# Validating quantum computers using randomized model circuits

Andrew W. Cross, Lev S. Bishop, Sarah Sheldon, Paul D. Nation, and Jay M. Gambetta

Phys. Rev. A **100**, 032328 — Published 20 September 2019

DOI: 10.1103/PhysRevA.100.032328

# Validating quantum computers using randomized model circuits

Andrew W. Cross,* Lev S. Bishop,† Sarah Sheldon, Paul D. Nation, and Jay M. Gambetta
*IBM T. J. Watson Research Center, Yorktown Heights, NY 10598*

We introduce a single-number metric, *quantum volume*, that can be measured using a concrete protocol on near-term quantum computers of modest size ($n \lesssim 50$), and measure it on several state-of-the-art transmon devices, finding values as high as 16. The quantum volume is linked to system error rates, and is empirically reduced by uncontrolled interactions within the system. It quantifies the largest random circuit of equal width and depth that the computer successfully implements. Quantum computing systems with high-fidelity operations, high connectivity, large calibrated gate sets, and circuit rewriting toolchains are expected to have higher quantum volumes. The quantum volume is a pragmatic way to measure and compare progress toward improved system-wide gate error rates for near-term quantum computation and error-correction experiments.

Recent quantum computing efforts have moved beyond controlling a few qubits, and are now focused on controlling systems with several tens of qubits [1–3]. In these noisy intermediate-scale quantum (NISQ) systems [4], performance of isolated gates may not predict the behavior of the system. Methods such as randomized benchmarking [5], state and process tomography [6], and gateset tomography [7] are valued for measuring the performance of operations on a few qubits, yet they fail to account for errors arising from interactions with spectator qubits [8, 9]. Given a system such as this, whose individual gate operations have been independently calibrated and verified, how do we measure the degree to which the system performs as a general purpose quantum computer? We address this question by introducing a single-number metric, the *quantum volume*, together with a concrete protocol for measuring it on near-term systems. Similar to how LINPACK [10] and improved benchmarks [11, 12], are used for comparing diverse classical computers, this metric is not tailored to any particular system, requiring only the ability to implement a universal set of quantum gates. With the concept of this metric being discussed elsewhere [13, 14], our focus here is on measuring this metric in near-term quantum devices.

The quantum volume protocol we present is strongly linked to gate error rates, and is influenced by underlying qubit connectivity and gate parallelism. It can thus be improved by moving toward the limit in which large numbers of well-controlled, highly coherent, connected, and generically programmable qubits are manipulated within a state-of-the-art circuit rewriting toolchain. High-fidelity state preparation and readout are also necessary. In this work, we evaluate the quantum volume of current IBM Q devices [1], and corroborate the results with simulations of the same circuits under a depolarizing error model. While we focus on transmon devices, the protocol can be implemented with any universal programmable quantum computing device.

* awcross@us.ibm.com
† lsbishop@us.ibm.com

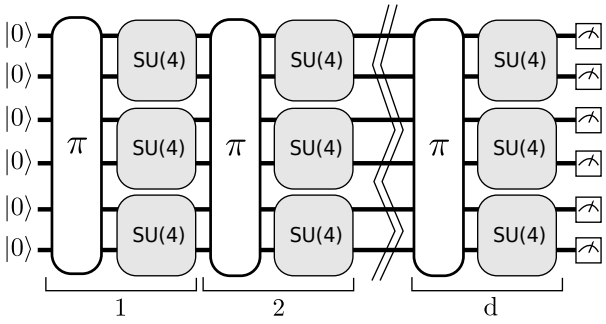


FIG. 1. **Model circuit.** A model circuit consists of $d$ layers of random permutations of the qubit labels, followed by random two-qubit gates. When the circuit width $m$ is odd, one of the qubits is idle in each layer. A final permutation can be applied to the labels of the measurement outcomes.

The quantum volume is based on the performance of random circuits with a fixed but generic form. It is well-known that quantum algorithms can be expressed as polynomial-sized quantum circuits built from two-qubit unitary gates [15]. Quantum algorithms are generally not random circuits. However, random circuits model generic state preparations, and are used as the basis of proposals for demonstrating quantum advantage [16]. In addition, circuits with a similar form appear in near-term algorithms like quantum adiabatic optimization algorithms [17] and variational quantum eigensolvers [18].

A *model circuit*, shown in Fig. 1, with depth $d$ and width $m$, is a sequence $U = U^{(d)} \dots U^{(2)} U^{(1)}$ of $d$ layers

$$U^{(t)} = U^{(t)}_{\pi_t(m'-1),\pi_t(m')} \otimes \cdots \otimes U^{(t)}_{\pi_t(1),\pi_t(2)}, \qquad (1)$$

each labeled by times $t = 1, \dots, d$ and acting on $m' = 2\lfloor n/2 \rfloor$ qubits. Each layer is specified by choosing a uniformly random permutation $\pi_t \in S_m$ of the $m$ qubit indices and sampling each $U^{(t)}_{a,b}$, acting on qubits $a$ and $b$, from the Haar measure on SU(4).

To define when a model circuit $U$ has been successfully implemented in practice, we use the *heavy output* generation problem [19]. The ideal output distribution is

$$p_U(x) = |\langle x|U|0\rangle|^2 \qquad (2)$$

where $x \in \{0,1\}^m$ is an observable bit-string. Consider the set of output probabilities given by the range of $p_U(x)$ sorted in ascending order $p_0 \leq p_1 \cdots \leq p_{2^m-1}$. The median of the set of probabilities is $p_{med} = (p_{2^{(m-1)}} + p_{2^{(m-1)}-1})/2$, and the heavy outputs are

$$H_U = \left\{x \in \{0,1\}^m \text{ such that } p_U(x) > p_{med}\right\}. \quad (3)$$

The heavy output generation problem is to produce a set of output strings such that more than two-thirds are heavy. The expected heavy output probability for an ideal device is asymptotically $(1 + \ln 2)/2 \sim 0.85$ [19], while it falls to $\sim 0.5$ if the device is completely depolarized.

To evaluate heavy output generation, we implement model circuits using the gate set provided by the target system. For example, the model circuit may need to be rewritten, not only to use the system's gate set, but also to respect the set of available interactions, which may require additional operations such as SWAP gates. The average gate fidelity [20] between $m$-qubit unitaries $U$ and $U'$ is

$$F_{\text{avg}}(U,U') = \frac{\left|\text{Tr}(U^\dagger U')\right|^2/2^m + 1}{2^m + 1}. \quad (4)$$

Given a model circuit $U$, a circuit-to-circuit transpiler finds an implementation $U'$ for the target system such that $1 - F_{\text{avg}}(U,U') \leq \epsilon \ll 1$. In many cases, the approximation error $\epsilon$ is limited by the selected classical precision within the transpiler (eg. for arithmetic to compute new gate angle parameters), but may be further increased if the hardware requires SU(4) to be approximated with a discrete set of available gates.

The transpiler is free to use all available tricks and hardware resources to implement $U'$ (e.g., taking great computational effort in finding an optimized $U'$, using extra qubits for gate teleportation or temporary storage, etc.). It may optimize over qubit placements by choosing the best region of the device. If it is practical to calibrate a very large gate set, and it happens to include an accurate implementation of $U$, the transpiler is free to use it. None of these approaches is expected to provide an asymptotic advantage, but may significantly improve practical performance. We do require that the transpiler make an honest attempt to implement $U$, and not merely choose a relatively simple operation far from $U$ that nevertheless produces the heavy outputs for $U$. The compilation routine for computing the quantum volume of IBM Q devices is described in Appendix A, and an approximation scheme given in Appendix B.

The observed distribution for an implementation $U'$ of model circuit $U$ is $q_U(x)$, and the probability of sampling a heavy output is

$$h_U = \sum_{x \in H_U} q_U(x). \quad (5)$$

To determine if a given output is heavy, we compute $H_U$ directly from $U$ using a method that scales exponentially[1] with $m$. The probability of observing a heavy output by implementing a randomly selected depth $d$ model circuit is $h_d = \int_U h_U dU$. Ideally, we would estimate this quantity using all of the qubits of a large device, but NISQ devices have appreciable error rates, so we begin with small model circuits and progress to larger ones. We are interested in the achievable model circuit depth $d(m)$ for a given model circuit width $m \in [n]$. We define the achievable depth $d(m)$ to be the largest $d$ such that we are confident $h_d > 2/3$ (See Appendix C for further discussion of confidence intervals). In other words,

$$h_1, h_2, \ldots, h_{d(m)} > 2/3 \text{ and } h_{d(m)+1} \leq 2/3. \quad (6)$$

Algorithm 1 provides pseudocode for testing when each $h_d > 2/3$.

**Algorithm 1** Check heavy output generation

```
function isHeavy(m, d; n_c ≥ 100, n_s)
    n_h ← 0
    for n_c repetitions do
        U ← random model circuit, width m, depth d
        H_U ← heavy set of U from classical simulation
        U' ← compiled U for available hardware
        for n_s repetitions do
            x ← outcome of executing U'
            if x ∈ H_U then n_h ← n_h + 1
    return (n_h − 2√(n_h(n_s − n_h/n_c)))/(n_c n_s) > 2/3
```

We desire a metric that is a single real number, as this enables straightforward comparison. Data $\{d(m)\}$ can be gathered by sweeping over values of $m$ and $d$. We are free to choose any function of this data $\{d(m)\}$ to capture how well a device performs. The quantum volume treats the width and depth of a model circuit with equal importance and measures the largest square-shaped (i.e., $m = d$) model circuit a quantum computer can implement successfully on average [13, 14]. We define the quantum volume $V_Q$ as

$$\log_2 V_Q = \underset{m}{\text{argmax}} \min(m, d(m)) \quad (7)$$

and take this definition going forward.

This definition differs from [13, 14] and loosely coincides with the complexity of classically simulating the model circuits. There are different ways to classically simulate the model quantum circuits. A straightforward wave-vector propagation approach requires exponential space and time $\sim 2^m$. A 'Feynman' algorithm uses linear

[1] For error rates as low at $10^{-4}$, we anticipate that model circuits $U$ that can be successfully implemented will involve few enough qubits and/or low enough depth to compute $H_U$ classically. For lower error rates than this, the quantum volume can be superseded by new volume metrics or modified so classical simulations are not necessary.

space $\sim dm$ but exponential time $\sim 4^{dm}$. It is possible to trade off time and space complexity in a smooth way [19]. Clever partitioning of circuits can achieve good parallelism and efficient use of distributed memory resources for particular supercomputer architectures [21–27]. Particular efforts for circuit partitioning and parallelism have been expended for circuits defined on a 2-dimensional square grid of qubits, where the state-of-the-art is $d = 40$ for a $9 \times 9$ grid [22].

One view of these methods is that they use heuristics to approach optimal variable elimination ordering for a tensor network calculation on the graph corresponding to the circuit. The time complexity scales exponentially with the treewidth of the circuit graph [28]. The treewidth is upper-bounded by $m$, and while there are specific circuits of depth $d = 4$ with expander graph structure for which the treewidth is $\Omega(m)$, heuristic estimation of the treewidth for some classes of random circuits [24, 25] indicates that the treewidth grows roughly as $d$. Therefore, we heuristically bound the treewidth of the model circuits as $\min(d, m)$, and since the simulation complexity grows exponentially with the treewidth, we define the quantum volume as $V_Q = 2^{\min(d,m)}$.

We have run quantum volume circuits on four IBM Q devices: 5-qubit *Tenerife* [29], 16-qubit *Melbourne* [30], 20-qubit *Tokyo*, and 20-qubit *System One*. We generate 200 circuits for $d = m$ with $m$ in $2, 3, 4$ to determine $V_Q$. The experimental results and comparison to simulated data for *Tokyo* and *System One* are given in Figs. 2 and 3 respectively, whereas a summary of results across all devices is in Table I. We note that the noisy simulation substantially over-estimates the performance, highlighting the value of system-level metrics such as quantum volume. In order to set a high confidence level that the experimental measurements of $h_d$ surpass the threshold, we repeat the experiments for $m = 2$ on *Tenerife* and $m = 3$ on *Tokyo* with 5000 circuits. This larger number of circuits has a strict threshold of $\hat{h}_d > 0.68$ for a 97.5% one-sided confidence interval (see Appendix C). From Table I we see that $\log_2 V_Q = 3$ for *Tokyo*, $\log_2 V_Q = 2$ for *Tenerife*, and $\log_2 V_Q < 2$ for *Melbourne*. Additional details about the devices used here are given in Appendix D.

We also compare circuits run on *Tokyo* with optimized compiling schemes. Table II presents $\hat{h}_d$ for $m = d = 4$ found with circuits optimized both by the KAK decomposition [31, 32] described in Appendix A and the approximate SU(4) decomposition described in Appendix B. The approximate decomposition takes the CX error rate as a parameter to determine acceptable approximation errors when synthesizing a circuit for an element of SU(4). We apply this decomposition assuming CX error rates of 0.01, 0.03, and 0.05 and compare the results. We find modest increases in $\hat{h}_d$ that correspond to the reduction in the total number of CX gates in the compiled circuits: the standard Qiskit Terra transpiler [**?** ] produces circuits with 28 CX gates on average, and we measure $\hat{h}_d = 0.614(0.003)$; KAK reduces the average number of CX gates to 21 and produces $\hat{h}_d = 0.632(0.005)$. The approximate SU(4) circuits introduce further gains with the best result of $\hat{h}_d = 0.649(0.005)$ achieved using circuits with a 1% CX error approximation.

Finally, we present the outcomes of the quantum volume circuits measured on *System One*. This device has the lowest gate error rates of all the devices measured, with single qubit gate errors a factor of four smaller and two qubit gate errors nearly half than those measured on *Tokyo*. These reduced error rates suggest *System One* should have the best performance of all the devices measured, and in fact we find the highest heavy output probabilities for $m = d > 3$ on this device as is evident in Table I. For the case $m = d = 4$ the results lie just below the threshold of $\hat{h}_d = 2/3$, and optimizing the circuits with both the KAK decomposition and the approximate SU(4) with 1% CX error yields $\hat{h}_d = 0.699(0.001)$.

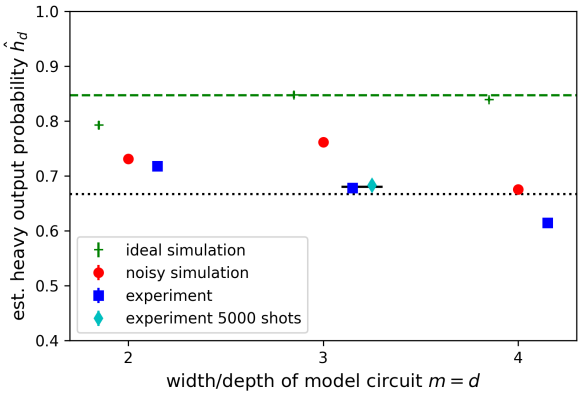


FIG. 2. Experimental data for square (width = depth) quantum volume circuits using the IBM Q 20-qubit device, *Tokyo*. The ideal simulation results are green plus signs. The noisy simulations, using a depolarizing noise model with average error rates from the qubits used on the device, are red circles. The experiments using 200 circuits are blue squares. The dotted line is the threshold of 2/3 for heavy output generation, and the dashed (green) line is the asymptotic ideal heavy output probability of $\frac{1+\ln 2}{2}$ [19], which the ideal simulations quickly approach. In order to set a high confidence level that $h_d$ surpasses the threshold, the point at $m = d = 3$ was repeated with 5000 circuits (cyan diamond). This number of shots corresponds to a stricter threshold of 0.68 indicated by the solid line at the experimental points for $m = 3$.

To understand how the quantum volume scales in a system with limited connectivity, as gate error probabilities decrease, we consider model circuits of width $m$ on a square grid of $m$ qubits. The $m$ qubits are arranged into the largest possible square, and extra qubits are added first to a new right column and then to a new bottom row. We approximate the achievable model circuit depth $\tilde{d}(m)$ by assuming independent stochastic errors, so that the computation fails with high probability when the model

| Circuit | Tenerife | Melbourne | Tokyo | System One |
|---|---|---|---|---|
| $m = d = 2$ | 0.685 (0.001)* | 0.638 (0.006) | 0.718 (0.006) | 0.711 (0.006) |
| $m = d = 3$ | 0.651 (0.006) | 0.641 (0.009) | 0.682 (0.002)* | 0.729 (0.007) |
| $m = d = 4$ | 0.516 (0.002) | 0.523 (0.002) | 0.614 (0.003) | 0.664 (0.004) |
| $m = d = 4\dagger$ | | | 0.649 (0.005) | 0.699 (0.001)** |
| $m = d = 5$ | | | | 0.601 (0.004) |

TABLE I. Experimentally estimated heavy output probabilities for four IBM Q devices: 5-qubit *Tenerife*, 16-qubit *Melbourne*, 20-qubit *Tokyo*, and 20-qubit *System One*, for circuits of equal width $m$ and depth $d$. For each $m$, 200 circuits were run on every device. The experiments ($*/**$) were repeated with (5000/1000) circuits to ensure a 97.5% one-sided confidence interval as descriped in Appendix C. $m = d = 4\dagger$ experiments used circuits optimized with the KAK and approximate SU(4) decompositions assuming a 1% CX error rate.

| | Standard | KAK | 1% approx. | 3% approx. | 5% approx. |
|---|---|---|---|---|---|
| Average # CX Gates | 28.1 | 21.0 | 17.7 | 16.1 | 15.1 |
| Noisy Simulation | 0.676 (0.003) | 0.687 (0.004) | 0.693 (0.004) | 0.692 (0.004) | 0.685 (0.005) |
| Experiment | 0.614 (0.003) | 0.632 (0.005) | 0.649 (0.005) | 0.647 (0.005) | 0.646 (0.005) |

TABLE II. Gate counts and heavy output probabilities for $m = d = 4$ circuits optimized with the KAK decomposition and the approximate SU(4) decompositons assuming CX error rates of 1%, 3%, and 5%. For each width/depth, 200 circuits were run on *Tokyo* and simulated using average error rates from *Tokyo*.

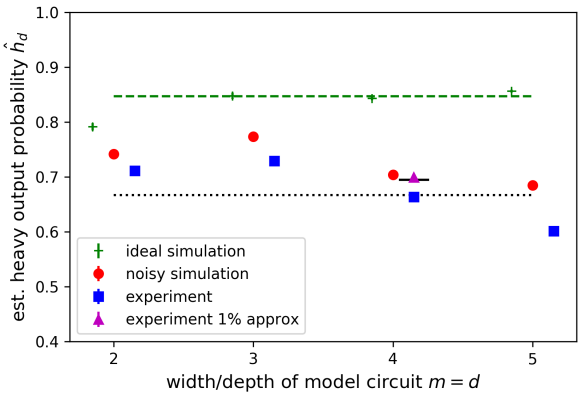


FIG. 3. Experimental data for square (width = depth) quantum volume circuits using the IBM Q System One 20-qubit device. As in Figure 2, the ideal simulation results are green plus signs, the noisy simulations are red circles, and the experiments are blue squares. Again, the dotted line is the threshold of 2/3 for heavy output generation, and the dashed (green) line is the asymptotic ideal heavy output probability. The additional point at $m = d = 4$ (magenta triangle) is not only repetition with more circuits but experimental results using optimized circuits with the KAK approximation, assuming 1% error gates. The experiments with optimized circuits were run with 1000 circuits. The threshold for this number of circuits is 0.695 and is indicated by the solid line at $m = 4$.

circuit volume (width times depth) satisfies

$$m\tilde{d}(m) \approx \frac{1}{\epsilon_{\text{eff}}(m)}. \tag{8}$$

We substitute an estimate of the mean effective error probability $\epsilon_{\text{eff}}(m)$ per two-qubit gate into this expression. This estimate $\epsilon_{\text{eff}}(m) = (a\sqrt{m} + b)\epsilon$ is proportional to the two-qubit gate error probability $\epsilon$, with a prefactor that is linear in $\sqrt{m}$. This factor fits the mean number of SWAPs necessary to bring a pair of qubits next to each other, apply the gate, and then return them to their original positions. It is twice the average shortest path length (minus one). We do a similar calculation for a loop of $m$ qubits and find $\epsilon_{\text{eff,loop}}(m) = (a'm+b')\epsilon$, which grows linearly with the number of qubits[2]. At a given error rate $\epsilon$, we can use these expressions to estimate the quantum volume, permitting $m$ to grow as needed.

| $\log_2 V_Q$ | All-to-All | Square Grid | Loop |
|---|---|---|---|
| 4 | 0.03 | 0.028 | 0.028 |
| 6 | 0.015 | 0.011 | 0.011 |
| 8 | 0.008 | 0.005 | 0.0047 |
| 12 | 0.0032 | 0.0015 | 0.0014 |

TABLE III. Estimates of the maximum permissible two-qubit error needed for quantum volume $V_Q$, with $\log_2 V_Q$ given in the leftmost column, for three coupling maps: all-to-all connectivity, square grid, and loop. The estimates are based on simulations using a depolarizing noise model with two-qubit error $\epsilon$ as given, single-qubit error equal to $\epsilon/10$, and perfect measurements.

To validate these estimates, we consider the influence of connectivity on quantum volume by simulating three coupling graphs for up to 12 qubits: all-to-all connectivity, square grid, and loop. We estimate the two-qubit

[2] For a square array, we find $a \approx 1.29$ and $b \approx -0.78$, and for a loop, we find $a' = 1/2$ and $b' \approx -0.45$.

| $\log_2 V_Q$ | 0% meas. error | 1% meas. error | 5% meas. error |
|---|---|---|---|
| 4 | 0.028 | 0.026 | 0.020 |
| 6 | 0.011 | 0.010 | 0.007 |
| 8 | 0.005 | 0.0045 | 0.0023 |
| 12 | 0.0015 | 0.00125 | 0.0002 |

TABLE IV. A comparison of the maximum permissible two-qubit error rate for $\log_2 V_Q$ of 4, 6, 8, and 12 for three values of the measurement error: 0%, 1%, and 5%. These simulations all use a square grid coupling map; the 0% measurement error column is identical to the square grid column of Table III.

gate error $\epsilon$ required for each coupling graph to obtain a $\log_2 V_Q$ of 4, 6, 8, and 12, assuming the single-qubit gate error is equal to $\epsilon/10$ (Table III). We run these simulations with no measurement error for all graphs, and for measurement errors of 0%, 1%, and 5% for the square grid (Table IV). The values for $\epsilon$ here correspond to 200 simulated circuits with a heavy output probability of $\hat{h}_d = 0.67 \pm 0.05$.

It is clear from Table III that all-to-all connectivity provides an advantage over the less-connected graph; $\log_2 V_Q$ of 12 is achievable with twice the two-qubit error rate (0.0032) of the square grid (0.0015) and the 12-qubit loop (0.0014). At the same time, there is little difference between the required two-qubit error rate for the square grid versus the loop graphs; the error rate for the loop is less than 7% lower than that of the square grid for the 12-qubit case. This relatively small difference is due to the small total number of qubits, since there is a significant asymptotic difference between loop and grid layouts. However, the difference may increase, even at small sizes, when using an optimal transpiler. All circuits for the simulations in Tables III and IV were compiled using the standard Qiskit Terra transpiler. Quantum volume estimates computed from Eq. 8 are consistent with these depolarizing noise simulations at error probabilities down to $\epsilon \approx 10^{-3}$, as shown in Fig. 4.

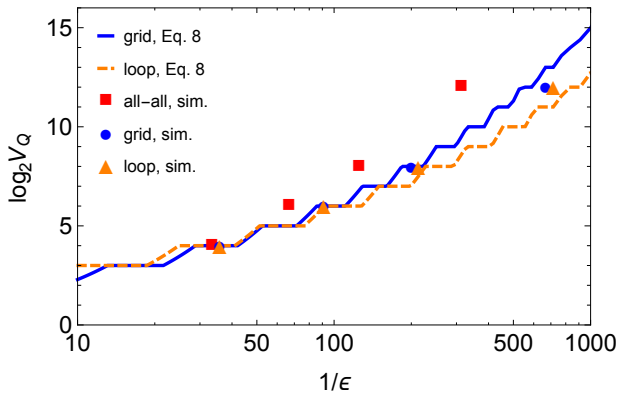


FIG. 4. The quantum volume increases as a function of inverse gate error $1/\epsilon$. This plot shows numerical simulation results from the top half of Table III together with estimates using the expression in Eq. 8 for grid and loop connectivities.

These simulations give an indication of how quantum volume measurements might look on different quantum computing architectures. Trapped ions, for instance, will benefit from having all-to-all connectivity. Typical trapped-ion systems have both two-qubit gate errors and measurement errors less than 0.01, which based on Table III should be sufficient to achieve $\log_2 V_Q = 6$ if not higher. Recently, trapped-ion experiments have demonstrated two-qubit gates with errors of 0.001 [34], indicating higher quantum volumes should be possible. However, multi-qubit experiments are susceptible to larger error rates than isolated two-qubit gates, due to correlated errors across many ions [35]. A measurement of quantum volume would give a reliable validation of multi-qubit trapped-ion systems. Similarly, we can infer that for superconducting devices, coupling maps with more connectivity should produce higher quantum volume, but only if additional coupling does not also introduce larger errors.

**Conclusion:** In this paper we expand on a previously presented metric, the quantum volume [13, 14], and show both a concrete specification and a method for benchmarking noisy intermediate-scale quantum devices. This metric takes into account all relevant hardware parameters. This includes the performance parameters (coherence, calibration errors, crosstalk, spectator errors, gate fidelity, measurement fidelity, initialization fidelity) as well as the design parameters such as connectivity and gate set. It also includes the software behind the circuit optimization. Additionally, the quantum volume is architecture-independent, and can be applied to any system that is capable of running quantum circuits. We implement this metric on several IBM Q devices, and find that we can successfully implement model circuits on up to $\log_2 V_Q = 4$ qubits, which corresponds to a quantum volume as high as $V_Q = 16$. We conjecture that systems with higher connectivity will have higher quantum volume given otherwise similar performance parameters.

From numerical simulations for a given connectivity, we find that there are two possible paths for increasing the quantum volume. Although all operations must improve to increase the quantum volume, the first path is to prioritize improving the gate fidelity above other operations, such as measurement and initialization. This sets the roadmap for device performance to focus on the errors that limit gate performance, such as coherence and calibration errors. The second path stems from the observation that, for these devices and this metric, circuit optimization is becoming important. We implemented various circuit optimization passes (far from optimal) and showed a measurable change in the experimental performance. In particular, we introduced an approximate method for NISQ devices, and used it to show experimental improvements.

We encourage the adoption of quantum volume as a primary performance metric, which we believe will allow the field to work together and focus efforts on the important factors to develop improved NISQ devices. To this

end, we have released a library for measuring quantum volume as an open-source component of Qiskit Ignis [33].

## ACKNOWLEDGMENTS

The authors acknowledge support from ARO under Contract No. W911NF-14-1-0124 and thank Sergey Bravyi, John A. Smolin, and Christopher J. Wood for informative discussions. We thank Antonio Córcoles, Abigail Cross, John Gunnels, David McKay, Travis Scholten, and Ted Yoder for valuable comments on the manuscript. We are grateful to the IBM Q team for their contributions to the systems and devices used in this work.

[1] "IBM Q Experience," https://quantumexperience.ng.bluemix.net/qx/experience, Last Accessed: 2018-11.
[2] N. Friis, O. Marty, C. Maier, C. Hempel, M. Holzäpfel, P. Jurcevic, M. Plenio, M. Huber, C. Roos, R. Blatt, and B. Lanyon, Phys. Rev. X **8**, 021012 (2018).
[3] C. Song, K. Xu, W. Liu, C. ping Yang, S.-B. Zheng, H. Deng, Q. Xie, K. Huang, Q. Guo, L. Zhang, P. Zhang, D. Xu, D. Zheng, X. Zhu, H. Wang, Y.-A. Chen, C.-Y. Lu, S. Han, and J.-W. Pan, Phys. Rev. Lett. **119**, 180511 (2017).
[4] J. Preskill, Quantum **2** (2018).
[5] E. Magesan, J. M. Gambetta, and J. Emerson, Physical Review A **85**, 042311 (2012).
[6] M. G. A. Paris and J. Řeháček, eds., *Quantum State Estimation*, Lecture Notes in Physics (Springer-Verlag, Berlin Heidelberg, 2004).
[7] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Physical Review A **87**, 062119 (2013); R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz, Nature Communications **8**, 14485 (2017).
[8] D. C. McKay, S. Sheldon, J. A. Smolin, J. M. Chow, and J. M. Gambetta, arXiv preprint **arXiv:1712.06550** (2017).
[9] M. Takita, A. W. Cross, A. D. Córcoles, J. M. Chow, and J. M. Gambetta, Phys. Rev. Lett. **119**, 180501 (2017).
[10] J. J. Dongarra, P. Luszczek, and A. Petitet, Concurrency and Computation: Practice and Experience **15**, 803 (2003).
[11] J. Dongarra, M. A. Heroux, and P. Luszczek, *HPCG Benchmark: a New Metric for Ranking High Performance Computing Systems*, Technical Report UT-EECS-15-736 (Electrical Engineering and Computer Science Department, Knoxville, Tennessee, 2015).
[12] M. Adams, *HPGMG 1.0: A Benchmark for Ranking High Performance Computing Systems*, Technical Report LBNL 6630E (LBNL, 2014).
[13] L. S. Bishop, S. Bravyi, A. Cross, J. M. Gambetta, and J. A. Smolin, "Quantum volume," (2017).
[14] N. Moll, P. Barkoutsos, L. S. Bishop, J. M. Chow, A. Cross, D. J. Egger, S. Filipp, A. Fuhrer, J. M. Gambetta, M. Ganzhorn, A. Kandala, A. Mezzacapo, P. Muller, W. Riess, G. Salis, J. Smolin, I. Tavernelli, and K. Temme, Quantum Sci. Technol. **3**, 030503 (2018).
[15] M. Nielsen and I. Chuang, *Quantum computation and quantum information* (Cambridge, 2000).
[16] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Nat. Phys. **14**, 595 (2018).
[17] E. Farhi, J. Goldstone, and S. Gutmann, arXiv preprint **arXiv:1411.4028** (2014).
[18] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New J. Phys. **18**, 023023 (2016); M.-H. Yung, J. Casanova, A. Mezzacapo, J. McClean, L. Lamata, A. Aspuru-Guzik, and E. Solano, Sci. Rep. **4** (2014).
[19] S. Aaronson and L. Chen, arXiv preprint **arXiv:1612.05903** (2016).
[20] M. Horodecki, P. Horodecki, and R. Horodecki, Physical Review A **60**, 1888 (1999).
[21] Z.-Y. Chen, Q. Zhou, C. Xue, X. Yang, G.-C. Guo, and G.-P. Guo, Science Bulletin **63**, 964 (2018), arXiv: 1802.06952.
[22] J. Chen, F. Zhang, C. Huang, M. Newman, and Y. Shi, arXiv:1805.01450 [quant-ph] (2018), arXiv: 1805.01450.
[23] R. Li, B. Wu, M. Ying, X. Sun, and G. Yang, arXiv:1804.04797 [quant-ph] (2018), arXiv: 1804.04797.
[24] E. Pednault, J. A. Gunnels, G. Nannicini, L. Horesh, T. Magerlein, E. Solomonik, and R. Wisnieff, arXiv:1710.05867 [quant-ph] (2017), arXiv: 1710.05867.
[25] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, and H. Neven, arXiv:1712.05384 [quant-ph] (2017), arXiv: 1712.05384.
[26] T. Häner, D. S. Steiger, M. Smelyanskiy, and M. Troyer, arXiv:1604.06460 [quant-ph] (2016), arXiv: 1604.06460.
[27] M. Smelyanskiy, N. P. D. Sawaya, and A. Aspuru-Guzik, arXiv:1601.07195 [quant-ph] (2016), arXiv: 1601.07195.
[28] I. Markov and Y. Shi, SIAM Journal on Computing **38**, 963 (2008).
[29] "5-qubit backend: IBM Q team, "IBM Q 5 Tenerife backend specification v1.3.0, (2018)." https://ibm.biz/qiskit-tenerife, Last Accessed: 2018-11.
[30] "16-qubit backend: IBM Q team, "IBM Q 16 Melbourne backend specification v1.0.0, (2018)." https://ibm.biz/qiskit-melbourne, Last Accessed: 2018-11.
[31] S. Bullock and I. Markov, Phys. Rev. A **68**, 012318 (2003).
[32] V. Shende, I. Markov, and S. Bullock, Phys. Rev. A **69**, 062321 (2004).
[33] H. Abraham, I. Y. Akhalwaya, G. Aleksandrowicz, T. Alexander, G. Alexandrowics, E. Arbel, A. Asfaw, C. Azaustre, P. Barkoutsos, G. Barron, L. Bello, Y. Ben-Haim, L. S. Bishop, S. Bosch, D. Bucher, CZ, F. Cabrera, P. Calpin, L. Capelluto, J. Carballo, C.-F. Chen, A. Chen, R. Chen, J. M. Chow, C. Claus, A. W. Cross, A. J. Cross, J. Cruz-Benito, C. Culver, A. D. Córcoles-Gonzales, S. Dague, M. Dartiailh, A. R. Davila, D. Ding, E. Dumitrescu, K. Dumon, I. Duran, P. Eendebak, D. Egger, M. Everitt, P. M. Fernández, A. Frisch, A. Fuhrer, J. Gacon, Gadi, B. G. Gago, J. M. Gambetta, L. Garcia, S. Garion, Gawel-Kus, L. Gil, J. Gomez-Mosquera, S. de la Puente González, D. Greenberg, J. A. Gunnels, I. Haide, I. Hama-

mura, V. Havlicek, J. Hellmers, Ł. Herok, H. Horii, C. Howington, W. Hu, S. Hu, H. Imai, T. Imamichi, R. Iten, T. Itoko, A. Javadi-Abhari, Jessica, K. Johns, N. Kanazawa, A. Karazeev, P. Kassebaum, V. Krishnan, K. Krsulich, G. Kus, R. LaRose, R. Lambert, J. Latone, S. Lawrence, P. Liu, P. B. Z. Mac, Y. Maeng, A. Malyshev, J. Marecek, M. Marques, D. Mathews, A. Matsuo, D. T. McClure, C. McGarry, D. McKay, S. Meesala, A. Mezzacapo, R. Midha, Z. Minev, P. Murali, J. Müggenburg, D. Nadlinger, G. Nannicini, P. Nation, Y. Naveh, Nick-Singstock, P. Niroula, H. Norlen, L. J. O'Riordan, S. Oud, D. Padilha, H. Paik, S. Perriello, A. Phan, M. Pistoia, A. Pozas-iKerstjens, V. Prutyanov, J. Pérez, Quintiii, R. Raymond, R. M.-C. Redondo, M. Reuter, D. M. Rodríguez, M. Ryu, M. Sandberg, N. Sathaye, B. Schmitt, C. Schnabel, T. L. Scholten, E. Schoute, I. F. Sertage, Y. Shi, A. Silva, Y. Siraichi, S. Sivarajah, J. A. Smolin, M. Soeken, D. Steenken, M. Stypulkoski, H. Takahashi, C. Taylor, P. Taylour, S. Thomas, M. Tillet, M. Tod, E. de la Torre, K. Trabing, M. Treinish, TrishaPe, W. Turner, Y. Vaknin, C. R. Valcarce, F. Varchon, D. Vogt-Lee, C. Vuillot, J. Weaver, R. Wieczorek, J. A. Wildstrom, R. Wille, E. Winston, J. J. Woehr, S. Woerner, R. Woo, C. J. Wood, R. Wood, S. Wood, J. Wootton, D. Yeralin, J. Yu, L. Zdanski, Zoufalc, azulehner, drholmie, fanizzamarco, kanejess, klinvill, merav aharoni, ordmoj, tigerjack, yang.luh, and yotamvakninibm, "Qiskit: An open-source framework for quantum computing," (2019).

[34] C. J. Ballance, T. P. Harty, N. M. Linke, M. A. Sepiol, and D. M. Lucas, Phys. Rev. Lett. **117**, 060504 (2016).

[35] T. Monz, P. Schindler, J. T. Barreiro, M. Chwalla, D. Nigg, W. A. Coish, M. Harlander, W. Hänsel, M. Hennrich, and R. Blatt, Phys. Rev. Lett. **106**, 130506 (2011).

[36] A. W. Cross, L. Bishop, J. Smolin, and J. M. Gambetta, arXiv preprint **arXiv:1707.03429** (2017).

[37] E. Cuthill and J. McKee, Proc. 24th Nat. Conf. ACM , 157 (1969).

[38] A. George and J. W. Liu, *Computer Solution of Large Sparse Positive Definite Systems* (Prentice-Hall, 1981).

[39] W. M. Chan and A. George, BIT **20**, 8 (1980).

[40] B. Kraus and J. I. Cirac, Physical Review A **63**, 062309 (2001).

[41] N. Khaneja, R. Brockett, and S. J. Glaser, Physical Review A **63**, 032308 (2001).

[42] P. Watts, J. Vala, M. M. Müller, T. Calarco, K. B. Whaley, D. M. Reich, M. H. Goerz, and C. P. Koch, Physical Review A **91**, 062306 (2015).

[43] F. Vatan and C. Williams, Physical Review A **69**, 032315 (2004).

[44] G. Vidal and C. M. Dawson, Physical Review A **69**, 010301 (2004).

[45] Y.-S. Zhang, M.-Y. Ye, and G.-C. Guo, Physical Review A **71**, 062331 (2005).

[46] P. Watts, M. O'Connor, J. Vala, P. Watts, M. O'Connor, and J. Vala, Entropy **15**, 1963 (2013).

[47] M. Musz, M. Kuś, and K. Życzkowski, Physical Review A **87**, 022111 (2013).

## Appendix A: Qiskit transpiler passes

Model circuits must be rewritten to use the gate set of the target system, while attempting to minimize any additional overhead that might result from the translation. The IBM Q systems used in this paper accept quantum circuits expressed by products of controlled-NOT (CNOT) gates and single-qubit gates [36]. The single-qubit gates are defined by

$$u_1(\lambda) = \mathrm{diag}(1, e^{i\lambda}) \tag{A1}$$

$$u_2(\phi, \lambda) = R_z(\phi + \pi/2) R_x(\pi/2) R_z(\lambda - \pi/2) \tag{A2}$$

$$u_3(\theta, \phi, \lambda) = R_z(\phi + 3\pi) R_x(\pi/2) R_z(\theta + \pi) R_x(\pi/2) R_z(\lambda) \tag{A3}$$

where $R_P(\theta) = \exp(-i\theta P/2)$ for a Pauli matrix $P \in \{X, Y, Z\}$. The available CNOT gates for a particular system are given in the form of a qubit connectivity graph $G = (V, E)$. Each vertex of $G$ represents a qubit and each (directed) edge represents a pair of qubits that can be coupled by gates.

We generate input model circuits by sampling and expanding each SU(4) gate to CNOT and single-qubit gates using the KAK decomposition [31, 32] implemented in Qiskit Terra (see also Appendix B). Each input circuit is then mapped to the target system and optimized using a sequence of circuit rewriting passes that are implemented in Qiskit Terra. These passes are named unrolling, CNOT reorientation, CNOT cancellation, single-qubit optimization, and swap mapping. All of the passes can be applied multiple times, but some passes, such as CNOT reorientation, have requirements that are ensured by other passes, such as swap mapping.

The unrolling pass is essentially a macro expansion that descends into each gate's hierarchical definition and rewrites that gate in terms of lower-level gates. In the setting of rewriting model circuits, the lower-level gate set is always the IBM Q gate set. For example, a Hadamard ($H$) gate is defined as $u_2(0, \pi)$ in the Qiskit Terra gate library, which is in the IBM Q gate set, and a SWAP gate is defined as $\mathrm{CNOT}_{a,b}\,\mathrm{CNOT}_{b,a}\,\mathrm{CNOT}_{a,b}$.

The CNOT reorientation pass examines each CNOT gate in the circuit and applies the identity

$$\mathrm{CNOT}_{c,t} = (H \otimes H)\,\mathrm{CNOT}_{t,c}(H \otimes H) \tag{A4}$$

if $(t, c)$ is a directed edge of $G$ but $(c, t)$ is not. The pass fails if neither $(c, t)$ nor $(t, c)$ are edges of $G$.

The CNOT cancellation pass collects sequences $\mathrm{CNOT}_{c,t}^m$ of CNOT gates with the same control and target qubits, and replaces them by $\mathrm{CNOT}_{c,t}$ if $m$ is odd or removes them from the circuit if $m$ is even.

The single-qubit optimization pass collects sequences of single-qubit gates on the same qubit and replaces each sequence by at most one single-qubit gate. Furthermore, the replacement is chosen in an attempt to minimize the number of physical pulses used to implement the gate; $u_1$ uses zero pulses, $u_2$ uses one pulse, and $u_3$ uses two

pulses. The algorithm composes the gates in sequence, rewriting each composed pair of gates as a new gate according to a handful of rewriting rules that follow from the definitions.

The swap mapping pass is the most involved of the fundamental passes within Qiskit Terra. This pass first partitions the input circuit into a sequence of layers such that each layer consists of gates that act on disjoint sets of qubits. The algorithm then acts layer by layer. For simplicity we will ignore single-qubit gates in the following discussion. Consider the gate $U = U_1 U_2 \dots U_m$ applied in a particular layer, where $U_1, \dots, U_m$ are pairwise disjoint two-qubit gates that may act on remote pairs qubits. When the mapping pass acts on this layer, it computes a quantum circuit $U'$ with the following properties:

1. $U'$ consists of nearest-neighbor gates with respect to the connectivity graph $G = (V, E)$
2. $U' = WU$ where $W$ is some permutation of the $n = |V|$ qubits
3. $U'$ has small depth, which the algorithm tries to minimize subject to the first two conditions

The algorithm to compute $U'$ consists of a sequence of rounds, each of which increases the depth of $U'$ by one. At the beginning of a round, the algorithm applies all gates $U_j$ that are nearest-neighbors and removes them from $U$. The rest of the round performs a greedy (randomized) optimization over swap gates to choose a depth-one swap circuit that brings pairs of qubits coupled by gates as close as possible.

The passes are applied in the following order for our standard compilation:

1. Unrolling pass
2. Swap mapping pass
3. Unrolling pass (to expand SWAP gates)
4. CNOT reorientation pass
5. CNOT cancellation pass
6. Unrolling pass (to expand Hadamard gates)
7. Single-qubit optimization pass

In our study of optimized model circuits, we apply the following optimization passes after the standard set of passes:

1. Two-qubit block collection pass
2. Two-qubit block optimization pass

The two-qubit block collection pass is an analysis pass that traverses the circuit's gates in topologically sorted order. Starting at each newly-discovered CNOT gate, the pass explores that gate's predecessors and ancestors to collect the largest block of previously unseen and contiguous gates acting on the control and target qubits. The pass continues in this manner and returns a collection of disjoint blocks. The two-qubit block optimization pass computes the unitary operation for each block, synthesizes a new sub-circuit (either exactly, using the KAK decomposition [31, 32], or approximately; see Appendix B), and replaces the block.

To further reduce the number of SWAP gates, we considered an optimization called the Local Ordering Circuit Optimization (LOCO), that permutes qubits such that those interacting via CNOT gates are as nearest-neighbor as possible in the circuit representation; the circuit is optimized for a linear nearest-neighbor topology. This method employs a weighted-variant of reverse Cuthill-Mckee ordering [37, 38] to reorder the sparse matrix $A_{ij}$, with non-zero elements counting the number of CNOT gate operations between qubits $i$ and $j$ in the circuit, so that its bandwidth is minimized. The matrix is symmetric as we do not consider the direction of the CNOTs. This reordering is efficient, having a runtime that is linear in the number of nonzero matrix elements [39]. To properly account for multiple CNOT interactions between qubits, the LOCO algorithm uses a weighted heuristic when reordering, that favors optimizing pairs of qubits with the largest number of repeated interactions over those with fewer gates between them. Input circuits whose bandwidth was reduced by LOCO were replaced with their optimized counterparts. Although this optimization did not lead to significant improvements for heavy output generation using small numbers of qubits, we expect SWAP optimizations such as these to further improve results for larger circuits mapped onto devices with limited connectivity.

## Appendix B: Approximate compiling

We can always decompose [40, 41] an arbitrary two-qubit unitary in the form

$$U = K_1 U_d(\alpha, \beta, \gamma) K_2, \tag{B1}$$

where $K_i = K_i^l \otimes K_i^r$ are products of single-qubit unitaries $K_i^{l,r}$, the two-qubit component is represented in terms of the *information content* $(\alpha, \beta, \gamma)$ as

$$U_d(\alpha, \beta, \gamma) = \exp[i(\alpha \sigma_x \otimes \sigma_x + \beta \sigma_y \otimes \sigma_y + \gamma \sigma_z \otimes \sigma_z)], \tag{B2}$$

and we can always restrict to the Weyl chamber $\pi/4 \geq \alpha \geq \beta \geq |\gamma|$. Let $U \sim V$ denote equivalence between $U$ and $V$ under local operations, implying equality of the information content of $U$ and $V$.

We can calculate a trace of the product of two $U_i = U_d(\alpha_i, \beta_i, \gamma_i)$ as

$$\begin{aligned} \mathrm{Tr}(U_c^\dagger U_t) = {} & 4\cos(\Delta_\alpha)\cos(\Delta_\beta)\cos(\Delta_\gamma) \\ & - 4i \sin(\Delta_\alpha)\sin(\Delta_\beta)\sin(\Delta_\gamma), \end{aligned} \tag{B3}$$

where

$$\Delta_\alpha = \alpha_c - \alpha_t, \tag{B4a}$$
$$\Delta_\beta = \beta_c - \beta_t, \tag{B4b}$$
$$\Delta_\gamma = \gamma_c - \gamma_t. \tag{B4c}$$

From this trace we may easily determine the average gate fidelity [20]

$$F_{\text{avg}}(U_c, U_t) = \frac{4 + \left|\text{Tr}(U_c^\dagger U_t)\right|^2}{20} \tag{B5}$$

and these expressions give also the maximal fidelity between arbitrary unitaries $U_{c,t} \in \text{SU}(4)$ after optimizing over local pre- and post-rotations [42]

$$\max_{K_1^l, K_1^r, K_2^l, K_2^r} F_{\text{avg}}\big[(K_1^l \otimes K_1^r) U_c (K_2^l \otimes K_2^r), U_t\big]. \tag{B6}$$

We are interested in decompositions of a target unitary $U_t \in \text{SU}(4)$ with the minimal number of applications of a fixed 'basis' gate $U_b$. It is obvious that with zero applications of the basis we can construct only non-entangling target unitaries $U_t \sim U_d(0,0,0)$, and with one application of the basis we can construct only target unitaries which are equivalent to the basis $U_t \sim U_d(\alpha_b, \beta_b, \gamma_b)$. For $U_b \sim \text{CNOT} \sim U_d(\pi/4, 0, 0)$ it is well-known [43, 44] that 3 applications of the basis is sufficient to cover all of SU(4). Zhang et al. [45] give decompositions using a more general 'super controlled' basis $U_b \sim U_d(\pi/4, \beta_b, 0)$, for any $\beta_b$, both an expansion with 3 applications of $U_b$ to decompose an arbitrary $U_t \sim U_d(\alpha_t, \beta_t, \gamma_t)$ and also an expansion using two applications of $U_b$ for a restricted target unitary $U_t \sim U_d(\alpha_t, \beta_t, 0)$, $\gamma_t = 0$ for any $\alpha_t$, $\beta_t$.

The above expansions are *exact* so that the constructed unitary $U_c$ satisfies

$$F_{\text{avg}}(U_t, U_c) = 1, \tag{B7}$$

but we can use eq. (B5) to find the average gate fidelity due to approximating general $U_t$ by fewer applications of the basis gate than is necessary for exact expansion. With zero applications of arbitrary $U_b$ we have:

$$U_c^{(0)} = K_{t,1} K_{t,2}, \tag{B8a}$$
$$F_{\text{avg}}^{(0)} = \Big[1 + 4\cos^2(\alpha_t)\cos^2(\beta_t)\cos^2(\gamma_t) + 4\sin^2(\alpha_t)\sin^2(\beta_t)\sin^2(\gamma_t)\Big]/5, \tag{B8b}$$

which is optimal. With one application of arbitrary $U_b$ we have:

$$U_c^{(1)} = K_{t,1} U_d(\alpha_b, \beta_b, \gamma_b) K_{t,2}, \tag{B8c}$$
$$F_{\text{avg}}^{(1)} = \Big[1 + 4\cos^2(\Delta_\alpha)\cos^2(\Delta_\beta)\cos^2(\Delta_\gamma) + 4\sin^2(\Delta_\alpha)\sin^2(\Delta_\beta)\sin^2(\Delta_\gamma)\Big]/5, \tag{B8d}$$

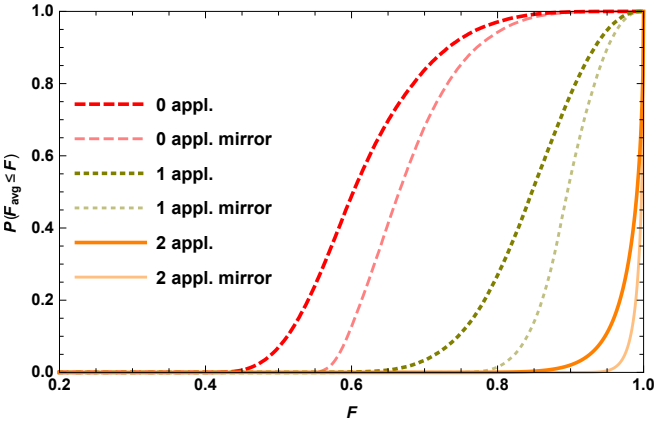


FIG. 5. (Color online) Average gate fidelity for random target gates in the Haar measure, for approximations using zero, one, or two applications of a 2-qubit super controlled basis gate, with and without freedom to mirror. These approximations are optimal for the case that the basis gate is equivalent to CNOT.

which is optimal. With two applications of super controlled $U_b \sim U_d(\pi/4, \beta_b, 0)$ we have:

$$U_c^{(2)} = K_{t,1} U_d(\alpha_t, \beta_t, 0) K_{t,2}, \tag{B8e}$$
$$F_{\text{avg}}^{(2)} = \Big[1 + 4\cos^2(\gamma_t)\Big]/5, \tag{B8f}$$

which is optimal for $U_b \sim \text{CNOT} \sim U_d(\pi/4, 0, 0)$ or $U_b \sim \text{DCNOT} \sim U_d(\pi/4, \pi/4, 0)$. For completeness, with 3 applications of super controlled $U_b$ there is no need to approximate and we have:

$$U_c^{(3)} = K_{t,1} U_d(\alpha_t, \beta_t, \gamma_t) K_{t,2} = U_t, \tag{B8g}$$
$$F_{\text{avg}}^{(3)} = 1, \tag{B8h}$$

which is clearly optimal.

There can be an additional freedom when expanding a two-qubit gate: in many cases it does not matter whether we implement $U_t$ or $U_{tm} = U_t \cdot \text{SWAP}$ since the latter differs merely by permutation of the output qubit labels. We call it the *mirror gate* of $U_t$ and its expansion is easily related to $U_t$:

$$U_{tm} \sim U_d\big(\pi/4 - |\gamma_t|, \pi/4 - \beta_t, \text{sgn}(\gamma_t)(\alpha_t - \pi/4)\big), \tag{B9}$$

making use of the sign function defined as $\text{sgn}(x) = -1$ for $x < 0$ and $\text{sgn}(x) = 1$ for $x \geq 0$. We can extend eqs. (B8) to give $i$-gate expansions of $U_{tm}$, $U_c^{(im)}$ with fidelities $F_{\text{avg}}^{(im)}$, defined by choosing to expand whichever of $U_t$ and $U_{tm}$ gives the better fidelity. For example, the 2-gate expansion has

$$F_{\text{avg}}^{(2m)} = \Big[1 + 4\cos^2\Big(\min\big[|\gamma_t|, |\alpha_t - \pi/4|\big]\Big)\Big]/5. \tag{B10}$$

Because of the mirroring action within the Weyl chamber, the expansion of the mirrored gate has best fidelity

exactly when the expansion of the unmirrored gate has worst fidelity, and vice versa. In addition to improving $F_{\text{avg}}$, the freedom to combine a SWAP operation may also allow reduction in the number of inserted SWAP gates during a 'swap mapping pass' as described in Appendix A.

It is interesting to investigate the expected infidelity of each of the approximate expansions of $U_t$, averaged over $U_t$ uniformly distributed within SU(4) in the Haar measure on the Weyl chamber [46, 47]

$$M(\alpha,\beta,\gamma) = \frac{24}{\pi}\Big[\cos(4\alpha)\cos(8\beta) + \cos(4\beta)\cos(8\gamma) + \cos(4\gamma)\cos(8\alpha) - \cos(8\alpha)\cos(4\beta) - \cos(8\beta)\cos(4\gamma) - \cos(8\gamma)\cos(4\alpha)\Big], \tag{B11}$$

allowing calculating the distribution of fidelities of the 2-basis gate approximation of eq. (B8f) for a random element of SU(4)

$$P(F_{\text{avg}}^{(2)} < F) = \cos^4(2z)\Big[\big(4z-\pi\big)\big(\cos(4z)-2\big) - 3\sin(4z)\Big]/\pi, \tag{B12}$$

where $z$ is defined by

$$\cos(z) = \frac{\sqrt{5F-1}}{2}, \tag{B13}$$

for $F > 3/5$, and

$$P(F_{\text{avg}}^{(2)} < F) = 0 \tag{B14}$$

for $F \leq 3/5$. Similarly, for the mirrored version eq. (B10)

$$P(F_{\text{avg}}^{(2m)} < F) = \cos(4z)\Big[\big(8z-\pi\big)\big(\cos(8z)-2\big) - 3\sin(8z)\Big]/\pi, \tag{B15}$$

for $z < \pi/8$, $F > 0.88$, and

$$P(F_{\text{avg}}^{(2m)} < F) = 0 \tag{B16}$$

for $z \geq \pi/8$.

The 2-basis gate approximations perform surprisingly well, with the median fidelities $F_{\text{avg}}^{(2)} = 0.99$, $F_{\text{avg}}^{(2m)} = 0.997$ comparing favorably to the typical 2-qubit gate fidelities for current quantum devices. The full distribution of fidelities for the zero-, one-, and two-gate approximations are plotted in Fig. 5, where the zero- and one-gate distributions are determined by random sampling.

By comparing $F_{\text{avg}}^{(i)}$ for all $i$ we can choose the best approximation for any given $U_t$. Specifically, if the basis gate $U_b$ may be implemented with average gate fidelity $F_b$ we can estimate the overall fidelity by multiplying the fidelity due to approximation with the fidelity due to the number of applications of $U_b$, and choose the expansion with the highest overall fidelity

$$F_{\text{best}} = \max_i F_{\text{avg}}^{(i)}(F_b)^i, \tag{B17a}$$

$$F_{\text{best}}^{(m)} = \max_i F_{\text{avg}}^{(im)}(F_b)^i. \tag{B17b}$$

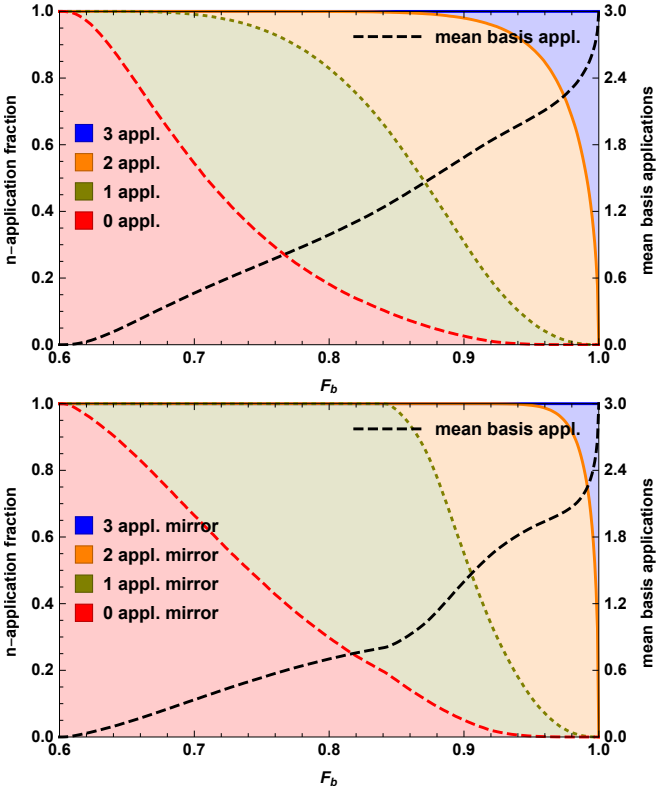


FIG. 6. (Color online) Number of basis gate applications used for approximating chosen for randomly chosen target gates in the Haar measure, choosing the approximation according to eq. (B17) as a function of the basis gate fidelity $F_b$. Fraction of cases with each number of applications is shown by shading (left axis) and the mean number of basis applications is shown by the dashed line (right axis). (a) without mirroring, as in eq. (B17a), (b) with mirroring, as in eq. (B17b)

The statistics of the number of basis gate applications for a randomly-generated ensemble of target gates are shown in Fig. 6. With a fairly noisy basis gate $F_b = 0.97$ and no mirroring, the best expansion by this method has 3 applications of the basis for 22%, two applications for 76%, one application for 2%, and zero applications for $< 0.1\%$ of targets, thus an average of 2.2 basis gate applications. With the freedom to mirror, three applications for 3%, two applications for 93%, one application for 4%, and zero applications for $< 0.1\%$ of targets, thus a mean of 2.0 basis gate applications. The resulting fidelity can be quoted as an 'effective fidelity' $F_e$ equal to the cube root of the mean of $F_{\text{best}}$, which we can interpret as the equivalent basis gate fidelity if we were to use only exact 3-gate expansions of random targets. We show in Fig. 7 the ratio of the effective infidelity $1 - F_e$

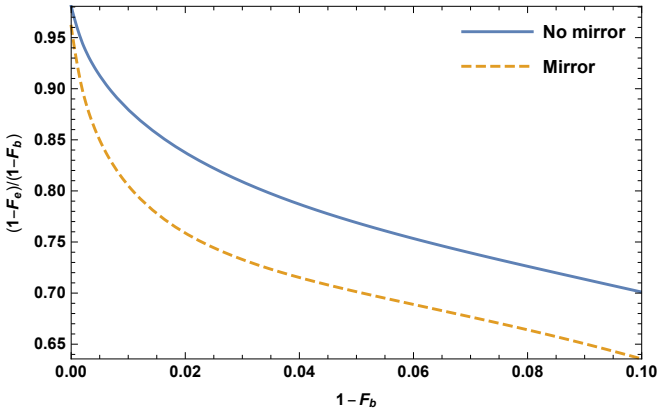


FIG. 7. (Color online) Effective infidelity ratio as a function of basis gate infidelity, with and without freedom to mirror.

to the basis gate infidelity $1 - F_b$, giving the factor by which the use of approximate expansions improves effective gate performance. For $F_b = 0.97$ we get $F_e = 0.976$, $F_e^{(m)} = 0.978$, reducing the infidelity by factors of 0.82 and 0.74 respectively.

For the volume measurements described in the main text, Table II, we implemented the approximate two-qubit block optimization compilation pass without mirroring, assuming fixed $1 - F_b$ of 1%, 3% or 5%. Because the 4 qubits chosen for the *System One* have a linear nearest-neighbor topology, we were able to implement a special-case optimization that replaces some gates by the corresponding mirrored gate in order to minimize the number of inserted SWAP gates for this topology. Using measured CNOT fidelities for each of the qubit pairs, implementing the mirror expansions, and combining the mirror choice with a swap-mapping pass for general topologies should allow future compiler-driven improvements in quantum volume.

## Appendix C: Confidence intervals for the heavy probability

To be confident with a finite number of trials that the heavy probability $h_d$ exceeds 2/3, we should set stricter threshold $t > 2/3$, requiring the estimated heavy probability $\hat{h}_d > t$ to claim success. This is a hypothesis test with null hypothesis $H_0 : h_d = 2/3$ and alternative hypothesis $H_1 : h_d > 2/3$. Drawing $n_c$ random model circuits of given width and depth, and executing each circuit $n_s$ times gives a total of $n_c n_s$ experiment outcomes, each of which is to be checked against simulation of the corresponding circuit to determine a count $n_h$ of heavy outcomes. We estimate $h_d$ in the natural way by the heavy fraction over these outcomes

$$\hat{h}_d = \frac{n_h}{n_c n_s}. \tag{C1}$$

For the purposes of making a conservative bound on the spread of $\hat{h}_d$ we analyze using the worst-case distribution where the heavy probability conditioned on each circuit is either zero or one. Thus, executing each circuit multiple times $n_s > 1$ (as is typically convenient to avoid reconfiguring experimental settings and allow recycling of simulation results) will generally narrow the observed fluctuations in $\hat{h}_d$ but, for fear of systematic errors we do not allow this to alter the threshold $t$. Under this worst-case assumption, $n_h/n_s$ is binomial distributed with parameter $n_c$.

While it would be straightforward to calculate numerically confidence intervals directly from the binomial distribution, because the interesting range of $\hat{h}_d$ is close to 2/3 where a normal approximation is valid, we instead require a minimum of $n_c = 100$ circuits and make a normal approximation to the binomial, and write the requirements for claiming success at a given width and depth

$$n_c \geq 100 \tag{C2}$$

$$\frac{n_h - z\sqrt{n_h(n_s - \frac{n_h}{n_c})}}{n_c n_s} > \frac{2}{3}, \tag{C3}$$

where we set $z = 2$ for a 97.5% '2-sigma' one-sided confidence interval. For example, to claim success with $n_c = 5000$ model circuits, the observed heavy fraction must exceed the threshold $t = 0.68$.

## Appendix D: Device parameters

We measured the quantum volume of four IBM Q devices: 5-qubit *Tenerife*, 16-qubit *Melbourne*, and 20-qubit *Tokyo*, and 20-qubit *System One*. The device connectivities are shown in Fig. 8, with the four qubits from each device that were used for the $m = d = 4$ experiments highlighted in grey boxes. Table V lists the average error rates for the set of qubits used in these experiments. These error rates were measured one day before the quantum volume experiments were performed. Fluctuations in these numbers can occur during the time scale of these experiments, but they are representative of the single-qubit, two-qubit, and measurement errors for each device. The data from Table V was also used in the noisy simulations of the quantum volume circuits in Table II.

| | Tenerife | Melbourne | Tokyo | System One |
|---|---|---|---|---|
| # Qubits | 5 | 16 | 20 | 20 |
| $\epsilon_{1Q}$ | $1.7 \times 10^{-3}$ | $1.6 \times 10^{-3}$ | $1.6 \times 10^{-3}$ | $0.4 \times 10^{-3}$ |
| $\epsilon_{CX}$ | $4.7 \times 10^{-2}$ | $3.4 \times 10^{-2}$ | $2.1 \times 10^{-2}$ | $1.1 \times 10^{-2}$ |
| $\epsilon_M$ | $5.8 \times 10^{-2}$ | $8.7 \times 10^{-2}$ | $3.0 \times 10^{-2}$ | $3.9 \times 10^{-2}$ |

TABLE V. Average error rates for the experimental devices: $\epsilon_{1Q}$ for single-qubit error rates, $\epsilon_{CX}$ for two-qubit error rates, and $\epsilon_M$ for measurement. The averages are taken over the set of qubits from each device that were used in the quantum volume experiments.

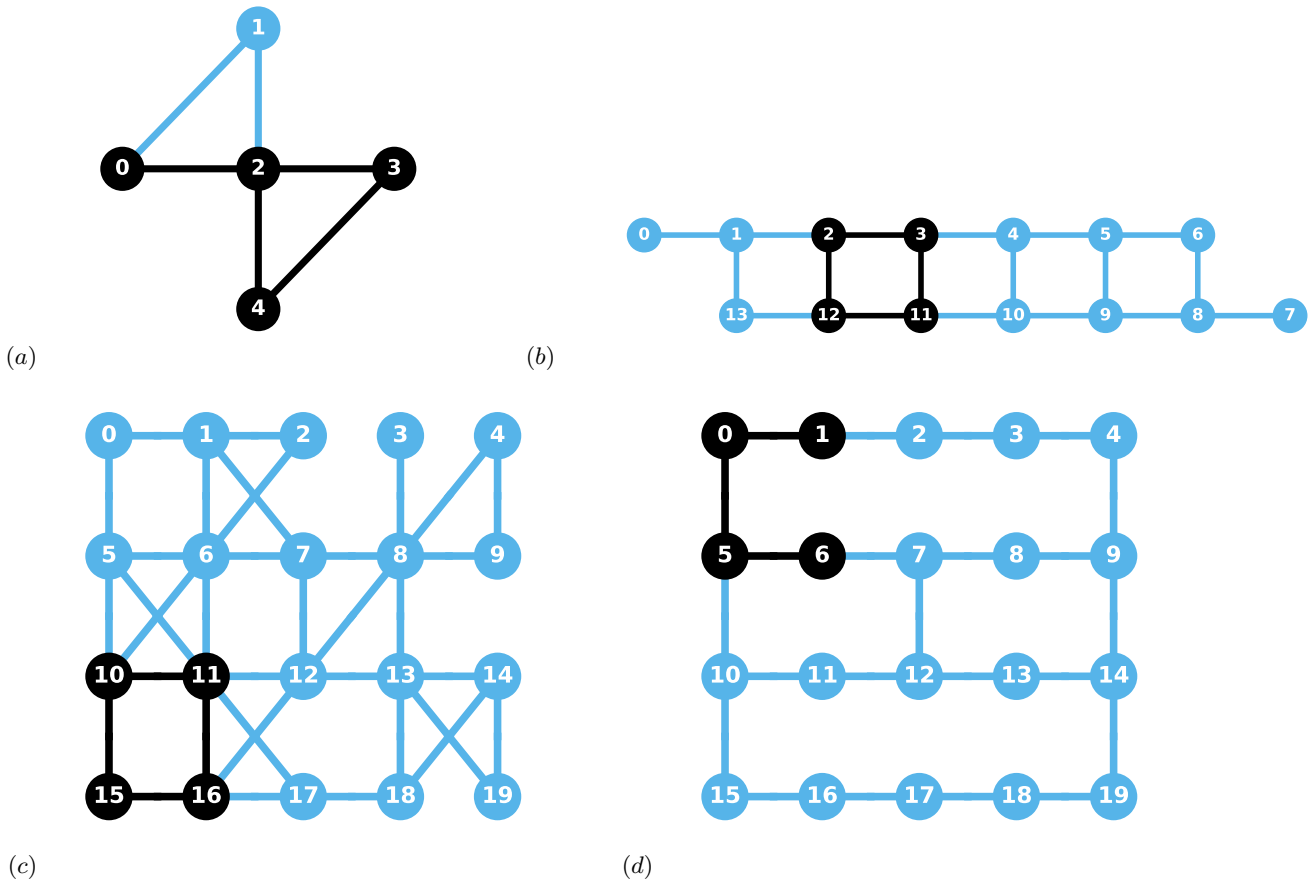


FIG. 8. Device diagrams used for the experimental data in Table I: (a) *Tenerife*, (b) *Melbourne*, (c) *Tokyo*, and (d) *IBM Q System One*. The highlighted qubits are those selected for the experiments discussed here. CX gates are available between pairs of qubits connected by a highlighted line.