npj | Quantum Information

www.nature.com/npjqi

ARTICLE OPEN

Entanglement across separate silicon dies in a modular superconducting qubit device

Alysson Gold[1]✉, J. P. Paquette[1], Anna Stockklauser[1], Matthew J. Reagor[1], M. Sohaib Alam[1], Andrew Bestwick[1], Nicolas Didier[1], Ani Nersisyan[1], Feyza Oruc[1], Armin Razavi[1], Ben Scharmann[1], Eyob A. Sete[1], Biswajit Sur[1], Davide Venturelli[2,3], Cody James Winkleblack[1], Filip Wudarski[2,3], Mike Harburn[1] and Chad Rigetti[1]

Assembling future large-scale quantum computers out of smaller, specialized modules promises to simplify a number of formidable science and engineering challenges. One of the primary challenges in developing a modular architecture is in engineering high fidelity, low-latency quantum interconnects between modules. Here we demonstrate a modular solid state architecture with deterministic inter-module coupling between four physically separate, interchangeable superconducting qubit integrated circuits, achieving two-qubit gate fidelities as high as 99.1 ± 0.5% and 98.3 ± 0.3% for iSWAP and CZ entangling gates, respectively. The quality of the inter-module entanglement is further confirmed by a demonstration of Bell-inequality violation for disjoint pairs of entangled qubits across the four separate silicon dies. Having proven out the fundamental building blocks, this work provides the technological foundations for a modular quantum processor: technology which will accelerate near-term experimental efforts and open up new paths to the fault-tolerant era for solid state qubit architectures.

npj Quantum Information (2021)7:142 ; https://doi.org/10.1038/s41534-021-00484-1

INTRODUCTION

Progress in quantum operations over multi-node networks could enable modular architectures spanning distances from the nanometer to the kilometer scale[1–5]. Heralded entanglement protocols, whereby entanglement is generated probabilistically, have now reached entanglement rates up to 200 Hz[6–12]. Superconducting systems have established direct exchange of quantum information over cryogenic microwave channels[13–18], which is particularly useful toward interconnects of intermediate range such as between dilution refrigerators. Yet, in the context of superconducting qubit based processors, none of these methods are likely to outperform local gates between qubits, which can achieve coupling rates in the tens of MHz and fidelities reaching 99.9%[19–23]. Importantly, modules consisting of closely spaced and directly coupled separate physical dies retain many of the benefits of distributed modular architectures without the challenge of remote entanglement. Increased isolation between modules reduces cross-talk and correlated errors, for example due to high energy background radiation[24–26], and by fabricating smaller units and selecting the highest yielding units for device assembly, higher device yield is achievable[27,28]. Mastering 3D integration and modular solid state architectures has thus been a long-standing objective[29–31].

We demonstrate herein a modular superconducting qubit device with direct coupling between physical modules. The device, which consists of four eight-qubit integrated circuits (QuICs) fabricated on individual dies and flip-chip bonded to a larger carrier chip, achieves coupling rates and entanglement quality approaching the state-of-the-art in intra-chip coupling.

RESULTS

Design of a modular superconducting qubit device

The multi-die device assembly is constructed through flip-chip bonding of four nominally identical dies to a larger carrier die as shown in Fig. 1a. The carrier chip assumes a similar role to the chip multiprocessor in a classical multi-core processor while also providing microwave shielding, circuitry to interface between the individual QuICs and signal routing for the device I/O. The smaller individual dies comprise the QuICs, each consisting of four flux tunable and four fixed transmon qubits[32], and corresponding readout resonators and flux control lines as shown in Fig. 1b. The readout is multiplexed with four qubits and resonators per readout line and qubits are driven through the readout on this test platform. Qubits are labelled with a letter for the die position from left to right and a number for the qubit position within the die, e.g., B6. Entangled pairs are labeled according to the adjacent qubits, e.g., B6-C1. The QuIC dies are designed to be identical in order to maximize fabrication yield and enable modular assembly. The benefits to fabrication yield are evident in considering the number of distinct permutations that exist for a single device assembly: for a wafer with 220 QuIC dies, there are over 2.2 billion possible unique device assemblies.

The device Hamiltonian is designed to enable two-qubit parametric gates[32–36] between pairs of qubits on separate dies (see Methods). Coupling between qubits on separate chips is mediated through capacitive couplers on the QuIC and the carrier side, resulting in a cross-chip, charge-charge interaction. The carrier chip contains couplers with paddles at each end which are positioned below similar paddle-shaped couplers extending from the qubits on the QuIC as shown in Fig. 1a. This is similar in concept to the vacuum gap capacitors used in superconducting lumped element LC resonators[37,38] and in coupling qubits to resonators[39]. There is no intentional coupling between qubits on the same die in this test platform so as to isolate the basic inter-chip coupling mechanism and avoid complexities arising from larger circuits such as frequency collisions and leakage. However, the qubit and coupler design can be adapted to a larger lattice with intra-chip connectivity.

[1]Rigetti Computing, 775 Heinz Ave, Berkeley, CA 94701, USA. [2]Quantum Artificial Intelligence Laboratory (QuAIL), NASA Ames Research Center, Moffett Field, CA 94035, USA. [3]USRA Research Institute for Advanced Computer Science (RIACS), Mountain View, CA 94043, USA. ✉email: agold@rigetti.com

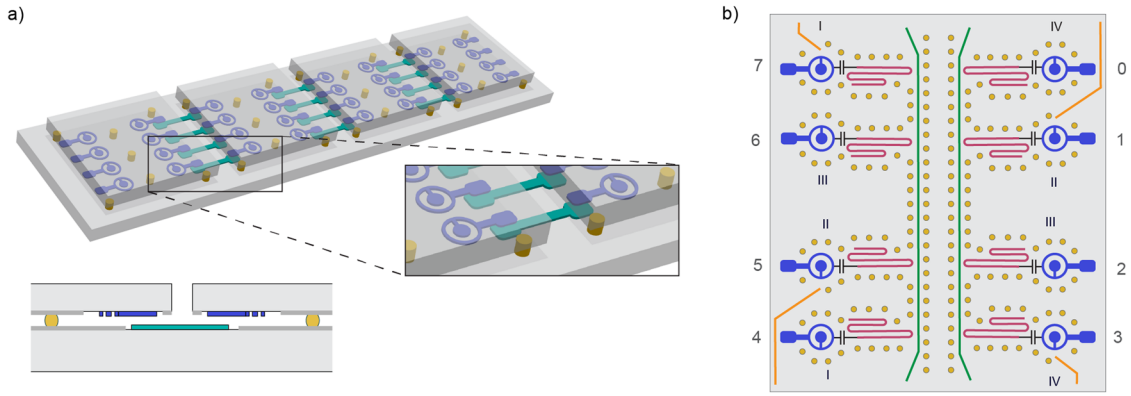

Fig. 1 Schematics (not to scale) illustrating the multi-core architecture. a Isometric view of the device assembly. The qubits (blue circular structures) are fabricated on the QuIC die and have one arm with a paddle-shaped coupler extending to the edge of the chip. The chips are flip-chip bonded onto the carrier chip using indium bump bonds (yellow) and the qubit couplers are aligned above couplers on the carrier chip (teal) as shown in the inset as well as the cross-sectional view. b False-colored schematic of a single QuIC including readout resonators and readout lines (magenta and green), indium bumps (yellow), flux bias lines (orange) and the qubits and paddles of the inter-chip couplers (blue). The physical qubits are labelled 0–7 while Roman numerals correspond to the design specification for the qubit (see Methods). Note that component geometries are not drawn exactly as fabricated: the schematic is intended primarily to depict the circuit layout as opposed to details into individual feature geometries.

Device fabrication, assembly, and validation

The QuIC chips are fabricated using standard lithographic techniques on a Si wafer, which is then diced to create individual dies. The Josephson junctions which form the SQUID loops of the transmon qubits are fabricated through double-angle evaporation of Al. Superconducting circuit components, including the Al pads for the Josephson junctions and Nb ground planes, signal routing and coplanar waveguides and resonators, are fabricated by pattern, deposition and lift-off steps[40].

Flip-chip bonding of the carrier and QuIC modules is accomplished through the deposition and patterning of indium bumps of height 6.5 μm and 40 μm diameter onto the carrier chip. The QuIC chips are flipped and aligned to the carrier before thermo-compression bonding, creating a superconducting bond between the carrier and QuIC chips. The fabrication process is described in detail in the Methods section. The indium bump heights post-bonding determine the vertical separation between the QuIC chips and carrier, as shown in Fig. 1a.

Importantly, the capacitance between the carrier and QuIC paddles is inversely proportional to the height of the indium bump bonds, h, as expected for a parallel plate capacitor. The bare coupling rate between qubits, g is directly proportional to this capacitance and thus follows the same dependence on h. Due to bonding process variation, the indium bump height spans a range of 1.5–4 μm. As shown in Fig. 2, this corresponds to a range for g of 8.8–26.1 MHz for the coupling rate.

Despite the range of anticipated couplings, the simulated fidelity for parametric gates in this design is relatively unaffected. Figure 2 shows simulation results of both the parametric CZ unitary gate error in the absence of loss and dephasing channels (coherent error) and the coherence-limited gate error (incoherent error) as a function of the bump height. The incoherent error is obtained assuming an ideal coherent exchange between the qubits while the coherent error takes into account the unwanted interactions arising from the capacitive coupling and flux modulation. For the coherence-limited fidelity calculation, we use a relaxation time, T_1, of 73/18 μs and a dephasing time, T_2, of 43/15 μs for fixed/tunable qubits. Over the full range of indium bump height expected from the bonding process, the predicted maximum achievable fidelity (taken as the minimum of the coherence-limited and unitary fidelity) varies from just under 99.0–99.5%. For an initial proof of concept, this range is acceptable; however, to push toward fidelities exceeding 99% or to employ this as part of a tunable coupler scheme[41,42], efforts will be needed to reduce the spread. Additional calibration of the force applied during the bonding process and design revisions to reduce the sensitivity of the coupling to bump height by changing the paddle geometry could reduce this variation further for gate schemes requiring a tighter tolerance. While attaining the required accuracy could be a challenge, we note that achieving higher coupling rates (higher than our targets here, which are limited by ZZ coupling) is in fact easier than with standard 2D (lateral) capacitive coupling, requiring simply increasing the paddle widths or reducing the target bump height.

The device assembly, designed and fabricated as described above, is measured in a dilution refrigerator at 10 mK. To assess the accuracy of the simulations and modeling conducted during device design, we characterize the device Hamiltonian (qubit and resonator frequencies, and coupling rates between elements) and compare with predictions from simulation. Qubit frequencies are within 2.1% of predicted values for the f_{01} transition and 11% for the qubit anharmonicity at zero applied flux bias (see Methods), demonstrating good agreement and indicating the inter-chip coupling technology does not impact the steady-state device physics in an unexpected manner.

Experimental determination of the bare coupling rates

The capacitive coupling formed by the inter-chip couplers results in a charge-charge interaction between the coupled qubits, q_1 and q_2. In the dispersive regime, where the detuning between qubits is large compared to the bare coupling rate (given here in radial units, $\tilde{g} = 2\pi g$), $|\omega_{01,1} - \omega_{01,2}| \gg \tilde{g}$, we can calculate the bare coupling rate from the measured dispersive shift, χ_{qq}. The relationship between χ_{qq} and $\tilde{g}$ is given by Eq. (1) for the general case of two flux-tunable transmons, which differs from the treatment of transmon-resonator dispersive shifts. Note here we are working in the transmon limit and the equation below is the result of a perturbative expansion of the Hamiltonian eigenstates and eigenenergies as a function of applied magnetic flux, following the treatment in ref. [35]. $E_{J,\mathrm{eff}}(\Phi)$ is the effective Josephson energy of the DC-SQUID, a function of the applied magnetic flux through the SQUID, Φ, and is defined, along

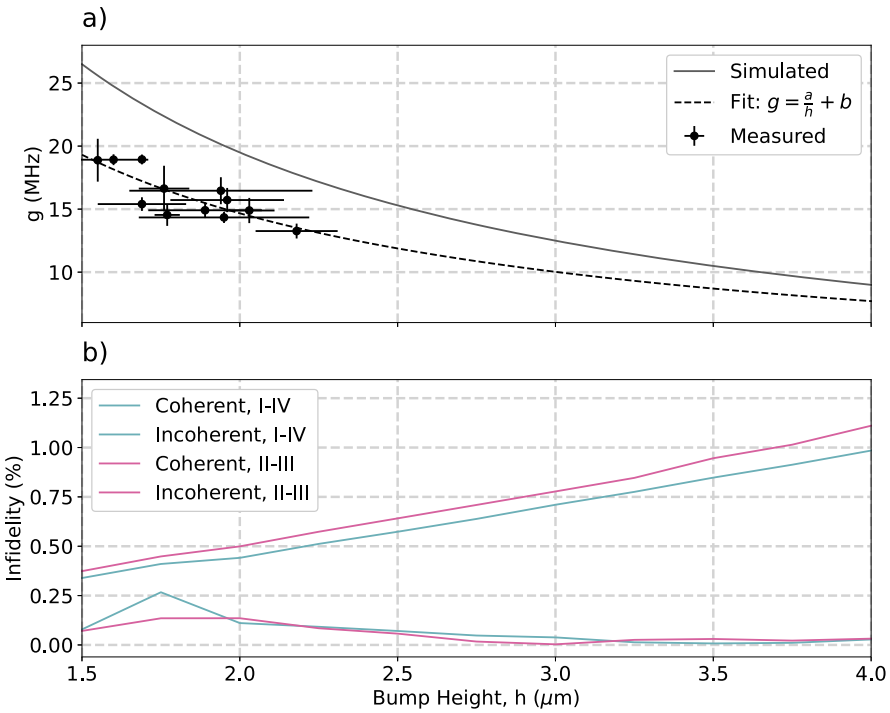

Fig. 2 Impact of post-bonding indium bump height on coupling rate and coherent and incoherent simulated gate errors. The fit parameters in (**a**) are $a = 27.9 \pm 6.4$ MHz μm and $b = 0.7 \pm 3.5$ MHz. The discrepancy between measurement and simulation and details on bump height measurements are discussed in the Methods section. There are only two distinct entangled qubit pairs in the designed Hamiltonian, the II–III and I–IV pairs as shown in Fig. 1b, both of which are simulated in (**b**). Error bars represent the measurement uncertainty on the bump height (*x*-axis) and coupling rate (*y*-axis).

with $\lambda(\Phi)$ and $\Lambda(\Phi)$ in the same reference.

$$\chi_{qq}(\Phi_1, \Phi_2) = 2\tilde{g}^2 \left[\frac{\mu^2_{01,1}(\Phi_1)\mu^2_{12,2}(\Phi_2)}{\omega_{01,1}(\Phi_1) - \omega_{12,2}(\Phi_2)} - \frac{\mu^2_{12,1}(\Phi_1)\mu^2_{01,2}(\Phi_2)}{\omega_{12,1}(\Phi_1) - \omega_{01,2}(\Phi_2)} \right], \tag{1}$$

$$\mu_{01}(\Phi) = \left[\frac{E_{J,\mathrm{eff}}(\Phi)}{E_{J,\mathrm{eff}}(0)} \right]^{1/4} \frac{\lambda(\Phi)}{\lambda(0)}, \tag{2}$$

$$\mu_{12}(\Phi) = \left[\frac{E_{J,\mathrm{eff}}(\Phi)}{E_{J,\mathrm{eff}}(0)} \right]^{1/4} \frac{\Lambda(\Phi)}{\lambda(0)}. \tag{3}$$

We measure the dispersive shift through time Ramsey measurements taken with the tunable qubit biased at its maximum frequency. In our modified time Ramsey measurement, an $X/2$ pulse is applied to a qubit, q_1, causing the qubit to precess about the equator. After a time delay, Δt, a Z pulse is applied which rotates the qubit through a phase $\phi = 2\pi\Delta t\delta f$, where δf is the detuning of the pulse frequency relative to the qubit frequency, f_{01} (the application of the Z pulse is a slight modification on the standard Ramsey sequence of an $X/2$ pulse followed by a delay and another $X/2$ pulse which enables better signal visibility)). Finally, another $X/2$ pulse is applied and the qubit state is measured. The resulting excited state visibility oscillates as a function of the time delay, reaching full visibility when the Z rotation perfectly offsets the phase accumulated from the precession during the delay time. From the period of the oscillations, the difference between the qubit frequency and the applied pulse frequency (already detuned from the expected qubit frequency by δf), and by consequence the qubit frequency itself as the applied pulse frequency is well defined by the control electronics, can be determined with high precision.

To measure χ_{qq}, the time Ramsey measurement is performed on a qubit q_1 with adjacent qubit q_2 in the ground state and again in the excited state. The difference between the f_{01} measured for q_1 at both points gives a precise measurement of χ_{qq}. The measurement is then performed in the opposite direction, with the state of q_1 varied while the frequency of q_2 is measured. The bare coupling rate calculated from the dispersive shift, as given by Eq. (1), should be equivalent in both directions for a pure ZZ interaction, to within the error of the measurement. While the design target for all couplers was 12 MHz for a 3 μm indium bump height post-bonding, the observed coupling rate varied across the chip from 13.26 ± 0.59 MHz to 18.94 ± 0.39 MHz (see Fig. 2). This was within the anticipated range due to indium bump height variation (see Methods for further details, in particular Table 1).

Cross-chip entangling gate performance

We calibrated and benchmarked gates on ten out of twelve inter-chip pairs. The remaining two pairs could achieve population transfer but due to frequency targeting error in the fabrication process, the gate modulation frequencies were outside of the frequency band of the control electronics and degraded AC flux control resulted in low fidelity. The primary benchmarking methods employed were two-qubit randomized benchmarking (RB) and interleaved randomized benchmarking (iRB)[43,44]. We quote the estimate from iRB when the RB protocol estimates an average gate fidelity of $\mathcal{F}_{\mathrm{RB}} \geq 92\%$, which bounds the iRB estimate to at most 20% of the reported gate error due to imperfect gate randomization[43]. Below this empirical threshold, the assumptions of the error model can lead to large uncertainty and an overestimate of the gate fidelity for iRB.

Table 2 provides a summary of the CZ gate fidelities measured for each of the ten pairs and compares them with the coherence limited fidelity (the maximum achievable fidelity predicted from the measured relaxation and dephasing times of the qubit pair). The coherence limited fidelity is computed from the T_1 and T_2 under modulation, $\tilde{T}_1$ and $\tilde{T}_2$, ie. the coherence times as measured while an AC flux bias is applied to the tunable qubit at the gate

modulation frequency, emulating conditions during gate operation. $\widetilde{T}_1$ and $\widetilde{T}_2$ for the tunable qubit in each pair, the limiting qubit in regards to coherence, are also recorded in the table.

Comparing the measured fidelities with the coherence limited fidelities, the fidelity is almost always limited by the qubit coherence suggesting the inter-chip coupling mechanism does not limit gate error directly. Furthermore, we have compared qubit coherence times for inter-chip-coupled qubits to a baseline of similar qubits that are not coupled by inter-chip couplers and coherence times do not appear to be limited by the inter-chip coupling technology itself (see Methods). This suggests that gate fidelities are limited by the same sources of error arising in monolithic devices and are not directly, or indirectly through impacts to qubit coherence, negatively impacted by the inter-chip coupling technology.

Apart from four gates with two level systems which could explain the increased dephasing under modulation, CZ gate fidelities were above 90% with 5 out of 12 gates demonstrating >95% measured fidelities. While Table 2 lists only CZ fidelities as CZ gates were within the AC flux control band for all pairs and they allowed a straightforward comparison to the coherence limited fidelity, the maximum gate fidelity measured was a 99.1% ± 0.5% iRB for an iSWAP gate on the C1-D6 pair, which we expand upon in Fig. 3.

Multi-die bell inequality violation

We now turn our attention to assessing the viability of a future modular quantum processor based on these techniques. Importantly for this analysis, the inter-chip connections on our test device are established by unique qubits, and, in addition, qubits fabricated on the same chip are not coupled. We thus investigate the simultaneous quality of two-qubit connections, for the three disjoint pairs. This step is important for assessing functional challenges toward leveraging non-local quantum states in larger scale algorithms. Following a tradition established for multi-node or modular experimental efforts in superconducting qubits[28,45–47], we design a test for the deterministic violation of a Bell inequality with this inter-chip platform. We describe a figure of merit $\langle \mathcal{W}_\Sigma \rangle = \sum_k \langle W_k \rangle$, where W_k is a witness to entanglement of connection k, applying the standard Bell observable for two-qubits,

$$\mathcal{W} = QS + RS + RT - QT, \tag{4}$$

with $Q = Z_n$, $R = X_n$, $S = \frac{X_m - Z_m}{\sqrt{2}}$, and $T = \frac{X_m + Z_m}{\sqrt{2}}$, taking qubits {$n$, m} across an inter-chip connection. For N disjoint Bell pairs, the total figure of merit is bounded by $\langle \mathcal{W}_\Sigma \rangle \leq 2N\sqrt{2}$ and signal above $\langle \mathcal{W}_\Sigma \rangle > 2N$ certifies that the network supports genuine entanglement over at least one connection simultaneously. Moreover, investigating the individual Bell signals $\langle W_k \rangle$ can test entanglement over each connection independently.

Our experimental procedure is shown in Fig. 4. We choose three connections that bridge all four chips in a disjoint pattern (A0-B7, B0-C7, C1-D6). We prepare the three pairs in an equal superposition of two-qubits: $|\Psi\rangle_k = (|00\rangle_k + |11\rangle_k)/\sqrt{2}$. Then, we measure the qubits in the $\langle ZZ \rangle$ basis or $\langle XX \rangle$ basis. A total of 100 experiments were run, having 10^4 shots per basis, collecting measurement data simultaneously for all pairs. A summary of results is in Table 3, where all three connections violate the Bell test by at least three standard deviations. With high confidence, therefore, the test platform supports simultaneous disjoint, pair-wise entanglement involving all four chips. In addition, our total figure of merit, $\langle \mathcal{W}_\Sigma \rangle = 6.651 \pm 0.067$ exceeds the classical bound by nearly ten standard deviations.

Table 1. Bare coupling rate, g, as measured from the qubit-qubit dispersive shift, χ_{qq} compared to simulated values given measured bump heights, h.

Pair	h (μm)	Meas. g (MHz)	Sim. g (MHz)
A0-B7	2.18 ± 0.13	13.26 ± 0.59	16.86 ± 1.05
A1-B6	2.03 ± 0.08	14.89 ± 1.00	18.44 ± 0.79
A2-B5	1.96 ± 0.18	15.72 ± 0.96	19.32 ± 1.96
A3-B4	1.94 ± 0.29	16.46 ± 1.07	19.60 ± 3.15
B0-C7	1.95 ± 0.27	14.34 ± 0.44	19.42 ± 2.89
B1-C6	1.89 ± 0.18	14.91 ± 0.62	20.11 ± 2.12
B2-C5	1.69 ± 0.14	15.40 ± 0.54	22.62 ± 2.07
B3-C4	1.76 ± 0.08	16.63 ± 1.81	21.29 ± 1.04
C0-D7	1.77 ± 0.04	14.54 ± 0.87	21.11 ± 0.51
C1-D6	1.69 ± 0.01	18.94 ± 0.39	22.29 ± 0.07
C2-D5	1.55 ± 0.01	18.88 ± 1.70	24.52 ± 0.16
C3-D4	1.60 ± 0.11	18.92 ± 0.43	23.70 ± 1.66

DISCUSSION

Concluding, we demonstrate that the flip-chip bonded, multi-die fabrication process with inter-chip coupler technology is capable of achieving high fidelity entanglement, including gate fidelities regularly exceeding 95% and up to 99% in the best case, and simultaneous entanglement between silicon dies violating the Bell test by over three standard deviations.

Future work should explore the potential benefits of this modular approach beyond the intrinsic advantages in regards to flexible device construction and yield. This includes increased isolation between qubits on separate physical die, an important factor particularly in developing robust hardware

Table 2. Measured fidelity for CZ gates compared with coherence limited fidelity.

Pair	$\widetilde{T}_1$ (μs)	$\widetilde{T}_2$ (μs)	Coherence limited fidelity (%)	Measured fidelity (%)	Gate time (ns)
A0-B7	24.57 ± 1.60	8.98 ± 0.92	98.58 ± 0.11	98.34 ± 0.31	152
A1-B6	9.35 ± 0.79	1.59 ± 0.12	92.07 ± 0.55	90.09 ± 0.51	148
A3-B4	11.56 ± 1.76	1.07 ± 0.11	88.84 ± 1.07	82.70 ± 0.78	164
B0-C7	26.67 ± 4.61	2.74 ± 0.31	96.74 ± 0.33	96.04 ± 0.72	128
B2-C5	4.59 ± 0.78	1.66 ± 0.21	88.17 ± 1.30	84.63 ± 0.92	176
B3-C4	7.16 ± 0.92	1.36 ± 0.11	97.40 ± 0.15	97.47 ± 0.94	116
C0-D7	1.52 ± 0.42	2.75 ± 0.37	90.81 ± 0.98	87.08 ± 0.59	284
C1-D6	14.51 ± 0.67	2.52 ± 0.24	96.92 ± 0.24	97.26 ± 0.29	108
C2-D5	7.49 ± 0.67	1.93 ± 0.14	77.25 ± 1.42	80.68 ± 0.98	468
C3-D4	30.72 ± 2.50	5.09 ± 0.65	98.20 ± 0.19	96.78 ± 1.73	116

The tunable qubit coherence times for the fixed - tunable pair with the tunable qubit under modulation, $\widetilde{T}_1$ and $\widetilde{T}_2$, are also listed.

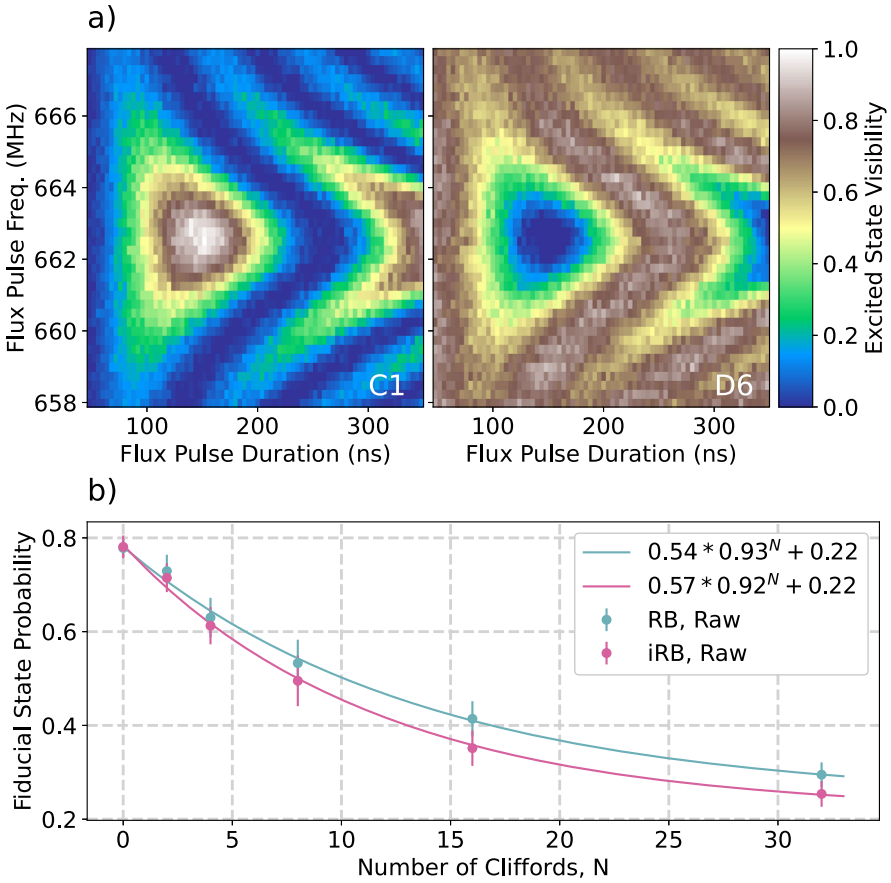

Fig. 3 Two-qubit gate data for an iSWAP gate on the C1-D6 pair. a Excited state visibility as a function of AC flux modulation frequency and pulse duration. Full population transfer occurs at agate modulation frequency of 662.5 MHz and 152 ns. **b** RB and iRB benchmarking data. Fitting the exponential decay for thestandard and interleaved benchmark gives an RB fidelity of 95.8% ± 0.2% and an iRB fidelity of 99.1% ± 0.5%. Error bars represent the 95% confidence bounds on the fitted fidelity for each measurement.

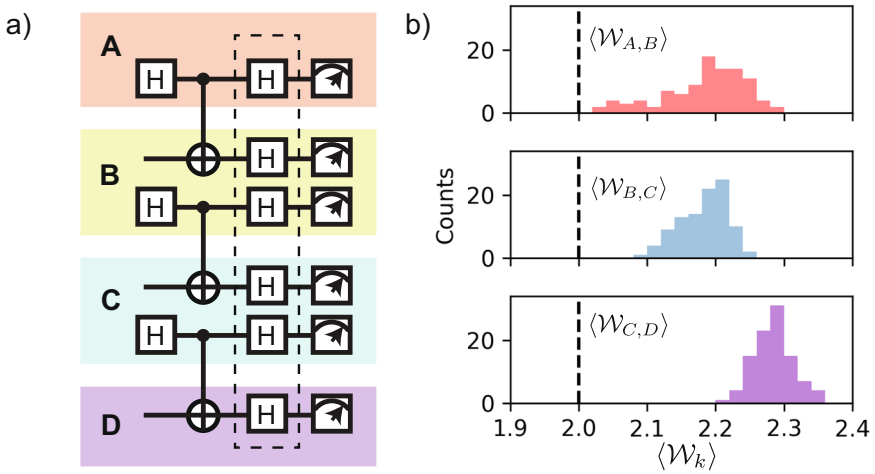

Fig. 4 Simultaneous Bell Inequality Violations. Disjoint, pair-wise entanglement is generated across chip boundaries via CNOT operations on inter-chip couplers compiled to CZ gates, with the optional basis change for evaluating either $\langle ZZ \rangle$ (no final Hadamard) or $\langle XX \rangle$ (with final Hadamard). No error mitigation or readout correction schemes were applied. **b** Histograms of the average values $\langle W_k \rangle$ across 100 individual experimental runs using 10^4 shots in each measurement basis. The dashed line indicates the classical limit of 2.0. Because all three outcomes fall in the quantum limit, we conclude that the test platform supports simultaneous disjoint, pair-wise entanglement involving all four chips.

suitable for near-term error correction schemes. Recently, the impacts of cosmic and background radiation on solid state devices have been of significant interest due to the correlated errors that result and the challenge these pose to fault tolerant quantum computing[24–26]. In this case and more generally, the physics of quasi-particle trapping and phonon propagation through superconducting qubit chips would be interesting to explore on multi-die devices. Phonons should collect on the boundaries of the individual die and not propagate to qubits on other dies, reducing correlated errors.

In addition, while the use of indium bump bonds means we cannot replace individual QuICs post-bonding, this modular architecture would have even higher yield benefits if QuIC selection could be done after measuring the QuICs cold, instead of after room-temperature metrology. This type of flexibility, which could come from either having a test set-up for individual QuICs or a configurable interconnect between the carrier die and QuICs, is also an interesting avenue for future study.

Finally, we note that the true impact of this technology will be in its integration with state-of-the-art processing architectures. With additional intra-chip circuitry and changes to the qubit topology on the individual QuICs, this technology can be extended to form a seamless modular quantum processor that is flexible in regards to the number and type of modules integrated and, with sensitivity to fabrication yield and intra-die cross-talk limited only by the module size, highly scalable. By enabling the fabrication of devices consisting of hundreds to thousands of qubits which are sufficiently isolated to mitigate correlated errors, this technology provides a clear path forward toward fault tolerant computing.

METHODS

Fabrication and bonding process

The carrier chip is composed of cavities etched in Si, coated with patterned superconducting metal, and indium bumps (deposited outside of the cavities on the flat outermost surface) which after bonding form the superconducting connection between carrier and QuICs. Carrier chips are fabricated from high resistance Si wafers. The fabrication flow starts with a photolithography process to create cavity patterns on wafers followed by a Bosch etch (DRIE) step to fabricate 24 μm deep pockets with vertical sidewalls. The surface is then conformally coated with a 560 nm-thick Nb/MoRe stack, deposited by sputtering (PVD), to form a continuous superconducting shield. Vertical cavity sidewalls are confirmed to have a continuous metal film connecting the top surface to the cavity bottom surface. A thin layer of MoRe alloy is deposited on top of Nb film to seal the Nb surface, enabling an oxide-free metal-to-metal interface for reliable electrical connection between the Nb device layer and the In bumps in the bonding areas.

The metal film stack is then patterned by a two-plane photolithography process followed by a reactive ion etching (RIE) step with a certain etching selectivity to the Si substrate. During the first exposure, focus and dose settings are selected to target the top wafer surface patterning, while in the second exposure settings are changed to target the cavity bottom surface only, which is 24 μm deeper. Once the Nb/MoRe stack etching is completed, a negative-tone photoresist lithography is used to transfer the In bump patterns onto the top metal surface by lift-off processing[48]. Electron-beam evaporation is used to deposit a 6.5 μm thick indium layer, producing a high quality, easy to lift-off film. An automatic lift-off tool that uses a combination of chemical cleaning and physical energy produced by high-pressure jets removes the In film from non-patterned areas, completing the process. No Josephson junction fabrication steps are required as no active components are located on the carrier chip. Future designs could include transmons in the carrier chip as part of, for example, a quantum bus to provide longer-range bus coupling between qubits instead of the direct coupling employed in this initial design.

To establish a superconducting connection between the carrier chip and the four separate QuIC chips, a flip-chip bonder is used. Each QuIC device is precisely aligned and thermo-compression bonded to the carrier chip. Prior to bonding, both carrier and QuIC chip surfaces are solvent cleaned followed by an atmospheric downstream plasma cleaning to chemically clear surfaces of native oxides. This process also temporarily passivates the In bumps from oxidation and helps to generate a strong chemical bond between In bumps and the corresponding pads. The multi-chip bonding process consists of sequential bonding of four separate QuICs to the designated locations on the carrier chip. For each bonding, the carrier and QuIC chips are aligned to each other with a horizontal accuracy of better than ± 2.5 μm. After the horizontal alignment is completed, a vertical parallelism adjustment is done using auto-collimation and laser-levelling methods with an accuracy of ± 0.5 μm. The ensuing thermo-compression process consists of three different phases. In the first phase, force and temperature values are gradually increased and stabilized. In the second phase, the actual thermo-compression bonding takes place for two minutes. To prevent thermal aging of QuICs that are already bonded, the carrier chip temperature is maintained at 30 °C, while the QuICs are heated to 70 °C only during bonding. In the final phase, the stack is cooled to 30 °C with a nitrogen flow. The same process is repeated sequentially for all the QuICs.

Device hamiltonian and parametric gates

To design the device Hamiltonian, the circuit parameters were extracted using quasi-static electromagnetic simulations and a positive second order representation[49] was used to solve the linearized circuit. The nonlinear effects of the Josephson junctions are subsequently accounted for through a perturbative treatment. The designed Hamiltonian parameters for this device are provided in Table 4, including the maximum and minimum f_{01} transition frequencies over the flux bias tuning range, the anharmonicity at the maximum of the tuning range, $\eta = f_{12,\max} - f_{01,\max}$, the frequency of the readout resonator coupled to the qubit, f_{RO}, and the qubit-readout resonator dispersive shift, $\chi_{q,RO}$. The coupling rate between qubit pairs was designed to be 12 MHz at a 3 μm indium bump height for all couplers.

Due to fabrication process variation, the Josephson junction width, and hence the Josephson energy, of each fabricated junction will differ slightly from the design target. Using the relationship between the room temperature conductance of a junction and its Josephson energy at cryogenic temperatures[50], a more accurate prediction for the device Hamiltonian can be obtained for fabricated devices using the same modelling process as during the initial device design, but replacing the target EJ values with the predicted EJ values from room temperature conductance measurements. Changes to the Josephson energy of the single junction in a fixed transmon or the two junctions in a DC-SQUID tunable transmon primarily impact the qubit frequencies, with little impact on qubit anharmonicities, readout resonator properties, or coupling rates.

We plot in Fig. 5 the design target, predicted and measured $f_{01,\max}$, demonstrating agreement between the predicted frequencies and those measured cold, to within ± 108 MHz or 2.2% in the worst case. The discrepancies are within the prediction error we expect due to uncertainty in the empirically determined linear coefficient relating room temperature conductance to inductance at cryogenic temperatures, and uncertainty in

Table 3. Results for joint Bell test marginals and total observable.

Observable	Qubits	Result
$\langle \mathcal{W}_{A,B} \rangle$	A0-B7	2.184 ± 0.060
$\langle \mathcal{W}_{B,C} \rangle$	B0-C7	2.183 ± 0.034
$\langle \mathcal{W}_{C,D} \rangle$	C1-D6	2.284 ± 0.029
$\langle \mathcal{W}_{\Sigma} \rangle$	Combined	6.651 ± 0.067

Table 4. Hamiltonian properties as designed.

Qubit	Flux Tunable?	$f_{01,\max}$ (MHz)	$f_{01,\min}$ (MHz)	η (MHz)	f_{RO} (MHz)	$(2\pi)^{-1}\chi_{q,RO}$ (MHz)
I	Fixed	3654	3654	− 190	7232	0.80
II	Tunable	5066	4266	− 200	7476	0.81
III	Fixed	3714	3714	− 190	7273	0.85
IV	Tunable	4946	4146	− 200	7425	0.81

The qubit numbering corresponds to that shown in Fig. 1b, where there are only four unique design targets which are repeated in inverted order on the opposite side of the chip.

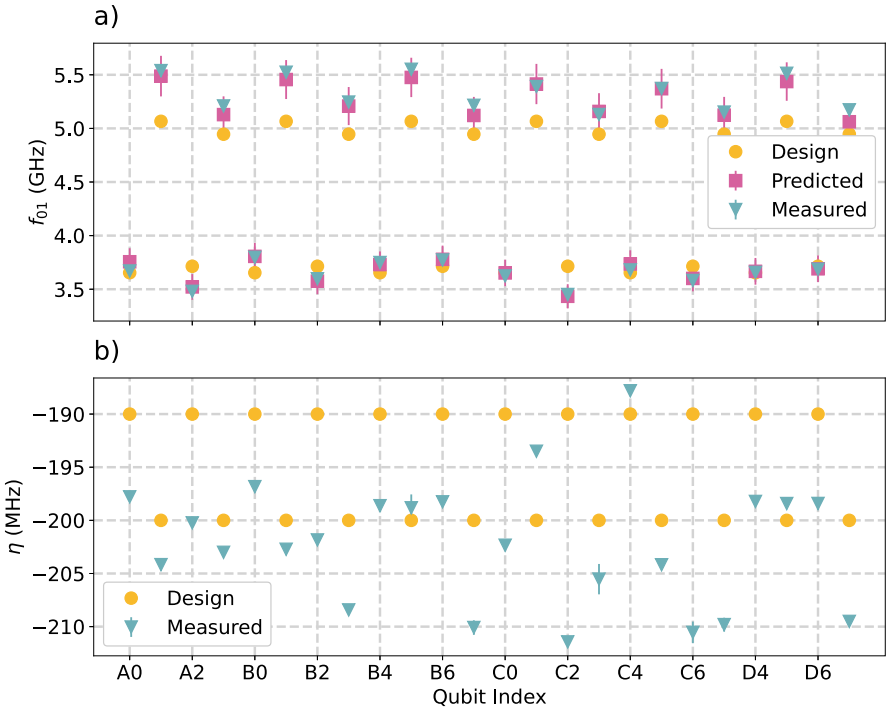

Fig. 5 Comparison of predicted and measured transition frequency, f_{01} and anharmonicity, η at zero applied flux bias. Note that while the design target is constant from one die to the next (all dies are designed to be identical), small variations in the predicted and measured qubit frequencies are present due to fabrication process variation, which can be measured at room temperature. Good agreement between the predicted and measured qubit frequencies, to within ± 108 MHz, is obtained. Anharmonicities show a systematic offset from the design targets but are still within 11% of the target value. Error bars on the predicted values represent the standard deviation in the values obtained from a Monte–Carlo simulation where the expected variation in the simulation inputs is accounted for (room temperature metrology error, error in the qubit charging energy, etc).

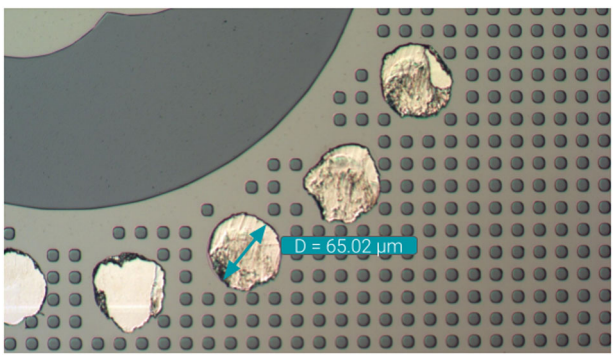

Fig. 6 Magnified image of a region of the carrier chip. This image shows the indium bumps post-bonding and post-shearing, a destructive process whereby a cut is made through the device separating the QuIC chips from the carrier chip. The diameter and height of the indium bumps are known pre-bonding, such that by measuring the bump diameter post-bonding, a rough estimate for the bump height can be obtained assuming the bump is cylindrical in both cases: $h_{\mathrm{post}} = h_{\mathrm{pre}}\left(D_{\mathrm{pre}}/D_{\mathrm{post}}\right)^2$.

the conductance measurement itself. The qubit anharmonicities are compared with the design target and are accurate to within 11%, demonstrating a systematic offset that will be corrected in future designs.

The device Hamiltonian is designed to enable parametric gates between one tunable (T) qubit coupled to one fixed qubit (F). In this scheme, an AC flux bias at RF frequency f_{p} is applied to the tunable qubit around its parking flux bias. Under flux modulation, the transmon frequency oscillates at harmonics of the modulation frequency around its time-averaged frequency $\bar{f}_{T,01}$. Transmon frequency modulation gives rise to sidebands at frequencies $\bar{f}_{T,01} + kf_{\mathrm{p}}$, separated by the modulation frequency around the average frequency. When the modulation frequency is tuned such as to align one sideband with the transition frequency of the fixed qubit, a coherent exchange takes place between the two qubits at a rate equal to the bare coupling strength renormalized by the sideband weight. When the tunable qubit is parked at the maximum of the tuning band, only even sidebands have a non-zero weight and the sideband $k = \pm 2$ is used. Entangling gates are then enacted by modulating the tunable qubit at half the average detuning between the qubits' transition frequencies. To obtain the iSWAP gate, the interaction between states $|01\rangle$ and $|10\rangle$ is activated at the modulation frequency $f_{\mathrm{p}} = |\bar{f}_{T,01} - f_{F,01}|/2 \equiv \Delta/2$ (with the convention $|FT\rangle$). For the CZ gate, the interaction between $|11\rangle$ and $|02\rangle$ is activated at $f_{\mathrm{p}} = (\Delta + \eta_T)/2$ (CZ_{02}) or between $|11\rangle$ and $|20\rangle$ at $f_{\mathrm{p}} = (\Delta - \eta_F)/2$ (CZ_{20}). The gate time is adjusted to provide a π rotation for iSWAP and 2π rotation for CZ between the corresponding two-qubit states.

Analysis of bump heights and coupling rates

After cryogenic tests were complete, the indium bump heights were measured at various locations across the device. This was done in a destructive manner by shearing the chip to separate the QuIC and the carrier chips, then measuring the indium bump diameter, as shown in Fig. 6. Working under the assumption that before and after bonding the indium bumps are approximately cylindrical in shape, and with the diameter and height known before bonding, the diameter as measured after shearing provides an estimate for the bump height post-bonding. The measured coupling rates could then be compared with post-hoc simulated coupling rates computed for each coupler based on the measured bump heights. We obtain qualitative agreement between the simulated and measured coupling rates, however the measured coupling rates are ~20% lower than expected from design. Future studies will look to address this discrepancy. Some of the reasons for the latter are the overestimation of the bump height size and inaccuracies in material properties assumed in the simulation. In addition the measured coupling values are effectively reduced by a term representing the next-nearest neighbor couplings to resonators, which exist in a higher band than the transmons.

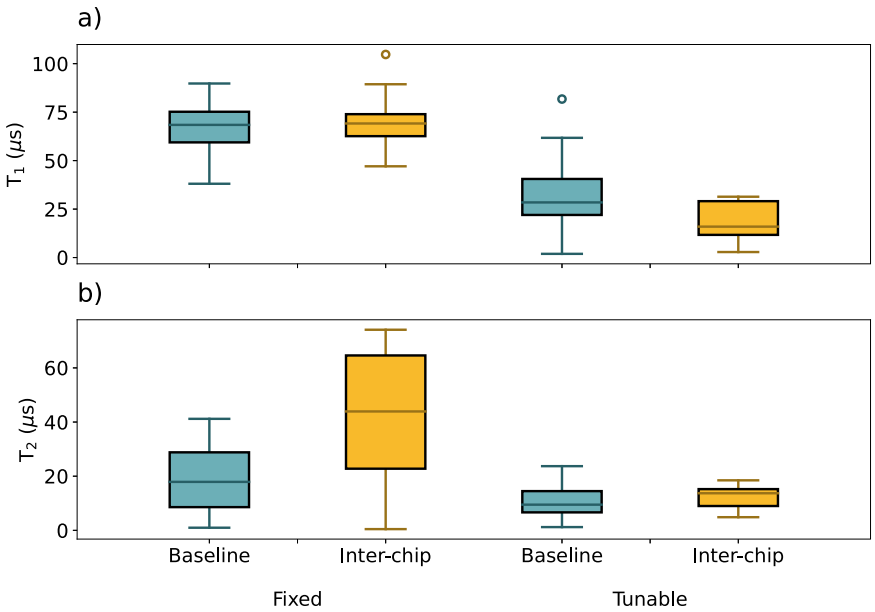

Fig. 7 Comparison of T_1 and T_2 for qubits with inter-chip coupling (inter-chip) compared to qubits on devices from similar wafers without inter-chipcouplers (baseline). The baseline data includes 4 fixed qubits and 27 tunable qubits while the inter-chip data includes 12 fixed and 12 tunable qubits, and for each qubit the measured values are averaged over several measurements taken over the course of a week. The boxes plotted depict the four inter-quartile ranges for the distribution of average values with outliers shown as hollow circles. Tunable qubits have lower T_1 and T_2 due to the coupling of the flux bias line to the qubit and are thus plotted separately here.

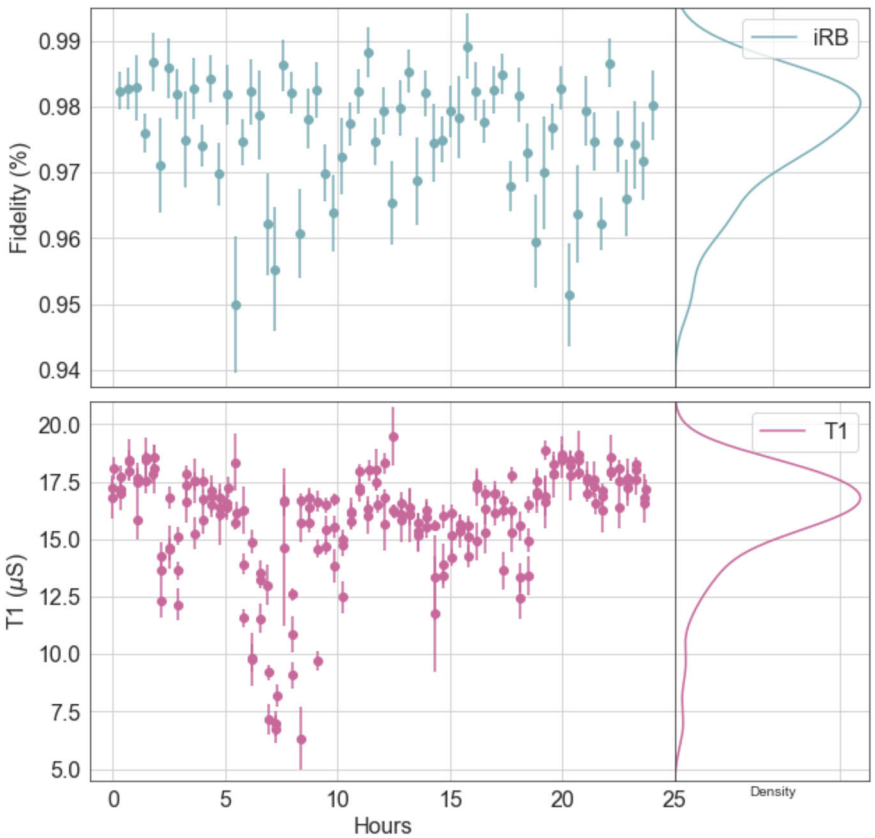

Fig. 8 Timeseries of the iRB fidelity and T_1 of the tunable qubit on the C1-D6 pair. The tunable qubit has a lower T_1 relative to the fixed qubit on the edge and is thus the limiting factor in regards to incoherent error. The repeated measurement sequence calibrates readout parameters, parks the tunable qubit at its maximum frequency, measures the coherence time and benchmarks the gate. Error bars represent the 95% confidence bounds on the fitted parameter (fidelity and T_1) for each measurement.

Impact of inter-chip couplers on qubit coherence and 2Q gate stability

An important question to address is whether inter-chip coupling exposes qubits to additional loss or dephasing channels relative to standard intra-chip lateral couplers. As the electric field between the paddles of the coupler passes through vacuum rather than a silicon substrate, no additional dielectric losses are expected. Furthermore, the galvanic connection across the carrier chip is small enough relative to the 3–8 GHz band of interest that it can be treated as a lumped element and is not expected to produce any additional resonant coupling between qubits and the electromagnetic environment (chip modes, package modes, etc.) up to frequencies in excess of 15 GHz.

It is thus not expected that these couplers should impact the qubit relaxation (T_1) or dephasing (T_2) times. This was reflected in experimental results, as shown in Fig. 7 where we compare T_1 and T_2 for a device with inter-chip coupling compared to devices from similar wafers with no coupling at all. No statistically significant difference in the T_1 and T_2 times was observed relative to the baseline.

In addition to comparing the coherence between inter-chip qubits and intra-chip qubits, we also assess the entangling gate stability and relaxation time, T_1 over a period of 24 h, as plotted for the C1-D6 pair in Fig. 8. For each data point plotted, the qubits are re-tuned, readout is calibrated, and the two-qubit gate is benchmarked using interleaved RB. Qubit re-tuning is done by parking the tunable qubit at its maximum frequency and calibrating the gate pulses for the fixed and tunable qubit. Readout calibration involves preparing ground and excited states to update the classifier that will be used to discriminate single shot measurements. T_1 decay is monitored through repeated coherence measurements taken after re-tuning and readout calibrations and before benchmarking. Temporal fluctuations show a drop in the gate fidelity when the T_1 falls below 10 μs, but generally it remains stable to within four percentage points with a distribution centered around 98%. Fluctuations in T_1 is an active research topic in the field of superconducting qubits[25,26].

DATA AVAILABILITY

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

Received: 11 March 2021; Accepted: 1 September 2021;
Published online: 28 September 2021

REFERENCES

1. Awschalom, D. et al. Development of Quantum Interconnects (QuICs) for Next-Generation Information Technologies. *PRX Quant* **2**, 017002 (2021).
2. Steffen, L. et al. Deterministic quantum teleportation with feed-forward in a solid state system. *Nature* **500**, 319–322 (2013).
3. Chou, K. S. et al. Deterministic teleportation of a quantum gate between two logical qubits. *Nature* **561**, 368–373 (2018).
4. Wan, Y. et al. Quantum gate teleportation between separated qubits in a trapped-ion processor. *Science* **364**, 875–878 (2019).
5. Hensen, B. et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).
6. Olmschenk, S. et al. Quantum teleportation between distant matter qubits. *Science* **323**, 486–489 (2009).
7. Moehring, D. L. et al. Entanglement of single-atom quantum bits at a distance. *Nature* **449**, 68–71 (2007).
8. Humphreys, P. C. et al. Deterministic delivery of remote entanglement on a quantum network. *Nature* **558**, 268–273 (2018).
9. Monroe, C. et al. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A* **89**, 022317 (2014).
10. Ritter, S. et al. An elementary quantum network of single atoms in optical cavities. *Nature* **484**, 195–200 (2012).
11. Zhong, C. et al. Proposal for Heralded Generation and Detection of Entangled Microwave-Optical-Photon Pairs. *Phys. Rev. Lett.* **124**, 010511 (2020).
12. Krastanov, S. et al. Optically Heralded Entanglement of Superconducting Systems in Quantum Networks. *Phys. Rev. Lett.* **127**, 040503 (2021).
13. Campagne-Ibarcq, P. et al. Deterministic Remote Entanglement of Superconducting Circuits through Microwave Two-Photon Transitions. *Phys. Rev. Lett.* **120**, 200501 (2018).
14. Magnard, P. et al. Microwave Quantum Link between Superconducting Circuits Housed in Spatially Separated Cryogenic Systems. *Phys. Rev. Lett.* **125**, 260502 (2020).
15. Axline, C. J. et al. On-demand quantum state transfer and entanglement between remote microwave cavity memories. *Nat. Phys.* **14**, 705–710 (2018).
16. Leung, N. et al. Deterministic bidirectional communication and remote entanglement generation between superconducting qubits. *NPJ Quant. Inf.* **5**, 1–5 (2019).
17. Zhong, Y. et al. Deterministic multi-qubit entanglement in a quantum network. *Nature* **590**, 571–575 (2021).
18. Kurpiers, P. et al. Deterministic quantum state transfer and remote entanglement using microwave photons. *Nature* **558**, 264–267 (2018).
19. Chen, M.-C. et al. Demonstration of Adiabatic Variational Quantum Computing with a Superconducting Quantum Coprocessor. *Phys. Rev. Lett.* **125**, 180501 (2020).
20. Foxen, B. et al. Demonstrating a Continuous Set of Two-Qubit Gates for Near-Term Quantum Algorithms. *Phys. Rev. Lett.* **125**, 120504 (2020).
21. Negîrneac, V. et al. High-Fidelity Controlled-*Z* Gate with Maximal Intermediate Leakage Operating at the Speed Limit in a Superconducting Quantum Processor. *Phys. Rev. Lett.* **126**, 220502 (2021).
22. Sung, Y. et al. Realization of High-Fidelity CZ and *ZZ* -Free iSWAP Gates with a Tunable Coupler. *Phys. Rev. X* **11**, 021058 (2021).
23. Stehlik, J. et al. Tunable Coupling Architecture for Fixed-frequency Transmons. *Phys. Rev. Lett.* **127**, 080505 (2021).
24. Wilen, C. et al. Correlated charge noise and relaxation errors in superconducting qubits. *Nature* **594**, 369–373 (2021).
25. Vepsäläinen, A. P. et al. Impact of ionizing radiation on superconducting qubit coherence. *Nature* **584**, 551–556 (2020).
26. Cardani, L. et al. Reducing the impact of radioactivity on quantum circuits in a deep-underground facility. *Nat. Commun.* **12**, 1–6 (2021).
27. Brink, M., Chow, J. M., Hertzberg, J., Magesan, E. & Rosenblatt, S. Device challenges for near term superconducting quantum processors: frequency collisions. In *2018 IEEE IEDM*, pp. 6.1.1–6.1.3 (2018). https://ieeexplore.ieee.org/document/8614500.
28. Dickel, C. et al. Chip-to-chip entanglement of transmon qubits using engineered measurement fields. *Phys. Rev. B* **97**, 064508 (2018).
29. Rosenberg, D. et al. 3D integrated superconducting qubits. *NPJ Quant. Inf.* **3**, 1–5 (2017).
30. Foxen, B. et al. Qubit compatible superconducting interconnects. *Quant. Sci. Technol.* **3**, 014005 (2017).
31. Brecht, T. et al. Multilayer microwave integrated quantum circuits for scalable quantum computing. *NPJ Quant. Inf.* **2**, 1–4 (2016).
32. Reagor, M. et al. Demonstration of universal parametric entangling gates on a multi-qubit lattice. *Sci. Adv.* **4**, eaao3603 (2018).
33. Bertet, P., Harmans, C. & Mooij, J. Parametric coupling for superconducting qubits. *Phys. Rev. B* **73**, 064512 (2006).
34. Niskanen, A. et al. Quantum coherent tunable coupling of superconducting qubits. *Science* **316**, 723–726 (2007).
35. Didier, N., Sete, E. A., daSilva, M. P. & Rigetti, C. Analytical modeling of parametrically modulated transmon qubits. *Phys. Rev. A* **97**, 022330 (2018).
36. Hong, S. S. et al. Demonstration of a parametrically activated entangling gate protected from flux noise. *Phys. Rev. A* **101**, 012302 (2020).
37. Cicak, K. et al. Low-loss superconducting resonant circuits using vacuum-gap-based microwave components. *Appl. Phys. Lett.* **96**, 093502 (2010).
38. Lecocq, F., Teufel, J. D., Aumentado, J. & Simmonds, R. W. Resolving the vacuum fluctuations of an optomechanical system using an artificial atom. *Nat. Phys.* **11**, 635–639 (2015).
39. Minev, Z. K. et al. Planar Multilayer Circuit Quantum Electrodynamics. *Phys. Rev. Appl.* **5**, 044021 (2016).
40. Nersisyan, A. et al. Manufacturing low dissipation superconducting quantum processors. In *2019 IEEE IEDM*, pp. 31.1.1–31.1.4 (2019). https://ieeexplore.ieee.org/document/8993458.
41. Xu, Y. et al. High-Fidelity, High-Scalability Two-Qubit Gate Scheme for Superconducting Qubits. *Phys. Rev. Lett.* **125**, 240503 (2020).
42. Yan, F. et al. Tunable Coupling Scheme for Implementing High-Fidelity Two-Qubit Gates. *Phys. Rev. Appl.* **10**, 054062 (2018).
43. Magesan, E. et al. Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking. *Phys. Rev. Lett.* **109**, 080505 (2012).
44. Knill, E. et al. Randomized benchmarking of quantum gates. *Phys. Rev. A* **77**, 012307 (2008).
45. Narla, A. et al. Robust Concurrent Remote Entanglement Between Two Superconducting Qubits. *Phys. Rev. X* **6**, 031036 (2016).
46. Campagne-Ibarcq, P. et al. Deterministic Remote Entanglement of Superconducting Circuits through Microwave Two-Photon Transitions. *Phys. Rev. Lett.* **120**, 200501 (2018).
47. Zhong, Y. P. et al. Violating Bell's inequality with remotely connected superconducting qubits. *Nat. Phys.* **15**, 741–744 (2019).
48. O'Brien, W. et al. Superconducting caps for quantum integrated circuits. Preprint at https://arxiv.org/abs/1708.02219 (2017).

49. Scheer, M. G. & Block, M. B. Computational modeling of decay and hybridization in superconducting circuits. Preprint at https://arxiv.org/abs/1810.11510 (2018).
50. Ambegaokar, V. & Baratoff, A. Tunneling Between Superconductors. *Phys. Rev. Lett.* **10**, 486–489 (1963).

ACKNOWLEDGEMENTS

The authors would like to thank A. Grassellino, A. Romanenko, L. Cardani, and R. McDermott for insightful discussions on the effects of cosmic and background radiation on qubit coherence. We thank the Rigetti fabrication team for manufacturing the device, the Rigetti technical operations team for fridge build out and maintenance, the Rigetti cryogenic hardware team for providing the chip packaging, the Rigetti quantum engineering team, in particular Alex Hill and Joseph Valery, for guidance during measurement and data analysis, and the Rigetti control systems and embedded software teams for creating the Rigetti AWG control system. This material is based upon work supported by Rigetti Computing and the Defense Advanced Research Projects Agency (DARPA) under agreement No. HR00112090058 and DARPA under IAA 8839, Annex 114.

AUTHOR CONTRIBUTIONS

A.G. and J.P.P. contributed equally to this work. A.G. led the team working on this project, analyzed the simulation and experimental data and wrote the paper apart from section E, J.P.P. drafted the initial device design, designed the package and measured the device, A.S. and M.H. developed the inter-chip coupling concept, M.R. conducted the Bell Test experiment and wrote section II-E, M.S.A., N.D., A.R., E.S., D.V. and F.W. assisted with theory and data analysis, A.B., A.N., F.O., B.S. and B.S. designed and fabricated the device, C.J.W. assisted with device measurement, and C.R. supervised this work.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Alysson Gold.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.