Optimized SWAP networks with equivalent circuit averaging for QAOA

Akel Hashim,[1,2,3,*] Rich Rines,[4,*] Victory Omole,[4] Ravi K. Naik,[1,3] John Mark Kreikebaum,[1,5,†] David I. Santiago,[3] Frederic T. Chong,[4,6] Irfan Siddiqi,[1,3,5] and Pranav Gokhale[4,‡]

[1]*Quantum Nanoelectronics Laboratory, Department of Physics, University of California at Berkeley, Berkeley, California 94720, USA*
[2]*Graduate Group in Applied Science and Technology, University of California at Berkeley, Berkeley, California 94720, USA*
[3]*Computational Research Division, Lawrence Berkeley National Lab, Berkeley, California 94720, USA*
[4]*Super.tech, a division of ColdQuanta, Chicago, Illinois 60615, USA*
[5]*Materials Sciences Division, Lawrence Berkeley National Lab, Berkeley, California 94720, USA*
[6]*University of Chicago, Chicago, Illinois 60637, USA*

(Received 11 November 2021; accepted 18 May 2022; published 11 July 2022)

The SWAP network is a qubit routing sequence that can be used to efficiently execute the Quantum Approximate Optimization Algorithm (QAOA). Even with a minimally connected topology on an n-qubit processor, this routing sequence enables $\mathcal{O}(n^2)$ operations to execute in $\mathcal{O}(n)$ steps. In this work, we optimize the execution of SWAP networks for QAOA through two techniques. First, we take advantage of an overcomplete set of native hardware operations [including 150-ns controlled-$\frac{\pi}{2}$ phase gates with up to 99.67(1)% fidelity] to decompose the relevant quantum gates and SWAP networks in a manner which minimizes circuit depth and maximizes gate cancellation. Second, we introduce equivalent circuit averaging, which randomizes over degrees of freedom in the quantum circuit compilation to reduce the impact of systematic coherent errors. Our techniques are experimentally validated at the Advanced Quantum Testbed through the execution of QAOA circuits for finding the ground state of two- and four-node Sherrington-Kirkpatrick spin-glass models with various randomly sampled parameters. We observe a ∼60% average reduction in error (total variation distance) for QAOA of depth $p = 1$ on four transmon qubits on a superconducting quantum processor.

DOI: 10.1103/PhysRevResearch.4.033028

I. INTRODUCTION

A key challenge for scaling near-term quantum computers to address practical problems is limited qubit connectivity. While qubit mapping techniques can mitigate this limitation, recent results suggest that any mismatch between hardware connectivity and the connectivity required for specific applications can erase the potential for a quantum speedup [1,2]. This poses a particular challenge for superconducting quantum hardware, which—despite the advantages of fast operation speed, high gate fidelity, and scalable fabrication—generally has the disadvantage of sparse nearest-neighbor qubit connectivity.

The SWAP network, introduced in Ref. [3] and studied further in Refs. [4,5], offers a promising path forward for coping with limited connectivity. In fact, the SWAP network requires only minimal linear connectivity between qubits; any additional qubit couplings are unnecessary. This property is well suited to superconducting qubits where it has the additional advantage of minimizing the effect of crosstalk due to frequency crowding [6].

Qubit routing in an n-qubit SWAP network follows a sequence of $n - 1$ steps, incurring $\mathcal{O}(n)$ total quantum circuit depth. This linear cost suffices to carry out all $\mathcal{O}(n^2)$ pair-wise interactions between qubits, even for linearly arranged qubits. By contrast, naive qubit routing approaches would require $\mathcal{O}(n^3)$ circuit depth to perform all of the necessary operations because each of the $\mathcal{O}(n^2)$ pair-wise interactions would be serialized and would incur an $\mathcal{O}(n)$ SWAP overhead. The quadratic advantage in circuit depth offered by SWAP networks persists even in comparison to state-of-the-art qubit routing [7].

There are numerous applications of SWAP networks, broadly corresponding to evolution under a fully connected Hamiltonian comprising interactions between all possible pairs or subsets of qubits [4]. Examples include the simulation of the Sherrington-Kirkpatrick spin-glass model [8], Max-Cut for use cases like VLSI circuit design [9], and k-means clustering on large datasets with coresets [7]. Furthermore, Hamiltonian evolution is at the heart of many noisy intermediate-scale quantum (NISQ) [10] algorithms such as the Quantum Approximate Optimization Algorithm (QAOA) [11] and its derivatives [12–14], making the implementation of SWAP networks invaluable for near-term applications. In addition, SWAP networks will be favorable for noise mitigation approaches involving virtual distillation [15], in which

*These authors contributed equally to this work.
†Now at Google Quantum AI, Mountain View, CA, USA.
‡pranav.gokhale@coldquanta.com

multiple copies of a quantum state can be arranged in parallel registers with linear connectivity [16].

Given the fundamental importance of SWAP networks to many quantum applications, it is important to fully optimize their execution. Here, we introduce and apply two compilation techniques that improve the performance of SWAP networks for QAOA. The first technique employs a richer gateset than enabled by standard quantum assembly (QASM) representation for circuit decomposition. The second technique, which we term equivalent circuit averaging (ECA), involves randomizing circuit decomposition over degrees of freedom in compilation to mitigate the impact of systematic coherent errors. Both of these techniques are validated at the Advanced Quantum Testbed (AQT) at Lawrence Berkeley National Laboratory.

The rest of this paper is organized as follows. Section II describes the Advanced Quantum Testbed's hardware. Section III presents our optimized gate decompositions for the Hadamard, SWAP, and ZZ-SWAP operations. Section IV presents results from cycle benchmarking of our optimized gate sequences. Section V examines the application of SWAP networks to QAOA, and Sec. VI introduces equivalent circuit averaging for this application. Section VII concludes. Appendixes A and B detail single-qubit and two-qubit parameters for the Advanced Quantum Testbed. Finally, Appendix C presents examples of the full QAOA SWAP network circuits that we executed.

II. ADVANCED QUANTUM TESTBED

The experiments in this work were performed on four fixed-frequency transmon [17] qubits (labeled Q4, Q5, Q6, and Q7; see Table III in Appendix A) on an eight-qubit superconducting quantum processor (AQT@LBNL Trailblazer8-v5.c2) at the Advanced Quantum Testbed [18]. The eight qubits are coupled to nearest neighbors via fixed-frequency resonators in a ring geometry, thus the four qubits used in this work have linear connectivity.

Arbitrary single-qubit SU(2) gates are typically implemented using physical $X_{\pi/2}$ gates (via resonant Rabi-driven pulses) and virtual Z_θ gates (via phase shifts between physical pulses) [19]

$$U(\alpha,\beta,\gamma) = Z_{\alpha-\pi/2}X_{\pi/2}Z_{\pi-\beta}X_{\pi/2}Z_{\gamma-\pi/2}, \tag{1}$$

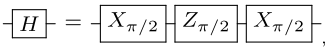
this $ZXZXZ$-decomposition reduces the time and complexity involved in calibrating and benchmarking single-qubit gates. The disadvantage is that every computational single-qubit gate is actually composed of two physical $X_{\pi/2}$ pulses, each 30 ns in duration; thus, every single-qubit gate (cycle) in a circuit takes 60 ns by default, even if the gate could be implemented with only a single $X_{\pi/2}$ pulse. This needlessly increases circuit depth, leaving the qubits more susceptible to decoherence.

Two-qubit entangling operations are achieved using a tunable ZZ-coupling via off-resonant drives [20,21] between neighboring qubits, which is used to implement controlled-Z (CZ) operations between adjacent qubit pairs, as well as controlled-S (CS) and controlled-$S^\dagger$ ($CS^\dagger$) gates. The duration of our two-qubit CZ gate is 200 ns, which is limited by the drive-induced decoherence discussed in Ref. [20]. However, because the CS or $CS^\dagger$ gate performs half the rotation of a CZ, it can be implemented in less time than the CZ. We calibrate and measure a process infidelity of $4.3(1)\times10^{-3}$ for a 150-ns CS gate between qubits (Q5, Q6), and process infidelities of $5.0(1)\times10^{-3}$ and $3.3(1)\times10^{-3}$ for 150 ns $CS^\dagger$ gates between qubits (Q4, Q5) and (Q6, Q7), respectively (see Table IV in Appendix B), which is ∼100 ns faster with an error rate that is ∼2× lower than previously measured for superconducting qubits [22].

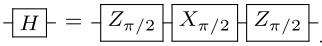
III. OPTIMIZED GATE DECOMPOSITIONS

A. Optimized Hadamard and SWAP

We first optimize decompositions for the Hadamard (H) and SWAP operations. The H gate has two equivalent decompositions using the $\{X_{\pi/2}, Z_\theta\}$ basis:



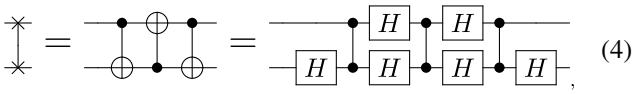
(2)



(3)

The standard $ZXZXZ$-decomposition of the H gate corresponds to Eq. (2). While this is a valid decomposition, Eq. (3) is preferable because it requires a single physical $X_{\pi/2}$ pulse instead of two. Therefore, the optimized Hadamard halves the duration of the gate, taking only 30 ns instead of 60 ns.

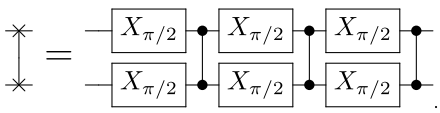
Next, we consider the SWAP operation. The standard decomposition of the SWAP is



(4)

where ⋮ indicates the CZ gate. The middle circuit represents the standard QASM decomposition of a SWAP, which involves three alternating CX gates. No further decompositions are possible until we go below the level of QASM [23]. Each CX decomposes to AQT's two-qubit CZ basis gate by invocation of the identity $CX(q_c,q_t) = H(q_t)CZ(q_c,q_t)H(q_t)$, where q_c (q_t) is the control (target) qubit. We immediately see that applying the optimized H leads to an improvement: the total SWAP duration is reduced by 4×30 ns $= 120$ ns, and the number of required $X_{\pi/2}$ physical pulses is halved from 12 to 6.

We can optimize even further by applying a transposition identity to move the bottom-right H to the top-left. This identity reduces the total SWAP duration by an additional 30 ns since the two "edge" H gates become parallelized. After annihilating all virtual rotations arising from Eq. (3) via commutation identities, we have the final optimized SWAP



(5)

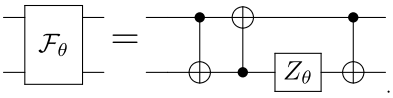
We deployed these optimized Hadamard and SWAP decompositions through the SUPERSTAQ platform [24], which can target AQT hardware. Section IV presents cycle benchmarking [25] results for these optimizations.

B. Background on ZZ-SWAP

The core operation needed in a QAOA SWAP network is the *ZZ*-SWAP gate, which is equivalent to a *ZZ* interaction followed by a SWAP operation and is defined as the unitary operation below with input parameter θ:

$$\mathcal{F}_\theta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & e^{i\theta} & 0 \\ 0 & e^{i\theta} & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6}$$

The standard QASM-decomposed quantum circuit implementation of $\mathcal{F}_\theta$ comprises three CX gates and a single-qubit Z_θ rotation [7]:



(7)

It is possible to boost performance beyond this decomposition by leveraging knowledge of the target hardware's underlying native gateset. For example, the authors of Ref. [26] compiled the *ZZ*-SWAP operation directly down to three native two-qubit Sycamore (SYC) gates, rather than recompiling each CX down to SYC gates. In a related manner, the authors of Ref. [27] developed an efficient parametric implementation of SWAP networks via access to a native $XY(\theta)$ gate (with duration independent of θ) and a native CZ gate.

However, in these examples, the total duration of the *ZZ*-SWAP operation is always constant, regardless of θ. This leaves room for improvement. For example, it was shown that access to a parametric CZ [i.e. CPHASE(ϕ)] yields significant improvements for the decomposition of many quantum operations [28]; in the next subsection we demonstrate that this is true for the *ZZ*-SWAP operation as well. However, there are experimental obstacles to tuning a high-fidelity parametric gate with variable duration. For example, the authors of Ref. [29] noted ramp effects at small θ for a parametric cross-resonance gate.

Rather than incurring the calibration overhead of a parametric gate with variable duration like CPHASE(ϕ), we instead focus on the optimization opportunities from an *overcomplete* discrete two-qubit gate set, which contains additional two-qubit gates beyond what is necessary for universal quantum computation. Concretely, we next examine the optimized $\mathcal{F}_\theta$ decompositions possible when we have access to both a CZ and $CS = \sqrt{CZ}$ gate, where the CS is faster than the CZ gate.

In typical applications, multiple *ZZ*-SWAP gates are arranged into a nearest-neighbor *ZZ*-SWAP *network* that carries out $t = 0, \ldots, n-1$ steps. Each step alternates between an odd and even pattern. At steps with odd t, each neighboring qubit pair with indices $(2k, 2k+1)$ for $k \in [0, \frac{n}{2}]$ is entangled in accordance with a target Hamiltonian and then SWAPped. Even-t steps perform this interaction for qubit pairs with indices $(2k+1, 2k+2)$. Note that each step is highly parallel, with $\sim n/2$ operations occurring simultaneously. A prototypical example is shown in Fig. 1, which implements the Hamiltonian evolution $e^{i\gamma H}$ corresponding to a Sherrington-Kirkpatrick spin-glass model $H = \sum_{i<j<n} J_{ij} Z_i Z_j$ on $n = 4$ nodes. The utility of a SWAP network is that it efficiently

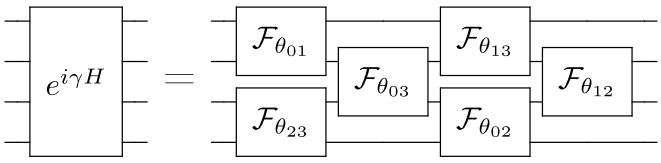


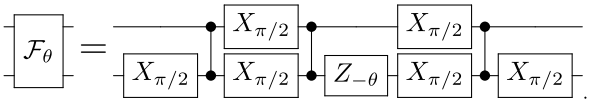
FIG. 1. SWAP network implementing the Hamiltonian evolution $e^{i\gamma H}$ for a four-node Sherrington-Kirkpatrick model $H = \sum_{i<j<4} J_{ij} Z_i Z_j$, where $\theta_{ij} = \gamma J_{ij}$. Note that the qubit order is reversed after the operation.

generates an all-to-all interaction with a linear-depth circuit of nearest-neighbor interactions, where each interaction is a single *ZZ*-SWAP gate corresponding to one of the commuting weight-2 terms in H.

C. Optimized ZZ-SWAP Gates

We now introduce optimized, θ-dependent decompositions of the *ZZ*-SWAP gate $\mathcal{F}_\theta$, taking advantage of an overcomplete two-qubit gate set consisting of both CZ and CS or $CS^\dagger$ gates.

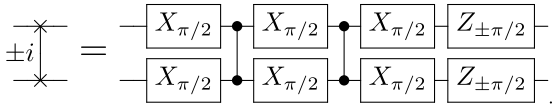
For $\theta \in \{0, \pi\}$, the *ZZ*-SWAP unitary in Eq. (6) is equivalent (up to virtual phases) to the standard SWAP gate, and so can be decomposed using the optimized SWAP described in Sec. III A. In the general case, using the optimized Hadamard [cf. Eq. (3)] and virtual Z rotations, the baseline *ZZ*-SWAP decomposition still requires three CZ and six $X_{\pi/2}$ gates



(8)

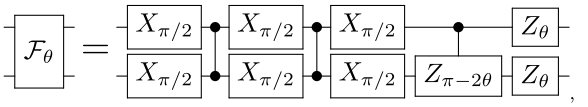
Unfortunately, in the general case the first and final $X_{\pi/2}$ gates in Eq. (8) cannot be parallelized as they are in the optimized SWAP, and so both contribute to the depth of the standalone *ZZ*-SWAP circuit. In Sec. III D, we will show that this overhead can be mitigated in the context of a full SWAP network for QAOA.

A second special case exists for $\theta = \pm\pi/2$, in which the *ZZ*-SWAP is equivalent to the iSWAP (iSWAP†) gate and requires just two CZ gates. Again employing the optimized Hadamard, it can be decomposed as



(9)

If we have access to a parameterized CPHASE(ϕ) gate, we can naturally generalize the optimized SWAP and iSWAP decompositions to all *ZZ*-SWAP circuits. Setting $\phi = \pi - 2\theta$,



(10)

correctly generates the three-CZ optimized SWAP for $\theta \in \{0, \pi\}$ and the two-CZ optimized iSWAP for $\theta = \pm\pi/2$. Assuming the gate time of any CPHASE(ϕ) gate to be proportional to ϕ for intermediary $0 \leqslant \phi \leqslant \pi$, the total time for $\mathcal{F}_\theta$ is the same as $2 + 2|\theta \bmod \pi - \pi/2|/\pi$ CZ gates.

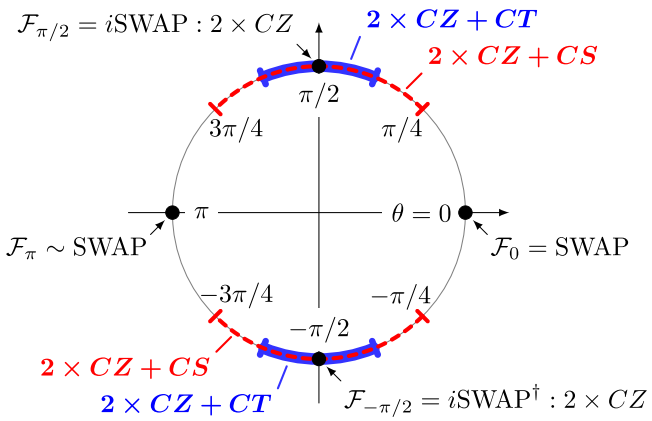
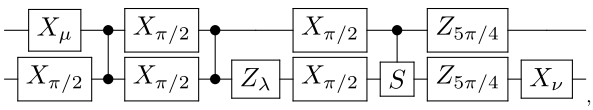


FIG. 2. CZ-depth of optimized $\mathcal{F}_\theta$ decompositions. The angle θ defines the ZZ-SWAP unitary, with $\theta = 0$ (π) corresponding to the standard SWAP gate (up to virtual phases) and $\theta = \pi/2$ ($-\pi/2$) corresponding to the iSWAP (iSWAP†) gate. All $\mathcal{F}_\theta$ can be implemented with three CZ gates using Eq. (8). However, only two CZ gates are needed for iSWAP and iSWAP† gates. For $\theta \in [\pm\pi/4,\ \pm 3\pi/4]$ (red), $\mathcal{F}_\theta$ can be implemented with two CZ and one CS gate; whereas for $\theta \in [\pm 3\pi/8,\ \pm 5\pi/8]$ (blue), $\mathcal{F}_\theta$ can be implemented with two CZ and one CT gate.

Absent a fully parameterized CPHASE(ϕ) operation, we can generalize Eqs. (8) and (10) to implement subsets of $\mathcal{F}_\theta$ using only discrete values of ϕ. The Cartan decomposition in SU(4) shows that any two-qubit gate is locally equivalent (i.e., equivalent up to a transformation involving single-qubit gates) to a unitary operator $e^{ic_1 XX + ic_2 YY + ic_3 ZZ}$, where the locally invariant coordinates (c_1, c_2, c_3) fully characterize the nonlocal properties of the gate [30]. The invariant coordinates of CPHASE(ϕ) and $\mathcal{F}_\theta$ are $(\phi/4, 0, 0)$ and $(\pi/4, \pi/4, |\pi - 2\theta|/4)$, respectively. The latter can therefore be constructed from two CZ gates and any one CPHASE(ϕ) with $\phi \geqslant |\pi - 2\theta|$, along with the appropriate single-qubit gates.

As shown in Fig. 2, by choosing $\phi = \pi/2$ (the CS gate) we can implement half of all possible $\mathcal{F}_\theta$ gates with the equivalent of 2.5 CZ gates (that is, two CZ gates and one CS gate). For any $\pi/4 \leqslant \theta \bmod \pi \leqslant 3\pi/4$, we have



(11)

where

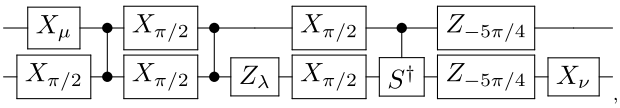
$$\lambda = \cos^{-1}(-\sqrt{2}\cos\theta) \cdot \text{sgn}(\sin\theta), \tag{12}$$

enforces the local equivalence between Eq. (11) and $\mathcal{F}_\theta$, and

$$\mu = \csc^{-1}(-\sqrt{2}\sin\theta), \tag{13}$$

$$\nu = -\cos^{-1}(\cot\theta), \tag{14}$$

provide the necessary local corrections. An equivalent decomposition can be constructed using a $CS^\dagger = \text{CPHASE}(-\pi/2)$ in place of the CS:



(15)

The X_μ and X_ν gates can each be implemented with two $X_{\pi/2}$ pulses and virtual phases using Eq. (1); however, this complexity can be mitigated within a full SWAP network for QAOA (Sec. III D).

Additional controlled-Z_ϕ operations for fixed values of ϕ would further refine the optimized $\mathcal{F}_\theta$ decomposition toward the lower bound provided by fully parameterized CPHASE(ϕ). For example, as shown in Fig. 2 one quarter of all possible $\mathcal{F}_\theta$ are reachable using a $CT = \text{CPHASE}(\pi/4) = \sqrt[4]{CZ}$ gate, and so the addition of a CT to the gateset would reduce the CZ depth of these decompositions to the equivalent of 2.25 CZ gates.

D. Optimized SWAP Networks for QAOA

We can further simplify the decomposition of ZZ-SWAP gates in the context of the larger SWAP network. Various discrete and continuous symmetries in the ZZ-SWAP operation result in the following degrees of freedom to its optimized decomposition:

(1) $\mathcal{F}_\theta = (1 \otimes X)\mathcal{F}_{-\theta}(X \otimes 1)$,
(2) $\mathcal{F}_\theta = (Z \otimes Z)\mathcal{F}_{\theta+\pi}$,
(3) $\mathcal{F}_\theta(q_0, q_1) = \mathcal{F}_\theta(q_1, q_0)$ (qubit interchange),
(4) $\mathcal{F}_\theta = \mathcal{F}_{-\theta}^\dagger$,
(5) $\mathcal{F}_\theta = (Z_{-\vartheta} \otimes Z_{-\varphi})\mathcal{F}_\theta(Z_\varphi \otimes Z_\vartheta)\ \forall\ \vartheta, \varphi \in \mathbb{R}$,

where ϑ, φ are continuous parameters. Symmetry 4 is useful only for $\mathcal{F}_\theta$ gates implemented using a CS or $CS^\dagger$, in which case it can be used to reverse the order of entangling gates (and corresponding single-qubit gates X_μ and X_ν) in the circuit. (For the 3-CZ decomposition of $\mathcal{F}_\theta$, the physical implementations of $\mathcal{F}_\theta$ and $\mathcal{F}_{-\theta}^\dagger$ are identical.) Because any $\mathcal{F}_{\pm\theta}$ admitting the CS decomposition in Eq. (11) also admits the $CS^\dagger$ decomposition in Eq. (15), we are always able to construct both $\mathcal{F}_\theta$ and $\mathcal{F}_{-\theta}^\dagger$ using either one of these gates. We therefore calibrate only one of $\{CS, CS^\dagger\}$ per pair of qubits (see Appendix B), and use symmetry 4 to reverse the order of the interactions (i.e., whether or not the CS or $CS^\dagger$ interaction precedes or succeeds the two CZ gates).

We use an automated scheduler which computes the set of logically equivalent decompositions generated by symmetries 1 to 4 for each gate in the network. For each decomposition, the continuous parameters ϑ, φ (symmetry 5) are determined analytically so as to minimize the cost of just the single-qubit gates immediately preceding the gate (the trailing $Z_{-\varphi}$ and $Z_{-\vartheta}$ can be absorbed into any subsequent gates using the same symmetry, so these parameters can be optimized for each gate independently). The scheduler searches for a sequence of these decompositions which minimizes overall circuit depth (corresponding to $X_{\pi/2}$ count in the critical path). Because the combinatorial search space grows exponentially in the number of ZZ-SWAP gates, finding an exact minimum is in general intractable. For QAOA we are required to generate and execute many circuits while varying classical parameters, making

TABLE I. Comparison of four-node $p = 1$ QAOA circuits generated with and without the optimized scheduling routines described in Sec. III D. The optimized scheduler minimizes the number and depth of $X_{\pi/2}$ gates in each circuit, and further reduces the depth of the $CZ + CS$ circuits relative to those using only the CZ decomposition.

		$X_{\pi/2}$ Count	
Decomposition	Scheduler	(total)	(crit. path)
CZ	Baseline	44	14
	Optimized	34	11
$CZ + CS$	Baseline	51	18
	Optimized	38	11

classical optimization time especially important. However, the regular structure of the SWAP network makes it possible to find nearly optimal sequences using simple heuristics and scalable local search procedures.

For networks using only the CZ decomposition of $\mathcal{F}_\theta$ [Eq. (8)], the two outer $X_{\pi/2}$ gates of each decomposition are independent of θ. Simple heuristics can then be used to find sequences which maximally annihilate these outer gates: specifically, we can use symmetry 3 to align pairs of $X_{\pi/2}$ between layers of ZZ-SWAPs, which can then be annihilated with X or Z gates injected via symmetry 1, 2, or 5. Applying this strategy to four-qubit $p = 1$ QAOA circuits, we find that we can reduce the number of $X_{\pi/2}$ gates both overall and in the critical path by over 20% (see Table I). Further, because the annihilation of these gates is independent of the rotation angles of each gate, this optimization does not need to be repeated as we vary the classical parameters in subsequent QAOA iterations.

Optimizing networks using the CS decomposition of $\mathcal{F}_\theta$ is more challenging because X_μ and X_ν both depend on θ and do no naturally annihilate one another. We begin with a simple scheduling pass which chooses from the equivalent decompositions of each ZZ-SWAP considering just the single-qubit gates immediately preceding it. Typically multiple such decompositions are equally good in terms of minimizing single-qubit gate cost; in this case we defer the selection until the decomposition of a subsequent gate on that qubit and then select one which allows for the best decompositions of the later gate. Though the resulting circuit is not guaranteed to be optimal, we find that by combining this localized search routine and simple heuristics to determine the order in which gates are expended we can achieve more than a 25% reduction in the total $X_{\pi/2}$ count and a 33% reduction in $X_{\pi/2}$ gates along the critical path for the four-node QAOA circuit (Table I). Notably, the resulting single-qubit gate complexity is comparable to that of the CZ-only circuits, indicating that we successfully mitigated the burden of X_μ and X_ν in Eqs. (11) and (15). The required optimization time is linear in the number of ZZ-SWAP gates and at most quadratic in the number of equivalent decompositions of each gate. In this case the optimal sequence of decompositions depends on each ZZ-SWAP's rotation angle, and so will have to be updated as classical parameters are varied between QAOA iterations. However, small changes will often only require recomputing the values ϑ, φ for each gate, significantly reducing this classical overhead as QAOA converges on optimal parameter values.

IV. CYCLE BENCHMARKING OF OPTIMIZED DECOMPOSITIONS

To benchmark the performance of the optimized pulse sequences relative to their standard decompositions, we utilize cycle benchmarking [25] (CB), a scalable protocol for measuring the performance of parallel gate cycles. Cycle benchmarking differs from randomized benchmarking [31–34] (RB) in two keys ways: (i) it utilizes Pauli twirling instead of Clifford twirling, which maps gate errors into stochastic Pauli channels (instead of a global depolarizing channel); (ii) CB benchmarks the performance of quantum gates performed in parallel, providing a measure of their performance in the context of multiqubit quantum algorithms. In contrast, benchmarking the individual constituent gates of multiqubit cycles was shown to be a poor predictor of the global performance of quantum circuits [35] due to the presence of coherent errors and crosstalk between qubits, and because such benchmarks fail to capture errors on (or incurred by) idling spectator qubits [36].

CB measures the process fidelity of a target cycle by preparing the system in a Pauli basis state (e.g., $XYIZ$ for four qubits), and measuring the exponential decay as a function of sequence depth. A separate exponential decay of the form Ap_P can be fit for each basis preparation and measurement state P (i.e., Pauli channel), where A is the state-preparation and measurement (SPAM) parameter and p the fit parameter. Much like interleaved randomized benchmarking [37] (IRB), in which the target gate is interleaved between random Clifford gates, CB interleaves the target cycle between cycles of random single-qubit Pauli gates. Therefore, CB measures the process fidelity of a dressed cycle, which contains the errors due to the interleaved target cycle as well as the Pauli twirling gates. The total process fidelity is the average over K Pauli channels

$$F = \frac{1}{K}\sum_{P\in\mathcal{P}} p_P, \tag{16}$$

where the number of Pauli channels $K = |\mathcal{P}| \leqslant 4^n$ (n qubits) in the set $\mathcal{P}$ that are sampled out of the full 4^n possible states sets the precision of the fidelity estimate [25]. The process infidelity of the dressed cycle is therefore given as $e_D = 1 - F$. To separate the infidelity of the target cycle from the twirling gates, we measure the CB fidelity of the "all-identity" reference cycle, which equates to benchmarking the average performance of only the Pauli twirling gates. Similar to IRB, we can use this to estimate the process infidelity of the target (T) cycle by taking the ratio of the process fidelities of the dressed (D) and reference (I) cycles,

$$e_T = \frac{d-1}{d}\left(1 - \frac{F_D}{F_I}\right), \tag{17}$$

where $d = 2^n$ is the dimension of the system. Using CB has been shown to tighten the upper and lower bounds on the fidelity estimate of the interleaved cycle relative to IRB [20], which can span orders of magnitude [38]. We use this method

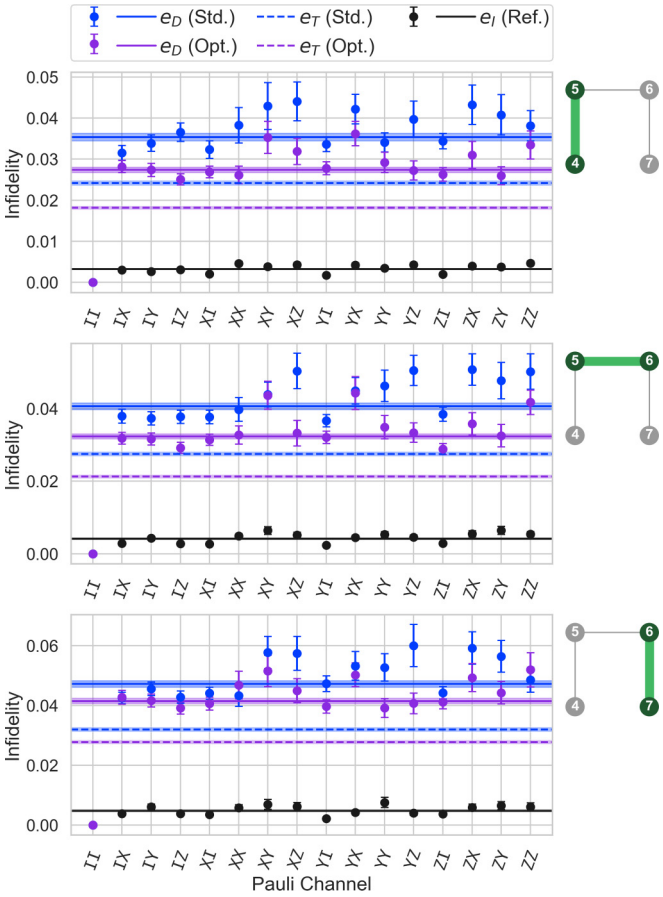


FIG. 3. Improved SWAP gates via gate-based optimizations. The process infidelity of the optimized SWAP gate between Q4 and Q5 (top), Q5 and Q6 (middle), and Q6 and Q7 (bottom) is lower than the process infidelity of the standard decomposition due to the elimination of unnecessary gates. The blue, purple, and black data points represent the Pauli infidelity $1 - p_P$ for each Pauli channel P for the standard (Std.) SWAP gate, optimized (Opt.) SWAP gate, and reference (Ref.) cycle, respectively. The solid blue (purple) line is the average process infidelity e_D of the dressed cycle for the standard (optimized) SWAP gate and the solid black line is the average process infidelity e_I of the Pauli twirling operators. The dashed blue (purple) line is the process infidelity e_T of target cycle (i.e., SWAP gate) estimated via Eq. (17) for the standard (optimized) SWAP sequence. The semi-transparent bands around the average process infidelities represent the 95% confidence interval of the estimates.

to estimate a target infidelities for CZ, CS, and $CS^\dagger$ gates (see Table IV in Appendix B).

In Fig. 3, we plot the CB results for the standard and optimized SWAP gates [see Eq. (5)] between all three qubit pairs. We see that the optimized target cycle infidelity e_T of the SWAP gates is reduced 25%, 23%, and 13% relative to the standard SWAP gate for (Q4, Q5), (Q5, Q6), and (Q6, Q7), respectively. This average improvement can generally be expected for circuits utilizing basic SWAP gates; however, further optimizations can be implemented in the context of full QAOA SWAP networks with the replacement of one of the CZs in the SWAP gate with a CS or $CS^\dagger$ gate, as outlined in the previous section. Furthermore, while the benchmarking results in Fig. 3 show improvements in the SWAP gates between all qubit pairs, they do not capture what improvements can be expected for cycles of gates in any four-qubit application. In Table II, we compare the benchmarked process infidelities of optimized gate cycles versus the standard decompositions for relevant cycles appearing in four-qubit QAOA SWAP networks (see Sec. V). These include the all-Hadamard cycle for basis preparation and converting CZs to CXs, the relevant multiqubit gate cycles appearing in the QAOA SWAP networks and the parallel SWAP cycles incorporating the optimizations outlined in the previous section. We see universal improvement in the target cycle infidelity e_T for the optimized cycles, with reductions in e_T ranging from 64% for the all-Hadamard cycle to 12% for the SWAP cycle.

In addition, we list the duration of each cycle and use this to approximate an upper bound on the error rate due to T_1. We see optimized cycles provide a reduction in the error due to T_1 ranging from 50% for the all-Hadamard cycle to 17% for the SWAP cycle. The remaining improvement is likely due to reduction in coherent errors. For example, the CS and $CS^\dagger$ gates perform a smaller rotation and thus generally require a smaller pulse amplitude, which can lead to less crosstalk on neighboring qubits. These results demonstrate that simple improvements in circuit decomposition and gate optimizations can lead to dramatic improvements in benchmarked gate and cycle performance. Next, we highlight how these fidelity improvements can lead to performance improvements in SWAP networks for QAOA.

V. APPLICATION BENCHMARKING OF QAOA

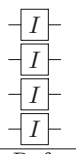
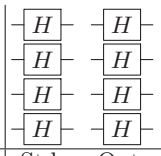
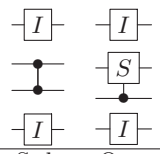
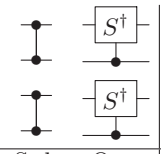
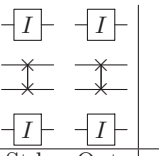
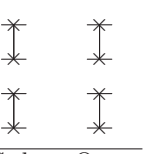
The Quantum Approximate Optimization Algorithm (QAOA) [11] describes a variational ansatz for solving combinatorial optimization problems described by an objective Hamiltonian H. QAOA is characterized by a hyperparameter p that specifies the depth of the ansatz. Specifically, the ansatz is $e^{i\beta_p B}e^{i\gamma_p H}\ldots e^{i\beta_1 B}e^{i\gamma_1 H}$, where $B = \sum_i X_i$ is a mixing Hamiltonian and $\vec{\gamma}$, $\vec{\beta}$ represent $2p$ classically optimized variational parameters. It is believed that QAOA is hard to approximate even at $p = 1$ and is therefore a leading candidate for demonstrations of quantum advantage [39]. We generate QAOA circuits corresponding to Sherrington-Kirkpatrick spin-glass model Hamiltonians with edge weights J_{ij} randomly selected from ± 1 (see Appendix C for the exact symbolic form of the circuits). Each $e^{i\gamma H}$ is then implemented with a network of ZZ-SWAP gates $\mathcal{F}_{\pm\gamma}$. Parameters β_i, γ_i are sampled uniformly from $[0, 2\pi)$.

In Fig. 4, we measure two-qubit ($p = 1$) and four-qubit ($p = 1$ and $p = 2$) QAOA circuits (see Appendix C for example circuits) for various angles γ and benchmark the performance using the total variation distance (TVD)

$$D(p, q) = \frac{1}{2}\sum_{x\in X}|p_x - q_x|, \tag{18}$$

where p_x is the probability of measuring a bit string x in a set X and q_x is the ideal (noiseless) probability. We see that the optimized (Opt.) circuits generally provide more

TABLE II. Benchmarked improvements in optimized cycles. All optimized (Opt.) cycles have a lower CB process infidelity than their respective standard (Std.) decompositions. The duration of each cycle is calculated using the two-qubit gate times listed in Table IV and 30 ns for every $X_{\pi/2}$ gate. An approximate upper bound on the error rate due to T_1 is calculated via $(1-\prod_{q\in\{4,5,6,7\}} e^{-t/T_{1,q}})/2$ using the duration t of each cycle listed above and the T_1 times listed in Table IV for each qubit q, which assumes that T_1 events are independent across all qubits. (The factor of $1/2$ accounts for the fact that, averaged over all possible input states, the qubits only spend half of the time in an excited state.)

Cycle											
Error Rate	Ref.	Std.	Opt.	Std.	Opt.	Std.	Opt.	Std.	Opt.	Std.	Opt.
e_I (10^{-3})	9.6(6)										
e_D (10^{-2})		1.5(1)	1.16(6)	2.11(7)	1.67(8)	3.4(1)	2.09(8)	9.6(7)	6.3(2)	11.7(4)	10.4(4)
e_T (10^{-2})		0.5(1)	0.19(8)	1.09(9)	0.68(9)	2.3(1)	1.07(9)	8.1(7)	5.1(2)	10.2(3)	9.0(4)
Reduction in e_T		64%		38%		53%		38%		12%	
Cycle Duration (ns)		60	30	200	150	200	150	840	690	840	690
T_1 Error $(10^{-2})\sim$		0.21	0.11	0.68	0.52	0.68	0.52	2.8	2.3	2.8	2.3
Reduction in T_1 Error		50%		25%		25%		17%		17%	

accurate performance relative to the standard (Std.) decompositions, reducing the average TVD from $D_{\text{Std.}} = 0.20(5)$ to $D_{\text{Opt.}} = 0.14(3)$ for four-qubit QAOA circuits of depth $p = 1$ and from $D_{\text{Std.}} = 0.23(4)$ to $D_{\text{Opt.}} = 0.22(6)$ for circuits of depth $p = 2$. For two-qubit networks, the optimized circuits outperform the standard circuits on average for qubits (Q5, Q6), but perform worse [equivalent] for (Q4, Q5) [(Q6, Q7)]. We conjecture that the failure of the optimized circuits to outperform the standard circuits for qubits (Q4, Q5) and (Q6, Q7) is due to systematic coherent errors whose impacts can dominate algorithm performance and are not accurately captured by randomized benchmarks (see the discussion in Sec. VI).

The parameter angle γ determines what gate optimizations can be implemented for each network, with $\pi/4 \leqslant \gamma \leqslant 3\pi/4$ and $5\pi/4 \leqslant \gamma \leqslant 7\pi/4$ defining the angles for which CS or $CS^\dagger$ gates can be used in place of CZ gates (see Fig. 2). The γ values are randomly chosen in Fig. 4, but they are seeded such that half of the circuits tested can take advantage of the CS or $CS^\dagger$ decomposition of ZZ-SWAP at $p = 1$. Values of β are chosen uniformly at random. We note that in regions where CS or $CS^\dagger$ gates can be used for two (four) qubits, the standard (optimized) decompositions outperform the optimized (standard) decompositions on average. This is likely due to the fact that the small improvements in two-qubit gate fidelities provided by the CS or $CS^\dagger$ gates (see Table IV) are unlikely to provide any significant performance guarantees for two-qubit circuits in the presence of coherent errors. In contrast, the large improvements in cycle fidelity for multiqubit ($n > 2$) cycles containing CS or $CS^\dagger$ gates (cf. Table II) are much more likely to provide robust performance guarantees for multiqubit circuits, even in the presence of coherent errors. These results demonstrate that simple changes to circuit decomposition and gate-based optimizations can lead to clear improvements in algorithm and application performance, especially in multiqubit circuits, highlighting the importance of smart compilers and more continuous gatesets in the NISQ era.

VI. EQUIVALENT CIRCUIT AVERAGING

One limitation of benchmarking the average performance of gates or cycles is that randomized benchmarks are not accurate predictors of the global performance of structured quantum circuits due to the presence of coherent errors [35]. When averaging over a twirling group, such as the Clifford (Pauli) group for RB (CB), all errors are converted into a global depolarizing (stochastic Pauli) channel. However, in actual quantum algorithms, the physical error mechanisms are more complex than depolarizing or Pauli channels, as coherent errors can interfere constructively or destructively from one cycle to the next. Therefore, while the optimized pulse sequences show clear improvements in cycle fidelity (cf. Table II) measured via CB, this does not always guarantee improvements in the performance of algorithms composed of these cycles. This can be seen in Fig. 4, in which the standard circuit decompositions occasionally outperform the optimized circuits for the four-qubit results and outperform the optimized circuits on average for the two-qubit results for qubits (Q4, Q5).

Being systematic in nature, coherent errors can, in theory, be measured and corrected via recalibration or added compensation pulses. However, the complexity of fully characterizing coherent errors (i.e., context-dependent rotation axes and angles [40]) on multiqubit processors that arise due to classical and quantum crosstalk is intractable, and no known scalable methods exist for doing so for systems with continuous single-qubit gate sets. Various methods exist for suppressing coherent errors, such as dynamical decoupling [41] and error-correcting composite pulse sequences [42], or randomization methods for "tailoring" them into stochastic noise, such as Pauli twirling [43–46], Pauli frame randomization [47–49], and randomized compiling [50,51]. However, these methods generally require the modification of single-qubit gates or the inclusion of more gates (e.g., in the case of dynamical decoupling and composite sequences), or require that the two-qubit gates in circuits are Clifford so that inverting Pauli operators can be efficiently computed and applied. Adopting

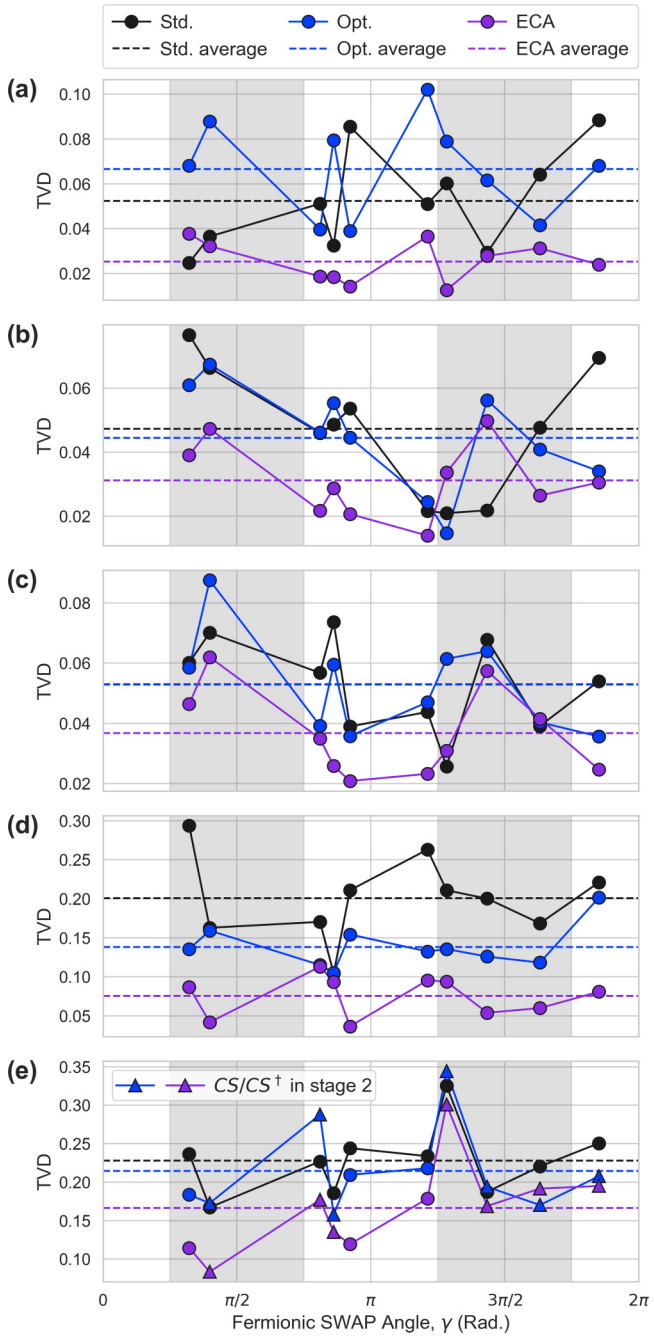


FIG. 4. Improved QAOA SWAP networks via gate-based optimizations. The TVD performance of QAOA networks of angle γ are plotted for qubits (a) (Q4, Q5), (b) (Q5, Q6), and (c) (Q6, Q7), and four-qubit circuits with (d) $p = 1$ and (e) $p = 2$ stages. The mixing parameter β is chosen at random for each network. For (b), (d), and (e), the optimized (blue, Opt.) circuits outperform the standard (black, Std.) on average (dashed lines). The ECA (purple) results consistently outperform both the standard and optimized circuits. The gray shaded regions define the angles for which CS or $CS^\dagger$ gates can be utilized. The results in (e) are plotted against the γ from stage 1 (the triangle markers denote circuits which utilize a CS or $CS^\dagger$ gate in stage 2). (Error bars on the TVD $\sim\mathcal{O}[10^{-3}]$ are smaller than the markers.)

these techniques would therefore require forgoing the circuit optimizations (and corresponding fidelity gains) employed so far in this work—both by necessitating additional $X_{\pi/2}$ pulses and precluding the use of non-Clifford CS and $CS^\dagger$ gates.

A similar strategy was proposed for circuit synthesis methods, in which systematic approximation errors are rendered incoherent by averaging over various circuits near a target unitary generated from ensembles of approximate decompositions [52,53]. We employ this general idea (with systematic errors in the physical gates taking the place of approximation errors) using the space of equivalent *ZZ*-SWAP decompositions generated by the degrees of freedom outlined in Sec. III D. By randomly sampling from these decompositions for each *ZZ*-SWAP gate, we can generate a set of randomized but logically equivalent circuits to average over. However, this unconstrained randomization would require forgoing the single-qubit gate reduction achieved by the optimized scheduler. Instead, we can generate logically equivalent circuits which preserve these gate-level optimizations by following the optimized scheduling procedure outlined in Sec. III D, but with randomness injected into the selection between the equally good decompositions of each gate. Though the constraints on randomization necessary to preserve circuit depth mean that we cannot make solid guarantees on the mitigation of coherent errors, we empirically find that averaging over equivalent circuits generated in this way is an effective strategy for systematic error mitigation. We call this strategy equivalent circuit averaging (ECA). The computational overhead of ECA scales linearly with both the number of logically equivalent circuits to be generated and the cost of optimized scheduling for each circuit (proportional to the number of *ZZ*-SWAP gates in the circuit).

For the circuits in Fig. 4, we generate $M = 20$ logically equivalent optimized circuits for each angle γ (see Appendix C for example circuits). To normalize shot statistics, we measure each equivalent circuit $s = S/M$ times and compute the union over all M results to obtain an equivalent statistical distribution for a circuit measured S times; $S = 10\,000$ and $s = 500$ for the results in Fig. 4. We see that ECA dramatically reduces the TVD on average in comparison to both the standard and optimized results for all of the two- and four-qubit QAOA SWAP network results, reducing the average TVD by $\sim 60\%[26\%]$ from $D_{\text{Std.}} = 0.20(5)$ to $D_{\text{ECA}} = 0.08(2)$ [$D_{\text{Std.}} = 0.23(4)$ to $D_{\text{ECA}} = 0.17(6)$] for the four-qubit $p = 1[p = 2]$ QAOA results, and providing the most accurate measured probability distribution in 88% of all of the two- and four-qubit circuits measured.

While the classical overhead of generating and measuring M logically equivalent circuits increases linearly in M, we observe significant improvements in the measured results. These results demonstrate that ECA is a useful tool for smart compilers which optimize circuit decomposition using various degrees of freedom, and is not limited to circuits only containing two-qubit Clifford gates, adding to the toolbox of randomization methods that can be employed in the NISQ era. We also note that averaging out systematic

errors is likely beneficial even at the expense of gate-level optimizations. The simplest version of ECA, which samples from the set of all logically equivalent decompositions of each *ZZ*-SWAP by randomly applying the symmetries in Sec. III D without consideration of circuit depth would have negligible computational overhead.

VII. CONCLUSION

Quantum compilers play a fundamental role in the translation of abstract quantum circuits to machine instructions in gate-based quantum computing. In the NISQ era, it is necessary to consider the balance between the calibration overhead for large gatesets and optimal circuit decomposition for quantum application performance. In this work, we show that utilizing a smart compiler for canceling unnecessary single-qubit gates is a simple method for improving the performance of quantum circuits. We further demonstrate that, by adding an additional two-qubit gate (*CS* or $CS^\dagger$) to our gateset for each qubit pair, we observe significant improvement in the benchmarked cycle and application performance. While our work focuses on SWAP networks and their application to QAOA, non-Clifford *CS* gates also find importance in universal quantum computation and magic-state distillation for fault-tolerance [54–56]. Furthermore, while we added additional two-qubit gates to our gateset, future work could explore the potential benefits of compiling all circuits down to high-fidelity two-qubit gates which are not fully entangling [57–59], thus removing the need to calibrate multiple two-qubit gates per pair of qubits.

Additionally, we introduce ECA to mitigate the impact of systematic coherent errors in non-Clifford circuits by utilizing the various degrees of freedom of quantum compilers to generate many logically equivalent circuits. Given the difficulty in characterizing and predicting the impact of coherent errors on algorithm performance, such a method negates the need for doing so by assuming that the average over many circuits will reduce the impact of coherent errors on the algorithm results. We demonstrate the effectiveness of this approach with our application benchmark results, in which we find that ECA improves the accuracy of the measured probability distribution for 88% of the randomly generated two- and four-qubit QAOA circuits.

While ECA was employed by taking advantage of the various degrees of freedom in networks of *ZZ*-SWAP gates, a more sophisticated search procedure would likely expand the applicability of our methods for scheduling and generating equivalent circuits for more general applications. We further imagine possible "hybrid" strategies in which ECA is combined with other randomization protocols (e.g., randomized compiling) for maximizing the ways in which logically equivalent circuits can be expressed, thus minimizing residual coherent errors. The cost of ECA (both classically and in terms of single-qubit gate optimization) in the general case and the degree to which it tailors noise in quantum systems (i.e., in the manner of other randomization methods which twirl over a specific gateset) are open questions, which we plan to explore in future work.

Finally, as described in Sec. III C, access to a parameterized CPHASE(ϕ) gate would minimize the *CZ* gate time for any *ZZ*-SWAP gate. The corresponding gate decomposition [Eq. (10)] also avoids the θ-dependent X_μ and X_ν gates in Eq. (11), allowing for more efficient gate cancellation and a greater opportunity for randomness in ECA. The experimental demonstration of this decomposition would be a natural extension of this work and would provide insight into the value of parameterized two-qubit gates for NISQ systems.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research Quantum Testbed Program under Contract No. DE-AC02-05CH11231. This material is also supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under Award No. DE-SC0021526. This work is funded in part by EPiQC, an NSF Expedition in Computing, under Grant No. CCF-1730449; in part by STAQ under grant NSF Phy-1818914; in part by NSF Grant No. 2110860; in part by the U.S. Department of Energy Office of Advanced Scientific Computing Research, Accelerated Research for Quantum Computing Program; and in part by NSF OMA-2016136 and in part based upon work supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers. A.H. acknowledges financial support from the National Defense Science & Engineering Graduate (NDSEG) Fellowship.

R.R. and P.G. devised the optimized circuit decompositions. A.H. conducted the experiments and analyzed the data. ECA was conceived by A.H. and developed by R.R. and P.G. R.K.N., D.I.S., F.T.C., and I.S. supervised all theoretical and experimental work. J.M.K. fabricated the sample. V.O. developed the SUPERSTAQ software interface to the AQT. A.H., R.R., V.O., and P.G. wrote the manuscript with input from all coauthors.

F.T.C., P.G., R.R., and V.O. have a financial interest in the SUPERSTAQ platform. F.T.C. is Chief Scientist for Quantum Software at ColdQuanta and an advisor to Quantum Circuits, Inc. All other authors declare no competing interest.

APPENDIX A: SINGLE-QUBIT PARAMETERS

Table III lists the relevant qubit parameters for the four transmon qubits used in this work. Qubit frequencies and anharmonicities are measured using Ramsey spectroscopy. Relaxation (T_1) and coherence (T_2^* and T_2^{echo}) times are extracted by fitting exponential decay curves to the excited state lifetime and Ramsey spectroscopy measurements (without and with an echo pulse), respectively. Readout fi-

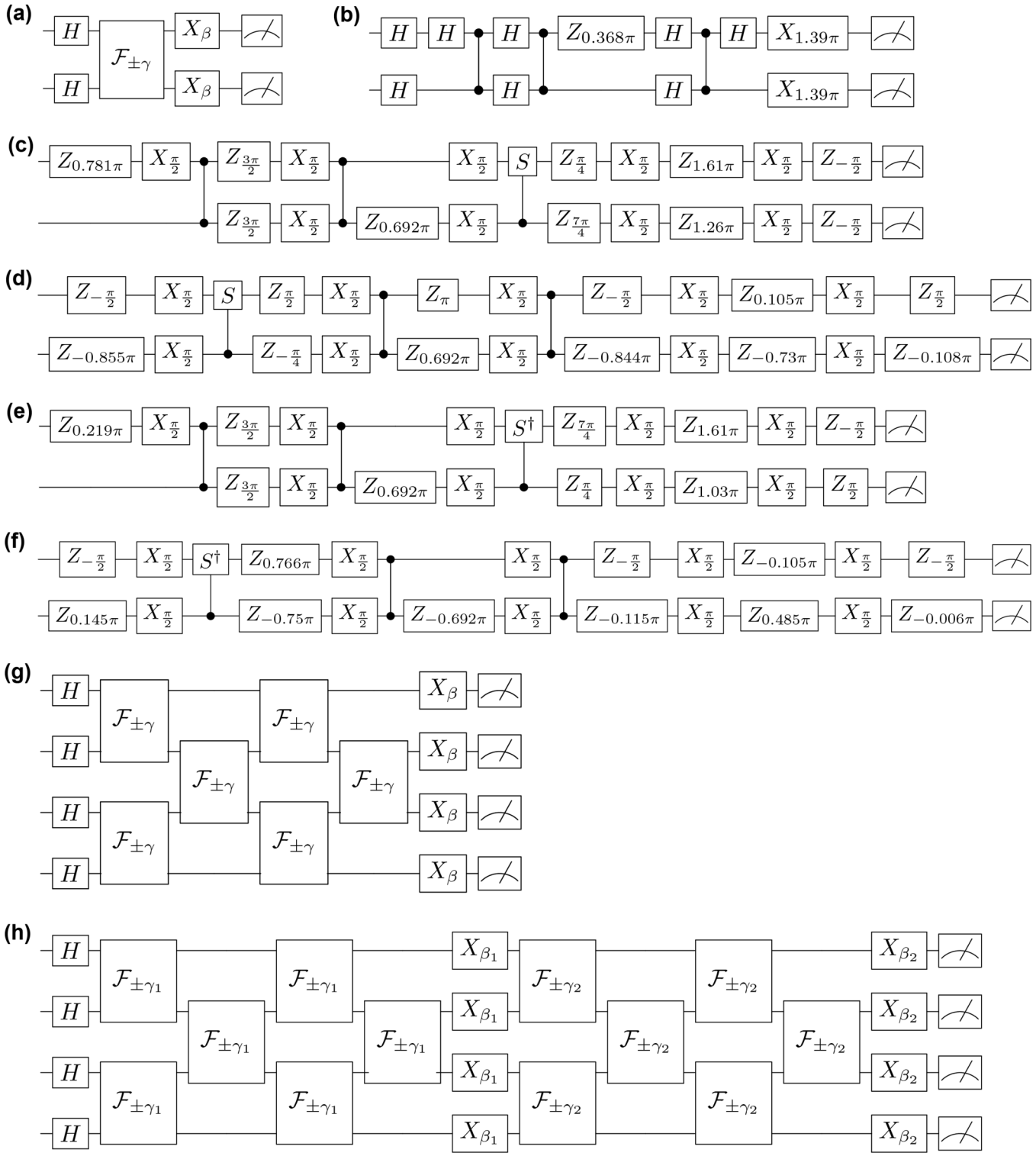


FIG. 5. Example SWAP networks for QAOA. Symbolic circuit representations of QAOA SWAP networks (of depth p) for (a) two qubits ($p = 1$), (g) four qubits ($p = 1$), and (h) four qubits ($p = 2$). (b) Baseline decomposition of a two-qubit QAOA SWAP network for a random choice of γ and β with three CZs. (c) Optimized decomposition of the circuit in (b) in terms of the native gateset utilizing a CS instead of a CZ. (d) Logically equivalent decomposition of the circuit in (c). (e) Optimized decomposition of the circuit in (b) in terms of the native gateset utilizing a $CS^\dagger$ instead of a CZ. (f) Logically equivalent decomposition of the circuit in (e).

delities [$P(0|0)$ and $P(1|1)$] are determined by performing ensemble measurements of the qubits prepared in $|0\rangle$ and $|1\rangle$ and classifying the results using a Gaussian mixture model fit to the in-phase (I) and quadrature (Q) heterodyne voltage signals. Error rates for single-qubit gates are measured using randomized benchmarking (RB) and simultaneous (Sim.) RB.

TABLE III. Single-qubit parameters.

	Q4	Q5	Q6	Q7
Qubit freq. (GHz)	5.254275	5.331004	5.490952	5.661671
Anharm. (MHz)	−275	−275	−271.35	−269
T_1 (μs)	60(5)	62(5)	52(4)	55(8)
T_2^* (μs)	36(5)	37(6)	36(6)	33(6)
T_2^{echo} (μs)	62(5)	73(7)	68(7)	54(6)
Readout $P(0\|0)$	0.999	0.995	0.995	0.995
Readout $P(1\|1)$	0.990	0.989	0.979	0.974
RB (10^{-3})	0.68(2)	1.01(2)	0.95(5)	0.67(1)
Sim. RB (10^{-3})	1.49(8)	2.5(2)	3.1(2)	2.4(2)

All error rates are defined in terms of the process infidelity $e_F = 1 - p$, where p is the exponential fit parameter in Ap^m for a sequence depth of m and SPAM parameter A. This is equivalent to the average gate infidelity $r(\mathcal{E})$,

$$e_F(\mathcal{E}) = r(\mathcal{E})\frac{d+1}{d}, \tag{A1}$$

where $d = 2^n$ is the system dimensionality (n qubits).

APPENDIX B: TWO-QUBIT GATE PARAMETERS

Table IV lists the parameters for the individual CZ, CS, and $CS^\dagger$ gates used in this work. All two-qubit gates are composed of square pulses with cosine ramps. The total gate duration of each pulse (including the ramps) are listed in Table IV; the fraction of the total gate duration for the ramp up and ramp down (individually) are specified under "Ramp fraction." Although the CS and $CS^\dagger$ gates can nominally be performed in half the duration of the CZ gates, the cosine ramps limit the minimum duration of the gates. Instead, the CS and $CS^\dagger$ gates are constructed to contain approximately half the total integrated area under the curve as the CZ gates, thus performing half of the conditional rotation as the CZ. This is only approximate since the conditional stark shift on each qubit will differ depending on the drive frequencies and amplitudes.

The choice of CS versus $CS^\dagger$ for each qubit pair was determined depending on the sign of the Stark-induced ZZ interaction (cf. Refs. [20,21]); it is more efficient (i.e., requires a smaller amplitude) to drive the CS rotation in one direction for some qubits, and in the opposite direction for other qubits, depending on the detuning of the drive signal. While the CZ can be benchmarked using RB, the CS and $CS^\dagger$ gates are non-Clifford, and therefore require either non-Clifford RB [22,54] or cycle benchmarking [25] (CB) with refocusing pulses (used in this work). Table IV lists the average process infidelity e_D of the dressed cycle (target gate plus Pauli twirling gates), as well as the inferred process infidelity e_T of the target gate alone [cf. Eq. (17)] using the measured CB process infidelity e_I of the "all-identity" reference cycle for each qubit pair. Two-qubit RB process infidelities e_F are also included for the CZ gates. While the fidelities of the individual two-qubit gates are useful for determining the quality of the gates in general, the process infidelities of the distinct parallel four-qubit cycles are more relevant to the application circuits presented in the body of this work. These values are listed in Table II of the main text.

TABLE IV. Two-qubit gate parameters.

Gate \ Qubits:		(Q4, Q5)	(Q5, Q6)	(Q6, Q7)
CZ	Duration (ns)	200	200	200
	Ramp fraction	0.3	0.3	0.3
	RB e_F (10^{-2})	1.9(1)	2.04(8)	1.95(6)
	CB e_D (10^{-2})	1.09(1)	1.05(1)	1.26(1)
	CB e_T (10^{-3})	5.8(1)	4.8(1)	5.9(2)
CS	Duration (ns)		150	
	Ramp fraction		0.4	
	CB e_D (10^{-2})		0.98(1)	
	CB e_T (10^{-3})		4.3(1)	
$CS^\dagger$	Duration (ns)	150		150
	Ramp fraction	0.4		0.4
	CB e_D (10^{-2})	0.98(1)		0.91(1)
	CB e_T (10^{-3})	5.0(1)		3.3(1)
Ref.	CB e_I (10^{-3})	3.24(5)	4.12(8)	4.8(1)

APPENDIX C: EXAMPLE QAOA SWAP NETWORK CIRCUITS

Example circuits for the QAOA SWAP networks presented in the main text can be seen in Fig. 5. This includes the symbolic representation of the two- and four-qubit QAOA SWAP networks of depth $p = 1$ and the four-qubit SWAP network of depth $p = 2$. An example two-qubit circuit is presented in Fig. 5(b) for a random choice of the two classical parameters γ and β. Additionally, the exact decompositions for this circuit in terms of CS and $CS^\dagger$ gates are included, as well as logically equivalent variants of each.

[1] D. S. Franca and R. Garcia-Patron, Limitations of optimization algorithms on noisy quantum devices, arXiv:2009.05532.

[2] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, Noise-induced barren plateaus in variational quantum algorithms, Nat. Commun. **12**, 6961 (2021).

[3] I. D. Kivlichan, J. McClean, N. Wiebe, C. Gidney, A. Aspuru-Guzik, G. K.-L. Chan, and R. Babbush, Quantum Simulation of Electronic Structure with Linear Depth and Connectivity, Phys. Rev. Lett. **120**, 110501 (2018).

[4] B. O'Gorman, W. J. Huggins, E. G. Rieffel, and K. B. Whaley, Generalized swap networks for near-term quantum computing, arXiv:1905.05118.

[5] T. Hagge, Optimal fermionic swap networks for Hubbard models, arXiv:2001.08324.

[6] Y. Ding, P. Gokhale, S. F. Lin, R. Rines, T. Propson, and F. T. Chong, Systematic crosstalk mitigation for superconducting qubits via frequency-aware compilation, in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* (IEEE, New York, 2020), pp. 201–214.

[7] T. Tomesh, P. Gokhale, E. R. Anschuetz, and F. T. Chong, Coreset clustering on small quantum computers, arXiv:2004.14970.

[8] E. Farhi, J. Goldstone, S. Gutmann, and L. Zhou, The quantum approximate optimization algorithm and the sherrington-kirkpatrick model at infinite size, arXiv:1910.08187.

[9] F. Barahona, M. Grötschel, M. Jünger, and G. Reinelt, An application of combinatorial optimization to statistical physics and circuit layout design, Oper. Res. **36**, 493 (1988).

[10] J. Preskill, Quantum computing in the nisq era and beyond, Quantum **2**, 79 (2018).

[11] E. Farhi, J. Goldstone, and S. Gutmann, A quantum approximate optimization algorithm, arXiv:1411.4028.

[12] S. Bravyi, A. Kliesch, R. Koenig, and E. Tang, Obstacles to Variational Quantum Optimization from Symmetry Protection, Phys. Rev. Lett. **125**, 260505 (2020).

[13] S. Hadfield, Z. Wang, B. O'Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, From the quantum approximate optimization algorithm to a quantum alternating operator ansatz, Algorithms **12**, 34 (2019).

[14] J. Wurtz and P. J. Love, Counterdiabaticity and the quantum approximate optimization algorithm, Quantum **6**, 635 (2022).

[15] B. Koczor, Exponential Error Suppression for Near-Term Quantum Devices, Phys. Rev. X **11**, 031057 (2021).

[16] W. J. Huggins, S. McArdle, T. E. O'Brien, J. Lee, N. C. Rubin, S. Boixo, K. B. Whaley, R. Babbush, and J. R. McClean, Virtual Distillation for Quantum Error Mitigation, Phys. Rev. X **11**, 041036 (2021).

[17] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, J. Majer, A. 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 (2007).

[18] AQT@LBL - SC Qubit Testbed, https://aqt.lbl.gov/, accessed: 2021-12-11.

[19] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, Efficient z gates for quantum computing, Phys. Rev. A **96**, 022330 (2017).

[20] B. K. Mitchell, R. K. Naik, A. Morvan, A. Hashim, J. M. Kreikebaum, B. Marinelli, W. Lavrijsen, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, Hardware-Efficient Microwave-Activated Tunable Coupling Between Superconducting Qubits, Phys. Rev. Lett. **127**, 200502 (2021).

[21] K. Wei, E. Magesan, I. Lauer, S. Srinivasan, D. Bogorin, S. Carnevale, G. Keefe, Y. Kim, D. Klaus, W. Landers *et al.*, Quantum crosstalk cancellation for fast entangling gates and improved multi-qubit performance, arXiv:2106.00675.

[22] S. Garion, N. Kanazawa, H. Landa, D. C. McKay, S. Sheldon, A. W. Cross, and C. J. Wood, Experimental implementation of non-clifford interleaved randomized benchmarking with a controlled-s gate, Phys. Rev. Research **3**, 013204 (2021).

[23] P. Gokhale, T. Tomesh, M. Suchara, and F. T. Chong, Faster and more reliable quantum swaps via native gates, arXiv:2109.13199.

[24] SuperstaQ Development Team, SuperstaQ: Connecting applications to quantum hardware, www.super.tech/about-superstaq (2021).

[25] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, Characterizing large-scale quantum computers via cycle benchmarking, Nat. Commun. **10**, 5347 (2019).

[26] M. P. Harrigan, K. J. Sung, M. Neeley, K. J. Satzinger, F. Arute, K. Arya, J. Atalaya, J. C. Bardin, R. Barends, S. Boixo *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, Nat. Phys. **17**, 332 (2021).

[27] D. M. Abrams, N. Didier, B. R. Johnson, M. P. da Silva, and C. A. Ryan, Implementation of the xy interaction family with calibration of a single pulse, Nat. Electron. **3**, 744 (2020).

[28] G. S. Barron, F. A. Calderon-Vargas, J. Long, D. P. Pappas, and S. E. Economou, Microwave-based arbitrary cphase gates for transmon qubits, Phys. Rev. B **101**, 054508 (2020).

[29] P. Gokhale, A. Javadi-Abhari, N. Earnest, Y. Shi, and F. T. Chong, Optimized quantum compilation for near-term algorithms with openpulse, in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* (IEEE, New York, 2020), pp. 186–200.

[30] J. Zhang, J. Vala, S. Sastry, and K. B. Whaley, Geometric theory of nonlocal two-qubit operations, Phys. Rev. A **67**, 042313 (2003).

[31] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B **7**, S347 (2005).

[32] 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 (2008).

[33] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. A **80**, 012304 (2009).

[34] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and Robust Randomized Benchmarking of Quantum Processes, Phys. Rev. Lett. **106**, 180504 (2011).

[35] T. Proctor, K. Rudinger, K. Young, E. Nielsen, and R. Blume-Kohout, Measuring the capabilities of quantum computers, Nat. Phys. **18**, 75 (2022).

[36] S. Krinner, S. Lazar, A. Remm, C. K. Andersen, N. Lacroix, G. J. Norris, C. Hellings, M. Gabureac, C. Eichler, and A. Wallraff, Benchmarking Coherent Errors in Controlled-Phase Gates due to Spectator Qubits, Phys. Rev. Appl. **14**, 024042 (2020).

[37] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking, Phys. Rev. Lett. **109**, 080505 (2012).

[38] A. Carignan-Dugas, J. J. Wallman, and J. Emerson, Bounding the average gate fidelity of composite channels using the unitarity, New J. Phys. **21**, 053016 (2019).

[39] E. Farhi and A. W. Harrow, Quantum supremacy through the quantum approximate optimization algorithm, arXiv:1602.07674.

[40] K. Rudinger, C. W. Hogle, R. K. Naik, A. Hashim, D. Lobser, D. I. Santiago, M. D. Grace, E. Nielsen, T. Proctor, S. Seritan *et al.*, Experimental characterization of crosstalk errors with simultaneous gate set tomography, PRX Quantum **2**, 040338 (2021).

[41] V. Tripathi, H. Chen, M. Khezri, K.-W. Yip, E. Levenson-Falk, and D. A. Lidar, Suppression of crosstalk in superconducting qubits using dynamical decoupling, arXiv:2108.04530.

[42] G. H. Low, T. J. Yoder, and I. L. Chuang, Optimal arbitrarily accurate composite pulse sequences, Phys. Rev. A **89**, 022341 (2014).

[43] M. R. Geller and Z. Zhou, Efficient error models for fault-tolerant architectures and the pauli twirling approximation, Phys. Rev. A **88**, 012314 (2013).

[44] Z. Cai, X. Xu, and S. C. Benjamin, Mitigating coherent noise using pauli conjugation, npj Quantum Inf. **6**, 17 (2020).

[45] C. Song, J. Cui, H. Wang, J. Hao, H. Feng, and Y. Li, Quantum computation with universal error mitigation on a superconducting quantum processor, Sci. Adv. **5**, eaaw5686 (2019).

[46] Y. Kim, C. J. Wood, T. J. Yoder, S. T. Merkel, J. M. Gambetta, K. Temme, and A. Kandala, Scalable error mitigation for noisy quantum circuits produces competitive expectation values, arXiv:2108.09197.

[47] E. Knill, Fault-tolerant postselected quantum computation: Threshold analysis, arXiv:quant-ph/0404104.

[48] O. Kern, G. Alber, and D. L. Shepelyansky, Quantum error correction of coherent errors by randomization, Eur. Phys. J. D **32**, 153 (2005).

[49] M. Ware, G. Ribeill, D. Ristè, C. A. Ryan, B. Johnson, and M. P. da Silva, Experimental pauli-frame randomization on a superconducting qubit, Phys. Rev. A **103**, 042604 (2021).

[50] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, Phys. Rev. A **94**, 052325 (2016).

[51] A. Hashim, R. K. Naik, A. Morvan, J.-L. Ville, B. Mitchell, J. M. Kreikebaum, M. Davis, E. Smith, C. Iancu, K. P. O'Brien *et al.*, Randomized Compiling for Scalable Quantum Computing on a Noisy Superconducting Quantum Processor, Phys. Rev. X **11**, 041039 (2021).

[52] M. B. Hastings, Turning gate synthesis errors into incoherent errors, arXiv:1612.01011.

[53] E. Campbell, Shorter gate sequences for quantum computing by mixing unitaries, Phys. Rev. A **95**, 042306 (2017).

[54] A. W. Cross, E. Magesan, L. S. Bishop, J. A. Smolin, and J. M. Gambetta, Scalable randomised benchmarking of non-clifford gates, npj Quantum Inf. **2**, 16012 (2016).

[55] J. Haah and M. B. Hastings, Codes and protocols for distilling t, controlled-s, and toffoli gates, Quantum **2**, 71 (2018).

[56] A. N. Glaudell, N. J. Ross, and J. M. Taylor, Optimal two-qubit circuits for universal fault-tolerant quantum computation, npj Quantum Inf. **7**, 103 (2021).

[57] T. Kim and B.-S. Choi, Efficient decomposition methods for controlled-r n using a single ancillary qubit, Sci. Rep. **8**, 5445 (2018).

[58] E. C. Peterson, L. S. Bishop, and A. Javadi-Abhari, Optimal synthesis into fixed xx interactions, Quantum **6**, 696 (2022).

[59] T. Satoh, S. Oomura, M. Sugawara, and N. Yamamoto, Pulse-engineered Controlled-V gate andits applications on superconducting quantum device, IEEE Trans. Quant. Eng. **3**, 3101610 (2022).