Measurement and Control of Quasiparticle Dynamics in a Superconducting Qubit

C. Wang,[1, *] Y. Y. Gao,[1] I. M. Pop,[1] U. Vool,[1] C. Axline,[1] T. Brecht,[1] R. W. Heeres,[1]
L. Frunzio,[1] M. H. Devoret,[1] G. Catelani,[2] L. I. Glazman,[1] and R. J. Schoelkopf[1]

[1]*Department of Applied Physics and Physics, Yale University, New Haven, CT 06520, USA*
[2]*Peter Grünberg Institut (PGI-2), Forschungszentrum Jülich, 52425 Jülich, Germany*
(Dated: October 27, 2014)

Superconducting circuits have attracted growing interest in recent years as a promising candidate for fault-tolerant quantum information processing. Extensive efforts have always been taken to completely shield these circuits from external magnetic field to protect the integrity of superconductivity. Surprisingly, here we show vortices can improve the performance of superconducting qubits by reducing the lifetimes of detrimental single-electron-like excitations known as quasiparticles. Using a contactless injection technique with unprecedented dynamic range, we quantitatively distinguish between recombination and trapping mechanisms in controlling the dynamics of residual quasiparticles, and show quantized changes in quasiparticle trapping rate due to individual vortices. These results highlight the prominent role of quasiparticle trapping in future development of superconducting qubits, and provide a powerful characterization tool along the way.

I. INTRODUCTION

Superconducting quantum circuits have made rapid progress[1] in realizing increasingly sophisticated quantum states[2,3] and operations[4,5] with high fidelity. Excitations of the superconductor, or quasiparticles (QP), can limit their performance by causing relaxation and decoherence, with the rate approximately proportional to the QP density[6]. Operating at 20 mK or lower temperature, superconducting aluminum in thermal equilibrium should have no more than one pair of quasiparticles for the volume of the Earth. However, a substantial background of quasiparticles has been observed in various devices from single electron[7] or Cooper-pair transistors[8], kinetic inductance detectors[9,10] to superconducting qubits[11–13]. A detailed understanding of the generation mechanism and dominant relaxation processes will eventually be necessary to suppress this small background of quasiparticles and continue the improvement of these devices.

Quasiparticle dynamics has been traditionally characterized by the "lifetime" of excess quasiparticles relaxing towards a steady state, τ_{ss}. A variety of techniques have been used to measure τ_{ss} in aluminum, including low-frequency sub-gap electrical[14,15] or thermal transport[16,17], resonance frequency shifts[18,19] and more recently, qubit energy decay[20,21]. Electron-phonon mediated pair-recombination has been established as the canonical mechanism of QP decay[22]. Single-quasiparticle loss mechanisms in the presence of quasiparticle "traps" such as normal metal contacts[7,16,17,23,24], engineered gap inhomogeneity[8,25,26], Andreev bound states[27] or magnetic field penetration[28,29] have also been studied.

Ascertaining the relative importance of quasiparticles to the decoherence of any superconducting qubit, versus other mechanisms such as dielectric loss or radiation, remains challenging. Ideally, one would like to probe and vary QP dynamics in a highly-coherent qubit, without requiring additional circuit elements or constraints that compromise its performance. Here we introduce such a technique, capable of measuring the QP dynamics of a qubit in operation and quantifying the processes of recombination, trapping, diffusion and background generation of quasiparticles. With this technique, we directly measure the trapping of quasiparticles by a single vortex, a basic property of superconductors. We demonstrate that QP trapping by vortices can suppress the background QP density, resulting in surprising net improvement of qubit coherence in spite of the well-known vortex flow dissipation[30].

In this Article, we present time-domain measurements of quasiparticle relaxation in 3D transmon qubits[31] over 2-3 orders of magnitude in density. We find the QP dynamics can be dominated by either recombination or trapping effects depending on the device geometry and the resultant presence or absence of vortices. We demonstrate strong *in-situ* control of QP dynamics by magnetic field, and measure an intrinsic single-vortex trapping "power" of $(6.7 \pm 0.5) \times 10^{-2}$ cm^{2}s^{-1} (*i.e.* $\tau_{ss} = 1$ s induced by a single vortex over an area of 6.7×10^{-2} cm^2). Improvements of relaxation time (T_1) and coherence time (T_{2E}) by more than a factor of 2 are observed in one geometric design of devices when the devices are cooled in a small magnetic field (10-200 mG). We measure a stray QP generation rate of about 1×10^{-4}s^{-1}, suggesting the long coherence time of the widely-adapted design of 3D transmons[12,31–33] may already be greatly assisted by unintentional vortices. Improved coherence by field-cooling, also being reported in superconducting planar resonators[34] and fluxonium qubits[35] during the preparation of our manuscript, can be definitively correlated with QP trapping rates of vortices as probed by our measurements of QP dynamics.

II. RESULTS

Injection and measurement of quasiparticles. In our experiment, we inject quasiparticles into two types of transmon qubits in a 3D circuit QED architecture[31] us-

ing only the existing microwave ports. Each transmon qubit comprises a single $\text{Al}/\text{AlO}_x/\text{Al}$ Josephson junction shunted by a large Al coplanar capacitor (electrodes) on a c-plane sapphire substrate. Type A devices are very similar to those in ref. 31 with a pair of large 500 μm $\times$ 250 μm electrodes (Fig. 1b). The electrodes of Type B devices are composed of a narrow (6 to 30 μm wide) coplanar gap capacitor and a pair of 80 μm $\times$ 80 μm "pads" (see Fig. 1c and Supplementary Note 1), but provide total capacitance and qubit Hamiltonian parameters very similar to Type A devices. Chips containing one or two qubits are mounted in 3D aluminum or copper rectangular waveguide cavities (Fig. 1a) and all measurements are done in an Oxford cryogen-free dilution refrigerator at base temperature of 15-20 mK, with magnetic field shielding, infrared shielding and filtering described in ref. 32. To inject quasiparticles, similarly to ref. 35 we apply a high-power microwave pulse at the bare cavity resonance frequency from the input port. The injection pulse creates about 10^5 circulating photons in the cavity, resulting in RF voltage across the Josephson junction that exceeds the superconducting gap, and produces $\sim 10^5$ quasiparticles per μs. The duration of the injection pulse is long enough (200-500 μs) so that the injected quasiparticles can fully diffuse within the device while the production and loss of quasiparticles reach a dynamic balance. (See Supplementary Notes 2 and 3 for analysis of QP injection and diffusion.)

We use the recovery of the energy relaxation time (T_1) of the qubit as a direct and calibrated probe of the decay of QP density. Standard microwave pulse sequences are applied to determine the qubit T_1 following a variable delay t after the QP injection (Fig. 1f and Supplementary Note 4), from which we extract the qubit relaxation rate $\Gamma = 1/T_1$ as a function of t as shown in Fig. 2. Despite possible heating due to the injection pulse, we find the effective temperature of the qubit and quasiparticle bath does not exceed 70 mK for the entire range of our measurement (Supplementary Note 5), therefore thermal generation of quasiparticles and spontaneous $|g\rangle \to |e\rangle$ transition of the qubit can be neglected. We use x_{qp} to represent the QP density near the Josephson junction normalized by the Cooper pair density ($n_{cp} \approx 4 \times 10^6$ μm^{-3} for aluminium). It is related to the measured qubit decay rate by $\Gamma(t) = Cx_{\text{qp}}(t) + \Gamma_{ex}$, where Γ_{ex} is a constant qubit decay rate due to non-quasiparticle dissipation mechanisms, and $C = \sqrt{2\omega_q\Delta/\pi^2\hbar}$ is a calculated[6] and confirmed[31] constant involving only the superconducting gap $\Delta(\approx 180~\mu eV)$ and the angular frequency of the transmon $\omega_q(\approx 2\pi \cdot 6$ GHz). Since Γ_{ex} is strictly bounded by the qubit relaxation rate without QP injection, $\Gamma_{ex} \leq \Gamma_0 = \Gamma(t \to \infty)$, for a significant range of the data, QP density can be approximated by $x_{\text{qp}}(t) \approx \Gamma(t)/C$.

Distinguishing between quasiparticle recombination and trapping. The dynamics of the quasiparticle density x_{qp} near the junction in the presence of recombination and trapping can be modeled by the following
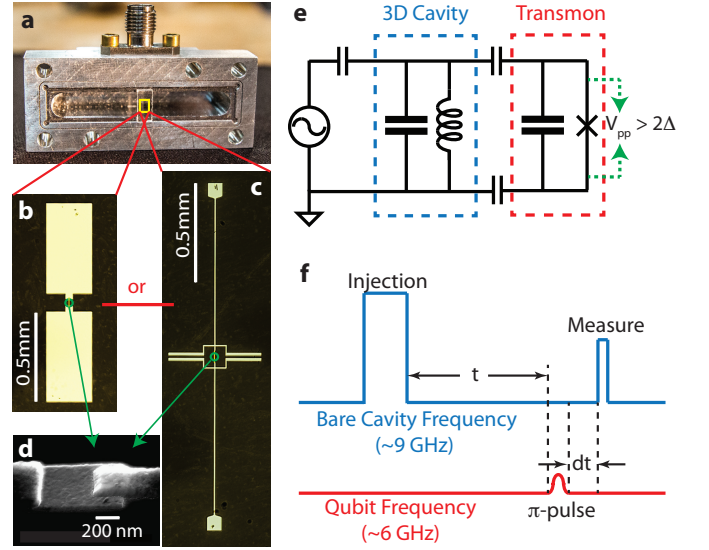

FIG. 1. **Device and measurement schematics. a**, Photograph image of a 3D aluminum cavity loaded with a transmon qubit. **b**, **c**, Optical images of a Type A device and a Type B device. **d**, Scanning electron microscope image of a Josephson junction located in between the two electrodes of a transmon. **e**, Effective circuit of the cavity-qubit system. Injection of quasiparticles is achieved by applying a resonant microwave voltage to the 3D cavity capacitively coupled to the junction. **f**, Pulse sequences for measuring the decay of quasiparticles based on qubit T_1. A π-pulse is applied to excite the qubit from $|g\rangle$ to $|e\rangle$ at a variable delay t after the injection pulse, followed by a readout pulse after another variable delay dt ($dt \ll t$).

equation (see Supplementary Note 6 for details):

$$\frac{dx_{\text{qp}}}{dt} = -rx_{\text{qp}}^2 - sx_{\text{qp}} + g \qquad (1)$$

The quadratic term describes the canonical QP recombination in pairs with a recombination constant r. The linear term describes trapping effects that localize or remove single quasiparticles from tunneling across the Josephson junction and inducing qubit relaxation. The effective trapping rate s depends not only on the property and density of the trapping sites, but also their geometric distribution and associated diffusion time scale. The constant term g describes QP generation rate by pair-breaking stray radiation or other unidentified sources[36]. If trapping is dominant ($s \gg rx_{\text{qp}}$ for most of the measured range of x_{qp}), the decay of x_{qp} follows an exponential function. This is a surprisingly good approximation for the $\Gamma(t)$ we measured in Device A1 (of Type A) (red fit in Fig. 2a). On the other hand, if recombination dominates, the decay of x_{qp} follows a hyperbolic cotangent function, with initially a steep $1/t$ decay crossing over to an exponential tail. We measure $\Gamma(t)$ in Device B1 (of Type B) strikingly close to this limit (green fit in Fig. 2b).

To analyze both recombination and trapping more quantitatively, we solve Eq. (1) analytically, yielding a
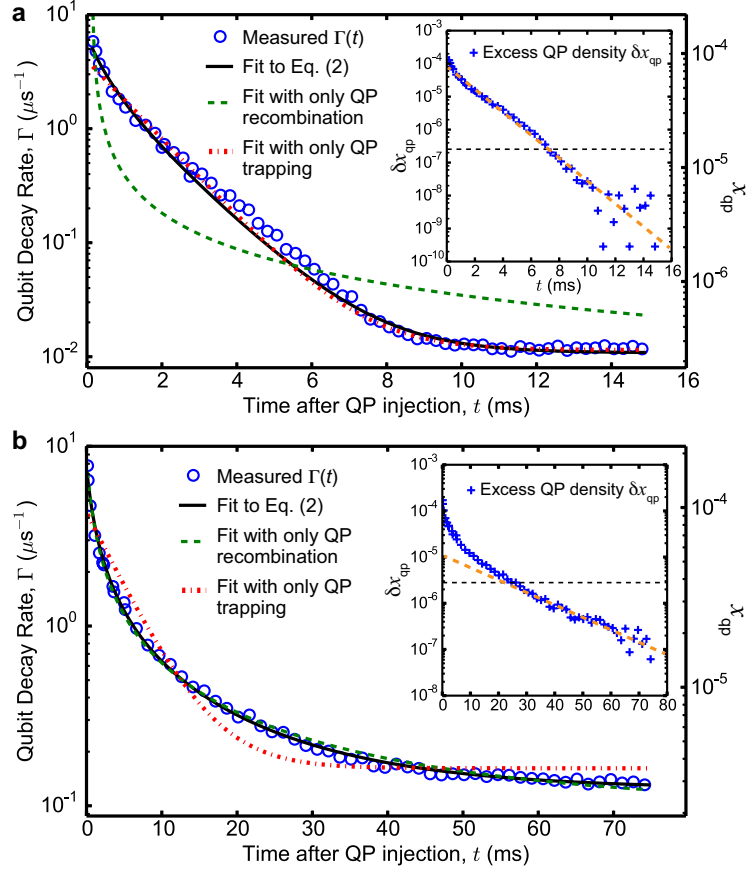
FIG. 2. **Quasiparticle decay dynamics characterized by the recovery functional forms of qubit decay rate.** Qubit energy decay rate Γ as a function of time t after quasiparticle injection pulses and fits to various functional forms for Device A1 (**a**, dominated by trapping) and Device B1 (**b**, dominated by recombination) in separate aluminum 3D cavities. For illustration purposes, we label $x_{\mathrm{qp}} = \Gamma/C$ on the right axes, neglecting qubit relaxation from non-QP mechanisms that have a stringent bound $\Gamma_{ex} \leq \Gamma_0$. Insets, replots of the $\Gamma(t)$ data in main panels in the form of excess QP density δx_{qp} after subtracting a background (whose magnitude Γ_0/C is shown by the black dashed line and is an upper bound of x_0). The magenta dashed line is guide to the eye of the exponential decay of δx_{qp} approaching the steady state.

four parameter fit for $\Gamma(t)$ (black curves in Fig. 2):

$$\Gamma(t) = C x_i \frac{1 - r'}{e^{t/\tau_{\mathrm{ss}}} - r'} + \Gamma_0 \qquad (2)$$

where x_i is the initial injected QP density, $\Gamma_0 = C x_0 + \Gamma_{ex}$ is the qubit relaxation rate without QP injection, consisting of contributions from both 'background QP density x_0 and other mechanisms. r' is a dimensionless fit parameter $(0 < r' < 1)$. Note that as $t \to \infty$, Eq. (2) approaches an exponential decay with time constant τ_{ss}. The recombination constant r and the trapping rate s can be determined from these fit parameters (Supplementary Note 6). For B1, a fit to Eq. (2) gives an exponential tail with $\tau_{\mathrm{ss}} = 18 \pm 2$ ms, a recombination constant $r = 1/(170 \pm 20 \text{ ns})$ and a weak trapping rate $s = 1/(30 \pm {}^{24}_{12}$ ms). For A1, we find $s \approx 1/\tau_{\mathrm{ss}} = 1/(1.5 \pm 0.1$ ms) and $r = 1/(105 \pm 30$ ns), with the trapping term dominating most of the measurement range $(x_{\mathrm{qp}} < 10^{-4})$.

The qualitative difference between the functional forms of QP decay in the two devices can be better illustrated

by plotting the excess quasiparticle density due to QP injection, $\delta x_{\mathrm{qp}} = x_{\mathrm{qp}}(t) - x_0$, as a function of time (Fig. 2 insets). For A1, the instantaneous QP decay rate indicated by the slope of $\delta x_{\mathrm{qp}}(t)$ remains equal to its steady-state value $(1/\tau_{\mathrm{ss}})$ even when the QP density is orders of magnitude higher than its background density (i.e. when $\delta x_{\mathrm{qp}}(t) \gg \Gamma_0/C \geq x_0$). For B1, the slope increases significantly when $\delta x_{\mathrm{qp}}(t) > x_0$.

Controlling quasiparticle dynamics by cooling in magnetic field. Why do the two devices with identical material properties and similar qubit properties differ so much in quasiparticle relaxation dynamics? We attribute this to the trapping effect from vortices, regions with diminished superconducting gap, in the large electrodes in A1 despite the small residual field $(B \sim 1\text{-}2$ mG) we are able to achieve with our magnetic shielding measures. On the other hand, Device B1 is likely free of vortices due to the much narrower geometry of the electrodes. To test this hypothesis, we repeat $\Gamma(t)$ measurements in B1 after cooling the device through the crit-

ical temperature (T_c) in a perpendicular magnetic field of variable magnitude B in either polarity. Indeed as B increases, we observe significant acceleration of QP decay with increasingly pronounced single-exponential characteristics, indicating enhanced QP trapping (Fig. 3a). In comparison, changing the applied magnetic field at 20 mK does not produce measurable changes in quasiparticle dynamics.

By fitting $\Gamma(t)$ to Eq. (2) at each cooling field, we find that: 1) the recombination constant r remains unchanged within fitting uncertainty, 2) the trapping rate s increases in discrete and near-equal steps for small magnetic fields ($B \lesssim 40$ mG) (Fig. 3c), 3) over a broader field range, s increases approximately linearly with B and saturates at $\gtrsim 1$ ms^{-1} at high field ($B \gtrsim 100$ mG) (Fig. 3b).

Our observed critical field threshold, B_k, where the trapping rate s starts to increase, corresponds to the expected entry of the first vortex in one of the 80 μm × 80 μm pads. B_k can be estimated based on a thermodynamic analysis of a vortex in a thin superconducting disk[37] together with consideration of vortex creation-annihilation kinetics[38], giving $B_k \sim 8$ mG, close to our observed values of 11 mG for B1 in an Al cavity, 14 mG and 10 mG for other two devices of Type B, B2 and B3, in Cu cavities. In comparison, B_k for Device A1 is estimated to be about 0.5 mG, lower than the estimated residual field (and its possible inhomogeneity) of our setup and therefore cannot be observed experimentally.

Quasiparticle trapping by individual vortices. The discrete trapping rates at small magnetic field are strongly suggestive of a fixed quasiparticle "trapping power" for each individual vortex. We define trapping power, P, by modeling a vortex as a point object at a 2D spatial coordinate $\vec{r}_0$ with a delta-function local trapping rate $P\delta(\vec{r} - \vec{r}_0)$. P could be more microscopically modeled as the product of QP trapping rate in the vortex core and an effective trapping area[28,34]. However, the "trapping power" representation offers the advantage of a general formulation without invoking any microscopic models. In the limit that diffusion is much faster than trapping, the total microscopic trapping power of N vortices, NP, manifests itself macroscopically as the product of the measured trapping rate and the total area, A, of the device, i.e. $sA = NP$. For a small number of vortices, we observe quantized changes of sA products in steps of ~ 0.06 cm^{2}s^{-1} consistent between all three systematically-measured Type B devices with up to 50% difference in device areas (Fig. 3c). In Fig. 3c we have subtracted a relatively small (zero-vortex) background trapping rate that varies from device to device (zero-field $s < 0.05$ ms^{-1} for B1 but ~ 0.18 ms^{-1} for B3), whose origin remains to be explored in future studies. Assuming each step corresponds to the entry of one vortex (which is stochastically most probable and also suggested by the widths of the steps), and adjusting for the finite speed of quasiparticle diffusion (Supplementary Note 6B), we determine trapping power $P = (6.7 \pm 0.5) \times 10^{-2}$ cm^{2}s^{-1}

as an intrinsic property of each individual vortex in our superconducting film.

The reduced step heights and the eventual saturation of s at higher magnetic field can be fully explained by QP diffusion using realistic geometric parameters of our device (Supplementary Note 6B). When there are a large number of vortices in the pads, the apparent trapping rate s is limited by the diffusion time for quasiparticles to reach the trapping pad from other regions of the device (Fig. 3e). The saturated trapping rate is higher for B2 because of the smaller volume of its gap capacitor. By fitting s as a function of B over a large range (Fig. 3b) for both devices, we determine the diffusion constant $D = 18 \pm 2$ cm^{2}s^{-1} for our Al film at 20 mK, consistent with the values measured in X-ray single-photon spectrometers adjusted for different temperature[25].

Quasiparticle recombination constant. Across seven devices (Supplementary Table 1) we measure recombination constants r in the range of $1/(170$ ns$)$ to $1/(80$ ns$)$, consistent with the theoretical electron-phonon coupling time of aluminum $\tau_0 = 438$ ns[22] adjusting for the phonon trapping effect[39]. Recombination-generated phonons can re-break Cooper-pairs before escaping into the substrate, reducing the effective recombination constant by a factor F, giving $r = 4(\frac{\Delta}{k_B T_c})^3/(F\tau_0) = 21.8/(F\tau_0)$. For our 80 nm bi-layer Al film on sapphire the best estimate of F is in the range of 5 to 10 considering the strong acoustic mismatch at the interface[40]. The reported values of r for Al in the literature range from $1/(120$ ns$)$ to $1/(8$ ns$)$[14,18,19,28,41]. The recombination constant we measured directly from the power-law decay characteristics is near the low end. It is comparable to $r = 1/(120$ ns$)$ extracted from DC steady-state injection measurements in extended Al films with similar thickness on sapphire[14].

Our measured value of r suggests the effect of recombination is extremely weak for the sparse QP background in typical superconducting qubits (recombination-resulted $\tau_{ss} \sim 50$ ms for $x_{qp} \sim 10^{-6}$). Even one single vortex (occupying less than one-millionth of the total area of a Type B device) can eliminate quasiparticles much faster than the intrinsic recombination process, as demonstrated by the drastically different QP decay curves for 0 (red), 1 (orange) and 2 (green) vortices in Fig. 3a.

Improved qubit coherence by vortices. In strong correlation with the reduced QP lifetime due to vortex trapping, we observe dramatic improvement of qubit coherence as a result of the suppressed background QP density. Improved qubit T_1 is already demonstrated by the lower background Γ_0 in Fig. 3a at higher cooling field. This "steady-state" T_1 of the qubit can also be measured separately without QP injection, which more than doubled from its zero-field value over a wide range of cooling magnetic field for both Devices B1 (Fig. 4a) and B2 (Fig. 4b). The coherence time with Hahn echo, T_{2E}, shows similar relative improvement because it is close to the limit of $2T_1$.
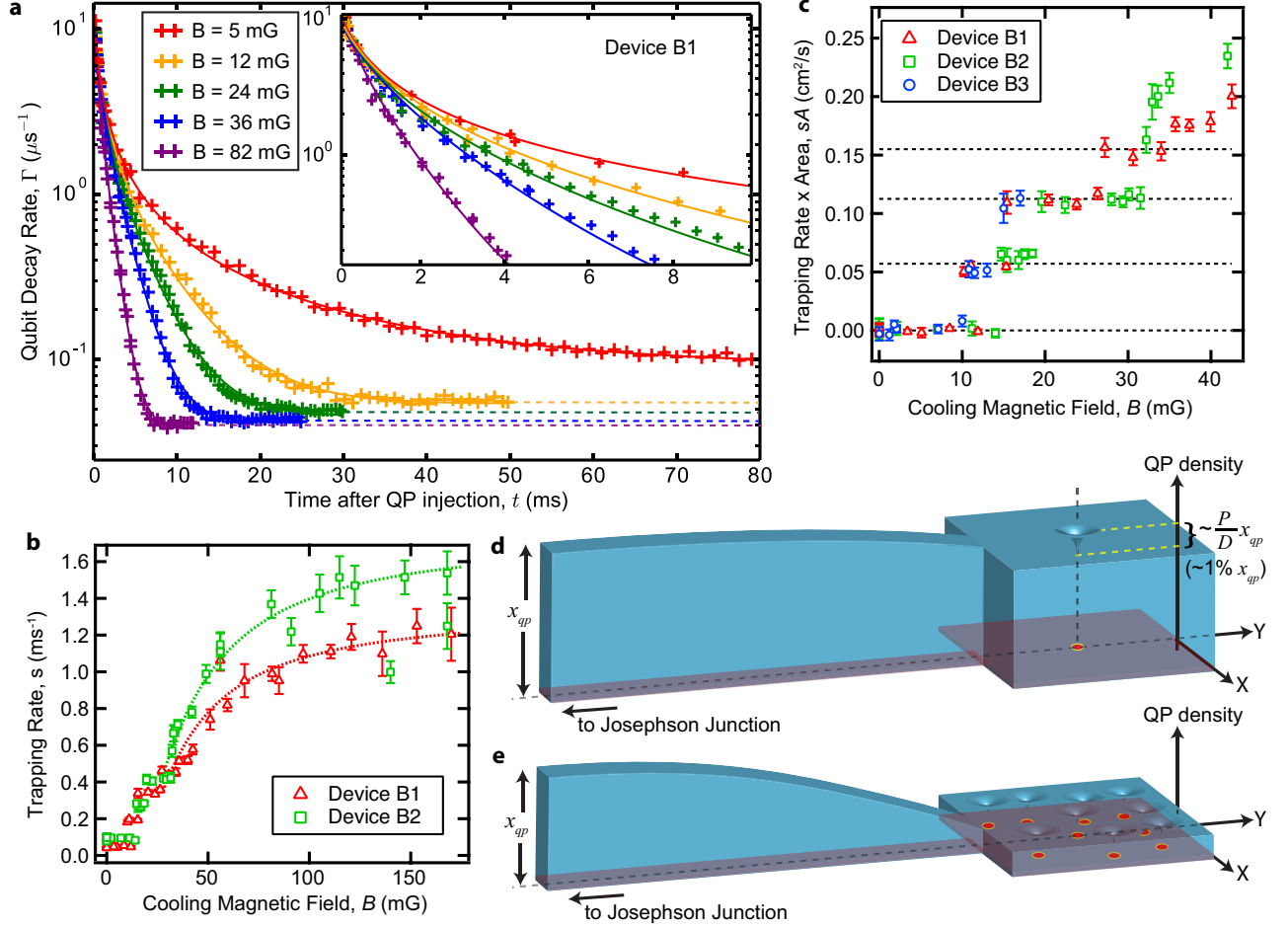
We can separate qubit loss mechanisms based on the

FIG. 3. **Quasiparticle trapping by vortices. a**, Qubit decay rate Γ as a function of time t after quasiparticle injection pulses for Device B1 cooled in a few selected magnetic fields B. Solid lines fits to Eq. (2). Dashed lines, background qubit decay rate indicating the "steady-state" T_1 of the qubit. Inset, zoom-in of the initial part of the decay. **b**, QP trapping rate s as a function of cooling magnetic field B for Devices B1 and B2. Each point represents data acquired after one field-cool thermal cycle (from above the T_c of aluminum), with the error bars showing the 1 s.d. of fluctuations for repetitive measurements within each thermal cycle. Dotted lines, fits to the QP trapping/diffusion model. **c**, QP trapping rate after subtracting a zero-vortex background and multiplying by the total area of the device, as a function of B for Devices B1, B2 and B3. Dashed lines are guides to the eye showing the discrete steps associated with 0, 1, 2 and 3 vortices. **d**, Cartoon schematic (not to scale) of the near-homogeneous spatial distribution of QP density, represented by height of the blue volume ("tank water level"), from the narrow wire (left) to the trapping pad (right) in a Type B device when only one vortex is present. QP density near the vortex displays a small ripple with a relative depth on the order of $P/D \approx 1\%$. **e**, Cartoon schematic (not to scale) of the QP density distribution with large number of vortices in the pad. The density at the trapping pad is much lower than x_{qp} near the junction.

linear relation between $1/T_1$ and τ_{ss} (noting $x_0 \approx g\tau_{\mathrm{ss}}$),

$$\frac{1}{T_1} = Cg\tau_{\mathrm{ss}} + \Gamma_{ex} \qquad (3)$$

where the intercept equals the qubit relaxation rate due to other loss mechanisms, and the slope reveals the stray QP generation rate g (Fig. 4c, d). In both B1 and B2, Γ_{ex} is likely dominated by dielectric loss at the $\mathrm{Al/Al_2O_3}$ interface under the gap capacitor[42]. We find QP generation rate $g \approx 0.7\text{-}1.3 \times 10^{-4}\mathrm{s}^{-1}$ in the two devices over several thermal cycles, or 0.3-0.6 quasiparticles created per ms for every $\mu\mathrm{m}^3$ of volume.

Vortex flow dissipation. The eventual decrease of T_1 and T_{2E} at high magnetic field ($B \gtrsim 200$ mG) can be attributed to vortices entering the gap capacitors where current density is high. Such dissipation due to vortex flow resistance has been well-known to degrade the quality factor of superconducting resonators and qubits[43,44]. Therefore precautions such as multilayer magnetic shielding, specialized non-magnetic hardware and honeycomb-style device designs have been widely employed in the community to avoid vortices. However, here we have shown that the benefit of vortices in suppressing non-equilibrium quasiparticles is not only relevant to practical
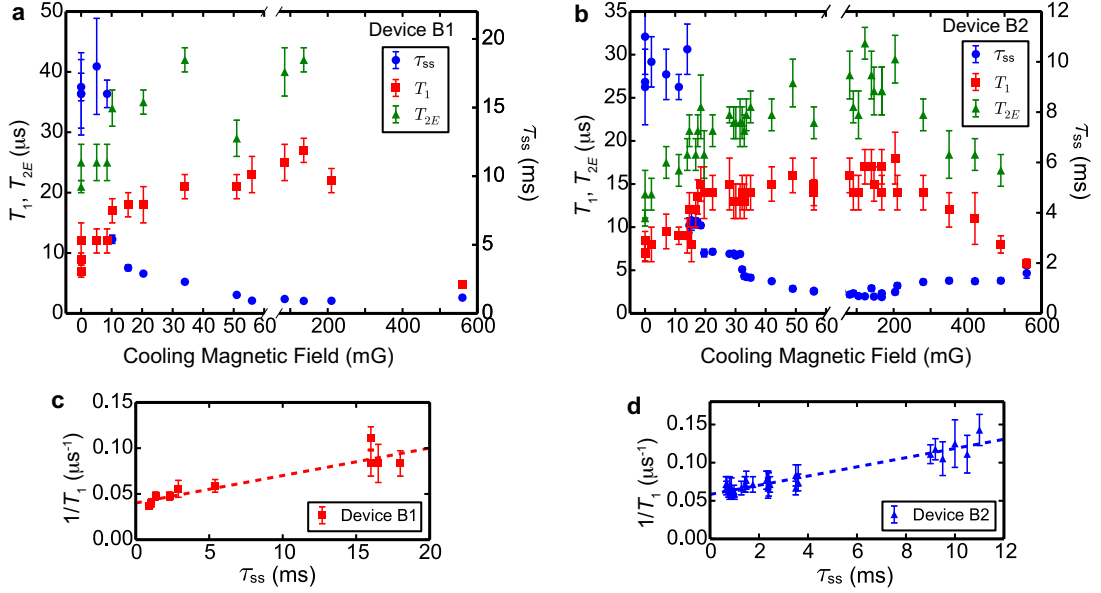
FIG. 4. **Improved qubit coherence by cooling in magnetic field and its correlation with reduced QP lifetime. a**, **b**, Qubit steady-state relaxation time T_1, coherence time with echo T_{2E} and quasiparticle lifetime τ_{ss} as a function of cooling magnetic field B for Devices B1 (**a**) and B2 (**b**). Each point represents data acquired after one field-cool thermal cycle, with the error bars showing the 1 s.d. of fluctuations for repetitive measurements within each thermal cycle. Note the discontinuous scale of the X-axes. **c**, **d**, $1/T_1$ versus τ_{ss} as varied by cooling in different magnetic fields for B1 (**c**) and B2 (**d**). The dashed lines are linear fits based on Eq. (3), giving $\Gamma_{ex} = 1/(26~\mu s)$, $g = 0.7 \times 10^{-4} s^{-1}$ for B1 and $\Gamma_{ex} = 1/(17~\mu s)$, $g = 1.3 \times 10^{-4} s^{-1}$ for B2 respectively.

devices, but can also significantly outweigh its negative impact if the locations of the vortices are optimized.

In device A1 (or most Type A devices, see Supplementary Table 1), our measured τ_{ss} at nominally zero field implies the presence of about 20 vortices based on our measured single-vortex trapping power, consistent with its geometry assuming a residue magnetic field of 1-2 mG. Introducing more vortices to Type A devices by cooling in an applied magnetic field can further reduce τ_{ss}, but no clear improvement of qubit T_1 is observed. We attribute it to the limited range that τ_{ss} can be varied (from its already small value at nominally zero field), and the vortex flow dissipation which clearly reduces T_1 at $B \gtrsim 30$ mG (Supplementary Note 8). Assuming both types of devices have similar g as they are shielded and measured in the same setup, it is plausible to infer that the long coherence times of the widely-adopted Type A 3D transmons would have been limited to much lower values without the assistance by QP trapping of unintentional vortices.

III. DISCUSSION

In our work, for the first time, the interaction between quasiparticles and a single vortex is measured. The single-vortex trapping power (P) is an intrinsic property of an aluminum film, and has the same dimension as the diffusion constant (D). The fact that we measure $P/D \approx 10^{-2}$ implies that a quasiparticle can diffuse

through a vortex with only a small ($\sim 1\%$) probability to be trapped. Despite a vortex being a topological defect with $\Delta = 0$ at its core, the spatial distribution of QP density is barely perturbed by the presence of a vortex, just like a small "ripple" on the order of 1% deep in a flat "sea" (Fig. 3d, see Supplementary Note 6E for calculation). For a uniform film extended in 2D space, this relatively homogeneous QP distribution holds for all practical length scales at any density of vortices, so the trapping rate s can be simply computed from the total trapping power. In the high magnetic field limit, this leads to $s \propto B$ with a linear coefficient $P/\phi_0 = 0.3$ $\mu s^{-1} G^{-1}$ based on our measured P, close to the fitted value of 0.5 $\mu s^{-1} G^{-1}$ in ref. 28.

Understanding and controlling the steady-state QP lifetime (τ_{ss}) may be important for a variety of superconducting devices. For superconducting qubits and resonators, achieving short τ_{ss} (either with vortices or potentially more efficient methods such as band-gap engineering or normal metal traps) to suppress the background QP density is desirable. In other devices such as kinetic inductance detectors[9,10], it may be desirable to obtain a long QP lifetime. Our measured τ_{ss} of 18 ms in a Type B device is so far the longest reported in aluminum. Previous experiments in aluminum films showed that the expected exponential increase of τ_{ss} with lower temperature saturates below 200 mK[14,15,18] to about 3 ms at most[10].

To shed light on the mechanisms limiting τ_{ss} in this

regime of extremely low QP density (without intentional QP traps), our technique allows quantitative separation between QP recombination and any residual trapping effects. Our quantification of the weak recombination of background quasiparticles and the single-vortex trapping extends our understanding of QP dynamics into the 10's of millisecond regime, and thus sets a more stringent bound on possible additional mechanisms in limiting the QP lifetimes.

To distinguish between recombination and trapping of quasiparticles, our analysis relies critically on large dynamic range in x_{qp} to achieve sufficient contrast between the functional forms of QP decay. A measurement near a steady state would observe the exponential tail of the decay, giving $\tau_{ss} = 1/(s + 2rx_0)$ without distinguishing between the two mechanisms. Such analysis of linear response has been traditionally carried out in time-domain measurements after photon pulses[15,19,23,25]. Noise spectroscopy measurements[15,19] by design are also limited to measuring the single time scale of τ_{ss}. The experiment of Lenander et al.[20] introduced the use of qubit T_1 to probe quasiparticle dynamics, but the achieved dynamic range was below a factor of 4. The dynamic range of 2-3 orders of magnitude in our experiment has been made possible by a long T_1 time in 3D transmons, the effectiveness of our microwave injection technique, and the geometric simplicity of an isolated aluminum island to eliminate out-diffusion.

A significant stray QP generation rate of about $1 \times 10^{-4}\mathrm{s}^{-1}$ has been measured in our study. Due to the large device volume and relatively long integration time, our measured g should be considered a spatial-temporal average of QP generation rate. Quite remarkably, it agrees within a factor of 3 with the average QP generation rate in a much smaller fluxonium qubit[35] with a much shorter τ_{ss} where there is evidence for the discreteness of QP numbers and QP generation events. This magnitude of QP generation rate, together with the weakness of recombination at low QP density, strongly suggests QP trapping should be an important ingredient for suppressing non-equilibrium x_{qp} in superconducting qubits and other devices to further improve performance. Our method of extracting g utilizing the sensitivity of Type B devices to QP generation should facilitate future identification of the QP generation source. The injection and measurement technique introduced in this work can be readily applied to nearly all cQED implementations without any modification to device structure or measurement circuit, and can play a crucial role in the future development of quantum circuits as a powerful probe of quasiparticle dynamics.

IV. ACKNOWLEDGEMENTS

V. AUTHOR CONTRIBUTIONS

C.W. and Y.Y.G. carried out the experiment and performed data analysis based on the model developed by C.W., G.C. and L.I.G. The experimental method was conceived and developed by I.M.P., U.V., C.W., Y.Y.G., M.H.D. and R.J.S. C.A., T.B. and R.W.H. provided further experimental contributions. Devices are fabricated by C.W. and L.F. C.W. and R.J.S. led the writing of the manuscript. All authors provided suggestions for the experiment, discussed the results and contributed to the manuscript.

VI. ADDITIONAL INFORMATION

Competing financial interests: The authors declare no competing financial interests.

[*] chen.wang@yale.edu

[1] Devoret, M. H. & Schoelkopf, R. J. Superconducting circuits for quantum information: an outlook. *Science* **339,** 1169-1174 (2013).

[2] Hofheinz, M. *et al.* Synthesizing arbitrary quantum states in a superconducting resonator. *Nature* **459,** 546-549 (2009).

[3] Vlastakis, B. *et al.* Deterministically encoding quantum information using 100-photon Schrödinger cat states. *Science* **342,** 607-610 (2013).

[4] Steffen, L. *et al.* Deterministic quantum teleportation with feed-forward in a solid state system. *Nature* **500,** 319-322 (2013).

[5] Chow, J. M. *et al.* Implementing a strand of a scalable fault-tolerant quantum computing fabric. *Nat. Commun.* **5,** 4015 (2014).

[6] Catelani, G., Schoelkopf, R. J., Devoret, M. H. & Glazman, L. I. Relaxation and frequency shifts induced by quasiparticles in superconducting qubits. *Phys. Rev. B* **84,** 064517 (2011).

[7] Knowles, H. S., Maisi, V. F. & Pekola, J. P. Efficiency of quasiparticle evacuation in superconducting devices. *Appl. Phys. Lett.* **100,** 262601 (2012).

[8] Aumentado, J., Keller, M. W., Martinis, J. M. & Devoret, M. H. Nonequilibrium quasiparticles and 2e periodicity in single-Cooper-pair transistors. *Phys. Rev. Lett.* **92,** 066802 (2004).

[9] Day, P. K., LeDuc, H. G., Mazin, B. A., Vayonakis, A. & Zmuidzinas, J. A. broadband superconducting detector suitable for use in large arrays. *Nature* **425**, 817-821 (2003).

[10] de Visser, P. J., Baselmans, J. J. A., Bueno, J., Llombart, N. & Klapwijk, T. M. Fluctuations in the electron system of a superconductor exposed to a photon flux. *Nat. Commun.* **5**, 4130 (2014).

[11] Sun, L. *et al.* Measurements of quasiparticle tunneling dynamics in a band-gap-engineered Transmon Qubit. *Phys. Rev. Lett.* **108**, 230509 (2012).

[12] Ristè, D. *et al.* Millisecond charge-parity fluctuations and induced decoherence in a superconducting transmon qubit. *Nat. Commun.* **4**, 1913 (2013).

[13] Pop, I. M. *et al.* Coherent suppression of electromagnetic dissipation due to superconducting quasiparticles. *Nature* **508**, 369-372 (2014).

[14] Gray, K. E. Steady state measurements of the quasiparticle lifetime in superconducting aluminium. *J. Phys. F: Met. Phys.* **1**, 290-308 (1971).

[15] Wilson, C. M., Frunzio, L. & Prober, D. E. Time-resolved measurements of thermodynamic fluctuations of the particle number in a nondegenerate Fermi gas. *Phys. Rev. Lett.* **87**, 067004 (2001).

[16] Ullom, J. N, Fisher, P. A. & Nahum, M. Measurements of quasiparticle thermalization in a normal metal. *Phys. Rev. B* **61**, 14839 (2000).

[17] Rajauria R. *et al.* Quasiparticle-diffusion-based heating in superconductor tunneling microcoolers. *Phys. Rev. B* **85**, 020505 (2012).

[18] Barends, R. *et al.* Quasiparticle relaxation in optically excited high-Q superconducting resonators. *Phys. Rev. Lett.* **100**, 257002 (2008).

[19] de Visser, P. J. *et al.* Number fluctuations of sparse quasiparticles in a superconductor. *Phys. Rev. Lett.* **106**, 167004 (2011).

[20] Lenander, M. *et al.* Measurement of energy decay in superconducting qubits from nonequilibrium quasiparticles. *Phys. Rev. B* **84**, 024501 (2011).

[21] Wenner, J. *et al.* Excitation of superconducting qubits from hot nonequilibrium quasiparticles. *Phys. Rev. Lett.* **110**, 150502 (2013).

[22] Kaplan, S. B. *et al.* Quasiparticle and phonon lifetimes in superconductors. *Phys. Rev. B* **14**, 4854-4873 (1976).

[23] Goldie, D. J., Booth N. E., Patel C. & Salmon G. L. Quasiparticle trapping from a single-crystal superconductor into a normal-metal film via the proximity effect. *Phys. Rev. Lett.* **64**, 954-957 (1990).

[24] Joyez, P., Lafarge, P., Filipe, A., Esteve, D. & Devoret, M. H. Observation of parity-induced suppression of Josephson tunneling in the superconducting single electron transistor. *Phys. Rev. Lett.* **72**, 2458-2461 (1994).

[25] Friedrich, S. *et al.* Experimental quasiparticle dynamics in a superconducting, imaging x-ray spectrometer. *Appl. Phys. Lett.* **71**, 3901-3903 (1997).

[26] Court N. A, Ferguson A. J., Lutchyn R. & Clark R. G. Quantitative study of quasiparticle traps using the single-Cooper-pair transistor. *Phys. Rev. B* **77**, 100501 (2008).

[27] Levenson-Falk, E. M., Kos, F., Vijay, R., Glazman, L. I. & Siddiqi, I. Single-quasiparticle trapping in aluminum nanobridge Josephson junctions. *Phys. Rev. Lett.* **112**, 047002 (2014).

[28] Ullom, J. N., Fisher, P. A. & Nahum, M. Magnetic field dependence of quasiparticle losses in a superconductor. *Appl. Phys. Lett.* **73**, 2494-2496 (1998).

[29] Peltonen, J. T., Muhonen, J. T., Meschke, M., Kopnin, N. B. & Pekola, J. P. Magnetic-field-induced stabilization of nonequilibrium superconductivity in a normal-metal/insulator/superconductor junction. *Phys. Rev. B* **84**, 220502 (2011).

[30] Bardeen, J. & Stephen, M. J. Theory of the motion of vortices in superconductors. *Phys. Rev.* **140**, A1197-A1207 (1965).

[31] Paik, H. *et al.* Observation of high coherence in Josephson junction qubits measured in a three-dimensional circuit QED architecture. *Phys. Rev. Lett.* **107**, 240501 (2011).

[32] Sears, A. P. *et al.* Photon shot noise dephasing in the strong-dispersive limit of circuit QED. *Phys. Rev. B* **86**, 180504 (2012).

[33] Vijay, R. *et al.* Stabilizing Rabi oscillations in a superconducting qubit using quantum feedback. *Nature* **490**, 77-80 (2012).

[34] Nsanzineza, I. & Plourde, B. L. T. Trapping a single vortex and reducing quasiparticles in a superconducting resonator. *Phys. Rev. Lett.* **113**, 117002 (2014).

[35] U. Vool, *et al.* Non-Poissonian quantum jumps of a fluxonium qubit due to quasiparticle excitations. Preprint at: http://arxiv.org/abs/1406.1769 (2014).

[36] Martinis, J. M., Ansmann, M. & Aumentado, J. Energy decay in superconducting Josephson-junction qubits from nonequilibrium quasiparticle excitations. *Phys. Rev. Lett.* **103**, 097002 (2009).

[37] Kogan, V. G., Clem, J. R. & Mints, R. G. Properties of mesoscopic superconducting thin-film rings: London approach. *Phys. Rev. B* **69**, 064516 (2004).

[38] Kuit, K. H. *et al.* Vortex trapping and expulsion in thin-film $YBa_2Cu_3O_{7-\delta}$ strips. *Phys. Rev. B* **77**, 134504 (2008).

[39] Rothwarf, A. & Taylor, B. N. Measurement of recombination lifetimes in superconductors. *Phys. Rev. Lett.* **19**, 27-30 (1967).

[40] Kaplan, S. B. Acoustic matching of superconducting films to substrates. *J. Low Temp. Phys.* **37**, 343-365 (1979).

[41] Wilson C. M. & Prober D. E. Quasiparticle number fluctuations in superconductors. *Phys. Rev. B* **69**, 094524 (2004).

[42] Wenner, J. *et al.* Surface loss simulations of superconducting coplanar waveguide resonators. *Appl. Phys. Lett.* **99**, 113513 (2011).

[43] Wang H. *et al.* Improving the coherence time of superconducting coplanar resonators. *Appl. Phys. Lett.* **95**, 233508 (2009).

[44] Song C. *et al.* Microwave response of vortices in superconducting thin films of Re and Al. *Phys. Rev. B* **79**, 174512 (2009).

Supplementary Information for "Measurement and Control of Quasiparticle Dynamics in a Superconducting Qubit"

C. Wang,[1, *] Y. Y. Gao,[1] I. M. Pop,[1] U. Vool,[1] C. Axline,[1] T. Brecht,[1] R. W. Heeres,[1]
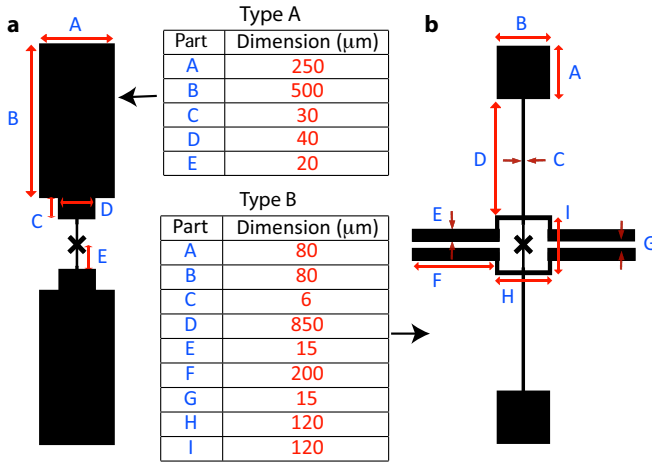L. Frunzio,[1] M. H. Devoret,[1] G. Catelani,[2] L. I. Glazman,[1] and R. J. Schoelkopf[1]

[1]*Department of Applied Physics and Physics, Yale University, New Haven, CT 06520, USA*
[2]*Peter Grünberg Institut (PGI-2), Forschungszentrum Jülich, 52425 Jülich, Germany*
(Dated: October 27, 2014)

SUPPLEMENTARY NOTE 1. DETAILS OF DEVICE GEOMETRY

The detailed dimensions of the electrodes of the two types of transmons are illustrated in Fig. S1. The width E and distance G of the coplanar gap capacitor in Type B are kept equal but varied from device to device. All the other dimensions are fixed. Device B1 has $E = G = 15$ μm, Device B2 has $E = G = 10$ μm, and Device B3 has $E = G = 30$ μm.

Type A

Part	Dimension (μm)
A	250
B	500
C	30
D	40
E	20

Type B

Part	Dimension (μm)
A	80
B	80
C	6
D	850
E	15
F	200
G	15
H	120
I	120

Supplementary Figure S1. **Detailed device geometry.** Dimensions of the electrodes of Type A (**a**) and Type B (**b**) transmons.

SUPPLEMENTARY NOTE 2. MODEL OF QUASIPARTICLE INJECTION

In our experiment, contactless injection of quasiparticles is realized by applying a high power microwave pulse at the bare cavity frequency to achieve an AC voltage $V_j > 2\Delta/e$ across the Josephson junction. The actual power used for injection is empirical without direct knowledge of V. We find three regimes separated by two critical power thresholds: 1) low power: there is negligible microwave transmission at the bare cavity frequency, because with the presence of Josephson inductance the cavity mode frequency is shifted. 2) medium power:

there is a drastic increase in transmission at the bare cavity frequency, and a small amount of QP's may be produced (affecting Γ by up to a few times) with strong sample-to-sample variations. 3) high power: transmitted power becomes linear as a function of input power, and a large number of QP's are produced (affecting Γ by up to 3 orders of magnitude). The first critical transition in transmission power is the mechanism of the widely-used Jaynes-Cummings readout,[1] and we believe the second transition in QP generation corresponds to $V_j \approx 2\Delta/e$. We typically use an injection power 1-4 dB above this second threshold power, which is about -70 to -55 dBm at the input port of the cavity based on our coarse estimate of the cable loss. Here we provide a brief calculation of the microwave power required to realize $V_j = 2\Delta/e$ based on a self-consistency argument.

Assuming the RMS voltage $V_j = 2\Delta/e$, the junction can be approximated by a normal resistor. We use a HFSS (Ansys) numerical package to simulate the quality factor of a lossless cavity containing our qubit structure with a tunnel junction resistance of R_j. This "internal" Q of the cavity found in the simulation is solely due to the dissipation at the normal-state junction resistor, which we denote by Q_j. For our experimental parameter of $R_j \approx 8$ kΩ, we find $Q_j \approx 1.1 \times 10^4$.

The power dissipated in the junction P_j can be calculated as $P_j \approx V_j^2/R_j = 4\Delta^2/e^2 R_j$. P_j is connected to the incoming power P_{in} by:

$$\frac{P_j}{P_{in}} = \frac{4Q_{tot}^2}{Q_{in}Q_j} \qquad (1)$$

where Q_{tot} is the total Q of the cavity, composed of coupling Q's of the input and output microwave ports Q_{in} and Q_{out}, the normal-state junction Q_j and Q from the cavity wall (and dielectric, etc.) Q_w:

$$\frac{1}{Q_{tot}} = \frac{1}{Q_{in}} + \frac{1}{Q_{out}} + \frac{1}{Q_j} + \frac{1}{Q_w} \qquad (2)$$

For our devices in aluminum cavity, we typically have $Q_{in} \approx 2 \times 10^6$, $Q_{out} \approx 1 \times 10^5$, $Q_w \gg 10^5$, so $Q_{tot} \approx Q_j$, therefore the input power required for $V_j = 2\Delta/e$ is:

$$P_{in} \approx \frac{Q_{in}}{4Q_j}P_j \approx 200\frac{\Delta^2}{e^2 R} \approx 10^{-9}\text{W} = -60\text{dBm} \qquad (3)$$

For our devices in copper cavity, we typically have $Q_{in} \approx 3 \times 10^5$, $Q_{out} \approx 1.5 \times 10^4$, $Q_w \approx 1.5 \times 10^4$, so
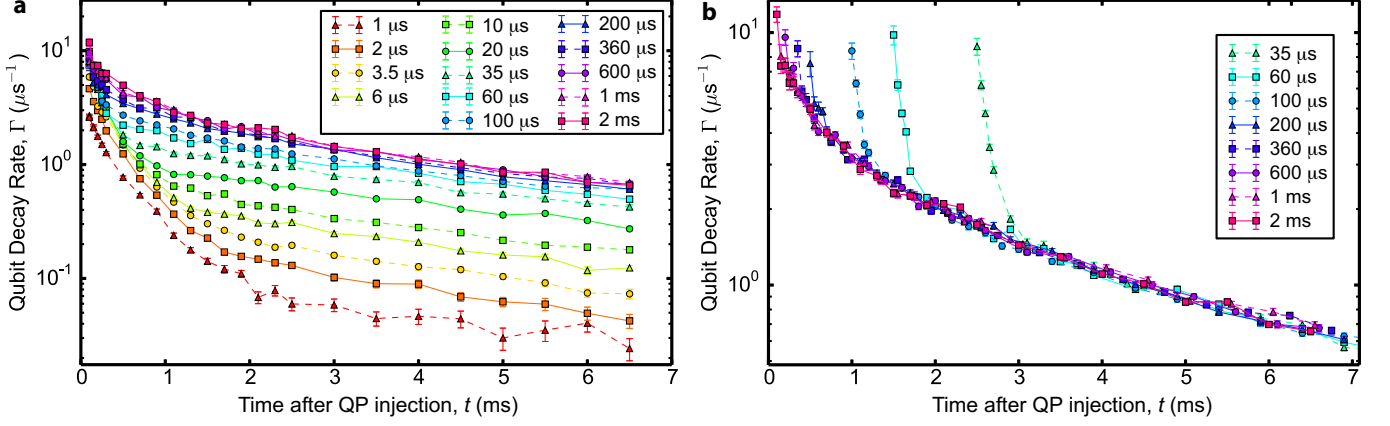
Supplementary Figure S2. **Pulse-length-dependence of measured quasiparticle decay dynamics. a**, $\Gamma(t)$ for Device B2 cooled in zero magnetic field following quasiparticle injection pulses with a series of different pulse length t_{inj}. A background qubit decay rate (Γ_0) has been subtracted. **b**, Selected $\Gamma(t)$ traces from **a** with each trace shifted horizontally by a time offset. Error bars represent one s.d. uncertainty in determining T_1 from individual exponential fits.

$Q_{tot} \approx 4.5 \times 10^3$, so:

$$P_{in} = \frac{Q_{in}Q_j}{4Q_{tot}^2}P_j \approx 200\frac{\Delta^2}{e^2 R} \approx 10^{-9}\text{W} = -60\text{dBm} \quad (4)$$

In both cases, we find the required input power similar to the injection power threshold we have observed.

So how many quasiparticles does the pulse produce? We make another order-of-magnitude estimate here by assuming each tunneling electron produces one pair of QP's. The number of tunneling electrons per second is given by $V_j/(R_j e)$. Therefore the QP injection rate is:

$$G \approx 2\frac{V_j}{R_j e} \approx \frac{4\Delta}{e^3 R_j} \approx 5 \times 10^5/\mu s \quad (5)$$

A better estimate may consider that the instantaneous voltage across the junction is only above $2\Delta/e$ for part of an oscillation period, resulting in a slightly lower injection rate.

SUPPLEMENTARY NOTE 3. INJECTION PULSE LENGTH AND THE ROLE OF DIFFUSION

Since the injected quasiparticles are concentrated near the Josephson junction, their spatial distribution is in general not homogenous. However, our measurement technique is only sensitive to the QP density near the junction. (Note that we have explicitly specified the QP density for the near-junction area in the definition of x_{qp}) Therefore, quasiparticles diffusing away from the junction will also appear as a decay of x_{qp}. This complexity can be analyzed in depth by measurements of QP decay following injection pulses with different length.

In Fig. S2a we show a large series of $\Gamma(t)$ curves for Device B2 cooled in zero field following injection pulses

with different pulse length t_{inj}. A background qubit decay rate (Γ_0) has been subtracted so that the $\Gamma(t)$ shown here is strictly due to the excess quasiparticles from the injection. We find: 1) For $t_{inj} \geq 6~\mu s$, all curves start from about the same Γ at $t = 100~\mu s$ (the first point of our measurement), therefore extending injection pulse to more than $6\mu s$ no longer increases QP density near the junction. 2) For $6~\mu s \leq t_{inj} \leq 600~\mu s$, the $\Gamma(t)$ curves quickly deviate from each other after the initial points. After a sufficiently long wait time t (e.g. see the last point of the curves $t = 6.5$ ms), we see substantially higher x_{qp} for longer injection pulses. This indicates a longer injection pulse injects more quasiparticles filling up remote areas of the device, while a shorter injection pulse results in a steeper initial decay due to quasiparticles diffusing away from the near-junction area. 3) As $t_{inj} \geq 600~\mu s$, the $\Gamma(t)$ curves are completely indistinguishable. In this regime, the QP injection has reached a dynamic balance with the QP diffusion and relaxation, which we call "saturation injection", and the entire spatial distribution of QP density no longer changes with increasing t_{inj}.

By shifting each $\Gamma(t)$ curve in Fig. S2a horizontally by a time offset, we find that all curves merge into a universal decay curve after some short-term behavior specific to each curve (Fig. S2b) (which, as described above, is due to different initial spatial distribution of quasiparticles). This universal curve is therefore the slowest eigenmode of the collective decay of quasiparticles in the entire electrodes, which is described by Eq. (1) in the main text. When recombination is dominant this mode is homogenous, and in the presence of trapping this mode may involve a steady spatial gradient of quasiparticles. The time it takes for each measured $\Gamma(t)$ to converge to this decay mode depends on the diffusion time within the electrodes, but the deviation can be minimized by choosing an initial spatial distribution (after the injection pulse) as close to the eigenmode as possible. Therefore, in or-

der to use the intuitive Eq. (1) in the main text without detailed simulation of the diffusion process, we choose $t_{inj} = 360$ μs for the data presented in the main text for Device B2, close to the state of saturation injection. Furthermore, when analyzing $\Gamma(t)$ data using Eq. (2) in the main text, we do not include data of $t < 200$ μs, allowing sufficient additional diffusion time for the QP distribution to approach the slowest decay mode we are interested in.

From Fig. S2 we can also estimate the time scale of diffusion in the device. We observe saturation injection with pulse length $\gtrsim 600$ μs. In addition, we find that after a short injection pulse such as $t_{inj} = 60$ μs it takes a similar amount of time ~ 500 μs for $\Gamma(t)$ to merge into the universal decay curve (Fig. S2b). Furthermore, the two numbers are in very good agreement with the saturated trapping rate $1/(600$ μs$)$ observed for this device [see Fig. 3b of the main text] which provides a reliable measure of the diffusion time scale (as modeled in Supplementary Note 6B).

## SUPPLEMENTARY NOTE 4. MEASUREMENT AND DATA PROCESSING METHODS

We use a combination of dispersive readout and high-power readout (using the Jaynes-Cummings nonlinearity[1]) to measure $\Gamma(t)$ shown in the main text. High-power readout provides much better signal-to-noise ratio (SNR), but we observe evidences that it may produce quasiparticles with mechanisms unclear (as we noted in Supplementary Note 2). Therefore we only perform one readout following each injection pulse using the pulse sequence shown in Fig. S3a. Any unintended impact of the readout on the QP system should be completely erased by the next saturation injection (see Supplementary Note 3). Dispersive readout has relatively poor SNR and requires much longer integration time, but it does not perturb quasiparticle dynamics significantly. Therefore, one can perform multiple T_1 measurements after each injection pulse as shown in Fig. S3b. This is an efficient method when a very long delay time t (10's of ms) is required for the measurement (e.g. when τ_{qp} is long).

Both readout methods measure the state of the qubit (excited state probability P_e) at a variable delay dt ($dt \ll t$) after a π-pulse following a given delay t after quasiparticle injection. For each t, we then fit P_e as a function of dt to a single exponential function, and hence extract the qubit relaxation rate $\Gamma = 1/T_1$ as a function of t. Representative qubit relaxation curves are shown in Fig. S4. For every $\Gamma(t)$ trace shown in the main text, we use high power readout in the range of 200 μs $\leq t < 8$ ms and dispersive readout for $t > 1$ ms. Data from both methods are combined together to be plotted and fit to Eq. (2), and in overlapping region we find the T_1 measured by both methods agree with each other within statistical uncertainty (Fig. S5).

For both readout methods, the measurement delays,
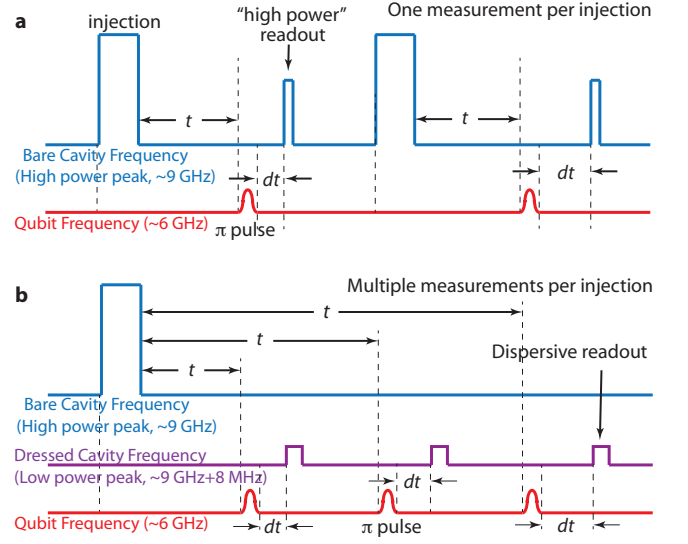
Supplementary Figure S3. **Pulse sequences for qubit T_1 measurements. a**, One measurement per injection using high power readout. **b**, Multiple measurement per injection using dispersive readout.
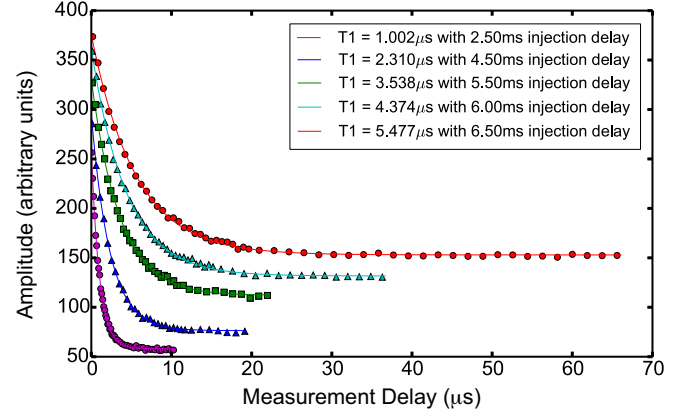
Supplementary Figure S4. **Representative qubit energy decay curves.** Qubit excited state probability P_e as a function of time dt after a π-pulse at a few different times (t) after QP injection measured by high power readout (on device A3 in an aluminum cavity). The curves are vertically offsetted for better visibility in the typical choice of dt's we use for different T_1's.

dt, are chosen to be close to the estimated qubit T_1 under the given condition within an order of magnitude. This is very important in a long automated measurement to maximize data-acquisition efficiency where the T_1 of the qubit varies by three orders of magnitude. Such an estimate is done by an educated initial guess and active feedback on time scale of minutes. When dispersive readout is used, because the readout signals corresponding to qubit $|g\rangle$ and $|e\rangle$ states are relatively stable, we further focus our measurement on the regime $T_1/2 < dt < T_1$ where P_e is maximally sensitive to the T_1 of the qubit
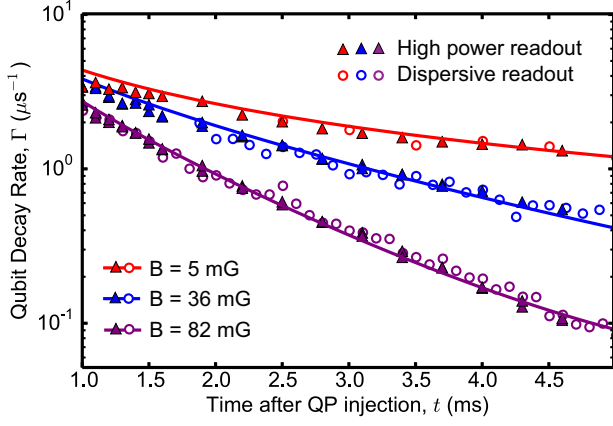
Supplementary Figure S5. **Comparison of qubit decay rates measured by different readout methods.** Qubit decay rate Γ measured by high power read-out (filled triangles) and dispersive read-out (hollow circles), plotted as a function of time t after quasiparticle injection pulses for Device B1 cooled in a few selected magnetic fields B. Data is shown for a limited range of t for a close comparison between the two read-out methods. Solid curves, fit to Eq. (2) of the main text using full range of $\Gamma(t)$ data.

(and periodically calibrate the signals for $|g\rangle$ and $|e\rangle$ states several times per hour).

When the qubit relaxation is dominated by injected quasiparticles, we find the measured T_1 is consistent under the same injection condition over long periods of time (*i.e.* several days). This indicates that our injection is quite stable. This can be seen from the inset of Fig. 3a in the main text where at small t the curves for different cooling fields (measured over the course of two weeks) start from approximately the same qubit decay rate Γ.

SUPPLEMENTARY NOTE 5. HEATING FROM THE INJECTION PULSE

Despite the power and length used for the QP injection pulse, we find the qubit is not heated significantly within the relevant time span of our measurements ($\gtrsim 200~\mu$s after the injection pulse). Since a qubit at an elevated temperature should have a relatively high probability to occupy the excited state due to thermal excitations, we can infer the temperature of the qubit at any given time after the QP injection pulse by measuring the excited state population using a protocol introduced in Ref. 2. Across different samples we find the qubit excited state population stays well below 5% (Fig. S6), indicating the qubit temperature does not exceed 80 mK for all time delays relevant to our measurement. This relatively modest heating effect is consistent with Ref. 3 where the qubit temperature after a high power microwave pulse can be more accurately measured in a fluxonium system. The thermal population of quasiparticles at 80 mK is expected to be many orders of magnitude below the non-
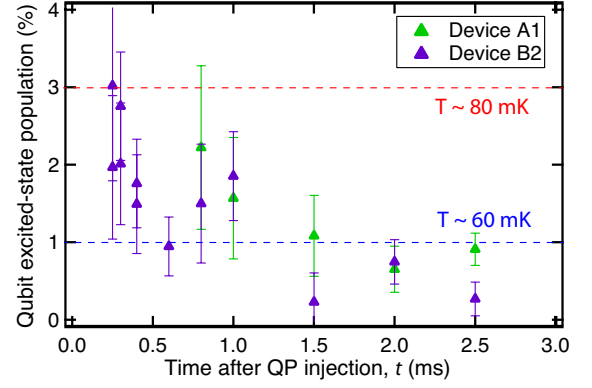
Supplementary Figure S6. **Heating from the quasiparticle injection pulse.** The $|e\rangle$-state population of the qubit is plotted as a function of time after QP injection. The qubit effective temperatures corresponding to excited-state populations of 3% and 1% are about 80 mK and 60 mK respectively. Error bars represent one s.d. statistical uncertainty in extracting the $|e\rangle$-state population from fitting the Rabi oscillations between $|e\rangle$ and the second excited state $|f\rangle$.

equilibrium x_{qp} observed in our device, and we further confirm that T_1 of our qubits shows no measurable decrease when the fridge temperature is raised from 20 mK to 100 mK as was shown in Ref. 4. Therefore the significantly reduced T_1 after the injection pulse is due to the excess amount of quasiparticles instead of an increase of temperature.

We further note that the excited state population of the qubit reflects the energy distribution of the quasiparticles when the qubit decay is dominated by quasiparticle tunneling (which is the case for majority of our measurements). While the density of quasiparticles shortly after the injection (*e.g.* $x_{\mathrm{qp}} \gtrsim 10^{-4}$ at $t = 200~\mu$s) would correspond to a thermal QP density at well over 200 mK, their energy distribution is predominantly near the gap with an effective temperature of no more than 80 mK. We attribute this to the much shorter time scale for quasiparticles to thermalize with the phonon bath (losing their energy in excess of Δ) compared with the time scale of recombination or trapping.

SUPPLEMENTARY NOTE 6. MODEL FOR QUASIPARTICLE DYNAMICS

Here we consider a phenomenological model to account for diffusion, recombination, and trapping of quasiparticles. In the presence of these processes, we write the (normalized) quasiparticle density as a function of both time and spatial coordinate, $x_{\mathrm{qp}}(t, \vec{R})$. Note that $x_{\mathrm{qp}}(t)$ defined in the main text is equal to $x_{\mathrm{qp}}^{\mathrm{JJ}}(t) = x_{\mathrm{qp}}(t, \vec{R} = 0)$ defined here if we choose the position of the Josephson junction as $\vec{R} = 0$.

The quasiparticle density obeys the following equa-

tion[5]:

$$\frac{\partial x_{\text{qp}}}{\partial t} = D\nabla^2 x_{\text{qp}} - r x_{\text{qp}}^2 - s_0 x_{\text{qp}} + g - P \sum_{i=1}^{N} x_{\text{qp}}\, \delta\left(\vec{R} - \vec{R}_i\right)$$

$$(6)$$

where D is the (effective) diffusion constant for quasiparticles, r is the recombination coefficient, s_0 is the trapping coefficient which accounts for small, homogeneous (across the device) residual trapping in the absence of vortices, and g is the generation rate. The factor P which we call "trapping power" accounts for the efficiency of the quasiparticles trapping in the normal core of a vortex; we will comment on the possible microscopic origin of P in Section D below. In a thin-film device, the quasiparticles density is homogeneous across the film's thickness, so $\vec{R}$ is a two-dimensional vector in the plane of the device (hereinafter we assume a simplified geometry depicted in Fig. S7), ∇^2 is the Laplace operator in two dimensions, and $\vec{R}_i$ is the coordinate of the i-th vortex. The total number of vortices trapped in the device $N = N_L + N_R$ is the sum of the number of vortices in the left and right pads.

There is no simple full analytical solution of the above non-linear partial differential equation; however, considerable simplifications are possible in experimentally relevant regimes. Below we first separately consider the cases in which vortices are absent or present in the pads, and then consider the combined effect of recombination and vortex trapping under certain approximations.

A. Evolution of quasiparticle population in the absence of vortices

If there are no vortices, the last term in Eq. (6) is absent. As quasiparticles are injected at the junction, they will initially diffuse away from the junction; then diffusion, recombination, and trapping all act to make the quasiparticle density uniform throughout the device, so at sufficiently long times we can drop the first term on the right hand side of Eq. (6) [see also the next section]. The remaining terms form an ordinary differential equation [also see Eq.(1) of the main text]:

$$\frac{dx_{\text{qp}}}{dt} = -r x_{\text{qp}}^2 - s_0 x_{\text{qp}} + g \tag{7}$$

that determines the time evolution of the quasiparticle density. The general solution of that equation can be written in the form

$$x_{\text{qp}}(t) = x_i \frac{1 - r'}{e^{t/\tau_{\text{ss}}} - r'} + x_0 \tag{8}$$

where $x_i = x_{\text{qp}}(t=0) - x_0$ is the is the normalized density of the injected quasiparticles,

$$x_0 = \frac{\sqrt{s_0^2 + 4gr} - s_0}{2r} \tag{9}$$

is the steady-state density, and

$$\frac{1}{\tau_{\text{ss}}} = 2r x_0 + s_0 \tag{10}$$

is the time constant for the exponential decay at long times $t \gg \tau_{\text{ss}}$. The dimensionless parameter r' (with $0 \leq r' < 1$) quantifies the deviation from a simple exponential decay.

Since the qubit decay rate Γ is proportional to x_{qp} (which is homogenous in this case), by fitting the time dependence of Γ after quasiparticle injection [see Fig. 2 in the main text] with the function

$$\Gamma(t) = C x_i \frac{1 - r'}{e^{t/\tau_{\text{ss}}} - r'} + \Gamma_0 \tag{11}$$

we can estimate (or put bounds on) the parameters entering Eq. (8). Then we can obtain the parameters s, r, and g via the relations

$$r = \frac{r'}{(1 - r')\tau_{\text{ss}} x_i} \tag{12}$$

$$s_0 = \frac{1}{\tau_{\text{ss}}}\left[1 - \frac{2r' x_0}{(1 - r') x_i}\right] \tag{13}$$

$$g = \frac{x_0}{\tau_{\text{ss}}}\left[1 - \frac{r' x_0}{(1 - r') x_i}\right] \tag{14}$$

The recombination constant r extracted from our measurements is in the range $1/(80\text{ ns})$-$1/(170\text{ ns})$ and agrees reasonably well with the theoretical estimate (see subsection "Quasiparticle recombination constant") $r = 21.8/(F\tau_0)$; here $\tau_0 = 438$ ns[7] characterizes the electron-phonon coupling in Al, and F is the phenomenological parameter associated with the phonon trapping and re-absorption in the Al film. Parameter F depends[8] on the phonon transmission coefficient η across the Al-sapphire boundary and on the ratio of the re-absorption length Λ to the Al film thickness d, and ranges from $F = 1/\eta$ at $d \ll \Lambda/4$ to $F = 4d/(\eta\Lambda)$ at $d \gg \Lambda/4$. The properly averaged[8] over the transversal and longitudinal phonons value of η yields $F \approx 5$. There is an uncertainty in the measured value of Λ, with reported values between 100 nm to 350 nm[8]; the factor $4d/\Lambda$ possibly brings an additional factor of ~ 2 in F, resulting in $F = 5 - 10$ quoted in the main text.

Note that in fitting Γ, only an upper bound on x_0 can be placed, $x_0 \leq \Gamma_0/C$, since Γ_0 can receive contributions from other (non-quasiparticle) decay mechanisms. Also taking into account $x_0 \geq 0$, we obtain a bound for s_0:

$$\frac{1}{\tau_{\text{ss}}}\left[1 - \frac{2r'\Gamma_0}{(1 - r') x_i C}\right] \leq s_0 \leq \frac{1}{\tau_{\text{ss}}} \tag{15}$$

For most of our experimental conditions, we find the bounded range for s is comparable or smaller than statistical uncertainties, therefore giving a relatively accurate determination of s_0. We also obtain an upper bound[11] for g:

$$g \leq \frac{(1 - r') x_i}{4r'\tau_{\text{ss}}} = \frac{1}{4r\tau_{\text{ss}}^2} \tag{16}$$

Using the parameters extracted for Device B1, the latter expression gives, $g < 2 \times 10^{-4} \text{s}^{-1}$, in excellent agreement with the value estimated using Eq. (3) in the main text [see also Fig. 4c and 4d], as well as with that recently estimated in an experiment with a fluxonium qubit.[3]

B. Trapping of quasiparticles by vortices

When the qubit is cooled in a sufficiently high magnetic field ($\gtrsim 10$ mG for Type B devices), superconducting vortices are trapped in the pads. The localization of the vortices far from the junction means that at all times the quasiparticle density is not uniform, so that in general the diffusion term in Eq. (6) cannot be neglected. On the other hand, if quasiparticle trapping is the most efficient mechanism for relaxing the quasiparticle density, we can neglect the small effects of generation and recombination and reduce Eq. (6) to

$$\frac{\partial x_{\text{qp}}}{\partial t} = D \nabla^2 x_{\text{qp}} - s_0 x_{\text{qp}} - P \sum_{i=1}^{N} x_{\text{qp}} \, \delta \left(\vec{R} - \vec{R}_i \right) \quad (17)$$

(The injection of quasiparticles is introduced by a time-dependent boundary condition, similar to Ref. 6.)

We can solve Eq. (17) based on a geometric model of the superconducting electrodes shown in Fig. S7 consisting of pads, gap capacitors and connecting wires. This level of complexity is required in order to properly account for the time scale of diffusion when there are unequal number of vortices in the two pads, and the resultant calculation is fairly complicated. The goal of the following calculation is to solve the spatial mode with the slowest decay rate s, that is a certain spatial distribution of x_{qp} that decays synchronously across the device, i.e. $x_{\text{qp}}(\vec{R}, t) = x_{\text{qp}}(\vec{R}, t = 0)e^{-st}$. One instructive limiting case to see the effect of vortices is when the diffusion time ($\sim L^2/D$) is much shorter than the decay time ($1/s$) so that the spatial gradient of x_{qp} in Eq. (17) can be neglected. Then by integrating Eq. (17) over the area A of the device one can immediately see: $sA = s_0 A + NP$. Therefore the decay rate multiplied by the device area is the macroscopic observable directly linked to the trapping power we defined. More generally, we find for any given number of vortices N_L and N_R on the two pads, this decay rate s can be numerically computed from Eqs. (37)-(39) derived below and compared with our measured data as shown in Fig. S9.

Our geometric model of the electrodes is composed of two square pads of area S_{pad}, each connected to the gap capacitor by a narrow wire of length L and width W, with $W \ll L$. Due to their narrow widths, we treat each wire as one-dimensional. The quasiparticle density, normalized by the density of Cooper pairs,[13] is denoted as $x_{\text{qp}}^L(t, y)$ [$x_{\text{qp}}^R(t, y)$] in the wire to the left [right] of the capacitor; the corresponding coordinate y runs from $-L$ to 0 [0 to L]. As a simple model for the gap capacitor, we use a thin wire (width W) extended to length h on both sides of the central wire; below we will discuss a more realistic model for the gap capacitor. We assume symmetry with respect to the horizontal line connecting the centers of the pads; then we only need to introduce two densities for the left and right upper halves of the capacitor, denoted by $x_{\text{qp}}^{VL}(t, y)$ and $x_{\text{qp}}^{VR}(t, y)$, respectively. Here coordinate y runs from 0 at the top to h at the junction with the horizontal wire. The two plates of the gap capacitors are at a distance $2l$ from each other, with the junction placed at the center of a short and thin wire ($l \ll h, L$). The quasiparticle density in this central wire is denoted by $x_{\text{qp}}^C(t, y)$, with the spatial coordinate $y \in [-l, l]$.

Since the qubit is affected only by quasiparticles in the junction vicinity, we are interested in the time evolution of $x_{\text{qp}}^C(t, 0)$. To find the decay rate of $x_{\text{qp}}^C(t, 0)$, we start by noting that except in the pads, the quasiparticle densities obey a simple diffusion equation [cf. Eq. (17)]

$$\frac{\partial x_{\text{qp}}^i}{\partial t} = D \frac{\partial^2 x_{\text{qp}}^i}{\partial y^2} - s_0 x_{\text{qp}}^i \quad (18)$$

with $i = L$, R, C, VL, or VR. The general solution to this equation can be written in the form

$$x_{\text{qp}}^i(t, y) = e^{-st} \left[\alpha^i \cos ky + \beta^i \sin ky \right], \quad (19)$$

where s and k are related by

$$s = Dk^2 + s_0. \quad (20)$$

Here we are interested in the spatial mode with the lowest decay rate; it is obtained by finding the smallest k that satisfies the boundary conditions which we now discuss.

First, we require that no current leaves the top (and bottom, by symmetry) of the vertical wires:

$$\left. \frac{\partial x_{\text{qp}}^{Vj}}{\partial y} \right|_{y=0} = 0 \quad (21)$$

where, from now on, $j = L$ or R. These conditions imply $\beta^{Vj} = 0$. Next, we require continuity of density at the two cross points:

$$x_{\text{qp}}^C(\pm l) = x_{\text{qp}}^j(0) = x_{\text{qp}}^{Vj}(h), \quad (22)$$

where in the first term the positive sign is to be used for $j = R$ and the negative one for $j = L$. From the last equality we find immediately
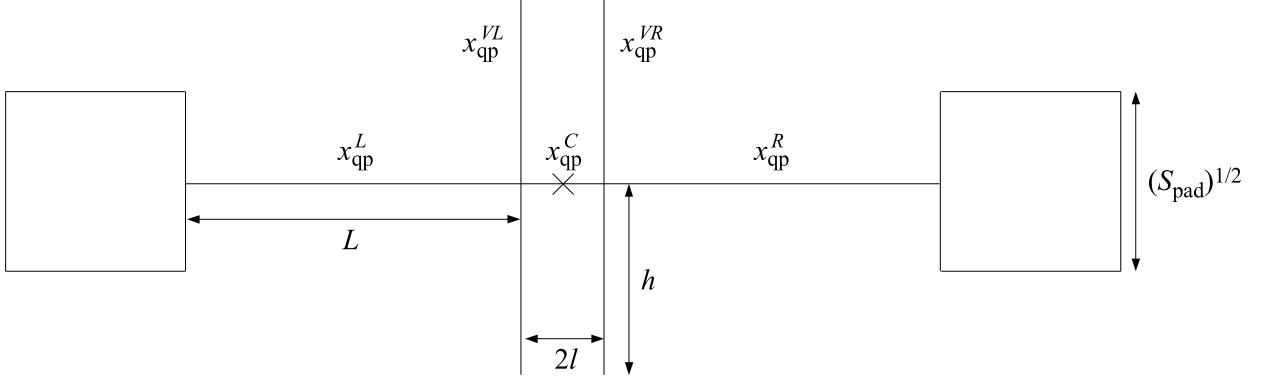
$$\alpha^{Vj} = \alpha^j / \cos kh. \quad (23)$$

Taking the sum and difference of the first equality in Eq. (22) with $j = R$ and $j = L$ gives

$$\alpha^C = \left(\alpha^R + \alpha^L \right) / 2 \cos kl, \quad \beta^C = \left(\alpha^R - \alpha^L \right) / 2 \sin kl. \quad (24)$$

We then require current conservation at the two cross points:

$$\left. \frac{\partial x_{\text{qp}}^L}{\partial y} \right|_{y=0} + 2 \left. \frac{\partial x_{\text{qp}}^{VL}}{\partial y} \right|_{y=h} = \left. \frac{\partial x_{\text{qp}}^C}{\partial y} \right|_{y=-l} \quad (25)$$

Supplementary Figure S7. **Simplified geometry used to model the Type B devices.** The thin lines represent wires of width W much smaller than their lengths. See Fig. S8 and the text after Eq. (37) for a more realistic model of the vertical capacitor pads.

for the left-side cross point and

$$\frac{\partial x_{\rm qp}^{C}}{\partial y}\bigg|_{y=l} + 2\frac{\partial x_{\rm qp}^{VR}}{\partial y}\bigg|_{y=h} = \frac{\partial x_{\rm qp}^{R}}{\partial y}\bigg|_{y=0} \qquad (26)$$

for the right one. Using the results above, these conditions can be written explicitly as

$$\beta^{L} - 2\alpha^{L}\tan kh = \frac{\alpha^{R} + \alpha^{L}}{2}\tan kl + \frac{\alpha^{R} - \alpha^{L}}{2}\cot kl \qquad (27)$$

$$-\frac{\alpha^{R} + \alpha^{L}}{2}\tan kl + \frac{\alpha^{R} - \alpha^{L}}{2}\cot kl - 2\alpha^{R}\tan kh = \beta^{R} \qquad (28)$$

and enable us to express $\beta^{L,R}$ in terms of $\alpha^{L,R}$.

We note that the boundary conditions discussed so far do not involve the pads, and let us write the densities in the left and right wires in the explicit forms

$$x_{\rm qp}^{L} = e^{-st}\bigg[\alpha^{L}\cos ky + \bigg(\frac{\alpha^{R} + \alpha^{L}}{2}\tan kl \qquad (29)$$
$$+\frac{\alpha^{R} - \alpha^{L}}{2}\cot kl + 2\alpha^{L}\tan kh\bigg)\sin ky\bigg]$$

$$x_{\rm qp}^{R} = e^{-st}\bigg[\alpha^{R}\cos ky + \bigg(-\frac{\alpha^{R} + \alpha^{L}}{2}\tan kl \qquad (30)$$
$$+\frac{\alpha^{R} - \alpha^{L}}{2}\cot kl - 2\alpha^{R}\tan kh\bigg)\sin ky\bigg]$$

where only the two coefficients α^{L} and α^{R} are yet to be determined. In the pads, we assume the diffusion time $S_{\rm pad}/D$ to be the shortest time scale (in particular, shorter than the inverse of the average trapping rate

$N_{L,R}P/S_{\rm pad}$), so that the densities in the pads, $x_{\rm qp}^{Pj}$, can be taken as uniform; the value of trapping power P we ultimately extract from the experiment justify this assumption. By integrating Eq. (6) over, *e.g.*, the right pad, we find

$$S_{\rm pad}\frac{\partial x_{\rm qp}^{PR}}{\partial t} = I_{W} - S_{\rm pad}s\,x_{\rm qp}^{PR} - N_{R}P\,x_{\rm qp}^{PR}, \qquad (31)$$

where I_{W} represent the quasiparticle current going from the pad into the wire. By current conservation we also have

$$I_{W} = -W D\frac{\partial x_{\rm qp}^{R}}{\partial y}\bigg|_{y=L}, \qquad (32)$$

while continuity requires $x_{\rm qp}^{PR} = x_{\rm qp}^{R}|_{y=L}$. Therefore, we arrive at the following boundary condition at position $y = L$:

$$\frac{\partial x_{\rm qp}^{R}}{\partial y} = -\frac{1}{WD}\bigg[S_{\rm pad}\bigg(\frac{\partial x_{\rm qp}^{R}}{\partial t} + sx_{\rm qp}^{R}\bigg) + N_{R}P\,x_{\rm qp}^{R}\bigg]\bigg|_{y=L}. \qquad (33)$$

Similarly, at the boundary with the left pad ($y = -L$) we have:

$$\frac{\partial x_{\rm qp}^{L}}{\partial y} = \frac{1}{WD}\bigg[S_{\rm pad}\bigg(\frac{\partial x_{\rm qp}^{L}}{\partial t} + sx_{\rm qp}^{L}\bigg) + N_{L}P\,x_{\rm qp}^{L}\bigg]\bigg|_{y=-L}. \qquad (34)$$

Substituting Eqs. (29) and (30) into Eqs. (33)-(34), and requiring the existence of a non-trivial solution, leads to an equation for k. Introducing dimensionless variable $z = kL$, this equation is:

$$\left[z\left(\tan z+f(z)\right)+\left(az^2-\bar{N}\frac{P\tau_D}{A_W}\right)\left(1-\tan z f(z)\right)\right]^2-\left(\frac{\Delta N}{2}\frac{P\tau_D}{A_W}\right)^2\left(1-\tan z f(z)\right)^2$$

$$-\frac{1}{4}\left[\tan\left(z\frac{l}{L}\right)+\cot\left(z\frac{l}{L}\right)\right]^2\left\{\left[z-\tan z\left(az^2-\bar{N}\frac{P\tau_D}{A_W}\right)\right]^2-\left(\frac{\Delta N}{2}\frac{P\tau_D}{A_W}\right)^2\tan^2 z\right\}=0 \tag{35}$$

with $\tau_D=L^2/D$ the diffusion time along the left and right wires, $A_W=LW$ the area of the wires (from a cross point to a pad), $a=S_{\mathrm{pad}}/A_W$, $\bar{N}=(N_L+N_R)/2$ the average number of vortices, $\Delta N=N_R-N_L$, and

$$f(z)=\frac{1}{2}\tan\left(z\frac{l}{L}\right)-\frac{1}{2}\cot\left(z\frac{l}{L}\right)+2\tan\left(z\frac{h}{L}\right). \tag{36}$$

We can simplify Eq. (35) as follows: even at large number of vortices, the variable k being as small as possible means that it is at most of order $1/L$, i.e., z is at most of order 1. Then we note that for the short wire inside the gap capacitor we have $l\ll L$ and thus, keeping only the leading term, in the limit $l\to0$ we find

$$\left[z-\tan z\left(az^2-\bar{N}\frac{P\tau_D}{A_W}\right)\right]\left[z\left(\tan z+2\tan\left(z\frac{h}{L}\right)\right)\right.$$
$$+\left(az^2-\bar{N}\frac{P\tau_D}{A_W}\right)\left(1-2\tan\left(z\frac{h}{L}\right)\tan z\right)\right]$$
$$+\left(\frac{\Delta N}{2}\frac{P\tau_D}{A_W}\right)^2\tan z\left(1-2\tan\left(z\frac{h}{L}\right)\tan z\right)=0. \tag{37}$$

As an improvement to our description of the device, we further model the gap capacitor by adding to the thin wires of width W and length h, wider sections of width W_c and length L_c, as shown in Fig. S8.
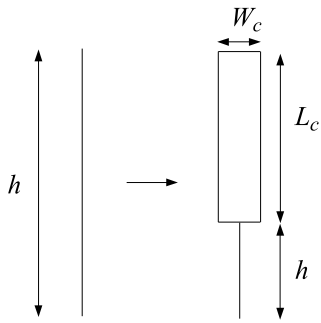
Supplementary Figure S8. **Improved model for the capacitor of the Type B devices.** To the the thin wire of width W and length h on the left, we add a wider wire of width W_c and length L_c to form the structure on the right.

Requiring again current conservation and continuity of the density, we find that this more realistic geometry can be incorporated into the formulas above by the substitu-

tion

$$\tan\left(z\frac{h}{L}\right)\to\frac{\cos\left(z\frac{L_c}{L}\right)\sin\left(z\frac{h}{L}\right)+\frac{W_c}{W}\sin\left(z\frac{L_c}{L}\right)\cos\left(z\frac{h}{L}\right)}{\cos\left(z\frac{L_c}{L}\right)\cos\left(z\frac{h}{L}\right)-\frac{W_c}{W}\sin\left(z\frac{L_c}{L}\right)\sin\left(z\frac{h}{L}\right)}. \tag{38}$$

Equation (37) with the substitution in Eq. (38) makes it possible to calculate the expected decay rates $1/\tau_{\mathrm{ss}}$ of the quasiparticle density based on the geometry of each device. Indeed, given the values of trapping power P and diffusion constant D (which determines the diffusion time τ_D), and the number of vortices in each pad, we can solve Eqs. (37)-(38) for z and then find s from

$$s=\frac{1}{\tau_D}z^2+s_0. \tag{39}$$

Equations (37)-(38) can be solved analytically only in limiting cases. For example, in the case of small trapping power $P\to0$, the leading order result for the decay rate is

$$s\simeq(N_L+N_R)\frac{P}{A}+s_0 \tag{40}$$

where A is the total area of the device. This expression explains why, after subtracting the device-dependent homogeneous trapping coefficient s_0 and multiplying by the total device area, the decay rates for the different devices are approximately equal and quantized. In the opposite limit of large trapping power (and/or large number of vortices), we find that when the area of the capacitor $A_c=2(L_cW_c+hL)$ is small compared to that of the central wire, $A_c\ll A_W$, the variable z to be substituted in Eq. (39) is given by

$$z\simeq\frac{\pi}{2}\frac{1}{1+A_c/A_W}. \tag{41}$$

This formula qualitatively explains why in Fig. 3b of the main text the decay rate of Device B2 (with a smaller gap capacitor) saturates at a higher value than the decay rate of Device B1 (with a bigger gap capacitor).

For a quantitative comparison with experiments, we numerically solve Eqs. (37)-(38) to find z, and then substitute this number into Eq. (39). In Fig. S9 we show with dashed lines the results of such calculations with $P=0.067$ cm^{2}s^{-1}, $D=18$ cm^{2}s^{-1}, and the number of vortices in each pad given to the right of each panel; the homogeneous trapping coefficient s_0 has been subtracted out to facilitate comparison among the devices. The results calculated with the same P and D at higher magnetic field has been plotted as the dashed line in Fig. 3b
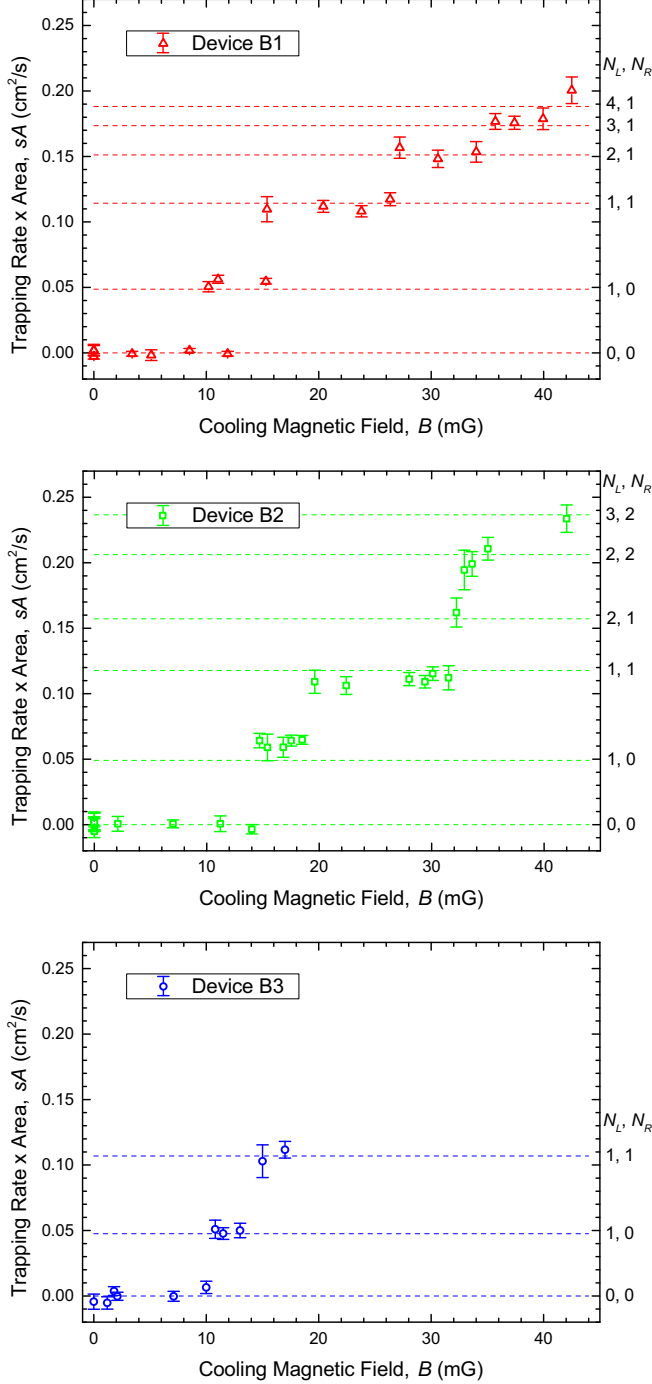
Supplementary Figure S9. **Fitting the discrete steps in quasiparticle trapping rates**. Normalized trapping rates for (top to bottom) devices B1, B2, and B3 as a function of cooling magnetic field, with error bars showing the range of fluctuations for repetitive measurements over each thermal cycle. Horizontal dashed lines: rates calculated as explained in the text for the number of vortices in the two pads listed by the two numbers to the right of each line. *e.g.* 21 represents $N_L = 2$, $N_R = 1$ or vice versa.

of the main text, where N_L and N_R are assumed to be equal and increasing linearly with magnetic field. The diffusion constant D is determined by the saturation level of s and the geometry of the device [see Eqs. (39) and (41) and $\tau_D = L^2/D$], and its uncertainty is determined to be $\sim 10\%$ (or ± 2 cm^{2}s^{-1}) based on the statistical fluctuation of s (Fig. 3b of the main text). (Note that this does not include possible error due to the simplified model of the actual device geometry, which by our estimate should not increase the overall uncertainty significantly.) The trapping power P is mostly determined by the height of the first 2-3 steps in sA, slightly adjusted for the finite speed of diffusion [see Eq. (40)]. We determine $\sim 7\%$ (or ± 0.005 cm^{2}s^{-1}) uncertainty on P based on the statistical fluctuation of s for the $(N_L, N_R) = (1,0)$ and $(1,1)$ states ($\sim 7\%$) and uncertainty propagated from D ($\sim 2\%$).

We note that the agreement between the predicted steps in sA and the measured data (Fig. S9) could be further improved by considering vortices entering the device two at a time (using the vortex number series $(N_L, N_R) = (0,0), (1,1), (2,2)$, etc. for the first few steps). Under this interpretation the trapping power for a single vortex would be a factor of two smaller, *i.e.* 0.034 cm^{2}s^{-1}, since each step change in sA corresponds to the entry of two vortices. While we cannot entirely exclude this possibility, we argue it is an unlikely scenario since such a high degree of symmetry seems unwarranted. The actual critical field B_k for the two trapping pads in each device should vary from each other to a similar degree as they vary from device to device, typically a few mG, which is larger than the $\lesssim 1$ mG increments we typically use in changing cooling magnetic field near B_k. Furthermore, variation of a few mG in B_k is consistent with the width of the first step in sA (labeled "10" in Fig. S9). The second step (labeled "11" in Fig. S9) being much wider than the first step also suggests a vortex number series of $(N_L, N_R) = (0,0), (1,0), (1,1), (2,1)$, etc. (plotted in Fig. S9 and assumed in the main text) as the more plausible scenario.

Without changing the assumption on vortex numbers, the agreement between theory and experiment could also be improved, especially for Device B2, by introducing a non-homogenous background trapping rate concentrated in one of the pads instead of using a homogenous trapping coefficient s_0. The origin of the background trapping rate remains to be studied in the future.

C. Evolution of quasiparticle population in the presence of vortex trapping and recombination

Now that we have analyzed the quasiparticle population dynamics in the presence of either recombination or trapping from vortices, here we attempt to approximately combine both results.

In the previous section we have solved the slowest linear decay mode in the presence of vortices (neglect-

ing quasiparticle recombination and generation) so that $x_{qp}(\vec{R}, t) = x_{qp}^{JJ} e^{-st} \psi(\vec{R})$, where $x_{qp}^{JJ} = x_{qp}(\vec{R} = 0)$ is the normalized quasiparticle density near the Josephson junction (simply quoted as x_{qp} in the main text), $\psi(\vec{R})$ describes the spatial mode (normalized to $\psi(\vec{R} = 0) = 1$) satisfying

$$D\nabla^2 \psi(\vec{R}) - s_0 \psi(\vec{R}) - P \sum_{i=1}^{N} \psi(\vec{R}) \delta\left(\vec{R} - \vec{R}_i\right) = -s\psi(\vec{R})$$
(42)

where s is the decay rate of the mode obtained from Eqs. (37)-(39).

Now we add recombination and generation to the dynamics assuming they are relatively weak, so the spatial mode is not altered significantly and x_{qp} can still be factorized as $x_{qp}(\vec{R}, t) = x_{qp}^{JJ}(t) \psi(\vec{R})$. Under this approximation we can use the linear decay term with rate s [the RHS of the Eq. (42)] to replace the diffusion and trapping terms [the LHS of the Eq. (42)] in Eq. (6) and get:

$$\frac{dx_{qp}^{JJ}}{dt} = -r(x_{qp}^{JJ})^2 - s x_{qp}^{JJ} + g$$
(43)

This is the ordinary differential equation for quasiparticle dynamics we have used in the main text [Eq. (1)]. We can then follow the same procedures in Section A to derive the relations between fit parameters of Eq. (2) of the main text and r, s and g, except that s_0 in Section A should be replaced by s. Therefore the trapping rates s extracted in the main text based on this equation in the presence of vortices corresponds to the linear decay rate solved in Section B.

D. Microscopic model for trapping power

In modelling QP trapping by a vortex, we treated the latter as a point-like object, see Eq. (6). In a more realistic (albeit still crude) model, we can expect trapping of quasiparticles to be effective over a finite region around the vortex core, and the trapping rate to be related to the quasiparticle relaxation by electron-electron (ee) and electron-phonon (ep) interactions. This can be taken into account via the replacement

$$P\delta(\vec{R} - \vec{R}_i) \rightarrow \frac{1}{\tau_n} \theta\left(R_c - \left|\vec{R} - \vec{R}_i\right|\right)$$
(44)

in Eq. (6). Here R_c is the radius of the vortex core modeled as a normal-state disk (R_c is of the order of the coherence length ξ), $\theta(R)$ is the Heaviside step function, and $1/\tau_n = 1/\tau_{ee} + 1/\tau_{ep}$ is the microscopic electron relaxation rate. After integrating Eq. (44) over the pad area, we can relate the trapping power P to these parameters:

$$P = \frac{1}{\tau_n} \pi R_c^2.$$
(45)

Estimates of the two contributions to $1/\tau_n$ can be performed using the theory of electron-electron interactions in disordered conductors[14] and the conventional theory of electron-phonon interaction in metals.[15] The material-specific input parameters (electron density of states and the deformation potential) for Al are known from literature, see, e.g., Ref. 7, and we know the parameters specific for our devices (the film thickness and diffusion constant) from the performed measurements. We find that the electron-electron and electron-phonon mechanisms contribute comparably to $1/\tau_n$ at a level $1/\tau_{ee} \approx 6 \times 10^6 \, \mathrm{s}^{-1}$, although it should be stressed that these rates are not accurately known[17]. There is also some uncertainty about the coherence length ξ, but the "dirty-limit" estimate $\xi \approx 100$ nm seems reasonable for a ~ 80 nm-thick film. Evaluation of P using Eq. (45) and assuming $R_c = \xi$ yields $P \approx 0.004 \, \mathrm{cm}^2\mathrm{s}^{-1}$, about 18 times lower than found in the experiment. However, the assumption $R_c = \xi$ may be underestimating the effective area where quasiparticles can be trapped. Indeed, a treatment of gap suppression near a vortex[16] gives $R_c \approx 2.7\xi \approx 270$ nm and using this estimate [also employed in a recent work, Ref. 9] we find P within a factor of 3 of the measured value. Also, it is not clear whether the different group velocities of quasiparticles in the superconductor and in the normal core of the vortex should affect the effective quasiparticle trapping rate[10]. Given the simplified nature of the model and the uncertainties about the values of parameters entering it, the disagreement between measured and estimated trapping powers by a factor of 3 to 18 is not surprising.

Recently, Nsanzineza and Plourde[9] used a similar phenomenological model of quasiparticle trapping from vortices to interpret an observed increase in quality factor of a superconducting resonator as a function of magnetic field. While Ref. 9 suggested their data is consistent with a microscopic electron relaxation rate of about $3 \times 10^6 \mathrm{s}^{-1}$, (similar to our estimate above,) we note the experiment can be better interpreted in terms of trapping power so that it is independent of microscopic assumptions. The implied trapping power $P \approx 0.024 \, \mathrm{cm}^2\mathrm{s}^{-1}$ is within a factor of 3 of our measured value, which is a decent agreement given that Ref. 9 does not have a direct probe of quasiparticle dynamics, and the extracted P (or τ_n) inevitably depends on a list of unknown parameters such as recombination rate, diffusion constant, generation rate, kinetic inductance fraction, etc. The quantization of trapping rates we have observed is a much more direct and accurate measurement of the quasiparticle trapping effect of a vortex, and should facilitate development of more quantitative microscopic theories addressing the vortex-quasiparticle interaction in superconductors.

E. Quasiparticle density distribution in the presence of vortices

Although the microscopic mechanisms remain to be further explored, the measured trapping power P gives us a quantitative understanding of the overall strength of the vortex-quasiparticle interaction. One important conclusion we can draw from our measured ratio of $P/D \approx 4 \times 10^{-3} \ll 1$ (based on $P = 0.067$ cm^{2}s^{-1}, $D = 18$ cm^{2}s^{-1}) is that the spatial distribution of QP density in a 2D extended film is close to homogenous even in the presence of vortices. This approximate homogeneity has been assumed for the quasiparticle density in each pad, x_{qp}^{Pj}, in subsection VI-B, and here we provide a justification.

We first consider one vortex at $\vec{R} = 0$ in a film that is extended in 2D for $|\vec{R}| < r$ with no other quasiparticle loss mechanisms within this area. We assume there is no QP generation source in the entire system, so there exists a slowest decay mode so that $x_{\mathrm{qp}}(\vec{R},t) = x_{\mathrm{qp}}(\vec{R})e^{-st}$, and

$$D\nabla^2 x_{\mathrm{qp}}(\vec{R}) - \frac{P}{\pi R_c^2} x_{\mathrm{qp}}(\vec{R})\,\theta(R_c - |\vec{R}|) = -s x_{\mathrm{qp}}(\vec{R}) \tag{46}$$

Here for simplicity we use a step function to model the trapping power as was done in the previous subsection, but the subsequent conclusion is qualitatively independent of this assumption. The value of s depends on the geometry and loss mechanisms[12] in the region of $|\vec{R}| > r$, but our analysis only relies on $s \geq 0$ which does not require knowledge of the $|\vec{R}| > r$ region. Equation (46) can be simplified in cylindrical coordinates (ρ, θ), considering rotational symmetry when $\rho \ll r$, $x_{\mathrm{qp}}(\vec{R}) = x_{\mathrm{qp}}(\rho)$,

$$D\frac{1}{\rho}\frac{d}{d\rho}\left[\rho\frac{dx_{\mathrm{qp}}}{d\rho}\right] = \frac{P}{\pi R_c^2} x_{\mathrm{qp}}\,\theta(R_c - \rho) - s x_{\mathrm{qp}} \tag{47}$$

Noting $s \ll P/(\pi R_c^2)$, we solve Eq. (47) to first order in P/D for $\rho \leq R_c$:

$$\frac{x_{\mathrm{qp}}}{x_{\mathrm{qp}}^0} \approx 1 + \frac{P}{4\pi D}\left(\frac{\rho}{R_c}\right)^2 < 1.001 \tag{48}$$

where $x_{\mathrm{qp}}^0 = x_{\mathrm{qp}}(\rho = 0)$ is the QP density at the center of the vortex. Noting $s \geq 0$, we solve Eq. (47) for $R_c < \rho < r$:

$$\frac{x_{\mathrm{qp}}}{x_{\mathrm{qp}}^0} \leq 1 + \frac{P}{2\pi D}\left[\frac{1}{2} + \ln\left(\frac{\rho}{R_c}\right)\right] < 1.01 \tag{49}$$

where we assume $r \sim 1$ μm to 1 mm as the practical (typical) length scale for an area of a film trapping a single vortex ($r \approx 80$ μm for the pads in Type B devices) and $R_c \sim \xi \sim 100$ nm. Therefore the QP density dip at the position of the vortex is no more than $\sim P/D$ (or 1% at most) in relative depth. Due to the weak logarithmic dependence on length scales, this conclusion is very insensitive to the microscopic model of vortices.

In the case of multiple vortices in the pad, as long as magnetic field and the quality of the film are roughly homogeneous, vortices are evenly distributed in the pad. At all magnetic fields one can divide the pad into sub-areas each containing one vortex and has aspect ratio of order of unity, and the spatial distribution of quasiparticle remains approximately uniform for the entire pad.

This approximate homogeneity of the QP density in extended 2D geometries is a direct consequence of the much shorter diffusion time scale L^2/D than the trapping time scale $A/(NP)$ (where L is the typical length between an arbitrary point of the superconducting film and its nearest vortex, and A is the total area of the film). In extended 2D geometries, $L^2 \sim A/N$. However, in quasi-1D geometries such as the thin connection wire between the pads in our Type B devices, it is possible to have $L^2 \gg A/N$, resulting in $L^2/D \gtrsim A/(NP)$. Therefore large gradient of quasiparticle density can be present along the wire (when the total trapping power in the pads is sufficiently large), while the distribution of quasiparticles in the pads remains approximately homogeneous.

SUPPLEMENTARY NOTE 7. TRANSMON FREQUENCY SHIFT DUE TO QUASIPARTICLES

In the main text we relied on the proportionality between qubit decay rate and quasiparticle density to study the time evolution of the latter. Indeed, neglecting other relaxation mechanisms, the qubit decay rate Γ due to low-energy quasiparticles is

$$\frac{\Gamma}{\omega} = x_{\mathrm{qp}}\frac{1}{\pi}\sqrt{\frac{2\Delta}{\hbar\omega}}\,, \tag{50}$$

where ω is the qubit frequency and x_{qp} the normalized quasiparticle density. Interestingly, the presence of quasiparticles also results in a change in the qubit frequency, as derived in Refs. 18 and 13 and measured in a phase qubit in Ref. 19 and a transmon in Ref. 4. Specifically, for a single-junction transmon the frequency change $\delta\omega$ due to low-energy quasiparticles is related to their density by

$$\frac{\delta\omega}{\omega} = -\frac{1}{2}x_{\mathrm{qp}}\left[\frac{1}{\pi}\sqrt{\frac{2\Delta}{\hbar\omega}} + 1\right]\,. \tag{51}$$

The first term in square brackets originates from quasiparticle tunneling events and is a manifestation of fluctuation-dissipation relations: it has the same cause as the one leading to the qubit decay, see Eq. (50). The second term, by contrast, is due to the suppression of the superconducting gap in the presence of quasiparticles

Taken together, Eqs. (50) and (51) predict a simple relation between $\delta\omega$ and Γ:

$$\delta\omega = -\frac{1}{2}\Gamma\left[1 + \pi\sqrt{\frac{\hbar\omega}{2\Delta}}\right]\,. \tag{52}$$
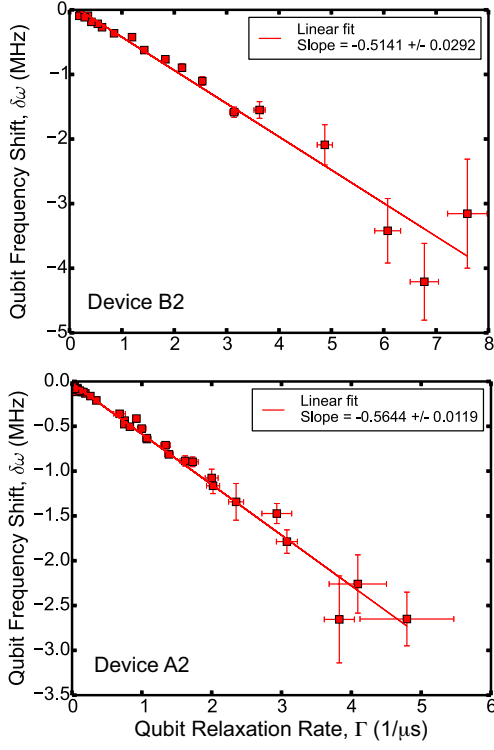
Supplementary Figure S10. **Transmon frequency shifts after quasiparticle injection.** Angular frequency shift $\delta\omega$ of two transmon devices after variable delays after quasiparticle injection pulse, plotted as a function of qubit relaxation rate, Γ after the same delay. Error bars represent one s.d. statistical uncertainties based on fitting qubit relaxation curves (for Γ) and Ramsey oscillation (for $\delta\omega$).

This relationship can be checked experimentally using our technique: after injecting quasiparticles, we can obtain the transition rate Γ by measuring the energy relaxation time T_1 of the qubit, as well as the change in its frequency via a T_2 Ramsey experiment. As shown in Fig. S10, our measurements in two devices display the expected proportionality between $\delta\omega$ and Γ. However, the slopes of the best-fit lines are lower than those predicted by Eq. (52) by a factor of approximately 1.7.

SUPPLEMENTARY NOTE 8. LOSSES DUE TO THE VORTEX FLOW RESISTANCE

Along with the beneficial effect of trapping nonequilibrium particles, vortices may cause electromagnetic losses

in a circuit, if microwaves excite the vortex motion. However, the current excited by the transmon electromagnetic mode in the vortex-trapping pads is negligible, due to the location of the pads in Type B devices. Therefore, the vortex-flow loss is minimized. This is evidenced by the relatively flat qubit T_1 vs. B dependence for a large range of magnetic field (40 mG $< B <$ 200 mG) despite increasing number of vortices. However, at even higher field ($B \gtrsim 200$ mG) qubit T_1 starts decreasing, which is indicative of losses due to vortices penetrating the gap capacitors where the current density is high.

We have also studied how the relaxation time, T_1, of type A qubits varies with the cooling magnetic field. For two of such qubits (device A1 and A2), we observe no significant changes in T_1 at small cooling fields ($\lesssim$ 30mG) and substantial decrease in T_1 at large cooling magnetic fields ($\gtrsim$ 30mG) (Fig. S11). Because Type A devices have very large electrodes and most likely have trapped vortices when cooled in nominally zero magnetic field, increasing magnetic field (in all regimes of B) leads to additional vortices throughout the entire big pads where current density can be either large or small depending on the specific locations. The decrease in T_1 is consistent with additional vortex-flow loss due to increasing number of vortices, and the relatively flat T_1 at low magnetic field can be explained as the combined effects of a small enhancement of quasiparticle trapping and a small additional vortex-flow loss.
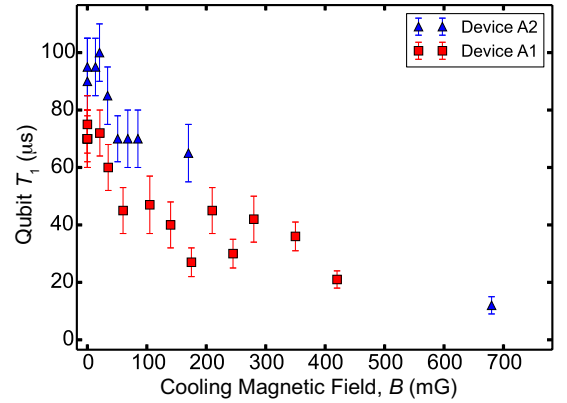
Supplementary Figure S11. **Cooling magnetic field dependence of Type A qubit T_1.** Qubit relaxation time, T_1 as a function of cooling magnetic field for device A1 and A2. The magnetic field for A2 may be overestimated by up to 50% in overall scale due to inhomogeneous field for the particular magnetic field coil configuration used for that cool-down. Error bars represent one s.d. of fluctuations for repetitive measurements (over several hours to one day) within each cycle.

* chen.wang@yale.edu

[1] M. D. Reed, L. DiCarlo, B. R. Johnson, L. Sun, D. I. Schuster, L. Frunzio, and R. J. Schoelkopf, *Phys. Rev. Lett.*, **105** 173601 (2010).

[2] K. Geerlings, Z. Leghtas, I. M. Pop, S. Shankar, L. Frunzio, R. J. Schoelkopf, M. Mirrahimi, and M. H. Devoret, *Phys.*

Supplementary Table S1. **Summary of the parameters of measured transmon devices.** Repeated cavity entries represent cavities containing two qubits. Repeated qubit entries represent qubits that have been sequentially mounted into different cavities in different experiment runs (fridge cool-downs). Quantities unmeasured are left blank. Note not all devices have been measured under varies cooling magnetic field due to the time-consuming nature of thermal cycling.

Qubit	Cavity	Cavity Freq (GHz)	Qubit Freq (GHz)	E_C[a] (MHz)	E_J[b] (GHz)	$\tau_{\text{ss}(B\sim0)}$ (ms)	$1/r_{(B\sim0)}$ (ns)	$T_{1(B\sim0)}$ (μs)	$T_{1(B\sim100mG)}$ (μs)
A1	Al1	9.18639	6.01059	341.0	14.79	1.5	105	75[c]	50[c]
A2	Al2	9.08685	6.06365	341.8	15.01	0.25	NA[d]	95	65
A3	Al3	8.33612	6.44094	339.0	16.95	1.6	130	75	
A4	Al3	8.33612	6.29598	335.0	16.41	2.8		70	
B1	Al4	9.26600	5.71200	291.7	15.45	18	170	9.5	25
B2	Al5	9.22704	5.75975	285.8	15.99			8.0	19
B2	Cu1	9.23181	5.67777	287.1	15.49	10	160	7.5	16
B3	Al6	9.24689	5.88310	294.6	16.19			20	30
B3	Cu2	9.24613	5.89244			6.5	90	18	
B4	Cu3	9.12054	6.25856			4.7	85	16	
B5	Cu3	9.12054	6.61310			13.5	80	8.5	

[a] E_C is measured by the transmon anharmonicity (frequency difference between the $|g\rangle \to |e\rangle$ and $|e\rangle \to |f\rangle$ transitions).
[b] E_J is calculated based on the transmon frequency relation $\omega_q = \sqrt{8E_J E_C} - E_C$.
[c] These are measured during a different cool-down from the data shown in Fig. 2 of the main text (where the average zero-field T_1 for the same device is slightly higher).
[d] We are unable to quantitatively determine r due to extremely strong trapping.

Rev. Lett. **110**, 120501(2013).

[3] U. Vool, I. M. Pop, K. Silwa, B. Abdo, C. Wang, T. Brecht, Y. Y. Gao, S. Shankar, M. Hatridge, G. Catelani, M. Mirrahimi, L. Frunzio, R. J. Schoelkopf, L. I. Glazman, and M. H. Devoret, *arXiv*:1406.1769 [quant-ph] (2014).

[4] H. Paik, D.I. Schuster, L.S. Bishop, G. Kirchmair, G. Catelani, A. P. Sears, B. R. Johnson, M. J. Reagor, L. Frunzio, L. I. Glazman, S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf, *Phys. Rev. Lett.* **107**, 240501 (2011).

[5] Also see Eq. (1) of Ref. 9 which is the steady-state version of this equation, and Eq. (1) of Ref. 10 which considers a special case with uniformly and continuously distributed vortices.

[6] S. Rajauria, H. Courtois, and B Pannetier, *Phys. Rev. B* **80**, 214521 (2009).

[7] S. B. Kaplan, C. C. Chi, D. N. Langenberg, J. J. Chang, S. Jafarey, and D. J. Scalapino, *Phys. Rev. B* **14** 4854 (1976).

[8] S. B. Kaplan, *Journal of Low Temperature Physics* **37**, 343 (1979).

[9] I. Nsanzineza, and B. L. T. Plourde, *arXiv*:1405.0256 [cond-mat.supr-con] (2014).

[10] J. N. Ullom, P. A. Fisher, and M. Nahum, *Appl. Phys. Lett.* **73**, 2494 (1998).

[11] If $\Gamma_0/C < 1/2\tau_{\text{ss}}r$, the more stringent bound $g < \Gamma_0/C(1-$ $\tau_{\text{ss}}r\Gamma_0/C)/\tau_{\text{ss}}$ holds.

[12] If $|\vec{R}| = r$ is the boundary of the superconducting film (requiring no quasiparticle flow), $s \approx P/A$. If an infinitely large reservoir is connected from the $|\vec{R}| > r$ area, $s \to 0$. Realistic conditions should be in between the two limits.

[13] G. Catelani, R. J. Schoelkopf, M. H. Devoret, and L. I. Glazman, *Phys. Rev. B* **84**, 064517 (2011).

[14] B. L. Altshuler, and A. G. Aronov, *Electron-electron interaction in disordered conductors*, p. 61 *in*: ed. by A.L. Efros and M. Pollak, Elsevier, Amsterdam 1985.

[15] A. A. Abrikosov, *Fundamentals of the theory of metals*, 630pp., North Holland, Amsterdam 1988.

[16] A. Golubov and E. Houwman, *Physica C* **205**, 147 (1993).

[17] J. N. Ullom, P. A. Fisher, and M. Nahum, *Phys. Rev. B* **61**, 14839 (2000).

[18] G. Catelani, J. Koch, L. Frunzio, R. J. Schoelkopf, M. H. Devoret, and L. I. Glazman, *Phys. Rev. Lett.* **106**, 077002 (2011).

[19] M. Lenander, H. Wang, R. C. Bialczak, E. Lucero, M, Mariantoni, M. Neeley, A. D. O'Connell, D. Sanks, M. Weides, J. Wenner, T. Yamamoto, Y. Yin, J. Zhao, A. N. Cleland, J. M. Martinis, *Phys. Rev. B* **84**, 024501 (2011).