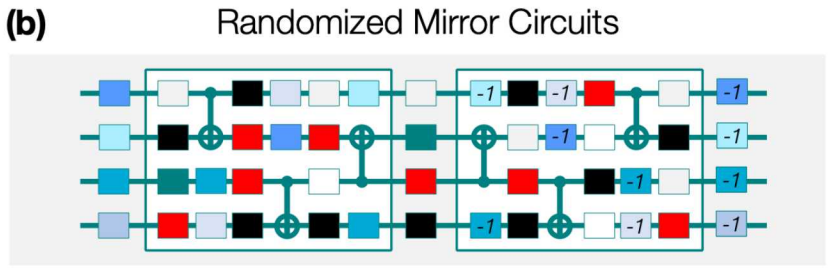
- We benchmarked 12 publicly accessible processors from IBM and Rigetti Computing.
- We ran varied width and depth randomized mirror circuits, with approximately exponentially spaced widths and depths.
- We benchmarked multiple set of qubits (device regions) for each width, testing a number of regions that scales linearly with device size (sub-linear scaling is also easily obtained). We'll show data for the best performing region.
- At each (width, depth) pair and for each device region of that width, we sampled 40 randomized mirror circuits[1], and we ran each one ~1000 times.

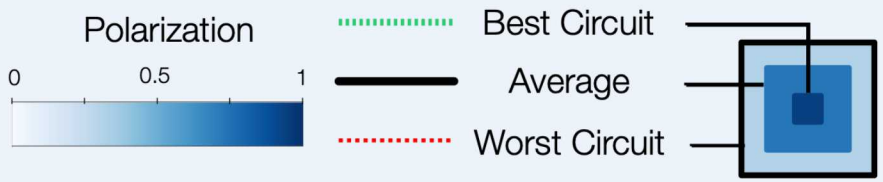
[1]The layer sampling consisted of uniformly selecting 1 two-qubit gate from the natively available two-qubit gates, applying it with 50% probability, and then applying uniformly random and independent one-qubit Clifford gates on all qubits that don't yet have a gate acting on them.

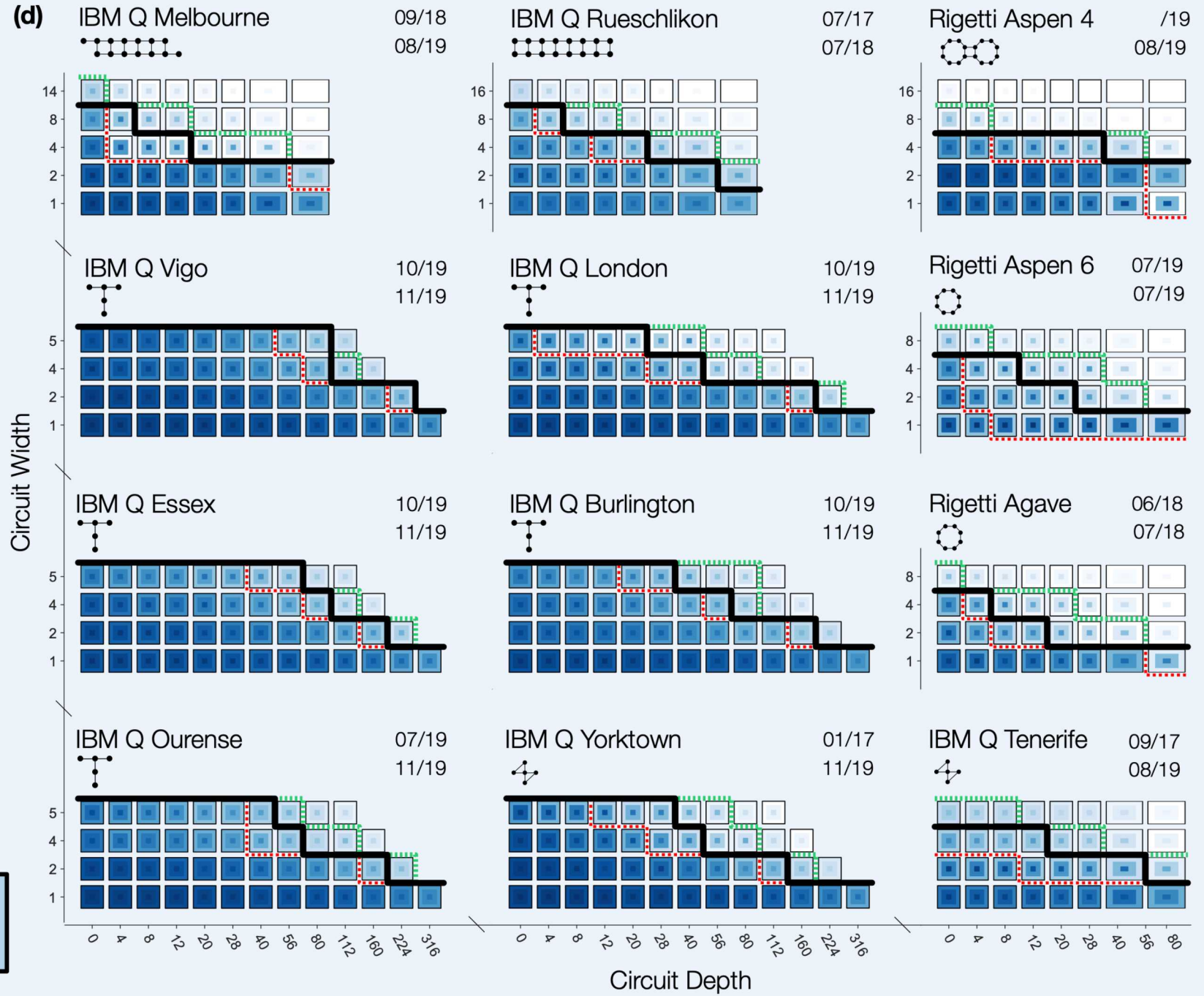
Experimental results

We summarize performance via "polarization" = $(P - 1/2^w) / (1 - 1/2^w)$ where P is the probability that the correct bit-string is output by circuit, and w is it's width.



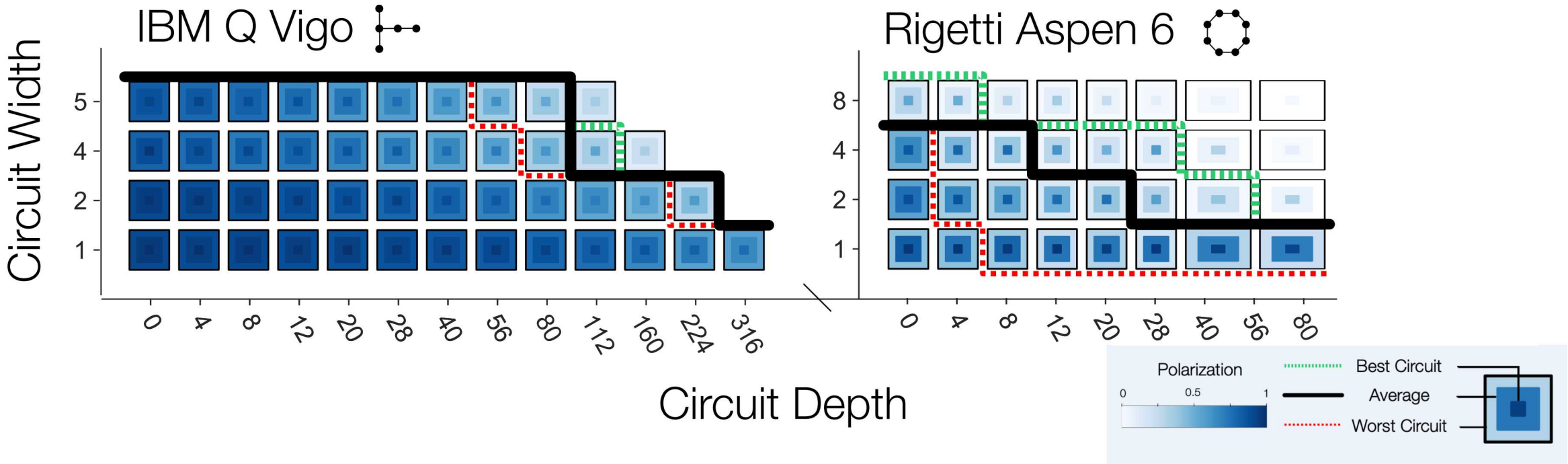






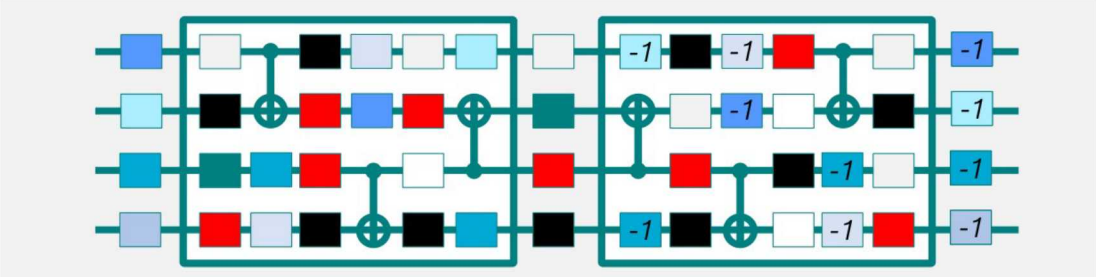
Mean performance on randomized circuit is not always predictive

- Some device's (e.g., Aspen 6) show massive variance in the performance of random circuits of the same depth & width. Other devices don't (e.g., IBM Q Vigo).
- So mean performance on randomized circuits at each {width, depth} is only sometimes a good predictor of hardware performance on a *randomly sampled* circuit.
- This is practically relevant as RB, cross-entropy benchmarking and quantum volume all summarize performance with a metric related to a mean over random circuits.

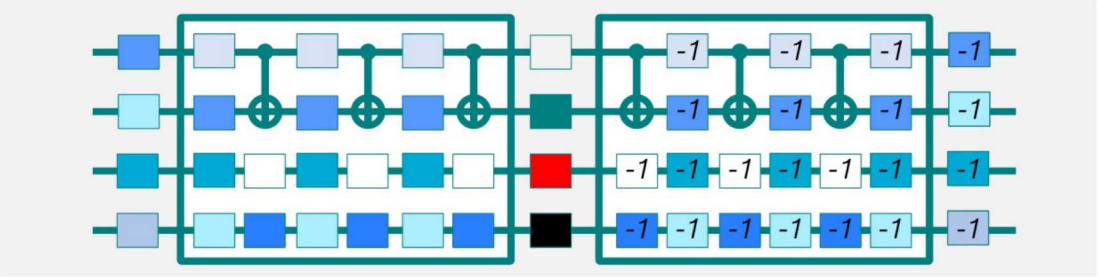


Variation in performance over circuits implies ***structured*** errors.

- Variation in the performance of equal {width, depth} circuits must be caused by deviations from simple i.i.d. depolarization. There is ***structure*** in the errors.
- Structured errors can be catastrophically amplified by circuits containing the "correct" structures.
- The simplest structure: different gates have different error rates. This is ubiquitous, mundane, and easily modelled.
- There are many more complex types of structure:
  - Unitary/calibration errors.
  - Biased stochastic noise.
  - Crosstalk.

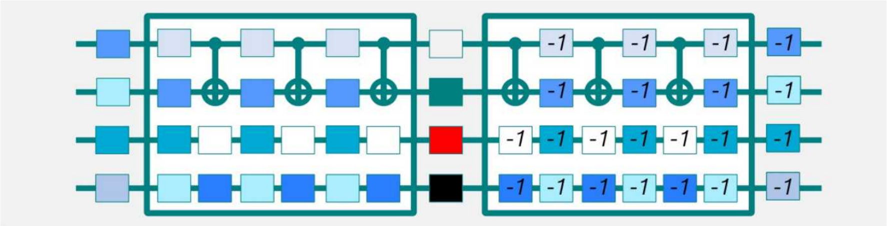


- Random circuits only contain structure by chance, so they're an inefficient probe for structured errors.
- Circuits with long-range order can amplifying errors.



Comparing performance on ordered and disordered circuits

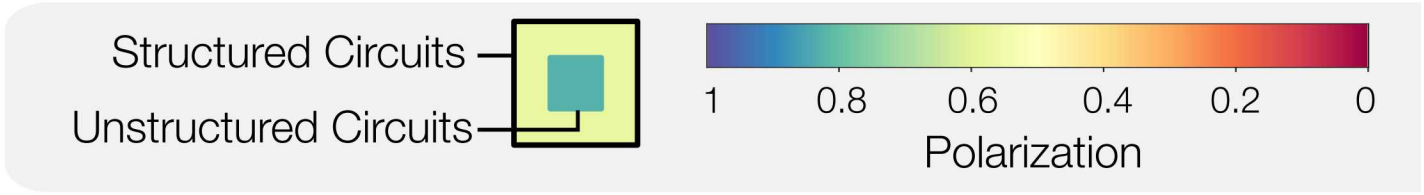
- To test the impact of structured errors, we ran similar ordered and disordered mirror circuits (of various widths and depths).

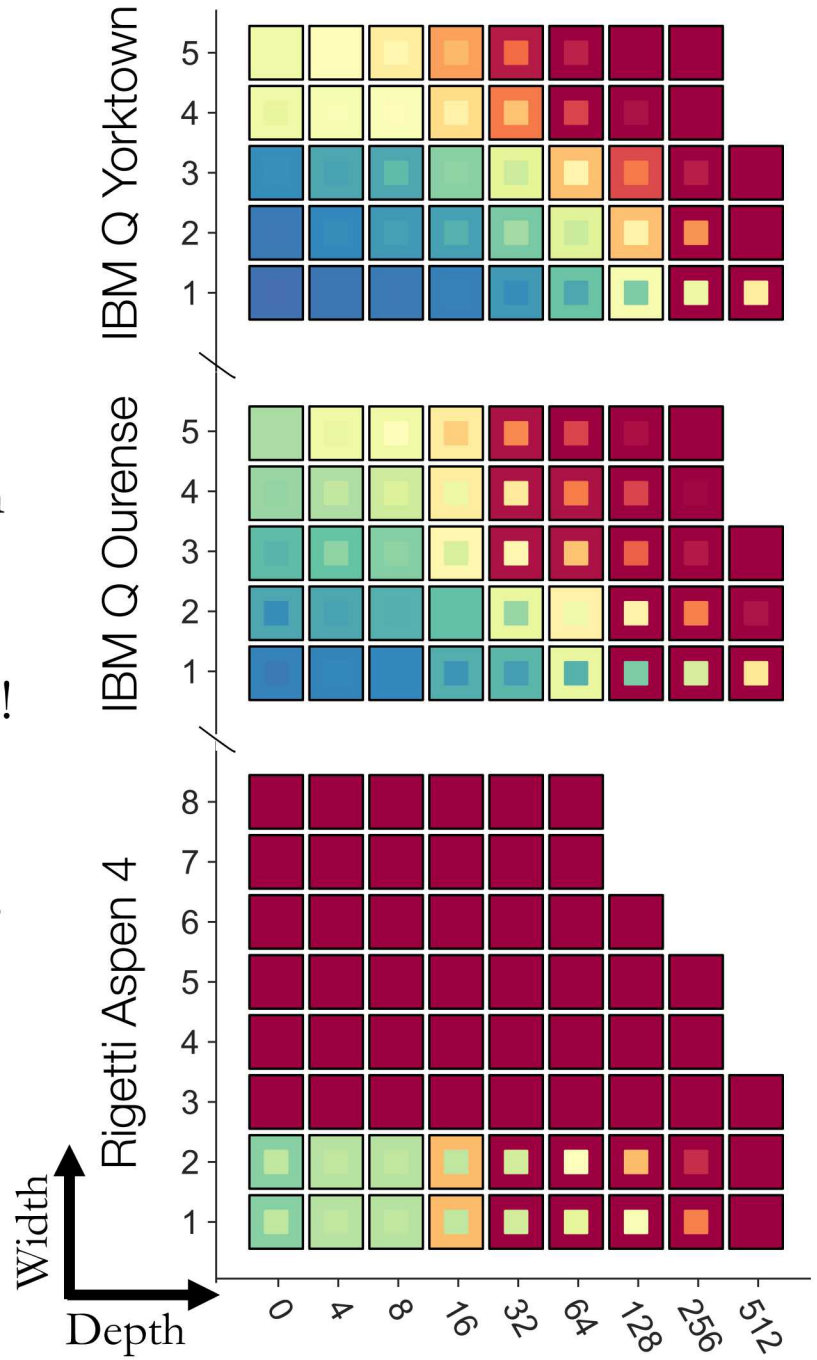


- The ordered circuits consisted of repeating a short unit cell.
- The ordered circuits have a two-qubit gate density fixed to ~1/8, and the disordered circuits have this density on average.
- We benchmarked one device region for each width, using the best region according to the device's calibration data.
- At each (width, depth) pair and for each device region of that width, we sampled 40 disordered and 40 ordered circuit, and we ran each one ~1000 times.

Comparing performance on ordered and disordered circuits

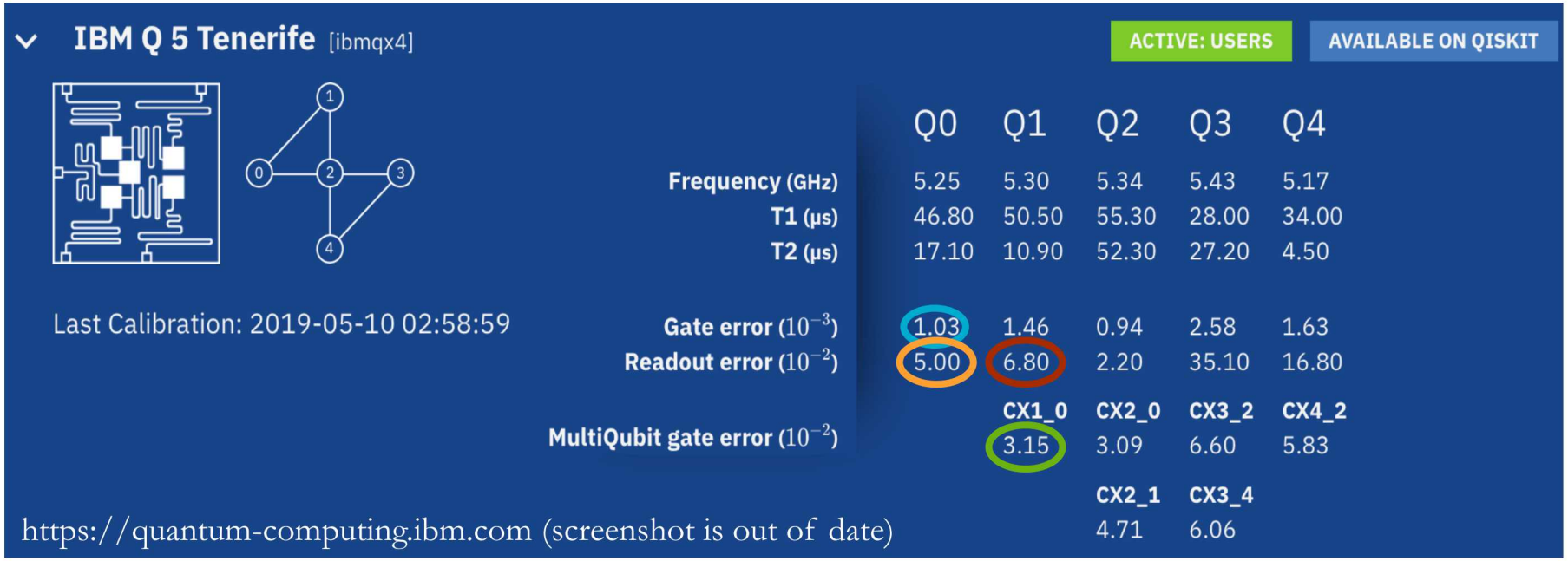
- Experimental results on 3 representative devices.
- Plot shows *worst-case* performance of structured/ordered circuits and unstructured/disordered circuits as a function of width and depth.
- Performance is typically *much* worse for structured circuits!
- As most algorithmic circuits are highly structured, this demonstrates that randomized benchmarks are unlikely to be predictive of NISQ applications.





Do a device's "error rates" predict it's capability set?

- Quantum computer "quality" is typically summarized by error rates (typically obtained via RB).



- Doesn't adding up the error rates of all the operations you use predict circuit performance?

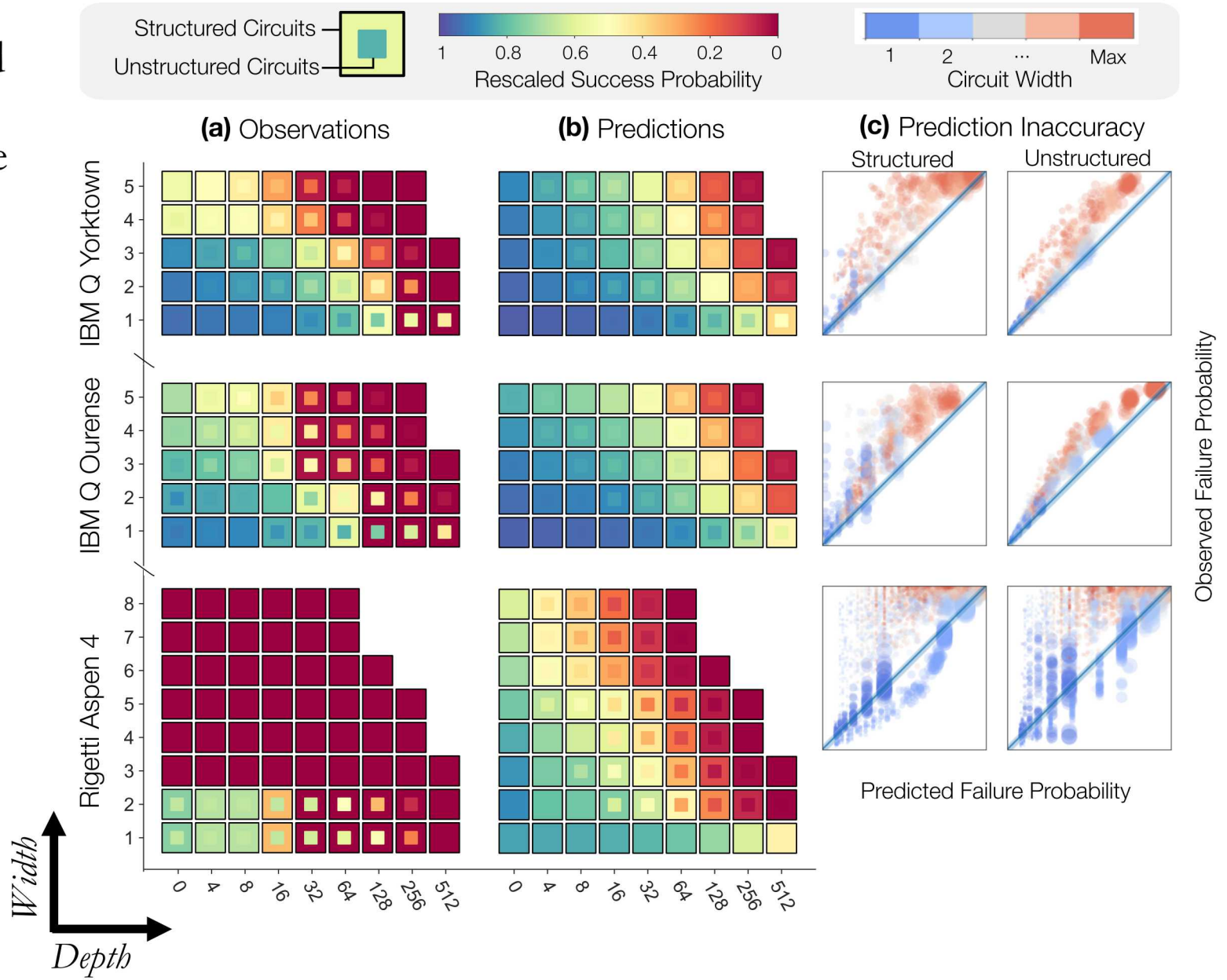
Example: the error rate of a circuit containing a 1-qubit gate on Q0 and a two-qubit gate on Q0-Q1 is (roughly[1]) predicted to be: 1 - (1-0.001)(1-0.03)*(1-0.05)*(1-0.07) ~ 15%.*

- This model *cannot* predict a difference between ordered and disordered circuits! But how bad is the inaccuracy?

[1]We actually use a depolarization model which differs from this calculation by some circuit-width-dependent factors.

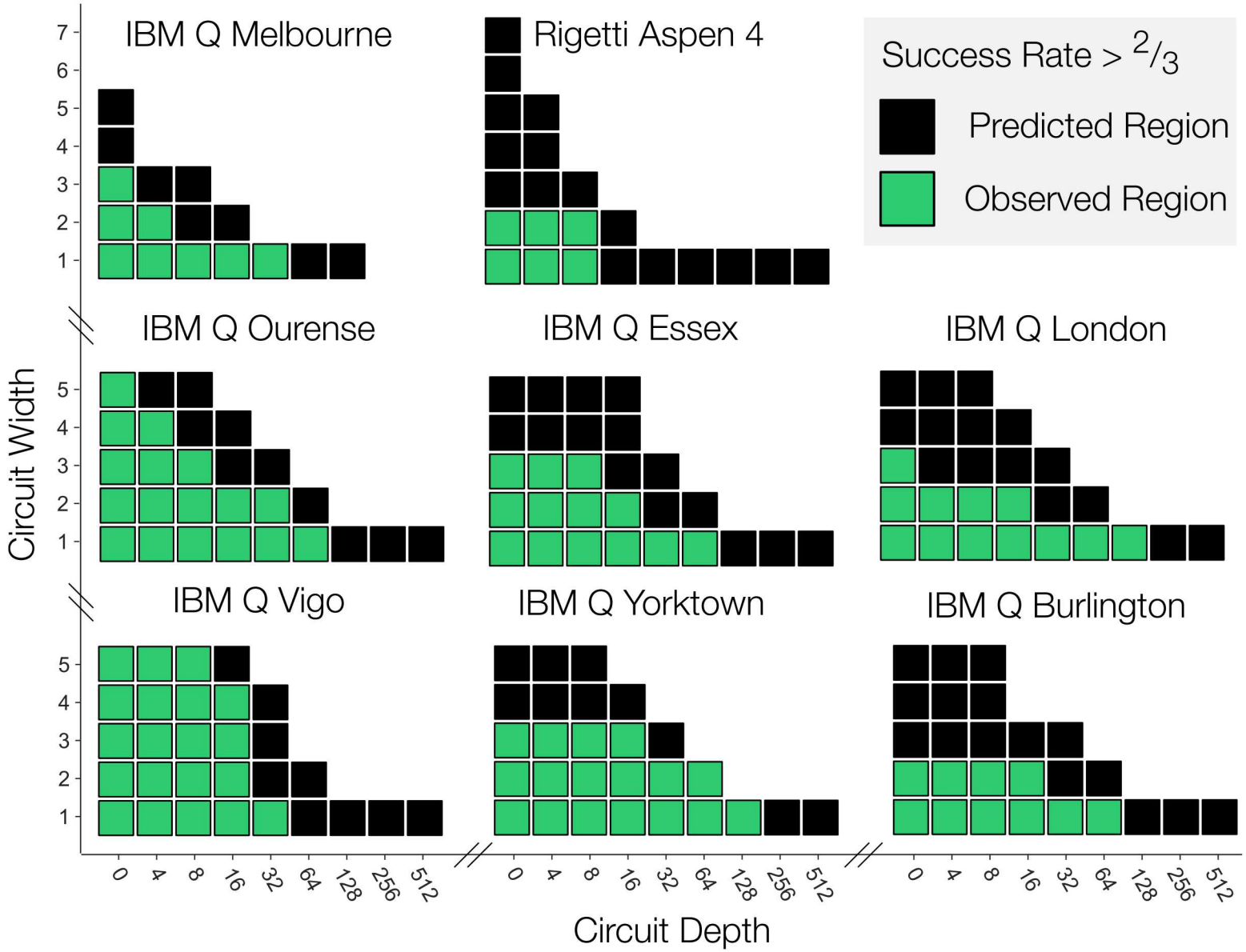
Do a device's "error rates" predict it's capability set?

- For each device, we calculated the predicted success probability of every circuit we ran.
- The predictions use the error rates provided by the manufacture.
- The predicted worst-case perform is drastically over-optimistic – a symptom of structured errors.
- Even best-fit error rates cannot predict the data (data not shown).



Predicted versus actual capabilities

- We can summarize the observed and actual perform via *worst-case* capability regions.
- This shows the region in which *every* circuit with a two-qubit gate density < ¼ should succeed in at least 2 in 3 runs.
- Observations are an upper-bound on this region.
- But the observations are still much more pessimistic than the predictions from gate & readout error rates.

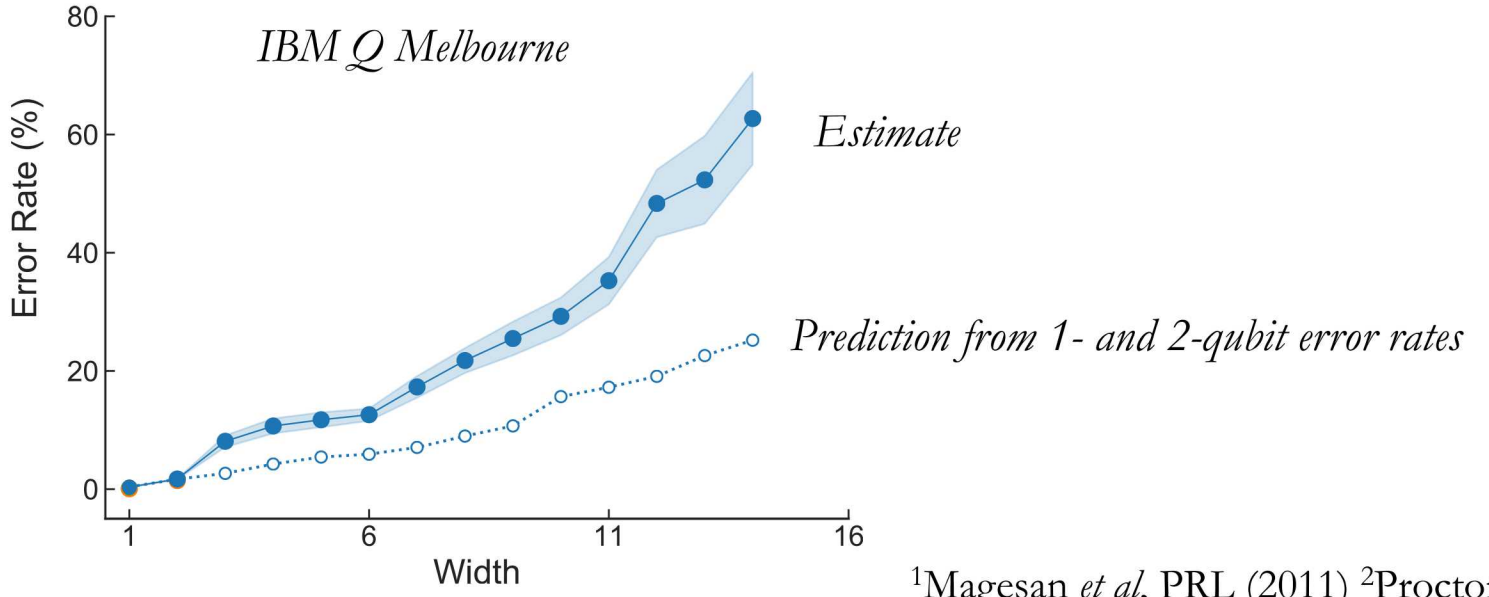


Bonus! Error rates from randomized mirror circuits

- Randomized mirror circuits can also be used as a traditional Randomized Benchmarking protocol, i.e., to estimate gate error rates.

- We can perform simple process on data, and fit it to an exponential to estimate average gate (more accurately, layer) error rates (which is, roughly speaking the gate's "infidelity").

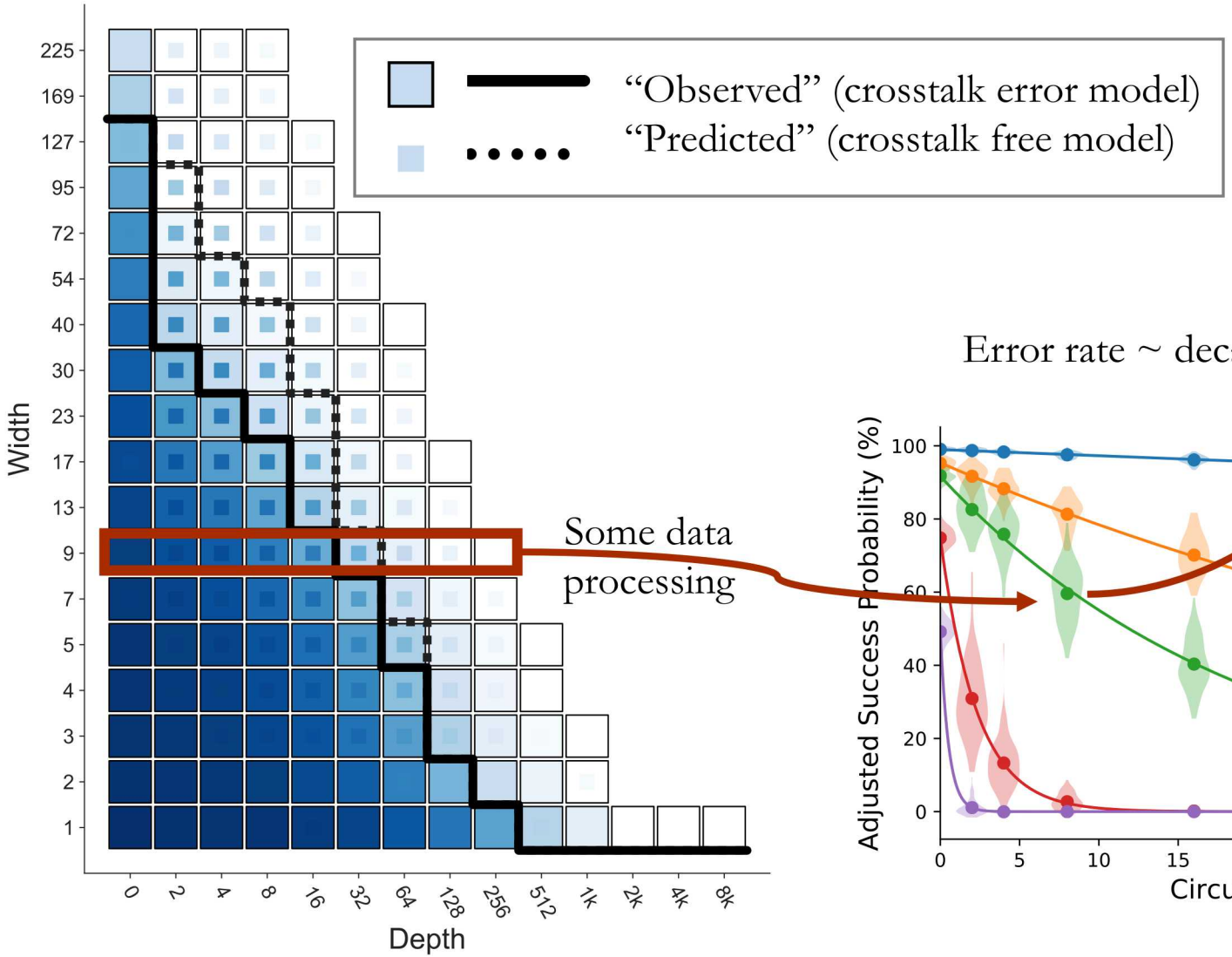
- This is an improvement on traditional RB[1] (and it's more scalable variant *direct* RB[2]) as it scales to 100s or 1000s of qubits!



[1]Magesan *et al*, PRL (2011) [2]Proctor *et al*, PRL (2019)

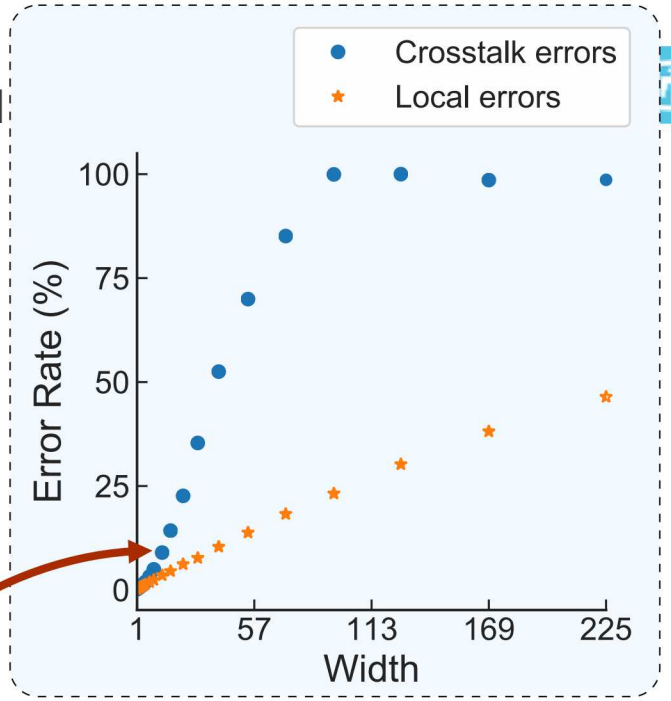
Bonus! Error rates from randomized mirror circu

Simulation of benchmarking of a 225Q lattice with long-range 2Q-gate crosstalk.

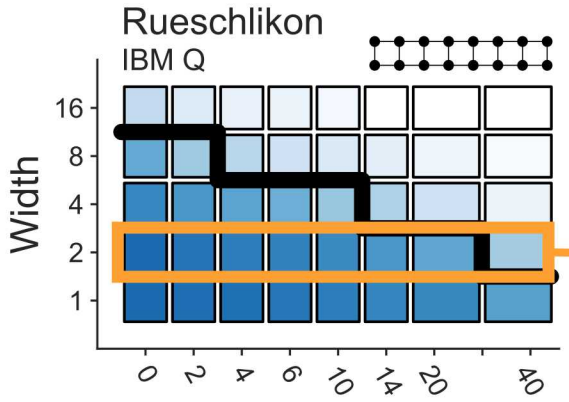


Some data processing

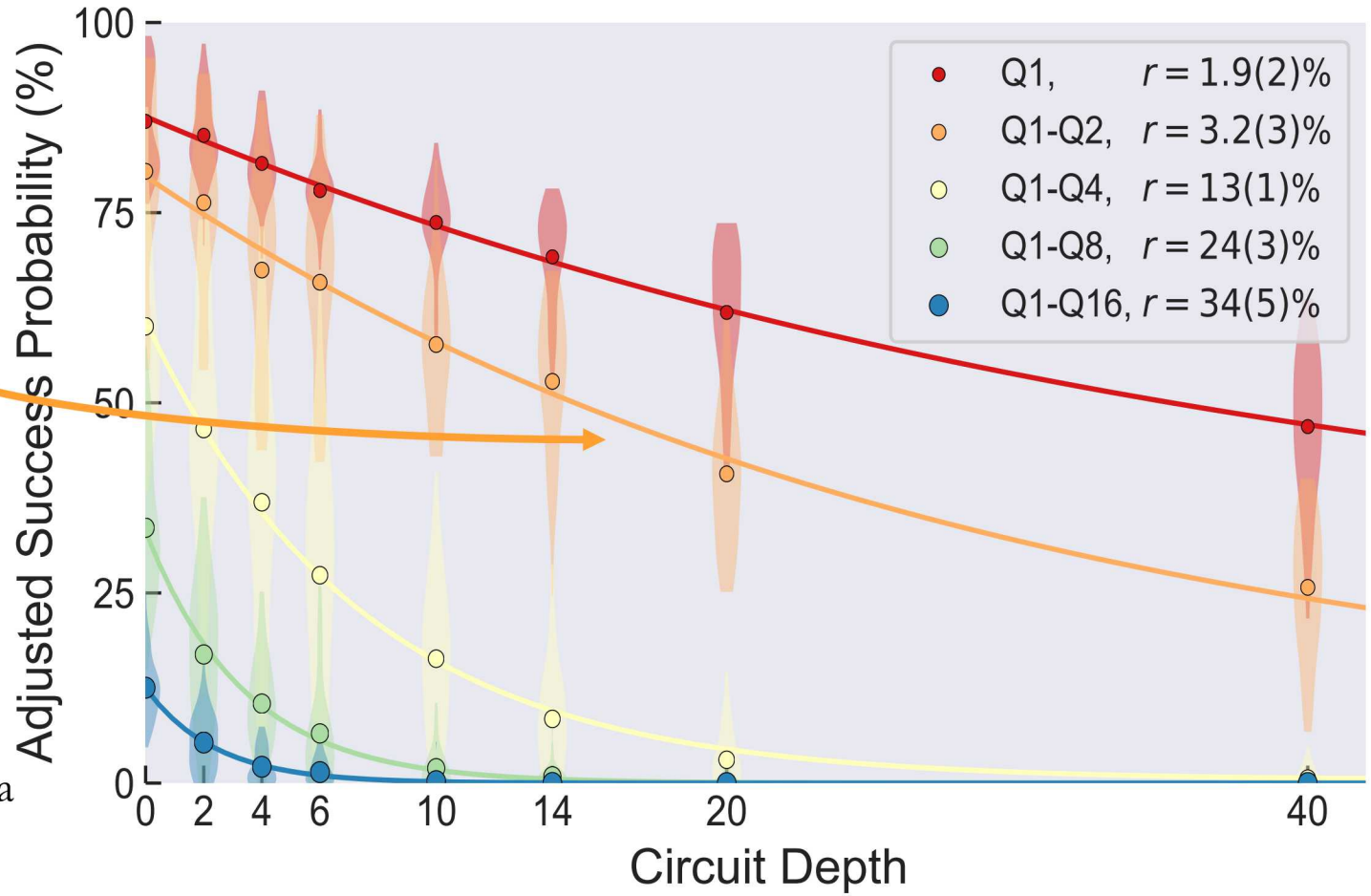
Error rate ~ decay rate



The RB analysis on IBM Q Rueschlikon



(almost)



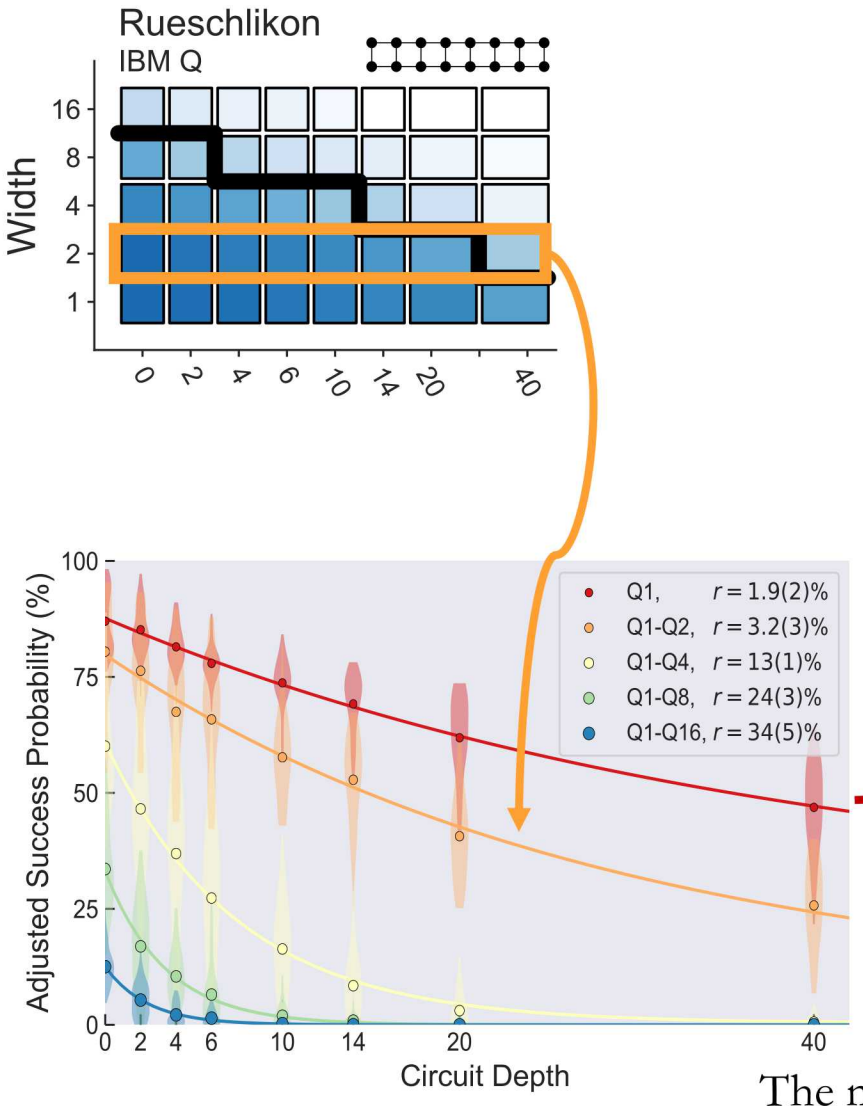
- The *adjusted success probability* is:

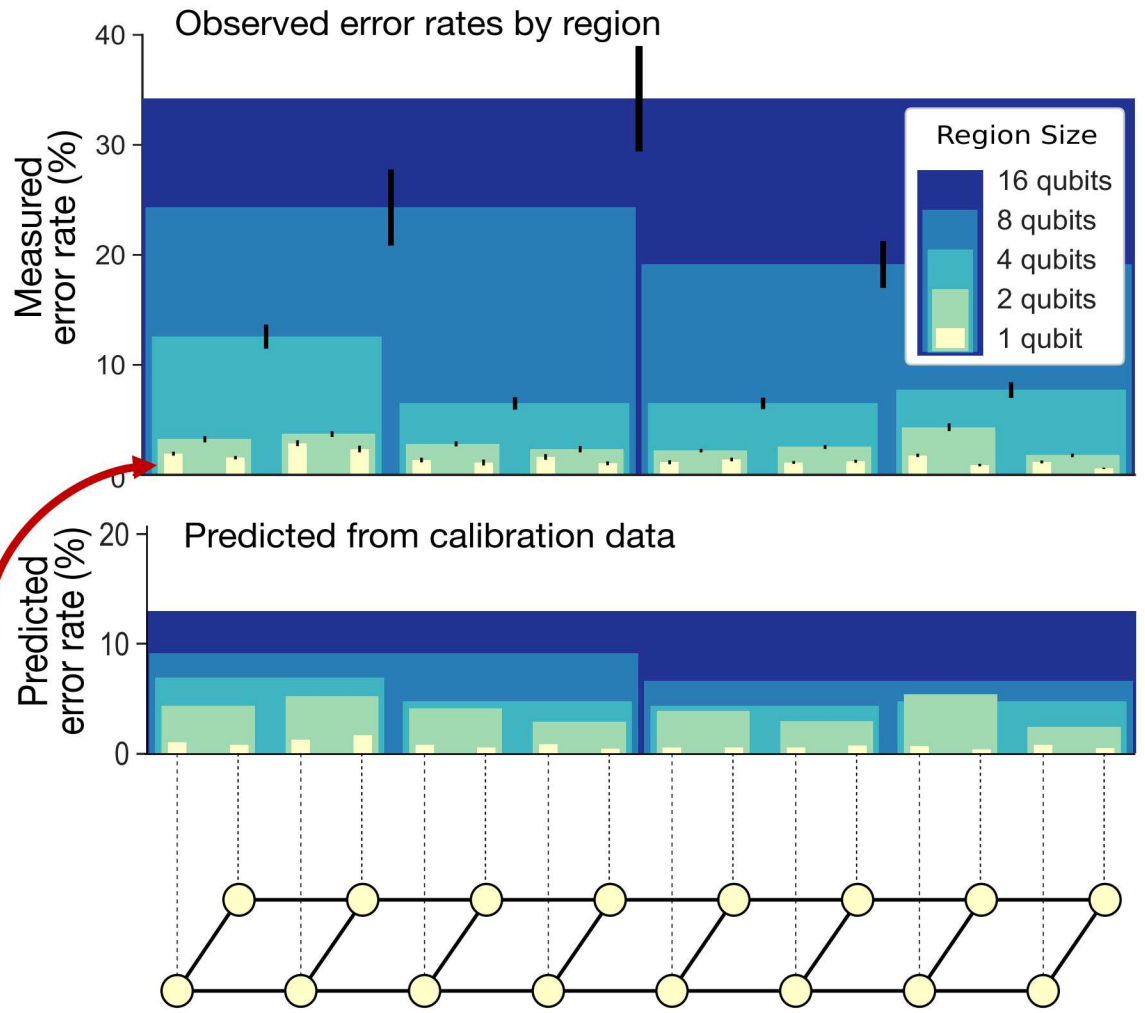
$$S = \sum_{k=0}^{w} \left(-\frac{1}{2}\right)^{k} h_k$$

where h_k is the rate that the output bit-string is a Hamming distance of k from the "target" bit-string.

- We've proven that S will decay exponentially under broad conditions, at a rate that's ~ the average error rate of a layer.
- We observe clear exponential decays on real hardware.

The RB analysis on IBM Q Rueschlikon





The measured error rate by region is **significantly** higher than the prediction **because of crosstalk. Their difference is a quantification of emergent error.**

Summary

- Quantum computer users want to know what programs a device can successfully run: it's *capability set*.
- We've introduced a *very* general class of circuits ("mirror circuits") for scalable benchmarking of a processor's capability set.
- Experiments on hardware from IBM and Rigetti show that:
  - Performance on randomized/disordered circuits is not predictive of performance on non-random circuits.
  - Per-gate error rates do not model the structure in errors, and so their predictions are egregiously inaccurate.
- Randomized or *ad hoc* benchmarks will not be sufficient to understand whether a near-term quantum computer can run a given application.
- Mirror circuits are a great foundation on which to build fast, scalable and well-motivated benchmarks.
- *Bonus!* Randomized mirror circuits facilitate truly scalable RB.

Thanks!

Many thanks to IBM Quantum Experience and Rigetti Computing for access to their quantum computing platforms, and technical help.

Get Your Capabilities Checked Now!

If you'd like to run mirror circuit benchmarks to test your hardware's capabilities:

- Look for postings to arXiv soon!
- Get in contact with me (tjproct@sandia.gov) or anyone at Sandia's QPL.
- Code for running experiments like these is in pyGSTi (*www.pygsti.info*).

Interested in a postdoc at Sandia's Quantum Performance Lab? Get in touch!

This work was 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. 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 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.