Supplementary Material for "A Cascaded Random Access Quantum Memory"

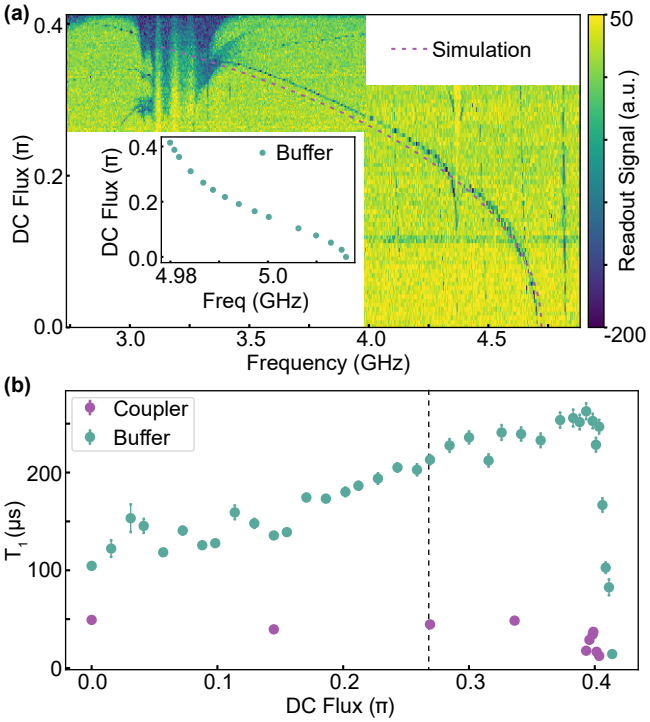
I. DEVICE PARAMETERS

A. Coupler Dispersion

We will first discuss the dispersion of the coupler and the buffer mode. Supplementary Figure 1(a) shows the coupler and buffer mode frequencies at different DC flux points, Φ_{DC}. For each value of Φ_{DC}, the buffer mode is first initialized to $|1\rangle$ using the transmon-buffer modes $|f0\rangle \leftrightarrow |g1\rangle$ sideband. The RF flux modulation is then turned on for $1\,\mu$s, followed by another $|f0\rangle \leftrightarrow |g1\rangle$ sideband that maps the buffer state into the transmon $\{|g\rangle, |f\rangle\}$ subspace. This state is subsequently mapped to the $\{|g\rangle, |e\rangle\}$ subspace using a transmon π pulse between $|f\rangle$ and $|e\rangle$. Blue regions in the plot indicate possible transitions between the buffer and other modes in the system. One of the most tunable traces corresponds to the coupler mode, which matches the black-box quantization results [1].

The T_1 data for the coupler and buffer modes is shown in Supplementary Figure 1(b). To measure buffer mode's T_1, we prepare $|1\rangle$ in the buffer mode using the $|f0\rangle \leftrightarrow |g1\rangle$ sideband with the transmon, wait for a varied time, then apply the $|f0\rangle \leftrightarrow |g1\rangle$ sideband again and readout the qubit state. Since the coupler mode lacks a dedicated dispersive readout, we use a cascaded readout scheme: prepare $|1\rangle$ in B, apply an RF flux-modulated sideband between $|B\rangle$ and the coupler, transfer the excitation to the coupler, wait for a varied time, reverse the mapping, and finally read out the qubit state.

When the coupler is in the flux filter stopband (see Supplementary Section VI), the buffer mode's T_1 increases as Φ_{DC} increases. This behavior is attributed to the growing frequency difference between the coupler and the buffer mode, which reduces the dressing of the buffer mode. For $\Phi_{\mathrm{DC}} > 0.4\pi$, the coupler frequency exits the filter stopband. Consequently, the buffer mode's T_1 drops significantly, and the coupler mode's T_1 becomes too short to be measured accurately with our FPGA electronics [2].



Supplementary Figure 1. Coupler DC flux sweep. (a) Coupler and buffer (inset) mode frequencies. At each DC flux point, the buffer mode is initially prepared in $|1\rangle$, followed by an RF-flux modulation frequency sweep. The blue colors in the plot indicate possible transitions. The black-box simulation of the coupler frequency is shown as the red dashed line. The X-axis shows the coupler frequency by subtracting the buffer mode frequency. The scan is performed in two parts due to the RF channel frequency limitations. (b) T_1 measurements of the coupler and buffer modes. All RAQM experiments are conducted with the coupler at the dashed line (0.269π).

B. Device Coherences

All modes' frequencies and coherences at the operation point $\Phi_{\mathrm{DC}} = 0.269\pi$ (dashed line in Supplementary Figure 1(b)) are shown in Supplementary Table 1. In our experiments, the T_1 of the storage modes is limited by the external drive pin length, resulting in an over-coupled regime where external coupling Q_c is smaller than the modes' intrinsic Q_i. We reduced the drive pin length in a subsequent cooldown to increase Q_c. In this new cooldown, we experimentally observed that 6 out of the 7 storage modes have T_1 exceeding 1 ms. This coherence time is comparable to a bare Aluminum flute cavity [3]. Further improvements would require different cavity materials, such as niobium [4]. Detailed coherence time in the new cooldown is shown in Supplementary Table 2.

C. Self-Kerrs and Cross-Kerrs

Supplementary Table 3 shows the cross-Kerrs and self-Kerrs measured in the experiments. The self-Kerrs K of the two buffer modes B_i's (including the unused B_2 mode in the experiments) are measured through two coherent cavity displacements gapped by a varied waiting time τ [3]. The displacement amplitude α is also varied. After each sequence, the population of B_i in the state $|0\rangle$ is measured, and the oscillation frequency, f_d, is extracted for each α. Under small-angle approximation $K|\alpha|^2 t \ll \pi$ [5], the self-Kerr coefficient is determined as

$\Phi_{dc} = 0.269\pi$	Symbol	Frequency/2π (GHz)	T_1 (μs)	T_R (μs)	T_{echo} (μs)
Qubit (Q)	ω_q	3.568	493 ± 13.1	259 ± 8.0	370 ± 14
Coupler (C)	ω_c	4.037	45 ± 0.8	0.39 ± 0.04	
Buffer 1	ω_{b1}	4.984	209 ± 5.1	75.5 ± 3.8	219 ± 7.0
Buffer 2	ω_{b2}	5.158	120 ± 1.3	70.6 ± 5.4	193 ± 6.7
Storage 1	ω_{s1}	5.333	358.3 ± 6.0	235.7 ± 7.1	551.6 ± 16.8
Storage 2	ω_{s2}	5.505	1254.8 ± 30.2	378.3 ± 15.1	1493.2 ± 53.5
Storage 3	ω_{s3}	5.681	799.0 ± 12.7	677.2 ± 16.8	1309.9 ± 41.8
Storage 4	ω_{s4}	5.860	597.4 ± 11.9	806.2 ± 23.7	972.4 ± 29.8
Storage 5	ω_{s5}	6.037	355.7 ± 5.5	591.8 ± 15.2	663.3 ± 22.7
Storage 6	ω_{s6}	6.229	589.5 ± 12.9	1071.1 ± 63.0	1048.4 ± 54.1
Storage 7	ω_{s7}	6.407	371.2 ± 6.4	663.5 ± 29.0	692.7 ± 22.7
Dump 1	ω_{d1}	7.297	0.73 ± 0.03		
Dump 2	ω_{d2}	7.252	1.81 ± 0.11		
Readout (R)	ω_r	8.051	0.05 ± 0.01		

Supplementary Table 1. Mode frequencies and coherence.

	Storage 1	Storage 2	Storage 3	Storage 4	Storage 5	Storage 6	Storage 7
T_1 (ms)	1.153 ± 0.020	1.165 ± 0.019	1.138 ± 0.016	1.180 ± 0.009	1.086 ± 0.016	1.019 ± 0.014	0.629 ± 0.007

Supplementary Table 2. The experimentally measured storage mode T_1 in a different cooldown. The cavity drive pin length was reduced to prevent storage modes from being over-coupled. S_7 has a lower T_1 because of its proximity to the on-chip filter's stop band edge, resulting in reduced protection.

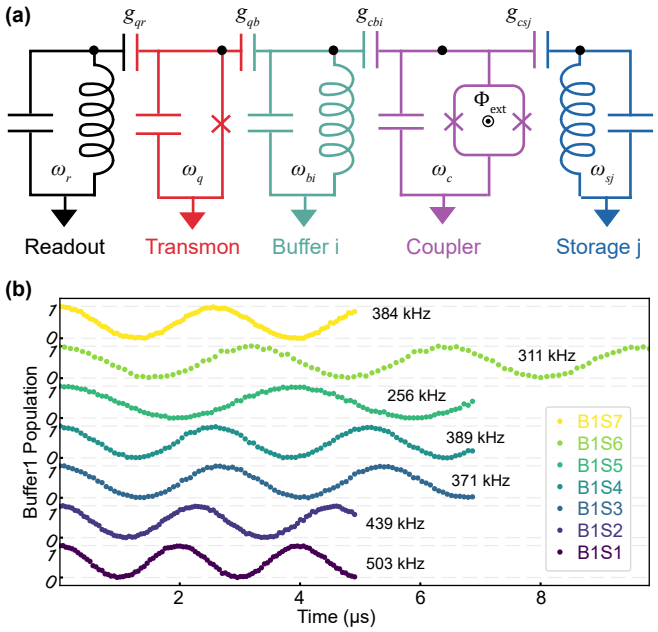
the fitted slope of f_d to α. S_i's self-Kerrs are measured similarly, with the displaced state prepared in B first and then swapped into S_i through the BS_i sideband. Cross-Kerrs between cavity modes are measured by performing Ramsey experiment on the target mode S_i with the spectator mode S_j initialized at $|0\rangle$ or $|1\rangle$. We measure the cross-Kerr strength between the coupler and $B_i(S_5)$ by performing cavity mode spectroscopy while the coupler is at $|g\rangle$ or $|e\rangle$. The cross-Kerrs between the coupler and other storage modes are measured through the BS_i sideband frequency difference when the coupler is at $|g\rangle$ or $|e\rangle$.

II. SYSTEM HAMILTONIAN

Supplementary Figure 2(a) shows the RAQM circuit diagram. The effective capacitance and inductance for different modes are calculated from Black-box quantization [1] using the admittance extracted through HFSS simulations. To calculate the memory sideband rate analytically, we first consider the subsystem containing B_i, coupler, and S_j for simplicity. The corresponding dressed frequencies are ω_{bi}, ω_c, and ω_{sj}. Assuming symmetric Josephson junctions with energy E_J, and a purely differential RF flux drive[6], the potential energy of the coupler SQUID is

$$U = 2E_J \cos(\varphi_{ext}) \cos(\varphi_c). \quad (1)$$

Here,$\varphi_{ext} = \pi\Phi_{ext}/\Phi_0$ is the external flux threading the SQUID loop and φ_c is the phase variable associated with the coupler mode in the dressed basis. Under the external



Supplementary Figure 2. (a) RAQM circuits diagram. For simplicity, we only show the coupling to the buffer mode B_i and the storage mode S_j. The coupler is capacitively coupled to all buffer, storage, and dump modes. (b) Experimentally measured B_1S_j sideband rates.

RF flux modulation of amplitude ϵ and at frequency ω_d as well as a continuous DC flux bias φ_{DC}:

$$\varphi_{ext} = \varphi_{DC} + \varepsilon \sin(\omega_d t), \quad (2)$$

Assuming ε to be small ($\varepsilon = 0.035 \ll 1$ from sim-

Kerrs (kHz)	Q	C	B_1	B_2
Q	-143000			
C	-	-57120		
B_1	-285	-1452	-6.8	
B_2	-271	-	-15	-5.1
S_1	-	-422	-3.336 ± 0.154	-
S_2	-	-412	-1.838 ± 0.171	-
S_3	-	-202	-1.192 ± 0.167	-
S_4	-	-62	-1.344 ± 0.173	-
S_5	-	-278	-0.600 ± 0.163	-
S_6	-	-192	-0.716 ± 0.172	-
S_7	-	-232	-0.525 ± 0.172	-
R	-320	-	-	-

Kerrs (kHz)	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	-0.841						
S_2	-0.178 ± 0.011	-0.062					
S_3	-0.076 ± 0.009	-0.044 ± 0.01	0.073				
S_4	-0.061 ± 0.009	-0.046 ± 0.009	-0.020 ± 0.009	0.954			
S_5	-0.037 ± 0.01	-0.046 ± 0.009	-0.038 ± 0.009	-0.016 ± 0.009	-0.466		
S_6	-0.043 ± 0.014	-0.017 ± 0.014	-0.041 ± 0.016	-0.012 ± 0.014	-0.019 ± 0.015	-0.194	
S_7	-0.005 ± 0.010	-0.030 ± 0.009	-0.007 ± 0.010	-0.025 ± 0.009	-0.009 ± 0.008	-0.011 ± 0.010	0.048

Supplementary Table 3. Mode self-Kerrs and cross-Kerrs at $\Phi_{\rm dc} = 0.269\pi$. Positive Kerrs might come from hybridizing to the coupler's higher differential mode.

ulations) , we can substitute 2 into 1 and expand that equation to first order in ε as:

$$U \approx 2E_J \left(\cos\varphi_{\rm DC} - \varepsilon \sin(\omega_d t)\sin\varphi_{\rm DC}\right)\cos\varphi_c \tag{3}$$

When the coupler is dispersively coupled to the other modes, in the appropriate rotating frame, φ_c can be approximated using annihilation operators:

$$\begin{aligned}\varphi_c = \varphi_{\rm zpf}(a_c e^{i\omega_c t} - \frac{g_{cbi}}{\omega_c - \omega_{bi}} a_{bi} e^{i\omega_{bi} t} \\ - \frac{g_{csj}}{\omega_c - \omega_{sj}} a_{sj} e^{i\omega_{sj} t} + h.c.)\end{aligned} \tag{4}$$

where $\varphi_{\rm zpf}$ is the zero point fluctuation of the phase variable. Here the dressed operators for the coupler mode include contributions from each mode, normalized by participation factors of the form g/Δ where g is the coupling strength and Δ is the detuning between the coupler and a given mode. We can plug Eq. 4 into Eq. 3 and expand the potential up to the second order in the participation factors. We choose $\omega_d = \omega_{bi} - \omega_{sj}$ to activate the beamsplitter interactions. Keeping only the time-independent terms, we have:

$$U \approx \varepsilon E_J \sin(\varphi_{\rm DC}) \frac{g_{cbi} g_{csj}}{(\omega_c - \omega_{bi})(\omega_c - \omega_{sj})} (a_{bi} a_{sj}^\dagger + h.c.), \tag{5}$$

which describes the beamsplitter rate. In experiments, we calibrate this beamsplitter gate by preparing a photon in the buffer mode and activating the beamsplitter drive for $\sim 100\mu$s (equivalent to ~ 50 swaps) for a range of frequencies. We choose the resonant frequency which maximizes the probability of finding the photon in the storage mode. The repeated sequence of swap gates helps us calibrate the pulse parameters in presence of coherent errors. For calibration in presence of bichromatic drive tones or strong stark shifts, we refer the reader to ref. [6]. Supplementary Figure 2(b) shows the sideband rates used in the RAQM experiments.

The cross-Kerr $\chi_{b_i s_j}$ between buffer mode B_i and storage mode S_j, and the modes' self-Kerrs $\{k_{bi}, k_{sj}\}$ can be calculated using the effective mode capacitance $\{C_{bi}, C_{sj}\}$ and inductance $\{L_{bi}, L_{sj}\}$ [1]:

$$\chi_{b_i s_j} = -2\sqrt{k_{bi} k_{sj}}, \tag{6}$$

$$k_{bi(sj)} = \frac{L_{bi(sj)}}{C_{bi(sj)}} \frac{C_c}{L_c} E_{C_c}, \tag{7}$$

$$L_c = \frac{\Phi_0^2}{4\pi^2 E_{Jc} \cos\left(\varphi_{\rm DC}\right)} \tag{8}$$

Here, E_{C_c} is the charging energy for the coupler, and C_c is the coupler's capacitance to ground.

III. ACTIVE RESET

The long coherence of the storage modes results in an extended system reset time ($5T_1 > 5$ ms). To speed up all the experiments, we implement an active reset protocol modularly, including buffer mode reset, qubit reset, and storage mode reset.

Supplementary Figure 3 (a) illustrates the sideband interactions between the buffer and dump modes. We separately prepare $|1\rangle$ in B_1 and B_2, then activate the coupler modulation at the difference frequencies of $B_1 D_1$ and $B_2 D_2$ for a variable time. The lossy dump modes,

	Symbol	Value
Readout length	t_r	$1.0\,\mu$s
Readout relaxation time	t_{gr}	$2.5\,\mu$s
Assignment fidelity ($\|g\rangle$)	P_{gg}	99.68%
Assignment fidelity ($\|e\rangle$)	P_{ee}	98.34%

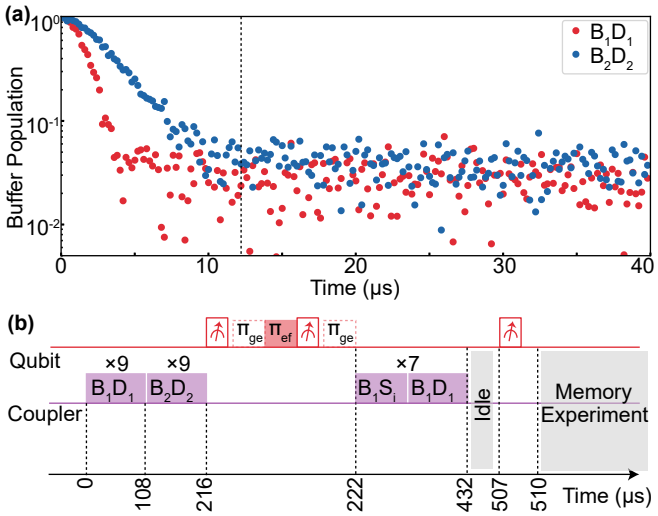
Supplementary Table 4. Qubit readout performance. The readout fidelity is estimated by averaging 20000 single-shot histograms.

acting as cold reservoirs, autonomously dissipate the received excitations, effectively cooling the buffer modes. The residual population in the buffer modes is swapped back to the qubit's $|f\rangle$ level, rotated to $|e\rangle$ level, and subsequently measured. The residual population of both buffer modes remains constant at $< 10\%$ after $12\,\mu$s of sideband interactions.

Supplementary Figure 3 (b) illustrates our RAQM active reset protocol, performed before the main experiments. The buffer modes are evacuated first. In the memory experiments, both the buffer modes and the qubit theoretically use, at most, the first excited states. Consequently, the protocol ensures state reset upto the second excitation for all modes. Since the qubit-B_i dispersive shift χ_{qb_i} is comparable to the sideband rate, we sequentially choose 9 different coupler RF modulations frequencies for each B_iD_i sideband, accounting for the dressing effects of the dispersive shifts. After evacuating both buffer modes, we sequentially reset the qubit's $|e\rangle$ and $|f\rangle$ states. First, we measure the qubit state using a $1\,\mu$s readout pulse, wait for $2\,\mu$s, and conditionally flip the qubit state if it is in the $|e\rangle$ state. Next, we apply a qubit $|f\rangle$-$|e\rangle$ π pulse and repeat the active reset for the $|e\rangle$ state. To reset the storage modes, we sequentially apply 7 swap pairs, transferring the population from S_i to B_1, then dump the population from B_1 to D_1 and autonomously evacuated. Finally, before starting the memory experiments, we idle the system for $75\,\mu$s so that any remaining low-coherence modes are fully reset. The experiments are initiated only when the qubit is confirmed in the $|g\rangle$. The qubit readout performance in our experiments is shown in Supplementary Table 4.

IV. SYSTEM TEMPERATURE

We benchmark the cavity's temperature using the Ramsey parity measurement [7]. To detect the thermal photon population in B_1, we repeatedly measure the parity information 80 times in a single experiment. Each measurement sequence consists of the following pulses: a transmon $\pi/2$ rotation, an idle period of π/χ, another transmon $\pi/2$ rotation, a readout pulse ($1\,\mu$s), and a readout evacuation period ($2.5\,\mu$s). Each experiment produces an 80-bit string indicating the transmon's state. Supplementary Figure 4(a) illustrates the string behavior when B_1 is initialized to $|1\rangle$.

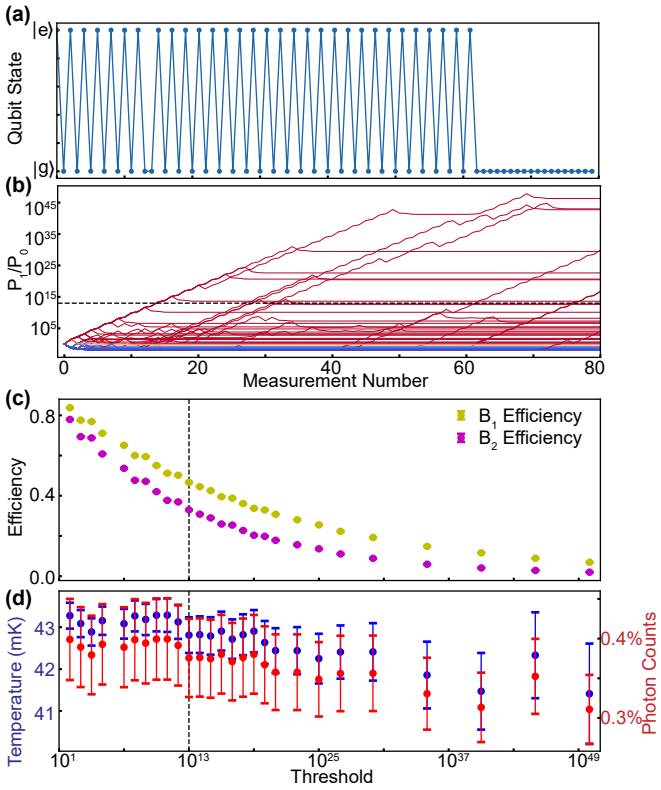


Supplementary Figure 3. RAQM active reset. (a) Buffer-Dump modes swap. Dash line (12μs) shows the time experiments used to evacuate buffer modes. (b) Active reset sequence. The sequence starts with evacuating buffer modes. Multiple buffer-dump swaps at different frequencies (dressed by χ_{qb_i}) are sequentially applied. Qubit state (up to $|f\rangle$ level) is actively reset. All Storage modes' populations are sequentially swapped into B_1, then evacuated through D_1. After $75\,\mu$s' idling period, the experiment starts when the qubit state is $|g\rangle$.

Given a specific type of transmon state string, we quantify whether the cavity is populated by applying the hidden Markov process [7] to calculate the cavity population ratio: $\frac{P_1}{P_0}$. Here, the cavity state probabilities P_0 and P_1 correspond to the cavity being in $|0\rangle$ and $|1\rangle$, respectively. These probabilities are reconstructed using the backward propagation algorithm based on the transmon state string. Starting from the beginning of the state string, the algorithm iteratively calculates the likelihood of observing the measured readout signal up to the current position, assuming the cavity was initially in $|0\rangle$ or $|1\rangle$. Each transmon state string is then converted into a probability string.

In Supplementary Figure 4(b), we analyzed 10^5 measurements taken when the system is in its thermal equilibrium state. A higher ratio in the plot indicates a greater likelihood of photon population in the buffer modes. As the threshold increases (dashed line), measurement errors are filtered out, isolating events corresponding to thermal photons.

After extracting the thermal photon events, we need to compensate for two key factors to avoid overestimating the system temperature: detection efficiency and event assignment. The detection efficiency η_i describes the probability of detecting a photon when B_i is in $|1\rangle$. In our experiments, we separately prepare $|1\rangle$ in B_1 and B_2, perform the Ramsey parity measurement to count cavity photon events, and approximate η_i using 2×10^5 experiments. In Supplementary Figure 4(c), fewer pho-



Supplementary Figure 4. Cavity temperature measurement. (a) The oscillatory readout signal indicates the presence of a photon in the cavity by continuously monitoring the cavity parity. (b) Each readout signal string is mapped to the cavity state probability through hidden Markov analysis. Red traces show the ratio change for different readout signals as strings are analyzed. The analyzed detection efficiency (c) and temperature (d) depend on the threshold. Here, we show the temperature for B_1 as an example. 2×10^5 measurements are used to approximate the detection efficiency and temperature.

ton events are captured as the threshold increases. Since both B_1 and B_2 are coupled to the transmon, thermal population in either cavity can trigger the detection. We assume that both modes have the same temperature T_c. Given the experimentally measured total photon counts δ and efficiency η_i, the temperature T_c required to explain the counts should satisfy:

$$\sum_{i=1,2} \frac{\eta_i}{-1 + \exp\left(\frac{\hbar\omega_{bi}}{k_B T_c}\right)} = \delta \tag{9}$$

Supplementary Figure 4(d) shows the corrected thermal photon counts in B_1 and the corresponding temperature. We choose 10^{13} as our threshold ($\eta_1 = 46.6\% \pm 0.2\%$, $\eta_2 = 33.0\% \pm 0.2\%$) and list the system temperature (steady state temperature T_{SS} and photon counts $\bar{n}_{ss}$, active reset temperature T_{ar} and photon counts $\bar{n}_{ar}$ in Supplementary Table 5. Qubit temperature is calculated by measuring $|e\rangle \leftrightarrow |f\rangle$ Rabi oscillation amplitude with and without initial π pulse between $|g\rangle$ and $|e\rangle$.

$\Phi_{\rm dc} = 0.269\pi$	$\bar{n}_{ss}(\%)$	T_{ss} (mK)	$\bar{n}_{ar}(\%)$	T_{ar} (mK)
Qubit	1.59 ± 0.12	41.2 ± 0.01	-	-
Coupler	1.18 ± 0.12	43.5 ± 0.3	4.96 ± 0.25	63.5 ± 0.3
Buffer 1	0.38 ± 0.05	42.8 ± 0.4	0.55 ± 0.06	45.8 ± 0.4
Buffer 2*	-	42.8 ± 0.4	-	45.8 ± 0.4
Storage 1	0.20 ± 0.03	41.3 ± 0.5	0.59 ± 0.06	49.9 ± 0.4
Storage 2	0.14 ± 0.02	40.4 ± 0.5	0.48 ± 0.06	49.4 ± 0.4
Storage 3	0.12 ± 0.02	40.8 ± 0.5	0.33 ± 0.04	47.8 ± 0.4
Storage 4	0.11 ± 0.02	41.4 ± 0.5	0.27 ± 0.03	47.5 ± 0.5
Storage 5	0.08 ± 0.02	40.6 ± 0.5	0.24 ± 0.03	48.1 ± 0.2
Storage 6	0.17 ± 0.02	46.7 ± 0.5	0.61 ± 0.05	58.5 ± 0.2
Storage 7	0.05 ± 0.01	40.9 ± 0.6	0.17 ± 0.02	48.1 ± 0.2
Dump 1	0.02 ± 0.004	40.8 ± 0.7	0.03 ± 0.005	42.8 ± 0.3
Dump 2	0.02 ± 0.004	41.4 ± 0.6	0.04 ± 0.008	45.1 ± 0.3

Supplementary Table 5. System temperature and thermal photon populations.
* Buffer 2's temperature is assumed to be the same as Buffer 1.

The temperature of the other storage and dump modes is measured with an additional π swap between corresponding buffer modes initially. The coupler's temperature is higher after the active reset, which can be mitigated by adding additional coupler reset through the dump modes or waiting longer ($> 75\,\mu$s) before the experiments start.

V. PHASE CALIBRATION FOR RAQM EXPERIMENTS

In the RAQM experiments, states are modified on the transmon side and transferred to B_1 and S_i. However, the beamsplitter interactions between buffer and storage modes naturally provide iSWAP instead of the SWAP gate ($\pi/2$ phase difference). The sideband interactions used for state transfer also induce Stark shifts in mode frequencies. As a deterministic frequency shift, this phase accumulation can all be corrected by adjusting the phases of all subsequent transmon gates [8], effectively acting as an additional virtual-Z gate.

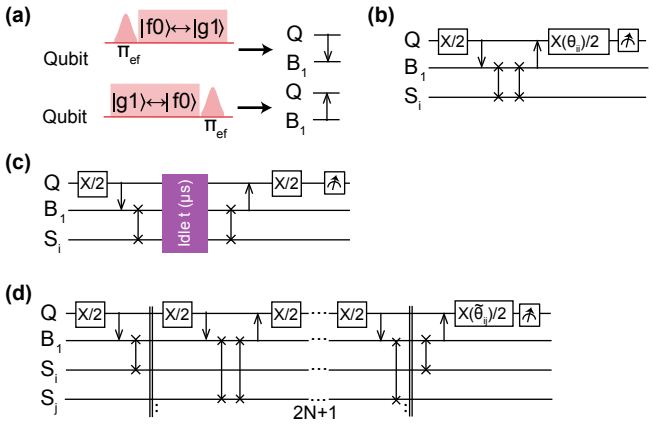
The virtual-Z phase is the sum of the following three components:

(a) Active-access phase (θ_{ii}): This is the phase accrued in S_i during its read and write operations.

(b) Inactive-access phase (θ_{ij}): This is the phase accrued in S_i during the read and write operations of S_j.

(c) Idle phase ($\theta_{\Delta_i}(t)$): When sidebands ($Q \leftrightarrow B_1$, $B_1 \leftrightarrow S_i$) are pulsed, the FPGA board tracks the Stark-shifted frequency f_{ss}. During the periods when the sidebands are off, the system's idle frequency f_{id} differs from the tracked frequency, causing all subsequent sideband pulse phases to deviate by $\Delta_i t_{\rm gap}$. Here $\Delta_i = f_{ss} - f_{id}$ is the total Stark-shift of S_i due to sidebands.

Supplementary Figure 5 shows the process of virtual-Z phase calibration. For simplicity, state read and write operations between the transmon and B_1 are represented using circuit symbols in Supplementary Figure 5(a). Supplementary Figure 5(b) shows the circuit to calibrate θ_{ii}.



Supplementary Figure 5. RAQM Phase calibration. Each transmon gate's virtual-Z phase comprises three components: the active-access phase θ_{ii}, the inactive-access phase θ_{ij} accumulated during storage read and write, and the time-dependent idle phase $\theta_{\Delta_i}(t)$ caused by hardware frequency tracking. (a) Pulse sequences for state transfer between transmon and B_1. (b) active-access phase θ_{ii} calibration for S_i. (c) Idle phase calibration extracts the Stark-shift Δ_i to S_i through the Ramsey-like experiments. (d) Inactive-access phase θ_{ij} calibration on S_i after S_j access operations. θ_{ij} are calculated based on the maximum contrast phase $\tilde{\theta}_i j$ and Δ_i. To improve calibration accuracy, $N+1$ writes and N reads are repeated.

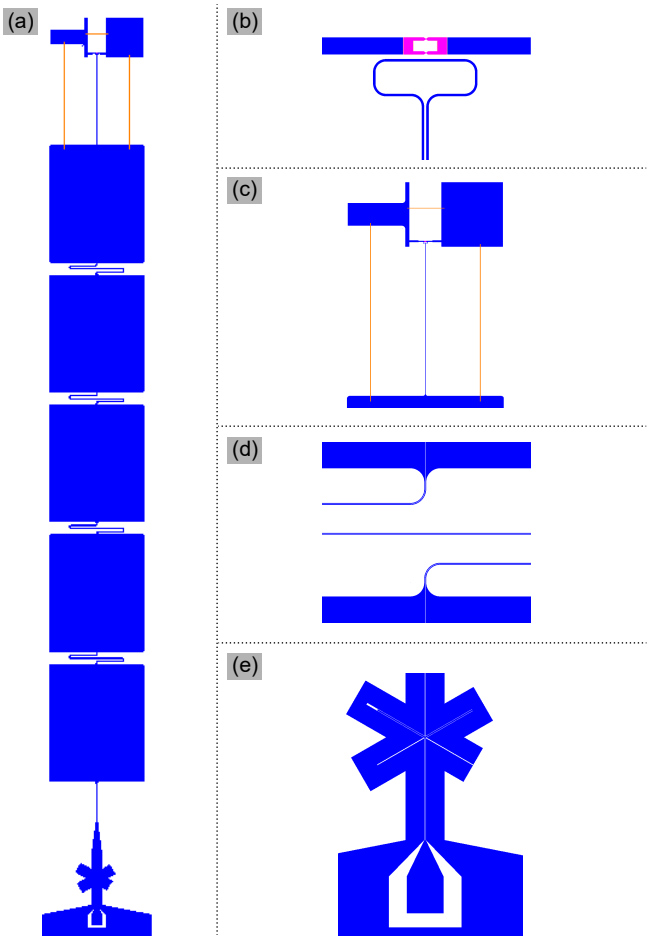
The phase of the second transmon gate is swept, and θ_{ii} is determined as the phase that maximizes the final $|e\rangle$ population. Supplementary Figure 5(c) shows the circuit to calibrate θ_{ii}. By sweeping the waiting time between two B_1S_i swaps, we fit the Ramsey signal and extract the Stark-shift frequency Δ_i. Supplementary Figure 5(d) shows the circuit to calibrate θ_{ij}. Since the inactive-access phase is small, we repeat the read and write operations on S_j for 5 times. The maximum contrast phase $\tilde{\theta}_{ij}$, determined by the maximum final $|e\rangle$ population in the transmon, can be used to calculate θ_{ij}:

$$\theta_{ij} = \frac{\tilde{\theta}_{ij} - \Delta_i t_{\text{gap}}}{5} \tag{10}$$

Here, t_{gap} is the time between the first and the last B_1S_i swaps. For the multiplexed RAM RB experiments, each S_i tracks its virtual-Z phase based on the RB sequences. The phases of all transmon gates are first calculated on a classical computer and then sent as instructions to the FPGA board to generate the corresponding pulses.

VI. FAST FLUX DELIVERY FOR HIGH-Q 3D CAVITIES

Flux control is essential in quantum computing for implementing high-fidelity two-qubit gate schemes and single transmon Z-control. While on-chip flux lines are commonly used in 2D systems, integrating flux lines in 3D cavities is challenging due to two primary factors:



Supplementary Figure 6. Fast flux delivery for 3D cavities. (a) Overview of the coupler chip, which comprises four main components from top to bottom: (b) the SQUID coupler with its associated flux loop, (c) shorting line structures (in orange) for ESD protection, (d) the edge-coupled microstrip filter, and (e) the double-Y balun. The blue areas indicate Nb metal regions, while pink highlights the Al SQUID junctions. The chip is inserted into the multimode cavity and wire-bonded to an SMA connector for DC and RF flux control.

(1) The absence of a well-defined ground plane on the chip.

(2) The widely distributed cavity mode field, which can easily couple to the flux line, substantially reduces the cavity T_1 without flux line protection.

In 3D systems, introducing flux lines while preserving cavity coherence remains a significant challenge. For DC flux tuning, external flux coils are commonly used, but their reliance on normal metals limits cavity coherence. Recently, on-chip flux pick-up loops [9] and magnetic hoses [10] have emerged as promising alternatives for DC flux biasing. In contrast, RF flux modulation has only been demonstrated using a coaxial stub cavity [6]. Despite these advancements, an on-chip flux line capable of supporting both DC and RF signals while maintain-

ing high coherence in a single cavity mode has yet to be realized.

Here, we realized a simple on-chip flux line design compatible with standard 3D cavities for both DC and RF flux delivery, with the following advantages:

(1) Ultra-wide (> 3 GHz) and deep stopband with a flexible flux line geometry.

(2) Compatible with coaxial connectors, one of the most widely used types for RF experiments.

(3) Requiring only a single standard superconducting chip, eliminating the need for additional circuit components.

(4) Free from external flux coils while maintaining high cavity coherence (> 1 ms) and low cavity thermal populations ($< 0.05\%$).

Our coupler chip design is shown in Supplementary Figure 6(a). It is composed of 4 parts (shown in Supplementary Figure 6(b), (c), (d), (e)): The SQUID coupler, the shorting lines, the edge-coupled microstrip filter, and the double-Y balun.

The SQUID coupler is a flux-tunable transmon with two pads that are capacitively coupled to both buffer and storage modes. The flux loop is designed to deliver a pure RF flux modulation to the SQUID. However, the flux loop also has a stray capacitive coupling to the coupler. One solution is to maximize the flux line's inductive coupling to the SQUID loop [11, 12]. In these references involving 2D designs, the flux line produced by such optimization is effective for all frequency ranges and helps mediate fast and clean two-mode squeezing sidebands at > 7 GHz. In this work, we can alternatively optimize the SQUID coupler geometries for low RF modulation frequencies < 2 GHz [6, 13]. We optimized the location of the flux loop using ANSYS HFSS simulations. In particular, we quantified the flux-to-charge drive ratio with the following equations: We define the two SQUID junctions as J_1 and J_2. When RF signals at frequency ω_{rf} are applied through the flux line, the induced currents along J_1 and J_2 are denoted $I_1(\omega_{rf})$ and $I_2(\omega_{rf})$, respectively [6]. For a pure flux drive:

$$I_1(\omega_{rf}) = -I_2(\omega_{rf}) \tag{11}$$

For a pure charge drive:

$$I_1(\omega_{rf}) = I_2(\omega_{rf}) \tag{12}$$

In the presence of both drives, maximizing the flux modulation amplitude is equivalent to minimizing the following ratio $r(\omega_{rf})$:

$$r(\omega_{rf}) = \left| \frac{I_1(\omega_{rf}) + I_2(\omega_{rf})}{I_1(\omega_{rf}) - I_2(\omega_{rf})} \right| \tag{13}$$

Following ref. [6], we optimize the flux-loop position and capacitor pad geometry to minimize $r(\omega_{rf})$. In practice, optimizing across the entire range of ω_{rf} is challenging. Since only RF modulation frequencies below 2.5 GHz are required, we focused our HFSS optimization on the low-frequency range.

Shorting lines are for ESD protection of fab on sapphire chips. They are removed later during packaging and are discussed in the Supplementary Section. X.

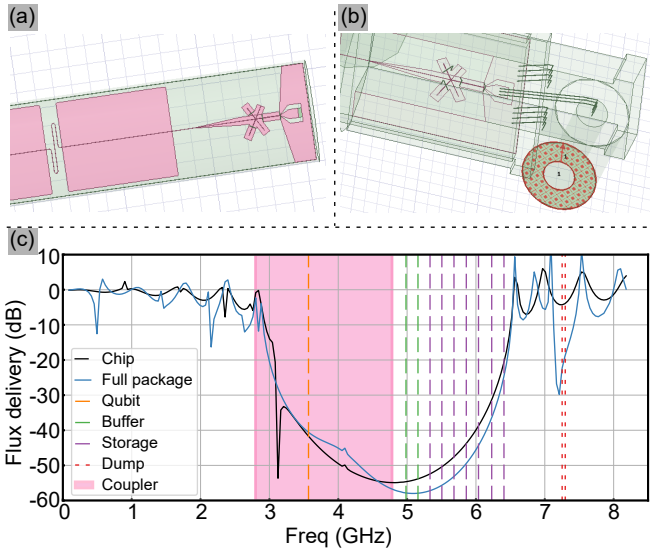
The edge-coupled microstrip filter [14] provides a wide stopband to protect all coherent elements in our RAQM. As shown in Supplementary Figure 7(c), the simulated response (blue curve) demonstrates this stopband. Similar to a Step-Impedance Low-Pass Filter (SIPF), the edge-coupled microstrip filter consists of alternating high- and low-impedance coplanar stripline (CPS) sections of equal electrical length. The stopband's width, depth, and center frequency are determined independently by the impedance ratio, the number of sections, and the electrical length. The analytic expression [15] can be used to quickly calculate the ABCD matrix of the filter and approximate the stopband shape.

The double-Y balun [16, 17] is used for connecting a balanced circuit (edge-coupled microstrip filter) to an unbalanced circuit (coaxial cable). This DC-compatible balun [18] contains six ports: two signal ports (a CPS port and a CPW port) and four dummy ports (CPS short, CPS open, CPW short, and CPW open). All six ports are matched to the same impedance (in our case, 50Ω) and electrical length λ. The signal at the CPW port is evenly distributed among the four dummy ports. Only the differential mode can propagate through the balun, while common-mode signals are rejected due to the opposite reflection phases of the port pairs [17]. The high-frequency cutoff of the balun is limited by its $\lambda/8$ mode. The open-circuit approximation also becomes invalid at higher frequencies.

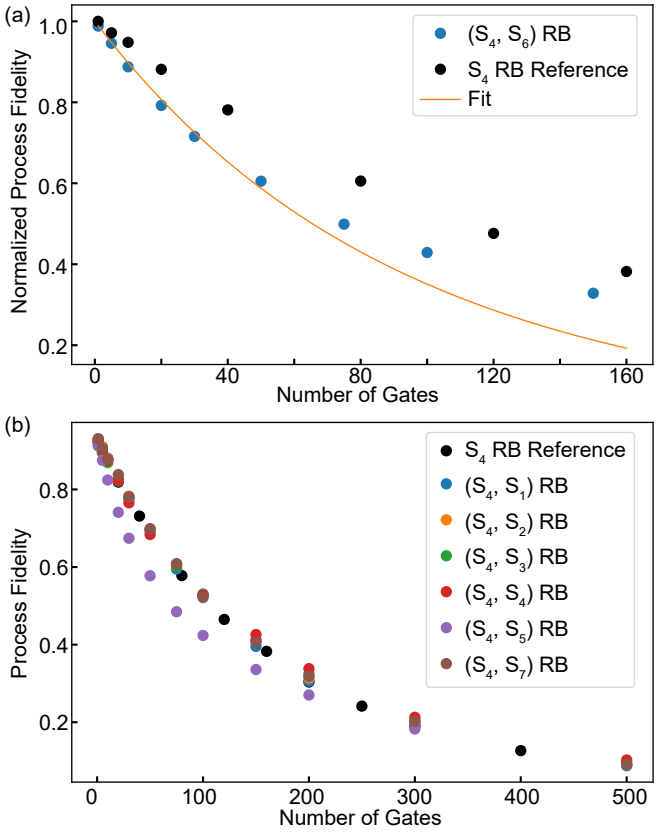
The flux filter works effectively when the input impedance is constant across all ω_{rf}. When coaxial cables transmit flux signals through the flux line, the double-Y balun helps suppress filter ripples caused by impedance mismatch on the package side. Supplementary Figure 7(a) and (b) show HFSS simulations for two cases: a perfect on-chip $50\,\Omega$ excitation port, and the entire package. The resulting RF modulation responses, depicted in Supplementary Figure 7(c) by the black and blue curves, demonstrate that the double-Y balun effectively aligns the two curves at low frequencies. More significant ripples appear in the blue curve at frequencies (> 7 GHz) higher than the balun range. Reducing the size of the double-Y balun can rapidly increase the cutoff frequency. In our design, the balun size (~ 0.5 mm) is constrained by the need for additional wirebonds. For two-mode squeezing interactions with a frequency around 10 GHz, directly fabricating airbridges on the balun and optimizing the design of CPS and CPW open ports will help achieve the desired frequency range.

VII. MEMORY CROSS-TALK BENCHMARKING

Crosstalk errors can degrade the stored state fidelity during RAQM operation. These errors depend on the



Supplementary Figure 7. HFSS-simulated response of the Balun-Filter. Simulation models include (a) the coupler chip alone and (b) the complete package. (c) RF flux modulation power delivery. Black and blue curves represent different models shown in (a) and (b), which align closely for frequencies below 7 GHz. This alignment indicates compatibility between the coaxial connectors and the coupler chip. Vertical lines denote experimentally measured mode frequencies.



Supplementary Figure 8. RB for state-dependent access error. RB is performed in the target mode S_i to benchmark the BS_i swap fidelity while the spectator mode S_j is occupied. The decay curve for each RB (target S_i, spectator S_j) is additionally averaged over 6 cases, one for each of the possible spectator occupation: $\{|0\rangle, |1\rangle, |+x\rangle, |-x\rangle, |+y\rangle, |-y\rangle\}$. (a) Semiclassical fit for the $(S_i, S_j) = (S_4, S_6)$ RB, reporting the additional swap error $\varepsilon_i^j = 0.938 \pm 0.060\%$. (b) Benchmarking $B_1S_i = B_1S_4$ swap fidelity in the presence of random population in S_j. The reference data refers to no population in the spectator mode.

stage of RAQM control for a given accessed mode S_i. The three stages are active access, idle, and inactive access. The dominant crosstalk errors during these stages are state-dependent access errors, many-body dephasing, and spectator access dephasing, respectively. We will present details on benchmarking these errors in the subsections below.

A. State-Dependent Access Error

The state-dependent access error occurs when fetching the target mode via the BS_i swap. When the other storage modes $S_j, (j \neq i)$ are occupied, the buffer-storage cross-Kerr interactions χ_{BS_j} can change the BS_i swap frequency randomly. Those randomly applied off-resonant swaps have lower fidelity and finally dephase the state in S_i after rounds of read/write. To benchmark this error, we perform randomized benchmarking (RB) introduced in the main text Section III between B_1 and S_i in the presence of populations in distinct spectator mode S_j. During the RAM RB experiments, the RAQM is populated with different Clifford states in the $\{|0\rangle, |1\rangle\}$ subspace. We thus perform 6 rounds of RB to check the BS_i swap fidelity with the 6 distinct Clifford states in S_j. The final RB decay curve is average over the 6 curves. We evaluated fidelities for 15 different gate depths in the range $[0, 500]$. For each depth, we chose 20 different sequences for each RB gate depth, with each sequence measured $1000 - 5000$ times depending on gate depth (more repetitions for larger gate depths).

To extract the state-dependent access error, we perform a semi-classical Markovian analysis based on the measured RB decay curve. This analysis captures the effect of decaying spectator populations, which causes the state-dependent access error to be lower for large gate depths. As shown in Figure 8 (b), after 300 gates ($\sim 450\mu$s), decay curves for all target-spectator S_iS_j pairs follow the same trend as the reference case. We will now outline the analysis procedure.

First, the fidelity of $B_1 \leftrightarrow S_i$ swap is state-dependent: when $|S_j\rangle = |0\rangle$ ($|1\rangle$), the fidelity is F_0 (F_1). At gate depth N, we label the population of $|S_j\rangle$ in the $|0\rangle$ ($|1\rangle$) state as $P_0(N)$ ($P_1(N)$). We label the initial population of $|S_j\rangle$ in $|1\rangle$ as ζ. After applying each gate, the probability of $|S_j\rangle$ staying at $|1\rangle$ is A. Ignoring photon excitation error, $|S_j\rangle$'s population distribution at gate depth N is:

$$\begin{cases} P_0(N) = 1 - A^N\zeta \\ P_1(N) = A^N\zeta \end{cases} \tag{14}$$

Here $A = \exp(-\frac{t_{B_1\leftrightarrow S_i}}{\tilde{T}_1^j})$ is calculated through two experimentally measurable quantities: average single dual-rail gate length $t_{B_1\leftrightarrow S_i}$, and S_j's T_1 in the presence of $B_1 \leftrightarrow S_i$ dual-rail operation $\tilde{T}_1^j$. In our analysis, however, we will approximate $\tilde{T}_1^j \approx T_1^j$ as S_j's raw T_1.The measured RB process fidelity $\tilde{F}(N)$ at each gate depth N is:

$$\begin{aligned} \tilde{F}(N) &= \prod_{j=1}^{N} \left(F_0 P_0(j) + F_1 P_1(j)\right) \\ &= F_0^N \prod_{j=1}^{N} \left(1 + (\frac{F_1}{F_0} - 1)A^j\zeta\right) \end{aligned} \tag{15}$$

We define $\varepsilon_i^j \equiv 1 - \frac{F_1}{F_0}$, whose absolute value is relative deviation of F_1 from F_0. Here, indices i, j denote the target and spectator mode, respectively. We expect ε_i^j to be a small quantity because the buffer-storage cross-Kerr interaction is a small quantity on the timescale of gates, i.e., $\chi_{BS_j} t_{B_1\leftrightarrow S_i} < 0.5\%$. Thus, we can make the approximation $1 - \varepsilon_i^j A^j\zeta \simeq \exp(-\varepsilon_i^j A^j\zeta)$ for small gate depths ($j t_{B_1\leftrightarrow S_i} \ll T_1^j$). Using this approximation, we have:

$$\begin{aligned} \tilde{F}(N) &\simeq F_0^N \exp(\varepsilon_i^j\zeta \sum_{j=1}^{N} A^j) \\ &= F_0^N \exp\left(\varepsilon_i^j\zeta \frac{A(1-A^N)}{1-A}\right) \end{aligned} \tag{16}$$

To extract ε_i^j, we use the above equation to fit the previously obtained decay curve to a depth up to 50 gates ($\sim 50 - 75\mu$s). We set $\zeta = 1/2$ as the decay curve averages over the spectator mode's population over the 6 cardinal states of its Bloch sphere. An example of this fitting is shown in Supplementary Figure 8 (a). The state-dependent access error for a swap gate is $\varepsilon_{ai}^j = (1 - F_0^i\varepsilon_i^j)^{3/2}$, which is plotted in the main text Fig. 4 (c). For most target-spectator mode pairs, the state-dependent access error is below $< 0.3\%$, while only a few pairs have an error rate as high as 1.396%.

B. Many-Body Dephasing Error

Many-body dephasing refers to S_i being dephased by states in other storage modes via the storage-storage cross-Kerr interactions. This error occurs during all RAQM control stages but it is the dominant crosstalk error during the idle stage. We quantify this error by measuring the Ramsey of the target mode S_i while other storage modes are occupied. The protocol is as follows: (1) Prepare other storage modes $\{S_j | j \neq i\}$ uniformly in one of the 6 basis states $\{|0\rangle, |1\rangle, |+x\rangle, |-x\rangle, |+y\rangle, |-y\rangle\}$ (2) Perform cavity Ramsey on the storage mode S_i. By fitting the Ramsey signal to exponentially decaying sinusoids, we attain 6 dephasing rates $\{\kappa_p\}$, one for each basis state. Since the cross-Kerr interactions between storage modes (≈ 50 Hz) are 2 orders of magnitude smaller than the modes' decoherence rates, the S_i's many-body dephasing rate Γ_{mbd}^i is much smaller than the modes' intrinsic dephasing rates. As the worst case, we take the many-body dephasing rate as the largest variation in dephasing rates measured among 6 cases:

$$\Gamma_{mbd}^i = \max\{\kappa_p\} - \min\{\kappa_p\}. \tag{17}$$

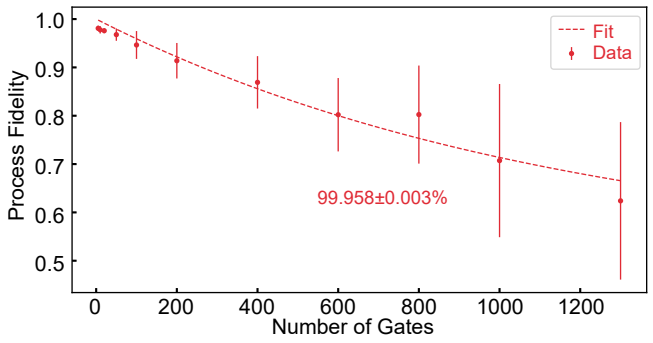
We measured this rate to be at most $\Gamma_{mbd}^i = 0.510 \pm 0.222$ kHz, which introduces at most 0.102% storage read/write error.

C. Spectator Access Dephasing

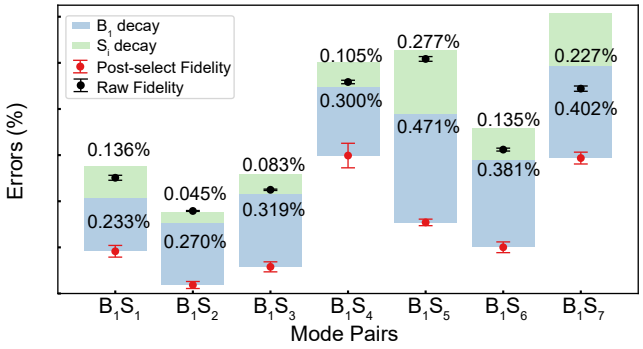
Spectator access dephasing occurs when we fetch states into the buffer mode from $S_i's$ spectator modes via BS_j swaps. Since we introduce random populations in the buffer mode, these populations effectively dephase the stored state in S_i via the χ_{BS_i} cross-Kerr interactions. To quantify this error, we measure the Ramsey of S_i mode in the presence of random $B_1 - S_j$ beamsplitter gates. We first initialize a distinct spectator mode S_j with a single photon. We then perform Ramsey experiment on S_i where the two $X_{\pi/2}$ pulses on S_i are separated by an RB sequence of BS_j beamsplitter gates. If we replace the RB sequence with an even number of BS_j swap gates, we would observe beating in S_i's Ramsey trace. This beating is due to the difference between χ_{BS_i} and $\chi_{S_iS_j}$ cross-Kerr interactions. However, the RB sequence of BS_j beamsplitter gates randomizes the population in buffer mode, rendering the beating incoherent. To extract the spectator access dephasing rate, we subtract S_i's intrinsic dephasing rate $\Gamma_{\phi,\mathrm{T2}}$ from the measured rate Γ_{meas}. The fitted error rate Γ_{meas} yields the effective error rate due to $B_1 - S_j$ gates as

$$\Gamma_{ai}^j = \Gamma_{\mathrm{meas}} - \Gamma_{\phi,\mathrm{T2}}$$

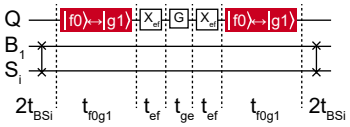
As shown in the main text Fig. 4 (d), the spectator access dephasing rate is below 3 kHz for most target-spectator pairs, while some have as high as 18.75 kHz. As RAQM size increases, each mode spends more time idling per memory flashing round. As a result, the modes become more sensitive to spectator access dephasing as we scale the size of RAQM.



Supplementary Figure 9. Transmon single qubit gate RB. The gate depths are limited by the RFSoC register memory.



Supplementary Figure 10. Error budget for B_1S_i swap fidelity. The blue and green parts represent the T_1 decay error from the B_1 and S_i during the swap.



Supplementary Figure 11. Pulse sequence of a storage access gate. X_{ef} refers to the π pulse on the transmon $\{|e\rangle, |f\rangle\}$ subspace. G represents a random gate on the transmon $\{|g\rangle, |e\rangle\}$ subspace with the same gate length t_{ge}.

VIII. SINGLE MODE CONTROL

Using the transmon-buffer 4-wave mixing interactions, we can perform universal gates on the buffer mode in the $\{|0\rangle, |1\rangle\}$ subspace. Each buffer mode gate is a cascaded gate composed of three periods: swapping the buffer state into the transmon, performing a single transmon gate, and swapping the transmon state back into the buffer.

The transmon-buffer swap is realized using the $|f0\rangle \leftrightarrow |g1\rangle$ sideband. The sideband is activated by charge driving with a frequency ω_{d4}. Here, $\omega_{d4} \approx 2\omega_q + \alpha - \omega_{bi}$ is chosen based on an experimental scan of the resonant swapping frequency. We use a Gaussian flat-top pulse with a $3\sigma(\sigma = 0.005\,\mu s)$ ramping as the pulse waveform.

After the swap, the buffer state $x|0\rangle + y|1\rangle$ is mapped to the transmon state $x|g\rangle + ye^{1j\phi_{d4}}|1\rangle$. Performing a single transmon gate requires a virtual phase correction to reverse the deterministic phase ϕ_{d4} coming from the sideband stark-shift. The phase cancellation procedure is similar to that described in Supplementary Section V. After performing the single transmon gate, the state is swapped back to the buffer. Supplementary Figure 9 shows the single transmon gate fidelity $(99.958\% \pm 0.003\%)$ measured through RB.

The single buffer gate length is $1.65\,\mu s$ with an average Clifford gate fidelity of 98.64%. This includes two $0.615\,\mu s$ buffer-transmon swaps, two 140 ns transmon $|f\rangle \leftrightarrow |e\rangle$ π pulses, and a 140 ns transmon $|e\rangle \leftrightarrow |g\rangle$ pulse. Compared to the small transmon-buffer χ-shift $(1/\chi_{qb} \sim 3.5\,\mu s)$, this is a cavity gate scheme beyond the χ-limit. Our experimental gate speed is limited by the room-temperature amplifier power on the drive lines.

IX. ERROR ANALYSIS

A. Single-mode error budget

In our buffer-storage Randomized Benchmarking experiment, each experimental point corresponds to the average of 30 random sequences, and each sequence is repeated from 1000 to 20000 times to get enough statistics. Each RB gate uses a Gaussian flat-top pulse with $3\sigma(\sigma = 0.005\,\mu s)$. Since we use parity measurement to extract the buffer-storage state after finishing an RB sequence, it is possible to misassign a higher fock state population in the readout. Noticing that the buffer populations in $|2\rangle$ are treated as $|0\rangle$, the measured raw RB fidelity is still accurate if we assume negligible buffer populations beyond $|3\rangle$. Based on our low system temperature (See Supplementary Section IV), the measurement misassignment error is not a major source of error in our system.

Supplementary Figure 10 shows the error budget for the swap fidelity of all B_1S_i pairs. The raw and post-select fidelities are all experimentally measured through randomized benchmarking. The blue and green parts represent the T_1 decay error from the B_1 and S_i during the swap. We use the B_1 and S_i's coherence during idle time (Supplementary Table 1) to calculate the decay error contribution during beamsplitter interactions. The remaining mismatch in fidelities could come from modes' T_1 fluctuations. The post-selected fidelity may be limited by coupler heating and cavity dephasing.

B. Full RAQM error budget

Here, we discuss the error budget for random read infidelity. Supplementary Figure 11 shows the pulse sequence for each S_i's access gate. Here, we label the gate lengths for a B_1S_i swap, a QB_1 $|f0\rangle \leftrightarrow |g1\rangle$ swap, a transmon $|e\rangle \leftrightarrow |f\rangle$ π pulse, and a random transmon

gate G on the $\{|g\rangle, |e\rangle\}$ subspace as $2t_{BS_i}$, t_{f0g1}, t_{ef}, and t_{ge}. While t_{BS_i} are dependent on S_i, the other gate lengths are fixed in our experiments: $t_{f0g1} = 0.606\,\mu$s and $t_{ef} = t_{ge} = 0.14\,\mu$s. Each S_i's access gate is divided into read, transmon processing, and write periods with length $t_{\rm read}$, $t_{\rm write}$, and $t_{\rm proc}$. Information is transferred from S_i to B_1 during the read period, while the write period reverses this process. Under this definition, we have:

$$t_{\rm read} = t_{\rm write} = 2t_{BS_i}$$
$$t_{\rm proc} = 2t_{f0g1} + 2t_{ef} + t_{ge}$$

The state-dependent access error, which happens during actively accessing S_i, is calculated using the measured worst-case swap infidelity ε^j_{ai}:

$$1 - \left(1 - \sum_{i=1;i\neq j}^{7} \varepsilon^j_{ai}\right)^{1/7}. \tag{18}$$

The coefficient accounts for the number of storage modes (7) in the multiplexed RAQM operation to avoid double counting (the same for the rest of the error budget discussions). Supplementary Table 6 shows the experimentally measured state-dependent access error.

Many-body dephasing occurs whenever S_i is populated with the target state. Using the many-body dephasing rate Γ^i_{mbd} extracted during the RAQM idle period, we can calculate the many-body dephasing error for S_i:

$$1 - \exp\left(-\frac{\Gamma^i_{mbd}}{7} \sum_{j=1;j\neq i}^{t} (t_{\rm read} + t_{\rm proc}/2)\right). \tag{19}$$

The spectator-access dephasing, which happens during the S_i's inactive-access period, is calculated using the measured S_j's spectator-access dephasing rate Γ^j_{ai}:

$$1 - \prod_{j=1;j\neq i}^{7} \left(1 - \Gamma^j_{ai}(t_{\rm read} + t_{f0g1})/7\right). \tag{20}$$

Details about the spectator-access dephasing rate are shown in Supplementary Table 7. For each pair's RB, we use 30 different sequences for each RB gate depth to get statistics.

The RAQM swap errors are scaled by the average number of swaps on S_i's random read operations during RAM experiments:

$$\left(1 - F_{ri}^{1/7}\right). \tag{21}$$

The decay and dephasing error of S_i includes all the inactive-access periods for S_i:

$$1 - \exp\left(-\frac{\Gamma^i_1}{7} \sum_{j=1;j\neq i}^{t} (t_{\rm read} + t_{\rm proc}/2)\right), \tag{22}$$

$$1 - \exp\left(-\frac{\Gamma^i_\phi}{7} \sum_{j=1;j\neq i}^{t} (t_{\rm read} + t_{\rm proc}/2)\right). \tag{23}$$

Γ^i_1 and Γ^i_ϕ represent the decay and pure dephasing rate for S_i. Supplementary Table 8 lists the error budget in detail. The average random read fidelities of different RAQM sizes are shown in Supplementary Table 9.

For the RAM RB experiments, we chose 30 different sequences for each RB gate depth, with each sequence measured 1000 times to get statistics. Since transmon's $\{|g\rangle, |e\rangle, |f\rangle\}$ are in the RAM RB, the process fidelity should stabilize at 1/3, given an evenly distributed readout contrast for all states. In our experiments, the readout signal difference in $\{|g\rangle, |e\rangle, |f\rangle\}$ causes the final process fidelity to be deviated from 1/3.

C. Master equation simulation

In this section, we perform a master-equation simulation to study the impact of decoherence on the fidelity of BS_i swap gates using QuTiP [19]. We focused on BS_2 gate, which had the lowest infidelity as shown in the main text Fig. 2 (e). We considered the following lab-frame Hamiltonian:

$$H_0 = \omega_b a_b^\dagger a_b + \omega_s a_s^\dagger a_s + H_c$$

where H_c contains interactions of the modes with the coupler

$$\begin{aligned} H_c &= \omega_{c0}\sqrt{|\cos(\varphi_{\rm ext}(t))|}a_c^\dagger a_c \\ &\quad + g_{cb}(a_c^\dagger a_b + h.c.) + g_{cs}(a_c^\dagger a_{si} + h.c.). \end{aligned} \tag{24}$$

Here ω_{c0} is the bare frequency of the SQUID coupler (at zero flux bias) and g_{ci} is the Jaynes-Cummings coupling between the SQUID coupler and the mode i. All parameters were extracted from experiments. Here, the flux through the SQUID loop is modulated as

$$\varphi_{\rm ext}(t) = \frac{\pi\Phi_{\rm ext}}{\Phi_0} = \varphi_{\rm DC} + \epsilon\cos(\omega t + \phi(t)) \tag{25}$$

using a RF drive with strength ϵ and frequency ω. The DC offset $\varphi_{\rm DC} = 0.269\pi$ is also the same as that used in experiments. We diagonalized H_0 in absence of the drive ($\epsilon = 0$) to obtain the dressed states in the dual rail subspace of two modes $|0\rangle_L = |\widetilde{100}\rangle, |1\rangle_L = |\widetilde{001}\rangle$. Here, the ket symbol follows the order $|\omega_b, \omega_c, \omega_s\rangle$. However, for numerical stability, we performed simulations in the rotating frame of the two modes and the coupler $U = \exp\{-i[\omega_b b^\dagger b + \omega_s s^\dagger s + \omega_c c^\dagger c)t]\}$ where $\omega_c = \omega_{c0}\sqrt{|\cos(\varphi_{\rm ext}(0))|}$ is the frequency of the coupler in the absence of RF modulation. The Hamiltonian is as follows:

$$\begin{aligned} H_{\rm rot} &= UH_0U^\dagger \\ &= [\omega_{c0}\sqrt{|\cos(\varphi_{\rm ext}(t))|} - \omega_c]a_c^\dagger a_c \\ &\quad + g_{cb}(e^{i\Delta_{cb}t}a_c^\dagger a_b + h.c.) + g_{cs}(e^{i\Delta_{cs}t}a_c^\dagger a_{si} + h.c.) \end{aligned} \tag{26}$$

Occupied modes	S_1	S_2	S_3	S_4	S_5	S_6	S_7
Active access BS_1		0.697% ±0.176%	0.178% ±0.039%	0.019% ±0.031%	0.162% ±0.02%	0.017% ±0.032%	0.068% ±0.019%
Active access BS_2	0.031% ±0.018%		1.111% ±0.248%	0.098% ±0.029%	0.049% ±0.023%	0.236% ±0.063%	0.205% ±0.034%
Active access BS_3	0.108% ±0.028%	0.086% ±0.054%		0.055% ±0.027%	0.03% ±0.031%	0.058% ±0.027%	0.067% ±0.045%
Active access BS_4	0.058% ±0.065%	0.088% ±0.06%	0.04% ±0.012%		0.064% ±0.048%	1.396% ±0.089%	0.125% ±0.065%
Active access BS_5	0.179% ±0.071%	0.058% ±0.032%	0.012% ±0.038%	0.068% ±0.044%		0.019% ±0.025%	0.056% ±0.049%
Active access BS_6	0.009% ±0.157%	0.257% ±0.107%	0.091% ±0.061%	0.321% ±0.122%	0.361% ±0.138%		0.122% ±0.045%
Active access BS_7	0.011% ±0.052%	0.618% ±0.018%	0.083% ±0.04%	0.05% ±0.026%	0.11% ±0.049%	0.014% ±0.025%	

Supplementary Table 6. RAQM state-dependent access error for storage-buffer swap.

Dephasing rate (kHz)	S_1	S_2	S_3	S_4	S_5	S_6	S_7
Inactive access BS_1		6.304 ±0.007	2.061 ±0.012	4.343 ±0.007	0.169 ±0.036	0.906 ±0.032	0.993 ±0.021
Inactive access BS_2	0.237 ±0.005		6.998 ±0.018	0.988 ±0.017	0.165 ±0.023	1.280 ±0.030	0.854 ±0.019
Inactive access BS_3	1.089 ±0.003	0.226 ±0.009		1.363 ±0.027	0.256 ±0.021	1.024 ±0.033	0.510 ±0.026
Inactive access BS_4	7.182 ±0.002	2.464 ±0.004	2.288 ±0.006		2.940 ±0.005	18.75 ±0.033	2.763 ±0.007
Inactive access BS_5	2.625 ±0.002	0.195 ±0.007	1.698 ±0.027	0.987 ±0.011		0.842 ±0.023	0.881 ±0.012
Inactive access BS_6	11.25 ±0.002	0.067 ±0.009	0.190 ±0.049	0.282 ±0.023	0.243 ±0.016		0.502 ±0.019
Inactive access BS_7	6.496 ±0.002	11.85 ±0.001	2.694 ±0.017	3.088 ±0.004	3.552 ±0.004	2.848 ±0.009	

Supplementary Table 7. RAQM spectator-access dephasing rate.

where the detunings are $\Delta_{ci} = \omega_c - \omega_i$.

To calibrate the beamsplitter gate $U_{\rm b.s}(0) = X/2$, we first selected the drive amplitude and frequency which maximized $\mathcal{F} = |\langle 0_L|1_L\rangle|^2$ after a $X/2 + X/2$ pulse sequence. Each $X/2$ is a flat top pulse with flat length $0.532\,\mu$s and a gaussian ramp of duration 15 ns and $\sigma = 5$ ns, which are the same parameters as that used in the experiment. Since the drive amplitude and frequency have a nonlinear relationship, we swept over both parameters to arrive at the optimal point ($\epsilon = 0.03505$, $\omega = 514.1731\,$MHz) which yielded $1 - \mathcal{F} < 10^{-5}$. Moreover, we added a global phase to convert the $\sqrt{\rm iSWAP}$ into $\sqrt{\rm SWAP}$ gate.

Finally, we performed a simulation of beam splitter Randomized Benchmarking [6] in the presence of decoherence. The density matrix of system evolves according to the Lindblad master equation as

$$\partial_t \rho = -i[H_{\rm rot}, \rho] + \sum_{i=b,c,s} \Gamma_{1,i}\mathcal{D}[a_i]\rho + \sum_{i=b,s} \Gamma_{2,i}\mathcal{D}[a_i^\dagger a_i]\rho \tag{27}$$

where $\mathcal{D}[L] = L\rho L^\dagger - (1/2)\{L^\dagger L, \rho\}$ is the Liouvillian superoperator, $\Gamma_{1,i} = 1/T_{1,i}$ is the decay rate and $(\Gamma_{2,i} = 1/T_{{\rm echo},i} - 1/2T_{1,i})$ is the pure dephasing rate. The decay and dephasing times were extracted from experiments. We did not include coupler dephasing as we weren't able to perform the Ramsey echo experiment on the coupler at $\phi_{\rm DC}$ due to large flux noise. We evaluated fidelities at 13 gate depths in the range $[1, 100]$ with 5 different sequences for each gate depth. The simulation results are shown in Table 10. These results show that while the raw fidelity of BS_2 swap gate is limited primarily by decay, its post-selected fidelity is limited by the pure dephasing of the two modes. Additionally, this simulation overestimates the gate infidelity by 25% due to numerical errors of the solver which make long simulations unfeasible.

X. ELECTROSTATIC DISCHARGE PROTECTION

Our coupler chip uses a sapphire substrate and lacks a ground plane, making it highly susceptible to Electro-

Storage mode	S_1	S_2	S_3	S_4	S_5	S_6	S_7
State-dependent access error	0.164%	0.245%	0.058%	0.255%	0.056%	0.167%	0.127%
Many-body dephasing	0.102%	0.046%	0.030%	0.026%	0.030%	0.028%	0.042%
Spectator-access dephasing	0.928%	0.705%	0.450%	0.324%	0.242%	0.734%	0.177%
Swaps	0.072%	0.051%	0.064%	0.132%	0.146%	0.089%	0.127%
Storage decay	0.559%	0.159%	0.245%	0.329%	0.525%	0.327%	0.493%
Storage dephasing	0.084%	0.054%	0.027%	0.038%	0.019%	0.020%	0.018%
Calculated random read infidelity	1.854%	1.181%	0.855%	1.019%	0.999%	1.309%	0.942%
Measured random read infidelity	1.171% ±0.041%	1.259% ±0.041%	1.112% ±0.042%	1.114% ±0.049%	1.232% ±0.066%	1.232% ±0.060%	1.192% ±0.056%

Supplementary Table 8. RAQM random read error budget.

RAQM size	Average random read infidelity
$[S_1]$	$0.192\% \pm 0.046\%$
$[S_2]$	$0.161\% \pm 0.037\%$
$[S_3]$	$0.299\% \pm 0.043\%$
$[S_4]$	$0.160\% \pm 0.035\%$
$[S_5]$	$0.473\% \pm 0.045\%$
$[S_6]$	$0.906\% \pm 0.051\%$
$[S_7]$	$0.590\% \pm 0.048\%$
$[S_1, S_2]$	$0.630\% \pm 0.025\%$
$[S_1, S_3]$	$0.612\% \pm 0.021\%$
$[S_1, S_4]$	$0.631\% \pm 0.023\%$
$[S_1, S_5]$	$1.130\% \pm 0.028\%$
$[S_1, S_6]$	$1.027\% \pm 0.025\%$
$[S_1, S_7]$	$0.588\% \pm 0.018\%$
$[S_2, S_3]$	$0.867\% \pm 0.021\%$
$[S_2, S_4]$	$0.545\% \pm 0.020\%$
$[S_2, S_5]$	$0.843\% \pm 0.025\%$
$[S_2, S_6]$	$0.975\% \pm 0.022\%$
$[S_2, S_7]$	$0.706\% \pm 0.019\%$
$[S_3, S_4]$	$0.619\% \pm 0.019\%$
$[S_3, S_5]$	$0.936\% \pm 0.028\%$
$[S_3, S_6]$	$0.936\% \pm 0.028\%$
$[S_3, S_7]$	$0.615\% \pm 0.049\%$
$[S_4, S_5]$	$0.572\% \pm 0.019\%$
$[S_4, S_6]$	$1.474\% \pm 0.040\%$
$[S_4, S_7]$	$0.755\% \pm 0.021\%$
$[S_5, S_6]$	$0.826\% \pm 0.020\%$
$[S_5, S_7]$	$1.143\% \pm 0.033\%$
$[S_6, S_7]$	$1.254\% \pm 0.027\%$
$[S_4, S_5, S_7]$	$0.783\% \pm 0.013\%$
$[S_1, S_2, S_3]$	$1.016\% \pm 0.017\%$
$[S_1, S_2, S_6]$	$0.993\% \pm 0.114\%$
$[S_1, S_2, S_3, S_4, S_5, S_6, S_7]$	$1.187\% \pm 0.019\%$

Supplementary Table 9. Average random read fidelity of different RAQM sizes.

static Discharge (ESD), significantly reducing fabrication yield. The coupler's dimensions (43×6 mm), the long flux line (~ 70 mm), and the large pads in the flux line exacerbate ESD issues, resulting in almost 0% yield of the coupler without ESD protections. The large pads can accumulate static charge during liftoff, and discharging generates a large current around the SQUID loop, damaging the coupler junctions.

To mitigate ESD issues, we employ two strategies: First, we increase the impedance of the high-impedance section, which broadens the filter stopband and reduces discharge current along the flux line. Second, we add three shorting lines around the coupler, creating a bypass circuit to protect the SQUID during fabrication. The shorting lines are carefully scratched with a clean tweezer during packaging, with the scratch size controlled to minimize the loss that could affect the coupler's T_1.

XI. DEVICE FABRICATION AND MEASUREMENT SETUP

Our Casaded random-accessed quantum memory has three components: the multimode flute cavity, the transmon chip, and the coupler chip.

The multimode cavity is fabricated from high-purity (99.9995%) Aluminum. Three rounds of chemical etching removed in total $\sim 200\mu$m surface using Transene Aluminum Etchant A.

For the transmon chip, we use a 430 μm thick C-plane HEMEX sapphire wafer annealed at 1200^oC for 2 hours as the substrate. 150 nm thick Tantalum film is sputtered at 800^oC as the ground plane. Large patterns, except Josephson junctions, are fabricated through photolithography with Heidelberg MLA 150 Direct Writer and fluorine-based dry-etching. The junction mask was fabricated with the Raith EBPG 5000+ E-Beam writer on a bi-layer resist (MMA EL11-950 PMMA A7). Transmon's Josephson junctions are of the Manhattan type. The junction mask was metalized in a Plassys electron beam evaporator with double-angle evaporation. The wafer was then diced and lifted off and packaged inside the multimode flute cavity.

For the coupler chip, we use a 530 μm thick EFG C-plane sapphire wafer annealed at 1200^oC for 2 hours as the substrate. 150 nm thick Niobium film is evaporated in Kurt J. Lesker E-beam Evaporator at room temperature as the ground plane. Large patterns, except Josephson junctions, are fabricated through photolithography with Heidelberg MLA 150 Direct Writer and Fluorine-based dry-etching. The junction mask was fabricated

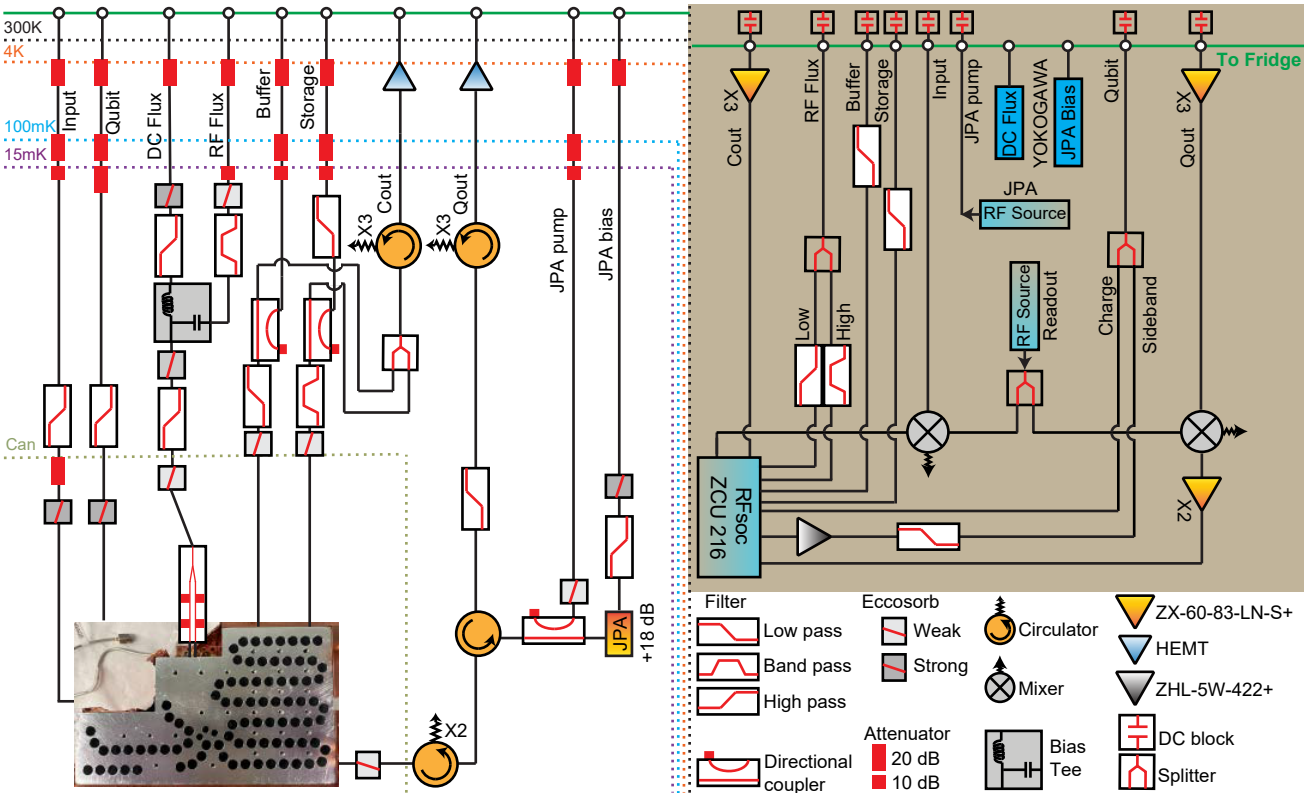
Errors	Raw Infidelity	Post Selected Infidelity
Dephasing $[\Gamma_{\phi,b}, \Gamma_{\phi,s}]$	$0.1095\% \pm 0.0019\%$	$0.0481\% \pm 0.0005\%$
Decay $[\Gamma_{1,b}, \Gamma_{1,s}, \Gamma_{1,c}]$	$0.4007\% \pm 0.0043\%$	$0.0039\% \pm 0.0001\%$
Decay and Dephasing $[\Gamma_{1,b}, \Gamma_{1,s}, \Gamma_{1,c}, \Gamma_{\phi,b}, \Gamma_{\phi,s}]$	$0.4408\% \pm 0.0052\%$	$0.0479\% \pm 0.0006\%$
Experiment	$0.35789\% \pm 0.00002\%$	$0.036890\% \pm 0.01520\%$

Supplementary Table 10. Master Equation Simulation of BS_2 swap gate using beam-splitter Randomized Benchmarking[6]

with Raith EBPG 5200+ E-Beam writer on a bi-layer resist (MMA EL13-950 PMMA A4). Coupler's Josephson junctions are Dolan bridge type. The junction mask was metalized in a Plassys electron beam evaporator with double-angle evaporation. The wafer was then diced and lifted off. Each shorting line structure mentioned in Supplementary Section X is scratched six times with an IPA-cleaned tweezer during packaging.

The multimode device is packaged inside a double layer μ-metal shielded sample can and installed inside a dilution fridge.

Supplementary Figure 12 shows the room and cryogenic temperature measurement chain. The device is mounted on the mixing chamber plate of a dilution refrigerator with a base temperature of 10 mK. All RF input signals are generated via the ZCU216 RFSoC board. Two DC sources (Yokogawa GS200) bias the DC flux of the coupler and the Josephson Parametric Amplifiers (JPA). One Signal Core SC5511A serves as the local oscillator (LO), split into two tones: one tone, after frequency up-conversion, functions as the readout pulse; the other is used for frequency down-conversion and is collected by the RFSoC ADC. A second Signal Core SC5511A provides the pump tone for the JPA, amplifying the qubit readout signal (labeled as Qout). The pulse sent to the qubit channel includes two components: the single-qubit drive and the qubit-buffer sideband drive, both synthesized directly through different RFSoC DAC channels. The sideband drive is further amplified with a ZHL-5W-422+ high-power amplifier at room temperature. Buffer and Storage modes are probed separately through a weakly coupled port machined at the multimode cavity; the signals are then combined and collected by the RFSoC ADC. Due to the RFSoC Nyquist frequency range, RF flux modulation is synthesized through two channels: low modulation frequency (0.1–1 GHz) and high modulation frequency (1–3 GHz). These two signals are combined at room temperature and then sent through the RF flux line, where they are combined with the DC flux at base temperature. The signal reaches the SMA connector, which is wire-bonded directly to the coupler chip. All RF lines have a DC block at the top of the fridge to avoid ground-loop, and all lines coming out of the device have Quantum Microwave eccosorbs placed as close as possible to the sample.



Supplementary Figure 12. Detailed cryogenic and room temperature measurement setup.

[1] S. E. Nigg, H. Paik, B. Vlastakis, G. Kirchmair, S. Shankar, L. Frunzio, M. H. Devoret, R. J. Schoelkopf, and S. M. Girvin, Black-box superconducting circuit quantization, Phys. Rev. Lett. **108**, 240502 (2012).

[2] L. Stefanazzi, K. Treptow, N. Wilcer, C. Stoughton, C. Bradford, S. Uemura, S. Zorzetti, S. Montella, G. Cancelo, S. Sussman, A. Houck, S. Saxena, H. Arnaldi, A. Agrawal, H. Zhang, C. Ding, and D. I. Schuster, The qick (quantum instrumentation control kit): Readout and control for qubits and detectors, Review of Scientific Instruments **93**, 10.1063/5.0076249 (2022).

[3] S. Chakram, K. He, A. V. Dixit, A. E. Oriani, R. K. Naik, N. Leung, H. Kwon, W.-L. Ma, L. Jiang, and D. I. Schuster, Multimode photon blockade, Nature Physics **18**, 879 (2022).

[4] A. E. Oriani, F. Zhao, T. Roy, A. Anferov, K. He, A. Agrawal, R. Banerjee, S. Chakram, and D. I. Schuster, Niobium coaxial cavities with internal quality factors exceeding 1.5 billion for circuit quantum electrodynamics (2024), arXiv:2403.00286 [quant-ph].

[5] K. S. Chou, J. Z. Blumoff, C. S. Wang, P. C. Reinhold, C. J. Axline, Y. Y. Gao, L. Frunzio, M. H. Devoret, L. Jiang, and R. J. Schoelkopf, Deterministic teleportation of a quantum gate between two logical qubits, Nature **561**, 368 (2018).

[6] Y. Lu, A. Maiti, J. W. O. Garmon, S. Ganjam, Y. Zhang, J. Claes, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, High-fidelity parametric beamsplitting with a parity-protected converter, Nature Communications **14**, 5767 (2023).

[7] A. V. Dixit, S. Chakram, K. He, A. Agrawal, R. K. Naik, D. I. Schuster, and A. Chou, Searching for dark matter with a superconducting qubit, Phys. Rev. Lett. **126**, 141302 (2021).

[8] M. Ganzhorn, G. Salis, D. J. Egger, A. Fuhrer, M. Mergenthaler, C. Müller, P. Müller, S. Paredes, M. Pechal, M. Werninghaus, and S. Filipp, Benchmarking the noise sensitivity of different parametric two-qubit gates in a single superconducting quantum computing platform, Phys. Rev. Res. **2**, 033447 (2020).

[9] B. J. Chapman, S. J. de Graaf, S. H. Xue, Y. Zhang, J. Teoh, J. C. Curtis, T. Tsunoda, A. Eickbusch, A. P. Read, A. Koottandavida, S. O. Mundhada, L. Frunzio, M. Devoret, S. Girvin, and R. Schoelkopf, High-on-off-ratio beam-splitter interaction for gates on bosonically encoded qubits, PRX Quantum **4**, 020355 (2023).

[10] F. Valadares, N.-N. Huang, K. T. N. Chu, A. Dorogov, W. Chua, L. Kong, P. Song, and Y. Y. Gao, On-demand transposition across light-matter interaction regimes in bosonic cqed, Nature Communications **15**, 5816 (2024).

[11] Z. Li, T. Roy, D. Rodríguez Pérez, K.-H. Lee, E. Kapit, and D. I. Schuster, Autonomous error correction of a single logical qubit using two transmons, Nature Communications **15**, 1681 (2024).

[12] Z. Li, T. Roy, Y. Lu, E. Kapit, and D. I. Schuster, Autonomous stabilization with programmable stabilized state, Nature Communications **15**, 6978 (2024).

[13] Y. Lu et al., Systematic construction of time-dependent Hamiltonians for microwave-driven Josephson circuits, in preparation.

[14] F. Zhao, Z. Li, A. V. Dixit, T. Roy, A. Vrajitoarea, R. Banerjee, A. Anferov, K.-H. Lee, D. I. Schuster, and A. Chou, A flux-tunable cavity for dark matter detection (2025), arXiv:2501.06882 [quant-ph].

[15] M. Kirschning and R. Jansen, Accurate wide-range design equations for the frequency-dependent characteristic of parallel coupled microstrip lines, IEEE Transactions on Microwave Theory and Techniques **32**, 83 (1984).

[16] V. Trifunovic and B. Jokanovic, Four decade bandwidth uniplanar balun, Electronics Letters **28**, 534 (1992).

[17] V. Trifunovic and B. Jokanovic, Review of printed marchand and double y baluns: characteristics and application, IEEE Transactions on Microwave Theory and Techniques **42**, 1454 (1994).

[18] J. B. Venkatesan and W. R. Scott, Investigation of the double-y balun for feeding pulsed antennas, in SPIE Defense + Commercial Sensing (2003).

[19] J. Johansson, P. Nation, and F. Nori, Qutip: An open-source python framework for the dynamics of open quantum systems, Computer Physics Communications **183**, 1760–1772 (2012).