Date of publication TBD, date of current version: 15 October, 2021.

Digital Object Identifier TBD

Stability of Noisy Quantum Computing Devices

SAMUDRA DASGUPTA[1,2], and TRAVIS S. HUMBLE[1,2]
[1]Quantum Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[2]Bredesen Center, University of Tennessee, Knoxville, USA

Corresponding author: Samudra Dasgupta (ORCID: 0000-0002-7831-745X)

ABSTRACT Noisy, intermediate-scale quantum (NISQ) computing devices offer opportunities to test the principles of quantum computing but are prone to errors arising from various sources of noise. However, fluctuations in the noise itself can lead to unstable devices that are unreliable and undermine the reproducibility of program results including performance benchmarks. Here we characterize the reliability of NISQ devices by quantifying the stability of essential performance metrics. Using the Hellinger distance, we quantify the similarity between experimental characterizations of several NISQ devices by comparing gate fidelities, duty cycles, and register addressability across multiple temporal and spatial scales. From observations collected over 22 months, we identify fluctuations in these metrics that characterize the instability of current NISQ devices and we conclude that consistent monitoring of such metrics are required to ensure reproducibility of quantum computing benchmarks.

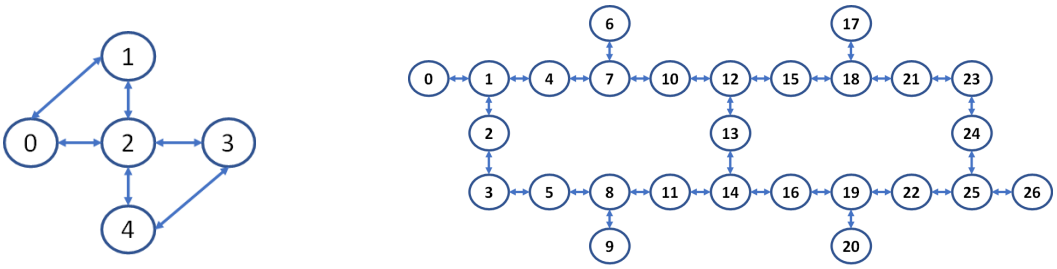
INDEX TERMS Quantum Computing, Device Characterization, Stability

I. INTRODUCTION

ON-GOING efforts to realize the principles of quantum computing have demonstrated high-fidelity control over quantum physical systems ranging from superconducting electronics [1], trapped ions [2], and silicon [3] among many others [4]. As these efforts aim for future fault-tolerant operation [5], they currently establish noisy, intermediate-scale quantum (NISQ) devices as a frontier for testing quantum computing under experimental conditions [6]. As first-in-kind platforms, NISQ computing devices enable design verification [7], device characterization [8], program validation [9], and a breadth of testing and evaluation for application performance [10]–[15] with several recent demonstrations exemplifying the milestone of quantum computational advantage [16]–[18].

A prominent feature of experimental NISQ computing is that noise and errors limit the behavior of these devices [19]. Characterization methods to quantify noise in NISQ devices inform benchmarking methods that assess the accuracy of a quantum computation [20]–[22]. Reliable benchmarks establish bounds on the statistical significance of the observed results and help clarify the conditions necessary for results to be reproducible. However, experimental characterizations of current NISQ devices reveal transient changes in the noise that undermine benchmarking effort [23]. Even when temporal drift and spatial variability are mitigated efficiently over short time scales, e.g., the duration of a single program execution [24]–[28], the presence of unstable noise sources impede efforts to reproduce experimental benchmarks and track progress in device performance. Subsequent methods to quantify and monitor device stability are therefore needed to assess the reproducibility of quantum computing benchmarks.

Here we address the stability of NISQ devices as a key concern in the reproducibility of quantum computing benchmarks. We define stability relative to fluctuations in noisy behavior across space and time, where a stable device is defined to exhibit noise that is consistent across these different scales within a desired tolerance. We present a suite of metrics by which to evaluate stability in current NISQ devices, and we experimentally quantify stability of several example devices. Device stability is directly related to the reproducibility of



(a) Yorktown **(b)** Toronto

FIGURE 1: Schematic layout of the (a) yorktown and (b) toronto devices produced by IBM. Circles denote register elements and edges denote connectivity of 2-qubit operations.

experimental results, but it is distinct from characterizing device fidelity. The latter assess how well a device metric meets design criteria and ignores the transient changes that gives rise to unstable behavior. The stability analysis presented here identifies when device metrics are stable and the experimental conditions under which benchmark results are reproducible.

TABLE 1: Device metrics for assessing DiVincenzo criteria

Metric (*Symbol*)	*Description*
Capacity (n)	Maximal amount of information that may be stored in the register.
Initialization Fidelity (F_{I})	Accuracy with which a fiducial register state is prepared.
Gate Fidelity (F_{G})	Accuracy with which a gate transforms the register state.
Duty Cycle (τ)	Ratio of gate duration to coherence time.
Addressability (F_{A})	Mutual information between register elements.

Motivated by the crucial difference between the computational accuracy of an executed program and the stability of the underlying device, our approach quantifies device stability to determine when, if ever, a device metric may be considered stable. Our approach borrows criteria originally proposed by DiVincenzo as minimal for the realization of quantum computing [29]. Table 1 presents each criterion as quantified by a corresponding metric alongside a representative symbol and definition. A variety of methods exist for estimating each metric, and we select specific approaches to demonstrate the role of stability in observed results. For example, we characterize the capacity n of the quantum register as the number of elements that can be addressed. We summarize approaches for the other metrics below with additional details found in Appendix A.

Initialization fidelity F_I quantifies the accuracy with which a target quantum state is prepared in the register, and we quantify this fidelity in terms of the observed error following readout. Sophisticated tomographic approaches are generally more informative, but we find that characterization in a single computational basis state, $|0\rangle^{\otimes n}$, yields insights into the transient error rate e_R. The initialization fidelity itself is then defined as

$$F_I = 1 - e_R \tag{1}$$

where the readout error e_R is the average over the probability of obtaining the measurement outcome 1 when the qubit is prepared in the $|0\rangle$ state and the probability of obtaining the measurement outcome 0 when the qubit is prepared in the $|1\rangle$ state. In each case, the error is described by a Bernoulli process parameterized by a probability p with a variance $p(1-p)$. Similarly, gate fidelity F_G measures the accuracy with which a quantum operation transforms the register state. We rely on randomized benchmarking of the 2-qubit Clifford group to track the error behavior of the CNOT operation [30]. This technique measures survival probability following a sequence of randomly selected Clifford elements and fits the resulting sequence of fidelity to a linear model that eliminates the influence of state preparation and measurement errors [31]. From the fitted parameters, the error per Clifford gate ϵ_G defines the given gate fidelity as

$$F_G = 1 - \epsilon_G \tag{2}$$

The duty cycle τ is defined as the ratio of register coherence time to gate duration. Whereas gate duration represents the amount of time to complete the intended transform, the coherence time represents the timescale over which the encoded quantum information decoheres [32]. Using a simplified model of exponential decay that neglects other processes, decoherence of a single-qubit is characterized by a time-dependent decay with the characteristic time T_2. Hence, the duty cycle is defined as

$$\tau = T_2/T_G \tag{3}$$

with T_G the duration for gate G. In practice, we observe both T_2 and T_G fluctuate in time with variability in gate time T_G generally much slower. When gate time increases, the number of quantum operations reliably performed within the coherence time decreases and, thus, the depth of the circuit that can be reliably executed decreases.

Finally, the addressability F_A quantifies the ability to measure register elements individually. We define addressability in terms of pair-wise correlations that arise during measurement in the computational basis, and we quantify this metric using the normalized mutual information η. Given the entropy $H(X)$ of a discrete random variable X, and

joint entropy $H(X,Y)$ of random variables X and Y, we define the mutual information $I(X,Y)$ between two random variables X and Y as

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{4}$$

The normalized mutual information η is then obtained by normalizing the mutual information $I(X,Y)$ with the average entropy $H_{avg}(X,Y) = (H(X) + H(Y))/2$ such that

$$\eta(X,Y) = \frac{I(X,Y)}{H_{avg}(X,Y)} \tag{5}$$

This yields the addressability

$$F_A = 1 - \eta \tag{6}$$

which is dependent on the measured state. In particular, the addressability is unity for separable states and vanishes for maximally entangled states due to correlated outcomes.

We evaluate the stability of each metric presented using statistical tests to measure variations in the distributions observed for NISQ devices. Our test uses the well-known Hellinger distance

$$d_H(P_i, P_j) = \sqrt{1 - BC(P_i, P_j)} \tag{7}$$

as a similarity measure between the i-th and j-th distributions, P_i and P_j. For either discrete or continuous metric values, the Bhattacharyya coefficient

$$BC(P_i, P_j) = \begin{cases} \int_x \sqrt{p_i(x)p_j(x)}dx & \text{(continuous)} \\ \sum_{x \in X} \sqrt{p_i(x)p_j(x)} & \text{(discrete)} \end{cases} \tag{8}$$

is bounded between 0 and 1. The Hellinger distance itself vanishes for identical distributions and approaches unity for distributions with no overlap.

The Hellinger distance provides insight into the fluctuations between the device metrics as well as the underlying quantum states. For example, the Hellinger distance between the measurement outcomes of a quantum circuit provides a lower bound on the distance between the underlying quantum states ρ and σ [33]. These distances coincide when the optimally discriminating measurement is the computational basis, such that we may characterize the reproducibility of a program in terms of the Hellinger distance between computational outcomes. Our stability analysis assess the variations in each device metric to describe how the underlying noisy quantum programs fluctuate [34].

II. RESULTS

We evaluate the device metrics of Table 1 with respect to temporal stability and spatial stability. Temporal stability evaluates changes in the distance between observed metrics at different times, while spatial stability evaluates how metrics vary across locations in the quantum register. We characterize in detail the stability of two superconducting transmon devices produced by IBM called yorktown and toronto . With the layouts shown in Fig. 1, these NISQ devices are leading examples of programmable platforms used for experimental validation of quantum computing applications. A detailed historical record quantifying the calibration of the $n = 5$ yorktown is accessible using the Qiskit toolset [30]. We collected daily calibration metrics of yorktown from 1 March 2019 to 30 December 2020. We also collected a second data set from the $n = 27$ toronto device to characterize short-term and long-length scale stability using the addressability metric. The second data set was collected directly from toronto on 11 December 2020 during periods of time 8:00-8:30am, 11:00-11:30am, 2:00-2:30pm, 5:00-5:30pm, 8:00-8:30pm, and 11:00-11:30pm (UTC-5).

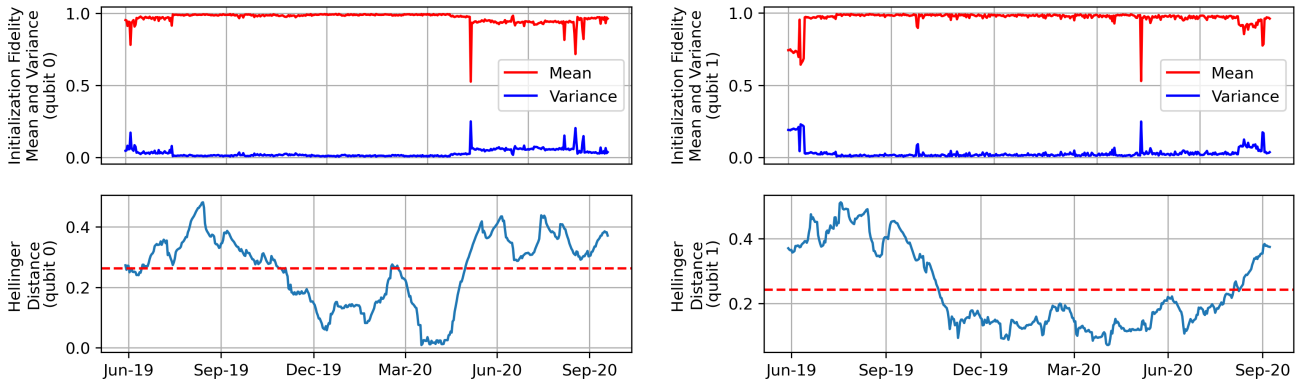
A. INITIALIZATION FIDELITY

We first evaluate temporal stability of the initialization fidelity for the yorktown device. Figures 2(a)-(e) shows the Hellinger distance over successive quarters for the initialization fidelity of all register elements from March 2019 to Dec 2020. The top panel of each figure shows the corresponding time series for F_I and its variance, while the bottom panel plots the corresponding Hellinger distance with the median distance for each element shown as a dashed red line to emphasize the aperiodic behavior. While there are long spans during which F_I itself is tightly controlled, the changes in the distance reveal the presence of large fluctuations in the underlying distributions. For example, the Hellinger distances shown in Figs. 2(a), (b), (d), and (e) demonstrate short-term stability near May 2020, but every element shows significant fluctuations in the distribution of the fidelity at all other times. Neither the value of the fidelity nor its variance track these instabilities in the device metric.

We also evaluate the spatial stability of the initialization fidelity. Here we use the $n = 27$ toronto device characterized at periodic intervals on a single day. Figure 2(f) plots the Hellinger distance for F_I between pairs of register elements, showing large differences between the distributions observed at different locations. While many distributions of the initialization fidelity within the toronto device are similar, others are well separated as indicated by the darker bands. The inset plots histograms for registers 0 and 11, which have a Hellinger distance of 1.0.

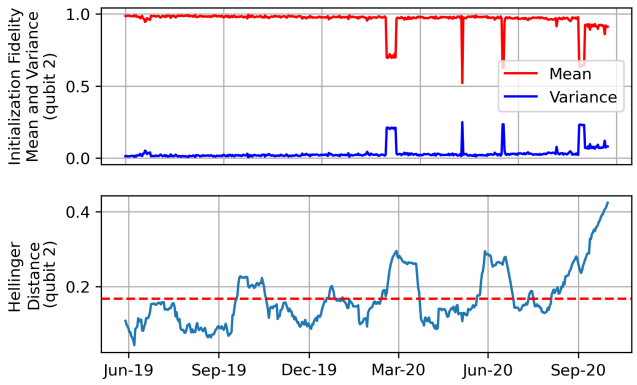
B. GATE FIDELITY

We next analyze stability of gate fidelity F_G for the CNOT gate in the yorktown device. The error per gate ϵ was retrieved from March 2019 to December 2020, and Figs. 3(a)-(e) plot the Hellinger distance for the CNOT fidelity between different register pairs. The stability of the fidelity is referenced to the initial distribution for F_G collected in March 2019, and the metric diverged sharply during the period June 2019 to December 2019. However, the metric F_G for all the register pairs fluctuated much less during the next 12 months. Whether this later behavior is sufficiently stable depends on the given application. Similar temporal analyses may be extended to all CNOT gates to generate a a detailed temporal map of the error model. However, it is clear from Figs. 3a-3e

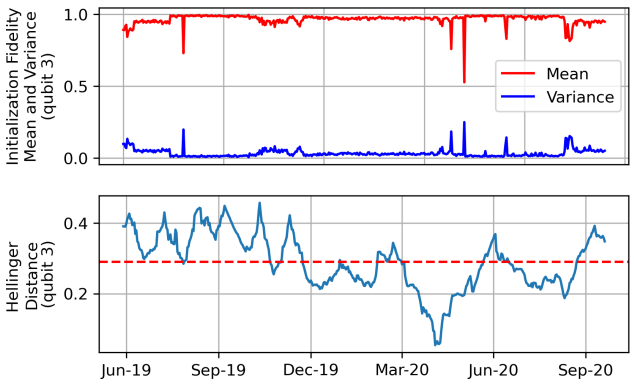


(a) Yorktown qubit 0

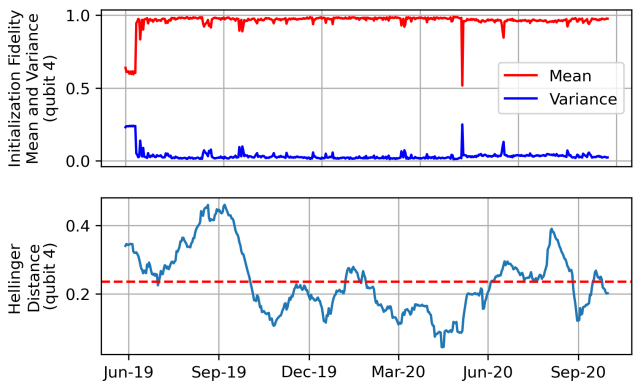
(b) Yorktown qubit 1



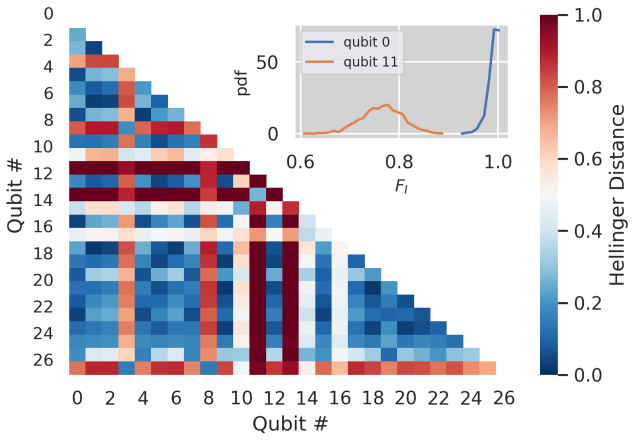
(c) Yorktown qubit 2



(d) Yorktown qubit 3



(e) Yorktown qubit 4



(f) Toronto (all qubits)

FIGURE 2: (a)-(e) Temporal stability of the initialization fidelity F_I of each register element in the yorktown device. The top panel shows the average F_I of the register with associated variance, and the bottom panel shows a running calculation of the Hellinger distance using a one-month window. The dashed red line is the median value. (f) Spatial stability of the initialization fidelity F_I for the toronto device sampled during 8:00-8:30 AM (UTC-5) on 11 December 2020. The heat map shows the Hellinger distance between histograms for each register pair. The inset shows the distributions of F_I for registers 0 and 11, which represent the largest distance observed of 1.0 due to minimal overlap.

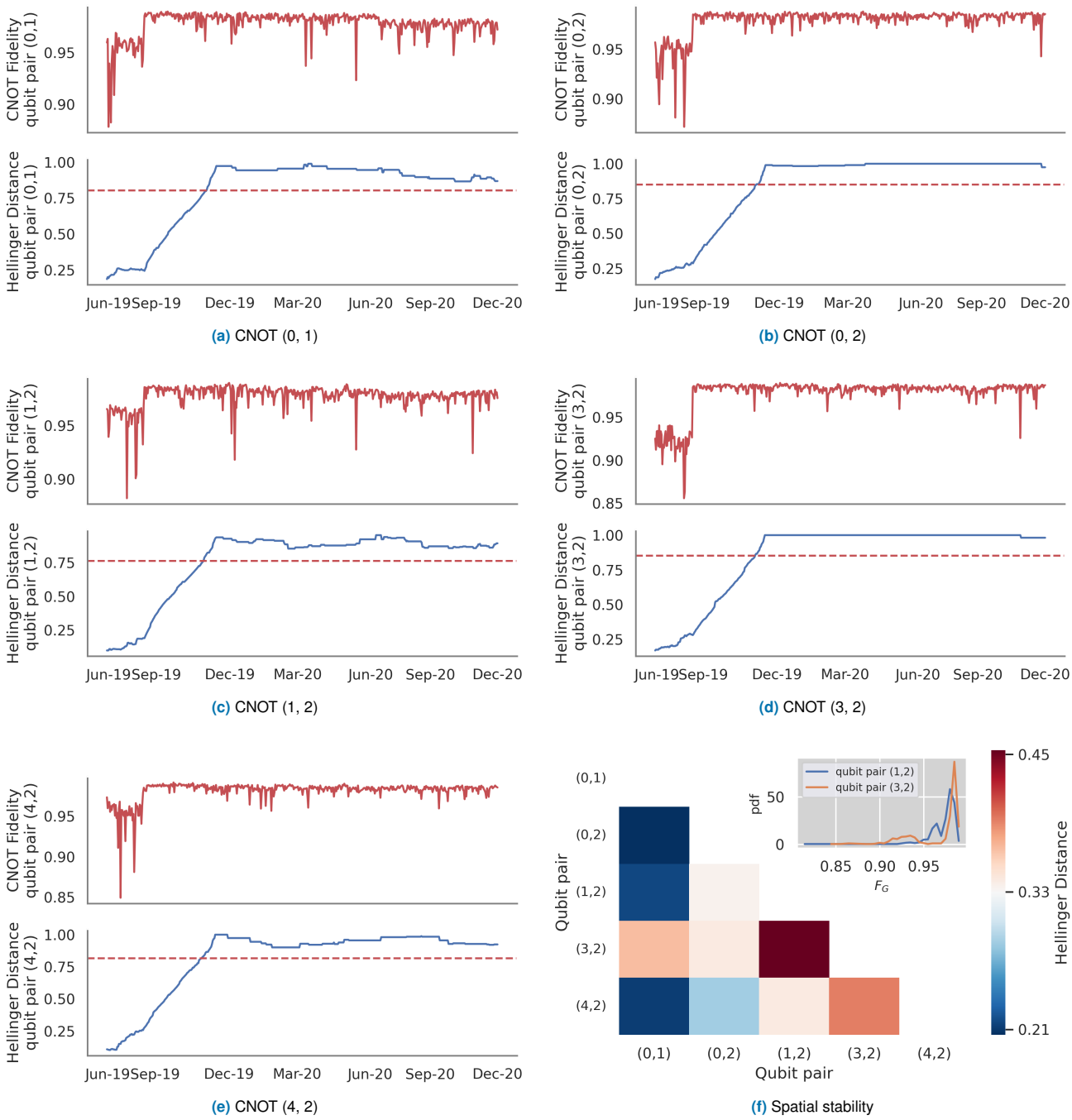


FIGURE 3: (a)-(e) Temporal stability of the gate fidelity F_G for the CNOT gate for sequential register pairs in the yorktown device from March 2019 to December 2020. The top panel shows the average F_G of the register pair and the bottom panel shows a running calculation of the Hellinger distance with respect to May 2019. The dashed red line is the median value. (f) Spatial stability of the gate fidelity F_G for the CNOT gates of the yorktown device from March 2019 to December 2020. The inset shows the distribution of gate fidelities for pairs (1,2) and (3,2), which yield a Hellinger distance 0.467.

that the noisy CNOT operations performed on March 2019 are different from those performed in September 2020.

The spatial stability of F_G for the yorktown device is also shown in Fig. 3f. The heatmap measures the distance between the fidelities describing pairs of CNOT operations. It corresponds to the metric distribution for the entire data period, though temporal subsets could be easily generated. The greatest distance was found between pairs (1,2) and (3,2), and the presence of register 3 generally correlated with larger distances. The inset shows the probability distributions for two register pairs, (1,2) and (3,2), which yielded a distance of 0.467. This comparison highlights that the distributions may be peaked near similar values and yet still starkly dissimilar.

C. DUTY CYCLE

Duty cycle monitors the ratio of coherence time to gate duration to yield a composite metric that estimates the number of operations that can be performed reliably. Figure 4a presents the time series of these different metrics for a CNOT operation in the yorktown device. We note that a CNOT operation within this device is composed from a sequence operations of native one- and two-qubit gates, and the reported duration is the complete duration of that sequence. The plots in Fig. 4(a)-(d) include the harmonic mean of the decoherence time T_2 for the different register pairs, the tunable duration of the CNOT gate, and the composed duty cycle metric. For register pair (0,1), for example, on approximately July 24, 2020, the T_2 time decreased sharply from 77 μs to 31 μs for register 0 and from 82 μs to 24 μs for register 1. A corresponding increase in the duration of the CNOT operation between registers 0 and 1, from approximately 370 ns to 441 ns, lead to an overall decrease but seemingly consistent value for the yorktown duty cycle. These changes in device parameters led to a sharp decrease in the duty cycle from 107.2 to 30.9 as seen in Fig. 4a.

Figure 4a also present the Hellinger distance for the duty cycle, calculated using histograms based on a running series of 3-month data. The dashed red line indicates the mean for the plotted series and highlights the fluctuations in the distance. However, it is notable that the distance decreases sharply (indicating stability of the metric) late in the series due to our use of a three-month moving average for the Hellinger distance.

D. ADDRESSABILITY

Addressability quantifies how well individual register measurements are differentiated, and we quantify these correlations between individual pairs of measurements using the normalized mutual information η and associated fidelity F_A introduced in Eq. (6). Figure 5 plots the addressability F_A of the toronto device when tested by encoding a fiducial separable state $|0,0\rangle$ in each register pair. Whereas ideal performance would yield $F_A = 1$ for each register pair, the heatmap highlights how the addressability F_A between register pairs varies across the entire 27-element register.

The inset compares the limits of this behavior by showing the lowest valued addressability of 0.8875 for register pair (23, 21) and the highest value of 0.9989 for register pair (11, 13). We highlight that the latter pair of registers also yields the largest distance between initialization fidelity shown in Fig. 2f.

A similar analysis of the addressability may be performed by first encoding a Bell-state within the register pair. As a maximally entangled state, the ideal addressabilty would $F_A = 0$ as the measurement outcomes should be strictly correlated. We limited this analysis to only nearest-neighbor register pairs as defined by the hardware connectivity, cf. Fig. 1b, in order to minimize the number of CNOT operations used in preparing the entangled state. The resulting addressability demonstrates similar variability across register pairs.

III. CONCLUSION

Fluctuations in device parameters represent a significant concern for the reproducibility of NISQ computing demonstrations. Many current experimental demonstrations rely on quantum circuits calibrated immediately prior to program execution and tuned during run-time. While this approach is successful for singular demonstrations, the resulting circuits and calibrations are implicitly dependent on the device parameters, which we have shown fluctuate significantly over time, space, and technology. Our results include analyses from two data sets. The first is collected over a period of 22 months, while the second is collected over half-hour intervals during a single day, which include data both near and far from device calibration points. Both analyses demonstrate clear fluctuations in the estimated metrics over the course of testing. Fluctuations on these key device metrics are specifically representative of the performance that users are experiencing, as our experimental data was taken on machines currently in operation.

The observed instability in these device metrics characterize experiments on today's NISQ devices in multiple ways. On time-scales of months and years, stability characterizes the ability to reproduce published experimental claims, while daily variations effect efforts to verify and validate individual experimental results. On the order of seconds and minutes, device instability modifies the underlying noise models, e.g., wide-sense stationary versus time-dependent processes, and impacts the confidence available from parameter estimation used for benchmarking quantum computing devices.

Our framework to evaluate the stability of NISQ devices monitors multiple metrics that connect to fundamental criteria for quantum computing ad shown in Table 1. These metrics may be continuously monitored by quantum computing device developers to characterize the stability of a device as well as to maintain the device within defined tolerances. We have shown how the Hellinger distance may be used to threshold when a device behaves in a sufficiently stable regime. Our results indicate clearly distinguishable differences in the Hellinger distance for several metrics and suggest the need for developing characterization protocols to

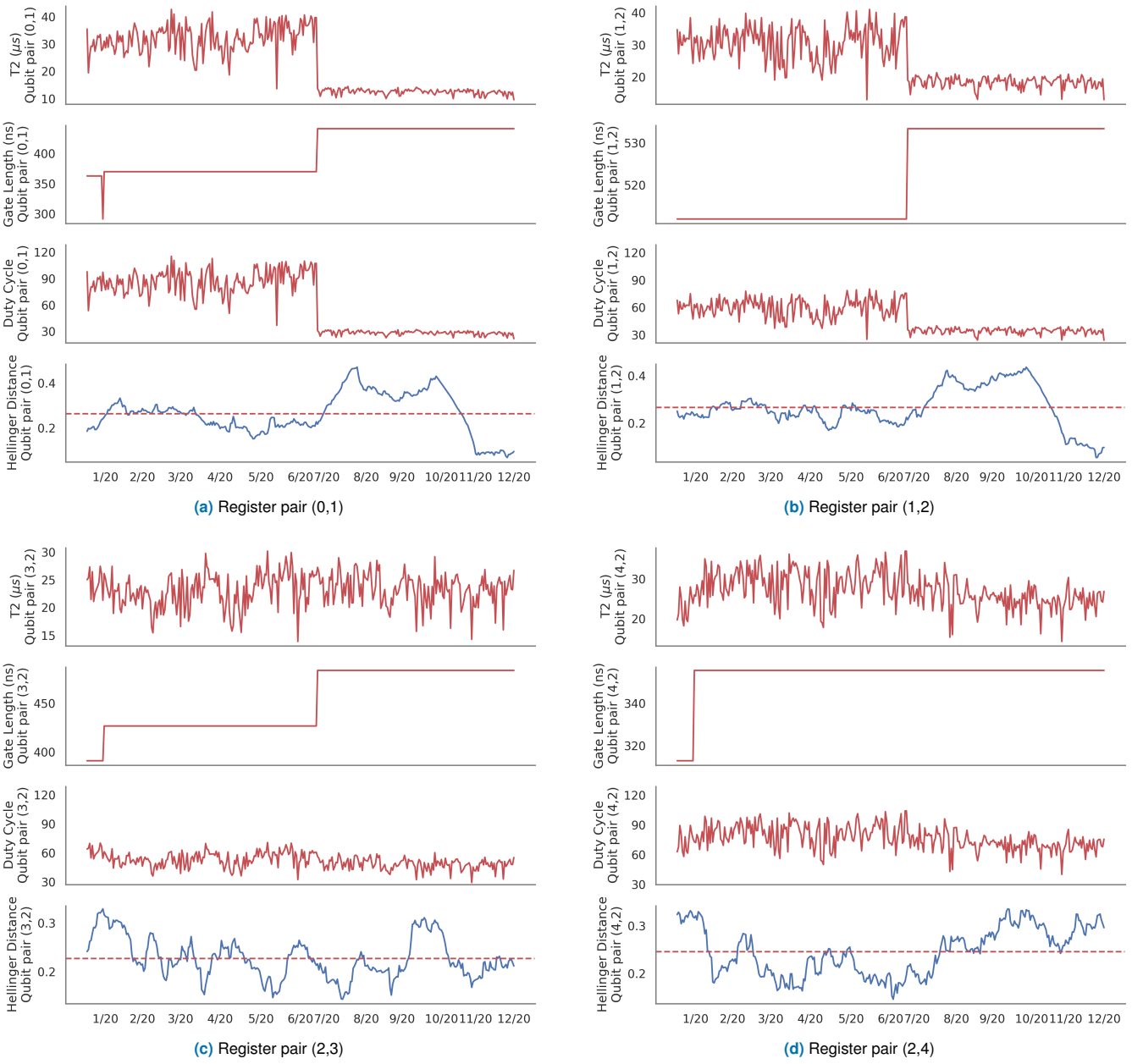


FIGURE 4: (a)-(d) Temporal stability of the CNOT duty cycle for sequential register pairs in the yorktown device. The top panel shows the harmonic mean of the register decoherence time T_2 for the elements, the upper-middle panel shows the gate duration T_G, the lower-middle panel plots the corresponding duty cycle τ, and the bottom panel presents the Hellinger distance for the duty cycle averaged over a one-month window. The dashed red line is the median value.

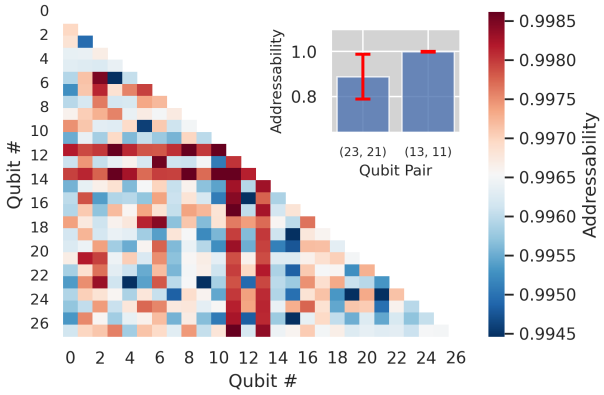


FIGURE 5: Addressability of register pairs in the toronto device sampled 08:00-08:30 AM (UTC-5) on 11 December 2020. This data corresponds to the register prepared in the separable fiducial state. The inset shows the range i.e. the lowest and highest values for addressability. The average of (23,21) is the lowest value at 0.887 while all other values lie in the range [0.992, 1). The outlier is the only value that does not appear in the plot.

provide stability analyses that aid in understanding the causes for reproducibility.

We have identified stability as a feature of fundamental importance to current testing and evaluation of NISQ devices. These analyses are readily extended to additional characteristics as well as metadata analysis of the calibration date, optimal pulse data, and other features about device operations. We are especially motivated by the long-term reproducibility of results from experimental quantum computer science, and in particular those using NISQ devices. Reproducibility hinges on the stable and reliable performance of the computing device. Without additional efforts to make current experimental results reproducible, the knowledge and insights gained from today's burgeoning field of quantum computer research may be undercut by low confidence in the reported results.

ACKNOWLEDGMENT

This work is supported by the Department of Energy (DOE) Office of Science, Early Career Research Program. This research used computing resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. SD thanks Peter Lockwood for support. A preliminary version of this work was published in the IEEE Quantum Computing Conference 2020 [35].

AUTHOR INFORMATION

These authors contributed equally: S. Dasgupta, T. S. Humble

REFERENCES

[1] Jonathan J Burnett, Andreas Bengtsson, Marco Scigliuzzo, David Niepce, Marina Kudra, Per Delsing, and Jonas Bylander. Decoherence benchmarking of superconducting qubits. npj Quantum Information, 5(1):1–8, 2019.

[2] Yong Wan, Daniel Kienzler, Stephen D Erickson, Karl H Mayer, Ting Rei Tan, Jenny J Wu, Hilma M Vasconcelos, Scott Glancy, Emanuel Knill, David J Wineland, et al. Quantum gate teleportation between separated qubits in a trapped-ion processor. Science, 364(6443):875–878, 2019.

[3] K. W. Chan, W. Huang, C. H. Yang, J. C. C. Hwang, B. Hensen, T. Tanttu, F. E. Hudson, K. M. Itoh, A. Laucht, A. Morello, and A. S. Dzurak. Assessment of a silicon quantum dot spin qubit environment via noise spectroscopy. Phys. Rev. Applied, 10:044017, Oct 2018.

[4] Travis S Humble, Himanshu Thapliyal, Edgard Munoz-Coreas, Fahd A Mohiyaddin, and Ryan S Bennink. Quantum computing circuits and devices. IEEE Design & Test, 36(3):69–94, 2019.

[5] Daniel Gottesman. Theory of fault-tolerant quantum computation. Physical Review A, 57(1):127, 1998.

[6] John Preskill. Quantum computing in the nisq* era and beyond. Bulletin of the American Physical Society, 64, 2019.

[7] Ye-Chao Liu, Jiangwei Shang, Xiao-Dong Yu, and Xiangdong Zhang. Efficient verification of quantum processes. Phys. Rev. A, 101:042315, Apr 2020.

[8] Robin Harper, Steven T Flammia, and Joel J Wallman. Efficient learning of quantum noise. Nature Physics, 16(12):1184–1188, 2020.

[9] Samuele Ferracin, Theodoros Kapourniotis, and Animesh Datta. Accrediting outputs of noisy intermediate-scale quantum computing devices. New Journal of Physics, 21(11):113038, 2019.

[10] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. Nature, 549(7671):242–246, 2017.

[11] Eugene F Dumitrescu, Alex J McCaskey, Gaute Hagen, Gustav R Jansen, Titus D Morris, T Papenbrock, Raphael C Pooser, David Jarvis Dean, and Pavel Lougovski. Cloud quantum computing of an atomic nucleus. Physical Review Letters, 120(21):210501, 2018.

[12] Cornelius Hempel, Christine Maier, Jonathan Romero, Jarrod McClean, Thomas Monz, Heng Shen, Petar Jurcevic, Ben P Lanyon, Peter Love, Ryan Babbush, et al. Quantum chemistry calculations on a trapped-ion quantum simulator. Physical Review X, 8(3):031022, 2018.

[13] Natalie Klco, Eugene F Dumitrescu, Alex J McCaskey, Titus D Morris, Raphael C Pooser, Mikel Sanz, Enrique Solano, Pavel Lougovski, and Martin J Savage. Quantum-classical computation of schwinger model dynamics using quantum computers. Physical Review A, 98(3):032331, 2018.

[14] Alexander J McCaskey, Zachary P Parks, Jacek Jakowski, Shirley V Moore, Titus D Morris, Travis S Humble, and Raphael C Pooser. Quantum chemistry as a benchmark for near-term quantum computers. npj Quantum Information, 5(1):1–8, 2019.

[15] Alessandro Roggero, Andy CY Li, Joseph Carlson, Rajan Gupta, and Gabriel N Perdue. Quantum computing for neutrino-nucleus scattering. Physical Review D, 101(7):074038, 2020.

[16] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. Nature, 574(7779):505–510, 2019.

[17] Google AI Quantum et al. Hartree-fock on a superconducting qubit quantum computer. Science, 369(6507):1084–1089, 2020.

[18] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, et al. Quantum computational advantage using photons. Science, 370(6523):1460–1463, 2020.

[19] John M Martinis. Qubit metrology for building a fault-tolerant quantum computer. npj Quantum Information, 1(1):1–3, 2015.

[20] Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. Nature Reviews Physics, 2(7):382–390, 2020.

[21] Robin Blume-Kohout and Kevin C. Young. A volumetric framework for quantum computer benchmarks. Quantum, 4:362, November 2020.

[22] Megan L Dahlhauser and Travis S Humble. Modeling noisy quantum circuits using experimental characterization. Physical Review A, 103(4):042603, 2021.

[23] Timothy Proctor, Melissa Revelle, Erik Nielsen, Kenneth Rudinger, Daniel Lobser, Peter Maunz, Robin Blume-Kohout, and Kevin Young. Detecting and tracking drift in quantum information processors. Nature communications, 11(1):1–9, 2020.

[24] Kristan Temme, Sergey Bravyi, and Jay M Gambetta. Error mitigation for short-depth quantum circuits. Physical Review Letters, 119(18):180509, 2017.
[25] Abhinav Kandala, Kristan Temme, Antonio D Córcoles, Antonio Mezzacapo, Jerry M Chow, and Jay M Gambetta. Error mitigation extends the computational reach of a noisy quantum processor. Nature, 567(7749):491–495, 2019.
[26] Michael R Geller. Rigorous measurement error correction. Quantum Science and Technology, 5(3):03LT01, 2020.
[27] Ellis Wilson, Sudhakar Singh, and Frank Mueller. Just-in-time quantum circuit transpilation reduces noise. In 2020 IEEE International Conference on Quantum Computing and Engineering (QCE), pages 345–355. IEEE, 2020.
[28] Kathleen E Hamilton and Raphael C Pooser. Error-mitigated data-driven circuit learning on noisy quantum hardware. Quantum Machine Intelligence, 2(1):1–15, 2020.
[29] David P DiVincenzo. The physical implementation of quantum computation. Fortschritte der Physik: Progress of Physics, 48(9-11):771–783, 2000.
[30] Gadi Aleksandrowicz, Thomas Alexander, Panagiotis Barkoutsos, Luciano Bello, Yael Ben-Haim, D Bucher, FJ Cabrera-Hernández, J Carballo-Franquis, A Chen, CF Chen, et al. Qiskit: An open-source framework for quantum computing. Accessed on: Mar, 16, 2019.
[31] Easwar Magesan, Jay M Gambetta, and Joseph Emerson. Scalable and robust randomized benchmarking of quantum processes. Physical Review Letters, 106(18):180504, 2011.
[32] J. Gambetta. Noise Amplification Squeezes More Computational Accuracy From TodayâĂŹs Quantum Processors, 2019 (accessed February 4, 2020).
[33] Marcin Jarzyna and Jan Kołodyński. Geometric approach to quantum statistical inference. IEEE Journal on Selected Areas in Information Theory, 1(2):367–386, 2020.
[34] Samudra Dasgupta and Travis S. Humble. Reproducibility in quantum computing. In 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pages 458–461, 2021.
[35] Samudra Dasgupta and Travis S Humble. Characterizing the stability of nisq devices. In 2020 IEEE International Conference on Quantum Computing and Engineering (QCE), pages 419–429. IEEE, 2020.
[36] https://www.ibm.com/quantum-computing/experience/, accessed Dec 2019 - Feb 2020.

B. APPENDIX

This appendix includes details on the data collection methods, the analysis methods and supplementary results from this analysis.

A. DATA COLLECTION

For the $n = 5$ yorktown device, we use calibration data from the publicly available data set provided by the IBM Quantum Experience [36]. This data set includes daily calibration data collected for the yorktown device over a period of 22 months from March 2019 to December 2020. This data set includes a total of 673 calibration records. The first record has a time stamp of 2019-02-27 06:56:32-05:00 and the last record has a time stamp of 2020-12-30 01:18:20-05:00. We list the available fields in each record below. We note a change in the records that occurred in September 2019 when the gate duration data became available. This yielded 445 records with gate duration data. Otherwise, each calibration record contains the following data identified by the indicated fields:

1) Date and time when the device properties were last updated as last_update_date
2) Register readout error estimates as, e.g. q0_readout_err
3) Date and time when a register readout was calibrated as, e.g. q0_readout_err_cal_time
4) CNOT gate error estimates as, e.g. cx01_gate_err
5) Date and time when a CNOT gate error measurement was calibrated as, e.g. cx01_gate_err_cal_time
6) Register decoherence time estimate, the unit of measure, and the date and time of the characterization as, e.g. q0_T2, q0_T2_unit, and q0_T2_cal_time
7) CNOT gate length and the unit of measure as, e.g. cx01_gate_length, cx01_gate_length_unit
8) Date and time when a CNOT gate length was calibrated as, e.g. cx01_gate_length_cal_time

For the $n = 27$ toronto device, we generated a second data set on 11 December 2020 that spans the periods of time 8:00-8:30am, 11:00-11:30am, 2:00-2:30pm, 5:00-5:30pm, 8:00-8:30pm, and 11:00-11:30pm (UTC-5). The collected data included the readout of all $n = 27$ register elements when the initial state is prepared to be in the all-zeros computational basis state. The total number of observations in this data set was $5,750,784$ register results. Each register element was sampled for $212,992$ sequential data points. We divided this temporal set into contiguous sets of 1000 data points each, e.g., the first set contains the first 1000 observations for each of the 27 registers as a 1000×27 binary matrix. Set 1 contains the next set of 1000 observations for each of the 27 qubits - another 1000×27 binary matrix, etc. For each data set, we calculated the addressability F_A for each of the possible 351 register pairs and each pair was characterized by 211 observations for F_A.

A similarly structured data set was constructed from the readout of all $n = 27$ register elements when the initial state of register pairs are prepared in the Bell-state triplet state. As above, we collected data from toronto following a canonical Bell-state preparation circuits applied to nearest-neighbor register locations. There are 28 such nearest-neighbor pairs in the toronto layout. We characterized the measurement outcomes of a Bell-state preparation circuit for each of these 28 pairs on December 11, 2020 between 11pm-11:30pm. Each experiment was recorded with 8192 observations.

We generated a third data set composed from data collected across a family of different IBM devices. These devices are labeled as yorktown, bogota, rochester, paris, and athens. The calibration records collected from the IBM Quantum Experience were used to construct a data set for each device consisting of the update time, the readout errors, the readout calibration time, the CNOT gate error rates and the CNOT gate calibration time. This was collected over the period of 27 February 2019 to 31 December 2020.

B. STABILITY ANALYSIS

For analysing device stability, we use the Hellinger distance between empirical distributions.

We monitor temporal stability by tracking temporal changes in the Hellinger distance between distributions at different times as

$$\mathcal{H}(t) = d_H(\mathcal{F}_X(t-\tau), \mathcal{F}_X(t)) \tag{9}$$

where $\mathcal{F}_X(t)$ is the distribution of metric X at time t and τ is a reference duration. For temporal stability analysis, the successive histograms $\mathcal{H}(t_n)$ and $\mathcal{H}(t_{n+1})$ are spaced 1 month apart to reveal changes over a moving 1-month window. We chose one-month as the separation between two successive histograms when analysing temporal stability to ensure that we are comparing the device characteristics after a sufficient amount of time has elapsed. However, this is a matter of design choice and can be customized to user needs. Each histogram bins the data collected over a 3-month period, such that there are 90 daily data values in each histogram. We chose 3 months ($\approx$ 90 data points) to ensure that the histogram has sufficient number of data points to help us deduce a distribution function. If the number of data points is too small then the distribution function will be unreliable. If too high then the time-variation of the distribution will be suppressed. For the temporal stability analysis of the gate fidelity, we track $\mathcal{H}(t)$ with respect to a reference time t_0 such that $\tau = t - t_0$ to reveal variations relative to when an experiment was first conducted (t_0).

We monitor spatial stability by tracking spatial changes in the Hellinger distance between distributions at different register locations as

$$\mathcal{H}(i,j) = d_H(\mathcal{F}_X(i), \mathcal{F}_X(j)) \tag{10}$$

where $\mathcal{F}_X(i)$ is the corresponding distribution of metric X at register i over all available temporal data. When calculating the distance between histograms of registers at different layout locations, the full time series available was always used.

Similarly, we monitor the stability of a metric between different devices by using register locations i and j that

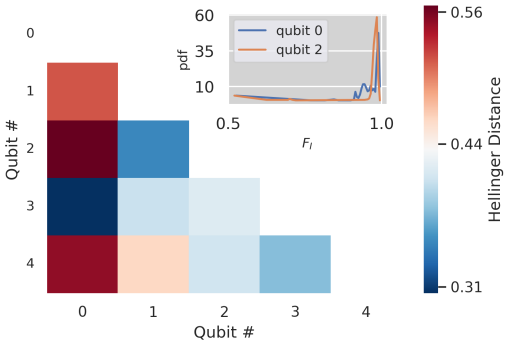


FIGURE A1: Spatial stability of the initialization fidelity for the yorktown device from May 2019 to December 2020, where the inset highlights the registers with the maximum distance.

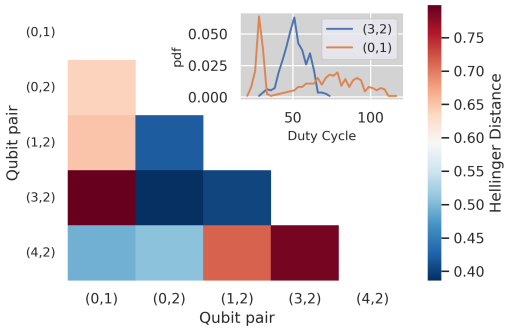


FIGURE A2: The spatial stability of the duty cycle τ for yorktown. The inset shows the experimental histograms for register pairs (0,1) and (2,3) which are separated by the largest Helligner distance of 0.789.

correspond to locations in different devices. For inter-device stability analysis, the full time series available was always used.

C. SUPPLEMENTAL FIGURES

Figure A1 plots the spatial stability analysis of initialize fidelity for the $n = 5$ yorktown device from the first data set. The data demonstrates variation in the distance between register pairs that range from approximately 0.35 to as larger as 0.55. The inset shows the distributions of F_I for the largest distance observed.

Figure A2 plots the spatial stability of the duty cycle for the $n = 5$ yorktown device from the first data set. The data demonstrates variation in the distance between register pairs that range from approximately 0.40 to as larger as 0.789. The inset shows the distributions of the duty cycle for the largest distance observed. Distributions for register pairs (0,1) and (3,2) showed a large distance of 0.789. The spatial stability pattern for the duty cycle does not match the pattern observed for the initialization fidelity in Fig. 3f, hinting at more different sources of errors for these transient behaviors.

Figure A3 plots the spatial stability of gate fidelity varies across the five devices from the third data set. The figure plots a pair-wise heatmap based on the Hellinger distance between gate fidelity distributions for these devices. The inset shows an example of the underlying distributions between the rochester and athens devices. These two devices demonstrate significant differences in distribution as indicated by a Hellinger distance of 1.0 (no overlap).

Figure A4 plots the normalized mutual information (NMI) for Bell-state encoded states. There only 28 register pairs that support direct preparation of a Bell state and the figure plots the NMI for those pairs that support these states. For this state, strictly correlated measurement outcomes are expected and the NMI should be maximal. The inset shows an example of the range of the observed NMI between the Bell pairs. We find the NMI between registers 12 and 15 was

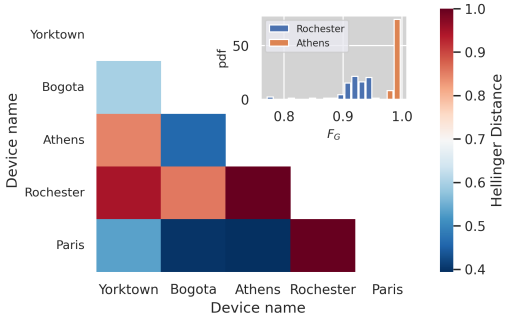


FIGURE A3: Inter-device stability of the gate fidelity F_G for the CNOT gate for a family of five devices for the time period Mar 2019 - Dec 2020. The inset shows the distribution of F_G for the CNOT gate from the athens and rochester devices which have the highest corresponding Hellinger distance of nearly 1.

least with a mean of $\eta = 0.14 \pm 0.014$, while the largest NMI was observed for registers 25 and 26 with a mean of $\eta = 0.84 \pm 0.023$.

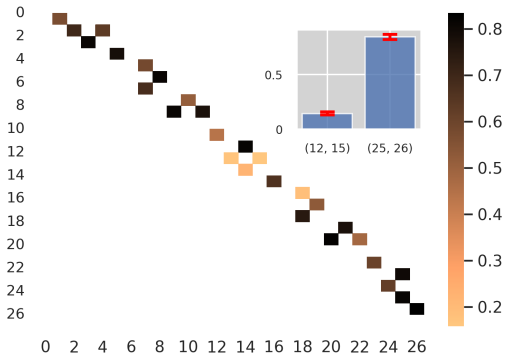


FIGURE A4: Normalized mutual information of register pairs in the toronto device sampled 11:00-11:30 PM (UTC-5) on 11 December 2020. Data corresponds to register pairs prepared in the Bell state and the inset shows the range of the lowest and highest values for the normalized mutual information.