Demonstration of quantum volume 64 on a superconducting quantum computing system

Petar Jurcevic, Ali Javadi-Abhari, Lev S. Bishop, Isaac Lauer, Daniela F. Bogorin, Markus Brink, Lauren Capelluto, Oktay Günlük, Toshinaro Itoko, Naoki Kanazawa, Abhinav Kandala, George A. Keefe, Kevin Kruslich, William Landers, Eric P. Lewandowski, Douglas T. McClure, Giacomo Nannicini, Adinath Narasgond, Hasan M. Nayfeh, Emily Pritchett, Mary Beth Rothwell, Srikanth Srinivasan, Neereja Sundaresan, Cindy Wang, Ken X. Wei, Christopher J. Wood, Jeng-Bang Yau, Eric J. Zhang, Oliver E. Dial, Jerry M. Chow, Jay M. Gambetta

***Abstract*—We improve the quality of quantum circuits on superconducting quantum computing systems, as measured by the quantum volume, 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 quantum volume 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.**

I. 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 towards machines capable of implementing circuits with a quantum advantage, meaning certain information processing tasks can be performed more efficiently or cost effectively than 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, quantum volume includes gate errors, measurement errors, the quality of the circuit compiler, and spectator errors. In Ref. [1] and later in Ref. [2], QV16 was measured on *ibmq_johannesburg* and a Honeywell quantum system, respectively. In Ref. [3] QV8 was measured for the Rigetti Aspen-4 quantum system. We recently increased *ibmq_johannesburg* to QV32 [4] 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 [5].

Here we demonstrate an increase in the quantum volume of an IBM quantum system by improving the Qiskit compiler [6], implementing excited state promoted (ESP) readout, shorter two-qubit gates, and adding dynamic decoupling to the idle qubits. These last two demonstrate the need for timing and pulse control in cloud quantum systems [7]. 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 $98.7\% (z > 2)$, see Fig. 2 a).

In section II we give an overview of the *ibmq_montreal* device, which is a 27-qubit IBM Quantum Falcon processor; in section III we discuss the improvements to the compile; in section IV we discuss the dynamical decoupling protocol; in V we discuss the faster implementation of the direct CNOT gate which extends the improved pulse control of [4]; in section VI we discuss the improvement in measurement fidelity by using a control pulse to promote the excited state to a higher level before measurement [8]. Finally in section VII we conclude the paper.

II. 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 [9]. A photo of this processor is shown in Fig. 1 a), and a schematic of its connectivity is shown in Fig. 1 b). A high connectivity layout, such as 'all-to-all', is preferable for random quantum circuits (such as QV circuits) as the average qubit-qubit distance is reduced; 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 swaps to emulate 'all-to-all' connectivity. In these systems, by using the techniques described in [4], we have measured a QV of 32 on the last 7 deployed systems [10] demonstrating the reliability of this architecture.

In this paper we achieve a QV64 on *ibmq_montreal*, which is one of the latest deployed IBM Quantum Falcon processors. The quantum volume circuits were run on a line of six qubits, Q16-Q19-Q22-Q25-Q24-Q23 (orange shaded qubits in Fig. 1 b). Individual qubit properties are shown in

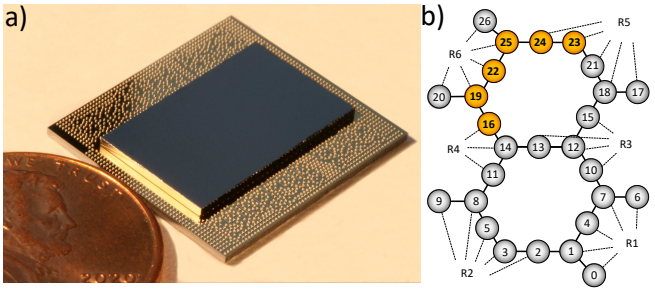

Fig. 1. a) Image of a representative IBM Quantum Falcon processor with a penny for scale. The lattice connectivity is defined through couplings on a top qubit die which is bump-bonded to a bottom interposer die for signal delivery and readout. b) Schematic of the 27-qubit (numbered 0 through 26) heavy-hex layout connectivity. Qubits used for the confirmed QV64 are shaded in orange. Dashed lines indicate collections of qubits that are multiplexed together for readout (labeled R1 to R6).

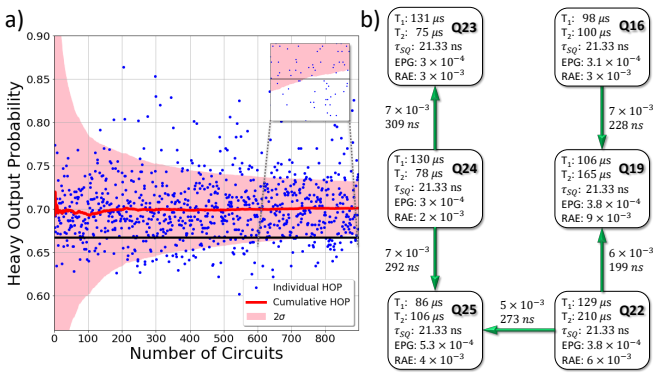

Fig. 2. a) One of two statistically confirmed QV64 runs. Here, a total of $\approx$ 900 random circuits are run. Inset: QV success criteria were reached $>$ 724 circuits. Blue: Heavy output probability for each individual circuit. Red: Cumulative Heavy output probability with shaded region $\pm 2\sigma$ as calculated per appendix C of [1]. Black: Quantum volume success threshold at 2/3. b) Qubits used in the successful QV64 measurement. EPG: error per gate (single-qubit) measured with RB. RAE: readout assignment error. τ_{SQ}: single-qubit gate duration. Natural two-qubit gate direction is shown in green, from control to target. Two-qubit gate error rates and gate durations are shown next to the corresponding qubit-qubit link.

Fig 2 b) with the following average values: $\mathrm{T}_1 = 113\,\mu$s, $\mathrm{T}_2 = 122\,\mu$s, error per single-qubit gate 2.8×10^{-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 [11].

The qubits are fixed-frequency transmons with frequencies $\approx$ 5 GHz. Single-qubit gates are driven resonantly with a microwave pulse of duration $\tau_{\mathrm{sq}} = 21.33\,\mathrm{ns}$. A DRAG pulse envelope [12] corrects σ_z-errors and signal dispersion due to wiring. Two-qubit gates are based on a cross-resonance scheme [13–15] with a target rotary pulse [4] and an additional offset pulse-shape on the target for implementing a direct (echoless) CNOT as described later. Two-qubit gate lengths are $\tau_{\mathrm{tq}} = 199 - 309\,\mathrm{ns}$.

III. 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 quantum volume 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 peep-hole 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.

a) Qubit layout and routing via Binary Integer Programming: We formulate qubit layout and routing as a binary integer programming (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:

$$C = K^d \prod_{j \in G} F_j^{\mathrm{best}} \prod_{j \in \bar{G}} \bar{F}_j^{\mathrm{best}} \prod_{j \in S} F_b^3, \tag{1}$$

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} [$\bar{F}_j^{\mathrm{best}}$] is the modeled fidelity of the best approximation to the target unitary making $i = 0, \ldots, 3$ uses of the entangling gate

$$F_j^{\mathrm{best}} = \max_i F_{i,j}^{\mathrm{avg}} (F_b)^i, \tag{2}$$

$$\bar{F}_j^{\mathrm{best}} = \max_i \bar{F}_{i,j}^{\mathrm{avg}} (F_b)^i, \tag{3}$$

and $F_{i,j}^{\mathrm{avg}}$ 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 [16]. Figure 3 shows the performance of this BIP pass in comparison to the state-of-the art SABRE algorithm [17] available in Qiskit, showing substantial improvement in both the mean and maximum number of uses of the entangling gate.

b) 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[18] (c.f. Section V). The Phase gate can be achieved with zero time and error [19]. 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 [20], 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 2-qubit gates [20, 21]. 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:

1. 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 a 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).

2. 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.

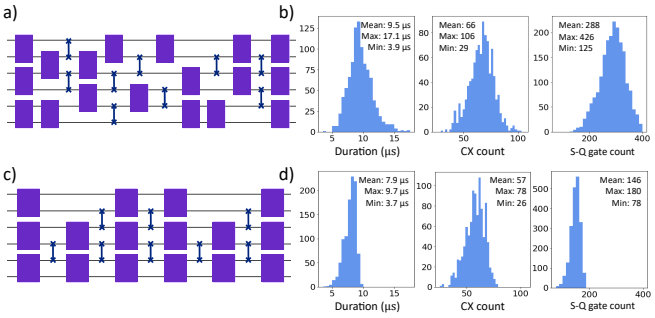

Fig. 3. Comparison of QV64 circuits transpiled to a line connectivity with a), b) the state-of-the-art Qiskit compiler and c), d) an improved transpilation method based on BIP and additional gate cancellations, see text. a) and c) show the same random example circuit, mapped with "SABRE" and mapped with 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 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.

3. 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 in compiling [22–24].

IV. 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 (Fig. 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 III for reducing the number of SWAP gates, SWAPs are not executed at once over the entire quantum register, as shown in 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 (Fig. 5). Secondly, "idle holes" can still arise even if no SWAP operations are required. While single-qubit gates

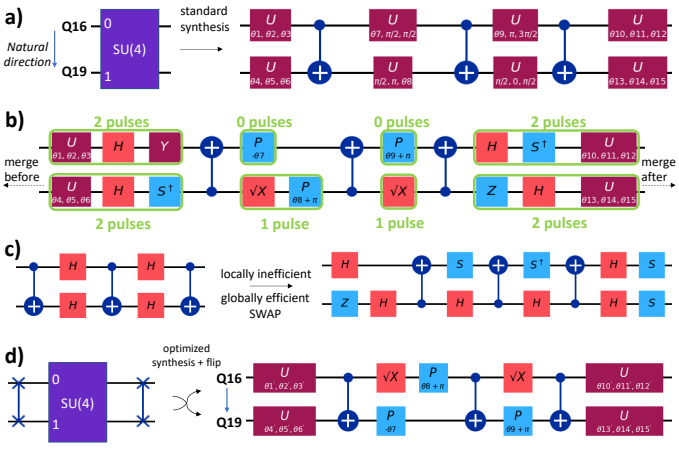

Fig. 4. a) Standard (Qiskit) decomposition of an SU(4) operator in terms of 3 CNOTs and layers of single-qubit rotations. Each U contributes 2 SQ pulses for a total of 16 pulses. When the expansion is not in the "natural" direction of the hardware CX, extra single qubit rotations will be involved. b) A new pulse-efficient SU(4) decomposition obtained constructively from the first (c.f. same 15 θ parameters). c) This decomposition applied to SWAPs creates more global efficiency. Even though more pulses are used locally (H), the "inner" pulses are reduced and the "outer" ones can merge with gates before and after the SWAP. d) If the direction of the above expansion is incorrect, synthesize a different "doubly-mirrored" operator, then flip it at the point of use. This has the same number of pulses but now in the correct direction.

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 [25–27], and in its simplest form, can be a single Hahn echo-pulse [28], refocusing the low-frequency noise spectrum acting on a unitary. Various decoupling sequences have been proposed [29–31], some with self-correcting properties [32, 33], others with non-equidistant temporal spacing [34], and hybrids combining both [35–37], in order to optimize the effective filter function.

For the successful QV64 measurement presented here, we used the sequence $\tau^{i,q}/2 - X_p - \tau^{i,q} - X_m - \tau^{i,q}/2$, with delays $\tau^{i,q} = \left(T^{i,q}_{\text{idle}} - 2 * T_{X_{p/m}}\right)/2$, where $T^{i,q}_{\text{idle}}$ is the ith 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 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. We found that the $X_p - X_m$ sequence gave the best average performance, when compared to higher order decoupling sequences. The interplay between various DD sequences and random circuits, such as QV circuits, is an open research focus.

V. 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 IV), 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 and 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 III, 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 Fig. 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 Fig. 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 Fig. 7 a), to ECR and ECR CX. This sequence demonstrates an improvement over previous direct CNOTs attempts [15] by leveraging our understanding of target rotary pulsing [4]. 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

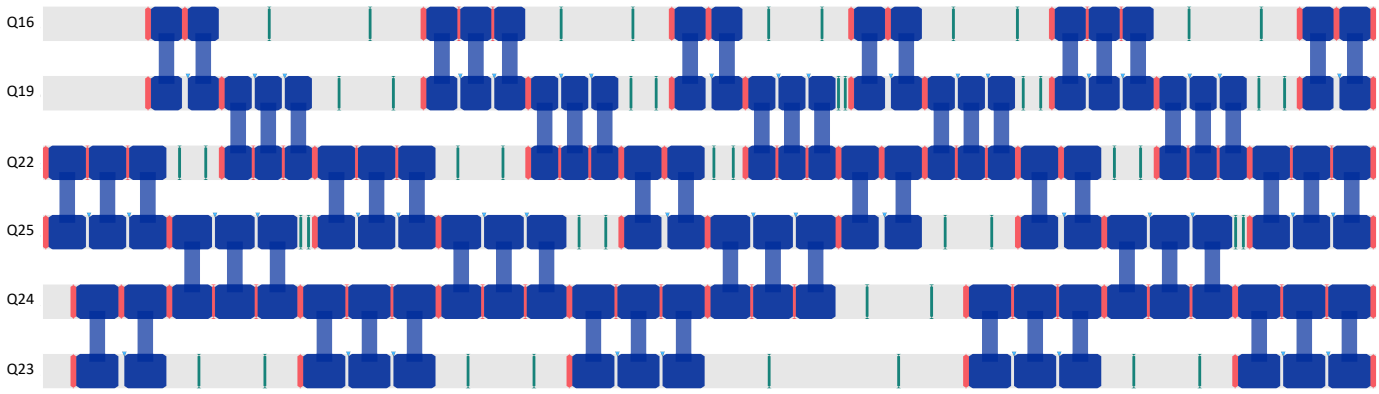

Fig. 5. An example QV64 circuit drawn as scheduled on the device. Two-qubit gates are depicted in blue, single-qubit gates in red with scaling proportional to their gate lengths. Grey areas indicate idle times on particular qubits. Dynamical decoupling pulses, in green, are placed symmetrically within idle times. Idle times range from half to six times a two qubit gate length.

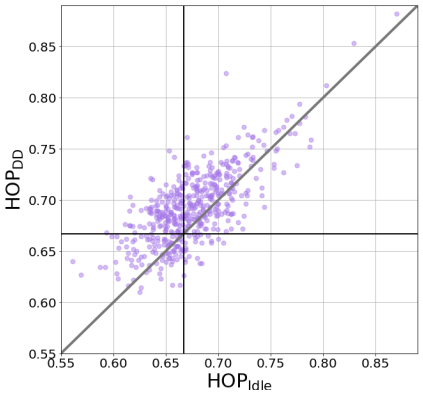

Fig. 6. Comparison between heavy output probabilities for the same circuits with and without dynamical decoupling. 2/3-threshold is indicated by the horizontal (vertical) black line. 72.8% of all circuits show larger HOP with decoupling, i.e. are above the grey line, with an average increase of 0.0178.

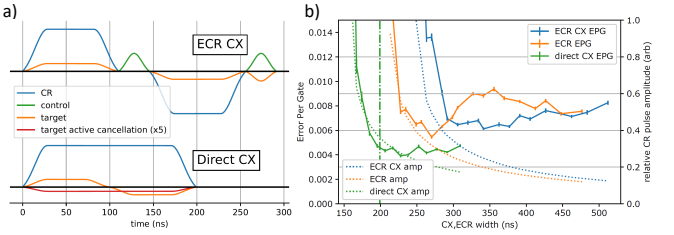

Fig. 7. a) Pulse envelope comparison between the echoed cross resonance (ECR) CX gate and the direct CX gate implementation of Control-Target C_{22}-T_{19} on *ibmq_montreal*. b) Error per gate vs. gate width for ECR CX (blue), ECR (orange), and direct CNOT (green). CR-drive signal amplitudes for the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

in Fig.7 b). All gate sequences – ECR, ECR CX, 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.

VI. State Initialization and Readout

Qubit-state initialization to a fiducial simple state and qubit-specific measurement are two out of five (plus two) necessary DiVinceno criteria for quantum computation [38]. While certain metrics are designed specifically to be insensitive to "state preparation and measurement" (SPAM) errors, e.g. randomized benchmarking [39, 40] and gate set tomography [41, 42], quantum volume was developed as a holistic system measure and hence is sensitive to SPAM-errors into account.

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 [43, 44]. IBM Quantum systems implement a similar unconditional reset scheme [45]. By measuring the readout matrix (Fig. 8 a)) we can obtain a reset error of $\mathcal{E}_{\mathrm{RS}} = 2.8 \times 10^{-2}$ for the six-qubit ground state $|0\ldots0\rangle$.

Single qubits are dispersively read out by transversely coupled transmission line cavities [46]. 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[47]. The signal is amplified with a quantum limited travelling wave parametric amplifier followed by a classical amplification chain. This standard procedure (SP) for typically deployed systems gives a total assignment error of $\mathcal{E}_{\mathrm{SP}} = 0.10$ 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 [8], where $|f\rangle$

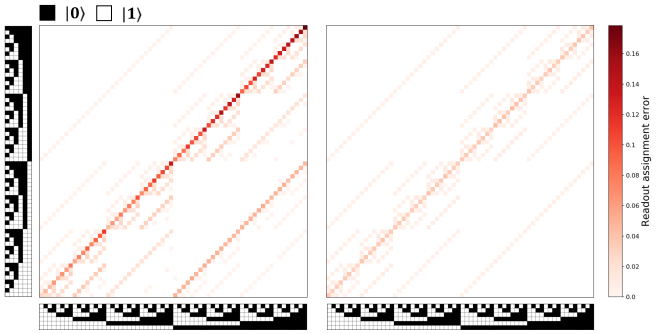

Fig. 8. Comparison of readout assignment-matrices. Color map indicates assignment error. Y-axis: Prepared six-qubit state vector encoded in black ($|0\rangle$) and white ($|1\rangle$). X-axis: Assigned six-qubit state vector. Left matrix: Standard procedure of the deployed system with $|000000\rangle$-state reset error $\mathcal{E}_{\mathrm{RS}} = 2.8 \times 10^{-2}$ and a total assignment error $\mathcal{E}_{\mathrm{SP}} = 0.10$. Right matrix: State-of-the-art excited state promotion (ESP) readout with $\mathcal{E}_{\mathrm{ESP}} = 3.5 \times 10^{-2}$ and $\mathcal{E}_{\mathrm{RS}} = 3.7 \times 10^{-2}$

is the second excited transmon state. The advantage of this population transfer is twofold. 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 $|e\rangle$-state [48], 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 [49]). This scheme effectively extends the $|1\rangle$ qubit-state lifetime and further reduces false $|0\rangle$ assignment due to T_1 decays. State discrimination is set with a linear discriminant analysis (LDA) 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 $\mathcal{E}_{\mathrm{ESP}} = 3.5 \times 10^{-2}$ with an initialization error of $\mathcal{E}_{\mathrm{RS}} = 3.7 \times 10^{-2}$, measured with the assignment matrix (Fig. 8 b)).

VII. Conclusion

In this paper we have shown an improvement in the quantum volume of a state-of-the-art superconducting quantum system. We measured a quantum volume of 64. This was reached through a combination of four factors: improving the Qiskit compiler, refinements to two-qubit gate and its calibration, adding in dynamical decoupling to mitigate noise affecting idle qubits, and the introduction of excited state promoted readout. The last two techniques were developed by having lower-in-the-stack access to how the pulses and gates that compromise 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 quantum volume. 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}.

Acknowledgement

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.

References

[1] A.W. Cross, L.S. Bishop, S. Sheldon, P.D. Nation, and J.M. Gambetta. Validating quantum computers using randomized model circuits. *Phys. Rev. A*, 100:032328, Sep 2019.

[2] J. M. Pino, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, M. S. Allman, M. Baldwin, C. H.and Foss-Feig, D. Hayes, K. Mayer, C. Ryan-Anderson, and B. Neyenhuis. Demonstration of the qccd trapped-ion quantum computer architecture. *arXiv:2003.01293*.

[3] Peter J Karalekas, Nikolas A Tezak, Eric C Peterson, Colm A Ryan, Marcus P da Silva, and Robert S Smith. A quantum-classical cloud platform optimized for variational hybrid algorithms. *Quantum Science and Technology*, 5(2):024003, apr 2020.

[4] Neereja Sundaresan, Isaac Lauer, Emily Pritchett, Easwar Magesan, Petar Jurcevic, and Jay M. Gambetta. Reducing unitary and spectator errors in cross resonance with optimized rotary echoes. *arXiv:2007.02925*.

[5] Honeywell claims to have built the highest-performing quantum computer available. https://phys.org/news/2020-06-honeywell-built-highest-performing-quantum.html, 2020.

[6] Qiskit: An open-source framework for quantum computing. https://qiskit.org/, 2020.

[7] David C McKay, Thomas Alexander, Luciano Bello, Michael J Biercuk, Lev Bishop, Jiayin Chen, Jerry M Chow, Antonio D Córcoles, Daniel Egger, Stefan Filipp, et al. Qiskit backend specifications for openqasm and openpulse experiments. *arXiv preprint arXiv:1809.03452*, 2018.

[8] F. Mallet, F. Ong, A. Palacios-Laloy, F. Nguyen, P. Bertet, D. Vion, and D. Esteve. Single-shot qubit readout in circuit quantum electrodynamics. *Nature Phys*, 5:791–795, 2009.

[9] Christopher Chamberland, Guanyu Zhu, Theodore J. Yoder, Jared B. Hertzberg, and Andrew W. Cross. Topological and subsystem codes on low-degree graphs with flag qubits. *Phys. Rev. X*, 10:011022, Jan 2020.

[10] IBM Quantum Experience. https://quantum-computing.ibm.com/, 2020.

[11] J.M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J.A. Smolin, J.M. Chow, C.A. Ryan, C. Rigetti, S. Poletto, T.A. Ohki, M.B. Ketchen, and M. Steffen. Characterization of addressability by simultaneous randomized benchmarking. *Phys. Rev. Lett.*, 109:240504, Dec 2012.

[12] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm. Simple pulses for elimination of leakage in weakly nonlinear qubits. *Phys. Rev. Lett.*, 103:110501, Sep 2009.

[13] G.S. Paraoanu. Microwave-induced coupling of superconducting qubits. *Phys. Rev. B*, 74:140504, Oct 2006.

[14] C. Rigetti and M. Devoret. Fully microwave-tunable universal gates in superconducting qubits with linear couplings and fixed transition frequencies. *Phys. Rev. B*, 81:134507, Apr 2010.

[15] J.M. Chow, A. D. Córcoles, J.M. Gambetta, C. Rigetti, B. R. Johnson, J.A. Smolin, J. R. Rozen, G.A. Keefe, M.B. Rothwell, M.B. Ketchen, and M. Steffen. Simple all-microwave entangling gate for fixed-frequency superconducting qubits. *Phys. Rev. Lett.*, 107:080502, Aug 2011.

[16] IBM ILOG Cplex. V12. 1: User's manual for cplex. *International Business Machines Corporation*, 46(53):157, 2009.

[17] Gushu Li, Yufei Ding, and Yuan Xie. Tackling the Qubit Mapping Problem for NISQ-Era Quantum Devices. *arXiv:1809.02573 [quant-ph]*, May 2019. arXiv: 1809.02573.

[18] S. Sheldon, E. Magesan, J.M. Chow, and J.M. Gambetta. Procedure for systematically tuning up cross-talk in the cross-resonance gate. *Phys. Rev. A*, 93:060302, Jun 2016.

[19] D.C. McKay, S. Filipp, A. Mezzacapo, E. Magesan, J.M. Chow, and J.M. Gambetta. Universal gate for fixed-frequency qubits via a tunable bus. *Phys. Rev. Applied*, 6:064007, Dec 2016.

[20] Farrokh Vatan and Colin Williams. Optimal quantum circuits for general two-qubit gates. *Physical Review A*, 69(3):032315, 2004.

[21] Vivek V Shende, Igor L Markov, and Stephen S Bullock. Minimal universal two-qubit controlled-not-based circuits. *Physical Review A*, 69(6):062321, 2004.

[22] Dmitri Maslov. Basic circuit compilation techniques for an ion-trap quantum machine. *New Journal of Physics*, 19(2):023035, 2017.

[23] Prakash Murali, Norbert Matthias Linke, Margaret Martonosi, Ali Javadi Abhari, Nhung Hong Nguyen, and Cinthia Huerta Alderete. Full-stack, real-system quantum computer studies: Architectural comparisons and design insights. In *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*, pages 527–540. IEEE, 2019.

[24] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Sergio Boixo, Michael Broughton, Bob B Buckley, David A Buell, et al. Quantum approximate optimization of non-planar graph problems on a planar superconducting processor. *arXiv preprint arXiv:2004.04197*, 2020.

[25] Lorenza Viola, Emanuel Knill, and Seth Lloyd. Dynamical decoupling of open quantum systems. *Phys. Rev. Lett.*, 82:2417–2421, Mar 1999.

[26] A. M. Souza, Gonzalo A. Álvarez, and Dieter Suter. Robust dynamical decoupling. *Phil. Trans. R. Soc. A*, 370:4748–4769, Oct 2012.

[27] Dieter Suter and Gonzalo A. Álvarez. Colloquium: Protecting quantum information against environmental noise. *Rev. Mod. Phys.*, 88:041001, Oct 2016.

[28] E. L. Hahn. Spin echoes. *Phys. Rev.*, 80:580–594, Nov 1950.

[29] H. Y. Carr and E. M. Purcell. Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Phys. Rev.*, 94:630–638, May 1954.

[30] S. Meiboom and D. Gill. Modified spin-echo method for measuring nuclear relaxation times. *Review of Scientific Instruments*, 29(8):688–691, 1958.

[31] A.A Maudsley. Modified carr-purcell-meiboom-gill sequence for nmr fourier imaging applications. *Journal of Magnetic Resonance (1969)*, 69(3):488 – 491, 1986.

[32] K. Khodjasteh and D. A. Lidar. Fault-tolerant quantum dynamical decoupling. *Phys. Rev. Lett.*, 95:180501, Oct 2005.

[33] Kaveh Khodjasteh and Daniel A. Lidar. Performance of deterministic dynamical decoupling schemes: Concatenated and periodic pulse sequences. *Phys. Rev. A*, 75:062310, Jun 2007.

[34] Götz S. Uhrig. Keeping a quantum bit alive by optimized π-pulse sequences. *Phys. Rev. Lett.*, 98:100504, Mar 2007.

[35] Götz S. Uhrig. Concatenated control sequences based on optimized dynamic decoupling. *Phys. Rev. Lett.*, 102:120502, Mar 2009.

[36] Alexandre M. Souza, Gonzalo A. Álvarez, and Dieter Suter. Robust dynamical decoupling for quantum computing and quantum memory. *Phys. Rev. Lett.*, 106:240501, Jun 2011.

[37] Gregory Quiroz and Daniel A. Lidar. Quadratic dynamical decoupling with nonuniform error suppression. *Phys. Rev. A*, 84:042328, Oct 2011.

[38] David P. DiVincenzo. The physical implementation of quantum computation. *Fortschritte der Physik*, 48(9-11):771–783, 2000.

[39] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland. Randomized benchmarking of quantum gates. *Phys. Rev. A*, 77:012307, Jan 2008.

[40] E. Magesan, J. M. Gambetta, and J. Emerson. Scalable and robust randomized benchmarking of quantum processes. *Phys. Rev. Lett.*, 106:180504, May 2011.

[41] Seth T. Merkel, Jay M. Gambetta, John A. Smolin, Stefano Poletto, Antonio D. Córcoles, Blake R. Johnson, Colm A. Ryan, and Matthias Steffen. Self-consistent quantum process tomography. *Phys. Rev. A*, 87:062119, Jun 2013.

[42] Robin B Blume-Kohout, John King Gamble, Erik Erik Nielsen, Jonathan Mizrahi, Jonathan D. Sterk, and Peter Maunz. Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit. *arXiv:1310.4492*, 2013.

[43] P. Magnard, P. Kurpiers, B. Royer, T. Walter, J.-C. Besse, S. Gasparinetti, M. Pechal, J. Heinsoo, S. Storz, A. Blais, and A. Wallraff. Fast and unconditional all-microwave reset of a superconducting qubit. *Phys. Rev. Lett.*, 121:060502, Aug 2018.

[44] D.J. Egger, M. Werninghaus, M. Ganzhorn, G. Salis, A. Fuhrer, P. Müller, and S. Filipp. Pulsed reset protocol for fixed-frequency superconducting qubits. *Phys. Rev. Applied*, 10:044030, Oct 2018.

[45] B. Vlastakis, K. Fung, and M. Steffen. *In preparation*,

2020.

[46] Jay Gambetta, Alexandre Blais, M. Boissonneault, A. A. Houck, D. I. Schuster, and S. M. Girvin. Quantum trajectory approach to circuit qed: Quantum jumps and the zeno effect. *Phys. Rev. A*, 77:012112, Jan 2008.

[47] Colm A. Ryan, Blake R. Johnson, Jay M. Gambetta, Jerry M. Chow, Marcus P. da Silva, Oliver E. Dial, and Thomas A. Ohki. Tomography via correlation of noisy measurement records. *Phys. Rev. A*, 91:022118, Feb 2015.

[48] Michael J. Peterer, Samuel J. Bader, Xiaoyue Jin, Fei Yan, Archana Kamal, Theodore J. Gudmundsen, Peter J. Leek, Terry P. Orlando, William D. Oliver, and Simon Gustavsson. Coherence and decay of higher energy levels of a superconducting transmon qubit. *Phys. Rev. Lett.*, 114:010501, Jan 2015.

[49] Jens Koch, Terri M. Yu, Jay Gambetta, A. A. Houck, D. I. Schuster, J. Majer, Alexandre Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf. Charge-insensitive qubit design derived from the cooper pair box. *Phys. Rev. A*, 76:042319, Oct 2007.