High-fidelity three-qubit \(i\)Toffoli gate for fixed-frequency superconducting qubits

Yosep Kim,\(^1\)† Alexis Morvan,\(^1\) Long B. Nguyen,\(^1\) Ravi K. Naik,\(^1,2\) Christian Jünger,\(^1\)
Larry Chen,\(^2\) John Mark Kreikebaum,\(^2,3\) David I. Santiago,\(^1,2\) and Irfan Siddiqi\(^1,2,3\)

\(^1\)Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
\(^2\)Department of Physics, University of California, Berkeley, California 94720, USA
\(^3\)Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

(Dated: July 3, 2023)

The development of noisy intermediate-scale quantum (NISQ) devices has extended the scope of executable quantum circuits\(^12\) with high-fidelity single- and two-qubit gates\(^34,6\). Equipping NISQ devices with three-qubit gates will enable the realization of more complex quantum algorithms\(^7,8\) and efficient quantum error correction protocols\(^10,12\) with reduced circuit depth. Several three-qubit gates have been implemented for superconducting qubits\(^12,14\), but their use in gate synthesis has been limited due to their low fidelity. Here, using fixed-frequency superconducting qubits, we demonstrate a high-fidelity \(i\)Toffoli gate based on two-qubit interactions, the so-called cross-resonance effect\(^15,16\). As with the Toffoli gate, this three-qubit gate can be used to perform universal quantum computation\(^17–19\). The \(i\)Toffoli gate is implemented by simultaneously applying microwave pulses to a linear chain of three qubits, revealing a process fidelity as high as 98.26(2)\%. Moreover, we numerically show that our gate scheme can produce additional three-qubit gates which provide more efficient gate synthesis than the Toffoli and \(i\)Toffoli gates. Our work not only brings a high-fidelity \(i\)Toffoli gate to current superconducting quantum processors but also opens a pathway for developing multi-qubit gates based on two-qubit interactions.

The three-qubit Toffoli gate is universal for reversible classical computation, enabling arbitrary Boolean operations over quantum registers\(^17\). Moreover, together with the Hadamard gate, it forms a universal quantum gate set\(^18,19\). Appending the Toffoli gate to a gate set consisting of the CNOT and single-qubit gates is practically helpful for reducing the overhead of gate synthesis. For instance, both the overall gate count and the circuit depth can be reduced by a factor of \(O(\log n)\) in Shor’s algorithm, which factorizes \(n\)-bit integers\(^7,8\). The oracle functions of Grover’s search algorithm\(^9\) and quantum error correction protocols\(^10,12\) can also be constructed with reduced resource costs. However, implementing a high-fidelity Toffoli gate is experimentally challenging; thus, the operation is usually realized by decomposing it into single- and two-qubit gates\(^20\). This decomposition requires at least five two-qubit gates for fully connected qubits\(^21,22\) and eight for nearest-neighbor connected qubits\(^23\), considerably reducing its fidelity.

Native Toffoli gates have been implemented by exploiting specific features of each quantum computing platform, such as the common vibrational mode of ion strings\(^24\), Rydberg blockade of neutral atoms\(^25\), and electron spin resonance of quantum dots\(^26\). In superconducting circuits, the second or third excited states of frequency-tunable qubits have been employed\(^12–14\), but the best fidelity was limited to \(\sim 87\%\) as higher energy levels are generally more susceptible to decoherence and state leakage. Although a high-fidelity Toffoli gate has been recently reported using all-to-all intermodal couplings of multimode superconducting circuits\(^27\), the multimode architecture restricts the scalability of the gate. Additionally, there are proposals based on the degeneracy of excited-state manifolds\(^28\), Hamiltonian engineering\(^29\), or strong Ising-type couplings\(^30\), which are challenging to implement experimentally.

Here, we demonstrate a high-fidelity \(i\)Toffoli gate by simultaneously driving three superconducting qubits at the same frequency. The simplicity of our scheme makes the gate calibration straightforward, and the gate can even be implemented on cloud-based quantum processors\(^31\). Similar to the Toffoli gate, the \(i\)Toffoli gate inverts a target qubit conditioned upon two control qubits but with a phase shift of \(\pi/2\). This three-qubit gate can also be used for arbitrary Boolean operations\(^17\) and forms a universal quantum gate set with the Hadamard gate\(^19\). To characterize our gate performance independently of state-preparation-and-measurement (SPAM) errors, we leverage cycle benchmarking\(^32\) and measure a process fidelity of 98.26(2)\%. By analyzing the error budget, we find that the fidelity could be further improved using recently developed schemes to suppress undesired longitudinal interactions\(^4,6\). In addition, our numerical study shows that our gate scheme is capable of generating additional three-qubit gates that synthesize random unitary circuits with a lower circuit depth compared to the Toffoli and \(i\)Toffoli gates.

Schematic of the \(i\)Toffoli gate

The \(i\)Toffoli gate is performed on the three fixed-frequency transmons\(^33\) shown in Fig. 1a. We use the left and right transmons as the control qubits and the center one as the target qubit, labeled \(Q_{c1}, Q_{c2}\), and \(Q_t\) accordingly. The device parameters and the experimental setup are presented in Methods and Supplementary Note 1, respectively. The \(i\)Toffoli gate is implemented by simultaneously applying microwave pulses to the three qubits at the target qubit frequency \(\omega_t\). The microwave tones on the control qubits \(Q_{c1,c2}\) induce the cross-resonance effect\(^15,16\) between \(Q_{c1,c2}\) and \(Q_t\) that manifests the interaction Hamiltonians \(H_{CR} \propto Z_{c1,c2}X_t\), while the tone on the target qubit results in \(X_t\) and \(Y_t\) rotations. \(X_k, Y_k\), and \(Z_k\) represent the Pauli operators for qubit \(Q_k\). Taken together, the Hamiltonian of the driven three-qubit system is,

\[
H = \alpha Z_{c1}X_t + \beta X_tZ_{c2} + \gamma X_t + \delta Y_t, \tag{1}
\]

where \(\alpha\) and \(\beta\) depend linearly on the amplitude of the control drives, and \(\gamma\) and \(\delta\) are determined by the target drive’s amplitude and phase. Here we have omitted \(Z_{c1}\) and \(Z_{c2}\) terms,
caused by the off-resonant control drives, as they can be removed at no cost using virtual Z gates.\textsuperscript{35}

A Toffoli-like gate is implemented if Q\textsubscript{c} undergoes a \(\pi\) rotation about the \(x\)-axis on the Bloch sphere for a certain control state and remains invariant otherwise. To explore the conditional evolution of Q\textsubscript{c}, we reduce the Hamiltonian \(H\) in Eq. (1) into the effective target Hamiltonian \(H^(k)_{t} \equiv \epsilon(k|H|k)\)\textsubscript{c} depending on control state \(|k)\)\textsubscript{c}:

\[
\begin{align*}
H_{t}^{00} &= (\alpha + \beta + \gamma)X_{t} + \delta Y_{t}, \\
H_{t}^{01} &= (\alpha - \beta + \gamma)X_{t} + \delta Y_{t}, \\
H_{t}^{10} &= (-\alpha + \beta + \gamma)X_{t} + \delta Y_{t}, \\
H_{t}^{11} &= (-\alpha - \beta + \gamma)X_{t} + \delta Y_{t}.
\end{align*}
\]

Although the conditional inverting operation can be realized for any control state with these Hamiltonians, in this work, we demonstrate the conditional operation for \(|00)\)\textsubscript{c}. Specifically, we consider the scheme in which Q\textsubscript{c} rotates by \(3\pi\) when the control state is \(|00)\)\textsubscript{c} and by \(2\pi\) otherwise. This can be accomplished by first imposing \(\alpha = \beta = \gamma\) to make the conditional Rabi frequencies of Q\textsubscript{c} the same except for \(|00)\): \(\Omega^{r}_{00} = \sqrt{\alpha^{2} + \delta^{2}}\) for \(|00)\)\textsubscript{c} and \(\Omega^{r}_{\text{oth}} = \sqrt{\alpha^{2} + \delta^{2}}\) for the other control states. Then, we can make \(\Omega^{r}_{00} = 1.5\Omega^{r}_{\text{oth}}\) with \(\delta = \sqrt{2}\pi/5\alpha\), and a Toffoli-like gate is realized by setting the gate duration to \(\tau = 2\pi/\Omega^{r}_{\text{oth}}\).

Figure 1b shows the conditional evolution of Q\textsubscript{c} from the initial state \(|0)\textsubscript{t}\). Due to the Pancharatnam-Berry geometric phase,\textsuperscript{36,37} the target qubit accumulates a \(3\pi/2\) phase for control state \(|00)\), and a \(\pi\) phase for the other control states, making the operation the \(\text{i}\)Toffoli gate rather than the conventional Toffoli gate. Although the rotation axis for \(|00)\)\textsubscript{c} is not aligned along the \(x\)-axis, this axis can effectively be transformed onto the \(x\)-axis by sandwiching the three-qubit operation with virtual Z gates on the target qubit. The circuit diagram of the \(\text{i}\)Toffoli operation is shown in Fig. 1c. The \(\text{i}\)Toffoli gate differs from the Toffoli gate only by a controlled-phase-gate phase shift to \(|00)\).

**Gate operation and geometric phase error**

We now turn to the experimental implementation of the \(\text{i}\)Toffoli gate. For the gate operation, as shown in Fig 2a, we use microwave pulses with a flat-top envelope and 30-nS cosine ramps. The pulses are applied at the target qubit frequency \(\omega_{t}\) for pulse duration \(\tau\). The gate calibration is conducted by tuning the amplitude and phase of each pulse to make the conditional Rabi frequencies \(\Omega^{r}_{00} = 1.5\Omega^{r}_{\text{oth}}\) to 1.5\(\Omega^{r}_{10} = 1.5\Omega^{r}_{11}\). The calibration procedure is detailed in Methods. Figure 2b shows the experimental results of the conditional Rabi oscillations starting from \(|0)\textsubscript{t}\), and the \(\text{i}\)Toffoli gate is implemented at \(\tau = 353\) ns.

The oscillation amplitudes of less than 1 in Fig. 2b indicate that the rotation axis is not in the \(xy\)-plane. The axial tilt is mainly due to parasitic longitudinal (ZZ) interactions between nearest-neighbor qubits.\textsuperscript{39,40} These undesired interactions add a conditional \(\Delta^{k}Z_{c}\) term to the effective target Hamiltonians in Eq. (2) for each control state \(|k)\)\textsubscript{c}. This is detrimental to the \(\text{i}\)Toffoli gate as a non-zero \(\Delta^{k}\) induces a different geometric phase (GP) for \(|0)\textsubscript{t}\) and \(|1)\textsubscript{t}\). Note that the GP is given by \(\Delta \phi = \frac{1}{2} \int_{C} (1 - \cos \theta) d\phi\) for a trajectory C on the Bloch sphere, where \(\theta\) is the angle between the state vector and the rotation axis and \(\phi\) denotes the rotation angle.\textsuperscript{46,37} Considering the ro-
For the $i$Toffoli gate, the three qubits are simultaneously driven using flat-top cosine-ramp pulses (blue) at the transition frequency of $Q_i$. To investigate the conditional evolution of $Q_i$, the control state of $Q_{i1}$ and $Q_{i2}$ is initialized in $|00\rangle_c$, $|01\rangle_c$, $|10\rangle_c$, and $|11\rangle_c$, and then the expectation value of $Q_i$ is measured in the Pauli $Z$ basis. The gate calibration procedure is detailed in Methods. After the gate calibration, the conditional Rabi oscillations of $Q_i$ initially in $|0\rangle$, are measured as functions of the pulse duration $\tau$. The gate duration of the $i$Toffoli gate (dashed line) is determined by fitting the oscillations.

The condition $\Delta^{00} = 0$ is satisfied if the drive tones are at the transition axis tilted by undesired interactions, $\cos \theta = \Delta^{kl}/\Omega_i^{kl}$ for $|0\rangle$ and $-\Delta^{kl}/\Omega_i^{kl}$ for $|1\rangle$. Thus, in the absence of qubit decoherence, the $i$Toffoli operation $U_{GP}$ maps the computational basis states as follows:

$$
|k\rangle_c |0\rangle_t \rightarrow e^{i\pi (1-\Delta^{kl}/\Omega_i^{kl})} |k\rangle_c |0\rangle_t,

|k\rangle_c |1\rangle_t \rightarrow e^{i\pi (1+\Delta^{kl}/\Omega_i^{kl})} |k\rangle_c |1\rangle_t,
$$

where $|k\rangle_c \in \{ |0\rangle_c, |1\rangle_c, |10\rangle_c, |11\rangle_c \}$. In the case of $|00\rangle_c$, $\Delta^{00}$ is presumed to be zero so that the target qubit's computational states are perfectly inverted with a phase shift of $3\pi/2$. The fidelity measured above is subject to state-preparation - and-measurement (SPAM) errors. The population leakage to the second excited state is measured to be less than 0.1% for all the qubits.

1000 Monte Carlo simulation runs taking into account statistical errors. The population leakage to the second excited state is measured to be less than 0.1% for all the qubits.

The fidelity measured above is subject to state-preparation-and-measurement (SPAM) errors. To extract a SPAM-error-free fidelity estimate of a quantum operation, interleaved randomized benchmarking (IRB) is typically used\textsuperscript{43-45}. However, IRB is impractical for the $i$Toffoli gate due to the complexity of three-qubit Clifford twirling\textsuperscript{46}. As an alternative, we perform cycle benchmarking (CB) which relies on Pauli twirling\textsuperscript{42}. The CB protocol has better scalability than IRB as the complexity of twirling operation is reduced. The circuit schematic of the CB protocol is depicted in Fig. 3b. The Pauli
twirling of the \( i \)Toffoli gate tailors the gate errors to stochastic Pauli error channels, and the initial state and measurement basis are chosen to pick up a Pauli channel using True-Q\(^37\).

We run CB for all three-qubit Pauli channels with depth \( m \in \{2, 4, 16, 32\} \) and 30 samples for each \( m \). In order to decouple SPAM errors from fidelity estimate, the fidelity \( p_k \) of each Pauli channel \( k \) is extracted by fitting the fidelity decay to an exponential model \( \exp(-\lambda t) \). However, since the fidelity still includes the error from Pauli-twirling operations, the Pauli-twirling error needs to be separately estimated by performing CB without the \( i \)Toffoli gate as a reference. Figure 3e shows the cumulative fidelity histograms of all three-qubit Pauli channels for the Pauli-twirled \( i \)Toffoli gate \( p_k^{(\text{ref})} \) and the reference \( p_k^{(\text{ref})} \). The fidelity of each Pauli channel is exhibited in Supplementary Note 2. From the results, the process fidelity of the \( i \)Toffoli gate is estimated to be 98.26(2)% by averaging \( p_k^{(\text{ref})} / p_k^{(\text{ref})} \) over all the Pauli channels\(^32\). The uncertainty represents the fitting errors of fidelity decays.

Additionally, to get insights into factors limiting improvement in process fidelity, we perform CB for \( i \)Toffoli gates having different gate durations and analyze the error budget. As the error sources, the geometric phase (GP) error and decoherence of the qubits are considered. The GP error is estimated according to Eq. (3), and the coherence-limited infidelity\(^48\) is evaluated using the relaxation (\( T_1 \)) and the Hahn-echo dephasing (\( T_2^{\text{echo}} \)) times of each qubit and the gate duration. See Methods for the calculation of coherence-limited infidelity. We experimentally observed that the \( T_2^{\text{echo}} \) of \( Q_{c1} \) and \( Q_{c2} \) are reduced under the \( i \)Toffoli drive relative to the bare coherence times. Accordingly, we calculated the coherence limits using both the bare coherence times and the coherence times under drive. The measurement sequences and data of the coherence times are presented in Supplementary Note 3. Figure 4 shows that the process fidelities obtained from CB are well explained by the estimated error budget. The results reveal that undesired longitudinal interactions, or the GP errors, increase with decreasing gate duration, limiting further reduction in gate duration and improvement in process fidelity.
Gate synthesis of three-qubit circuits

Finally, we numerically explore gate synthesis of three-qubit circuits to demonstrate the utility of the \textit{i}Toffoli gate. To this end, we consider an ansatz circuit \( V \) depicted in Fig. 5a with depth \( m \). The gate synthesis of a target circuit \( U \) is conducted by optimizing the ansatz parameters to minimize the distance between \( V \) and \( U \), or their infidelity \( 1 - F(V, U) \). The simulation results in Fig. 5b show that both Toffoli and \textit{i}Toffoli gates can be used to synthesize arbitrary three-qubit Clifford circuits at \( m_{\text{Cliff}} = 6 \) and Haar-random unitary circuits at \( m_{\text{Haar}} = 10 \). Additionally, the results indicate that the ansatz circuit with the \textit{i}Toffoli gate can synthesize more various target circuits at lower depths than the Toffoli gate.

Furthermore, we estimate threshold depths \( m_{\text{Cliff}} \) and \( m_{\text{Haar}} \) of our gate scheme by varying \( Y_i \)'s amplitude \( \delta \) in the Hamiltonian of Eq. (1). For this simulation, the other Hamiltonian amplitudes and the gate duration are fixed to the same values used for implementing the \textit{i}Toffoli gate. As \( Y_i \) is the only non-commuting term in the Hamiltonian, it plays a key role in preventing the Hamiltonian operation from being a simple combination of single- and two-qubit gates. Interestingly, the results in Fig. 5e reveal that, over a range of \( \delta \) values, our gate scheme can produce three-qubit gates which synthesize random unitary circuits at \( m_{\text{Cliff}} = m_{\text{Haar}} = 9 \). That is, while the structured Toffoli and \textit{i}Toffoli gates have advantages for Clifford circuits, these alternative three-qubit gates could be beneficial for implementing random unitary circuits.

Discussion and outlook

We have demonstrated a single-step \textit{i}Toffoli gate scheme based on the cross-resonance (CR) effect with fixed-frequency transmon qubits. As the CR effect can readily be implemented between capacitively coupled superconducting qubits\textsuperscript{15,16}, our scheme is applicable to a number of superconducting architectures including cloud-based quantum processors\textsuperscript{31}. In our gate scheme, the non-commuting term \( Y_i \) in Eq. (1) plays a key role in preventing the full Hamiltonian operation from being decomposable into simple consecutive operations. By varying the amplitude of the \( Y_i \) term, we have numerically found that this gate scheme can produce additional three-qubit gates providing more efficient gate synthesis than both the Toffoli and \textit{i}Toffoli gates. We hope our work will trigger further studies to develop diverse multi-qubit gates derived from two-qubit interactions involving non-commuting Hamiltonian operators.

To characterize a 353-ns \textit{i}Toffoli gate without state-preparation-and-measurement (SPAM) errors, we have leveraged cycle benchmarking and measured a process fidelity of 98.26(2)\%. The infidelity is primarily due to decoherence of the qubits and undesired longitudinal (ZZ) interactions. By analyzing the error budget, we have found that such undesired interactions limit a further improvement in process fidelity. Recently, the multi-path coupling scheme\textsuperscript{3} and the microwave-induced ZZ tuning scheme\textsuperscript{5,6} have been demonstrated to suppress ZZ interactions and improve the fidelity of two-qubit CR gates. These schemes will directly benefit the fidelity of our CR-based \textit{i}Toffoli gate as well. We expect that our high-fidelity three-qubit gate will open a pathway for executing complex quantum circuits, such as Shor’s and Grover’s algorithms\textsuperscript{5,9}, and for exploring efficient quantum error correction protocols\textsuperscript{10-12}.

* Correspondence to Y.K. (yosep9201@gmail.com)

References

1. Preskill, J. Quantum computing in the NISQ era and beyond. \textit{Quantum} \textbf{2}, 79 (2018).
2. Bharti, K. et al. Noisy intermediate-scale quantum (NISQ) algorithms. Preprint at https://arxiv.org/abs/2101.08448 (2021).
3. Sung, Y. et al. Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler. \textit{Phys. Rev. X} \textbf{11}, 021058 (2021).
4. Kandala, A. et al. Demonstration of a high-fidelity CNOT for fixed-frequency transmons with engineered ZZ suppression. Preprint at https://arxiv.org/abs/2011.07050 (2020).
5. Mitchell, B. K. et al. Hardware-efficient microwave-activated tunable coupling between superconducting qubits. Preprint at https://arxiv.org/abs/2105.05384 (2021).
6. Wei, K. X. et al. Quantum crosstalk cancellation for fast entangling gates and improved multi-qubit performance. Preprint at https://arxiv.org/abs/2106.00675 (2021).
7. Häner, T., Roetteler, M. & Svore, K. M. Factoring using 2n+2 microwave-induced ZZ tuning scheme with linear couplings and fixed transition frequencies. \textit{Phys. Rev. B} \textbf{81}, 134507 (2010).
8. Gidney, C. & Ekerä, M. How to factor 2048-bit RSA integers in 8 hours using 20 million noisy qubits. \textit{Quantum} \textbf{5}, 433 (2021).
9. Figgatt, C. et al. Complete 3-Qubit Grover search on a programmable quantum computer. \textit{Nat. Commun.} \textbf{8}, 1918 (2017).
10. Yoder, T. J., Takagi, R. & Chuang, I. L. Universal fault-tolerant gates on concatenated stabilizer codes. \textit{Phys. Rev. X} \textbf{6}, 031039 (2016).
11. Paetznick, A. & Reichardt, B. W. Universal fault-tolerant quantum computation with only transversal gates and error correction. \textit{Phys. Rev. Lett.} \textbf{111}, 090505 (2013).
12. Reed, M. D. et al. Realization of three-qubit quantum error correction with superconducting circuits. \textit{Nature} \textbf{482}, 382-385 (2012).
13. Fedorov, A., Steffen, L., Baur, M., da Silva, M. P. & Wallraff, A. Implementation of a Toffoli gate with superconducting circuits. \textit{Nature}, \textbf{481}, 170-172 (2012).
14. Hill, A. D., Hodson, M. J., Didier, N. & Reagor, M. J. Realization of arbitrary doubly-controlled quantum phase gates. Preprint at https://arxiv.org/abs/2108.01652 (2021).
15. Rigetti, C. & Devoret, M. Fully microwave-tunable universal quantum computing. \textit{Rev. Mod. Phys.} \textbf{89}, 105002 (2017).
16. Toffoli, T. \textit{Reversible Computing} (Lect. Notes Computer Sci. 85, Springer, 1980).
17. Shai, Y. Both Toffoli and controlled-NOT need little help to do universal quantum computing. \textit{Quantum Inf. Comput.} \textbf{3}, 84-92 (2003).
18. Aharonov, D. A simple proof that Toffoli and Hadamard are quan-
tum universal. Preprint at https://arxiv.org/abs/quant-ph/0301040 (2003).
20 Figgatt, C. et al. Parallel entangling operations on a universal ion-trap quantum computer. Nature 572, 368-372 (2019).
21 Barenco, A. et al. Elementary gates for quantum computation. Phys. Rev. A 52, 3457-3467 (1995).
22 Yu, N., Duan, R. & Ying, M. Five two-qubit gates are necessary for implementing the Toffoli gate. Phys. Rev. A 88, 010304(R) (2013).
23 Smith, E. et al. LEAP: Scaling numerical optimization based synthesis using an incremental approach. Preprint at https://arxiv.org/abs/2106.11246 (2021).
24 Monz, T. et al. Realization of the quantum Toffoli gate with trapped ions. Phys. Rev. Lett. 102, 040501 (2009).
25 Levine, H. et al. Parallel implementation of high-fidelity multiqubit gates with neutral atoms. Phys. Rev. Lett. 123, 170503 (2019).
26 Hendrickx, N. W. et al. A four-qubit germanium quantum processor. Nature 591, 580–585 (2021).
27 Roy, T. et al. Programmable superconducting processor with native three-qubit gates. Phys. Rev. Appl. 14, 014072 (2020).
28 Khazali, M. & Mølmer, K. Fast multiqubit gates by adiabatic evolution in interacting–state manifolds of Rydberg atoms and superconducting circuits. Phys. Rev. X 10, 021054 (2020).
29 Zabedinejad, E., Ghosh, J. & Sanders, B. C. High-fidelity single-shot Toffoli gate via quantum control. Phys. Rev. Lett. 15, 200502 (2015).
30 Rasmussen, S. E., Groenland, K., Gerritsma, R., Schoutens, K. & Zinner, N. T. Single-step implementation of high-fidelity n-bit Toffoli gates. Phys. Rev. A 101, 022308 (2020).
31 IBM Quantum: https://quantum-computing.ibm.com/
32 Erhard, A. et al. Characterizing large-scale quantum computers via cycle benchmarking. Nat. Commun. 10, 5347 (2019).
33 Koch, J. et al. Charge insensitive qubit design derived from the Cooper pair box. Phys. Rev. A 76, 042319 (2007).
34 Wallraff, A. et al. Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. Nature 431, 162-167 (2004).
35 McKay, D. C., Wood, C. J., Sheldon, S., Chow, J. M. & Gambetta, J. M. Efficient Z gates for quantum computing. Phys. Rev. A 96, 022330 (2017).
36 Jordan, T. F. Berry phases for partial cycles. Phys. Rev. A 38, 1590–1592 (1988).
37 Cho, Y.-W. et al. Emergence of the geometric phase from quantum measurement back-action. Nat. Phys. 15, 665-670 (2019).
38 Selinger, P. Quantum circuits of T-depth one. Phys. Rev. A 87, 042302 (2013).
39 Tripathi, V., Khazali, M. & Korotkov, A. N. Operation and intrinsic error budget of a two-qubit cross-resonance gate. Phys. Rev. A 100, 012301 (2019).
40 Magesan, E. & Gambetta, J. M. Effective Hamiltonian models of the cross-resonance gate. Phys. Rev. A 101, 052308 (2020).
41 Nielsen, M. A. A simple formula for the average gate fidelity of a quantum dynamical operation. Phys. Lett. A, 303, 249-252 (2002).
42 Nachman, B., Urbanek, M., de Jong, W. A. & Bauer, C. W. Unfolding quantum computer readout noise. npj Quantum Inf. 6, 84 (2020).
43 Magesan, E., Gambetta, J. M. & Emerson, J. Scalable and robust randomized benchmarking of quantum processes. Phys. Rev. Lett. 106, 180504 (2011).
44 Magesan, E. et al. Efficient measurement of quantum gate error by interleaved randomized benchmarking. Phys. Rev. Lett. 109, 080505 (2012).
45 Morvan, A. et al. Qutrit randomized benchmarking. Phys. Rev. Lett. 126, 210504 (2021).
46 McKay, D. C., Sheldon, S., Smolin, J. A., Chow, J. M. & Gambetta, J. M. Three qubit randomized benchmarking. Phys. Rev. Lett. 122, 200502 (2019).
47 Beale, S. J. et al. True-Q: https://trueq.quantumbenchmark.com (2020).
48 Dawkins, H., Wallman, J. & Emerson, J. Combining $T_1$ and $T_2$ estimation with randomized benchmarking and bounding the diamond distance. Phys. Rev. A 102, 022220 (2020).
49 Johansson, J. R., Nation, P. D. & Nori, F. QuTiP 2: A Python framework for the dynamics of open quantum systems. Comp. Phys. Comm. 184, 1234-1240 (2013).
50 Sheldon, S., Magesan, E., Chow, J. M. & Gambetta, J. M. Procedure for systematically tuning up cross-talk in the cross-resonance gate. Phys. Rev. A 93, 060302(R) (2016).

Methods

Device parameters

The transition frequencies and anharmonicities of transmon qubits \{Q, Q\} are measured as \{5.254, 5.331, 5.491\} GHz and \{-277.1, -272.1, -271.8\} MHz, respectively, using Ramsey spectroscopy. Their coherence times are estimated through the measurements of the excited state lifetime and the Hahn echo decay: $T_1 = \{70(7), 61(3), 57(9)\}$ μs and $T_2 = \{60(5), 75(9), 66(9)\}$ μs. To obtain the static longitudinal (ZZ) interaction strength between pairs of qubits, conditional Ramsey measurements are conducted, yielding $\zeta_{QZ} / 2\pi = 96$ kHz for \{Q, Q\} and $\zeta_{ZZ} / 2\pi = 171$ kHz for \{Q, Q\}. The readout frequencies are measured as \{6.564, 6.623, 6.679\} GHz, and the readout fidelities are obtained as $P(0|0) = \{0.998, 0.997, 0.996\}$ and $P(1|1) = \{0.948, 0.982, 0.981\}$, where $P(x|y)$ is the probability that the qubit in state $y$ is measured to be in state $x$.

Calibration procedure

As shown in Fig 2a, individual microwave pulses are applied to all three qubits. Using the effective amplitude $a_k$ and phase $\phi_k$ of each pulse, we can recast the three-qubit Hamiltonian in Eq. (1) into,

$$H = a_{c1} \cos \phi_{c1} Z_1Z_4 + a_{c1} \sin \phi_{c1} Z_1Y_1 + a_{c2} \cos \phi_{c2} X_2Z_2 + a_{c2} \sin \phi_{c2} Y_2Z_2 + a_1 \cos \phi X_1 + a_1 \sin \phi Y_1.$$

In order to implement the $\iota$-Toffoli gate, the pulse parameters need to be calibrated to satisfy $\alpha = \beta = \gamma = \sqrt{5/2}\delta$ in terms of Eq. (1). Whether the condition is satisfied can be checked by measuring the conditional Rabi frequencies $\Omega_{Q1}^\delta$ of $Q$, initially in $|0\rangle$. The calibration procedure of the $\iota$-Toffoli gate is as follows:

1. Set $\alpha$: Fix $a_{c1}$ and $\phi_{c1}$ of the first control drive. The gate duration will be inversely proportional to the amplitude $a_{c1}$.
2. Set $\beta = \alpha$: Tune $a_{c2}$ and $\phi_{c2}$ of the second control drive to make $\Omega_{Q1}^\delta = \Omega_{Q1}^\alpha$.
3. Remove ZY: Sweep $\phi_{c1}$ and $\phi_{c2}$ simultaneously with the same offset to remove $Z_1Y_1$ and $Y_1Z_2$. The amplitudes are estimated by performing Hamilton tomogra-
Note that the simultaneous phase sweep will not change the condition of $\Omega_{q}^{01} = \Omega_{q}^{10}$.

4. Set $\gamma = \alpha$: Tune $\phi_{t}$ of the target drive with $\phi_{t} = 0$, for $X_{t}$, to achieve $\Omega_{t}^{11} = \Omega_{t}^{01}$.

5. Set $\delta = \sqrt{27/5}\alpha$: Add additional target drive for $Y_{t}$ with $\phi_{t} = \pi/2$ and tune the amplitude to make $\Omega_{t}^{00} = 1.5\Omega_{t}^{01}$.

If needed, the calibration procedure is repeated until $\Omega_{t}^{00} = 1.5\Omega_{t}^{01} = 1.5\Omega_{t}^{10} = 1.5\Omega_{t}^{11}$ is fulfilled. Even if there are undesired longitudinal interactions, the calibration procedure can be conducted in the same manner. Finally, the gate duration is determined so that $Q_{t}$ rotates by $3\pi$ for control state $|00\rangle_{c}$.

**Coherence-limited process fidelity**

A Pauli transfer matrix (PTM) of a Pauli-twirled single-qubit decoherence channel is written as\(^{48}\),

$$
\mathbf{E} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & \sqrt{1 - \gamma_{1}} \sqrt{1 - \gamma_{2}} & 0 & 0 \\
0 & 0 & \sqrt{1 - \gamma_{1}} \sqrt{1 - \gamma_{2}} & 0 \\
0 & 0 & 0 & 1 - \gamma_{1}
\end{pmatrix},
$$

where $\gamma_{1} = 1 - e^{-\Gamma_{1}\tau}$ and $\gamma_{2} = 1 - e^{-2\Gamma_{2}\tau}$ represent the degree of relaxation and dephasing for gate duration $\tau$, respectively. The relaxation and dephasing rates of a qubit can be experimentally estimated by measuring the coherence times: $\Gamma_{1} = 1/T_{1}$ and $\Gamma_{2} = 1/T_{2}^{{\text{echo}}} - 1/2T_{1}$. In the absence of crosstalk errors, the PTM of a Pauli-twirled 3-qubit channel is simply given by the tensor product $\mathbf{E}^{\otimes 3}$, and the coherence-limited process fidelity is calculated as,

$$
F_{cl} = \frac{1}{d^{2}} \text{Tr}[\mathbf{E}^{\otimes 3}]
$$

$$
= \frac{1}{d^{2}} \prod_{q=1}^{3} \left( 2 + 2 \sqrt{1 - \gamma_{1}^{(q)}} \sqrt{1 - \gamma_{2}^{(q)} - \gamma_{1}^{(q)}} \right),
$$

where $\gamma_{1}^{(q)}$ and $\gamma_{2}^{(q)}$ denote the decoherence of qubit $q$, and $d = 8$ is the Hilbert space dimension.

**Acknowledgements**

The authors gratefully acknowledge the conversations and insights of J. Wallman and I. Hincks. This work was supported by the Quantum Testbed Program of the Advanced Scientific Computing Research Division, Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.
SUPPLEMENTARY NOTE 1 - EXPERIMENTAL SETUP

A detailed experimental schematic is shown in Fig. S1. We implement the $i$Toffoli gate on the 8-qubit superconducting quantum processor used in Ref. [1]. Specifically, in this work, we employ the top-left three transmon qubits in the processor image. The processor is operated in a BlueFors XLD dilution refrigerator with a base temperature of 10 mK and magnetically shielded with Copper (Cu), Tin (Sn), and Mu-metal cylinders.

Qubit control and readout IF pulses are generated using a Keysight PXI Arbitrary Waveform Generator (AWG) at a sample rate of 1 GS/s. The IF pulses are then upconverted using a 5.415 GHz tone for control pulses and a 6.83 GHz tone for readout pulses via IQ mixers. Two Keysight MXG N5183B separately generate these local oscillator (LO) tones. To suppress thermal photons transmitted from higher-temperature stages, RF attenuators are installed at each stage and K&L low-pass filters are additionally used at the base stage.

After the readout pulses are dispersively coupled with qubit states, the readout signals are amplified by a traveling wave parametric amplifier (TWPA) at 10 mK, a high-electron mobility transistor amplifier (HEMT) at 4K, and room temperature amplifiers. The TWPA is pumped by a 16.5 dBm, 7.42 GHz tone (Hittite HMC M2100). To analyze the readout outcomes, the amplified readout signals are downconverted to IF signals with a 6.83 GHz LO tone and digitized by an Alazar ADC at a sample rate of 1 GS/s.

SUPPLEMENTARY NOTE 2 - CYCLE BENCHMARKING AND PAULI TRANSFER MATRIX

We characterize the quantum operation of a 353-ns $i$Toffoli gate using two methods. First, as described in the main text, we run CB for all three-qubit Pauli channels with depth $m \in \{2, 4, 16, 32\}$ and 30 samples for each $m$. The fidelity $p_k$ of each Pauli channel $k$ is extracted by fitting the fidelity decay to the exponential model $Ap_k^m$. This allows decoupling the fidelity from state-preparation-and-measurement (SPAM) errors. The Pauli fidelities of the Pauli-twirled $i$Toffoli gate $p_k^{(itof)}$ and the reference $p_k^{(ref)}$ are shown in Figure S2. The reference fidelities $p_k^{(ref)}$ represent the
errors from Pauli-twirling operations. By deducting the Pauli-twirling errors, the SPAM-error-free Pauli fidelities of the \(i\)Toffoli gate can be obtained as \(p_k^{(i\text{Toff})}/p_k^{(\text{ref})}\). We estimate process fidelity by averaging Pauli fidelities over all three-qubit Pauli channels [2]. Consequently, the process fidelities of the Pauli-twirled \(i\)Toffoli, the reference, and the \(i\)Toffoli gates are obtained as 97.51(2)\%, 99.24(1)\%, and 98.26(2)\%, respectively. The uncertainties represent the fitting errors of the fidelity decay.

![Graph showing infidelity vs. Pauli decay term](image)

**Fig. S2.** Detailed cycle benchmarking results of a 353-ns \(i\)Toffoli gate and the reference. The uncertainty of each data is obtained from the fitting error. The horizontal solid lines represent the averaged infidelities of the Pauli-twirled \(i\)Toffoli gate and the reference.

Second, in order to see the quantum operation of a 353-ns \(i\)Toffoli gate more explicitly, the Pauli transfer matrix (PTM) is reconstructed. To this end, we prepare 64 three-qubit input states by applying single-qubit operations \{I, X\(_{\pi/2}\), X\(_{\pi}\), Y\(_{\pi/2}\)\} for each qubit. Then, we measure the output states and reconstruct the PTM from the data via maximum likelihood estimation method [3]. The readout errors in the data are corrected based on the readout fidelities [4]. Figures S3 and S4 show the reconstructed PTM, \(\mathcal{E}_{\text{exp}}\), and the difference with the ideal PTM, \(\mathcal{E}_{\text{ideal}}\). The process fidelity is evaluated to be \(\mathcal{F}_{\text{PTM}} = \text{Tr}[\mathcal{E}_{\text{ideal}}\mathcal{E}_{\text{exp}}]/d^2 = 97.1(8)\%\) where \(d = 8\) is the Hilbert space dimension [2]. The uncertainty is obtained by performing 1000 Monte Carlo simulation runs taking into account statistical errors.
Fig. S3. Experimentally reconstructed Pauli transfer matrix of a 353-ns iToffoli gate, $\mathcal{E}_{\text{exp}}$, via maximum likelihood estimation method. The readout errors are corrected based on the readout fidelities. The process fidelity is estimated to be 97.1(8)% with respect to the ideal gate.
Fig. S4. The difference between the ideal iToffoli gate’s Pauli transfer matrix (PTM) and the experimentally reconstructed PTM in Fig. S3, $\mathcal{E}_{\text{ideal}} - \mathcal{E}_{\text{exp}}$. The largest element difference is 0.055. Note that the color bar is rescaled to a range between -0.1 to 0.1 for clarity.
As discussed in the main text, decoherence of qubits is one of the main error sources in our iToffoli gate, along with geometric phase errors. When implementing the iToffoli gate, we drive control qubits at off-resonance frequency and the drive shifts the qubit frequencies. The frequency shift could induce faster energy relaxation (lower $T_1$) of control qubits if the qubit frequencies come into resonance with strongly coupled two-level systems [5]. On the other hand, a faster dephasing effect (lower $T_2$) under off-resonant drive has also been experimentally observed [6].

To investigate these decoherence effects under the iToffoli drive, we measure the energy relaxation rate $\Gamma_1 = 1/T_1$ and the Hahn-echo dephasing rate $\Gamma_2^{\text{echo}} = 1/T_2^{\text{echo}}$ of control qubits as functions of a relative iToffoli drive amplitude. Figures S5a and b depict the pulse sequences for the measurements, respectively. From the experimental results shown in Fig. S5c, we observe that $\Gamma_2^{\text{echo}}$ of Q$_c1$ and Q$_c2$ gradually increases as relative drive amplitude increases, whereas $\Gamma_1$ is relatively independent to the drive amplitude except the peaks at relative amplitudes of 0.5 and 0.9. The peaks are suspected to be due to frequency collisions caused by frequency shifts. Note that the iToffoli gates with gate durations of 482, 353, 280, and 243 ns, in Fig. 4 of the main text, are implemented at relative amplitudes of 0.2, 0.3, 0.4, and 0.5, respectively.

Fig. S5. Decoherence of control qubits under iToffoli drive. Pulse sequences used to measure, a, the energy relaxation rate $\Gamma_1 = 1/T_1$ and b, the Hahn-echo dephasing rate $\Gamma_2^{\text{echo}} = 1/T_2^{\text{echo}}$ of the first control qubit Q$_c1$. The decoherence rates of the second control qubit Q$_c2$ are separately measured in the same manner. c, The measurement results of $\Gamma_1$ and $\Gamma_2^{\text{echo}}$ versus relative drive amplitudes. The error bars represent statistical fluctuations.

[1] Hashim, A. et al. Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor. Preprint at https://arxiv.org/abs/2010.00215 (2020).
[2] Nielsen, M. A. A simple formula for the average gate fidelity of a quantum dynamical operation. Phys. Lett. A, 303, 249-252 (2002).
[3] Chow, J. M. et al. Universal quantum gate set approaching fault-tolerant Thresholds with Superconducting Qubits. Phys. Rev. Lett. 109, 060501 (2012).
[4] Nachman, B., Urbanek, M., de Jong, W. A. & Bauer, C. W. Unfolding quantum computer readout noise. npj Quantum Inf. 6, 84 (2020).
[5] Carroll, M., Rosenblatt, S., Jurcevic, P., Lauer, I. & Kandala, A. Dynamics of superconducting qubit relaxation times. Preprint at https://arxiv.org/abs/2105.15201 (2021).
[6] Mitchell, B. K. et al. Hardware-efficient microwave-activated tunable coupling between superconducting qubits. Preprint at https://arxiv.org/abs/2105.05384 (2021).