

PAPER • OPEN ACCESS

Demonstration of quantum volume 64 on a superconducting quantum computing system

To cite this article: Petar Jurcevic *et al* 2021 *Quantum Sci. Technol.* **6** 025020

View the [article online](#) for updates and enhancements.

You may also like

- [Inclusion of quasi-vertex views in a brain-dedicated multi-pinhole SPECT system for improved imaging performance](#)

Benjamin Auer, Navid Zeraatkar, Justin C Goding et al.

- [Electrical equivalent model of a long dielectric barrier discharge plasma jet for endoscopy](#)

Orianne Bastin, Max Thulliez, Teo Serra et al.

- [THE UBV COLOR EVOLUTION OF CLASSICAL NOVAE. II. COLOR-MAGNITUDE DIAGRAM](#)

Izumi Hachisu and Mariko Kato



Easy-to-use and Helium-3 free
cryogenics solutions

LEARN MORE

Quantum Science and Technology



PAPER

OPEN ACCESS

RECEIVED
4 September 2020

REVISED
25 January 2021

ACCEPTED FOR PUBLICATION
10 February 2021

PUBLISHED
17 March 2021

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the
title of the work, journal
citation and DOI.



Demonstration of quantum volume 64 on a superconducting quantum computing system

Petar Jurcevic¹ , Ali Javadi-Abhari¹, Lev S Bishop¹, Isaac Lauer¹, Daniela F Bogorin¹, Markus Brink¹, Lauren Capelluto¹, Oktay Günlük¹, Toshinari Itoko², Naoki Kanazawa², Abhinav Kandala¹, George A Keefe¹, Kevin Krsulich¹, William Landers¹, Eric P Lewandowski¹, Douglas T McClure¹, Giacomo Nannicini¹, Adinath Narasgond¹, Hasan M Nayfeh¹, Emily Pritchett¹, Mary Beth Rothwell¹, Srikanth Srinivasan¹, Neereja Sundaresan¹, Cindy Wang¹, Ken X Wei¹, Christopher J Wood¹ , Jeng-Bang Yau¹, Eric J Zhang¹ , Oliver E Dial¹, Jerry M Chow¹ and Jay M Gambetta¹

¹ IBM Quantum, IBM T J Watson Research Center, Yorktown Heights, NY 10598, United States of America

² IBM Quantum, IBM Research Tokyo, 19-21 Nihonbashi Hakozaiki-cho, Chuo-ku, Tokyo, 103-8510, Japan

E-mail: petar.jurcevic@ibm.com

Keywords: quantum volume, transmon qubits, dynamical decoupling, quantum circuit compilation, two qubit gates, random quantum circuits

Abstract

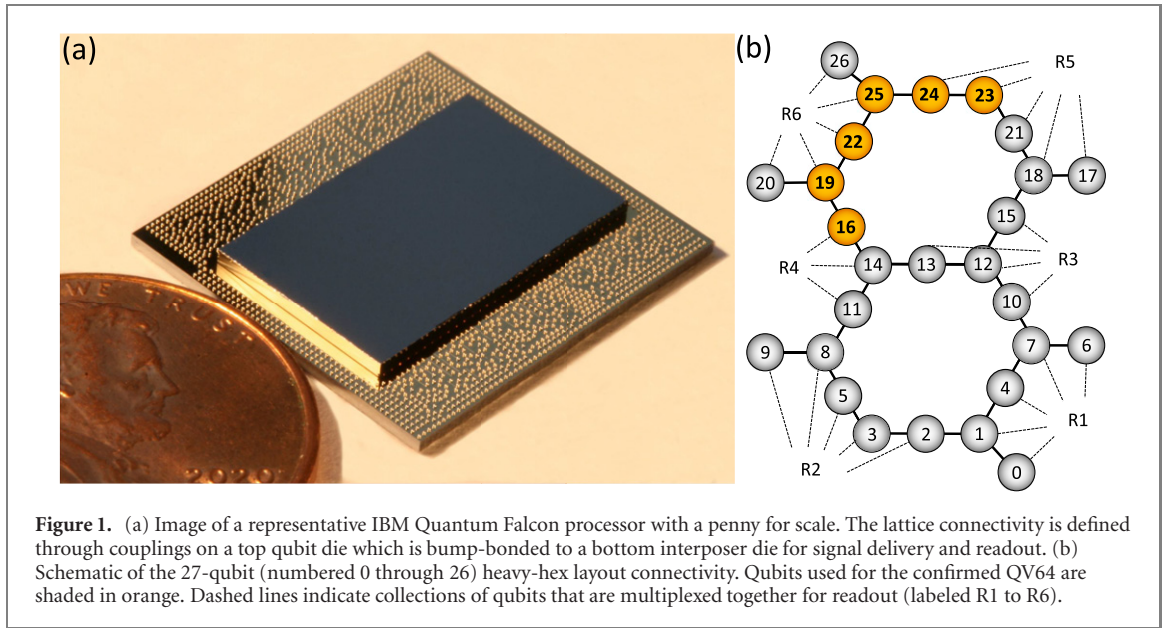
We improve the quality of quantum circuits on superconducting quantum computing systems, as measured by the quantum volume (QV), with a combination of dynamical decoupling, compiler optimizations, shorter two-qubit gates, and excited state promoted readout. This result shows that the path to larger QV systems requires the simultaneous increase of coherence, control gate fidelities, measurement fidelities, and smarter software which takes into account hardware details, thereby demonstrating the need to continue to co-design the software and hardware stack for the foreseeable future.

1. Introduction

Quantum computing is a new kind of computing, using the same physical rules that atoms follow in order to manipulate information. At this fundamental level, quantum computers execute quantum circuits—like a classical computer's logical circuits—but now using the physical phenomena of superposition, entanglement, and interference to implement mathematical calculations that are out of reach for even our most advanced supercomputers.

As we progress toward machines capable of implementing circuits with a quantum advantage, meaning certain information processing tasks can be performed more efficiently or cost effectively than with classical circuits, quantum volume (QV) [1] serves as a holistic benchmark for quantum systems indicating the size of the quantum circuits that can be run on them. Sensitive to improvements in many aspects of device performance, QV includes gate errors, measurement errors, the quality of the circuit compiler, and spectator errors exemplifying the importance of co-design between software and hardware [2–4]. In reference [1] and later in reference [5], QV16 was measured on *ibmq_johannesburg* and a Honeywell quantum system, respectively. In reference [6] QV8 was measured for the Rigetti Aspen-4 quantum system. We recently increased *ibmq_johannesburg* to QV32 [7] by improving our physical understanding of the two-qubit cross-resonance gate and using rotary echo pulses to reduce gate and spectator errors. Finally in unpublished work Honeywell has claimed to measure QV64 [8].

Here we demonstrate an increase in the QV of an IBM quantum system by improving the Qiskit compiler [9], implementing excited state promoted (ESP) readout, shorter two-qubit gates, and adding dynamic decoupling to the idle qubits. These last three demonstrate the need for timing and pulse control in cloud quantum systems [10]. While individually not one of these improvements is enough to allow *ibmq_montreal* to reach QV64, when combined we achieve QV64 with a heavy output probability (HOP) of



$0.701 \pm 0.031 (> 2/3 \pm 2\sigma)$ with a confidence interval of 98.744% ($z = 2.25$), see figure 2(a). The QV test requires exceeding $2/3$ HOP by a 97.725% ($z = 2$) confidence interval.

In section 2 we give an overview of the *ibmq_montreal* device, which is a 27-qubit IBM Quantum Falcon processor; in section 3 we discuss the improvements to the compiler; in section 4 we discuss the dynamical decoupling protocol; in section 5 we discuss the faster implementation of the direct CNOT gate which extends the improved pulse control of [7]; in section 6 we discuss the improvement in measurement fidelity by using a control pulse to promote the excited state to a higher level before measurement [11]. Finally in section 7 we conclude the paper.

2. Quantum system—*ibmq_montreal*

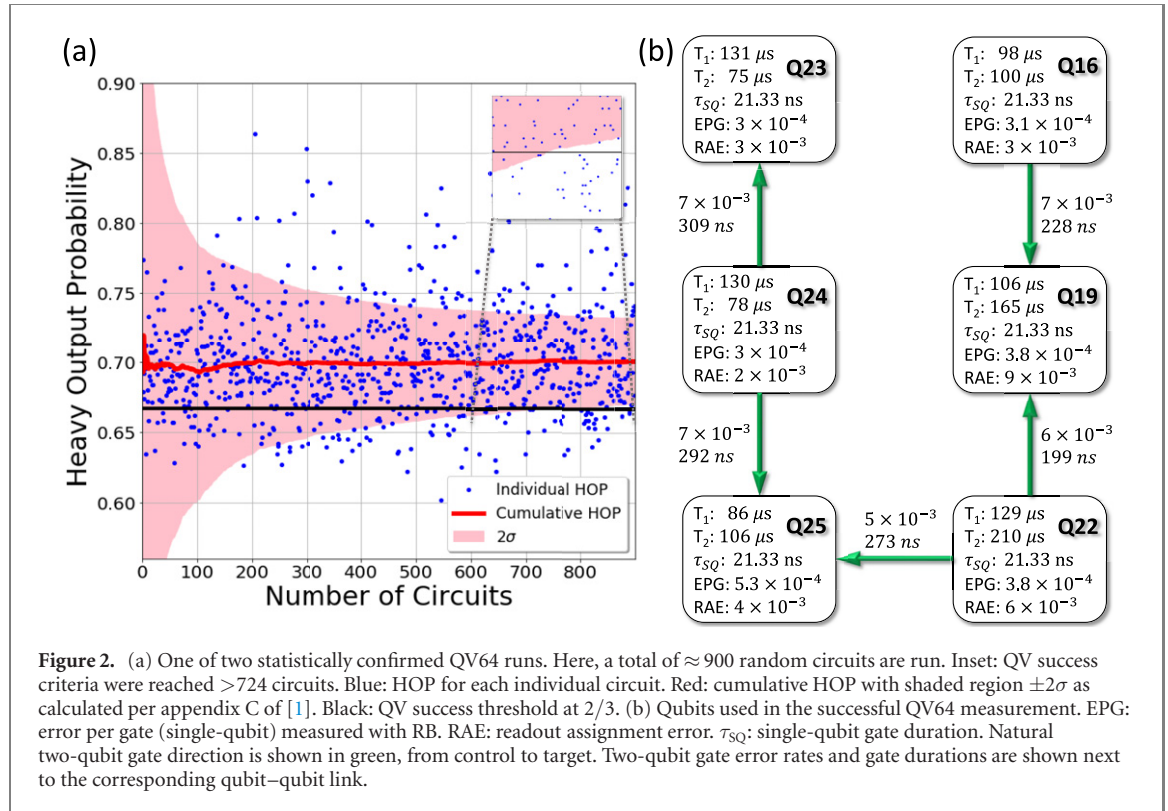
The device studied in this work is from the recent series of IBM Quantum Falcon processors, which consist of 27 qubits arranged in a lattice designed for a distance-3 hybrid Bacon-Shor-surface code [12]. A photo of this processor is shown in figure 1(a), and a schematic of its connectivity is shown in figure 1(b). A high connectivity layout, such as ‘all-to-all’, is preferable for random quantum circuits (such as QV circuits) in order to minimize the average qubit–qubit distance; however, additional edges in the connectivity increase the chance of frequency collision, cross-talk, and spectator errors. The IBM Quantum Falcon processor is a compromise, preserving a connectivity efficient for a logical qubit while simultaneously reducing detrimental effects of collisions and cross-talk without excessive insertion of SWAP gates to emulate ‘all-to-all’ connectivity. In these and related systems, by using the techniques described in [7], we have measured a QV of 32 on the last 7 deployed systems [13] demonstrating the reliability of this architecture.

In this paper we achieve QV64 on *ibmq_montreal*, which is one of the latest deployed IBM Quantum Falcon processors. The QV circuits were run on a line of six qubits, Q16–Q19–Q22–Q25–Q24–Q23 (orange shaded qubits in figure 1(b)). Individual qubit properties are shown in figure 2(b) with the following average values: $T_1 = 113 \mu\text{s}$, $T_2 = 122 \mu\text{s}$, error per single-qubit gate 3.8×10^{-4} , error per two-qubit gate 6.4×10^{-3} , and single-qubit readout assignment error 6.0×10^{-3} . Gate errors were measured with simultaneous single-qubit and individual two-qubit randomized benchmarking [14].

The qubits are fixed-frequency transmons with frequencies ≈ 5 GHz. Single-qubit gates are driven resonantly with a microwave pulse of duration $\tau_{\text{sq}} = 21.33$ ns. A DRAG pulse envelope [15] corrects σ_z -errors and signal dispersion due to wiring. Two-qubit gates are based on a cross-resonance scheme [16–18] with a target rotary pulse [7] and an additional offset pulse-shape on the target for implementing a direct (echoless) CNOT as described later. Two-qubit gate lengths are $\tau_{\text{tq}} = 199$ –309 ns.

3. Compiler

Circuit compilation is a substantial part of quantum computation. Here we report improvements in the state-of-the-art Qiskit compiler to achieve reductions in the number of gates which results in circuits with shorter depths. The compilation of a QV circuit for a superconducting processor can be roughly broken



down into two stages. The first stage is to map the circuit to the hardware's qubit connectivity constraints. At the conclusion of this step, each circuit will consist of a series of $SU(4)$ gates on the available links, as well as the overhead of routing qubit information on the physical fabric, usually in the form of SWAPs. The second step consists of local expansions to the native gates of the hardware and optimizations. We introduce new compiler passes to improve both stages, and leverage existing passes in the Qiskit compiler throughout to achieve further reductions where possible: approximate synthesis, commutative cancellation, and peephole optimization of single-qubit and two-qubit chains of gates.

It is worth noting that the particular passes reported here have general utility beyond QV. Qubit mapping and routing is ubiquitous in compiling for limited-connectivity architectures, and $SU(4)$ synthesis has broad use in peephole optimization of sequential two-qubit gates.

3.1. Qubit layout and routing via binary integer programming (BIP)

We formulate qubit layout and routing as a BIP problem, which we are able to solve to optimality. We choose as the cost function, C , the effective fidelity, modeled as the product of the fidelity of all the implemented gates:

$$C = K^d \prod_{j \in G} F_j^{\text{best}} \prod_{j \in \bar{G}} \bar{F}_j^{\text{best}} \prod_{j \in S} F_b^3, \quad (1)$$

where K is a factor penalizing circuits with high depth d ; G ($\bar{G}$) are the set of gates that are mapped directly (mapped with mirroring—combining SWAP with a gate); and S is the set of added SWAP gates. Here, F_b is the gate fidelity of the available entangling gate (which must be applied 3 times to implement SWAP), F_j^{best} ($\bar{F}_j^{\text{best}}$) is the modeled fidelity of the best approximation to the target unitary making $i = 0, \dots, 3$ uses of the entangling gate

$$F_j^{\text{best}} = \max_i F_{ij}^{\text{avg}} (F_b)^i, \quad (2)$$

$$\bar{F}_j^{\text{best}} = \max_i \bar{F}_{ij}^{\text{avg}} (F_b)^i, \quad (3)$$

and F_{ij}^{avg} is the average gate fidelity due to approximating the j th gate with i uses of the entangling gate [1, appendix B].

The freedom to implement either a gate or its mirror allows elimination of many explicit SWAP gates, and by restricting the number of candidate SWAP insertion sites we are able to reduce the size of the BIP problem such that it can be solved to optimality in around one second per circuit, using optimization software such as CPLEX [19]. Figure 3 shows the performance of this BIP pass in comparison to the

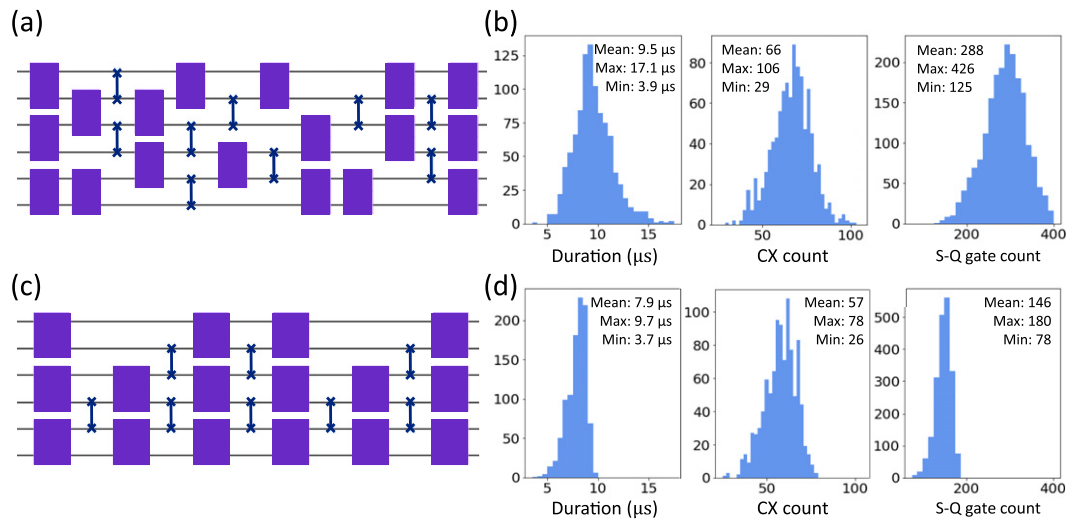


Figure 3. Comparison of QV64 circuits transpiled to a line connectivity with (a) and (b) the state-of-the-art Qiskit compiler and (c) and (d) an improved transpilation method based on BIP and additional gate cancellations, see text. (a) and (c) show the same random example circuit, mapped with ‘SABRE’ and mapped with BIP, respectively. The purple boxes represent the random SU(4) and SWAPs are indicated in blue. (b) and (d) show statistics of 2000 circuits using both methods. CX count improvements are due to improved mapping, and S-Q (single-qubit) count improvements are the result of pulse-efficient compilation. Both contribute to shorter durations. We assume basis gate fidelity $F_b = 0.99$ for the approximate SU(4) expansion in all cases. If the native gate is ECR (rather than direct-CX), we get additional 7% reduction in mean duration by targeting the native gate and absorbing local pre-rotations.

state-of-the art SABRE algorithm [20] available in Qiskit, showing substantial improvement in both the mean and maximum number of uses of the entangling gate.

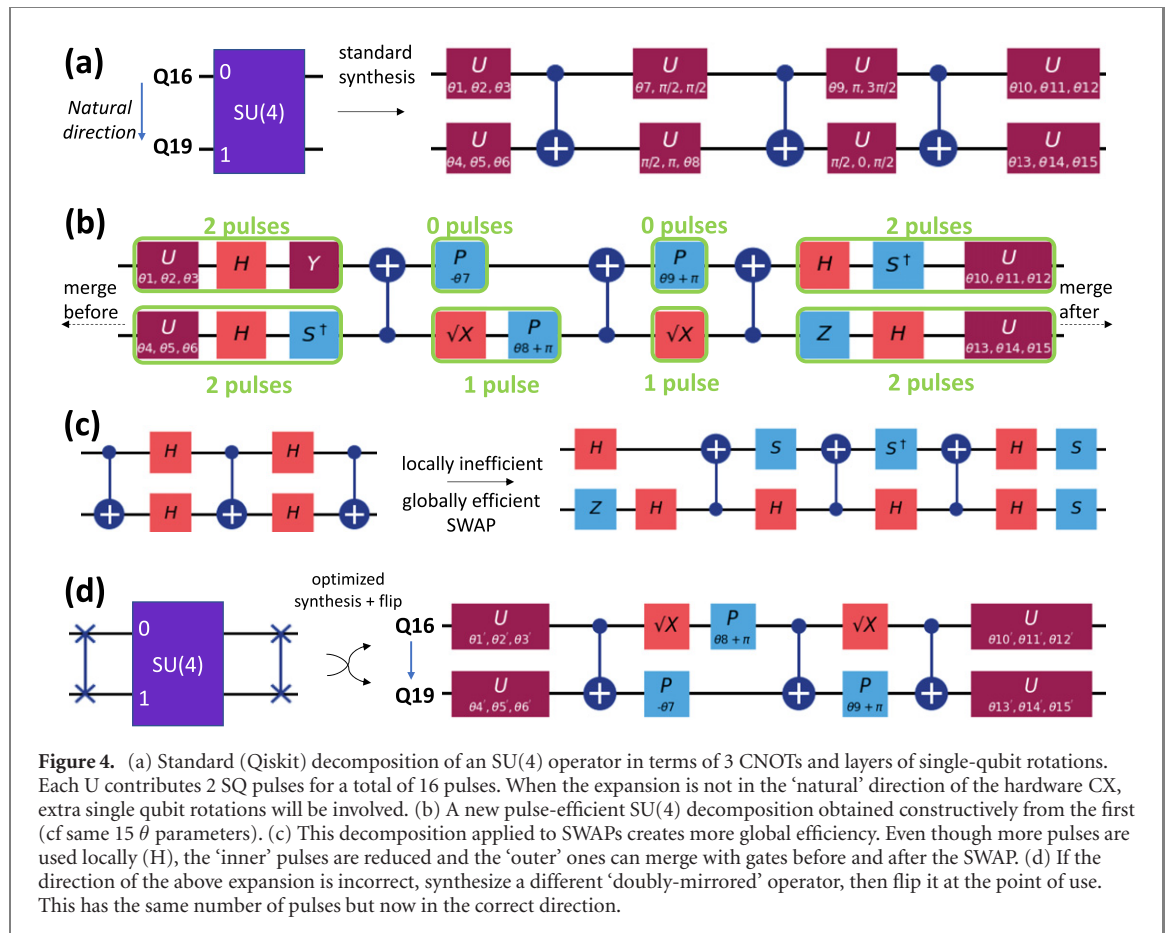
3.2. Pulse-efficient SU(4) decomposition

The *ibmq_montreal* device has the following native gate set for achieving universal quantum computation: Ctrl-X (CX), Sqrt-X (SX) and Phase(θ). The CX gate itself can be implemented directly or be created using an echo cross-resonance (ECR) pulse [21] (cf section 5). The Phase gate can be achieved with zero time and error [22]. We refer to any gate that is one pulse (i.e. equivalent to an SX by a pre-/post-phase) as a single-qubit (SQ) gate (e.g. Hadamard). A generic single-qubit operation (U) can be achieved with at most 2 SQ pulses.

Given the CX, SQ or ECR, SQ set of native pulses, we aim to minimize them during the expansion of each SU(4) and SWAP. A second goal is to expand them in a way that creates further opportunities for optimization. It is known that any SU(4) can be implemented using at most 3 CX gates [23], and 2 CX gates suffice for many useful approximations (e.g. at 99% fidelity) [1] (cf Figure 4(a)). ECR is locally equivalent to CX, so it has the same requirements. While the question of ‘optimal’ SU(4) decomposition has been extensively studied, the optimality criteria has usually been the number of two-qubit gates [23, 24]. To extract ultimate performance, we are also interested in minimizing the total number of pulses and the duration.

Our approach is based on three strategies:

- Circuit simplification to reduce redundant pulses: starting from a Qiskit synthesis of an arbitrary SU(4), we apply repeated circuit identities to the result to reduce its cost. This gives us a constructive SU(4) decomposition, depicted in figure 4(b), which is optimal in the number of pulses (by a simple parameter counting argument). This decomposition has another advantage, in that 8 out of 10 single-qubit pulses are placed on the outside of the structure. Given that 2 SQ pulses suffice for any aggregate single-qubit operation, this creates an opportunity for merging with preceding and following layers of SU(4) in the circuit. One surprising consequence of this decomposition is that for the special case of an SWAP operation, the decomposition is locally less efficient than a textbook expansion; however globally it is more efficient as it creates more opportunities for cancellation (figure 4(c)). We arrive at similar pulse-efficient decompositions targeting the ECR gate, and also for approximated SU(4)s that use 2 CX instead of 3 (omitted for brevity).
- Decomposition in the natural gate direction: while the device software is easily capable of implementing a CX gate in both directions, in reality there is a preferred gate direction in terms of speed and error on the hardware. The other direction is achieved by local pre- and post-rotations. The



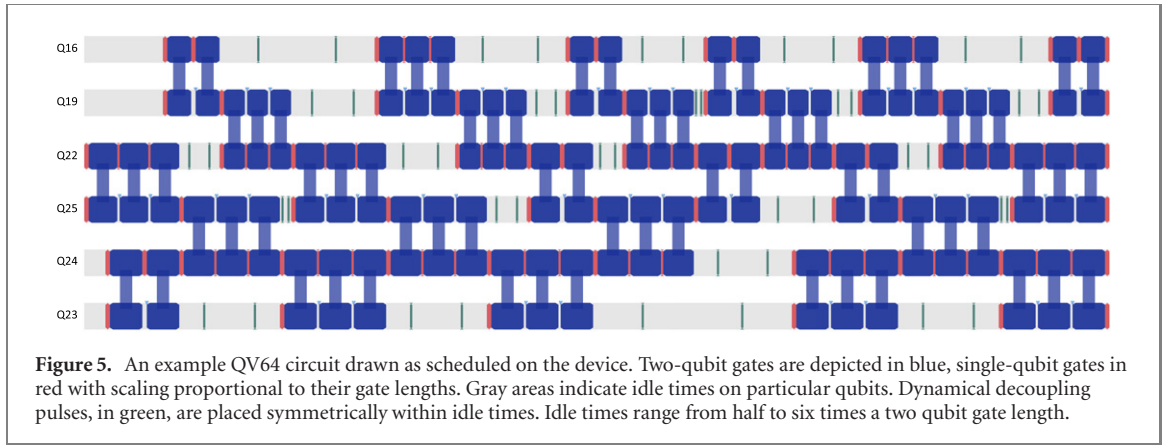
same is true for ECR gates. By querying the device for its natural direction, we can expand each $SU(4)$ and SWAP in the correct direction in the compiler, avoiding further cost down the road. To synthesize a general $SU(4)$ when the logical and physical directions are mismatched, we employ a trick of double mirroring (adding SWAPs before and after the $SU(4)$). The doubly-mirrored $SU(4)$ implements a different operator, where the middle two rows and middle two columns are swapped. We perform a pulse-efficient synthesis on the doubly-mirrored operator, but apply it in the circuit with the reverse order of qubits. This will ensure the original operator is implemented, but also now with the correct physical gate direction (figure 4). Double-mirroring creates a locally equivalent gate, so any approximation to the original $SU(4)$ still holds with the same error bounds.

- (c) Decomposition to native gate: if a direct CX is not available, we compile to the fundamental two-qubit interaction available. In the case of ECR, this saves us the extra single-qubit pulses involved in creating a CX. This demonstrates the benefit of removing simplifying abstraction barriers in the exposed gate set to gain efficiency [25–27].

4. Dynamical decoupling

When quantum circuits are mapped to physical hardware, not all physical gates can be performed simultaneously. Gate execution-times can vary significantly, not only between single- and two-qubit gates, but also between individual qubits and qubit-pairs. In addition, architecture-specific gate schemes and connectivity determine which and how many gates can be executed in parallel.

An analysis of QV64 circuits mapped to a line of transmon-qubits reveals idle times that are a significant portion of the total circuit duration (figure 5). Two main effects create these idle slots. Firstly, a line configuration with nearest-neighbor gates requires a total of 7.3 SWAPs on average per QV circuit. In the optimal layout and routing choice (section 3) for reducing the number of SWAP gates, SWAPs are not executed at once over the entire quantum register, as shown in figure 3(c). When the basis two-qubit gate is a local equivalent of CNOT, this creates ‘idle holes’ for the duration of three two-qubit gates (figure 5). Secondly, ‘idle holes’ can still arise even if no SWAP operations are required. While single-qubit gates are tuned with identical durations across the entire register, two-qubit gate durations depend on qubit



frequencies and coupling, differing by a factor of 1.5–2 between the fastest and the slowest gates. Given that two-qubit gates are $\approx 10\times$ longer than single-qubit gates, these differences accumulate over the course of the computation, opening up additional temporal gaps when individual qubits sit idle.

Ideally idle qubits would evolve the identity operation; however, this is executed far from perfectly in realistic architectures. While thermal relaxation and white noise dephasing lead to dissipative information loss, cross-talk and unwanted non-local spectator interactions lead to local and non-local unitary errors, respectively. In addition, non-Markovian noise sources such as charge noise lead to non-white dephasing. All three error sources are detrimental as circuits become larger, i.e., wider and deeper. Dynamical decoupling is a thoroughly discussed error mitigation technique [28–30], and in its simplest form, can be a single Hahn echo-pulse [31], refocusing the low-frequency noise spectrum acting on a unitary. Various decoupling sequences have been proposed [32–34], some with self-correcting properties [35, 36], others with non-equidistant temporal spacing [37], and hybrids combining both [38–40], in order to optimize the effective filter function. Recently, dynamical decoupling has been shown to improve single-qubit states and an entangled two-qubit state on a Rigetti and IBM quantum computer [41].

For the successful QV64 measurement presented here, we used the sequence $\tau^{iq}/2 - X_p - \tau^{iq} - X_m - \tau^{iq}/2$, with delays $\tau^{iq} = (T_{\text{idle}}^{iq} - 2 * T_{X_{p/m}}) / 2$, where T_{idle}^{iq} is the i th idle length on qubit q , and $T_{X_{p/m}}$ is the duration of one echo pulse with $X_{p,m}$ being a π -pulse around x -axis with positive/negative sense of rotation.

Figure 6 shows a comparison of identical QV-circuits run with (DD) and without (idle) dynamical decoupling. Dynamical decoupling with $X_p - X_m$ sequences improves 72.8% of all circuits in this run, i.e. $\text{HOP}_{\text{DD}} > \text{HOP}_{\text{idle}}$, with an average HOP increase of 0.0178. The interplay between various DD sequences and random circuits, such as QV circuits, is an open research focus.

5. Direct CX gate

Even with state-of-the-art compiling, QV64 circuits consist of a total of 57 two-qubit gates and 146 single-qubit gates on average. Any improvement in gate speed can significantly reduce the circuit duration compared to the coherence times. However, the optimal gate speed for running a circuit is in general not the speed that maximizes the fidelity of the individual gates. In particular, qubits experience idle times in a multi-qubit circuit (see section 4), and the fidelity of the identity operation during these idle times is not captured in the single-qubit or two-qubit randomized benchmarking fidelities often used to characterize quantum systems. Finding the optimal trade-off between individual gate fidelity and circuit fidelity is currently open research, in addition to characterizing which errors are enhanced by driving gates faster. Here we focus on techniques to reduce two-qubit gate durations, but note that small increases in the speed of either single- or two-qubit gates can significantly impact the performance of QV64 circuits.

As mentioned in section 3, an immediate way to ‘speed up’ two-qubit gates is to incorporate into the circuit compilation any pre-/post-single-qubit rotations needed to get from the native ECR gate to a CX or CNOT. We compare the standard echoed cross-resonance gate ECR CX, shown at the top of figure 7(a), to an ECR gate in which single qubit rotations are compiled separately, reducing the two-qubit gate duration to only the entangling portion of the gate. The errors of ECR CX and ECR, measured by two-qubit randomized benchmarking, are shown in figure 7(b) as a function of the two-qubit gate duration.

Two-qubit gates can be further sped up by finding high-fidelity alternatives to the echo pulse sequence, effectively removing another single-qubit gate from the total two-qubit gate duration. We compare an example of a ‘direct’ echo-free CX pulse sequence, shown at the bottom of figure 7(a), to ECR and ECR CX.

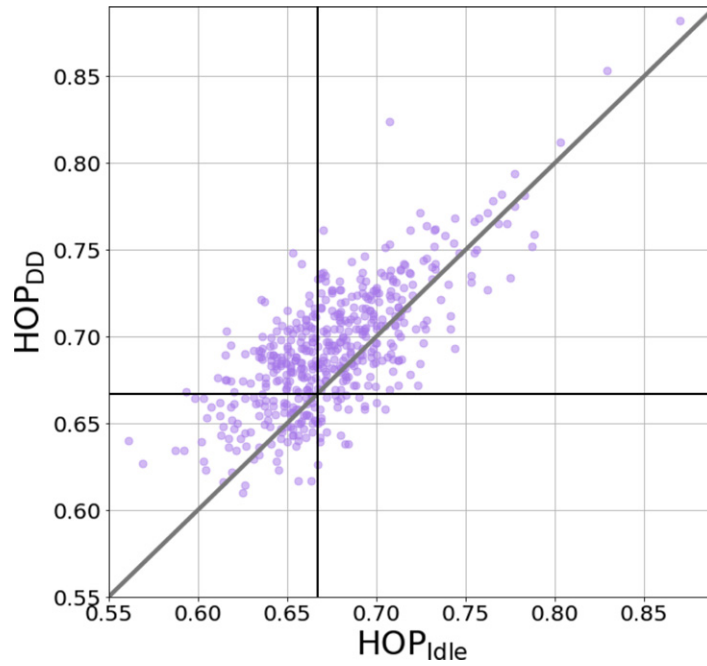


Figure 6. Comparison between heavy output probabilities for the same circuits with and without dynamical decoupling. 2/3-threshold is indicated by the horizontal (vertical) black line. 72.8% of all circuits show larger HOP with decoupling, i.e. are above the gray line, with an average increase of 0.0178.

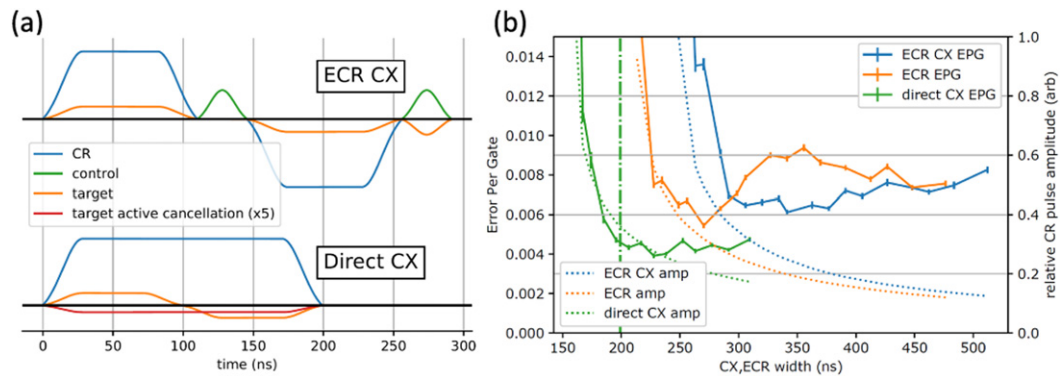
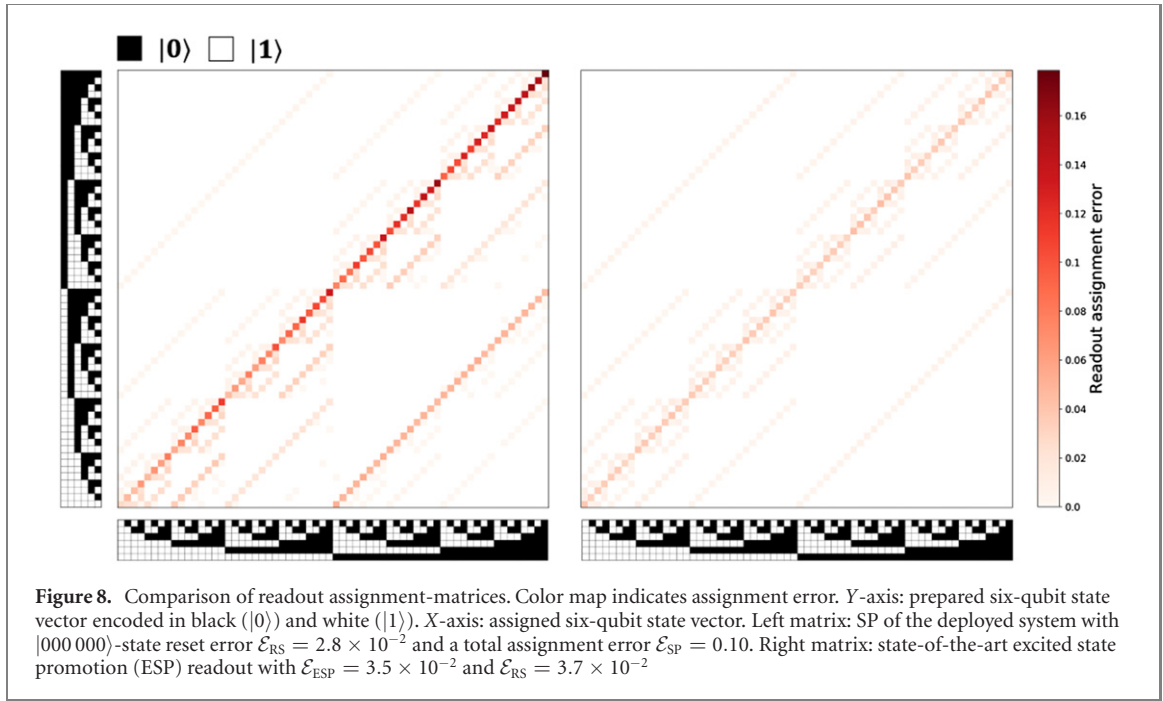


Figure 7. (a) Pulse envelope comparison between the echoed cross resonance (ECR) CX gate implementation of control-target $C_{22}-T_{19}$ on *ibmq_montreal*. (b) Error per gate vs gate width for ECR CX (blue), ECR (orange), and direct CX (green). CR-drive signal amplitudes for the various gate versions and gate widths are shown by the dotted lines. Vertical dashed line indicates the direct CX gate width used for QV64.

This sequence demonstrates an improvement over previous direct CNOT attempts [18] by leveraging our understanding of target rotary pulsing [7]. The resonant drive of the target is implemented as the sum of two parts, an active cancellation tone and a target rotary tone that are symmetric and antisymmetric over the CR pulse, respectively. The active cancellation tone cancels IX terms in the native CR Hamiltonian and any IY terms due to classical crosstalk, while the target rotary pulse can be used to reduce unwanted ZZ and ZY.

The impact of reducing the total gate duration is clearly evidenced by a reduction of two-qubit gate error, as shown in figure 7(b). All gate sequences—ECR CX, ECR, and direct CX—experience a sudden loss of fidelity with increasing pulse amplitude, but the direct CX experiences this break down at a much shorter gate time. We note that reducing the gate duration below that which minimizes two-gate error as measured by randomized benchmarking can increase the HOP of a QV circuit, showing the importance of balancing circuit optimization with gate optimization. For our successful demonstration of QV64 we used a direct CX gate duration of 199 ns, which is shorter than that which minimizes the two-qubit gate error.



6. State initialization and readout

Qubit-state initialization to a fiducial simple state and qubit-specific measurement are two out of five (plus two) necessary DiVincenzo criteria for quantum computation [42]. While certain metrics are designed specifically to be insensitive to ‘state preparation and measurement’ (SPAM) errors, e.g. randomized benchmarking [43, 44] and gate set tomography [45, 46], QV was developed as a holistic system measure and hence is sensitive to SPAM-errors.

In its simplest form, qubit initialization or reset is done passively by waiting multiple T_1 relaxation times before every new computational cycle in order to let the qubit thermalize with its surrounding bath. With ever-increasing coherence times, thermal relaxation protocols impractically limit the computational repetition rate. Various active reset schemes have been proposed and experimentally demonstrated [47, 48]. IBM Quantum systems implement a similar unconditional reset scheme [49]. By measuring the readout matrix (figure 8(a)) we can infer a reset error of $\mathcal{E}_{\text{RS}} = 2.8 \times 10^{-2}$ for the six-qubit ground state $|0 \dots 0\rangle$.

Single qubits are dispersively read out by transversely coupled transmission line cavities [50]. The I - Q trajectories of each measurement signal are integrated with a filter function weighting the initial signal more heavily, hence reducing the sensitivity to T_1 events during measurement [51]. The signal is amplified with a quantum limited traveling wave parametric amplifier followed by a classical amplification chain. This standard procedure (SP) for typically deployed systems gives a total assignment error of $\mathcal{E}_{\text{SP}} = 0.10$ for all 2^6 states.

In order to further boost readout we have implemented excited state promotion (ESP) by applying an additional π -pulse between the first and second excited transmon states $|1\rangle \rightarrow |f\rangle$ before each measurement pulse [11, 52], where $|f\rangle$ is the second excited transmon state. The advantage of this population transfer is threefold. Firstly, the dispersive χ -shift between $|0\rangle \leftrightarrow |f\rangle$ is stronger leading to a larger separation of the signals in the I - Q plane. Secondly, even though the $|f\rangle$ -state has a lifetime half of the $|1\rangle$ -state [53], the qubit excitation has to decay twice $|f\rangle \rightarrow |1\rangle \rightarrow |0\rangle$ (while a two-photon decay $|f\rangle \rightarrow |0\rangle$ is strongly suppressed [54]). This scheme effectively extends the $|1\rangle$ qubit-state lifetime and further reduces false $|0\rangle$ assignment due to T_1 decays. Lastly, the relaxation through an intermediate state leads to a sub-Poissonian relaxations statistics improving the readout error even in the absence of an increase in signal-to-noise ratio [55]. State discrimination is set with a linear discriminant analysis between the states $|0\rangle$ and $|f\rangle$ in the I - Q plane. In order to reset the extended qutrit system, we adapt our reset protocol in the following way: reset— $\pi_{|f\rangle \rightarrow |1\rangle}$ —reset. This state-of-the-art readout reduces the total assignment error to $\mathcal{E}_{\text{ESP}} = 3.5 \times 10^{-2}$ with an initialization error of $\mathcal{E}_{\text{RS}} = 3.7 \times 10^{-2}$, measured with the assignment matrix (figure 8(b)).

7. Conclusion

In this paper we have shown an improvement in the QV of a state-of-the-art superconducting quantum system. We reached a QV of 64 through a combination of four factors: improvements of the Qiskit compiler, refinements to the two-qubit gate and its calibration, addition of dynamical decoupling to mitigate noise affecting idle qubits, and introduction of excited state promoted readout. The last three techniques were developed by having lower-in-the-stack access to how the pulses and gates that comprise quantum circuits are defined before being sent to control the qubits. Furthermore, we note that optimizing the fidelity of quantum circuits is not equivalent to optimizing the gates and confirms the need for circuit benchmarks like QV. This type of hardware-aware approach to make improvements to circuit performance is a hallmark of the current era of noisy quantum systems which we expect to continue until we can achieve error rates in the range of 10^{-4} .

Acknowledgment

We thank all those that contributed to the hardware system delivery and implementation, including the IBM Microelectronics Research Laboratory and Central Scientific Services teams as well as the worldwide team who designed, built and tested the custom control electronics. We further acknowledge all the work from the broader IBM Quantum team who helped this effort across the full stack. The 27-qubit chip design used in this study is also being used in the IARPA LogiQ program. There is some shared learning for characterization of devices which we acknowledge support through the ARO Grant No. W911NF-16-1-0114.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Petar Jurcevic  <https://orcid.org/0000-0003-1234-6386>
 Christopher J Wood  <https://orcid.org/0000-0001-7606-7349>
 Eric J Zhang  <https://orcid.org/0000-0003-4278-3763>
 Jay M Gambetta  <https://orcid.org/0000-0002-4620-0978>

References

- [1] Cross A W, Bishop L S, Sheldon S, Nation P D and Gambetta J M 2019 Validating quantum computers using randomized model circuits *Phys. Rev. A* **100** 032328
- [2] Vuduc R and Czechowski K 2012 Toward a theory of algorithm-architecture co-design *High Performance Computing for Computational Science-VECPAR 2012* ed M Daye, O Marques and K Nakajima vol 7851 (Springer) pp 4–8
- [3] Brown K R, Kim J and Monroe C 2016 Co-designing a scalable quantum computer with trapped atomic ions (arXiv:1602.02840)
- [4] Nam Y and Maslov D 2019 Low-cost quantum circuits for classically intractable instances of the Hamiltonian dynamics simulation problem *npj Quantum Inf.* **5** 44
- [5] Pino J M *et al* 2020 Demonstration of the QCCD trapped-ion quantum computer architecture (arXiv:2003.01293)
- [6] Karalekas P J, Tezak N A, Peterson E C, Ryan C A, da Silva M P and Smith R S 2020 A quantum-classical cloud platform optimized for variational hybrid algorithms *Quantum Sci. Technol.* **5** 024003
- [7] Sundaresan N, Lauer I, Pritchett E, Magesan E, Jurcevic P and Gambetta J M 2020 Reducing unitary and spectator errors in cross resonance with optimized rotary echoes (arXiv:2007.02925)
- [8] Yirka B 2020 Honeywell claims to have built the highest-performing quantum computer available <https://phys.org/news/2020-06-honeywell-built-highest-performing-quantum.html>
- [9] Qiskit: An open-source framework for quantum computing 2021 <https://qiskit.org/>
- [10] McKay D C *et al* 2018 Qiskit backend specifications for openQASM and openpulse experiments (arXiv:1809.03452)
- [11] Mallet F, Ong F R, Palacios-Laloy A, Nguyen F, Bertet P, Vion D and Esteve D 2009 Single-shot qubit readout in circuit quantum electrodynamics *Nat. Phys.* **5** 791–5
- [12] Chamberland C, Zhu G, Yoder T J, Hertzberg J B and Cross A W 2020 Topological and subsystem codes on low-degree graphs with flag qubits *Phys. Rev. X* **10** 011022
- [13] IBM Quantum Experience 2020 <https://quantum-computing.ibm.com/>
- [14] Gambetta J M *et al* 2012 Characterization of addressability by simultaneous randomized benchmarking *Phys. Rev. Lett.* **109** 240504
- [15] Motzoi F, Gambetta J M, Reberstrost P and Wilhelm F K 2009 Simple pulses for elimination of leakage in weakly nonlinear qubits *Phys. Rev. Lett.* **103** 110501
- [16] Paraoanu G S 2006 Microwave-induced coupling of superconducting qubits *Phys. Rev. B* **74** 140504

- [17] Rigetti C and Devoret M 2010 Fully microwave-tunable universal gates in superconducting qubits with linear couplings and fixed transition frequencies *Phys. Rev. B* **81** 134507
- [18] Chow J M et al 2011 Simple all-microwave entangling gate for fixed-frequency superconducting qubits *Phys. Rev. Lett.* **107** 080502
- [19] IBM CPLEX Optimizer 2019 CPLEX User's Manual <https://www.ibm.com/analytics/cplex-optimizer>
- [20] Li G, Ding Y and Yuan X 2019 Tackling the qubit mapping problem for NISQ-era quantum devices (arXiv:1809.02573)
- [21] Sheldon S, Magesan E, Chow J M and Gambetta J M 2016 Procedure for systematically tuning up cross-talk in the cross-resonance gate *Phys. Rev. A* **93** 060302
- [22] McKay D C, Filipp S, Mezzacapo A, Magesan E, Chow J M and Gambetta J M 2016 Universal gate for fixed-frequency qubits via a tunable bus *Phys. Rev. Appl.* **6** 064007
- [23] Vatan F and Williams C 2004 Optimal quantum circuits for general two-qubit gates *Phys. Rev. A* **69** 032315
- [24] Shende V V, Markov I L and Bullock S S 2004 Minimal universal two-qubit controlled-not-based circuits *Phys. Rev. A* **69** 062321
- [25] Maslov D 2017 Basic circuit compilation techniques for an ion-trap quantum machine *New J. Phys.* **19** 023035
- [26] Murali P, Linke N M, Martonosi M, Abhari A J, Nguyen N H and Alderete C H 2019 Full-stack, real-system quantum computer studies: architectural comparisons and design insights 2019 ACM/IEEE 46th Annual Int. Symp. on Computer Architecture (ISCA) (IEEE) pp 527–40
- [27] Arute F et al 2020 Quantum approximate optimization of non-planar graph problems on a planar superconducting processor (arXiv:2004.04197)
- [28] Viola L, Knill E and Lloyd S 1999 Dynamical decoupling of open quantum systems *Phys. Rev. Lett.* **82** 2417–21
- [29] Souza A M, Álvarez G A and Suter D 2012 Robust dynamical decoupling *Phil. Trans. R. Soc. A* **370** 4748–69
- [30] Suter D and Álvarez G A 2016 Colloquium: protecting quantum information against environmental noise *Rev. Mod. Phys.* **88** 041001
- [31] Hahn E L 1950 Spin echoes *Phys. Rev.* **80** 580–94
- [32] Carr H Y and Purcell E M 1954 Effects of diffusion on free precession in nuclear magnetic resonance experiments *Phys. Rev.* **94** 630–8
- [33] Meiboom S and Gill D 1958 Modified spin-echo method for measuring nuclear relaxation times *Rev. Sci. Instrum.* **29** 688–91
- [34] Maudsley A A 1986 Modified Carr–Purcell–Meiboom–Gill sequence for nmr fourier imaging applications *J. Magn. Reson.* **69** 488–91
- [35] Khodjasteh K and Lidar D A 2005 Fault-tolerant quantum dynamical decoupling *Phys. Rev. Lett.* **95** 180501
- [36] Khodjasteh K and Lidar D A 2007 Performance of deterministic dynamical decoupling schemes: concatenated and periodic pulse sequences *Phys. Rev. A* **75** 062310
- [37] Uhrig G S 2007 Keeping a quantum bit alive by optimized π -pulse sequences *Phys. Rev. Lett.* **98** 100504
- [38] Uhrig G S 2009 Concatenated control sequences based on optimized dynamic decoupling *Phys. Rev. Lett.* **102** 120502
- [39] Souza A M, Álvarez G A and Suter D 2011 Robust dynamical decoupling for quantum computing and quantum memory *Phys. Rev. Lett.* **106** 240501
- [40] Quiroz G and Lidar D A 2011 Quadratic dynamical decoupling with nonuniform error suppression *Phys. Rev. A* **84** 042328
- [41] Pokharel B, Anand N, Fortman B and Lidar D A 2018 Demonstration of fidelity improvement using dynamical decoupling with superconducting qubits *Phys. Rev. Lett.* **121** 220502
- [42] DiVincenzo D P 2000 The physical implementation of quantum computation *Fortschr. Phys.* **48** 771–83
- [43] Knill E et al 2008 Randomized benchmarking of quantum gates *Phys. Rev. A* **77** 012307
- [44] Magesan E, Gambetta J M and Emerson J 2011 Scalable and robust randomized benchmarking of quantum processes *Phys. Rev. Lett.* **106** 180504
- [45] Merkel S T, Gambetta J M, Smolin J A, Poletto S, Córcoles A D, Johnson B R, Ryan C A and Steffen M 2013 Self-consistent quantum process tomography *Phys. Rev. A* **87** 062119
- [46] Blume-Kohout R, Gamble J K, Nielsen E, Mizrahi J, Sterk J D and Manuz P 2013 Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit (arXiv:1310.4492)
- [47] Magnard P et al 2018 Fast and unconditional all-microwave reset of a superconducting qubit *Phys. Rev. Lett.* **121** 060502
- [48] Egger D J, Werninghaus M, Ganzhorn M, Salis G, Fuhrer A, Müller P and Filipp S 2018 Pulsed reset protocol for fixed-frequency superconducting qubits *Phys. Rev. Appl.* **10** 044030
- [49] Vlastakis B, Fung K and Steffen M 2020 still in preparation
- [50] Gambetta J, Blais A, Boissonneault M, Houck A A, Schuster D I and Girvin S M 2008 Quantum trajectory approach to circuit QED: quantum jumps and the Zeno effect *Phys. Rev. A* **77** 012112
- [51] Ryan C A, Johnson B R, Gambetta J M, Chow J M, da Silva M P, Dial O E and Ohki T A 2015 Tomography via correlation of noisy measurement records *Phys. Rev. A* **91** 022118
- [52] Elder S S et al 2020 High-fidelity measurement of qubits encoded in multilevel superconducting circuits *Phys. Rev. X* **10** 011001
- [53] Peterer M J et al 2015 Coherence and decay of higher energy levels of a superconducting transmon qubit *Phys. Rev. Lett.* **114** 010501
- [54] Koch J et al 2007 Charge-insensitive qubit design derived from the Cooper pair box *Phys. Rev. A* **76** 042319
- [55] D'Anjou B and Coish W A 2017 Enhancing qubit readout through dissipative sub-Poissonian dynamics *Phys. Rev. A* **96** 052321