Demonstrating two-qubit gates at the quantum speed limit using superconducting qubits

Joel Howard,[1] Alexander Lidiak,[1] Casey Jameson,[1] Bora Basyildiz,[2] Kyle Clark,[1] Tongyu Zhao,[3] Mustafa Bal,[3,4] Junling Long,[3] David P. Pappas,[5,6] Meenakshi Singh,[1,*] and Zhexuan Gong[1,†]

[1]*Department of Physics, Colorado School of Mines, Golden, Colorado 80401, USA*
[2]*Department of Computer Science, Colorado School of Mines, Golden, Colorado 80401, USA*
[3]*Department of Physics, University of Colorado, Boulder, Colorado 80309, USA*
[4]*Superconducting Quantum Materials and Systems Center, Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA*
[5]*National Institute of Standards and Technology, Boulder, Colorado 80305, USA*
[6]*Rigetti Computing, 775 Heinz Ave, Berkeley CA 94701*

The speed of elementary quantum gates, particularly two-qubit gates, ultimately sets the limit on the speed at which quantum circuits can operate. In this work, we experimentally demonstrate commonly used two-qubit gates at nearly the fastest possible speed allowed by the physical interaction strength between two superconducting transmon qubits. We achieve this quantum speed limit by implementing experimental gates designed using a machine learning inspired optimal control method. Importantly, our method only requires the single-qubit drive strength to be moderately larger than the interaction strength to achieve an arbitrary two-qubit gate close to its analytical speed limit with high fidelity. Thus, the method is applicable to a variety of platforms including those with comparable single-qubit and two-qubit gate speeds, or those with always-on interactions.

Introduction.— Increasing the speed of elementary quantum gates boosts the "clock speed" of a quantum computer. For noisy, intermediate-scale quantum computers [1] with finite coherence times [2, 3], speeding up single- and two-qubit quantum gates also increases the circuit depth needed for solving useful computational problems [4, 5]. In most experimental platforms, single-qubit gates are achieved via electromagnetic fields that drive individual qubit transitions. The maximum speed of these gates is limited by the strength of the driving fields [6, 7] and is usually much faster than the decoherence rate of the qubit. A two-qubit entangling gate, necessary for universal quantum gates, can however only operate at a speed proportional to the interaction strength between the qubits [8–10], which is typically much weaker than available single-qubit drive strengths and cannot be easily increased.

Assuming a limited interaction strength, one can analytically obtain the maximum speed for any particular two-qubit gate in the limit of arbitrarily fast single-qubit gates [11, 12]. In practice, all single-qubit gates have a finite speed, and in some platforms, they are not much faster than two-qubit gates [13–15]. The speed limits of two-qubit gates in such realistic scenarios have not been studied. Therefore, we seek to both theoretically and experimentally investigate these practical speed limits, which could guide the optimal design of quantum gates in future quantum computers.

In this letter, we report a new method for designing two-qubit gates that are speed optimized, and we demonstrate the method experimentally using superconducting transmon qubits. Importantly, we find that the protocol for achieving fastest two-qubit gates in Refs. [11, 12] can be far from optimal with a finite single-qubit gate time. Our method differs from previous protocols [11–13] in that we apply single-qubit drives simultaneously with the two-qubit interaction, a crucial strategy for speed optimization. We optimize the pulse shapes of the single-qubit drives using a method that combines the well-known GRAPE algorithm [16, 17] with state-of-art machine learning techniques. Remarkably, our method can achieve maximally entangling two-qubit gates close to their analytical speed limits found in Refs. [11, 12] with rather modest single-qubit drive strengths, thus largely eliminating the impractical assumption of infinitely fast single-qubit gates. Same applies to the SWAP gate we implemented, which is crucial for remote quantum gates but notoriously hard to achieve with a short gate sequence [18]. In addition, we do not require the interaction to be turned off, making the method suitable for platforms with always-on coupling [13–15]. Finally, our speed-optimized gates achieve the same level of fidelity as that achieved previously on the same hardware [13], thus our method does not sacrifice gate fidelity for speed.

Experimental setup.—Our experimental platform consists of strongly-coupled fixed-frequency superconducting transmon qubits with static capacitive couplings in a hanger readout geometry [19]. An intrinsic silicon substrate is used on which aluminum oxide tunnel junctions are fabricated via an overlap technique [20]. The remaining circuit components are made of niobium. The full chip design and corresponding circuit model is shown in Fig. 1. The two transmon qubits' transition frequencies are $5.10, 5.26$ MHz, anharmonicities are $270, 320$ MHz, T_1 decay times are $40, 21$ μs and T_2^* decay times are $12, 10$ μs, respectively. In the rotating frame of the two qubit frequencies and assuming $\hbar = 1$, the static Hamiltonian of the two qubits can be written as [13, 21]

$$H_0 = g(\sigma_1^z + \sigma_2^z + \sigma_1^z\sigma_2^z) \tag{1}$$

where $g \approx 2\pi \times 1.75$MHz represents a fixed Ising coupling strength between the qubits. To interact with the qubits, we deliver two microwave drives – resonant with each qubit's transition frequency – simultaneously through the feedline. Each of the drive fields contain two adjustable quadratures (X or Y), and can be described by the drive Hamiltonian.

$$H_1(t) = \sum_{\gamma=x,y}\sum_{i=1,2} \Omega_i^\gamma(t)\tilde{\sigma}_i^\gamma \tag{2}$$

where $\Omega_i^{x,y}(t)$ denotes the Rabi frequency of the drive resonant with qubit i's transition in the X or Y quadrature at time t. For perfect single-qubit drives $\tilde{\sigma}_i^\gamma = \sigma_i^\gamma$. However, due to the strong Ising coupling between the two qubits in H_0, the drive strength on one qubit is dependent on the other qubit's state, resulting in $\tilde{\sigma}_1^\gamma = \sigma^\gamma \otimes (|0\rangle\langle 0| + r_2 |1\rangle\langle 1|)$ and $\tilde{\sigma}_2^\gamma = (|0\rangle\langle 0| + r_1 |1\rangle\langle 1|) \otimes \sigma^\gamma$, with $r_1 \approx 1.1$ and $r_2 \approx 0.7$ for our current chip. We note that with a weaker coupling strength or with a tunable coupler [22, 23], both r_1 and r_2 can be made closer to or equal to 1.

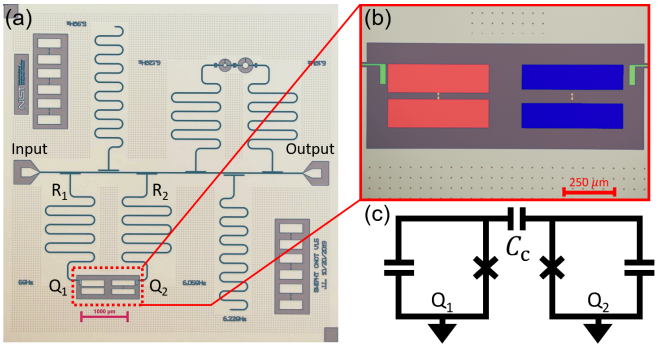

Figure 1. (a) Optical micrograph of the experimental chip including qubits, readout resonators, test Josephson junctions, and test resonators. (b) Zoomed-in view of the two floating qubits. Each qubit consists of two identical pads (red for the left qubit and blue for the right qubit) and a Josephson junction connecting the two pads. Each qubit is coupled to its own readout resonator (green). (c) Grounded circuit model of the capacitively coupled qubits.

Analytical speed limit.—In the limit of arbitrarily strong single-qubit drives, i.e. $\Omega_{\max} \equiv \max |\Omega_{1,2}^{x,y}(t)| \to \infty$, one can derive an analytical speed limit for any target two-qubit unitary with the above-mentioned static Hamiltonian H_0 and control Hamiltonian H_1. Note that in this limit, the speed limit is only well defined when $r_1 = r_2 = 1$, since otherwise H_1 will lead to arbitrarily strong interactions. As detailed in [12], any two-qubit target unitary U can always be decomposed as

$$U = (U_1 \otimes U_2) U_d (V_1 \otimes V_2), \quad U_d = e^{-i \sum_{\gamma=x,y,z} \lambda_\gamma \sigma^\gamma \otimes \sigma^\gamma} \tag{3}$$

where U_1, V_1 (U_2, V_2) are some single-qubit gates on the first (second) qubit, and $\lambda_{x,y,z} \in [-\frac{\pi}{4}, \frac{\pi}{4}]$, with their exact values obtainable with the knowledge of U [24]. To obtain the analytical speed limit, we assume that single qubit gates are arbitrarily fast and thus take a negligible amount of time. The total gate time for implementing U therefore reduces to the time spent on realizing U_d, which is responsible for any entanglement generation. Based on our H_0, together with instantaneous single-qubit rotations, U_d can be realized with a minimum time of:

$$T_{\min} = \frac{|\lambda_x| + |\lambda_y| + |\lambda_z|}{g} = \begin{cases} \pi/(4g) & \text{CNOT} \\ 3\pi/(4g) & \text{SWAP} \\ 3\pi/(8g) & \sqrt{\text{SWAP}} \end{cases} \tag{4}$$

Eq. (4) is the analytical speed limit for a two-qubit gate U. Its values for the CNOT, SWAP, and $\sqrt{\text{SWAP}}$ gate are also shown in the equation (see [24] for detailed derivations).

Optimal control.—In practice, single-qubit gate speeds are limited by finite drive strengths and the analytical speed limit in Eq. (4) does not apply. To our best knowledge, no analytical speed limit has been found with finite single-qubit gate time. If we still follow Eq. (3) for realizing a target two-qubit gate, the resulting gate time can be much longer than $T_{\min}$, especially for our system. To realize universal single-qubit gates needed in Eq. (3), we use a two-axis gate (TAG) protocol first developed in Ref. [13], which employs an analytically obtained 3-segment drive pulse for $\Omega_{1,2}^{x,y}(t)$ to exactly cancel the effects of static interaction for any values of r_1 and r_2. A single-qubit gate implemented this way has a gate time of at least $\pi/(2g)$ [21]. Apart from the controlled-Z gate that can be directly realized via an evolution of H_0 over a time $t = \pi/(4g) \approx 71.4$ns [13, 25, 26], any two-qubit gate design that involves the use of single-qubit gate(s) realized via TAG requires a gate time of at least $T_{\min} + \pi/(2g)$ (since single-qubit gates cannot shorten $T_{\min}$), and is thus far from optimal.

Consequently, to approach the analytical speed limit with finite $\Omega_{\max}$ (or finite single qubit gate time), we adopt an alternative approach that avoids the use of any single-qubit gate and try to generate the target two-qubit gate directly. Specifically, we directly optimizes the pulse shapes $\Omega_{1,2}^{x,y}(t)$ in our control Hamiltonian $H_1(t)$ in order to minimize the gate time for achieving a certain target gate with sufficiently high fidelity. For a given set of pulse shape functions $\Omega_{1,2}^{x,y}(t)$, we numerically find the evolution operator $\mathcal{U} = \mathcal{T} e^{-i \int_0^T [H_0 + H_1(t)] dt}$ where $\mathcal{T}$ denotes the time ordered integral and T denotes the total evolution time. Next, we calculate the average gate fidelity between the target unitary U and the evolution operator $\mathcal{U}$ using [27]

$$F = \frac{1}{5} + \frac{1}{20} \sum_j \text{Tr}\left(U U_j U^\dagger \mathcal{U} U_j \mathcal{U}^\dagger\right) \tag{5}$$

where $U_j \in \{\sigma^\gamma \otimes \sigma^{\gamma'}\}$ and $\sigma^\gamma \in \{\sigma^x, \sigma^y, \sigma^z, I\}$. For efficient numerical optimization, we will assume $\Omega_i^\gamma(t)$ is an M-segment piece-wise function, i.e. $\Omega_i^\gamma(t) = \Omega_{i,m}^\gamma$ for $t \in [\frac{m-1}{M}T, \frac{m}{M}T]$. Our goal is to maximize F over all possible values of $\{\Omega_{i,m}^\gamma\}$ subjected to the constraints $|\Omega_{i,m}^\gamma| \le \Omega_{\max}$ for a given time T. The numerical speed limit T_F is then defined as the minimum T that can achieve $F > 1 - \epsilon$, where ϵ is the infidelity we can tolerate (set to 1% in the following).

Since F is a highly nonlinear function of $\{\Omega_{i,m}^\gamma\}$, simple numerical optimization methods will not work well in finding the global maximum of F. Here we develop a new method that combines the standard GRAPE algorithm [17] with state-of-art machine learning techniques. Using the backward propagation method in the widely used machine learning library PyTorch [28], we calculate the gradients of F over each pulse parameter $\Omega_{i,m}^\gamma$ automatically. We then perform a stochastic gradient descent (SGD) algorithm with the Nesterov Momentum method [29] to maximize F over the pulse parameters. To avoid obtaining only a local maximum for F, we repeat each

gradient descent process with a large number (up to 200) of random seeds used for both initialization and SGD, and then select the global maximum among all repetitions.

To benchmark our numerical optimization method, we choose the target gate to be the CNOT gate and find the above-mentioned numerical speed limit T_F for $F = 99\%$ as a function of $\Omega_{\max}$. We set $M = 16$, which allows the calculation to be done within a few hours on a small HPC cluster. Further increasing M does not lead to noticeable improvements. As shown in Fig. 2, we clearly see that as $\Omega_{\max}/g$ increases, T_F approaches the analytical speed limit $T_{\min}$, indicating that the optimization succeeded in reaching the theoretical speed limit. Importantly, the maximum single-qubit drive strength $\Omega_{\max}$ does not need to be significantly larger than the interaction strength g to get close to the analytical speed limit. For example, setting $\Omega_{\max} = 3g$ already gives us a minimum gate time of $1.05T_{\min}$ with $F > 99\%$. We also note from Fig. 2 that our method significantly outperforms the standard GRAPE algorithm in the widely used QuTiP software [30].

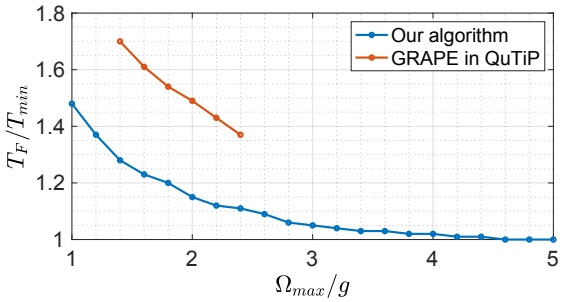

Figure 2. The minimum time T_F (in units of $T_{\min}$) it takes to achieve a CNOT gate of $F > 99\%$ as a function of $\Omega_{\max}$ (in units of g) using either our optimization algorithm (blue curve) or the GRAPE algorithm in QuTiP (red curve). Both algorithms use $M = 16$ segments of the drive pulses and 200 random restarts. The GRAPE algorithm in QuTiP can only reach $F > 99\%$ for a small range of $\Omega_{\max}$ values, and even for such values the minimum gate time achieved is noticeably longer than that from our algorithm.

Experimental Results.— We now proceed to demonstrate the speed limits of the two-qubit CNOT, SWAP, and $\sqrt{\text{SWAP}}$ gates experimentally. The procedure for this is as follows. First, for each gate, the total evolution time T is varied from 0 to $\gtrsim T_{\min}$ in 20 steps and the optimized pulse sequence is obtained numerically for each value of T. Next, this pulse sequence is applied to the transmon qubits experimentally by modulating the microwave drive signals. Finally, the average gate fidelity F is measured at time T by performing a quantum process tomography (QPT) [31]. Our QPT involves applying 36 different pre-rotations to an initial state with both qubits in the state $|0\rangle$, applying the optimized pulse sequence for time T, and then measuring 9 different Pauli operator (see [24] for details), resulting in 324 different experimental protocols, each of which is further repeated 500 times to ensure low statistical errors. After correcting the state preparation and measurement (SPAM) errors as well as performing a maximum likelihood estimation to ensure a completely positive and trace-preserving quantum map [24], the QPT allows us to find a Pauli transfer matrix [32] for the corresponding quantum process, which can be further used to infer F [24]. This process allows us to find the value of T above which we can get sufficiently high gate fidelity. Such T is the experimental speed limit for the target gate.

There are several experimental limitations in this procedure. First, as strong microwave drives can heat up the superconducting qubits and cause decoherence, we only send microwave pulses of at most $2\pi \times 6$ MHz in Rabi frequency, roughly 3 times the coupling strength g. But as we have shown in Fig. 2, this limitation should not prevent us from getting very close to the analytical speed limit. A more noticeable limitation is that we can only generate smoothly varying pulse shapes that approximate the segmented (and thus discontinuous) pulse shapes used in the numerical optimization. As the number of segments M increases, this approximation deteriorates while the gate speed increases (and eventually converges). For our setup, we choose $M = 4$ for the experiment as a sweet spot for balancing the error and speed. We note that this limitation can be addressed by numerically optimizing smooth pulse shape functions (such as a train of Gaussian envelopes), although such optimization is computationally challenging. Finally, with $r_1, r_2 \neq 1$ experimentally, our single-qubit drives will induce a small amount of extra interaction that would in principle allow us to go above the analytical speed limit for sufficiently large $\Omega_{\max}$. Our numerical optimizer accounts for this artifact. The amount of speedup over the scenario of $r_{1,2} = 1$ varies for different target gates.

Our experimental results are shown in Fig. 3. The measured gate fidelity F (red curves) closely matches the one obtained from the numerical simulation of the experiment with no error (blue curves). The deviations between the two grow as the gate fidelity gets close to 1 for reasons we discuss in the next section. For the CNOT gate, we were able to achieve $F \approx 96.5\%$ experimentally with a gate time of $T \approx 1.32T_{\min}$ (Fig. 3a). We emphasize that this outperforms the CNOT gate implemented using the SWIPHT protocol [33] performed on the same hardware ($F \approx 94.6\%$ for a gate time of $1.87T_{\min}$ [13]), which is a specially optimized protocol for our hardware. The highest fidelity we achieve is $F \approx 98.3\%$ at time $T \approx 1.84T_{\min}$, with $F \approx 99.8\%$ theoretically.

For the SWAP and $\sqrt{\text{SWAP}}$ gates, the extra interactions caused by non-unity r_1 and r_2 values noticeably speed up the gates beyond the analytical speed limits in Eq. (4). For the SWAP gate (Fig. 3b), we obtain an experimental gate fidelity of $F \approx 94.8\%$ at $T \approx 0.82T_{\min}$, where theoretically we can obtain $F \approx 99.9\%$. The highest fidelity we can achieve experimentally is $F \approx 95.9\%$ at $T \approx 1.01T_{\min}$, where theoretically $F \approx 99.997\%$. A SWAP gate with such a short time is hard to achieve via a traditional gate sequence using a universal gate set, making our method particularly useful given the importance of SWAP gates in many quantum algorithms [18].

Finally, for the $\sqrt{\text{SWAP}}$ gate (Fig. 3c), we obtain an experimental gate fidelity of $F \approx 95.4\%$ at $T \approx 1.03T_{\min}$,

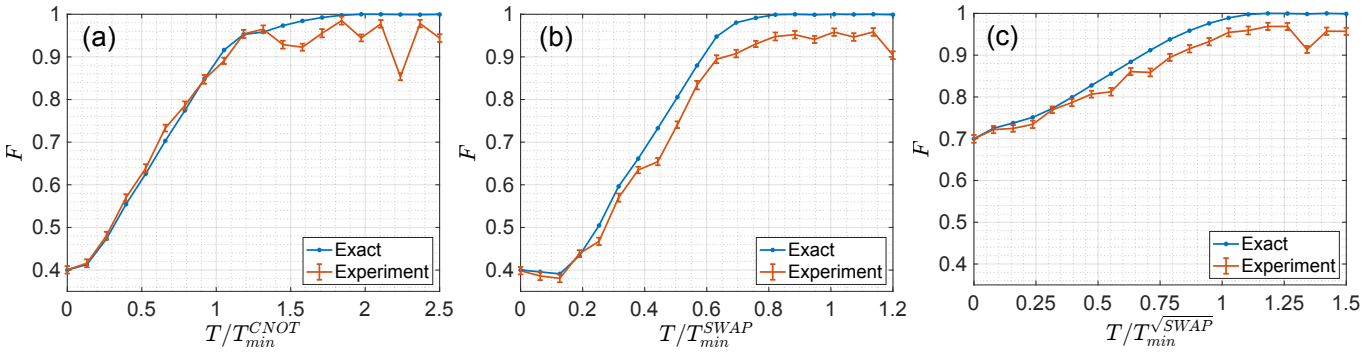

Figure 3. Experimental measurements of the average gate fidelity F using optimized 4-segment drive pulses , with the target gate being (a) CNOT, (b) SWAP, and (c) $\sqrt{\text{SWAP}}$. The red curves represent experimental measurements while the blue curves represent the exact numerical calculation of F without considering any experimental error. $\Omega_{\text{max}} = 6$MHz for the CNOT gate and $\Omega_{\text{max}} = 5$MHz for SWAP and $\sqrt{\text{SWAP}}$ gate. The error bars represent a upper bound on the statistical error of the mean for 500 repeated measurements at each point [24].

where theoretically $F \approx 98.9\%$. The highest fidelity we can achieve experimentally is $F \approx 97.0\%$ at $T \approx 1.18T_{\text{min}}$, with $F \approx 99.999\%$ in theory. For all gates, the demonstrated experimental speed limits are reasonably close to the analytical speed limits. Higher experimental fidelities can be achieved with future hardware improvements.

Error analysis.— We have calculated the fidelity between the experimental process and the exact time evolution operator $\mathcal{U}$ for each point in Fig. 3, which is in general $> 95\%$ [24]. As seen from Fig. 3, the experimental errors get larger at large values of T. This is possibly due to the following reasons. First, the qubits decohere as time increases. This is evidenced by our measurement of a dark evolution fidelity that drops from $\approx 99.3\%$ to $\approx 94.4\%$ from $T = 0$ to $T \sim \pi/g$ [24]. Second, when T is large enough to allow the numerically optimized F to approach 1, imperfect calibration or fluctuations on the microwave drive amplitudes or phases tend to have a larger effect on the gate fidelity than that with shorter T [24].

There are also systematic errors coming from energy levels and couplings not represented in our Hamiltonians H_0 and H_1. For example, higher excited states outside the qubit subspace may be occupied during the time evolution. Their occupation may not be negligible in particular because of an accidental near-degeneracy between our $|00\rangle - |10\rangle$ transition (5.093GHz) and the $|00\rangle - |02\rangle$ two-photon transition (10.213GHz). In addition, an off-resonant transition from the $|01\rangle$ to $|02\rangle$ state may also play a role due to the limited anharmonicity of our transmon qubits.

Conclusion and Outlook.— We have demonstrated quantum speed limits for various two-qubit gates experimentally. The gates are implemented via a new optimal control method and are much faster than those implemented using a combination of single qubit gates and a native two-qubit gate on our platform. Our method is useful for designing speed-optimized two-qubit gates in a variety of quantum computing platforms, especially those with non-negligible single-qubit gate times.

An important future direction is to generalize this work to a multi-qubit scenario where additional qubits are used to speed up a two-qubit gate. Previous work has shown that significant scaling speedups can be obtained in performing remote quantum gates or preparing useful many-body entangled states [34, 35] with long-range interacting qubits. However, questions regarding the speed limit of entangling gates when interactions are strongly long-ranged are still largely open [36]. Such interactions play important roles in quantum information scrambling [37] and the development of fully-connected quantum computers [38]. Another interesting direction is to study the speed limit of entangling gates when higher excited states outside the qubit subspace are utilized [39], where experimental and analytical results are both lacking.

This work is jointly supervised by MS and ZXG. We thank NIST Boulder for hosting the experiment and the HPC center at Colorado School of Mines for providing computational resources. We acknowledge funding support from the NSF RAISE-TAQS program CCF-1839232, the NSF Triplets program DMR-1747426, the NSF NRT program DGE-2125899, and the W. M. Keck Foundation.

* msingh@mines.edu
† gong@mines.edu

[1] J. Preskill, Quantum **2**, 79 (2018).
[2] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, Annual Review of Condensed Matter Physics **11**, 369 (2020).
[3] A. P. M. Place *et al.*, Nature Communications **12**, 1779 (2021).
[4] M. H. Devoret and R. J. Schoelkopf, Science **339**, 1169 (2013).
[5] F. Arute *et al.*, Nature **574**, 505 (2019).
[6] M. R. Frey, Quantum Information Processing **15**, 3919 (2016).
[7] N. Malossi, M. G. Bason, M. Viteau, E. Arimondo, D. Ciampini, R. Mannella, and O. Morsch, Journal of Physics: Conference Series **442**, 012062 (2013).
[8] V. M. Schäfer, C. J. Ballance, K. Thirumalai, L. J. Stephenson, T. G. Ballance, A. M. Steane, and D. M. Lucas, Nature **555**, 75 (2018).
[9] L.-M. Duan, Phys. Rev. Lett. **93**, 100502 (2004).
[10] A. M. Steane, G. Imreh, J. P. Home, and D. Leibfried, New Journal of Physics **16**, 053049 (2014).

[11] G. Vidal, K. Hammerer, and J. I. Cirac, Phys. Rev. Lett. **88**, 237902 (2002).

[12] B. Kraus and J. I. Cirac, Phys. Rev. A **63**, 062309 (2001).

[13] J. Long, T. Zhao, M. Bal, R. Zhao, G. S. Barron, H.-s. Ku, J. A. Howard, X. Wu, C. R. H. McRae, X.-H. Deng, G. J. Ribeill, M. Singh, T. A. Ohki, E. Barnes, S. E. Economou, and D. P. Pappas, arXiv: 2103.12305 (2021).

[14] X. Li, Y. Ma, J. Han, T. Chen, Y. Xu, W. Cai, H. Wang, Y.-P. Song, Z.-Y. Xue, Z.-Q. Yin, and L. Sun, Phys. Rev. Applied **10**, 054009 (2018).

[15] S. A. Caldwell *et al.*, Phys. Rev. Applied **10**, 034050 (2018).

[16] J. Werschnik and E. K. U. Gross, Journal of Physics B: Atomic, Molecular and Optical Physics **40**, R175 (2007).

[17] N. Khaneja *et al.*, Journal of Magnetic Resonance (2005).

[18] IBM, "The open science prize: Solve for swap gates and graph states," (2021).

[19] K. O'Brien, V. Ramasesh, A. Dove, J. M. Kreikebaum, J. Colless, and I. Siddiqi, *Quantum Information and Measurement (QIM) 2017*, Quantum Information and Measurement (QIM) 2017 , QF6A.4 (2017).

[20] X. Wu, J. L. Long, H. S. Ku, R. E. Lake, M. Bal, and D. P. Pappas, Applied Physics Letters **111**, 032602 (2017).

[21] J. Long, *Superconducting Quantum Circuits for Quantum Information Processing*, Ph.D. thesis, University of Colorado, Boulder (2020).

[22] P. Mundada, G. Zhang, T. Hazard, and A. Houck, Phys. Rev. Applied **12**, 054023 (2019).

[23] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavsson, and W. D. Oliver, Phys. Rev. Applied **10**, 054062 (2018).

[24] See the Supplemental Material for details.

[25] M. C. Collodo, J. Herrmann, N. Lacroix, C. K. Andersen, A. Remm, S. Lazar, J.-C. Besse, T. Walter, A. Wallraff, and C. Eichler, Phys. Rev. Lett. **125**, 240502 (2020).

[26] G. S. Barron, F. A. Calderon-Vargas, J. Long, D. P. Pappas, and S. E. Economou, Phys. Rev. B **101**, 054508 (2020).

[27] M. A. Nielsen, Physics Letters A **303**, 249 (2002).

[28] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," (Curran Associates Inc., Red Hook, NY, USA, 2019).

[29] Y. E. Nesterov, in *Dokl. akad. nauk Sssr*, Vol. 269 (1983) pp. 543–547.

[30] J. Johansson, P. Nation, and F. Nori, Computer Physics Communications **183**, 1760 (2012).

[31] Grove, "Histogram based tomography," (2016).

[32] J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, M. B. Ketchen, and M. Steffen, Phys. Rev. Lett. **109**, 060501 (2012).

[33] S. E. Economou and E. Barnes, Phys. Rev. B **91**, 161405(R) (2015).

[34] Z. Eldredge, Z.-X. Gong, J. T. Young, A. H. Moosavian, M. Foss-Feig, and A. V. Gorshkov, Phys. Rev. Lett. **119**, 170503 (2017).

[35] M. C. Tran, C.-F. Chen, A. Ehrenberg, A. Y. Guo, A. Deshpande, Y. Hong, Z.-X. Gong, A. V. Gorshkov, and A. Lucas, Phys. Rev. X **10**, 031009 (2020).

[36] A. Y. Guo, M. C. Tran, A. M. Childs, A. V. Gorshkov, and Z.-X. Gong, Phys. Rev. A **102**, 010401(R) (2020).

[37] N. Lashkari, D. Stanford, M. Hastings, T. Osborne, and P. Hayden, Journal of High Energy Physics **2013**, 22 (2013).

[38] N. M. Linke, D. Maslov, M. Roetteler, S. Debnath, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe, Proceedings of the National Academy of Sciences **114**, 3305 (2017).

[39] S. Ashhab, F. Yoshihara, T. Fuse, N. Yamamoto, A. Lupascu, and K. Semba, Phys. Rev. A **105**, 042614 (2022).

Supplementary Material for "Demonstrating two-qubit entangling gates at the quantum speed limit using superconducting qubits"

Joel Howard,[1] Alexander Lidiak,[1] Casey Jameson,[1] Bora Basyildiz,[2] Kyle Clark,[1] Tongyu Zhao,[3] Mustafa Bal,[3,4] Junling Long,[3] David P. Pappas,[5,6] Meenakshi Singh,[1,*] and Zhexuan Gong[1,†]

[1]*Department of Physics, Colorado School of Mines, Golden, Colorado 80401, USA*
[2]*Department of Computer Science, Colorado School of Mines, Golden, Colorado 80401, USA*
[3]*Department of Physics, University of Colorado, Boulder, Colorado 80309, USA*
[4]*Superconducting Quantum Materials and Systems Center, Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA*
[5]*National Institute of Standards and Technology, Boulder, Colorado 80305, USA*
[6]*Rigetti Computing, 775 Heinz Ave, Berkeley CA 94701*

In this supplementary material, we provide additional details on our experimental hardware, how to obtain the analytical speed limits, our quantum process tomography, and the error analysis.

I. EXPERIMENTAL HARDWARE, CALIBRATION, AND CHARACTERIZATION

Our experimental device is operated at 10mK in an Bluefors LD dilution refrigerator. Full schematics of the experimental setup are shown in Fig. 1. All qubit drive and readout microwave tones are delivered via the feedline, which has an output amplification chain of a Raytheon BBN Josephson parametric amplifier (JPA) preamp at base, high-electron-mobility transistor (HEMT) amplifier at 4K, and a high-gain room-temperature amplifier.

Each experimental cycle consists of a state initialization, a time evolution under the engineered Hamiltonian flanked by process tomography rotations [1] and followed by a heterodyne state readout (see Fig. 1). The state initialization occurs by waiting $500\mu s \approx 12T_1$ between two experimental cycles, which is long enough to guarantee that each qubit is in the $|0\rangle$ state. All gates consist of microwave tones from a Holzworth HS9008B pulse shaped by a BBN arbitrary pulse sequencer (APS) quadrature modulation scheme. Readout consists of a simultaneous 2μs probe (Agilent N5183Ms) of the two readout resonators to detect shifts in their frequencies due to their respective qubit states. The I/Q components of the readout signal shift are extracted via down conversion and a digital lock-in routine with a reference tone. They are then used to identify the two-qubit states as $|00\rangle, |01\rangle, |10\rangle, |11\rangle$ via a classification algorithm using support vector machines. Total state preparation and measurement errors, quantified by the basis-state preparation confusion matrix [1] stayed under 5%, with the errors dominated by readout errors associated with qubit state relaxation during the measurement.

The computational subspace spectrum of the transmons was determined via a combination of spectroscopy (directly probing excitations with a $10\mu s$ square pulse) and Ramsey experiments (driving 2 MHz off-resonant, running a typical Ramsey sequence, and noting the deviation of the fitted frequency from 2MHz) [2]. The strength of the drive fields on the qubits was inferred from the frequency of Rabi oscillations of the excited state population incurred by driving at uniform strength for a linearly increasing duration. The linearity of the pulse shaping quadrature channels on the pulse sequencer was characterized by measuring the Rabi oscillation frequency resulting from a sweep over pulse amplitudes, analyzing it via Fourier filtering [3], and correcting for it at the software level.

II. OBTAINING ANALYTICAL SPEED LIMITS

We provide details on how to obtain the analytical speed limit T_{min} defined in Eq. (4) of the main text. Given a two-qubit unitary operator U, the key step is to find the decomposition of U into $(U_1 \otimes U_2)U_d(V_1 \otimes V_2)$ where $U_d = e^{-i\sum_{\gamma=x,y,z}\lambda_\gamma \sigma^\gamma \otimes \sigma^\gamma}$ with $\lambda_{x,y,z} \in [-\frac{\pi}{4}, \frac{\pi}{4}]$, and U_1, V_1 (U_2, V_2) are some single-qubit gates on the first (second) qubit. This decomposition is non-trivial, and the detailed procedure can be found in Ref. [4]. Here we provide the results of the decomposition for the three target gates we studied in the table below. The values of $\lambda_{x,y,z}$ directly lead to T_{min} values for the three target gates shown in Eq. (4) of the main text. Note that an overall phase difference is tolerated for the decomposition of U.

* msingh@mines.edu
† gong@mines.edu

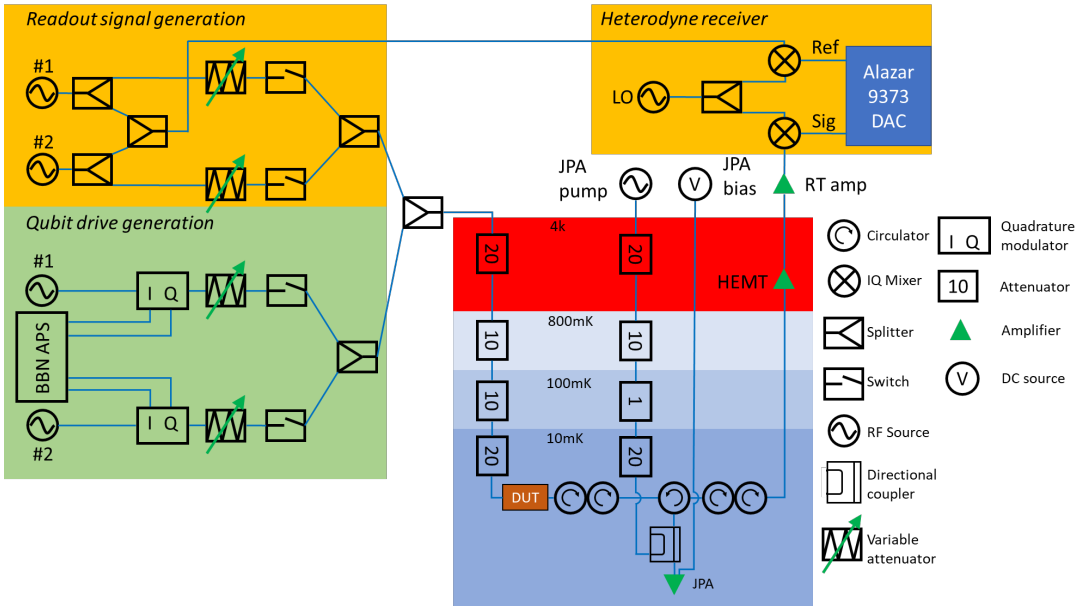

FIG. 1. Our experimental setup composed of qubit drives and heterodyne state readout. Each qubit drive (green section) is shaped via quadrature modulation by a BBN APS1 and Polyphase Microwave AM4080A. Readout (orange) consists of digitally locking in the signal passing through the device with a reference and extracting I/Q shifts used to classify the ground/excited states.

Target gate	U_1	U_2	V_1	V_2	λ_x	λ_y	λ_z
CNOT	$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & i\\ -1 & i\end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}$	$\frac{\pi}{4}$	0	0
SWAP	$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$	$\frac{\pi}{4}$	$\frac{\pi}{4}$	$\frac{\pi}{4}$
$\sqrt{\text{SWAP}}$	$\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$	$\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$	$\frac{\pi}{8}$	$-\frac{\pi}{8}$	$-\frac{\pi}{8}$

III. PROCESS TOMOGRAPHY AND SPAM ERROR CORRECTION

We perform a two-qubit quantum process tomography (QPT) via a standard protocol [1, 5] based on the measurement of a Pauli transfer matrix $\mathcal{R}$ defined as

$$\mathcal{R}_{i,j} \equiv \mathrm{Tr}\left[P_i \mathcal{E}(P_j)\right], \quad P_i \in \{I, X, Y, Z\} \otimes \{I, X, Y, Z\} \tag{1}$$

where $\mathcal{E}(\cdot)$ denotes a quantum map of the process being measured.

For an ideal two-qubit gate U, we can calculate the 16×16 Pauli transfer matrix $\mathcal{R}$ numerically using Eq. (1) (see Fig. 2 for examples). To measure $\mathcal{R}$ experimentally, we need to perform single-qubit rotations before and after the quantum process [6]. Here we apply 9 post-rotations $R_k \in \{I, X_{-\pi/2}, Y_{\pi/2}\}^{\otimes 2}$ and 36 pre-rotations $R_l \in \{Y_{\pi/2}, Y_{-\pi/2}, X_{-\pi/2}, X_{\pi/2}, I, X_\pi\}^{\otimes 2}$ [5], for a total of 324 different sequences, each of which is repeated for 500 times to suppress statistical noise in the measurement. The single-qubit rotations here are achieved via the 'two-axis gate' protocol [3], and they are fined tuned to ensure single-qubit gate fidelities up to 99.1% [7].

Each single experiment returns a measurement outcome as one of the four two-qubit basis states $|j\rangle \in \{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$. We then group the outcomes of all 162,000 experiments using a tensor n_{jkl} which counts the number of measurement outcome state $|j\rangle$ for the k^{th} post-rotation and l^{th} pre-rotation. To take into account possible measurement errors, we separately measure a 4×4 confusion matrix $\mathcal{P}$ where $\mathcal{P}_{i,j}$ denotes the probability of obtaining measurement outcome as the state $|i\rangle$ if we initialize the two qubits in state $|j\rangle$.

To infer the Pauli transfer matrix $\mathcal{R}$ from the measurement outcomes tensor n_{jkl}, we further invoke a maximum likelihood

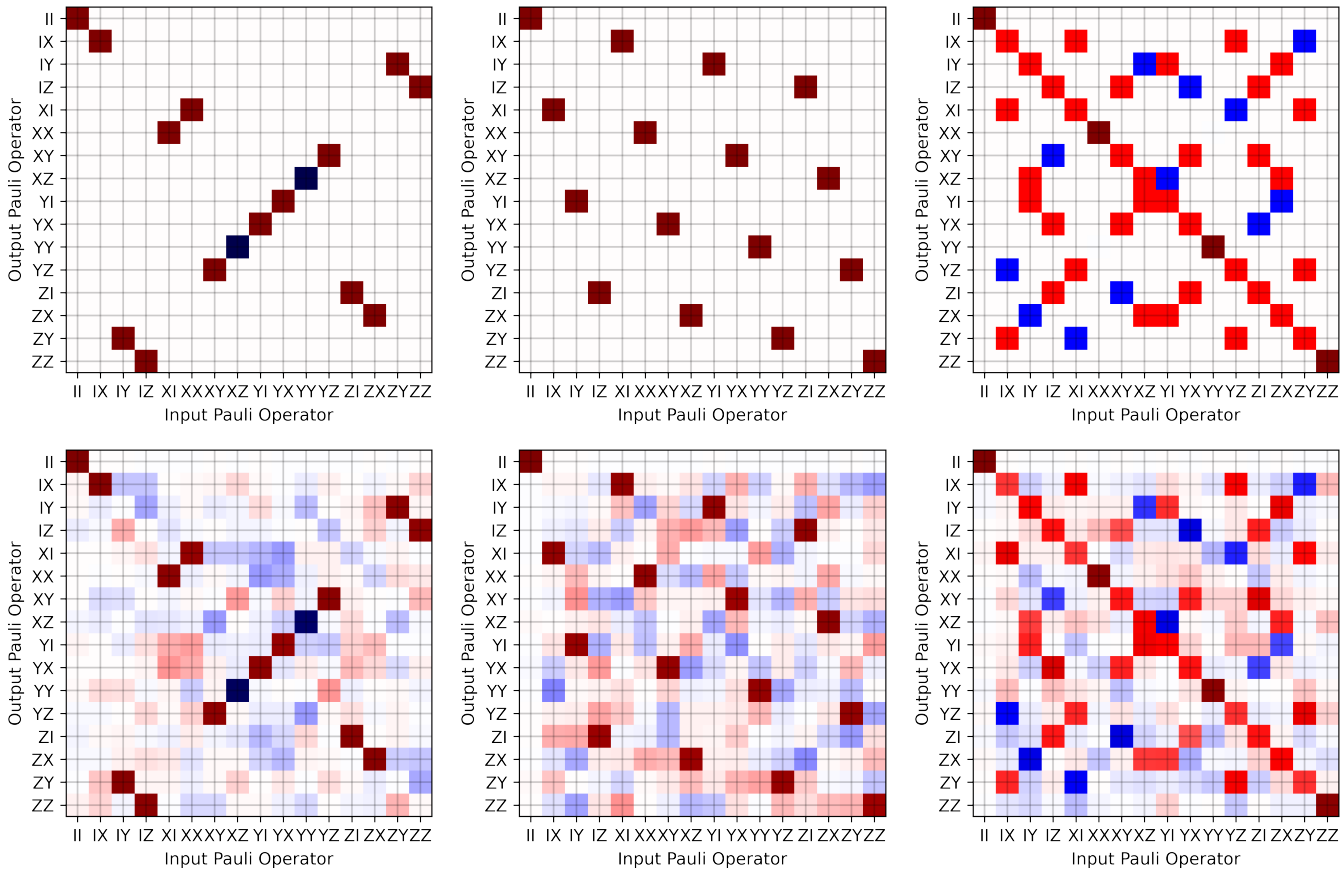

FIG. 2. Quantum Process Tomography based on the Pauli transfer matrix for the three target gates we performed experimentally. Top row from left to right: Pauli transfer matrices for an ideal CNOT gate, SWAP gate, and $\sqrt{\text{SWAP}}$ gate. Bottom row from left to right: Examples of SPAM-error-corrected Pauli transfer matrices obtained experimentally for the CNOT , SWAP, and $\sqrt{\text{SWAP}}$ gates, with average gate fidelities being 95.6%, 93.1%, and 95.7% respectively.

estimation (MLE) method with the following log-likelihood function of $\mathcal{R}$ [1]:

$$\log L(\mathcal{R}) = \sum_{j,k,l} n_{jkl} \log \left(\sum_{m,n=0}^{15} B_{jklmn} \mathcal{R}_{mn} \right) \tag{2}$$

$$B_{jklmn} = \sum_{m',n'=0}^{15} \sum_{j'=0}^{3} \mathcal{P}_{jj'} \left\langle j' \right| P_{m'} \left| j' \right\rangle (\mathcal{R}_k)_{m'm} (\mathcal{R}_l)_{nn'} \operatorname{Tr}(P_{n'} \rho_0) \tag{3}$$

where $\mathcal{R}_k$ and $\mathcal{R}_l$ are the Pauli transfer matrices for the above-defined post-rotation unitary R_k and pre-rotation unitary R_l respectively. $\rho_0 = |00\rangle\langle 00|$ is our initial state and $P_{m'}, P_{n'}$ are defined in Eq. (1). The experimental Pauli transfer matrix $\mathcal{R}$ is then obtained by maximizing $\log L(\mathcal{R})$ under the constraint that $\mathcal{R}$ represents a trace-preserving and completely positive quantum map [1, 5].

Before we perform QPT for the speed-optimized two-qubit gates, we first perform the above QPT procedure for a zero-time evolution to obtain a Pauli transfer matrix $\mathcal{R}_I^{\text{exp}}$, which without state preparation and measurement (SPAM) errors should represent an identity quantum map. The measured $\mathcal{R}_I^{\text{exp}}$ can be used to correct the SPAM errors for a quantum process we perform with nonzero time evolution, whose measured Pauli transfer matrix is denoted by $\mathcal{R}_U^{\text{exp}}$ for a target unitary U.

To achieve SPAM error correction, we first obtain the process matrix χ_I^{exp} and χ_U^{exp} for the corresponding Pauli transfer matrix $\mathcal{R}_I^{\text{exp}}$ and $\mathcal{R}_U^{\text{exp}}$ respectively (by inverting Eq. (5) below) [6]. The process matrix for a quantum map $\mathcal{E}$ is defined via $\mathcal{E}(\rho) = \sum_{m,n} \chi_{mn} P_m \rho P_n$ with P_m, P_n defined in Eq. (1). Second, we obtain the SPAM error corrected process matrix $\chi_U^{\text{corrected}}$ for the target process using

$$\chi_U^{\text{corrected}} = T^{-1}(T\chi_U^{\text{exp}}T^\dagger - V\chi_I^{\text{exp}}V^\dagger + \chi_I^{\text{exp}})(T^\dagger)^{-1}, \tag{4}$$

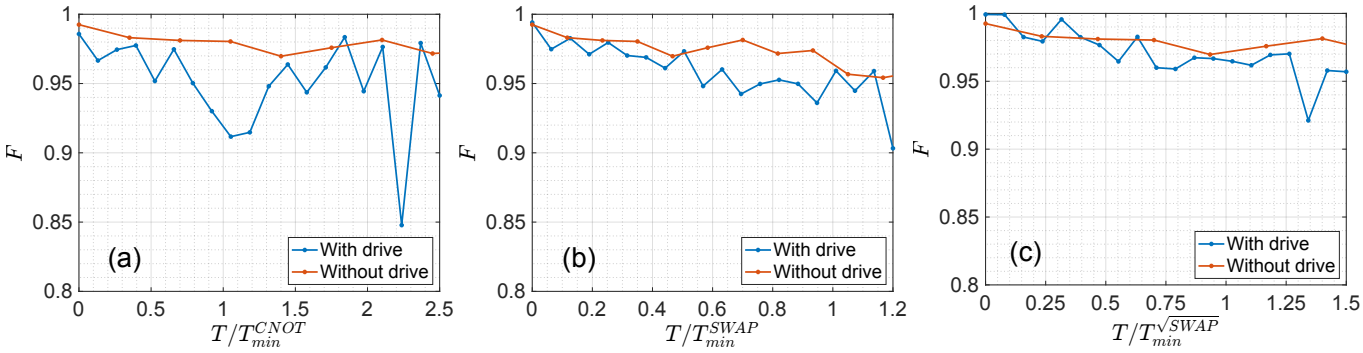

FIG. 3. Average gate fidelity F (blue curves) between the experimental quantum process and the corresponding exact time evolution operator generated by the optimized pulse shapes for a gate time T, with the target gate being CNOT (a), SWAP (b), and $\sqrt{\text{SWAP}}$ (c). The horizontal axis is normalized by the corresponding analytical speed limit for each target gate (see Eq. (4) of the main text). The red curves (identical among the subplots but with different time ranges) reveal the average gate fidelity between a dark evolution quantum process over time T and the corresponding exact time evolution operator without error and dissipation.

where $T_{mn} = \text{Tr}\big(P_m P_n U^\dagger\big)/4$ and $V_{mn} = \text{Tr}(P_m P_n)/4$ [8]. Finally, we convert the error corrected process matrix $\chi_U^{\text{corrected}}$ to the error corrected Pauli transfer matrix $\mathcal{R}_U^{\text{corrected}}$ via

$$\mathcal{R}_{ij} = \sum_{m,n=0}^{15} \chi_{mn}\, \text{Tr}\{P_i P_m P_j P_n\}. \tag{5}$$

Examples of such error-corrected Pauli transfer matrix $\mathcal{R}_U^{\text{corrected}}$ for U being the CNOT, SWAP, or $\sqrt{\text{SWAP}}$ gate are shown in Fig. 2. Finally, $\mathcal{R}_U^{\text{corrected}}$ allows us to find the average gate fidelity F to the target gate U via [9]

$$F = \frac{4\,\text{Tr}\big(\mathcal{R}_U^{\text{corrected}} \mathcal{R}_U^{\text{ideal}}\big) + 1}{5}. \tag{6}$$

where $\mathcal{R}_U^{\text{ideal}}$ represents the Pauli transfer matrix of the ideal target gate U.

IV. ADDITIONAL ERROR ANALYSIS

One error source shared among all our data points is the statistical error due to quantum measurements. To quantify this error, we simulate additional measurements by adding a Gaussian distributed random noise with zero mean and unity standard deviation on each Pauli operator measured during our QPT. This allows us to set an upper bound on the statistical error of the mean that would be obtained on re-performing the full experiment with all other error sources held fixed. As shown in Fig. 3 of the main text, this statistical error on the measured F is less than 1% in all cases.

To better quantify and understand other experimental errors in Fig. 3 of the main text, we calculated the average gate fidelity F between the experimental quantum process and the corresponding exact unitary generated by the engineered Hamiltonian with optimized pulse shapes for each value of gate time T and each target gate demonstrated. As shown by the blue curves in Fig. 3, the infidelity due to experimental error sources are in general within 5%, and we observe a general trend of increasing error as we increase T. To find out potential error sources, we first performed quantum process tomography for the dark evolution of the system (i.e. only under the static Hamiltonian H_0 in the main text) with varying evolution time T and calculate the average gate fidelity of such process with regard to the exact evolution operator e^{-iH_0T}. We find such fidelity also decreases as T increases (see red curves in Fig. 3), from about 99% at $T = 0$ to about 95% at the longest gate time ($1.2T_{\text{min}}^{\text{SWAP}}$) we explored. We attribute the error sources due to the dark evolution to state relaxation (due to finite T_1 and T_2 times), imperfect calibration or fluctuations of the coupling strength g in H_0, and the fact that H_0 is only an approximation of the actual superconducting circuit [3].

To explain why the fidelity in Fig. 3 is noticeably lower when driving the qubits, we numerically simulate the effects of imperfect calibration or noises on the optimized pulse shapes (either for amplitudes or phases) by adding random perturbations to each optimized pulse parameter $\Omega_{i,m}^{\gamma}$ (see main text). We expect such perturbations to be present in our experimental setup with magnitudes of a few percent of Ω_{max}. The simulated average gate fidelity F is shown in Fig. 4, where all other parameters are identical to the exact F curves in Fig. 3 of the main text. We see that our optimization method is robust to small amount (1%) of noises on the pulse shapes, but for larger noises (5%), we can lose a few percent of fidelity compared to the exact cases, and the fidelity lost is larger when F approaches unity, consistent with the behavior of the experimentally measured F in Fig. 3 of

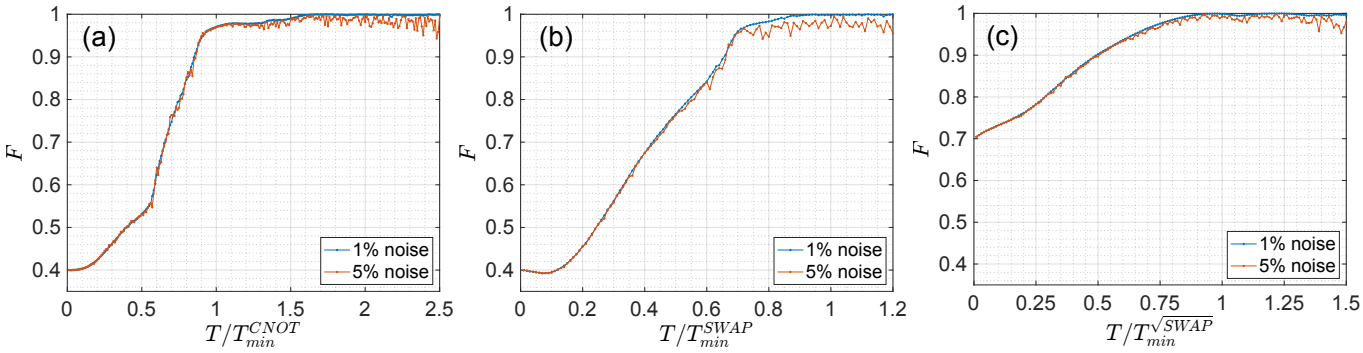

FIG. 4. Average gate fidelity F calculated using the optimized pulse shapes in Fig. 3 of the main text but with random Gaussian noise added to each pulse shape parameter $\Omega_{i,m}^{\gamma}$ (see main text) with the target gate being CNOT (a), SWAP (b), and $\sqrt{\text{SWAP}}$ (c). The blue (red) curves correspond to a standard deviation of the Gaussian noise at $0.01\Omega_{\text{max}}$ ($0.05\Omega_{\text{max}}$).

the main text. We suspect that the combined effects of the error sources in the dark evolution and the engineered pulse shapes, as well as errors caused by leakage to non-qubit states, account for the majority of experimental errors we observe.

[1] Grove, "Histogram based tomography," (2016).
[2] N. F. Ramsey, Phys. Rev. **78**, 695 (1950).
[3] J. Long, *Superconducting Quantum Circuits for Quantum Information Processing*, Ph.D. thesis, University of Colorado, Boulder (2020).
[4] B. Kraus and J. I. Cirac, Phys. Rev. A **63**, 062309 (2001).
[5] J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, M. B. Ketchen, and M. Steffen, Phys. Rev. Lett. **109**, 060501 (2012).
[6] D. Greenbaum, arXiv: 1509.02921 (2015).
[7] J. Long, T. Zhao, M. Bal, R. Zhao, G. S. Barron, H.-s. Ku, J. A. Howard, X. Wu, C. R. H. McRae, X.-H. Deng, G. J. Ribeill, M. Singh, T. A. Ohki, E. Barnes, S. E. Economou, and D. P. Pappas, arXiv: 2103.12305 (2021).
[8] X. Y. Han, T. Q. Cai, X. G. Li, Y. K. Wu, Y. W. Ma, Y. L. Ma, J. H. Wang, H. Y. Zhang, Y. P. Song, and L. M. Duan, Phys. Rev. A **102**, 022619 (2020).
[9] M. A. Nielsen, Physics Letters A **303**, 249 (2002).