General method for extracting the quantum efficiency of dispersive qubit readout in circuit QED

C. C. Bultink,[1,2] B. Tarasinski,[1,2] N. Haandbæk,[3] S. Poletto,[1,2] N. Haider,[1,4] D. J. Michalak,[5] A. Bruno,[1,2] and L. DiCarlo[1,2]

[1)] *QuTech, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands*
[2)] *Kavli Institute of Nanoscience, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands*
[3)] *Zurich Instruments AG, Technoparkstrasse 1, 8005 Zürich, Switzerland*
[4)] *Netherlands Organisation for Applied Scientific Research (TNO), P.O. Box 155, 2600 AD Delft, The Netherlands*
[5)] *Components Research, Intel Corporation, 2501 NW 229th Ave, Hillsboro, OR 97124, USA*

(Dated: 16 November 2017)

We present and demonstrate a general 3-step method for extracting the quantum efficiency of dispersive qubit readout in circuit QED. We use active depletion of post-measurement photons and optimal integration weight functions on two quadratures to maximize the signal-to-noise ratio of non- steady-state homodyne measurement. We derive analytically and demonstrate experimentally that the method robustly extracts the quantum efficiency for arbitrary readout conditions in the linear regime. We use the proven method to optimally bias a Josephon traveling-wave parametric amplifier and to quantify the different noise contributions in the readout amplification chain.

Many protocols in quantum information processing, like quantum error correction[1,2], require rapid interleaving of qubit gates and measurements. These measurements are ideally nondemolition, fast, and high fidelity. In circuit QED[3–5], a leading platform for quantum computing, nondemolition readout is routinely achieved by off-resonantly coupling a qubit to a resonator. The qubit-state-dependent dispersive shift of the resonator fundamental is inferred by measuring the resonator response to an interrogating pulse using homodyne detection. A key element setting the speed and fidelity of dispersive readout is the quantum efficiency[6], which quantifies how the signal-to-noise ratio is degraded with respect to the limit imposed by quantum vacuum fluctuations.

In recent years, the use of superconducting parametric amplifiers[7–11] as the front end of the readout amplification chain has boosted the quantum efficiency towards unity, leading to readout infidelity on the order of one percent[12,13] in individual qubits. Most recently, the development of traveling-wave parametric amplifiers[14,15] (TWPAs) has extended the amplification bandwidth from tens of MHz to several GHz and with sufficient dynamic range to readout tens of qubits. For characterization and optimization of the amplification chain, the ability to robustly determine the quantum efficiency is an important benchmark.

A common method for quantifying the quantum efficiency η that does not rely on calibrated noise sources compares the information obtained in a weak qubit measurement (expressed by the signal-to-noise-ratio SNR) to the dephasing of the qubit (expressed by the decay of the off-diagonal elements of the qubit density matrix)[16], $\eta = \frac{\mathrm{SNR}^2}{4\gamma_\mathrm{m}}$, with $\mathrm{e}^{-\gamma_\mathrm{m}} = \frac{|\rho_{01}(T)|}{|\rho_{01}(0)|}$. Previous work[17–19] has been restricted to fast resonators driven under specific symmetry conditions such that information is contained in only one quadrature of the output field and in steady state. To allow in-situ calibration of η in multi-qubit devices under development[20–24], a method is desirable that does not rely on either of these conditions.

In this Letter, we present and demonstrate a general 3-step method for extracting the quantum efficiency of linear dispersive readout in cQED. Our method disposes with previous requirements in both the dynamics and the phase space trajectory of the resonator field, while requiring two easily met conditions: the depletion of resonator photons post measurement[25,26], and the ability to perform weighted integration of both quadratures of the output field[27,28]. We experimentally test the method on a qubit-resonator pair with a Josephson TWPA (JTWPA)[14] at the front end of the amplification chain. To prove the generality of the method, we extract a consistent value of η for different readout drive frequencies and drive envelopes. Finally, we use the method to optimally bias the JTWPA and to quantify the different noise contributions in the readout amplification chain.

We first derive the method, obtaining experimental boundary conditions. For a measurement in the linear dispersive regime of cQED, the internal field $\alpha(t)$ of the readout resonator, driven by a pulse with envelope $\varepsilon f(t)$ and detuned by Δ from the resonator center frequency, is described by[29,30]

$$\frac{\partial \alpha_{|0\rangle/|1\rangle}}{\partial t} = -i\varepsilon f(t) - i(\Delta \pm \chi)\alpha(t) - \frac{\kappa}{2}\alpha(t), \quad (1)$$

where κ is the resonator linewidth and 2χ is the dispersive shift. The upper (lower) sign has to be chosen for the qubit in the ground $|0\rangle$ [excited $|1\rangle$] state. We study the evolution of the SNR and the measurement-induced dephasing as a function of the drive amplitude ε. We find that the SNR scales linearly, $\mathrm{SNR} = a\varepsilon$, and that coherence elements exhibit a Gaussian dependence,

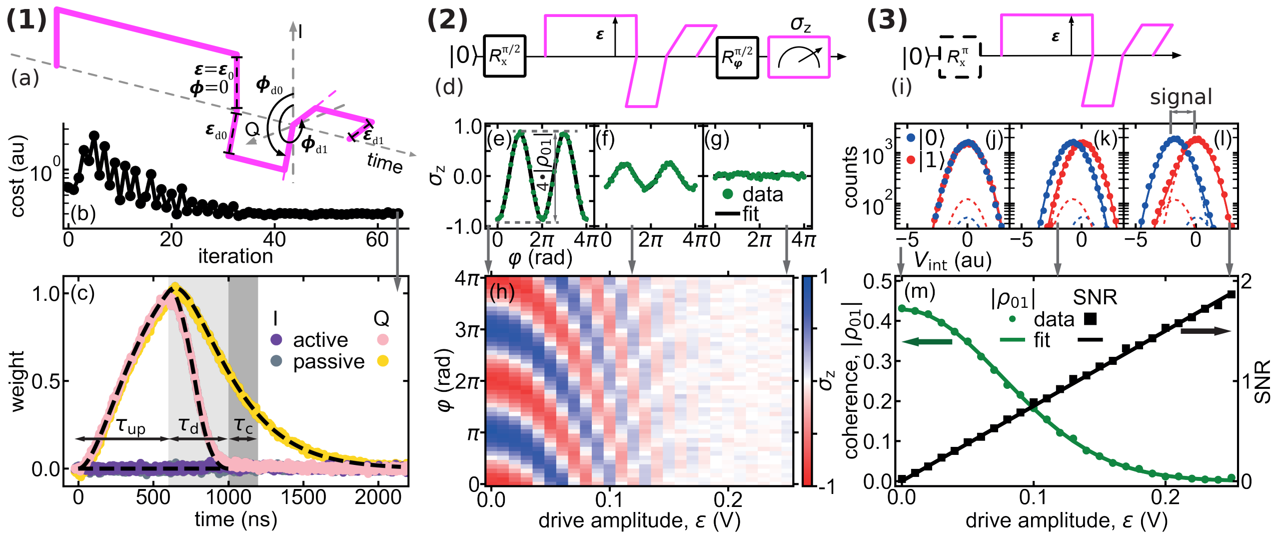

FIG. 1. The 3-step method for extracting the quantum efficiency. **(1)** Active depletion tuneup and optimal integration weights calibration. (a) The measurement pulse consists of a ramp-up of duration $\tau_{\mathrm{up}} = 600$ ns, fixed phase $\phi = 0$ and amplitude ε (fixed during tuneup to $\varepsilon = \varepsilon_0 = 0.25$ V) and two 200 ns depletion segments ($\tau_{\mathrm{d}} = 400$ ns) with each a tunable phase (ϕ_{d0}, ϕ_{d1}) and amplitude ($\varepsilon_{\mathrm{d0}}$, $\varepsilon_{\mathrm{d1}}$). (b) Cost function evolution during numerical optimization of the four depletion parameters. The cost function uses the rms of the optimal weight functions during $\tau_{\mathrm{c}} = 200$ ns. (c) Optimal weight functions for the in-phase quadrature I and out-of-phase quadrature Q for the optimized depletion pulse ($\varepsilon_{\mathrm{d0}} = 1.64\varepsilon$, $\varepsilon_{\mathrm{d1}} = 0.39\varepsilon$, $\phi_{\mathrm{d0}} = 0.993\pi$ rad, $\phi_{\mathrm{d1}} = -0.016\pi$ rad). As a reference, weight functions are shown for passive depletion ($\varepsilon_{\mathrm{d0}} = \varepsilon_{\mathrm{d1}} = 0$ V). The weight functions show the dynamics of the information gain during readout and the effect of the active photon depletion (light grey area). **(2)** Study of dephasing under variable-strength weak measurement. (d) Pulse sequence. The measurement pulse, globally scaled with ε, is embedded in a fixed-length ($\tau_{\mathrm{int}} = 1100$ ns) Ramsey sequence with final strong fixed-amplitude measurement. The azimuthal angle φ of the final $\pi/2$ rotation is swept from 0 to 4π to discern deterministic phase shifts and dephasing. (e-g) Observed Ramsey fringes at $\varepsilon = 0.0, 0.12, 0.25$ V, respectively. The coherence $|\rho_{01}|$ is extracted by fitting each fringe with the form $\sigma_{\mathrm{z}} = 2\,|\rho_{01}|\cdot\cos(\varphi + \varphi_0)$. (h) Ramsey fringes as a function of ε. **(3)** Study of signal-to-noise ratio of variable-strength weak measurement. (i) Pulse sequence to measure single-shot readout histograms. (j-l) Histograms of 2^{15} shots at $\varepsilon = 0.0, 0.12, 0.25$ V, respectively. The qubit is prepared in $|0\rangle$ without (blue) and in $|1\rangle$ with a π pulse (red). Each measurement record is integrated in real time with the weight functions of (c) during $\tau_{\mathrm{int}} = 1100$ ns to obtain V_{int}. Each histogram (markers) is fitted with the sum of two Gaussian functions (solid lines) to consent relaxation during the measurement and imperfect qubit preparation (dashed lines are the single Gaussian functions). From the fits we get the signal, distance between the main Gaussian for $|0\rangle$ and $|1\rangle$, and noise, their average standard deviations. (m) Quantum efficiency extraction. Coherence data is fitted with the form $|\rho_{01}| = b \cdot \mathrm{e}^{-\varepsilon^2/2\cdot\sigma^2}$ and signal-to-noise data with the form $\mathrm{SNR} = a\cdot\varepsilon$. From the best fits we extract $\eta_{\mathrm{e}} = a^2\cdot\sigma^2/2 = 0.165 \pm 0.002$.

$|\rho_{01}(T,\varepsilon)| = |\rho_{01}(T,0)|\,\mathrm{e}^{-\frac{\varepsilon^2}{2\sigma_{\mathrm{m}}^2}}$, with a and σ_{m} dependent on κ, χ, Δ, and $f(t)$. Furthermore, we find (Supplementary material[31])

$$\eta = \frac{\mathrm{SNR}^2}{4\gamma_{\mathrm{m}}} = \frac{\sigma_{\mathrm{m}}^2 a^2}{2} \tag{2}$$

provided two conditions are met. The conditions are: i) optimal integration functions[27,28] are used to optimally extract information from both quadratures, and ii) the intra-resonator field vanishes at the beginning and end; i.e., photons are depleted from the resonator post-measurement.

To meet these conditions, we introduce a three-step experimental method. First, numerically tuneup active photon depletion and calibration of the optimal integration weights. Second, obtain the measurement-induced dephasing of variable-strength weak measurement by including the pulse within a Ramsey sequence. Third, measure the SNR of variable-strength weak measurement from single-shot readout histograms.

We test the method on a cQED test chip containing seven transmon qubits with dedicated readout resonators, each coupled to one of two feedlines[31]. We present data for one qubit-resonator pair, but have verified the method with other pairs in this and other devices. The qubit is operated at its sweetspot frequency $f_{\mathrm{q}} = 5.070$ GHz, where the measured energy relaxation and echo dephasing times are $T_1 = 15$ μs and $T_{2,\mathrm{echo}} = 26$ μs, respectively. The resonator has a low-power fundamental at $f_{\mathrm{r},|0\rangle} = 7.852400$ GHz ($f_{\mathrm{r},|1\rangle} = f_{\mathrm{r},|0\rangle} + \chi/\pi = 7.852295$ GHz) for qubit in $|0\rangle$ ($|1\rangle$), with linewidth $\kappa/2\pi = 1.4$ MHz. The readout pulse generation and readout signal integration are performed by single-sideband mixing. Pulse-envelope generation, demodulation and signal processing are performed by a Zurich Instruments UHFLI-QC with 2 AWG channels and 2 ADC channels running at 1.8 GSample/s with 14− and 12−bit resolution, respectively.

In the first step, we tune up the depletion steps and calibrate the optimal integration weights. We use a measurement ramp-up pulse of duration $\tau_{\mathrm{up}} = 600$ ns, followed by a photon-depletion counter pulse[25,26] consisting of two steps of 200 ns duration each, for a total depletion time $\tau_{\mathrm{d}} = 400$ ns [Figs. 1(a,b)]. To successfully deplete without relying on symmetries that are specific to a readout frequency at the midpoint between ground and excited state resonances (i.e., $\Delta = 0$), we vary 4 parameters of the depletion steps (as in Ref. 26). Here, we optimize the amplitude and phase of both depletion steps using the Nelder-Mead algorithm with a cost function that penalizes non-zero averaged transients for both $|0\rangle$ and $|1\rangle$ during a $\tau_{\mathrm{c}} = 200$ ns time window after the depletion (cost function provided in Ref. 31). From the averaged transients of the finally obtained measurement pulse, we extract the optimal integration weights given by[27,28] the difference between the averaged transients for $|0\rangle$ and $|1\rangle$ [Fig. 1(c)]. The success of the active depletion is evidenced by the nulling at the end of τ_{d}. In this initial example, we connect to previous work by setting $\Delta = 0$, leaving all measurement information in one quadrature.

We next use the tuned readout to study its measurement-induced dephasing and SNR to finally extract η. We measure the dephasing by including the measurement-and-depletion pulse in a Ramsey sequence and varying its amplitude, ε [Figs. 1(d-h)]. By varying the azimuthal angle of the second qubit pulse, we allow distinguishing dephasing from deterministic phase shifts and extract $|\rho_{01}|$ from the amplitude of the fitted Ramsey fringes. The SNR at various ε is extracted from single-shot readout experiments preparing the qubit in $|0\rangle$ and $|1\rangle$ [Figs. 1(i-l)]. We use double Gaussian fits in both cases, neglecting measurement results in the spurious Gaussians to reduce corruption by imperfect state preparation and residual qubit excitation and relaxation. As expected, as a function of ε, $|\rho_{01}|$ decreases following a Gaussian form and the SNR increases linearly [Fig. 1(m)]. The best fits to both dependencies give $\eta_{\mathrm{e}} = 0.165 \pm 0.002$. Note that both dephasing and SNR measurements include ramp-up, depletion and an additional $\tau_{\mathrm{buffer}} = 100$ ns, making the total integration window $\tau_{\mathrm{int}} = \tau_{\mathrm{up}} + \tau_{\mathrm{d}} + \tau_{\mathrm{buffer}} = 1100$ ns.

We next demonstrate the generality of the method by extracting η as a function of the readout drive frequency. We repeat the method at fifteen readout drive detunings over a range of 2.8 MHz $\sim \kappa/\pi \sim 14\chi/\pi$ around $\Delta = 0$ [Figs. 2(a,b)]. Furthermore, we compare the effect of using optimal weight functions versus square weight functions, and the effect of using active versus passive photon depletion. The square weight functions correspond to a single point in phase space during τ_{int}, with unit amplitude and an optimized phase maximizing SNR. We satisfy the 0-photon field condition by depleting the photons actively with $\tau_{\mathrm{int}} = 1100$ ns (as in Fig. 1) or passively by waiting with $\tau_{\mathrm{int}} = 2100$ ns. When information is extracted from both quadratures using optimal weight functions, we measure an average $\eta_{\mathrm{e}} = 0.167$ with 0.04 standard deviation. The extracted optimal integration functions in the time domain [Figs. 2(c,d)] show how the resonator returns to the vacuum for both active and passive depletion. Square weight functions are not able to track the measurement dynamics in the time domain (even at $\Delta = 0$), leading to a reduction in η_{e}. Figures 2(e,f) show the weight functions in phase space. The opening of the trajectories with detuning illustrates the rotating optimal measurement axis during measurement and leads to a further reduction of increase of η_{e} when square weight functions are used. The dynamics and the η_{e} dependence on Δ are excellently captured by our linear model, which uses $\eta = 0.1670$ and the sepa-

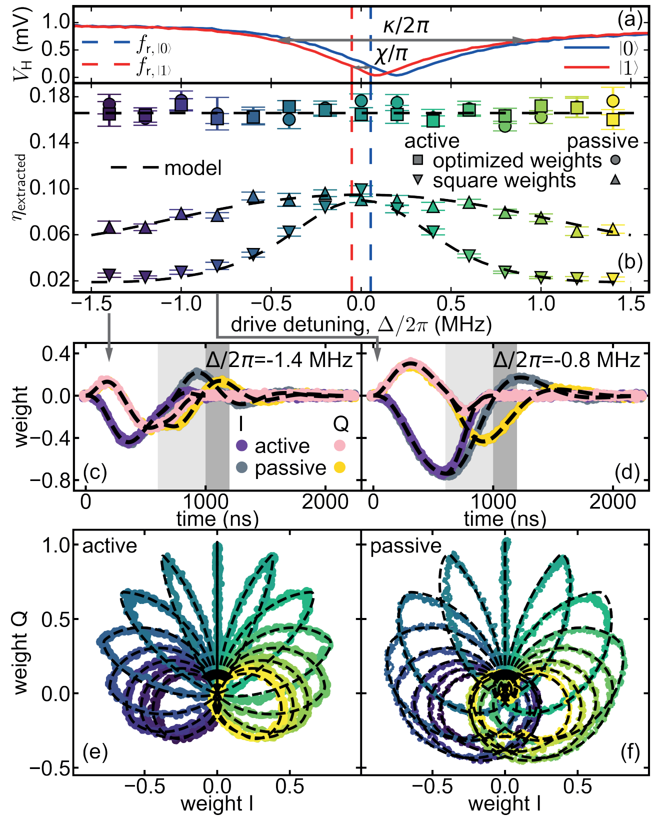

FIG. 2. (a) Pulsed feedline transmission near the low- power resonator fundamentals. The qubit is prepared in $|0\rangle$ without (blue) and in $|1\rangle$ with a π pulse (red). The data fits $\kappa/2\pi =$ 1.4 MHz and $f_{\mathrm{r},|0\rangle} = 7.852400$ GHz ($f_{\mathrm{r},|1\rangle} = 7.852295$ GHz), indicated by the dashed vertical lines. (b) Quantum efficiency extraction as a function of Δ using the pulse timings and 3-step method of Fig. 1. We use both the active depletion ($\tau_{\mathrm{int}} = 1100$ ns, $\varepsilon_{\mathrm{d0}} = \varepsilon_{\mathrm{d0,opt}}$, $\varepsilon_{\mathrm{d1}} = \varepsilon_{\mathrm{d1,opt}}$) and passive depletion schemes ($\tau_{\mathrm{int}} = 2100$ ns, $\varepsilon_{\mathrm{d0}} = \varepsilon_{\mathrm{d1}} = 0$) and assess the benefit of optimal weights to standard square integration weights. (c,d) Optimal weight functions for I and Q at $\Delta/2\pi = -1.4$ MHz, -0.8 MHz [as in Fig. 1(b)]. (e, f) Parametric plot of the optimal weight functions at all measured Δ [marker colors correspond to (a)]. Dashed black curves (b-f) are extracted from a linear model (see main text).

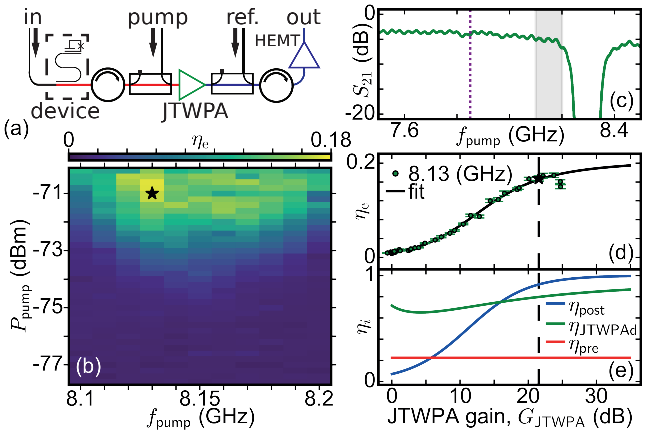

FIG. 3. JTWPA pump tuneup to maximize the quantum efficiency and amplification chain modeling. (a) Simplified setup diagram, showing the input paths for the readout signal carrying the information on the qubit state and the added pump tone biasing the JTWPA. Both microwave tones are combined in the JTWPA amplifying the small readout signal. (b) η_e as a function of pump power and frequency. (c) CW low-power transmission of the JTWPA showing the dip in transmission due to the dispersion feature near 8.3 GHz and low-power insertion loss of ~ 4.0 dB near $f_{r,|0\rangle}$ (dashed vertical line). The grey area indicates the frequency range of (b). S_{21} is obtained by measuring and comparing the output power when selecting the pump input or the reference input (input lines are duplicates and calibrated up to the directional couplers at room temperature). (d) Parametric plot of η_e at $f_{pump} = 8.13$ GHz and independently measured JTWPA gain. The fit (line) uses a 3-stage model with $\eta(G_{JTWPA}) = \eta_{pre} \cdot \eta_{JTWPAd}(G_{JTWPA}) \cdot \eta_{post}(G_{JTWPA})$ [model details in the main text]. (e) Plots of the best-fit η_{pre}, $\eta_{JTWPAd}(G_{JTWPA})$ and $\eta_{post}(G_{JTWPA})$. The stars (b, d) and vertical dashed lines (d, e) indicate ($P_{pump} = -71.0$ dBm, $f_{pump} = 8.13$ GHz, $\eta = 0.1670$, $G_{JTWPA} = 21.6$ dB) used throughout the experiment.

rately calibrated κ and χ [Fig. 2(a)]. Furthermore, the matching of the dynamics and depletion pulse parameters[31] when using active photon depletion confirm the numerical optimization techniques.

To further test the robustness of the method to arbitrary pulse envelopes, we have used a measurement-and-depletion pulse envelope $f(t)$ resembling a typical Dutch skyline. The pulse envelope outlines five canal houses, of which the first three ramp up the resonator and the latter two are used as the tunable depletion steps. Completing the three steps, we extract[31] $\eta_e = 0.165 \pm 0.005$, matching our previous value to within error.

We use the proven method to optimally bias the JTWPA and to quantify the different noise contributions in the readout chain. To this end, we map η_e as a function of pump power and frequency, just below the dispersive feature of the JTWPA, finding the maximum $\eta_e = 0.1670$ at ($P_{pump} = -71.0$ dBm, $f_{pump} = 8.13$ GHz) [Figs. 3(a-c)]. We next compare the obtained η_e at the optimal bias frequency to independent low-power measurements of the JTWPA gain G_{JTWPA} we find $G_{JTWPA} = 21.6$ dB at the optimal bias point. We fit this parametric plot with a 3-stage model, with noise contributions before, in and after the JTWPA, $\eta(G_{JTWPA}) = \eta_{pre} \cdot \eta_{JTWPAd}(G_{JTWPA}) \cdot \eta_{post}(G_{JTWPA})$. The parameter η_{pre} captures losses in the device and the microwave network between the device and the JTWPA and is therefore independent of G_{JTWPA}. The JTWPA has a distributed loss along the amplifying transmission line, which is modeled as an array of interleaved sections with quantum-limited amplification and sections with attenuation adding up to the total insertion loss of the JTWPA (as in Ref. 14). Finally, the post-JTWPA amplification chain is modeled with a fixed noise temperature, whose relative noise contribution diminishes as G_{JTWPA} is increased. The best fit [Figs. 3(d,e)] gives $\eta_{pre} = 0.22$, consistent with 50% photon loss due to symmetric coupling of the resonator to the feedline input and output, an attenuation of the microwave network between device and JTWPA of 2 dB and residual loss in the JTWPA of 27%. We fit a distributed insertion loss of the JTWPA of 4.6 dB, closely matching the separate calibration of 4.2 dB [Fig. 3(c)]. Finally, we fit a noise temperature of 2.6 K, close to the HEMT amplifier's factory specification of 2.2 K.

We identify room for improving η_e to ~ 0.5 by implementing Purcell filters with asymmetric coupling[19,32] (primarily to the output line) and decreasing the insertion loss in the microwave network, by optimizing the setup for shorter and superconducting cabling between device and JTWPA.

In conclusion, we have presented and demonstrated a general 3-step method for extracting the quantum efficiency of linear dispersive qubit readout in cQED. We have derived analytically and demonstrated experimentally that the method robustly extracts the quantum efficiency for arbitrary readout conditions in the linear regime. This method will be used as a tool for readout performance characterization and optimization.

See supplementary material for the derivation of Eq. (2) and additional figures.

ACKNOWLEDGMENTS

We thank W. D. Oliver for providing the JTWPA, N. K. Langford for experimental contributions, M. A. Rol for software contributions, and C. Dickel and F. Luthi for discussions. This research is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the U.S. Army Research Office Grant No. W911NF-16-1-0071. Additional funding is provided by Intel Corporation and the ERC Synergy Grant QC-lab. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

[1] D. P. DiVincenzo, Physica Scripta **2009**, 014020 (2009).
[2] B. M. Terhal, Rev. Mod. Phys. **87**, 307 (2015).
[3] A. Blais, R.-S. Huang, A. Wallraff, S. M. Girvin, and R. J. Schoelkopf, Phys. Rev. A **69**, 062320 (2004).
[4] A. Wallraff, D. I. Schuster, A. Blais, L. Frunzio, R.-S. Huang, J. Majer, S. Kumar, S. M. Girvin, and R. J. Schoelkopf, Nature **431**, 162 (2004).
[5] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, J. Majer, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, Phys. Rev. A **76**, 042319 (2007).
[6] A. A. Clerk, M. H. Devoret, S. M. Girvin, F. Marquardt, and R. J. Schoelkopf, Reviews of Modern Physics **82**, 1155 (2010), 0810.4729.
[7] M. A. Castellanos-Beltran, K. D. Irwin, G. C. Hilton, L. R. Vale, and K. W. Lehnert, Nat. Phys. **4**, 929 (2008).
[8] R. Vijay, M. H. Devoret, and I. Siddiqi, Rev. Sci. Instrum. **80**, 111101 (2009).
[9] N. Bergeal, F. Schackert, M. Metcalfe, R. Vijay, V. E. Manucharyan, L. Frunzio, D. E. Prober, R. J. Schoelkopf, S. M. Girvin, and M. H. Devoret, Nature **465**, 64 (2010).
[10] J. Y. Mutus, T. C. White, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, E. Jeffrey, J. Kelly, A. Megrant, C. Neill, P. J. J. O'Malley, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, K. M. Sundqvist, A. N. Cleland, and J. M. Martinis, Appl. Phys. Lett. **104**, 263513 (2014).
[11] C. Eichler, Y. Salathe, J. Mlynek, S. Schmidt, and A. Wallraff, Phys. Rev. Lett. **113**, 110502 (2014).
[12] J. E. Johnson, C. Macklin, D. H. Slichter, R. Vijay, E. B. Weingarten, J. Clarke, and I. Siddiqi, Phys. Rev. Lett. **109**, 050506 (2012).
[13] D. Ristè, J. G. van Leeuwen, H.-S. Ku, K. W. Lehnert, and L. DiCarlo, Phys. Rev. Lett. **109**, 050507 (2012).
[14] C. Macklin, K. O'Brien, D. Hover, M. E. Schwartz, V. Bolkhovsky, X. Zhang, W. D. Oliver, and I. Siddiqi, Science **350**, 307 (2015).
[15] M. R. Vissers, R. P. Erickson, H. S. Ku, L. Vale, X. Wu, G. C. Hilton, and D. P. Pappas, Appl. Phys. Lett. **108**, 012601 (2016).
[16] This definition of quantum effiency applies to phase-preserving amplification where the unavoidable quantum noise from the idler mode of the amplifier is included in the quantum limit. Using this definition, $\eta = 1$ corresponds to an ideal phase preserving amplification.
[17] M. Hatridge, S. Shankar, M. Mirrahimi, F. Schackert, K. Geerlings, T. Brecht, K. Sliwa, B. Abdo, L. Frunzio, S. Girvin, R. Schoelkopf, and M. Devoret, Science **339**, 178 (2013).
[18] S. Abay, D. Persson, H. Nilsson, F. Wu, H. Q. Xu, M. Fogelström, V. Shumeiko, and P. Delsing, Phys. Rev. B **89**, 214508 (2014).
[19] E. Jeffrey, D. Sank, J. Y. Mutus, T. C. White, J. Kelly, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. Megrant, P. J. J. O'Malley, C. Neill, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, Phys. Rev. Lett. **112**, 190504 (2014).
[20] Q. Liu, M. Li, K. Dai, K. Zhang, G. Xue, X. Tan, H. Yu, and Y. Yu, Applied Physics Letters **110**, 232602 (2017).
[21] M. Reagor, C. B. Osborn, N. Tezak, A. Staley, G. Prawiroatmodjo, M. Scheer, N. Alidoust, E. A. Sete, N. Didier, M. P. da Silva, E. Acala, J. Angeles, A. Bestwick, M. Block, B. Bloom, A. Bradley, C. Bui, S. Caldwell, L. Capelluto, R. Chilcott, J. Cordova, G. Crossman, M. Curtis, S. Deshpande, T. E. Bouayadi, D. Girshovich, S. Hong, A. Hudson, P. Karalekas, K. Kuang, M. Lenihan, R. Manenti, T. Manning, J. Marshall, Y. Mohan, W. O'Brien, J. Otterbach, A. Papageorge, J. P. Paquette, M. Pelstring, A. Polloreno, V. Rawat, C. A. Ryan, R. Renzas, N. Rubin, D. Russell, M. Rust, D. Scarabelli, M. Selvanayagam, R. Sinclair, R. Smith, M. Suska, T. W. To, M. Vahidpour, N. Vodrahalli, T. Whyland, K. Yadav, W. Zeng, and C. T. Rigetti, arXiv:1706.06570 (2017).
[22] M. Takita, A. W. Cross, A. D. Córcoles, J. M. Chow, and J. M. Gambetta, Phys. Rev. Lett. **119**, 180501 (2017).
[23] C. Neill, P. Roushan, K. Kechedzhi, S. Boixo, S. V. Isakov, V. Smelyanskiy, R. Barends, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. Fowler, B. Foxen, R. Graff, E. Jeffrey, J. Kelly, E. Lucero, A. Megrant, J. Mutus, M. Neeley, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, H. Neven, and J. M. Martinis, arXiv:1709.06678 (2017).
[24] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, Phys. Rev. Applied **8**, 034021 (2017).
[25] D. T. McClure, H. Paik, L. S. Bishop, M. Steffen, J. M. Chow, and J. M. Gambetta, Phys. Rev. Appl. **5**, 011001 (2016).
[26] C. C. Bultink, M. A. Rol, T. E. O'Brien, X. Fu, B. C. S. Dikken, C. Dickel, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, R. N. Schouten, and L. DiCarlo, Phys. Rev. Appl. **6**, 034008 (2016).
[27] C. A. Ryan, B. R. Johnson, J. M. Gambetta, J. M. Chow, M. P. da Silva, O. E. Dial, and T. A. Ohki, Phys. Rev. A **91**, 022118 (2015).
[28] E. Magesan, J. M. Gambetta, A. D. Córcoles, and J. M. Chow, Phys. Rev. Lett. **114**, 200501 (2015).
[29] J. Gambetta, A. Blais, D. I. Schuster, A. Wallraff, L. Frunzio, J. Majer, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, Phys. Rev. A **74**, 042318 (2006).
[30] A. Frisk Kockum, L. Tornberg, and G. Johansson, Phys. Rev. A **85**, 052318 (2012).
[31] See supplementary material.
[32] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Pechal, M. Mondal, M. Oppliger, C. Eichler, and A. Wallraff, Phys. Rev. Appl. **7**, 054020 (2017).

SUPPLEMENTARY MATERIAL FOR "GENERAL METHOD FOR EXTRACTING THE QUANTUM EFFICIENCY OF DISPERSIVE QUBIT READOUT IN CIRCUIT QED"

This supplement provides additional sections and figures in support of claims in the main text. Section I derives Eq. 2 and Sec. II provides the cost function used for the optimization of depletion pulses. Figure S1 supplies the optimized depletion pulse parameters as a function of Δ and the SNR and coherence as a function of the drive amplitude and Δ. Figure S2 shows the extraction of η_e for an alternative pulse shape. Finally, Fig. S3 provides a full wiring diagram and a photograph of the device.

I. DERIVATION OF EQUATION 2

In general, the measured homodyne signal consists of an in-phase (I) and in-quadrature (Q) components, given by[30]

$$V_{\mathrm{I},|i\rangle}(t) = V_0\left(\sqrt{2\kappa\eta}\,\mathrm{Re}(\alpha_{|i\rangle}(t)) + n_\mathrm{I}(t)\right),$$
$$V_{\mathrm{Q},|i\rangle}(t) = V_0\left(\sqrt{2\kappa\eta}\,\mathrm{Im}(\alpha_{|i\rangle}(t)) + n_\mathrm{Q}(t)\right).$$

Here, V_0 is an irrelevant scale factor and n_I, n_Q are continuous, independent Gaussian white noise terms with unit variance, $\langle n_j(t)n_k(t')\rangle = \delta_{j,k}\delta(t-t')$, while the internal resonator field $\alpha_{|i\rangle}$ follows Eq. (1) for $i \in \{0,1\}$. In the shunt resonator arrangement used on the device for this work, we should in fact include an additional term describing the directly transmitted measurement pulse. We omitted this term here, as it is independent of qubit state, and thus irrelevant for the following, as we will exclusively encounter the signal difference between qubit states.

For state discrimination, the homodyne signals are each multiplied with the optimal weight functions[2,3], given by the difference of the averaged signals, then summed and integrated over the measurement window:

$$V_{\mathrm{int},|i\rangle} = \int V_{\mathrm{I},|i\rangle}\langle V_{\mathrm{I},|1\rangle} - V_{\mathrm{I},|0\rangle}\rangle + V_{\mathrm{Q},|i\rangle}\langle V_{\mathrm{Q},|1\rangle} - V_{\mathrm{Q},|0\rangle}\rangle dt.$$

The signal S we define as the absolute separation between the average V_int for $|1\rangle$ and $|0\rangle$:

$$S = \langle V_{\mathrm{int},|1\rangle} - V_{\mathrm{int},|0\rangle}\rangle$$
$$= 2\kappa\eta V_0^2 \int \left|\alpha_{|1\rangle} - \alpha_{|0\rangle}\right|^2 dt.$$

The noise N, defined as the standard deviation of $V_{\mathrm{int},|\mathrm{i}\rangle}$, is independent of $|i\rangle$ and given by

$$N^2 = \langle \delta^2 V_\mathrm{int}\rangle$$
$$= V_0^2\left\langle \int \left(\langle V_{\mathrm{I},|1\rangle} - V_{\mathrm{I},|0\rangle}\rangle n_\mathrm{I} + \langle V_{\mathrm{Q},|1\rangle} - V_{\mathrm{Q},|0\rangle}\rangle n_\mathrm{Q}\right)^2 dt\right\rangle$$
$$= 2\kappa\eta V_0^2 \int \left|\alpha_{|1\rangle} - \alpha_{|0\rangle}\right|^2 dt,$$

where we used the white noise property of $n_{\mathrm{I,Q}}(t)$.

The SNR is then given by

$$\mathrm{SNR} = \frac{S}{N} = \sqrt{2\kappa\eta \int \left|\alpha_{|1\rangle} - \alpha_{|0\rangle}\right|^2 dt}.$$

Note that the $\alpha_{|i\rangle}$ scale linearly with the amplitude ε due to the linearity of Eq. (1), so that the SNR scales linearly with ε as well.

The measurement pulse leads to measurement-induced dephasing. Experimentally, the dephasing can be quantified by including the measurement pulse in a Ramsey sequence. The coherence elements of the qubit density matrix are reduced due to the pulse as[1]

$$|\rho_{01}(\varepsilon)| = \mathrm{e}^{-\gamma_\mathrm{m}}\,|\rho_{01}(\varepsilon = 0)|,$$

where

$$\gamma_\mathrm{m} = 2\chi \int \mathrm{Im}(\alpha_{|0\rangle}\alpha^*_{|1\rangle})dt.$$

Thus, γ_m scales with ε^2, and the coherence elements decay as a Gaussian in ε.

We now show that the γ_m and SNR are related by the simple relation Eq. (2), independent of resonator and pulse parameters. For that, we need to make use of constraint (ii), namely that the resonator fields $\alpha_{|i\rangle}$ vanish at the beginning and end of the integration window. We then start with the identity

$$0 = \int \partial_t \left|\alpha_{|0\rangle} - \alpha_{|1\rangle}\right|^2 dt$$
$$= 2\int \mathrm{Re}\left((\alpha^*_{|1\rangle} - \alpha^*_{|0\rangle})\partial_t(\alpha_{|1\rangle} - \alpha_{|0\rangle})\right) dt,$$

where the first equality holds since we integrate a differential, and the boundary vanishes by requirement (ii).

We insert Eq. (1) into this expression, obtaining

$$\mathrm{Re}\int \left(\alpha^*_{|1\rangle} - \alpha^*_{|0\rangle}\right) \times$$
$$\left(\left(-i\Delta - \frac{\kappa}{2}\right)\left(\alpha_{|1\rangle} - \alpha_{|0\rangle}\right) - i\chi\left(\alpha_{|1\rangle} + \alpha_{|0\rangle}\right)\right)dt = 0.$$

Rearranging and dropping purely imaginary terms, after a few lines of manipulation, we obtain

$$\frac{\kappa}{2}\int \left|\alpha_{|1\rangle} - \alpha_{|0\rangle}\right|^2 dt$$
$$= -\,\mathrm{Re}\left(i\chi \int \left(\alpha_{|1\rangle} + \alpha_{|0\rangle}\right)\left(\alpha^*_{|1\rangle} - \alpha^*_{|0\rangle}\right)dt\right)$$
$$= -\,\mathrm{Re}\left(i\chi \int \left(|\alpha_{|1\rangle}|^2 - |\alpha_{|0\rangle}|^2 + 2i\mathrm{Im}(\alpha_{|0\rangle}\alpha^*_{|1\rangle})\right)dt\right)$$
$$= 2\chi \int \mathrm{Im}(\alpha_{|0\rangle}\alpha^*_{|1\rangle})dt.$$

The first and last term are proportional to SNR^2 and γ_m, respectively. We have thus shown that the SNR, when defined with optimal integration weights, and the measurement-induced dephasing γ_m are related by Eq. (2), which is independent of the resonator parameters κ, χ, and the functional form $\varepsilon f(t)$ of the drive.

II. OPTIMIZATION COST FUNCTION FOR DEPLETION TUNEUP

Here we give the cost function that is used for numerical tuneup of the depletion pulse parameters. The transients are obtained by preparing the qubit in $|0\rangle$ ($|1\rangle$) and averaging the time-domain homodyne voltages $V_{\mathrm{I},|0\rangle}$ and $V_{\mathrm{Q},|0\rangle}$ ($V_{\mathrm{I},|1\rangle}$ and $V_{\mathrm{Q},|1\rangle}$) of the transmitted measurement pulse for 2^{15} repetitions. The cost function consists of four different terms. The first two null the transients in both quadratures post-depletion. The last two additionally penalize the difference between the transients for $|0\rangle$ and $|1\rangle$ with a tunable factor d. In the experiment, we found reliable convergence of the depletion tuneup for $d = 10$.

$$
\begin{aligned}
cost = &\sqrt{\int_{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}}^{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}+\tau_{\mathrm{c}}} \langle V_{\mathrm{I},|0\rangle}(t)\rangle^2 + \langle V_{\mathrm{Q},|0\rangle}(t)\rangle^2 dt} \\
&+ \sqrt{\int_{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}}^{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}+\tau_{\mathrm{c}}} \langle V_{\mathrm{I},|1\rangle}(t)\rangle^2 + \langle V_{\mathrm{Q},|1\rangle}(t)\rangle^2 dt} \\
&+ d\sqrt{\int_{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}}^{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}+\tau_{\mathrm{c}}} \langle V_{\mathrm{I},|1\rangle}(t) - V_{\mathrm{I},|0\rangle}(t)\rangle^2 dt} \\
&+ d\sqrt{\int_{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}}^{\tau_{\mathrm{up}}+\tau_{\mathrm{d}}+\tau_{\mathrm{c}}} \langle V_{\mathrm{Q},|1\rangle}(t) - V_{\mathrm{Q},|0\rangle}(t)\rangle^2 dt.
\end{aligned}
$$

[1] A. Frisk Kockum, L. Tornberg, and G. Johansson, Phys. Rev. A **85**, 052318 (2012).

[2] C. A. Ryan, B. R. Johnson, J. M. Gambetta, J. M. Chow, M. P. da Silva, O. E. Dial, and T. A. Ohki, Phys. Rev. A **91**, 022118 (2015).

[3] E. Magesan, J. M. Gambetta, A. D. Córcoles, and J. M. Chow, Phys. Rev. Lett. **114**, 200501 (2015).

[4] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, Phys. Rev. Applied **8**, 034021 (2017).

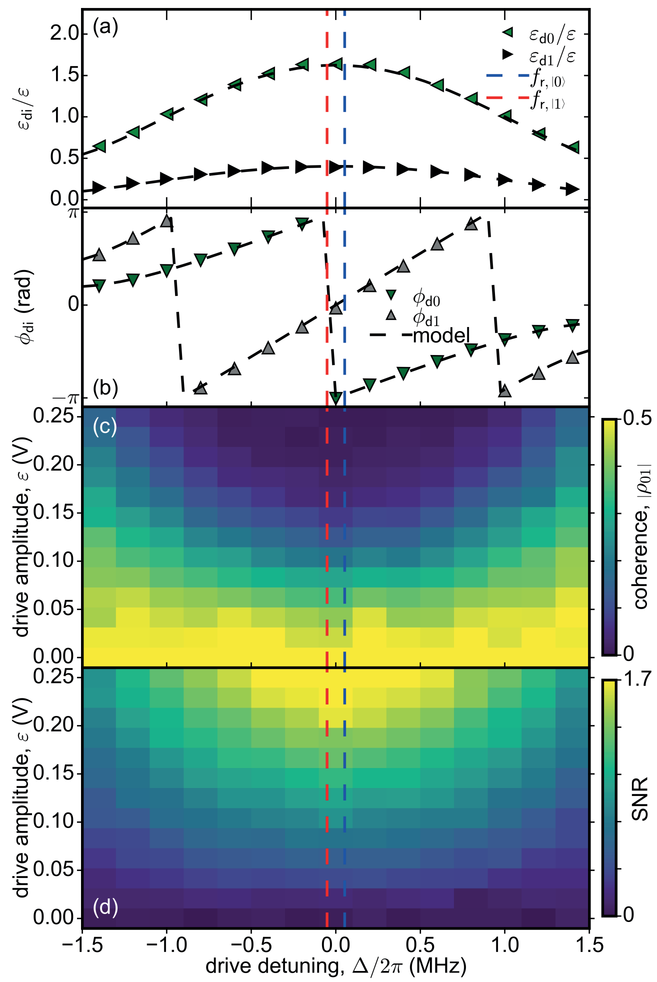

FIG. S1. Depletion pulse parameters, coherence and SNR as a function of detuning. (a,b) Depletion pulse parameters from the depletion optimizations used in Fig. 2. Dashed vertical lines indicate $f_{\mathrm{r},|0\rangle}$ (blue) and $f_{\mathrm{r},|1\rangle}$ (red). Dashed black curves are extracted from a linear model (see main text). Coherence (c) and SNR (d) as a function of drive amplitude and detuning. At non-zero ε, SNR is maximal (coherence is minimal) at the midpoint frequency $\Delta = 0$ and decreases (increases) with detuning.

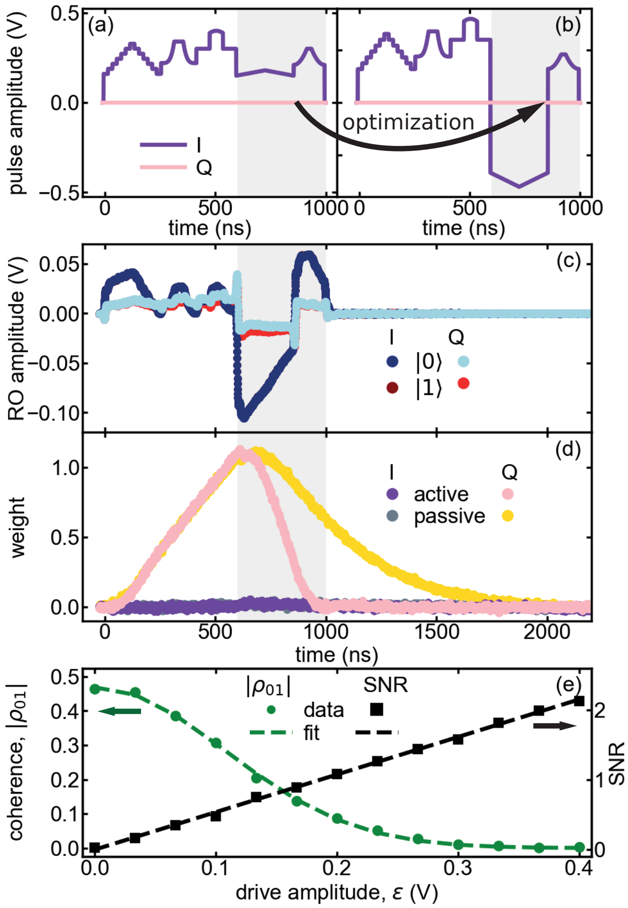

FIG. S2. The 3-step method for quantum efficiency extraction with a pulse envelope consisting of seventeenth-century Dutch canal house façade outlines. (a) Pulse envelope with five façades, of which the first three ramp up the resonator with duration $\tau_{\mathrm{up}} = 600$ ns, fixed phase $\varphi = 0$ and amplitude ε (fixed during tuneup to $\varepsilon = \varepsilon_0 = 0.4$ V) and the last two are 240 ns and 160 ns depletion segments ($\tau_{\mathrm{d}} = 400$ ns) with each a tunable phase and amplitude. (b) Optimized depletion pulse with $\varepsilon_{\mathrm{d0}} = 1.68\varepsilon$, $\varepsilon_{\mathrm{d1}} = 0.58\varepsilon$, $\phi_{\mathrm{d0}} = 1.005\pi$ rad, $\phi_{\mathrm{d1}} = 0.007\pi$ rad. (c) Averaged feedline transmission of the optimized depletion pulse. The qubit is prepared in $|0\rangle$ (blue) and in $|1\rangle$ (red). (d) Optimal weight functions extracted for the depletion pulse (purple) and as a reference, weight functions are shown for passive depletion ($\varepsilon_{\mathrm{d0}} = \varepsilon_{\mathrm{d1}} = 0$ V). (d) Quantum efficiency extraction using 13 values of ε. The best-fit values give $\eta_{\mathrm{e}} = 0.167 \pm 0.005$.

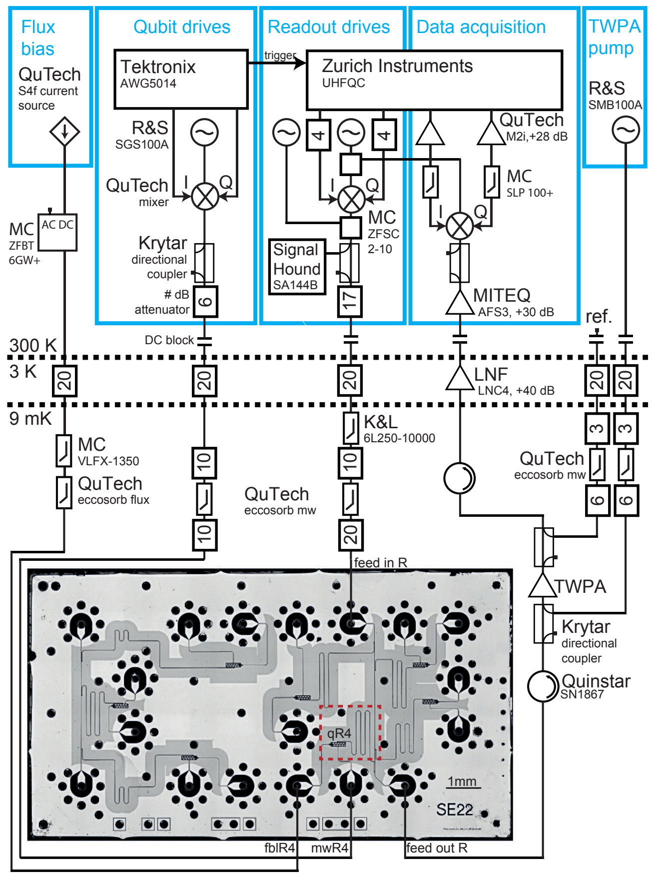

FIG. S3. Photograph of cQED chip (identical design as the one used) and complete wiring diagram of electronic components inside and outside the ^{3}He/^{4}He dilution refrigerator (Leiden Cryogenics CF-CS81). The test chip contains seven transmon qubits individually coupled to dedicated microwave drive lines, flux bias lines and readout resonators. The three (four) resonators on the left (right) side couple capacitively to the left (right) feedline traversing the chip from top to bottom. All 18 connections are made from the back side of the chip and reach the front through vertical coax lines[4]. Each vertical coax line consists of a central through-silicon via (TSV) that carries the signal and seven surrounding TSVs acting as shield connecting the front and back side ground planes. Other, individual TSVs interconnect front side and back side ground planes to eliminate chip modes.