High-efficiency measurement of an artificial atom embedded in a parametric amplifier

A. Eddins,1,2 J.M. Kreikebaum,1,2 D.M. Toyli,1,2 E.M. Levenson-Falk,1,2 A. Dove,1,2 W.P. Livingston,1,2 B.A. Levitan,3 L.C.G. Govia,4 A.A. Clerk,4 and I. Siddiqi1,2

1 Quantum Nanoelectronics Laboratory, Department of Physics, University of California, Berkeley CA 94720, USA.
2 Center for Quantum Coherent Science, University of California, Berkeley CA 94720, USA.
3 Department of Physics, McGill University, Montreal, Quebec H3A 2T8, Canada
4 Institute for Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA

(Dated: June 15, 2018)

A crucial limit to measurement efficiencies of superconducting circuits comes from losses involved when coupling to an external quantum amplifier. Here, we realize a device circumventing this problem by directly embedding a two-level artificial atom, comprised of a transmon qubit, within a flux-pumped Josephson parametric amplifier. Surprisingly, this configuration is able to enhance dispersive measurement without exposing the qubit to appreciable excess backaction. This is accomplished by engineering the circuit to permit high-power operation that reduces information loss to unmonitored channels associated with the amplification and squeezing of quantum noise. By mitigating the effects of off-chip losses downstream, the on-chip gain of this device produces end-to-end measurement efficiencies of up to 80%. Our theoretical model accurately describes the observed interplay of gain and measurement backaction, and delineates the parameter space for future improvement. The device is compatible with standard fabrication and measurement techniques, and thus provides a route for definitive investigations of fundamental quantum effects and quantum control protocols.

I. INTRODUCTION

The sum of interactions between a quantum system and all environmental channels produces a continuous flow of quantum information into the environment, causing dephasing at a rate $\Gamma_\phi$. For a two-level qubit described by $\sigma_z$ and measured along that axis, one may define the fraction of this information flux experimentally captured per unit time to be the measurement efficiency $\eta_{\text{meas}} = \Gamma_{\text{meas}}/2\Gamma_\phi$, a critical parameter for continuous quantum measurements, where $\Gamma_{\text{meas}}$ represents the rate at which the experimentalist learns about $\sigma_z$ and is defined such that $\eta_{\text{meas}}$ ranges from 0 to 1.

Maximizing this efficiency for superconducting qubit measurements requires multiple stages of cryogenic amplification to boost information-bearing quantum microwave above the noise floor of room-temperature electronics. The use of off-chip superconducting parametric amplifiers for the first gain stage has enabled a variety of experiments investigating quantum measurement dynamics [1–6], with improvements in efficiency reported using multi-junction circuits [7]. However, prior to amplification the information encoded in the field is extremely fragile, such that in these configurations $\sim$30% of the information is dissipated in lossy microwave circulators and other components en route to the amplifier, lowering the ceiling on $\eta_{\text{meas}}$. Interest in surmounting this limitation has helped spur recent progress in the development of superconducting circulators and directional amplifiers [8–14]. Alternatively, high efficiency may be realized by strongly measuring a second, ancillary qubit and resonator mode (e.g. [15]).

Here we develop a minimal circuit architecture providing on-chip parametric gain through integration of a standard Josephson parametric amplifier (JPA) with the qubit in a configuration we dub the Qubit Parametric Amplifier (QPA), removing virtually all pre-amplification loss. Previous demonstrations with on-chip amplifiers have leveraged the bifurcation dynamics of a nonlinear resonator [16–18]. In contrast, the QPA implements on-chip the parametric mode of operation that has been widely applied in continuous measurements of qubits. This scheme presents a novel challenge, as the in-situ microwave amplification and squeezing opens a parasitic measurement channel inducing excess dephasing. We model and characterize this backaction in detail, and successfully mitigate it via a weakly nonlinear design permitting fast measurement, producing steady-state efficiencies as high as $\eta_{\text{meas}} = 0.80$ with direction for further improvement.

A schematic of our experiment appears in Fig. 1(a). The QPA consists of a transmon qubit [19] dispersively coupled to a JPA. A microwave readout tone at frequency $\omega_{\text{QPA}}$ reflects off the QPA, acquiring qubit-state information. A pump tone of the form $\cos(2(\omega_{\text{QPA}}t + \Phi))$ applied to the pump port of the QPA concurrent with the readout modulates the QPA resonance frequency, producing on-chip phase-sensitive amplification of the measurement field. Adjusting the phase of the pump tone relative to the readout tone changes which field quadrature is amplified and which is squeezed. The output of the QPA is then routed by microwave circulators to additional amplification stages including a second, off-chip JPA and a superconducting Josephson Traveling Wave Parametric Amplifier (JTWPA) [20] en route to room-temperature demodulation and digitization. By acting as a phase-sensitive preamplifier before the JTWPA, which necessarily adds at least half a photon of noise in standard phase-preserving operation, the off-chip JPA reduces the amount of on-chip gain required for high efficiency.
A circuit diagram and false-color photographs of the QPA are shown in Fig. 1(b,c). The on-chip JPA design is similar to that of some off-chip JPs [21,22], consisting of an interdigitated capacitor in parallel with a combination of geometric and Josephson inductance to form an LC resonator (purple) whose frequency $\omega_{\text{QPA}}$ tunes with the flux applied through the pair of SQUID loops. A superconducting coil housed below the chip enables static tuning of $\omega_{\text{QPA}}$, while a pump applied via the flux-line (cyan) modulates $\omega_{\text{QPA}}$ to produce parametric gain. Some variation in circuit parameters occurred as data were acquired over the course of multiple cooldowns; we give representative parameter values here, and list precise values for each dataset in Appendix A.2. The QPA resonator has a zero-flux frequency of $\omega_{\text{QPA, max}}/2\pi = 6.970$ GHz; we tuned this down to $\omega_{\text{QPA}}/2\pi \leq 6.740$ GHz to increase the modulation amplitude produced by the flux-pump. Coupling capacitors and a 180° microwave hybrid couple the resonator to the readout transmission line with an effective $\kappa_{\text{ext}}/2\pi = 25.7$ MHz $\gg \kappa_{\text{int}}/2\pi$. The transmon qubit (red) resonates at $\omega_q/2\pi = 4.271$ GHz and is capacitively coupled to the on-chip JPA with dispersive interaction strength $\chi/2\pi = 1.9$ MHz, with the convention that the AC Stark shift changes $\omega_q$ by $2\chi n$. The paddle design of the qubit is chosen to reduce loss due to electromagnetic participation of the surface-vacuum interface, and a floating radiation shield (white) suppresses radiative decay of the qubit into other environmental modes. The measured lifetime $T_1 = 4.2(8) \mu$s is near the expected Purcell-decay limited value $T_1 \approx 6 \mu$s, which could be improved in future designs via integration of a Purcell filter [23].

In the dispersive approximation and in the frame rotating at $\omega_{\text{QPA}}$, the internal QPA dynamics can be described by the Hamiltonian

$$\hat{H}_{\text{QPA}} \approx \frac{\hbar}{2}(\Delta + 2\chi(\hat{a}^{\dagger}\hat{a} + 1/2))\hat{\sigma}_z + \frac{i\lambda}{2}(\hat{a}^{\dagger2} - \hat{a}^2), \quad (1)$$

with $\Delta = \omega_q - \omega_{\text{QPA}}$. The first of the two terms is the familiar dispersive Hamiltonian that also describes the more common case of readout using a linear resonator. The second term describes the on-chip, phase-sensitive gain process, where $\lambda$ is set by the flux-pump strength and would equal the rate of squeezing if there were no dissipation ($\kappa = 0$). A succinct theoretical analysis of the system is given in Appendix B with further details available in [24].

II. MEASUREMENT BACKACTION WITH ON-CHIP GAIN

As in conventional qubit measurement setups, the dispersive interaction encodes information about the $\hat{\sigma}_z$ component of the qubit state into the mean value of one quadrature of the output field, which we refer to as the signal quadrature $Q$. During this measurement the qubit is dephased at a rate $\Gamma \varphi \geq \Gamma_{\varphi, QL}$, where $\Gamma_{\varphi, QL}$ represents quantum-limited backaction. High $\eta_{\text{meas}}$ requires this inequality to be nearly saturated. Here, however, on-chip amplification drives a second, parasitic measurement process in which $\hat{\sigma}_z$ information is encoded in other statistical moments of the output field. This dephasing mechanism is predicted to be independent of the mean field
in the resonator, making it distinct from effects in resonantly current-pumped systems \[25^{–30}\]. A rough heuristic model describes the parasitic measurement in two steps: the phase-sensitive on-chip gain squeezes the microwave vacuum noise, and the resultant output squeezed state is rotated in phase by the dispersive interaction, encoding \(\hat{\sigma}_z\) information in the covariance of the output-field quadratures. These moments are largely not detected downstream, in part due to the fragility of the moments with respect to losses, and in part because the phase-sensitive following JPA typically deamplifies this information. As the parasitic measurement increases \(\Gamma_\phi\) without increasing the room-temperature SNR, it lowers \(\eta_{\text{meas}}\).

Starting from Eq. 1, one can derive an expression for the parasitic dephasing rate (Appendix B, Eq. 2):

\[
\Gamma_{\phi, \text{parasitic}} = \frac{1}{2} \text{Re} \left( \sqrt{D(-\lambda)} + \sqrt{D(\lambda)} \right) - \frac{\kappa}{2} - 1/T_2^*.
\]

Here \(\lambda\) is related to the on-chip gain by

\[
\lambda = \frac{\kappa \sqrt{G_{\text{QPA}} - 1}}{2 \sqrt{G_{\text{QPA}} + 1}}.
\]

we have defined \(D(\lambda) = (\kappa/2 + \lambda + i\chi)^2 - 2i\chi\lambda\), and \(T_2^*\) is an empirical parameter describing dephasing absent any applied drives. Several metrics are available to parameterize the gain dynamics; \(G_{\text{QPA}}\) indicates the (phase-preserving) power gain experienced by a tone slightly detuned from \(\omega_{\text{QPA}}\), which we measure directly using a vector network analyzer.

We characterize \(\Gamma_{\phi, \text{parasitic}}\) via Ramsey oscillations of the qubit simultaneous with on-chip gain for several values of \(G_{\text{QPA}}\). Absent any gain (Fig. 3(a)), we observe \(1/T_2^* = 0.23(7)\ \mu\text{s}^{-1}\). Applying pump power produces squeezed vacuum inside the QPA, causing \(\Gamma_{\phi}\) to increase significantly (Fig. 3(b,c)). The inset ellipses of Fig. 3 indicate the quadrature variances and covariance (kurtosis is not shown) of the QPA output field predicted using equations from \[24\] and experimental parameters, colored red or blue depending on the qubit state. Increasing \(G_{\text{QPA}}\) decreases the overlap of the ellipses, speeding up the parasitic measurement. We find good agreement with the predictions of our zero-free parameter model over a range of \(G_{\text{QPA}}\) values as plotted in Fig. 3(d), supporting the validity of the model and indicating the absence of any comparable additional dephasing mechanism for these operating conditions.

A second set of Ramsey experiments illuminates how varying on-chip gain modifies backaction during an applied weak measurement. Our theory analysis (Appendix B, Eq. 2) predicts the total dephasing to vary according to

\[
\Gamma_{\phi} = \frac{2\chi^2\kappa^2 P_{\text{in}}}{\hbar \omega_{\text{QPA}}} \left( \frac{\cos^2 \Phi}{|D(-\lambda)|^2} + \frac{\sin^2 \Phi}{|D(\lambda)|^2} \right) + \Gamma_{\phi, \text{parasitic}},
\]

where \(P_{\text{in}}\) is the power of the measurement tone incident to the QPA, and \(\Gamma_{\phi, \text{parasitic}}\) is, notably, still given by Eq. 2. Absent any on-chip gain (\(\lambda = 0\)), Eq. 4 can be approximated by the more standard expression \(\Gamma_{\phi} = 8\chi^2 \tilde{n}/\kappa + O(\tilde{\chi}^4)\), describing dephasing induced as \(\hat{\sigma}_z\) information is encoded in the phase of the QPA output field (Fig. 3(a), theory). Experimentally, with \(G_{\text{QPA}} = 0\ \text{dB}\) we observe dephasing at rate \(\Gamma_{\phi} = 0.49\ \mu\text{s}^{-1}\) (Fig. 3(d)), from which we infer \(P_{\text{in}} = -142\ \text{dBm}\) for this choice of drive. Keeping \(P_{\text{in}}\) fixed, we apply a flux-pump such that \(G_{\text{QPA}} = 3\ \text{dB}\). Fig. 3(b,c) show the expected output fields when the on-chip gain is aligned with \(\Phi = 0\) or orthogonal to \(\Phi = \pi/2\) the signal quadrature. The signal size \((\langle Q_z \rangle - \langle Q_y \rangle)\) is nearly constant in all three cases:
since the input measurement drive lies along \( I \) while the output signal lies along \( Q \), the net effects of amplification and deamplification approximately cancel. In contrast, the noise fluctuations do get amplified (squeezed), such that the SNR at the QPA output depends on \( \Phi \). Since amplifying the signal quadrature squeezes the photon number fluctuations in the conjugate quadrature which cause dephasing, we expect amplifier mode (\( \Phi = 0 \)) to minimize \( \Gamma_\phi \) and squeezing mode (\( \Phi = \pi/2 \)) to maximize it, in agreement with the comparison of Figs. 3(e,f). Results of additional Ramsey measurements shown in Figure 3(g) reveal the full dependence of \( \Gamma_\phi \) on \( \Phi \) and \( G_{\text{QPA}} \), and verify the predictions of our theory model (Eq. 4).

We note that squeezing mode is also of interest as a means of improving SNR by reducing the quantum fluctuations of the output measurement field [21, 31, 32]; similar in-situ squeezing generation has been demonstrated in a recent optical experiment [33].

We henceforth focus exclusively on amplifier mode (\( \Phi = 0 \)). The primary benefit of amplifier mode is that the noise floor of the QPA output signal quadrature is increased with minimal information loss, which enables greater overall efficiency by making the SNR insensitive to noise added downstream. A secondary effect is the deamplification of the mean field without deamplification of the signal; an interesting question is whether this effect, perhaps combined with injected orthogonally-squeezed vacuum, might enable a greater dispersive signal size for fixed mean intra-resonator photon number.

### III. MEASUREMENT EFFICIENCY

The total measurement efficiency is the product of on-chip efficiency and the efficiency of the rest of the measurement chain: \( \eta_{\text{meas}} = \eta_{\text{QPA}} \eta_{\text{rest}} \). Increasing on-chip gain, \( G_{\text{QPA}} \), increases \( \eta_{\text{rest}} \) as the amplified signal quadrature becomes robust to losses, but decreases \( \eta_{\text{QPA}} \) due to the parasitic measurement discussed in Section II such that there is an optimal \( G_{\text{QPA}} \) value maximizing \( \eta_{\text{meas}} \) for a given measurement drive. We can write the efficiency as the ratio of empirical quantities, \( \eta_{\text{meas}} = \Gamma_{\text{meas}}/2\Gamma_\phi \).

Here \( \Gamma_{\text{meas}} = \frac{1}{4} \text{SNR}^2/4 \) is the rate at which the square of the room-temperature voltage SNR increases with integration time, or equivalently the rate at which \( \sigma_z \) information is acquired by our digitizer.

In order to determine steady-state \( \eta_{\text{meas}} \) as a function of \( G_{\text{QPA}} \) and measurement drive \( |\alpha_{\text{in}}| = \sqrt{P_{\text{in}}/\hbar \omega_{\text{QPA}}} \), we extracted both \( \Gamma_{\text{meas}} \) and \( \Gamma_\phi \) using the pulse sequence shown in Fig. 4(a). The flux-pump was switched on in advance such that the mean intra-QPA field was deamplified along \( I \) at all times, reducing the intra-QPA circulating power produced by a strong measurement drive and thus helping to minimize undesired nonlinear processes in the QPA. Next, a continuous measurement tone was turned on and sufficient time allowed to pass to ensure the cavity had stabilized before a \( \pi/2 \) pulse was applied to initiate Ramsey evolution. After the second \( \pi/2 \), \( |\alpha_{\text{in}}| \) was increased to perform projective readout. The ensemble-averaged readout results for variable Ramsey evolution time were used to determine the dephasing rate \( \Gamma_\phi \), as in the right inset of Fig. 4(a). Integrating the steady-state weak-measurement record from the longest Ramsey evolution yields histograms (left inset) from which we determine \( \text{SNR}(t) \) and thus \( \eta_{\text{meas}} \).

(b) Measurement efficiency (\( \eta_{\text{meas}} \))

(c) \( \eta_{\text{rest}} \)

(d) \( \eta_{\text{QPA}} \)

FIG. 4. (a) Ramsey sequence used to determine \( \eta_{\text{meas}} \). Continuous measurement of the qubit causes dephasing at rate \( \Gamma_\phi \) (right inset). Sampling the output measurement field during the Ramsey evolution yields histograms (left inset) from which we determine \( \text{SNR}(t) \) and thus \( \eta_{\text{meas}} \). (b) Inferred \( \eta_{\text{meas}} \) values as a function of \( G_{\text{QPA}} \) and continuous measurement amplitude, \( |\alpha_{\text{in}}| \). Settings marked with an asterisk produced a spurious third histogram peak containing > 1% of the total events, which we attribute to large intra-QPA fields inducing transitions to the third transmon level. The dashed orange box surrounds the optimal operating region averaging \( \eta_{\text{meas}} = 80\% \). (c) The on-chip efficiency \( \eta_{\text{QPA}} \), inferred by comparing \( \Gamma_\phi \) and \( \Gamma_\phi_{\text{parasitic}} \), decreases with \( G_{\text{QPA}} \) and increases with \( |\alpha_{\text{in}}| \). (d) The off-chip efficiency, inferred by dividing \( \eta_{\text{QPA}} \) by \( \eta_{\text{QPA}} \), increases with \( G_{\text{QPA}} \) and remains roughly constant for the shown \( |\alpha_{\text{in}}| \) values. Horizontal lines mark mean values to guide the eye.
circulators and other off-chip components. The bottom edge of the box is defined by the decrease in \( \eta_{\text{meas}} \) associated with \( \Gamma_\phi, \text{parasitic} \) becoming a larger fraction of \( \Gamma_\phi \) as \( |\alpha_{\text{in}}| \) is decreased. The other two sides of the box are marked by the onset of non-ideal behavior evidenced by a third peak appearing in the measurement histograms. This spurious peak is not fully understood, but seems to involve population of the third transmon level driven by large intra-QPA photon numbers occurring at too low or too high \( G_{\text{QPA}} \) (corresponding to a large mean field or a large field variance, respectively) or at too high \( |\alpha_{\text{in}}| \).

With the assumption that \( \Gamma_\phi, \text{parasitic} \) remains independent of \( |\alpha_{\text{in}}| \) at this operating point, we can express the on-chip efficiency \( \eta_{\text{QPA}} \) in terms of empirical dephasing rates as \( \eta_{\text{QPA}} = 1 - \Gamma_\phi, \text{parasitic}/\Gamma_\phi \), from which we calculate the values shown in Fig. 4(c). In the absence of on-chip gain, \( \eta_{\text{QPA}} \) approaches unity as \( \Gamma_\phi, \text{parasitic} = 1/T_2 \ll \Gamma_\phi \). As gain is increased, \( \Gamma_\phi, \text{parasitic} \) increases, resulting in lower \( \eta_{\text{QPA}} \): as the measurement strength is increased, the parasitic dephasing becomes less significant, increasing \( \eta_{\text{QPA}} \). Calculating further, we can divide these values by the \( \eta_{\text{meas}} \) values in Fig. 4(b) to estimate \( \eta_{\text{rest}} \), shown in Fig. 4(d). This plot is restricted to lower-power operating conditions in which the device was better behaved; over this domain, we infer that information loss downstream of the QPA decreases with \( G_{\text{QPA}} \) and is approximately independent of \( |\alpha_{\text{in}}| \), supporting our previous assumption that \( \Gamma_\phi, \text{parasitic} \) is likewise independent of \( |\alpha_{\text{in}}| \). It is encouraging that the calculated values of \( \eta_{\text{rest}} \) approach 1 at \( G_{\text{QPA}} = 4 \text{ dB} \), though the current device did not permit increasing \( |\alpha_{\text{in}}| \) sufficiently to maximally benefit from this much gain. We expect this dynamic range ceiling, and thus \( \eta_{\text{meas}} \), may be raised by reducing \( \chi/\kappa \) and/or increasing the number of JPA SQUIDs [35, 36], suppressing deleterious Kerr effects not included in our model.

IV. CONCLUSION

We have characterized the measurement backaction on a qubit dispersively coupled to a parametric amplifier flux-pumped for gain, demonstrated how on-chip gain can mitigate off-chip sources of information loss, and observed steady-state efficiency \( \eta_{\text{meas}} \) up to 80%. Going forward, incremental improvements in \( \eta_{\text{meas}} \) may be achieved by further weakening device nonlinearity as discussed above to permit a larger measurement drive \( |\alpha_{\text{in}}| \) and thus greater \( \eta_{\text{QPA}} \). A more dramatic improvement might be realized by probing the device stroboscopically. In a linear readout resonator, stroboscopic measurement has been shown to eliminate the undesired squeezing rotations caused by dispersive coupling [37]; realizing the analogous effect in the QPA would close the parasitic dephasing channel such that increasing \( G_{\text{QPA}} \) would boost \( \eta_{\text{rest}} \) without degrading \( \eta_{\text{QPA}} \), even for small \( |\alpha_{\text{in}}| \). Other potential near-term experiments include exploration of effects of on-chip gain on initial transients when switching on a measurement, and an investigation of whether combining amplifier mode with injected orthogonally-squeezed vacuum enables greater dispersive readout SNR for a fixed intra-resonator photon number. Longer term, we envision the QPA to be an enabling technology for applications demanding signal-to-noise ratios approaching the quantum limit, such as measurement-based quantum feedback [38–40] or further studies of individual quantum trajectories, perhaps with extension to multi-qubit experiments via a chip layout similar to [17] followed by a broadband off-chip amplifier.

ACKNOWLEDGMENTS

The authors thank R. Vijay, S. Hacohen-Gourgy, E. Flurin, and L. Martin for useful discussions, and thank MIT Lincoln Labs for fabrication of the JTWPA with support from the LPS and IARPA. Work was supported by the Army Research Office under Grant No. W911NF-14-1-0078. A.A.C. and L.C.G.G. acknowledge support from the AFOSR MURI FA9550-15-1-0029. A.E. acknowledges support from the Department of Defense through the NDSEG fellowship program.

Appendix A: Experimental Methods

1. Device fabrication

The device used for this paper is patterned on > 8000 \( \Omega \)-cm intrinsic Si using photolithography and subsequent plasma etching of 100 nm thick e-beam evaporated Al. Al/AlO\(_x\)/Al junctions for the transmon and SQUIDs are defined in separate steps with e-beam lithography and subsequent double-angle evaporation. Adhering to the constraints of the pre-defined bond pads and QPA digitated capacitor, we were able to reduce the radiative loss of the qubit by introducing electrically-floating metal shielding around the qubit capacitor paddles (white in Fig. 1(c)). The chip is mounted on OFHC copper enclosed in an Al package, surrounded by cryoperm, and mounted to the base stage of a dilution refrigerator at \( \leq 35 \text{ mK} \). A copper wire fed through a high aspect-ratio hole in the aluminum helps thermalize the interior OFHC mount to the base stage. Flux to tune the QPA was applied with an off-chip coil.

2. Circuit parameters

Circuit parameters changed slightly after thermal cycling, and also as the QPA was flux-biased to different operating frequencies \( \omega_{\text{QPA}} \). The table lists precise parameter information corresponding to the three data figures.
of the main text. The change in $\chi$ between cooldowns is not well understood.

|                      | Fig. 2 | Fig. 3 | Fig. 4 |
|----------------------|--------|--------|--------|
| Cooldown             | B      | A      | B      |
| $\omega_{\text{QPA}}/2\pi$ (GHz) | 6.740  | 6.740  | 6.700  |
| $\kappa/2\pi$ (MHz) | 25.4   | 25.7   | 28.6   |
| $\omega_0/2\pi$ (GHz) | 4.274  | 4.274  | 4.271  |
| $\chi/2\pi$ (MHz)   | 1.9    | 1.7    | 2.0    |

3. Detailed wiring diagram

Figure 5 shows a complete diagram of the experiment. An experimental overview appears in (a), with individual subsystems detailed in (b) and a component legend given in (c). Most centrally, a microwave generator at $\omega_{\text{QPA}}$ (estimated by measuring the QPA resonance frequency while Rabi-driving the qubit) is split to drive four phase-sensitive processes: dispersive qubit measurement, on-chip amplification in the QPA, off-chip amplification in the JPA, and room-temperature demodulation of the output measurement signal. The amplification processes are highly sensitive to changes in applied pump power, such that the small change in insertion-loss associated with adjusting the phase of a phase-shifter would problematically change the amplifier gain. Several technical solutions are possible; for the QPA, we use a spectrum analyzer to ensure power-flatness as we programatically step the phase of the flux-pump. For the JPA, we change the effective amplification phase by phase-shifting all other tones at $\omega_{\text{QPA}}$ while leaving the JPA pump unchanged. To realize amplifier-mode operation of the QPA, the phase of the QPA pump was first chosen to minimize $\Gamma_\phi$, and then the phase of the off-chip JPA pump was adjusted to maximize SNR. Stark shifts were measured in advance for all $G_{\text{QPA}}, |\alpha_{in}|$ settings used to produce the data in Fig. 4 and qubit pulse frequencies and amplitudes were programatically adjusted accordingly at each setting. Small superconducting coils (not shown) were used to apply dc-flux biases to the QPA and JPA to tune their resonance frequencies. A vector network analyzer was used to characterize device resonance frequencies and gains. Cross-talk effects of the JPA flux-pump on the JTWPA became apparent at high JTWPA gain despite the intermediary low-pass filter, degrading JTWPA performance. These effects were suppressed by operating the JTWPA at reduced gain ($\sim 15$ dB).

4. $T_1$ vs on-chip gain

We briefly investigated the effect of on-chip gain on the qubit lifetime $T_1$. Results are shown in Fig. 4. The data suggest a small decrease in $T_1$ when gain is turned on, with no clear dependence as gain is further increased.

5. Amplifier-mode dephasing vs $P_{in}$

Extending the measurement-backaction data presented in Fig. 3, we fixed the QPA pump phase to operate the QPA in amplifier mode ($\Phi = 0$) and recorded $\Gamma_\phi$ for variable on-chip gain and measurement strength $|\alpha_{in}|$. The results are displayed alongside the theory prediction of Eq. 4 in Fig. 6. Good agreement is seen for low measurement strength and on-chip gain. At high drive strengths or high gains, excess dephasing is observed intermittently, i.e. for some experimental executions. At intermediate drive strength, near the center of the plot, this undesired behavior appears to be reduced as $G_{\text{QPA}}$ increases from $\sim 1$ dB to $\sim 2 - 3$ dB.

Appendix B: Theoretical derivations

1. Dephasing with on-chip gain

Our goal is to derive an analytic expression for the qubit dephasing rate in the long-time limit, without as-
\[ \dot{\hat{\rho}} = -i \left[ \hat{H}, \hat{\rho} \right] + \kappa D [\hat{a}] \hat{\rho} + \frac{1}{T_1} D [\hat{\sigma}_-] \hat{\rho} + \frac{1}{2T_2} D [\hat{\sigma}_z] \hat{\rho}, \]  
\[ \hat{H} = \frac{i\lambda}{2} (\hat{a}^\dagger \hat{a}^2 - \hat{a}^2) + \chi \hat{a}^\dagger \hat{a} \hat{\sigma}_z + \sqrt{\kappa} \left( \alpha_{in} \hat{a}^\dagger + \alpha_{in}^* \hat{a} \right), \]  

assuming that the dispersive coupling $\chi$ is weak. We start

\[ \hat{\rho} = \sum_{\mu, \nu \in \{\uparrow, \downarrow\}} \hat{\rho}_{\mu\nu} \otimes |\mu\rangle\langle\nu| \]  

with the master equation

\[ \hat{\rho} = \text{QPA dynamics of } \hat{H}_{\text{QPA}} \text{ from Eq. 1, and the coherent measurement drive on the QPA resonator, characterized by the drive amplitude } \alpha_{in}. \]

The dephasing rate quantifies the decay of the qubit coherence in the long time limit, described by the decay of the qubit off-diagonal matrix elements of the full density matrix. If we write the full density matrix as

\[ \hat{\rho} = \sum_{\mu, \nu \in \{\uparrow, \downarrow\}} \hat{\rho}_{\mu\nu} \otimes |\mu\rangle\langle\nu| \]  

where $\hat{\rho}_{\mu\nu}$ is an operator on the resonator Hilbert space and $|\downarrow\rangle$ and $|\uparrow\rangle$ are the ground and excited states of the qubit, then the qubit dephasing rate is fully captured by the evolution of the part of the density matrix proportional to $|\uparrow\rangle\langle\downarrow|$ (or its Hermitian conjugate). Thus, we
are interested in the evolution of the operator $\hat{\rho}_{\uparrow\downarrow}$. As is standard, we define the dephasing rate as
\begin{equation}
\Gamma_\Phi = \lim_{t \to \infty} \frac{-\ln(\text{Tr}[\hat{\rho}_{\uparrow\downarrow}(t)])}{t},
\end{equation}
which captures the exponential decay of the qubit coherence in the long-time limit.

From Eq. (B3) we calculate the evolution equation for the operator $\hat{\rho}_{\uparrow\downarrow}$:
\begin{equation}
\dot{\hat{\rho}}_{\uparrow\downarrow} = \left[ \frac{\lambda}{2} \left( \hat{a}^\dagger \hat{a}^\dagger - \hat{a} \hat{a} \right) - \sqrt{\kappa} (\hat{\alpha}_\text{in} \hat{a}^\dagger - \hat{\alpha}_\text{in}^* \hat{a}) , \hat{\rho}_{\uparrow\downarrow} \right] 
- i\chi \{ \hat{a}^\dagger \hat{a} , \hat{\rho}_{\uparrow\downarrow} \} + \kappa D[\hat{a}] \hat{\rho}_{\uparrow\downarrow} - \left( \frac{1}{2T_1} + \frac{1}{2T_2} \right) \hat{\rho}_{\uparrow\downarrow}.
\end{equation}

Note that this equation is not trace preserving, as it does not describe evolution of a valid density matrix. Extending beyond the results of Ref. [24], we have included the effect of qubit relaxation ($T_1$) in the evolution of $\hat{\rho}_{\uparrow\downarrow}$, and while this means the evolution of the qubit is no longer QND, the resulting equation for $\hat{\rho}_{\uparrow\downarrow}$ remains closed on itself, and can be solved analytically.

The first step in solving Eq. (B5) is to remove the exponential decay caused by the qubit incoherent dynamics, and we do so by defining $\hat{\rho}'_{\uparrow\downarrow} = e^{i\hat{\rho}'_{\uparrow\downarrow}} \hat{\rho}_{\uparrow\downarrow}$, where $T_2 = 2T_1 T_2 / \left( 2T_1 + T_2 \right)$ introduced in the main text describes the intrinsic dephasing of the qubit. The evolution equation for $\hat{\rho}'_{\uparrow\downarrow}$ describes the qubit dephasing due to interaction with the resonator, and has the same form as Eq. (B5) but without the last term (proportional to $1/T_2^2$).

To solve the evolution equation for $\hat{\rho}'_{\uparrow\downarrow}$, it is more convenient to move to the Wigner representation, and obtain a partial differential equation for $W_{\uparrow\downarrow}(x, p; t)$, the Wigner function representation of $\hat{\rho}'_{\uparrow\downarrow}$. As Eq. (B5) contains terms at most quadratic in $\hat{a}$ and $\hat{a}^\dagger$ it is possible to solve this PDE with a Gaussian ansatz.

The Gaussian ansatz reduces Eq. (B5) to a set of coupled ODE’s for the means, variances and overall norm of $\hat{\rho}'_{\uparrow\downarrow}$. After solving these in steady state (see Ref. [24] for further details), with the coherent drive defined by
\begin{equation}
\alpha_{\text{in}} = \sqrt{\frac{P_{\text{in}}}{\hbar \Omega_{\text{QPA}}} (\cos(\Phi) + i \sin(\Phi))},
\end{equation}
we can then use Eq. (B4) to define the dephasing rate, which gives the expression found in Eq. (4) (By defining $|\alpha_{\text{in}}|$ in terms of $\Phi$ here, we implicitly fix the phase of the pump, in contrast to the convention used in the main text).

2. Measurement rate with on-chip gain

We now briefly outline the theoretical calculations of the measurement rate, and for further details the interested reader should consult chapter 3 of Ref. [24]. From standard input-output theory, the Heisenberg-Langevin equation for the resonator operator $\hat{a}$ in a frame rotating at the bare resonator frequency $\omega_{\text{QPA}}$ is
\begin{equation}
\dot{\hat{a}} = \left( -i\chi \hat{\sigma}_z - \frac{\kappa}{2} \right) \hat{a} + \Lambda \hat{a}^\dagger - \sqrt{\kappa} \hat{a}_{\text{in}},
\end{equation}
where $\hat{a}_{\text{in}}$ is the input field to the resonator. In our case, for dispersive measurement of the qubit this is ideally a coherent state, such that $\langle \hat{a}_{\text{in}} \rangle = \alpha_{\text{in}}$, with $\alpha_{\text{in}}$ defined in Eq. (B6). For the purposes of the intra-resonator dynamics and calculation of the measurement rate we can treat the qubit operator $\hat{\sigma}_z$ as a classical real variable $\sigma = \pm 1$, corresponding to the ground or excited state of the qubit in the $\hat{\sigma}_z$ basis. Doing so allows us to solve Eq. (B7) exactly, and from this solution extract the measurement rate.

We consider two modes of operation for the QPA, “amplifier mode”, where the input field aligns with the direction of squeezing ($\Phi = 0$), and “squeezer mode”, where the input field aligns with the direction of amplification ($\Phi = \pi/2$). In amplifier mode, we use the QPA to amplify the size of the signal created by the qubit, which comes at the cost of also amplifying the noise in the cavity output field. In squeezer mode, we use the QPA to squeeze the noise in the quadrature containing qubit information, and while the noise can be heavily squeezed, the signal produced by the qubit is only squeezed at most by a factor of two. For a fixed input photon flux, the steady-state intra-resonator photon number is not the same for both modes of operation, but is independent of the qubit state in both cases.

The output field is related to the input field by the standard input-output relation $\hat{a}_{\text{out}} = \hat{a}_{\text{in}} + \sqrt{\kappa} \hat{a}$, and from the output mode we define the measured signal operator by
\begin{equation}
\hat{Q}(t) = \frac{e^{-i\delta} \hat{a}_{\text{out}} + e^{i\delta} \hat{a}_{\text{out}}^\dagger}{\sqrt{2}},
\end{equation}
where the angle $\delta$ parameterizes the quadrature measured. We must choose the measured quadrature such that it is out-of-phase with the input coherent signal (as the qubit information will be contained in the out-of-phase quadrature), such that $|\delta - \Phi| = \pi/2$.

As we are interested in the long-time limit of the QPA dynamics, rather than the SNR we will calculate the measurement rate, defined by
\begin{equation}
\Gamma_{\text{meas}} \equiv \lim_{\tau \to \infty} \frac{\text{SNR}^2(\tau)}{2\tau} = \frac{1}{4} \left( \langle \hat{Q} \rangle_\uparrow - \langle \hat{Q} \rangle_\downarrow \right)^2 \left( S_{\Omega_{\text{QPA}}}^2(0) + S_{\Omega_{\text{QPA}}}^2(0) \right),
\end{equation}
where $\langle \cdot \rangle_\nu$ indicates that the expectation value is taken with respect to the cavity in steady-state and the qubit in state $|\nu\rangle$ for $\nu \in \{\uparrow, \downarrow\}$, corresponding to $\sigma = \pm 1$ respectively. $S_{\Omega_{\text{QPA}}}^2[\rho]$ is the symmetrized noise power of the detected quadrature at frequency $\omega$, defined in the standard way [24].
The measurement rate will depend on what mode the QPA is operated in (i.e. the angle $\Phi$), and for our two operation modes the measurement rates are

$$\Gamma_{\text{meas}}^{\text{amp}} = \frac{\chi^2|\alpha|^2}{\kappa} \left( \frac{\chi^2}{\kappa} + \lambda^2 \right) + \bar{n}_{\text{add}}$$  \hspace{1cm} (B10)

$$\Gamma_{\text{meas}}^{\text{sqz}} = \frac{\chi^2|\alpha|^2}{\kappa} \left( \frac{\chi^2}{\kappa} + \lambda^2 \right) + \bar{n}_{\text{add}}$$  \hspace{1cm} (B11)

where we have added by hand a noise term $\bar{n}_{\text{add}}$ to quantify noise added to the signal downstream of the QPA. For a fair comparison we parametrize the rates in terms of a constant intra-resonator photon number $|\alpha|^2$, which we note requires different input photon flux for the two operation modes.

From Eqs. (B10) and (B11) we see that in both modes of operation the output contains amplified noise, no matter the value of $\lambda$. While this is by design in amplifier mode, in squeezer mode it is naively unexpected, and is a result of interaction with the qubit. The dispersive interaction results in a qubit-dependent phase shift on the field exiting the resonator, such that it no longer perfectly interferes with the promptly reflected field. The effect of this is a mixing of the squeezed and amplified noise, such that all quadratures contain noise contributions from both.

However, for very small $\chi/\kappa$, squeezer mode operation does not suffer from this unwanted mixed-in amplified noise, as can be seen when we write the measurement rates to leading order in $\chi/\kappa$

$$\Gamma_{\text{meas}}^{\text{amp}} \approx \frac{2\chi^2|\alpha|^2(1 + \sqrt{G_0})^2}{\kappa(\kappa + 2\bar{n}_{\text{add}})}$$  \hspace{1cm} (B12)

$$\Gamma_{\text{meas}}^{\text{sqz}} \approx \frac{2\chi^2|\alpha|^2(1 + 1/\sqrt{G_0})^2}{\kappa(1/\kappa + 2\bar{n}_{\text{add}})}$$  \hspace{1cm} (B13)

where we have defined $\sqrt{G_0} = (\kappa/2 + \lambda)/(\kappa/2 - \lambda)$, with $G_0 = 1$ for zero gain. Both measurement rates should be contrasted with the zero gain measurement rate

$$\Gamma_{\text{meas}}^0 = \frac{2\chi^2|\alpha|^2}{\kappa(\kappa + 2\bar{n}_{\text{add}})}$$

found for a standard linear-resonator setup, or when the QPA is operated with zero gain.

In the ideal limit, where $\bar{n}_{\text{add}} = 0$, amplifier mode offers little to no advantage over zero gain, as both the signal and noise are amplified by the same factor at large gain, which can be seen by comparing Eq. (B12) for $\bar{n}_{\text{add}} = 0$ to Eq. (B14). However, in this case squeezer mode can be advantageous, as in the large gain limit the noise is drastically reduced, while the signal is relatively unaffected. In particular, for $\chi/\kappa \ll 1$, comparing Eq. (B13) for $\bar{n}_{\text{add}} = 0$ to Eq. (B14) we see that the measurement rate is enhanced by a large factor proportional to $G_0$. Accounting for effects beyond first order in $\chi/\kappa$ by using the full expression of Eq. (B11), we find that the squeezer mode measurement rate is enhanced by the factor $\Gamma_{\text{meas}}^{\text{sqz}}/\Gamma_{\text{meas}}^0 = \kappa/\chi$ at the optimal value of $\lambda$.

Conversely, in the non-ideal situation where $\bar{n}_{\text{add}}$ is large, squeezer mode offers no advantage, as the noise can never be reduced below the noise floor set by $\bar{n}_{\text{add}}$, as clearly indicated by Eq. (B13). In this situation amplifier mode is beneficial, as by amplifying both the signal and noise leaving the QPA, the output becomes insensitive to noise added downstream (concretely, $G_0 > \bar{n}_{\text{add}}$ in the denominator of Eq. (B12). As shown in the main text, this is the mode of operation we find gives the greatest efficiency for our current setup.

---

* Author to whom correspondence should be addressed: aeddins@berkeley.edu

† Current address: Department of Physics & Astronomy, University of Southern California, Los Angeles CA 90089, USA

‡ Current address: Raytheon BBN Technologies, 10 Moulton St., Cambridge, MA 02138, USA

[1] R. Vijay, C. Macklin, D. H. Slichter, S. J. Weber, K. W. Murch, R. Naik, A. N. Korotkov, and I. Siddiqi, “Stabilizing Rabi oscillations in a superconducting qubit using quantum feedback,” Nature 490, 77–80 (2012).

[2] K. W. Murch, S. J. Weber, C. Macklin, and I. Siddiqi, “Observing single quantum trajectories of a superconducting quantum bit,” Nature 502, 211–214 (2013).

[3] S. J. Weber, A. Chantamri, J. Dressel, A. N. Jordan, K. W. Murch, and I. Siddiqi, “Mapping the optimal route between two quantum states,” Nature 511, 570.

[4] Shay Hacohen-Gourari, Leigh S Martin, Emmanuel Fluin, Vinay V Ramasesh, K Birgitta Whaley, and Irfan Siddiqi, “Quantum dynamics of simultaneously measured non-commuting observables,” Nature 538, 491 (2016).

[5] P. Campagne-Ibarcq, P. Six, L. Bretheau, A. Sarlette, M. Mirrahimi, P. Rouchon, and B. Huard, “Observing Quantum State Diffusion by Heterodyne Detection of Fluorescence,” Physical Review X 6, 011002 (2016).

[6] Q. Ficheux, S. Jezouin, Z. Leghtas, and B. Huard, “Dynamics of a qubit while simultaneously monitoring its relaxation and dephasing,” arXiv:1711.01208 (2017).

[7] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Fechel, M. Mondal, M. Opplinger, C. Eichler, and A. Wallraff, “Rapid High-Fidelity Single-Shot Dispersive Readout of Superconducting Qubits,” Physical Review Applied 7, 054020.
fiers,” Physical Review Applied 8, 054030 (2017).

[37] Andrew Eddins, Sydney Schreppler, David M. Toyli, Leigh S. Martin, Shay Hacohen-Gourgy, Luke C. G. Govia, Hugo Ribeiro, Aashish A. Clerk, and Irfan Siddiqi, “Stroboscopic qubit measurement with squeezed illumination,” (2017).

[38] Hanhan Li, Alireza Shabani, Mohan Sarovar, and K. Birgitta Whaley, “Optimality of qubit purification protocols in the presence of imperfections,” Physical Review A 87, 032334 (2013).

[39] Leigh Martin, Felix Motzoi, Hanhan Li, Mohan Sarovar, and K. Birgitta Whaley, “Deterministic generation of remote entanglement with active quantum feedback,” Physical Review A 92, 062321 (2015).

[40] Leigh Martin, Mahrud Sayrafi, and K Birgitta Whaley, “What is the optimal way to prepare a Bell state using measurement and feedback?” Quantum Science and Technology 2, 044006 (2017).

[41] D. Wahyu Utami and A. A. Clerk, “Entanglement dynamics in a dispersively coupled qubit-oscillator system,” Physical Review A 78, 042323 (2008).