Quantum computing architectures rely on classical electronics for control and readout. Employing classical electronics in a feedback loop with the quantum system allows to stabilize states, correct errors and to realize specific feedforward-based quantum computing and communication schemes such as deterministic quantum teleportation. These feedback and feedforward operations are required to be fast compared to the coherence time of the quantum system to minimize the probability of errors. We present a field programmable gate array (FPGA) based digital signal processing system capable of real-time quadrature demodulation, determination of the qubit state and generation of state-dependent feedback trigger signals. The feedback trigger is generated with a latency of 110 ns with respect to the timing of the analog input signal. We characterize the performance of the system for an active qubit initialization protocol based on dispersive readout of a superconducting qubit and discuss potential applications in feedback and feedforward algorithms.

I. INTRODUCTION

Recent quantum physical research is directed towards gaining experimental control of large-scale, strongly-interacting quantum systems such as trapped ions [1] and solid-state devices [2]. The ultimate goal is to realize a quantum computer [3, 4] with a large number of quantum bits (qubits) which may outperform classical computers for certain computational tasks [7–11]. However, quantum systems do not act as stand-alone components but must be combined with classical electronics to control inputs such as microwave pulses or external magnetic fields and to record and analyze the output signals [12]. Analyzing the output signals in real time can be advantageous to condition input signals on prior measurement results and therefore realize a feedback loop with the quantum system [13].

Quantum feedback schemes [14] make use of the results of quantum measurements to act back onto the quantum state of the system within its coherence time. Experimental realizations of quantum feedback have shown that it is possible to prepare and stabilize non-classical states of electromagnetic fields in optical [15, 16] and microwave [17] cavities, and to enhance the precision of phase measurements using an adaptive homodyne scheme [18].

The first demonstrations of feedback protocols with superconducting qubits showed active initialization of qubits into their ground state [19] and the stabilization of Rabi and Ramsey oscillations [20, 21]. Further recent feedback experiments with superconducting qubits demonstrated the deterministic preparation of entangled two-qubit states [22, 23], the reversal of measurement-induced dephasing [24], and the stabilization of arbitrary single-qubit states by continuously observing the spontaneous emission from a qubit [25].

Quantum feedforward schemes are closely related to quantum feedback schemes. In quantum feedforward schemes one part of a quantum system is measured while the action takes place on another part of the quantum system. A prominent example for a feedforward scheme is the quantum teleportation protocol [26], which has been realized with active feedforward in quantum optics setups [27, 28], in molecules using nuclear magnetic resonance [31], trapped ions [32, 33], atomic ensembles [34] and solid-state qubits [35, 36].

The feedback latency is commonly defined as the time required for a single feedback round, i.e. the time between the beginning of the measurement of the state and the completion of the feedback action onto the state. A general requirement to achieve high success probabilities in quantum feedback schemes is that the feedback latency is much shorter than the timescale on which the quantum state decoheres.

Analog feedback schemes such as those reported in Refs. [20, 25] feature feedback latencies on the order of 100 ns, where the latencies are limited by analog bandwidth and delays in the cables in the cryogenic setups. However, analog signal processing circuits have limited flexibility. The flexibility can be improved by using a digital signal processing (DSP) unit in the feedback loop, which can be implemented on a central processing unit (CPU) or on a field programmable gate array (FPGA) [37]. CPU-based DSP systems offer versatile and convenient programming at the cost of several microseconds latency [17] due to the delays introduced by the digital input and output of the signal, which is too slow to achieve very low error probabilities for feedback operations on superconducting qubits.

In this paper, we describe an FPGA-based feedback-capable signal analyzer which allows for real-time digital demodulation of a dispersive readout signal [38, 40] and the generation of a qubit-state-dependent trigger with...
input–to–output latency of 110 ns. Our signal analyzer is therefore among the fastest feedback-capable digital signal analyzers reported so far [21, 22, 11, 33]. The capabilities of our signal analyzer enabled the feedforward action in the deterministic quantum teleportation experiment presented in Ref. [35]. In this paper, we illustrate the use of the feedback signal analyzer in a feedback loop for qubit initialization [19] and experimentally characterize its latency and performance.

The paper is organized as follows: In Sec. II we present an overview of a typical feedback loop in which our instrument is used and analyze the feedback latency. In Sec. III we discuss the implementation of the digital signal processing on the FPGA and analyze the processing latencies. Finally, in Sec. IV we experimentally characterize the performance of the feedback loop. In the appendices, we provide more details about our experimental setup and our implementation of the digital signal processing on the FPGA.

II. OVERVIEW OF THE FEEDBACK LOOP

In this section, we explain the elements of a typical feedback loop shown in Fig. 1(a). We designed the feedback loop to issue pulses onto a superconducting qubit inside a dilution refrigerator conditioned on a measurement of the qubit state by analog and digital signal processing using cryogenic and room-temperature electronics. We first discuss the elements of the detection scheme and the actuator electronics and then present the latencies of the feedback loop. We provide a detailed description of our experimental setup in App. A.

A. Principle of the detection scheme

We consider the dispersive readout of the state of transmon qubits [44, 45] with typical frequencies \( \omega_q/(2\pi) \approx 4–6 \text{ GHz} \) for the transition between the ground \( |g\rangle \) and first excited state \( |e\rangle \). We couple a microwave resonator to the qubit [green box in Fig. 1(a)] with a frequency difference between qubit and resonator designed to be in the dispersive regime [35, 59].

In our experimental realization of the feedback loop (see Sec. IV), the qubit transition frequency is \( \omega_q/(2\pi) = 6.143 \text{ GHz} \) and the center resonator frequency amounts to \( \omega_r/(2\pi) = 7.133 \text{ GHz} \) with dispersive coupling rate \( \chi/(2\pi) \approx 1.1 \text{ MHz} \) between the qubit and the resonator. Depending on whether the qubit is in state \( |g\rangle \) or \( |e\rangle \), we observe the dispersively shifted resonator frequency at \( \omega_r \pm \chi \) respectively.

The qubit-state-dependent frequency shift leads to a state-dependent resonator response when the resonator is probed with a microwave pulse. In the dispersive readout scheme, high-fidelity quantum nondemolition readout [13] is achieved when probing the resonator with power \( \kappa\langle \hat{n} \rangle \hbar \omega_r \approx 10^{-16} \text{ W} \) such that the steady-state average photon number \( \langle \hat{n} \rangle \) in the resonator is on the order of 1–10 microwave photons [39, 40–50]. Due to the low power, it is essential to connect the output of the resonator to a Josephson parametric amplifier (JPA) [47, 61, 68] to be able to discern the qubit-state-dependent resonator response within a single repetition of the experiment and in a time shorter than the qubit lifetime. Other schemes involve the direct coupling of a qubit to a Josephson bifurcation amplifier [29, 92], autoresonant oscillator [63] or parametric oscillator [64] to be able to discern the qubit state with a higher microwave power.

For simplicity, we consider the case where the resonator is probed with a microwave pulse with frequency \( \omega_r \) and...
square envelope. The scheme considered here could be extended to include more sophisticated pulse shapes \cite{14,49,60} which increase the speed and fidelity of the readout as well as the speed of the reset of the intra-resonator field.

We employ the complex representation of the signal $I(t) + iQ(t) = A(t) \exp[\phi(t)]$ where $A(t)$ and $\phi(t)$ are the time-dependent amplitude and phase of the signal at frequency $\omega_r$. Upon transmission of the readout pulse with frequency close to resonance, the time-dependent in-phase $I(t)$ and quadrature $Q(t)$ components of the signal follow an exponential rise towards steady-state values starting at time $t_0$ after the onset of the readout pulse as illustrated in Fig. 1(b) \cite{61}. The steady-state values depend on whether the qubit is in state $|g\rangle$ (blue curve) or state $|e\rangle$ (red curve). The trajectories of the readout signal in the two-dimensional plane spanned by $I$ and $Q$ as sketched in Fig. 1(c) start at the center of the plane which corresponds to zero amplitude and move into two different directions depending on the qubit state $|g\rangle$ (blue curve) or $|e\rangle$ (red curve).

The signal is subject to noise added by passive and active components \cite{65}. Therefore we apply a linear filter to the signal with the goal to attenuate noise frequency components while keeping the frequency components that contain the signal \cite{24,40,16,69}. In particular, we apply a moving average filter which is advantageous in terms of the signal processing latency (see Sec. III.C). The moving average is equivalent to an unweighted integration of the original signal in a particular integration window starting at a variable time $t_1$ and ending at time $t_2 = t_1 + \tau_i$ [see Fig. 1(b) and Fig. 1(c)], where $\tau_i$ is a constant integration time. We define the total readout duration as the time difference $\tau_{RO} = t_2 - t_0$ between the onset of the readout pulse and the end of the integration window. In the experiment presented in Sec. IV we used an integration window of $\tau_i = 40$ ns and a readout duration of $\tau_{RO} = (105 \pm 2)$ ns.

In the absence of transitions between qubit states during the integration time, the statistical distribution of the integrated signal, when the experiment is repeated many times, is expected to be represented by two Gaussian-shaped peaks in a histogram of the $I$ component [Fig. 1(d)]. In the presence of qubit state transitions during the readout, the distributions corresponding to the states $|g\rangle$ and $|e\rangle$ are expected to be non-Gaussian with an increased overlap \cite{40,46}. We discern the states $|g\rangle$ and $|e\rangle$ of the qubit by comparing the $I$ signal to a threshold value $I_t$ [dashed line in Fig. 1(d)]. The fidelity of the readout depends on the signal-to-noise ratio of the readout signal \cite{46,49}. To maximize the readout fidelity, we optimize the integration window and threshold value $I_t$.

B. Implementation of the detection scheme

The readout pulse is issued by the static control hardware [gray box in Fig. 1(a)]. Simultaneously, the static control hardware sends a trigger [tr in Fig. 1(a)] to the FPGA to synchronize the digital signal processing with the readout pulse.

We use an analog detection chain [yellow box in Fig. 1(a)] containing amplifiers with a total gain of approximately 120 dB (see App. A) to detect the signal at the output of the resonator. In addition, the detection chain uses analog down-conversion electronics to convert the readout signal to an intermediate frequency $\omega_{IF}$ compatible with the sampling rate $f_s = 100$ MS/s of our DSP unit. We choose an intermediate frequency at a quarter of the sampling frequency, i.e. $\omega_{IF}/(2\pi) = f_s/4 = 25$ MHz, which allows for efficient digital down-conversion (see Sec. III.C). Note that in principle it is possible to directly demodulate the signal into its $I$ and $Q$ components in the analog signal processing but this requires the $I$ and $Q$ component of the signal to be digitized using two separate analog-to-digital converter (ADC) channels \cite{11,70,71}. The separate digitization of the $I$ and $Q$ components is sensitive to mismatches between the conversion-loss and reference level which lead to a distortion of the digitized complex signal. In contrast, down-conversion to an intermediate frequency in the range of 10 MHz to 1 GHz avoids low-frequency noise, DC offsets and requires only one ADC channel at the cost of a reduced bandwidth \cite{11,70,71}.

We implement the digital signal processing on a Xilinx Virtex-4 FPGA mounted on a commercial DSP unit by Nallatech, Inc. (BenADDA-V4 \cite{72}) [blue box in Fig. 1(a)] which includes an ADC with sampling rate $f_s = 100$ MS/s and 14-bit voltage resolution. In a first step, the DSP digitally demodulates the signal [labeled as demod. in Fig. 1(a)]. The state discrimination module [state det. in Fig. 1(a)] then compares the filtered $I$ signal at time $\tau_{RO}$ to the threshold $I_t$ to determine the qubit state from the demodulated signal. Depending on the determined qubit state, a feedback trigger [fb in Fig. 1(a)] is sent from the FPGA to the actuator electronics.

C. Actuator

The actuator is realized with an arbitrary waveform generator (AWG). When it receives the feedback trigger, the AWG generates a feedback pulse with a sampling rate of 1 GHz. In our experiment, the actuator pulse (AP) has a duration of $\tau_{AP} = 28$ ns and uses the derivative removal by adiabatic gate (DRAG) technique \cite{72,73} to prevent transitions to higher-excited states of the transmon outside of the subspace spanned by the states $|g\rangle$ and $|e\rangle$. We typically generate the actuator pulse with a carrier frequency of 100–300 MHz limited by the bandwidth of the AWG and analog mixer. In the experiment presented in Sec. IV we chose a carrier frequency of 100 MHz for the
actuator pulse. We use an analog mixer to up-convert the actuator pulse to the qubit transition frequency, which is typically in the range of 4–6 GHz. Forwarding this pulse to the qubit realizes a conditional quantum gate on the qubit closing the feedback loop.

**D. Latencies**

We define the latency $\tau_{FB}$ of the feedback loop [Fig. 1(a)] as the time from the beginning of the readout pulse until the completion of the feedback pulse, i.e.

$$\tau_{FB} \equiv \tau_{EL,tot} + \tau_{RO} + \tau_{AP}, \quad (1)$$

where $\tau_{EL,tot}$ is the total electronic delay of the signal in the analog and digital components and cables of the feedback loop, $\tau_{RO}$ the readout duration (see Sec. II A) and $\tau_{AP} = 28$ ns is the length of the actuator pulse (see Sec. II C). We measured the total electronic delay $\tau_{EL,tot} = (219 \pm 2)$ ns in-situ by changing the up-conversion frequency of the feedback pulse to the resonance frequency of the readout resonator and adjusting the amplitude of the pulse. The resonant feedback pulse is transmitted through the resonator which makes it possible to determine the timing of the feedback pulse relative to the readout pulse. By adding up the contributions according to Eq. (1) we infer a feedback latency of $\tau_{FB} = (352 \pm 3)$ ns.

The electronic delay

$$\tau_{EL,tot} \equiv \tau_{proc} + \tau_{ADC,DIO} + \tau_{AWG} + \tau_{G,tot}, \quad (2)$$

can be broken up into accumulated contributions. The signal processing, which we implemented in the FPGA, introduces a processing delay of three clock cycles $\tau_{proc} = 30$ ns (see Sec. II A). The feedback trigger is delayed by $\tau_{proc} + \tau_{ADC,DIO} = (110 \pm 3)$ ns with respect to the analog input signal, where $\tau_{ADC,DIO}$ is the delay introduced by the ADC and digital interfaces (see App. B).

By subtracting the separately determined quantities $\tau_{proc}$, $\tau_{ADC,DIO}$ and $\tau_{AWG}$ from the total electronic delay $\tau_{EL,tot}$ we estimate the inferred total group delay $\tau_{G,tot} = (69 \pm 7)$ ns in the cables and analog components. We expect the total cable length connecting the analog and digital components to be the dominant contribution to the inferred group delay. The inferred group delay corresponds to an approximate total cable length of 14 m considering an effective dielectric constant $\epsilon_{eff} \approx 2$ for the coaxial cables with PTFE dielectric. This inferred total cable length is consistent with the experimental setup. The cable length in our setup could be reduced further by placing the individual components of the feedback loop closer to each other which can be achieved, for example, by placing the FPGA and control electronics inside the dilution refrigerator [74,76].

**III. FPGA-BASED DIGITAL SIGNAL PROCESSING**

In this section, we describe our digital signal processing (DSP) circuit which we implemented on the Virtex–4 FPGA. To derive feedback triggers, the DSP circuit (Fig. 2) determines the qubit state by digital demodulation of the readout signal (see Sec. II). We start by discussing the digitization and synchronization of the input signal. Next, we discuss the signal processing features of each block and the corresponding latencies. Details of the FPGA implementation of each signal processing block are discussed in App. C. We analyze the FPGA timing and resource usage for the implementation of the DSP circuit on the Xilinx Virtex–4, Virtex–6 and Virtex–7 FPGA in App. D.

**A. Digitization of the input signal**

Before entering the DSP circuit, the readout signal is digitized by an external ADC chip which samples the signal with rate $f_s = 100$ MS/s. Typical readout signals are sine waves with qubit-state-dependent amplitude and phase as shown in Fig. 3(a). We parameterize the time-dependent voltage at the input of the ADC as

$$V_{ADC}(t) = A(t) \cos(\omega_{IF} t + \phi(t))$$

$$= \frac{A(t)}{2} \left( e^{i(\omega_{IF} t + \phi(t))} + e^{-i(\omega_{IF} t + \phi(t))} \right). \quad (3)$$

As discussed in Sec. II B we choose an intermediate frequency of $\omega_{IF}/(2\pi) = f_s/4 = 25$ MHz for the readout signal after analog down-consultation (see Sec. II B) which is a useful choice for digital demodulation as discussed below. The time-dependent amplitude $A(t)$ is proportional to the amplitude $A(t)$ of the field at the output of the resonator scaled by the gain of the analog detection chain and conversion loss of the mixer.

The ADC samples the signal $V_{ADC}(t_n)$ at discrete times $t_n = n/f_s = n \times 10$ ns with index $n$. The ADC encodes the input voltage range of approximately $\pm 1$ V as 14-bit fixed-point binary values. The fixed-point representation leads to a discretization step size of $2^{-13} \approx 0.12$ mV. A trigger pulse (tr) is provided together with the analog signal via a separate digital input of the FPGA to mark the onset of the readout pulse.

**B. Pipelined processing**

We designed the DSP circuit to process the signal from the ADC in a pipelined manner. The signal from the ADC is initially buffered in a register implemented by synchronous D-flip-flops (ADC $z^{-1}$ block in Fig. 2) which forward the value of the signal at each event of a rising edge of the sampling clock to the next processing element in the pipeline.
A separate trigger input (tr, orange lines in Fig. 2) marks the beginning of each experimental repetition. In order to synchronize the trigger with the ADC signal, the trigger initially goes through six pipelined registers ($z^{-6}$ in Fig. 2), which compensate the difference in delay between the ADC line and trigger line. To synchronize the signal processing with the sampling clock, we insert further pipelined registers into the signal and trigger lines at specific points in the circuit (blue dashed lines in Fig. 2).

C. Digital demodulation

As discussed in Sec. 11B, we digitally demodulate the readout signal to obtain the $I$ and $Q$ components of the signal. Digital demodulation is achieved by digital frequency down-conversion which involves digital mixing of the signal with a digital reference oscillator followed by digital low-pass filtering to remove noise and unwanted sideband frequency components [71].

1. Digital mixing

In the first part of the digital demodulation circuit (yellow box in Fig. 2), we implement a digital mixing method [70, 71] (digital mixer in Fig. 2) to obtain a sideband at zero frequency. In the digital mixer, the input signal $V_{ADC}$ as defined in Eq. (3), is multiplied with a complex exponential with down-conversion frequency $\omega_{IF}$ to obtain a complex output signal $S_m$:

$$S_m(t_n) = V_{ADC} \times e^{-j \omega_{IF} t_n} = \frac{A(t)}{2} \left( e^{j \phi(t)} + e^{-j(2 \omega_{IF} t + \phi(t))} \right).$$

The action of the multiplication is to generate two sidebands corresponding to the two complex exponentials in Eq. (4): one is corresponding to the complex signal $I + iQ = A(t)e^{j\phi(t)}/2$ and the other leads to oscillations with frequency $2\omega_{IF}$ of the output signals of the mixer [Fig. 3 b)]. The complex signal $I + iQ$ is the basis on which we determine the state of the qubit after filtering out the oscillating sideband (see following sections).

In practice, the real ($\text{Re}[S_m]$) and imaginary ($\text{Im}[S_m]$) parts of the output signal of the mixer are computed separately by multiplying the input signal with a discrete cosine to obtain the real part and with a discrete negative sine to obtain the imaginary part. The FPGA implementation of the digital mixer is described in App. C1. For $\omega_{IF}/(2\pi) = f_s/4$, the digital mixer introduces a latency of less than one clock cycle (10 ns) due to its multiplierless implementation [70, 71]. Since the output signal of
the mixer is registered by synchronous D–flip–flops, the effective latency is one clock cycle. For synchronization, the trigger signal (tr) is delayed by one clock cycle [$z^{-1}\text{tr}$ in Fig. 3(b)].

2. Digital low-pass filter

The second essential part of the digital down–conversion circuit is a digital low-pass filter, which extracts the $I$ and $Q$ components from the signals $\Re[S_m]$ and $\Im[S_m]$ by removing the sideband spectral components oscillating at frequency $2\omega_{\text{IF}}$ [71]. We implement the digital low-pass filter as a finite impulse response (FIR) filter [71] which is a discrete convolution of the digital signal with a finite sequence of filter coefficients. By matching the filter coefficients (integration weights) to the expected resonator response, it is possible to optimize the single-shot readout fidelity [24, 40, 42, 45, 46, 59]. While our DSP circuit in principle allows for 40-point FIR filters with arbitrary filter coefficients, a moving average is the simplest type of FIR low-pass filter which is possible to implement without multipliers and therefore has a reduced processing latency and uses less FPGA resources than a more general FIR filter. The FPGA implementation of the moving average module is described in App. C2.

The moving average (FIR filter in Fig. 2) is applied separately to the real part ($\Re[S_m]$) and imaginary part ($\Im[S_m]$) of the complex output signal of the digital mixer, $S_m$, leading to

$$I(t_n) + iQ(t_n) = \frac{1}{l} \sum_{k=n-l+1}^{n} S_m(t_k), \quad (5)$$

which is a discrete convolution with a square window of length $l$. In the limit of negligible modulation bandwidth, the moving average filters a sinusoidal perfectly if the window length $l$ is a multiple of the oscillation period. In the case of $\omega_{\text{IF}}/(2\pi) = f_s/4$, the periodicity of the unwanted terms at $2\omega_{\text{IF}}$ is equal to two discrete samples. Therefore any window length which spans an even number of samples is suitable to filter out the $2\omega_{\text{IF}}$ sideband.

The output of the moving average with window length $l = 4$ is shown in Fig. 3(c). The $I$ and $Q$ signals at the output of the moving average show a smooth ramp towards a steady-state value. In the simulated signals shown in Fig. 3 an appropriate global phase offset has been chosen such that the difference between the traces corresponding to the $|g\rangle$ and $|e\rangle$ state is maximized in the $I$ component of the signal (see Sec. II A).

The moving average module has a latency of one clock cycle. The trigger is delayed accordingly by one additional clock cycle ($z^{-1}z^{-1}\text{tr} = z^{-2}\text{tr}$) for synchronization.

D. Offset subtraction and scaling

Following the FIR filter block, the $I$ and $Q$ signals enter blocks which perform offset subtraction and scaling of the signal (green boxes in Fig. 2). The main purpose of offset subtraction is to set a threshold value as described in Sec. III E. Moreover, offset subtraction and scaling allows to make best use of the fixed range and resolution used for recording histograms (see Sec. III F).

The outputs of the offset subtraction and scaling blocks are described by

$$\hat{I}(t_n) \equiv m_I(I(t_n) - c_I) \quad (6)$$
$$\hat{Q}(t_n) \equiv m_Q(Q(t_n) - c_Q), \quad (7)$$

where $c_I$ and $c_Q$ are offsets in the I/Q plane and $m_I$ and $m_Q$ are multiplication factors. We determine the parameters ($c_I, c_Q$) and ($m_I, m_Q$) in a calibration measurement. The latencies of the offset subtraction and scaling blocks are less than one clock cycle and no synchronous D–flip–flops are used.

E. State discrimination module

The state discrimination module (red box in Fig. 2) determines the state of the qubit based on the preprocessed input signals $\hat{I}$ and $\hat{Q}$. Due to the offset subtraction, the threshold value for state discrimination can be kept fixed at zero which simplifies the FPGA implementation of the state discrimination module as discussed in App. C4.

The readout time $\tau_{\text{RO}}$ relative to the onset of the readout pulse (see Sec. II A) is specified with a variable delay of $d$ clock cycles after the detection of the trigger signal, i.e. $d \times 10\text{ ns} = \tau_{\text{RO}}$. In the example shown in Fig. 3(c), the $|g\rangle$ and $|e\rangle$ states of the qubit are discriminated based on a threshold value (thick horizontal bar) defined for the $I$ signal at a time $t = 160\text{ ns}$ which is $d = 14$ clock cycles after the detection of the trigger signal $z^{-2}\text{tr}$. The simulated $I$ signals corresponding to the $|0\rangle$ [blue curve in Fig. 3(c)] and $|1\rangle$ [red curve in Fig. 3(c)] state are well distinguishable at the time when the threshold is checked, such that the state of the qubit can be determined successfully even in presence of noise (see Sec. IV). The state discrimination module either issues the feedback trigger [red curve in Fig. 3(d)] or does not issue the feedback trigger [blue curve in Fig. 3(d)] based on the determined qubit state.

Our DSP circuit provides the possibility to derive a second feedback trigger (fb2 in Fig. 2) based on both the in-phase ($\hat{I}$) or quadrature ($\hat{Q}$) signal components. For example, in the quantum teleportation protocol [26] the states of two qubits at the sender’s location are measured in order to perform a state-dependent rotation on a qubit at the receiver’s location. In our experimental realization of the teleportation protocol as discussed in Ref. [35], we discriminated the states of the two sender’s qubits based on two threshold values defined for the $I$ and $Q$ signals.
Based on the outcome of comparing the $I$ and $Q$ signals to the two threshold values, we issued two independent trigger signals to two separate AWGs in order to implement a conditional operation on the receiver’s qubit \[35\].

\section*{F. Histogram module}

The histogram module records how often the values of the signals $I$ and $Q$ obtained from a specific integration window fall into a particular histogram bin when the experiment is repeated many times. The bins are defined by subdividing the signal range from -1 to +1 into typically 128 bins. From the histogram, an estimate of the probability density function of the signal at the specified times is obtained.

We typically repeat the experiment $10^5$–$10^7$ times to obtain standard deviations of less than a part per thousand for the counts in each histogram bin. Storing the histogram of the signal needs less memory than storing the value of the signal in each repetition if the number of repetitions exceeds the number of histogram bins. The histogram module therefore allows for data reduction at the time when the data is recorded.

We have used the histogram module in previous experiments to characterize the quantum statistics of microwave radiation emitted from circuit QED systems \[77–80\]. In the context of feedback experiments, we record histograms to obtain the probabilities of observing a particular qubit state in two consecutive qubit readouts as described in Sec. IV.

We update the histogram at the same time as the state discrimination module determines the qubit state in order to analyze the readout fidelity and feedback performance (see Sec. IV). We synchronize the state discrimination module and the histogram module using a marker signal (fBTime in Fig. 2) which is sent from the state discrimination module to the histogram module. We use an external Zero Bus Turnaround (ZBT) Random Access Memory (RAM) (see Fig. 2) to store the histogram. When the recording of the histogram is completed, we transfer the histogram to the host computer via the interface. The implementation details of the histogram module are described in App. C5

\section*{IV. QUBIT STATE INITIALIZATION EXPERIMENT}

In this section, the functionality of the presented DSP circuit is demonstrated in the context of a qubit state initialization experiment. In the experiment we use the feedback loop to reset the state of a superconducting qubit \[19, 81, 83\] (see App. E) deterministically into its ground state, independent of its initial state. We correlate the outcomes of two consecutive qubit measurements in order to separate out the different effects such as the qubit lifetime and readout fidelity which contribute to the overall performance of the feedback protocol.

We choose the repetition period 10$\mu$s of the experiment to be longer than the qubit lifetime $T_1 \approx 1.4\mu$s, such that the qubit is approximately in thermal equilibrium with its environment at the beginning of each experimental repetition. We observe a finite thermal population $P_{\text{therm}} \approx 7\%$ of the excited state $|e\rangle$ due to the elevated effective temperature of about 114 mK of the system on which the experiments were performed (see App. F).

In order to test the feedback protocol, we prepare an equal superposition of the computational states $|g\rangle$ and $|e\rangle$ of the superconducting qubit. This choice of initial state will ideally lead to equal probabilities to find the states $|g\rangle$ and $|e\rangle$ when the qubit is measured. Preparing an equal superposition as an initial state will therefore test the feedback actuator for both computational states $|g\rangle$ and $|e\rangle$ of the qubit. An additional data set (App. F) shows that the feedback scheme can also be used to reduce the thermal population of the excited state \[19, 82, 84\], providing an additional benchmark for our feedback loop.

Ideally, we consider the case when the qubit is initialized in the state $|g\rangle$ corresponding to the Bloch vector pointing to the upper pole of the Bloch sphere [stage 1 in Fig. 4(a)]. A microwave pulse at frequency $\omega_0$ [green line in Fig. 1(b)] is applied to the qubit to realize a $\pi/2$ rotation which brings the qubit into the superposition state $|+\rangle \equiv (|g\rangle + |e\rangle)/\sqrt{2}$ corresponding to a Bloch vector pointing at the equator of the Bloch sphere [stage 2 in Fig. 4(a)].

When the qubit initially is in state $|e\rangle$, for example due to the non-zero temperature of the system, the effect of the $\pi/2$ rotation is to prepare the state $|-\rangle \equiv (|g\rangle - |e\rangle)/\sqrt{2}$ which is an equal superposition of $|g\rangle$ and $|e\rangle$ with a different phase. The states $|+\rangle$ and $|-\rangle$ are expected to lead to an identical distribution of outcomes in the state detection.

In the experiment, directly after the preparation of the initial state, at time $\tau_{\text{M1}} = 0$, the state of the qubit is measured with a readout pulse of length 160 ns (see M1 in Fig. 4) applied to the resonator. The dispersive readout projects the state of the qubit into either the ground or excited state corresponding to the upper and lower pole of the Bloch sphere [stage 3 in Fig. 4(a)]. The DSP (see Sec. I1) extracts the in-phase component $I_1$ during the readout pulse M1. We filter the signal $I_1$ with a moving average of four consecutive samples, corresponding to an integration window [blue region M1 in Fig. 4(b)] of 40 ns. We extracted the time $\tau_{\text{RO}} \approx 105$ ns of the end of the integration window \[85\] relative to the beginning of the readout pulse by fitting a theoretical model to the switch-on dynamics of the readout signal in a time-resolved measurement \[46\].

The histogram of $I_1$ [blue dots in Fig. 4(c)] reveals two Gaussian peaks corresponding to the distributions of the in-phase signal for the qubit being in state $|g\rangle$
FIG. 4. (a) Quantum circuit depicting the experimental protocol to test the feedback routine. The state at each stage of the protocol is represented on the Bloch sphere (blue). The horizontal black line indicates the evolution of the qubit state over time. Double arrows (⇒) represent the flow of classical information. The sequence of operations is: a $\pi/2$ rotation (green box) of the Bloch vector about an equatorial axis, a first projective measurement (M1), a conditional $\pi$ rotation (red dashed box) that depends on the feedback trigger (fb) determined by the digital signal processing (DSP) and a second projective measurement (M2) of the qubit state. (b) Pulse scheme showing the timing of microwave pulses applied to the qubit (green trace), the pulses depends on the feedback trigger (fb) determined by the digital signal processing (DSP) and a second projective measurement $\pi$ the Bloch vector about an equatorial axis, a first projective measurement (M1), a conditional $\pi$ rotation (red dashed box) of the qubit state with characteristic time $T_1$ mark the integration windows of the measurements M1 and M2. The time offset $\tau_{RO}$ marks the time from the beginning of each readout pulse to the end of the corresponding integration window, $\tau_{FB}$ marks the delay in the feedback electronics and $\tau_{RO,tot}$ marks the total feedback latency as defined in the main text. (c) Histograms of the in-phase signal $I_1$ obtained from the first readout pulse M1 (blue dots) and in-phase signal $I_2$ obtained from the second readout pulse M2 (orange dots) for the case when the feedback actuator is disabled. The dashed line marks the feedback threshold. For M1 and M2 the percentage of counts on the right side of the threshold is indicated. (d) The same type of histograms as in (c) but with the feedback actuator enabled. (e) Two-dimensional histogram with 128 × 128 bins counting the combined outcomes of the first readout $I_1$ (horizontal axis) and second readout $I_2$ (vertical axis) for the case when the feedback actuator is disabled. The plane is divided into four regions ($R_{GG}$, $R_{GE}$, $R_{EG}$, $R_{EE}$) separated by the threshold (dashed lines). The percentage of counts relative to the total count is indicated in each quadrant. Red lines are contour lines marking specific counts of $\{0.5, 1, 2\} \times 10^3$. (f) The same type of two-dimensional histogram as in (e) but with the feedback actuator enabled.

or $|e\rangle$. The measured initial excited state probability $P[E_1]_{fb\; off} = 46.06(3)\%$, is the fraction of counts of values $I_1$ above the threshold value $I_t = 16\; mV$ [dashed line in Fig. 4(c)] relative to the total count $C_{tot} \approx 2097152$ of measurements.

With a master equation $[86]$ we simulate the decay of the qubit state with characteristic time $T_1 = 1.4\; \mu s$ during the time of the $\pi/2$ pulse and the readout up to the center of the integration window [see Fig. 4(b)]. Furthermore we take into account a bias of the measured probabilities towards 50% due to the finite readout error of 3% (see App. [C]). From the master equation simulation we obtain an expected excited state probability $P[E_1]_{sim} = 47.07\%$ in the first measurement M1 which agrees reasonably with the measured probability $P[E_1]_{fb\; off}$ (see above). A source of systematic errors is measurement-induced mixing $[87]$. An additional reason for the systematic deviation of the measured probability from the simulated probability is that the chosen threshold value $I_t = 16\; mV$ deviates from the value $I_{t,\; opt} \approx 13\; mV$ which optimizes readout fidelity (see App. [C]). This offset leads to a bias of the observed probabilities towards the ground state in addition to a systematic bias due to state transitions during the integration time $[40]$.

The feedback loop is configured to deterministically prepare the state $|g\rangle$ [stage 4 in Fig. 4(a)]. The feedback pulse, inducing a $\pi$ rotation of the Bloch vector of the
qubit, turns the state \(|e\rangle\) into \(|g\rangle\) and vice versa. Thus, the feedback \(\pi\) pulse is issued only if the first measurement \(M1\) revealed the qubit to be in state \(|e\rangle\). The \(\pi\) pulse [red dashed line in Fig. 4(b)] arrives at the qubit with delay of \(\tau_{\text{EF, tot}}\) (see Sec. II D) conditioned on the readout result of \(M1\).

For verification, a second readout pulse (\(M2\) in Fig. 4) is applied to the qubit at the time \(t_{M2} = 360\) ns directly after the arrival of the feedback pulse at the qubit. The difference between \(t_{M2}\) and the beginning of the first readout pulse corresponds to the total feedback latency \(\tau_{\text{FB}}\) (see Sec. II D). We recorded histograms of \(I_2\), which is the filtered in-phase component of the signal at time \(t_{M2} + \tau_{\text{RO}}\). When the feedback actuator is disabled, the histogram of \(I_2\) [orange dots in Fig. 4(c)] shows reduced counts on the right side of the threshold with an excited state probability of \(P[E_2]_{\text{fb off}} = 34.97(3)\%\). Extending the master equation simulation introduced above to include the full pulse sequence up to the second readout pulse, we obtain \(P[E_2]_{\text{fb off, sim}} = 37.89\%\) in reasonably good agreement with the measured value. The state decay between \(M1\) and \(M2\), which leads to the observed reduction in the excited state population, causes errors in the feedback action as discussed below.

When the experiment is repeated with the feedback actuator enabled, the double-peaked histogram obtained from the first readout \(I_1\) [blue dots in Fig. 4(d)] is approximately identical to the case without feedback, as expected, with the measured excited state probability \(P[E_1]_{\text{fb on}} = 46.02(3)\%\) agreeing with \(P[E_1]_{\text{fb off}}\) within the statistical error bars. After the feedback pulse, in the histogram of \(I_2\) [orange dots in Fig. 4(d)], the measured excited state probability is significantly reduced to \(P[E_2]_{\text{fb on}} = 13.23(2)\%\). This probability compares reasonably well with the simulated value of \(P[E_2]_{\text{fb on, sim}} = 10.50\%\) obtained from the master equation simulation introduced above. We attribute the difference between the measured and simulated value of \(P[E_2]_{\text{fb on}}\) to measurement-induced mixing and the deviation of the feedback threshold from the optimal value (see above).

To obtain a figure of merit for the feedback protocol that is independent of characteristics such as state decay and temperature of the quantum system, we study correlations between the outcomes of the two readout pulses \(M1\) and \(M2\). From the two-dimensional histograms [Fig. 4(e,f)] with axes \(I_1\) and \(I_2\), we obtain experimental probabilities to observe a specific range \(R\) of two consecutive measurement outcomes \((I_1, I_2)\). The probabilities \(P[R_{xy}]\) correspond to observing the qubit in state \(x\) with the first readout pulse and consecutively in state \(y\) with the second readout pulse. These probabilities are obtained from the normalized counts in the four quadrants \((R_{GG}, R_{GE}, R_{EG}, R_{EE})\) separated by the threshold [dashed lines in Fig. 4(e,f)].

When the feedback is enabled, the measured probability \(P[R_{EE}]_{\text{fb on}} = 11.57(2)\%\) [Fig. 4(f)] corresponds to the unwanted event of the state \(|e\rangle\) being observed consecutively with both readout pulses. We explain the dominant contribution to \(P[R_{EE}]_{\text{fb on}}\) by state decay between the first readout pulse and the feedback pulse. The probability of state decay between the first and second readout pulse is extracted from a reference measurement of \(P[R_{EE}]_{\text{fb off}} = 12.97(2)\%\) [Fig. 4(e)] when the feedback is disabled. The probabilities \(P[R_{EE}]_{\text{fb on}}\) and \(P[R_{EE}]_{\text{fb off}}\) are close to each other since the conditional \(\pi\) pulse swaps the state \(|g\rangle\) with \(|e\rangle\) before the second readout pulse. The corresponding simulated probabilities \(P[R_{EE}]_{\text{fb on, sim}} = 8.32\%\) and \(P[R_{EE}]_{\text{fb off, sim}} = 11.37\%\) (see Tab. I) are in reasonable agreement with the experimental values considering the sources of systematic errors as discussed above.

The measured probability \(P[R_{EE}]_{\text{fb on}} = 1.67(1)\%\) [Fig. 4(f)] of a transition from state \(|g\rangle\) to \(|e\rangle\) when the feedback loop is enabled is close to the reference value \(P[R_{EE}]_{\text{fb off}} = 1.88(1)\%\) [Fig. 4(e)] when the feedback is disabled. This shows that the state is correctly left unchanged when the qubit is already in state \(|g\rangle\). A possible reason for the small systematic deviation of \(P[R_{EE}]_{\text{fb on}}\) from \(P[R_{EE}]_{\text{fb off}}\), which is on the order of 0.2%, could be drifts in the experimental parameters such as the phase of the readout signal.

In summary, the probabilities of the combined events (Tab. I) show that in the feedback protocol the \(\pi\) pulse is applied only when it is intended and that the probability of the unwanted events in region \(R_{EE}\) is limited by state decay between the first measurement and the feedback pulse.

### V. CONCLUSIONS AND DISCUSSION

We developed a low-latency FPGA-based digital signal processing unit for quantum feedback and feedforward applications such as the qubit initialization scheme presented in this paper and the deterministic quantum teleportation realized in Ref. 35.

Our experimental results show that the feedback loop performs as expected. The total feedback latency amounts to \(\tau_{\text{FB}} = (352 \pm 3)\) ns determined by the sum

|               | Feedback off | Feedback on |
|---------------|--------------|-------------|
|               | exp. sim.    | exp. sim.   |
| \(P[R_{GG}]\) | 52.06(3)%    | 50.74%      |
| \(P[R_{EE}]\) | 1.88(1)%     | 2.18%       |
| \(P[R_{EG}]\) | 12.97(2)%    | 11.37%      |
| \(P[R_{EB}]\) | 33.10(3)%    | 35.71%      |
of ADC latency, processing latency, AWG latency, cable delays, readout time and feedback pulse duration. To reduce the probability of state decay between the state detection and the feedback action, the ratio $r \equiv \tau_B/T_1$ of the feedback latency to the qubit lifetime $T_1$ needs to be reduced. Since the probability of state decay is expected to be proportional to $1 - \exp(-r)$, a $T_1$ time of about 40 $\mu$s would be needed to achieve error probabilities of less than 1% in one iteration of the feedback scheme presented in this work. Conversely, with the longest $T_1$ times achievable with state-of-the-art superconducting circuits of up to approximately 100 $\mu$s, feedback latencies of less than 100 ns would be needed to reduce the error probability to less than one part per thousand. In the present work, we demonstrated digital processing latencies on the order of 30 ns, which are among the shortest latencies reported for FPGA-based signal analyzers in the context of superconducting qubits. Simultaneously, the usage of advanced readout strategies enables a shorter optimal readout times. Shorter latencies for analog-to-digital conversion and cable delays may be achievable by using custom-made circuit boards which work at cryogenic temperatures or by on-chip logical elements.

Low latency feedback loops may play a role in realizing future quantum computers, where a key ingredient is quantum error correction in which error syndromes of a quantum error correction code are detected by repetitive measurements. The syndrome measurements are designed to keep track of unwanted bit flip and phase errors. In this context it is essential to have a flexible low latency classical processing unit to process the error syndromes without causing additional delay for the quantum processor. A large set of quantum error correction codes may work with a passive ‘Pauli frame’ update, however, it still remains an open question whether some level of correction and qubit reset using active feedback is preferable. Therefore, having a low latency signal processor with feedback capabilities as presented in this work, will be instrumental for scaling up quantum technologies.

ACKNOWLEDGMENTS

The authors would like to thank Deniz Bozyigit for initial contributions to the FPGA firmware. The authors further acknowledge useful discussions with Johannes Heinsoo, Sebastian Krinner, Markus Oppliger and Lars Steffen.

The authors acknowledge financial support by the National Centre of Competence in Research Quantum Science and Technology (NCCR QSIT), a research instrument of the Swiss National Science Foundation (SNSF), by the Swiss Federal Department of Economic Affairs, Education and Research through the Commission for Technology and Innovation (CTI), by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the U.S. Army Research Office grant W911NF-16-1-0071 and by ETH Zurich. 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.

Appendix A: Experimental setup

The device under test (DUT, green box in Fig. 5) is a superconducting circuit with one superconducting transmon qubit. The DUT is thermalized to the 20 mK stage of a dilution refrigerator (purple box in Fig. 5).

Single-qubit quantum gates are realized by driving transitions between the ground and first excited state of the transmon by applying microwave pulses through a dedicated microwave line (port A in Fig. 5). The microwave line is thermalized by attenuators at three temperature stages $T = (4 \text{ K}, 100 \text{ mK}, 20 \text{ mK})$. The attenuators reduce the signal and noise coming from the room-temperature electronics and add Johnson-Nyquist noise at their respective temperature stages, thereby reducing the effective temperature of the microwave radiation in the cable. The qubit pulses for static control (grey box in Fig. 5) are generated by AWG 1 and up-converted to microwave frequencies using an I/Q mixer driven by a local oscillator (LO) signal from microwave generator MWG 1.

Readout of the qubits is realized by a pulsed measurement of the transmission of microwaves through a coplanar waveguide resonator (CPWR). The readout pulse is applied to the CPWR through the resonator drive line (port B in Fig. 5). The readout pulses are also generated by AWG 1. An I/Q mixer with an LO signal from MWG 2 allows for shaping the readout pulses which can be useful to achieve faster ring-up and ring-down of the intra-cavity field. In order to adjust the power range for the resonator drive a variable attenuator is used at the RF output of the mixer.

The transmitted signal is directed through an isolator, circulator, and directional coupler to a Josephson parametric amplifier (JPA) based on a $\lambda/4$ resonator shunted with an array of SQUID loops. The isolators and circulators protect the DUT from pump leakage and thermal noise. The pump tone needed to achieve a gain of approximately 20 dB in the JPA is derived via splitters from the same microwave generator MWG 2 as is used for the readout pulses which reduces drifts of relative phase between the two signals. Low phase noise is essential if the JPA is operated in a phase-sensitive mode. The pump signal (port E in Fig. 5) is combined with the signal from the resonator through a directional coupler.

Both the signal and pump tone are reflected from the
JPA. To avoid saturation of the subsequent amplifiers, we destructively interfere the reflected pump tone with a cancellation tone applied to the directional coupler (port D in Fig. 5). The phase and amplitude of the cancellation tone are adjusted using a variable phase shifter and attenuator.

After amplification by the JPA, the signal is passed via isolators which attenuate reversely propagating radiation towards a high-electron-mobility transistor (HEMT) amplifier to further amplify the signal with a gain of 40 dB before it exits the dilution refrigerator (port C in Fig. 5).

In the detection electronics (yellow box in Fig. 5) at room temperature, the signal is amplified further using low-noise microwave amplifiers. In order to reduce noise below the frequencies of interest, the signal is high-pass filtered with a cut-off frequency of about 4 GHz. The carrier frequency of typically 7 GHz is converted down to an intermediate frequency (IF) using an analog I/Q mixer and a separate microwave generator, MWG 3, for the LO signal. The IF signal at the I output of the mixer is further amplified using an IF amplifier and low-pass filters are used to suppress noise outside the detection bandwidth of the ADC (50 MHz). Attenuators between the amplifiers and the mixer are used to suppress stand-

FIG. 5. Schematic of the experimental setup used for quantum feedback. Cyan arrows point into the direction of the signal flow in the feedback loop. The device under test (DUT) is a superconducting circuit comprised of a qubit coupled to a coplanar waveguide resonator (CPWR). The color scheme of the blocks DUT (green), static control (grey), detector (yellow), Nallatech BenADDA-V4 card (blue) and actuator (red) corresponds with Fig. 1 in the main text. In addition, the three different shades of purple in the dilution refrigerator (purple box) indicate the temperature stages (20 mK, 100 mK and 4 K) to which the corresponding components are thermalized. The signal ports at the dilution refrigerator are labeled with letters in circles (A-E).
After amplification and analog down-conversion, the signal is digitized by the ADC and forwarded to the FPGA on the Nallatech BenADDA-V4TM card. The digital signal processing (DSP) circuit which we implemented on the FPGA generates a feedback trigger conditioned on the digitized and processed signal (see Sec. III).

The feedback trigger is forwarded to AWG 2 which is part of the actuator electronics (red box in Fig. 5). When receiving the feedback trigger, AWG 2 generates a pulse which is up-converted to the qubit frequency, typically at 5–6 GHz, using an I/Q mixer and LO from microwave generator MWG 4. Bias-tees allow to compensate unwanted DC offsets of the I/Q inputs of the mixer in order to suppress LO leakage. The up-converted microwave pulses are forwarded to the qubit (port A in Fig. 5).

All AWGs, MWGs, as well as ADC and DSP clocks are synchronized to a 10 MHz sine wave from an SRS FS725 rubidium frequency standard.

Appendix B: Latency of analog to digital conversion and digital input

The ADC latency and digital input–output latencies of the FPGA are inferred from the timing relative to the input trigger and feedback trigger. When the variable delay in the state discrimination module (see App. C 4) is set to \( d = 1 \) clock cycle, we measure the delay from the trigger input to the feedback trigger with an oscilloscope to be \( \tau_{\text{tr–fb}} = 110\,\text{ns} \pm 3\,\text{ns} \). Since the input trigger is synchronized with the digitized signal from the ADC in the DSP circuit, we infer that the ADC delay and digital input–output delay is \( \tau_{\text{ADC, DIO}} = \tau_{\text{tr–fb}} - \tau_{\text{proc}} = 80\,\text{ns} \pm 3\,\text{ns} \).

The delay \( \tau_{\text{ADC, DIO}} \) has several contributions which we did not determine individually. The pipelined architecture of the AD6645 ADC introduces a delay of four clock cycles \( (40\,\text{ns}) \) and a latency of one additional clock cycle \( (10\,\text{ns}) \) to transfer the digitized signal from the ADC to the FPGA where it is registered in a synchronous D–flip–flop. Further delays are expected to contribute to \( \tau_{\text{ADC, DIO}} \) due to the routing of the digital signal on the BenADDA-V4TM board as well as pad–to–flip–flop and flip–flop–to–pad delays on the FPGA (see App. D 1).

Appendix C: Implementation details of digital signal processing blocks

Here we specify implementation details of the blocks of the DSP circuit presented in Sec. III which are relevant for the processing latency.

1. Digital mixer

The cosine and sine signals, \( \cos(\omega_{\text{IF}}t_n) \) and \( -\sin(\omega_{\text{IF}}t_n) \), for digital mixing are typically generated either using a lookup table with precomputed values or by an iterative algorithm and then multiplied with two copies of the signal as shown in Fig. 6(a). While these methods work for arbitrary frequencies \( \omega_{\text{IF}} \), a simplified method exists for the special case when \( \omega_{\text{IF}} \) equals a quarter of the sampling rate, i.e. \( \omega_{\text{IF}}/(2\pi) = f_s/4 \) [70, 71].

In the \( f_s/4 \) case, the periodic sequences for the cosine and negative sine are simply \( (1, 0, -1, 0) \) and \( (0, -1, 0, 1) \) respectively [70]. Since multiplication with 0, 1 and \(-1\) is trivial, we replace the multipliers by counter-driven multiplexers (MUX) that periodically switch between four inputs as shown in Fig. 5(b). The 2-bit repeating counter (CNT) iterates through a sequence of four values \( (0, 1, 2, 3) \), jumping to the next value in every clock cycle and restarting from 0 after it has reached 3. The output of the counter is forwarded to the selection (sel) input of the multiplexers (MUX). The selection input of the multiplexers determine which of the four inputs \((\text{in}0, \text{in}1, \text{in}2, \text{in}3)\) of the multiplexers are forwarded to their output. The four inputs of the multiplexer for the real part \((\text{Re}[S_m])\) correspond to multiplying the signal with \( (1, 0, -1, 0) \) while the inputs of the multiplexer for the imaginary part \((\text{Im}[S_m])\) correspond to multiplication with \( (0, -1, 0, 1) \).

2. Moving average

In the following, we discuss how to implement the moving average (circuit shown in Fig. 6(c)), which is the simplest type of FIR filter, with a processing latency of less than one clock cycle \( (10\,\text{ns}) \). The moving average is applied in parallel to the real and imaginary parts of the output \( S_m \) of the mixer, i.e. two copies of the circuit shown in Fig. 6(c) are implemented with outputs \( I \) and \( Q \) respectively.

The first step in the circuit for computing the moving average, as shown in Fig. 6(c), is to fan out the input signal into two branches. One branch \( b \) is delayed by a variable delay \( (z^{-i}) \) of \( l \) clock cycles while no operation is performed on the other branch \( a \), i.e. \( b_m = a_{m-l} \). A subtractor then computes the difference \( a - b \) between the values of the two branches which is forwarded to an accumulator \([+= \text{ in Fig. 6(c)}]\). In every clock cycle, the accumulator adds the value at its input to the sum stored internally and forwards the updated sum to the output. Therefore the output \( S_{\text{acc}i,n} \) at clock cycle \( n \) of the accumulator is the sum of all input samples up to clock cycle.
\[ n - 1, \text{i.e.} \]
\[
S_{\text{accu},n} = \sum_{m=0}^{n-1} (a_m - b_m) = \sum_{m=0}^{n-1} (a_m - a_{m-l}) \\
= \sum_{m=l}^{n-1} a_m + \sum_{m=0}^{n-l-1} (a_m - a_m) - \sum_{m=l}^{n-1} a_m \\
= \sum_{m=n-l}^{n-1} a_m, \quad (C1) \]

where the last equality holds assuming that all input samples with negative index are equal to zero, i.e. \( a_m = 0 \) for \( m < 0 \). To make sure that this assumption holds true, we initialize the registers of the variable delay and the accumulator to zero. As depicted in Fig. 6(c), an additional adder (+) adds the most recent value of the difference \( a_n - b_n \) at the input of the accumulator to its output and a constant factor of \( 1/l \) normalizes the moving average. Thus, the final signal at the output of the moving average module \( S_{\text{MA}} \) is
\[
S_{\text{MA},n} = \frac{1}{l} (S_{\text{accu},n} + a_n - b_n) \\
= \frac{1}{l} \left( \sum_{m=n-l}^{n-1} a_m + a_n - a_{n-l} \right) \\
= \frac{1}{l} \sum_{m=n-l}^{n} a_m, \quad (C2) \]

As opposed to the sums in Eq. \((C1)\), which stop at index \( n - 1 \), the final sum in Eq. \((C2)\) includes the most recent sample with index \( n \), which shows that the additional adder reduces the effective processing latency to less than one clock cycle.

3. Preprocessing module

Offset subtraction \((-c \text{ in Fig. } 6\text{d})\) is implemented with lookup tables (LUTs). The parameter \( c \) is configurable via the interface with the host computer (indicated by dashed arrows). The multiplication \((\times n \text{ in Fig. } 6\text{d})\) is implemented without the use of actual multipliers but rather uses bit shift operations, which are effective multiplications with powers of two. Avoiding the allocation of multipliers reduces hardware resource consumption and leads to reduced processing latencies. The multiplication is made configurable using multiplexers to choose...
between different bit shift operations. The bit shift operation is chosen via the host computer interface.

4. State discrimination module

The state discrimination module determines the qubit state and provides feedback triggers based on the sign bits $x$ and $y$ of the preprocessed signals $\tilde{I}$ and $\tilde{Q}$ as shown in Fig. 6(e). The sign bits of $\tilde{I}$ and $\tilde{Q}$ are 0 if the respective signal is positive and 1 if it is negative, as depicted in Fig. 6(f). Due to the prior offset subtraction, determining the sign bits of $\tilde{I}$ and $\tilde{Q}$ is equivalent to comparing the $I$ and $Q$ signals each to an arbitrary threshold value. Two lookup tables (LUT) define the binary feedback with two independent bits $L_{xy}^{(1)}$ and $L_{xy}^{(2)}$ which are selected based on the two sign bits $x$ and $y$ as depicted in Fig. 6(g). The entries of the LUT can be set via the host computer interface [dashed arrows in Fig. 6(e)]. In the example shown in Fig. 6(h), the value of the feedback bit is 1 if and only if $x = 0$ corresponding to a non-negative value of the $I$ component of the signal.

The input trigger signal is used as a reference for the timing of the feedback triggers relative to the onset of the readout pulse. As shown in Fig. 6(e), the trigger first enters a rising edge detection block. The output of the rising edge detection block is 1 if and only if the input binary value of the trigger was 0 in the previous clock cycle and 1 in the present clock cycle. The output of the rising edge detection is delayed with a variable delay $\tau_{\text{d}}$ where $d$ is the number of clock cycles (each being 10 ns) corresponding to the readout time $\tau_{\text{RO}}$, i.e. $d \times 10 \text{ ns} = \tau_{\text{RO}}$. The parameter $d$ can be set via the host computer interface. The output of the variable delay, to which we refer as the fbTime marker, marks the specific time at which the feedback pulse is provided. To assert that the feedback triggers are issued at the correct time, the feedback triggers fb and fb2 are based on the AND operation of the output of the LUT and the fbTime marker.

5. Histogram module

The histogram module is important to assess the feedback performance and to calibrate the experimental setup. Here we explain how our multi-dimensional histogram module is implemented. The histogram module has different operational modes. We first introduce the circuit for recording two-dimensional histograms as shown in Fig. 7(a). The input signals $\tilde{I}$ and $\tilde{Q}$ are rounded to 7-bit fixed point numbers which means that the full range of ±1V is subdivided into $2^7 = 128$ bins. The 7-bit fixed-point representations of $\tilde{I}$ and $\tilde{Q}$ are concatenated into a 14-bit address of the histogram bin which stores the number of occurrences of the combination of values ($\tilde{I}, \tilde{Q}$). The “increase count” block manages the communication with the ZBT RAM in order to increase the stored count whenever the enable flag (en) is active. For feedback experiments, the enable flag is derived from the fbTime marker such that the histogram is updated when the feedback decision is made (see Sec. III).

In the correlation mode of our histogram module, a buffer [Fig. 7(b)] stores the value of $\tilde{I}$ at every reception of the fbTime marker. The buffered signal $\tilde{I}_1$ is combined with the most recent signal $\tilde{I}_2$ to record the probability to observe a specific combination $\tilde{I}_1$ and $\tilde{I}_2$ in two consecutive readouts (see Sec. IV). In addition a segment counter [seg cnt in Fig. 7(b)] allows to distinguish alternating experimental scenarios, such as when the feedback is enabled or disabled alternately in consecutive runs of the experiment. In the correlation mode, the value of $Q$ is in principle not needed but is an additional useful piece of information. In order to make use of the total amount of $2^{25}$ bits (4 MB) available space in the ZBT RAM, we reduce the Q dimension to 5 bits and concatenate the values ($\tilde{I}_2, \tilde{Q}, \tilde{I}_1, \text{seg}$) into a 21-bit address with a word size of 16 bits to store the counts as presented in Fig. 7(b).

In the time-resolved mode, a time counter [time cnt in
Fig. [7] is used to add time as an additional dimension of the histogram. The time counter starts upon the reception of the fbTime marker and the enable flag is held active for up to 16 clock cycles. As for the generation mode there is a segment counter [seg cut in Fig. [7]c] which allows to discern different consecutive scenarios such as when the qubit is prepared in the |g⟩ or |e⟩ state alternately. This makes the time-resolved mode useful for calibration tasks such as finding the optimal qubit read-out time by observing the separation of the distributions of I and Q values for the states |g⟩ [blue histogram in Fig. [7]d] and |e⟩ [red histogram in Fig. [7]d] over time.

Appendix D: FPGA timing and resource analysis

1. FPGA timing analysis

Using the Xilinx ISE® tool suite [100], we extracted information about the timing of the signal processing for our present implementation of the DSP circuit in the Virtex–4 (xc4vsx35-10ff668) and for future implementations on the Virtex–6 (xc6vlx240t-1ff1156) and Virtex–7 (xc7vx485t-2ffg1761c) FPGA. In App. D2 we present the corresponding FPGA resource allocations.

We define the pad–to–pad delay τp–p as the delay the digitized signal encounters in the path from the signal input pads of the FPGA to the feedback trigger output pad. For the full implementation on the Virtex–4 FPGA (“V–4 full” in Tab. II), the predicted pad–to–pad delay amounts to

\[
τ_{p–p} \equiv τ_{p–f} + τ_{proc} + τ_{f–p} = 1.5 \text{ ns} + 30 \text{ ns} + 3.8 \text{ ns} = 35.3 \text{ ns},
\]

where \( τ_{proc} = 30 \text{ ns} \) is the processing latency of three pipeline stages (see blue dashed lines in Fig. 2). Moreover, the pad–to–flip–flop delay \( τ_{p–f} = 1.5 \text{ ns} \) is the maximum delay from the ADC input pads to the D–flip–flops of the first pipelined register. Furthermore, the flip–flop–to–pad delay \( τ_{f–p} = 3.8 \text{ ns} \) is the maximum delay from the flip–flops of the last pipelined register to the output pad of the feedback trigger. The pad–to–flip–flop \( τ_{p–f} \) and flip–flop–to–pad \( τ_{f–p} \) delays are expected to contribute to the digital input and output delay \( τ_{ADC,DIO} \) (see Sec. IID).

A clock period analysis shows that the minimum clock period due to the timing of the signals between two pipelined registers amounts to \( T_{\text{min}} = 6.7 \text{ ns} \) which corresponds to a maximum clock frequency of \( f_{\text{max}} = 149 \text{ MHz} \). Increasing the clock frequency in a pipelined architecture is however only beneficial when the sampling rate of the ADC is also increased. Instead, removing pipeline stages in the signal path can lead to a further decrease in processing time as long as the minimal clock period is larger than the sampling period, i.e. \( T_{\text{min}} \geq 1/f_s \).

As a first step towards a future optimization of the processing and pad–to–pad delay, we separately simulated the implementation of what we consider the core feedback functionality of the DSP circuit which only includes the \( f_s/4 \) mixer, the moving average, the offset subtraction and scaling modules and the state discrimination module. For the implementation of the core DSP circuit we keep only two pipelined registers, one at the ADC input and one at the feedback outputs fb and fb2. Therefore the processing latency amounts to one clock cycle. In order to optimize the pad-to-pad delay, we optimized first the register-to-register delay, which determines the maximal clock frequency. Afterwards, we optimize the pad-to-register and register-to-pad delays. Assuming that the sampling rate is equal to the maximal clock frequency, we obtain a pad–to–pad delay of 4 ns + 9.7 ns + 7 ns = 20.7 ns [c.f. Eq. (D1)] for the Virtex–4 implementation, 3 ns + 6.2 ns + 5.5 ns = 14.7 ns [c.f. Eq. (D1)] for the Virtex–6 implementation (“V–6 core” in Tab. II), and 4 ns + 5.3 ns + 5 ns = 14.3 ns for the Virtex–7 implementation (“V–7 core” in Tab. II). These results show that a further reduction of the latency introduced by the DSP from 35.3 ns to 14.3 ns is possible by an optimized implementation of the core functionalities and using a recent FPGA. We therefore consider the integration of a recent FPGA into our experimental setup as possible future work.

2. FPGA resource analysis

Here we report the FPGA resource allocation for the full design implemented on the Virtex–4 FPGA and compare it to the resources needed to implement the core functionality consisting of the \( f_s/4 \) mixer (App. C1), the moving average (App. C2), the preprocessing module (App. C3) and the state discrimination module (App. C4). The analysis of the resource allocation is done for the implementation of the core design on the Virtex–4, Virtex–6 and Virtex–7 FPGAs corresponding to the timing analysis performed in App. D1.

The resource usage is summarized in Tab. III. The full design (V–4 full) uses \( n_{DFF} = 15’312 \) D–flip–flops corresponding to 49% of the total number of D–flip–flops and \( n_{LUT} = 18’361 \) four-input lookup tables (LUT) which is 59% of the available LUTs on the Virtex–4 FPGA. The majority of resources in the full design is consumed

|                | \( τ_{p–p} \) | \( τ_{proc} \) | \( τ_{clk} \) | \( τ_{clk,min} \) | \( f_{max} \) |
|----------------|-------------|-------------|-------------|----------------|-------------|
| V–4 full       | 35.3 ns     | 30 ns       | 10 ns       | 6.7 MHz        |             |
| V–4 core       | 20.7 ns     | 10 ns       | 9.7 ns      | 9.7 MHz        | 103 MHz     |
| V–6 core       | 14.7 ns     | 6.2 ns      | 6.2 ns      | 6.2 MHz        | 161 MHz     |
| V–7 core       | 14.3 ns     | 5.3 ns      | 5.3 ns      | 5.3 MHz        | 188 MHz     |

TABLE II. Summary of the simulated FPGA timings: the pad–to–pad delay \( τ_{p–p} \) from the data input to the feedback trigger output, the processing time \( τ_{proc} \), the chosen clock period \( τ_{clk} \), the minimum clock period \( τ_{clk,min} \) and the maximum clock frequency \( f_{max} \). See main text for details.
TABLE III. FPGA resource summary specifying the allocated number of D–flip–flops \(n_{\text{DFF}}\), number of LUTs \(n_{\text{LUT}}\) and number of dedicated DSP slice resources \(n_{\text{DSP}}\). Percentages are relative (rel.) to the total amount of resources on the corresponding FPGA. See main text for details.

|           | \(n_{\text{DFF}}\) (rel.) | \(n_{\text{LUT}}\) (rel.) | \(n_{\text{DSP}}\) (rel.) |
|-----------|-----------------------------|-----------------------------|-----------------------------|
| V–4 full  | 15’312 (49%)                 | 18’361 (59%)                | 184 (95%)                   |
| V–4 core  | 361 (1%)                     | 535 (2%)                    | 0                           |
| V–6 core  | 372 (< 1%)                   | 369 (< 1%)                  | 0                           |
| V–7 core  | 371 (< 1%)                   | 509 (< 1%)                  | 0                           |

by the flexibility in the signal processing, such as the phase-adjustable mixer and the FIR filter with arbitrary coefficients and the possibility to record histograms. In addition, the full design includes hardware modules for interfacing with PC and ZBT memory. To implement the added flexibility in the signal processing, the full design requires \(n_{\text{DSP}} = 184\) dedicated DSP slice resources, which contain multipliers and adders.

The core design (V–4 core, V–6 core and V–7 core in Tab. III) implements only a subset of the functionality to maintain the minimal requirements for the DSP operations. Therefore, the number of D–flip–flops \(n_{\text{DFF}}\) and LUTs \(n_{\text{LUT}}\) is reduced by almost two orders of magnitude compared to the full implementation. In addition, the implementation of the core functionality does not require dedicated DSP slices of the FPGA \(n_{\text{DSP}}\) in Tab. III since no multipliers are used in the blocks of the core design. The numbers \(n_{\text{DFF}}\) and \(n_{\text{LUT}}\) vary depending on whether the core design is implemented for the Virtex–4 (V–4 core), Virtex–6 (V–6 core) or Virtex–7 (V–7 core) FPGA, as displayed in Tab. III. We ascribe the variations of resource usage among the implementations of the core design to differences in the slice and LUT structure between the respective FPGA models. Different slice and LUT structures result in differences of the resource optimization in the mapping process using the Xilinx ISE software.

Appendix E: Experimental parameters

The superconducting transmon qubit [44] has a resonance frequency \(\omega_0/(2\pi) = 6.148\) GHz corresponding to the transition between the ground and first excited state and an anharmonicity of \(\alpha = -401\) MHz. The qubit is capacitively coupled to a \(\lambda/2\) coplanar waveguide resonator with a coupling strength \(g/(2\pi) \approx 65\) MHz. We measure a fundamental mode resonance frequency of \(\omega_r/(2\pi) = 7.133\) GHz defined as the center of the dispersively shifted resonance frequencies for the qubit states \(|g\rangle\) and \(|e\rangle\). We measure a linewidth of \(\kappa/(2\pi) = 6.3\) MHz of the resonator. The qubit shows an exponential energy relaxation with time constant \(T_1 \approx 1.4\) μs. We choose an experiment repetition period of 10 μs, which for the given \(T_1\), is sufficient to obtain a residual out-of-equilibrium excited state population of 0.1%. The measured thermal equilibrium excited state probability is \(P_{\text{therm}} \approx 7\%\) (see App. [F]).

The envelope of the microwave pulses for qubit rotations is Gaussian with \(\sigma = 7\) ns, truncated symmetrically at \(\pm 2\sigma\) as seen in the pulse scheme Fig. [4a] and uses the DRAG technique [72, 101] to avoid errors due to the presence of states outside the qubit subspace.

From pulsed spectroscopy we observe a dispersive shift of the resonator frequency \(\omega_r^{(g)}(\psi)\) for the qubit in the ground \(|g\rangle\) or excited state \(|e\rangle\) of

\[
2\chi \equiv \omega_r^{(e)} - \omega_r^{(g)} = -2.2\text{ MHz} \times 2\pi.
\]

We choose the frequency of the resonator drive pulses for dispersive readout at the center between the two shifted resonator frequencies, i.e. \(\omega_1 \equiv (\omega_r^{(e)} + \omega_r^{(g)})/2\). The amplitude of the readout pulse is chosen such that the expected steady-state mean phonon number is \(\langle n_{\text{readout}} \rangle \approx 10\), which we calibrated by measuring the ac Stark shift [102] of the qubit frequency when a continuous coherent drive is applied to the resonator.

Appendix F: Reduction of thermal excited state population

A possible application of active feedback initialization is to temporarily reduce the excited state population when the qubit is initialized in the thermal equilibrium with its environment [19, 82, 84]. In order to test the performance of our feedback loop for reducing the thermal excited state population, we omit the \(\pi/2\) pulse in the beginning of the protocol presented in Sec. [IV] such that the expected input state is a mixed state described by the density matrix

\[
\rho_{\text{therm}} \equiv (1 - P_{\text{therm}})|g\rangle\langle g| + P_{\text{therm}}|e\rangle\langle e|,
\]

where \(P_{\text{therm}}\) is the excited state population when the system is in thermal equilibrium with its environment.

As discussed in Sec. [IV] the qubit state is measured by two successive readout pulses M1 and M2. When the feedback actuator is disabled, the measured histogram of the in-phase component \(I_1\) of the signal during M1 (blue dots in Fig. [8a]) is almost identical to the histogram of the in-phase component \(I_2\) of the signal during M2 (orange dots in Fig. [8a]). From counting the values on the right side of the threshold [dashed line in Fig. [8a]], we obtain corresponding thermal excited state probabilities of \(P[I_1]|_{\text{off}} = 8.21(2)\%\) for the first measurement M1 and \(P[E_2]|_{\text{off}} = 8.18(2)\%\) for the second measurement M2. This indicates that, without conditioning on the measurement outcome, the measurement leaves the thermal steady state \(\rho_{\text{therm}}\) unperturbed. Note that the overlap readout error (see App. [C]) biases the measured probabilities towards 50%. Taking this bias into account, we infer a thermal excited state population
of $P_{\text{therm}} \approx 7\%$ from the measured probabilities $\mathbb{P}[E_1]$ and $\mathbb{P}[E_2]$. The inferred thermal excited state population $P_{\text{therm}}$ corresponds to a temperature of a bosonic environment of $T_{\text{env}} \approx 114$ mK. The effective temperature $T_{\text{env}}$ is close to the measured base temperature of the dilution refrigerator which for the presented experiment was 90 mK instead of the typical temperature of 20 mK due to problems with the cryogenic setup.

When feedback is enabled, the excited state probability in the second measurement amounts to $\mathbb{P}[E_2]_{\text{fb on}} = 5.43(2)\%$, as obtained from the histogram of $I_2$ [orange dots in Fig. 8(b)], is reduced compared to the excited state probability in the first measurement $\mathbb{P}[E_1]_{\text{fb on}} = 8.29(2)\%$ obtained from the histogram of $I_1$ [blue dots in Fig. 8(b)], showing that a reduction of the thermal excited state population is possible with our feedback loop. The measured probability $\mathbb{P}[E_2]_{\text{fb on}}$ is in reasonably good agreement with the simulated value of $\mathbb{P}[E_2]_{\text{fb on, sim}} = 5.24\%$ obtained from a master equation simulation using the same model and parameters as discussed in Sec. IV.

We recorded two-dimensional histograms of the values $I_1$ and $I_2$ for the case when feedback is disabled and enabled as shown in Fig. 8(c) and Fig. 8(d) respectively. The measured relative counts in the four regions ($R_{GG}$, $R_{GE}$, $R_{EG}$, $R_{EE}$) of the two-dimensional histograms show the swapping of the probabilities $\mathbb{P}[R_{EG}]$ and $\mathbb{P}[R_{EE}]$ and the invariance of the probabilities $\mathbb{P}[R_{GG}]$ and $\mathbb{P}[R_{GE}]$ under the feedback action as discussed in Sec. IV. The histogram of the signal in the region $R_{EE}$ for the "feedback off" case [Fig. 8(c)] matches well with the histogram in region $R_{EE}$ for the "feedback on" case [Fig. 8(d)]. In particular, the corresponding probabilities $\mathbb{P}[R_{EE}]_{\text{fb off}} = 2.66(1)\%$ and $\mathbb{P}[R_{EE}]_{\text{fb on}} = 2.79(2)\%$ match reasonably well, which shows that the feedback pulse is applied when the state $|e\rangle$ is detected in the first measurement. The experimentally observed probabilities are in reasonably good agreement with the simulation results $\mathbb{P}[R_{EE}]_{\text{fb off, sim}} = 1.98\%$ and $\mathbb{P}[R_{EE}]_{\text{fb on, sim}} = 1.45\%$ (Tab. IV) considering the sources of systematic errors as discussed in Sec. IV. We observe that the histogram in the region $R_{EE}$ in Fig. 8(d) is double-peaked, which is a consequence of the readout error since the tail of the distribution associated with the $|g\rangle$ state extends into the region $R_{EE}$.

Furthermore, the histograms in the region $R_{GE}$ match for both the "feedback off" [Fig. 8(c)] and the "feedback on" case [Fig. 8(d)]. The probabilities of $\mathbb{P}[R_{GE}]_{\text{fb off}} = 2.63(1)\%$ and $\mathbb{P}[R_{GE}]_{\text{fb on}} = 2.64(1)\%$ agree within the statistical error bars, which shows that the feedback pulse is not applied when the state $|g\rangle$ is detected in the first measurement.

The data presented here serves as a further experimental benchmark of our implementation of the feedback scheme and illustrates the use of two-dimensional

![Figure 8](image-url)

**FIG. 8.** Same type of histograms as presented in Fig. 4 for the scenario when the initial $\pi/2$ pulse is omitted. (a) Histograms of the in-phase signal in the first measurement $I_1$ (blue dots) and second measurement $I_2$ (orange dots) when feedback is disabled. The dashed line marks the feedback threshold. Percentages are the summed counts of occurrences above the threshold relative to the total count $C_{\text{tot}} = 2\,997\,152$ for the signal in M1 and M2 respectively. (b) The same type of histograms as in (a) but with feedback enabled. (c) Two-dimensional histogram with $128 \times 128$ bins as a function of the in-phase signal in the first measurement $I_1$ versus the in-phase signal in the second measurement $I_2$ with feedback disabled. Red lines are contour lines marking specific counts of $(0.05, 1, 2, 4) \times 10^3$. In each region ($R_{GG}$, $R_{GE}$, $R_{EG}$, $R_{EE}$) separated by the threshold (dashed lines) the percentage of counts relative to the total count is indicated. (d) The same type of two-dimensional histogram as in (c) but with feedback enabled.

| Feedback Off | Feedback On |
|--------------|-------------|
| exp. sim. | exp. sim. |
| $\mathbb{P}[R_{GG}]$ | 89.16(2)\% | 88.01\% | 89.07(2)\% | 88.01\% |
| $\mathbb{P}[R_{GE}]$ | 2.63(1)\% | 3.79\% | 2.64(1)\% | 3.79\% |
| $\mathbb{P}[R_{EG}]$ | 2.66(1)\% | 1.98\% | 5.50(2)\% | 6.75\% |
| $\mathbb{P}[R_{EE}]$ | 5.55(2)\% | 6.22\% | 2.79(1)\% | 1.45\% |

**TABLE IV.** Experimental (exp.) and simulated (sim.) probabilities $\mathbb{P}[R_{xy}]$ of the events to observe the qubit in state $x$ in the first measurement and in state $y$ in the second measurement when the feedback is either disabled or enabled. The simulated values are obtained from a master equation simulation. See text for details.
histograms to get insight into processes that lead to the observed excited state probabilities.

Appendix G: Readout fidelity

The readout is calibrated in a separate calibration step where either no pulse or a π pulse is applied to the qubit prior to the measurement. A threshold check, as described in Sec. [1], leads either to the result $G$ corresponding to the qubit state $|g\rangle$ or $E$ corresponding to $|e\rangle$. The single-shot readout fidelity is defined as

$$ F_r = 1 - |\mathbb{P}[E|\text{“no pulse”}] - \mathbb{P}[G|\text{“π pulse”}]|, $$

where $\mathbb{P}[E|\text{“no pulse”}]$ represents the conditional probability of obtaining the result $E$ when no pulse has been applied whereas $\mathbb{P}[G|\text{“π pulse”}]$ represents the conditional probability of obtaining the result $G$ when a π pulse has been issued. For a fixed moving average window length of $l = 4$ digital samples (40 ns), the single-shot fidelity reaches a maximal value of $F_r = 77\%$ at time $\tau_{RO} \approx 105 \text{ ns}$ relative to the onset of the readout pulse. We expect the contributions to the readout infidelity to be

$$ 1 - F_r \approx 2P_{\text{therm}} + P_{\text{decay}} + P_{\text{overlap}}, \quad (G2) $$

where $P_{\text{therm}} \approx 7\%$ is the initial excited state population in thermal equilibrium (see App. [1]), $P_{\text{decay}} \approx 1 - \exp(-\tau_{RO}/T1) \approx 6\%$ is the error due to the decay of the $|e\rangle$ state and $P_{\text{overlap}} \approx 3\%$ is the probability of misidentification of the qubit state due to overlap of the probability density functions for the signals corresponding to the $|g\rangle$ and $|e\rangle$ state. We extracted the contributions to the readout infidelity from fits to recorded histograms using methods similar to the ones described in Ref. [4].

[1] R. Blatt and D. Wineland, Nature 453, 1008 (2008)
[2] S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf, Phys. Scr. T137, 014012 (2009)
[3] D. Kielpinski, C. Monroe, and D. Wineland, Nature 417, 709 (2002)
[4] M. Devoret and R. J. Schoelkopf, Science 339, 1169 (2013)
[5] J. Kelly, R. Barends, A. G. Fowler, A. Megrant, E. Jeffrey, T. C. White, D. Sank, J. Y. Mutus, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, I.-C. Hou, C. Neill, P. J. J. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, Nature 519, 66 (2015)
[6] B. Lekitsch, S. Weid())
[7] M. A. Nielsen and I. L. Chuang, Quantum Computation and Quantum Information (Cambridge University Press, 2000).
[8] J. Preskill, in 25th Solvay Conference on Physics, October 19-22, 2011, Brussels, Belgium. (2012).
[9] T. F. Ronnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, Science 345, 420 (2014)
[10] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, arXiv:1608.00263v3 (2017)
[11] S. Jordan, “Quantum algorithm zoo,” http://math.nist.gov/quantum/zoo/ (2017), accessed: 2017-05-17.
[12] D. J. Reilly, npj Quantum Information 1, 15011 (2015)
[13] H. Wiseman and G. Milburn, Quantum Measurement and Control (Cambridge University Press, 2010).
[14] J. Zhang, Y.-X. Liu, R.-B. Wu, K. Jacobs, and F. Nori, Physics Reports 679, 1 (2017)
[15] W. P. Smith, J. E. Reiner, L. A. Orozco, S. Kuhr, and H. M. Wiseman, Phys. Rev. Lett. 89, 136301 (2002)
[16] J. E. Reiner, W. P. Smith, L. A. Orozco, H. M. Wiseman, and J. Gambetta, Phys. Rev. A 70, 023819 (2004)
[17] C. Sayrin, I. Dotsenko, X. Zhou, B. Paeudercer, T. Rybarczyk, S. Gleyzes, P. Rouchon, M. Mirrahimi, H. Amini, M. Brune, J.-M. Raimond, and S. Haroche, Nature 477, 73 (2011)
[18] M. A. Armen, J. K. Au, J. K. Stockton, A. C. Doherty, and H. Mabuchi, Phys. Rev. Lett. 89, 136302 (2002)
[19] D. Ristè, C. B. Tilton, K. W. Lehnert, and L. DiCarlo, Phys. Rev. Lett. 109, 240502 (2012)
[20] R. Vijay, C. Macklin, D. H. Slichter, S. J. Weber, K. W. Murch, R. Naik, A. N. Korotkov, and I. Siddiqi, Nature 490, 77 (2012)
[21] P. Campagne-Ibarcq, E. Flurin, N. Roch, D. Darson, P. Morin, M. Mirrahimi, M. H. Devoret, F. Mallet, and B. Huard, Phys. Rev. X 3, 021008 (2013)
[22] D. Ristè, M. Dukalski, C. A. Watson, G. de Lange, M. J. Tiggelman, Y. M. Blanter, K. W. Lehnert, R. N. Schouten, and L. DiCarlo, Nature 502, 350 (2013)
[23] Y. Liu, S. Shankar, N. Ofek, M. Hatridge, A. Narla, K. M. Siwa, L. Frunzio, R. J. Schoelkopf, and M. H. Devoret, Phys. Rev. X 6, 011022 (2016)
[24] G. de Lange, D. Riste, M. J. Tiggelman, C. Eichler, L. Tornberg, G. Johansson, A. Wallraff, R. N. Schouten, and L. DiCarlo, Phys. Rev. Lett. 112, 080501 (2014)
[25] P. Campagne-Ibarcq, S. Jezouin, N. Cottet, P. Six, B. Huard, Phys. Rev. Lett. 117, 060502 (2016)
[26] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, Phys. Rev. Lett. 70, 1895 (1993)
[27] D. Bouwmeester, J. W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger, Nature 390, 575 (1997)
[28] T. B. Pittman, B. C. Jacobs, and J. D. Franson, Phys. Rev. A 66, 052305 (2002)
[29] S. Giacomini, F. Sciarrino, E. Lombardi, and F. De Martini, Phys. Rev. A 66, 030302 (2002)
[30] X.-S. Ma, T. Herbst, T. Scheidl, D. Wang, S. Kropatschek, W. Naylor, B. Wittmann, A. Mech,
It is in principle possible to shorten the 160 ns duration of the readout pulse to match the end of the integration window at $t = 105$ ns but we keep the length of the readout pulse constant to simplify the calibration procedure.

H. J. Carmichael, *Statistical Methods in Quantum Optics 1: Master Equations and Fokker-Planck Equations*, 2nd ed. (Springer, Berlin, 2002).

D. H. Slichter, R. Vijay, S. J. Weber, S. Boutin, M. Boissonneault, J. M. Gambetta, A. Blais, and I. Siddiqi, Phys. Rev. Lett. 109, 153601 (2012)