Restless Tuneup of High-Fidelity Qubit Gates

M. A. Rol,1,2 C. C. Bultink,1,2 T. E. O’Brien,3 S. R. de Jong,1,2 L. S. Theis,4 X. Fu,1 F. Luthi,1,2 R. F. L. Vermeulen,1,2 J. C. de Sterke,5,6 A. Bruno,1,2 D. Deurloo,6,1 R. N. Schouten,1,2 F. K. Wilhelm,7 and L. DiCarlo1,2

1QuTech, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands
2Kavli Institute of Nanoscience, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands
3Instituut-Lorentz for Theoretical Physics, Leiden University, P.O. Box 9506, 2300 RA Leiden, The Netherlands
4Theoretical Physics, Saarland University, 66123 Saarbrucken, Germany
5Topic Embedded Systems B.V., P.O. Box 440, 5680 AK Best, The Netherlands
6Netherlands Organisation for Applied Scientific Research (TNO), P.O. Box 155, 2600 AD Delft, The Netherlands
7Center for Quantum Devices, Kavli Institute of Nanoscience, Delft University of Technology, Netherlands

(Received 22 November 2016; revised manuscript received 18 February 2017; published 24 April 2017)

We present a tuneup protocol for qubit gates with tenfold speedup over traditional methods reliant on qubit initialization by energy relaxation. This speedup is achieved by constructing a cost function for Nelder-Mead optimization from real-time correlation of nondemolition measurements interleaving gate operations without pause. Applying the protocol on a transmon qubit achieves 0.999 average Clifford fidelity in one minute, as independently verified using randomized benchmarking and gate-set tomography. The adjustable sensitivity of the cost function allows the detection of fractional changes in the gate error with a nearly constant signal-to-noise ratio. The restless concept demonstrated can be readily extended to the tuneup of two-qubit gates and measurement operations.

DOI: 10.1103/PhysRevApplied.7.041001

I. INTRODUCTION

Reliable quantum computing requires the building blocks of algorithms, quantum gates, to be executed with low error. Strategies aiming at quantum supremacy without error correction [1,2] require \( \sim 10^3 \) gates, and thus gate errors \( \sim 10^{-3} \). Concurrently, a convincing demonstration of quantum fault tolerance using the 17- and 49-qubit surface-code encoding [3,4] under development by several groups worldwide requires gate errors one order of magnitude below the \( \sim 10^{-2} \) threshold of surface code [5,6].

The quality of qubit gates depends on qubit coherence times and the accuracy and precision of the pulses realizing them. With the exception of a few systems known with metrological precision [7], pulsing requires meticulous calibration by closed-loop tuning, i.e., pulse adjustment based on experimental observations. Numerical optimization algorithms have been implemented to solve a wide range of tuning problems with a cost-effective number of iterations [8–13]. However, relatively little attention has been given to quantitatively exploring the speed and robustness of the algorithms used. This becomes crucial with more complex and precise quantum operations, as the number of parameters and requisite precision of calibration grow.

Though many aspects of tuning qubit gates are implementation independent, some details are specific to physical realizations. Superconducting transmon qubits are a promising hardware for quantum computing, with gate times already exceeding coherence times by 3 orders of magnitude. Conventional gate tuneup relies on qubit initialization, performed passively by waiting several times the qubit energy-relaxation time \( T_1 \), or actively through feedback-based reset [14]. Passive initialization becomes increasingly inefficient as \( T_1 \) steadily increases [15,16], while a feedback-based reset is technically involved [17].

In this Letter, we present a gate-tuneup method that dispenses with \( T_1 \) initialization and achieves tenfold speedup over the state of the art [9] without active reset. Restless tuneup exploits the real-time correlation of quantum-nondemolition (QND) measurements to interleave gate operations without pause, and the evaluation of a cost function for numerical optimization with adjustable sensitivity at all levels of gate fidelity. This cost function is obtained from a simple modification of the gate sequences of conventional randomized benchmarking (CRB) to penalize both gate errors within the qubit subspace and any leakage from it. We quantitatively match the signal-to-noise ratio of this cost function with a model that includes measured \( T_1 \) fluctuations. Restless tuneup robustly achieves \( T_1 \)-dominated gate fidelity of 0.999, verified using both CRB with \( T_1 \) initialization and a first implementation of gate-set tomography (GST) [18] in a superconducting qubit. While this performance matches that of conventional tuneup, restless tuneup is tenfold faster and converges in one minute.

II. RESTLESS CONCEPT AND SPEEDUP

In many tuneup routines [Fig. 1(a)], the relevant information from the measurements can be expressed as the fraction \( \varepsilon \) of nonideal outcomes (\( m_n \)). In conventional gate
tuneup, a qubit is repeatedly initialized in the ground state $|0\rangle$, driven by a set of gates $\{G\}$ whose net operation is ideally identity, and measured [Fig. 1(b)]. The conventional cost function is the raw infidelity,

$$
\epsilon_C = \frac{1}{N} \sum_{n=1}^{N} (m_n \neq 0).
$$

The central idea of restless tuning [Fig. 1(c)] is to remove the time-costly initialization step, by measuring the correlation between subsequent QND measurements and interleaving gate operations without any rest [19]. For example, when the net ideal gate operation is a bit flip, we can define the error fraction

$$
\epsilon_R = \frac{1}{N} \sum_{n=2}^{N} (m_n = m_{n-1}). \quad (1)
$$

We demonstrate the restless tuneup of derivative-removal-by-adiabatic-gate (DRAG) pulses [20] on the transmon qubit recently reported in Ref. [12] (a summary of device parameters is in Ref. [21]). We choose DRAG pulses (duration $\tau_p = 20$ ns) for their proven ability to reduce gate error and leakage [26,27] with few-parameter analytic pulse shapes. These pulses consist of Gaussian ($G$) and derivative of Gaussian ($D$) envelopes of the in- and quadrature-phase components of a microwave drive at the transition frequency $f$ between qubit levels $|0\rangle$ and $|1\rangle$. These components are generated using four channels of an arbitrary waveform generator (AWG), frequency up-conversion by sideband modulation of one microwave source, and two in-phase–quadrature ($I$–$Q$) mixers. The $G$ and $D$ components are combined inside a vector switch matrix (VSM) [28] (details in Ref. [21]). A key advantage of this scheme using four channels is the ability to independently set the $G$ and $D$ amplitudes ($A_G$ and $A_D$, respectively), without uploading new waveforms to the AWG.

To measure the speedup obtained from the restless method, we must take the complete iteration into account. The traditional iteration of a tuneup routine involves the following: (1) setting parameters (four channel amplitudes on a Tektronix 5014 AWG); (2) acquiring $N = 8000$ measurement outcomes; (3) sending the measurement outcomes to the computer and processing them; and (4) miscellaneous overhead that includes determining the parameters for the next iteration, as well as saving and plotting data. In Fig. 1(d), we visualize these costs for an example optimization experiment. We intentionally penalize the restless method by choosing a large number of gates ($\sim 550$). Even in these conditions, restless sequences reduce the acquisition time from 1.60 to 0.12 s. However, the improvement in total time per iteration (from 1.98 to 0.50 s) is modest due to 0.38 s of overhead.

We take two steps to reduce overhead. The 0.23 s required to send all measurement outcomes to the computer and then calculate the error fraction is reduced to $< 1$ ms by calculating the fraction in real time, using the same field-programmable gate-array system that digitizes and processes the raw measurement signals into bit outcomes. The 0.09 s required to set the four channel amplitudes in the AWG is reduced to 1 ms by setting $A_G$ and $A_D$ in the VSM. With these two technical improvements, the remaining overhead is dominated by the miscellaneous contributions (40 ms). This reduces the total time per restless (conventional) iteration to 0.16 s (1.64 s).

III. RESTLESS RANDOMIZED BENCHMARKING AS COST FUNCTION

A quantity of common interest in gate tuneup is the average Clifford fidelity $F_{Cl}$, which is typically measured using CRB. $\{G\}$ consists of sequences of $N_{Cl}$ random Clifford gates, including a final recovery Clifford gate that makes the ideal net operation identity. Following [29], we compose the 24 single-qubit Clifford gates from the set of $\pi$ and $\pm\pi/2$ rotations around the $x$ and $y$ axes, which requires an average of 1.875 gates per Clifford. Gate errors make $\epsilon_C$ increase with $N_{Cl}$ as [30,31].
Here, $A$ and $B$ are constants determined by state-preparation-and-measurement (SPAM) error, and $1 - F_C$ is the standard deviation of $\epsilon_C$ for different $N_C$ and fitting Eq. (2). However, for tuning it is sufficient to optimize $\epsilon_C$ at one choice of $N_C$, because $\epsilon_C(N_C)$ decreases monotonically with $F_C$ [9].

In the presence of leakage, CRB sequences and $\epsilon_C$ are not ideally suited for restless tuneup. Typically, there is significant overlap in the readout signals from the first (|1⟩) and second (|2⟩) excited state of a transmon. A transmon in |2⟩ can produce a string of identical measurement outcomes until it relaxes back to the qubit subspace. If the ideal net operation is identity, the measurement outcomes can be indistinguishable from ideal behavior. Although the leakage on single-qubit gates is typically small ($10^{-5}$–$10^{-3}$ per Clifford for the range of $A_D$ considered [27,28]), a simple modification to the sequence allows penalizing leakage. By choosing the recovery Clifford for restless randomized benchmarking (RRB) sequences so that the ideal net operation of $\{G\}$ is a bit flip, leakage produces an error. This simple modification makes $\epsilon_R$ a better cost function.

We now examine the suitability of the restless scheme for optimization (Fig. 2). Plots of the average $\epsilon_R(N_C)$ [$\epsilon_R(N_C)$] at various $F_C$ (controlled via $A_G$) behave similarly to $\epsilon_C$ in CRB. Furthermore, $\epsilon_R$ is minimized at the same $A_G$ as $\epsilon_C$, with only a shallower dip because of SPAM. The $(A_G, A_D)$ landscapes for both cost functions [Figs. 2(b) and 2(d)] are smooth around the optimum, making them suitable for numerical optimization. The fringes far from the optimum arise from the limited number of seeds (always 200) used to generate the randomized-benchmarking sequences. Note that while the landscapes are visually similar, the difference in time required to map them is striking: ~50 min for $\epsilon_C$ vs $\epsilon_R < 5$ min for $\epsilon_R$ at $N_C = 300$.

The sensitivity of $\epsilon_R$ to the tuning parameters depends on both the gate fidelity and $N_C$. This can be seen in the variations between curves in Fig. 2(a). In order to quantify this sensitivity, we define a signal-to-noise ratio. For signal we take the average change in the error fraction, $\Delta \bar{\epsilon}_R = \bar{\epsilon}_R(F_C^0) - \bar{\epsilon}_R(F_C^1)$, from $F_C^0$ to $F_C^1 \approx \frac{1}{2} + \frac{1}{2} F_C^0$ (halving the infidelity). For noise we take $\sigma_{\bar{\epsilon}_R}$, the average standard deviation of $\epsilon_R$ between $F_C^0$ and $F_C^1$. We find that the maximal signal-to-noise ratio remains ~15 for an optimal choice of $N_C$ that increases with $F_C^0$ (Fig. 3 and details in Ref. [21]). This allows tuning in logarithmic time, since reducing error rates $p \to p/2^{M}$ requires only $M$ optimization steps.

A simple model describes the measurement outcomes as independent and binomially distributed with error probability $\epsilon_R$, as per Eq. (2) with $\epsilon_C \to \epsilon_R$. This model captures all the essential features of the signal. However, it only quantitatively matches the noise at high $N_C$. Experiment shows an increase in noise at low $N_C$. In this range, $\epsilon_R$ is dominated by SPAM, which is primarily due to $T_1$. We surmise that the increase stems from $T_1$ fluctuations [32] during the acquisition of statistics in these RRB experiments. To test this.

![FIG. 2. (a) Average error fraction of RRB for different $F_C$ vs $N_C$. (b) $\epsilon_C$ and $\epsilon_R$ as a function of $A_G$ for $N_C = 80$ and $N_C = 300$. The curves are denoted by a dashed line in (c),(d), (c), (d) $\epsilon$ for $N_C = 300$ as a function of $A_G$ and $A_D$. White circles indicate minimal $\epsilon$. The total acquisition time is shown at the bottom right.](image)

![FIG. 3. (a) Signal $\Delta \bar{\epsilon}_R$ for a halving of the gate infidelity, plotted as a function $N_C$ at $F_C^1 \sim 0.989$ (red), 0.996 (green), and 0.998 (blue). (b) Noise dependence on $N_C$ at the same fidelity levels. Added curves are obtained from the two models described in the main text.](image)
hypothesis, we develop an extensive model incorporating $T_1$ fluctuations into the calculation of both signal and noise [21]. We find good agreement with experimental results using independently measured values of $T_1$ and $\sigma_{T_1}$. The good agreement confirms the nondemolition character of the measurement previously reported in Ref. [12].

IV. PERFORMANCE AS A TUNEUP PROTOCOL

Following its validation, we now employ $\varepsilon_R$ in a two-step numerical optimization protocol (Fig. 4). We choose the Nelder-Mead algorithm [33] as it is derivative free and easy to use, requiring only the specification of a starting point and initial step sizes. The first step using $\varepsilon_R(N_{C1} = 80)$ ensures convergence even when starting relatively far from the optimum, while the second step using $\varepsilon_R(N_{C1} = 300)$ fine tunes the result. We test the optimization for four realistic starting deviations from the optimal parameters $(A_D^{\text{opt}}, A_G^{\text{opt}})$. $A_G$ is chosen at both approximately 6% above and below $A_G^{\text{opt}}$, selected as a worst-case estimate from a Rabi oscillation experiment. $A_D$ is chosen at both approximately half and double $A_D^{\text{opt}}$. The initial step sizes are $\Delta A_G \approx -0.03A_G^{\text{opt}}$, $\Delta A_D \approx -0.25A_D^{\text{opt}}$ for the first step, and $\Delta A_G \approx -0.01A_G^{\text{opt}}$, $\Delta A_D \approx -0.08A_D^{\text{opt}}$ for the second step.

We assess the accuracy of the above optimization and compare to traditional methods. A CRB experiment [Fig. 4(c)] following two-parameter restless optimization indicates $F_{\text{Cl}} = 0.9991$. This value matches the average achieved by both restless and conventional tuneups for the different starting conditions. We also implement GST to independently verify results obtained using CRB. From the process matrices we extract the average GST Clifford fidelity, $F_{\text{GST}}^{(T_1)} = 0.9990 \pm 0.0005$ (0.9990 ± 0.00003) for restless (conventional) tuneup [21], consistent with the value obtained from CRB.

The robustness of the optimization protocol is tested by interleaving tuneups with CRB and $T_1$ measurements over eleven hours (summarized in Table I, and detailed in Ref. [21]). Both tuneups reliably converge to $F_{\text{Cl}} = 0.9991$, close to the $T_1$ limit [34]:

$$ F_{\text{Cl}}^{(T_1)} \approx \frac{1}{6} (3 + 2e^{-\tau_c/2T_1} + e^{-\tau_c/T_1}) = 0.9994, $$ (3)

with $\tau_c = 1.875\tau_p$. However, restless tuneup converges in one minute, while conventional tuneup requires eleven.

It remains to test how restless tuneup behaves as additional parameters are introduced. Many realistic scenarios also require tuning the drive frequency $f$. As a worst case, we take an initial detuning of ±250 kHz. The initial step size in the first (second) step is 100 kHz (50 kHz). The three-parameter optimization converges to $F_{\text{Cl}} = 0.9990 \pm 0.0001$ for both restless and conventional tuneups. We attribute the slight decrease in $F_{\text{Cl}}$ achieved by three-parameter optimization to the observed reduction in average $T_1$.

V. SUMMARY

In summary, we develop an accurate and robust tuneup method achieving a tenfold speedup over the state of the art [9]. This speedup is achieved by avoiding qubit initialization by relaxation, and by using real-time correlation of measurement outcomes to build the cost function for numerical optimization. We apply the restless concept to
the tuneup of Clifford gates on a transmon qubit, reaching a $T_1$-dominated fidelity of 0.999 in one minute, verified by conventional randomized benchmarking and gate-set tomography. We show experimentally that the method can detect fractional reductions in gate error with nearly constant signal-to-noise ratio. An interesting next direction is to develop an algorithm that makes optimal use of this tunable sensitivity while maintaining the demonstrated robustness. The enhanced speed combined with the generic nature of the optimizer would also allow exploring other, more generic nonadiabatic gates without analytic pulse shapes, in a fashion analogous to optimal control theory [35,36]. Immediate next experiments will extend the restless concept to the tuneup of two-qubit controlled-phase gates [37,38] exploiting interactions with noncomputational states [39], in which leakage errors often dominate ($\sim 10^{-3}$). In this context, we anticipate that the RRB modification and the $e_R$ cost function will prove essential to reaching 0.999 fidelity. Finally, we also envision applying the restless concept to the simultaneous tuneup of single-qubit gates in the many-qubit setting (e.g., a logical qubit).

ACKNOWLEDGEMENTS

We thank R. Sagastizabal for experimental assistance, C. Dickel, J. Helsen, and S. Poletto for discussions, A. Johnson for support with Microsoft QCoDeS, and K. Ruddinger, E. Nielsen, and R. Blume-Kohout for support with GST/pyGSTi. 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 provided by the ERC Synergy Grant QC-lab, the China Scholarship Council (X. F.) and Microsoft Corporation by the ERC Synergy Grant QC-lab, the China Scholarship Council (X. F.) and Microsoft Corporation (IARPA), via the U.S. Army Research Office Grant No. W911NF-16-1-0071. Additional funding provided (IARPA), via the U.S. Army Research Office Grant No. W911NF-16-1-0071. Additional funding provided (IARPA), via the U.S. Army Research Office Grant No. W911NF-16-1-0071. Additional funding provided.

[18] Robin Blume-Kohout, John King Gamble, Erik Nielsen, Kenneth Rudinger, Jonathan Mizrahi, Kevin Fortier, and Peter Maunz, Demonstration of qubit operations below a

[19] Except $3.25 \mu s$ needed for passive depletion of photons left over from the 1-$\mu$s measurement [12].


[34] E. Magesan (private communication).


