Mitigating Leakage and Noise in Superconducting Quantum Computing

Battistel, F.

DOI
10.4233/uuid:07d93422-d07c-492e-8d65-592344e01936

Publication date
2022

Document Version
Final published version

Citation (APA)
Battistel, F. (2022). *Mitigating Leakage and Noise in Superconducting Quantum Computing*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:07d93422-d07c-492e-8d65-592344e01936

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

MITIGATING LEAKAGE AND NOISE IN SUPERCONDUCTING QUANTUM COMPUTING

MITIGATING LEAKAGE AND NOISE IN SUPERCONDUCTING QUANTUM COMPUTING

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 27 january 2022 om 15:00 uur

door

Francesco BATTISTEL

Master of Science in Theoretical and Mathematical Physics,
Ludwig-Maximilians Universität & Technische Universität München, Duitsland,
geboren te Pordenone, Italië.

Dit proefschrift is goedgekeurd door de promotoren

Prof.dr. B.M. Terhal
Prof.dr. L. DiCarlo

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr. B.M. Terhal,	Technische Universiteit Delft, *promotor*
Prof.dr. L. DiCarlo,	Technische Universiteit Delft, *promotor*
Onafhankelijke leden:	
Prof.dr. K.R. Brown	Duke University, Verenigde Staten
Prof.dr. S. Filipp	Technische Universität München / Walther-Meißner-Institute, Duitsland
Prof.dr. M. Walter	Universiteit van Amsterdam
Prof.dr. Y.M. Blanter	Technische Universiteit Delft
Prof.dr. L.M.K. Vandersypen,	Technische Universiteit Delft, reservelid

Keywords: superconducting qubits, leakage, quantum error correction, gates

Printed by: Gildeprint

ISBN 978-94-6384-285-3

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

CONTENTS

SUMMARY

Computers are used all over the place to perform tasks ranging from sending an email to running some complicated numerical simulation. That is brilliant of course, because computers enable us to solve a lot of problems in the world in this way. At the same time, for some of those problems, not even powerful supercomputers are enough to get the result of the computation in any reasonable amount of time. An alternative that might be able to solve some of these problems very quickly are quantum computers. The operations performed by a quantum computer need to be faithful in order to get the right result of the quantum computation. However, nowadays quantum computers are fairly noisy, severely limiting their range of applicability in the near future.

Various methods for quantum error correction have been developed, showing that, if error rates are below a certain threshold, one can make the computation as error-free as desired. However, while quantum error correction is starting to be tested in experiments, its performance has been mostly studied with respect to idealized error models. Furthermore, quantum error correction comes at the price of a substantial overhead in number of qubits and number of operations, especially if error rates are just barely below threshold. From a different perspective, error-mitigation techniques that do not need the full machinery of quantum correction have been put forward, fostering hope that noisy near-term devices might run useful applications even without quantum error correction. However, in either case the physical error rates of the fundamental operations are still high.

In this thesis we focus on achieving lower error rates for some of the fundamental operations in a quantum computer, specifically for superconducting qubits, and we demonstrate the beneficial impact of these results on quantum error correction in a realistic setting.

We develop error models that are physically motivated for superconducting qubits (reviewed in Chapter 2), based on the noise sources to which they are sensitive (reviewed in Chapter 3). The major elements of novelty in our models are the inclusion of leakage, quasi-static flux noise, and distortions of electronic signals.
In Chapter 6 we discuss a flux-pulsing technique for controlled-phase gates, named Net Zero. In the first part, we show that the characteristic zero-integral feature protects from long-timescale distortions, echoes out flux noise and uses leakage interference to mitigate leakage, leading to a fast, high-fidelity gate. In the second part, we introduce an updated version of Net Zero, called Sudden Net Zero, that maintains the same advantages and adds easiness of tuneup and straightforward conditional-phase tunability.

Diagnosing errors is crucial for correcting them and tuning up gates. In Chapter 7 we introduce Spectral Quantum Tomography, a tomographic method that can provide detailed information about errors in single- and two-qubit gates, in a way that is independent of state-preparation and measurement errors. In particular, we investigate the footprint of relaxation and dephasing, as well as leakage and non-Markovian noise.

Leakage outside of the qubit computational subspace is particularly damaging for quantum error correcting codes, in particular stabilizer codes (reviewed in Chapter 4). Leakage-reduction units (reviewed in Chapter 5) can bring a leaked qubit back to the computational subspace, thus restoring part of the loss in performance. Based on the error model developed for two-qubit gates, we study the effect of leakage in quantum error correction using realistic density-matrix simulations.
In Chapter 8 we use hidden Markov models to detect leakage in a transmon-qubit-based surface code and improve the logical fidelity by post-selection. The detection is based on recognizing patterns in the stabilizer measurements that can likely be attributed to leakage.
In Chapter 9 we introduce a hardware-efficient leakage-reduction scheme to directly remove leakage in a scalable way that does not require extra qubits or time, leading to a reduction of the logical error rate. In particular, we propose two separate leakage-reduction units tailored for data and ancilla qubits, respectively. For data qubits, we apply a microwave pulse that transfers leakage to its dedicated readout resonator, where it quickly decays into the environment. For ancilla qubits, we use a microwave pulse that maps the leaked state to a computational state.

These techniques for two-qubit gates, tomography and leakage mitigation contribute to reducing the error rates, benefiting quantum error correction as well as near-term devices. In the Conclusion we give an outlook on the potential challenges in superconducting quantum computing, including tunable couplers, real-time decoding and physical error rates in large devices.

SAMENVATTING

Computers worden voortdurend gebruikt om taken uit te voeren die variëren van het verzenden van een email tot het doen van een complexe numerieke simulatie. Dat is natuurlijk waardevol, want computers stellen ons in staat om op deze wijze veel problemen in de wereld op te lossen. Tegelijkertijd kunnen voor bepaalde problemen zelfs krachtige supercomputers niet het resultaat van een berekening verkrijgen in een aanvaardbaar tijdsbestek. Kwantumcomputers vormen een alternatief om sommigen van deze problemen erg snel op te lossen. The operaties die door een kwantumcomputer worden uitgevoerd moeten betrouwbaar zijn om het juiste resultaat van de berekening te verkrijgen. Hedendaagse kwantumcomputers zijn echter relatief gevoelig voor ruis, iets dat hun toepassingsgebied in de nabije toekomst sterk beperkt.

Er zijn verschillende methoden voor kwantumfoutcorrectie ontwikkeld. Deze laten zien dat wanneer de foutgraad beneden een bepaalde drempelwaarde ligt, de berekening zo foutvrij gemaakt kan worden als men wilt. Ondanks dat kwantumfoutcorrectie in een experimentele omgeving getest begint te worden, is hun prestatie tot op heden voornamelijk getest aan de hand van geïdealiseerde foutmodellen. Daarnaast gaat kwantumfoutcorrectie gepaard met een substantiële overhead in termen van het aantal qubits en het aantal operaties, voornamelijk wanneer de foutgraad amper beneden de drempelwaarde ligt. Als alternatief zijn er methodes voor foutmitigatie voorgedragen die niet het volledige kwantumfoutcorrectie mechanisme nodig hebben, wat hoop biedt dat kwantumapparaten op korte termijn zelfs waardevolle berekeningen kunnen uitvoeren zonder kwantumfoutcorrectie. In beide gevallen is de fysieke foutgraad van de fundamentele operaties echter nog steeds hoog.

In dit proefschrift focussen we ons op het behalen van lage foutgraden voor een aantal van de fundamentele operaties van een kwantumcomputer, in het specifiek voor supergeleidende qubits, en demonstreren we het bevorderlijke effect van deze resultaten op kwantumfoutcorrectie in een realistische omgeving.

We ontwikkelen fysisch gemotiveerde foutmodellen voor supergeleidende qubits (besproken in Hoofdstuk 2), gebaseerd op de bronnen van ruis waarvoor zij gevoelig zijn (besproken in Hoofdstuk 3). The belangrijkste nieuwe elementen in onze modellen zijn het includeren van leakage effecten, quasi-statische flux ruis, en vervormingen van elektronische signalen.

In Hoofdstuk 6 bespreken we een flux-puls methode voor controlled-phase gates, genaamd Net Zero. In het eerste deel laten we zien dat de karakteristieke nul-integraal eigenschap bescherming biedt tegen vervormingen op een lange tijdschaal, flux ruis uit echoot en leakage interferentie gebruikt om leakage te verminderen, wat leidt tot een snelle en betrouwbare gate. In het tweede deel introduceren we een bijgewerkte versie van Net Zero, genaamd Sudden Net Zero. Deze behoudt dezelfde voordelen, en voegt ook vergemakkelijking van de afstelling en rechtstreekse conditional-phase afstembaarheid toe.

Het diagnostiseren van fouten is cruciaal voor het herstellen ervan en het afstellen van gates. In Hoofdstuk 7 introduceren we Spectrale Kwantum Tomografie; een tomografische methode die gedetailleerde informatie geeft over fouten in enkele- en twee-qubit gates op een wijze die onafhankelijk is van toestandsbereiding en meetfouten. Specifiek onderzoeken we het spoor van relaxatie en defasering, en van leakage en niet-Markoviaanse ruis.

Leakage buiten de qubit computationele subruimte is bijzonder schadelijk voor kwantumfoutcorrectie codes, in het bijzonder voor stabilisator codes (besproken in Hoofdstuk 4). Leakage-verminderingseenheden (besproken in Hoofdstuk 5) zijn in staat een door leakage getergde qubit terug te brengen naar de computationele subruimte, en daarmee een deel van de prestatieafname te compenseren. We bestuderen het effect van leakage in kwantumfoutcorrectie aan de hand van het foutmodel ontwikkeld voor twee-qubit gates, gebruikmakend van realistische dichtheidsmatrix simulaties.

In Hoofdstuk 8 gebruiken we verborgen Markov modellen om leakage te detecteren in een surface code (gebaseerd op transmon qubits) en verbeteren we de logische betrouwbaarheid aan de hand van naselectie. De detectie is gebaseerd op het herkennen van patronen in de stabilisator metingen die waarschijnlijk toegeschreven kunnen worden aan leakage.

In Hoofdstuk 9 introduceren we een hardware-efficiënte en leakage-reducerende methode om direct leakage te elimineren op een schaalbare wijze die niet additionele qubits of tijd nodig heeft, wat leidt tot een reductie van de logische foutgraad. We stellen in het bijzonder twee verschillende leakage-verminderingseenheden voor, die specifiek voor data qubits en voor ancilla qubits gemaakt zijn. Voor data qubits passen we een microgolf puls toe die leakage overdraagt aan zijn toegewezen uitlees resonator, waarna het snel vervalt naar de omgeving. Voor ancilla qubits gebruiken we een microgolf puls die een door leakage getergde toestand afbeeldt op een toestand uit de computationele subruimte.

Deze methodes voor twee-qubit gates, tomografie en leakage mitigatie dragen bij aan de reductie van foutgraden, wat voordelen biedt voor zowel kwantumfoutcorrectie als korte-termijn kwantumapparaten. In de Conclusie geven we een vooruitzicht op de potentiële uitdagingen voor supergeleidende kwantum computers, waaronder afstelbare koppelaars, real-time decoderen en fysieke foutgraden in grotere apparaten.

1

INTRODUCTION

1.1. QUANTUM COMPUTING LANDSCAPE

While the theory of quantum mechanics was developed already in the mid 1920s, it took a much longer time to conceive the notion of quantum computing. The foundations were laid in the 1980s and 1990s, with roots as far back as the 1970s, by physicists and computer scientists such as Richard P. Feynman [1], Paul A. Benioff, David E. Deutsch and Charles H. Bennett. Quantum key distribution introduced by Bennett and Brassard in 1984 [2] was one of the first examples showing that quantum mechanics could allow for the execution of tasks that are impossible using only classical resources, even though in the context of communication rather than computing. A major boost to the field of quantum computing was the factoring algorithm by Peter W. Shor in 1994 [3], which demonstrated that quantum computers could have significant practical implications. Realizing that noise was an obstacle to quantum computing that needed to be dealt with, Shor also introduced the first quantum error correcting code [4] and contributed to the emerging field of fault tolerance [5]. While practical levels of noise are still one of the most important concerns nowadays, Shor's code pioneered the following research efforts in quantum error correction and fault tolerance [6]. These succeeded in proving that it is indeed possible to achieve fault tolerance if error rates are below a certain threshold [7–9], paving the way for experimental endeavors in building a functional quantum computer.

In 2021 quantum computing is a fast-growing enterprise. Not only universities and research institutes are trying to develop this new technology, but also large companies. Startups are being founded thanks to the support of governments and private investors as well. The effort is not only concerned with quantum computing, which is the focus of this thesis (mostly with superconducting qubits), but also quantum communication and sensing. The quantum-computing platforms that are being developed in experiment range from superconducting qubits to trapped ions, photons, spin qubits, Majorana qubits and nitrogen-vacancy centers.

In the midst of all of this, it is not yet clear what will be the killer application for quantum computing, if any. As loading large amounts of classical data on quantum hardware seems a daunting problem [10], the most likely applications will involve relatively

low amounts of data but a sufficiently long computation to be intractable by classical means. In particular, quadratic quantum speedups seem to not be enough to beat classical computers in any reasonable timescale, whereas at least cubic or quartic speedups (or of course exponential speedups) have a much better chance [11]. One of the most important candidate areas for applications is quantum chemistry [12]. Shor's factoring algorithm [3] is also an application to break widely-used RSA public-key cryptosystems, but new quantum-resistant cryptographic systems are being developed and deployed to inhibit its usefulness against encrypted data generated today or in the future [13]. An ironic outcome of quantum computing might be better classical algorithms taking inspiration from the quantum ones. It is not only a matter of return on investment for the business sector, but quantum computers might also be a great tool for science to discover new physics and chemistry. This applies both to direct scientific discoveries like demonstrating the mechanisms of high-temperature superconductivity [14], as well as to fundamental understanding of decoherence and quantum mechanics at large scales.

The most important milestone reached so far by quantum computing is probably the demonstration of so-called "quantum supremacy" [15]. That is, the execution of a task which, as far as we know, cannot be executed efficiently by a classical computer, even though the considered quantum circuits are fairly shallow. However, this task (sampling from random quantum circuits) does not provide any practical quantum advantage, contrarily to the sought-after applications discussed above. An important goal towards reaching practical quantum advantages is to scale up quantum error correction, which is an active area in experimental research [16–18]. This is quite challenging as quantum error correction requires error rates to be below threshold. Furthermore, these error rates should be way below threshold to reduce an otherwise-large overhead in terms of number of qubits and number of operations. Error mitigation [19] corresponds to a series of techniques being developed to mitigate certain kinds of noise in current processors, without the full machinery of quantum error correction. While a future quantum computer might use both error correction and mitigation [20–22], it not clear whether error mitigation can suffice alone.

The dangers ahead on the way to useful quantum computing are significant. The most recent estimates require tens of thousands to millions of qubits [11, 23] to solve quantum-chemistry problems or break cryptography on a significant scale. While these estimates represent orders of magnitude of improvement over previous ones, we cannot naively extrapolate past accomplishments into the future. On the hardware side, coherence times and error rates have improved by orders of magnitude as well [24], but there might emerge limiting factors that are difficult to eliminate. Regarding the financial support, many governments and venture funds have made their first big bet on quantum computing and a few more might do so too. At the same time, we have to deliver on those expectations if we want a second bet to follow up.

Most challenges can be summarized in a few words: lowering the error rates and scaling up. There is actually a third aspect, which is to keep the error rates low while scaling up. Error rates need to be lowered in hardware first and then reduced as much as possible with quantum error correction and mitigation. Furthermore, all operations should be way below threshold, ideally by a few orders of magnitude. Scalability pertains many aspects, among which are fabrication (high-yield production with on-target pa-

rameters), calibration (automatized routines), classical electronics and heat load near the chip, connectivity to and within the chip and potentially across different chips [25]. Keeping error rates low and scaling up requires to be able to think modularly about qubits, operations and control. To this end, filters and tunable couplers between the qubits and/or the external world will be of crucial importance (see also Chapter 10).

1.2. THESIS OUTLINE

In this thesis we focus on the first challenge, that is, mitigating noise at the hardware level and lowering the error rates of some of the fundamental operations. We then study the beneficial impact on quantum error correction, with respect to a realistic noise model. Among the considered noise sources, a common thread in the following chapters is an often-neglected aspect that plagues quantum error correction, which is leakage. In superconducting transmon qubits, leakage comes mostly from the conditional-phase gate (see Section 2.8.3). Other peculiar realistic noise sources that we include in our modeling are low-frequency flux noise (see Section 3.2.1) and distortions of the electronic signals (see Section 3.2.5).

The first few chapters (2-5) provide an introduction to the concepts that are useful for the understanding of the following ones, which constitute the bulk of my research.

Chapter 2 introduces superconducting qubits, as well as techniques to control and measure them. Since qubits and these techniques have often been developed in parallel with the understanding of noise, it is hard to separate their description from an introduction to noise in superconducting qubits. However, in Chapter 2 we have chosen to only briefly describe certain features of the noise where necessary, whereas we have postponed a detailed description of noise to Chapter 3.
In Chapter 2 we start with a description of the physics of the Josephson junction, which is the fundamental element allowing superconducting quantum computing to exist. We give a physical motivation for the transmon qubit (transmon in short), describing how it developed from the Cooper-pair box to counter charge noise. Note that most of this thesis, especially Chapters 6, 8 and 9, is focused on transmons. In Chapter 2, only after briefly describing other superconducting qubits, we formalize the mathematical recipe of circuit quantization, which can be used to derive the Hamiltonian of any circuit and qubit. We then discuss various methods for reset (particularly relevant for the leakage-reduction units in Chapter 9), readout, single-qubit gates and two-qubit gates (particularly relevant for the gating scheme in Chapter 6).

In **Chapter 3** we broadly discuss noise in superconducting qubits. First, we give an overview of the major sources of noise. Then, in relation to that, we summarize the noise models that we use for the Lindblad simulations (used in Chapters 6, 8 and 9) and for the density-matrix simulations (used in Chapters 8 and 9), with more details being presented in Chapters 6 and 8, respectively. Finally, we discuss gate-benchmarking tools such as process tomography and randomized benchmarking (relevant for Chapter 7 and partially Chapter 6).

Chapter 4 introduces quantum error correction in general (so not only based on superconducting qubits). We describe the surface code (relevant for Chapters 8 and 9), as well as fault tolerance and decoding, mostly in the context of strictly two-level systems.

In **Chapter 5** we discuss the interplay between quantum error correction and leak-

age outside of the two levels forming the qubit computational subspace. We review the literature that studied the effect of leakage on the performance of quantum error correction, as well as the leakage-reduction units that were introduced (relevant for Chapter 8 and especially Chapter 9). Some of the previous work had a general scope, whereas some was focused on either superconducting transmon qubits or trapped-ion qubits. We conclude Chapter 5 with a comparison to the leakage models and results in this thesis.

In **Chapter 6** we introduce two variations of a technique for the controlled-phase gate, that we call Net Zero. In transmons, there are three main ways to perform a controlled-phase gate (reviewed in Section 2.8.3). Among those, baseband flux pulsing, although it is the fastest approach, had the disadvantage of being susceptible to long-timescale distortions, making the gate not repeatable. Net Zero divides a baseband flux pulse into two halves with opposite polarity, removing the DC component of the pulse, thus helping to suppress long-timescale distortions. On top of that, the symmetry of the Net Zero pulse allows for echoing out quasi-static components of flux noise, leading to a high-fidelity gate. Furthermore, Net Zero allows for destructive interference of leakage in analogy to a Mach-Zehnder interferometer. The error model and the Lindblad simulations that I developed match the experimental results and have been useful to reach the best performance in experiment, as well as to analyze the error budget of the gate.
In the second part of Chapter 6, we introduce Sudden Net Zero, whose simpler (and faster) pulse shape preserves all features above and adds easiness of tuneup and conditional-phase tunability.

In **Chapter 7** we describe Spectral Quantum Tomography. In experiment, gates are often tuned up by repeating a gate for a variable number of times and measuring the evolution of a certain error signature. Spectral tomography is the answer to the question "How can I extract the maximum amount of information by repeating a gate multiple times?". In particular, spectral tomography provides the maximum amount of information that depends on the gate only, i.e. that it is insensitive to state-preparation and measurement errors. Unlike randomized benchmarking, spectral tomography is non-scalable as it can be reasonably applied only to single- and two-qubit gates. However, contrasted to a single number for the average fidelity in randomized benchmarking, spectral tomography produces detailed information about relaxation, dephasing, leakage, non-Markovianity and other kinds of errors.

Chapter 8 provides a way to detect leakage in a transmon-based surface code. Stabilizer codes are not designed to correct leakage errors. Leakage has thus a disruptive effect, because a leaked data qubit effectively reduces the code distance and a leaked ancilla-qubit effectively disables the parity-check unit while spreading errors to nearby data qubits. However, these effects produce characteristic patterns in the measured syndrome, which can be used to detect leakage. We use hidden Markov models, one for each qubit, to detect leakage based on the history of measured syndromes on neighboring stabilizers. We train and benchmark these models with respect to density-matrix simulations of the distance-3 rotated surface code. We show that post-selecting out runs where leakage was detected allows us to significantly improve the logical performance of the code.

In **Chapter 9** we introduce two leakage-reduction units (LRUs) for a transmon-based surface code. The relaxation rate of the transmon sets the average time spent by a qubit in the leaked state. Hence, leakage is local in space and time for a sufficiently large code,

leading to the existence of a threshold, albeit it is expected to be low (assuming that regular error rates are sufficiently below threshold). Using LRUs to quickly bring a qubit back to the computational subspace can lead to a higher threshold. Unlike previous proposals that require extra qubits, gates or time, we introduce a hardware-efficient scheme with no overhead that uses two separate LRUs for data and ancilla qubits. For data qubits, using Lindblad simulations we study a microwave pulse on the transmon. The pulse trades two excitations on the transmon for one in its dedicated readout resonator, where the excitation quickly decays to the feedline environment. For ancilla qubits, we consider a $|1\rangle \leftrightarrow |2\rangle$ π pulse on the transmon, conditioned on the declaration of a $|2\rangle$. Using density-matrix simulations of the distance-3 rotated surface code, we significantly restore the logical performance of the code, even if the LRUs are implemented with limited fidelity.

We conclude by outlining in **Chapter 10** the most critical issues which, in my opinion, need to be solved for superconducting quantum computing to succeed. Among these, I highlight in particular the problem of crosstalk and the benefits of tunable couplers, the issues with implementing real-time decoding, as well as scalability issues regarding physical error rates, dilution refrigerators and the chip itself.

REFERENCES

[1] J. Preskill, *Quantum computing 40 years later,* (2021), arXiv:2106.10522 [quant-ph] .

[2] C. H. Bennett and G. Brassard, *Quantum cryptography: Public key distribution and coin tossing,* in *Proceedings of IEEE International Conference on Computers, Systems, and Signal Processing* (India, 1984) p. 175.

[3] P. W. Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,* SIAM Journal on Computing **26**, 1484 (1997).

[4] P. W. Shor, *Scheme for reducing decoherence in quantum computer memory,* Phys. Rev. A **52**, R2493 (1995).

[5] P. Shor, *Fault-tolerant quantum computation,* in *Proceedings of 37th Conference on Foundations of Computer Science* (1996) pp. 56–65.

[6] D. Gottesman, *Quantum Error Correction and Fault-Tolerance,* (2005), arXiv:quant-ph/0507174 [quant-ph] .

[7] D. Aharonov and M. Ben-Or, *Fault-tolerant quantum computation with constant error rate,* SIAM Journal on Computing **38**, 1207 (2008).

[8] E. Knill, R. Laflamme, and W. H. Zurek, *Resilient quantum computation: error models and thresholds,* Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **454**, 365–384 (1998).

[9] A. Kitaev, *Fault-tolerant quantum computation by anyons*, Vol. 303 (Elsevier BV, 2003) p. 2–30.

[10] J. Preskill, *Quantum Computing in the NISQ era and beyond,* Quantum **2**, 79 (2018).

[11] R. Babbush, J. R. McClean, M. Newman, C. Gidney, S. Boixo, and H. Neven, *Focus beyond quadratic speedups for error-corrected quantum advantage,* PRX Quantum **2** (2021), 10.1103/prxquantum.2.010103.

[12] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, and et al., *Quantum chemistry in the age of quantum computing,* Chemical Reviews **119**, 10856–10915 (2019).

[13] R. A. Perlner and D. A. Cooper, *Quantum resistant public key cryptography: A survey,* in *Proceedings of the 8th Symposium on Identity and Trust on the Internet*, IDtrust '09 (Association for Computing Machinery, New York, NY, USA, 2009) p. 85–93.

[14] *The Hubbard model at half a century,* Nature Physics **9**, 523 (2013).

[15] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, *Quantum supremacy using a programmable superconducting processor,* Nature **574**, 505–510 (2019).

[16] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, N. Lacroix, G. J. Norris, M. Gabureac, C. Eichler, and A. Wallraff, *Repeated quantum error detection in a surface code,* Nature Physics **16**, 875–880 (2020).

[17] J. F. Marques, B. M. Varbanov, M. S. Moreira, H. Ali, N. Muthusubramanian, C. Zachariadis, F. Battistel, M. Beekman, N. Haider, W. Vlothuizen, A. Bruno, B. M. Terhal, and L. DiCarlo, *Logical-qubit operations in an error-detecting surface code,* Nature Physics (2021).

[18] Z. Chen, K. J. Satzinger, J. Atalaya, A. N. Korotkov, A. Dunsworth, D. Sank, C. Quintana, M. McEwen, R. Barends, P. V. Klimov, S. Hong, C. Jones, A. Petukhov, D. Kafri, S. Demura, B. Burkett, C. Gidney, A. G. Fowler, A. Paler, H. Putterman, I. Aleiner, F. Arute, K. Arya, R. Babbush, J. C. Bardin, A. Bengtsson, A. Bourassa, M. Broughton, B. B. Buckley, D. A. Buell, N. Bushnell, B. Chiaro, R. Collins, W. Courtney, A. R. Derk, D. Eppens, C. Erickson, E. Farhi, B. Foxen, M. Giustina, A. Greene, J. A. Gross, M. P. Harrigan, S. D. Harrington, J. Hilton, A. Ho, T. Huang, W. J. Huggins, L. B. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, K. Kechedzhi, S. Kim, A. Kitaev, F. Kostritsa, D. Landhuis, P. Laptev, E. Lucero, O. Martin, J. R. McClean, T. McCourt, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Newman, M. Y. Niu, T. E. O'Brien, A. Opremcak, E. Ostby, B. Pató, N. Redd, P. Roushan, N. C. Rubin, V. Shvarts, D. Strain, M. Szalay, M. D. Trevithick, B. Villalonga, T. White, Z. J. Yao,

P. Yeh, J. Yoo, A. Zalcman, H. Neven, S. Boixo, V. Smelyanskiy, Y. Chen, A. Megrant, J. Kelly, and G. Q. AI, *Exponential suppression of bit or phase errors with cyclic error correction,* Nature **595**, 383 (2021).

[19] S. Endo, Z. Cai, S. C. Benjamin, and X. Yuan, *Hybrid quantum-classical algorithms and quantum error mitigation,* Journal of the Physical Society of Japan **90**, 032001 (2021).

[20] M. Lostaglio and A. Ciani, *Error mitigation and quantum-assisted simulation in the error corrected regime,* (2021), arXiv:2103.07526 [quant-ph] .

[21] C. Piveteau, D. Sutter, S. Bravyi, J. M. Gambetta, and K. Temme, *Error mitigation for universal gates on encoded qubits,* (2021), arXiv:2103.04915 [quant-ph] .

[22] Y. Suzuki, S. Endo, K. Fujii, and Y. Tokunaga, *Quantum error mitigation for fault-tolerant quantum computing,* (2021), arXiv:2010.03887 [quant-ph] .

[23] C. Gidney and M. Ekerå, *How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits,* Quantum **5**, 433 (2021).

[24] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, *Superconducting qubits: Current state of play,* Annual Review of Condensed Matter Physics **11**, 369 (2020).

[25] G. J. N. Alberts, M. A. Rol, T. Last, B. W. Broer, C. C. Bultink, M. S. C. Rijlaarsdam, and A. E. Van Hauwermeiren, *Accelerating quantum computer developments,* EPJ Quantum Technology **8**, 18 (2021).

2

SUPERCONDUCTING QUBITS

Superconducting qubits [1–4] are one of the most promising platforms for quantum computing nowadays. In this chapter we review some of the developments towards the transmon and other types of superconducting qubits, as well as typical methods for their initialization, control and readout.

2.1. JOSEPHSON JUNCTION

The fundamental circuit element allowing superconducting quantum computing in the first place is the Josephson junction. Here we provide the basic intuition about it, while detailed information can be found in [5, 6]. Standard superconductivity, at least at low temperature, is well explained by BCS theory, where electrons do not travel independently but form Cooper pairs. Breaking a Cooper pair requires a certain amount of energy, referred to as the superconducting gap. The Josephson effect corresponds to the quantum tunneling of Cooper pairs between two superconductor pieces separated by a weak link. In practice, the weak link forming such a Josephson junction can be either made of an insulating material (S-I-S junction), or a piece of normal metal (S-N-S junction).

As the two pieces of superconductor face each other, they also create a small capacitor. The circuit diagram for a real Josephson junction is thus given in Fig. 2.1(a). For the moment we assume that this is the only electrical circuit we have. Each half of the circuit, corresponding to one of the two superconducting pieces, is an "island" not directly connected to anything else (except that one is grounded in practice; furthermore, usually in a larger circuit only one island is really isolated from other circuitry, if there is any island at all). The total number of Cooper pairs on both islands is fixed, but the number in one of them, N, can vary due to Josephson tunneling. Quantum mechanically this number corresponds to an operator $\hat{N} = \sum_{N=-\infty}^{+\infty} N |N\rangle \langle N|$.

The tunneling of Cooper pairs can be described via the Hamiltonian

$$H_J = -\frac{E_J}{2} \sum_{N=-\infty}^{+\infty} (|N\rangle \langle N+1| + |N+1\rangle \langle N|), \tag{2.1}$$

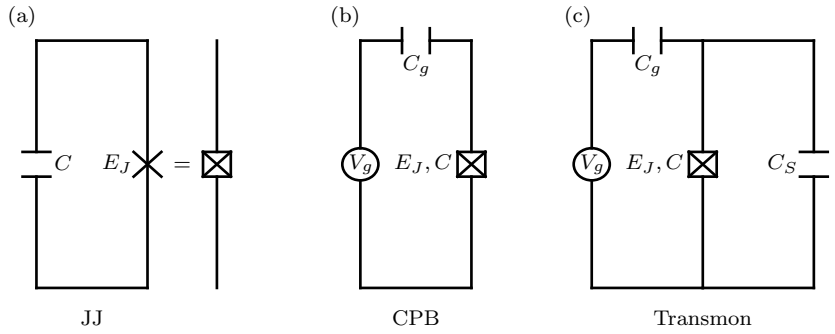

Figure 2.1: Circuit diagrams for a (a) Josephson junction (see Section 2.1), (b) Cooper-pair box (see around Eq. (2.4)), (c) transmon (see after Eq. (2.4)).

where E_J is the energy associated with one Cooper pair tunneling and it depends on the material and characteristics of the junction. We can introduce the phase states

$$|\phi\rangle = \sum_{N=-\infty}^{+\infty} e^{iN\phi} |N\rangle \tag{2.2}$$

and an operator $e^{i\hat{\phi}} = 1/(2\pi) \int_0^{2\pi} d\phi\, e^{i\phi} |\phi\rangle\langle\phi|$, whose eigenstates are the $\{|\phi\rangle\}$. One can easily check that $e^{i\hat{\phi}} |N+1\rangle = |N\rangle$. Consequently, $e^{i\hat{\phi}} = \sum_{N=-\infty}^{+\infty} |N\rangle\langle N+1|$ and $H_J = -E_J \cos\hat{\phi}$, where the phase ϕ can be interpreted as the superconducting-phase difference across the junction. Furthermore, one can compute that $[\hat{\phi}, \hat{N}] = i$, with the cautiousness that only periodic functions of the compact variable ϕ are well-defined, like $e^{i\hat{\phi}}$. Hence, $\hat{\phi}$ and $\hat{N}$ behave like position and momentum, respectively.

To get the full Hamiltonian of the junction, the energy stored in the capacitor is given by $H_C = \hat{Q}^2/2C$, where C is the intrinsic junction capacitance and Q is the charge. As $Q = 2eN$ for Cooper pairs, the overall Hamiltonian of a Josephson junction, $H = H_C + H_J$, is given by

$$H = 4E_C \hat{N}^2 - E_J \cos\hat{\phi}, \tag{2.3}$$

defining $E_C = e^2/2C$.

2.2. TRANSMON

COOPER-PAIR BOX

In the limit where $E_J \ll E_C$, the eigenstates of the Josephson-junction Hamiltonian in Eq. (2.3) are approximately the eigenstates of the charge operator $2e\hat{N}$. If one connects a voltage source V_g to the Josephson junction, one can bias the number of Cooper pairs on the island, as well as generate superpositions of charge eigenstates. This device, known as the Cooper-Pair Box (CPB) [6], is shown in Fig. 2.1(b), where V_g is connected to the Josephson junction by a capacitance C_g. The bias is modeled by an offset charge $N_g =$

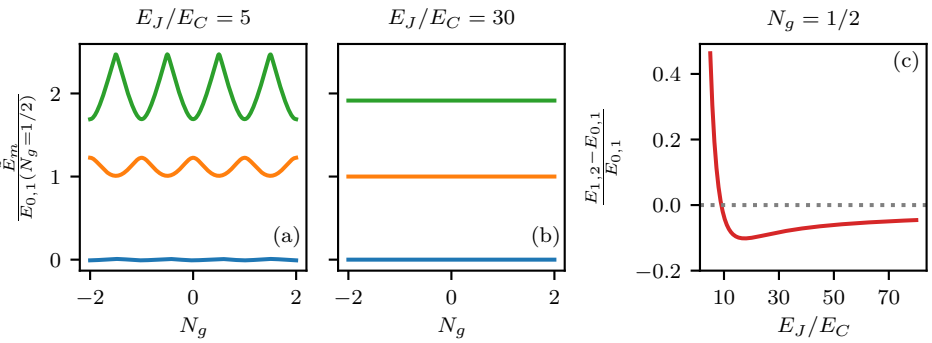

Figure 2.2: Spectrum and relative anharmonicity. (a,b) Spectrum for a CPB ($E_J/E_C = 5$) and for a transmon ($E_J/E_C = 30$), respectively, as a function of the charge offset N_g. Let $E_m \in \{E_0, E_1, E_2\}$ be the energies of the three lowest eigenstates of Eq. (2.4). We plot $\tilde{E}_m \coloneqq E_m - \int_0^1 dN_g\, E_0(N_g)$ relative to the transition frequency $E_{0,1} \coloneqq E_1 - E_0$ at $N_g = 1/2$ (where this transition frequency is minimal). One can see that the $\{E_m\}$ are insensitive to N_g in the transmon regime. (c) Relative anharmonicity $\frac{E_{1,2}-E_{0,1}}{E_{0,1}}$ versus E_J/E_C, where $E_{i,j} \coloneqq E_j - E_i$, evaluated here at $N_g = 1/2$. One can see that the anharmonicity is relatively low and that the system becomes more harmonic with increasing E_J/E_C.

$C_g V_g/2e$, thus giving the Hamiltonian

$$H_{\text{CPB}} = 4E_C(\hat{N} - N_g)^2 - E_J \cos\hat{\phi}. \tag{2.4}$$

CHARGE NOISE

In practice, any effective voltage source in the environment due to charges in the material can couple to the island and make N_g fluctuate (charge noise). As this holds even if there is no purposefully placed voltage source, the Hamiltonian in Eq. (2.3) is an idealization in the absence of charge noise. Randomly hopping charges cause fluctuations in N_g. As shown in Fig. 2.2(a), the energy of the CPB eigenstates varies strongly with N_g. As these fluctuations in N_g happen uncontrollably, they lead to decoherence of the CPB, which had typical dephasing times of at most a few nanoseconds [4].

TRANSMON

The solution to the problem of charge noise has been to introduce a large capacitance C_S shunting the Josephson junction. This design, shown in Fig. 2.1(c), is known as the transmon [7]. While the transmon Hamiltonian is formally the same as in Eq. (2.4), in this case $E_C = e^2/[2(C + C_S)]$. For large C_S one can enter the regime with $E_J \gg E_C$, where the dependence of the spectrum on N_g is suppressed (see Fig. 2.2(b)). Compared to the CPB, the eigenstates of the Hamiltonian are no longer approximate eigenstates of the charge operator but of the phase operator $e^{i\hat{\phi}}$. These eigenstates (see Eq. (2.2)) have a broad distribution over charge states. Intuitively, a hopping charge causes a small shift in such a distribution over charge states (rather than jumps between phase eigenstates). In practice, this shift is negligible for $E_J/E_C \gtrsim 30$ and leads to the flat dependence of the transmon eigenenergies on N_g (see Fig. 2.2(b)), thus giving protection against charge noise.

TRANSMON ENERGY LEVELS

To better understand the transmon eigenstates, we expand $\cos\hat{\phi}$ around 0 up to 4th order. This approximation is based on the fact that the lowest-energy phase eigenstate has $\phi = 0$ and that we expect this to be approximately the ground state in the transmon regime. The Hamiltonian then takes the approximate form

$$H_{\text{transmon}} \approx 4E_C\hat{N}^2 + \frac{E_J}{2}\hat{\phi}^2 - \frac{E_J}{24}\hat{\phi}^4, \tag{2.5}$$

where we can neglect N_g in the transmon regime as discussed above. The first two terms in Eq. (2.5) constitute a quantum harmonic oscillator where $\hat{N}$ and $\hat{\phi}$ take the role of position and momentum, respectively. Explicitly, we can introduce annihilation and creation operators b and $b^\dagger$ via

$$\hat{N} = \frac{i}{\sqrt{2}}\left(\frac{E_J}{8E_C}\right)^{1/4}(b^\dagger - b) \tag{2.6}$$

$$\hat{\phi} = \frac{1}{\sqrt{2}}\left(\frac{E_J}{8E_C}\right)^{-1/4}(b^\dagger + b). \tag{2.7}$$

Then

$$H_{\text{transmon}} = \sqrt{8E_JE_C}\left(b^\dagger b + \frac{1}{2}\right) - \frac{E_C}{12}\left(b + b^\dagger\right)^4. \tag{2.8}$$

In the following we neglect the $+1/2$ term in Eq. (2.8) since it is a constant energy shift for the whole Hamiltonian. The last, non-linear term in Eq. (2.8) turns this harmonic oscillator into a (slightly) anharmonic one, for which the energy levels are not equally spaced (see Fig. 2.2(b,c)). Analytically, we can expand $(b + b^\dagger)^4$. This expansion contains terms that are "energy conserving", like $b^\dagger b^\dagger bb$, i.e. that contain twice b and twice $b^\dagger$ (in different orders), while all the other terms are "energy non-conserving". We apply the Rotating Wave Approximation (RWA), which consists of neglecting all the energy non-conserving terms. Then, using the commutation relationship $[b, b^\dagger] = 1$ (identity) to rearrange the b's and $b^\dagger$'s, one can get

$$H^{\text{RWA}}_{\text{transmon}} = \omega b^\dagger b + \frac{\alpha}{2}b^\dagger b^\dagger bb, \tag{2.9}$$

where

$$\omega = \sqrt{8E_JE_C} - E_C, \tag{2.10}$$

$$\alpha = -E_C \tag{2.11}$$

are the transmon frequency and anharmonicity, respectively (within the given approximations). We label the eigenstates of $b^\dagger b$ as $|0\rangle, |1\rangle, |2\rangle \ldots$ Since $b^\dagger b^\dagger bb = b^\dagger b(b^\dagger b - 1)$, these are the eigenstates of $H^{\text{RWA}}_{\text{transmon}}$ as well. Labeling the energy of $|i\rangle$ as ω_i, we can rewrite Eq. (2.9) as

$$H^{\text{RWA}}_{\text{transmon}} = \sum_{i=0}^{+\infty}\omega_i\,|i\rangle\langle i| \tag{2.12}$$

with

$$\omega_0 = 0, \qquad \omega_1 = \omega, \qquad \omega_2 = 2\omega + \alpha, \qquad \omega_3 = 3\omega + 3\alpha \tag{2.13}$$

for the lowest energy levels. Hence the transition frequencies $\omega_{i,j}$ between $|i\rangle$ and $|j\rangle$ are

$$\omega_{0,1} = \omega \qquad \omega_{1,2} = \omega + \alpha \qquad \omega_{2,3} = \omega + 2\alpha. \tag{2.14}$$

The anharmonicity thus directly expresses the degree of unequal spacing between energy levels, compared to a harmonic oscillator. In particular, since α is negative (see Eq. (2.11)), $\omega_{1,2} < \omega_{0,1}$.

We briefly note that the 4th-order expansion, together with the RWA, can be somewhat inaccurate especially in the vicinity of avoided crossings between energy levels [8]. Instead, a 6th-order expansion without performing the RWA matches the exact result in a more accurate way [8].

DISCUSSION ABOUT THE ANHARMONICITY

A finite anharmonicity is crucial to be able to use the two lowest-energy states as a qubit. Indeed, in a perfectly harmonic oscillator there is no way to control the states $|0\rangle$ and $|1\rangle$ independently from $|2\rangle$ (or even higher states). However, typical transmons have frequencies in the range $\omega/2\pi \sim$ 3-8 GHz and anharmonicities in the range $\alpha/2\pi \sim -$150-400 MHz, due to practical limitations [9]. A relatively low anharmonicity is the price to pay for insensitivity to charge noise. While it still allows the use of transmons as qubits, the issue is that this limits the speed to execute gates and requires carefully-engineered pulses to avoid leakage to higher excited states (see Sections 2.7 and 2.8).

2.2.1. TUNABLE TRANSMON

The transmon frequency can be made tunable by using two Josephson junctions in parallel (SQUID), rather than a single one (see Fig. 2.3(a)). This creates a loop that allows to apply an external magnetic field Φ_e for tuning. The Hamiltonian is

$$H_{\text{SQUID}} = 4E_C\hat{N}^2 - E_{J_1}\cos\hat{\phi} - E_{J_2}\cos(\hat{\phi} + \phi_e), \tag{2.15}$$

where $\phi_e = 2\pi\Phi_e/\Phi_0$ and $\Phi_0 = h/2e$ is the magnetic flux quantum (it can be derived using circuit quantization as described in Section 2.4). Using trigonometric identities, Eq. (2.15) can be rewritten as [7]

$$H_{\text{SQUID}} = 4E_C\hat{N}^2 - E_{J_\Sigma}(\phi_e)\cos(\hat{\phi} - \varphi_0), \tag{2.16}$$

where

$$E_{J_\Sigma}(\phi_e) \coloneqq (E_{J_1} + E_{J_2})\sqrt{\cos^2\left(\frac{\phi_e}{2}\right) + d^2\sin^2\left(\frac{\phi_e}{2}\right)} \tag{2.17}$$

$$\tan\varphi_0 = d\tan\left(\frac{\phi_e}{2}\right) \tag{2.18}$$

and $d = \left|E_{J_1} - E_{J_2}\right|/(E_{J_1} + E_{J_2})$ is the degree of asymmetry. For a time-independent ϕ_e, the shift φ_0 can be removed by a change of variables, leading to the Hamiltonian

$$H_{\text{SQUID}} = 4E_C\hat{N}^2 - E_{J_\Sigma}(\phi_e)\cos\hat{\phi}. \tag{2.19}$$

If ϕ_e is time dependent, the transmon eigenstates change in a time-dependent manner via φ_0. However, this change is small unless d is large and ϕ_e gets close to π. For example, if $d = 0.02$ and $\phi_e = 1.4$ rad (a typical value for CZ; see Fig. 2.6), then $\varphi_0 \approx 1$ deg only.

Equation (2.19) is equivalent to the Hamiltonian of a single-junction transmon (see Eq. (2.4)) but with a tunable Josephson energy, thus the results in Section 2.2 apply by simply replacing E_J with $E_{J_\Sigma}(\phi_e)$. If the junctions are symmetric, i.e. if $E_{J_1} = E_{J_2} \equiv E_J \implies d = 0$, Equation (2.17) simplifies to

$$E_{J_\Sigma}(\phi_e) = 2E_J \left|\cos(\phi_e/2)\right|. \tag{2.20}$$

Based on Eq. (2.10), it follows that

$$\omega(\phi_e) + E_C \propto \sqrt{\left|\cos(\phi_e/2)\right|}. \tag{2.21}$$

One can observe that the transmon frequency decreases when ϕ_e goes from 0 to π and that it never exceeds the value at $\phi_e = 0$. These observations also hold for any $d \neq 0$ as can be understood from Eq. (2.17) since $d \in [0,1]$ (see also Fig. 2.3(b)). We note that for low-asymmetry SQUIDs ($d \ll 1$) and ϕ_e close to π, $E_{J_\Sigma}(\phi_e)$ in Eq. (2.17) can become comparable or lower than E_C, thus going outside of the transmon regime. In that case, one cannot simply insert Eq. (2.17) into Eq. (2.10) to get the transmon frequency ω. To avoid these issues, in Fig. 2.3(b) we compute ω by full diagonalization of Eq. (2.4) at $N_g = 1/2$, as done for Fig. 2.2. Recall that $N_g = 1/2$ corresponds to the sweetspot as a function of N_g when outside of the transmon regime.

The advantages of frequency tunability are multiple: avoiding frequency collisions on a multi-qubit chip, avoiding two-level systems strongly coupled to the qubit (see Section 3.2.1), mitigating crosstalk (see Section 3.2.7), allowing flux-based single- and two-qubit gates (see Sections 2.7 and 2.8). On the other hand, magnetic-field fluctuations (flux noise) introduce an important dephasing mechanism that is absent in fixed-frequency transmons (given the use of air-bridges; see Fig. 2.4). We say more about flux noise in Section 3.2.1. Here we only introduce the concept of a sweetspot, that is, a point ϕ_e^* where

$$\frac{\partial \omega}{\partial \Phi_e}\Big|_{2\pi\frac{\Phi_e}{\Phi_0}\equiv\phi_e^*} = 0. \tag{2.22}$$

One can understand that sweetspots are first-order insensitive to fluctuations in Φ_e, since the transmon frequency ω around these points does not change as a function of Φ_e, up to first order. Independently of d, the "main" sweetspot is at $\phi_e^* = 0$, corresponding to the maximum ω. Focusing on the range $[0, 2\pi]$ by periodicity, there is also a second sweetspot at $\phi_e^* = \pi$ (see Fig. 2.3(b)), corresponding to the minimum of ω, where the specific value of this minimum $\omega_{\min}$ varies with d. At $d = 0$, $\omega_{\min} = 0$ and it is thus too low to control the transmon with microwave pulses. If instead d is large enough so that $\omega_{\min}/2\pi \gtrsim 3$ GHz, then one can use standard microwave pulses and can in principle exploit the flux-insensitivity at the second sweetspot. Asymmetric SQUIDs can thus be advantageous if, for example, one fabricates all transmons in a chip in the same way, but then biases half of them to the second sweetspot (this approach is currently taken by ETH Zürich in their surface code). As a second example, one could make the second

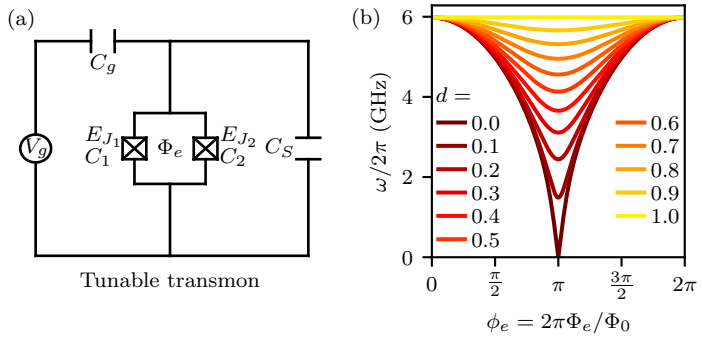

Figure 2.3: Tunable transmons. (a) The circuit diagram for a tunable transmon (see Section 2.2.1). (b) The transmon frequency ω as a function of the external flux Φ_e for various degrees of asymmetry d (see after Eq. (2.18)). While ω follows closely Eq. (2.10) with E_J given in Eq. (2.17) for large d and/or ϕ_e not too close to π, to avoid the issues discussed after Eq. (2.21), we obtain ω by diagonalizing directly Eq. (2.4) at $N_g = 1/2$.

sweetspot align with the interaction frequency in a flux-based two-qubit gate, mitigating dephasing during the gate. However, it is still hard to fabricate junctions which precisely hit their target parameters, thus also limiting the use of asymmetric SQUIDs.

2.2.2. STARMON

There exist multiple physical designs for a transmon. The design used in the DiCarlo lab is the so-called starmon (see Fig. 2.4). Two large capacitor plates face each other in the middle, forming the transmon shunting capacitance. The SQUID that provides the non-linearity is a relatively small element in between the two plates. The exact size of the Josephson junctions determines the Josephson energy, which in turn sets the transmon frequency. The flux line reaches the starmon on the SQUID side to provide flux tunability. The starmon also has four larger arms to which the bus resonators can connect, allowing two-qubit gates with other starmons (see Section 2.8). The microwave drive line and the readout resonator connect to one smaller arm each. Each readout resonator has a dedicated Purcell filter (see Section 2.6.1) for faster readout while protecting qubit coherence.

The four starmons shown in Fig. 2.4 are part of a Surface-7 chip developed in the DiCarlo lab. In line with the operation scheduling in Ref. [12] for the surface code, there are three sets of frequency bands for the transmon qubit frequencies. The target parameters for these transmon qubit frequencies are around 6.7, 6.0, 4.9 GHz, respectively. The readout resonators and Purcell filters have frequencies around 7-7.8 GHz, whereas the bus resonators around 20 GHz in the current design.

2.3. OTHER SUPERCONDUCTING QUBITS

The transmon belongs to the family of charge qubits, where the name comes from the fact that in the CPB the charge is a good quantum number (meaning that CPB eigenstates are approximately charge eigenstates; see Section 2.2). As discussed in Section 2.2, this is not

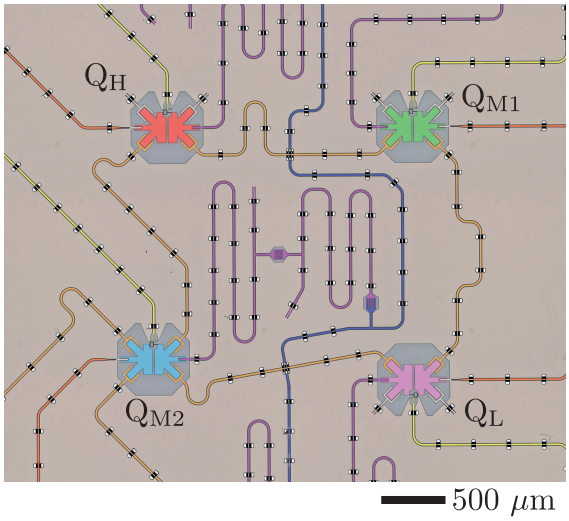

Figure 2.4: Optical image [10] zoomed in to four transmons of the seven-transmon device in Section 6.12 and Ref. [11]. False colors are added to help identify circuit elements. Transmons Q_H (red) and Q_L (pink) each connect to Q_{M1} (green) and Q_{M2} (cyan) using dedicated coupling bus resonators for each pair (light orange). Each transmon has a flux-control line for two-qubit gating (yellow), a microwave-drive line mostly for single-qubit gating (dark orange), and a dispersively-coupled resonator with Purcell filter for readout (purple). The readout-resonator/Purcell-filter pair for Q_{M2} is visible at the center of the image. The vertically running common feedline (blue) connects to all Purcell filters, enabling simultaneous readout of the four transmons by frequency multiplexing. Air-bridge crossovers enable the routing of all input and output lines to the edges of the chip and enable to ground the whole plane as well, effectively breaking loops that could otherwise be a souce of flux noise and flux crosstalk.

really the case for transmons, but due to the similar design they are usually categorized together. Here we discuss a few other families of superconducting qubits.

FLUX QUBIT

The other main family of superconducting qubits are flux qubits [13] and their evolutions. In a way, the flux qubit is an evolution of a SQUID, where instead of two junctions in a loop one uses three of them, one of which is smaller than the others. The introduction of a third junction changes the potential profile in a relevant way compared to a SQUID. Defining as γ the ratio between the E_J of the large and small junctions, for $\gamma > 1$ the Hamiltonian can be approximated as [2]

$$H \approx 4E_C N^2 - E_J \cos(\phi + \phi_e) - 2\gamma E_J \cos(\phi/2). \tag{2.23}$$

If $1 < \gamma < 2$, the potential profile qualitatively changes from a single well (like for a SQUID) to a double well in $[0, 2\pi]$. Each of these wells hosts one of the qubit eigenstates, which can be interpreted as corresponding approximately to current rotating in one or the other direction in the loop. In this regime the flux qubit is known as the persistent-current flux qubit.

CSFQ (CAPACITIVELY-SHUNTED FLUX QUBIT)

In the regime $\gamma > 2$, the interpretation of rotating currents no longer holds and the potential well is again only one. To improve coherence times, in this regime a shunting capacitance has been added such that $E_J \gg E_C$ (similarly to how a shunting capacitance was added to the CPB to get the transmon; see Section 2.2), leading to the so-called CSFQ [14]. The CSFQ features long coherence times ($T_1 \sim \mathcal{O}(100)\ \mu$s) and a positive anharmonicity (typically 500 MHz) that can be larger than the transmon in absolute value. It has been proposed to combine transmons and CSFQs in a single device to mitigate residual ZZ crosstalk (see Section 3.2.7) thanks to their opposite-sign anharmonicities [15].

FLUXONIUM

As it emerges from the discussion above, there is no clear-cut division between the two families of charge and flux qubits. There is rather a spectrum from CPB to transmon, SQUID, CSFQ and flux qubit. The fluxonium [16] features elements from both families, with a complex set of both transmon-like and flux-like transitions. Structurally, fluxonium is a loop with one small junction and with, instead of two larger junctions as in the flux qubit, many ($M = \mathcal{O}(100)$) junctions in series. The Hamiltonian is seemingly similar to Eq. (2.23):

$$H \approx 4E_C N^2 - E_J \cos(\phi + \phi_e) - M\gamma E_J \cos(\phi/M) \tag{2.24}$$

$$\approx 4E_C N^2 - E_J \cos(\phi + \phi_e) + \frac{E_L}{2}\phi^2 \tag{2.25}$$

with $E_L = (\gamma/M)E_J$, where the last approximation is motivated by the large M. Thus effectively, the junctions simply form a linear inductor. The reason why a "normal" geometric inductor is not used directly is that fluxonium requires a large inductance (corresponding to a small E_L, since $E_L = 4\pi^2\Phi_0^2/L$ for an inductor with inductance L). However, it is hard to realize geometric inductors with large inductance (and small parasitic capacitance).

An alternative to an array of Josephson junctions is to use a material with high kinetic inductance [17]. Different fluxonium designs then depend on the relative scale of E_C, E_J and E_L. So far, fluxonium exhibits record coherence times ($T_1, T_2 > 1$ ms [18]), thanks to the fact that the qubit eigenstates are engineered to be fairly isolated from each other, each in its own potential well. The flip side is that both single- and two-qubit gates are difficult to implement, although promising gating techniques are being developed [19, 20]. Furthermore, since the qubit frequency can be as low as ~ 500 MHz, in general single-qubit gates cannot be simply implemented with standard microwave drives (see Section 2.7) because the electronics cannot generate pulses with such a low frequency.

2.4. CIRCUIT QUANTIZATION

In Section 2.2 we could derive the Hamiltonian of the transmon due to the relatively simple structure of the circuit. Here we discuss the general procedure to get the Hamiltonian of an arbitrary electrical circuit [6], which can be used to derive the Hamiltonian of any other superconducting qubit (we use this procedure explicitly in Section 2.8.1). We limit ourselves to the case of a lossless circuit, although dissipation can also be included [6].

Consider a network of two-terminal circuit elements (capacitors, inductors, Josephson junctions...). The points where two or more of these elements connect are called nodes. Each connection between two nodes, independently of the specific element, is a branch. We assign a certain orientation (chosen arbitrarily) to each branch. One can associate a voltage $V_b(t)$ and a current $I_b(t)$ through each branch b, whose sign is determined by the orientation of the branch. In the absence of time-varying magnetic fields the circuit is conservative and the voltages and currents across different branches follow the Kirchhoff rules, namely, $\sum_{b\in\text{loop}} V_b = 0$ and $\sum_{b\in\text{node}} I_b = 0$. Due to these constraints, $\{V_b\}$ and $\{I_b\}$ are not independent variables. To solve for the classical dynamics, as well as for quantization, it is important to identify a set of independent variables. Furthermore, for a large circuit it is useful to have a general method to write the Kirchhoff laws for all nodes and loops.

While one could make different choices of independent variables, here we proceed as follows [6]. First, one can introduce a branch charge Q_b and a generalized branch flux Φ_b as

$$Q_b(t) = \int_{-\infty}^{t} dt'\, I_b(t') \tag{2.26}$$

$$\Phi_b(t) = \int_{-\infty}^{t} dt'\, V_b(t'). \tag{2.27}$$

As a note, elements whose V_b (resp. I_b) is solely a function of Q_b (Φ_b) are called capacitive (inductive). Then we choose a spanning tree of the network, i.e. a loop-free subset of branches such that each node is part of the tree. We also assign an orientation to each fundamental loop l, i.e. any loop that is obtained by adding a single branch to the spanning tree. We allow for the presence of (time-independent) external fluxes Φ_l^e, for which it holds

$$\sum_{b\in l} R_{bl}\Phi_b = \Phi_l^e, \tag{2.28}$$

where $R_{bl} = \pm 1$ depending on whether the branch and loop orientations are parallel or anti-parallel (see [21] for circuit quantization with time-dependent external fluxes). We pick any node and call this the "ground" node Φ_g. We can define node fluxes

$$\Phi_n = \sum_{b \in \text{path } g \to n} S_{bn} \Phi_b, \tag{2.29}$$

where the path is along the tree and $S_b = \pm 1$ depending on the relative orientation of the branch and the path. Note that these $\{\Phi_n\}$ are independent and any Φ_b can be expressed as a function of them, also using Eq. (2.28) to formally assign the external flux to the branch outside the tree. Furthermore, as $\dot{\Phi}_b = V_b$, by construction the Kirchhoff voltage laws are automatically satisfied when expressing them in terms of the voltages at the nodes $V_n := \dot{\Phi}_n$.

To systematically address the Kirchhoff current laws (and to later move to the Hamiltonian formalism for quantization), we consider the Lagrangian formalism. As we have chosen $\mathbf{\Phi}_n$ as the independent variables (bold face indicates vectors obtained by stacking all the variables) the Lagrangian is given in terms of $\mathbf{\Phi}_n$ and $\dot{\mathbf{\Phi}}_n$ as $\mathscr{L}(\mathbf{\Phi}_n, \dot{\mathbf{\Phi}}_n) = \mathscr{T}(\dot{\mathbf{\Phi}}_n) - \mathscr{U}(\mathbf{\Phi}_n)$, where $\mathscr{T}$ and $\mathscr{U}$ are the kinetic and potential energies, respectively. The most common energy terms, associated with capacitors, inductors and Josephson junctions are

$$C\dot{\Phi}_b^2/2, \qquad \Phi_b^2/2L, \qquad -E_J \cos \Phi_b, \tag{2.30}$$

respectively, where the branch variables need to be expressed in terms of the node variables. Note that with these terms the kinetic part of the Lagrangian is linear and can be expressed as

$$\mathscr{T}(\dot{\mathbf{\Phi}}_n) = \dot{\mathbf{\Phi}}_n^T \mathbf{C} \dot{\mathbf{\Phi}}_n/2, \tag{2.31}$$

where the capacitance matrix $\mathbf{C}$ is constructed from all the capacitances in the circuit. The Euler-Lagrange equations, i.e.

$$\frac{d}{dt}\left(\frac{\partial \mathscr{L}}{\partial \dot{\Phi}_n}\right) - \left(\frac{\partial \mathscr{L}}{\partial \Phi_n}\right) = 0, \tag{2.32}$$

one for each Φ_n, are then the equations of motion of the system. Note that taking these derivatives of the energy terms given in Eq. (2.30) always gives a current, so here the Euler-Lagrange equations are precisely the Kirchhoff current laws.

To proceed towards quantization we need to switch to the Hamiltonian formalism, because only then we can introduce operators and commutation relations. First, the generalized momenta are given by

$$q_n = \frac{\partial \mathscr{L}}{\partial \dot{\Phi}_n}. \tag{2.33}$$

Then the Hamiltonian is defined as

$$\mathscr{H}(\mathbf{\Phi}_n, \mathbf{q}_n) = \dot{\mathbf{\Phi}}_n(\mathbf{\Phi}_n, \mathbf{q}_n) \cdot \mathbf{q}_n - \mathscr{L}\big(\mathbf{\Phi}_n, \dot{\mathbf{\Phi}}_n(\mathbf{\Phi}_n, \mathbf{q}_n)\big), \tag{2.34}$$

where $\cdot$ indicates the scalar product. Note that one is required to express $\dot{\mathbf{\Phi}}_n$ in terms of $\mathbf{\Phi}_n$ and $\mathbf{q}_n$, which in general requires that the Lagrangian is concave with respect to the $\dot{\mathbf{\Phi}}_n$. In practice, here it means that $\mathbf{C}$ has to be invertible. If it is not, this pathological case is usually due to a Φ_n that does not have dynamics (i.e. there is no $\dot{\Phi}_n$ in the Euler-Lagrange equations) and thus can be algebraically removed using the Euler-Lagrange equations themselves. Then one obtains a lower dimensional $\mathbf{C}$ which is invertible. Finally, one can "promote" $\mathbf{\Phi}_n$ and $\mathbf{q}_n$ to operators and impose the canonical commutation relations $[\Phi_n, q_n] = i\hbar$ for each node, getting the quantum Hamiltonian of the system. In the rest of this thesis we use the convention that $\hbar = 1$.

2.5. RESET

Quantum computation requires the capability to initialize qubits in a known pure state, usually taken to be $|0\rangle$, the ground state of the system (here a transmon). This has to occur with high fidelity, needs to be fast for some applications and might involve resetting not just $|1\rangle$ but also $|2\rangle$ or even higher levels to $|0\rangle$. Furthermore, reset (also called initialization) can be unconditional or conditional on the measurement outcome (feedback). Here we discuss the pros and cons and features of various reset methods (relaxation, feedback [22, 23], all-microwave [24–26] and flux pulsing [27]). Other reset methods, such as the quantum circuit refrigerator [28] and flux modulation [29], are not discussed in detail here.

2.5.1. RELAXATION

The simplest approach lets the transmon passively relax to the ground state, on a time scale set by the relaxation time T_1 (see Section 3.1). Generally, waiting for $\sim 10T_1$ is more than enough. This method does not really require any calibration and is unconditional. However, it is relatively slow and it will (hopefully!) get worse in future devices with longer T_1's. We note that a precondition for this method to work is that the excitation rate from $|0\rangle$ to $|1\rangle$ is low, otherwise the steady state is not close to a pure state. More precisely, assume that the transmon is in thermal equilibrium with a bath at temperature T. The occupancy for each state $|i\rangle$ is given by $Z_i/Z =: n_i$, where $Z_i = e^{-\frac{\hbar\omega_i}{k_B T}}$ and $Z = \sum_i Z_i$ is the partition function, with $\hbar\omega_i$ the energy of state $|i\rangle$. For a transmon, $\omega_0 = 0$, $\omega_1 = \omega$ and $\omega_2 = 2\omega + \alpha$, where ω and α are the frequency and anharmonicity, respectively, as discussed in Section 2.2. Considering only the first three levels,

$$Z = 1 + e^{-\frac{\hbar\omega}{k_B T}} + e^{-\frac{\hbar(2\omega+\alpha)}{k_B T}}. \tag{2.35}$$

The lowest stage of the dilution refrigerator (fridge in short) where the chip is kept has a temperature of about $T \sim 20$ mK. Hence, for a common frequency $\omega/2\pi = 6$ GHz, one has $\hbar\omega \gg k_B T$. In this limit we can approximate $Z \approx 1 + e^{-\frac{\hbar\omega}{k_B T}}$, since $e^{-\frac{\hbar(2\omega+\alpha)}{k_B T}}$ in Eq. (2.35) is negligible compared to $e^{-\frac{\hbar\omega}{k_B T}}$ (recall that α is low). In other words, as thermal fluctuations cause only few excitations to $|1\rangle$ because of the relatively large energy gap, there are even less excitations to $|2\rangle$ (or to even higher levels, which we had already neglected in the

analysis above). Then the so-called average number of residual excitations is

$$n_1 \approx \frac{1}{e^{\frac{\hbar\omega}{k_B T}} + 1}. \tag{2.36}$$

For example, for $\omega/2\pi = 6$ GHz and $T = 20$ mK one has $n_1 \approx 6 \cdot 10^{-7}$. However, commonly $n_1 = 10^{-3}$-10^{-2} in the lab, which is not necessarily too high for practical purposes, but definitively higher than based on the thermalization picture. This means that the effective transmon temperature is rather at 42-64 mK, potentially because of quasi-particles (due to cosmic rays) and because of the connections of the transmon to higher stages of the fridge and to the room-temperature electronics, but it is not fully understood.

We briefly note that for fluxonium (see Section 2.3), the two-level-system approximation before Eq. (2.36) still holds because the anharmonicity is positive and large. Hence, we can apply Eq. (2.36) to fluxonium as well. However, since the frequency is much lower, on the order of 500 MHz (see Section 2.3), even with $T = 20$ mK one finds $n_1 \approx 0.23$. This value is huge, implying that passive relaxation methods cannot be used to reset fluxonium to $|0\rangle$ within any good approximation.

2.5.2. FEEDBACK

This method consists of measuring the qubit in the $\{|0\rangle, |1\rangle, |2\rangle, \ldots\}$ basis and applying a corrective gate depending on the measurement outcome. E.g., if one measures $|1\rangle$, then one applies an X gate (also called a π pulse) to bring the state back to $|0\rangle$. This method is useful for applications where one needs to measure a qubit mid-circuit and re-use it right away (furthermore, beyond reset, conditional feedback is useful to do a certain operation depending on the outcome, for example in quantum error correction). Conditional-feedback reset has been experimentally implemented for $|0\rangle$ and $|1\rangle$ [22, 23], but not $|2\rangle$. However, due to the challenges in the implementation, conditional feedback is not a common feature found in experimental settings at the moment. The issue is to process the measurement outcome in the classical electronics (e.g. via an FPGA [30]) and feed back the corrective operation in a sufficiently short amount of time ($\lesssim 500$ ns). In that way, the rest of the computation does not have to be delayed, with otherwise detrimental effects in terms of coherence.

In Chapter 9 we assume that one can apply conditional feedback to map a measured $|2\rangle$ to $|1\rangle$, thus constituting a leakage-reduction unit (π-LRU; see Section 9.3.2).

2.5.3. ALL-MICROWAVE RESET

References [24–26] introduced an unconditional, active reset scheme that resets both $|1\rangle$ and $|2\rangle$ to $|0\rangle$. In Chapter 9 we adapt this scheme to a leakage-reduction unit (res-LRU; see Section 9.2) and we discuss it in detail. Here we provide a short summary of the reset scheme as introduced in Refs. [24–26].

The fundamental elements are the readout resonator and one or two transmon drives. The resonator acts as an energy sink due to its designed strong coupling to the feedline environment, while the drives moves excitations from the transmon to the resonator. In one variation of the scheme [24, 25], a main, microwave drive on resonance with the $|20\rangle \leftrightarrow |01\rangle$ transition (notation: $|$transmon,resonator$\rangle$) is used to trade leakage on the transmon for a photon in the resonator. This drive induces oscillations between $|20\rangle$

and $|01\rangle$ and is calibrated to stop the first time that $|20\rangle$ is fully depleted. Then a second microwave drive on resonance with the transmon $|1\rangle \leftrightarrow |2\rangle$ transition is used to swap the population in $|1\rangle$ onto the depleted $|2\rangle$. Finally, the main drive pulse is repeated a second time, leaving the transmon ideally in $|0\rangle$. Note that technically the drive amplitude shifts the eigenfrequencies (see Eq. (9.62)), thus changing the resonance condition, such that the drive frequency has to be calibrated accordingly.

In the second variant [26], the two drives are applied at the same time, which allows depletion in a single step, although calibration is harder. Furthermore, in Ref. [26] the authors keep the drives on for a few oscillations, rather than stopping at the first minimum. This approach solves potential timing issues and is more resilient to crosstalk (see Section 9.6.3). A fidelity to $|0\rangle$ of 99% has been shown in 280 ns, and of 99.8% in 500 ns, compared to 98.3% in 210 ns in Ref. [25] (in the latter case, an overhead of 2 μs should actually be added to account for the slow decay rate of their resonator).

2.5.4. FLUX PULSE

An unconditional reset scheme that uses a flux pulse is given in Ref. [27]. The transmon is brought on resonance with its dedicated readout resonator to directly exploit the capacitive coupling between them. The pulse is realized with the "fast-adiabatic" technique [31]. The pulse parameters are chosen such that, not just the transmon population in $|1\rangle$ is transferred to the resonator, but also in $|2\rangle$ and even $|3\rangle$, passing through a cascade of avoided crossings (e.g. $|10\rangle \mapsto |01\rangle$ or $|20\rangle \mapsto |11\rangle \mapsto |02\rangle$, with the notation $|$transmon,resonator$\rangle$). After a carefully chosen time such that the exchange is completed, the transmon is moved below the resonator frequency, also in a fast-adiabatic fashion, allowing the resonator population to decay (so that e.g. $|01\rangle \to |00\rangle$ or $|02\rangle \to |00\rangle$). Then the transmon is brought back as fast as possible (< 2 ns) to its sweetspot frequency. Overall the protocol has been shown to reach $> 99\%$ fidelity to $|0\rangle$ in 250 ns, starting from either $|1\rangle$, $|2\rangle$ and $|3\rangle$. In particular, the best fidelity is 99.8% when starting from $|1\rangle$.

As the transmon sweeps a 2.5 GHz range in frequency, this method can be applied only in the presence of tunable couplers, otherwise the transmon frequency would very likely be on resonance with some transition on nearby transmons, even if briefly. Furthermore, as transmons are usually operated at their maximal-frequency sweetspot, this method requires that the chip is designed with all readout-resonator frequencies to be below the transmon frequencies in order to flux tune to resonant transitions (we note that this is the opposite of the current approach in the DiCarlo lab; see Section 2.2.2). This was not a requirement of the reset methods discussed above.

2.6. READOUT

Here we discuss the theory of dispersive readout, in particular how the resonator frequency depends on the transmon state. Then we briefly outline how the resonator frequency is measured in experiment and how the signals are treated to actually declare a measurement outcome.

2.6.1. DISPERSIVE READOUT

The most common way of measuring the state of a transmon is the so-called dispersive readout, via its dedicated readout resonator. The Hamiltonian of the coupled transmon-resonator system is given by

$$H = H^r + H^q + H^c \tag{2.37}$$

$$H^r = \omega^r a^\dagger a \tag{2.38}$$

$$H^q = \omega^q b^\dagger b + \frac{\alpha}{2}(b^\dagger)^2 b^2 \tag{2.39}$$

$$H^c = g(ab^\dagger + a^\dagger b), \tag{2.40}$$

where ω^r and ω^q are the resonator and transmon frequencies, respectively, α is the transmon anharmonicity, g corresponds to the capacitive coupling, a and b are the annihilation operators for the resonator and transmon, respectively. The transmon Hamiltonian H^q is derived within the approximations described in Section 2.2. An expression for the coupling g depending on the basic circuit parameters $(E_C, E_J, \ldots)$ is given in Eq. (2.67).

As in Eq. (2.12), one can rewrite H^q as

$$H^q = \sum_{j=0}^{+\infty} \omega_j^q \, |j\rangle\langle j|, \tag{2.41}$$

where $\omega_0^q = 0$, $\omega_1^q = \omega^q$, $\omega_2^q = 2\omega^q + \alpha$ and in general $\omega_j^q = j\omega^q + \alpha j(j-1)/2$. We can introduce the transition frequencies $\omega_{j,j+1}^q = \omega_{j+1}^q - \omega_j^q$ and the detunings

$$\Delta_j \coloneqq \omega_{j,j+1}^q - \omega^r. \tag{2.42}$$

We assume to be in the dispersive regime, which can be expressed as $g/\Delta \ll 1$ with $\Delta \coloneqq \Delta_0$, or more precisely as

$$\lambda_j \sqrt{n} \coloneqq \frac{g\sqrt{j+1}}{\Delta_j}\sqrt{n} \ll 1, \tag{2.43}$$

where n is the number of photons in the resonator (recall that $a = \sum_{n=1}^{+\infty} \sqrt{n}\,|n-1\rangle\langle n|$), and where we assume that this holds at least up to $j = 2$ and for low-enough n. One can then use a Schrieffer-Wolff transformation $e^S = \sum_{j,n=0}^{+\infty} |jn\rangle\langle jn|_D$ [32–35] (reviewed also in Section 9.5.1) to effectively capture the action of H^c, where $\{|jn\rangle\}$ are the eigenstates of $H^0 \coloneqq H^r + H^q$ (the "bare" basis) and $\{|jn\rangle\}_D$ are the eigenstates of $H^0 + H^c = H^r + H^q + H^c$ (the "dressed" basis). Specifically, here we use $S \approx S_1$, where S_1 is a first-order transformation in $\lambda_j\sqrt{n}$, for which $e^{S_1}|jn\rangle_D \approx |jn\rangle$. Here [7]

$$S_1 = \Big(\sum_{j=0}^{+\infty} \lambda_j (a\,|j+1\rangle\langle j|) - \text{h.c.}\Big), \tag{2.44}$$

for which

$$e^S H e^{-S} \approx e^{S_1} H e^{-S_1} \approx H + [S_1, H] \approx \tilde{H} \coloneqq \sum_{j=0}^{+\infty} \left(\omega_j^q + \chi_{j-1,j}\right) |j\rangle\langle j| + \left(\omega^r + \sum_{j=0}^{+\infty} \left(\chi_{j-1,j} - \chi_{j,j+1}\right) |j\rangle\langle j|\right) a^\dagger a, \quad (2.45)$$

where we have introduced the partial dispersive shifts

$$\chi_{j,j+1} \coloneqq \frac{g^2(j+1)}{\Delta_j} \quad (2.46)$$

and where we have neglected a double-excitation-exchange term since it is proportional to α, which is relatively low for transmons.

For the moment we restrict Eq. (2.45) to the first two levels of the transmon:

$$\begin{aligned} \tilde{H}|_{\mathscr{C}} &= \left(\omega^q + \chi_{0,1}\right) |1\rangle\langle 1| \\ &\quad + \left(\omega^r - \chi_{0,1} |0\rangle\langle 0| + \left(\chi_{0,1} - \chi_{1,2}\right) |1\rangle\langle 1|\right) a^\dagger a, \quad (2.47) \\ &= -\frac{\omega^q + \chi_{0,1}}{2} Z \\ &\quad + \left(\omega^r - \frac{\chi_{1,2}}{2} + \left(-\chi_{0,1} + \frac{\chi_{1,2}}{2}\right) Z\right) a^\dagger a, \quad (2.48) \end{aligned}$$

where we have used $|0\rangle\langle 0| = (I + Z)/2$ and $|1\rangle\langle 1| = (I - Z)/2$. We can introduce the shifted qubit frequency $\tilde{\omega}^q = \omega^q + \chi_{0,1}$, the shifted resonator frequency $\tilde{\omega}^r = \omega^r - \chi_{1,2}/2$ and the dispersive shift

$$\begin{aligned} \chi &= -\chi_{0,1} + \frac{\chi_{1,2}}{2} \quad (2.49) \\ &= \frac{g^2 \alpha}{\Delta(\Delta + \alpha)}. \quad (2.50) \end{aligned}$$

Then

$$\tilde{H}|_{\mathscr{C}} = -\frac{\tilde{\omega}^q}{2} Z + \left(\tilde{\omega}^r + \chi Z\right) a^\dagger a. \quad (2.51)$$

One can see that the dispersive shift is the amount by which $\tilde{\omega}^r$ is shifted in one direction or the other depending on the qubit being in $|0\rangle$ or $|1\rangle$ ($\tilde{\omega}^r \mapsto \tilde{\omega}^r \pm \chi$, respectively).

MEASUREMENT OF STATE $|2\rangle$

If one also wants to measure $|2\rangle$, it is important to know by how much $\tilde{\omega}^r$ is changed if the transmon is in $|2\rangle$. This quantity, which we call $\chi^{(2)}$, is not easily found in the literature to my knowledge. From Eq. (2.45) it can be computed to be:

$$\begin{aligned} \chi^{(2)} &= \omega^r + (\chi_{1,2} - \chi_{2,3}) - \tilde{\omega}^r = \frac{3\chi_{1,2}}{2} - \chi_{2,3} \quad (2.52) \\ &= \frac{3g^2\alpha}{(\Delta + \alpha)(\Delta + 2\alpha)}. \quad (2.53) \end{aligned}$$

The ratio between the dispersive shifts is then

$$\frac{\chi^{(2)}}{\chi} = \frac{3\Delta}{\Delta + 2\alpha}. \tag{2.54}$$

First, note that it is 1 for $\Delta = \alpha$, in which case it would not be possible at all to distinguish $|1\rangle$ and $|2\rangle$ in the measurement. Luckily, commonly $\alpha \sim -300$ MHz and $\Delta \sim -1$ GHz, for which $\chi^{(2)} \approx 2\chi$. If instead $\Delta \sim +1$ GHz, then $\chi^{(2)} \approx 7\chi$. These are just some examples but one can understand that $\chi^{(2)}$ can vary widely.

PURCELL

On the one hand, the coupling term H^c in Eq. (2.40) enables dispersive readout. On the other hand, the coupling opens a relaxation channel for the transmon via the resonator, known as the Purcell effect [36]. This is particularly troublesome because the relaxation rate κ of the resonator is large by design, whereas one strives to keep the relaxation rate $1/T_1$ of the transmon as small as possible. The rate κ is large because, as in Section 2.6.2, it sets the time for the probe pulse to be returned from the resonator. It cannot be too large either, because the interaction needs to last long enough to collect information about the transmon state. Indeed, maximal contrast is achieved when $\chi = \kappa/2$ [2], where in practice these parameters are on the order of a few MHz, allowing for a measurement time on the order of a few hundreds of nanoseconds.

A common solution to mitigate the Purcell effect is to use a Purcell filter [37, 38], which consists of another resonator placed between the feedline and the readout resonator itself. The Purcell filter is designed to alter the environmental impedance such that the coupling to the environment is suppressed at the transmon frequency, while it remains strong at the readout-resonator frequency.

2.6.2. READOUT OF A TRANSMON IN EXPERIMENT

MEASURING THE RESONATOR FREQUENCY

Thanks to the dispersive coupling (see Eq. (2.51)), one can measure the transmon by collecting information about the resonator frequency. Here we sketch the experimental measurement process [2, 3, 39]. One sends a signal that populates the resonator with photons and that gradually leaks back out. The state populating the resonator acquires a phase-shift which depends on the difference of the carrier frequency of the probe and the resonator frequency (which depends on the qubit state). This phase-shift is converted in the measurement of the outgoing amplified signal in the IQ plane, where the voltage quadratures are called in-phase (I) and quadrature (Q). The final location in the IQ plane thus provides information about the transmon state. To be more precise, the information about the transmon state leaks out in time from the readout resonator, hence the final voltage V in the IQ plane is obtained by integration with appropriate weights [40]. For each given state ($|0\rangle$, $|1\rangle$, $|2\rangle$...), one is not expected to measure always the same value of V because of fundamental Heisenberg uncertainty relations. Rather, the measured values of V generally follow a Gaussian distribution in the IQ plane.

MEASUREMENT CALIBRATION

The measurement is pre-calibrated to be able to associate to each value of V a likelihood of corresponding to a $|0\rangle$, $|1\rangle$ or $|2\rangle$ (or even more states in principle). Many instances of

each state are prepared and measured to identify where the Gaussian distributions $\mathcal{N}_j$ of each state are located (one can also observe relaxation and residual excitations in this calibration). Then one can divide the IQ plane into maximum-likelihood regions, one per each state. The procedure to associate a likelihood to each state is described in further detail in Section 8.11.1.

IMPORTANCE OF MEASURING $|2\rangle$

We note that $|2\rangle$ if often neglected in measurements as the qubit states are generally more relevant. However, as leakage can occur, measuring $|2\rangle$ is important to detect leakage and possibly apply a correction to bring it back to the computational subspace (see Chapters 8 and 9). Furthermore, one often does not record both quadratures but only a linear combination of them, corresponding to the axis in the IQ plane that passes through the centers of $\mathcal{N}_0$ and $\mathcal{N}_1$. In particular, one records the projection of V onto this axis. Then, optimal separation corresponds to separating the projected distributions by a threshold value placed in the middle. While this requires to save less data, in general the price is a partial loss of information, since $|2\rangle$ cannot be measured as accurately as it would from having data on the full IQ plane. It is still possible to extract information about $|2\rangle$ only if (luckily) the projection of $\mathcal{N}_2$ onto the combined quadrature does not significantly overlap with any of the other two.

AMPLIFICATION

The more the photons that populated the resonator, or the longer the probe pulse (with fewer photons), the more separated the Gaussian distributions $\{\mathcal{N}_j\}$ are in general, allowing for a measurement with higher signal-to-noise ratio. However, the dispersive regime is valid only until the photon number $\bar{n}$ is limited, i.e. until the approximation in Eq. (2.43) holds (also as long as other approximations leading to Eq. (2.51) hold, like the RWA). In particular, $\bar{n}$ should be smaller than the critical photon number $n_{\text{cr}} = \Delta^2/4g^2$ [7]. For current superconducting-qubit control stacks, n_{cr} is generally too low to be properly detected, so a long amplification chain to higher fridge stages is needed. Lots of work [2] has been dedicated to amplifiers to avoid the introduction of unnecessary noise, especially because phase-insensitive amplifiers inevitably introduce some noise due to the standard quantum limit. Finally, we note that from a fundamental-physics point of view, it is unclear where exactly in this chain the transmon "collapses" onto an eigenstate and is really measured.

MULTIPLEXED READOUT

In a chip with multiple qubits, using slightly different frequencies for the readout resonators allows simultaneous measurement via a single feedline [41]. This is because probe pulses at different frequencies can be mixed together. There are limits in bandwidth, though, as well as in how close the frequencies can be, so that, for example at the companies IBM and Rigetti, a good tradeoff seems to have 8 resonators per feedline.

2.7. SINGLE-QUBIT GATES

XY CONTROL

A microwave-drive line is generally used to perform single-qubit gates for superconducting qubits. This line is capacitively coupled to the qubit and a voltage $V_d(t)$ is applied to perform different gates. The drive Hamiltonian is $H_d = i\mathscr{E}(t)(b - b^\dagger)$, where $\mathscr{E}(t)$ is proportional to $V_d(t)$ and b is the annihilation operator, here for a transmon. It can be shown [2] that it is possible to implement arbitrary rotations around any axis in the equator of the Bloch sphere, depending on the choice of $V_d(t)$. In particular, let

$$V_d(t) = V_0 s(t)\Big(I\sin(\omega_d t) + Q\cos(\omega_d t)\Big) \tag{2.55}$$

with $I = \cos(\phi)$ and $Q = \sin(\phi)$, where ω_d and ϕ are the drive frequency and phase, respectively, V_0 is a constant, $s(t)$ is an envelope function, I and Q are called the in-phase and quadrature components of the voltage, respectively. The axis of rotation in the equator is determined by ϕ and the angle of rotation by the drive power together with the gate time, e.g., $\phi = 0$ gives rotations around X and $\phi = \pi/2$ around Y.

Z ROTATIONS

One can either implement Z rotations in a physical way, e.g. by fluxing a tunable transmon slightly away from its sweetspot, or in a "virtual" way. This is achieved by appropriately changing the drive phase [2] in the AWG (Arbitrary Waveform Generator) generating $V_d(t)$. The advantage of virtual Z rotations is that they cost almost no time and their fidelity is nominally unity as they are performed "in software" and merged with other pulses. While XY control alone allows to implement any single-qubit gate by applying multiple rotations in sequence, combining XY control with (virtual) Z rotations allows to reduce the number of operations.

DRAG TO MITIGATE LEAKAGE AND PHASE ERRORS

As transmons have a relatively small anharmonicity α, $V_d(t)$ can contain components that are on resonance with the $|1\rangle \leftrightarrow |2\rangle$ transition and not just $|0\rangle \leftrightarrow |1\rangle$, due to the smoothing envelope $s(t)$. This can cause leakage to $|2\rangle$, as well as (coherent) phase errors due to the repulsion of $|1\rangle$ and $|2\rangle$ in the presence of the drive. The DRAG technique (Derivative Reduction by Adiabatic Gate) [42] allows to solve these issues. In its original implementations it allows one to reduce either leakage or phase errors. However, extensions of the DRAG technique allow one to reduce both errors at the same time [2]. The overall result is that single-qubit gates routinely reach average gate fidelities $\gtrsim 99.8\%$ [43].

2.8. TWO-QUBIT GATES

We discuss the approximations underlying the Hamiltonian of two transmons coupled via a bus resonator (the kind of system considered in the experiments and numerical simulations described in Chapter 6). We then discuss the energy spectrum of the effective two-transmon system, focusing on the avoided crossings as a function of flux, which can be used to perform a two-qubit gate. Finally, we discuss the three major ways of implementing a two-qubit gate with transmons: baseband flux pulsing (used in Chapter 6), parametric driving and the cross-resonance gate.

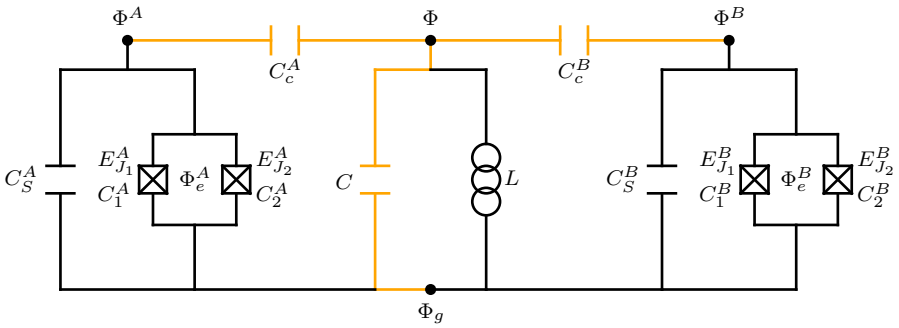

Figure 2.5: The circuit for two tunable transmons (see Section 2.2.1) coupled via a bus resonator. We introduce node fluxes Φ^A, Φ^B, Φ and ground $\Phi_g \equiv 0$. The external magnetic fluxes are Φ_e^A and Φ_e^B. In orange our choice for the spanning tree of the circuit.

2.8.1. COUPLED TRANSMONS

We apply the rules of circuit quantization (see Section 2.4) to the circuit in Fig. 2.5. Each SQUID ($X = A$ or B) has two Josephson junctions ($j = 1$ or 2) characterized by $E_{J_j}^X$, two intrinsic capacitances C_j^X and a shunting capacitance C_S^X. Let $C_\Sigma^X = C_1^X + C_2^X + C_S^X$. Furthermore, each loop is pierced by an external magnetic flux Φ_e^X and is capacitively connected to the resonator by a capacitance C_c^X. The resonator in the middle has capacitance C and inductance L. Considering the spanning tree in Fig. 2.5, the Lagrangian of the circuit is

$$\begin{aligned}\mathscr{L} = &\sum_{X=A,B} \frac{1}{2} C_\Sigma^X \left(\dot{\Phi}^X\right)^2 + E_{J_1}^X \cos(\phi^X - \phi_e^X) + E_{J_2}^X \cos\phi^X \\ &+ \frac{1}{2} C \dot{\Phi}^2 - \frac{\Phi^2}{2L} \\ &+ \sum_{X=A,B} \frac{1}{2} C_c^X \left(\dot{\Phi}^X - \dot{\Phi}\right)^2, \end{aligned} \tag{2.56}$$

where $\phi^X = 2\pi\Phi^X/\Phi_0$ and $\phi_e^X = 2\pi\Phi_e^X/\Phi_0$ with Φ_0 being the flux quantum, and where we have assumed $\dot{\Phi}_e^X = 0$. The kinetic terms (those containing a time derivative) can be rewritten in terms of a capacitive matrix $\mathbf{C}$ (see Eq. (2.31)). To move to the Hamiltonian formalism, one needs to invert $\mathbf{C}$. Assuming that $C_c^X \ll \min\{C_\Sigma^X, C\}$, we approximate $\mathbf{C}^{-1}$ up to second-order in C_c^A, C_c^B, leading to the Hamiltonian

$$\begin{aligned}\mathscr{H} \approx &\sum_{X=A,B} \frac{1}{2\tilde{C}_\Sigma^X} (Q^X)^2 - E_{J_1}^X \cos\left(\phi^X - \phi_e^X\right) - E_{J_2}^X \cos\phi^X \\ &+ \frac{1}{2\tilde{C}} Q^2 + \frac{\Phi^2}{2L} \\ &+ \sum_{X=A,B} \frac{1}{\tilde{C}_c^X} Q^X Q + \frac{1}{\tilde{C}_J} Q^A Q^B, \end{aligned} \tag{2.57}$$

where the momenta are defined as $Q^X := \frac{\partial \mathscr{L}_1}{\partial \dot{\Phi}^X}$ and $Q := \frac{\partial \mathscr{L}_1}{\partial \dot{\Phi}}$, and where

$$\frac{1}{\tilde{C}_\Sigma^X} = \frac{1}{C_\Sigma^X} - \frac{C_c^X}{(C_\Sigma^X)^2} + \left(\frac{1}{(C_\Sigma^X)^3} + \frac{1}{C(C_\Sigma^X)^2}\right)(C_c^X)^2 \tag{2.58}$$

$$\frac{1}{\tilde{C}} = \frac{1}{C} - \frac{C_c^A + C_c^B}{C^2} + \frac{1}{C}\left(\frac{(C_c^A)^2}{(C_\Sigma^A)^2} + \frac{(C_c^B)^2}{(C_\Sigma^B)^2}\right) + \frac{(C_c^A + C_c^B)^2}{C^3} \tag{2.59}$$

$$\frac{1}{\tilde{C}_c^X} = \frac{C_c^X}{CC_\Sigma^X} - \left(\frac{1}{C(C_\Sigma^X)^2} + \frac{1}{C^2 C_\Sigma^X}\right)(C_c^X)^2 - \frac{C_c^A C_c^B}{C^2 C_\Sigma^X} \tag{2.60}$$

$$\frac{1}{\tilde{C}_J} = \frac{C_c^A C_c^B}{CC_\Sigma^A C_\Sigma^B}. \tag{2.61}$$

We note that the term $Q^A Q^B/\tilde{C}_J$ in Eq. (2.57) is already a direct exchange term between the two transmons, even before considering the dispersive regime (see after Eq. (2.70)). This term can easily be forgotten if one tries to directly write the Hamiltonian of the circuit in Fig. 2.5, instead of starting with the Lagrangian as prescribed by circuit quantization (see Section 2.4). The effect of this term is to alter the expression for the effective coupling between the two transmons and, for example, it plays a role in the design of tunable couplers [44].

We proceed with a few manipulations of $\mathscr{H}$ in Eq. (2.57). For the resonator one can introduce the standard harmonic-oscillator operators as

$$Q = i\sqrt{\frac{\tilde{C}\omega_r}{2}}(a^\dagger - a) \tag{2.62}$$

$$\Phi = \frac{1}{\sqrt{2\tilde{C}\omega_r}}(a^\dagger + a), \tag{2.63}$$

where $\omega_r = 1/\sqrt{L\tilde{C}}$ is the resonator frequency (note that it depends on the renormalized capacitance $\tilde{C}$ and not C). Then the resonator free Hamiltonian is simply $\omega_r(a^\dagger a + 1/2)$, where the $1/2$ term is ignored in the following since it is a constant energy shift.
Regarding the SQUIDs, as described in Eq. (2.19), the cosine terms can be combined as $E_{J_\Sigma}^X(\phi_e^X)\cos\phi^X$, where we make the dependence of $E_{J_\Sigma}^X$ on ϕ_e^X explicit. Then, as in Eqs. (2.6) and (2.7), in the transmon regime we can introduce similar operators as for the resonator:

$$N^X = \frac{i}{\sqrt{2}}\left(\frac{E_{J_\Sigma}^X(\phi_e^X)}{8E_{C_\Sigma}^X}\right)^{1/4}(b_X^\dagger - b_X) \tag{2.64}$$

$$\phi^X = \frac{1}{\sqrt{2}}\left(\frac{E_{J_\Sigma}^X(\phi_e^X)}{8E_{C_\Sigma}^X}\right)^{-1/4}(b_X^\dagger + b_X), \tag{2.65}$$

where $N^X = Q^X/2e$ and $E_{C_\Sigma}^X = e^2/(2\tilde{C}_\Sigma^X)$. In the same approximations as leading to Eq. (2.9) (cosine up to 4th order and rotating-wave approximation), the free transmon Hamiltonian is $\omega^X(\phi_e^X)\, b_X^\dagger b_X + \frac{\alpha^X}{2}(b_X^\dagger)^2 b_X^2$, with transmon frequency $\omega^X(\phi_e^X) = \sqrt{8E_{J_\Sigma}^X(\phi_e^X)E_{C_\Sigma}^X} - E_{C_\Sigma}^X$ and anharmonicity $\alpha^X = -E_{C_\Sigma}^X$.

Combining the results above, one gets

$$\begin{aligned}\mathscr{H} \approx & \sum_{X=A,B} \omega^X(\phi_e^X)\, b_X^\dagger b_X + \frac{\alpha^X}{2}(b_X^\dagger)^2 b_X^2 \\ & + \omega_r a^\dagger a \\ & - \sum_{X=A,B} g^X(\phi_e^X)\left(b_X^\dagger - b_X\right)\left(a^\dagger - a\right) \\ & - \tilde{J}^{AB}(\phi_e^A,\phi_e^B)\left(b_A^\dagger - b_A\right)\left(b_B^\dagger - b_B\right),\end{aligned} \tag{2.66}$$

where

$$g^X(\phi_e^X) := (\tilde{C}_c^X)^{-1} e\sqrt{\tilde{C}\omega_r}\left(\frac{E_{J_\Sigma}^X(\phi_e^X)}{8E_{C_\Sigma}^X}\right)^{1/4}, \tag{2.67}$$

$$\tilde{J}^{AB}(\phi_e^A,\phi_e^B) := (2e)^2 \frac{1}{2\tilde{C}_J}\left(\frac{E_{J_\Sigma}^X(\phi_e^A)}{8E_{C_\Sigma}^A}\right)^{1/4}\left(\frac{E_{J_\Sigma}^B(\phi_e^B)}{8E_{C_\Sigma}^B}\right)^{1/4}. \tag{2.68}$$

We now perform the rotating-wave approximation for the energy non-conserving terms in Eq. (2.66), i.e. we neglect those terms containing two creation or two annihilation operators, like $b_X^\dagger a^\dagger$ and $b_A b_B$. We get

$$\begin{aligned}\mathscr{H} \approx & \sum_{X=A,B} \omega^X(\phi_e^X)\, b_X^\dagger b_X + \frac{\alpha^X}{2}(b_X^\dagger)^2 b_X^2 \\ & + \omega_r a^\dagger a \\ & + \sum_{X=A,B} g^X(\phi_e^X)\left(b_X^\dagger a + b_X a^\dagger\right) + \tilde{J}^{AB}(\phi_e^A,\phi_e^B)\left(b_A^\dagger b_B + b_A b_B^\dagger\right).\end{aligned} \tag{2.69}$$

The free transmon Hamiltonian (first line in Eq. (2.69)) can be rewritten as $\sum_{j=0}^{+\infty}\omega_j^X |j\rangle\langle j|_X$, similarly to Eq. (2.12), where $\{|j\rangle_X\}$ are transmon eigenstates and $\{\omega_j^X\}$ the eigenfrequencies.

Similarly to the discussion after Eq. (2.43) in Section 2.6.1 about dispersive readout, we assume to be in the dispersive regime, i.e. that the parameters

$$\lambda_j^X\sqrt{n} := \frac{g^X\sqrt{j+1}}{\Delta_j^X}\sqrt{n}, \tag{2.70}$$

satisfy $\lambda_j^X\sqrt{n} \ll 1$ (at least up to $j=2$ and for low n), where $\Delta_j^X = \omega_{j,j+1}^X - \omega_r$ and $\omega_{j,j+1}^X$ is the transition frequency from transmon level $|j\rangle_X$ to $|j+1\rangle_X$. Then we apply a Schrieffer-Wolff transformation e^S [32–35] (reviewed also in Section 9.5.1) to capture effectively the action of $\sum_{X=A,B} g^X(\phi_e^X)\left(b_X^\dagger a + b_X a^\dagger\right)$ in Eq. (2.69). Specifically, here we use $S \approx S_1$, where S_1 is a first-order transformation in $\lambda_j\sqrt{n}$. Here [7]

$$S_1 = \left(\sum_{X=A,B}\sum_{j=0}^{+\infty}\lambda_j^X\left(a|j+1\rangle\langle j|_X\right) - \text{h.c.}\right). \tag{2.71}$$

Then effectively

$$e^{S}\mathcal{H}e^{-S} \approx e^{S_1}\mathcal{H}e^{-S_1} \approx \mathcal{H} + [S_1, \mathcal{H}] \tag{2.72}$$

$$\begin{aligned}\approx \tilde{\mathcal{H}} := &\sum_{X=A,B}\sum_{j=0}^{+\infty}\left(\omega_j^X(\phi_e^X) + \chi_{j-1,j}^X\right)|j\rangle\langle j|_X \\ &+\left(\omega_r + \sum_{X=A,B}\sum_{j=0}^{+\infty}\left(\chi_{j-1,j}^X - \chi_{j,j+1}^X\right)|j\rangle\langle j|_X\right)a^\dagger a \\ &+\sum_{jk=0}^{+\infty} J_{jk}^{AB}(\phi_e^A,\phi_e^B)\sqrt{j+1}\sqrt{k+1}\Big(|j+1,k\rangle\langle j,k+1|_{AB} + \text{h.c.}\Big),\end{aligned} \tag{2.73}$$

where

$$J_{jk}^{AB}(\phi_e^A,\phi_e^B) := \tilde{J}^{AB}(\phi_e^A,\phi_e^B) + \frac{g^A(\phi_e^A)\,g^B(\phi_e^B)}{2}\Big(\frac{1}{\Delta_j^A} + \frac{1}{\Delta_j^B}\Big) \tag{2.74}$$

and where we have omitted the explicit dependence of Δ_j^X and of the partial dispersive shifts $\chi_{j,j+1}^X := \frac{(g^X)^2(j+1)}{\Delta_j}$ on ϕ_e^X.

Under the assumption that the resonator is in the ground state $|0\rangle$, since $a^\dagger a|0\rangle = 0$ one finally gets

$$\begin{aligned}\tilde{\mathcal{H}} \approx &\sum_{X=A,B}\sum_{j=0}^{+\infty}\tilde{\omega}_j^X(\phi_e^X)|j\rangle\langle j|_X \\ &+\sum_{jk=0}^{+\infty} J_{jk}^{AB}(\phi_e^A,\phi_e^B)\sqrt{j+1}\sqrt{k+1}\Big(|j+1,k\rangle\langle j,k+1|_{AB} + \text{h.c.}\Big),\end{aligned} \tag{2.75}$$

where we introduce $\tilde{\omega}_j^X(\phi_e^X) = \omega_j^X(\phi_e^X) + \chi_{j-1,j}^X$. Beside a rescaled transmon frequency $\tilde{\omega}^X(\phi_e^X) = \tilde{\omega}_1^X(\phi_e^X)$, we can also introduce a rescaled anharmonicity $\tilde{\alpha}^X = \alpha^X + \chi_{1,2}^X - 2\chi_{0,1}^X$ and rewrite Eq. (2.75) (exactly up to $j = 2$) as

$$\begin{aligned}\tilde{\mathcal{H}} = &\sum_{X=A,B}\tilde{\omega}^X(\phi_e^X)\,b_X^\dagger b_X + \frac{\tilde{\alpha}(\phi_e^X)}{2}(b_X^\dagger)^2 b_X^2 \\ &+\sum_{jk=0}^{+\infty} J_{jk}^{AB}(\phi_e^A,\phi_e^B)\sqrt{j+1}\sqrt{k+1}\Big(|j+1,k\rangle\langle j,k+1|_{AB} + \text{h.c.}\Big),\end{aligned} \tag{2.76}$$

where we make the dependence of $\tilde{\alpha}^X$ on ϕ_e^X explicit (via $\chi_{j,j+1}^X$).

Equation (2.76) is often presented as

$$\tilde{\mathcal{H}} \approx \sum_{X=A,B}\tilde{\omega}^X(\phi_e^X)\,b_X^\dagger b_X + \frac{\tilde{\alpha}}{2}(b_X^\dagger)^2 b_X^2 + J^{AB}(b_A b_B^\dagger + b_A^\dagger b_B) \tag{2.77}$$

where $J^{AB} = \frac{g^A g^B}{2}(\frac{1}{\Delta^A} + \frac{1}{\Delta^B})$. In other words, beyond Eq. (2.76) one performs the further approximations that 1) the dependence of J_{jk}^{AB} on ϕ_e^X can be neglected; 2) $\Delta_j^X \approx \Delta_0^X \equiv$

$\Delta^X = \omega^X - \omega_r$, leading to J^{AB}_{jk} being independent of j,k, specifically $J^{AB}_{jk} = J^{AB}_{00} \equiv J^{AB}$; 3) $\tilde{J}^{AB}$ in Eq. (2.68) can be neglected; 4) the dependence of $\tilde{\alpha}$ on ϕ_e^X can be neglected. We note that approximation 2) is generally good for transmons since the anharmonicity is small (this approximation is exact if $\alpha = 0$).
Within these approximations one has

$$J^{AB}(b_A b_B^\dagger + b_A^\dagger b_B) = J^{AB}(|01\rangle\langle 10| + \text{h.c.}) + \sqrt{2}J^{AB}(|11\rangle\langle 02| + \text{h.c.}) + \ldots, \tag{2.78}$$

i.e. the couplings in the one- and two-excitation manifolds are expected to differ by a $\sqrt{2}$-factor. However, one is generally aware that the prediction of this $\sqrt{2}$ factor is only approximate. As a consequence, in experiments these couplings, as well as frequencies and anharmonicities, are directly measured (at the ϕ_e^X of interest) to avoid mismatches with the approximations used in the theory.
We remark that in the numerical simulations in Chapter 6 (see around Eq. (6.11)), apart for all other approximations in this section, we also consider approximations 2), 3), 4) and partially 1), meaning that we take into account that Δ^X, but not g^X, varies with ϕ_e^X. Nevertheless, we still find a very good match with the experimental results (see Fig. 6.4 in particular). We attribute this precisely to the fact that in the simulations we used the parameters measured in experiment at the flux values that are relevant for the two-qubit gate.

While measurements of parameters in experiment can provide direct insight into the full Hamiltonian, the first-principle analysis performed in this section shows that one should at least not take for granted certain relationships between parameters that are predicted by approximate formulas, like the $\sqrt{2}$ factor in Eq. (2.78). Furthermore, for g^X and J^{AB}_{jk} we have made explicit the dependence on ϕ_e as well as on other circuit parameters (see Eq. (2.67) and Eq. (2.74), respectively). In particular, we highlight the dependence of g^X on ω_r. This fact has been exploited in a recent update of the surface-code chips in the DiCarlo lab: bus resonators used to have a target frequency of $\omega_r/2\pi = 8.5$ GHz, but this was changed to 20 GHz. Naively, based on Eq. (2.74), one only expects J^{AB}_{jk} to decrease since Δ^X_j increases. To keep the same two-qubit gate speed (i.e. the same value for the coupling) one would need to increase the coupling capacitances C_c^X, which are already relatively large in current devices. However, the increase in g^X with ω_r helped to keep the same two-qubit-gate speed without having to significantly increase the coupling capacitances to compensate.

2.8.2. AVOIDED CROSSINGS

We now discuss the spectrum of two coupled transmons as a function of the applied flux (see Fig. 2.6). We consider the coupling to be mediated by a bus resonator, as in Section 2.8.1, but a direct capacitive coupling would give a similar spectrum as well. In particular, we consider Eq. (2.77) for simplicity. We assume that only one transmon is fluxed, specifically the one with higher frequency (here we assume it is transmon B), whereas the other one stays at its sweetspot ($\phi_e^A = 0$).

As discussed after Eq. (2.17), the frequency of a tunable transmon decreases when ϕ_e^B increases from 0 to π. If there would be no coupling, the energy levels would simply cross at those points where $\omega_1^B(\phi_e^B \equiv \phi_e^{i\text{SWAP}}) = \omega_1^A$ and $\omega_2^B(\phi_e^B \equiv \phi_e^{\text{CZ}}) = \omega_1^A + \omega_1^B(\phi_e^B \equiv \phi_e^{\text{CZ}})$,

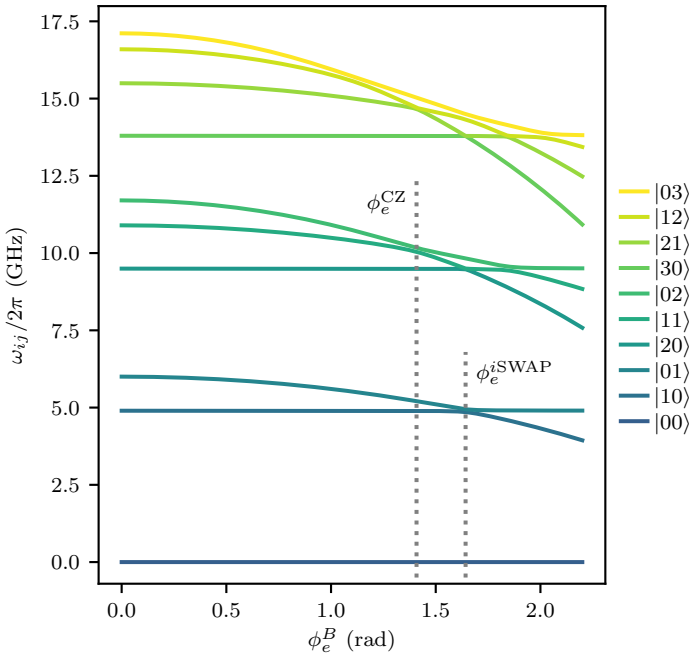

Figure 2.6: Spectrum of two coupled transmons. We consider Eq. (2.77) with $\tilde{\omega}^A(0)/2\pi = 4.9$ GHz, $\tilde{\omega}^B(0)/2\pi = 6$ GHz, $\tilde{\alpha}^A/2\pi = \tilde{\alpha}^B/2\pi = -300$ MHz and $J^{AB}/2\pi = 45$ MHz. We plot the frequency ω_{ij} of level $|ij\rangle$ as a function of the external flux ϕ_e^B, while $\phi_e^A \equiv 0$. The interaction points to perform CZ or iSWAP are marked by vertical dotted lines.

among others (see also Fig. 8.1). The latter condition for the CZ gate can also be rewritten as $\omega_1^B(\phi_e^B \equiv \phi_e^{\mathrm{CZ}}) = \omega_1^A - \alpha^B$. The values $\phi_e^{i\mathrm{SWAP}}$ and ϕ_e^{CZ} are called the interaction points for the respective gates. In other words, these points correspond to the center of the so-called avoided crossings between levels $|01\rangle$ and $|10\rangle$ and between $|02\rangle$ and $|11\rangle$, respectively. The presence of a non-zero coupling makes these levels interact and opens a gap equal to twice the coupling. To see this, consider Eq. (2.77) restricted to e.g. the subspace $\mathscr{S} = \mathrm{span}\{|11\rangle, |02\rangle\}$:

$$\tilde{\mathscr{H}}|_{\mathscr{S}} = \begin{pmatrix} \omega_{11}(\phi_e^B) & \sqrt{2}J^{AB} \\ \sqrt{2}J^{AB} & \omega_{02}(\phi_e^B). \end{pmatrix} \tag{2.79}$$

At $\phi_e^B = \phi_e^{\mathrm{CZ}}$, by definition one has $\omega_{11}(\phi_e^{\mathrm{CZ}}) = \omega_{02}(\phi_e^{\mathrm{CZ}})$, which we simply denote as ω^{CZ}. The eigenvalues are $\omega^{\mathrm{CZ}} - \sqrt{2}J^{AB}$ and $\omega^{\mathrm{CZ}} + \sqrt{2}J^{AB}$ and the corresponding eigenvectors are $\overline{|11\rangle} := (|11\rangle + |02\rangle)/\sqrt{2}$ and $\overline{|02\rangle} := (|11\rangle - |02\rangle)/\sqrt{2}$. The difference between the eigenvalues is thus $2(\sqrt{2}J^{AB})$. This "level repulsion" is stronger at the center of the avoided crossing (at ϕ_e^{CZ} or $\phi_e^{i\mathrm{SWAP}}$) and it becomes weaker away from it, although it is never really 0 (this is at the origin of the residual ZZ crosstalk discussed in Section 3.2.7).

Since J^{AB} induces two-qubit interactions, these avoided crossings can be exploited in various ways to implement two-qubit gates between transmons. In the following section we describe the CZ in detail and we briefly mention the iSWAP.

2.8.3. THREE ALTERNATIVE METHODS TO IMPLEMENT THE CZ GATE

In this section we introduce baseband flux pulsing (used in Chapter 6) in a detailed way, whereas we discuss a few specific points about parametric driving and the cross-resonance gate.

A controlled-phase or control-Z or CZ gate is defined in the computational subspace as

$$\mathrm{CZ} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{2.80}$$

i.e. $|11\rangle$ acquires a so-called conditional phase $\phi_{2\mathrm{Q}} = \pi$ ($e^{i\pi} = -1$), while the other computational states remain unchanged. In general, a gate of the form

$$U = \begin{pmatrix} e^{i\phi_{00}} & 0 & 0 & 0 \\ 0 & e^{i\phi_{01}} & 0 & 0 \\ 0 & 0 & e^{i\phi_{10}} & 0 \\ 0 & 0 & 0 & e^{i\phi_{11}} \end{pmatrix} \tag{2.81}$$

can be brought to the form

$$U' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{i\phi_{2\mathrm{Q}}} \end{pmatrix} \tag{2.82}$$

via single-qubit Z rotations (modulo a global phase), where

$$\phi_{2Q} = \phi_{11} - \phi_{01} - \phi_{10} + \phi_{00}. \tag{2.83}$$

This shows that ϕ_{2Q} is the relevant quantity that characterizes the CZ. Then U' is a generalized CZ with arbitrary conditional phase. If one deals with qutrits and not qubits, one can exploit higher excited states, in particular $|02\rangle$ and its avoided crossing with $|11\rangle$ (see Section 2.8.2), to implement a CZ in the computational subspace.

ALTERNATIVE 1: BASEBAND FLUX-BASED CZ

Here one uses a relatively slow ("baseband") flux pulse to tune the flux to ϕ_e^{CZ}, i.e. to the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing, the so-called interaction point for CZ (see Section 2.8.2). There are two "extreme" variants to implement a CZ using baseband flux. In the first, adiabatic, variant the flux is tuned very slowly such that $|11\rangle$, which is the eigenstate of $\tilde{\mathscr{H}}$ (see Eq. (2.79)) at $\phi_e^B = 0$, can evolve into $\overline{|11\rangle} = (|11\rangle + |02\rangle)/\sqrt{2}$ at the avoided crossing, which is the instantaneous eigenstate at that point (see after Eq. (2.79)). The coupling pushes down the frequency of $\overline{|11\rangle}$ by an amount $\sqrt{2}J^{AB}$ (see after Eq. (2.79)) compared to the case with no coupling (while $|01\rangle$ and $|10\rangle$ are approximately unchanged). It follows that $\overline{|11\rangle}$ acquires a conditional phase equal to $J^{AB}\,t$ (or more precisely $J_{01}^{AB}\,t$; see Eq. (2.74)), where t is the time spent at the avoided crossing. Then $\overline{|11\rangle}$ is adiabatically brought back to $|11\rangle$ by slowly turning off the applied flux. More precisely, since, as mentioned in Section 2.8.2, the coupling does have an effect not only at the avoided crossing, one chooses a pulse with total duration T such that the conditional phase

$$\phi_{2Q} \coloneqq \int_0^T dt\, \zeta\big(\phi_e(t)\big) \tag{2.84}$$

is $\phi_{2Q} = \pi$, where

$$\zeta\big(\phi_e(t)\big) = \omega_{11}\big(\phi_e(t)\big) - \omega_{01}\big(\phi_e(t)\big) - \omega_{10}\big(\phi_e(t)\big) + \omega_{00}\big(\phi_e(t)\big) \tag{2.85}$$

is the instantaneous ZZ-coupling strength. We note that $\zeta(0)$ quantifies the residual ZZ coupling discussed in Section 3.2.7. The quantities $\int_0^T dt\,\omega_{01}(t) = \phi_{01}$ and $\int_0^T dt\,\omega_{10}(t) = \phi_{10}$ are single-qubit phases that can easily be removed in practice (see Section 2.7) to ideally give a CZ as defined in Eq. (2.80).

The second "extreme" baseband variant is fully diabatic, i.e. the flux is turned on as fast as possible from 0 to ϕ_e^{CZ}, without letting $|11\rangle$ track the instantaneous eigenstate of the Hamiltonian. Then $|11\rangle$ rotates in the subspace spanned by $\overline{|11\rangle}$ and $\overline{|02\rangle}$ (defined after Eq. (2.79)). Specifically, at half the evolution, $|11\rangle$ fully transforms into $|02\rangle$ (i.e. intermediate leakage is maximal) and then goes back to $-|11\rangle$ (in the rotating frame of the qubits for this explanation), as desired for CZ. The minus sign is due to the fact that a 2π rotation is equal to $-I$ for a two-level system. Finally, the flux is turned off again as fast as possible. In the lab frame, single-qubit phases are also acquired in the process but they can be removed as for the adiabatic approach. We use this diabatic approach in the Sudden Net Zero CZ gate introduced in Section 6.12.

There are also other baseband variants [31, 45] to do the CZ that fall in-between these two "extreme" approaches. In particular, the "fast-adiabatic" approach [31] tries

to combine adiabaticity and speed by optimizing the pulse shape such that it is fast (although not as fast as in the fully diabatic case) but without letting $|11\rangle$ to significantly leak to $|02\rangle$. The pulse shape is discussed in Section 6.11.2. We use this approach in the original Net Zero CZ gate (see Section 6.1). In that case we actually use such a high speed that the intermediate leakage is significant (similarly to a diabatic gate), but leakage interference enables to have low leakage overall. Finally, for all variants discussed above, in general one does not need to reach exactly the point ϕ_e^{CZ}, but one can either undershoot or overshoot by some amount. In that case, the effective coupling is less strong, leading to a somewhat slower CZ.

iSWAP. We note that the iSWAP gate can be implemented in a similarly diabatic way (or high-speed fast-adiabatic way) but using the $|01\rangle \leftrightarrow |10\rangle$ avoided crossing instead of $|11\rangle \leftrightarrow |02\rangle$. Modulo single-qubit phases, the rotation of one state into the other leads to the generalized $i\mathrm{SWAP}_\theta$, with

$$i\mathrm{SWAP}_\theta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta) & -i\sin(\theta) & 0 \\ 0 & -i\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.86}$$

where θ is proportional to the time spent at (or close to) the avoided crossing. In particular, $\theta = \pi/2$ gives the canonical iSWAP. We note that, since $\phi_e^{\mathrm{CZ}} < \phi_e^{i\mathrm{SWAP}}$ (see Fig. 2.6), one crosses the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing on the way to $|01\rangle \leftrightarrow |10\rangle$, although rather quickly. Still, this gives rise to the issue that iSWAP gates might produce a certain amount of conditional phase [46], although techniques with tunable couplers have been developed to solve this issue [47].

ALTERNATIVE 2: PARAMETRIC DRIVING

This alternative also uses flux to implement a CZ, but it is conceptually different. In alternative 1, a "baseband" pulse is used, i.e. there are only relatively low-frequency components since in general the pulse goes to the avoided crossing and back just once. Instead, in parametric driving [48, 49] the flux is modulated at a relatively high frequency ($\omega_m \sim \mathcal{O}(200)$ MHz [50]). Specifically, $\phi_e^B(t) = \Omega_m \cos(\omega_m t + \theta_m)$, where Ω_m, ω_m and θ_m are the modulation amplitude, frequency and phase, respectively. Choosing the right Ω_m and ω_m [50] allows to effectively activate the $|11\rangle \leftrightarrow |02\rangle$ interaction to implement a CZ. However, the effective coupling is significantly smaller than the bare one $\sqrt{2}J^{AB}$, leading to longer gate times compared to baseband pulses. The technical advantage of parametric driving is that the pulse is inherently robust to long-timescale distortions (see Section 3.2.5) since it does not contain low-frequency components.

ALTERNATIVE 3: CROSS-RESONANCE GATE

The cross-resonance gate [51, 52] does not use flux for its implementation, but it is a purely microwave approach. It does not exploit the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing either, thus it can be applied to strictly two-level systems, as well as anharmonic oscillators such as transmons. The concept of the gate is to drive one qubit at the frequency of the other, where the drive Hamiltonian is $H_d^B = \Omega(t)\left(b_B e^{-i\omega^A t} + b_B^\dagger e^{i\omega^A t}\right)$, with $\Omega(t)$ the drive amplitude, and where we assume that e.g. qubit B is being driven. In the presence of a

flip-flop coupling J^{AB} like in Eq. (2.77), the result is that one induces Rabi oscillations of the undriven qubit with a frequency that depends on the state of the driven one. The effective coupling thus is a ZX term. Such a term does not give rise directly to either a CZ or a CNOT (control-X), but it can easily be transformed into either one via single-qubit gates. A complication is that spurious terms are produced [34], especially IX and ZI terms, apart for the desired ZX term. However, this issue has been addressed by using a pulse scheme with echoing pulses designed to cancel the spurious terms [53, 54].

Since a flux line is not required, an advantage of the cross-resonance gate is that it is compatible with fixed-frequency transmons, which have longer coherence times compared to tunable ones, thanks to not being sensitive to flux noise. Besides, the flux line is not required at all if one uses fixed-frequency transmons, allowing to remove this component from the chip and to reduce the number of external control lines per qubit. Of course the downside is to lose all the pros of tunability (see Section 2.2.1). Furthermore, the effective ZX coupling is generally small, even when strong drives are used. This means that cross-resonance gates usually take more time than baseband-flux approaches, partially undermining the benefit of longer coherence times, although there seems to be potential for improvement [55]. Furthermore, the strong drive can enable unwanted exchanges between other levels of the same transmons or of nearby transmons, especially since the anharmonicity is relatively low. This can lead to an increased error rate or leakage [56], requiring careful design of the chip (as well as good frequency targeting in fabrication [57]) to avoid frequency collisions. This has led IBM to consider an architecture where qubits have at most three neighbors, and to develop custom quantum error correcting codes beyond the surface code [58] (see Chapter 4 for quantum error correcting codes).

REFERENCES

[1] G. Wendin and V. S. Shumeiko, *Superconducting Quantum Circuits, Qubits and Computing,* (2005), arXiv:cond-mat/0508729 [cond-mat.supr-con] .

[2] P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *A quantum engineer's guide to superconducting qubits,* Applied Physics Reviews **6**, 021318 (2019).

[3] A. Blais, A. L. Grimsmo, S. Girvin, and A. Wallraff, *Circuit quantum electrodynamics,* Reviews of Modern Physics **93** (2021), 10.1103/revmodphys.93.025005.

[4] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, *Superconducting qubits: Current state of play,* Annual Review of Condensed Matter Physics **11**, 369 (2020).

[5] J. M. Martinis and K. Osborn, *Superconducting qubits and the physics of Josephson junctions,* (2007).

[6] U. Vool and M. Devoret, *Introduction to quantum electromagnetic circuits,* International Journal of Circuit Theory and Applications **45**, 897 (2017).

[7] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, J. Majer, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Charge-insensitive qubit design derived from the Cooper pair box,* Phys. Rev. A **76**, 042319 (2007).

[8] O. Hefti, *CPHASE analysis in the eigenmode approach in presence of ZZ crosstalk,* Master Thesis, TU Delft (2020).

[9] N. K. Langford, *Circuit QED - Lecture notes,* (2013), arXiv:1310.1897 [quant-ph] .

[10] V. Negîrneac, H. Ali, N. Muthusubramanian, F. Battistel, R. Sagastizabal, M. S. Moreira, J. F. Marques, W. J. Vlothuizen, M. Beekman, C. Zachariadis, N. Haider, A. Bruno, and L. DiCarlo, *High-fidelity controlled-Z gate with maximal intermediate leakage operating at the speed limit in a superconducting quantum processor,* Phys. Rev. Lett. **126**, 220502 (2021).

[11] J. F. Marques, B. M. Varbanov, M. S. Moreira, H. Ali, N. Muthusubramanian, C. Zachariadis, F. Battistel, M. Beekman, N. Haider, W. Vlothuizen, A. Bruno, B. M. Terhal, and L. DiCarlo, *Logical-qubit operations in an error-detecting surface code,* Nature Physics (2021).

[12] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, *Scalable quantum circuit and control for a superconducting surface code,* Phys. Rev. Appl. **8**, 034021 (2017).

[13] J. E. Mooij, T. P. Orlando, L. Levitov, L. Tian, C. H. van der Wal, and S. Lloyd, *Josephson persistent-current qubit,* Science **285**, 1036 (1999).

[14] F. Yan, S. Gustavsson, A. Kamal, J. Birenbaum, A. P. Sears, D. Hover, T. J. Gudmundsen, D. Rosenberg, G. Samach, S. Weber, J. L. Yoder, T. P. Orlando, J. Clarke, A. J. Kerman, and W. D. Oliver, *The flux qubit revisited to enhance coherence and reproducibility,* Nature Communications **7**, 12964 (2016).

[15] J. Ku, X. Xu, M. Brink, D. C. McKay, J. B. Hertzberg, M. H. Ansari, and B. L. T. Plourde, *Suppression of unwanted ZZ interactions in a hybrid two-qubit system,* Phys. Rev. Lett. **125**, 200504 (2020).

[16] V. E. Manucharyan, J. Koch, L. I. Glazman, and M. H. Devoret, *Fluxonium: Single cooper-pair circuit free of charge offsets,* Science **326**, 113 (2009).

[17] N. Maleeva, L. Grünhaupt, T. Klein, F. Levy-Bertrand, O. Dupre, M. Calvo, F. Valenti, P. Winkel, F. Friedrich, W. Wernsdorfer, and et al., *Circuit quantum electrodynamics of granular aluminum resonators,* Nature Communications **9** (2018), 10.1038/s41467-018-06386-9.

[18] A. Somoroff, Q. Ficheux, R. A. Mencia, H. Xiong, R. V. Kuzmin, and V. E. Manucharyan, *Millisecond coherence in a superconducting qubit,* (2021), arXiv:2103.08578 [quant-ph] .

[19] Q. Ficheux, L. B. Nguyen, A. Somoroff, H. Xiong, K. N. Nesterov, M. G. Vavilov, and V. E. Manucharyan, *Fast logic with slow qubits: Microwave-activated controlled-Z gate on low-frequency fluxoniums,* Phys. Rev. X **11**, 021026 (2021).

[20] H. Xiong, Q. Ficheux, A. Somoroff, L. B. Nguyen, E. Dogan, D. Rosenstock, C. Wang, K. N. Nesterov, M. G. Vavilov, and V. E. Manucharyan, *Arbitrary controlled-phase gate on fluxonium qubits using differential ac-Stark shifts,* (2021), arXiv:2103.04491 [quant-ph] .

[21] X. You, J. A. Sauls, and J. Koch, *Circuit quantization in the presence of time-dependent external flux,* Physical Review B **99** (2019), 10.1103/physrevb.99.174512.

[22] D. Ristè, C. C. Bultink, K. W. Lehnert, and L. DiCarlo, *Feedback control of a solid-state qubit using high-fidelity projective measurement,* Physical Review Letters **109** (2012).

[23] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, J. Heinsoo, J.-C. Besse, M. Gabureac, A. Wallraff, and C. Eichler, *Entanglement stabilization using ancilla-based parity detection and real-time feedback in superconducting circuits,* npj Quantum Information **5** (2019), 10.1038/s41534-019-0185-4.

[24] S. Zeytinoğlu, M. Pechal, S. Berger, A. A. Abdumalikov, A. Wallraff, and S. Filipp, *Microwave-induced amplitude- and phase-tunable qubit-resonator coupling in circuit quantum electrodynamics,* Physical Review A **91** (2015).

[25] D. Egger, M. Werninghaus, M. Ganzhorn, G. Salis, A. Fuhrer, P. Müller, and S. Filipp, *Pulsed reset protocol for fixed-frequency superconducting qubits,* Phys. Rev. Applied **10**, 044030 (2018).

[26] P. Magnard, P. Kurpiers, B. Royer, T. Walter, J.-C. Besse, S. Gasparinetti, M. Pechal, J. Heinsoo, S. Storz, A. Blais, and A. Wallraff, *Fast and unconditional all-microwave reset of a superconducting qubit,* Phys. Rev. Lett. **121**, 060502 (2018).

[27] M. McEwen, D. Kafri, Z. Chen, J. Atalaya, K. J. Satzinger, C. Quintana, P. V. Klimov, D. Sank, C. Gidney, A. G. Fowler, F. Arute, K. Arya, B. Buckley, B. Burkett, N. Bushnell, B. Chiaro, R. Collins, S. Demura, A. Dunsworth, C. Erickson, B. Foxen, M. Giustina, T. Huang, S. Hong, E. Jeffrey, S. Kim, K. Kechedzhi, F. Kostritsa, P. Laptev, A. Megrant, X. Mi, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Niu, A. Paler, N. Redd, P. Roushan, T. C. White, J. Yao, P. Yeh, A. Zalcman, Y. Chen, V. N. Smelyanskiy, J. M. Martinis, H. Neven, J. Kelly, A. N. Korotkov, A. G. Petukhov, and R. Barends, *Removing leakage-induced correlated errors in superconducting quantum error correction,* (2021), arXiv:2102.06131 [quant-ph] .

[28] H. Hsu, M. Silveri, V. Sevriuk, M. Möttönen, and G. Catelani, *Charge dynamics in quantum-circuit refrigeration: thermalization and microwave gain,* (2021), arXiv:2107.04278 [cond-mat.mes-hall] .

[29] Y. Zhou, Z. Zhang, Z. Yin, S. Huai, X. Gu, X. Xu, J. Allcock, F. Liu, G. Xi, Q. Yu, H. Zhang, M. Zhang, H. Li, X. Song, Z. Wang, D. Zheng, S. An, Y. Zheng, and S. Zhang, *Rapid and unconditional parametric reset protocol for tunable superconducting qubits,* (2021), arXiv:2103.11315 [quant-ph] .

[30] P. Reinhold, *Controlling Error-Correctable Bosonic Qubits*, Ph.D. thesis, Yale University (2019).

[31] J. M. Martinis and M. R. Geller, *Fast adiabatic qubit gates using only σ_z control,* Phys. Rev. A **90**, 022307 (2014).

[32] J. R. Schrieffer and P. A. Wolff, *Relation between the Anderson and Kondo Hamiltonians,* Phys. Rev. **149**, 491 (1966).

[33] S. Bravyi, D. P. DiVincenzo, and D. Loss, *Schrieffer - Wolff transformation for quantum many-body systems,* Annals of Physics **326**, 2793 (2011).

[34] E. Magesan and J. M. Gambetta, *Effective Hamiltonian models of the cross-resonance gate,* Physical Review A **101** (2020).

[35] M. Boissonneault, J. M. Gambetta, and A. Blais, *Dispersive regime of circuit QED: Photon-dependent qubit dephasing and relaxation rates,* Physical Review A **79** (2009).

[36] S. Haroche and J. Raimond, *Exploring the Quantum: Atoms, Cavities, and Photons*, Oxford Graduate Texts (Oxford University Press, 2006).

[37] M. D. Reed, B. R. Johnson, A. A. Houck, L. DiCarlo, J. M. Chow, D. I. Schuster, L. Frunzio, and R. J. Schoelkopf, *Fast reset and suppressing spontaneous emission of a superconducting qubit,* Applied Physics Letters **96**, 203110 (2010).

[38] E. Jeffrey, D. Sank, J. Y. Mutus, T. C. White, J. Kelly, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. Megrant, P. J. J. O'Malley, C. Neill, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, *Fast accurate state measurement with superconducting qubits,* Phys. Rev. Lett. **112**, 190504 (2014).

[39] S. Barzanjeh, D. P. DiVincenzo, and B. M. Terhal, *Dispersive qubit measurement by interferometry with parametric amplifiers,* Physical Review B **90** (2014), 10.1103/physrevb.90.134515.

[40] K. Reuer, *Real time, single shot, dispersive readout of superconducting qubits using a field progammable gate array,* Master Thesis, ETH Zürich (2018).

[41] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, and C. Eichler, *Rapid high-fidelity multiplexed readout of superconducting qubits,* Phys. Rev. Appl. **10**, 034040 (2018).

[42] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, *Simple pulses for elimination of leakage in weakly nonlinear qubits,* Phys. Rev. Lett. **103**, 110501 (2009).

[43] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo, *Restless tuneup of high-fidelity qubit gates,* Phys. Rev. Applied **7**, 041001 (2017).

[44] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Tunable Coupling Scheme for Implementing High-Fidelity Two-Qubit Gates,* Physical Review Applied **10**, 054062 (2018).

[45] D. Guéry-Odelin, A. Ruschhaupt, A. Kiely, E. Torrontegui, S. Martínez-Garaot, and J. Muga, *Shortcuts to adiabaticity: Concepts, methods, and applications,* Reviews of Modern Physics **91** (2019), 10.1103/revmodphys.91.045001.

[46] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Z. Chen, K. Satzinger, R. Barends, F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, S. Boixo, D. Buell, B. Burkett, Y. Chen, R. Collins, E. Farhi, A. Fowler, C. Gidney, M. Giustina, R. Graff, M. Harrigan, T. Huang, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, P. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, E. Lucero, J. McClean, M. McEwen, X. Mi, M. Mohseni, J. Y. Mutus, O. Naaman, M. Neeley, M. Niu, A. Petukhov, C. Quintana, N. Rubin, D. Sank, V. Smelyanskiy, A. Vainsencher, T. C. White, Z. Yao, P. Yeh, A. Zalcman, H. Neven, J. M. Martinis, and Google AI Quantum, *Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms,* Physical Review Letters **125** (2020).

[47] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler,* Phys. Rev. X **11**, 021058 (2021).

[48] S. A. Caldwell, N. Didier, C. A. Ryan, E. A. Sete, A. Hudson, P. Karalekas, R. Manenti, M. P. da Silva, R. Sinclair, E. Acala, N. Alidoust, J. Angeles, A. Bestwick, M. Block, B. Bloom, A. Bradley, C. Bui, L. Capelluto, R. Chilcott, J. Cordova, G. Crossman, M. Curtis, S. Deshpande, T. E. Bouayadi, D. Girshovich, S. Hong, K. Kuang, M. Lenihan, T. Manning, A. Marchenkov, J. Marshall, R. Maydra, Y. Mohan, W. O'Brien, C. Osborn, J. Otterbach, A. Papageorge, J.-P. Paquette, M. Pelstring, A. Polloreno, G. Prawiroatmodjo, V. Rawat, M. Reagor, R. Renzas, N. Rubin, D. Russell, M. Rust, D. Scarabelli, M. Scheer, M. Selvanayagam, R. Smith, A. Staley, M. Suska, N. Tezak, D. C. Thompson, T.-W. To, M. Vahidpour, N. Vodrahalli, T. Whyland, K. Yadav, W. Zeng, and C. Rigetti, *Parametrically activated entangling gates using transmon qubits,* Phys. Rev. Applied **10**, 034050 (2018).

[49] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M. P. da Silva, and B. R. Johnson, *Demonstration of a parametrically activated entangling gate protected from flux noise,* Physical Review A **101** (2020).

[50] N. Didier, E. A. Sete, M. P. da Silva, and C. Rigetti, *Analytical modeling of parametrically modulated transmon qubits,* Phys. Rev. A **97**, 022330 (2018).

[51] C. Rigetti and M. Devoret, *Fully microwave-tunable universal gates in superconducting qubits with linear couplings and fixed transition frequencies,* Phys. Rev. B **81**, 134507 (2010).

[52] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, *Simple all-microwave entangling gate for fixed-frequency superconducting qubits,* Phys. Rev. Lett. **107**, 080502 (2011).

[53] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, *Procedure for systematically tuning up cross-talk in the cross-resonance gate,* Physical Review A **93**, 060302 (2016).

[54] N. Sundaresan, I. Lauer, E. Pritchett, E. Magesan, P. Jurcevic, and J. M. Gambetta, *Reducing unitary and spectator errors in cross resonance with optimized rotary echoes,* PRX Quantum **1** (2020), 10.1103/prxquantum.1.020318.

[55] S. Kirchhoff, T. Keßler, P. J. Liebermann, E. Assémat, S. Machnes, F. Motzoi, and F. K. Wilhelm, *Optimized cross-resonance gate for coupled transmon systems,* Physical Review A **97**, 042348 (2018).

[56] V. Tripathi, M. Khezri, and A. N. Korotkov, *Operation and intrinsic error budget of a two-qubit cross-resonance gate,* Phys. Rev. A **100**, 012301 (2019).

[57] J. B. Hertzberg, E. J. Zhang, S. Rosenblatt, E. Magesan, J. A. Smolin, J.-B. Yau, V. P. Adiga, M. Sandberg, M. Brink, J. M. Chow, and J. S. Orcutt, *Laser-annealing Josephson junctions for yielding scaled-up superconducting quantum processors,* npj Quantum Information **7**, 129 (2021).

[58] C. Chamberland, G. Zhu, T. J. Yoder, J. B. Hertzberg, and A. W. Cross, *Topological and subsystem codes on low-degree graphs with flag qubits,* Physical Review X **10** (2020), 10.1103/physrevx.10.011022.

3

NOISE IN SUPERCONDUCTING QUBITS

3.1. OVERALL MEASURES OF DECOHERENCE: T_1 AND T_2

Before discussing specific noise mechanisms for superconducting qubits, here we cover the two major parameters that are used to characterize the quality of any kind of qubit.

Consider a Markovian system, where the correlation time of the noise is considered to be much shorter than the typical timescale of the system dynamics [1]. In this case one can describe the evolution via the Lindblad equation

$$\dot{\rho} = \mathscr{L}(\rho) = -i[H,\rho] + \sum_j L_j \rho L_j^\dagger - \frac{1}{2}\{L_j^\dagger L_j, \rho\}, \tag{3.1}$$

where H is the system Hamiltonian, $\{L_j\}$ are the quantum jump operators and $\mathscr{L}$ is the overall Lindbladian. First, let $L_1 = \sigma_-/\sqrt{T_1}$ for a qubit (or $L_1 = a/\sqrt{T_1}$ for a harmonic oscillator, but in this section we focus on qubits) be the only jump operator. Then, assuming that H is simply the free Hamiltonian, $H = -\omega Z/2$, the initial density matrix evolves as

$$\rho(0) = \begin{pmatrix} 1-p & \beta \\ \beta^* & p \end{pmatrix} \quad \mapsto \quad \rho(t) = \begin{pmatrix} 1-pe^{-\frac{t}{T_1}} & \beta e^{i\omega t} e^{-\frac{t}{2T_1}} \\ \beta^* e^{-i\omega t} e^{-\frac{t}{2T_1}} & pe^{-\frac{t}{T_1}} \end{pmatrix}. \tag{3.2}$$

One can see that the population in $|1\rangle$ decays towards the ground state $|0\rangle$ with a time constant given by T_1. Thus the jump operator σ_- models qubit relaxation and T_1 is called the relaxation time. Note also that the off-diagonal terms decay with a time constant $2T_1$ (apart for a phase factor which could be removed going to a reference frame rotating at the qubit frequency). This is referred to as relaxation-induced dephasing and corresponds to the fact that superposition states lose phase coherence when relaxation occurs. The jump operator σ_+ would instead model qubit excitation, although it is often neglected as the excitation rate is usually much lower than the relaxation rate, given the low temperature

at which qubits are kept (~ 20 mK), relative to their frequency (3-8 GHz for transmons). See Section 2.5.1 for more information about residual excitations.

If the only jump operator is instead $L_2 = \sqrt{1/(2T_\phi)}\,Z$, the density matrix evolves as

$$\rho(0) = \begin{pmatrix} 1-p & \beta \\ \beta^* & p \end{pmatrix} \quad \mapsto \quad \rho(t) = \begin{pmatrix} 1-p & \beta e^{i\omega t} e^{-\frac{t}{T_\phi}} \\ \beta^* e^{-i\omega t} e^{-\frac{t}{T_\phi}} & p \end{pmatrix}. \tag{3.3}$$

In this case, only the off-diagonal elements decay with time constant T_ϕ, corresponding to so-called pure dephasing. The combination of relaxation and pure dephasing makes the off-diagonal elements decay at a rate $1/T_2 := 1/2T_1 + 1/T_\phi$, where T_2 is referred to as the decoherence time.

While T_1 and T_2 are commonly used to characterize qubits, the validity of the model above depends on how Markovian the environment is in experiment. Regarding relaxation due to charge noise, since the latter is suppressed at least for transmons (see Section 2.2), we focus on dephasing processes, which affect T_ϕ (and in turn T_2). Non-Markovian processes are the ones that are slow relative to the dynamics of the system, like the low-frequency components of $1/f$ noise (see Section 3.2.1). In a quasi-static approximation, a parameter like the qubit frequency is assumed to stay constant within a certain period (e.g. idling, or during a gate) but to vary across repetitions (of the experiment or the gate), following a certain distribution. If this distribution is Gaussian, the off-diagonal elements decay as a Gaussian $e^{-(t/T_\phi^G)^2}$, instead of as an exponential e^{-t/T_ϕ} (there is also a factor $e^{-t/2T_1}$ for all cases so we do not write it explicitly in this discussion). In general, the decay due to dephasing is $e^{-\chi_N(t)}$ for a generic function $\chi_N(t)$, where the N is used to refer to the number of echo pulses applied [2] (echo pulses are discussed below). This function might fall somewhere in-between exponential and Gaussian decay, but it can also be more general. An expression for $e^{-\chi_N(t)}$ is given by [2]

$$e^{-\chi_N(t)} = \exp\Big(-\frac{t^2}{2}\frac{\partial\omega}{\partial\lambda}\int_{-\infty}^{+\infty} g_N(\omega,t)S(\omega)\,d\omega\Big), \tag{3.4}$$

where λ is the noise parameter (for example, it is the flux in the case of flux noise), ω is the qubit frequency, g_N is an appropriate filtering function and $S(\omega)$ is the spectral density of the noise. The latter is defined as

$$S(\omega) = \int_{-\infty}^{+\infty} dt\,\langle\lambda(t)\lambda(0)\rangle e^{-i\omega t}, \tag{3.5}$$

i.e. it is the Fourier transform of the noise autocorrelation function $\langle\lambda(t)\lambda(0)\rangle$. We remark that this is the so-called bilateral spectral density and it is defined without a $1/\pi$ or $1/(2\pi)$ factor in front, to which one needs to be careful when comparing different references.

Slow noise components can often be mitigated using various techniques. The simplest one is an echo experiment during an idling period, in which a π pulse (i.e. a bit flip) is applied to the qubit in the middle of the evolution, as well as at the end. This perfectly cancels out quasi-static noise because any drift in one direction is compensated by the same drift in the opposite one. If T_2 is measured in an echo experiment it is usually

called T_2^E, while if it is measured in a standard Ramsey experiment (so no echo pulse) it is called T_2^*. If the noise is slow but not "slow-enough" compared to the frequency of the echo pulses, a single pair of echo pulses cancels out only a subset of the noise spectrum. In other words, $g_1(t,\omega)$ [2] in Eq. (3.4) filters out only certain components of $S(\omega)$. In that case, to improve the filtering effect, one can use more frequent echoing pulses. Clearly $T_2^E \geq T_2^*$ so it is generally good to use echo pulses. However, they cannot be applied during gates or measurements of the given qubit, thus one needs more specialized techniques, such as the Net Zero technique that we introduce in Chapter 6 for flux-based gates.

3.2. PHYSICAL NOISE SOURCES IN SUPERCONDUCTING QUBITS

3.2.1. TWO-LEVEL SYSTEMS

In this section we provide a phenomenological description of noise due to two-level systems (TLSs), highlighting first the connection between random telegraph noise and $1/f$ noise.

RANDOM TELEGRAPH NOISE

Consider first a single TLS and assume that it randomly switches from one state to the other, and vice versa, with rates $\gamma^{\uparrow}$ and $\gamma^{\downarrow}$, respectively. This is referred to as random telegraph noise. For simplicity, assume that $\gamma^{\uparrow} = \gamma^{\downarrow} \equiv \gamma$. Consider that the TLS is in one specific state at $t = 0$. Let $p(t)dt$ be the probability that the TLS remains in that state for a time t, and then jumps to the other state between t and $t + dt$. In other words, $p(t)$ is the distribution of switching times. To derive an expression for $p(t)$, let us discretize t as $t = ndt$. By definition of rate, γdt is the probability that there is a change of state during dt. It follows that

$$p(t)dt = (1-\gamma dt)^n \gamma dt = (1-\gamma dt)^{\frac{t}{dt}} \gamma dt. \tag{3.6}$$

Simplifying dt on both sides and taking the limit, one gets

$$p(t) = \lim_{dt\to 0} (1-\gamma dt)^{\frac{t}{dt}} \gamma = \gamma e^{-\gamma t}. \tag{3.7}$$

Thus $p(t)$ follows an exponential decay.

Let λ in Eq. (3.5) be $\lambda(t) = \sum_j \lambda_j^{\pm}(t)$, where the sum runs over TLSs and the $\{\lambda_j^{\pm}\}$ are the values associated with the two states of a TLS. It has been proven [3] that the presence of many TLSs with exponentially decaying $p(t)$ and with decay rates $\gamma \in [\gamma_1, \gamma_2]$ produces a noise spectral density (see Eq. (3.5)) given by

$$S(f) = \frac{A}{f} \tag{3.8}$$

with A constant, for all the frequencies $f = \omega/(2\pi)$ such that $\gamma_1 \ll f \ll \gamma_2$. This is called $1/f$ noise and it is found to be ubiquitous not just in quantum devices but also in e.g. classical electronics. The minimal set of assumptions required (i.e. an ensemble of TLSs with broadly distributed rates) determines the ubiquity of $1/f$ noise, but at the same time makes it hard to exactly identify which physical systems are responsible for it (see below). We note that the physical systems may have more than two levels, in which case one can

substitute the rate γ with a sum of rates towards multiple states, while still getting the same form for $p(t)$ in Eq. (3.7). If then one of these rates is much larger than the others, one can even neglect the other levels and focus just on the two with the largest γ.

CHARGE NOISE

Regarding superconducting qubits, TLSs [4, 5] can couple to the qubits either via the charge operator (charge noise) or via their magnetic field (flux noise). In the first case these dielectric TLSs can cause qubit relaxation by swapping an excitation from the qubit to the TLS and subsequently dispersing it in the environment [6]. These TLSs lead to a reduction in T_1 and are actually considered the current limiting factor that sets T_1. Furthermore, they are considered to be responsible for the relatively large fluctuations of the measured values of T_1 in time [6]. If E_J/E_C is not within the transmon regime, these TLSs also produce fluctuations in the energy levels, which lead to a reduction of T_ϕ as well. Indeed, the transmon was introduced over the Cooper-pair box to counter this deleterious effect of charge noise (see Section 2.2).

One can make a distinction between TLSs that are strongly coupled to the qubit and those that are weakly coupled to it [6]. The latter affect T_1 as an ensemble independently of the qubit frequency (in the way described above), whereas a single TLS of the former type can strongly reduce T_1 if the TLS frequency matches the qubit frequency. These strongly coupled TLSs are fairly common and constitute a major challenge to coherence times, especially in multi-qubit devices. One solution is to warm up and cool down again the fridge (or only the chip, if possible), as it is observed that this usually changes the location and distribution of the TLSs. However, it is not a scalable solution. Another solution is to use flux-tunable qubits, so that one can tune them to a slightly different frequency to avoid resonance with TLSs. However, frequency tunability comes at the cost of potential frequency crowding, additional control circuitry and introduces sensitivity to flux noise (see below), especially when qubits have to be moved far away from their sweetspot(s).

FLUX NOISE

Magnetic TLSs cause flux noise for qubits that use an externally controlled magnetic field, like tunable transmons (see Section 2.2.1), if they are physically located close enough to the SQUID loop itself. Depending on whether the magnetic field points in one direction or switches to the other, the magnetic flux through the SQUID loop is slightly altered. The combination of many TLSs then causes fluctuations of the qubit frequency with $1/f$ spectral density, as described around Eq. (3.8), leading to a reduction in T_ϕ away from the sweetspot, as quantified in the following. The sensitivity $s(\Phi)$ to flux noise is defined as

$$s(\Phi) := \frac{1}{2\pi}\frac{\partial\omega(\Phi)}{\partial\Phi}, \tag{3.9}$$

where $\omega/2\pi$ is the qubit frequency and Φ is the magnetic flux through the SQUID loop. By definition (see Eq. (2.22)), $s(\Phi) = 0$ at the sweetspots, in particular at $\Phi = 0$, thus T_ϕ is (first-order) insensitive to flux noise at that point. If the transmon is not at its sweetspot, in experiment it is typically observed that the dephasing rate $\Gamma_\phi = 1/T_\phi$ is linearly increasing

with $s(\Phi)$ (see e.g. Fig. 6.9), at least for a not too large Φ (after which second-order effects may start to play a role). For an echo experiment, an analytical formula is given by [7]

$$\Gamma_\phi^E(\Phi) = 2\pi\sqrt{\ln 2}\sqrt{A}\,s(\Phi). \tag{3.10}$$

By fitting the experimental measurements of $\Gamma_\phi^E(\Phi)$, one can use this formula to extract the value of $\sqrt{A}$. For a Ramsey experiment, a linear dependence is also typically observed (see e.g. Fig. 6.9, even though it is not necessarily guaranteed), where the numerical coefficient of the linear increase with $s(\Phi)$ is larger (whereas $\sqrt{A}$ is the same). This can be attributed to the fact that an echo pulse mitigates the "slow-enough" components of $1/f$ noise (see Section 3.1), while this filtering effect is not present in a Ramsey experiment.

UNKNOWN IDENTITY OF TLSS

Dielectric TLSs are understood to correspond to defects or charge traps that reside at interfaces between dielectrics, the junction tunnel barrier, the substrate or any combination thereof [2]. However, the details and the specific compounds involved are unclear and an active research topic. Similarly, magnetic TLSs are associated with magnetic dipoles at the superconducting metal surfaces, but their specific identity is unknown. One common suspect are residual contaminants due to the fabrication process [8, 9]. Thus new fabrication techniques are regularly being developed trying to improve qubit coherence times. Identifying the actual nature of TLSs would greatly help to steer these efforts towards the most relevant direction.

3.2.2. QUASI-PARTICLES

At finite temperature, the Cooper pairs in a superconductor can be broken into two separate electrons despite the protection provided by the superconducting gap. This happens as the Gibbs state has excitations on top of the Cooper-pair condensate due to thermal fluctuations. These unpaired electrons are called quasi-particles. Depending on the type of qubit, they are responsible for a reduction in T_1, e.g. for transmons, or T_2 as well [2].

The density of broken Cooper pairs, commonly expressed as $N_{\text{broken}}/N_{\text{pair}}$, is expected to decay exponentially with temperature. In particular, it is expected to be $< 10^{-24}$ at 20-40 mK [10]. However, it is typically found to be around 10^{-8}-10^{-6} [11, 12], that is, many orders of magnitude larger than expected. This discrepancy is not well understood and it is not explained by the theory of superconductivity alone. It might be partially due to radioactive impurities or cosmic rays [10, 13, 14] (see Section 3.2.3 below), since both release relatively large amounts of energy on the chip, potentially breaking many Cooper pairs.

3.2.3. COSMIC RAYS AND RADIOACTIVITY

Recent work [10, 13, 14] has identified cosmic rays and radioactive impurities in materials within the fridge as a threatening source of errors in superconducting qubits, due to the quasi-particles produced by the energy deposited by these events. The upper bound on T_1 set by cosmic radiation has been estimated around 4 ms [10]. While no superconducting qubit has so far reported such a high number for T_1, losses due to dielectric TLSs and other

sources have been steadily diminishing to the point that such T_1 seems to be achievable in the near future [15]. Cosmic rays might thus soon become the limiting error source in setting coherence times and will need to be addressed.

Reducing radioactive impurities and moving the fridge to an underground facility has been shown to improve the quality factor of superconducting resonators [13] and the same is expected for qubits. Furthermore, high-energy cosmic rays have been identified as a source of catastrophic error bursts [14], corresponding to chip-wide correlated errors at a rate of approximately 1 event every 10 s. These errors are too extended in space and time to be handled by quantum error correction, constituting a roadblock to fault-tolerance and to any computation that cannot be completed within just a few seconds (unless the chip is really large, or a modular design is developed, with multiple chips connected within the fridge or across different fridges). This further motivates research of shielding methods, as well as techniques to reduce radioactive impurities in materials.

3.2.4. PHOTON-SHOT NOISE

Transmons are commonly coupled to their readout resonator for measurements (see Section 2.6). Dispersive readout is based on the fact that the transmon state shifts the resonator frequency depending on the state of the qubit, as manifested by Eq. (2.51). That equation, in the computational subspace, can also be rewritten as

$$\tilde{H}|_{\mathscr{C}} = \left(-\frac{\tilde{\omega}^q}{2} + \chi a^\dagger a\right) Z + \tilde{\omega}^r a^\dagger a, \tag{3.11}$$

i.e. it can be reinterpreted as the qubit frequency being shifted depending on the resonator state. As a consequence, fluctuations in the photon number in the resonator (known as photon-shot noise) lead to a reduction of T_ϕ, since the qubit phase cannot be tracked precisely.

3.2.5. DISTORTIONS OF ELECTRONIC SIGNALS

Classical electronic signals need to be sent from the room-temperature control electronics to deep inside the fridge where the qubits are located. Many effects might distort the target shape of the signal, such as limited waveform-generator bandwidth, high-pass bias tees, low-pass filters, impedance mismatches, skin effect, on-chip response... Here we focus in particular on distortions in the flux line, where a voltage pulse $V_{\mathrm{AWG}}(t)$ induces a current, which in turn generates a magnetic flux $\Phi(t)$ threading the SQUID loop of the qubit. The flux is generally used to activate a two-qubit gate (mainly CZ and iSWAP) by tuning the qubit frequency (see Section 2.8.3).

Distortions can be described as a linear time-invariant system that transduces voltage to flux and is characterized by its impulse response $h(t)$, where

$$\Phi(t) = h * V_{\mathrm{AWG}}(t) = \int_{-\infty}^{+\infty} d\tau\, h(t-\tau) V_{\mathrm{AWG}}(\tau) \tag{3.12}$$

with $*$ indicating convolution. The impulse response would ideally be a Dirac delta function, $h(\tau) = c\delta(\tau)$, where c is the (constant) conversion factor from voltage to flux (which needs to be calibrated in experiment). However, relative to the gate duration T_g, $h(\tau)$ generally contains short-, medium- and long-timescale distortions, which correspond to $h(\tau)$

being non-zero for $\tau \ll T_g$ (even on the sub-nanosecond scale), $\tau \sim T_g$ and $\tau \gg T_g$ (up to tens of microseconds), respectively. We can also translate these features from the time domain to the frequency domain by applying the Fourier transform $\mathscr{F}$ to Eq. (3.12), getting

$$\hat{\Phi}(\omega) = \hat{h}(\omega)\hat{V}_{\mathrm{AWG}}(\omega) \tag{3.13}$$

thanks to the convolution theorem, where

$$f(t) \overset{\mathscr{F}}{\mapsto} \hat{f}(\omega) := \int_{-\infty}^{+\infty} dt\, f(t) e^{-i\omega t}. \tag{3.14}$$

In the frequency domain, ideally $\hat{h}(\omega) = c$, i.e. all frequency components in $\hat{V}_{\mathrm{AWG}}$ are reproduced as intended in $\hat{\Phi}$ (modulo a constant rescaling). The definitions of short-, medium- and long-timescale distortions translate to $\hat{h}(\omega)$ having peaks on high-, medium- or low-frequency components (relative to $2\pi/T_g$), respectively. Hence, certain parts of the spectrum of the flux pulse are altered by an unequal distribution of weights in $\hat{h}(\omega)$.

Distortions can be characterized experimentally using the qubit as a probe [16]. Thanks to this characterization, one can pre-distort the desired flux $\Phi_{\mathrm{target}}(t)$ with a best estimation $\tilde{h}^{-1}$ of h^{-1}, i.e. one sets

$$V_{\mathrm{AWG}}(t) = \tilde{h}^{-1} * \Phi_{\mathrm{target}}(t). \tag{3.15}$$

Applying the Fourier transform and combining this equation with Eq. (3.13), we get

$$\hat{\Phi}(\omega) = \hat{r}(\omega)\hat{\Phi}_{\mathrm{target}}(\omega), \tag{3.16}$$

where

$$\hat{r}(\omega) = \hat{\tilde{h}}^{-1}(\omega)\hat{h}(\omega) \tag{3.17}$$

quantifies the remaining distortions after corrections. From Eq. (3.16) it is clear that, if $\tilde{h}^{-1} = h^{-1}$, then one would have $\hat{r}(\omega) = 1$ and $\Phi(t) = \Phi_{\mathrm{target}}(t)$. However, one can usually correct only short- and medium-timescale distortions with so-called finite and infinite impulse-response filters. Instead, long-timescale distortions are, first, hard to quantify and, second, they are not compatible with real-time execution of operations in a fully programmable quantum computer (see also Section 6.2). Pre-distortions to be applied on a gate would depend on the history of previous gates, which is a challenging and non-scalable solution.

Baseband-flux two-qubit gates (see Section 2.8.3) are generally the fastest among two-qubit gates in transmons, because the qubit is fluxed right to the interaction point (and back), for a total time that just slightly exceeds the fundamental speed limit ($\pi/(\sqrt{2}J^{AB})$ for the CZ) set by the exchange coupling J^{AB} (see Eq. (2.77)). At the same time, in their conventional form [17] they are among the most sensitive to long-timescale distortions, since the spectral density of the pulse has a large weight on low-frequency components. Based on the discussion after Eq. (3.17), $\hat{r}(\omega)$ multiplies these low-frequency components with unequal factors, thus significantly distorting the pulse. Because of these long-timescale distortions, this problem has been avoided either by removing the flux line altogether

and doing purely microwave two-qubit gates, namely the cross-resonance gate [18, 19] (see Section 2.8.3 for a brief introduction), or by resorting to parametric driving that modulates the qubit frequency via fast oscillations of the flux [20, 21] (see also Section 2.8.3). In the latter case, since the pulse spectral density is peaked at relatively high ω, whereas it is basically 0 at low ω, it does not matter if $\hat{r}(\omega)$ distorts those low-ω components.

In Chapter 6 we introduce the Net Zero technique and apply it to baseband-flux gates. A single, positive-flux pulse is replaced by two symmetric halves with positive and negative flux. The dependence of the transmon frequency on flux (see Eq. (2.21)) is the same for positive and negative flux, leading to the same effect on the Hamiltonian. Among other advantages, the key zero-integral feature removes the DC component and the very-low frequency components of $\Phi(t)$ (see Section 6.3). In this way, long-timescale distortions are strongly suppressed while keeping the same gate speed of baseband-flux gates. See Chapter 6 for more information.

3.2.6. LEAKAGE

The name superconducting qubits is misleading because they are never actually two-level systems. Rather, they are comprised of many levels, where only two can be approximately considered as a qubit, thanks to the anharmonicity. However, due to a low anharmonicity and/or the explicit use of non-qubit states for operations, it is possible that higher excited states (usually $|2\rangle$ or also $|3\rangle$) get populated. This is called leakage and it is an often-neglected error source in superconducting qubits. Remarkably, two-qubit gates are often characterized only by their fidelity, but leakage is not quantified, based on the assumption that it is "low". However, even a seemingly low amount of leakage can have a significant impact e.g. on the logical performance of the surface code (see Section 8.3).

We amply discuss leakage in Sections 2.7, 2.8, 3.3 and 3.4.3 and Chapter 5. Specifically, in Section 2.7 we mentioned how the DRAG pulsing technique reduces leakage in single-qubit gates to practically negligible levels. In Section 2.8 we discussed conditional-phase gates, which are the major source of leakage in superconducting qubits, due to the use of an avoided crossing between a computational and a leaked state. In Section 3.3 we discuss our studies of leakage and in particular the effective model (based on Lindblad simulations) that we use in the density-matrix simulations of the surface code. In Section 3.4.3 we highlight a randomized-benchmarking protocol with a modification to estimate the leakage rate. Finally, in Chapter 5 we discuss in great detail the previous literature on leakage and how to deal with it using leakage-reduction units. We do not discuss measurement-induced leakage (see e.g. Ref. [22] and the readout histograms in Ref. [23]) as we assume that it is less strong than leakage from the two-qubit gates.

3.2.7. CROSSTALK

Crosstalk might refer to a variety of phenomena that, broadly speaking, occur when an operation or the state of part of the system affects a different part which should be isolated in principle. Crosstalk might come from purely classical effects or it can be an undesired feature of the system Hamiltonian. Here we briefly discuss classical crosstalk due to the lines coupled to the qubit, as well as residual ZZ crosstalk.

CLASSICAL MICROWAVE AND FLUX CROSSTALK

Phenomenologically, the radiation produced by the microwave drive line of a qubit might affect other qubits and induce some undesired extra driving. The design and isolation of the drive lines is important to minimize this effect. One can also apply compensation pulses to other microwave lines to actively counteract this kind of crosstalk.

The current flowing into the flux line can spill into nearby patches of the chip. As the connections between qubits define closed loops in the chip topology, this current can circulate there and induce magnetic flux in neighboring qubits. It is customary to use airbridges [24], i.e. small pieces of metal over the various lines and connections, in order to break such loops and ground the entire chip, allowing these currents to flow away.

RESIDUAL ZZ CROSSTALK

Here we consider qubits to be parked at their sweetspots (or anyway far away from avoided crossings used to perform two-qubit gates). In this case, ideally, the frequency of a qubit should not be affected by whether a neighboring qubit is in state $|0\rangle$ or $|1\rangle$. In other words, one should have $\omega_{11} = \omega_{01} + \omega_{10}$ (let $\omega_{00} = 0$). However, if two transmons A and B are coupled with a fixed coupling J^{AB} (e.g. via a bus resonator; see Eq. (2.77)), one can derive from Eq. (2.79) that ω_{11} shifts by an amount

$$\zeta \approx \frac{2(J^{AB})^2}{\left|\omega^A - \omega^B - \alpha^B\right|} \tag{3.18}$$

in the limit $J^{AB} \ll \left|\omega^A - \omega^B - \alpha^B\right|$, where $\omega^A = \omega_{10}$ and $\omega^B = \omega_{01}$. Here for simplicity we have assumed that only $|02\rangle$ is sufficiently coupled to $|11\rangle$ (so B is by far the higher-frequency qubit), whereas $|20\rangle$ is further off-resonant. Furthermore, technically also $|01\rangle$ and $|10\rangle$ interact with each other, even though with a lower coupling, but one can redefine the working qubits as dressed qubits and this also renormalizes their qubit frequencies.

In practice [23], $\zeta/2\pi$ can reach up to ~ 3 MHz and it is uncommon to find $\zeta/2\pi < \mathcal{O}(100)$ kHz without using any additional circuitry. If one is operating a surface code, for a quantum error-correction cycle of duration 800 ns, this naively means that one performs up to 1.6 extra CZ gates per qubit pair during that time (which is equivalent to 0.4 CZ gates since $\mathrm{CZ}^2 = I$). The effect is, however, not so catastrophic, since one can take this into account and tune the parity-check unit (see Fig. 4.1(b,c)) as a single block [23], rather than tuning the CZ gates individually, or one can use echo pulses (as mentioned below) or dynamical decoupling [25].

Given the increasing size of quantum processors, residual ZZ crosstalk has received increasing attention in the field. One simple, hardware-efficient mitigation approach corresponds to using echo pulses on one qubit during the execution of the quantum error-correction cycle [23, 26]. The echo pulses do not just help to mitigate decoherence but they can partially revert the effect of residual ZZ crosstalk as well. This approach can be applied to arbitrary circuits to some extent. To instead solve this issue altogether, many tunable couplers have been developed [27–32]. The so-called gmon [27] was based on an inductive tunable coupler, however, most recent proposals use a capacitive tunable coupler [29], whose first example was introduced in Ref. [28].

Tunability comes at the price of additional components and control lines on the chip. However, in my opinion tunable couplers are essential to be able to scale up quantum

processors with no or little crosstalk. Furthermore, even with current tunable couplers, it actually seems difficult to achieve the $\zeta = 0$ condition for all pairs of qubits simultaneously, since the frequency tuning of one coupler might slightly affect the other at the circuit-Hamiltonian level. It is thus crucial to further develop tunable couplers.

3.3. NOISE MODELS IN THIS THESIS

Here we give an overview of the two main noise models developed and considered in the numerical simulations in this thesis. Furthermore, we provide some information about the implementation of the simulations themselves. The purpose of this section is to summarize and tie together the multiple uses and descriptions of the simulations that are contained in Chapter 6, Chapter 8 (specifically in Section 8.11.6) and Chapter 9 (specifically in Section 9.2) for the Lindblad simulations (sometimes referred to as full-trajectory simulations) and in Chapter 8 and Chapter 9 (especially in Section 9.3) for the density-matrix simulations. Extensive details can be found in Section 6.11.3 for the Lindblad simulations and in Sections 8.10.1 and 8.10.2 for the density-matrix simulations.

3.3.1. LINDBLAD SIMULATIONS

In the Lindblad simulations we simulate the full dynamics of the system, using the Lindblad equation. The Hamiltonian can in principle be arbitrary, although in practice we consider either the Hamiltonian of two coupled transmons (see Eq. (6.11)) or of a transmon capacitively coupled to its readout resonator (see Eq. (9.1)). We consider relaxation (T_1), dephasing (T_ϕ and quasi-static flux noise), leakage from the CZs and distortions (for an introduction to noise see Sections 3.1 and 3.2).

RELAXATION

Relaxation is modeled in an effective way as in Section 3.1, thus we do not simulate the TLSs, quasi-particles or cosmic rays that lead to the considered values of T_1, as this would be computationally unfeasible and goes beyond the purpose of these noise models.

A nuance is the basis in which relaxation should take place, e.g. in the case of two coupled transmons (the same applies to fast dephasing, as described below). Namely, one could consider the bare basis $\{|ij\rangle = |i\rangle \otimes |j\rangle\}$ of each transmon independently, so e.g. $|10\rangle \rightarrow |00\rangle$ upon a relaxation event on the first transmon. On the other hand, one could consider the dressed basis $\{|ij\rangle_D\}$ (see after Eq. (2.43)), that is, the basis that diagonalizes $H_0 + H_c$, where H_0 is the bare Hamiltonian and H_c is the coupling term. In this case e.g. $|10\rangle_D \rightarrow |00\rangle_D$. Specifically, we consider the dressed basis at the sweetspot, where qubits are generally parked for most of the time. We assume that the environment has time to "learn" that the eigenstates of the system are the dressed ones and thus couples to them, causing transitions between those rather than the bare ones. In general, when one performs measurements in experiment, one is indeed expected to measure features of the dressed states [33, 34]. While being fluxed to perform e.g. a CZ, the qubit is not anymore at the sweetspot and the dressed basis changes as a function of the flux. However, we still consider relaxation in the sweetspot dressed basis, because these fluxing periods are short and we assume that the environment does not have time to adjust.

DEPHASING

As far as dephasing is concerned, we do not simulate the full $1/f$ flux noise, nor we model Markovian sources in detail (the only exception is that we explicitly include photon-shot noise in Chapter 9). Instead, we try to capture the most important features in a computationally feasible way. Specifically, we take as a reference the gate time T_g and we consider all frequency components of the noise with $\omega > 2\pi/T_g$ as "fast" (Markovian), whereas we consider those with $\omega < 2\pi/T_g$ as quasi-static. We assume that the former are captured by the measured T_2^E and we include them in the simulations via the Lindblad equation (see around Eq. (3.3). Furthermore, either using Eq. (3.10) or directly the measured experimental values, we keep into account that T_2^E varies as a function of the flux Φ. The difference with Section 3.1 is that in the following chapters we do not consider qubits but qutrits, thus the jump operator cannot be just proportional to Z. The jump operator $\sqrt{2/T_\phi^E}\, a^\dagger a$ produces a quadratic dependence of T_ϕ^E on the level number, e.g. $|2\rangle$ dephases with time constant $T_\phi^E/2^2$, whereas the jump operators given in Eqs. (6.16) to (6.18) produce a linear dependence. As $1/f$ flux noise gives a linear dependence, in Chapters 6 and 8 we use that set of jump operators.

Regarding the quasi-static components, we model them as a stochastic constant shift $\Delta\Phi$ of the applied flux $\Phi(t)$. We assume that $\Delta\Phi$ follows a Gaussian distribution $p_\sigma(\Delta\Phi)$ centered in 0 and with a certain standard deviation σ. We choose σ such that the combination of fast (quantified by T_2^E) and quasi-static components produces the measured T_2^* when simulating a Ramsey experiment (see Fig. 6.9). For each $\Delta\Phi$ we integrate the Lindblad equation $\dot\rho = \mathscr{L}(\rho)$ to compute the time-evolution superoperator

$$\mathscr{P}_{T_g} := \mathscr{T}\, e^{\int_0^{T_g} dt'\, \mathscr{L}_{t'}}, \tag{3.19}$$

where $\mathscr{T}$ is the time-ordering operator. Then we average this as

$$\mathscr{P}_{T_g}^{\text{av}} = \int_{-\infty}^{+\infty} d(\Delta\Phi)\; p_\sigma(\Delta\Phi)\cdot \mathscr{P}_{T_g}(\Delta\Phi). \tag{3.20}$$

In practice, we discretize this integral and cut it at $\pm 5\sigma$. Furthermore, to compute each $\mathscr{P}_{T_g}$ we discretize the time evolution as

$$\mathscr{P}_{T_g} \simeq e^{\delta t \mathscr{L}_{T_g-\delta t}}\, e^{\delta t \mathscr{L}_{T_g-2\delta t}} \dots e^{\delta t \mathscr{L}_{2\delta t}}\, e^{\delta t \mathscr{L}_{\delta t}}\, e^{\delta t \mathscr{L}_0}, \tag{3.21}$$

where δt is chosen to be sufficiently small to approximate the exact evolution well. Note that in the numerics with this method one has to compute matrix exponentials, which in practice requires to diagonalize the exponent. In terms of physical time required by the numerics, we have found this method to outperform *qutip*-based methods that directly solve the differential equation $\dot\rho = \mathscr{L}(\rho)$, as long as the required δt is not too small. The discretization in Eq. (3.21) can be made smarter by an adaptive method that varies δt depending on the speed at which the flux changes, however, we have not explored this.

DISTORTIONS

The distortions are taken into account by distorting $\Phi_{\text{target}}(t)$ with the experimentally measured impulse response $\tilde h^{-1} * h(t)$ that takes into account pre-distortions (see Fig. 6.8).

LEAKAGE

As we consider qutrits (or even more levels in Chapter 9), leakage is automatically included in our simulations. Note that leakage can simply be due to the unitary dynamics, while noise can enhance it further (or also diminish it in case of relaxation). In Sections 6.9 and 6.19 we study the dependence of leakage on noise.

CROSSTALK

Residual ZZ crosstalk is naturally included in the simulations of two transmons, however, we do not really study crosstalk with respect to other neighboring systems (transmons, resonators...) since those are not included in the simulations. Only in Section 9.6.3 we effectively account for residual ZZ crosstalk by shifting directly the transmon frequency, rather than including neighboring transmons in the simulations.

3.3.2. DENSITY-MATRIX SIMULATIONS

Here we discuss the simulations [35, 36] of the density matrix of a full chip. In this thesis we have considered Surface-17, that is, the distance-3 rotated surface code under development in the DiCarlo lab (see Fig. 9.3 and Ref. [37]). We consider relaxation (T_1), dephasing (T_ϕ) and we introduce an effective model for the CZ that reproduces the main features observed in the Lindblad simulations, including leakage.

SIZE OF THE DENSITY MATRIX

As the cost of these simulations scales exponentially with the number of qubits, Surface-17 is the maximum we could study with the aid of a few GPUs. Of the 17 qubits in total, the density matrix keeps track of only 10 at any point, that is, the 9 data qubits and 1 ancilla qubit at the time. This is possible thanks to the fact that stabilizers commute and that we assume that each ancilla qubit is perfectly projected onto a computational state by the measurement. The latter allows to store the post-measurement ancilla-qubit state in a classical register rather than the density matrix (because the ancilla qubit is in a product state with the rest of the system). If n_{qb} transmons are actual qubits and n_{qt} are qutrits, the density matrix has size $4^{n_{\text{qb}}} \otimes 9^{n_{\text{qt}}}$. We further reduce this to $4^{n_{\text{qb}}} \otimes 5^{n_{\text{qt}}}$ by considering a model for leakage (see below) where the leakage subspace is decohered with respect to the computational subspace, in which case four off-diagonal matrix elements are 0. This model is motivated by the fact that stabilizer measurements tend to decohere leakage relatively fast (see Chapter 8).

In the numerics we also store the density matrix at the end of each quantum error-correction cycle so that we can analyze this data at any point later. The main quantities we extract are the evolution of the leakage population and of the logical fidelity. However, only the diagonal entries of this density matrix are necessary for these purposes, so we store only those entries to save a considerable amount of space.

PAULI TRANSFER MATRIX

We use the Pauli Transfer Matrix formalism (PTM) to represent states and operations (superoperators). This is numerically convenient because PTMs compose multiplicatively. The density matrix ρ can be written as a vector $\vec{\rho}$ with entries given by

$$\rho_i = \text{Tr}(P_i \rho), \tag{3.22}$$

where $\{P_i\}$ are n-qubit Pauli operators in the case of qubits ($d = 2$) or Gell-Mann matrices in the case of qutrits ($d = 3$). Gates, characterized in general by a superoperator $\mathscr{R}$, can be represented as a PTM matrix R with entries

$$R_{ij} = \frac{1}{d}\mathrm{Tr}\big(P_j \mathscr{R}(P_i)\big). \tag{3.23}$$

Measurements can also be implemented by applying a different PTM depending on the measurement outcome, which is sampled with probability determined by Born's rule from the current density matrix.

RELAXATION AND DEPHASING

In the case of a transmon idling for a period t, we apply the PTM $R_{\downarrow,t}$ obtained by integrating the Lindblad equation (Eq. (3.1)) for a time t. As Hamiltonian we use the bare transmon Hamiltonian

$$H = \omega b^\dagger b + \frac{\alpha}{2}(b^\dagger)^2 b^2, \tag{3.24}$$

where ω is the frequency, α is the anharmonicity and b is the annihilation operator, restricted to 2 levels if the transmon is modeled as a qubit or 3 as a qutrit ($\alpha = 0$ for qubits). The jump operators are $L_{\mathrm{amp}} = b/\sqrt{T_1}$ for relaxation (also called amplitude damping), and $\{L_{\mathrm{deph},i}\}$ given in Eqs. (6.16) and (6.18) for dephasing (or phase damping). Furthermore, we vary T_ϕ based on the frequency at which a transmon is fluxed compared to its sweetspot frequency, using Eq. (3.10) with $\sqrt{A} = 4\ \mu\Phi_0$. To make this concrete, we set $T_\phi = 30\ \mu$s at the sweetspot, which then decreases up to 6-8 μs during a CZ gate.

In the case of a gate, amplitude and phase damping are accounted for by symmetrically applying two periods of idling around the unitary operation R_{gate} (modeled as instantaneous):

$$R_{\downarrow,t_{\mathrm{gate}}/2} R_{\mathrm{gate}} R_{\downarrow,t_{\mathrm{gate}}/2}. \tag{3.25}$$

We note that R_{gate} can either be the ideal operation (as we do for single-qubit gates) or a parametrized model including unitary errors (as we do for the CZ, see below). We choose this modeling in which gates are parametrized directly by their performance parameters, rather than simulating the full Lindblad evolution, because we want to study the performance of the surface code versus the imperfections of its constituents. All the system and pulse variables, which one would need to specify in a Lindblad simulation, are unnecessary for this purpose. Furthermore, a parametrized model easily allows to perform scans over those performance parameters.

In the case of measurements we proceed similarly by applying

$$R_{\downarrow,t_{\mathrm{m}}/2} R_{\mathrm{proj}} R_{\downarrow,t_{\mathrm{m}}/2}, \tag{3.26}$$

where R_{proj} is the ideal projector on either $|0\rangle$, $|1\rangle$ or $|2\rangle$ depending on which is the measurement outcome. The latter is sampled from the current density matrix according to Born's rule.

CZ MODEL

We develop a parametrized error model for R_{gate} for the CZ gate, that reproduces the features observed from the Lindblad simulations. As discussed in general above, the usefulness of the parametrized model is that one can specify e.g. how much leakage the CZ has without having to specify all the system parameters (frequencies, pulse parameters...) and having to simulate the full dynamics until one gets the desired result. The parameters of the model are the phases acquired by the basis states and the CZ average leakage probability L_1 (see below).

For a two-qutrit gate there are 9 phases that can be imparted to the 9 basis states, meaning one applies the unitary matrix $U = \exp[i\,\text{diag}(\phi_{00},\phi_{01},\phi_{02},\phi_{10},\phi_{11},\phi_{12},\phi_{20},\phi_{21},\phi_{22})]$. One is always a global phase, so we can set $\phi_{00}=0$. For a CZ, $\phi_{01}=\phi_{10}=0$ and $\phi_{11}=\pi$. In particular, in this way the conditional phase $\phi_{2Q}=\phi_{11}-\phi_{01}-\phi_{10}+\phi_{00}$ (defined in Eq. (2.83)) is $\phi_{2Q}=\pi$, as desired for a CZ. We discuss the other phases below.

In these simulations, leakage is generated by applying the unitary V such that

$$|11\rangle \mapsto \sqrt{1-4L_1}\,|11\rangle + e^{i\phi}\sqrt{4L_1}\,|02\rangle\,, \tag{3.27}$$

$$|02\rangle \mapsto \sqrt{1-4L_1}\,|02\rangle - e^{-i\phi}\sqrt{4L_1}\,|11\rangle\,, \tag{3.28}$$

where L_1 is the average leakage probability (or rate), which by definition [38] (see also Eq. (3.43)) is averaged over the 4 computational states. Since there is almost no leakage from $|00\rangle$, $|01\rangle$ and $|10\rangle$, the leakage population escaping from $|11\rangle$ is equal to $4L_1$, as given in Eq. (3.27). Thus in total for the CZ we apply the PTM R_{gate} corresponding to the unitary VU.

If leakage states tend to decohere fast (which we observe to be the case due to the stabilizer measurements; see Section 8.10.2), then we can set the coherences between the computational and leakage subspaces to 0 after a CZ. As a result, ϕ in Eqs. (3.27) and (3.28) is irrelevant and three of the five phases involving a leaked state (ϕ_{02}, ϕ_{20}, ϕ_{21}, ϕ_{12}, ϕ_{22}) are global phases, whereas only two linear combinations affect the dynamics of the system. These two, that we call leakage conditional phases, are $\phi^{\mathscr{L}}_{\text{stat}} := \phi_{02}-\phi_{12}$ and $\phi^{\mathscr{L}}_{\text{flux}} := \phi_{20}-\phi_{21}$, i.e. they are the phases acquired by the lower and higher frequency qubit of the pair, respectively, when interacting with a leaked qubit. These are relevant (after a qubit has leaked) because, for example, if a data qubit in $(|0\rangle+|1\rangle)/\sqrt{2}$ performs a CZ with a lower-frequency ancilla qubit in $|2\rangle$, the state becomes

$$\frac{1}{\sqrt{2}}|2\rangle\otimes(|0\rangle+|1\rangle) \mapsto \frac{e^{i\phi_{20}}}{\sqrt{2}}|2\rangle\otimes(|0\rangle+e^{-i\phi^{\mathscr{L}}_{\text{flux}}}|1\rangle). \tag{3.29}$$

Thus ϕ_{20} is a global phase, whereas $\phi^{\mathscr{L}}_{\text{flux}}$ corresponds to a Z-rotation error on a data qubit.

If instead an ancilla qubit in $(|0\rangle+|1\rangle)/\sqrt{2}$ performs a CZ with a higher-frequency data qubit in $|2\rangle$, the state becomes

$$\frac{1}{\sqrt{2}}(|0\rangle+|1\rangle)\otimes|2\rangle \mapsto \frac{e^{i\phi_{02}}}{\sqrt{2}}(|0\rangle+e^{-i\phi^{\mathscr{L}}_{\text{stat}}}|1\rangle)\otimes|2\rangle\,. \tag{3.30}$$

Thus ϕ_{02} is a global phase, whereas the measurement outcome (after a Hadamard) is partially randomized depending on $\phi^{\mathscr{L}}_{\text{stat}}$. Anticipating some results in Chapter 8, the mea-

Method	Detailed information	SPAM-resistant	Scalability
Process tomography	Yes	No	No
Randomized benchmarking	No	Yes	Yes
Spectral tomography	Yes	Yes	No

Table 3.1: Summary of the main features of the gate-benchmarking methods described in Section 3.4 (process tomography and randomized benchmarking) and in Chapter 7 (spectral tomography). Detailed information refers as to whether only few quantities, such as the average gate fidelity, can be extracted, compared to specific information about errors in the gate. SPAM-resistant refers to insensitivity to State-Preparation and Measurement errors (SPAM). Scalability is about the practical applicability to gates acting on an increasing number of qubits.

surement outcome is fully randomized anyway by the anti-commutation effect (see Section 8.11.2), however, the parity of the superchecks is indeed partially randomized by $\phi^{\mathscr{L}}_{\text{stat}}$ (see Fig. 8.10).

3.4. GATE-BENCHMARKING TOOLS

In this section we discuss the advantages and disadvantages of a few of the most widespread methods to evaluate the performance of quantum gates (summarized in Table 3.1). Process tomography and randomized benchmarking are compared to spectral tomography in Chapter 7, whereas randomized benchmarking with leakage modification is used in Chapter 6 to characterize the controlled-phase gate.

3.4.1. PROCESS TOMOGRAPHY

We begin this section with state tomography, because process tomography can be reduced to state tomography with an extra constraint (see below).

PRELIMINARY: STATE TOMOGRAPHY

The most straightforward way to measure a quantum state ρ is to measure an (over)complete set of observables, from which the density matrix can be fully reconstructed. Here we consider qubits. A typical choice is to measure all the Pauli operators (except the identity). For example, for a qubit it means measuring it in the X, Y and Z basis (of course, one choice of basis per preparation of ρ), as one gets the Bloch vector in this way. In general, recall that in the Pauli Transfer Matrix formalism (PTM; see after Eq. (3.22)) one can write

$$\rho = \frac{1}{2^n} \sum_{i=0}^{2^n-1} \rho_i P_i \tag{3.31}$$

for n qubits, where $\rho_i = \text{Tr}(P_i \rho) = \langle P_i \rangle$ as in Eq. (3.22). In other words, the average values $\{\langle P_i \rangle\}$ of the Pauli operators provide a representation of ρ.

Recall that a density matrix ρ is defined by the properties $\text{Tr}(\rho) = 1$ and $\rho \geq 0$, which we refer to as the physicality constraints. In practice, the issue is that experimental measurements always have a statistical uncertainty, potentially leading to unphysical density matrices. Ensuring only the constraint $\text{Tr}(\rho) = 1$ would be easy since one could simply divide the measured ρ by its trace. However, ensuring $\rho \geq 0$ is more involved. Here we consider the maximum likelihood method [39] that searches for the physical

density matrix ρ_{ph} that is closest to the measured one, in the sense of best reproducing the measurement results. Thus one solves the optimization problem

$$\begin{aligned} \min \sum_{i=0}^{2^n-1} \left|\rho_i^{\text{meas}} - \text{Tr}(P_i \rho_{\text{ph}})\right|^2 \\ \text{subject to} \quad \text{Tr}(\rho_{\text{ph}}) = 1 \\ \rho_{\text{ph}} \geq 0. \end{aligned} \tag{3.32}$$

An implementation of an algorithm solving this convex problem can be found in Qiskit [40] via the function *cvx-fit*, based on the convex-optimization Python package *cvxpy*.

FROM STATE TO PROCESS TOMOGRAPHY

A quantum gate is ideally a unitary, whereas a noisy implementation is a superoperator $\mathscr{R}$ in general. As given in Eq. (3.23), one can associate a PTM R to such a gate. By definition, one can measure the PTM by measuring the average values of all Paulis after "preparing" each Pauli. The latter means that half of the time one prepares the $+1$-eigenstate of that Pauli, the other half the -1-eigenstate, and subtracts the two average values at the end for each output Pauli.

A superoperator $\mathscr{R}$ is defined by the properties of trace preservation and complete positivity (TPCP). In the PTM representation, the TP condition is easy to express as $R_{0j} = (1,0,\ldots,0)$, however the CP condition is not straightforward. It is thus convenient to use the Choi representation. The Choi state $\rho^{\mathscr{R}}$ is defined as

$$\rho^{\mathscr{R}} = \frac{1}{2^{2n}} \sum_{ij} R_{ij} P_j^T \otimes P_i. \tag{3.33}$$

Note that we have introduced an auxiliary subsystem of the same dimension as the original one. The TPCP conditions imply that $\rho^{\mathscr{R}}$ is a density matrix (so $\text{Tr}(\rho^{\mathscr{R}}) = 1$ and $\rho^{\mathscr{R}} \geq 0$) with the additional constraint that $\text{Tr}_1(\rho^{\mathscr{R}}) = I_2/2^n$, where Tr_1 denotes the trace over the first, auxiliary subsystem. Similarly to state tomography, the statistical uncertainty in the measurements might lead to an unphysical Choi state. The maximum-likelihood physical $\rho_{\text{ph}}^{\mathscr{R}}$ can then be found by solving the optimization problem in Eq. (3.32) with the extra constraint that $\text{Tr}_1(\rho_{\text{ph}}^{\mathscr{R}}) = I_2/2^n$, using the same methods [39–41].

AVERAGE GATE FIDELITY

Let $\mathscr{R}$ be an (arbitrary) quantum channel and $\mathscr{U}$ a unitary quantum channel. The average fidelity $F(\mathscr{R},\mathscr{U})$ between the two channels is defined as

$$F(\mathscr{R},\mathscr{U}) = \int_{\text{Haar}} d\psi \left(\mathscr{U}(|\psi\rangle\langle\psi|)\right)^\dagger \mathscr{R}(|\psi\rangle\langle\psi|). \tag{3.34}$$

As unitary channels are invertible, it holds that $F(\mathscr{R},\mathscr{U}) = F(\mathscr{U}^\dagger\mathscr{R},\mathscr{I}) =: F(\mathscr{U}^\dagger\mathscr{R})$, where $\mathscr{I}$ is the identity channel. When $\mathscr{U}$ is an ideal target gate and $\mathscr{R}$ is a noisy implementation, we refer to F as the average gate fidelity. One can show [42] that

$$F(\mathscr{R}) = \frac{1}{d_1+1}\left(\frac{\text{Tr}(R)}{d_1} + 1\right), \tag{3.35}$$

where R can be either the PTM or Liouville representation of $\mathscr{R}$, and where d_1 is the Hilbert-space dimension. Here $d_1 = 2^n$ for n qubits. In the presence of leakage, the definition of F is generalized in Eq. (3.47).

PROS AND CONS OF PROCESS TOMOGRAPHY

Process tomography can be used to extract the average gate fidelity as

$$F = \frac{2^n + \mathrm{Tr}(R_{\mathrm{ideal}}^{\dagger} R_{\mathrm{ph}})}{2^n(2^n+1)}, \tag{3.36}$$

where R_{ideal} is the PTM of the ideal unitary gate and R_{ph} is the PTM corresponding to the Choi state found by the optimization above. However, process tomography provides much more information since it provides indeed the whole PTM of the gate, which allows one to evaluate its effect on any state. For example, in the case of single-qubit gates, as they correspond to rotations, one can estimate deviations from the intended angle or axis of rotation.

Process tomography has, however, two main disadvantages. The first is that it is sensitive to state-preparation and measurement errors (SPAM). Indeed, one cannot really discern whether the gate is faulty, or whether the state preparation or the measurement used to evaluate the PTM are faulty. As a consequence, in experiment one cannot decide which of these operations need to be improved the most.
The second disadvantage is that process tomography is not scalable. The number of Paulis that need to be "prepared" and measured (and thus the time to do that) scales exponentially with the number of qubits. While 4-qubit tomography was implemented in Ref. [23], it is not practically feasible to consider systems with a larger number of qubits.

3.4.2. RANDOMIZED BENCHMARKING

In its standard version, randomized benchmarking [43, 44] is not a tool to characterize the performance of a single gate, but rather of a set of gates forming a group (technically, it needs to be a 2-design). Typically one considers the Clifford group, getting an average fidelity for all the gates in this group.

The randomized-benchmarking protocol prescribes as follows:

1. Randomly choose m Clifford gates $\mathscr{C}_1, \ldots, \mathscr{C}_m$ and construct the sequence $\mathscr{G}_m = \mathscr{C}_{m+1} \circ \mathscr{C}_m \circ \ldots \circ \mathscr{C}_1$, where $\mathscr{C}_{m+1} := \mathscr{C}_1^{\dagger} \circ \ldots \circ \mathscr{C}_m^{\dagger}$;

2. Prepare the qubit(s) in a fixed initial state ρ_0 (assume that $\rho_0 = |0\ldots0\rangle\langle 0\ldots0|$ as it is usually the case). Apply the gates in the sequence one by one, obtaining $\rho_m = \mathscr{G}_m(\rho_0)$;

3. Perform a measurement to estimate the recovery probability $p_0(\rho_m) := \mathrm{Tr}(M_0\rho_m)$ up to a suitable precision, where the measurement operator is $M_0 = |0\ldots0\rangle\langle 0\ldots0|$. Note that if all gates are ideal $\rho_m = \rho_0$ and $p_0 = 1$;

4. Repeat steps 1-3 for N times for different random sequences to obtain the empirical average $p_0(m) := \mathbb{E}_{\mathscr{G}_m}[p_0(\rho_m)]$;

5. Repeat steps 1-4 for different m's;

6. Perform a fit to the decay model $p_0(m) = A_0 + C_0\lambda_2^m$.

Then the average gate fidelity over the group can be estimated as

$$F = \frac{1 + (d_1 - 1)\lambda_2}{d_1}, \tag{3.37}$$

where d_1 is the dimension of the (multi-)qubit Hilbert space.

ASSUMPTIONS ON THE NOISE

The validity of Eq. (3.37) hangs on a few assumptions. The most important is that the noise is gate-independent, i.e. that all noisy gates $\tilde{\mathscr{C}}_j$ can be written as

$$\tilde{\mathscr{C}}_j = \mathscr{E} \circ \mathscr{C}_j \tag{3.38}$$

for the same noise channel $\mathscr{E}$. This assumption is not really expected to be satisfied in practice, although some weak gate dependence can be allowed [44]. Other assumptions are Markovianity (implicit in Eq. (3.38) since we have assumed that noise was a TPCP map), the fact that the quality of the measurement is independent of the sequence (length), and the fact that there is no leakage. In Section 3.4.3 we review a modified protocol that removes the last assumption and even gives an estimate for leakage.

INTERLEAVED RANDOMIZED BENCHMARKING

Interleaved randomized benchmarking [45] was introduced to evaluate the average gate fidelity of a specific Clifford gate $\bar{\mathscr{C}}$, rather than an average over the whole group. One first executes the standard randomized-benchmarking protocol as given above, getting a certain value for λ_2. Then one executes the same protocol except for replacing the sequences in step 1 by $\mathscr{G}_m^{\text{Int}} = \mathscr{C}_{m+1} \circ \bar{\mathscr{C}} \circ \mathscr{C}_m \circ \bar{\mathscr{C}} \circ \ldots \circ \bar{\mathscr{C}} \circ \mathscr{C}_1$, where now $\mathscr{C}_{m+1}$ is chosen to invert all the previous Cliffords and not only the random ones. One then gets a value for λ_2^{Int}. Under the assumption that the average error channel is depolarizing, the average gate fidelity of $\bar{\mathscr{C}}$ is then estimated as

$$F = \frac{1 + (d_1 - 1)\frac{\lambda_2^{\text{Int}}}{\lambda_2}}{d_1}. \tag{3.39}$$

The error E on this estimation should not be extracted from the fit routine but it is

$$E = \min \begin{cases} \frac{(d_1-1)\left|\lambda_2 - \frac{\lambda_2^{\text{Int}}}{\lambda_2}\right| + (1-\lambda_2)}{d_1} \\ \frac{2(d_1^2-1)(1-\lambda_2)}{\lambda_2 d_1^2} + \frac{4\sqrt{1-\lambda_2}\sqrt{d_1^2-1}}{\lambda_2}. \end{cases} \tag{3.40}$$

In other words, the 'true' average gate fidelity of $\bar{\mathscr{C}}$ lies in $[F - E, F + E]$.

PROS AND CONS OF RANDOMIZED BENCHMARKING

With respect to process tomography (see Section 3.4.1), randomized benchmarking has two main advantages. First, the repetition of many gates (between the preparation and

measurement steps) magnifies the gate errors, compared to the SPAM errors that remain fixed. This makes randomized benchmarking SPAM-resistant (similarly to how spectral tomography in Chapter 7 is SPAM-resistant as well). The second main advantage is that randomized benchmarking is efficient with respect to the number of qubits and thus scalable. Furthermore, it is also efficient in terms of real time required to perform a randomized-benchmarking experiment (see Section 7.3.2 for an estimate of the required resources).

The downside is that randomized benchmarking provides only average information about a gate, namely the average gate fidelity over the whole group or for a specific gate, as well as the average leakage rate in the modified protocol presented in Section 3.4.3 below. However, no specific error diagnosis is possible.

3.4.3. RANDOMIZED BENCHMARKING WITH LEAKAGE MODIFICATION

The modified randomized-benchmarking protocol to characterize leakage as well was introduced in Ref. [38].

DEFINITIONS

We divide the overall Hilbert space of the qudits involved in the gate into the computational subspace $\mathscr{X}_1$ and the leakage subspace $\mathscr{X}_2$ (denoted as $\mathscr{C}$ and $\mathscr{L}$, respectively, in the rest of this thesis), with projectors Π_1 and Π_2, respectively. Let $d_1 = \dim \mathscr{X}_1$ and $d_2 = \dim \mathscr{X}_2$. We define the leakage of a density matrix as

$$\mathbb{L}(\rho) = \mathrm{Tr}(\Pi_2 \rho) = 1 - \mathrm{Tr}(\Pi_1 \rho). \tag{3.41}$$

For a superoperator $\mathscr{R}$, we define the average leakage rate L_1 and the average seepage rate L_2 as

$$L_1 := \int_{|\psi_1\rangle \in \mathscr{X}_1} d\psi_1 \, \mathbb{L}\big(\mathscr{R}(|\psi_1\rangle\langle\psi_1|)\big) = \mathbb{L}\Big(\mathscr{R}\Big(\frac{\Pi_1}{d_1}\Big)\Big) \tag{3.42}$$

$$= 1 - \frac{1}{d_1} \sum_{i=0}^{d_1-1} \mathrm{Tr}\big(\Pi_1 \mathscr{R}(|i\rangle\langle i|)\big), \tag{3.43}$$

$$L_2 := \int_{|\psi_2\rangle \in \mathscr{X}_2} d\psi_2 \, \mathbb{L}\big(\mathscr{R}(|\psi_2\rangle\langle\psi_2|)\big) = 1 - \mathbb{L}\Big(\mathscr{R}\Big(\frac{\Pi_2}{d_2}\Big)\Big) \tag{3.44}$$

$$= \frac{1}{d_2} \sum_{i=d_2}^{d_1+d_2-1} \mathrm{Tr}\big(\Pi_1 \mathscr{R}(|i\rangle\langle i|)\big). \tag{3.45}$$

Thus, in particular L_1 is the leakage averaged over the basis states in the computational subspace.

The average gate fidelity in the computational subspace is then defined as

$$F(\mathscr{R}) := \int_{|\psi_1\rangle \in \mathscr{X}_1} d\psi_1 \, \langle\psi_1|\mathscr{R}(|\psi_1\rangle\langle\psi_1|)|\psi_1\rangle. \tag{3.46}$$

One can also express it as [38]

$$F(\mathscr{R}) = \frac{1}{d_1+1}\left(\frac{\mathrm{Tr}(R_{\Pi_1}R)}{d_1} + 1 - L_1\right). \tag{3.47}$$

$$= \frac{1}{d_1+1}\left(\frac{\sum_k |\mathrm{Tr}(\Pi_1 K_k)|^2}{d_1} + 1 - L_1\right), \tag{3.48}$$

where R_{Π_1} is the PTM of the projector Π_1 and where the $\{K_k\}$ are the Kraus operators of $\mathscr{R}$.

3

PROTOCOL

The randomized-benchmarking protocol with leakage modification prescribes as follows:

1. Randomly choose m Clifford gates $\mathscr{C}_1, \ldots, \mathscr{C}_m$ and construct the sequence $\mathscr{G}_m = \mathscr{C}_{m+1} \circ \mathscr{C}_m \circ \ldots \circ \mathscr{C}_1$, where $\mathscr{C}_{m+1} := \mathscr{C}_1^\dagger \circ \ldots \circ \mathscr{C}_m^\dagger$;

2. Prepare the qubit(s) in a fixed initial state ρ_0 (assume that $\rho_0 = |0\ldots0\rangle\langle 0\ldots0|$ as it is usually the case). Apply the gates in the sequence one by one, obtaining $\rho_m = \mathscr{G}_m(\rho_0)$;

3. Measure the probabilities $p_j(\rho_m) := \mathrm{Tr}(M_j\rho_m)$ up to a suitable precision, where the measurement operators $\{M_j\}$ correspond to a projective measurement over computational states, with $j \in \{0, \ldots, d_1 - 1\}$.
   Obtain an estimate of the population in $\mathscr{X}_1$ as $p_{\mathscr{X}_1}(\rho_m) = \sum_{j=0}^{d_1-1} p_j(\rho_m) = \mathrm{Tr}(\Pi_1\rho_m)$;

4. Repeat steps 1-3 for N times for different random sequences to obtain the empirical averages $p_0(m) := \mathbb{E}_{\mathscr{G}_m}[p_0(\rho_m)]$ and $p_{\mathscr{X}_1}(m) := \mathbb{E}_{\mathscr{G}_m}[p_{\mathscr{X}_1}(\rho_m)]$;

5. Repeat steps 1-4 for different m's;

6. Perform a fit to the decay model $p_{\mathscr{X}_1}(m) = A + B\lambda_1^m$ and $p_0(m) = A_0 + B_0\lambda_1^m + C_0\lambda_2^m$.

Then the average gate fidelity and the average leakage and seepage rates over the group can be estimated as

$$L_1 = (1-A)(1-\lambda_1) \tag{3.49}$$

$$L_2 = A(1-\lambda_1) \tag{3.50}$$

$$F = \frac{1 + (d_1-1)\lambda_2 - L_1}{d_1}. \tag{3.51}$$

Compared to the standard randomized-benchmarking protocol in Section 3.4.2, this protocol differs in step 3, in the fact that in step 4 one estimates not only $p_0(m)$ but also $p_{\mathscr{X}_1}(m)$, and in the fit model. The crucial difference lies in step 3, in which one does not measure only the population in $|0\rangle$, but also in all other computational states. Implicitly, step 3 assumes that the measurement distinguishes these states from the leakage subspace, i.e. that there is a measurement operator $M_{\mathscr{X}_2}$ that returns a distinct value for "leaked". However, measurements in transmons (see Sections 2.6.2 and 8.11.1) typically provide the readout $|0\rangle \to$ "0", $|1\rangle \to$ "1" and $|2\rangle \to$ "1", thus one cannot distinguish a $|1\rangle$ from the leaked state $|2\rangle$ (assume that this is the only state in the leakage subspace).

In either way, this does not allow one to perform step 3. To go around this issue, one can use the procedure introduced in Ref. [46] and reviewed in Section 6.11.7. In short, in this procedure one doubles the number of experiments, performing half of them in the normal way, and performing the other half by appending a π pulse (X gate) just before the measurement. The π pulse maps $|0\rangle$ to $|1\rangle$ and vice versa, but leaves $|2\rangle$ invariant. Roughly speaking, this leads to differences in the measurement statistics that allow one to estimate $p_2(\rho_m) = 1 - p_{\mathscr{X}_1}(\rho_m)$.

ASSUMPTIONS ON LEAKAGE

Beside the other assumptions outlined in Section 3.4.2 for randomized benchmarking, the results in this section require that averaging over Clifford sequences also averages out coherences between the computational and leakage subspace. If this assumption is violated, one might observe oscillations in $p_{\mathscr{X}_2}(m)$, however, for small amounts of leftover coherence, one does not observe oscillations but tends to overestimate L_1 [38].

In Chapter 6 we use this protocol (in its interleaved version) to characterize the CZ. Recall that the CZ is a Clifford gate. Furthermore, as it is the primary two-qubit gate in a transmon architecture, random Cliffords are compiled out of CZs and single-qubit Clifford gates, with an average of 1.5 CZs per Clifford [47, 48]. Leakage during the CZ (see Section 2.8.2) is coherent, thus *per se* the coherences are not 0. However, away from the interaction point (during single-qubit gates or idling steps) the phase of $|2\rangle$ evolves fast and it is not tracked, leading to depolarization of the leakage subspace. We expect that averaging over Cliffords leads to a further suppression of these coherences. This is corroborated by the fact that in experiment we do not observe oscillations of $p_{\mathscr{X}_2}(m)$ (see Figs. 6.5 and 6.18 and more plots in Ref. [49]). Based on the considerations above, in Chapter 6 we assume that the amount of leftover coherence is negligible.

REFERENCES

[1] H. P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press, Oxford, 2002).

[2] P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *A quantum engineer's guide to superconducting qubits,* Applied Physics Reviews **6**, 021318 (2019).

[3] E. Milotti, *1/f noise: a pedagogical review,* (2002), arXiv:physics/0204033 [physics.class-ph] .

[4] C. Wang, C. Axline, Y. Y. Gao, T. Brecht, Y. Chu, L. Frunzio, M. H. Devoret, and R. J. Schoelkopf, *Surface participation and dielectric loss in superconducting qubits,* Applied Physics Letters **107**, 162601 (2015).

[5] O. Dial, D. T. McClure, S. Poletto, G. A. Keefe, M. B. Rothwell, J. M. Gambetta, D. W. Abraham, J. M. Chow, and M. Steffen, *Bulk and surface loss in superconducting transmon qubits,* Superconductor Science and Technology **29**, 044001 (2016).

[6] P. V. Klimov, J. Kelly, Z. Chen, M. Neeley, A. Megrant, B. Burkett, R. Barends, K. Arya, B. Chiaro, Y. Chen, A. Dunsworth, A. Fowler, B. Foxen, C. Gidney, M. Giustina,

R. Graff, T. Huang, E. Jeffrey, E. Lucero, J. Y. Mutus, O. Naaman, C. Neill, C. Quintana, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, T. C. White, S. Boixo, R. Babbush, V. N. Smelyanskiy, H. Neven, and J. M. Martinis, *Fluctuations of energy-relaxation times in superconducting qubits,* Phys. Rev. Lett. **121**, 090502 (2018).

[7] F. Luthi, T. Stavenga, O. W. Enzing, A. Bruno, C. Dickel, N. K. Langford, M. A. Rol, T. S. Jespersen, J. Nygård, P. Krogstrup, and L. DiCarlo, *Evolution of nanowire transmon qubits and their coherence in a magnetic field,* Phys. Rev. Lett. **120**, 100502 (2018).

[8] P. Kumar, S. Sendelbach, M. Beck, J. Freeland, Z. Wang, H. Wang, C. Y. Clare, R. Wu, D. Pappas, and R. McDermott, *Origin and reduction of 1/f magnetic flux noise in superconducting devices,* Phys. Rev. Appl. **6**, 041001 (2016).

[9] S. de Graaf, A. Adamyan, T. Lindström, D. Erts, S. Kubatkin, A. Y. Tzalenchuk, and A. Danilov, *Direct identification of dilute surface spins on Al_2O_3 : Origin of flux noise in quantum circuits,* Physical Review Letters **118** (2017), 10.1103/physrevlett.118.057703.

[10] A. P. Vepsäläinen, A. H. Karamlou, J. L. Orrell, A. S. Dogra, B. Loer, F. Vasconcelos, D. K. Kim, A. J. Melville, B. M. Niedzielski, J. L. Yoder, and et al., *Impact of ionizing radiation on superconducting qubit coherence,* Nature **584**, 551–556 (2020).

[11] S. Gustavsson, F. Yan, G. Catelani, J. Bylander, A. Kamal, J. Birenbaum, D. Hover, D. Rosenberg, G. Samach, A. P. Sears, S. J. Weber, J. L. Yoder, J. Clarke, A. J. Kerman, F. Yoshihara, Y. Nakamura, T. P. Orlando, and W. D. Oliver, *Suppressing relaxation in superconducting qubits by quasiparticle pumping,* Science **354**, 1573 (2016).

[12] K. Serniak, M. Hays, G. de Lange, S. Diamond, S. Shankar, L. Burkhart, L. Frunzio, M. Houzet, and M. Devoret, *Hot nonequilibrium quasiparticles in transmon qubits,* Physical Review Letters **121** (2018), 10.1103/physrevlett.121.157701.

[13] L. Cardani, F. Valenti, N. Casali, G. Catelani, T. Charpentier, M. Clemenza, I. Colantoni, A. Cruciani, G. D'Imperio, L. Gironi, and et al., *Reducing the impact of radioactivity on quantum circuits in a deep-underground facility,* Nature Communications **12** (2021), 10.1038/s41467-021-23032-z.

[14] M. McEwen, L. Faoro, K. Arya, A. Dunsworth, T. Huang, S. Kim, B. Burkett, A. Fowler, F. Arute, J. C. Bardin, A. Bengtsson, A. Bilmes, B. B. Buckley, N. Bushnell, Z. Chen, R. Collins, S. Demura, A. R. Derk, C. Erickson, M. Giustina, S. D. Harrington, S. Hong, E. Jeffrey, J. Kelly, P. V. Klimov, F. Kostritsa, P. Laptev, A. Locharla, X. Mi, K. C. Miao, S. Montazeri, J. Mutus, O. Naaman, M. Neeley, C. Neill, A. Opremcak, C. Quintana, N. Redd, P. Roushan, D. Sank, K. J. Satzinger, V. Shvarts, T. White, Z. J. Yao, P. Yeh, J. Yoo, Y. Chen, V. Smelyanskiy, J. M. Martinis, H. Neven, A. Megrant, L. Ioffe, and R. Barends, *Resolving catastrophic error bursts from cosmic rays in large arrays of superconducting qubits,* (2021), arXiv:2104.05219 [quant-ph] .

[15] A. Somoroff, Q. Ficheux, R. A. Mencia, H. Xiong, R. V. Kuzmin, and V. E. Manucharyan, *Millisecond coherence in a superconducting qubit,* (2021), arXiv:2103.08578 [quant-ph] .

[16] M. A. Rol, L. Ciorciaro, F. K. Malinowski, B. M. Tarasinski, R. E. Sagastizabal, C. C. Bultink, Y. Salathe, N. Haandbaek, J. Sedivy, and L. DiCarlo, *Time-domain characterization and correction of on-chip distortion of control pulses in a quantum processor,* Applied Physics Letters **116**, 054001 (2020).

[17] J. M. Martinis and M. R. Geller, *Fast adiabatic qubit gates using only σ_z control,* Phys. Rev. A **90**, 022307 (2014).

[18] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, *Simple all-microwave entangling gate for fixed-frequency superconducting qubits,* Phys. Rev. Lett. **107**, 080502 (2011).

[19] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, *Procedure for systematically tuning up cross-talk in the cross-resonance gate,* Physical Review A **93**, 060302 (2016).

[20] S. A. Caldwell, N. Didier, C. A. Ryan, E. A. Sete, A. Hudson, P. Karalekas, R. Manenti, M. P. da Silva, R. Sinclair, E. Acala, N. Alidoust, J. Angeles, A. Bestwick, M. Block, B. Bloom, A. Bradley, C. Bui, L. Capelluto, R. Chilcott, J. Cordova, G. Crossman, M. Curtis, S. Deshpande, T. E. Bouayadi, D. Girshovich, S. Hong, K. Kuang, M. Lenihan, T. Manning, A. Marchenkov, J. Marshall, R. Maydra, Y. Mohan, W. O'Brien, C. Osborn, J. Otterbach, A. Papageorge, J.-P. Paquette, M. Pelstring, A. Polloreno, G. Prawiroatmodjo, V. Rawat, M. Reagor, R. Renzas, N. Rubin, D. Russell, M. Rust, D. Scarabelli, M. Scheer, M. Selvanayagam, R. Smith, A. Staley, M. Suska, N. Tezak, D. C. Thompson, T.-W. To, M. Vahidpour, N. Vodrahalli, T. Whyland, K. Yadav, W. Zeng, and C. Rigetti, *Parametrically activated entangling gates using transmon qubits,* Phys. Rev. Applied **10**, 034050 (2018).

[21] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M. P. da Silva, and B. R. Johnson, *Demonstration of a parametrically activated entangling gate protected from flux noise,* Physical Review A **101** (2020).

[22] D. Sank, Z. Chen, M. Khezri, J. Kelly, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. Fowler, E. Jeffrey, E. Lucero, A. Megrant, J. Mutus, M. Neeley, C. Neill, P. J. J. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, T. White, J. Wenner, A. N. Korotkov, and J. M. Martinis, *Measurement-induced state transitions in a superconducting qubit: Beyond the rotating wave approximation,* Phys. Rev. Lett. **117**, 190503 (2016).

[23] J. F. Marques, B. M. Varbanov, M. S. Moreira, H. Ali, N. Muthusubramanian, C. Zachariadis, F. Battistel, M. Beekman, N. Haider, W. Vlothuizen, A. Bruno, B. M. Terhal, and L. DiCarlo, *Logical-qubit operations in an error-detecting surface code,* Nature Physics (2021).

[24] Z. Chen, A. Megrant, J. Kelly, R. Barends, J. Bochmann, Y. Chen, B. Chiaro, A. Dunsworth, E. Jeffrey, J. Y. Mutus, and et al., *Fabrication and characterization of aluminum airbridges for superconducting microwave circuits,* Applied Physics Letters **104**, 052602 (2014).

[25] V. Tripathi, H. Chen, M. Khezri, K.-W. Yip, E. M. Levenson-Falk, and D. A. Lidar, *Suppression of crosstalk in superconducting qubits using dynamical decoupling*, (2021), arXiv:2108.04530 [quant-ph] .

[26] C. C. Bultink, T. E. O'Brien, R. Vollmer, N. Muthusubramanian, M. W. Beekman, M. A. Rol, X. Fu, B. Tarasinski, V. Ostroukh, B. Varbanov, A. Bruno, and L. DiCarlo, *Protecting quantum entanglement from leakage and qubit errors via repetitive parity measurements*, Science Advances **6**, eaay3050 (2020).

[27] Y. Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, B. Campbell, Z. Chen, B. Chiaro, and et al., *Qubit architecture with high coherence and fast tunable coupling*, Physical Review Letters **113** (2014), 10.1103/physrevlett.113.220502.

[28] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Tunable Coupling Scheme for Implementing High-Fidelity Two-Qubit Gates*, Physical Review Applied **10**, 054062 (2018).

[29] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler*, Phys. Rev. X **11**, 021058 (2021).

[30] B. K. Mitchell, R. K. Naik, A. Morvan, A. Hashim, J. M. Kreikebaum, B. Marinelli, W. Lavrijsen, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, *Hardware-efficient microwave-activated tunable coupling between superconducting qubits*, (2021), arXiv:2105.05384 [quant-ph] .

[31] E. A. Sete, A. Q. Chen, R. Manenti, S. Kulshreshtha, and S. Poletto, *Floating tunable coupler for scalable quantum computing architectures*, Physical Review Applied **15** (2021), 10.1103/physrevapplied.15.064063.

[32] J. Stehlik, D. Zajac, D. Underwood, T. Phung, J. Blair, S. Carnevale, D. Klaus, G. Keefe, A. Carniol, M. Kumph, and et al., *Tunable coupling architecture for fixed-frequency transmon superconducting qubits*, Physical Review Letters **127** (2021), 10.1103/physrevlett.127.080505.

[33] M. Khezri, J. Dressel, and A. N. Korotkov, *Qubit measurement error from coupling with a detuned neighbor in circuit QED*, Phys. Rev. A **92**, 052306 (2015).

[34] J. C. Pommerening and D. P. DiVincenzo, *What is measured when a qubit measurement is performed on a multiqubit chip*, Phys. Rev. A **102**, 032623 (2020).

[35] The quantumsim package can be found at `https://quantumsim.gitlab.io/`.

[36] T. E. O'Brien, B. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Information **3** (2017), 10.1038/s41534-017-0039-x.

[37] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, *Scalable quantum circuit and control for a superconducting surface code,* Phys. Rev. Appl. **8**, 034021 (2017).

[38] C. J. Wood and J. M. Gambetta, *Quantification and characterization of leakage errors,* Phys. Rev. A **97**, 032306 (2018).

[39] J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, M. B. Ketchen, and M. Steffen, *Universal quantum gate set approaching fault-tolerant thresholds with superconducting qubits,* Phys. Rev. Lett. **109**, 060501 (2012).

[40] M. S. ANIS, H. Abraham, AduOffei, R. Agarwal, G. Agliardi, M. Aharoni, I. Y. Akhalwaya, G. Aleksandrowicz, T. Alexander, M. Amy, *et al.*, *Qiskit: An open-source framework for quantum computing,* (2021), 10.5281/zenodo.2573505.

[41] J. de Jong, *Implementation of a fault-tolerant SWAP operation on the IBM 5-qubit device,* Master's thesis, Technical University of Delft (2019).

[42] M. A. Nielsen, *A simple formula for the average gate fidelity of a quantum dynamical operation,* Physics Letters A **303**, 249 (2002).

[43] E. Magesan, J. M. Gambetta, and J. Emerson, *Scalable and robust randomized benchmarking of quantum processes,* Phys. Rev. Lett. **106**, 180504 (2011).

[44] E. Magesan, J. M. Gambetta, and J. Emerson, *Characterizing quantum gates via randomized benchmarking,* Phys. Rev. A **85**, 042311 (2012).

[45] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, *Efficient measurement of quantum gate error by interleaved randomized benchmarking,* Phys. Rev. Lett. **109**, 080505 (2012).

[46] S. Asaad, C. Dickel, S. Poletto, A. Bruno, N. K. Langford, M. A. Rol, D. Deurloo, and L. DiCarlo, *Independent, extensible control of same-frequency superconducting qubits by selective broadcasting,* npj Quantum Inf. **2**, 16029 (2016).

[47] A. D. Córcoles, J. M. Gambetta, J. M. Chow, J. A. Smolin, M. Ware, J. Strand, B. L. T. Plourde, and M. Steffen, *Process verification of two-qubit quantum gates by randomized benchmarking,* Phys. Rev. A **87**, 030301 (2013).

[48] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and J. M. Martinis, *Superconducting quantum circuits at the surface code threshold for fault tolerance.* Nature **508**, 500 (2014).

[49] V. Negîrneac, H. Ali, N. Muthusubramanian, F. Battistel, R. Sagastizabal, M. S. Moreira, J. F. Marques, W. J. Vlothuizen, M. Beekman, C. Zachariadis, N. Haider, A. Bruno, and L. DiCarlo, *High-fidelity controlled-Z gate with maximal intermediate leakage*

operating at the speed limit in a superconducting quantum processor, Phys. Rev. Lett. **126**, 220502 (2021).

4

QUANTUM ERROR CORRECTION

As discussed in Chapter 3, many noise sources affect the quality of superconducting qubits and operations performed on them. Improving their quality is of fundamental importance towards building a functional quantum computer. Running a quantum computation for a few days or weeks will require error rates below $\sim 10^{-12}$. However, it seems unlikely that the field of quantum computing will manage to implement a qubit with such a low physical error rate, at least in the next few years. This might have been the thought also at the beginning of the classical-computing era, before the invention of nowadays trustworthy transistors, but we do not know if quantum computing will ever follow the same course. Thus, large quantum processors will likely need to use quantum error correction (QEC) [1–4] , discussed in this chapter, to implement logical qubits with very low logical error rates.

In short, in QEC the logical information is spread redundantly across many physical qubits to help protecting it, assuming that the noise acts locally. Then information about errors is extracted without destroying the encoded state and is used to try to correct these errors. We comment on the fact that QEC can be interpreted as a process that lowers the entropy of the system by collecting information about the errors, while the entropy tends to increase over time due to the accumulation of errors. If the entropy is kept low enough, even accounting for the entropy added by imperfect error correction, one can extend the quantum computation.

Except for Section 4.6, in the following we consider qubits and an independent Pauli error model.

4.1. QUANTUM ERROR CORRECTING CODES

For a system with Hilbert space $\mathscr{H}$, a quantum error correcting code is defined as a subspace $\mathscr{H}_L \subseteq \mathscr{H}$, also called the logical subspace. A QEC code is characterized by three main parameters:

- the number of physical qubits $n = \log_2(\dim \mathscr{H})$;
- the number of logical qubits $k = \log_2(\dim \mathscr{H}_L)$;

- the code distance d, defined as the minimum weight of a logical operator, i.e. the minimum number of qubits which non-trivially support an operator $O_L \neq \Pi_L$ such that $[O_L, \Pi_L] = 0$, where Π_L is the projector onto $\mathcal{H}_L$ (and is thus the logical identity within $\mathcal{H}_L$).

One of the difficult parts in defining a useful QEC code is to find a logical subspace such that the distance is large. Even more importantly, one wants to identify a family of QEC codes $\{\mathcal{H}_L^{(n)}\}$ such that d grows substantially with n (and k remains at least fixed or rather grows too).

We note that these are subspace quantum error correcting codes, but one can also define subsystem codes, in which $\mathcal{H}_L = \mathcal{H}_{L,\text{sys}} \otimes \mathcal{H}_G$. There, $\mathcal{H}_{L,\text{sys}}$ is the logical subsystem, whereas $\mathcal{H}_G$ contains logical qubits that have been "sacrificed" and are referred to as gauge qubits. Errors on the gauge qubits are irrelevant, so this might help protecting the remaining logical qubits.

4.2. STABILIZER CODES

The most studied type of QEC codes are stabilizer codes. The n-qubit Pauli group $\mathcal{G}_n$ is the group of tensor products of n Pauli operators (I, X, Y, Z), each multiplied by either $\{1, -1, i, -i\}$. A subgroup $\mathcal{S}_n$ of $\mathcal{G}_n$ is a stabilizer group if it is abelian and does not contain $-I^{\otimes n}$. $\mathcal{S}_n$ can be characterized by a set of independent generators $\{S_1, \ldots, S_{n-k}\}$. This set is not unique but it is usually chosen as the set where each generator has at most a certain weight. Each element of $\mathcal{S}_n$ is a stabilizer (although one often uses the word stabilizer to specifically refer to the generators, also called parity checks). A stabilizer code is the simultaneous $+1$-eigenspace of all the stabilizers. In particular, Π_L can be written as

$$\Pi_L = (I + S_1)\ldots(I + S_{n-k})/2^{n-k}. \tag{4.1}$$

Stabilizer codes are built to deal with Pauli errors. While these errors are not necessarily the most common per se, the Knill-Laflamme conditions [5] ensure that any linear combination of correctable errors is also correctable. Thus, in particular, if all Pauli errors up to a certain weight are correctable, then any error up to that weight is correctable as well. However, if there are more levels above the qubit levels and the state can leak to those levels, then the correctability of Pauli errors does not imply that leakage errors are correctable as well (see Section 4.6).

Any given (multi-qubit) Pauli P satisfies either $[P, S_j] = 0$ or $\{P, S_j\} = 0$ for each S_j. Thus if we were to measure S_j we would find either the measurement outcome $s_j = +1$ or $s_j = -1$, respectively, since

$$S_j(P|\psi_L\rangle) = \pm P S_j |\psi_L\rangle = \pm P|\psi_L\rangle = s_j(P|\psi_L\rangle). \tag{4.2}$$

The collection of all s_j is called the syndrome $\vec{s}$. Since different Paulis might give a different syndrome, measuring stabilizers provides some information about errors. However, any two Paulis that differ by the application of a stabilizer, or of a logical operator, lead to the same syndrome $\vec{s}$, thus it is not immediate what should be the best correction P_c to be applied given a certain syndrome. This is the task of the decoder (see Section 4.4). In general, what might go wrong is that $P_c P$ is a logical operator, different from the identity, which is undetectable by definition and leads to corruption of the logical information.

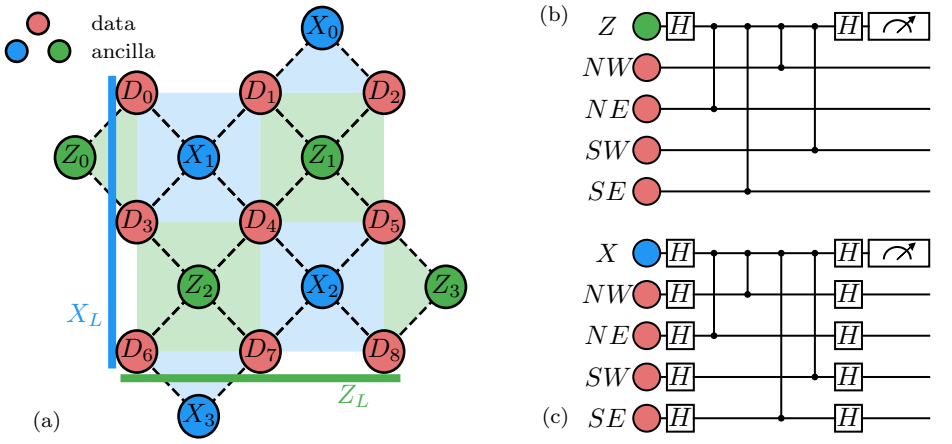

Figure 4.1: The surface code. (a) A small (distance-3) instance of the (rotated) surface code. Larger distances can be obtained by extending the pattern of X-type (blue) and Z-type (green) stabilizers. We comment that, in the scheme in Ref. [8], data qubits (red) have two different frequencies (see Fig. 8.2). Two representatives of the logical operators are displayed via either a vertical line ($X_L = X_0 X_3 X_6$) or horizontal line ($Z_L = Z_6 Z_7 Z_8$). Other representatives are obtained by multiplying these two with stabilizers. (b,c) Parity-check units for Z (b) and X (c) ancilla qubits with fault-tolerant ordering of the CZ gates.

In practice, one uses some extra qubits on the chip, called ancilla qubits, to measure the stabilizers. While in principle one could engineer a multi-qubit interaction (between > 2 qubits), this is challenging in experiment. Thus the stabilizers are generally measured by performing a series of two-qubit gates between the ancilla and one data qubit at the time. These gates are designed to collect the syndrome bit onto the ancilla qubit, which is subsequently measured to reveal it. To perform these gates, the ancilla qubits need to have a physical connection to the data qubits. Since long-distance connections are not really implementable in many quantum-computing platforms, including superconducting qubits (but not trapped ions [6]), this requirement poses locality constraints on the stabilizer codes that one may consider. It has been proven [7] that local 2-dimensional stabilizer codes (in Euclidean space) have a distance limited to $d \sim \mathcal{O}(\sqrt{n/k})$. Thus it is not linear in n as it would ideally be (see Section 4.1). However, this does not preclude the existence of a threshold (see Section 4.5) for e.g. the surface code, for which $d = \sqrt{n}$, although this limits k to be a constant in that case.

4.2.1. SURFACE CODE

One of the most well-known stabilizer codes is the surface code [9] (see Fig. 4.1), which is the flat version of Kitaev's toric code [10]. The surface code is the privileged choice in many experimental groups due to the 2D layout with nearest-neighbor interactions, which makes it straightforwardly amenable to implementation. Each ancilla qubit is connected to 4 data qubits in the bulk and to 2 at the boundary. In its original version, the stabilizers are products of only either X's or Z's. In this way one uses half of the stabilizers to correct for Z errors and the other half for X errors. Since $Y \sim XZ$, Y errors

are handled by correcting X and Z independently. We note that a recent version of the surface code, the so-called $XZZX$ code [11], considers only one kind of stabilizer which is made of half X's and half Z's. The impact of such a relatively simple modification leads to significant advantages whenever the noise is biased towards X or Z errors and if the two-qubit gates preserve the bias [12].

The logical operators of the surface code correspond to a product of X's along a full vertical edge for X_L (see Fig. 4.1(a)), and to a product of Z's along a full horizontal edge for Z_L (as well as any other representation obtained by multiplying these with a stabilizer). Note that upon multiplication by X-type stabilizers, X_L continues to connect the upper and lower boundaries is a snake-like form (similarly for Z_L). Because of this invariant feature upon deformation, the surface code falls within the class of topological codes. One could say that the protection provided by the surface code originates precisely from this topological property: a logical operator must cross the lattice from one side to the other, leading to a distance $d = \sqrt{n}$ for a square grid of n qubits.

4.3. FAULT TOLERANCE

QEC aims at achieving a logical error rate that is lower than the physical error rate. To do so, it introduces more qubits, gates and measurements. Since in practice each of these building blocks is error-prone itself, the question is whether QEC removes more errors than it introduces. The aim of the theory of fault tolerance [13] is to develop circuits that are resilient to each component being faulty, whether that component is (logical) initialization, gates, measurements or the QEC circuit itself. If a component is implemented in a non-fault-tolerant way, there is no point in coding because the logical operation will be at least as faulty as at the physical level. Instead, if all operations are fault tolerant, there is a chance that the logical fidelity is better than at the physical level, provided that the physical error rates are low enough (i.e. below threshold, or at least pseudo-threshold; see Section 4.5). That means that there is enough redundancy and capability for correcting errors in each step, in such a way that errors do not spread excessively and a computation can be carried on for as long as needed. Furthermore, fault tolerance is generally an asymptotic statement, meaning that it is still possible that, if the QEC code is too small, a fault-tolerant operation might perform worse than the corresponding physical one.

This notion of fault tolerance is rather ambitious and constitutes a long-term goal for quantum computing. Because of this, one often finds restrictive notions of fault tolerance in papers and talks, in such a way that we can claim that *some* fault tolerance has been reached. Many times fault tolerance is defined as: a single fault should not lead to a logical error. In that case, it is important that a single fault does not equate to a single-qubit error. Indeed, for two-qubit gates a single fault should refer to any two-qubit error following the two-qubit gate. This is relevant because a typical failure mode for a two-qubit gate corresponds to errors on the entangling part of the operation. Hence, considering only single-qubit errors after a two-qubit gate would be too optimistic.

4.4. DECODING

Focusing on stabilizer codes (see Section 4.2), the syndrome history collected in the QEC cycles constitutes the information that one can use to correct for errors. Note that a syndrome bit might be faulty itself, e.g. if a readout error can occur when measuring the ancilla qubit. Below we outline the decoding process, where we discuss first the maximum-likelihood decoder and then other decoders.

MAXIMUM-LIKELIHOOD DECODER

The general idea of a decoder is to find the (best) correction such that the error is corrected without producing a logical error. To make the assessment of this task (numerically) meaningful, one assumes that the last syndrome measurement $\vec{s}_f$ is noiseless (even though this is never the case in experiment), such that the state after correction is in the logical subspace. Otherwise, if the state is not in the logical subspace, one cannot decide whether a logical error has occurred or not. The maximum-likelihood decoder is the decoder which for each syndrome history $\mathbf{s} = \{\vec{s}_1, \ldots, \vec{s}_f\}$ finds a correction $P_c^*(\mathbf{s})$ where

$$P_c^*(\mathbf{s}) = \underset{P_c}{\operatorname{argmax}} \sum_{P:\Pi_L P_c P \Pi_L = \Pi_L} \mathbb{P}(P|\mathbf{s}), \tag{4.3}$$

with P_c being compatible with $\vec{s}_f$. The condition $\Pi_L P_c P \Pi_L = \Pi_L$ means that $P_c P$ acts as the logical identity on the logical subspace, i.e. it is a product of stabilizers. The computation of $\mathbb{P}(P|\mathbf{s})$ requires knowledge of the error model. Since there are 2^{n-k} stabilizers in the stabilizer group, the fact that the sum runs over an exponential number of terms implies that it might be inefficient to compute the most-likely correction $P_c^*(\mathbf{s})$.

MWPM DECODER AND OTHERS

Because of the inefficiency of the maximum-likelihood decoder, many other decoders [14–16] have been proposed, where the algorithm is computationally efficient. Each of these decoders is generally tailored to a certain class of QEC codes. Here we focus on the surface code. The question then is how good such a decoder is, particularly compared to the maximum-likelihood one. Thus, a good QEC code does not only possess a large distance, as discussed in Section 4.1, but should also admit an efficient decoder with good performance. Here the performance is defined in terms of minimizing the logical infidelity $\mathscr{E}_L = 1 - \mathscr{F}_L$ (or maximizing the logical fidelity $\mathscr{F}_L$), i.e. the probability that the correction leads to a logical error, computed as a weighted average over all possible syndromes.

In theoretical works the error model is often chosen to be an independent Pauli error model, in which Pauli errors of possibly any type occur for each qubit or gate independently with a certain probability p. In this case one can define a logical error rate ε_L, where here we consider the definition in Ref. [17]. In particular, $\mathscr{F}_L$ evolves as a function of the QEC-cycle number n_c as [17]

$$\mathscr{F}_L[n_c] = \frac{1}{2}\left(1 + (1 - 2\varepsilon_L)^{n_c}\right). \tag{4.4}$$

Roughly speaking, in an independent model the weight of errors follows a binomial distribution, peaked around pn_q, where n_q is the number of qubits. If pn_q is low enough,

most errors have a relatively low weight and they constitute the largest contribution to the sum in Eq. (4.3). For the surface code, the Minimum-Weight Perfect-Matching decoder (MWPM) [14, 17, 18] is a decoder which precisely approximates that sum by just the probability of the lowest-weight error. The latter can be found using Edmond's algorithm, which has polynomial complexity in the number of qubits and is thus efficient. MWPM has been shown to perform well for the surface code, and we use it as well in Chapters 8 and 9 (see Refs. [17, 18] for more information about our specific implementation). Other decoders either use more complex strategies to better approximate the sum in Eq. (4.3) or use a simple algorithm that runs faster (e.g. the Union-Find decoder [15]) while maintaining good performance.

In the density-matrix simulations of the surface code (see Section 3.3.2) we also consider the Upper-Bound decoder (UB). In this case we directly draw information about errors from the density matrix. Since an actual decoder is only allowed to access the syndrome measurements, UB provides an upper bound to the performance of any decoder. For a detailed description see Section 9.7.1.

REAL-TIME DECODING

A major, somewhat unexplored issue is that decoders will need to run in real time, in parallel with the computation, such that errors can be corrected on the fly. While this can in principle be avoided for computations that only use Clifford operations, useful universal computations are not purely made of Cliffords [1]. In particular, Clifford circuits can be simulated classically in an efficient way (Gottesman-Knill theorem). Thus decoders do not need to be only "efficient" in a theoretical computer-science way, but they need to be efficient in real time as well. We discuss this further in Section 10.2.

4.5. THRESHOLD

We have mentioned in Section 4.1 that it is good if a family of QEC codes $\{\mathscr{H}_{\mathrm{L}}^{(n_q)}\}$ has a distance d that grows significantly with n_q. It is even better if there exists a so-called (error) threshold (see below). For simplicity, in this exposition we do not distinguish between errors in e.g. the gates or measurements, but we assume a single overall p that parametrizes all error probabilities. The threshold p_{th} is the physical error rate such that, for a given decoder, the logical infidelity $\mathscr{E}_{\mathrm{L}}^{(n_q)}[n_c]$ after n_c QEC cycles satisfies

$$\lim_{n_q \to +\infty} \mathscr{E}_{\mathrm{L}}^{(n_q)}[n_c] = 0 \tag{4.5}$$

for every $p < p_{\mathrm{th}}$, where one takes $n_c = d$ for the surface code, and in particular

$$\mathscr{E}_{\mathrm{L}}^{(n_q)}[n_c = d] \sim \mathrm{poly}(d)\, e^{-cd}, \tag{4.6}$$

where $c \sim \log(p_{\mathrm{th}}/p)$. The reason for taking $n_c = d$ is that, when measurements can be faulty, one needs to repeat them for at least a certain number of times to correct for those errors. Note that the threshold is a combined property of the code and the decoder. Regarding the logical error rate $\varepsilon_{\mathrm{L}}^{(n_q)}$, defined via Eq. (4.4), its dependence on d must be

$$\varepsilon_{\mathrm{L}}^{(n_q)} \sim \frac{\mathrm{poly}(d)}{d}\, e^{-cd} \tag{4.7}$$

to give Eq. (4.6). Indeed, based on Eq. (4.4), one has $\mathscr{E}_{\mathrm{L}}^{(n_q)}[n_c] = \frac{1}{2}\left(1-\left(1-2\varepsilon_{\mathrm{L}}^{(n_q)}\right)^d\right) \sim d\varepsilon_{\mathrm{L}}^{(n_q)} \sim \mathrm{poly}(d)\, e^{-cd}$.

Typically, three kinds of error models are considered: incoming noise (Pauli errors inserted on data qubits at the beginning of each QEC cycle), phenomenological noise (classical readout errors on top of incoming noise), and circuit-level noise (Pauli errors inserted after each gate on both data and ancilla qubits). Regarding the surface code, for incoming noise the threshold is about 10.9% [14] for the maximum-likelihood decoder and 10.3% for MWPM [19]. Instead for phenomenological noise it drops to about 2.9% [20] and for circuit-level noise to 0.9% for MWPM [19], although the threshold has been found to vary in the range 0.5-1.1% for different variations of circuit-level noise [21]. Because of this, one needs to be careful when comparing thresholds across different papers and numerics (as well as the possible different definitions of logical fidelity and logical error rate). Nevertheless, 1% is generally taken as the minimum error probability that each operation is required to reach in experiment. Even though $\mathscr{E}_{\mathrm{L}}^{(n_q)}[n_c]$ decreases exponentially with p for $p < p_{\mathrm{th}}$ (see Eq. (4.6)), to really reap the benefits of QEC one needs at least $p = 0.1\%$ or even 0.01% or less. Then n_q does not need to be unreasonably large to lower the logical infidelity below a desired level.

One of the holy grails of current intermediate-scale processors is thus to get all error rates below threshold. If this continues to hold when scaling up the system, then one can increase the system size until the logical infidelity drops below any desired value.

Pseudo-threshold. We briefly mention that one can also define the notion of pseudo-threshold, which is not a property of a family of codes but of a code with a fixed size within the family. In that case, the pseudo-threshold p_{pseu} is defined as the physical error probability such that the corresponding physical error rate ϵ satisfies $\epsilon(p = p_{\mathrm{pseu}}) = \varepsilon_{\mathrm{L}}$. Hence, for $p < p_{\mathrm{pseu}}$, $\epsilon(p) < \varepsilon_{\mathrm{L}}$ and coding is advantageous at the given code size.

4.6. BEYOND (INDEPENDENT) PAULI ERRORS

So far we have considered QEC codes that assume qubits and we have assumed that the errors are all Paulis and in particular that they are independent for each location. If the independence assumption is not fulfilled, i.e. if errors can be correlated, the performance of the decoder can be undermined. This can be due to a more difficult estimation of $\mathbb{P}(P|\mathbf{s})$ (see Section 4.4), as well as due to high-weight errors being relatively more likely than if errors were independent. As a consequence, an efficient decoder like MWPM might not perform so well anymore because the lowest-weight correction is less of a good guess. Correlated errors can be due to e.g. crosstalk between qubits, especially ZZ crosstalk (see Section 3.2.7), or due to leakage lasting for many QEC cycles (see Chapters 5, 8 and 9). Data-qubit leakage effectively reduces the code distance and spreads Z-rotation errors onto neighboring ancilla qubits, while ancilla-qubit leakage temporarily disables parity checks and spreads Z-rotation errors onto neighboring data qubits (see Section 8.4).

If the qubit assumption is not fulfilled and specifically if there can be leakage to higher levels, it is a priori unclear whether the threshold does not become unreasonably low and fault tolerance cannot be achieved in practice. In particular, while protecting against Pauli errors (up to a certain weight) guarantees protection against any error within the qubit subspace (see Section 4.2), it does not necessarily mean that the protection is extended

to leakage errors. In Chapter 8 we study the dynamics of leakage in a transmon-based surface code. We have indeed seen that leakage severely affects the logical infidelity (see Fig. 8.2), even though we still expect that a threshold should exist and even though the use of leakage-reduction units can restore a good amount of the lost performance (see Section 9.1 for a broader discussion). Previous work about the presence of leakage in QEC and how to remove it is reviewed in Chapter 5.

REFERENCES

[1] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information* (2000).

[2] D. Gottesman, *Stabilizer codes and quantum error correction*, Ph.D. thesis, California Institute of Technology (1997), arXiv:quant-ph/9705052 [quant-ph] .

[3] D. A. Lidar and T. A. Brun, *Quantum Error Correction* (2013).

[4] B. M. Terhal, *Quantum error correction for quantum memories,* Rev. Mod. Phys. **87**, 307 (2015).

[5] E. Knill, R. Laflamme, and L. Viola, *Theory of quantum error correction for general noise,* Physical Review Letters **84**, 2525–2528 (2000).

[6] K. A. Landsman, Y. Wu, P. H. Leung, D. Zhu, N. M. Linke, K. R. Brown, L. Duan, and C. Monroe, *Two-qubit entangling gates within arbitrarily long chains of trapped ions,* Phys. Rev. A **100**, 022332 (2019).

[7] S. Bravyi, D. Poulin, and B. Terhal, *Tradeoffs for reliable quantum information storage in 2D systems,* Physical Review Letters **104** (2010), 10.1103/physrevlett.104.050503.

[8] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, *Scalable quantum circuit and control for a superconducting surface code,* Phys. Rev. Appl. **8**, 034021 (2017).

[9] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation,* Phys. Rev. A **86**, 032324 (2012).

[10] A. Kitaev, *Fault-tolerant quantum computation by anyons*, Vol. 303 (Elsevier BV, 2003) p. 2–30.

[11] J. P. Bonilla Ataides, D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, *The XZZX surface code,* Nature Communications **12** (2021), 10.1038/s41467-021-22274-1.

[12] S. Puri, L. St-Jean, J. A. Gross, A. Grimm, N. E. Frattini, P. S. Iyer, A. Krishna, S. Touzard, L. Jiang, A. Blais, and et al., *Bias-preserving gates with stabilized cat qubits,* Science Advances **6**, eaay5901 (2020).

[13] D. Gottesman, *Quantum Error Correction and Fault-Tolerance,* (2005), arXiv:quant-ph/0507174 [quant-ph] .

[14] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, *Topological quantum memory,* Journal of Mathematical Physics **43** (2002), 10.1063/1.1499754.

[15] N. Delfosse and N. H. Nickerson, *Almost-linear time decoding algorithm for topological codes,* (2017), arXiv:1709.06218 [quant-ph] .

[16] X. Ni, *Neural network decoders for large-distance 2D toric codes,* Quantum **4**, 310 (2020).

[17] T. E. O'Brien, B. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise,* npj Quantum Information **3** (2017), 10.1038/s41534-017-0039-x.

[18] T. E. O'Brien, B. M. Varbanov, and S. T. Spitz, *qgarden,* (2019).

[19] A. G. Fowler, A. C. Whiteside, and L. C. L. Hollenberg, *Towards practical classical processing for the surface code,* Physical Review Letters **108** (2012), 10.1103/physrevlett.108.180501.

[20] C. Wang, J. Harrington, and J. Preskill, *Confinement-Higgs transition in a disordered gauge theory and the accuracy threshold for quantum memory,* Annals of Physics **303**, 31–58 (2003).

[21] A. M. Stephens, *Fault-tolerant thresholds for quantum error correction with the surface code,* Phys. Rev. A **89**, 022321 (2014).

5

LEAKAGE AND QUANTUM ERROR CORRECTION

5.1. PREVIOUS WORK

In this chapter we review the papers that have studied the impact of leakage on quantum error correction and how to deal with it. A summary of the relevant features of the considered models and results can be found in Table 5.1.

5.1.1. LEAKAGE-REDUCTION UNITS (LRUS)

Below we define, on the one hand, "proper" Leakage-Reduction Units (LRUs) and, on the other hand, ancilla-qubit reset schemes that serve the same purpose of leakage reduction, even though the rest of this section focuses mostly on LRUs.

Consider a quantum system with Hilbert space $\mathscr{H} = \mathscr{C} \oplus \mathscr{L}$, where $\mathscr{C}$ denotes the computational subspace and $\mathscr{L}$ the leakage subspace. Let $\Pi_{\mathscr{C}}$ and $\Pi_{\mathscr{L}}$ be the projectors on the respective subspaces. An (ideal) LRU is a quantum channel $\mathscr{E}_{\text{LRU}}$ on density matrices ρ in $\mathscr{H}$, with the following properties:

1. if $\rho = \Pi_{\mathscr{C}}\rho\Pi_{\mathscr{C}}$ (i.e. if $\Pi_{\mathscr{L}}\rho\Pi_{\mathscr{L}} = 0$), then $\mathscr{E}_{\text{LRU}}(\rho) = \rho$;
2. if $\Pi_{\mathscr{L}}\rho\Pi_{\mathscr{L}} \neq 0$, then $\Pi_{\mathscr{L}}\mathscr{E}_{\text{LRU}}(\rho)\Pi_{\mathscr{L}} = 0$.

A non-ideal LRU fulfills either one or both of these conditions in an approximate way. Condition 1 above can be relaxed either by requiring that $\rho' =: \mathscr{E}_{\text{LRU}}(\rho) \approx \rho$ while still $\rho' = \Pi_{\mathscr{C}}\rho'\Pi_{\mathscr{C}}$, or the LRU might mistakenly induce some leakage itself, i.e. $\text{Tr}(\Pi_{\mathscr{L}}\rho'\Pi_{\mathscr{L}}) \equiv \epsilon \ll 1$. Condition 2 above can be relaxed by requiring that $\text{Tr}\big(\Pi_{\mathscr{L}}\mathscr{E}_{\text{LRU}}(\rho)\Pi_{\mathscr{L}}\big) < \text{Tr}(\Pi_{\mathscr{L}}\rho\Pi_{\mathscr{L}})$, since this leads to an exponential reduction of the leakage population upon repeated applications of the LRU.

Paper	Leak. model	Gate	Code	LRU	Threshold	Platform
Aliferis05 [1]	local		concatenated	general, quantum teleport.	exists	
Suchara14 [2]	stochastic	SealedDepo	toric	Full, Partial, Quick, Heralded	exists	
Fowler13 [3]	stochastic	SealedDepo	repetition	Quick	exists	
Ghosh13 [4]	coherent	CZ			n.a.	supercond.
Ghosh14 [5]	coherent	CZ	surface	Quick		supercond.
Brown18 [6]	stochastic	SealedDepo	toric	Quick	exists	trapped ions
Brown19A [7]	stochastic	SealedDepo & MS	surface	No LRU (only anc. q. reset)	exists	trapped ions
Brown19B [8]	stochastic	SealedDepo & MS	subsystem	Quick	exists	trapped ions
Brown20 [9]	stochastic	SealedDepo	toric	Quick	exists	
Varbanov20 [10]	density matrix	CZ	surface	Only leakage detection	n.a.	supercond.
Battistel21 [11]	density matrix	CZ	surface	res-LRU and π-LRU	n.a.	supercond.

Table 5.1: Summary of the papers that study the detrimental effects of leakage on quantum error correction. Definitions can be found in Chapter 5. Regarding the leakage error models, in short: in a local model, the only assumption is that leakage is local; in a stochastic model, states in the computational subspace jump stochastically to the leakage subspace, but no superposition is allowed; in a coherent model, leakage is a coherent process and superpositions are allowed; the density-matrix model is in-between the coherent and stochastic one since leakage occurs coherently and spreads coherent errors, and then we depolarize the leakage subspace (furthermore, regular errors are also modeled coherently).

A reset scheme (see Section 2.5 for examples in superconducting qubits) is an operation where a qubit is mapped to (usually) the ground state $|0\rangle$, independently of its initial state in a given subspace. The subspace can be only $\mathscr{C}$, but it can also include $\mathscr{L}$ [12–15]. In the latter case the reset also removes leakage, although it is not a LRU according to the definition above, since it does not fulfill condition 1 at all. While data qubits are reset at the beginning of a quantum computation, generally only ancilla qubits can be reset mid-circuit without destroying encoded information. Specifically, ancilla qubits can be reset only right after being measured, since e.g. in QEC the parity-check information has already been extracted and the ancilla qubit is in a product state with respect to the rest of the system (assuming that measurements are projective to a good approximation).

5.1.2. THRESHOLD THEOREM FOR CONCATENATED CODES WITH LRUS

The foundational work on LRUs has been done by Aliferis and Terhal [1], who considered concatenated codes. In this work, the only assumption on leakage is that it is local, which is generally the case for current quantum-computing platforms. They have shown that fault-tolerant quantum computation is still possible in the presence of leakage, if appropriate LRUs are employed, even if they are non-ideal.

Since a leaked state generally does not contain any logical information (except if leakage is generated by a unitary that maps $\mathscr{C}$ to $\mathscr{L}$), a LRU converts a leakage error into a regular error in the computational subspace. By using LRUs, it is shown in Ref. [1] that a concatenated code with a certain threshold in the no-leakage case still possesses a threshold if leakage is present. However, this threshold is lower because the regular error rate is effectively higher due to the combination of leakage and LRUs.

We remark that to realize a LRU it is not necessary to know whether the state is supported on the leakage subspace or not. However, leakage detection, i.e. having the possibility to experimentally distinguish a leaked state, would be beneficial [2].

5.1.3. TOPOLOGICAL CODES AND LRUS

The threshold theorem proved by Aliferis and Terhal holds for concatenated codes, whereas topological codes were not considered. A similar threshold theorem for topological codes has not been proven so far. Suchara, Cross and Gambetta [2] numerically showed that a threshold does exist for topological codes in the presence of leakage, in the case of the toric code. Of course this holds with respect to the specific noise and leakage models that they considered (described below). In particular, certain (reasonable) assumptions are made on leakage to keep the simulation efficient. However, these models are still insightful about a threshold for a fully realistic noise model. In this section we first focus on the leakage model and the steps to efficient simulations in general, and then we provide specific details on the error models in Ref. [2].

Step 1. Efficiency-wise it is necessary to consider only states that are mixtures of computational and leakage states, i.e. no superpositions between the two should be allowed by the leakage noise model. In this way, in a Pauli error model for regular errors, one can track the Pauli frame (which is efficient) and a "leakage frame" as described in Ref. [2], which is also shown to be efficient. Hence, a stochastic leakage model is considered. Let $p_\uparrow$ be the probability of leakage and $p_\downarrow$ the probability of seepage. We consider qutrits here, i.e. $\mathscr{L}$ is one-dimensional and corresponds to $|2\rangle$. The stochastic

leakage and seepage channels are respectively defined as:

$$\mathscr{E}_{\uparrow}(\rho) = (1-p_{\uparrow})\rho + p_{\uparrow}\,|2\rangle\langle 2|\,, \tag{5.1}$$

$$\mathscr{E}_{\downarrow}(\rho) = (1-p_{\downarrow})\rho + p_{\downarrow}\Big(\Pi_{\mathscr{C}}\rho\Pi_{\mathscr{C}} + \langle 2|\rho|2\rangle\,\frac{\Pi_{\mathscr{C}}}{2}\Big), \tag{5.2}$$

where ρ is a single-qubit (reduced) density matrix. In Ref. [2] $\mathscr{E}_{\uparrow}$ is applied on each qubit after a two-qubit gate, whereas $\mathscr{E}_{\downarrow}$ is applied after both single- and two-qubit gates. Technically, $\mathscr{E}_{\downarrow}$ is a LRU, even though it is very poor since $p_{\downarrow}$ is usually quite low. In particular, $\mathscr{E}_{\downarrow}$ resembles T_1 relaxation, except for the fact that relaxation brings $|2\rangle$ down to $|1\rangle$, rather than to the maximally mixed state $\Pi_{\mathscr{C}}/2$ as in this model. It is also clear that the leakage mechanism given by $\mathscr{E}_{\uparrow}$ does not generate superpositions of computational and leaked states, as required for efficient simulations.

Step 2. Next, one needs to specify how gates act on leaked qubits. Specifically, as in Ref. [2], one requires that the (ideal) gates are "sealed". Roughly speaking, a sealed gate is defined as a quantum channel $\mathscr{U}_n$ acting on n-qubit density matrices, such that

- it does not generate superpositions between computational and leaked states of any individual qubit;
- non-leaked incoming qubits remain non-leaked after the ideal gate.

Note that only the first condition is required to make the simulation efficient. The second one is chosen because potentially there is no threshold or it is very low if two-qubit gates spread leakage, meaning that two qubits are leaked after the ideal gate when only one was leaked before.
More precisely, a sealed single-qubit gate $\mathscr{U}_1$ takes the form $\mathscr{U}_1(\rho) = U_1\rho U_1^{\dagger}$, i.e. it acts as a unitary U_1, such that

$$U_1 = U_{\mathscr{C}} \oplus U_{\mathscr{L}}. \tag{5.3}$$

A sealed two-qubit gate $\mathscr{U}_2$ is a probabilistic mixture of unitaries $U_2^{(i)}$, i.e. it has Kraus operators $\{\sqrt{p_i}U_2^{(i)}\}$ where p_i is the probability of applying $U_2^{(i)}$, such that

$$U_2^{(i)} = U_{\mathscr{C}\otimes\mathscr{C}} \oplus U^{(i)}_{\mathscr{C}\otimes\mathscr{L}} \oplus U^{(i)}_{\mathscr{L}\otimes\mathscr{C}} \oplus U_{\mathscr{L}\otimes\mathscr{L}}. \tag{5.4}$$

That is, $\mathscr{U}_2$ fulfills the conditions for a sealed gate by not mixing the four subspaces in this equation.

Step 3. To complete the construction of this leakage model, one only has to give a specific expression for $\mathscr{U}_1$ and $\mathscr{U}_2$, especially with respect to the behavior when a qubit is leaked. In Ref. [2], for single-qubit gates they pick $U_{\mathscr{C}}$ (see Eq. (5.3)) to be the ideal operation, whereas $U_{\mathscr{L}} = I_{\mathscr{L}}$, i.e. it acts as the identity on $\mathscr{L}$.
For two-qubit gates, they pick $U_{\mathscr{C}\otimes\mathscr{C}}$ (see Eq. (5.4)) to be the ideal operation, whereas $U_{\mathscr{L}\otimes\mathscr{L}}$ is just a global phase for qutrits, so it is irrelevant. Furthermore, they choose $U^{(i)}_{\mathscr{C}\otimes\mathscr{L}} = \sqrt{\frac{1}{4}}P^{(i)}_{\mathscr{C}} \otimes I_{\mathscr{L}}$ (and similarly for $U^{(i)}_{\mathscr{L}\otimes\mathscr{C}}$), where $P^{(i)}_{\mathscr{C}} \in \{I, X, Y, Z\}$. In other words, a non-leaked qubit interacting with a leaked one is completely depolarized, since applying uniformly random Pauli errors is equivalent to complete depolarization. Here we refer to this kind of sealed gates as SealedDepo gates. Note that if SealedDepo gates involving

one leaked qubit are applied to multiple non-leaked qubits, the spread Pauli errors are uncorrelated. In particular, then, SealedDepo gates produce the worst-case *uncorrelated* noise on non-leaked qubits. However, sealed gates producing correlated noise (if e.g. $U^{(i)}_{\mathscr{C}\otimes\mathscr{L}} \equiv U_{\mathscr{C}\otimes\mathscr{L}}$ is a fixed unitary for all gates) can have a more detrimental effect than SealedDepo gates on the performance of a QEC code (but not necessarily [8]).

OVERALL NOISE MODEL

Here we discuss the noise locations for both regular and leakage errors in Ref. [2]. If a two-qubit gate acts on two non-leaked qubits, it acts as the ideal operation and it is followed by a Pauli channel with overall probability p_P of applying any two-qubit Pauli. If a two-qubit gate acts on a leaked and on a non-leaked qubit, a certain $U^{(i)}_{\mathscr{C}\otimes\mathscr{L}}$ is sampled according to the $\{p_i\}$ and applied to the non-leaked qubit. In the relatively unlikely case in which the two-qubit gate acts on two leaked qubits, the gate does nothing. In any case, $\mathscr{E}_\uparrow$ and $\mathscr{E}_\downarrow$ are applied on both qubits after applying the Pauli errors (if any).

Single-qubit gates are followed by a Pauli channel if they act on a non-leaked qubit, with the same probability p_P of applying any single-qubit Pauli. Then only $\mathscr{E}_\downarrow$ is applied. The same applies to idling steps.

Measurements are either supposed to declare a $|2\rangle$ as a $|1\rangle$, or to be able to distinguish these two states (in which case an increase in performance of the QEC code is observed). Furthermore, the measurement can report an incorrect outcome with probability $p_m \equiv p_P$.

LRU SCHEMES

Suchara, Cross and Gambetta [2] also introduced various schemes for applying LRUs in quantum error correction, and they found non-zero thresholds for all of them in the case of the toric code. The threshold is then a function of both the leakage probability $p_\uparrow$ and the regular error probability p_P (and weakly depends on $p_\downarrow$).
We note that they found a zero threshold if LRUs are not used. Furthermore, with respect to their model in Eq. (5.2), we expect that a threshold would exist if $p_\downarrow$ is large enough, but in Ref. [2] $p_\downarrow$ is relatively small (as it is expected for relaxation).

- Full-LRU scheme: this is the scheme where LRUs are applied to both qubits (so either data or ancilla qubits) after each two-qubit gate, as in Ref. [1].
  In Ref. [2], the specific LRU considered consists of preparing a "fresh" qubit q_f in $|0\rangle$, then performing a SWAP with the target qubit q_t to which the LRU is conceptually applied (note that this scheme requires many extra qubits on the chip, as well as a non-trivial connectivity potentially); after this, q_t is reset to $|0\rangle$, independently of its state being $|0\rangle$, $|1\rangle$ or $|2\rangle$, while q_f continues to be used in the quantum computation. We note that the SWAP is compiled in terms of three CNOTs, where the first one can be dropped since the control is on q_f (in $|0\rangle$) and thus this CNOT acts as the identity.

Assume that the action of the SWAP is ideally

$$|00\rangle \mapsto |00\rangle \tag{5.5}$$
$$|01\rangle \mapsto |10\rangle \tag{5.6}$$
$$|10\rangle \mapsto |01\rangle \tag{5.7}$$
$$|11\rangle \mapsto |11\rangle \tag{5.8}$$
$$|2\,j\rangle \mapsto |2\,j\rangle \tag{5.9}$$
$$|j\,2\rangle \mapsto |j\,2\rangle\,, \tag{5.10}$$

i.e. in particular that it does not swap leakage. If q_t was not leaked, after the SWAP it is in state $|0\rangle$ (thus the reset effectively does nothing), while its state has been correctly swapped onto q_f. If instead q_t was leaked, after the SWAP the "fresh", outgoing qubit $q_o \equiv q_f$ is still non-leaked, while the leakage on q_t is removed by the reset. If the SWAP is prone to regular or possibly leakage errors, q_o has a chance to be leaked. However, if the error rates are low, the LRU still reduces leakage in the system with high probability.
We note that this LRU does not rely on a stochastic leakage model to be effective. That is, even if q_t is in a superposition of computational and leaked states, q_o is fully non-leaked (modulo imperfections in the operations involved). However, the latter does not mean that q_o is error free. In the worst case that q_t was fully in $|2\rangle$, q_o is in $|0\rangle$, thus when it is re-entangled with the rest of the data qubits by the subsequent parity-check measurements, it is affected by a uniformly random Pauli error. If q_t was only partially leaked, q_o is still affected by some Pauli error but with a distribution that depends on the leakage.

- Partial-LRU scheme: LRUs are only applied to data qubits at the end of each QEC cycle, whereas ancilla qubits are reset after being measured (thus still removing leakage from the ancilla qubits).
  The less frequent use of these LRUs (described above) reduces the number of extra qubits and operations required, compared to the Full-LRU scheme, but it still requires one extra qubit per data qubit (although one could actually use only one extra qubit at the cost of serializing the application of the LRUs, as long as the connectivity allows for it).

- Quick-LRU scheme: data and ancilla qubits are swapped at the end of each QEC cycle, then the qubits taking the role of ancilla qubits are reset. Note that in this way one does not need any extra qubits. The SWAP is also compiled as three CNOTs, where here the first one cancels with the last one of the parity-check unit [2], so that overall only one CNOT is effectively added to the ones in the parity-check unit.
  As each qubit takes the role of ancilla qubit every two QEC cycles, leakage can last for at most two QEC cycles, again assuming that the SWAP does not swap the leakage as well. Specifically, if leakage is on the qubit taking the role of ancilla qubit, then leakage is removed immediately by the reset. If instead leakage is on the qubit taking the role of data qubit, then one needs to wait until the following QEC cycle (in which, then, the qubit takes the role of ancilla qubit) for the reset to remove this

leakage from the system.
We refer to the Quick-LRU also as the swap-LRU.

Reference [3] followed similar lines as in Ref. [2], considering a Quick-LRU scheme with respect to a stochastic leakage model and SealedDepo gates. In this case, the code being studied was the repetition code, whose performance was found to be similarly damaged by leakage, but a threshold still existed when using the Quick-LRU.

5.1.4. STUDIES OF COHERENT LEAKAGE IN SUPERCONDUCTING QUBITS

Ghosh and Fowler [4, 5] considered a coherent leakage model rather than a stochastic one. The CZ in superconducting qubits is modeled in a realistic way, taking into account the avoided crossings in the spectrum of two coupled transmons (see Section 2.8.2), as well as the functioning of a baseband flux-based CZ (see Section 2.8.3). In particular, the coherent exchange between $|11\rangle$ and $|02\rangle$ is considered. Higher excited states, namely $|3\rangle$, are not included in the description.

In Ref. [4] the simple case of a single data qubit measured by an ancilla qubit is considered (no other qubits). They find that the leakage conditional phases (defined in Section 3.3.2) affect the capability to detect leakage. In particular, when the leakage conditional phase is ≈ 0, they observe that the ancilla qubit does not detect leakage at all since it always reports the same measurement outcome, as if there was no error. This effect is nicknamed "leakage paralysis". However, the authors fail to notice that this effect disappears if there is more than one data qubit in the parity-check unit (due to the anticommutation effect; see Section 8.11.2). In Ref. [5] weight-2 stabilizers of a surface code are considered, but no larger-scale study is carried out.

5.1.5. STUDIES OF STOCHASTIC LEAKAGE IN TRAPPED IONS

A series of papers by N. and K. Brown et al. [6–8] has studied the effect of leakage on QEC based on trapped-ion qubits. They consider Quick-LRUs with respect to a stochastic leakage model with sealed gates (so their numerical simulations are efficient). The only aspect of the leakage error model that is tailored to trapped ions is the fact that the gates are not SealedDepo, but a realistic set of $\{U^{(i)}_{\mathscr{C}\otimes\mathscr{L}}, U^{(i)}_{\mathscr{L}\otimes\mathscr{C}}\}$ and $\{p_i\}$ is used for the Mølmer-Sørensen gate (MS) for a non-leaked qubit interacting with a leaked qubit. Those are extracted via a Pauli-twirl approximation of the coherent action of the Mølmer-Sørensen gate.

References [6, 7] considered two different species of trapped ions and studied a surface (or toric) code made up of either one or both of these species. One species cannot leak but has low coherence times, whereas the other one can leak but has long coherence times. If the surface code is made up of only one species, they find that leakage is so detrimental that it is preferable to use the non-leakage-prone ions, despite their higher susceptibility to regular errors. However, the best solution is a mixed-species surface code, where the non-leakage-prone ions are used as data qubits and the leakage-prone ions as ancilla qubits. In this way, regular ancilla-qubit errors are less frequent and ancilla-qubit leakage is removed by the reset. Hence, there is no need to swap data and ancilla qubits or develop any other LRU for the data qubits.
The authors comment that the superiority of the mixed-species scheme holds only for

Sealed MS gates but not for SealedDepo gates (even though this is less relevant for trapped ions). In the latter case, they find (surprisingly) that ancilla-qubit leakage is particularly damaging, even more than data-qubit leakage, presumably because of the errors spread by a leaked ancilla qubit. Thus these errors counterbalance the benefits, described above, of the mixed-species scheme.

Reference [8] studied subsystem codes, specifically Bacon-Shor codes and the subsystem surface code, finding that they are somewhat more resistant to leakage compared to subspace codes, specifically the surface code. This result is attributed to the fact that the parity checks of those subsystem codes have lower weight than in the subspace surface code, thus limiting the spread of correlated errors. Some improvements in the threshold, especially for Bacon-Shor codes, are found for Sealed MS gates, compared to the case of complete depolarization (SealedDepo gates).

5.1.6. DATA- VERSUS ANCILLA-QUBIT LEAKAGE AND CRITICAL LEAKAGE LOCATIONS

N. Brown, A. Cross and K. Brown [9] further studied leakage in the toric code with respect to a stochastic leakage model with SealedDepo gates and Quick-LRUs. First, as in Ref. [7], they found that ancilla-qubit leakage has a worse impact on the performance than data-qubit leakage. The second main finding was that leakage in the first CNOT of the parity-check unit is significantly more damaging than in the following CNOTs.

To my understanding, the fact that ancilla-qubit leakage is more damaging is related to the fact that in a SealedDepo gate the leaked qubit can spread errors of any type (X, Y, Z). Specifically, the point is that e.g. a Z-type ancilla qubit can spread X errors to nearby data qubits (i.e. the error of type *opposite* to the parity check). Based on the circuits in Fig. 4.1(b,c), one can get convinced that this is not possible when only regular errors are present. An issue with opposite-type errors is, for example, that if a Z-type ancilla qubit spreads 3 or 4 Z errors, by multiplication with a stabilizer these are respectively equivalent to 1 or 0 Z errors. However, this weight reduction is not possible when a Z-type ancilla qubit spreads 3 or 4 X errors. The first CNOT is then particularly problematic because there is no scheduling that can avoid spreading many errors.

In superconducting qubits we have found that a leaked qubit spreads Z rotations that depend on the leakage conditional phases (see Section 3.3.2). Parity checks of Z type thus spread Z-like errors and checks of X type spread X-like errors thanks to the Hadamard gates. That is, only same-type errors are spread and we do not observe data- or ancilla-qubit leakage to be significantly worse than the other in our error model. Indeed, in Fig. 9.5 one can see that using LRUs for only the data or ancilla qubits lowers the logical error rate by a comparable amount.

5.2. COMPARISON WITH WORK IN THIS THESIS

In Chapters 8 and 9 we consider the distance-3 instance of the (rotated) surface code (Surface-17), based on superconducting transmons. In particular, for the CZ we consider the coherent error model we developed (see Sections 3.3.2 and 8.2) based on the full Lindblad simulations of the gate (see Section 3.3.1). This model for the CZ is inserted in density-matrix simulations of Surface-17, where we include also T_1 and frequency-

dependent T_2 (see Section 3.3.2).

The main advantage of density-matrix simulations is that we can accurately study the evolution of leakage and how it interacts with stabilizer quantum error correction. We observe the anticommutation of parity checks in the presence of a leaked data qubit (see Section 8.11.2), as well as the effect of coherent Z rotations spread by a leaked ancilla qubit (see Section 8.4). Furthermore, we can test leakage detection via Hidden Markov Models in a realistic setting (see Chapter 8), as well as the benefits of LRUs (see Chapter 9). In particular, we introduce the res-LRU and π-LRU for data and ancilla qubits, respectively, where the res-LRU (see Section 9.2) consists of a microwave pulse applied on a data transmon and the π-LRU (see Section 9.3.2) consists of a $|1\rangle \leftrightarrow |2\rangle$ π-pulse on the ancilla qubit conditioned on the readout declaration of a $|2\rangle$. Compared to the Quick-LRU (or swap-LRU), these LRUs do not require additional qubits or hardware elements, nor extra QEC-cycle time (see Section 9.1 for a broader discussion).

We note that in the density-matrix simulations we decohere the leakage states after the CZs, motivated by the fact that stabilizer measurements tend to decohere leakage relatively fast (see Section 8.10.2). While this means that the leakage error model in the density-matrix simulations is not completely coherent, the spread Z-rotation errors are coherent in the computational subspace of the non-leaked qubit (regular relaxation errors are coherent as well), allowing us to observe the effects above and speeding up the simulations.

The disadvantage of the density-matrix simulations is that, despite our optimization efforts, they are intrinsically inefficient as the size of the density matrix grows exponentially with the number of transmons. It follows that with these methods we cannot study the threshold of the surface code in the presence of our leakage model and LRUs. We leave to future work the development of stochastic simulations that incorporate the realistic elements we studied as much as possible (see Section 9.4 for more details).

REFERENCES

[1] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults,* Quantum Info. Comput. **7**, 139 (2007).

[2] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code,* Quantum Info. Comput. **15**, 997 (2015).

[3] A. G. Fowler, *Coping with qubit leakage in topological codes,* Phys. Rev. A **88**, 042308 (2013).

[4] J. Ghosh, A. G. Fowler, J. M. Martinis, and M. R. Geller, *Understanding the effects of leakage in superconducting quantum-error-detection circuits,* Phys. Rev. A **88**, 062329 (2013).

[5] J. Ghosh and A. G. Fowler, *Leakage-resilient approach to fault-tolerant quantum computing with superconducting elements,* Phys. Rev. A **91**, 020302 (2015).

[6] N. C. Brown and K. R. Brown, *Comparing zeeman qubits to hyperfine qubits in the context of the surface code:* $^{174}Yb^{+}$ *and* $^{171}Yb^{+}$, Phys. Rev. A **97**, 052301 (2018).

[7] N. C. Brown and K. R. Brown, *Leakage mitigation for quantum error correction using a mixed qubit scheme,* Phys. Rev. A **100**, 032325 (2019).

[8] N. C. Brown, M. Newman, and K. R. Brown, *Handling leakage with subsystem codes,* New Journal of Physics **21**, 073055 (2019).

[9] N. C. Brown, A. Cross, and K. R. Brown, *Critical faults of leakage errors on the surface code,* in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)* (2020) pp. 286–294.

[10] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. DiCarlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code,* npj Quantum Information **6** (2020), 10.1038/s41534-020-00330-w.

[11] F. Battistel, B. Varbanov, and B. Terhal, *Hardware-efficient leakage-reduction scheme for quantum error correction with superconducting transmon qubits,* PRX Quantum **2**, 030314 (2021).

[12] S. Zeytinoğlu, M. Pechal, S. Berger, A. A. Abdumalikov, A. Wallraff, and S. Filipp, *Microwave-induced amplitude- and phase-tunable qubit-resonator coupling in circuit quantum electrodynamics,* Physical Review A **91** (2015).

[13] D. Egger, M. Werninghaus, M. Ganzhorn, G. Salis, A. Fuhrer, P. Müller, and S. Filipp, *Pulsed reset protocol for fixed-frequency superconducting qubits,* Phys. Rev. Applied **10**, 044030 (2018).

[14] P. Magnard, P. Kurpiers, B. Royer, T. Walter, J.-C. Besse, S. Gasparinetti, M. Pechal, J. Heinsoo, S. Storz, A. Blais, and A. Wallraff, *Fast and unconditional all-microwave reset of a superconducting qubit,* Phys. Rev. Lett. **121**, 060502 (2018).

[15] M. McEwen, D. Kafri, Z. Chen, J. Atalaya, K. J. Satzinger, C. Quintana, P. V. Klimov, D. Sank, C. Gidney, A. G. Fowler, F. Arute, K. Arya, B. Buckley, B. Burkett, N. Bushnell, B. Chiaro, R. Collins, S. Demura, A. Dunsworth, C. Erickson, B. Foxen, M. Giustina, T. Huang, S. Hong, E. Jeffrey, S. Kim, K. Kechedzhi, F. Kostritsa, P. Laptev, A. Megrant, X. Mi, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Niu, A. Paler, N. Redd, P. Roushan, T. C. White, J. Yao, P. Yeh, A. Zalcman, Y. Chen, V. N. Smelyanskiy, J. M. Martinis, H. Neven, J. Kelly, A. N. Korotkov, A. G. Petukhov, and R. Barends, *Removing leakage-induced correlated errors in superconducting quantum error correction,* (2021), arXiv:2102.06131 [quant-ph] .

6

NET ZERO CONDITIONAL-PHASE GATES

Conditional-phase (CZ) gates in transmons can be realized by flux pulsing computational states towards resonance with non-computational ones. In the first part of this chapter we present a 40 ns *CZ gate based on a bipolar flux pulse suppressing leakage (0.1%) by interference and approaching the speed limit set by exchange coupling. This pulse harnesses a built-in echo to enhance fidelity (99.1%) and is robust to long-timescale distortion in the flux-control line, ensuring repeatability. Numerical simulations matching experiment show that fidelity is limited by high-frequency dephasing and leakage by short-timescale distortion.*

Simple tuneup of fast two-qubit gates is essential for the scaling of quantum processors. In the second part of this chapter, we introduce the sudden variant (SNZ) of the Net Zero scheme realizing controlled-Z (CZ) gates by flux control of transmon frequency. SNZ CZ gates realized in a multi-transmon processor operate at the speed limit of transverse coupling between computational and non-computational states by maximizing intermediate leakage. Beyond speed, the key advantage of SNZ is tuneup simplicity, owing to the regular structure of conditional phase and leakage as a function of two control parameters. SNZ is compatible with scalable schemes for quantum error correction and adaptable to generalized conditional-phase gates useful in intermediate-scale applications.

The first part of this chapter has been published in Phys. Rev. Lett. **123**, 120502 (2019) [1]. The second part of this chapter has been published in Phys. Rev. Lett. **126**, 220502 (2021) [2]. F. B. realized the simulations and contributed to the presented concepts and the development of the error model. Furthermore, F. B. contributed extensively to the writing of the first part and provided input and feedback on the writing of the second part.

6.1. PART 1: FAST, HIGH-FIDELITY CONDITIONAL-PHASE GATE EXPLOITING LEAKAGE INTERFERENCE IN WEAKLY ANHARMONIC SUPERCONDUCTING QUBITS

6.2. INTRODUCTION

A steady increase in qubit counts [3–6] and operation fidelities [7–11] allows quantum computing platforms using monolithic superconducting quantum hardware to target outstanding challenges such as quantum advantage [12–14], quantum error correction (QEC) [15–19], and quantum fault tolerance (QFT) [20, 21]. All of these pursuits require two-qubit gates with fidelities exceeding 99%, fueling active research.

There are three main types of two-qubit gates in use for transmon qubits (see also Section 2.8.3), all of which harness exchange interactions between computational states ($|ij\rangle, i, j \in \{0,1\}$) or between computational and non-computational states (i or $j \geq 2$), mediated by a coupling bus or capacitor. Cross-resonance gates [10, 22] exploit the exchange interaction between $|01\rangle$ and $|10\rangle$ using microwave-frequency transversal drives. Parametric gates [9, 23] employ radio-frequency longitudinal drives, specifically flux pulses modulating the qubit frequency, to generate sidebands of resonance between $|01\rangle$ and $|10\rangle$ for iSWAP or between $|11\rangle$ and $|02\rangle$ or $|20\rangle$ for conditional phase (CZ). The oldest approach [24, 25] uses baseband flux pulses to tune $|11\rangle$ into near resonance with $|02\rangle$ to realize CZ. Either because they explicitly use non-computational states, or because of frequency crowding and the weak transmon anharmonicity, the three approaches are vulnerable to leakage of information from the computational subspace. Leakage is very problematic in applications such as QEC, complicating the design of error decoders and/or demanding operational overhead to generate seepage [26–30], generally reducing the error thresholds for QFT. This threat has motivated the design of fast-adiabatic pulses [31] to mitigate leakage and architectural choices in qubit frequency and coupler arrangements [32] to explicitly avoid it. Surprisingly, many recent demonstrations [9, 10, 33] of two-qubit gates place emphasis on reaching or approaching 99% fidelity without separately quantifying leakage.

Although baseband flux pulsing produces the fastest two-qubit gates to date (30 – 45 ns), two challenges have kept it from becoming the de facto two-qubit gating method. First, because the pulse displaces one qubit 0.5 – 1 GHz below its flux-symmetry point, i.e., the sweetspot, the sensitivity to flux noise increases dephasing and impacts fidelity. The second challenge is non-atomicity. If uncompensated, distortions in the flux-control lines originating from limited waveform-generator bandwidth, high-pass bias tees, low-pass filters, impedance mismatches, on-chip response, etc., can make the action of a pulse depend on the history of flux pulses applied. To date, predistortion corrections have been calculated in advance, requiring prior knowledge of the timing of all the flux-pulse-based operations required by the quantum circuit, and significant waveform memory. This standard practice is incompatible with real-time determination and execution of operations, as is required for control flow and feedback in a fully programmable quantum computer [34, 35].

In this Letter, we introduce a fast (40 ns), low-leakage (0.1%), high-fidelity (99.1%), and repeatable flux-pulse-based CZ gate suitable for a full-stack quantum computer

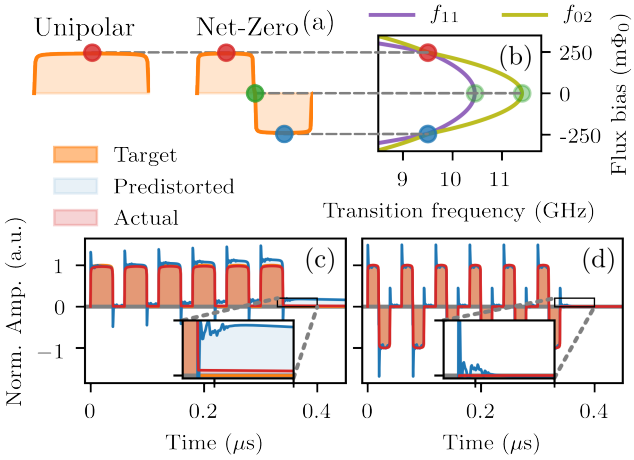

Figure 6.1: (a) Schematic representation of unipolar and NZ pulses that tune into resonance with (b) $|11\rangle \leftrightarrow |02\rangle$ in order to perform CZ gates. Repeated applications of unipolar (c) and NZ (d) CZ pulses showing the target (orange), predistorted (blue), and actual (red) waveforms for an imperfect distortion correction. The insets in (c) and (d) show the differing accumulation in the required predistortion correction.

executing operations in real time on transmon-based quantum hardware. These attractive characteristics are enabled by a zero-average bipolar flux-pulsing method, nicknamed Net-Zero (NZ), which uses the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing twice. Harnessing the analogy to a Mach-Zehnder interferometer, NZ exploits destructive interference to minimize leakage to $|02\rangle$ while approaching the speed limit set by the exchange coupling in the two-excitation manifold. The flux symmetry of the transmon Hamiltonian makes the phases acquired by the pulsed qubit first-order insensitive to low-frequency flux noise, increasing fidelity relative to a unipolar pulse. Crucially, the zero-average characteristic makes NZ insensitive to long-timescale distortions remaining in the flux-control line after real-time pre-compensation, making the CZ gate repeatable. Detailed numerical simulations supplied with calibrated experimental parameters and direct measurement of short-timescale distortions show an excellent match to experiment, and indicate that fidelity is limited by high-frequency flux noise while leakage is dominated by remaining short-timescale distortions.

6.3. NET-ZERO CONCEPT

The ideal CZ gate (see also Section 2.8.3) is described by the transformation:

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{i\phi_{01}} & 0 & 0 \\ 0 & 0 & e^{i\phi_{10}} & 0 \\ 0 & 0 & 0 & e^{i\phi_{11}} \end{pmatrix}, \tag{6.1}$$

in the computational basis $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$, where the single-qubit phases ϕ_{01} and ϕ_{10} are even multiples of π and the conditional phase defined by $\phi_{2Q} = \phi_{11} - \phi_{01} - \phi_{10}$ is

an odd multiple of π. A CZ gate of total duration $T_{CZ} = T_{2Q} + T_{1Q}$ can be realized in two steps. First, a strong flux pulse on the higher frequency qubit moves $|11\rangle$ into the avoided crossing with $|02\rangle$ and back to acquire ϕ_{2Q}. Next, simultaneous weaker pulses on both qubits adjust the single-qubit phases. We compare two types of flux pulses, the (unipolar) pulse introduced in [31] and the NZ pulse [Fig. 6.1(a)]. The NZ pulse consists of two back-to-back unipolar pulses of half the duration and opposite amplitude. Experiments are performed on a pair of flux-tunable transmons described in Section 6.11.1.

Because of distortions (see also Section 3.2.5), the waveform $V_{AWG}(t)$ specified in an arbitrary waveform generator (AWG) does not result in the qubit experiencing the targeted flux $\Phi_{target}(t)$. These distortions can be described as a linear time-invariant system that transduces voltage to flux and is characterized by its impulse response $h(t)$. To measure $h(t)$ at the qubit, we employ the Cryoscope technique that we introduce in [36]. We then use it to construct an inverse filter $\tilde{h}^{-1}$, known as a predistortion correction, to compensate the distortions. By performing a convolution of the desired signal $\Phi_{target}(t)$ with $\tilde{h}^{-1}$, the qubit experiences the pulse

$$\Phi(t) = h * V_{AWG}(t) = h * (\tilde{h}^{-1} * \Phi_{target})(t). \tag{6.2}$$

The predistortion corrections are performed using a combination of real-time filters implemented in a Zurich Instruments HDAWG and a short (20 ns) FIR filter implemented offline.

By eliminating the DC component of the pulse, NZ CZ gates are resilient to long-timescale distortions [37]. Because the transmon Hamiltonian is symmetric with respect to the sweetspot, it is possible to use both positive and negative amplitudes to perform a CZ gate [Fig. 6.1(b)] while satisfying the zero-average condition

$$\int_0^{T_{CZ}} \Phi_{target}(t')dt' = 0. \tag{6.3}$$

If Eq. (6.3) holds, the DC component is zero and the components in the Fourier transform $\Phi_{target}(\omega)$ at frequencies $\omega \lesssim \frac{2\pi}{T_{CZ}}$ are suppressed. Writing Eq. (6.2) in the Fourier domain: $\Phi(\omega) = \mathcal{H}(\omega) \cdot \tilde{\mathcal{H}}^{-1}(\omega) \cdot \Phi_{target}(\omega)$, it follows that if $\Phi_{target}(\omega)$ does not contain any components at $\omega < \frac{2\pi}{T_{CZ}}$, then $\Phi(\omega)$ does not depend on any components of $\mathcal{H}(\omega)$ at frequencies $\omega < \frac{2\pi}{T_{CZ}}$. As a consequence, the required corrections for NZ pulses do not accumulate, eliminating the need for accurate long-timescale distortion corrections and the resulting history-dependent errors [Fig. 6.1(d)].

6.4. REPEATABILITY

To measure the repeatability of CZ gates, the phase (ϕ_{01}) acquired by the pulsed qubit during a CZ gate is measured as a function of the separation time T_{Sep} between pulses (Fig. 6.2). Because of the detuning from the sweetspot, a small change in amplitude during the pulse leads to a significant change in frequency. This makes the acquired phase sensitive to distortions. We observe that not correcting distortions leads to significant phase errors (~ 80 deg). Correcting distortions using a predistortion filter keeps the error small (< 10 deg) for the first 500 ns but shows history-dependent behavior for longer timescales.

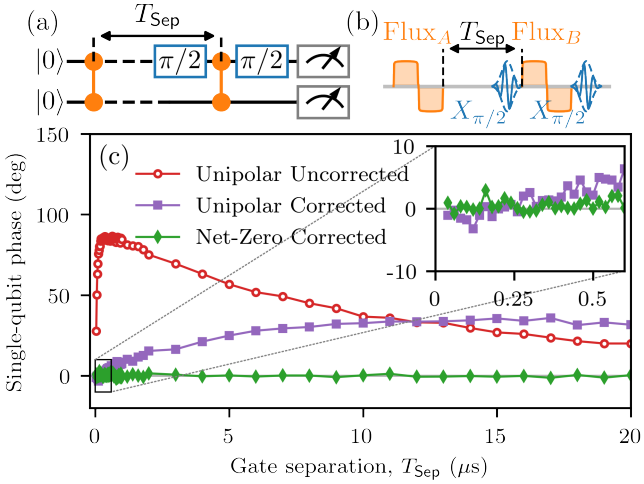

Figure 6.2: History dependence of flux pulses. Circuit (a) and pulses (b) used to measure the phase acquired during a pulse as a function of separation time T_{Sep} to another pulse. Pulses are calibrated to correspond to CZ gates. (c) Acquired single-qubit phase for unipolar pulses without (red), and with (purple) predistortion corrections and NZ pulses with predistortion corrections (green).

Using NZ pulses in combination with a predistortion filter eliminates all history dependence. Hence, we conclude that NZ pulses are robust against remaining long-timescale distortions.

6.5. ECHO EFFECT

We next investigate a built-in echo effect that provides protection against flux noise. Because the derivative of the flux arc is equal and opposite in sign at the positive and negative halves of the NZ pulse, we expect ϕ_{01} and ϕ_{2Q} to be first-order insensitive to low-frequency flux noise. As a test, we measure the dependence of ϕ_{2Q} on an applied DC flux offset for both a unipolar and NZ CZ gate [Fig. 6.3]. As shown in Fig. 6.3(b), ϕ_{2Q} is first-order (second-order) sensitive for a unipolar (NZ) pulse. We have also measured how the dephasing time depends on the detuning for both a square flux pulse and two half-square flux pulses with opposite sign (see Section 6.11.3). We find that the dephasing rate is significantly reduced when the opposite-sign flux pulses are used, confirming that NZ pulses have a built-in echo effect.

6.6. EXPERIMENT-SIMULATION MATCH

The pulse shape is intended to minimize leakage and is described by two parameters (see Section 6.11.2). Parameter θ_f is a measure of the flux at the middle of the unipolar pulse, and at the middle of each half of NZ. States $|11\rangle$ and $|02\rangle$ are resonant at $\theta_f = \pi/2$. Parameter λ_2 tunes the sharpness of the pulse rise and fall. We follow [38] in defining the leakage (L_1) of an operation as the average probability that a random computational state leaks out of the computational subspace.

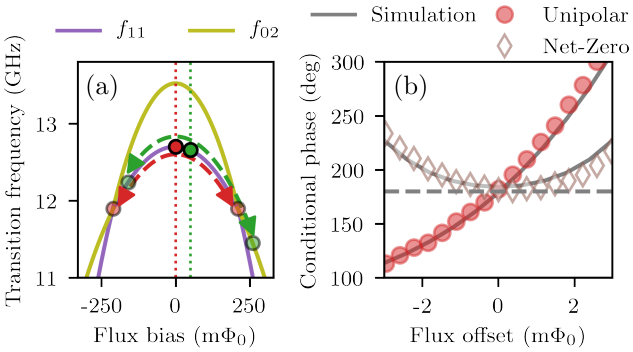

Figure 6.3: Echo effect in NZ pulses. (a) Level diagram showing the effect of a drift in flux on a NZ pulse: a NZ pulse will move to the interaction point on both sides (red); when the bias is offset (green), one side will overshoot while the other side will undershoot the interaction point, canceling the acquired extra phase. (b) Measured dependence of conditional phase on applied DC flux offset for both NZ (diamond) and unipolar (circles) $T_{CZ} = 60$ ns pulses ($T_{2Q} = 40$ ns). Solid lines correspond to simulation (see Section 6.11.3), dashed line indicates 180 deg. The unipolar (NZ) is first-order (second-order) sensitive to the applied offset.

In order to gain insight into how ϕ_{2Q} and L_1 depend on the pulse shape, we perform an experiment and compare this to simulations. The conditional oscillation experiment (Fig. 6.4) consists of a Ramsey-like experiment that allows us to measure ϕ_{2Q} and estimate L_1. This experiment measures the phase acquired during an (uncalibrated) CZ gate by the target qubit ($q_{\text{targ.}}$) while either leaving the control qubit ($q_{\text{contr.}}$) in the ground state, or adding an excitation to $q_{\text{contr.}}$. The difference between the phase acquired when $q_{\text{contr.}}$ is in $|0\rangle$ and when $q_{\text{contr.}}$ is in $|1\rangle$ gives ϕ_{2Q}. If leakage from $|11\rangle$ to $|02\rangle$ occurs, $q_{\text{contr.}}$ is in $|0\rangle$ when the second π pulse is applied, adding, instead of removing, an excitation to $q_{\text{contr.}}$. The leakage probability L_1 can be estimated as $\widetilde{L_1} = m/2$, where m is the population difference on the control qubit between both variants of the experiment. Because of relaxation effects, $\widetilde{L_1}$ slightly overestimates L_1.

The simulations model the system realistically and allow us to extract ϕ_{2Q}, L_1 and the average gate fidelity F for a single application of the gate (see Section 6.11.3). The pulse is modeled as a trajectory in a two-qutrit Hamiltonian. The noise model accounts for relaxation and dephasing effects as well as the effect of remaining distortions. The latter are measured using the Cryoscope technique [36]. For the dephasing we take into account the different timescales on which flux noise acts as well as the measured dependence on the flux bias.

6.7. LEAKAGE INTERFERENCE

Both experiment and simulation show a fringe of low leakage [Fig. 6.4(b,d)]. This fringe can be understood as "leakage interference" between $|11\rangle$ and $|02\rangle$ by analogy to a Mach-Zehnder interferometer (see Section 6.11.6). Such analogy has been exploited in a variety of platforms [39–43] to demonstrate coherent control of a single qubit by showing Stückelberg oscillations [44] as a consequence of periodic driving of the qubit into an avoided crossing. Here we pulse in-and-out of $|11\rangle \leftrightarrow |02\rangle$ twice to realize low-leakage two-qubit

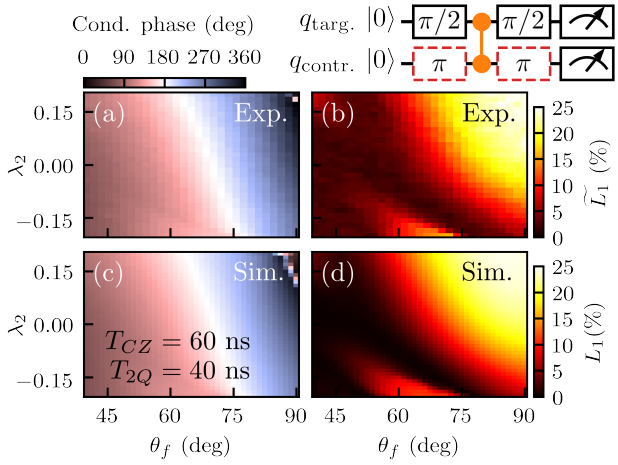

Figure 6.4: Conditional phase (a, c) and leakage (b, d) for a $T_{\text{CZ}} = 60$ ns ($T_{2Q} = 40$ ns) NZ flux pulse as a function of pulse parameters θ_f and λ_2 for both experiment (a, b) and simulation (c, d). The conditional phase increases with θ_f and λ_2, since both of these have the effect of making the pulse spend more time close to the interaction point. Leakage tends to increase significantly with larger values of θ_f with the exception of a diagonal fringe.

gates. The states $|11\rangle$ and $|02\rangle$ correspond to two paths of the interferometer. The first part of the NZ pulse (red in Fig. 6.1) corresponds to the first (imbalanced) beamsplitter. In general, after the first beamsplitter most of the population remains in $|11\rangle$ but part is transferred to $|02\rangle$. Pulsing through the sweetspot (green in Fig. 6.1) corresponds to the arms of the interferometer. The two paths are detuned by ~ 800 MHz, causing a phase to be acquired before the paths are recombined at the second half of the NZ pulse (blue in Fig. 6.1) corresponding to the second beamsplitter. The phase difference between the two paths will cause interference that either enhances or suppresses the leakage to $|02\rangle$.

6.8. PERFORMANCE

Given the good correspondence between experiment and simulation (Fig. 6.4), we can use simulations to explore the parameter space $(\theta_f, \lambda_2, T_{2Q})$ to find the shortest T_{2Q} enabling a high-fidelity, low-leakage CZ gate. The minimum CZ gate duration is fundamentally limited by the coupling strength J_2 as the time required to acquire 180 degrees of conditional phase at the avoided crossing: $T_{2Q} \geq \frac{\pi}{J_2} = 25$ ns. We find a $T_{2Q} = 28$ ns NZ pulse using leakage interference to achieve low leakage. The use of interference is demonstrated by the fact that the corresponding half pulse displays high leakage (see Section 6.11.5). We append $T_{1Q} = 12$ ns flux pulses on both qubits to correct the single-qubit phases, making the total duration of the phase-corrected CZ gate $T_{\text{CZ}} = 40$ ns. We ensure that these phase-correction pulses satisfy Eq. (6.3) and have a sufficiently low amplitude to not affect ϕ_{2Q} and L_1 significantly.

We characterize the performance of the CZ gate using an interleaved randomized benchmarking protocol [7, 45] with modifications that allow us to quantify leakage [38, 46] (see also Sections 3.4.3 and 6.11.7). The randomized benchmarking sequences are based

on 300 random seeds. For each seed, every data point is measured 104 times. We measure an average gate fidelity $F = 99.10\% \pm 0.16\%$ and leakage $L_1 = 0.10\% \pm 0.07\%$ for the NZ pulse with $T_{CZ} = 40$ ns [Fig. 6.5(a,b)]. We could not perform similar measurements for the unipolar pulse since this gate is not repeatable, as demonstrated in Fig. 6.2.

6.9. LIMITING NOISE SOURCES

It is possible to investigate the limits to the performance of the NZ CZ using simulation (see Section 6.11.3) and compare to the unipolar CZ, even though this is not possible in experiment since the unipolar CZ lacks the required characteristic of being repeatable. We simulate these gates for a range of different error models [Fig. 6.5(c,d)]. For each we optimize over θ_f and λ_2 to find the lowest ε and the corresponding L_1. A first observation is that the infidelity ($\varepsilon = 1 - F$) of the NZ gate does not significantly increase when the low-frequency flux-noise components are included, whereas this does affect the unipolar pulse. It appears that the difference in ε between the unipolar and NZ pulses for the full model can be attributed completely to this effect. This observation is consistent with the echo effect demonstrated in Fig. 6.3. Looking at the L_1 error budgets, L_1 is limited by short-timescale distortions. This is understandable as minimizing L_1 requires the pulse to follow a precise trajectory. Distortions also increase ε through L_1 (see Section 6.11.3). The simulations also indicate that dephasing causes leakage. This can be understood as dephasing effectively corresponds to an uncertainty in the energy levels. The simulated L_1 is larger than the measured L_1. This could be explained in two ways, either the distortions are less severe than our estimate, or the simulations, only concerned with a single application of the gate, do not take into account all the relevant effects. Specifically, because the population in the leakage subspace does not completely decohere, this population can seep back into the computational subspace due to an interference effect (similar to that in the NZ pulse itself) at subsequent applications of the gate. Because the first CZ gate cannot benefit from this coherence, the simulations, which only deal with a single CZ gate, slightly overestimate the effective leakage.

6.10. CONCLUSIONS

In summary, we have demonstrated a flux-based CZ gate for transmon qubits that is fast, low-leakage, high-fidelity and repeatable. The gate is realized using a bipolar Net-Zero flux pulse that harnesses leakage interference to achieve speed while maintaining low leakage. The NZ pulse exploits the flux symmetry of the pulsed transmon to build in an echo effect on its single-qubit phase and the conditional phase, increasing fidelity relative to a unipolar pulse. Finally, the action of the NZ pulse is robust to long-timescale distortions in the flux-control line remaining after real-time pre-compensation, enabling the repeatability of the CZ gate. These features make the realized NZ CZ gate immediately useful in high-circuit-depth applications of a full-stack quantum computer in which a controller issues operations to execute on the quantum hardware in real time. For example, Ref. [18] uses NZ CZ gates to stabilize two-qubit entanglement by multi-round indirect parity measurements. Future work will incorporate NZ CZ gates into our scheme [32] to realize a surface-code-based logical qubit [20] with monolithic transmon-cQED quantum hardware.

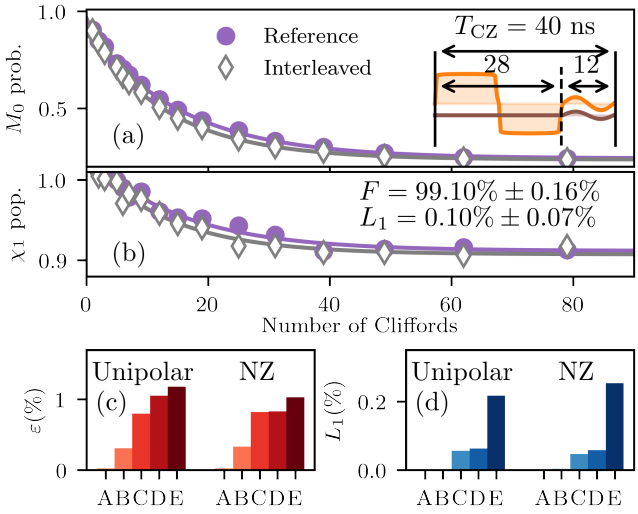

Figure 6.5: Interleaved randomized benchmarking with leakage modification and simulated performance using different error models for a T_{CZ} = 40 ns NZ CZ gate (T_{2Q} = 28 ns), schematically shown in the diagram. (a) Survival probability M_0 of recovering $|00\rangle$ for reference and interleaved two-qubit randomized benchmarking sequence. (b) Population in the computational subspace $\mathscr{X}_1$. Simulated ε (c) and L_1 (d) for different error models (see Section 6.11.3) for T_{CZ} = 40 ns unipolar and NZ pulses (T_{2Q} = 28 ns). The error models (*A* to *E*) contain: no noise (*A*), relaxation (*B*), all Markovian noise components (*C*), Markovian and quasi-static flux noise components (*D*) and all noise components including distortions (*E*).

6.11. METHODS

This section contains detailed information on the experimental protocols and the simulations performed in the first part of this chapter. Section 6.11.1 provides relevant device parameters. Section 6.11.2 describes the parametrization used for the unipolar and NZ pulses. Section 6.11.3 describes the simulations in detail. Section 6.11.4 and Section 6.11.7 describe protocols used to characterize the flux pulses. Section 6.11.5 investigates the limitations of the CZ gate. Section 6.11.6 discusses the Mach-Zehnder interferometer analogy in detail.

6.11.1. DEVICE PARAMETERS

All experiments were performed on a circuit-QED quantum chip containing three starmon-type [32] transmon qubits, labeled q_H, q_M, and q_L. Pairs q_H-q_M and q_M-q_L are coupled by separate bus resonators. Each qubit has a microwave drive line for single-qubit gating, a flux-bias line for local and ns-timescale control of the qubit frequency, and dedicated, fast readout resonators with Purcell protection for the qubits. The readout resonators are coupled to a common feedline, allowing independent readout of the three qubits by frequency multiplexing.

In the first part of this chapter we focus on the transmon pair q_H-q_M. We have achieved similar performance (fidelity, leakage and gate time) for the pair q_M-q_L. Relevant device parameters are given in Table 6.1.

Parameter	q_L	q_M	q_H
$\omega/2\pi$ operating point (GHz)	5.02	5.79	6.87
$\omega/2\pi$ sweetspot (GHz)	5.02	5.79	6.91
$\alpha/2\pi$ (MHz)	-300	-300	-331
$J_1/2\pi$ avoided crossing (MHz)	17.2		14.3
T_1 (μs)	31.8	15.2	19.2
T_2^* operating point (μs)	14.0	14.8	3.2
T_2^E operating point (μs)	33.8	19.4	14.7
$\sim \omega_{bus}/2\pi$ (GHz)	8.5		8.5

Table 6.1: Parameters of the three-transmon device: qubit frequency (ω), anharmonicity (α), exchange coupling between $|01\rangle$ and $|10\rangle$ (J_1), dephasing times (T_1, T_2^*, T_2^E) and bus-resonator frequency (ω_{bus}). Experiments in the first part of this chapter are performed with the pair q_H-q_M. q_H is operated 40 MHz below its sweetspot to minimize interaction with a spurious two-level system right at the sweetspot frequency.

6.11.2. FLUX PULSE PARAMETRIZATION

Unipolar and NZ pulses are based on the Martinis-Geller parametrization for fast-adiabatic gates [31]. This parametrization is determined by the Hamiltonian [Eq. (6.11)] projected onto a two-dimensional subspace. In the case of the CZ gate, this subspace is spanned by the states $|11\rangle$ and $|02\rangle$. The projected Hamiltonian, $H_{subspace}$, takes the form

$$H_{subspace} = \begin{pmatrix} \frac{\epsilon}{2} & J_2 \\ J_2 & -\frac{\epsilon}{2} \end{pmatrix}, \tag{6.4}$$

where $\epsilon = \omega_{|02\rangle} - \omega_{|11\rangle}$ is the bare detuning between $|11\rangle$ and $|02\rangle$ and J_2 is their coupling. The detuning ϵ is controlled by flux whereas J_2 is considered to be constant. We define the angle θ as

$$\theta \equiv \arctan\left(\frac{2J_2}{\epsilon}\right). \tag{6.5}$$

Note that $\theta = \pi/2$ at $\epsilon = 0$.

The waveform is expressed as a series

$$\theta(\tau(t)) = \theta_i + \sum_{j=1}^{N} \lambda_j \left(1 - \cos\left(\frac{2\pi \cdot j \cdot \tau(t)}{T_{2Q}}\right)\right), \tag{6.6}$$

where T_{2Q} is the pulse duration, θ_i corresponds to the detuning at the operating point and τ is proper time, which is related to real time t through $t(\tau) = \int_0^\tau d\tau' \sin\big(\theta(\tau')\big)$.

We truncate the series to $N = 2$. We make use of the relation between the angle at the middle of the unipolar pulse (θ_f) and the odd λ coefficients

$$\theta_f \equiv \theta(T_{2Q}/2) = \theta_i + 2 \sum_{j\,\text{odd}}^{N} \lambda_j, \tag{6.7}$$

to define the entire waveform using three parameters: θ_f, λ_2, and T_{2Q}. A NZ pulse is a sequence of two concatenated unipolar pulses, each lasting $T_{2Q}/2$ time and with the same θ_f and λ_2.

There are a few more transformations required in order to have a waveform in terms of the flux $\Phi_{\text{target}}(t)$ [Fig. 6.6]:

$$\theta(t) \mapsto \epsilon(t) \mapsto \omega_{q_\text{H}}(t) \mapsto \Phi_{\text{target}}(t). \tag{6.8}$$

The first transformation uses Eq. (6.5): $\epsilon(t) = 2J_2/\tan\theta(t)$. The second one uses the fact that by definition $\epsilon(t) = \omega_{|02\rangle}(t) - \omega_{|11\rangle}(t) = \omega_{q_\text{H}}(t) + \alpha_{q_\text{H}} - \omega_{q_\text{M}}$. The qubit frequency depends on flux according to the formula

$$\omega_{q_\text{H}}(\Phi) = (\omega^0_{q_\text{H}} - \alpha_{q_\text{H}})\sqrt{\left|\cos\left(\frac{\Phi}{\Phi_0}\pi\right)\right|} + \alpha_{q_\text{H}}, \tag{6.9}$$

where $\omega^0_{q_\text{H}}$ is the sweetspot frequency and α_{q_H} the anharmonicity, reported in Table 6.1. We refer to this relation between frequency and flux as the flux arc. The flux arc has been measured in the experiment and we find that it matches well with Eq. (6.9). We invert Eq. (6.9) to convert $\omega_{q_\text{H}}(t) \mapsto \Phi_{\text{target}}(t)$. Since $\omega_{q_\text{H}}(\Phi) = \omega_{q_\text{H}}(-\Phi)$, there is a positive and a negative solution for every value of ω_{q_H}. In the case of a unipolar pulse, we always consider the positive solution, whereas, in the case of a NZ pulse, the first and second half of the pulse use the positive and negative solutions, respectively. Changes that are clearly visible in the θ parametrization correspond to only a small change in the applied flux. This provides intuition why even a small distortion of the applied flux can have a relatively large effect on the gate quality.

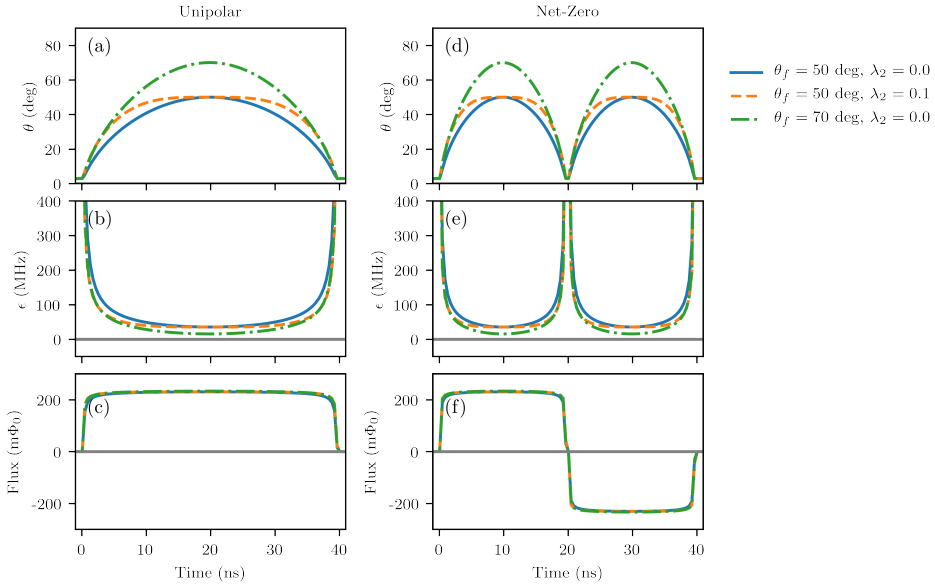

Figure 6.6: Unipolar (a-c) and NZ pulses (d-e) represented in terms of θ (a, d), bare detuning ϵ (b, e) and flux Φ (c, f). The center of the unipolar pulse is controlled by θ_f, while λ_2 controls the sharpness of rise and fall of the pulse.

6

6.11.3. SIMULATION STRUCTURE

The simulations model the system, consisting of two coupled transmons, using a two-qutrit Hamiltonian. One of the two transmons, namely q_H, is actively pulsed into resonance according to the pulse parametrization described in Section 6.11.2. The simulations (Fig. 6.7) include distortions, relaxation and flux-dependent dephasing effects. The error model also includes a distinction between Markovian (fast) and non-Markovian (slow) noise in order to accurately model dephasing effects. The simulations are used to calculate the propagator or time-evolution superoperator, from which the quantities of interest - fidelity, leakage and conditional phase - are extracted.

SYSTEM HAMILTONIAN

The system is composed of two transmons coupled via a bus resonator. We exclude the resonator from the model by making the assumption that it always remains in its ground state (it is excited only "virtually"). We restrict each transmon to its first three energy levels. In the dispersive regime, in the rotating-wave approximation, the Hamiltonian is given by

$$H(t) = \omega_{q_\mathrm{M}} a^\dagger_{q_\mathrm{M}} a_{q_\mathrm{M}} + \frac{\alpha_{q_\mathrm{M}}}{2}(a^\dagger_{q_\mathrm{M}})^2 a^2_{q_\mathrm{M}} + \omega_{q_\mathrm{H}}(\Phi(t))\, a^\dagger_{q_\mathrm{H}} a_{q_\mathrm{H}} + \frac{\alpha_{q_\mathrm{H}}}{2}(a^\dagger_{q_\mathrm{H}})^2 a^2_{q_\mathrm{H}} \tag{6.10}$$

$$+ J_1(\Phi(t))\,(a_{q_\mathrm{M}} a^\dagger_{q_\mathrm{H}} + a^\dagger_{q_\mathrm{M}} a_{q_\mathrm{H}}), \tag{6.11}$$

where only the higher-frequency transmon (q_H) is actively fluxed. Here a_{q_i} is the annihilation operator restricted to the first three energy levels, ω_{q_i} and α_{q_i} are the qubit

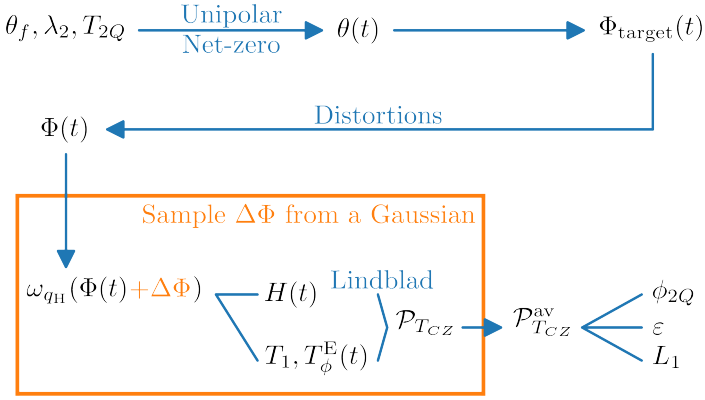

Figure 6.7: The parameters θ_f, λ_2 and the gate time T_{2Q} determine either a unipolar pulse or a NZ pulse in terms of $\theta(t)$, see Eq. (6.6). $\theta(t)$ is converted into $\Phi_{\text{target}}(t)$ thorough various transformations described in Section 6.11.2. Pulse distortions are applied by convolution to compute $\Phi(t)$ experienced by the qubit. The solution of the Lindblad equation is the time-evolution superoperator $\mathcal{P}_{T_{CZ}}$. Averaging over a Gaussian distribution for the quasi-static flux bias $\Delta\Phi$, we obtain the average superoperator $\mathcal{P}^{av}_{T_{CZ}}$. From that any quantity of interest can be computed, in particular the conditional phase ϕ_{2Q}, the average gate infidelity ε and the leakage L_1.

frequency and anharmonicity, respectively, and J_1 is the coupling. The coupling is weakly flux-dependent since $J_1(\Phi) \approx \frac{g_{q_M} g_{q_H}}{2}(\Delta_{q_M}^{-1} + \Delta_{q_H}^{-1}(\Phi))$, with g_{q_i} the coupling of q_i to the bus resonator and $\Delta_{q_i} \approx \omega_{\text{bus}} - \omega_{q_i} \gg g_{q_i}$ given the parameters in Table 6.1. When we generate the flux pulse according to Section 6.11.2, we consider $J_2 = \sqrt{2}J_1$ to be constant and J_2 equal to its measured value at the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing, whereas in the simulations we take into account the dependence of J_1 and J_2 on Φ.

DISTORTIONS

The flux pulse at the qubit is subject to distortions altering the shape of the waveform as experienced by the qubit. Distortions are described as a linear time-invariant system fully characterized by the impulse response h of the system. We best compensate such distortions by predistorting the desired pulse $\Phi_{\text{target}}(t)$ with an impulse response $\tilde{h}^{-1}$ designed to invert h. Then, the actual pulse $\Phi(t)$ experienced by the qubit is given by

$$\Phi(t) = (h * V_{\text{AWG}})(t) = (h * (\tilde{h}^{-1} * \Phi_{\text{target}}))(t) = ((\tilde{h}^{-1} * h) * \Phi_{\text{target}})(t), \tag{6.12}$$

where $*$ denotes convolution. The distortions remaining after applying $\tilde{h}^{-1}$ are determined by measuring the step response $s(t) = \int_0^t dt' (\tilde{h}^{-1} * h)(t')$ (Fig. 6.8) using the Cryoscope technique [36]. The impulse response extracted from these data is used to distort the pulses in simulations.

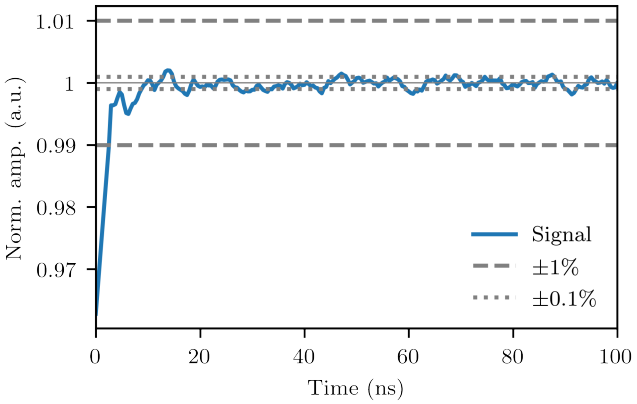

Figure 6.8: Step response at the qubit after applying distortion corrections, measured using the Cryoscope technique [36]. The impulse response extracted from this experiment is used to distort the pulses in the simulations. In the case of perfect distortion corrections, the normalized amplitude would have value 1 for all times larger than zero.

NOISE MODEL

There are two major error sources in superconducting qubits: relaxation and flux noise. The latter has a power spectral density $S_f \sim A/f$, where f is frequency and $\sqrt{A}$ is a constant of the order of 10 $\mu\Phi_0$, with Φ_0 the flux quantum. S_f contains both high-frequency and low-frequency components: we phenomenologically distinguish high and low frequencies depending on whether they are larger or smaller than $1/T_{\mathrm{CZ}}$. Relaxation and high-frequency flux-noise components are Markovian noise processes since they act on a timescale shorter than the gate time. On the other hand, the low-frequency flux-noise components determine a non-Markovian noise process, since they induce correlations across different gates.

We perform two experiments to quantify the strength of the dephasing affecting q_{H}: a Ram-Z and an Echo-Z experiment [Fig. 6.9]. In these experiments, the dephasing times $T^*_{2,q_{\mathrm{H}}}(\Phi)$ and $T^{\mathrm{E}}_{2,q_{\mathrm{H}}}(\Phi)$, respectively, at different flux sensitivities $\frac{1}{2\pi}\frac{\partial\omega_{q_{\mathrm{H}}}}{\partial\Phi}$ are measured while applying a flux pulse. In the Ram-Z experiment, this flux pulse is square. In the Echo-Z experiment, the flux pulse consists of two square half pulses that detune the qubit by the same amount in magnitude but with opposite-sign sensitivity. We perform these experiments for a range of fluxes. The experimental data for q_{H} is represented in Fig. 6.9. On the other hand, the static qubit q_{M} is always operated at the sweetspot. Therefore, we only use the measured Ramsey and Echo dephasing times at the sweetspot, reported in Table 6.1. The relaxation times $T_{1,q_{\mathrm{H}}}$ and $T_{1,q_{\mathrm{M}}}$, are also reported in this table.

We assume that the low-frequency flux-noise components are echoed out in an Echo-Z experiment. In other words, we assume that T_{1,q_i}, $T^{\mathrm{E}}_{2,q_i}(\Phi)$ quantify the strength of the Markovian noise. On the other hand, we assume that T_{1,q_i}, $T^*_{2,q_i}(\Phi)$ quantify the strength of the overall noise (both Markovian and non-Markovian). The strength of the non-Markovian noise alone cannot be extracted directly from the experiment. However, in the following we explain the model that we use fitting the experimental data (Fig. 6.9). In this way we can simulate separately both the Markovian and non-Markovian noise,

and obtain a realistic simulation of the system.

Model of Markovian noise.
A Markovian evolution is modeled with the Lindblad equation

$$\dot{\rho}(t) = -i[H(t), \rho(t)] + \sum_{j,q_i} \left(c_{j,q_i}(t)\rho(t)c^{\dagger}_{j,q_i}(t) - \frac{1}{2}\{c^{\dagger}_{j,q_i}(t)c_{j,q_i}(t), \rho(t)\} \right) =: \mathscr{L}_t\big(\rho(t)\big), \tag{6.13}$$

where $\mathscr{L}_t$ is the time-dependent Lindbladian defined by the Hamiltonian [Eq. (6.11)] and by the jump operators $\{c_{j,q_i}(t)\}$ specified in Eqs. (6.14) and (6.16) to (6.18) below.

To model relaxation, we use the jump operator

$$c_{0,q_i} = \sqrt{\frac{1}{T_{1,q_i}}}\, a_{q_i}. \tag{6.14}$$

To model pure dephasing, we first define a pure-dephasing time

$$T^{\mathrm{E}}_{\phi,q_i}(\Phi) = \left(\frac{1}{T^{\mathrm{E}}_{2,q_i}(\Phi)} - \frac{1}{2T_{1,q_i}} \right)^{-1}, \tag{6.15}$$

Ignoring relaxation-induced dephasing in this paragraph, the coherence $\langle 0|\rho_{q_i}(\Phi)|1\rangle$ decays as $e^{-t/T^{\mathrm{E}}_{\phi,q_i}(\Phi)}$, where ρ_{q_i} is the qutrit reduced density matrix. In Fig. 6.9 we see that the decay rates have a linear dependence on the flux sensitivity. Ignoring the anharmonicity, the frequency of the $|2\rangle$ state is twice the frequency of the $|1\rangle$ state, therefore, the sensitivity of the $|2\rangle$ state is twice as high. Given these two observations, we assume that $\langle 0|\rho_{q_i}(\Phi)|2\rangle \propto e^{-t/(T^{\mathrm{E}}_{\phi,q_i}(\Phi)/2)}$ and $\langle 1|\rho_{q_i}(\Phi)|2\rangle \propto e^{-t/T^{\mathrm{E}}_{\phi,q_i}(\Phi)}$. We find that such decay rates can be realized by the following jump operators

$$c_{1,q_i}(\Phi(t)) = \sqrt{\frac{8}{9T^{\mathrm{E}}_{\phi,q_i}(\Phi(t))}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}_{q_i}, \tag{6.16}$$

$$c_{2,q_i}(\Phi(t)) = \sqrt{\frac{2}{9T^{\mathrm{E}}_{\phi,q_i}(\Phi(t))}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}_{q_i}, \tag{6.17}$$

$$c_{3,q_i}(\Phi(t)) = \sqrt{\frac{2}{9T^{\mathrm{E}}_{\phi,q_i}(\Phi(t))}} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}_{q_i}. \tag{6.18}$$

Instead, if one would use only

$$c'_{1,q_i}(\Phi(t)) = \sqrt{\frac{2}{T^{\mathrm{E}}_{\phi,q_i}(\Phi(t))}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}_{q_i}, \tag{6.19}$$

which produces the same Lindbladian as $c''_{1,q_i}(\Phi(t)) = \sqrt{2/T^{\mathrm{E}}_{\phi,q_i}(\Phi(t))}\, a^\dagger_{q_i} a_{q_i}$, then one would get $\langle 0|\rho_{q_i}(\Phi)|2\rangle \propto e^{-t/(T^{\mathrm{E}}_{\phi,q_i}(\Phi)/4)}$ and $\langle 1|\rho_{q_i}(\Phi)|2\rangle \propto e^{-t/T^{\mathrm{E}}_{\phi,q_i}(\Phi)}$. This means that Eq. (6.19) would be the correct modeling if the decay rates in Fig. 6.9 would depend quadratically on the sensitivity, but they do not.

The formal solution of Eq. (6.13) is given by

$$\rho(t) = \mathscr{T}\, e^{\int_0^t dt'\, \mathscr{L}_{t'}} \big(\rho(0)\big), \tag{6.20}$$

where $\mathscr{T}$ is the time-ordering operator. We call $\mathscr{P}_{T_{\mathrm{CZ}}} := \mathscr{T}\, e^{\int_0^{T_{2Q}} dt'\, \mathscr{L}_{t'}}$ the propagator or time-evolution superoperator, evaluated up to the gate time T_{CZ}, which includes an idling time T_{1Q} to account for the noise during the single-qubit phase correction pulses. The propagator $\mathscr{P}_{T_{\mathrm{CZ}}}$ can be computed by solving the differential Eq. (6.13), or as

$$\mathscr{P}_{T_{\mathrm{CZ}}} \simeq e^{\delta t \mathscr{L}_{T_{\mathrm{CZ}}-\delta t}}\, e^{\delta t \mathscr{L}_{T_{\mathrm{CZ}}-2\delta t}} \dots e^{\delta t \mathscr{L}_{2\delta t}}\, e^{\delta t \mathscr{L}_{\delta t}}\, e^{\delta t \mathscr{L}_0}, \tag{6.21}$$

for a sufficiently small δt. In the simulations we use $\delta t = 0.1$ ns. In the Liouville representation, this equation is a product of matrices. We find that this method is an order of magnitude faster than using the qutip [47] differential equation solver.

Model of non-Markovian noise.
We model the low-frequency flux-noise components as quasi-static. Since the static qubit q_{M} is always operated at the sweetspot, where the sensitivity to flux noise is zero, we apply this model only to q_{H}. We assume that the qubit experiences a random, fixed flux offset $\Delta\Phi$ during the execution of a gate, but that $\Delta\Phi$ varies across different gates. For $\Delta\Phi \ll 1$, the effect of such offset on the pulse trajectory can be approximated at first order as $\omega_{q_{\mathrm{H}}}(\Phi(t)+\Delta\Phi) \approx \omega_{q_{\mathrm{H}}}(\Phi(t)) + \frac{\partial\omega_{q_{\mathrm{H}}}(\Phi(t))}{\partial\Phi}\Delta\Phi$, where $\frac{1}{2\pi}\frac{\partial\omega_{q_{\mathrm{H}}}(\Phi)}{\partial\Phi}$ is the flux sensitivity. Using Eq. (6.9) we can see that $\frac{\partial\omega_{q_{\mathrm{H}}}(\Phi)}{\partial\Phi} = -\frac{\partial\omega_{q_{\mathrm{H}}}(-\Phi)}{\partial\Phi}$. In the case of a NZ pulse, this implies that first-order frequency variations in the first half of the pulse are canceled by an equal and opposite variation in the second half, resulting in an echo effect.

We take the probability distribution p_σ of $\Delta\Phi$ to be Gaussian $p_\sigma(\Delta\Phi) = e^{-(\Delta\Phi)^2/(2\sigma^2)}/(\sqrt{2\pi}\sigma)$, where σ is the standard deviation of the Gaussian. Averaging over this distribution, we get the final propagator

$$\mathscr{P}^{\mathrm{av}}_{T_{\mathrm{CZ}}} = \int_{-\infty}^{+\infty} d(\Delta\Phi)\; p_\sigma(\Delta\Phi)\cdot\mathscr{P}_{T_{\mathrm{CZ}}}(\Delta\Phi), \tag{6.22}$$

which gives the time evolution including all the noise sources in the model, both Markovian and non-Markovian.

The standard deviation σ is not directly measured in the experiment. Instead, we fit this model to the experiment simulating a Ram-Z and Echo-Z experiment for q_{H} (Fig. 6.9). We vary the value of σ while keeping the Markovian noise model described above fixed. We find that the value $\sigma = 55\ \mu\Phi_0$ best fits both the Ram-Z and Echo-Z data at the same time. This is the value we use in all the simulations in the first part of this chapter.

QUANTITIES OF INTEREST

To quantify the quality of the CZ gate, we are interested in computing the conditional phase, the leakage and the average gate fidelity from the propagator $\mathscr{P}^{\mathrm{av}}_{T_{\mathrm{CZ}}}$. In the following, we summarize their definitions for a generic superoperator $\mathscr{P}$.

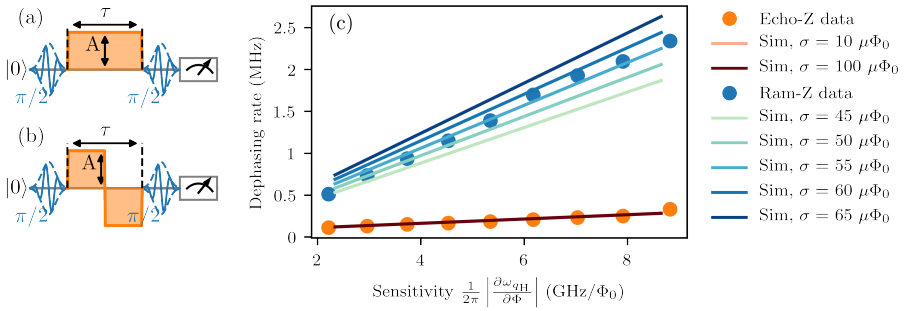

Figure 6.9: Comparison of experimental data and simulation (c) for the Ram-Z (a) and Echo-Z (b) experiments. In the Ram-Z (Echo-Z) experiment, the dephasing time is measured using a (two-half) square flux-pulse(s). All simulated curves include the effects of both the Markovian and non-Markovian noise. Only the strength of the non-Markovian noise [Eq. (6.22)], quantified by σ, is varied, while the strength of the Markovian noise, quantified by T_{1,q_H} and $T^{E}_{\phi,q_H}(\Phi)$, is kept fixed. We see that the value $\sigma = 55\,\mu\Phi_0$ best fits the Ram-Z data. It fits the Echo-Z data as well, given that the simulated curves are equal even for σ's that differ by an order of magnitude. This agrees with the intuition that the non-Markovian noise is echoed-out in an Echo-Z experiment.

We call $\mathscr{X}_1$ the computational subspace, spanned by the 2-qubit energy levels $|00\rangle$, $|01\rangle$, $|10\rangle$ and $|11\rangle$ at the operating point. The phases acquired by those states under the action of $\mathscr{P}$ are computed as

$$e^{i\phi_{ij}} = \frac{\langle ij|\mathscr{P}\big(|ij\rangle\langle 00|\big)|00\rangle}{\big|\langle ij|\mathscr{P}\big(|ij\rangle\langle 00|\big)|00\rangle\big|}, \tag{6.23}$$

where $i, j \in \{0,1\}$. If $\mathscr{P}$ is unitary, that is, $\mathscr{P}(\rho) = U\rho U^\dagger$ for some unitary U, then Eq. (6.23) reduces to $e^{i\phi_{ij}} = \frac{\langle ij|U|ij\rangle}{|\langle ij|U|ij\rangle|}$, and, if U is diagonal, then we simply have $U|ij\rangle = e^{i\phi_{ij}}|ij\rangle$. The phase ϕ_{00} of the ground state can be set to 0. The single-qubit phases are given by ϕ_{01} and ϕ_{10}. The conditional phase ϕ_{2Q} is defined as the phase acquired by the target qubit conditional on the state of the control qubit and it is given by

$$\phi_{2Q} = \phi_{11} - \phi_{10} - \phi_{01}. \tag{6.24}$$

Note that ϕ_{2Q} is invariant under single-qubit Z rotations.

We follow the definitions in [38] for leakage, seepage and average gate fidelity (see also Section 3.4.3). The leakage of a superoperator $\mathscr{P}$ is defined as

$$\begin{aligned} L_1 &= 1 - \int_{\psi_1\in\mathscr{X}_1} d\psi_1\ \mathrm{Tr}_{\mathscr{X}_1}\Big(\mathscr{P}\big(|\psi_1\rangle\langle\psi_1|\big)\Big) \\ &= 1 - \frac{1}{\dim\mathscr{X}_1}\sum_{i,j\in\{0,1\}}\mathrm{Tr}_{\mathscr{X}_1}\Big(\mathscr{P}\big(|ij\rangle\langle ij|\big)\Big). \end{aligned} \tag{6.25}$$

The quantity L_1 represents the average probability that a random computational state leaks out of $\mathscr{X}_1$.

The seepage of a superoperator $\mathscr{P}$ is defined as

$$L_2 = 1 - \int_{\psi_2\in\mathscr{X}_2} d\psi_2\ \mathrm{Tr}_{\mathscr{X}_2}\Big(\mathscr{P}\big(|\psi_2\rangle\langle\psi_2|\big)\Big), \tag{6.26}$$

where $\mathscr{X}_2$ is the leakage subspace.

The average gate fidelity, evaluated in the computational subspace, between $\mathscr{P}$ and a target unitary U is defined as

$$F = \int_{\psi_1 \in \mathscr{X}_1} d\psi_1 \, \langle\psi_1| U^\dagger \mathscr{P}\big(|\psi_1\rangle\langle\psi_1|\big) U |\psi_1\rangle \tag{6.27}$$
$$= \frac{\dim \mathscr{X}_1 (1 - L_1) + \sum_k \left|\mathrm{Tr}_{\mathscr{X}_1}(U^\dagger A_k)\right|^2}{\dim \mathscr{X}_1 (\dim \mathscr{X}_1 + 1)},$$

where the $\{A_k\}$ are the Kraus operators of $\mathscr{P}$. The average gate infidelity is defined as $\varepsilon = 1 - F$. We can see from Eq. (6.27) that F is affected by L_1. There are two contributions: one is explicit in the first term at the numerator, the other is implicit in the second term and is due to the fact that the Kraus operators of a leaky superoperator are in general different from the ones of a non-leaky superoperator. For a two-qubit gate, the explicit contribution to ε is equal to $L_1/5$, whereas the implicit one is evaluated numerically.

6.11.4. CONDITIONAL OSCILLATION EXPERIMENT

The conditional oscillation experiment (Fig. 6.10) can be used to measure the single-qubit phases (ϕ_{01} and ϕ_{10}) and the conditional phase (ϕ_{2Q}), and to estimate the leakage (L_1) defined in Eq. (6.25). In the conditional oscillation experiment, two variants of the same experiment are performed. In the first variant (Off), the target qubit ($q_{\text{targ.}}$) is rotated onto the equator of the Bloch sphere by a $\pi/2$ pulse and the control qubit ($q_{\text{contr.}}$) is left in the ground state. After that, a flux pulse is applied that is intended to perform a CZ gate. A recovery $\pi/2$ rotation, performed around an axis in the equatorial plane forming an angle ϕ with the X axis, is applied to $q_{\text{targ.}}$ before measuring the state of both qubits simultaneously. In the second variant (On), $q_{\text{contr.}}$ is rotated into the excited state before applying the CZ gate. Then, $q_{\text{contr.}}$ is pulsed back to the ground state before measuring both qubits.

The conditional phase ϕ_{2Q} can be extracted directly from the phase of the oscillations and corresponds to the difference in phase between the oscillations (Figure 6.10). The single-qubit phase ϕ_{10} (ϕ_{01}) can be measured by letting q_{M} (q_{H}) take the role of $q_{\text{targ.}}$ and correspond directly to the measured phase of $q_{\text{targ.}}$ in the Off variant.

The quantity denoted by m in Figure 6.10 is called the missing fraction. In the idealized case in which there is no noise and no leakage to other levels, we calculate $L_1 = m_{\text{idealized}}/2$. We see numerically that such relation approximately holds in the complete modeling with noise. Therefore, we define a leakage estimator $\widetilde{L_1} = m/2$, where m is the measured value. Due to relaxation effects, $\widetilde{L_1}$ generally overestimates L_1. The advantage of estimating the leakage with $\widetilde{L_1}$ rather than with a randomized benchmarking experiment (Section 6.11.7) is that it is much faster. In this way we can quickly acquire a scan of the leakage landscape to find pulse parameters giving a low-leakage CZ gate. Further characterization is then carried out with randomized benchmarking.

6.11.5. OPTIMAL PERFORMANCE

Using simulations, it is possible to find the optimal parameters (θ_f and λ_2) for a given T_{2Q} in order to perform a CZ gate. We optimize over the infidelity ε. In Fig. 6.11, the minimal

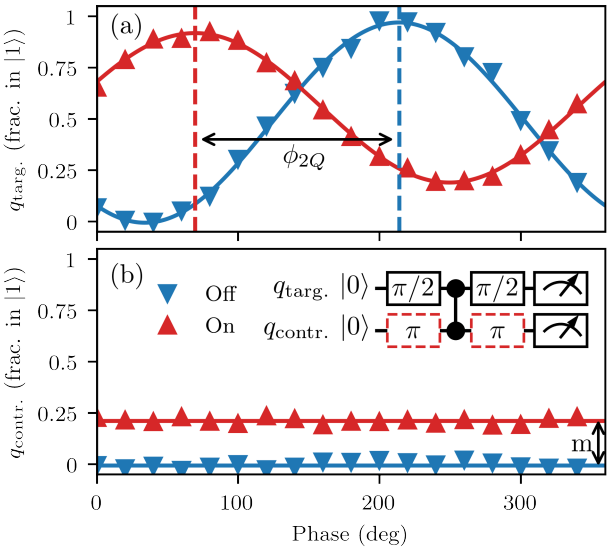

Figure 6.10: The conditional oscillation experiment described in Section 6.11.4.

infidelity ε and the corresponding leakage L_1 are shown as a function of T_{2Q}. Contrary to all the other figures in the first part of this chapter, the simulations shown in Fig. 6.11 do not include the effect of distortions. The shortest duration for which a NZ pulse with low leakage and high fidelity can be performed is $T_{2Q} = 28$ ns, close to the speed limit of $T_{2Q} = 25$ ns, set by the interaction strength. The difference in minimal infidelity between the unipolar and the NZ pulse is attributed to the built-in echo effect that makes the NZ pulse resilient to low-frequency flux-noise components. Unipolar pulses with good performance could in principle be realized slightly faster ($T_{2Q} = 26$ ns) than NZ pulses, due to the fact that NZ needs ~ 2 ns to sweep from one avoided crossing to the other in the middle of the pulse, during which no conditional phase is accumulated. We remark that we can study the performance of a single application of the unipolar CZ gate in simulation, but that this is not representative of the performance in the experiment since the unipolar pulse is not repeatable as demonstrated in Section 6.4.

The simulated landscape of the shortest duration ($T_{2Q} = 28$ ns) high-fidelity low-leakage NZ pulse is compared to experiment in Fig. 6.12. There is a relatively large region of low leakage at high θ_f (90-130 deg) that can be found in both simulation and experiment. The $T_{2Q} = 28$ ns pulses described in Section 6.8 are operated at the marked point ($\theta_f = 125$ deg, $\lambda_2 = -0.1$).

6.11.6. NET-ZERO PULSES AS A MACH-ZEHNDER INTERFEROMETER

To better understand the working of a NZ pulse, it is helpful to draw an analogy to a Mach-Zehnder interferometer [39–44]. In a NZ pulse, the trajectory first approaches the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing at positive flux amplitude, then it sweeps through the sweetspot, and it finally goes in and out of the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing at negative

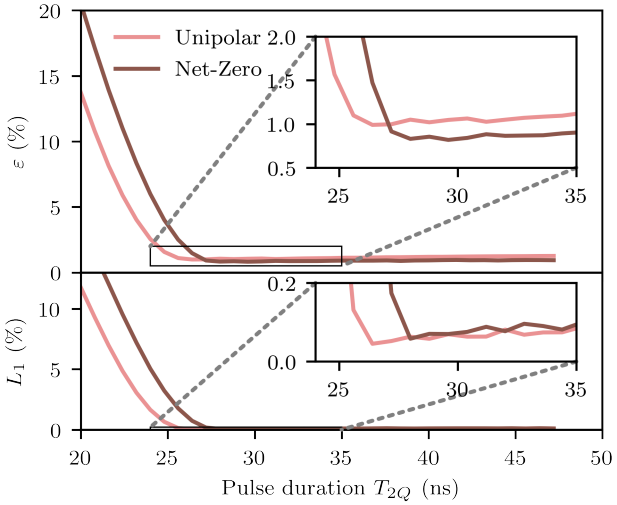

Figure 6.11: Minimal infidelity (ε), optimized over θ_f and λ_2 for a fixed T_{2Q} (T_{1Q} = 12 ns for all T_{2Q}), and leakage (L_1) evaluated at the minimal infidelity. Contrary to all other figures in the first part of this chapter, the simulations shown here do not include distortions because we want to quantify the intrinsic optimal performance of unipolar and NZ pulses against Markovian and non-Markovian noise. We see that both ε and L_1 decrease fast approaching the speed limit $\pi/J_2 \sim 25$ ns. Then NZ achieves lower infidelity and we can attribute this to the echo effect. We can use these simulations to find that the minimal T_{2Q} to realize a high-fidelity, low-leakage NZ pulse is T_{2Q} = 28 ns.

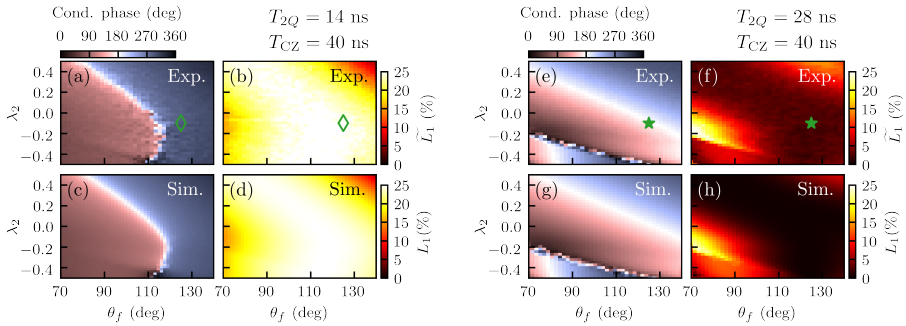

Figure 6.12: Matching of experimental (a,b;e,f) and simulated (c,d;g,h) landscapes of conditional phase and leakage as a function of the parameters θ_f and λ_2 of a T_{2Q} = 14 ns unipolar (a,b,c,d) pulse and of a T_{2Q} = 28 ns NZ (e,f,g,h) pulse. The T_{2Q} = 28 ns NZ pulse consists of two concatenated T_{2Q} = 14 ns unipolar pulses with opposite polarity. Phase corrections are appended to get a total length T_{CZ} = 40 ns. We find that the matching is excellent in both cases. The star (green) marks the point (θ_f = 125 deg, λ_2 = −0.1) used in the interleaved randomized benchmarking experiment described in Section 6.8. A transparent diamond (green) marks the corresponding point for the T_{2Q} = 14 ns unipolar pulse. Given that the T_{2Q} = 14 ns unipolar pulse does not show regions of low leakage, we conclude that the broad area of low leakage for the T_{2Q} = 28 ns NZ pulse is a fringe of destructive leakage interference. We have verified this also by varying the interference condition and observing this fringe move across the landscape, similarly to Fig. 6.14 and as described in Section 6.11.6.

flux amplitude. We argue that those three parts of the pulse correspond respectively to an (unbalanced) beamsplitter, to the arms of an interferometer, and to another (identical) beamsplitter. We make a few idealizations in this analysis. Namely, we ignore the weak coupling to other states and we consider a purely unitary process. Moreover, there is not a clear-cut separation between the beamsplitters, where the qubits are strongly coupled, and the arms of the interferometer, where they are effectively uncoupled. However, since the sweep in the middle is very fast, for the sake of this model it does not really matter where the line is drawn.

In general, a unipolar pulse has the following effect on the $|11\rangle$ state

$$|11\rangle \mapsto e^{i\phi_{2Q}^{\text{half}}}\sqrt{1-\alpha^2}\,|11\rangle + \alpha\,|02\rangle\,, \tag{6.28}$$

where $\alpha \in \mathbb{R}$ and $\alpha^2 = 4L_1^{\text{half}}$ (assuming no leakage to other states). In other words, during the first half of a NZ pulse, $|11\rangle$ acquires a certain conditional phase ϕ_{2Q}^{half} and it can also leak to $|02\rangle$, for example if the parameters of the pulse are not properly chosen or if the pulse is too short.

Unitarity implies that $|02\rangle \mapsto \alpha\,|11\rangle - e^{-i\phi_{2Q}^{\text{half}}}\sqrt{1-\alpha^2}\,|02\rangle$. Overall, modulo a global phase, this amounts to the unitary

$$B_1 = \begin{pmatrix} e^{i\phi_{2Q}^{\text{half}}}\sqrt{1-\alpha^2} & \alpha \\ \alpha & -e^{-i\phi_{2Q}^{\text{half}}}\sqrt{1-\alpha^2} \end{pmatrix}, \tag{6.29}$$

which is a beamsplitter that also imparts a conditional phase.

During the sweep across the sweetspot, $|11\rangle$ and $|02\rangle$ quickly acquire a relative phase φ due to the large energy gap between them (~ 800 MHz). We can formalize this with the unitary

$$P_\varphi = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\varphi} \end{pmatrix}, \tag{6.30}$$

which is a phase shifter.

The second beamsplitter, B_2, is equal to B_1 due to the symmetry of the pulse. The total evolution is given by

$$B_2 P_\varphi B_1 = B_1 P_\varphi B_1 = \begin{pmatrix} e^{i2\phi_{2Q}^{\text{half}}}\left((1-\alpha^2)+\alpha^2 e^{i\tilde{\varphi}}\right) & \alpha\sqrt{1-\alpha^2}\,e^{i\phi_{2Q}^{\text{half}}}(1-e^{i\tilde{\varphi}}) \\ \alpha\sqrt{1-\alpha^2}\,e^{i\phi_{2Q}^{\text{half}}}(1-e^{i\tilde{\varphi}}) & \alpha^2+(1-\alpha^2)e^{i\tilde{\varphi}} \end{pmatrix}, \tag{6.31}$$

where $\tilde{\varphi} := \varphi - 2\phi_{2Q}^{\text{half}}$. We are interested in the first matrix element because it gives the leakage L_1^{NZ} and conditional phase ϕ_{2Q}^{NZ} at the end of a NZ pulse. Explicitly

$$L_1^{\text{NZ}} = \left(\alpha^4 + (1-\alpha^2)^2 + 2\alpha^2(1-\alpha^2)\cos\tilde{\varphi}\right)/4, \tag{6.32}$$

$$\phi_{2Q}^{\text{NZ}} = 2\phi_{2Q}^{\text{half}} + \arctan\left(\frac{\alpha^2\sin\tilde{\varphi}}{(1-\alpha^2)+\alpha^2\cos\tilde{\varphi}}\right).. \tag{6.33}$$

There are two cases in which L_1^{NZ} can be made zero. The first one is when $\alpha^2 = 0$. This is when the half pulse has zero leakage in the first place. We refer to this case as the adiabatic

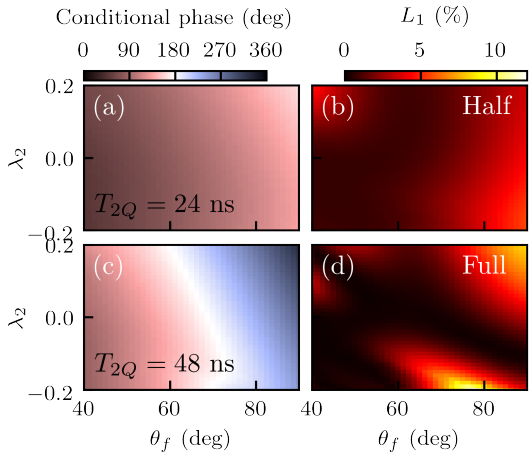

Figure 6.13: Simulation of conditional phase and leakage landscapes as a function of the parameters θ_f, λ_2 of a half (a,b) and full (c,d) $T_{2Q} = 48$ ns NZ pulse ($T_{CZ} = 60$ ns). The half pulse consists of only the first part of the NZ pulse, which is effectively a $T_{2Q} = 24$ ns unipolar pulse ($T_{CZ} = 60$ ns). Naively one may expect both the conditional phase and the leakage of the full pulse to be approximately twice that of the half-pulse. However, this is not the case for the leakage. In (b) we see a low-leakage area due to the adiabaticity of the pulse. We find this low-leakage area in (d) as well. However, an interference fringe is visible that does not occur for the half pulse.

condition. The second case is when $\alpha^2 \neq 0$ but $\tilde{\varphi} = (2k+1)\pi$, with k an integer. We refer to this second case as the interference condition. We point out that, in either case, the second term in Eq. (6.33) is zero, which implies that $\phi_{2Q}^{\mathrm{NZ}} = 2\phi_{2Q}^{\mathrm{half}}$ whenever $L_1^{\mathrm{NZ}} = 0$. As a consequence, the speed limit to do a NZ CZ with low leakage is the same as for the unipolar pulse (π/J_2). We also note that if $L_1^{\mathrm{half}} = \alpha^2/4$ is large and if L_1^{NZ} is low, it follows that the latter must result from destructive interference of leakage.

It is possible to explore both the adiabatic and interference conditions for low leakage in the simulations (Fig. 6.13). When performing a $T_{2Q} = 24$ ns unipolar (Half NZ) pulse, only the adiabatic condition can be used to achieve a low leakage. This condition is visible as the dark region in [Fig. 6.13(b)]. When simulating a $T_{2Q} = 48$ ns (Full) NZ pulse, a low-leakage fringe is visible [Fig. 6.13(d)] corresponding to the interference condition.

The position of the interference fringe should depend on the time between the two halves of the pulse. This can be explored by adding a buffer time Δt between the two halves of the pulse in simulation. For a $T_{2Q} = 40$ ns $+ \Delta t$ pulse, the fringe can be seen to move over the leakage landscape (Fig. 6.14). The period corresponds to the expected period of $\sim 1/800$ MHz $= 1.25$ ns.

6.11.7. LEAKAGE MODIFICATION FOR RANDOMIZED BENCHMARKING

Leakage out of the computational subspace is determined using the protocol introduced in [38], which constitutes a modification of the randomized benchmarking protocol (see also Section 3.4.3).

To determine the populations in the ground (g), first-excited (e), and second-excited

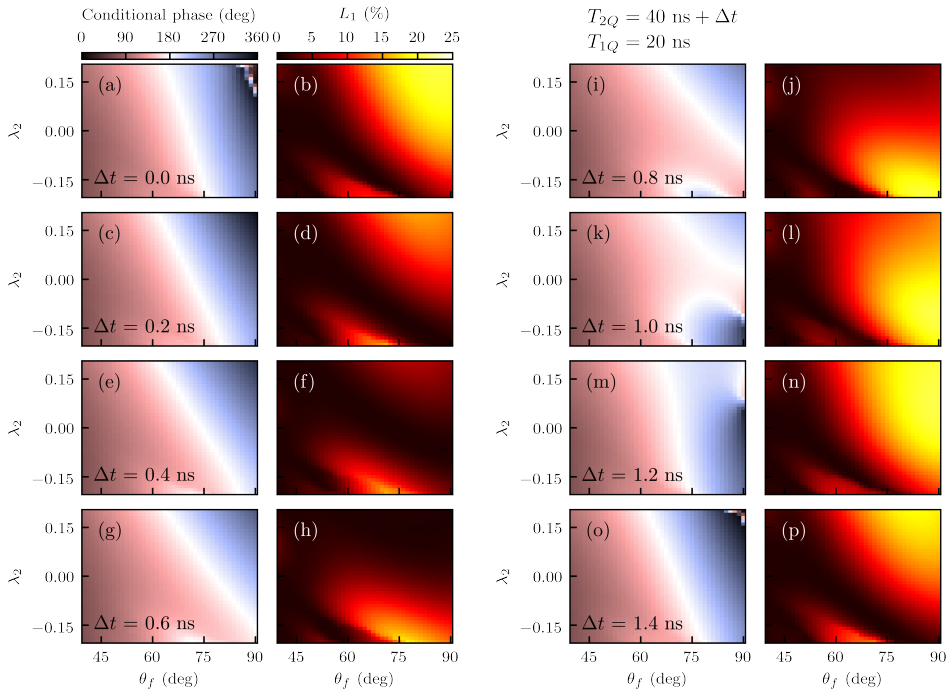

Figure 6.14: Moving interference fringes. To observe the effect of changing the length of the arms of the interferometer, a buffer (Δt) is added between the first and second part of the strong NZ pulse ($T_{2Q} = 40$ ns $+ \Delta t$, $T_{1Q} = 20$ ns) in simulation. The low-leakage fringe can clearly be seen to move over the landscape.

(f) states we follow the procedure described in [46]. In this procedure, a given experiment is performed in two different variants: once in the normal way, giving signal S_I, and once with a π pulse on the $g-e$ transition appended at the end of the sequence just before the measurement, giving signal S_X. When the respective reference signals V_0, V_1, and V_2 of a transmon qubit prepared in the g, e and f state are known, the respective populations of the g and e states, P_0 and P_1, can be extracted using

$$\begin{bmatrix} V_0 - V_2 & V_1 - V_2 \\ V_1 - V_2 & V_0 - V_2 \end{bmatrix} \begin{bmatrix} P_0 \\ P_1 \end{bmatrix} = \begin{bmatrix} S_\mathrm{I} - V_2 \\ S_\mathrm{X} - V_2 \end{bmatrix}, \tag{6.34}$$

under the assumption that higher-excited levels are unpopulated (in other words, $P_0 + P_1 + P_2 = 1$, where P_2 is the population in the f state).

Following [38], we fit the population $P_{\mathscr{X}_1}$ in the computational subspace $\mathscr{X}_1$ to a single exponential

$$P_{\mathscr{X}_1}(N_\mathrm{Cl.}) = A + B\lambda_1^{N_\mathrm{Cl.}}, \tag{6.35}$$

where $N_\mathrm{Cl.}$ is the number of Cliffords. The average leakage (L_1) and seepage (L_2) rates [Eqs. (6.25) and (6.26)] per Clifford can then be estimated as

$$L_1^\mathrm{Cl.} = (1-A)(1-\lambda_1), \tag{6.36}$$

$$L_2^\mathrm{Cl.} = A(1-\lambda_1). \tag{6.37}$$

Using the fitted value of λ_1, the survival probability M_0 is then fitted to a double exponential of the form

$$M_0(N_\mathrm{Cl.}) = A_0 + B_0\lambda_1^{N_\mathrm{Cl.}} + C_0\lambda_2^{N_\mathrm{Cl.}}. \tag{6.38}$$

The average gate infidelity per Clifford $\varepsilon^\mathrm{Cl.}$ is given by

$$\varepsilon^\mathrm{Cl.} = 1 - \frac{1}{d_1}\left[(d_1-1)\lambda_2 + 1 - L_1\right], \tag{6.39}$$

with $d_1 = \dim \mathscr{X}_1$. We note that if the leakage is weak ($\lambda_1 \ll \lambda_2$ and $B \ll A$), this reduces to the conventional randomized benchmarking formula. We refer to this experiment as the reference sequence.

This method is used in combination with interleaved randomized benchmarking [45] to extract the average gate infidelity (ε^CZ) and leakage (L_1^CZ) per CZ gate

$$\varepsilon^\mathrm{CZ} = 1 - \frac{1-\varepsilon^\mathrm{Int.}}{1-\varepsilon^\mathrm{Cl.}}, \tag{6.40}$$

$$L_1^\mathrm{CZ} = 1 - \frac{1-L_1^\mathrm{Int.}}{1-L_1^\mathrm{Cl.}}, \tag{6.41}$$

where $\varepsilon^\mathrm{Int.}$ ($L_1^\mathrm{Int.}$) stands for the average gate fidelity (leakage) in the interleaved sequence of the interleaved randomized benchmarking experiment.

6.12. PART 2: HIGH-FIDELITY CONTROLLED-Z GATE WITH MAXIMAL INTERMEDIATE LEAKAGE OPERATING AT THE SPEED LIMIT IN A SUPERCONDUCTING QUANTUM PROCESSOR

6.13. INTRODUCTION

Superconducting quantum processors have recently reached important milestones [48], notably the demonstration of quantum supremacy on a 53-transmon processor [49]. On the path to quantum error correction (QEC) and fault tolerance [20], recent experiments have used repetitive parity measurements to stabilize two-qubit entanglement [18, 19] and to perform surface-code quantum error detection in a 7-transmon processor [50]. These developments have relied on two-qubit controlled-phase (CPhase) gates realized by dynamical flux control of transmon frequency, harnessing the transverse coupling J_2 between a computational state $|11\rangle$ and a non-computational state such as $|02\rangle$ [24, 25]. Compared to other implementations, e.g., cross-resonance using microwave-frequency pulses [10] and parametric radio-frequency pulsing [9], baseband flux pulses achieve the fastest controlled-Z (CZ) gates (a special case of CPhase), operating near the speed limit $t_{\text{lim}} = \pi/J_2$ [51].

Over the last decade, baseband flux pulsing for two-qubit gating has evolved in an effort to increase gate fidelity and to reduce leakage and residual ZZ coupling. In particular, leakage became a main focus for its negative impact on QEC, adding complexity to error-decoder design and requiring hardware and operational overhead to seep [26–30]. To reduce leakage from linear-dynamical distortion in flux-control lines and limited time resolution in arbitrary waveform generators (AWGs), unipolar square pulses [25, 52] have been superseded by softened counterparts [7, 15] based on fast-adiabatic theory [31]. In parallel, coupling strengths have reduced to $J_2/2\pi \sim 10-20$ MHz to mitigate residual ZZ coupling, which affects single-qubit gates and idling at bias points, and produces crosstalk from spectator qubits [53]. Many groups are actively developing tunable coupling schemes to suppress residual coupling without incurring slowdown [54–58].

A main limitation to the fidelity of flux-based CPhase gates is dephasing from flux noise, as one qubit is displaced 0.5−1 GHz below its flux-symmetry point (i.e., sweetspot [59]) to reach the $|11\rangle$-$|02\rangle$ resonance. To address this limitation, in Section 6.3 introduced a bipolar variant [termed Net Zero (NZ)] of the fast-adiabatic scheme, which provides a built-in echo reducing the impact of low-frequency flux noise. The double use of the transverse interaction also reduces leakage by destructive interference, as understood by analogy to a Mach-Zehnder interferometer (MZI). Finally, the zero-average characteristic avoids the buildup of long-timescale distortions in the flux-control lines, significantly improving gate repeatability. NZ pulsing has been successfully used in several recent experiments [18, 50, 60], elevating the state of the art for CZ gate fidelity to $99.72 \pm 0.35\%$ [48]. However, NZ suffers from complicated tuneup, owing to the complex dependence of conditional phase and leakage on fast-adiabatic pulse parameters. This limits the use of NZ for two-qubit gating as processors grow in qubit count.

In this Letter, we introduce the sudden variant (SNZ) of the NZ scheme implementing CZ, which offers two advantages while preserving the built-in echo, destructive leakage interference, and repeatability characteristic of conventional NZ (CNZ). First, SNZ

operates at the speed limit of transverse coupling by maximizing intermediate leakage to the non-computational state. The second and main advantage is greatly simplified tuneup: the landscapes of conditional phase and leakage as a function of two pulse parameters have regular structure and interrelation, easily understood by exact analogy to the MZI. We realize SNZ CZ gates among four pairs of nearest neighbors in a seven-transmon processor and characterize their performance using two-qubit interleaved randomized benchmarking (2QIRB) with modifications to quantify leakage [38, 61]. The highest performance achieved from one 2QIRB characterization has $99.93 \pm 0.24\%$ fidelity and $0.10 \pm 0.02\%$ leakage. SNZ CZ gates are fully compatible with scalable approaches to QEC [32]. The generalization of SNZ to arbitrary CPhase gates is straightforward and useful for optimization [62], quantum simulation [63], and other noisy intermediate-scale quantum (NISQ) applications [64].

6.14. SUDDEN NET ZERO CONCEPT

A flux pulse to the $|11\rangle$-$|02\rangle$ interaction implements the unitary

$$U_{\mathrm{CPhase}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & e^{i\phi_{01}} & 0 & 0 & 0 \\ 0 & 0 & e^{i\phi_{10}} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{1-4L_1}e^{i\phi_{11}} & \sqrt{4L_1}e^{i\phi_{02,11}} \\ 0 & 0 & 0 & \sqrt{4L_1}e^{i\phi_{11,02}} & \sqrt{1-4L_1}e^{i\phi_{02}} \end{pmatrix}$$

in the $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle, |02\rangle\}$ subspace, neglecting decoherence and residual interaction between far off-resonant levels. Here, ϕ_{01} and ϕ_{10} are the single-qubit phases, $\phi_{11} = \phi_{01} + \phi_{10} + \phi_{2Q}$, where ϕ_{2Q} is the conditional phase, and L_1 is the leakage, The ideal CZ gate simultaneously achieves $\phi_{01} = \phi_{10} = 0 \pmod{2\pi}$, $\phi_{2Q} = \pi \pmod{2\pi}$ (phase condition PC), and $L_1 = 0$ (leakage condition LC), with arbitrary ϕ_{02}.

The SNZ CZ gate is realized with two square half pulses with equal and opposite amplitude $\pm A$ and duration $t_p/2$ each. To understand its action, consider first the ideal scenario with perfectly square half pulses (infinite bandwidth), infinite time resolution, $t_p = t_{\mathrm{lim}}$, and $A = 1$ (corresponding to $|11\rangle$ and $|02\rangle$ on resonance). The unitary action of each complete half pulse (rising edge, steady level, and falling edge combined) implements one of two beamsplitters in the MZI analogy: BS1 fully transmits $|11\rangle$ to $-i|02\rangle$ (producing maximal intermediate leakage), and BS2 fully transmits $-i|02\rangle$ to $-|11\rangle$, yielding an ideal CZ gate. SNZ adds an idling period t_ϕ between the half pulses to perfect the analogy to the MZI, allowing accrual of relative phase ϕ between $|02\rangle$ and $|11\rangle$ in between the beamsplitters.

6.15. EASINESS OF TUNE-UP: THEORY

The key advantage of SNZ over CNZ is the straightforward procedure to simultaneously meet PC and LC. To appreciate this, consider the landscapes of ϕ_{2Q} and L_1 as a function of A and t_ϕ [Fig. 6.15(c, d)] in this ideal scenario. The landscapes have a clear structure and link to each other. The L_1 landscape shows a vertical leakage valley at $A = 1$ arising from perfect transmission at each beamsplitter (LC1), and also two vertical valleys arising from perfect reflection (LC2). Leakage interference gives rise to additional diagonal valleys

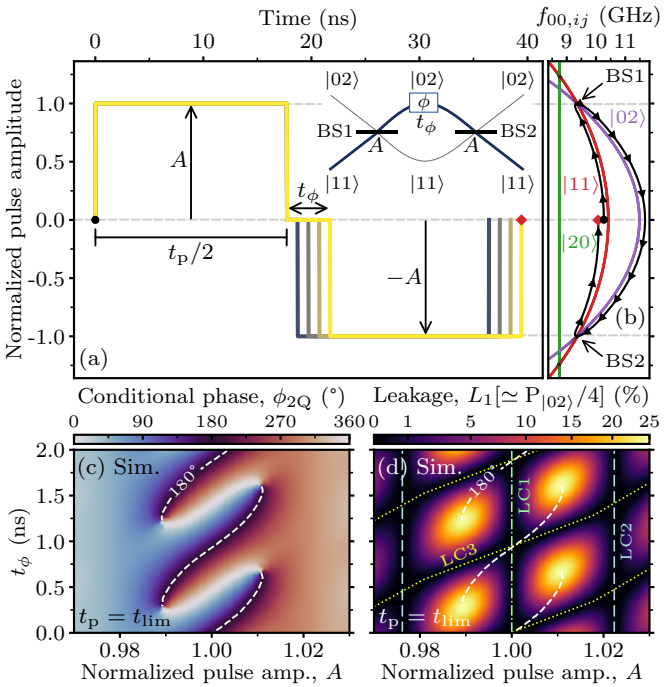

Figure 6.15: Numerical simulation of an ideal SNZ pulse (infinite bandwidth and time resolution) using parameters for pair Q_L-Q_{M2} (see Table 6.2). (a) Schematic of the ideal SNZ flux pulse, with $t_p = t_{lim}$ and variable A and t_ϕ. The amplitude A is normalized to the $|11\rangle$-$|02\rangle$ resonance. Inset: MZI analogy for $A = 1$. (b) Transition frequency from $|00\rangle$ to levels $|i\,j\rangle$ in the two-excitation manifold as a function of instantaneous pulse amplitude. (c, d) Landscapes of conditional phase ϕ_{2Q} (b) and leakage L_1 (c) as a function of A and t_ϕ.

(LC3). Crucially, juxtaposing the $\phi_{2Q} = 180°$ contour shows that PC is met periodically, at the crossing of LC1 and LC3 valleys, where $\Delta_{02}^{\max} t_\phi = 0 \pmod{2\pi}$ ($\Delta_{02}^{\max}$ is the detuning between $|02\rangle$ and $|11\rangle$ at the bias point). This regular leakage landscape therefore provides useful crosshairs for simultaneously achieving PC and LC. We note that $\phi_{2Q}(t_\phi)$ changes monotonically along the LC1 valley, allowing for CPhase gates with arbitrary ϕ_{2Q}. We leave this generalization for future work.

There are practical reasons to include t_ϕ in experiment: any flux-pulse distortion remaining from the first half pulse (e.g., due to finite pulse decay time) will break the symmetry between BS1 and BS2. Due to the time resolution t_s of the AWG used for flux control, ϕ can only increment in steps of $-\Delta_{02}^{\max} t_s$. Typically $\Delta_{02}^{\max}/2\pi = 0.5{-}1$ GHz and $t_s \sim 1$ ns, so the number of intermediate sampling points only provides coarse control. For fine control, we propose to use the amplitude $\pm B$ of the first and last sampling points during t_ϕ (see Section 6.21.1).

6.16. EASINESS OF TUNE-UP: EXPERIMENT

We now turn to the experimental realization of SNZ CZ gates between nearest-neighbor pairs among four transmons. High- and low-frequency transmons (Q_H and Q_L, respectively) connect to two mid-frequency transmons (Q_{M1} and Q_{M2}) using bus resonators dedicated to each pair [connectivity diagram shown in Fig. 6.18(a) inset]. Each transmon has a flux-control line for two-qubit gating, a microwave-drive line for single-qubit gating, and dedicated readout resonators [11, 18] (see Section 2.2.2 for details). Table 6.3 provides a summary of measured parameters for the four transmons. Each transmon is statically flux-biased at its sweetspot to counter residual offsets. Flux pulsing is performed using a Zurich Instruments HDAWG-8 ($t_s = 1/2.4$ ns). Following prior work [36], we compensate the bandwidth-limiting effect of attenuation in the flux-control coaxial line (skin effect) and cryogenic reflective and absorptive low-pass filters using real-time digital filters in the AWG. In this way, we produce on-chip flux waveforms with rise time t_{rise} on par with that of the AWG (0.5 ns).

We exemplify the tuneup of SNZ using pair Q_L-Q_{M2} (Fig. 6.16). We first identify t_{lim} for the $|11\rangle$-$|02\rangle$ interaction and amplitude A bringing the two levels on resonance. Both are extracted from the characteristic chevron pattern of $|2\rangle$-population $P_{|2\rangle}$ in Q_{M2} as a function of the amplitude and duration of a unipolar square flux pulse acting on $|11\rangle$ [Fig. 6.16(a)]. The symmetry axis corresponds to $A = 1$. The difference in consecutive pulse durations achieving $P_{|2\rangle}$ maxima along this axis gives an accurate estimate of t_{lim} unaffected by initial transients. We set $t_p \equiv 2n t_s$, where n is the number of sampling points achieving the first $P_{|2\rangle}$ maximum. Using the measured positive difference $t_p - t_{\text{lim}}$ and numerical simulation (data not shown), we estimate $t_{\text{rise}} \approx 0.5$ ns. Next, we use standard conditional-oscillation experiments (see also Section 6.11.4) to measure the landscapes of ϕ_{2Q} and leakage estimate $\tilde{L}_1$ for SNZ pulses over amplitude ranges $A \in [0.9, 1.1]$ and $B \in [0, A]$, keeping $t_\phi \gtrsim 3 t_{\text{rise}}$. As expected, the landscape of $\tilde{L}_1$ [Fig. 6.16(c)] reveals a vertical valley at $A = 1$ and a diagonal valley. Juxtaposing the $\phi_{2Q} = 180°$ contour from Fig. 6.16(b), we observe the matching of PC at the crossing of these valleys, in excellent agreement with a numerical two-qutrit simulation [Fig. 6.16(d)].

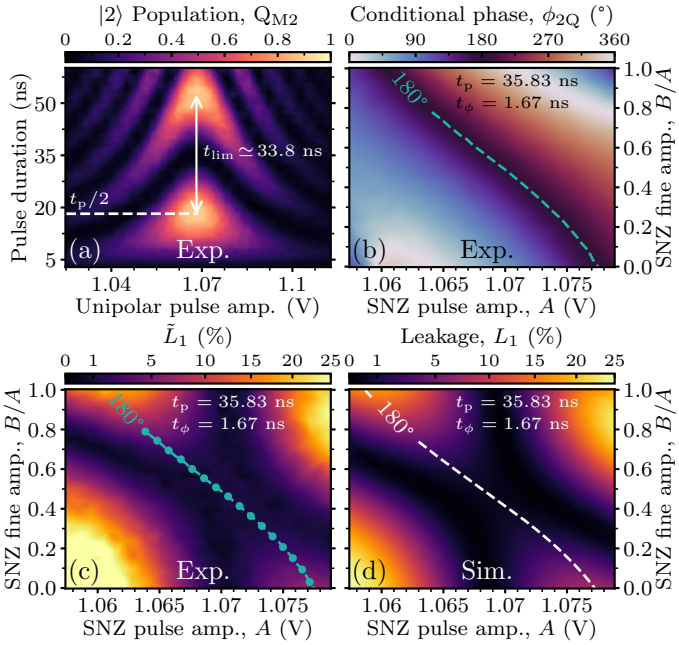

Figure 6.16: Calibration of the SNZ pulse for pair Q_L-Q_{M2} and comparison to simulation. (a) $|2\rangle$-state population of Q_{M2} as a function of the amplitude and duration of a unipolar square pulse making $|11\rangle$ interact with $|02\rangle$. (b,c) Landscapes of conditional phase ϕ_{2Q} and leakage estimate $\tilde{L}_1$ as a function of SNZ pulse amplitudes A and B, with $t_\phi = 1.67$ ns. The juxtaposed $\phi_{2Q} = 180°$ contour runs along the opposite diagonal compared to Fig. 6.15(b,c) because increasing B (which decreases Δ_{02}) changes ϕ in the opposite direction from t_ϕ. Data points marked with dots are measured with extra averaging for examination in Fig. 6.17. (d) Numerical simulation of leakage L_1 landscape and $\phi_{2Q} = 180°$ contour with parameters and flux-pulse distortions from experiment. All landscapes (also in Fig. 6.17) are sampled using an adaptive algorithm based on [65].

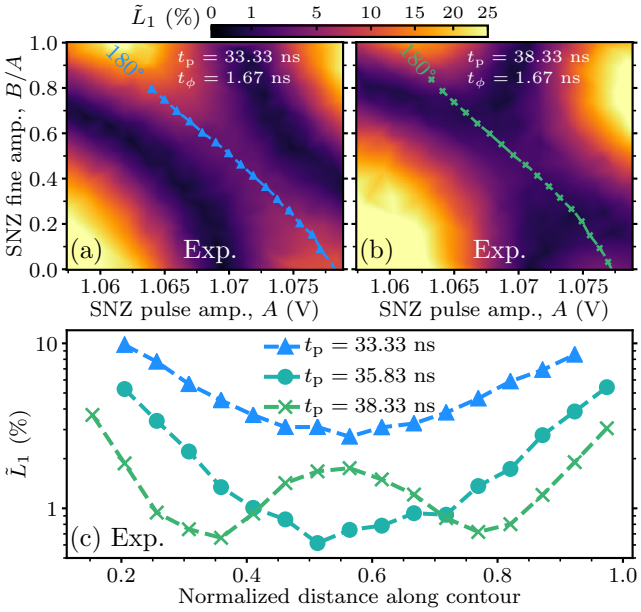

Figure 6.17: (a,b) Landscapes of the leakage estimate $\tilde{L}_1$ for intentionally short and long SNZ half pulses on Q_{M2}. (c) Extracted $\tilde{L}_1$ along the $\phi_{2Q} = 180°$ contours from (a), (b), and Fig. 6.16(c).

6

6.17. ROBUSTNESS TO PULSE DISCRETIZATION

Experimentally, due to the discreteness of t_s, it is unlikely to precisely match $t_p/2$ to the half-pulse duration that truly maximizes $P_{|2\rangle}$. To understand the consequences, we examine the ϕ_{2Q} and $\tilde{L}_1$ landscapes for SNZ pulses upon intentionally changing t_p by $\pm 6t_s$ (Fig. 6.17). While the PC contour remains roughly unchanged in both cases, there are significant effects on $\tilde{L}_1$. In both cases, we observe that $\tilde{L}_1$ lifts at the prior crossing of LC1 and LC3 valleys where $\phi_{2Q} = 180°$. For too-short pulses [Fig. 6.17(a)], there remain two valleys of minimal $\tilde{L}_1$, but these are now curved and do not cross $\phi_{2Q} = 180°$. For too-long pulses [Fig. 6.17(b)], there are also two curved valleys. Crucially, these cross the $\phi_{2Q} = 180°$ contour, and it remains possible to achieve PC and minimize leakage at two (A, B) settings. Extracting $\tilde{L}_1$ along the $\phi_{2Q} = 180°$ contours [Fig. 6.17(c)] confirms that too-long pulses can achieve the same minimal $\tilde{L}_1$ as when using the nominal t_p. The impossibility to achieve minimal leakage at $\phi_{2Q} = 180°$ for too-short pulses manifests the speed limit set by J_2. In turn, the demonstrated possibility to do so for too-long pulses (even overshooting by several sampling points) proves the viability of the SNZ pulse in practice.

6.18. PERFORMANCE

With these insights, we proceed to tune the remaining SNZ CZ gates following similar procedures. We use final weak bipolar pulses of total duration $t_{1Q} = 10$ ns to null the single-qubit phases in the frame of microwave drives. Since our codeword-based control

Parameter	Q_{M1}-Q_H	Q_{M2}-Q_H	Q_L-Q_{M1}	Q_L-Q_{M2}
t_{lim} (ns)	31.0	27.6	38.4	33.8
t_p, t_ϕ (ns)	32.50, 2.92	29.10, 3.75	40.83, 1.25	35.83, 1.67
t_{total} (ns)	45.42	42.91	52.08	47.50
Interaction	|11⟩-|02⟩	|11⟩-|02⟩	|11⟩-|20⟩	|11⟩-|02⟩
Parked qubit	Q_{M2}	Q_{M1}	–	–
Avg. F (%)	98.89 ± 0.35	99.54 ± 0.27	93.72 ± 2.10	97.14 ± 0.72
Avg. L_1 (%)	0.13 ± 0.02	0.18 ± 0.04	0.78 ± 0.32	0.63 ± 0.11
Max. F (%)	99.77 ± 0.23	99.93 ± 0.24	99.15 ± 1.20	98.56 ± 0.70
Min. L_1 (%)	0.07 ± 0.04	0.10 ± 0.02	0.04 ± 0.08	0.41 ± 0.10

Table 6.2: Summary of SNZ CZ pulse parameters and achieved performance for the four transmon pairs. Single-qubit phase corrections are included in t_{total}. Gate fidelities and leakage are obtained from 2QIRB keeping the other two qubits in |0⟩. Statistics (average and standard deviation) are taken from repeated 2QIRB runs (see [2] for technical details). The maximum F and minimum L_1 quoted are not necessarily from the same run.

electronics has a 20 ns timing grid, and 40 ns < $t_{total} = t_p + t_\phi + t_{1Q}$ < 60 ns for all pairs, we allocate 60 ns to every CZ gate. Some pair-specific details must be noted. Owing to the frequency overlap of Q_{M1} and Q_{M2}, implementing CZ between Q_H and Q_{M1} (Q_{M2}) requires a bipolar parking flux pulse on Q_{M2} (Q_{M1}) during the SNZ pulse on Q_H [32, 50]. For most pairs, we employ the |11⟩-|02⟩ interaction, which requires the smallest flux amplitude (reducing the impact of dephasing from flux noise) and does not require crossing any other interaction. However, for Q_L-Q_{M1}, we cannot reliably use this interaction as there is a flickering two-level system (TLS) overlapping with the |0⟩-|1⟩ transition in Q_{M1} at this amplitude [2]. For this pair, we therefore employ the |11⟩-|20⟩ interaction. Here, SNZ offers a side benefit: it crosses the Q_{M1}-TLS, |11⟩-|02⟩, and |01⟩-|10⟩ resonances as suddenly as possible, minimizing population exchange.

Table 6.2 summarizes the timing parameters and performance attained for the four SNZ CZ gates. The CZ gate fidelity F and leakage L_1 are extracted using a 2QIRB protocol [38]. For each pair, we report the average and standard deviation of both based on at least 10 repetitions of the protocol spanning more than 8 h [2]. Several observations can be drawn. First, CZ gates involving Q_H perform better on average than those involving Q_L. This is likely due to the shorter t_{lim} and correspondingly longer time 60 ns − t_p spent near the sweetspot. Additionally, the frequency downshifting required of Q_H to interact with the mid-frequency transmons is roughly half that required of the latter to interact with Q_L. This reduces the impact of dephasing from flux noise during the pulse. Not surprisingly, performance is worst for Q_L-Q_{M1}. Here, the pulse must downshift Q_{M1} the most to reach the distant |11⟩-|20⟩ interaction, increasing dephasing from flux noise. Also, there may be residual exchange at the crossed resonances. Overall, there is significant temporal variation in performance as gleaned by repeated 2QIRB characterizations. We believe this reflects the underlying variability of qubit relaxation and dephasing times and flux offsets, which however were not tracked simultaneously. In addition to having the best average performance, pair Q_{M2}-Q_H displays the maximum F of 99.93 ± 0.24% (Fig. 6.18) extracted from a single 2QIRB characterization. To the best of our knowledge, this is the highest CZ fidelity extracted from one 2QIRB characterization in a multi-transmon processor.

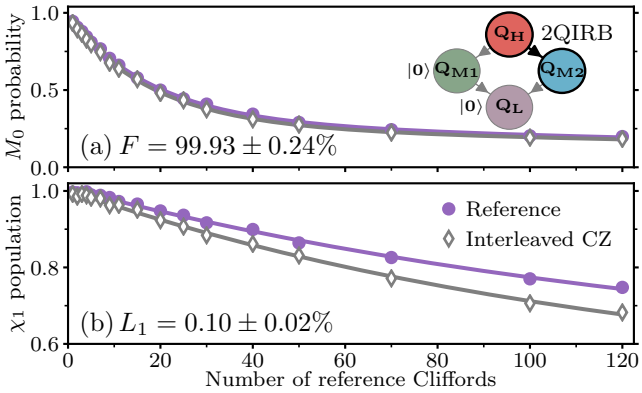

Figure 6.18: Best SNZ CZ gate performance achieved from a single run of 2QIRB. (a) Reference and CZ-interleaved return probability M_0 to $|00\rangle$ and (b) population in the computational space χ_1 as a function of the number of two-qubit Cliffords in the reference curve. Errors bars in F and L_1 are obtained from the uncertainty of exponential-decay fits.

6.19. LIMITING NOISE SOURCES

To understand the dominant sources of infidelity $\varepsilon = 1 - F$ and leakage, we run numerical simulations (similarly to Section 6.11.3), for both SNZ and CNZ, with experimental input parameters for pair Q_{M2}-Q_H. We dissect an error budget versus various models finding similar contributions for both gates (see Section 6.21.2). Nevertheless, the results suggest that SNZ slightly outperforms CNZ, likely due to a shorter time spent away from the sweetspot during the fixed 60 ns allocated for both variants. This confirms that the temporary full transfer from $|11\rangle$ to $|02\rangle$ does not compromise the gate fidelity.

6.20. CONCLUSION

In summary, we have proposed and realized high-fidelity CZ gates using the sudden version of the Net Zero bipolar fluxing scheme. SNZ CZ gates operate ever closer to the speed limit of transverse coupling by maximizing intermediate leakage to the non-computational state. Control architectures without a timing grid will benefit most from the speedup of SNZ over CNZ by reducing total gate time and thereby minimizing the impact of decoherence. A demonstrated second key advantage of SNZ over CNZ is ease of tuneup, owing to the simple structure of error landscapes as a function of pulse parameters. Harnessing the tuning simplicity, we already employ SNZ CZ gates in the Starmon-5 processor publicly available via the QuTech Quantum Inspire platform [66]. Moving forward, the compatibility of SNZ with our scalable scheme [32] for surface coding makes SNZ our choice for CZ gates for quantum error correction. Finally, the straightforward extension of SNZ to arbitrary conditional-phase gates will find immediate use in NISQ applications.

Data availability: Interested readers can reproduce our figures by using the processed data of the figures. The processed data can be found at https://github.com/DiCarloLab-Delft/High_Fidelity_ControlledZ_Data/.

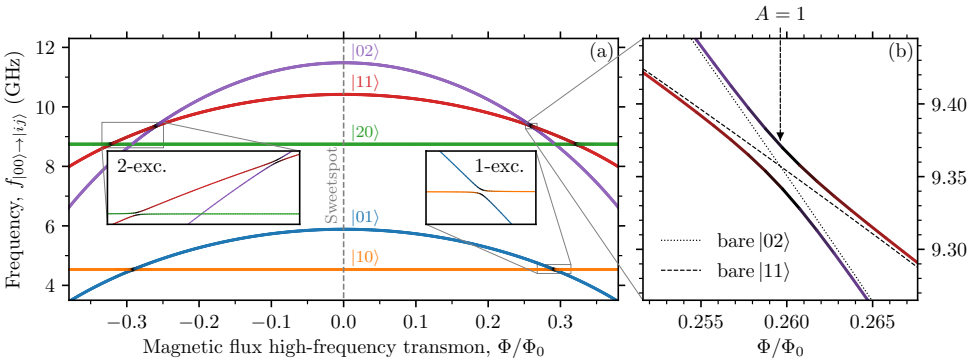

Figure 6.19: (a) Transition frequencies from $|00\rangle$ to levels $|ij\rangle$ in the one- and two-excitation manifold for transmon pair Q_L-Q_{M2} as a function of magnetic flux through the SQUID loop of Q_{M2}, normalized to the flux quantum Φ_0. Insets: Zoom-ins to the avoided crossings in the (left) two-excitation and (right) one-excitation manifolds. (b) Zoom-in to the $|11\rangle$-$|02\rangle$ avoided crossing occurring at $A = 1$. The minimum frequency splitting corresponds to $1/t_{\text{lim}} = J_2/\pi$.

6.21. METHODS

6.21.1. COMPARISON OF CONVENTIONAL NZ PULSES AND SNZ PULSES

This section highlights the main differences between CNZ pulses and the SNZ pulses introduced here. For reference, Fig. 6.19 illustrates the relevant energy-level structure for a pair of coupled transmons (here Q_L and Q_{M2}) as a function of magnetic flux on the higher-frequency transmon (here Q_{M2}). CNZ and SNZ CZ gates both exploit the avoided crossing in the two-excitation manifold between the computational state $|11\rangle$ and a non-computational state. Most often this non-computational state is $|02\rangle$ as reaching the avoided crossing requires the smallest flux-pulse amplitude and does not require passing through any other avoided crossings. In contrast, reaching the $|11\rangle$-$|20\rangle$ avoided crossing requires passing through the $|01\rangle$-$|10\rangle$ avoided crossing in the one-excitation manifold.

CNZ implements a CZ gate based on two back-to-back half strong flux pulses [Fig. 6.20(a)] of duration $t_p/2$ each, applied on the higher-frequency transmon. Typically, $t_p/t_{\text{lim}} \sim 1.1-1.6$. The strong half pulses are formally parametrized as in [31]. For the purposes of illustration, here we can loosely lump this parametrization as affecting the amplitude ($\pm A$) and curvature (A') of the strong half pulses. Immediately following the strong pulse, weak bipolar pulses of duration t_{1Q} are applied on both the higher- and lower-frequency transmons with amplitudes $\pm C$ and $\pm D$, respectively, in order to null the single-qubit phases acquired by each. Typically, $t_{1Q} = 10$ ns. In CNZ there is no intermediate idling period between the strong half pulses, so the analogy to the MZI is not exact [Fig. 6.20(c)]. During tuneup, one searches the (A, A') space to achieve a conditional phase (PC) of π by only affecting the unitary action of the two beamsplitters. Because for typical t_p CNZ produces significant leakage at the first strong pulse, achieving minimal leakage relies on meeting LC3 (leakage interference). The structure of the $\phi_{2Q}(A, A')$ and $L_1(A, A')$ landscapes and especially their interrelation are not straightforward, so the search for an (A, A') setting satisfying both PC and LC3 is not easily guided.

The SNZ pulses introduced here [Fig. 6.20(b)] differ in two key ways. First, the strong

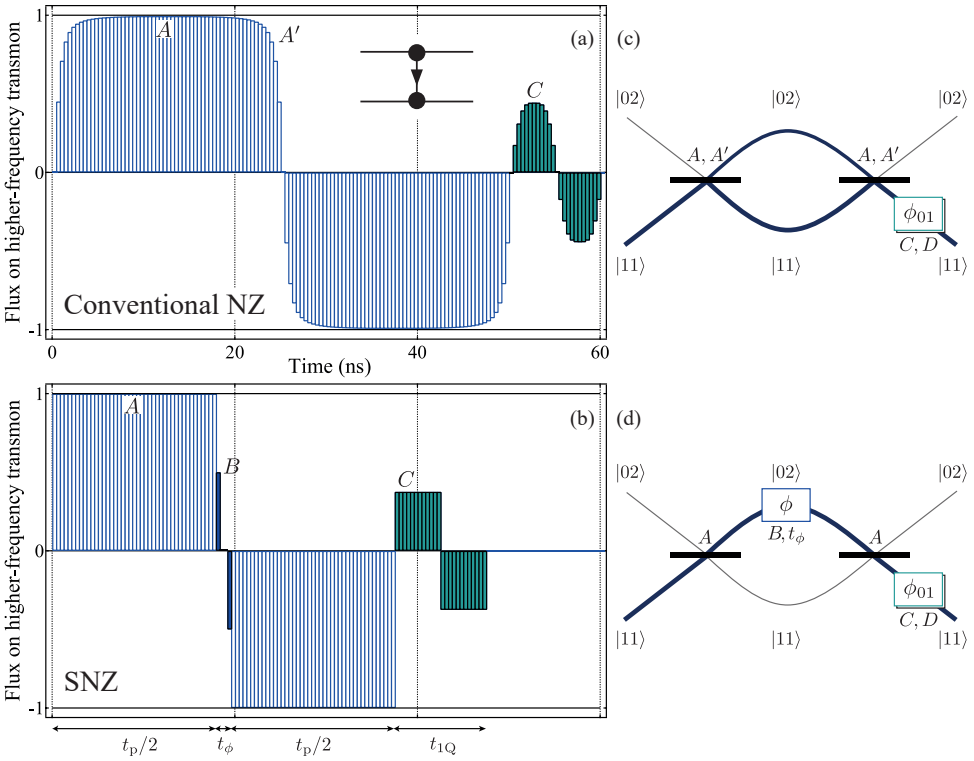

Figure 6.20: Comparison of conventional NZ and SNZ pulses for CZ gates. (a) Conventional NZ CZ pulses consist of two back-to-back strong half pulses of duration $t_p/2$ each, followed by two weak back-to-back half pulses of duration $t_{1Q}/2$ each on the higher-frequency qubit. The amplitude ($\pm A$) and curvature (A') of the strong pulses are jointly tuned to set the conditional phase ϕ_{2Q} at minimal leakage L_1, while the amplitude $\pm C$ of the weak pulses is used to null the single-qubit phase on the higher-frequency transmon. Weak pulses (amplitude $\pm D$) on the lower-frequency qubit (not shown here) are also used to null its single-qubit phase. (b) In SNZ, the strong pulses are replaced by square pulses with t_p as close as possible to t_{lim} but not shorter. Also, an intermediate idling period t_ϕ is added to accrue relative phase ϕ between $|02\rangle$ and $|11\rangle$. The amplitude $\pm B$ of the first and last sampling points in t_ϕ and the number of intermediate zero-amplitude points provide fine and coarse control of this relative phase, respectively. SNZ CZ gates also use weak bipolar pulses (now square) of total duration t_{1Q} to null single-qubit phases. (c) The MZI analogy for conventional NZ pulses is incomplete. Each strong half pulse implements a beamsplitter (ideally identical) with scattering parameters affected by A and A'. However, there is no possibility to independently control the relative phase in the two arms between the beamsplitters. (d) The MZI analogy is exact for SNZ pulse. The scattering at the beamsplitters is controlled by A and the relative phase ϕ is controlled finely using B and coarsely using t_ϕ.

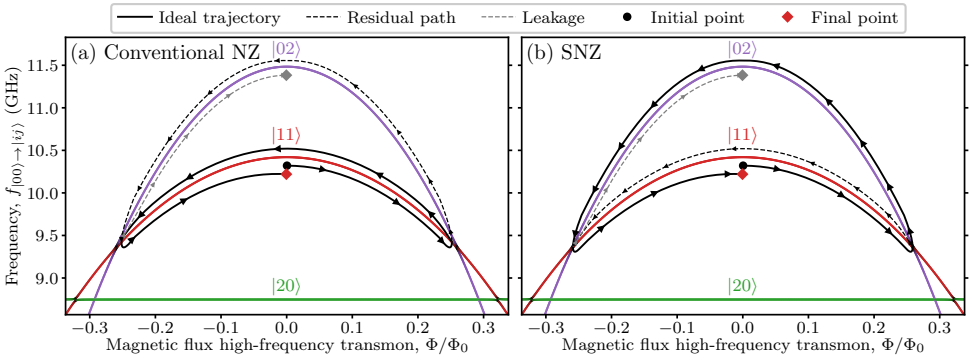

Figure 6.21: Schematic comparison of the trajectory of level populations in the two-excitation manifold for CNZ and SNZ strong pulses acting on |11⟩. Note that in both cases, most of the time is spent close or at the |11⟩-|02⟩ avoided crossing. (a) Trajectory for a CNZ pulse. (b) Trajectory for an SNZ pulse.

half pulses are replaced by square half pulses each with duration $t_p/2$ maximizing the transfer from |11⟩ to |02⟩ and vice versa. Second, an intermediate idling period t_ϕ is added to accrue relative phase ϕ between |02⟩ and |11⟩, perfecting the analogy to the MZI [Fig. 6.20(d)]. We use the amplitude $\pm B$ of the first and last sampling points in t_ϕ and the number of intermediate zero-amplitude points to achieve fine and coarse control of ϕ, respectively. As in CNZ, we use weak bipolar pulses on both transmons (also with $t_{1Q} = 10$ ns) to null the single-qubit phases. During tuneup, we search the (A, B) space to achieve $\phi_{2Q} = 180°$. As shown in the main text, the SNZ pulse design gives a very simple structure to the $\phi_{2Q}(A, B)$ and $L_1(A, B)$ landscapes. Crucially, the crossing point of LC1 and LC3 leakage valleys matches $\phi_{2Q} = 180°$. This simplicity of tuneup is the key advantage of SNZ over CNZ.

Another advantage of SNZ over CNZ is the reduced total time $t_{total} = t_p + t_\phi + t_{1Q}$ required to achieve a CZ gate. However, due to the 20 ns timing grid of our control electronics and the transverse coupling strengths in our device, this speedup is insufficient to reduce the total time allocated per CZ gate from 60 to 40 ns. Nonetheless, in SNZ, the fluxed transmon spends more time at its sweetspot, which reduces the dephasing due to flux noise.

Figure S3 illustrates the qualitative difference in the trajectory of level populations in the two-excitation manifold implemented by strong CNZ and SNZ pulses acting on the |11⟩ state. A CNZ pulse [Fig. 6.21(a)] uses the interaction point fast-adiabatically, keeping most population in |11⟩ after the first strong half pulse. In contrast, a SNZ pulse [Fig. 6.21(b)] uses the interaction suddenly to transfer most (ideally all) of the population to |02⟩ with the first strong half pulse.

6.21.2. SIMULATION RESULTS FOR SNZ AND CONVENTIONAL NZ CZ GATES VERSUS DIFFERENT ERROR MODELS

To identify the dominant sources of infidelity $\varepsilon = 1 - F$ and leakage for SNZ CZ gates, we perform a two-qutrit numerical simulation for pair Q_{M2}-Q_H with incremental addition of measured error sources [Fig. 6.22], as in Section 6.9. The simulation incrementally adds:

	Q_H	Q_{M1}	Q_{M2}	Q_L
Qubit frequency at sweetspot, $\omega_q/2\pi$ (GHz)	6.4329	5.7707	5.8864	4.5338
Transmon anharmonicity, $\alpha/2\pi$ (MHz)	-280	-290	-285	-320
Readout frequency, $\omega_r/2\pi$ (GHz)	7.4925	7.2248	7.0584	6.9132
Relaxation time, T_1 (μs)	37 ± 1	40 ± 1	47 ± 1	66 ± 1
Ramsey dephasing time, T_2^* (μs)	38 ± 1	49 ± 1	47 ± 1	64 ± 1
Echo dephasing time, T_2 (μs)	54 ± 2	68 ± 1	77 ± 1	94 ± 2
Residual qubit excitation, (%)	1.4	1.2	4.3	1.7
Best readout fidelity, F_{RO} (%)	99.1	98.5	99.4	97.8

Table 6.3: Summary of frequency, coherence, residual excitation, and readout parameters of the four transmons. The statistics of coherence times for each transmon are obtained from 30 repetitions of standard time-domain measurements [67] taken over ~4 h. The residual excitation is extracted from double-Gaussian fits of single-shot readout histograms with the qubit nominally prepared in $|0\rangle$. The readout fidelity quoted is the average assignment fidelity [68], extracted from single-shot readout histograms after mitigating residual excitation by post-selection on a pre-measurement.

(A) no noise; (B) relaxation; (C) Markovian dephasing; (D) dephasing from quasi-static flux noise; and (E) flux-pulse distortion. The experimental inputs for models B, C and D combine measured qubit relaxation time T_1 at the bias point, and measured echo and Ramsey dephasing times (T_2 and T_2^*) as a function of qubit frequency. The input to E consists of a final Cryoscope measurement of the flux step response using all real-time filters. The simulation suggests that the main source of ε is Markovian dephasing (as in Section 6.9), while the dominant contribution to L_1 is low-frequency flux noise. The latter contrasts with Section 6.9, where simulation identified flux-pulse distortion as the dominant leakage source. We identify two possible reasons for this difference: in the current experiment, the $1/f$ low-frequency flux noise is ~4 times larger (in units of $\Phi_0/\sqrt{\text{Hz}}$) and the achieved flux step response is noticeably sharper. Finally, we use the simulation to compare performance of SNZ to conventional NZ CZ. For the latter, we fix $t_\phi = 0$, $t_{1Q} = 60\,\text{ns} - t_p$, and use the fast-adiabatic pulse shape and $t_p = 45.83$ ns optimized by simulation. Overall, the error sources contribute very similarly to the error budget for both cases. The marginally higher overall performance found for SNZ is likely due to the increased time spent at the sweetspot during the gate time.

We emphasize that our two-qutrit simulation includes 9 energy levels (from $|00\rangle$ to $|22\rangle$). Therefore, it also captures leakage to $|20\rangle$ ($|02\rangle$) when using the $|11\rangle$-$|02\rangle$ ($|11\rangle$-$|20\rangle$) avoided crossing. For pair Q_{M2}-Q_H, for which we use $|11\rangle$-$|02\rangle$, the simulation gives a final $|20\rangle$-population of 0.005%, merely 4% of the total leakage L_1.

Finally, we use this numerical simulation with full error model E to illustrate that the SNZ pulse preserves the resilience of the conventional NZ scheme to low-frequency (quasi-static) flux offsets. Figure S10 shows that the single-qubit phase of the fluxed higher-frequency qubit (Q_H) and leakage L_1 are second-order sensitive to the offset. The conditional phase ϕ_{2Q} shows a very weak first-order dependence at zero offset. Numerical simulations with model D (not shown) show that the negative shift of the local maximum

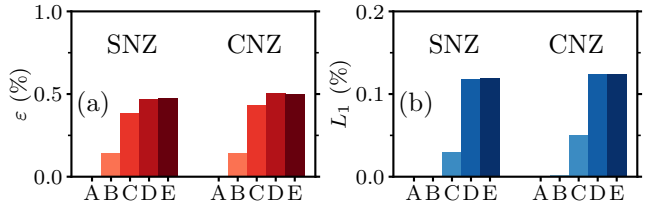

Figure 6.22: Error budgets for infidelity ε (a) and leakage L_1 (b) obtained by a numerical simulation (similarly to Section 6.9) of the Q_{M2}-Q_H SNZ CZ gate with parameters in Fig. 6.18 and for a conventional NZ gate with optimized parameters (see text for details). The simulation incrementally adds errors using experimental input parameters for this pair: (A) no noise; (B) relaxation; (C) Markovian dephasing; (D) dephasing from quasi-static flux noise; and (E) flux-pulse distortion.

in ϕ_{2Q} originates from the finite flux-pulse rise time.

REFERENCES

[1] M. A. Rol, F. Battistel, F. K. Malinowski, C. C. Bultink, B. M. Tarasinski, R. Vollmer, N. Haider, N. Muthusubramanian, A. Bruno, B. M. Terhal, and L. DiCarlo, *Fast, high-fidelity conditional-phase gate exploiting leakage interference in weakly anharmonic superconducting qubits,* Phys. Rev. Lett. **123**, 120502 (2019).

[2] V. Negîrneac, H. Ali, N. Muthusubramanian, F. Battistel, R. Sagastizabal, M. S. Moreira, J. F. Marques, W. J. Vlothuizen, M. Beekman, C. Zachariadis, N. Haider, A. Bruno, and L. DiCarlo, *High-fidelity controlled-Z gate with maximal intermediate leakage operating at the speed limit in a superconducting quantum processor,* Phys. Rev. Lett. **126**, 220502 (2021).

[3] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. Schuyler Fried, S. Hong, P. Karalekas, C. B. Osborn, A. Papageorge, E. C. Peterson, G. Prawiroatmodjo, N. Rubin, C. A. Ryan, D. Scarabelli, M. Scheer, E. A. Sete, P. Sivarajah, R. S. Smith, A. Staley, N. Tezak, W. J. Zeng, A. Hudson, B. R. Johnson, M. Reagor, M. P. da Silva, and C. Rigetti, *Unsupervised Machine Learning on a Hybrid Quantum Computer,* arXiv:1712.05771 (2017).

[4] W. Knight, *IBM raises the bar with a 50-qubit quantum computer,* https://www.technologyreview.com/s/609451/ibm-raises-the-bar-with-a-50-qubit-quantum-computer/.

[5] J. Kelly, *A preview of Bristlecone, Google's new quantum processor,* https://ai.googleblog.com/2018/03/a-preview-of-bristlecone-googles-new.html.

[6] Intel, *The future of quantum computing is counted in qubits,* https://newsroom.intel.com/news/future-quantum-computing-counted-qubits/.

[7] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and J. M.

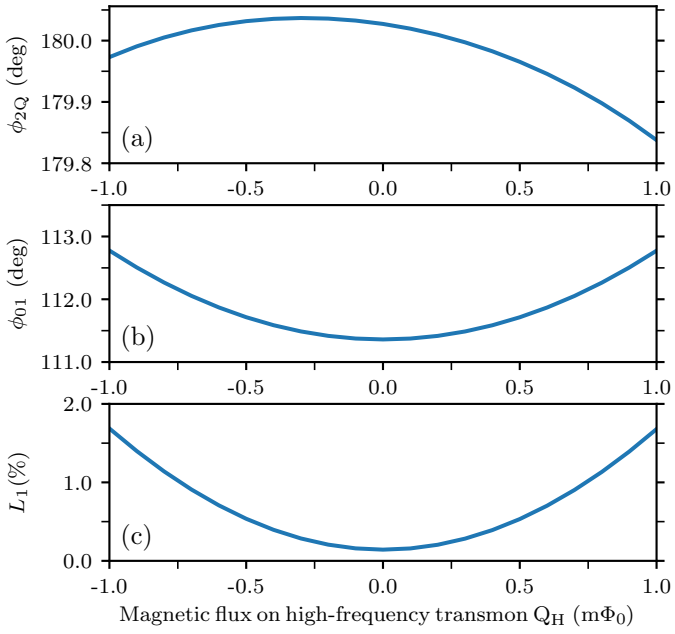

Figure 6.23: Numerical simulation (using model E and pair Q_{M2}-Q_H) of the dependence of conditional phase ϕ_{2Q}, single-qubit phase ϕ_{01}, and leakage L_1 on quasi-static flux noise.

Martinis, *Superconducting quantum circuits at the surface code threshold for fault tolerance.* Nature **508**, 500 (2014).

[8] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo, *Restless tuneup of high-fidelity qubit gates,* Phys. Rev. Applied **7**, 041001 (2017).

[9] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M. P. da Silva, and B. R. Johnson, *Demonstration of a parametrically activated entangling gate protected from flux noise,* Phys. Rev. A **101**, 012302 (2020).

[10] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, *Procedure for systematically tuning up cross-talk in the cross-resonance gate,* Phys. Rev. A **93**, 060302(R) (2016).

[11] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, and C. Eichler, *Rapid high-fidelity multiplexed readout of superconducting qubits,* Phys. Rev. App. **10**, 034040 (2018).

[12] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, *Characterizing quantum supremacy in near-term devices,* Nat. Phys. **14**, 595 (2018).

[13] C. Neill, P. Roushan, K. Kechedzhi, S. Boixo, S. V. Isakov, V. Smelyanskiy, A. Megrant, B. Chiaro, A. Dunsworth, K. Arya, R. Barends, B. Burkett, Y. Chen, Z. Chen, A. Fowler, B. Foxen, M. Giustina, R. Graff, E. Jeffrey, T. Huang, J. Kelly, P. Klimov, E. Lucero, J. Mutus, M. Neeley, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, H. Neven, and J. M. Martinis, *A blueprint for demonstrating quantum supremacy with superconducting qubits,* Science **360**, 195 (2018).

[14] S. Bravyi, D. Gosset, and R. König, *Quantum advantage with shallow circuits,* Science **362**, 308 (2018).

[15] J. Kelly, R. Barends, A. G. Fowler, A. Megrant, E. Jeffrey, T. White, D. Sank, J. Mutus, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, I.-C. Hoi, C. Neill, P. J. J. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, A. N. Cleland, J. Wenner, and J. M. Martinis, *State preservation by repetitive error detection in a superconducting quantum circuit,* Nature **519**, 66 (2015).

[16] D. Ristè, S. Poletto, M. Z. Huang, A. Bruno, V. Vesterinen, O. P. Saira, and L. DiCarlo, *Detecting bit-flip errors in a logical qubit using stabilizer measurements,* Nat. Commun. **6**, 6983 (2015).

[17] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, J. M. Chow, and J. M. Gambetta, *Demonstration of weight-four parity measurements in the surface code architecture,* Phys. Rev. Lett. **117**, 210505 (2016).

[18] C. C. Bultink, T. E. O'Brien, R. Vollmer, N. Muthusubramanian, M. W. Beekman, M. A. Rol, X. Fu, B. Tarasinski, V. Ostroukh, B. Varbanov, A. Bruno, and L. DiCarlo, *Protecting quantum entanglement from leakage and qubit errors via repetitive parity measurements,* Science Advances **6** (2020), 10.1126/sciadv.aay3050.

[19] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, J. Heinsoo, J.-C. Besse, M. Gabureac, A. Wallraff, and C. Eichler, *Entanglement stabilization using ancilla-based parity detection and real-time feedback in superconducting circuits,* npj Quantum Information **5**, 1 (2019).

[20] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation,* Phys. Rev. A **86**, 032324 (2012).

[21] J. M. Martinis, *Qubit metrology for building a fault-tolerant quantum computer,* npj Quantum Inf. **1**, 15005 (2015).

[22] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, *Simple all-microwave entangling gate for fixed-frequency superconducting qubits,* Phys. Rev. Lett. **107**, 080502 (2011).

[23] S. A. Caldwell, N. Didier, C. A. Ryan, E. A. Sete, A. Hudson, P. Karalekas, R. Manenti, M. P. da Silva, R. Sinclair, E. Acala, N. Alidoust, J. Angeles, A. Bestwick, M. Block, B. Bloom, A. Bradley, C. Bui, L. Capelluto, R. Chilcott, J. Cordova, G. Crossman, M. Curtis, S. Deshpande, T. E. Bouayadi, D. Girshovich, S. Hong, K. Kuang, M. Lenihan, T. Manning, A. Marchenkov, J. Marshall, R. Maydra, Y. Mohan, W. O'Brien, C. Osborn, J. Otterbach, A. Papageorge, J.-P. Paquette, M. Pelstring, A. Polloreno, G. Prawiroatmodjo, V. Rawat, M. Reagor, R. Renzas, N. Rubin, D. Russell, M. Rust, D. Scarabelli, M. Scheer, M. Selvanayagam, R. Smith, A. Staley, M. Suska, N. Tezak, D. C. Thompson, T.-W. To, M. Vahidpour, N. Vodrahalli, T. Whyland, K. Yadav, W. Zeng, and C. Rigetti, *Parametrically activated entangling gates using transmon qubits,* Phys. Rev. Applied **10**, 034050 (2018).

[24] F. W. Strauch, P. R. Johnson, A. J. Dragt, C. J. Lobb, J. R. Anderson, and F. C. Wellstood, *Quantum logic gates for coupled superconducting phase qubits,* Phys. Rev. Lett. **91**, 167005 (2003).

[25] L. DiCarlo, J. M. Chow, J. M. Gambetta, L. S. Bishop, B. R. Johnson, D. I. Schuster, J. Majer, A. Blais, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, *Demonstration of two-qubit algorithms with a superconducting quantum processor,* Nature **460**, 240 (2009).

[26] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults,* Quantum Info. Comput. **7**, 139 (2007).

[27] J. Ghosh, A. G. Fowler, J. M. Martinis, and M. R. Geller, *Understanding the effects of leakage in superconducting quantum-error-detection circuits,* Phys. Rev. A **88**, 062329 (2013).

[28] A. G. Fowler, *Coping with qubit leakage in topological codes,* Phys. Rev. A **88**, 042308 (2013).

[29] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code,* Quantum Info. Comput. **15**, 997 (2015).

[30] J. Ghosh and A. G. Fowler, *Leakage-resilient approach to fault-tolerant quantum computing with superconducting elements,* Phys. Rev. A **91**, 020302(R) (2015).

[31] J. M. Martinis and M. R. Geller, *Fast adiabatic qubit gates using only σ_z control,* Phys. Rev. A **90**, 022307 (2014).

[32] R. Versluis, S. Poletto, N. Khammassi, B. M. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, *Scalable quantum circuit and control for a superconducting surface code,* Phys. Rev. App. **8**, 034021 (2017).

[33] Y. Wang, Y. Li, Z. Yin, and B. Zeng, *16-qubit IBM universal quantum computer can be fully entangled,* npj Quantum Inf. **4**, 46 (2018).

[34] X. Fu, M. A. Rol, C. C. Bultink, J. van Someren, N. Khammassi, I. Ashraf, R. F. L. Vermeulen, J. C. de Sterke, W. J. Vlothuizen, R. N. Schouten, C. G. Almudever, L. DiCarlo, and K. Bertels, *An experimental microarchitecture for a superconducting quantum processor,* in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-50 '17 (ACM, New York, NY, USA, 2017) pp. 813–825.

[35] X. Fu, L. Riesebos, M. A. Rol, J. van Straten, J. van Someren, N. Khammassi, I. Ashraf, R. F. L. Vermeulen, V. Newsum, K. K. L. Loh, J. C. de Sterke, W. J. Vlothuizen, R. N. Schouten, C. G. Almudever, L. DiCarlo, and K. Bertels, *eQASM: An executable quantum instruction set architecture,* in *Proceedings of 25th IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, 2019) pp. 224–237.

[36] M. A. Rol, L. Ciorciaro, F. K. Malinowski, B. M. Tarasinski, R. E. Sagastizabal, C. C. Bultink, Y. Salathe, N. Haandbaek, J. Sedivy, and L. DiCarlo, *Time-domain characterization and correction of on-chip distortion of control pulses in a quantum processor,* App. Phys. Lett. **116**, 054001 (2020).

[37] B. R. Johnson, *Controlling Photons in Superconducting Electrical Circuits*, PhD Dissertation, Yale University (2011).

[38] C. J. Wood and J. M. Gambetta, *Quantification and characterization of leakage errors,* Phys. Rev. A **97**, 032306 (2018).

[39] W. D. Oliver, Y. Yu, J. C. Lee, K. K. Berggren, L. S. Levitov, and T. P. Orlando, *Mach-zehnder interferometry in a strongly driven superconducting qubit,* Science **310**, 1653 (2005).

[40] M. Sillanpää, T. Lehtinen, A. Paila, Y. Makhlin, and P. Hakonen, *Continuous-time monitoring of Landau-Zener interference in a Cooper-Pair Box,* Phys. Rev. Lett. **96**, 187002 (2006).

[41] J. R. Petta, H. Lu, and A. C. Gossard, *A coherent beam splitter for electronic spin states,* Science **327**, 669 (2010).

[42] A. Chatterjee, S. N. Shevchenko, S. Barraud, R. M. Otxoa, F. Nori, J. J. L. Morton, and M. F. Gonzalez-Zalba, *A silicon-based single-electron interferometer coupled to a fermionic sea,* Phys. Rev. B **97**, 045405 (2018).

[43] J. Zhou, P. Huang, Q. Zhang, Z. Wang, T. Tan, X. Xu, F. Shi, X. Rong, S. Ashhab, and J. Du, *Observation of time-domain Rabi oscillations in the Landau-Zener regime with a single electronic spin,* Phys. Rev. Lett. **112**, 010503 (2014).

[44] S. Shevchenko, S. Ashhab, and F. Nori, *Landau–Zener–Stückelberg interferometry,* Physics Reports **492**, 1 (2010).

[45] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, *Efficient measurement of quantum gate error by interleaved randomized benchmarking,* Phys. Rev. Lett. **109**, 080505 (2012).

[46] S. Asaad, C. Dickel, S. Poletto, A. Bruno, N. K. Langford, M. A. Rol, D. Deurloo, and L. DiCarlo, *Independent, extensible control of same-frequency superconducting qubits by selective broadcasting,* npj Quantum Inf. **2**, 16029 (2016).

[47] J. Johansson, P. Nation, and F. Nori, *QuTiP 2: A Python framework for the dynamics of open quantum systems,* Computer Physics Communications **184**, 1234 (2013).

[48] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, *Superconducting qubits: Current state of play,* Annual Review of Condensed Matter Physics **11**, 369 (2020).

[49] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, *Quantum supremacy using a programmable superconducting processor,* Nature **574**, 505 (2019).

[50] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, N. Lacroix, G. J. Norris, M. Gabureac, C. Eichler, and A. Wallraff, *Repeated quantum error detection in a surface code,* Nat. Phys. **16**, 875 (2020).

[51] R. Barends, C. M. Quintana, A. G. Petukhov, Y. Chen, D. Kafri, K. Kechedzhi, R. Collins, O. Naaman, S. Boixo, F. Arute, K. Arya, D. Buell, B. Burkett, Z. Chen, B. Chiaro,

A. Dunsworth, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, T. Huang, E. Jeffrey, J. Kelly, P. V. Klimov, F. Kostritsa, D. Landhuis, E. Lucero, M. McEwen, A. Megrant, X. Mi, J. Mutus, M. Neeley, C. Neill, E. Ostby, P. Roushan, D. Sank, K. J. Satzinger, A. Vainsencher, T. White, J. Yao, P. Yeh, A. Zalcman, H. Neven, V. N. Smelyanskiy, and J. M. Martinis, *Diabatic gates for frequency-tunable superconducting qubits,* Phys. Rev. Lett. **123**, 210501 (2019).

[52] L. DiCarlo, M. D. Reed, L. Sun, B. R. Johnson, J. M. Chow, J. M. Gambetta, L. Frunzio, S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf, *Preparation and measurement of three-qubit entanglement in a superconducting circuit,* Nature **467**, 574 (2010).

[53] S. Krinner, S. Lazar, A. Remm, C. Andersen, N. Lacroix, G. Norris, C. Hellings, M. Gabureac, C. Eichler, and A. Wallraff, *Benchmarking coherent errors in controlled-phase gates due to spectator qubits,* Phys. Rev. App. **14**, 024042 (2020).

[54] Y. Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, B. Campbell, Z. Chen, B. Chiaro, A. Dunsworth, E. Jeffrey, A. Megrant, J. Y. Mutus, P. J. J. O'Malley, C. M. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, M. R. Geller, A. N. Cleland, and J. M. Martinis, *Qubit architecture with high coherence and fast tunable coupling,* Phys. Rev. Lett. **113**, 220502 (2014).

[55] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Tunable coupling scheme for implementing high-fidelity two-qubit gates,* Phys. Rev. App. **10**, 054062 (2018).

[56] P. Mundada, G. Zhang, T. Hazard, and A. Houck, *Suppression of qubit crosstalk in a tunable coupling superconducting circuit,* Phys. Rev. App. **12**, 054023 (2019).

[57] M. C. Collodo, J. Herrmann, N. Lacroix, C. K. Andersen, A. Remm, S. Lazar, J.-C. Besse, T. Walter, A. Wallraff, and C. Eichler, *Implementation of conditional phase gates based on tunable ZZ interactions,* Phys. Rev. Lett. **125**, 240502 (2020).

[58] Y. Xu, J. Chu, J. Yuan, J. Qiu, Y. Zhou, L. Zhang, X. Tan, Y. Yu, S. Liu, J. Li, F. Yan, and D. Yu, *High-fidelity, high-scalability two-qubit gate scheme for superconducting qubits,* Phys. Rev. Lett. **125**, 240503 (2020).

[59] J. A. Schreier, A. A. Houck, J. Koch, D. I. Schuster, B. R. Johnson, J. M. Chow, J. M. Gambetta, J. Majer, L. Frunzio, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Suppressing charge noise decoherence in superconducting charge qubits,* Phys. Rev. B **77**, 180502(R) (2008).

[60] M. Kjaergaard, M. E. Schwartz, A. Greene, G. O. Samach, M. O. A. Bengtsson, C. M. McNally, J. Braumüller, D. K. Kim, P. Krantz, M. Marvian, A. Melville, B. M. Niedzielski, Y. Sung, R. Winik, J. Yoder, D. Rosenberg, S. L. K. Obenland, . T. P. Orlando, I. Marvian, S. Gustavsson, and W. D. Oliver, *A quantum instruction set implemented on a superconducting quantum processor,* ArXiv:2001.08838 (2020).

[61] E. Magesan, J. M. Gambetta, and J. Emerson, *Characterizing quantum gates via randomized benchmarking,* Phys. Rev. A **85**, 042311 (2012).

[62] N. Lacroix, C. Hellings, C. K. Andersen, A. Di Paolo, A. Remm, S. Lazar, S. Krinner, G. J. Norris, M. Gabureac, J. Heinsoo, A. Blais, C. Eichler, and A. Wallraff, *Improving the performance of deep quantum optimization algorithms with continuous gate sets,* PRX Quantum **1**, 110304 (2020).

[63] R. Barends, L. Lamata, J. Kelly, L. García-Álvarez, A. Fowler, A. Megrant, E. Jeffrey, T. White, D. Sank, J. Mutus, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, I.-C. Hoi, C. Neill, P. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, J. Wenner, E. Solano, and J. Martinis, *Digital quantum simulation of fermionic models with a superconducting circuit,* Nat. Commun. **6**, 7654 (2015).

[64] J. Preskill, *Quantum Computing in the NISQ era and beyond,* Quantum **2**, 79 (2018).

[65] B. Nijholt, J. Weston, J. Hoofwijk, and A. Akhmerov, *Adaptive: parallel active learning of mathematical functions,* (2019).

[66] *QuTech Quantum Inspire,* `https://www.quantum-inspire.com/`.

[67] P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *A quantum engineer's guide to superconducting qubits,* App. Phys. Rev. **6**, 021318 (2019).

[68] C. C. Bultink, B. Tarasinski, N. Haandbaek, S. Poletto, N. Haider, D. J. Michalak, A. Bruno, and L. DiCarlo, *General method for extracting the quantum efficiency of dispersive qubit readout in circuit QED,* App. Phys. Lett. **112**, 092601 (2018).

7

SPECTRAL QUANTUM TOMOGRAPHY

We introduce spectral quantum tomography, a simple method to extract the eigenvalues of a noisy few-qubit gate, represented by a trace-preserving superoperator, in a SPAM-resistant fashion, using low resources in terms of gate sequence length. The eigenvalues provide detailed gate information, supplementary to known gate-quality measures such as the gate fidelity, and can be used as a gate diagnostic tool. We apply our method to one- and two-qubit gates on two different superconducting systems available in the cloud, namely the QuTech Quantum Infinity and the IBM Quantum Experience. We discuss how cross-talk, leakage and non-Markovian errors affect the eigenvalue data.

This chapter has been published in npj Quantum Inf. **5**, 74 (2019) [1]. F. B. contributed to the development of the theoretical concepts presented and performed the simulations on non-Markovianity. Furthermore, F. B. contributed to the writing, especially in Sections 7.2.2, 7.5 and 7.7.3.

7.1. INTRODUCTION

A central challenge on the path towards large-scale quantum computing is the engineering of high-quality quantum gates. To achieve this goal, many methods which accurately and reliably characterize quantum gates have been developed (see also Section 3.4). Some of these methods are scalable, meaning that they require an effort which scales polynomially in the number of qubits on which the gates act. Scalable protocols, such as randomized benchmarking [2–9] necessarily give a partial characterization of the gate quality, for example an average gate fidelity. Other protocols such as robust tomography [10] or gate-set tomography [11, 12] trade scalability for a more detailed characterization of the gate. A desirable feature of all the above protocols is that they are resistant to state-preparation and measurement (SPAM) errors. The price of using SPAM-resistant (scalable) methods is that these protocols have significant experimental complexity and/or require assumptions on the underlying hardware to properly interpret their results.

In this work we present spectral quantum tomography, a simple non-scalable method that extracts spectral information from noisy gates in a SPAM-resistant manner. To process the tomographic data and obtain the spectrum of the noisy gate, we rely on the matrix-pencil technique, a well-known classical signal processing method. This technique has been advocated in [9] in the context of randomized benchmarking and has also been used in [13] for processing data in the algorithm of quantum phase estimation. It has also been used, under the phrase 'linear systems identification', in [14] to predict the time evolution of quantum systems. While the matrix pencil technique leads to explicitly useful estimates of eigenvalues and their amplitudes, we note that the same underlying idea is used in the method of "delayed vectors" which has been proposed in [15] to assess the dimensionality of a quantum system from its dynamics. This "delayed vectors" approach has been applied to assess leakage in superconducting devices in [16].

The spectral information of a noisy gate $\mathscr{S}$, which approximates some target unitary U, is given by the eigenvalues of the so-called Pauli transfer matrix representing $\mathscr{S}$. These eigenvalues, which are of the form $\lambda = \exp(-\gamma)\exp(i\phi)$, contain information about the quality of the implemented gate. Intuitively, the parameter γ captures how much the noisy gate deviates from unitarity due to entanglement with an environment, while the angle ϕ can be compared to the rotation angles of the targeted gate U. Hence ϕ gives information about how much one over- or under-rotates. The spectrum of $\mathscr{S}$ can also be related to familiar gate-quality measures such as the average gate fidelity and the unitarity. Moreover, in the case of a noisy process modeled by a Lindblad equation, the spectrum can be easily related to the more familiar notions of relaxation and dephasing times.

The main advantage of spectral quantum tomography is its simplicity, requiring only the (repeated) application of a single noisy gate $\mathscr{S}$, as opposed to the application of a large set of gates as in randomized benchmarking, gate-set tomography and robust tomography. Naturally, simplicity and low-cost come with some drawback, namely the method does not give information about the eigenvectors of the noisy gate, such as the axis around which one is rotating. However, information about the eigenvectors is intrinsically hard to extract in a SPAM-resistant fashion since SPAM errors can lead to additional rotations [17]. Another feature of spectral quantum tomography is that it can be used to extract signatures of non-Markovianity, namely the phenomenon where the noisy gate $\mathscr{S}$ depends on the context in which it is applied (e.g. time of application,

whether any gates have been applied before it). As we show in this chapter, our method can be used to detect various types of non-Markovian effects such as coherent revivals, parameter drifts, and Gaussian-distributed time-correlated noise. It is also possible to distinguish non-Markovian effects from qubit leakage. For these reasons we believe that spectral quantum tomography adds a useful new tool to the gate-characterization toolkit. The method could also have future applications in assessing the performance of logical gates in a manner which is free of logical state preparation and measurement errors, see the Discussion Section 7.6.1.

7.2. EIGENVALUES OF TRACE-PRESERVING COMPLETELY POSITIVE (TPCP) MAPS

Take a unitary gate U on a d-dimensional space with $U|\psi_j\rangle = e^{i\phi_j}|\psi_j\rangle$. The corresponding TPCP map $\mathcal{S}_U(\rho) = U\rho U^\dagger$ has one trace-full eigenvector, namely I with eigenvalue 1, as well as d^2-1 traceless eigenvectors. In particular, there are d^2-d traceless eigenvectors of the form $|\psi_j\rangle\langle\psi_l|$ for $j \neq l$ with eigenvalues $\exp(i(\phi_j-\phi_l))$, and $d-1$ traceless eigenvectors of the form $|\psi_1\rangle\langle\psi_1| - |\psi_j\rangle\langle\psi_j|$ for $j = 2,\ldots,d$ with eigenvalue 1.

For general TPCP maps it is convenient to use the Pauli transfer matrix formalism. For an n-qubit system ($d = 2^n$) consider the normalized set of Pauli matrices P_μ for $\mu = 0,\ldots,N$ with $N+1 = 4^n = d^2$, where $P_0 = I/\sqrt{2^n}$ and the normalization is chosen such that $\mathrm{Tr}[P_\mu P_\nu] = \delta_{\mu\nu}$. For a TPCP map $\mathcal{S}$ acting on n qubits, the Pauli transfer matrix is then defined as

$$S_{\mu\nu} = \mathrm{Tr}\big[P_\mu \mathcal{S}(P_\nu)\big], \quad \mu,\nu = 0,\ldots,N. \tag{7.1}$$

The form of the Pauli transfer matrix S is [18]

$$\mathcal{S} \leftrightarrow S = \begin{pmatrix} 1 & 0 \\ \mathbf{s} & T^{\mathcal{S}} \end{pmatrix}, \tag{7.2}$$

where $T^{\mathcal{S}}$ is a real $N\times N$ matrix and $\mathbf{s}$ is a N-dimensional column vector. The 1 and 0's in the top row of the Pauli transfer matrix are due to the fact that $\mathcal{S}$ is trace-preserving. For a unital $\mathcal{S}$ which obeys $\mathcal{S}(I) = I$, the vector $\mathbf{s} = 0$.

A few properties are known of the eigenvalue-eigenvector pairs of S, i.e. the pairs $(\lambda, \vec{\mathbf{v}})$ with $S\mathbf{v} = \lambda\mathbf{v}$:

- The eigenvalues of S are 1 and the eigenvalues of $T^{\mathcal{S}}$ since the solutions of the equation $\det(S-\lambda I) = 0$ are the solutions of the equation $(1-\lambda)\det(T^{\mathcal{S}}-\lambda I) = 0$.
- The eigenvalues of S, and thus the eigenvalues of $T^{\mathcal{S}}$, come in complex-conjugate pairs. This is true because $T^{\mathcal{S}}$ is a real matrix.
- The eigenvalues of $T^{\mathcal{S}}$ (or S for that matter) have modulus less than 1, i.e. $|\lambda| \leq 1$ (see e.g. Proposition 6.1 in [19]).

If $T^{\mathcal{S}}$ is *diagonalizable* as a matrix, it holds that $T^{\mathcal{S}} = VDV^{-1}$ where D is a diagonal matrix and V a similarity transformation. Generically, $T^{\mathcal{S}}$ will be diagonalizable, in which case there are N eigenvalue-eigenvector pairs for T. A sufficient condition for

diagonizability is, for example, that all the eigenvalues of $T^{\mathcal{S}}$ are distinct. In Section 7.7.1 we give examples and discuss what it means if $T^{\mathcal{S}}$ is not diagonalizable.

For some simple single-qubit channels we can explicitly compute the spectrum. For instance, for a single-qubit depolarizing channel with depolarizing probability p, the eigenvalues of the sub-matrix $T^{\mathcal{S}}$ of the Pauli transfer matrix are $\{1-p, 1-p, 1-p\}$. For a single qubit amplitude-damping channel with damping rate p they are $\{\sqrt{1-p}, \sqrt{1-p}, 1-p\}$ [12].

7.2.1. RELATION TO GATE-QUALITY MEASURES

The eigenvalues of the Pauli transfer matrix of a noisy gate $\mathcal{S}$ can be related to several other known measures of gate quality such as the average gate fidelity $\mathcal{F}(\mathcal{S}, U)$, the gate unitarity $u(\mathcal{S})$ and, for a single qubit ($n = 1$), the gate unitality.

The average gate fidelity is defined as $\mathcal{F}(\mathcal{S}, U) = \int d\phi \langle\phi| U^\dagger \mathcal{S}(|\phi\rangle\langle\phi|) U |\phi\rangle$. This fidelity relates directly to the entanglement fidelity $\mathcal{F}_{\text{ent}}(\mathcal{S}, U)$ via $\mathcal{F} = \frac{\mathcal{F}_{\text{ent}} d+1}{d+1}$ [20], where the entanglement fidelity is defined as

$$\mathcal{F}_{\text{ent}}(\mathcal{S}, U) = \text{Tr}\big[I \otimes U |\Psi\rangle\langle\Psi| I \otimes U^\dagger (I \otimes \mathcal{S})(|\Psi\rangle\langle\Psi|)\big],$$

where $|\Psi\rangle = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} |i, i\rangle$ is a maximally entangled state. Using that $|\Psi\rangle\langle\Psi| = \frac{1}{d} \sum_{\mu=0}^{N} P_\mu \otimes P_\mu$ and $U P_\mu U^\dagger = \sum_\kappa T_{\mu\kappa}^{U^\dagger} P_\kappa$ we can write

$$\mathcal{F}_{\text{ent}}(\mathcal{S}, U) = \frac{1}{d^2} \sum_\mu \text{Tr}\big[U P_\mu U^\dagger \mathcal{S}(P_\mu)\big] = \frac{1}{d^2}\big(1 + \text{Tr}\big[T^{U^\dagger} T^{\mathcal{S}}\big]\big).$$

Thus for the (entanglement) fidelity of a noisy gate $\mathcal{S}$ with respect to the identity channel $U = I$, one has $\mathcal{F}_{\text{ent}}(\mathcal{S}, I) = \frac{1}{d^2}(1 + \sum_i \lambda_i)$, implying a direct relation to the spectrum $\{\lambda_i\}$ of $T^{\mathcal{S}}$. A more interesting relation is how the eigenvalues of $T^{\mathcal{S}}$ bound the fidelity with respect to a targeted gate U. In Section 7.7.2 we prove that the entanglement fidelity can be upper bounded as

$$\mathcal{F}_{\text{ent}}(\mathcal{S}, U) \leq \frac{1}{d^2}\left[1 + (d^2-1)\left(\sqrt{1 - \frac{\sum_j |\lambda_j|^2}{d^2-1}} + \xi_{\max}\right)\right], \tag{7.3}$$

where $\xi_{\max} = \frac{1}{d^2-1} |\sum_j \lambda_j^{\text{ideal}} \lambda_j^*|$ with λ_j^{ideal} the eigenvalues of T^U with U the targeted unitary, ordered such that the sum $|\sum_j \lambda_j^{\text{ideal}} \lambda_j^*|$ is maximal.

This upper bound is not particularly tight, but for the case of a single qubit we can make a much stronger numerical statement, see Section 7.7.2.

Another measure of gate quality, namely the unitarity or the coherence of a channel [6] on a d-dimensional system, is defined as

$$u(\mathcal{S}) = \frac{d}{d-1} \int d\phi \, \text{Tr}\big[[\mathcal{S}'(|\phi\rangle\langle\phi|)]^\dagger \mathcal{S}'(|\phi\rangle\langle\phi|)\big], \tag{7.4}$$

where $\mathcal{S}'(\rho) \coloneqq \mathcal{S}(\rho) - \text{Tr}[\mathcal{S}(\rho)] I/\sqrt{d}$. A more convenient but equivalent definition is

$$u(\mathcal{S}) = \frac{1}{d^2-1} \text{Tr}\big[T^{\mathcal{S}\dagger} T^{\mathcal{S}}\big] = \frac{1}{d^2-1} \sum_i \sigma_i (T^{\mathcal{S}})^2, \tag{7.5}$$

where $\{\sigma_i\}$ are the singular values of the matrix $T^{\mathscr{S}}$. The unitarity captures how close the channel is to a unitary gate. A lower bound on the unitarity is given by Proposition 2 in Ref. [17]:

$$u(\mathscr{S}) \geq \frac{1+\sum_{i=1}^{d^2-1}|\lambda_i|^2-d}{d(d-1)}, \tag{7.6}$$

where $\{\lambda_i\}$ are the eigenvalues of $T^{\mathscr{S}}$. For a single qubit, an upper bound on the unitarity can also be given in terms of a non-convex optimization problem, see Section 7.7.2.

The unitality of a TPCP map is defined as $1-||\mathbf{s}||^2$ with $\mathbf{s}$ in Eq. (7.2). Specifically, for single-qubit channels one can derive the bound [17]

$$||\mathbf{s}||^2 \leq 1-|\lambda_1|^2-|\lambda_2|^2-|\lambda_3|^2+2\lambda_1\lambda_2\lambda_3. \tag{7.7}$$

7.2.2. RELATION TO RELAXATION AND DEPHASING TIMES

We consider the eigenvalues of a superoperator induced by a simple Lindblad equation modeling relaxation and decoherence of a driven qubit, as an example. We have a Lindblad equation with time-independent Lindbladian $\mathscr{L}$:

$$\dot{\rho} = \mathscr{L}(\rho). \tag{7.8}$$

The formal solution of Eq. (7.8) is given by $\rho(t) = e^{t\mathscr{L}}(\rho(t=0))$, where $e^{t\mathscr{L}}$ is a TPCP map for every t. We are interested in the total evolution after a certain gate time τ and set $\mathscr{S}_\tau = e^{\tau\mathscr{L}}$. We assume a simple model in which a qubit evolves according to a Hamiltonian $H = (h_x X + h_y Y + h_z Z)/2$ and is subject to relaxation and pure dephasing processes, according to the Lindbladian:

$$\mathscr{L}(\rho) = -i[H,\rho] + \frac{1}{T_1}\left(\sigma_-\rho\sigma_+ - \frac{1}{2}\{\sigma_+\sigma_-,\rho\}\right) + \frac{1}{2T_\phi}(Z\rho Z-\rho).$$

We define the relaxation respectively dephasing rates $\Gamma_1 = 1/T_1$ and $\Gamma_2 = 1/T_2 = 1/(2T_1) + 1/T_\phi$. The Pauli transfer matrix $L^{\mathscr{L}}$ of $\mathscr{L}$ then takes the form

$$L^{\mathscr{L}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -\Gamma_2 & h_z & h_y \\ 0 & -h_z & -\Gamma_2 & h_x \\ \Gamma_1 & -h_y & -h_x & -\Gamma_1 \end{pmatrix}. \tag{7.9}$$

We will denote the eigenvalues of $L^{\mathscr{L}}$ by Ω_j for $j \in \{0,\ldots,3\}$ and the eigenvalues of $\mathscr{S}_\tau$ by λ_j for $j \in \{0,\ldots,3\}$. As expected, $\Omega_0 = 0$ implying that $\lambda_0 = e^0 = 1$ is an eigenvalue of $\mathscr{S}_\tau$. The other three eigenvalues of $L^{\mathscr{L}}$ can be found from the 3×3 sub-matrix in the lower-right corner. Here we consider some simple cases.

Case 1: $h_x = h_y = h_z = 0$. In this case, for $j = 1,2,3$ the three eigenvalues of $\mathscr{L}$ and $\mathscr{S}_\tau$ are clearly

$$\Omega_j \in \{-\Gamma_2,-\Gamma_2,-\Gamma_1\},$$
$$\lambda_j \in \{e^{-\Gamma_2\tau}, e^{-\Gamma_2\tau}, e^{-\Gamma_1\tau}\},$$

thus relating directly to the relaxation and dephasing rates.

Case 2: $h_x = h_y = 0$. In this case we have

$$\Omega_j \in \{-\Gamma_2 + ih_z, -\Gamma_2 - ih_z, -\Gamma_1\},$$
$$\lambda_j \in \{e^{-\Gamma_2\tau}e^{ih_z\tau}, e^{-\Gamma_2\tau}e^{-ih_z\tau}, e^{-\Gamma_1\tau}\},$$

where we have separated the decaying part of the λ_j (corresponding to the real part of the Ω_j) and their phases (corresponding to the imaginary part). If we work in the rotating frame of the qubit, h_z can be understood as an over-rotation along the Z-axis, which would appear in the spectrum as an extra phase imparted to two of the eigenvalues. Again we see that the decaying part of the eigenvalues directly relates to the relaxation and dephasing rates.

Case 3: $h_y = h_z = 0$. This case shows that over-rotations can also modify the decay strength of the eigenvalues. We analyze the eigenvalues as a function of h_x. From $L^{\mathscr{L}}$ in Eq. (7.9) we see that $\Omega_1(h_x) = -\Gamma_2$ for all h_x. For the other eigenvalues we have

$$\Omega_{2,3}(h_x) = -\frac{1}{2}\Big(\Gamma_1 + \Gamma_2 \pm \sqrt{(\Gamma_1 - \Gamma_2)^2 - 4h_x^2}\Big). \tag{7.10}$$

We see that if $|h_x| < |\Gamma_1 - \Gamma_2|/2 \equiv h_x^{\mathrm{cr}}$, only the moduli of λ_2 and λ_3 are affected as compared to *Case 1*, in other words, λ_2 and λ_3 only decay with no extra phases. On the contrary, the phases of these eigenvalues becomes non-zero when the driving is sufficiently strong: $|h_x| > h_x^{\mathrm{cr}}$. It implies that if we look at the dynamics induced by the Lindblad equation, real oscillations, not only decay, will be present as a function of τ. Hence these two scenarios represent respectively the overdamped ($|h_x| < h_x^{\mathrm{cr}}$) and underdamped regime ($|h_x| > h_x^{\mathrm{cr}}$), similar to the dynamics of a vacuum-damped qubit-oscillator system, see e.g. Ref. [21]. At $|h_x| = h_x^{\mathrm{cr}}$, the system is critically damped and $L^{\mathscr{L}}$ does not have 4 linearly-independent eigenvectors, meaning that the Pauli transfer matrix of $\mathscr{S}_\tau$ is not diagonalizable. In this case the dynamics also has a linear dependence on t besides the exponential decay with t, see the discussion in Section 7.7.1.

7.3. SPECTRAL TOMOGRAPHY

In this section we describe the spectral tomography method, which estimates the eigenvalues of $\mathscr{S}$, where $\mathscr{S}$ is a TPCP implementation of a targeted unitary gate.

We model state-preparation errors as a perfect preparation step followed by an unknown TPCP map $\mathscr{N}_{\mathrm{prep}}$. Similarly, measurement errors are modeled by a perfect measurement preceded by an unknown TPCP map $\mathscr{N}_{\mathrm{meas}}$. We assume that when we apply the targeted gate k times, an accurate model of the resulting noisy dynamics is $\mathscr{S}^k$. The spectral tomography method can be applied without this assumption but the interpretation of its results is more difficult, see Section 7.5 for a discussion. The method works by constructing the following *signal* function, for $k = 0, 1, \ldots, K$ for some fixed K:

$$g(k) = \sum_{\mu=1}^{N} \mathrm{Tr}\big[P_\mu \mathscr{N}_{\mathrm{meas}} \circ \mathscr{S}^k \circ \mathscr{N}_{\mathrm{prep}}(P_\mu)\big]. \tag{7.11}$$

Gathering the data to estimate $g(k)$ requires (1) picking a traceless n-qubit Pauli P_μ, (2) preparing an n-qubit input state in one of the 2^n basis states corresponding to this chosen

Pauli, (3) applying the gate k times and measuring in the same chosen Pauli basis, and (4) repeating (1-3) over different Pauli's, basis states and experiments to get good statistics. As in standard process tomography [22], one takes linear combinations of the estimated probabilities for the outcomes to construct an estimator of a Pauli operator on a Pauli input. This gives an estimate of $g(k)$ for a fixed k. Repeating this process for $k \in \{0,\dots,K\}$ we reconstruct the entire signal function. In Section 7.3.2 we discuss the cost of doing these experiments as compared to randomized benchmarking.

Let us now examine how $g(k)$ depends on the eigenvalues of the matrix T. When there are no SPAM errors, that is, $\mathcal{N}_{\text{meas}}$ and $\mathcal{N}_{\text{prep}}$ are identity channels, we have

$$g^{\text{NO SPAM}}(k) = \sum_{\mu=1}^{N} (T^k)_{\mu\mu} = \text{Tr}\big[T^k\big] = \sum_{j=1}^{N} \lambda_j^k, \tag{7.12}$$

where $\{\lambda_j\}$ are the eigenvalues of T. The last step in this equality follows directly when T is diagonalizable, but it can alternatively be proved using the so-called Schur triangular form of T (we give this proof in Section 7.7.1).

When $\mathcal{N}_{\text{meas}}$ and $\mathcal{N}_{\text{prep}}$ are not identity channels, we have

$$g(k) = \text{Tr}\big[T_{\text{meas}}\, T^k\, T_{\text{prep}}\big] = \text{Tr}\big[A_{\text{SPAM}} D^k\big] = \sum_{j=1}^{N} A_j \lambda_j^k, \tag{7.13}$$

where T_{meas} and T_{prep} are respectively the T-submatrices of the Pauli transfer matrix of $\mathcal{N}_{\text{meas}}$ and $\mathcal{N}_{\text{prep}}$. Here we assume that $T = VDV^{-1}$ is diagonalizable and the matrix $A_{\text{SPAM}} = V^{-1} T_{\text{prep}} T_{\text{meas}} V$ captures the SPAM errors. One may expect that A_{SPAM} is close to the identity matrix in the typical case of low SPAM errors, in particular one may expect that $A_j \neq 0$ for all j so that all eigenvalues of T are present in the signal $g(k)$.

In principle, one could take more tomographic data and consider a full matrix-valued signal $c_{\mu\nu}(k) = \text{Tr}\big[P_\mu \mathcal{N}_{\text{meas}} \circ \mathcal{S}^k \circ \mathcal{N}_{\text{prep}}(P_\nu)\big]$ instead of only Eq. (7.11). This requires doing many more experiments and there is no clear advantage in terms of the ability to determine the spectrum.

7.3.1. SIGNAL ANALYSIS OR MATRIX-PENCIL METHOD FOR EXTRACTING EIGENVALUES

In this section we review the classical signal-processing method which reconstructs, from the (noisy) signal $g(k) = \sum_{j=1}^{N} A_j \lambda_j^k$ for $k = 0,\dots,K$, an estimate for the eigenvalues λ_j and the amplitudes A_j. Note that we have $g(k) \in \mathbb{R}$ due to Eq. (7.11). Not surprisingly, this signal-processing method has been employed and reinvented in a variety of scientific fields. We implement the so-called ESPRIT analysis described in Ref. [23], but see also Ref. [24]. In the context of spectral tomography we know that the signal $g(k)$ will in principle contain N eigenvalues (which are possibly degenerate). However, we can vary the number of eigenvalues we use to fit the signal to see whether a different choice than N gives a significantly better fit. This is relevant in particular when the implemented gate contains leakage or non-Markovian dynamics, see Section 7.5.

We require at least $K \geq 2N-2$ in order to determine the eigenvalues accurately. This implies that for a single-qubit gate with $N = 3$ we need at least $K = 4$ and for a two-qubit

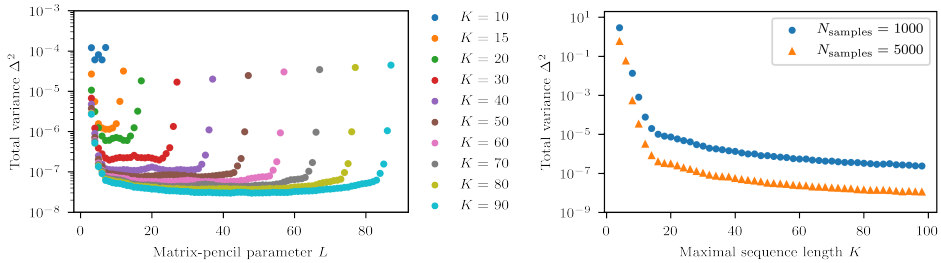

Figure 7.1: Preliminary study of the numerical accuracy of the matrix-pencil method as a function of L, K and N_{samples}. **(Left)** We use the matrix-pencil method with different L's and K's to estimate the eigenvalues of a random single-qubit channel, for $N_{\text{samples}} = 1000$. On the vertical axis we give the variance in the estimate of the eigenvalues: $\Delta^2 = \frac{1}{3}(\sum_{j=1}^{N=3} |\lambda_j - \lambda_j^{\text{est}}|^2)$. We see that, as long as the matrix-pencil parameter L is chosen away from 0 or K, the accuracy of the reconstructed signal is nearly independent of L. Furthermore, we see that higher K's can achieve a lower Δ^2. **(Right)** We generate a random single-qubit channel and set $L = K/2$. We plot Δ^2 as a function of K for two different values of $N_{\text{samples}} = 1000$ and $N_{\text{samples}} = 5000$, showing how a larger N_{samples} suppresses the total variance. We see that for constant N_{samples} the accuracy of the method increases rapidly at first when K is increased, but it increases more slowly if K is already large. This can be explained by the fact that the signal decreases exponentially in K and so data points for large K have much lower signal-to-noise ratio. For both figures, random channels were generated using QuTip's random TPCP map functionality, and measurement noise was approximated by additive Gaussian noise with standard deviation equal to $1/\sqrt{N_{\text{samples}}}$.

gate with $N = 15$ we need at least $K = 28$. However, the signal $g(k)$ has sampling noise due to a bounded N_{samples} and in practice it is good to choose K larger than strictly necessary to make the reconstruction more robust against noise. We study the effect of varying K in Fig. 7.1 (left panel).

The method goes as follows and relies on picking a so-called pencil parameter L.

Let us assume for now that each $g(k)$ is learned without sampling noise. One constructs a $(K-L+1)\times(L+1)$-dimensional data matrix Y as

$$Y = \begin{pmatrix} g(0) & g(1) & \dots & g(L) \\ g(1) & g(2) & \dots & g(L+1) \\ g(2) & \vdots & & \vdots \\ \vdots & & & \vdots \\ g(K-L) & \dots & \dots & g(K) \end{pmatrix} = \sum_{j=1}^{N} A_j \begin{pmatrix} 1 & \lambda_j & \dots & \lambda_j^L \\ \lambda_j & \lambda_j^2 & \dots & \lambda_j^{L+1} \\ \lambda_j^2 & \vdots & & \vdots \\ \vdots & & & \vdots \\ \lambda_j^{K-L} & \dots & \dots & \lambda_j^K \end{pmatrix}. \tag{7.14}$$

Note that $\text{rank}(Y) \leq N$ since Y is a sum of at most N rank-1 matrices when there are N eigenvalues. Consider two submatrices of Y: the matrix G_0 is obtained from Y by deleting the last column of Y, while the matrix G_1 is obtained by deleting the first column of Y. When $L = \frac{K}{2}$, the matrices G_0 and G_1 are square matrices of dimension $M = \frac{K}{2} + 1$. For this choice of L, the smallest value of K so that $M = N$ is $2N-2$. We seek a time-shift matrix $\mathfrak{T}$ such that $\mathfrak{T}G_0 = G_1$. When $M \geq N$, there certainly exists a matrix $\mathfrak{T}$ such that for

all $j \in \{1,\ldots,N\}$:

$$\mathfrak{T}\begin{pmatrix} 1 \\ \lambda_j \\ \vdots \\ \lambda_j^M \end{pmatrix} = \lambda_j \begin{pmatrix} 1 \\ \lambda_j \\ \vdots \\ \lambda_j^M \end{pmatrix}. \tag{7.15}$$

Furthermore, if G_0^{-1} exists, which is the case when $\text{rank}(G_0) = M$, this matrix $\mathfrak{T}$ will be uniquely given as $G_1 G_0^{-1}$. Hence, in this case there is a unique matrix $\mathfrak{T}$, obtained by constructing $G_1 G_0^{-1}$ from the data, which is guaranteed to have $\{\lambda_j\}$ as eigenvalues. When the pencil parameter $L > \frac{K}{2}$, one needs to ensure that there are at least N rows of the matrix Y: if not, $\mathfrak{T}$ would be of dimension less than N, not giving N eigenvalues. This implies $K \geq N + L - 1$.

The general method for a non-square Y which includes an additional sampling-noise reduction step then goes as follows. The choice for N in the procedure can be varied from its minimal value equal to $d^2 - 1$ to a larger value, depending on a goodness-of-fit.

1. Construct a singular-value decomposition of the matrix Y, i.e. $Y = R_1 \Sigma R_2^T$ and replace the diagonal matrix Σ by a diagonal matrix Σ_{clean} with only the largest N singular values. Let $Y_{\text{clean}} = R_1 \Sigma_{\text{clean}} R_2^T$. This step is to reduce sampling noise.

2. Take the submatrices G_0 and G_1 of Y_{clean}.

3. Compute $\mathfrak{T} = G_1 G_0^+$ where G_0^+ is the Moore-Penrose pseudo-inverse of the matrix G_0 so that $\mathfrak{T}$ is a matrix with at most N non-zero eigenvalues.

4. Compute the eigenvalues of $\mathfrak{T}$: these will be the estimates λ_j^{est} of λ_j for all $j \in \{1,\ldots N\}$. Formally, the linear matrix pencil is $G_0 - \lambda G_1$ and the eigenvalues of this matrix pencil, i.e. the values where $\det(G_0 - \lambda G_1) = 0$, are the λ_j^{est}.

We have first applied this method on the signal $g(k)$ of a randomly chosen single-qubit channel: by varying K and L we want to understand the role of the matrix-pencil parameter L and the choice for a larger K. The results are shown in Fig. 7.1 (left panel). Note that the chosen K's are quite far above the bound $K \geq N + L - 1$ to effectively suppress sampling noise. For each K there is a flat region in L where Δ^2 is roughly constant. In the remainder we will choose $L = K/2$, putting ourselves in the middle of this region. Fig. 7.1 (right panel) shows how increasing N_{samples} lowers the total variance of the estimated eigenvalues.

An additional processing step is the determination of the (complex) amplitudes $\{A_j\}$. Viewing $g(k)$ as a set of $K+1$ inner products between the vector $(A_1,\ldots,A_N)$ and the linearly-independent vectors $(\lambda_1^k, \lambda_2^k, \ldots, \lambda_N^k)$, it is clear that, given perfect knowledge of $g(k)$, the $\{A_j\}$ are uniquely determined when $K + 1 \geq N$. Since $g(k)$ is known with sampling noise, the $\{A_j\}$ can be found by solving the least-squares minimization problem $\min_{A_j} \sum_k |g(k) - \sum_j A_j (\lambda_j^{\text{est}})^k|^2$. The optimal values in this minimization A_j^{est} and λ_j^{est} together form the reconstructed signal $g^{\text{est}}(k)$ and the error is given by

$$\epsilon_N^{\text{rms}} = \left(\frac{1}{K+1} \sum_{k=0}^{K} |g(k) - g_N^{\text{est}}(k)|^2 \right)^{1/2}. \tag{7.16}$$

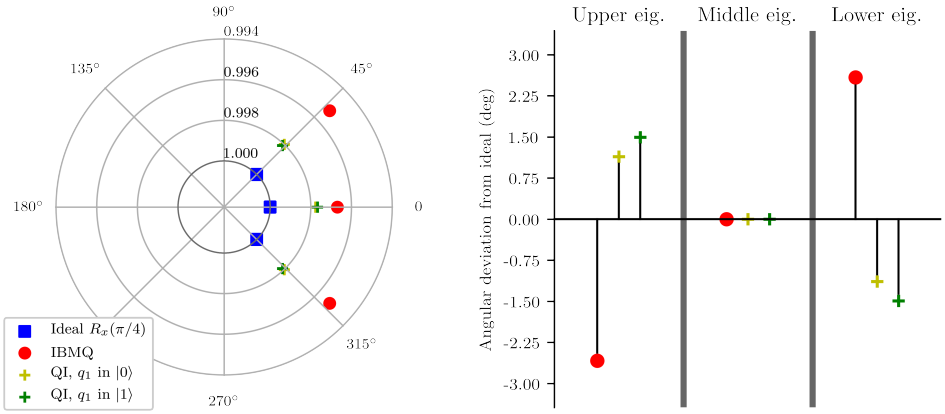

Figure 7.2: **(Left)** Spectral footprints for single-qubit $R_x(\pi/4)$ gates on the *ibmqx4* (IBMQ) and the Quantum Infinity (QI) chips at $K = 50$, $L = 30$ and $N_{\text{samples}} = 8192$. The modulus of the eigenvalues is plotted in the radial direction and in particular it decreases from the center to the outside and it is equal to 1 on the (most inner) black circumference. The angular coordinate corresponds to the phase of the eigenvalues. **(Right)** Precise value of the deviation of the phases of the three eigenvalues from the ideal ones.

7.3.2. RESOURCES

It is interesting to consider the amount of experiments that must be done to perform spectral quantum tomography. One must estimate the function $g(k)$ defined in Eq. (7.13). This reconstruction process requires running $2^n \times N \times (K+1)$ different experiments and repeating each experiment N_{samples} times. For a single-qubit gate we need $6(K+1)$ experiments, while for a two-qubit gate we need $60(K+1)$. Note that while the number of experiments scales exponentially with qubit number (not surprising for a tomographic protocol), the number of experiments needed for performing spectral tomography on single and two-qubit gates is comparable to the number of experiments that must be performed in randomized benchmarking on one or two qubits (which provides only average gate information). In randomized benchmarking one must sample M random sequences for each sequence length $k \in [0:K]$, yielding $M \times (K+1)$ experiments. This M is independent of the number of qubits [25]. In experiments M is often chosen between $M \approx 40$ [26, 27] at the low end and $M \approx 150$ at the higher end [28]. Values of K reported in randomized benchmarking experiments are also comparable to (or even higher than, see [26] where $K \approx 300$ is considered) the values of K used for single and two qubit spectral tomography (see Section 7.4).

7.4. SPECTRAL TOMOGRAPHY ON TWO SUPERCONDUCTING CHIPS

We have executed the spectral tomography method on a single-qubit $\pi/4$ rotation around the X-axis: $R_x(\pi/4) = \exp(-i\pi X/8)$. For this gate the ideal matrix $T^{R_x(\pi/4)}$ should have eigenvalues 1, $\exp(i\pi/4)$ and $\exp(-i\pi/4)$. We execute this gate on two different systems available in the cloud: the two-qubit Quantum Infinity provided by the DiCarlo group at QuTech (for internal QuTech use) and the *ibmqx4* (IBM Q5 Tenerife) available at `https://quantumexperience.ng.bluemix.net/qx/editor`. The results of this experiment

are shown in Fig. 7.2 (left panel) in a polar plot which we refer to as the 'spectral footprint' of the gate. For clarity, in Fig. 7.2 (right panel) we have also plotted the phase deviation from ideal for the implemented gates.

On the two-qubit (q_0 and q_1) Quantum Infinity chip, we perform the single-qubit gate experiment on q_0 twice to study cross-talk: in one case the undriven qubit q_1 on the chip is in state $|0\rangle$, in the other case q_1 is in state $|1\rangle$. Since the residual off-resonant qubit coupling, mediated through a common resonator, is non-zero, we observe a small difference between these two scenarios, see Fig. 7.2. For the Quantum Infinity chip, when q_1 is $|0\rangle$ we estimate $\lambda_j^{\text{est}} \in \{0.691+0.719i, 0.691-0.719i, 0.997\}$, while $\lambda_j^{\text{est}} \in \{0.687+0.7239i, 0.687-0.724i, 0.998\}$ when q_1 is $|1\rangle$. Using the single-qubit fidelity bound given in Section 7.7.2, we can compute that the fidelity with respect to the targeted gate $R_x(\pi/4)$ can be no more than 0.999 regardless of the state of q_1. We can also compute upper and lower bounds on the unitarity (see Section 7.2 and Section 7.7.2) which yields $0.994 \leq u \leq 0.996$ regardless of the state of q_1.

Regarding the *ibmqx4* chip, the data are taken when all other qubits are in state $|0\rangle$. The reconstructed eigenvalues $\lambda_j^{\text{est}} \in \{0.735+0.671i, 0.735-0.671i, 0.996\}$ turn out to be lower in magnitude. From these numbers we can conclude that the fidelity to the target gate is no higher than 0.998 and the unitarity lies between 0.988 and 0.991.

For all these numbers a two-way 95% confidence interval (for both real and imaginary parts) deviates by less than 0.005 from the quoted values. The confidence intervals are obtained through a Wild resampling bootstrap with Gaussian kernel [29].

We have considered whether the data can be better fitted with more than $N = 3$ eigenvalues. For each experiment we fit the data using N eigenvalues with $N \in \{4, \ldots 15\}$ and we test whether there is a significant increase in goodness-of-fit using a standard F-test [30, Section 2.1.5]. For no experiment and value of N does the resultant p-value drop below 0.05, leading us to conclude that increasing the number of eigenvalues does not significantly increase the accuracy of the fit.

We have also executed a two-qubit CNOT gate on *ibmqx4* (Fig. 7.3). The T matrix of the ideal CNOT gate has 15 eigenvalues and a very degenerate spectrum: 6 eigenvalues are equal to -1 and 9 eigenvalues are equal to 1, but our data, taking $K = 50$, shows that a best fit is obtained using 4 instead of 2 eigenvalues. Using an F-test shows that the goodness-of-fit is significantly improved using 4 eigenvalues rather than 2 or 3, whereas adding more eigenvalues beyond 4 does not significantly improve the goodness-of-fit ($p > 0.05$). We have not tried using larger K (which may lead to a resolution of more eigenvalues) since this would break the requirement that our experiments are executed as a single job performed in a short amount of time on the IBM Quantum Experience. The eigenvalues are $\lambda_j^{\text{est}} \in \{0.939+0.059i, 0.938-0.059i, -0.961+0.067i, -0.961-0.067i\}$, all with a 95% confidence interval smaller than $\pm 3 \times 10^{-3}$ for both real and imaginary parts. It is important to note that these 4 eigenvalues, coming in 2 complex-conjugate pairs, *cannot be the spectrum* of a two-qubit TPCP map $\mathcal{S}$, for the following reasons. As observed in Section 7.2, the submatrix $T^{\mathcal{S}}$ of the Pauli transfer matrix of $\mathcal{S}$ is a real matrix of odd ($4^2 - 1 = 15$) dimension. Since any complex eigenvalues of a real matrix come in conjugate pairs, $T^{\mathcal{S}}$ must have at least one real eigenvalue. Moreover, the data cannot be explained by allowing for leakage, as any eigenvalues associated to a small amount of leakage must have small associated amplitude, as we discuss in Section 7.5. This is not

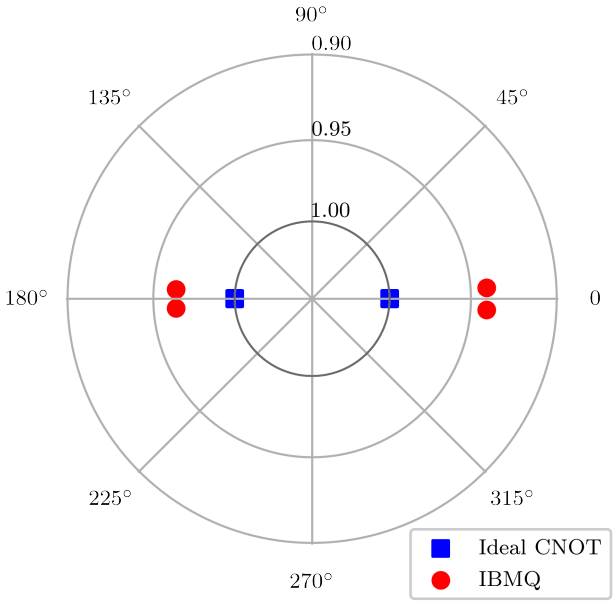

Figure 7.3: Spectral footprint of the CNOT gate for $K = 50$ and $N_{\text{samples}} = 8192$. Even though the CNOT gate has only two (degenerate) eigenvalues, we find that the spectrum of the noisy gate can be best described using 4 distinct eigenvalues. The fact that none of them are real suggests that the data cannot be due to the repeated execution of the same noisy gate. In Section 7.7.3 we propose a simple coherent non-Markovian model that offers a possible mechanism for the absence of real eigenvalues.

the case for the eigenvalues plotted in Fig. 7.3 as all their amplitudes have comparable magnitude $A^{\text{est}} \in \{3.34 - 1.70i, 3.34 + 1.70i, 1.57 + 0.91i, 1.57 - 0.91i\}$. In Section 7.7.3 we propose a simple model based on a frame mismatch accumulation that qualitatively reproduces these eigenvalues. This model is not stochastic but coherent, and it violates the assumption that the applied CNOT gate can be fully modeled as a TPCP map. A possible physical mechanism producing a frame mismatch accumulation can be a drift in an experimental parameter.

We do not compute bounds on the fidelity or unitarity of the CNOT gate since the bounds in Section 7.2.1 do not apply when the evolution is non-Markovian.

7.5. LEAKAGE AND NON-MARKOVIAN NOISE

In this section we consider how spectral tomography behaves under error models that violate the assumptions that go into Eq. (7.13).

Three common mechanisms for gate inaccuracy are (1) cross-talk, meaning the gate depends on or affects the state of other “spectator" qubits, (2) leakage, meaning that the dynamics of the gate acts outside of the computational qubit subspace and (3) non-Markovian dynamics, meaning that the assumption that k applications of the noisy gate are equal to $\mathscr{S}^k$ for some TPCP map $\mathscr{S}$ is incorrect. Characterizing gates with respect to

these features is important for assessing their use in multi-gate/multi-qubit devices for the purpose of quantum error correction or plainly reliable quantum computing [5].

One can see that all three scenarios are due to the dynamics taking place in a larger Hilbert space than the targeted computational qubit space. In the case of leakage, the larger space is an extension of the computational space, while in the other cases the larger space is the tensor product of the computational space with the state space of spectator qubits (1), as explored in Section 7.4, or other quantum or classical degrees of freedom in the environment (3).

7.5.1. LEAKAGE

Let us consider how gate leakage affects the signal $g(k)$, making the analysis for one or two *qutrits*. One can choose an operator basis for the qutrit space such as the basis of the 8 traceless (normalized) Gell-Mann matrices $\sigma_\mu^{\rm GM}$ for $\mu = 1,\dots,8$, together with the normalized identity $\sigma_0^{\rm GM} = \frac{1}{\sqrt{3}} I_3$. For a single qutrit, we can consider the 'Pauli' transfer matrix in this Gell-Mann basis, i.e. $S_{\mu\nu}^{\rm GM} = {\rm Tr}[\sigma_\mu^{\rm GM} \mathcal{S}(\sigma_\nu^{\rm GM})]$ and its submatrix $T^{\rm GM}$.

For a single qutrit, the signal $g^{\rm NO\ SPAM}(k)$ in Eq. (7.12) then equals ${\rm Tr}_{\rm comp}[(T^{\rm GM})^k]$ where ${\rm Tr}_{\rm comp}[A]$ represents the trace over a 3×3 submatrix of A, corresponding to the Gell-Mann matrices which act like X, Y, and Z in the two-dimensional computational space. In other words, we can see the matrix $T^{\rm GM}$ as being composed of blocks:

$$T^{\rm GM} = \begin{pmatrix} T_{\rm comp} & T_{\rm seep} \\ T_{\rm leak} & T_{\rm beyond} \end{pmatrix}, \tag{7.17}$$

where the upper-left block is the sub-matrix whose trace we take in $g^{\rm NO\ SPAM}(k)$. In the absence of other noise sources, $T^{\rm GM}$ corresponds to the evolution of a unitary gate and (assuming it is diagonalizable) it can be diagonalized by a rotation V as $T^{\rm GM} = VDV^{-1}$, where D is a diagonal matrix with all the eigenvalues $\{\lambda_j\}$. If we assume that leakage is low, meaning that $T_{\rm leak}$ and $T_{\rm seep}$ have small norm of $O(\epsilon)$, then at lowest order in ϵ the diagonalizing transformation V will be block-diagonal, i.e. $V \approx V_{\rm comp} \oplus V_{\rm beyond}$. This means that $g^{\rm NO\ SPAM}(k) = {\rm Tr}_{\rm comp}\left[(T^{\rm GM})^k\right] = {\rm Tr}_{\rm comp}\left[VD^kV^{-1}\right] \approx \sum_{j=1}^3 \lambda_j^k + O(\epsilon)$. Thus, at lowest order, the signal will have large amplitude on 3 relevant eigenvalues of the spectrum of $T^{\rm GM}$ and these eigenvalues could have been perturbatively shifted from their ideal location by low leakage. If the leakage is stronger, we can more generally write

$$g^{\rm NO\ SPAM,\ LEAK}(k) = \sum_{j=1}^{8} \tilde{A}_j \lambda_j^k,\ \tilde{A}_j = \langle \sigma_j | V^{-1} \Pi_{\rm comp} V | \sigma_j \rangle . \tag{7.18}$$

Here $|\sigma_j\rangle$ is a vector notation for one of the 8 Gell-Mann matrices σ_j and $\Pi_{\rm comp}$ is the projector onto the basis spanned by the 3 Gell-Mann matrices which are the Paulis in the computational space. From this expression we see that the effect of leakage is the contribution of more eigenvalues to the signal $g(k)$. For low leakage we may expect three dominant eigenvalues with relatively large amplitude $\tilde{A}_j$ and five eigenvalues with small amplitude.

For a gate on two qutrits, identical remarks apply, except that an additional basis transformation is required from the orthogonal Gell-Mann basis to the computational

qubit Pauli basis in order to keep the same division of T^{GM} as in Eq. (7.17). If we have two qutrits, the 80-dimensional traceless subspace is spanned by the matrices $\sigma_\mu^{\mathrm{GM}} \otimes \sigma_\nu^{\mathrm{GM}}$ for $\mu,\nu = 0,\ldots,8$ except $\mu = \nu = 0$. The issue is related to terms such as $\sigma_0^{\mathrm{GM}} \otimes \sigma_{\nu\neq 0}^{\mathrm{GM}}$ since the normalization of the qutrit identity ($\sigma_0^{\mathrm{GM}} = \frac{1}{\sqrt{3}} I_3$) is different from the normalization of the qubit identity ($P_0 = \frac{1}{\sqrt{2}} I_2$). This suggests that for two qutrits it is better to write T^{GM} in a basis which includes the Pauli matrices in the computational subspace ($P_\mu \otimes P_\nu$ for $\mu,\nu = 0,\ldots,3$ except $\mu = \nu = 0$) as a sub-basis. For two qutrits, the signal may then contain up to 80 eigenvalues of which all but 15 are expected to have small amplitude in case of low leakage.

7.5.2. NON-MARKOVIANITY: TIME-CORRELATED NOISE

Non-Markovian behavior of a gate can be due to temporal correlations in the classical or quantum environment of the driven qubit(s). Abstractly, we can include the environment in the gate action so that the evolution for each gate application is a unitary given by some U_{total} acting on system and environment. We can expand the Pauli transfer matrix of U_{total} in a Pauli basis for system and environment and view T_{comp} as a sub-block of T_{total}, similar as in the case of leakage. Diagonalizing T_{total} and taking the trace over the computational space will result in an expression such as Eq. (7.18). For example, an additional spectator or environment qubit can lead to a signal $g(k)$ of a single-qubit gate having contributions from 15 eigenvalues. Choosing a sufficiently large K may allow one to resolve these eigenvalues, even those with small amplitude.

A more malicious, but physically reasonable, form of classical non-Markovian noise makes gate-parameters temporally correlated. In order to numerically study the effect of non-Markovian noise, we consider a toy example in which a perfect CZ gate is followed by a rotation around the X axis on one qubit. For a series of k repetitions of a perfect CZ gate, we assume that each one is followed by the same rotation $R_x(\phi)$ acting always on the same qubit. We assume that the angle ϕ is Gaussian-distributed with mean 0 and standard deviation σ: $\mathbb{P}_\sigma(\phi) = \exp(-\phi^2/2\sigma^2)/\sqrt{2\pi}\sigma$. The time evolution for k repetitions is then given by

$$\mathcal{S}_k(\rho) = \int_{-\infty}^{+\infty} d\phi\, \mathbb{P}_\sigma(\phi) \big(R_x(\phi)\,\mathrm{CZ}\big)^k \rho \big(\mathrm{CZ}\, R_x(\phi)^\dagger\big)^k. \tag{7.19}$$

Note that $\mathcal{S}_k \neq (\mathcal{S}_1)^k$ since this noise is correlated across multiple repetitions of the gate. Furthermore, we assume perfect state-preparation and measurement. In this case, one can represent the noisy gate by some unitary U_{total} acting on the two qubits and on a classical state in a Gaussian stochastic mixture of angles ϕ. The continuous nature of this classical environment state leads to a lack of a hard cut-off on the number of eigenvalues in $g(k)$.

We apply the matrix-pencil method to the corresponding signal $g^{\mathrm{NO\ SPAM}}(k)$ and we use an F-test to determine the optimal number of eigenvalues for each σ (Fig. 7.4). For $\sigma = 22.9°$ and $K = 50$ we find eigenvalues with modulus clearly larger than 1. Those are unphysical but not excluded by the matrix-pencil method. We expect that such $|\lambda^{\mathrm{est}}| > 1$ disappear when considering a longer signal, since $g(k)$ does not increase exponentially in k. In other words, this is a sign that the signal contains more spectral content than

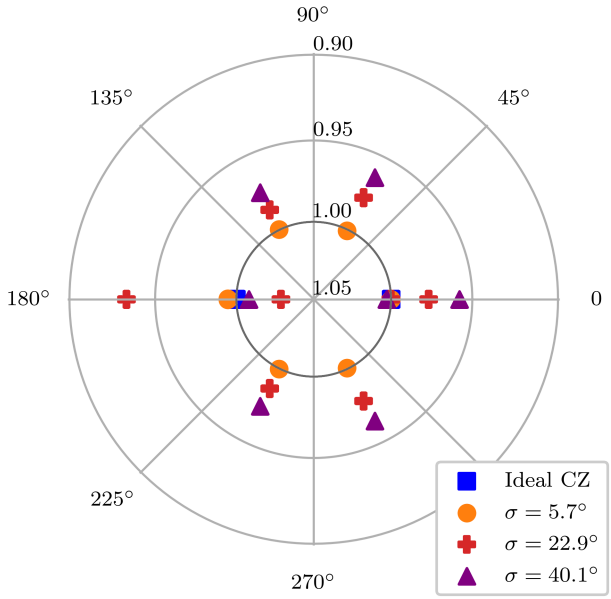

Figure 7.4: Spectral footprint of a simulated CZ gate affected by non-Markovian noise quantified by σ, see Eq. (7.19). For each σ we use an F-test (p-value 0.01) to find the number of eigenvalues that best fit the simulated $g^{\text{NO SPAM}}(k)$ with $K = 50$. We find respectively 7, 12 and 11 eigenvalues for $\sigma = 5.7°, 22.9°, 40.1°$ (here we show only the eigenvalues with modulus greater than 0.9). We observe eigenvalues with modulus larger than 1 if σ is sufficiently large. These results are qualitatively stable if we add a small amount of sampling noise.

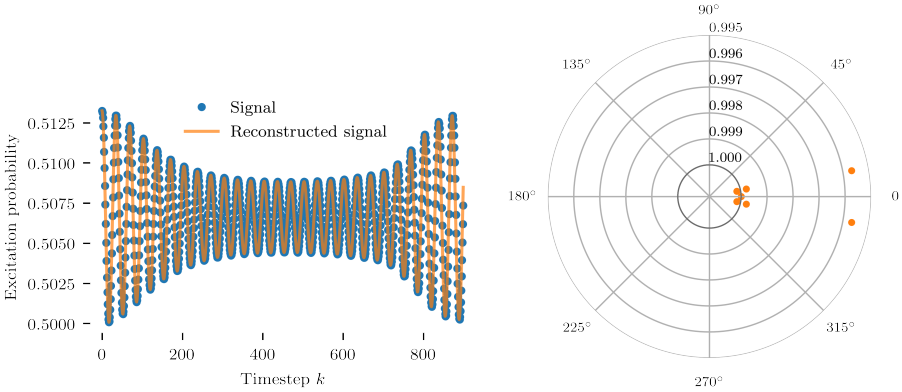

Figure 7.5: Study of the reviving signal given in Eq. (7.20) for $k \cdot \Omega \delta t = k \cdot 0.05$, $\bar{n} = 5$ and $K = 900$. We find that the reviving signal is well reconstructed by a fit with 15 eigenvalues, some of which are distinctly separated as can be seen in the spectral footprint. Some of the eigenvalues are estimated to be larger than 1. This is another example in which the matrix-pencil method gives unphysical eigenvalues in the presence of non-Markovian behavior (revivals here, time-correlated parameters in Fig. 7.4).

can be resolved from the time scale set by K. Indeed, for $\sigma = 22.9°$ we have made the same analysis for larger K's up to $K = 200$ and we find that those eigenvalues get closer and closer to 1. If instead we fix $K = 50$ and consider different σ's, we find that for a low σ (e.g. 5.7°) unphysical eigenvalues are not present (Fig. 7.4), whereas for $\sigma > 22.9°$ (e.g. 40.1°) they get again closer and closer to 1. This latter fact can be understood by noting that increasing σ is analogous to enlarging the time scale set by K, as the characteristic time scale of dephasing gets shorter for a fixed K. Based on these observations, we conclude that there is a certain intermediate time scale at which eigenvalues larger than 1 are extrapolated from the data in the presence of sufficiently-strong non-Markovian noise of the kind described in this section. Section 7.7.3 discusses a model with a different kind of time-correlation leading to a spectral footprint which is incompatible with that of a TPCP map.

7.5.3. NON-MARKOVIANITY: COHERENT REVIVALS

In order to better understand the occurrence of eigenvalue estimates $|\lambda^{\text{est}}| > 1$, we apply the matrix-pencil method on a signal (of a somewhat different physical origin), which has a revival over the time period set by K.

It is well-known that in the exchange of energy between a two-level atom with a bosonic mode, the Rabi oscillations of the two-level atom are subject to temporal revivals. These revivals are due to the fact that the bosonic driving field is not purely classical, but rather gets entangled with the state of the qubit via the Jaynes-Cummings interaction. In particular, for a coherent driving field with coherent amplitude α with average photon number $\bar{n} = |\alpha|^2$, the probability for the atom to be excited equals (see Section 3.4.3 in [21]):

$$P_e(t) = \frac{1}{2} + \frac{1}{2} \sum_{n=0}^{\infty} p_\alpha(n) \cos(\Omega t \sqrt{n+1}), \tag{7.20}$$

with $p_\alpha(n) = \exp(-|\alpha|^2)\frac{|\alpha|^{2n}}{n!}$. We consider $\bar{n} = 5$ and sample the damped oscillatory function $P_e(t) - \frac{1}{2}$ at regular intervals $k\Omega\delta t$ with $\Omega\delta t = 0.05$ and $k = 0, \ldots, K = 900$. The signal function $g(t) = P_e(t) - \frac{1}{2}$ contains eigenvalues equal to $\lambda_n = \exp(\pm i\, 0.05\sqrt{n+1})$ with amplitudes according to the Poisson distribution $p_\alpha(n)$ with mean photon number $\bar{n}$.

We observe that the matrix-pencil method finds eigenvalues larger than 1, see Fig. 7.5, which contribute significantly ($p < 0.01$ via F-test) to the reconstructed signal. We can understand this feature of eigenvalues exceeding 1 as a way in which the matrix-pencil method handles revivals: the signal has more spectral content than what can be resolved from the window of time given by K, in particular there is no hard cut-off on the number of eigenvalues which contribute. We have observed that an analysis of the signal over a longer period of time, that is, a larger K up to $K = 5000$, gives eigenvalues whose norm converges to at most 1.

7.6. DISCUSSION

We have introduced spectral quantum tomography, a simple method that uses tomographic data of the repeated application of a noisy quantum gate to reconstruct the spectrum of this quantum gate in a manner resistant to SPAM errors. We have experimentally validated our method on one- and two-qubit gates and have also numerically investigated its behavior in the presence of temporally-correlated non-trivial error models.

The effective upshot of leakage and non-Markovian noise is that the signal will have more spectral content than what can be resolved given a chosen sequence length K, leading to unphysical features in the spectrum such as an eigenvalue estimate larger than 1, or the absence of a real eigenvalue. Even though we have seen in our examples that a physical spectrum can be regained by going to larger K, depending on the noise model, this convergence may be very slow requiring much data-taking time. Hence these unphysical features are useful markers for deviations from our model of repeated TPCP qubit maps $\mathscr{S}^k$. We view it as an open question how well one can reliably distinguish different sources of deviations.

7.6.1. LOGICAL SPECTRAL QUANTUM TOMOGRAPHY

An interesting application of the spectral tomography method could be the assessment of logical gates on encoded quantum information in a SPAM-resistant fashion. In this logical scenario (for, say, a single logical qubit), one first prepares the eigenstates of the logical Pauli operators $\overline{X}, \overline{Y}$ and $\overline{Z}$. One then applies a unit of error-correction $k = 0, \ldots, K$ times: a single unit could be, say, the repeated error correction for L rounds of a distance-L surface code. Or a unit is the application of a fault-tolerant logical gate, e.g. by means of code-deformed error correction or a transversal logical gate followed by a unit of error correction. After k units one measures the logical Pauli operators fault-tolerantly, and repeats experiments to obtain the logical signal $\overline{g}(k)$. Studying the spectral features of such logical channel will give information about the efficacy of the quantum error correction unit and/or the applied logical gate while departures from the code space or a need to time-correlate syndrome data beyond the given QEC unit can show up as leakage and non-Markovian errors.

7.7. METHODS

7.7.1. SINGLE-QUBIT CASE WITH NON-DIAGONALIZABLE MATRIX T

In general, a matrix T can be brought to Jordan normal form by a similarity transformation, i.e. $T = VJV^{-1}$ with $J = \oplus_i J_i$ where each Jordan block J_i is of the form

$$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}, \tag{7.21}$$

see e.g. Theorem 3.1.11 in Ref. [31]. T is diagonalizable when each Jordan block is fully diagonal.

An example of a non-diagonalizable Lindblad superoperator on a single qubit has been constructed in Ref. [32]. Using this, one can easily get a single-qubit superoperator $\mathscr{S}$ for which the traceless block of the Pauli transfer matrix is a non-diagonalizable matrix T as follows. Let $\mathscr{S}(\rho) = \exp(\mathscr{L}\epsilon)(\rho) \approx \rho + \epsilon\mathscr{L}(\rho) + O(\epsilon^2)$ with $\mathscr{L}(\rho) = -i[\frac{yZ}{2}, \rho] + \mathscr{D}[(2x)^{1/2}\sigma_-](\rho) + \mathscr{D}[y^{1/2}X](\rho)$ with $\mathscr{D}[A](\rho) = A\rho A^\dagger - \frac{1}{2}\{A^\dagger A, \rho\}$ and real parameters $x, y \geq 0$. This implies that $\mathscr{S}$ has the 4×4 Pauli transfer matrix

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-\epsilon x & -\epsilon y & 0 \\ 0 & \epsilon y & 1-\epsilon(x+2y) & 0 \\ 2\epsilon x & 0 & 0 & 1-2\epsilon(x+y) \end{pmatrix} + O(\epsilon^2).$$

Taking some small ϵ and $x \neq 0$, one can check that the submatrix T does not have 3 eigenvectors and it has a pair of degenerate eigenvalues, so T is not diagonalizable. When we take $x = 0$, $\mathscr{S}$ is unital, that is $\mathscr{S}(I) = I$, and the submatrix T is not diagonalizable either.

Even though a matrix T is not always diagonalizable, there still exists the so-called Schur triangular form for any matrix T [31]. This form says that $T = W(D+E)W^\dagger$ with W a unitary matrix, D a diagonal matrix with the eigenvalues of T, and E a strictly upper-triangular "nilpotent" matrix with non-zero entries only above the diagonal. Since the $N \times N$ matrix E is strictly upper-triangular, one has $\mathrm{Tr}\left[D^i E^j\right] = 0$ for all $j \neq 0$. If we use this form in Eq. (7.12), one obtains for any k

$$g^{\text{NO SPAM}}(k) = \mathrm{Tr}\left[T^k\right] = \mathrm{Tr}\left[(D+E)^k\right] = \mathrm{Tr}\left[D^k\right], \tag{7.22}$$

since any product of the form $D^{l_1}E^{l_2}D^{l_3}\ldots E^{l_m}$ with some non-zero $l_i > 0$ is a matrix with zeros on the diagonal. In case of SPAM errors and non-diagonalizable T we consider

$$g(k) = \mathrm{Tr}\left[W^\dagger T_{\text{prep}}\, T_{\text{meas}}\, W(D+E)^k\right], \tag{7.23}$$

where $W^\dagger T_{\text{prep}}\, T_{\text{meas}}\, W$ is not the identity matrix due SPAM errors, implying that $g(k)$ can depend on E and have a non-exponential dependence on k. Thus, in the special case of a non-diagonalizable matrix T, the signal $g(k)$ would not have the dependence on the eigenvalues as in Eq. (7.13).

In particular, we can examine the physically-interesting non-diagonalizable *Case 3* in Section 7.2.2 in this light, taking $h_y = h_z = 0$ and a critical $h_x^{\rm cr} = \frac{\Gamma_1 - \Gamma_2}{2}$. The dynamics of the Lindblad equation after time t induces a superoperator $\mathscr{S}_t$ which will have the following action on the Pauli operators:

$$\begin{aligned}
\mathscr{S}_t(X) &= \exp(-\Gamma_2 t)X, \\
\mathscr{S}_t(Y) &= \exp(-(\Gamma_1+\Gamma_2)t/2)\left[(1+th_x^{\rm cr})Y - h_x^{\rm cr}Z\right], \\
\mathscr{S}_t(Z) &= \exp(-(\Gamma_1+\Gamma_2)t/2)\left[h_x^{\rm cr}tY + (1-h_x^{\rm cr}t)Z\right].
\end{aligned}$$

Here we can note the linear dependence on t due to the system being critically damped. If we consider the signal $g(t) = \sum_\mu \mathrm{Tr}[P_\mu \mathscr{S}_t(P_\mu)]$ we see that this linear dependence on t drops out in accordance with Eq. (7.22), i.e. this trace only depends on the eigenvalues and has an exponential dependence on t. In the presence of SPAM errors, some of the linear dependence could still be observable for such critically-damped system. In addition, coefficients such as $c_{\mu\nu}(t) = \mathrm{Tr}[P_\mu \mathscr{S}_t(P_\nu)]$ can depend linearly on t, making such tomographic data less suitable to extract eigenvalue information.

7.7.2. UPPER BOUND ON THE ENTANGLEMENT FIDELITY WITH THE TARGETED GATE

In this section we show how to relate the eigenvalues of the Pauli transfer matrix of a TPCP map $\mathscr{S}$ to an upper bound on the entanglement fidelity (and hence the average gate fidelity) with the targeted unitary gate U. Naturally, one can only expect to obtain an upper bound on the gate fidelity, since the eigenvalues do not provide information about the eigenvectors of $\mathscr{S}$. If the actual eigenvectors deviate a lot from ideal, the actual gate fidelity could be very low, so one can certainly not derive a lower bound on the fidelity based on the eigenvalues.

Lemma 1. *Let the eigenvalues of the $N \times N$ matrix $T^{\mathscr{S}}$ be $\{\lambda_i\}_{i=1}^N$ with $N = d^2 - 1$ for a d-dimensional system. Let U be the targeted gate with eigenvalues $\{\lambda_i^{\rm ideal}\}_{i=1}^N$ and let there be permutation π of i-th eigenvalue λ_i which maximizes $|\sum_i \lambda_{\pi(i)}^* \lambda_i^{\rm ideal}|$ so that $0 \le \xi_{\max} = \max_\pi \frac{1}{N}|\sum_i \lambda_{\pi(i)}^* \lambda_i^{\rm ideal}| \le 1$. The entanglement fidelity $\mathscr{F}_{\rm ent}(U,\mathscr{S}) = \frac{1}{N+1}(1 + \mathrm{Tr}[T^{U^\dagger} T^{\mathscr{S}}])$ is upper bounded as*

$$\mathscr{F}_{\rm ent}(U,\mathscr{S}) \le \frac{1}{N+1}\left(1 + N\sqrt{u(\mathscr{S}) - \frac{\sum_j |\lambda_j|^2}{N}} + N\xi_{\max}\right), \tag{7.24}$$

where $u(\mathscr{S})$ is the unitarity of $\mathscr{S}$.

Proof. We write $T^{\mathscr{S}}$ in Schur triangular form as $T^{\mathscr{S}} = W(D^{\mathscr{S}} + E)W^\dagger$ with W a unitary matrix, $D^{\mathscr{S}}$ a diagonal matrix with the eigenvalues of $T^{\mathscr{S}}$, and E a strictly upper-triangular "error" matrix with non-zero entries only above the diagonal [31]. Using the Cauchy-Schwartz inequality one has

$$\begin{aligned}
\mathrm{Tr}[T^{U^\dagger} T^{\mathscr{S}}] \le \mathrm{Tr}[T^{U^\dagger} W D^{\mathscr{S}} W^\dagger] + \\
(\mathrm{Tr}[E^\dagger E])^{1/2} (\mathrm{Tr}[T^{U^\dagger} T^U])^{1/2}.
\end{aligned} \tag{7.25}$$

Note that for a unitary gate U, $T^{U^\dagger} = (T^U)^T = (T^U)^\dagger$ and $T^{U^\dagger}T^U = I$ implying that T is an orthogonal matrix with unit singular values. We thus have $(\mathrm{Tr}[T^{U^\dagger}T^U])^{1/2} = \sqrt{N}$. One has $\mathrm{Tr}[T^{\mathscr{S}^\dagger}T^{\mathscr{S}}] = \mathrm{Tr}[(D^{\mathscr{S}}+E)^\dagger(D^{\mathscr{S}}+E)] = \mathrm{Tr}[(D^{\mathscr{S}^\dagger}D^{\mathscr{S}}+E^\dagger E)]$, using the strict upper-triangularity of E. In other words, $\mathrm{Tr}[E^\dagger E] = \mathrm{Tr}[T^{\mathscr{S}^\dagger}T^{\mathscr{S}}] - \sum_i |\lambda_i|^2$ where λ_i are the eigenvalues of $T^{\mathscr{S}}$. Recognizing that $\frac{1}{N}\mathrm{Tr}[T^{\mathscr{S}^\dagger}T^{\mathscr{S}}] = u(\mathscr{S})$, we obtain an upper bound on the second term in Eq. (7.25).

Now let's upper bound the first term in Eq. (7.25) for unknown unitary W. Assume w.l.o.g. that T^U and $D^{\mathscr{S}}$ are diagonal in the same basis (the additional rotation between these eigenbases can be absorbed into W). Let $T^U = \sum_i \lambda_i^{\text{ideal}} P_i$ and $D^{\mathscr{S}} = \sum_i \lambda_i P_i$ with orthogonal projectors P_i and $\sum_i P_i = I$. Define the matrix M with entries $M_{ij} = \mathrm{Tr}[P_i W P_j W^\dagger]$. The matrix M is doubly-stochastic, since $\sum_i M_{ij} = 1 = \sum_j M_{ij}$ which implies that $M = \sum_m q_m \pi_m$ with $q_m \geq 0, \sum_m q_m = 1$ (Birkhoff-von Neumann theorem [31]) with permutation matrix π_m. With these facts and the convention $\langle i|\lambda^{\mathscr{S}}\rangle = \lambda_i$ we can bound

$$|\mathrm{Tr}[T^{U^\dagger}WD^{\mathscr{S}}W^\dagger]| \leq \sum_m q_m |\langle \lambda^{\text{ideal}}|\pi_m|\lambda^{\mathscr{S}}\rangle| \leq N\xi_{\max}.$$

These bounds together then lead to Eq. (7.24). □

An immediate corollary of Theorem 1 is

$$\mathscr{F}_{\text{ent}}(U,\mathscr{S}) \leq \frac{1}{N+1}\left(1 + N\sqrt{1 - \frac{\sum_j |\lambda_j|^2}{N}} + N\xi_{\max}\right), \tag{7.26}$$

since $u(\mathscr{S}) \leq 1$ for TPCP maps. However, this is in general not a very strong upper bound on the fidelity.

We can do better in the single-qubit case by realizing that there are strong relations between the singular values σ_i of $T^{\mathscr{S}}$ and the absolute values of the eigenvalues $|\lambda_i|$ of $T^{\mathscr{S}}$. Ordering both the singular values and the eigenvalue magnitudes in descending order, we have the following (weak Majorization) inequalities for arbitrary matrices

$$\prod_{i=1}^N \sigma_i = \prod_{i=1}^N |\lambda_i|, \tag{7.27}$$

$$\sum_{i=1}^F \sigma_i \geq \sum_{i=1}^F |\lambda_i|, \quad F \in \{1,\ldots,N-1\}. \tag{7.28}$$

For single-qubit channels we can also impose TPCP constraints to the singular values of the channel. In particular we have [33, Eq. (4)]

$$\sigma_i \leq 1, \ \ \forall i \in \{0,1,2,3\}, \tag{7.29}$$

$$\sigma_1 + \sigma_2 \leq 1 + \sigma_3. \tag{7.30}$$

Using these relations we can upper bound the unitarity of a single-qubit channel $\mathscr{S}$, given

its eigenvalues, using the optimization:

$$\begin{aligned} \underset{\sigma_1,\sigma_2,\sigma_3}{\text{minimize}} \quad & u(\mathcal{S}) = \frac{1}{3}(\sigma_1^2+\sigma_2^2+\sigma_3^2) \\ \text{subject to} \quad & \sigma_1\sigma_2\sigma_3 = |\lambda_1||\lambda_2||\lambda_3|, \\ & 1 \geq \sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0, \\ & \sigma_1+\sigma_2 \leq 1+\sigma_3, \\ & \sigma_1+\sigma_2 \geq |\lambda_1|+|\lambda_2|, \\ & \sigma_1+\sigma_2+\sigma_3 \geq |\lambda_1|+|\lambda_2|+|\lambda_3|. \end{aligned}$$

This is a non-convex optimization problem in three variables, for which a global minimum can be numerically computed given $\lambda_1, \lambda_2, \lambda_3$. This gives an upper bound on the unitarity of $\mathcal{S}$ and hence on the entanglement fidelity of $\mathcal{S}$ to the target unitary U. In the main text we use this optimization to give non-trivial upper bounds on the fidelities of single-qubit gates realized on superconducting chips and analyzed using the spectral tomography method.

7.7.3. FRAME MISMATCH ACCUMULATION

In Section 7.4 we noted that the data gathered for the CNOT gate cannot be explained by a model of a noisy TPCP map $\mathcal{S}$ repeated k times. Here we propose a simple coherent model that qualitatively reproduces the features observed in Fig. 7.3 and we call this the frame mismatch accumulation model. Let $\mathcal{S}_0$ be a TPCP map that is a good approximation of the targeted gate applied exactly once (in the main text this was the CNOT) and let V be some unitary. In the frame mismatch accumulation model we assume that k consecutive applications of the gate are equal to:

$$\mathcal{S}_k = \prod_{i=0}^{k} \left(V^\dagger\right)^i \mathcal{S}_0 V^i = (V^\dagger)^{k+1}(V\mathcal{S}_0)^k. \tag{7.31}$$

Intuitively, this can be interpreted as an increasing mismatch between the frame in which $\mathcal{S}_0$ was defined and the frame in which the gate was implemented at the i-th repetition, up to $i = k$.

We apply this model to a CNOT gate, choosing $\mathcal{S}_0$ to be an ideal CNOT gate and choosing $V = \exp(-i\frac{\theta}{2} I \otimes Y)$ with $\theta = 0.05$ deg. In the case of the cross-resonance CNOT gate performed on *ibmqx4*, this may correspond to an imperfect cancellation of the $I \otimes Y$ term [34]. In Fig. 7.6 we see that this example closely reproduces the eigenvalues shown in Fig. 7.3. At the same time, we note that the qualitative features observed in Fig. 7.6 do not depend on the choice of the rotation axis of V (for either qubit), as long as the rotation does not commute with $\mathcal{S}_0$ (which would leave the gate unaffected by the frame mismatch).

Experimental data gathered for Figs. 7.2 and 7.3, as well as an implementation of the matrix pencil algorithm can be found online at https://doi.org/10.5281/zenodo.2613856.

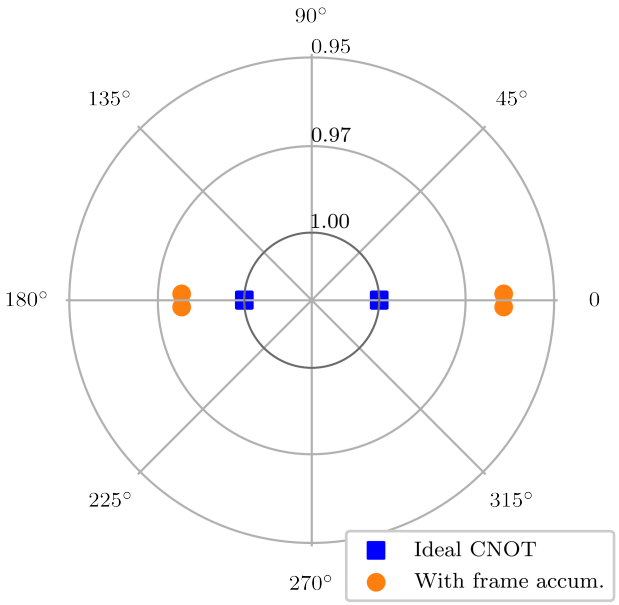

Figure 7.6: Spectral footprint of a simulated CNOT gate affected by frame mismatch accumulation, for $K = 50$. The shown eigenvalues are $\{0.9636+0.03276i, 0.9636-0.0327i, -0.9804+0.0495i, -0.9804-0.0495i\}$, qualitatively matching the experimentally-measured eigenvalues shown in Fig. 7.3 and, critically, matching the lack of real eigenvalues observed in Fig. 7.3.

REFERENCES

[1] J. Helsen, F. Battistel, and B. M. Terhal, *Spectral quantum tomography,* npj Quantum Information **5** (2019), 10.1038/s41534-019-0189-0.

[2] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, *Randomized benchmarking of quantum gates,* Phys. Rev. A **77**, 012307 (2008).

[3] E. Magesan, J. M. Gambetta, and J. Emerson, *Scalable and robust randomized benchmarking of quantum processes,* Phys. Rev. Lett. **106**, 180504 (2011).

[4] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, *Efficient measurement of quantum gate error by interleaved randomized benchmarking,* Phys. Rev. Lett. **109**, 080505 (2012).

[5] C. J. Wood and J. M. Gambetta, *Quantification and characterization of leakage errors,* Phys. Rev. A **97**, 032306 (2018).

[6] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, *Estimating the coherence of noise,* New Journal of Physics **17**, 113020 (2015).

[7] B. Dirkse, J. Helsen, and S. Wehner, *Efficient unitarity randomized benchmarking of few-qubit Clifford gates,* Phys. Rev. A **99**, 012315 (2019).

[8] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, *Characterizing large-scale quantum computers via cycle benchmarking,* Nature Communications **10**, 5347 (2019).

[9] E. Onorati, A. H. Werner, and J. Eisert, *Randomized benchmarking for individual quantum gates,* Phys. Rev. Lett. **123**, 060501 (2019).

[10] S. Kimmel, M. P. da Silva, C. A. Ryan, B. R. Johnson, and T. Ohki, *Robust extraction of tomographic information via randomized benchmarking,* Phys. Rev. X **4**, 011050 (2014).

[11] R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz, *Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography,* Nature Communications **8** (2017), 10.1038/ncomms14485.

[12] D. Greenbaum, *Introduction to quantum gate set tomography,* (2015), arXiv:1509.02921 [quant-ph] .

[13] T. E. O'Brien, B. Tarasinski, and B. M. Terhal, *Quantum phase estimation of multiple eigenvalues for small-scale (noisy) experiments,* New Journal of Physics **21**, 023022 (2019).

[14] R. S. Bennink and P. Lougovski, *Quantum process identification: a method for characterizing non-Markovian quantum dynamics,* New Journal of Physics **21**, 083013 (2019).

[15] M. M. Wolf and D. Perez-Garcia, *Assessing quantum dimensionality from observable dynamics,* Phys. Rev. Lett. **102**, 190504 (2009).

[16] A. Strikis, A. Datta, and G. C. Knee, *Quantum leakage detection using a model-independent dimension witness,* Phys. Rev. A **99**, 032328 (2019).

[17] L. Rudnicki, Z. Puchała, and K. Zyczkowski, *Gauge invariant information concerning quantum channels,* Quantum **2**, 60 (2018).

[18] M.-B. Ruskai, S. Szarek, and E. Werner, *An analysis of completely-positive trace-preserving maps on M2,* Linear Algebra and its Applications **347**, 159 (2002).

[19] M. Wolf, *Quantum channels and operations guided tour,* https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MichaelWolf/QChannelLecture.pdf.

[20] M. Horodecki, P. Horodecki, and R. Horodecki, *General teleportation channel, singlet fraction, and quasidistillation,* Phys. Rev. A **60**, 1888 (1999).

[21] S. Haroche and J.-M. Raimond, *Exploring the Quantum: Atoms, Cavities, and Photons* (Oxford Univ. Press, Oxford, 2006).

[22] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, 2000).

[23] T. K. Sarkar and O. Pereira, *Using the matrix pencil method to estimate the parameters of a sum of complex exponentials,* IEEE Antennas and Propagation Magazine **37**, 48 (1995).

[24] D. Potts and M. Tasche, *Parameter estimation for nonincreasing exponential sums by Prony-like methods,* Linear Algebra and its Applications **439**, 1024 (2013), https://doi.org/10.1016/j.laa.2012.10.036 .

[25] J. Helsen, J. J. Wallman, S. T. Flammia, and S. Wehner, *Multiqubit randomized benchmarking using few samples,* Phys. Rev. A **100**, 032304 (2019).

[26] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and J. M. Martinis, *Superconducting quantum circuits at the surface code threshold for fault tolerance,* Nature **508**, 500 (2014).

[27] X. Xue, T. F. Watson, J. Helsen, D. R. Ward, D. E. Savage, M. G. Lagally, S. N. Coppersmith, M. A. Eriksson, S. Wehner, and L. M. K. Vandersypen, *Benchmarking gate fidelities in a* Si/SiGe *two-qubit device,* Phys. Rev. X **9**, 021011 (2019).

[28] C. J. Ballance, T. P. Harty, N. M. Linke, M. A. Sepiol, and D. M. Lucas, *High-fidelity quantum logic gates using trapped-ion hyperfine qubits,* Phys. Rev. Lett. **117**, 060504 (2016).

[29] C.-F. J. W. *et al., Jackknife, bootstrap and other resampling methods in regression analysis,* The Annals of Statistics **14**, 1261 (1986).

[30] G. Seber and C. Wild, *Nonlinear regression* (John Wiley & Sons, Hoboken, NJ, 2003).

[31] R. A. Horn and C. R. Johnson, *Matrix Analysis* (cup, 1985).

[32] M. S. Sarandy and D. A. Lidar, *Adiabatic approximation in open quantum systems,* Phys. Rev. A **71**, 012331 (2005).

[33] M. M. Wolf and D. Perez-Garcia, *The inverse eigenvalue problem for quantum channels,* (2010), arXiv:1005.4545 [quant-ph] .

[34] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, *Procedure for systematically tuning up cross-talk in the cross-resonance gate,* Phys. Rev. A **93**, 060302 (2016).

8

LEAKAGE DETECTION FOR A TRANSMON-BASED SURFACE CODE

Leakage outside of the qubit computational subspace, present in many leading experimental platforms, constitutes a threatening error for quantum error correction (QEC) for qubits. We develop a leakage-detection scheme via Hidden Markov models (HMMs) for transmon-based implementations of the surface code. By performing realistic density-matrix simulations of the distance-3 surface code (Surface-17), we observe that leakage is sharply projected and leads to an increase in the surface-code defect probability of neighboring stabilizers. Together with the analog readout of the ancilla qubits, this increase enables the accurate detection of the time and location of leakage. We restore the logical error rate below the memory break-even point by post-selecting out leakage, discarding less than half of the data for the given noise parameters. Leakage detection via HMMs opens the prospect for near-term QEC demonstrations, targeted leakage reduction and leakage-aware decoding and is applicable to other experimental platforms.

This chapter has been published in npj Quantum Inf. **6**, 102 (2020) [1]. F. B. performed full-trajectory simulations and theoretical derivations and contributed to the development of the theoretical concepts presented. Furthermore, F. B. contributed extensively to the writing.

8.1. INTRODUCTION

Recent advances in qubit numbers [2–5], as well as operational [6–14] and measurement [15–17] fidelities have enabled leading quantum computing platforms, such as superconducting and trapped-ion processors, to target demonstrations of quantum error correction (QEC) [18–24] and quantum advantage [3, 25–27]. In particular, two-dimensional stabilizer codes, such as the surface code, are a promising approach [24, 28] towards achieving quantum fault tolerance and, ultimately, large-scale quantum computation [29]. One of the central assumptions of textbook QEC is that any error can be decomposed into a set of Pauli errors that act within the computational space of the qubit. In practice, many qubits such as weakly-anharmonic transmons, as well as hyperfine-level trapped ions, are many-level systems which function as qubits by restricting the interactions with the other excited states. Due to imprecise control [13, 30, 31] or the explicit use of non-computational states for operations [6, 7, 10, 12, 32–36], there exists a finite probability for information to leak from the computational subspace. Thus, leakage constitutes an error that falls outside of the domain of the qubit stabilizer formalism. Furthermore, leakage can last over many QEC cycles, despite having a finite duration set by the relaxation time [37]. Hence, leakage represents a menacing error source in the context of quantum error correction [18, 37–44], despite leakage probabilities per operation being smaller than readout, control or decoherence error probabilities [7, 9, 10, 45].

The presence of leakage errors has motivated investigations of its effect on the code performance and of strategies to mitigate it. A number of previous studies have focused on a stochastic depolarizing model of leakage [39, 41–44], allowing to explore large-distance surface codes and the reduction of the code threshold using simulations. These models, however, do not capture the full details of leakage, even though a specific adaptation has been used in the case of trapped-ion qubits [42–44]. Complementary studies have considered a physically realistic leakage model for transmons [37, 40], which was only applied to a small parity-check unit due to the computational cost of many-qutrit density-matrix simulations. In either case, leakage was found to have a strong impact on the performance of the code, resulting in the propagation of errors, in the increase of the logical error rate and in a reduction of the effective code distance. In order to mitigate these effects, there have been proposals for the introduction of leakage reduction units (LRUs) [38, 40, 41, 46] beyond the natural relaxation channel, for modifications to the decoding algorithms [18, 39, 41], as well as for the use of different codes altogether [43]. Many of these approaches rely on the detection of leakage or introduce an overhead in the execution of the code. Recently, the indirect detection of leakage in a 3-qubit parity-check experiment [21] was realized via a Hidden Markov Model (HMM), allowing for subsequent mitigation via post-selection. Given that current experimental platforms are within reach of quantum-memory demonstrations, detailed simulations employing realistic leakage models are vital for a comprehensive understanding of the effect of leakage on the code performance, as well as for the development of a strategy to detect leakage without additional overhead.

In this work we demonstrate the use of computationally efficient HMMs to detect leakage in a transmon implementation of the distance-3 surface code (Surface-17) [47]. Using full-density-matrix simulations [28, 48] we first show that repeated stabilizer measurements sharply project data qubits into the leakage subspace, justifying the use of

classical HMMs with only two hidden states (computational or leaked) for leakage detection. We observe a considerable increase in the surface-code defect probability of neighboring stabilizers while a data or ancilla qubit is leaked, a clear signal that may be detected by the HMMs. For ancilla qubits, we further consider the information available in the analog measurement outcomes, even when the leaked state $|2\rangle$ can be discriminated from the computational states $|0\rangle$ and $|1\rangle$ with limited fidelity. We demonstrate that a set of two-state HMMs, one HMM for each qubit, can accurately detect both the time and the location of a leakage event in the surface code. By post-selecting on the detected leakage, we restore the logical performance of Surface-17 below the memory break-even point, while discarding less than half of the data for the given error-model parameters. Finally, we outline a minimal set of conditions for our leakage-detection scheme to apply to other quantum-computing platforms. Although post-selection is not scalable due to an exponential overhead in the number of required experiments, these results open the prospect for near-term demonstrations of fault tolerance even in the presence of leakage. Furthermore, HMM-based leakage detection enables the possibility of scalable leakage-aware decoding [18, 41] and real-time targeted application of LRUs [38, 40, 41].

8.2. LEAKAGE ERROR MODEL

We develop an error model for leakage in superconducting transmons, for which two-qubit gates constitute the dominant source of leakage [6, 7, 10, 12, 13, 30–35], while single-qubit gates have negligible leakage probabilities [9, 45]. We thus focus on the former, while the latter is assumed to induce no leakage at all. We assume that single-qubit gates act on a leaked state as the identity. Measurement-induced leakage is also assumed to be negligible.

We use full-trajectory simulations to characterize leakage in the Net-Zero implementation (see Section 6.3) of the controlled-phase gate (CZ), considered as the native two-qubit gate in a transmon-based Surface-17, with experimentally targeted parameters (see Table 8.1 and Table 8.2). This gate uses a flux pulse such that the higher frequency qubit (Q_{flux}) is fluxed down from its sweetspot frequency ω_{max} to the vicinity of the interaction frequency $\omega_{\text{int}} = \omega_{\text{stat}} - \alpha$, where ω_{stat} is the frequency of the other qubit (Q_{stat}), which remains static, and α is the transmon anharmonicity. The inset in Fig. 8.1 **a** shows a schematic diagram of the frequency excursion taken by Q_{flux}. A (bipolar) 30 ns pulse tunes twice the qubit on resonance with the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing, corresponding to the interaction frequency ω_{int}. This pulse is followed by a pair of 10 ns single-qubit phase-correction pulses. The relevant crossings around ω_{int} are shown in Fig. 8.1 **a** and are all taken into account in the full-trajectory simulations. The two-qubit interactions give rise to population exchanges towards and within the leakage subspace and to the phases acquired during gates with leaked qubits, which we model as follows.

The model in Fig. 8.1 **b** considers a general CZ rotation, characterized by the two-qubit phase ϕ_{11} for state $|11\rangle$ and $\phi = 0$ for the other three computational states. The single-qubit relative phases ϕ_{01} and ϕ_{10} result from imperfections in the phase corrections. The conditional phase is defined as $\phi_{2Q} = \phi_{11} - \phi_{01} - \phi_{10} + \phi_{00}$, which for an ideal CZ is $\phi_{2Q} = \pi$. In this chapter, we assume $\phi_{00} = \phi_{01} = \phi_{10} = 0$ and $\phi_{2Q} = \phi_{11} = \pi$. We set $\phi_{02} = -\phi_{11}$ in the rotating frame of the qutrit, as it holds for flux-based gates [36].

Interactions between leaked and non-leaked qubits lead to extra phases, which we call

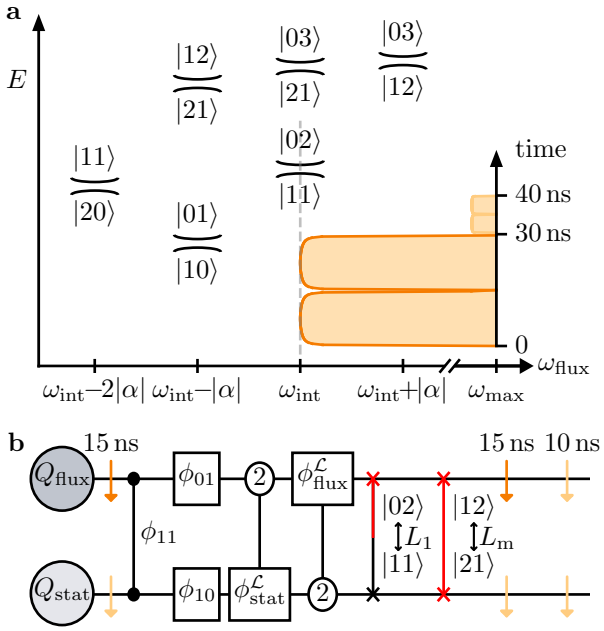

Figure 8.1: Schematic of the relevant interactions and the CZ error model for two transmons, a higher frequency one Q_{flux} and a lower frequency one Q_{stat}. The inset of **a** shows the frequency excursion taken by Q_{flux} from its sweetspot frequency ω_{max} to the interaction frequency ω_{int}, corresponding to the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing, followed by weaker single-qubit phase-correction pulses. During this excursion, the frequency ω_{stat} of Q_{stat} remains static at $\omega_{\text{stat}} = \omega_{\text{int}} - |\alpha|$, where α is the anharmonicity. **a** Sketch of all the considered avoided crossings, with the two-qubit system energy E on the vertical axis versus the frequency ω_{flux} of Q_{flux} on the horizontal axis. **b** The parametrized CZ error model. An ideal CZ is constructed with the two-qubit phase ϕ_{11} and the single-qubit phases ϕ_{01} and ϕ_{10}. It is followed by single-qubit rotations with phases $\phi^{\mathscr{L}}_{\text{flux}}$ and $\phi^{\mathscr{L}}_{\text{stat}}$, conditioned on the other transmon being leaked, and by the SWAP-like exchanges with leakage probability L_1 and leakage-mobility probability L_{m} (see Section 8.2 for precise definitions). Relaxation and decoherence, indicated by the orange arrows, are taken into account as well.

leakage conditional phases. We consider first the interaction between a leaked Q_{flux} and a non-leaked Q_{stat}. In this case the gate restricted to the $\{|02\rangle, |12\rangle\}$ subspace has the effect $diag\left(e^{i\phi_{02}}, e^{i\phi_{12}}\right)$, which up to a global phase corresponds to a Z rotation on Q_{stat} with an angle given by the leakage conditional phase $\phi^{\mathscr{L}}_{\text{stat}} := \phi_{02} - \phi_{12}$. Similarly, if Q_{stat} is leaked, then Q_{flux} acquires a leakage conditional phase $\phi^{\mathscr{L}}_{\text{flux}} := \phi_{20} - \phi_{21}$. These rotations are generally non-trivial, i.e., $\phi^{\mathscr{L}}_{\text{stat}} \neq \pi$ and $\phi^{\mathscr{L}}_{\text{flux}} \neq 0$, due to the interactions in the 3-excitation manifold which cause a shift in the energy of $|12\rangle$ and $|21\rangle$ (see Section 8.11.6). If the only interaction leading to non-trivial $\phi^{\mathscr{L}}_{\text{stat}}$, $\phi^{\mathscr{L}}_{\text{flux}}$ is the interaction between $|12\rangle$ and $|21\rangle$, then it can be expected that $\phi_{12} = -\phi_{21}$ in the rotating frame of the qutrit, leading to $\phi^{\mathscr{L}}_{\text{stat}} = \pi - \phi^{\mathscr{L}}_{\text{flux}}$.

Leakage is modeled as an exchange between $|11\rangle$ and $|02\rangle$, i.e., $|11\rangle \mapsto \sqrt{1-4L_1}\,|11\rangle + e^{i\phi}\sqrt{4L_1}\,|02\rangle$ and $|02\rangle \mapsto -e^{-i\phi}\sqrt{4L_1}\,|11\rangle + \sqrt{1-4L_1}\,|02\rangle$, with L_1 the leakage probability [49]. We observe that the phase ϕ and the off-diagonal elements $|11\rangle\langle 02|$ and $|02\rangle\langle 11|$ do not affect the results presented in this work, so we set them to 0 for computational efficiency (see Section 8.10.2). The SWAP-like exchange between $|12\rangle$ and $|21\rangle$ with probability L_{m}, which we call leakage mobility, as well as the possibility of further leaking to $|3\rangle$, are analyzed in Section 8.11.6.

The described operations are implemented as instantaneous in the *quantumsim* package (introduced in Ref. [48]), while the amplitude and phase damping experienced by the transmon during the application of the gate are symmetrically introduced around them, indicated by light-orange arrows in Fig. 8.1 **b**. The dark-orange arrows indicate the increased dephasing rate of Q_{flux} far away from ω_{max} during the Net-Zero pulse. The error parameters considered in this work are summarized in Section 8.10.2. In particular, unless otherwise stated, L_1 is set to 0.125% and $\phi^{\mathscr{L}}_{\text{flux}}$ and $\phi^{\mathscr{L}}_{\text{stat}}$ are randomized for each qubit pair across different batches consisting of 2×10^4 or 4×10^4 runs of 20 or 50 QEC cycles, respectively. This choice is motivated by our expectation that these phases are determined by the frequencies and anharmonicities of the two transmons as well as by the parameterization of the flux pulse implementing each CZ between the pair, which is fixed when tuning the gate experimentally. Since $\phi^{\mathscr{L}}_{\text{flux}}$ and $\phi^{\mathscr{L}}_{\text{stat}}$ have not been characterized in experiment, we instead choose to randomized them in order to capture an average behavior.

Some potential errors are found to be small from the full-trajectory simulations of the CZ gate and thus are not included in the parametrized error model. The population exchange between $|01\rangle \leftrightarrow |10\rangle$, with coupling J_1, is suppressed ($< 0.5\%$) since this avoided crossing is off-resonant by one anharmonicity α with respect to ω_{int}. While $|12\rangle \leftrightarrow |21\rangle$ is also off-resonant by α, the coupling between these two levels is stronger by a factor of 2, hence potentially leading to a larger population exchange (see Section 8.11.6). The $|11\rangle \leftrightarrow |20\rangle$ crossing is 2α away from ω_{int} and it thus does not give any substantial phase accumulation or population exchange ($< 0.1\%$). We have compared the average gate fidelity of CZ gates simulated with the two methods and found differences below $\pm 0.1\%$, demonstrating the accuracy of the parametrized model.

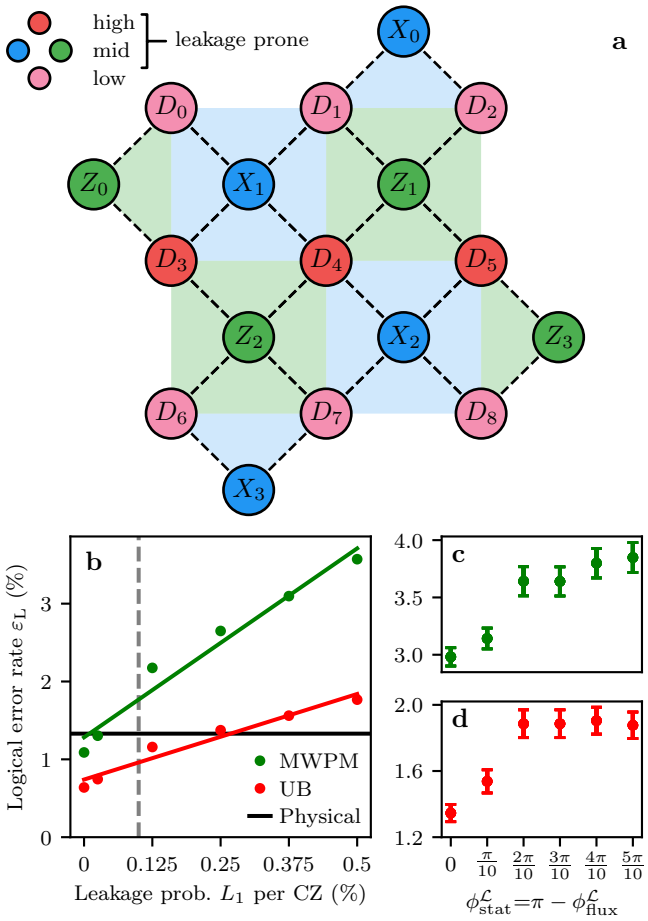

Figure 8.2: **a** Schematic overview of the Surface-17 layout [47]. Pink (resp. red) circles with D labels represent low- (high-) frequency data qubits, while blue (resp. green) circles with X (Z) labels represent ancilla qubits of intermediate frequency, performing X-type (Z-type) parity checks. **b** Dependence of the logical error rate ε_L on the leakage probability L_1 for a MWPM decoder (green) and for the decoding upper bound (red). The black solid line shows the physical error rate of a single transmon qubit. The dashed line corresponds to the recently achieved L_1 in experiment (see Section 6.8). Logical error rate ε_L for MWPM (**c**) and upper bound (**d**) as a function of the leakage conditional phases $\phi^{\mathcal{L}}_{\text{flux}}$ and $\phi^{\mathcal{L}}_{\text{stat}}$ (for $L_1 = 0.5\%$). Here, these phases are not randomized but fixed to the given values across all runs. The logical error rates are extracted from an exponential fit of the logical fidelity over 20 QEC cycles and averaged over 5 batches of 2×10^4 runs for **b** and one batch of 2×10^4 runs for **c,d**. Error bars correspond to 2 standard deviations estimated by bootstrapping (not included in **b** due the error bars being smaller than the symbol size).

8.3. EFFECT OF LEAKAGE ON THE CODE PERFORMANCE

We implement density-matrix simulations [48] to study the effect of leakage in Surface-17 (Fig. 8.2). We follow the frequency arrangement and operation scheduling proposed in Ref. [47], which employs three qubit frequencies for the surface-code lattice, arranged as shown in Fig. 8.2 **a**. The CZ gates are performed between the high-mid and mid-low qubit pairs, with the higher frequency qubit of the pair taking the role of Q_{flux} (see Fig. 8.1). Based on the leakage model in Section 8.2, only the high and mid frequency qubits are prone to leakage (assuming no leakage mobility). Thus, in the simulation those qubits are included as three-level systems, while the low-frequency ones are kept as qubits. Ancilla-qubit measurements are modeled as projective in the $\{|0\rangle, |1\rangle, |2\rangle\}$ basis and ancilla qubits are not reset between QEC cycles. As a consequence, given the ancilla-qubit measurement $m[n]$ at QEC cycle n, the syndrome is given by $m[n] \oplus m[n-1]$ and the surface-code defect $d[n]$ by $d[n] = m[n] \oplus m[n-2]$. For the computation of the syndrome and defect bits we assume that a measurement outcome $m[n] = 2$ is declared as $m[n] = 1$. The occurrence of an error is signaled by $d[n] = 1$. To pair defects we use a minimum-weight perfect-matching (MWPM) decoder, whose weights are trained on simulated data without leakage [28, 50] and we benchmark its logical performance in the presence of leakage errors. The logical qubit is initialized in $|0\rangle_{\text{L}}$ and the logical fidelity is calculated at each QEC cycle, from which the logical error rate ε_{L} can be extracted [28].

Figure 8.2 **b** shows that the logical error rate ε_{L} is sharply pushed above the memory break-even point by leakage. We compare the MWPM decoder to the decoding upper bound (UB), which uses the complete density-matrix information to infer a logical error. A strong increase in ε_{L} is observed for this decoder as well. Furthermore, the logical error rate has a dependence on the leakage conditional phases for both decoders, as shown in Fig. 8.2 **c,d**.

8.4. PROJECTION AND SIGNATURES OF LEAKAGE

We now characterize leakage in Surface-17 and the effect that a leaked qubit has on its neighboring qubits. From the density matrix (DM), we extract the probability $p^{\mathscr{L}}_{\text{DM}}(Q) = \mathbb{P}(Q \in \mathscr{L}) = \langle 2|\rho_Q|2\rangle$ of a qubit Q being in the leakage subspace $\mathscr{L}$ at the end of a QEC cycle, after the ancilla-qubit measurements, where ρ_Q is the reduced density matrix of Q.

In the case of data-qubit leakage, $p^{\mathscr{L}}_{\text{DM}}(Q)$ sharply rises to values near unity, where it remains for a finite number of QEC cycles (on average 16 QEC cycles for the considered parameters, given in Table 8.1). We refer to this sharp increase of $p^{\mathscr{L}}_{\text{DM}}(Q)$ as projection of leakage. An example showing this projective behavior (in the case of qubit D_4), as observed from the density-matrix simulations, is reported in Fig. 8.3 **a**. This is the typical behavior of leakage, as shown in Fig. 8.3 **b** by the bi-modal density distribution of the probabilities $p^{\mathscr{L}}_{\text{DM}}(Q)$ for all the high-frequency data qubits Q. As data-qubit leakage is associated with defects on the neighboring ancilla qubits (due to the use of the $|02\rangle \leftrightarrow |11\rangle$ crossing by the CZ gates) and with the further propagation of defects in the following QEC cycles (as shown below), we attribute the observed projection to a back-action effect of the repetitive stabilizer measurements (see Section 8.11.3). Given this projective behavior, we identify individual events by introducing a threshold $p^{\mathscr{L}}_{\text{th}}(Q)$, above which a qubit is considered as leaked. Here we focus on leakage on D_4, sketched in Fig. 8.3 **c**.

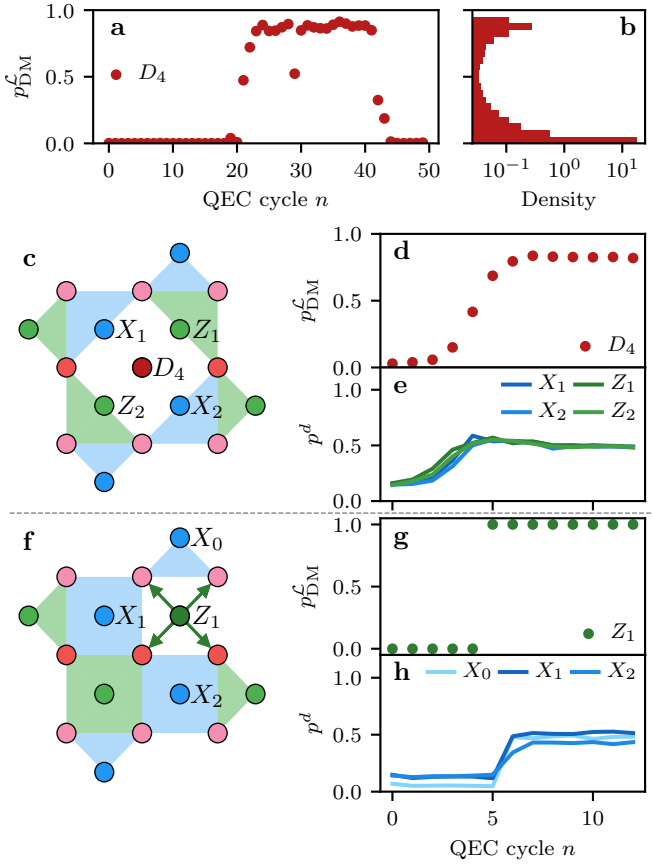

Figure 8.3: Projection and signatures of data-qubit leakage (**a**-**e**) and ancilla-qubit leakage (**f**-**h**). **a** Example realization of a data-qubit leakage event, extracted from the density-matrix simulations. **b** Density histogram of all data-qubit leakage probabilities over 20 bins, extracted over 4×10^4 runs of 50 QEC cycles each. **c**-**e** Signatures of data-qubit leakage. **c** Sketch of how leakage on a data qubit, e.g. D_4, alters the interactions with neighboring stabilizers, leading to their anti-commutation (see Section 8.11.2). **d** The average projection of the leakage probability $p^{\mathcal{L}}_{\mathrm{DM}}$ of D_4 over all events, where this probability is first below and then above a threshold of $p^{\mathcal{L}}_{\mathrm{th}} = 0.5$ for at least 5 and 8 QEC cycles, respectively. **e** The average number of defects on the neighboring stabilizers of D_4 over the selected rounds, showing a jump when leakage rises above $p^{\mathcal{L}}_{\mathrm{th}}$. **f**-**h** Signatures of ancilla-qubit leakage. **f** Sketch of how leakage on an ancilla qubit, e.g. Z_1, effectively disables the stabilizer check and probabilistically introduces errors on the neighboring data qubits. **g** We select realizations where Z_1 was in the computational subspace for at least 5 QEC cycles, after which it was projected into $|2\rangle$ by the readout and remained in that state for at least 5 QEC cycles. **h** The corresponding defect rate on neighboring stabilizers during the period of leakage. The error bars, which were estimated by bootstrapping, are smaller than the symbol sizes.

Based on a threshold $p^{\mathscr{L}}_{\text{th}}(D_4) = 0.5$, we select leakage events and extract the average dynamics shown in Fig. 8.3 **d**. Leakage grows over roughly 3 QEC cycles following a logistic function, reaching a maximum probability of approximately 0.8. We observe this behavior for all three high-frequency data qubits D_3, D_4, D_5.

We observe an increase in the defect probability of the neighboring ancilla qubits during data-qubit leakage. We extract the probability p^d of observing a defect $d = 1$ on the neighboring stabilizers during the selected data-qubit leakage events, as shown in Fig. 8.3 **e**. As $p^{\mathscr{L}}_{\text{DM}}(D_4)$ reaches its maximum, p^d goes to a constant value of approximately 0.5. This can be explained by data-qubit leakage reducing the stabilizer checks it is involved in to effective weight-3 anti-commuting checks, illustrated in Fig. 8.3 **c** and as observed in Ref. [21]. This anti-commutation leads to some of the increase in ε_{L} for the MWPM and UB decoders in Fig. 8.2 **b**. Furthermore, we attribute the observed sharp projection of leakage (see Fig. 8.3 **d**) to a back-action effect of the measurements of the neighboring stabilizers, whose outcomes are nearly randomized when the qubit is leaked (see Section 8.11.2 and Section 8.11.3). The weight-3 checks can also be interpreted as gauge operators, whose pairwise product results in weight-6 stabilizer checks, which can be used for decoding [51–54], effectively reducing the code distance from 3 to 2.

We also find a local increase in the defect probability during ancilla-qubit leakage. For ancilla qubits, $p^{\mathscr{L}}_{\text{DM}}$ is defined as the leakage probability after the ancilla projection during measurement. Since in the simulations ancilla qubits are fully projected, $p^{\mathscr{L}}_{\text{DM}}(Q) = 0, 1$ for an ancilla qubit Q, allowing to directly obtain the individual leakage events, as shown in Fig. 8.3 **g**. We note that an ancilla qubit remains leaked for 17 QEC cycles on average for the considered parameters (given in Table 8.1). We extract p^d during the selected events, as shown in Fig. 8.3 **h**. In the QEC cycle after the ancilla qubit leaks, p^d abruptly rises to a high constant value. We attribute this to the Z rotations acquired by the neighboring data qubits during interactions with the leaked ancilla qubit, as sketched in Fig. 8.3 **f** and described in Section 8.2. The angle of rotation is determined by $\phi^{\mathscr{L}}_{\text{flux}}$ or $\phi^{\mathscr{L}}_{\text{stat}}$, depending on whether the leaked ancilla qubit takes the roles of Q_{stat} or Q_{flux}, respectively (see Section 8.10.1 for the scheduling of operations). In the case of Z-type parity checks, these phase errors are detected by the X-type stabilizers. In the case of X-type checks, the phase errors on data qubits are converted to bit-flip errors by the Hadamard gates applied on the data qubits, making them detectable by the Z-type stabilizers. Furthermore, while the ancilla qubit is leaked, the corresponding stabilizer measurement does not detect any errors on the neighboring data qubits, effectively disabling the stabilizer, as sketched in Fig. 8.3 **f**. This, combined with the spread of errors, defines the signature of ancilla-qubit leakage and explains part of the observed increase in ε_{L} for the MWPM and UB decoders in Fig. 8.2 **b**.

For both data and ancilla qubits, a leakage event is correlated with a local increase in the defect rate, albeit due to different mechanisms. However, interpreting the spread of defects as signatures of leakage suggests the possible inversion of the problem, allowing for effective leakage detection.

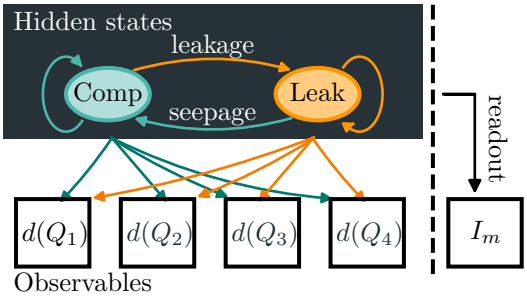

Figure 8.4: Schematic representation of an HMM for leakage detection. For both ancilla and data qubits only two hidden states are considered, corresponding to the qubit being either in the computational (teal) or leakage subspace (orange). Transitions between these states occur each QEC cycle, depending on the leakage and seepage probabilities. The state-dependent observables are the defects $d(Q)$ on the neighboring stabilizers. For ancilla qubits, the in-phase component I_m of the analog measurement is also used as an observable.

8.5. HIDDEN MARKOV MODELS

We use a set of HMMs, one HMM for each leakage-prone qubit, to detect leakage. This approach is similar to what recently demonstrated in a 3-qubit parity-check experiment [21], but we use simpler HMMs to make them computationally efficient. In general, an HMM (see Fig. 8.4 and Section 8.10.3) models the time evolution of a discrete set of hidden states, the transitions between which are assumed to be Markovian. At each time step a set of observable bits is generated with state-dependent emission probabilities. Depending on the observed outcomes, the HMM performs a Bayesian update of the predicted probability distribution over the hidden states.

We apply the concept of HMMs to leakage inference and outline their applicability for an accurate, scalable and run-time executable leakage-detection scheme. This is made possible by two observations. The first is that both data- and ancilla-qubit leakage are sharply projected (see Section 8.4) to high $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$. This justifies the use of classical HMMs with only two hidden states, corresponding to the qubit being in the computational or leakage subspace.

The second observation is the sharp local increase in p^d associated with leakage (see Section 8.4), which we identify as the signature of leakage. This allows us to consider only the defects on the neighboring stabilizers as relevant observables and to neglect correlations between pairs of defects associated with qubit errors. In the case of ancilla-qubit leakage, in addition to the defects, we consider the state information obtained from the analog measurement as input to the HMMs. Each transmon is dispersively coupled to a dedicated readout resonator. The state-dependent shift in the single-shot readout produces an output voltage signal, with in-phase and quadrature components (see Section 8.11.1).

The transition probabilities between the two hidden states are determined by the leakage and seepage probabilities per QEC cycle, which are, to lowest order, a function only of the leakage probability L_1 per CZ gate and of the relaxation time T_1 (see Section 8.10.3). We extract the state-dependent emission probabilities from simulation. When a qubit is not leaked, the probability of observing a defect on each of the neighboring stabilizers is determined by regular errors. When a data qubit is leaked, the defect probability is fixed to

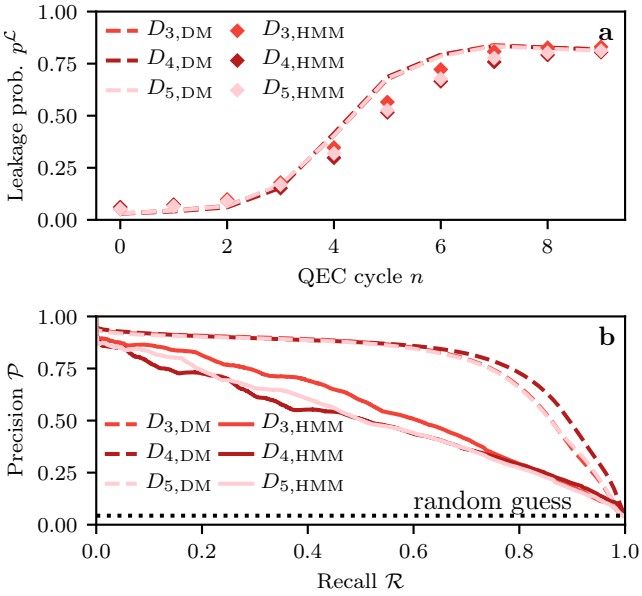

Figure 8.5: **a** Average response in time of the HMMs (diamonds) to leakage, in comparison to the actual leakage probability extracted from the density-matrix simulations (dashed lines). The average is computed by selecting single realizations where $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ was below a threshold $p^{\mathscr{L}}_{\mathrm{th}} = 0.5$ for at least 5 QEC cycles and then above it for 5 or more rounds. Error bars, estimated by bootstrapping, are smaller than the symbol sizes. **b** Precision-recall curves for the data qubits over 4×10^4 runs of 50 QEC cycles each using the HMM predictions (solid) and the leakage probability from the density matrix (dashed). The dotted line corresponds to a random guess classifier for which $\mathscr{P}$ is equal to the fraction of leakage events (occurring with probability given by the density matrix) over all QEC cycles and runs.

a nearly constant value by the stabilizer anti-commutation, while when an ancilla qubit is leaked, this probability depends on $\phi^{\mathscr{L}}_{\mathrm{flux}}$ and $\phi^{\mathscr{L}}_{\mathrm{stat}}$. Furthermore, the analog measurement outcome can be used to extract a probability of the transmon being in $|0\rangle$, $|1\rangle$ or $|2\rangle$ using a calibrated measurement (see Section 8.7 and Section 8.11.1).

8.6. DATA-QUBIT LEAKAGE DETECTION

We assess the ability of the data-qubit HMMs to accurately detect both the time and the location of a leakage event. We recall that these HMMs take the defects on neighboring stabilizers as input. The average temporal response $p^{\mathscr{L}}_{\mathrm{HMM}}(Q)$ of the HMMs to an event is shown in Fig. 8.5 and compared to the leakage probabilities $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ extracted from the density-matrix simulation. Events are selected when crossing a threshold $p^{\mathscr{L}}_{\mathrm{th}}$, as described in Section 8.4, and the response is averaged over these events. For the data-qubit HMMs, the response $p^{\mathscr{L}}_{\mathrm{HMM}}(Q)$ closely follows the probability $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ from the density matrix, reaching the same maximum leakage probability but with a smaller logistic growth rate. This slightly slower response is expected to translate to an average delay of about 1 QEC cycles in the detection of leakage.

We now explore the leakage-detection capability of the HMMs. The precision $\mathscr{P}$ of

an HMM tracking leakage on a qubit Q is defined as

$$\mathscr{P}_{\mathrm{HMM}}(Q) = \mathbb{P}\left(Q \in \mathscr{L} \mid p^{\mathscr{L}}_{\mathrm{HMM}}(Q) > p^{\mathscr{L}}_{\mathrm{th}}(Q)\right) \tag{8.1}$$

and can be computed as

$$\mathscr{P}_{\mathrm{HMM}}(Q) = \frac{\sum_i p^{\mathscr{L}}_{\mathrm{DM}}(Q,i)\,\theta\left[p^{\mathscr{L}}_{\mathrm{HMM}}(Q,i) - p^{\mathscr{L}}_{\mathrm{th}}(Q)\right]}{\sum_i \theta\left[p^{\mathscr{L}}_{\mathrm{HMM}}(Q,i) - p^{\mathscr{L}}_{\mathrm{th}}(Q)\right]}, \tag{8.2}$$

where i runs over all runs and QEC cycles and θ is the Heaviside step function. The precision is then the fraction of correctly identified leakage events (occurring with probability given by the density matrix), over all of the HMM detections of leakage. The recall $\mathscr{R}$ of an HMM for a qubit Q is defined as

$$\mathscr{R}_{\mathrm{HMM}}(Q) = \mathbb{P}\left(p^{\mathscr{L}}_{\mathrm{HMM}}(Q) > p^{\mathscr{L}}_{\mathrm{th}}(Q) \mid Q \in \mathscr{L}\right), \tag{8.3}$$

and can be computed as

$$\mathscr{R}_{\mathrm{HMM}}(Q) = \frac{\sum_i p^{\mathscr{L}}_{\mathrm{DM}}(Q,i)\,\theta\left[p^{\mathscr{L}}_{\mathrm{HMM}}(Q,i) - p^{\mathscr{L}}_{\mathrm{th}}(Q)\right]}{\sum_i p^{\mathscr{L}}_{\mathrm{DM}}(Q,i)}. \tag{8.4}$$

The recall is the fraction of detected leakage by the HMM over all leakage events (occurring with probability given by the density matrix). The precision-recall (PR) of an HMM (see Fig. 8.5 **b**) is a parametric curve obtained by sweeping $p^{\mathscr{L}}_{\mathrm{th}}(Q)$ and plotting the value of $\mathscr{P}$ and $\mathscr{R}$. Since the PR curve is constructed from $p^{\mathscr{L}}_{\mathrm{HMM}}(Q)$ over all QEC cycles and runs, it quantifies the detection ability in both time and space. The detection ability of an HMM manifests itself as a shift of the PR curve towards higher values of $\mathscr{P}$ and $\mathscr{R}$ simultaneously. We define the optimality $\mathscr{O}(Q)$ of the HMM corresponding to qubit Q as

$$\mathscr{O}(Q) = \mathrm{AUC}_{\mathrm{HMM}}(Q)\,/\mathrm{AUC}_{\mathrm{DM}}(Q), \tag{8.5}$$

where $\mathrm{AUC}_{\mathrm{HMM}}(Q)$ is the area under the PR curve of the HMM and $\mathrm{AUC}_{\mathrm{DM}}(Q)$ is the area for the optimal model that predicts leakage with probability $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$, achieving the best possible $\mathscr{P}_{\mathrm{DM}}$ and $\mathscr{R}_{\mathrm{DM}}$. An average optimality of $\mathscr{O}(Q) \approx 67.0\%$ is extracted for the data-qubit HMMs. Given the few QEC-cycle delay in the data-qubit HMM response to leakage, the main limitation to the observed HMM optimality $\mathscr{O}(Q)$ is false detection when a neighboring qubit is leaked (see Section 8.11.4).

8.7. ANCILLA-QUBIT LEAKAGE DETECTION

We now assess the ability of the ancilla-qubit HMMs to accurately detect both the time and the location of a leakage event. These HMMs take as observables the defects on the neighboring stabilizers at each QEC cycle as well as the analog measurement outcome of the ancilla qubit itself.

We first consider the case when the HMMs rely only on the increase in the defect probability p^d and show their PR curves in Fig. 8.6 **a,b**. Given that projective measurements are used in the simulations, $\mathrm{AUC}_{\mathrm{DM}}(Q) = 1$ for ancilla qubits. Bulk ancilla qubits

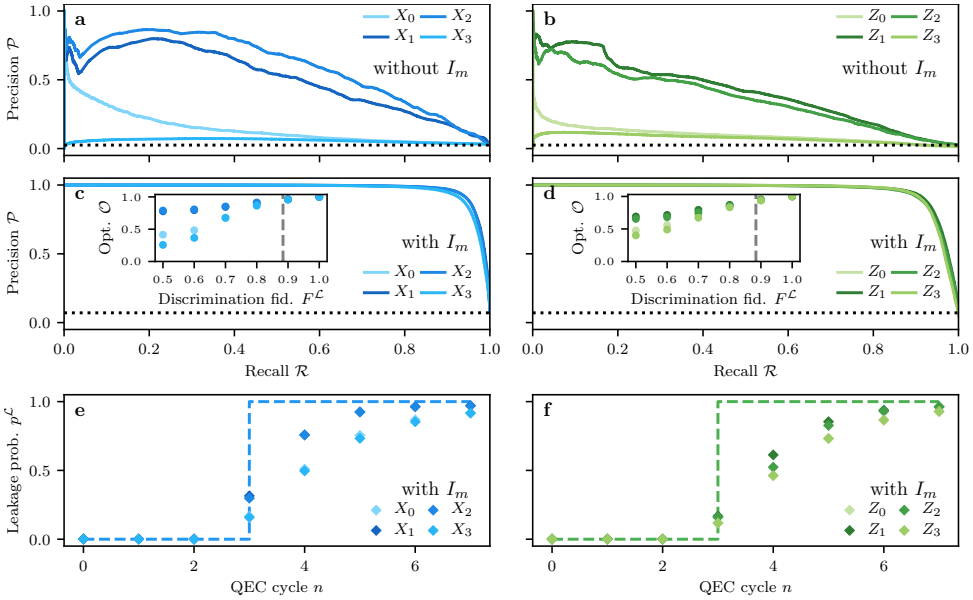

Figure 8.6: **a-d** Precision-recall curves for the ancilla-qubit HMMs over 4×10^4 runs of 50 QEC cycles each. In **a,b** the HMMs rely only on the observed defects on the neighboring stabilizers. In **c-f** the HMMs further get the in-phase component I_m of the analog readout as input, from which $p_m^{\mathcal{L}}$ is extracted. The dotted line corresponds to a random guess classifier for which $\mathcal{P}$ is equal to the fraction of leakage events over all QEC cycles and runs. As ancilla-qubit leakage is directly measured, $\mathcal{P}_{\mathrm{DM}} = 1$ for all values of $\mathcal{R}$ (not shown). Insets in **c,d**: the HMM optimality $\mathcal{O}$ as a function of the discrimination fidelity $F^{\mathcal{L}}$ between $|1\rangle$ and $|2\rangle$. The corresponding error bars (extracted over 2×10^4 runs of 20 QEC cycles each) are smaller than the symbol size. The vertical dashed line corresponds to the experimentally measured $F^{\mathcal{L}} = 88.4\%$. **e,f** Average response in time of the ancilla-qubit HMMs (diamonds) to leakage, in comparison to the actual leakage probability extracted directly from the readout (dashed), extracted over 4×10^4 runs of 50 QEC cycles each. The average is computed by selecting single realizations where the qubit was in the computational subspace for at least 3 QEC cycles and then in the leakage subspace for 5 or more.

have a moderate $\mathcal{O}(Q) \approx 47\%$, while boundary ancilla qubits possess nearly no ability to detect leakage. We attribute this to the boundary ancilla qubits having only a single neighboring stabilizer, compared to bulk ancilla qubits having 3 in Surface-17. The HMMs corresponding to pairs of same-type (X or Z) bulk ancilla qubits exhibit visibly different PR curves (see Fig. 8.6 **a,b**), despite the apparent symmetry of Surface-17. This symmetry is broken by the use of high- and low-frequency transmons as data qubits, leading to differences in the order in which an ancilla qubit interacts with its neighboring data qubits (see Ref. [47] and Fig. 8.8), together with the fact that CZs with $L_1 \neq 0$ do not commute in general. In addition to a low $\mathcal{O}(Q)$, the errors propagated by the leaked ancilla qubits (and hence the signatures of ancilla-qubit leakage) depend on $\phi^{\mathscr{L}}_{\text{stat}}$ and $\phi^{\mathscr{L}}_{\text{flux}}$ (randomized in the simulations). The values of these phases generally lead to different p^d than the ones parameterizing the HMM. The latter is extracted based on the average p^d observed over the runs (see Section 8.10.3). In the worst-case (for leakage detection), these phases can lead to no errors being propagated onto the neighboring data qubits, resulting in the undetectability of leakage. The mismatch between the fluctuating p^d (over $\phi^{\mathscr{L}}_{\text{stat}}$ and $\phi^{\mathscr{L}}_{\text{flux}}$) and the average value hinders the ability of the ancilla-qubit HMMs to detect leakage. Even if these phases were individually controllable, tuning them to maximize the detection capability of the HMMs would also lead to an undesirable increase in ε_{L} of a (leakage-unaware) decoder (see Fig. 8.2).

To alleviate these issues, we consider the state-dependent information obtained from the analog measurement outcome. The discrimination fidelity between $|1\rangle$ and $|2\rangle$ is defined as

$$F^{\mathscr{L}} = 1 - \frac{\mathbb{P}(1 \mid 2) + \mathbb{P}(2 \mid 1)}{2}, \tag{8.6}$$

where $\mathbb{P}(i \mid j)$ is the conditional probability of declaring the measurement outcome i given that the qubit has been prepared in state $|j\rangle$, assuming that no excitation or relaxation occur during the measurement (accounted for in post-processing). Here we assume that $\mathbb{P}(0 \mid 2) = \mathbb{P}(2 \mid 0) = 0$, as observed in experiment (see Fig. 8.9). We focus on the discrimination fidelity as in our simulations relaxation is already accounted for in the measurement outcomes (see Section 8.10.1). We extract $F^{\mathscr{L}}$ from recent experimental data [21], where the readout pulse was only optimized to discriminate between the computational states. By taking the in-phase component of the analog measurement, for each state $|j\rangle$ a Gaussian distribution $\mathcal{N}_j$ is obtained, from which we get $F^{\mathscr{L}} = 88.4\%$ (see Section 8.11.1).

In order to emulate the analog measurement in simulation, given an ancilla-qubit measurement outcome $m \in \{0,1,2\}$, we sample the in-phase response I_m from the corresponding distribution $\mathcal{N}_m$. The probability of the ancilla qubit being leaked given I_m is computed as

$$p^{\mathscr{L}}_m = \frac{\mathcal{N}_2(I_m)}{\sum_{j\in\{0,1,2\}} \mathcal{N}_j(I_m)}. \tag{8.7}$$

The ancilla-qubit HMMs use the sampled responses I_m, in combination with the observed defects, to detect leakage.

The PR curves of the HMMs using the analog readout are shown in Fig. 8.6 **c,d**, from which an average $\mathcal{O}(Q) \approx 97\%$ can be extracted for the ancilla-qubit HMMs. The temporal

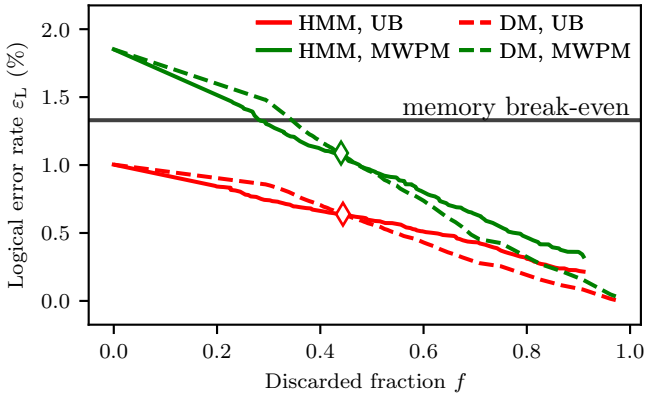

Figure 8.7: Improvement in the logical error rate ε_L via post-selecting on the detection of leakage for a MWPM decoder (green) and the decoder upper bound (red). The post-selection is based on the probabilities predicted by the HMMs (solid) or on those extracted from the density-matrix simulation (dashed), for 2×10^4 runs of 20 QEC cycles each. The physical error rate of a single transmon qubit under decoherence is also given (solid black). Detection of leakage allows for the restoration of the performance of the MWPM decoder, reaching the memory break-even point by discarding about $\approx 28\%$ of the data. The logical error rates obtained from simulations without leakage (and without post-selection) are indicated by diamonds.

responses of the HMMs to leakage are compared to the leakage probabilities extracted from measurement in Fig. 8.6 **e,f**, showing a relatively sharp response to a leakage event, with an expected delay in the detection of at most 2 QEC cycles. While $F^{\mathscr{L}} = 88.4\%$ might suggest an even sharper response, this is not the case as the HMM update depends on both the prior $p^{\mathscr{L}}_{\text{HMM}}$ (which is low when the qubit is not leaked) and on the likelihood of the sampled I_m together with the observed defects on the neighboring ancilla qubits (Section 8.10.3). While the initial response is not immediately high, given a (not too) low threshold, corresponding to a high $\mathscr{R}$, the HMMs still achieve a high $\mathscr{P}$, leading to the high $\mathscr{O}$ observed (see Fig. 8.6 **c**, **d**). A further benefit of the inclusion of the analog-measurement information is that the detection capability of the HMMs is now largely insensitive to the fluctuations in $\phi^{\mathscr{L}}_{\text{stat}}$ and $\phi^{\mathscr{L}}_{\text{flux}}$.

We explore $\mathscr{O}(Q)$ as a function of $F^{\mathscr{L}}$, as shown in the inset of Fig. 8.6 **c,d**. To do this, we model $\mathscr{N}_j$ for each state as symmetric and having the same standard deviation, in which case $F^{\mathscr{L}}$ is a function of their signal-to-noise ratio only (see Section 8.11.1). At low $F^{\mathscr{L}}$ $\left(\lesssim 60\%\right)$ the detection of leakage is possible but limited, especially for the boundary ancilla qubits. On the other hand, even at moderate values of $F^{\mathscr{L}}$ $(\approx 80\%)$, corresponding to experimentally achievable values, ancilla-qubit leakage can be accurately identified for both bulk and boundary ancilla qubits. Furthermore, relying solely on the analog measurements would allow for the potential minimization of the error spread associated with ancilla-qubit leakage, given controllability over $\phi^{\mathscr{L}}_{\text{stat}}$ and $\phi^{\mathscr{L}}_{\text{flux}}$, without compromising the capability of the HMMs to detect leakage.

8.8. IMPROVING CODE PERFORMANCE VIA POST-SELECTION

We use the detection of leakage to reduce the logical error rate ε_L via post-selection on leakage detection, with the selection criterion defined as

$$\max_{Q,n} p^{\mathscr{L}}(Q,n) \geq p^{\mathscr{L}}_{\text{th}}(Q)\,. \tag{8.8}$$

We thus post-select any run for which the leakage probability of any qubit exceeds the defined threshold in any of the QEC cycles. We note that post-selection is not scalable for larger-scale QEC, due to an exponential overhead in the number of required experiments, however, it can be useful for a relatively small code such as Surface-17. Furthermore, note that, while the criterion above is insensitive to overestimation of the leakage probability due to a leaked neighboring qubit (see Section 8.11.4), it is sensitive to the correct detection of leakage in the first place and to the HMM response in time (especially for short-lived leakage events).

We perform the multi-objective optimization

$$\begin{aligned} &\min_{p^{\mathscr{L}}_{\text{th}}(Q)} && \big(f, \varepsilon_L\big)\,, \\ &\text{subject to} && 0.02 \leq p^{\mathscr{L}}_{\text{th}}(Q) \leq 1, \end{aligned}$$

where f is the fraction of discarded data. The inequality constraint on the feasible space is helpful for the fitting procedure, required to estimate ε_L. This optimization uses an evolutionary algorithm (*NGSA-II*), suitable for conflicting objectives, for which the outcome is the set of lowest possible ε_L for a given f. This set is known as the Pareto front and is shown in Fig. 8.7 for both the MWPM and UB decoders. In Fig. 8.7 we also compare post-selection based on the HMMs against post-selection based on the density-matrix simulation. Here we use the predictions of the HMMs which include the analog measurement outcome with the experimentally extracted $F^{\mathscr{L}}$ (see Section 8.7). The observed agreement between the two post-selection methods proves that the performance gain is due to discarding runs with leakage instead of runs with only regular errors. The performance of the MWPM decoder is restored below the quantum memory break-even point by discarding $f \approx 28\%$. The logical error rates extracted from simulations without leakage are achieved by post-selection of $f \approx 44\%$ of the data for both the MWPM and UB decoders, when leakage is included.

8.9. DISCUSSION

We have investigated the effects of leakage and its detectability using density-matrix simulations of a transmon-based implementation of Surface-17. Data and ancilla qubits tend to be sharply projected onto the leakage subspace, either indirectly by a back-action effect of stabilizer measurements for data qubits or by the measurement itself for ancilla qubits. During leakage, a large, but local, increase in the defect rate of neighboring qubits is observed. For data qubits we attribute this to the anti-commutation of the involved stabilizer checks, while for ancilla qubits we find that it is due to an interaction-dependent spread of errors to the neighboring qubits. We have developed a low-cost and scalable approach based on HMMs, which use the observed signatures together with the

analog measurements of the ancilla qubits to accurately detect the time and location of leakage events. The HMM predictions are used to post-select out leakage, allowing for the restoration of the performance of the logical qubit below the memory break-even point by discarding less than half of the data (for such a relatively small code and for the given noise parameters), opening the prospect of near-term QEC demonstrations even in the absence of a dedicate leakage-reduction mechanism.

A few noise sources have not been included in the simulations. First, we have not included readout-declaration errors, corresponding to the declared measurement outcome being different from the state in which the ancilla qubit is projected by the measurement itself. These errors are expected to have an effect on the performance of the MWPM decoder, as well as a small effect on the observed optimality of the HMMs. We have also ignored any crosstalk effects, such as residual couplings, cross-driving or dephasing induced by measurements on other qubits. While the presence of these crosstalk mechanisms is expected to increase the error rate of the code, it is not expected to affect the HMM leakage-detection capability. We have assumed measurements to be perfectly projective. However, for small deviations, we do not expect a significant effect on the projection of leakage and on the observation of the characteristic signatures.

We now discuss the applicability of HMMs to other quantum-computing platforms subject to leakage and determine a set of conditions under which leakage can be efficiently detected. First, we assume single- and two-qubit gates to have low leakage probabilities, otherwise QEC would not be possible in general. In this way, single- and two-qubit leakage probabilities can be treated as perturbations to block-diagonal gates, with one block for the computational subspace $\mathscr{C}$ and one for the leakage subspace $\mathscr{L}$. We focus on the gates used in the surface code, i.e., CZ and Hadamard H (or $R_Y(\pi/2)$ rotations or equivalent gates). We consider data-qubit leakage first. We have observed that it is made detectable by the leakage-induced anti-commutation of neighboring stabilizers. The only condition ensuring this anti-commutation is that H acts as the identity in $\mathscr{L}$ or that it commutes with the action of CZ within the leakage block (see Section 8.11.2), regardless of the specifics of such action. Thus, data-qubit leakage is detectable via HMMs if this condition is satisfied. In particular, it is automatically satisfied if $\mathscr{L}$ is 1-dimensional. We now consider ancilla-qubit leakage. Clearly, ancilla-qubit leakage detection is possible if the readout discriminates computational and leakage states perfectly or with high fidelity. If this is not the case, the required condition is that leaked ancilla qubits spread errors according to non-trivial leakage conditional phases, constituting signatures that can be used by an HMM. If even a limited-fidelity readout is available, it can still be used to strengthen this signal, as demonstrated in Section 8.7. An issue is the possibility of the readout to project onto a superposition of computational and leakage subspaces. In that case, the significance of ancilla-qubit leakage is even unclear. However, for non-trivial leakage conditional phases, we expect a projection effect to the leakage subspace by a back-action of the stabilizer measurements, due to leakage-induced errors being detected onto other qubits, similarly to what observed for data qubits.

The capability to detect the time and location of a leakage event demonstrated by the HMMs could be used in conjunction with leakage-reductions units (LRUs) [38]. These are of fundamental importance for fault tolerance in the presence of leakage, since in Ref. [41] a threshold for the surface code was not found if dedicated LRUs are not used to

reduce the leakage lifetime beyond the one set by the relaxation time. While the latter constitutes a natural LRU by itself, we do not expect it to ensure a threshold since, together with a reduction in the leakage lifetime, it leads to an increase in the regular errors due to relaxation. A few options for LRUs (see also Chapter 5) in superconducting qubits are the swap scheme introduced in Ref. [37], or the use of the readout resonator to reset a leaked data-qubit into the computational subspace, similarly to Refs. [55, 56] [which developed into the res-LRU in Chapter 9]. An alternative is to use the $|02\rangle \leftrightarrow |11\rangle$ crossing to realize a "leakage-reversal" gate that exchanges the leakage population in $|02\rangle$ to $|11\rangle$. An even simpler gate would be a single-qubit π pulse targeting the $|1\rangle \leftrightarrow |2\rangle$ transition [which developed into the π-LRU in Chapter 9]. All these schemes introduce a considerable overhead either in hardware (swap, readout resonator), or time (swap, readout resonator, leakage-reversal gate), or they produce leakage when they are applied in the absence of it (leakage-reversal gate, π pulse). Thus, all these schemes would benefit from the accurate identification of leakage, allowing for their targeted application, reducing the average circuit depth and minimizing the probability of inadvertently inducing leakage. We also note that the swap scheme, in conjunction with a good discrimination fidelity for $|2\rangle$, could be used for detecting leakage not only on ancilla qubits but also on data qubits by alternatively measuring them. Still, this scheme would require 5 extra qubits for Surface-17 and would make the QEC-cycle time at least $\sim 50\%$ longer, together with more gate and idling errors, thus requiring much better physical error rates to achieve the same logical error rate in near-term experiments.

We discuss how decoders might benefit from the detection of leakage. Modifications to MWPM decoders have been developed for the case when ancilla-qubit leakage is directly measured [18, 41], and when data-qubit leakage is measured in the LRU circuits [41]. Further decoder modifications might be developed to achieve a lower logical error rate relative to a leakage-unaware decoder, by taking into account the detected leakage and the probability of leakage-induced errors, as well as the stabilizer information that can still be extracted from the superchecks (see Section 8.11.2). In the latter case, a decoder could switch back and forth from standard surface-code decoding to e.g. the partial subsystem-code decoding in Refs. [51–53]. Given control of the leakage conditional phases, the performance of this decoder can be optimized by setting $\phi^{\mathscr{L}}_{\text{stat}} = \pi$ and $\phi^{\mathscr{L}}_{\text{flux}} = 0$, minimizing the spread of phase errors on the neighboring data qubits by a leaked ancilla qubit, as well as the noise on the weight-6 stabilizer extraction in the case of a leaked data qubit (see Section 8.11.2). Given a moderate discrimination fidelity of the leaked state, this is not expected to compromise the detectability of leakage, as discussed in Section 8.7. At the same time, for such a decoder we expect the improvement in the logical error rate to be limited in the case of low-distance codes such as Surface-17, as single-qubit errors can result in a logical error. This is because leakage effectively reduces the code distance, either because a leaked data qubit is effectively removed from the code, or because of the fact that a leaked ancilla qubit is effectively disabled and in addition spreads errors onto neighboring data qubits. Large codes, for which leakage could be well tolerated (depending on the distribution of leakage events), cannot be studied with density-matrix simulations, as done in this work for Surface-17. However, the observed sharp projection of leakage and the probabilistic spread of errors justify the stochastic treatment of this error [41]. Under the assumption that amplitude and phase damping can be modeled

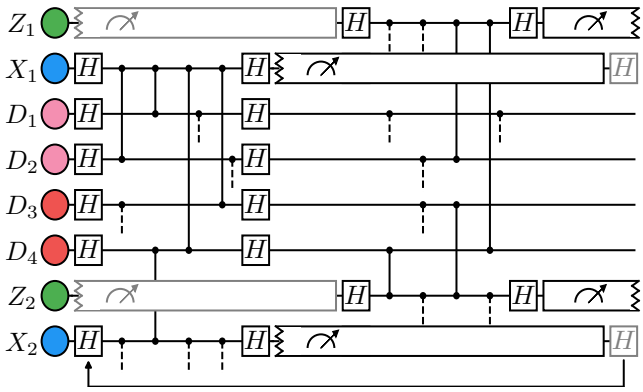

Figure 8.8: The quantum circuit for a single QEC cycle employed in simulation, for the unit-cell scheduling defined in [47]. The qubit labels and frequencies correspond to the lattice arrangement shown in Fig. 8.2. Gray elements correspond to operations belonging to the previous or the following QEC cycle. The X-type parity checks are performed at the start of the cycle, while the Z-type parity checks are executed immediately after the Z-type stabilizer measurements from the previous cycle are completed. The duration of each operation is given in Table 8.1. The arrow at the bottom indicates the repetition of QEC cycles.

stochastically as well, we expect that the performance of decoders and LRUs in large surface codes can be well approximated in the presence of leakage.

8.10. METHODS

8.10.1. SIMULATION PROTOCOL

For the Surface-17 simulations we use the open-source density-matrix simulation package *quantumsim* [28], available at [48]. For decoding we use a MWPM decoder [28], for which the weights of the possible error pairings are extracted from Surface-17 simulations via adaptive estimation [50] without leakage ($L_1 = 0$) and an otherwise identical error model (described in Section 8.10.2).

The logical performance of the surface code as a quantum memory is the ability to maintain a logical state over a number of QEC cycles. We focus on the Z-basis logical $|0\rangle_L$, but we have observed nearly identical performance for $|1\rangle_L$. We have not performed simulations for the X-basis logical states $|\pm\rangle_L = \frac{1}{\sqrt{2}}(|0\rangle_L \pm |1\rangle_L)$, as previous studies did not observe a significant difference between the two bases [28]. The state $|0\rangle_L$ is prepared by initializing all data qubits in $|0\rangle$, after which it is maintained for a fixed number of QEC cycles (maximum 20 or 50 in this work), with the quantum circuit given in Fig. 8.8. The first QEC cycle projects the logical qubit into a simultaneous eigenstate of the X- and Z-type stabilizers [29], with the Z measurement outcomes being +1 in the absence of errors, while the X outcomes are random. The information about the occurred errors is provided by the stabilizer measurement outcomes from each QEC cycle, as well as by a Z-type stabilizer measurements obtained by measuring the data qubits in the computational basis at the end of the run. This information is provided to the MWPM decoder, which estimates the logical state at the end of the experiment by tracking the Pauli frame. For decoding, we assume that the $|2\rangle$ state is measured as a $|1\rangle$, as in most current experiments.

Parameter	Value
Relaxation time T_1	30 μs
Sweetspot dephasing time $T_{\phi,\max}$	60 μs
High-freq. dephasing time at interaction point $T_{\phi,\text{int}}$	8 μs
Mid-freq. dephasing time at interaction point $T_{\phi,\text{int}}$	6 μs
Mid-freq. dephasing time at parking point $T_{\phi,\text{park}}$	8 μs
Low-freq. dephasing time at parking point $T_{\phi,\text{park}}$	9 μs
Single-qubit gate time t_{single}	20 ns
Two-qubit interaction time t_{int}	30 ns
Single-qubit phase-correction time t_{cor}	10 ns
Measurement time t_{m}	600 ns
QEC-cycle time t_{c}	800 ns

Table 8.1: The parameters for the qubit decoherence times and for the gate, measurement and QEC-cycle durations used in the density-matrix simulations.

In Section 8.7 we considered the discrimination of $|2\rangle$ in readout, which can be used for leakage detection. While this information can be also useful for decoding, we do not consider a leakage-aware decoder in this work.

The logical fidelity $F_{\text{L}}(n)$ at a final QEC cycle n is defined as the probability that the decoder guess for the final logical state matches the initially prepared one. The logical error rate ε_{L} is extracted by fitting the decay as

$$F_{\text{L}}(n) = \frac{1}{2}\left[1 + (1-2\varepsilon_{\text{L}})^{n-n_0}\right], \tag{8.10}$$

where n_0 is a fitting parameter (usually close to 0) [28].

8.10.2. ERROR MODEL AND PARAMETERS

In the simulations we include qubit decoherence via amplitude-damping and phase-damping channels. The time evolution of a single qubit is given by the Lindblad equation

$$\frac{d\rho}{dt} = -i\left[H,\rho\right] + \sum_i L_{\text{i}}\rho L_{\text{i}}^\dagger - \frac{1}{2}\left\{L_{\text{i}}^\dagger L_{\text{i}},\rho\right\}, \tag{8.11}$$

where H is the transmon Hamiltonian

$$H = \omega a^\dagger a + \frac{\alpha}{2}(a^\dagger)^2 a^2, \tag{8.12}$$

with a the annihilation operator, ω and α the qubit frequency and anharmonicity, respectively, and L_{i} the Lindblad operators. Assuming weak anharmonicity, we model amplitude

damping for a qutrit by

$$L_{\text{amp}} = \sqrt{\frac{1}{T_1}}a. \tag{8.13}$$

The $|2\rangle$ lifetime is then characterized by a relaxation time $T_1/2$. Dephasing is described by

$$L_{\text{deph},1} = \sqrt{\frac{8}{9T_\phi}}\begin{pmatrix}1&0&0\\0&0&0\\0&0&-1\end{pmatrix}, \tag{8.14}$$

$$L_{\text{deph},2} = \sqrt{\frac{2}{9T_\phi}}\begin{pmatrix}1&0&0\\0&-1&0\\0&0&0\end{pmatrix}, \tag{8.15}$$

$$L_{\text{deph},3} = \sqrt{\frac{2}{9T_\phi}}\begin{pmatrix}0&0&0\\0&1&0\\0&0&-1\end{pmatrix}, \tag{8.16}$$

leading to a dephasing time T_ϕ between $|0\rangle$ (resp. $|1\rangle$) and $|1\rangle$ ($|2\rangle$), and to a dephasing time $T_\phi/2$ between $|0\rangle$ and $|2\rangle$ (see Section 6.11.3). The Lindblad equation is integrated for a time t to obtain an amplitude- and phase-damping superoperator $R_{\downarrow,t}$, expressed in the Pauli Transfer Matrix representation. For a gate R_{gate} of duration t_{gate}, decoherence is accounted by applying $R_{\downarrow,t_{\text{gate}}/2}R_{\text{gate}}R_{\downarrow,t_{\text{gate}}/2}$. For idling periods of duration t_{idle}, $R_{\downarrow,t_{\text{idle}}}$ is applied.

For single-qubit gates we only include the amplitude and phase damping experienced over the duration t_{single} of the gate. These gates are assumed to not induce any leakage, motivated by the low leakage probabilities achieved [9, 45], and to act trivially in the leakage subspace. For two-qubit gates, namely the CZ, we further consider the increased dephasing rate experienced by qubits when fluxed away from their sweetspot. In superconducting qubits, flux noise shows a typical power spectral density $S_f = A/f$, where f is the frequency and $\sqrt{A}$ is a constant. In this chapter we consider $\sqrt{A} = 4\ \mu\Phi_0$, where Φ_0 is the flux quantum. Both low- and high-frequency components are contained in S_f, which we define relative to the CZ gate duration t_{CZ}. Away from the sweetspot frequency $\omega_{\max}$, a flux-tunable transmon has first-order flux-noise sensitivity $D_\phi = \frac{1}{2\pi}\left|\frac{\partial\omega}{\partial\Phi}\right|$. The high-frequency components are included as an increase in the dephasing rate $\Gamma_\phi = 1/T_\phi$ (compared to the sweetspot), given by $\Gamma_\phi = 2\pi\sqrt{\ln 2 A}D_\phi$ [57], while the low-frequency components are not included due to the built-in echo effect of Net-Zero pulses (see Section 6.5). High- and mid-frequency qubits are fluxed away to different frequencies, with dephasing rates computed with the given formula. Furthermore, during a few gates low-frequency qubits are fluxed away to a "parking" frequency in order to avoid unwanted interactions [47]. The computed dephasing times at the interaction point are given in Table 8.1. For the CZ gates, we include this increased dephasing during the time t_{int} spent at the interaction point, while for the phase-correction pulses of duration t_{cor} we consider

the same dephasing time as at the sweetspot. We do not include deviations in the ideal single-qubit phases of the CZ gate $\phi_{01} = 0$ and $\phi_{10} = 0$ and the two-qubit phase $\phi_{11} = \pi$, under the assumption that gates are well tuned and that the low-frequency components of the flux noise are echoed out (see Section 6.4).

We now consider the coherence of leakage in the CZ gates, which in the rotating frame of the qutrit is modeled as the exchanges

$$|11\rangle \mapsto \sqrt{1-4L_1}\,|11\rangle + e^{i\phi}\sqrt{4L_1}\,|02\rangle\,, \tag{8.17}$$

$$|02\rangle \mapsto \sqrt{1-4L_1}\,|02\rangle - e^{-i\phi}\sqrt{4L_1}\,|11\rangle\,, \tag{8.18}$$

with L_1 the leakage probability [49]. The phase ϕ can lead to an interference effect between consecutive applications of the CZ gate across pairs of data and ancilla qubits. In terms of the full density matrix, the dynamics of Eqs. (8.17) and (8.18) leads to a coherent superposition of computational and leaked states

$$\rho = \left(\begin{array}{c|c} \rho^{\mathscr{C}} & \rho^{\mathrm{coh}} \\ \hline \rho^{\mathrm{coh}} & \rho^{\mathscr{L}} \end{array}\right), \tag{8.19}$$

where $\rho^{\mathscr{C}}$ (resp. $\rho^{\mathscr{L}}$) is the density matrix restricted to the computational (leakage) subspace, while ρ^{coh} are the off-diagonal elements between these subspaces. We observe that varying the phase ϕ does not have an effect on the dynamics of leakage or on the logical error rate. We attribute this to the fact that each ancilla qubit interacts with a given data qubit only once during a QEC cycle and it is measured at the end of it (and as such it is dephased). Thus, the ancilla-qubit measurement between consecutive CZ gates between the same pair prevents any interference effect. Furthermore, setting $\rho^{\mathrm{coh}} = 0$, does not affect the projection and signatures of leakage nor the logical error rate (at least for the logical state prepared in the Z basis), leading to an incoherent leakage model. We attribute this to the projection of leakage itself, which leaves the qubit into a mostly incoherent mixture between the computational and leakage subspaces. In the harmonic rotating frame, $|2\rangle$ is expected to acquire an additional phase during periods of idling, proportional to the anharmonicity. However, following the reasoning presented above, we also believe that this phase is irrelevant.

An incoherent leakage model offers significant computational advantage for density-matrix simulations. For the case where $\rho_{\mathrm{coh}} \neq 0$, the size of the stored density matrix at any time is $4^6 \times 9^4$ (6 low-frequency data qubits, 3 high-frequency data qutrits plus 1 ancilla qutrit currently performing the parity check). Setting $\rho_{\mathrm{coh}} = 0$ reduces the size of the density matrix to $4^6 \times 5^4$, since for each qutrit only the $|2\rangle\langle 2|$ matrix element is stored in addition to the computational subspace. Thus, for the simulations in this work we rely on an incoherent model of leakage.

Measurements of duration t_{m} are modeled by applying $R_{\downarrow,t_{\mathrm{m}}/2} R_{\mathrm{proj}} R_{\downarrow,t_{\mathrm{m}}/2}$, where $R_{\downarrow,t_{\mathrm{m}}/2}$ are periods of amplitude and phase damping and R_{proj} is a projection operator. This projector is chosen according to the Born rule and leaves the ancilla qubit in either $|0\rangle$, $|1\rangle$ or $|2\rangle$. We do not include any declaration errors, which are defined as the measurement outcome being different from the state of the ancilla qubit immediately after the projection. Furthermore, we do not include any measurement-induced leakage, any decrease in the relaxation time via the Purcell effect or any measurement-induced dephasing via broadband sources. We do not consider non-ideal projective measurements

(leaving the ancilla in a superposition of the computational states) due to the increased size of the stored density matrix that this would lead to.

8.10.3. HMM FORMALISM

An HMM describes the time evolution of a set $S = \{s\}$ of not directly observable states s (i.e., "hidden"), over a sequence of independent observables $o = \{o_i\}$. At each time step n the states undergo a Markovian transition, such that the probability $p^s[n]$ of the system being in the state s is determined by the previous distribution $p^s[n-1]$ over all $s \in S$. These transitions can be expressed via the transition matrix A, whose elements are the conditional probabilities $A_{s,s'} := \mathbb{P}\left(s[n] = s \mid s[n-1] = s'\right)$. A set of observables is then generated with state-dependent probabilities $B_{o_i[n],s} := \mathbb{P}(o_i[n] = o_i \mid s[n] = s)$. Inverting this problem, the inference of the posterior state probabilities $p^s[n]$ from the realized observables is possible via

$$p^s[n] = \mathbb{P}(s[n] \mid o[n], o[n-1], \ldots, o[1]) \tag{8.20}$$

$$= \frac{\mathbb{P}(o[n] \mid s[n])\, p^s_{\text{prior}}[n]}{\mathbb{P}(o[n])} \tag{8.21}$$

$$= \frac{\prod_i \mathbb{P}(o_i[n] \mid s[n])\, p^s_{\text{prior}}[n]}{\prod_i \mathbb{P}(o_i[n])} \tag{8.22}$$

$$= \frac{\prod_i B_{o_i[n],s}\, p^s_{\text{prior}}[n]}{\sum_{s'} \prod_i B_{o_i[n],s'}\, p^{s'}_{\text{prior}}[n]}, \tag{8.23}$$

where $p^s_{\text{prior}}[n]$ is the prior probability

$$p^s_{\text{prior}}[n] = \sum_{s'} A_{s,s'}\, p^{s'}[n-1]. \tag{8.24}$$

We define $B_{o[n],s} = \prod_i B_{o_i[n],s}$, which for discrete o_i constitute the entries of the emission matrix B. In addition to the transition and emission probabilities, the initial state probabilities $p^s[n=0]$ are needed for the computation of the evolution.

In the context of leakage detection, we consider only two hidden states, $S = \{\mathscr{C}, \mathscr{L}\}$, namely whether the qubit is in the computational ($\mathscr{C}$) or the leakage subspace ($\mathscr{L}$). The transition matrix is parameterized in terms of the leakage and seepage probabilities per QEC cycle. The leakage probability is estimated as $\Gamma_{\mathscr{C}\to\mathscr{L}} \approx N_{\text{flux}} L_1$ (for low L_1), where N_{flux} is in how many CZ gates the qubit is fluxed during a QEC cycle and L_1 is the leakage probability per CZ gate. The seepage probability is estimated by $\Gamma_{\mathscr{L}\to\mathscr{C}} \approx N_{\text{flux}} L_2 + \left(1 - e^{\frac{t_c}{T_1/2}}\right)$, where t_c is the QEC cycle duration and T_1 the relaxation time (see Table 8.1), while L_2 is the seepage contribution from the gate, where $L_2 = 2L_1$ due to the dimensionality ratio between $\mathscr{C}$ and $\mathscr{L}$ for a qubit-qutrit pair [49]. The transition matrix A is then given by

$$A = \begin{pmatrix} 1 - \Gamma_{\mathscr{C}\to\mathscr{L}} & \Gamma_{\mathscr{L}\to\mathscr{C}} \\ \Gamma_{\mathscr{C}\to\mathscr{L}} & 1 - \Gamma_{\mathscr{L}\to\mathscr{C}} \end{pmatrix}. \tag{8.25}$$

We assume that all qubits are initialized in $\mathscr{C}$, which defines the initial state distribution $p^{\mathscr{C}}[n=0] = 1$ used by the HMMs.

The HMMs consider the defects $d(Q_i) \equiv d_i$ on the neighboring ancilla qubits Q_i at each QEC cycle, occurring with probability p^{d_i}, as the observables for leakage detection. Explicitly, the emission probabilities are parameterized in terms of the conditional probabilities $B_{d_i[n],s} = \mathbb{P}(d_i[n] \mid s)$ of observing a defect when the modeled qubit is in $s = \mathscr{C}$ or $s = \mathscr{L}$. We extract $B_{d_i[n],\mathscr{C}}$ directly from simulation, by averaging over all runs and all QEC cycles, motivated by the possible extraction of this probability in experiment. While this includes runs when the modeled qubit was leaked, we observe no significant differences in the HMM performance when we instead post-select out these periods of leakage, which we attribute to the low L_1 per CZ gate. We extract $B_{d_i[n],\mathscr{L}}$ from simulation over the QEC cycles when the leakage probability $p^{\mathscr{L}}_{\mathrm{DM}}(Q_i)$ as observed from the density matrix is above a threshold of $p^{\mathscr{L}}_{\mathrm{th}} = 0.5$. In the case of ancilla-qubit leakage, $B_{d_i[n],\mathscr{L}}$ depends on the values of the leakage conditional phases $\phi^{\mathscr{L}}_{\mathrm{stat}}$ and $\phi^{\mathscr{L}}_{\mathrm{flux}}$. Thus, in the case of randomized leakage conditional phases, the HMMs are parameterized by the average $B_{d_i[n],\mathscr{L}}$. In the case of data-qubit leakage, the extracted $B_{d_i[n],\mathscr{L}}$ is ≈ 0.5 regardless of the leakage conditional phases, as expected from the anti-commuting stabilizers (see Section 8.4).

For ancilla-qubit leakage detection, the analog measurement outcome I_m can be additionally considered as an observable, in which case $o = \{d_i, I_m\}$. In this case, the state-dependent probability is further parametrized by $B_{I_m[n],\mathscr{C}} = \mathbb{P}(I_m[n] \mid \mathscr{C}) = \mathscr{N}_0(I_m[n]) + \mathscr{N}_1(I_m[n])$ and by $B_{I_m[n],\mathscr{L}} = \mathbb{P}(I_m[n] \mid \mathscr{L}) = \mathscr{N}_2(I_m[n])$, where $\mathscr{N}_i$ are the Gaussian distributions of the analog responses in the IQ plane, projected along a rotated in-phase axis I, following the same treatment as in Section 8.11.1.

8.11. SUPPLEMENTAL MATERIAL

8.11.1. TRANSMON MEASUREMENTS IN EXPERIMENT

We consider the measurements of transmons in experiment [21], which is enabled by the dispersive coupling between a transmon and a dedicated readout resonator. The resonator is connected to a common feedline via a dedicated Purcell filter [17]. Measurement is performed by applying a readout pulse to the feedline, populating the resonator with photons. Each transmon induces a state-dependent shift of the frequency of the readout resonator, changing the amplitude and phase of the outgoing photons. This outgoing signal is amplified and the in-phase (I) and quadrature (Q) components are extracted. For calibration of the single-shot readout, the transmon is prepared in either $|0\rangle$, $|1\rangle$ or $|2\rangle$ and subsequently measured. Repeating this experiment characterizes the spread of the I and Q components of each state $|i\rangle$, which typically follow a two-dimensional Gaussian distribution $\mathscr{N}_i$ with mean $\vec{\mu}_i$ and standard deviation $\vec{\sigma}_i$ in the IQ plane [17, 58], as exemplified in Fig. 8.9 **a**.

Given an analog measurement of I and Q, the probability of a transmon being in state $|i\rangle$ can be expressed as

$$\mathbb{P}(i \mid I, Q) = \frac{\mathbb{P}(I, Q \mid i)\,\mathbb{P}(i)}{\mathbb{P}(I, Q)}, \tag{8.26}$$

where

$$\mathbb{P}(I, Q) = \sum_{j \in \{0,1,2\}} \mathbb{P}\big(I, Q \mid j\big)\,\mathbb{P}\big(j\big). \tag{8.27}$$

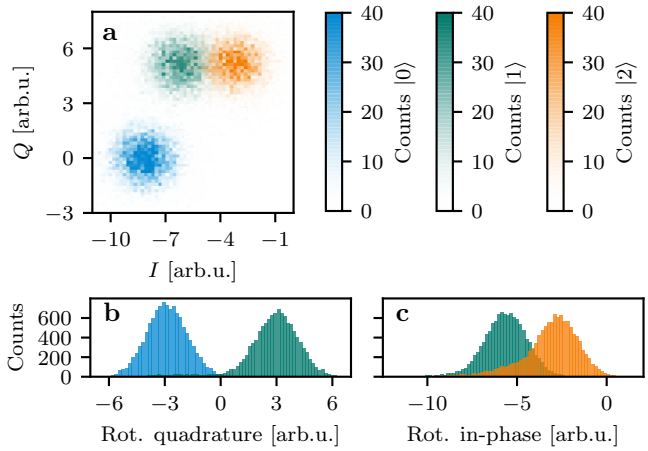

Figure 8.9: The analog measurement of transmons as extracted from experiment. **a** Histograms of the in-phase I and quadrature Q components of the measured readout for a transmon prepared in $|0\rangle$, $|1\rangle$ or $|2\rangle$. **b** The histograms of the responses for the transmon initially prepared in $|0\rangle$ or $|1\rangle$, projected along the rotated quadrature maximizing the discrimination fidelity $F^{01} = 99.6\%$. **c** The histograms of the responses for the transmon initialized in $|1\rangle$ or $|2\rangle$, projected along the I axis, in which case discrimination is achieved with a fidelity $F^{12} = 88.4\%$.

We assume that the prior state probabilities are equally likely. Furthermore, given the typically observed Gaussian distributions, it holds that $\mathbb{P}(I, Q \mid i) = \mathcal{N}_i(I, Q)$, which leads to

$$\mathbb{P}(i \mid I, Q) = \frac{\mathcal{N}_i(I, Q)}{\sum_{j\in\{0,1,2\}} \mathcal{N}_j(I, Q)}. \tag{8.28}$$

In experiment, one is typically interested in discriminating between pairs of states $|i\rangle$ and $|j\rangle$, for which the discrimination fidelity is defined as

$$F^{ij} = 1 - \mathbb{P}(j \mid i)\,\mathbb{P}(i) - \mathbb{P}(i \mid j)\,\mathbb{P}(j), \tag{8.29}$$

where $\mathbb{P}(i \mid j)$ is the probability of declaring a measurement outcome i given a prepared state $|j\rangle$, under the assumption of no excitation or relaxation during the measurement (accounted for in post-processing), and where $\mathbb{P}(i)$ is the prior probability of the qubit being in state $|i\rangle$. Hence, the discrimination fidelity corresponds to the probability of correctly declaring the projected state. We focus on the discrimination fidelity as in our simulations relaxation is already accounted for in the measurement outcomes (see Section 8.10.1). We assume $\mathbb{P}(i) = \mathbb{P}(j) = \frac{1}{2}$, which leads to

$$F^{ij} = 1 - \frac{\mathbb{P}(j \mid i) + \mathbb{P}(i \mid j)}{2}. \tag{8.30}$$

This can be related to the signal-to-noise ratio $\mathrm{SNR} = \left|\vec{\mu}_i - \vec{\mu}_j\right| / 2\sigma$, assuming symmetric

Gaussian distributions, as

$$F^{ij} = 1 - \frac{1}{2}\mathrm{erfc}\left(\frac{\mathrm{SNR}}{\sqrt{2}}\right). \tag{8.31}$$

The IQ response can be projected onto the axis joining the centers of a pair of two-dimensional Gaussian distributions, allowing to consider a single quadrature while maximizing the discrimination fidelity. Without loss of generality, we consider this optimal axis to be along I. This allows to express Eq. (8.28) as

$$\mathbb{P}(i \mid I) = \frac{\mathcal{N}_i(I)}{\sum_{j\in\{0,1,2\}} \mathcal{N}_j(I)}, \tag{8.32}$$

where $\mathcal{N}_i(I)$ is the marginal of $\mathcal{N}_i(I,Q)$. In experiment, in order to declare a binary measurement outcome, a threshold value for I is introduced, separating the regions for declaring either outcome. Following this approach, for a 3-outcome measurement, three projection axes are needed in general. However, since the Gaussian distributions for $|1\rangle$ and $|2\rangle$ are typically well-separated from the one for $|0\rangle$, it is possible to use only two axes, i.e., one to discriminate $|0\rangle$ from $|1\rangle$, and one to further discriminate $|2\rangle$ from the rest. For the measurement calibration from experiment [21], shown in Fig. 8.9 **a**, the discrimination between $|0\rangle$ and $|1\rangle$ can be achieved by projecting the analog responses along a rotated quadrature axis which maximizes the discrimination fidelity $F^{01} = 99.6\%$. Discriminating between $|1\rangle$ and $|2\rangle$ is performed with $F^{12} = 88.4\%$ by projecting along a rotated in-phase axis, maximizing this fidelity.

8.11.2. LEAKAGE-INDUCED ANTI-COMMUTATION

We study the behavior of neighboring stabilizers in the presence of a leaked data qubit. We focus on a parity-check operator in the bulk of the surface code. For the frequency scheme of Fig. 8.2, this involves two leakage-prone high-frequency transmons and two low-frequency transmons, modeled as qutrits and qubits, respectively. The ancilla qubit used to perform the parity checks is leakage prone as well. However, here we do not consider this possibility, given the low probability of a pair of neighboring data and ancilla qubits to be leaked simultaneously.

We consider the CZ for transmons described in Section 8.2, without including any decoherence. In the limit of the leakage probability $L_1 \to 0$ (and leakage mobility $L_m \to 0$), for an ancilla qubit A and a high-frequency data qubit D, the CZ can be decomposed as

$$|0\rangle\langle 0|_A \otimes \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}_D + |1\rangle\langle 1|_A \otimes \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -e^{-i\phi^{\mathscr{L}}_{\mathrm{stat}}} \end{pmatrix}_D \tag{8.33}$$

$$=: |0\rangle\langle 0|_A \otimes \tilde{I}_D + |1\rangle\langle 1|_A \otimes \tilde{Z}_D. \tag{8.34}$$

Note that $\tilde{I}|_{\mathscr{C}} = I$ and $\tilde{Z}|_{\mathscr{C}} = Z$, where I and Z are the standard identity and Pauli Z operators, respectively, and $\mathscr{C}$ is the qubit computational subspace. For a CZ between an ancilla qubit and a low-frequency data qubit, it simply holds $|0\rangle\langle 0|_A \otimes I_D + |1\rangle\langle 1|_A \otimes Z_D$. Small values of L_1, as observed in experiment (see Section 6.8), can be treated as a perturbation to this.

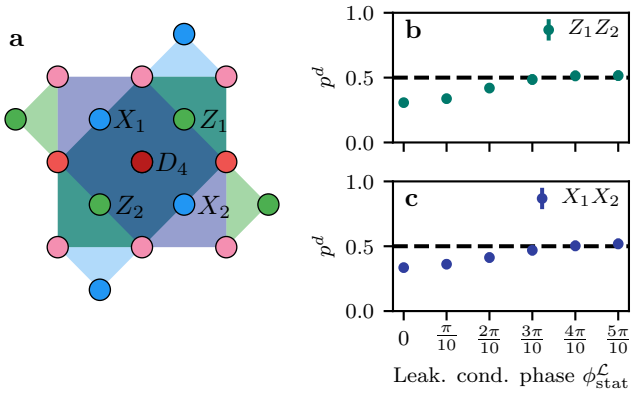

Figure 8.10: The effects of data-qubit leakage on the stabilizers of the code. **a** Sketch of how data-qubit leakage in the bulk (e.g. on D_4) effectively defines weight-3 gauge operators, whose product forms a weight-6 X-type (purple) or Z-type (teal) "supercheck" stabilizer, in addition to the standard weight-2 X-type (blue) and Z-type (green) stabilizers. **b,c** The average probability p^d of observing a defect on the supercheck stabilizers during leakage on D_4 (defined by the leakage probability being above a threshold of 0.5) as a function of the leakage conditional phase $\phi^{\mathcal{L}}_{\text{stat}}$.

For a parity-check measurement, the back-action on the state of the data qubits is given by either one of two operators, depending on the outcome. In the case of a Z-type parity-check unit, these operators are given by

$$M^Z_{\pm} = \frac{\tilde{I}_{abcd} \pm \tilde{Z}_{abcd}}{2}, \tag{8.35}$$

where $\tilde{I}_{abcd} \coloneqq \tilde{I}_a\tilde{I}_bI_cI_d$ and $\tilde{Z}_{abcd} \coloneqq \tilde{Z}_a\tilde{Z}_b\tilde{Z}_c\tilde{Z}_d$. Under the assumption that single-qubit gates, namely the Hadamard gate, do not induce any leakage and act trivially on the leakage subspace, for the X-type parity-check unit these operators are instead given by

$$M^X_{\pm} = \frac{\tilde{I}_{abcd} \pm \tilde{X}_{abcd}}{2}, \tag{8.36}$$

where $\tilde{X}_{abcd} \coloneqq \tilde{X}_a\tilde{X}_bX_cX_d$ and

$$\tilde{X} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -e^{-i\phi^{\mathcal{L}}_{\text{stat}}} \end{pmatrix}, \tag{8.37}$$

in which case $\tilde{X}|_{\mathscr{C}} = X$ with X the standard Pauli operator.

The X-type and Z-type parity checks commute if and only if $M^Z_{\pm}$ and $M^X_{\pm}$ commute, as it holds

$$[M^Z_{\pm}, M^X_{\pm}] = \frac{1}{4}[\tilde{Z}_{abcd}, \tilde{X}_{abcd}] \tag{8.38}$$

(and also $[M_\pm^Z, M_\pm^X] = -[M_\pm^Z, M_\mp^X]$). To compute the commutator we first evaluate

$$[\tilde{Z}, \tilde{X}] = 2i \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{8.39}$$

$$\{\tilde{Z}, \tilde{X}\} = 2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & e^{-2i\phi_{\text{stat}}^{\mathscr{L}}} \end{pmatrix}. \tag{8.40}$$

It follows that

$$\{\tilde{Z}, \tilde{X}\}\Big|_{\mathscr{C}} = \{Z, X\} = 0, \tag{8.41}$$

$$[\tilde{Z}, \tilde{X}]\Big|_{\mathscr{L}} = 0, \tag{8.42}$$

when restricted to the computational and leakage subspaces, respectively. Notice that, when all data qubits are in the computational subspace, it holds

$$\{\tilde{Z}_{abcd}, \tilde{X}_{abcd}\}\Big|_{\mathscr{C}} = \{Z_{abcd}, X_{abcd}\} = 0 \tag{8.43}$$

as in the standard qubit case, where $Z_{abcd} = Z_a Z_b Z_c Z_d$ and $X_{abcd} = X_a X_b X_c X_d$. Furthermore, we note that Eq. (8.42) holds solely because H acts trivially in $\mathscr{L}$ (as we assume) and it would continue to hold as long as H commutes with CZ on $\mathscr{L}$.

We now consider the case in which one of the high-frequency data qubits is in $\mathscr{L}$ (say a) and the remaining ones are in $\mathscr{C}$. In this case

$$\begin{aligned} \{\tilde{Z}_{abcd}, \tilde{X}_{abcd}\}\Big|_{\mathscr{L}_a} &= \left\{-e^{-i\phi_{\text{stat}}^{\mathscr{L}}} Z_{bcd}, -e^{-i\phi_{\text{stat}}^{\mathscr{L}}} X_{bcd}\right\} \\ &= e^{-i\phi_{\text{stat}}^{\mathscr{L}}} \{Z_{bcd}, X_{bcd}\} = 0. \end{aligned} \tag{8.44}$$

This shows that, in the presence of data-qubit leakage, $M_\pm^Z$ and $M_\pm^X$ do not commute. In particular, $\tilde{Z}_{abcd}$ and $\tilde{X}_{abcd}$ anti-commute and this result is independent of the leakage conditional phase. Furthermore, it holds

$$M_\pm^Z|_{\mathscr{L}_a} = \frac{I_{bcd} \pm e^{-i\phi_{\text{stat}}^{\mathscr{L}}} Z_{bcd}}{2} \tag{8.45}$$

and similarly for $M_\pm^X|_{\mathscr{L}_a}$.

For $\phi_{\text{stat}}^{\mathscr{L}} = 0, \pi$, $M_\pm^X|_{\mathscr{L}_a}$ are projectors onto the $\pm$-eigenspaces of Z_{bcd} or X_{bcd}, constituting effective weight-3 parity checks. In this case the anti-commutation [Eq. (8.44)] leads to fully randomized ancilla-qubit measurement outcomes, corresponding to a probability $p^d = 50\%$ of observing a defect each QEC cycle on each of the neighboring stabilizers. However, the product of two weight-3 same-type checks is a weight-6 stabilizer of the surface code, thus the product of the two ancilla-qubit measurement outcomes corresponds to the parity of the 6 data qubits involved. In particular, the stabilizer group can be redefined as including the standard weight-4 checks which do not involve the leaked qubit, together with the defined weight-6 "superchecks", while the weight-3 checks

are gauge operators [51–54], as illustrated in Fig. 8.10 **a**. For the superchecks to be correctly obtained, both X-type gauge operators need to be measured before any of the two Z-type gauge operators (or viceversa), which already holds true for the circuit schedule we consider [47]. In the case of a leaked qubit on the boundary, only one supercheck operator can be defined (for a rotated surface code, this is a weight-4 X- or Z-type boundary supercheck), while the other one must be ignored for decoding [51–53]. In the case of one leaked data-qubit in Surface-17, the minimum weight of a dressed logical operator is 2, reducing the code distance by 1. For example, if D_4 is leaked, two X errors on D_2 and D_7 constitute a logical X. In a larger surface code, the reduction of the distance depends on the number of leaked qubits, as well as their distribution on the lattice [51].

In the general case where $\phi^{\mathscr{L}}_{\text{stat}} \neq 0, \pi$, while the anti-commutation still holds, $M^Z_{\pm}|_{\mathscr{L}_a}$ and $M^X_{\pm}|_{\mathscr{L}_a}$ are not projectors and thus the ancilla-qubit measurement outcomes are not fully randomized, which is expected to have an effect on the observed p^d. However, in the simulations $p^d \approx 50\%$ for both the case when $\phi^{\mathscr{L}}_{\text{stat}}$ is randomized across runs (see Fig. 8.3) or when it is fixed, independently of the specific value. Since the defects d are computed as $d[n] = m[n] \oplus m[n-2]$, where $m[n]$ is the measurement outcome at QEC cycle n, even a moderate imbalance between the probabilities of measuring $m[n] = 0$ and $m[n] = 1$ (fluctuating across QEC cycles) can lead a defect probability $p^d \approx 50\%$. Furthermore, the phase rotations depending on $\phi^{\mathscr{L}}_{\text{stat}}$ affect the measurement of each of the two weight-3 gauge operators independently, which in turn undermines the correct extraction of the weight-6 stabilizer parity. This effect is observed in Fig. 8.10 **b,c**, where in the case of $\phi^{\mathscr{L}}_{\text{stat}} = 0, \pi$ the observed defect probability roughly corresponds to the expected one from a weight-6 check (relative to the observed one for the standard weight-4 and weight-2 checks in the absence of leakage), while a higher defect probability is observed otherwise, reaching up to 50%. Hence, the control of $\phi^{\mathscr{L}}_{\text{stat}}$ in experiment would be beneficial for decoding in the presence of data-qubit leakage whenever the superchecks are considered.

8.11.3. PROJECTION OF DATA-QUBIT LEAKAGE BY STABILIZER-MEASUREMENT BACK-ACTION

In this section we discuss how leakage is projected by the stabilizer measurements and in particular by the observed defects. First, we consider a simple 3-qubit parity-check circuit, for which an analytical formula can be derived for the projection of leakage after the observation of a single defect. We consider the circuit in Fig. 8.11 **a**. An ancilla qubit A is used to measure the stabilizer ZZ on two data qubits Q_1, Q_2. This is the same circuit as for one of the boundary Z-type ancilla qubits in Surface-17. Here we consider the initial state of the two qubits to be the Bell state $|\Phi^+\rangle_{Q_1Q_2} = (|00\rangle + |11\rangle)/\sqrt{2}$, that is, the +1-eigenstate for both ZZ and XX. For simplicity, the CZs are considered ideal apart from the one between A and Q_1 which has a leakage probability L_1 for Q_1, hence only this qubit can leak. To emulate relaxation in the actual system, we consider an incoming X error occurring with probability p on Q_1 (Z errors are not detected by a ZZ measurement, so we do not consider them here). Prior to the measurement, the system can be either in state

$$|\psi_1\rangle_{Q_1Q_2A} = \frac{1}{2\sqrt{2}}\Big(\big[2|00\rangle + (1+a)|11\rangle + b|21\rangle\big]|0\rangle + \big[(1-a)|11\rangle + b|21\rangle\big]|1\rangle\Big), \tag{8.46}$$

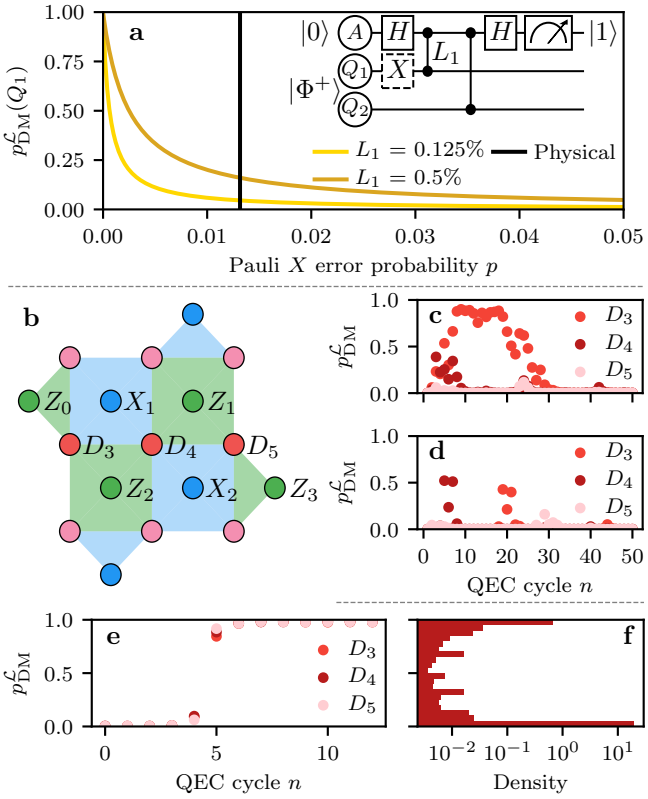

Figure 8.11: Projection of data-qubit leakage. **a** Inset: an ancilla qubit A, initialized in $|0\rangle$, measures ZZ on two data qubits Q_1, Q_2, initialized in the Bell state $|\Phi^+\rangle$, and we assume that the measurement projects A onto $|1\rangle$ (thus resulting in a defect here). All operations are noiseless except for a leakage probability L_1 in the first CZ. A Pauli X error occurs with probability p on Q_1. Main plot: post-measurement leakage probability $p^{\mathcal{L}}_{\mathrm{DM}}(Q_1)$ versus p. The black vertical line corresponds to the physical error rate of a transmon in the Surface-17 simulations. **b** Schematic overview of the Surface-17 layout, where pairs of high-frequency data qubits share two ancilla qubits as nearest neighbors. **c-d** Example realizations of data-qubit leakage projections, extracted from the density-matrix simulations. For each run we plot $p^{\mathcal{L}}_{\mathrm{DM}}$ for all three high-frequency data qubits. **e** The average projection of the leakage probability $p^{\mathcal{L}}_{\mathrm{DM}}$ of all three high-frequency data qubits in the absence of relaxation and decoherence (D_3 and D_4 are mostly obscured by D_5). This average is computed by selecting realizations where $p^{\mathcal{L}}_{\mathrm{DM}}(Q)$ was below a threshold $p^{\mathcal{L}}_{\mathrm{th}} = 0.5$ for at least 5 QEC cycles and then above it for 8 or more cycles. **f** Density histogram of all data-qubit leakage probabilities over 20 bins, in the absence of relaxation and decoherence, extracted over 2×10^4 runs of 20 QEC cycles each. Error bars, estimated by bootstrapping, are smaller than the symbol sizes.

with probability $1-p$, where $a=\sqrt{1-4L_1}$ and $b=\sqrt{4L_1}$, or in state

$$|\psi_2\rangle_{Q_1Q_2A}=\frac{1}{2\sqrt{2}}\Big(\big[(1-a)\,|10\rangle+b\,|20\rangle\big]\,|0\rangle+\big[(1+a)\,|10\rangle+2\,|01\rangle+b\,|20\rangle\big]\,|1\rangle\Big), \tag{8.47}$$

with probability p.

Here the measurement of the ancilla qubit in $|1\rangle$ leads to the observation of a defect. In that case, the back-action of this measurement gives the overall density-matrix:

$$\begin{aligned}\rho|_{|1\rangle}=\frac{1}{2(1-a)+4p(1+a)}\Big(\quad&(1-p)\big[(1-a)^2\,|11\rangle\langle 11|+b^2\,|21\rangle\langle 21|\big]\\&p\big[(1+a)^2\,|10\rangle\langle 10|+4\,|01\rangle\langle 01|+b^2\,|20\rangle\langle 20|\\&+2(1+a)(|10\rangle\langle 01|+|01\rangle\langle 10|)\big]\Big),\end{aligned} \tag{8.48}$$

where we have set the off-diagonal terms containing a $|2\rangle$ to 0, consistently with the simulations in this work (in any case, they do not matter for the present discussion), see Section 8.10.2. Tracing out Q_2, the leakage probability of Q_1 is

$$p^{\mathscr{L}}_{\mathrm{DM}}(Q_1)=\frac{4L_1}{2(1-\sqrt{1-4L_1})+4p(1+\sqrt{1-4L_1})}, \tag{8.49}$$

where the denominator is just the probability of observing a defect. Thus, the product of this probability and of $p^{\mathscr{L}}_{\mathrm{DM}}(Q_1)$ is a constant equal to $4L_1$. This means that the average leakage probability of Q_1, sampled over many measurements, is expected to grow towards the steady state proportionally to L_1, as observed in Section 8.11.8 for Surface-17. However, Eq. (8.49), plotted in Fig. 8.11 **a**, shows that $p^{\mathscr{L}}_{\mathrm{DM}}(Q_1)$, conditioned on the observation of a defect, can be much higher than L_1. In particular, when $p\to 0$, Q_1 becomes (almost) fully leaked. This is due to the fact that, if there are no regular Pauli errors causing defects, but leakage is possible and leads to defects (here due to the use of the $|11\rangle\leftrightarrow|02\rangle$ avoided crossing), then the observation of a defect indicates that the qubit is leaked. When p is larger, the projection of leakage is less sharp since it is more likely that a defect is caused by a regular error rather than by leakage. For example, for p equal to the physical error rate considered in the Surface-17 simulations ($T_1=T_2=30\ \mu$s), indicated by a black line in Fig. 8.11 **a**, $p^{\mathscr{L}}_{\mathrm{DM}}(Q_1)=4.5\%$, which is still much larger than $L_1=0.125\%$.

Since the error model we consider for Surface-17 is more realistic and there are more leakage-prone interactions between qubits, we further analyze the data-qubit leakage projection using numerics. We first focus on the behavior observed across individual realizations, where $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ of any of the data qubits sharply increases. An example of a leakage event of D_3 is shown in Fig. 8.11 **c**, where $p^{\mathscr{L}}_{\mathrm{DM}}(D_3)$ is sharply projected to a high value. However, during the initial projection, $p^{\mathscr{L}}_{\mathrm{DM}}(D_4)$ simultaneously rises to values around 0.5, where it remains for a few QEC cycles. We attribute this uncertainty to the fact that ancilla qubits X_1 and Z_2 are nearest neighbors of both data qubits, as illustrated in Fig. 8.11 **b**. The observation of defects on either one or both of these ancilla qubits can be roughly equally likely to be due to either data qubit being leaked. As leakage is projected via a back-action effect of the observation of defects, unambiguous defect observations lead to finite $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ of both data qubits. A second example of a realization

of data-qubit leakage is shown in Fig. 8.11 **d**, where both D_3 and D_4 exhibit sharp and brief projections to $p^{\mathscr{L}}_{\mathrm{DM}}(Q) \approx 0.5$ at different QEC cycles. These jumps can be either due to very short-lived leakage events, or due to the observations of multiple defects, which can eventually be attributed to one or more regular errors, but which also have a significant overlap with the signatures of leakage of D_3 or D_4, respectively. We note that we have observed multiple instances of the example realizations discussed above. Thus across individual realizations of leakage, $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ for the high-frequency data qubits is not always monotonically increasing (resp. decreasing) to high (low) probabilities in the case of a qubit leaking outside of (relaxing back to) the computational subspace. Similarly, there are fluctuations in $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ throughout leakage events across individual realizations. The observed bi-modal density distribution shown in Fig. 8.3 **b** shows that these small jumps and fluctuations are relatively rare, which we attribute to the repetitive stabilizer measurements and the observed strong signatures of leakage (see Section 8.4). To make the selection of leakage events (in Fig. 8.3 **d**, 8.5 **a**, 8.10 **b**, **c** 8.11 **e**, 8.12 **a**, **b**) less sensitive to such fluctuations, we apply a Savitzky-Golay filter with a window length of 5 QEC cycles and a first-order polynomial for the sample fitting. This filter smooths out the traces, to which we then apply our selection criterion. However, when computing the average projection from the selected realizations, we do not use the smoothed leakage probabilities, but directly the values extracted from simulation.

We finally analyze how the projection of data-qubit leakage in Surface-17 is affected by the physical error rates considered in this work. Figure 8.11 **e** shows that, in the absence of relaxation and decoherence ($T_1 = T_2 = \infty$), the average $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ of any of the high-frequency data qubits is projected to near unity in two QEC cycles whenever a qubit leaks. This projection is sharper than in the case with relaxation and decoherence, shown in Fig. 8.3 **d**, in agreement with the expectation based on Fig. 8.11 **a** for $p = 0$ and Eq. (8.49). The density distribution of all $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$ of the three high-frequency data qubits, shown in Fig. 8.11 **f**, while highly bi-modal is still supported on intermediate values between 0 and 1 of $p^{\mathscr{L}}_{\mathrm{DM}}(Q)$, contrarily to what Fig. 8.11 **a** would suggest for $p = 0$. We attribute this to the uncertainty associated with the observations of ambiguous defects through the leakage events, as suggested by Fig. 8.11 **c**-**d**.

8.11.4. HMM ERROR BUDGET

In this section we explore the limiting factors behind the remaining suboptimality of the HMMs presented in this chapter. The HMMs consider the probability of observing a defect at a given QEC cycle on each stabilizer independently, thus they do not take into account the correlations between defects due to regular errors. Data-qubit errors or hook errors (which are data-qubit errors propagated due to a single ancilla-qubit error during the parity-check circuit) give rise to a pair of correlated defects on different stabilizers either in the same QEC cycle or in two consecutive ones. Ancilla-qubit errors or measurement-declaration errors instead give rise to pairs of correlated defects on the same stabilizer and for one or two QEC cycles, respectively. As the HMMs take an increase in the defect probability as a signature of leakage, this is expected to result in the HMMs overestimating the probability of the tracked qubit being leaked. In addition, each HMM only takes the defects on the neighboring stabilizers as observables. Despite each HMM sharing observables with the neighboring ones, the probability of leakage at each

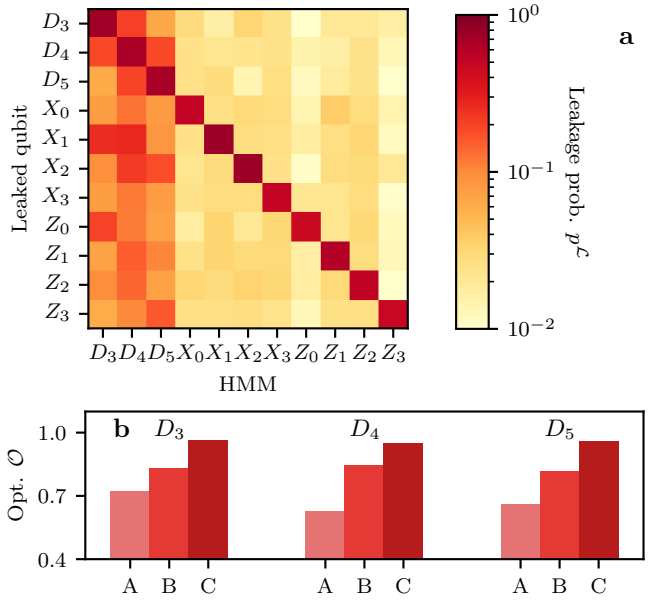

Figure 8.12: The crosstalk between the HMMs. **a** Average responses of all HMMs 1 QEC cycle after a given qubit leaks. We select individual realizations where the leakage probability $p^{\mathcal{L}}$ is first below and then above a threshold $p^{\mathcal{L}}_{\text{th}} = 0.5$ for 5 and 8 QEC cycles, respectively. **b** The extracted data-qubit HMM optimality $\mathcal{O}$. A: optimality of the HMMs including all error sources. B: runs where ancilla-qubit leakage was present (according to density matrix) are discarded. C: leakage on any of the other data qubits (not tracked by the given HMM) is discarded as well.

QEC cycle is estimated independently by each HMM. While this choice minimizes the computational overhead, as a result each HMM is additionally prone to overestimating the probability of leakage when a neighboring qubit is leaked instead (leading to an increased defect probability observed on only a subset of the stabilizers taken as observables by the HMM). The HMMs can be expanded to account for these limitations, either by increasing the number of hidden states to model regular errors [21] or by expanding the set of observables to include next-nearest neighbor stabilizers, in order to account for leakage on neighboring qubits, in which case the HMMs would be still local and hence scalable. As either solution would increase the complexity and overhead of the models, we evaluate the contributions of each of these limitation to the detection capabilities of the HMMs.

We first focus on the overestimation of the leakage probability predicted by the HMMs in the presence of leakage on a neighboring qubit, which we refer to as "HMM crosstalk". We consider the detection scheme taking into account the analog measurements (with the currently achieved experimental discrimination fidelity $F^{\mathscr{L}}$, see Section 8.7). The average responses of all HMMs to leakage events on any qubit and the predicted leakage probability 1 QEC cycle after detection (defined by the predicted probability crossing a threshold of 0.5) are shown in Fig. 8.12 **b**. The responses of the neighboring HMMs immediately (1-2 QEC cycles) after crossing this threshold is indicative of the likelihood of leakage being declared on a neighboring qubit (based on the extracted HMM responses shown in Figs. 8.5 and 8.6). Across individual runs, these parasitic responses can lead to false detections. Ancilla-qubit HMMs are insensitive to leakage on other data or ancilla qubits (see Fig. 8.12). We attribute this to the use of the analog measurement outcomes which discriminate between $|1\rangle$ and $|2\rangle$ with moderate fidelity and between $|0\rangle$ and $|2\rangle$) with very high fidelity. Instead, data-qubit HMMs are prone to overestimating the response in the case of leakage on other qubits. The crosstalk is proportional to the number of shared observables between the pairs of HMMs and depends on the expected defect probabilities during leakage by each model.

We further break down the relative contributions to the optimality $\mathscr{O}$ (defined in Section 8.6) of each of the data-qubit HMMs due to the crosstalk, shown in Fig. 8.12 **b**. Post-selecting out runs where ancilla-qubit leakage is detected from the density matrix increases the average $\mathscr{O}$ of the three data-qubit HMMs from $\mathscr{O} \approx 67.0\%$ to $\mathscr{O} \approx 83.3\%$. Further post-selecting out events where leakage is detected on any of the other data qubits (which are not tracked by the given HMM) increases the average optimality to $\mathscr{O} \approx 95.9\%$. The larger contribution from neighboring data-qubit leakage is consistent with the higher crosstalk (see Fig. 8.12 **a**) between data-qubit HMMs relative to the ancilla-qubit ones and constitutes the dominant limitation behind the HMM optimality. We attribute the remaining suboptimality to the presence of regular errors, caused by qubit relaxation and dephasing, and to the parametrization of the transition and output probabilities.

8.11.5. AN ALTERNATIVE SCHEME FOR ENHANCING ANCILLA-QUBIT LEAKAGE DETECTION

We consider an alternative scheme (to the one considering the analog measurement outcomes) allowing for enhancing ancilla-qubit leakage detection beyond that achievable by only considering the increase in the defect probability on neighboring stabilizers. In this scheme a π pulse is applied to each ancilla qubit every other QEC cycle, accounted

Parameter	D_{low}	D_{mid}	D_{high}
$\omega/2\pi$ at sweet spot (GHz)	4.9	6.0	6.7
$\alpha/2\pi$ (MHz)	−300	−300	−300
$J_1/2\pi$ at int. point (MHz)	15		15

Table 8.2: Parameters used in the CZ full-trajectory simulations, with α the anharmonicity and J_1 the coupling. We follow the frequency scheme of [47] with the arrangement shown in Fig. 8.2.

for in post-processing. Under the assumption that a π rotation has a trivial effect on a leaked qubit, the post-processed measurement outcomes (in the absence of errors) would show a flip every other QEC cycle during the period of leakage, which corresponds to a defect every QEC cycle. This scheme would require minimal overhead, as these rotations can be integrated with the existing single-qubit gates applied to the ancilla qubits at the start of each QEC cycle. A downside is that ancilla qubits would spend more time in the first excited state on average, increasing the effect of amplitude damping. We have not simulated this scheme, but we have investigated it entirely in post-processing by only applying flips to the measurement outcomes during periods of ancilla-qubit leakage (as extracted from the density matrix). Although this does not capture the increase in the ancilla-qubit error rate due to amplitude damping, we expect that it captures the effect of the scheme on the detection of leakage.

The average HMM optimality for the bulk X and Z ancilla qubits is $\mathcal{O}(X) \approx 64.9\%$ and $\mathcal{O}(Z) \approx 50.3\%$, respectively. For the boundary X and Z ancilla qubits, it is $\mathcal{O}(X) \approx 73.9\%$ and $\mathcal{O}(Z) \approx 46.4\%$, respectively. This constitutes an increase in optimality relative to the scheme relying only on the observed defects (see Fig. 8.6 **a,b**). However, the artificially induced defects on leaked ancilla qubits lead to the increase in the crosstalk between ancilla- and data-qubit HMMs. This has the effect of lowering the average data-qubit HMM optimality from $\mathcal{O}(D) \approx 67.0\%$ (see Section 8.6) to $\mathcal{O}(D) \approx 31.2\%$. While such scheme may be beneficial for the post-selection-based scheme defined in Section 8.8 (as in that case leakage detected on any qubits leads to discarding the run), it would be detrimental for leakage-aware decoding or targeted leakage-reduction units as these rely on the accurate detection in both time and space.

8.11.6. SECOND-ORDER LEAKAGE EFFECTS

In this section we consider exchanges between states in the leakage subspace as a result of a CZ gate acting on an already leaked qubit. We focus on the exchange between $|12\rangle$ and $|21\rangle$, referred to as "leakage mobility" in Section 8.2. We also expand the model to include $|3\rangle$ on the fluxing qubit and consider the exchange between $|03\rangle$ and $|12\rangle$, which we call "superleakage".

The Hamiltonian of two transmons dispersively coupled via a bus resonator in the

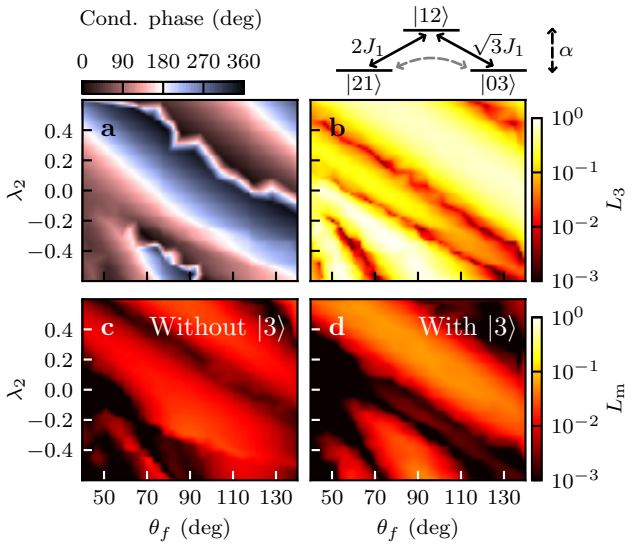

Figure 8.13: Heatmaps obtained from CZ full-trajectory simulations, including (**a,b,d**) and not including (**c**) $|3\rangle$ in the Hilbert space of the fluxing qubit. The conditional phase ϕ_{2Q} (**a**), L_1 (**b**) and L_m (**c,d**) are plotted versus the flux-pulse parameters (see Ref. Section 6.11.2 for definitions). The interaction point is located at $\theta_f = 90$ deg. The inset (top-right) schematically shows the direct and effective couplings between levels in the 3-excitation manifold at the interaction point. The states $|03\rangle$ and $|21\rangle$ are on resonance, while $|12\rangle$ is detuned by one anharmonicity α.

rotating-wave approximation is given by

$$\begin{aligned} H(t) = \omega_{\text{stat}} a^\dagger_{\text{stat}} a_{\text{stat}} + \frac{\alpha_{\text{stat}}}{2} (a^\dagger_{\text{stat}})^2 a^2_{\text{stat}} \\ + \omega_{\text{flux}}(\Phi(t))\, a^\dagger_{\text{flux}} a_{\text{flux}} + \frac{\alpha_{\text{flux}}}{2} (a^\dagger_{\text{flux}})^2 a^2_{\text{flux}} \\ + J_1(\Phi(t))\, (a_{\text{stat}} a^\dagger_{\text{flux}} + a^\dagger_{\text{stat}} a_{\text{flux}}), \end{aligned} \tag{8.50}$$

where a is the annihilation operator, ω and α are the qubit frequency and anharmonicity, respectively, and J_1 is the effective coupling mediated by virtual excitations through the bus resonator. We assume that this Hamiltonian is a valid approximation up to the included states. For this Hamiltonian, multiple avoided crossings are found when sweeping ω_{flux}, as schematically shown in Fig. 8.1. We perform full-trajectory simulations (following the same structure as in Section 6.11.3, excluding distortions and quasi-static flux noise) using the parameters reported in Table 8.1 and Table 8.2. Note that extending these simulations to $|3\rangle$ does not affect the leakage probability L_1 from the computational ($\mathscr{C}$) to the leakage subspace ($\mathscr{L}$), nor the fidelity within $\mathscr{C}$.

We define the superleakage probability L_3 as

$$L_3 := |\langle 03| \mathscr{S}_{\text{CZ}}(|12\rangle\langle 12|)|03\rangle|^2, \tag{8.51}$$

where $\mathscr{S}_{\text{CZ}}$ is the superoperator corresponding to the simulated noisy CZ. L_3 can be high depending on the specific parameters of the flux pulse and of the system, as Fig. 8.13 **b** shows for the high-mid qubit pair, even when $\phi_{2Q} = \pi$ (see Fig. 8.13 **a**). We attribute this to the avoided crossing between $|12\rangle \leftrightarrow |03\rangle$ occurring at $\omega_{\text{int}} + |\alpha|$, where ω_{int} is the fluxing-qubit frequency at the interaction point. For fast-adiabatic flux pulses [34] (with respect to the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing), pulsing the higher frequency qubit to the interaction point results in the near-diabatic passage through $|12\rangle \leftrightarrow |03\rangle$, inducing a Landau-Zener transition in which a small but finite population is transferred from $|12\rangle$ to $|03\rangle$. At the CZ interaction point, the off-resonant interaction between $|12\rangle$ and $|03\rangle$ leads to a further population exchange, with a coupling strength $\sqrt{3}J_1$. Compared to the off-resonant exchange between $|01\rangle$ and $|10\rangle$, this interaction is stronger by a factor $\sqrt{3}$, which can lead to large values of L_3 when combined with the initial population transfer to $|03\rangle$ on the way to the avoided crossing. Furthermore, the phases acquired during the two halves of a Net-Zero pulse can lead to interference Section 6.11.6, increasing or decreasing the $|12\rangle \leftrightarrow |03\rangle$ exchange population. Including the $|12\rangle \leftrightarrow |03\rangle$ crossing leads also to differences in the values of the leakage conditional phases.

We now focus on leakage mobility, which occurs with probability L_{m}, defined as

$$L_{\text{m}} := |\langle 21| \mathscr{S}_{\text{CZ}}(|12\rangle\langle 12|)|21\rangle|^2. \tag{8.52}$$

If $|3\rangle$ is not included, L_{m} takes small but non-negligible values, as shown in Fig. 8.13 **c**. We attribute this to the off-resonant interaction between $|12\rangle$ and $|21\rangle$, with coupling strength $2J_1$. Even though this coupling is stronger than for $|12\rangle \leftrightarrow |03\rangle$, L_{m} is generally smaller than L_3 due to the fluxing qubit not passing through the $|12\rangle \leftrightarrow |21\rangle$ avoided crossing (located at $\omega_{\text{int}} - |\alpha|$) on its way to the CZ interaction point. Including $|3\rangle$, L_{m} can take higher values, as shown in Fig. 8.13, which we associate to a two-excitation

exchange between $|03\rangle$ and $|21\rangle$, virtually mediated by $|12\rangle$. While this is a two-excitation process, $|21\rangle$ and $|03\rangle$ are on resonance at the interaction point, in which case the effective coupling can be estimated as the product of the bare couplings divided by the detuning with $|12\rangle$, i.e.

$$\frac{1}{2\pi}\frac{(2J_1)(\sqrt{3}J_1)}{\alpha} \approx 2.6\,\text{MHz}, \tag{8.53}$$

in analogy to the excitation exchange between a pair of transmons mediated virtually via the bus resonator. Since $|03\rangle$ and $|21\rangle$ are on resonance exactly at the interaction point only when $\alpha_{\text{flux}} = \alpha_{\text{stat}}$, differences in the anharmonicities affect the strength of this exchange.

8.11.7. EFFECTS OF LEAKAGE MOBILITY AND SUPERLEAKAGE ON LEAKAGE DETECTION AND CODE PERFORMANCE

We include leakage mobility in simulations, exploring the range of leakage-mobility probabilities $L_{\text{m}} \in [0, 1.5\%]$ for a fixed leakage probability $L_1 = 0.125\%$ and randomized leakage-conditional phases $\phi^{\mathscr{L}}_{\text{stat}}$ and $\phi^{\mathscr{L}}_{\text{flux}}$ (see Section 8.2). Due to constraints imposed by the size of the density matrix, we only include leakage mobility between the high-frequency data qubits and the ancilla qubits. Thus, we have neglected the possibility of leakage being transferred to the low-frequency data qubits.

Leakage mobility has a negligible effect on the logical performance of the code and the optimality of the HMMs. This is because leakage mobility is only significant in the case of an already leaked qubit, which occurs with a low probability across QEC cycles, given the low L_1 per CZ gate. Thus, the leakage swapping between neighboring qubits can be considered as a second-order effect and has a negligible impact on the logical error rate and HMM optimality extracted from the simulations. We also observe that the average duration of a leakage event on a given qubit is reduced in the presence of leakage mobility.

We now consider the effect of superleakage (see Section 8.11.6) on the logical fidelity and the detection of leakage. We have not performed Surface-17 simulations including $|3\rangle$ on any qubit, since this increases the simulation cost prohibitively. Superleakage is a result of the coherent exchange between $|03\rangle$ and $|12\rangle$, thus individual events are accompanied by a bit flip on a neighboring qubit. The frequency of these events is proportional to the superleakage probability L_3. Superleakage can result in an increase in the observed defect probabilities, increasing the logical error rate of the code, especially without modifications of the decoder to take this into account [18]. However, we do not expect superleakage to significantly affect the detection of leakage. This is because in the case of a leaked data qubit, the anti-commutation of the neighboring stabilizers still holds, leading to a defect probability of 0.5 regardless of the qubit being in $|2\rangle$ or $|3\rangle$ (under the assumption that single-qubit gates act trivially on the leakage subspace). In the case of a leaked ancilla qubit, the propagated bit flips due to superleakage can be considered as a signature of leakage, in addition to the phase errors due to the leakage conditional phases.

8.11.8. LEAKAGE STEADY STATE IN THE SURFACE CODE

Given leakage and seepage probabilities per QEC cycle, it is expected that each qubit in the surface code equilibrates to a steady-state leakage population after many QEC cycles.

Here we do not consider leakage mobility, which is generally small (see Section 8.11.6), allowing to consider a model for a single qubit. We construct a Markovian model to estimate the steady-state populations $p^{\mathscr{C}}$ (resp. $p^{\mathscr{L}}$) in the computational subspace $\mathscr{C}$ (leakage subspace $\mathscr{L}$).

We define $\Gamma_{i\to j}$ as the population-transfer probabilities per QEC cycle. The populations are subject to the constraint $p^{\mathscr{C}} + p^{\mathscr{L}} = 1$. The rate of change of these populations is given by the exchanges from and to each subspace:

$$\begin{aligned} \dot{p}^{\mathscr{C}} &= -p^{\mathscr{C}}\Gamma_{\mathscr{C}\to\mathscr{L}} + p^{\mathscr{L}}\Gamma_{\mathscr{L}\to\mathscr{C}}, \\ \dot{p}^{\mathscr{L}} &= p^{\mathscr{C}}\Gamma_{\mathscr{C}\to\mathscr{L}} - p^{\mathscr{L}}\Gamma_{\mathscr{L}\to\mathscr{C}}. \end{aligned} \tag{8.54}$$

The steady-state condition is $\dot{p}^i = 0$ for $i = \mathscr{C}, \mathscr{L}$, resulting in the steady-state populations p^i_{ss}:

$$\begin{aligned} p^{\mathscr{C}}_{ss} &= \frac{\Gamma_{\mathscr{L}\to\mathscr{C}}}{\Gamma_{\mathscr{C}\to\mathscr{L}} + \Gamma_{\mathscr{L}\to\mathscr{C}}}, \\ p^{\mathscr{L}}_{ss} &= \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}}{\Gamma_{\mathscr{C}\to\mathscr{L}} + \Gamma_{\mathscr{L}\to\mathscr{C}}}. \end{aligned} \tag{8.55}$$

Considering the CZ error model in Section 8.2, for a qubit it approximately holds that

$$\Gamma_{\mathscr{C}\to\mathscr{L}} \approx N_{\text{flux}} L_1, \tag{8.56}$$

$$\Gamma_{\mathscr{L}\to\mathscr{C}} \approx N_{\text{flux}} L_2 + (1 - e^{-\frac{t_c}{T_1/2}}), \tag{8.57}$$

where N_{flux} is in how many CZ gates the qubit is fluxed during a QEC cycle, t_c is the duration of a QEC cycle and L_1 (resp. L_2) is the average leakage (seepage) probability between $\mathscr{C}$ and $\mathscr{L}$ [49]. The use of the average leakage and seepage probabilities per gate is justified for the surface code because, in the case of data-qubit leakage, ancilla qubits are put in an equal superposition during the parity checks, while, in the case of ancilla-qubit leakage, data qubits are in simultaneous entangled eigenstates of the code stabilizers. The seepage probability [Eq. (8.57)] has one contribution from the unitary CZ-gate interaction and one from relaxation during the entire QEC cycle. Regarding the gate contribution, one has $L_2 = 2L_1$ due to the dimensionality ratio between $\mathscr{C}$ and $\mathscr{L}$ for a qubit-qutrit pair [49].

The expected steady-state populations in the simulations can be now computed. We focus on high-frequency data qubits since the low-frequency ones cannot leak without leakage mobility. We have $N_{\text{CZ}} = N_{\text{flux}} = 4$ (for D_4) or 3 (for D_3, D_5), $L_1 = 0.125\%$, $t_c = 800$ ns and $T_1 = 30$ μs. The result is $p^{\mathscr{L}}_{ss}(D_4) = 7.5\%$ and $p^{\mathscr{L}}_{ss}(D_3) = p^{\mathscr{L}}_{ss}(D_5) = 5.9\%$. Furthermore, Eq. (8.54) can be solved to find that the time evolution of $p^{\mathscr{L}}$ towards the steady state is

$$p^{\mathscr{L}}(n) = \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}}{\Gamma_{\mathscr{C}\to\mathscr{L}} + \Gamma_{\mathscr{L}\to\mathscr{C}}}(1 - e^{-(\Gamma_{\mathscr{C}\to\mathscr{L}} + \Gamma_{\mathscr{L}\to\mathscr{C}})n}), \tag{8.58}$$

where n is the QEC cycle number, shown in Fig. 8.14 for the three high-frequency data qubits. We find a good agreement (within error bars) between these predictions and the average leakage population extracted from the density matrix (see Fig. 8.14).

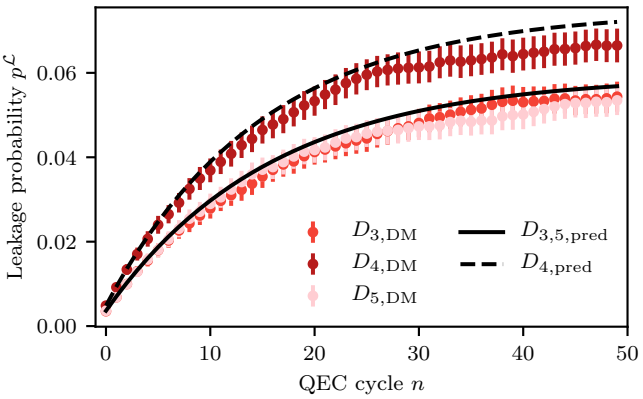

Figure 8.14: Evolution of the average leakage population $p^{\mathcal{L}}$ towards the steady state over 50 QEC cycles for the high-frequency data qubits in Surface-17. The leakage populations extracted from the density-matrix simulation (dots) agree well with the predicted one (black lines). The extracted populations are averaged over 4×10^4 runs. Error bars correspond to 2 standard deviations estimated by bootstrapping.

We now extend the model to the $|3\rangle$ state, despite the fact that we have not included it in simulation due to computational constraints. To do this, we divide the leakage subspace $\mathscr{L}$ into the sub-parts $\mathscr{L}_2$ and $\mathscr{L}_3$ corresponding to leakage in $|2\rangle$ and $|3\rangle$, respectively. The rate equations [Eq. (8.54)] are extended to

$$\begin{aligned}
\dot{p}^{\mathscr{C}} &= -p^{\mathscr{C}}\Gamma_{\mathscr{C}\to\mathscr{L}_2} + p^{\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{C}},\\
\dot{p}^{\mathscr{L}_2} &= p^{\mathscr{C}}\Gamma_{\mathscr{C}\to\mathscr{L}_2} - p^{\mathscr{L}_2}(\Gamma_{\mathscr{L}_2\to\mathscr{C}} + \Gamma_{\mathscr{L}_2\to\mathscr{L}_3}) + p^{\mathscr{L}_3}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2},\\
\dot{p}^{\mathscr{L}_3} &= p^{\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3} - p^{\mathscr{L}_3}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2}.
\end{aligned} \tag{8.59}$$

The steady-state populations $\{p^i_{ss}\}$ then become:

$$\begin{aligned}
p^{\mathscr{C}}_{ss} &= \frac{\Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2}}{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}},\\
p^{\mathscr{L}_2}_{ss} &= \frac{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2}}{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}},\\
p^{\mathscr{L}_3}_{ss} &= \frac{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}}{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}}.
\end{aligned} \tag{8.60}$$

In addition to Eqs. (8.56) and (8.57), in this model we have

$$\Gamma_{\mathscr{L}_2\to\mathscr{L}_3} \approx N_{\mathrm{flux}} L_3/2, \tag{8.61}$$

$$\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} \approx N_{\mathrm{flux}} L_3/2 + (1 - e^{-\frac{t_c}{T_1/3}}). \tag{8.62}$$

The factor of $1/2$ in Eq. (8.61) comes from the fact that superleakage from $\mathscr{L}_2$ to $\mathscr{L}_3$ is possible only when the qubit pair performing the CZ is in $|12\rangle$ and not in $|02\rangle$. For $L_3 = 10\%$, for example, the expected steady-state populations are $p^{\mathscr{L}_2}_{ss}(D_4) = 7.1\%$, $p^{\mathscr{L}_3}_{ss}(D_4) = 5.1\%$

and $p_{ss}^{\mathscr{L}_2}(D_3) = p_{ss}^{\mathscr{L}_2}(D_5) = 5.7\%$, $p_{ss}^{\mathscr{L}_3}(D_3) = p_{ss}^{\mathscr{L}_3}(D_5) = 3.8\%$. While $p_{ss}^{\mathscr{L}_2}$ is almost unchanged with respect to the case without superleakage, $p_{ss}^{\mathscr{L}_3}$ has a comparable magnitude to $p_{ss}^{\mathscr{L}_2}$, suggesting that superleakage needs to be taken into account in optimizing the surface-code performance over many QEC cycles.

REFERENCES

[1] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. DiCarlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code,* npj Quantum Information **6** (2020), 10.1038/s41534-020-00330-w.

[2] A. D. C'orcoles, A. Kandala, A. Javadi-Abhari, D. T. McClure, A. W. Cross, K. Temme, P. D. Nation, M. Steffen, and J. M. Gambetta, *Challenges and opportunities of near-term quantum computing systems,* Proceedings of the IEEE **108**, 1338 (2020).

[3] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, and et al., *Quantum supremacy using a programmable superconducting processor,* Nature **574**, 505–510 (2019).

[4] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Caldwell, N. Didier, E. Schuyler Fried, S. Hong, P. Karalekas, C. B. Osborn, A. Papageorge, E. C. Peterson, G. Prawiroatmodjo, N. Rubin, C. A. Ryan, D. Scarabelli, M. Scheer, E. A. Sete, P. Sivarajah, R. S. Smith, A. Staley, N. Tezak, W. J. Zeng, A. Hudson, B. R. Johnson, M. Reagor, M. P. da Silva, and C. Rigetti, *Unsupervised Machine Learning on a Hybrid Quantum Computer,* arXiv:1712.05771 (2017).

[5] K. A. Landsman, Y. Wu, P. H. Leung, D. Zhu, N. M. Linke, K. R. Brown, L. Duan, and C. Monroe, *Two-qubit entangling gates within arbitrarily long chains of trapped ions,* Phys. Rev. A **100**, 022332 (2019).

[6] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and J. M. Martinis, *Superconducting quantum circuits at the surface code threshold for fault tolerance.* Nature **508**, 500 (2014).

[7] R. Barends, C. M. Quintana, A. G. Petukhov, Y. Chen, D. Kafri, K. Kechedzhi, R. Collins, O. Naaman, S. Boixo, F. Arute, K. Arya, D. Buell, B. Burkett, Z. Chen, B. Chiaro, A. Dunsworth, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, T. Huang, E. Jeffrey, J. Kelly, P. V. Klimov, F. Kostritsa, D. Landhuis, E. Lucero, M. McEwen, A. Megrant, X. Mi, J. Mutus, M. Neeley, C. Neill, E. Ostby, P. Roushan, D. Sank, K. J. Satzinger, A. Vainsencher, T. White, J. Yao, P. Yeh, A. Zalcman, H. Neven, V. N. Smelyanskiy, and J. M. Martinis, *Diabatic gates for frequency-tunable superconducting qubits,* Phys. Rev. Lett. **123**, 210501 (2019).

[8] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo, *Restless tuneup of high-fidelity qubit gates,* Phys. Rev. Applied **7**, 041001 (2017).

[9] Z. Chen, J. Kelly, C. Quintana, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Lucero, E. Jeffrey, A. Megrant, J. Mutus, M. Neeley, C. Neill, P. J. J. O'Malley, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, T. C. White, A. N. Korotkov, and J. M. Martinis, *Measuring and suppressing quantum state leakage in a superconducting qubit,* Phys. Rev. Lett. **116**, 020501 (2016).

[10] M. A. Rol, F. Battistel, F. K. Malinowski, C. C. Bultink, B. M. Tarasinski, R. Vollmer, N. Haider, N. Muthusubramanian, A. Bruno, B. M. Terhal, and L. DiCarlo, *Fast, high-fidelity conditional-phase gate exploiting leakage interference in weakly anharmonic superconducting qubits,* Phys. Rev. Lett. **123**, 120502 (2019).

[11] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Z. Chen, K. Satzinger, R. Barends, F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, S. Boixo, D. Buell, B. Burkett, Y. Chen, R. Collins, E. Farhi, A. Fowler, C. Gidney, M. Giustina, R. Graff, M. Harrigan, T. Huang, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, P. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, E. Lucero, J. McClean, M. McEwen, X. Mi, M. Mohseni, J. Y. Mutus, O. Naaman, M. Neeley, M. Niu, A. Petukhov, C. Quintana, N. Rubin, D. Sank, V. Smelyanskiy, A. Vainsencher, T. C. White, Z. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis (Google AI Quantum), *Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms,* Phys. Rev. Lett. **125**, 120504 (2020).

[12] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M. P. da Silva, and B. R. Johnson, *Demonstration of a parametrically activated entangling gate protected from flux noise,* Phys. Rev. A **101**, 012302 (2020).

[13] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, *Procedure for systematically tuning up cross-talk in the cross-resonance gate,* Physical Review A **93**, 060302 (2016).

[14] T. P. Harty, D. T. C. Allcock, C. J. Ballance, L. Guidoni, H. A. Janacek, N. M. Linke, D. N. Stacey, and D. M. Lucas, *High-fidelity preparation, gates, memory, and readout of a trapped-ion quantum bit,* Phys. Rev. Lett. **113**, 220501 (2014).

[15] E. Jeffrey, D. Sank, J. Y. Mutus, T. C. White, J. Kelly, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. Megrant, P. J. J. O'Malley, C. Neill, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, *Fast accurate state measurement with superconducting qubits,* Phys. Rev. Lett. **112**, 190504 (2014).

[16] C. C. Bultink, M. A. Rol, T. E. O'Brien, X. Fu, B. C. S. Dikken, C. Dickel, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, R. N. Schouten, and L. DiCarlo, *Active resonator reset in the nonlinear dispersive regime of circuit QED,* Phys. Rev. Appl. **6**, 034008 (2016).

[17] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, and C. Eichler, *Rapid high-fidelity multiplexed readout of superconducting qubits,* Phys. Rev. Appl. **10**, 034040 (2018).

[18] J. Kelly, R. Barends, A. Fowler, A. Megrant, E. Jeffrey, T. White, D. Sank, J. Mutus, B. Campbell, Y. Chen, *et al.*, *State preservation by repetitive error detection in a superconducting quantum circuit,* Nature **519**, 66 (2015).

[19] D. Ristè, S. Poletto, M. Z. Huang, A. Bruno, V. Vesterinen, O. P. Saira, and L. DiCarlo, *Detecting bit-flip errors in a logical qubit using stabilizer measurements,* Nat. Commun. **6**, 6983 (2015).

[20] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, J. M. Chow, and J. M. Gambetta, *Demonstration of weight-four parity measurements in the surface code architecture,* Phys. Rev. Lett. **117**, 210505 (2016).

[21] C. C. Bultink, T. E. O'Brien, R. Vollmer, N. Muthusubramanian, M. W. Beekman, M. A. Rol, X. Fu, B. Tarasinski, V. Ostroukh, B. Varbanov, A. Bruno, and L. DiCarlo, *Protecting quantum entanglement from leakage and qubit errors via repetitive parity measurements,* Science Advances **6**, eaay3050 (2020).

[22] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, J. Heinsoo, J.-C. Besse, M. Gabureac, A. Wallraff, and C. Eichler, *Entanglement stabilization using ancilla-based parity detection and real-time feedback in superconducting circuits,* npj Quantum Information **5** (2019).

[23] V. Negnevitsky, M. Marinelli, K. K. Mehta, H.-Y. Lo, C. Flühmann, and J. P. Home, *Repeated multi-qubit readout and feedback with a mixed-species trapped-ion register,* Nature **563**, 527 (2018).

[24] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, N. Lacroix, G. J. Norris, M. Gabureac, C. Eichler, and A. Wallraff, *Repeated quantum error detection in a surface code,* Nature Physics , 1 (2020).

[25] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, *Characterizing quantum supremacy in near-term devices,* Nature Physics **14**, 595–600 (2018).

[26] C. Neill, P. Roushan, K. Kechedzhi, S. Boixo, S. V. Isakov, V. Smelyanskiy, A. Megrant, B. Chiaro, A. Dunsworth, K. Arya, R. Barends, B. Burkett, Y. Chen, Z. Chen, A. Fowler, B. Foxen, M. Giustina, R. Graff, E. Jeffrey, T. Huang, J. Kelly, P. Klimov, E. Lucero, J. Mutus, M. Neeley, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, H. Neven, and J. M. Martinis, *A blueprint for demonstrating quantum supremacy with superconducting qubits,* Science **360**, 195 (2018).

[27] S. Bravyi, D. Gosset, and R. König, *Quantum advantage with shallow circuits,* Science **362**, 308 (2018).

[28] T. O'Brien, B. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise,* npj Quantum Inf. **3** (2017).

[29] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation,* Phys. Rev. A **86**, 032324 (2012).

[30] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, *Simple all-microwave entangling gate for fixed-frequency superconducting qubits,* Phys. Rev. Lett. **107**, 080502 (2011).

[31] V. Tripathi, M. Khezri, and A. N. Korotkov, *Operation and intrinsic error budget of a two-qubit cross-resonance gate,* Phys. Rev. A **100**, 012301 (2019).

[32] F. W. Strauch, P. R. Johnson, A. J. Dragt, C. J. Lobb, J. R. Anderson, and F. C. Wellstood, *Quantum logic gates for coupled superconducting phase qubits,* Phys. Rev. Lett. **91**, 167005 (2003).

[33] L. DiCarlo, J. M. Chow, J. M. Gambetta, L. S. Bishop, B. R. Johnson, D. I. Schuster, J. Majer, A. Blais, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, *Demonstration of two-qubit algorithms with a superconducting quantum processor,* Nature **460**, 240 (2009).

[34] J. M. Martinis and M. R. Geller, *Fast adiabatic qubit gates using only σ_z control,* Phys. Rev. A **90**, 022307 (2014).

[35] S. A. Caldwell, N. Didier, C. A. Ryan, E. A. Sete, A. Hudson, P. Karalekas, R. Manenti, M. P. da Silva, R. Sinclair, E. Acala, N. Alidoust, J. Angeles, A. Bestwick, M. Block, B. Bloom, A. Bradley, C. Bui, L. Capelluto, R. Chilcott, J. Cordova, G. Crossman, M. Curtis, S. Deshpande, T. E. Bouayadi, D. Girshovich, S. Hong, K. Kuang, M. Lenihan, T. Manning, A. Marchenkov, J. Marshall, R. Maydra, Y. Mohan, W. O'Brien, C. Osborn, J. Otterbach, A. Papageorge, J.-P. Paquette, M. Pelstring, A. Polloreno, G. Prawiroatmodjo, V. Rawat, M. Reagor, R. Renzas, N. Rubin, D. Russell, M. Rust, D. Scarabelli, M. Scheer, M. Selvanayagam, R. Smith, A. Staley, M. Suska, N. Tezak, D. C. Thompson, T.-W. To, M. Vahidpour, N. Vodrahalli, T. Whyland, K. Yadav, W. Zeng, and C. Rigetti, *Parametrically activated entangling gates using transmon qubits,* Phys. Rev. Applied **10**, 034050 (2018).

[36] J. Ghosh, A. Galiautdinov, Z. Zhou, A. N. Korotkov, J. M. Martinis, and M. R. Geller, *High-fidelity controlled-σ^Z gate for resonator-based superconducting quantum computers,* Phys. Rev. A **87**, 022309 (2013).

[37] J. Ghosh, A. G. Fowler, J. M. Martinis, and M. R. Geller, *Understanding the effects of leakage in superconducting quantum-error-detection circuits,* Phys. Rev. A **88**, 062329 (2013).

[38] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults,* Quantum Info. Comput. **7**, 139 (2007).

[39] A. G. Fowler, *Coping with qubit leakage in topological codes,* Phys. Rev. A **88**, 042308 (2013).

[40] J. Ghosh and A. G. Fowler, *Leakage-resilient approach to fault-tolerant quantum computing with superconducting elements,* Phys. Rev. A **91**, 020302 (2015).

[41] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code,* Quantum Info. Comput. **15**, 997 (2015).

[42] N. C. Brown and K. R. Brown, *Comparing zeeman qubits to hyperfine qubits in the context of the surface code:* $^{174}Yb^{+}$ *and* $^{171}Yb^{+}$, Phys. Rev. A **97**, 052301 (2018).

[43] N. C. Brown, M. Newman, and K. R. Brown, *Handling leakage with subsystem codes,* New Journal of Physics **21**, 073055 (2019).

[44] N. C. Brown and K. R. Brown, *Leakage mitigation for quantum error correction using a mixed qubit scheme,* Phys. Rev. A **100**, 032325 (2019).

[45] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, *Simple pulses for elimination of leakage in weakly nonlinear qubits,* Phys. Rev. Lett. **103**, 110501 (2009).

[46] D. Hayes, D. Stack, B. Bjork, A. C. Potter, C. H. Baldwin, and R. P. Stutz, *Eliminating leakage errors in hyperfine qubits,* Phys. Rev. Lett. **124**, 170501 (2020).

[47] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, *Scalable quantum circuit and control for a superconducting surface code,* Phys. Rev. Appl. **8**, 034021 (2017).

[48] The quantumsim package can be found at `https://quantumsim.gitlab.io/`.

[49] C. J. Wood and J. M. Gambetta, *Quantification and characterization of leakage errors,* Phys. Rev. A **97**, 032306 (2018).

[50] S. T. Spitz, B. Tarasinski, C. W. J. Beenakker, and T. E. O'Brien, *Adaptive weight estimator for quantum error correction in a time-dependent environment,* Advanced Quantum Technologies **1**, 1800012 (2018).

[51] J. M. Auger, H. Anwar, M. Gimeno-Segovia, T. M. Stace, and D. E. Browne, *Fault-tolerance thresholds for the surface code with fabrication errors,* Phys. Rev. A **96**, 042316 (2017).

[52] S. Nagayama, A. G. Fowler, D. Horsman, S. J. Devitt, and R. V. Meter, *Surface code error correction on a defective lattice,* New Journal of Physics **19**, 023050 (2017).

[53] T. M. Stace and S. D. Barrett, *Error correction and degeneracy in surface codes suffering loss,* Phys. Rev. A **81**, 022317 (2010).

[54] S. Bravyi, G. Duclos-Cianci, D. Poulin, and M. Suchara, *Subsystem surface codes with three-qubit check operators,* Quantum Info. Comput. **13**, 963–985 (2013).

[55] D. Egger, M. Werninghaus, M. Ganzhorn, G. Salis, A. Fuhrer, P. Müller, and S. Filipp, *Pulsed reset protocol for fixed-frequency superconducting qubits,* Phys. Rev. Applied **10**, 044030 (2018).

[56] P. Magnard, P. Kurpiers, B. Royer, T. Walter, J.-C. Besse, S. Gasparinetti, M. Pechal, J. Heinsoo, S. Storz, A. Blais, and A. Wallraff, *Fast and unconditional all-microwave reset of a superconducting qubit,* Phys. Rev. Lett. **121**, 060502 (2018).

[57] F. Luthi, T. Stavenga, O. W. Enzing, A. Bruno, C. Dickel, N. K. Langford, M. A. Rol, T. S. Jespersen, J. Nygård, P. Krogstrup, and L. DiCarlo, *Evolution of nanowire transmon qubits and their coherence in a magnetic field,* Phys. Rev. Lett. **120**, 100502 (2018).

[58] D. Sank, Z. Chen, M. Khezri, J. Kelly, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. Fowler, E. Jeffrey, E. Lucero, A. Megrant, J. Mutus, M. Neeley, C. Neill, P. J. J. O'Malley, C. Quintana, P. Roushan, A. Vainsencher, T. White, J. Wenner, A. N. Korotkov, and J. M. Martinis, *Measurement-induced state transitions in a superconducting qubit: Beyond the rotating wave approximation,* Phys. Rev. Lett. **117**, 190503 (2016).

9

Hardware-Efficient Leakage-Reduction Scheme for Quantum Error Correction with Superconducting Transmon Qubits

Leakage outside of the qubit computational subspace poses a threatening challenge to quantum error correction (QEC). We propose a scheme using two leakage-reduction units (LRUs) that mitigate these issues for a transmon-based surface code, without requiring an overhead in terms of hardware or QEC-cycle time as in previous proposals. For data qubits we consider a microwave drive to transfer leakage to the readout resonator, where it quickly decays, ensuring that this negligibly disturbs the computational states for realistic system parameters. For ancilla qubits we apply a $|1\rangle \leftrightarrow |2\rangle$ π *pulse conditioned on the measurement outcome. Using density-matrix simulations of the distance-3 surface code we show that the average leakage lifetime is reduced to almost* 1 *QEC cycle, even when the LRUs are implemented with limited fidelity. Furthermore, we show that this leads to a significant reduction of the logical error rate. This LRU scheme opens the prospect for near-term scalable QEC demonstrations.*

This chapter has been published in PRX Quantum **2**, 030314 (2021) [1]. F. B. performed the study and the writing with input from all co-authors.

9.1. INTRODUCTION

Quantum computing has recently reached the milestone of quantum supremacy [2] thanks to a series of improvements in qubit count [3, 4], gate fidelities [5–16] and measurement fidelities [17–19]. The next major milestones include showing a quantum advantage [20–23] and demonstrating quantum error correction (QEC) [4, 24–32]. Errors accumulate over time in a quantum computer, leading to an entropy increase which severely hinders the accuracy of its output. Thus QEC is necessary to correct errors and remove entropy from the computing system. If the overall physical error rate is below a certain noise threshold for a given QEC-code family, the logical error rate decreases exponentially with the code distance d at the price of a poly(d) overhead, thus allowing to extend the computational time. Recently, small-size instances of error-detecting [30, 31] and error-correcting [4] codes have been experimentally realized. To further demonstrate fault tolerance it is crucial to scale up these systems and show that larger distance codes consistently lead to lower logical error rates than smaller distance codes [32].

Leakage outside of the computational subspace [9–11, 13, 33–38], present in leading quantum-computing platforms such as superconducting qubits and trapped ions, poses a particularly threatening challenge to fault tolerance [24, 39–49]. Leakage can increase entropy by making measurement outcomes no longer point to the underlying errors and can effectively reduce the code distance (see Section 8.11.2). Furthermore, leakage can last for many QEC cycles [41], making operations on a leaked qubit fail and possibly spread correlated errors through the code [32, 40, 47]. In particular, leakage falls outside the stabilizer formalism of QEC as it cannot be decomposed in terms of Pauli errors. Stabilizer codes [50, 51] and their decoders are thus typically ill-suited to deal with leakage, leading to a significant increase of the logical error rate [43, 46] (see also Fig. 8.2). If the average leakage lifetime $l_{\text{avg}}^{\mathscr{L}}$, that is, the average number of QEC cycles that a qubit stays leaked (after leaking in the first place), fulfills $l_{\text{avg}}^{\mathscr{L}} = \mathcal{O}(1)$ QEC cycles and $l_{\text{avg}}^{\mathscr{L}} \ll d$, then for low-enough error rates a threshold is likely to exist [40] as leakage would have a relatively local effect in space and time. Due to a finite energy-relaxation time, leakage does indeed last for $l_{\text{avg}}^{\mathscr{L}} = \mathcal{O}(1)$ QEC cycles. However, in practice it is important how large $l_{\text{avg}}^{\mathscr{L}}$ is, since if it is low the noise threshold is expected to be higher. Shortening the relaxation time to reduce $l_{\text{avg}}^{\mathscr{L}}$ is not effective as this increases the physical error rate as well.

A leakage-reduction unit (LRU) [39, 40, 42, 43, 48, 49, 52, 53] is an operation introducing a seepage mechanism besides that of the relaxation channel. A LRU converts leakage into regular (Pauli) errors and shortens the average leakage lifetime, ideally to 1 QEC cycle. As discussed above, this is expected to lead to a higher noise threshold, but not as high as for the case without leakage, since the leakage rate effectively adds to the regular error rate thanks to the LRU. As an alternative to the use of LRUs, post-selection based on leakage detection has been adopted (see Section 8.8) as a near-term method to reduce the logical error rate. While leakage detection could also be used to apply LRUs in a targeted way, post-selection is not scalable. By shortening the lifetime to $l_{\text{avg}}^{\mathscr{L}} = \mathcal{O}(1) \ll d$, the use of LRUs is instead a scalable approach.

In its imperfect experimental implementation a LRU can either introduce extra Pauli errors or mistakenly induce leakage on a non-leaked qubit. Furthermore, in the context of the surface code the LRUs investigated so far [42, 43, 48] introduce an overhead in terms of hardware and QEC-cycle time. Specifically, these LRUs are variants of the swap-LRU,

in which the qubits are swapped at the end of each QEC cycle, taking alternatively the role of data and ancilla qubits. In this way every qubit is measured every 2 QEC cycles. The core of the swap-LRU is the fact that the measured qubits are reset to the computational subspace after the measurement. This can be accomplished by a scheme which unconditionally maps $|1\rangle$ and $|2\rangle$ (and possibly $|3\rangle$ [49]) to $|0\rangle$ [54–56], or conditionally using real-time feedback [29, 57]. Under the standard assumption that the SWAP gates swap the states of two qubits only if none of them is leaked (which does not necessarily hold in experiment [49]), $l^{\mathscr{L}}_{\text{avg}}$ is ideally shortened to 2 QEC cycles. On the downside, for the pipelined surface-code scheme in [58], the pipeline is disrupted as qubits cannot be swapped until the measurement and reset operations are completed, leading overall to an increase up to 50% of the QEC-cycle time depending on the reset time. The extra CZ gates, needed to implement the SWAPs, can cause additional errors or leakage as the CZ is the major source of leakage in transmons [9–11, 13, 33–36]. Moreover, in the surface code an extra row of qubits is needed to perform all the SWAPs [42], which is a non-negligible overhead in the near term. All these issues increase the physical error rate by a considerable amount, thus requiring to increase the system size to compensate for that (assuming that the error rates are still below threshold).

In this chapter we propose two separate LRUs for data and ancilla qubits which use resources already available on chip, namely the readout resonator for data qubits (res-LRU) and a $|1\rangle \leftrightarrow |2\rangle$ π pulse conditioned on the measurement outcome for ancilla qubits (π-LRU). In particular, the use of the res-LRU avoids the necessity to swap data and ancilla qubits to be able to reset the data qubits. The res-LRU is a modification of the two-drive scheme in [54–56] to a single drive to deplete only the population in $|2\rangle$ but not $|1\rangle$, making it a LRU rather than a reset scheme. We additionally show that this negligibly affects the coherence within the computational subspace in an experimentally accessible regime, with a low probability of mistakenly inducing leakage as long as the thermal population in the readout resonator is relatively small. This allows us to unconditionally use res-LRU in the surface code in every QEC cycle. In the pipelined scheme [58] the res-LRU easily fits within the time in which the data qubits are idling while the ancilla qubits are finishing to be measured. As the π-LRU can be executed as a short pulse at the end of the measurement time with real-time feedback, our LRU scheme overall does not require any QEC-cycle time overhead. Using density-matrix simulations [47, 51, 59] of the distance-3 surface code (Surface-17), we show that the average leakage lifetime is reduced to almost 1 QEC cycle when res-LRU and π-LRU with realistic performance are employed. Furthermore, compared to the case without LRUs, the logical error rate is reduced by up to 30%. The proposed res-LRU and π-LRU can be straightforwardly adapted to QEC-code schemes other than [58] and the res-LRU is potentially applicable to superconducting qubits with higher anharmonicity than transmons. The demonstrated reduction serves as evidence of scalability for our LRU scheme, even though we cannot estimate a noise threshold as we have simulated only one size of the surface code. To explore larger codes it is necessary to use less computationally expensive simulations [24, 40, 43] that use a simplified version of our error model at the cost of losing some information contained in the density matrix. Furthermore, to optimize the noise threshold the LRUs can be supplied with a leakage-aware decoder [24, 40, 43, 60–62] that uses measurement information about leakage to better correct leakage-induced correlated errors.

9.2. READOUT-RESONATOR LRU

The readout resonator has been used [54–56] to reset a transmon qubit to the $|0\rangle$ state, depleting the populations in $|1\rangle$ and $|2\rangle$. Targeting the $|20\rangle \leftrightarrow |01\rangle$ transition, with the notation $|\text{transmon, resonator}\rangle$, those populations are swapped onto the readout resonator, where they quickly decay due to the strong coupling to the transmission-line environment. Ref. [54] uses two drives simultaneously while Refs. [55, 56] use these drives in a three-step process. Here we adapt these techniques to use a single drive in a single step to deplete the population in $|2\rangle$ only.

A LRU is defined [39] as an operation such that 1) the incoming leakage population is reduced after the application of the LRU, 2) the induced leakage when applied to a non-leaked state is ideally 0 (see also Section 5.1.2). We thus ensure below that not only leakage is reduced but also that the effect that the drive has on a non-leaked transmon is as small as possible.

9.2.1. TRANSMON-RESONATOR HAMILTONIAN

We consider a transmon capacitively coupled to a resonator and to a dedicated microwave drive line. The resonator possibly employs a Purcell filter which we do not include explicitly. In a frame rotating at the transmon-drive frequency ω_d for both the resonator and the transmon, the Hamiltonian is time-independent and is given by

$$H = H_0 + H_c + H_d \tag{9.1}$$

$$H_0 = \delta^r a^\dagger a + \delta^q b^\dagger b + \frac{\alpha}{2}(b^\dagger)^2 b^2 \tag{9.2}$$

$$H_c = g(ab^\dagger + a^\dagger b) \tag{9.3}$$

$$H_d = \frac{\Omega}{2}(e^{i\phi} b + e^{-i\phi} b^\dagger) \tag{9.4}$$

where a and b are the creation operators for the resonator and the transmon, respectively; $\delta^r = \omega_r - \omega_d$ and $\delta^q = \omega_q - \omega_d$ with ω_r and ω_q the resonator and transmon frequencies, respectively; $\alpha < 0$ is the transmon anharmonicity; g corresponds to the capacitive coupling; Ω and ϕ are the transmon-drive amplitude and phase, respectively. The phase is not relevant for the results in this chapter and we fix it to $\phi = 0$.

We can qualitatively understand (see Fig. 9.1(a)) that H contains an effective coupling $\tilde{g}$ between $|20\rangle$ and $|01\rangle$. If ω_d matches the transition frequency between the "bare" $|20\rangle$ and $|01\rangle$, these two states are degenerate in the rotating frame and they are connected by two paths (at lowest order) via either $|11\rangle$ or $|10\rangle$. If $\Delta := \omega_q - \omega_r \gg g$ and $\delta^q \gg \Omega$, then $|11\rangle$ and $|10\rangle$ are occupied only "virtually" and one gets purely an effective $|20\rangle \leftrightarrow |01\rangle$ coupling. Modulo a constant term, in the 2D subspace $\mathscr{S} = \text{span}\{|20\rangle, |01\rangle\}$ we can write H in Eq. (9.1) as $H|_{\mathscr{S}} \equiv -\eta(\omega_d)Z/2 + \tilde{g}(\omega_d)X$ for an appropriate function η (an approximation can be extracted from Eq. (9.62)). As a function of ω_d this Hamiltonian gives rise to a $|20\rangle \leftrightarrow |01\rangle$ avoided crossing centered at a frequency ω_d^* (see Fig. 9.1(b)) where $\eta(\omega_d^*) = 0$. The energy separation at the center of the avoided crossing is then $2\tilde{g}(\omega_d^*)$.

In order to quantitatively study the action of H, we unitarily transform it using a Schrieffer-Wolff transformation e^S [63–66]. Let $\{|ij\rangle_D\}$ be the basis of eigenvectors

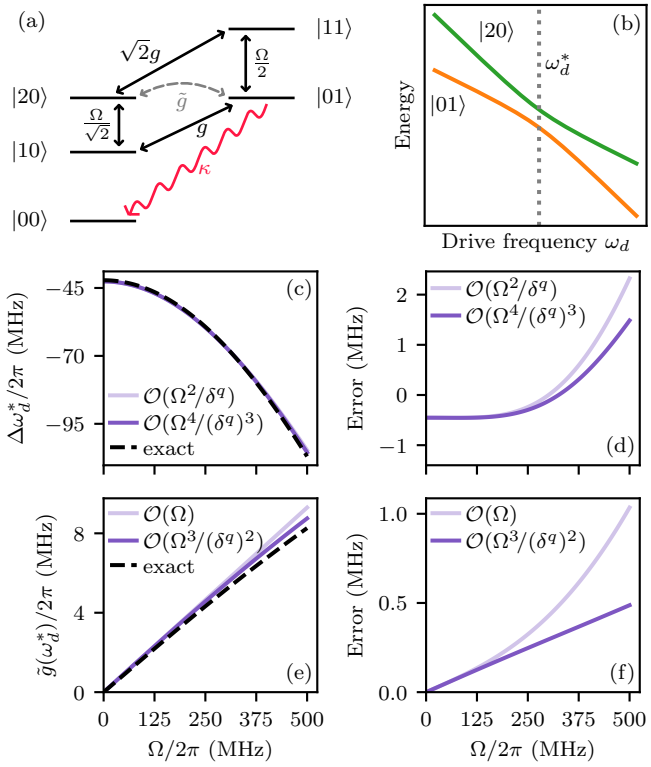

Figure 9.1: Concept of the readout-resonator LRU. (a) The state $|20\rangle$ (with the notation $|\text{transmon, resonator}\rangle$) is connected to $|01\rangle$ by two main paths via either $|11\rangle$ or $|10\rangle$, due to the capacitive coupling g or the transmon-drive amplitude Ω, respectively. This generates an effective coupling $\tilde{g}$ which can be used to swap $|20\rangle \leftrightarrow |01\rangle$. The latter quickly decays to $|00\rangle$ due to the typically high coupling κ of the readout resonator to the transmission-line environment, overall removing leakage from a leaked transmon. (b) In the rotating frame of the drive, $|20\rangle$ and $|01\rangle$ show an avoided crossing as a function of the drive frequency ω_d, centered at ω_d^*. The effective coupling $\tilde{g}(\omega_d^*)$ is equal to half the energy separation at that point. (c),(e) $\Delta\omega_d^* := \omega_d^* - (2\omega_q + \alpha - \omega_r)$ and $\tilde{g}(\omega_d^*)$ are respectively evaluated either exactly by full numerical diagonalization of H in Eq. (9.1), or by approximate analytical formulas (see Section 9.2.1 and Section 9.5) for the parameters in Table 9.1. The absolute errors with respect to the exact curves are shown in (d),(f) respectively.

of $H_0 + H_c$ (the transmon-resonator "dressed" basis). In the dispersive regime ($g \ll \Delta$), with respect to a 1st-order Schrieffer-Wolff transformation S_1 in the perturbation parameter g/Δ, such that $e^{-S_1}|ml\rangle \approx |ml\rangle_D$, we get (see Section 9.5)

$$H^D := e^S H e^{-S} \approx e^{S_1} H e^{-S_1} \tag{9.5}$$

$$= H_0^D + H_{d1}^D + H_{d2}^D \tag{9.6}$$

with

$$H_0^D = \Big(\delta^r - \sum_{m=0}^{\infty} \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m|\Big) a^\dagger a + \sum_{m=1}^{\infty} \Big(m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}}\Big) |m\rangle\langle m| \tag{9.7}$$

$$H_{d1}^D = \frac{\Omega e^{i\phi}}{2} b + \text{h.c.} \tag{9.8}$$

$$H_{d2}^D = \frac{\Omega e^{i\phi}}{2} \Big(a \sum_{m=0}^{\infty} \frac{g\Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| + a^\dagger \sum_{m=0}^{\infty} \frac{g\alpha\sqrt{m+1}\sqrt{m+2}}{\Delta_m \Delta_{m+1}} |m\rangle\langle m+2| \Big) + \text{h.c.}, \tag{9.9}$$

where $\Delta_m := \Delta + \alpha m$ and $\{|m\rangle\}$ are transmon states. H_0^D is diagonal and contains the dispersive shifts, H_{d1}^D is the transmon drive now in the unitarily transformed frame, H_{d2}^D contains an indirect resonator drive and couplings of the kind $a^\dagger |m\rangle\langle m+2|$. In particular, for $m = 0$ in Eq. (9.9) we get a lowest order approximation of $\tilde{g}$:

$$\tilde{g} \approx \frac{\Omega g \alpha}{\sqrt{2}\Delta(\Delta + \alpha)}. \tag{9.10}$$

Notice that at this order there is no dependence on ω_d. Furthermore, $\tilde{g}$ would vanish for $\alpha = 0$, since the two paths in Fig. 9.1(a) fully destructively interfere in that case. Since α is low for transmons, one can expect that Ω needs to be relatively large for $\tilde{g}$ to be substantial.

For the drive to be most effective it is important that ω_d matches ω_d^*. If $g = 0 = \Omega$, there is no avoided crossing but $|20\rangle$ and $|01\rangle$ simply cross at $\omega_{d,0}^* \equiv 2\omega_q + \alpha - \omega_r$ as can be straightforwardly computed from H_0 in Eq. (9.2). This value is shifted due to the capacitive coupling (as can be seen from Eq. (9.7)), as well as due to the possibly strong drive. For $g \neq 0$ and $\Omega \neq 0$ one can either compute ω_d^* by full numerical diagonalization of H and find the avoided crossing as a function of ω_d, or one can find an (approximate) analytical expression. For the latter we use another Schrieffer-Wolff transformation (rather than the resolvent method in Ref. [55], which does not give the full Hamiltonian) to account for the effect of the transmon drive H_{d1}^D and to compute ω_d^* up to order $\Omega^4/(\delta^q)^3$, see Section 9.5. We also use this transformation to compute $\tilde{g}$ up to order $\Omega^3/(\delta^q)^2$. Figures 9.1(c),(e) compare the analytical approach with the exact numerical results for $\Delta\omega_d^* = \omega_d^* - \omega_{d,0}^*$ and $\tilde{g}(\omega_d^*)$, respectively, given the parameters in Table 9.1. We consider 6 energy levels for

Parameter	Transmon	Readout resonator
Frequency $\omega/2\pi$	6.7 GHz	7.8 GHz
Anharmonicity $\alpha/2\pi$	−300 MHz	n.a.
Coupling $g/2\pi$	135 MHz	
Avg. photon number $\bar{n}$	n.a.	0.005
Relaxation time T_1	30 μs	16 ns ($\kappa/2\pi = 10$ MHz)
Dephasing time T_2	30 μs (flux noise)	32 ns

Table 9.1: Parameters used both in the analysis and Lindblad simulations of the readout-resonator LRU, similar to the experimental ones in Ref. [28]. The transmon parameters correspond to the target parameters of a high-frequency data qubit in Section 9.3.

the transmon and 3 for the resonator as we see that the exact curves converge for such choice. In Fig. 9.1(c)(d) we see that the two approximations are both pretty good, while in Fig. 9.1(e)(f) we see that Eq. (9.10) deviates by up to 1 MHz from the exact value at high Ω and that the absolute error with respect to the exact $\tilde{g}(\omega_d^*)$ scales in a seemingly quadratic way. Instead, the higher order approximation stays closer to the exact curve and the error scales linearly. We expect that the remaining gap would be mostly filled by considering also higher orders in g/Δ in the first Schrieffer-Wolff transformation, since increasing only the order of approximation in Ω/δ^q does not provide a significant improvement in Fig. 9.1(d).

9.2.2. PERFORMANCE OF THE READOUT-RESONATOR LRU

Given the theoretical understanding of the transmon-resonator system, we devise a pulse to minimize the population in $|2\rangle$ on a leaked transmon. We consider the pulse shape

$$\Omega(t) = \begin{cases} \Omega \sin^2(\pi \frac{t}{2t_{\text{rise}}}) & \text{for} \quad 0 \le t \le t_{\text{rise}} \\ \Omega & \text{for} \quad t_{\text{rise}} \le t \le t_{\text{p}} - t_{\text{rise}} \\ \Omega \sin^2(\pi \frac{t_{\text{p}}-t}{2t_{\text{rise}}}) & \text{for} \quad t_{\text{p}} - t_{\text{rise}} \le t \le t_{\text{p}} \end{cases} \tag{9.11}$$

similarly to Ref. [55], where t_{p} is the total pulse duration, at a fixed frequency $\omega_d(t) = \omega_d$. Hence, there are four parameters to optimize over, i.e. $\Omega, \omega_d, t_{\text{p}}$ and t_{rise}. We fix $t_{\text{rise}} = 30$ ns since we observe that this strongly suppresses non-adiabatic transitions out of the manifold of interest: for example, $|20\rangle$ is coupled to $|10\rangle$ by the drive but they are quite off-resonant, so only a fast pulse can cause "non-virtual" transitions between them. Indeed, for $t_{\text{rise}} \lesssim 10$ ns there appear ripples (for an example see Ref. [55]) in e.g. the $|20\rangle$ and $|10\rangle$ populations when the drive is turned on and off, leading to a reduction in performance. We expect that an improved pulse shape can shorten t_{rise}. However, we do not explore this given the long maximum t_{p} allowed in our surface-code scheme ($t_{\text{p}} \le T_{\text{slot}} = 440$ ns, see Section 9.3.1).

We use Lindblad simulations of the transmon-resonator system to optimize over Ω, ω_d

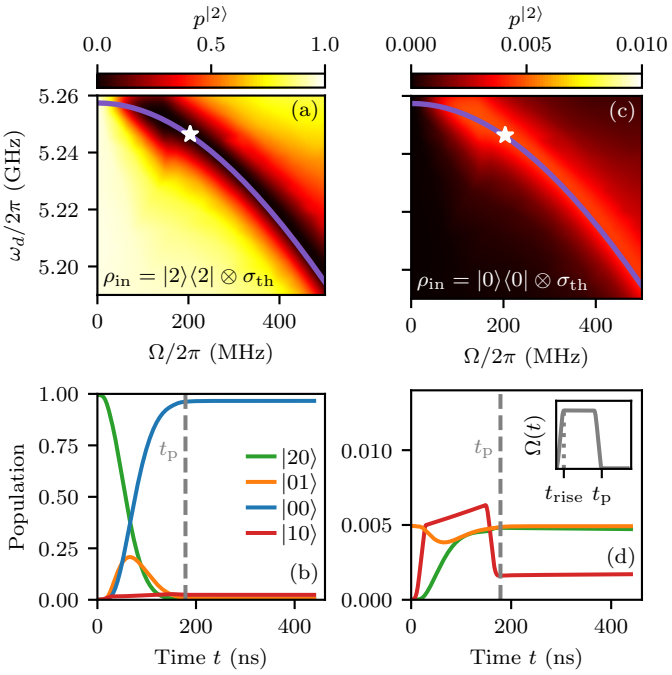

Figure 9.2: Lindblad simulations of the transmon-resonator system for the readout-resonator LRU. In (a),(b) the initial state is $|2\rangle\langle 2| \otimes \sigma_{\text{th}}$, while in (c),(d) it is $|0\rangle\langle 0| \otimes \sigma_{\text{th}}$, where σ_{th} is the resonator thermal state. (a),(c) Transmon leakage population $p^{|2\rangle} = \langle 2|\,\text{Tr}_r(\rho(T_{\text{slot}}))\,|2\rangle$ at the end of the time slot of $T_{\text{slot}} = 440$ ns. For each choice of (Ω, ω_d) we optimize the total pulse duration $t_{\text{p}} \leq T_{\text{slot}}$ to minimize $p^{|2\rangle}$ given the initial state $|2\rangle\langle 2| \otimes \sigma_{\text{th}}$, for fixed $t_{\text{rise}} = 30$ ns. The white star indicates the chosen operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_{\text{p}} = 178.6$ ns) with $p^{|2\rangle}_{\text{op.}} \approx 0.5\%$ in (a). The induced leakage in (c) is $p^{|2\rangle} \approx 0.48\%$ at the operating point. The purple line corresponds to the higher order estimate of the optimal drive frequency ω_d^* as a function of Ω (see Fig. 9.1(c)). The heatmaps are sampled using the *adaptive* package [67]. (b),(d) Time evolution of the populations in a few selected states for the operating point. The vertical dashed line indicates the used t_{p}. The inset in (d) shows a schematic of the pulse $\Omega(t)$.

and t_p. The Lindblad equation is given by

$$\dot{\rho} = -i\left[H^D, \rho\right] + \sum_j \left(L_j \rho L_j^\dagger - \frac{1}{2}\{L_j^\dagger L_j, \rho\}\right) \tag{9.12}$$

with $\{L_j\}$ the quantum jump operators. We express (and solve) this equation in the exact unitarily transformed frame. That is, while in Section 9.2.1 we have used a first-order Schrieffer-Wolff transformation e^{S_1} (see Eq. (9.5)), in the numerics we compute the full transformation e^S (see also Eq. (9.5)). In this way we find the basis that exactly diagonalizes $H_0 + H_c$ and express H_d in this basis as well, without any further Schrieffer-Wolff transformation like in Section 9.2.1. In other words, the simulations reproduce the dynamics under the Hamiltonian in Eqs. (9.1) to (9.4) without any approximation.

The Hamiltonian parameters are the same as in Section 9.2.1 and are reported in Table 9.1, including the noise parameters. In particular, while we neglect the transmon thermal population, we include it for the resonator since it determines the leakage that the pulse induces when the transmon was not leaked, as we discuss below. The resonator thermal state is given by [68]

$$\sigma_{\text{th}} \approx \left(1 - \frac{\bar{n}}{1+2\bar{n}}\right)|0\rangle\langle 0| + \frac{\bar{n}}{1+2\bar{n}}|1\rangle\langle 1| \tag{9.13}$$

for low average photon number $\bar{n}$. We consider dressed relaxation and dephasing, as given below, assuming that this is a good model in the dispersive regime. In the unitarily rotated frame, the employed jump operators $\{L_j\}$ are explicitly given by

$$\frac{1}{\sqrt{T_1^r}} a = \sqrt{\kappa}\, a, \quad \sqrt{\frac{\bar{n}}{1+\bar{n}}}\sqrt{\kappa}\, a^\dagger, \quad \sqrt{\frac{2}{T_\phi^r}}\, a^\dagger a, \tag{9.14}$$

$$\frac{1}{\sqrt{T_1^q}} b, \quad \sqrt{\frac{2}{T_\phi^q}}\, b^\dagger b, \tag{9.15}$$

where $T_\phi = (1/T_2 - 1/2T_1)^{-1}$ and where we consider 6 energy levels for the transmon and 3 for the resonator. Note that e.g. for a, going back to the original frame it holds that $e^{-S} a e^S = \sum_{l=0}^{1} \sqrt{l+1}\,|l\rangle_D\langle l+1|_D = a_D$ by definition of e^S, corresponding indeed to relaxation in the dressed basis. By considering dressed relaxation and dephasing, the effective relaxation time T_1^q of the transmon is not shortened by the fact that it is coupled to a lossy resonator (Purcell effect). We assume that this is a good approximation also during driving as the drive couples eigenstates which mostly have the same number of excitations in the resonator (except for $|20\rangle$ and $|01\rangle$ when the drive is near-resonant with this transition and causes a strong mixing of these states). We thus mimic the use of a Purcell filter but without including it in the simulations since that would increase the Hilbert-space dimension in a computationally expensive way.

For each choice of (Ω, ω_d) we optimize t_p such that, given the initial state $|2\rangle\langle 2| \otimes \sigma_{\text{th}}$, the leakage population $p^{|2\rangle} = \langle 2|\,\text{Tr}_r(\rho(T_{\text{slot}}))|2\rangle$ at the end of the available time slot is minimized (see Fig. 9.2(a)). The states $|20\rangle$ and $|01\rangle$ approximately form a two-level system with additional damping from $|01\rangle$ to $|00\rangle$, thus the drive effectively induces

damped Rabi oscillations [69] between them. Oscillations occur only for $\tilde{g} > \kappa/4$ [69] (underdamped regime), while for $\tilde{g} = \kappa/4$ (critical regime) or $\tilde{g} < \kappa/4$ (overdamped regime) the populations in $|20\rangle$ and $|01\rangle$ simply decay in an exponential-like way without forming any minimum. For the parameters in Table 9.1 the critical drive amplitude that gives $\tilde{g} = \kappa/4$ is $\Omega_{\text{cr}}/2\pi \approx 143$ MHz. Thus for $\Omega \leq \Omega_{\text{cr}}$ the best strategy is to drive until $p^{|2\rangle}$ reaches a (low) practically-stable value (which is in general not 0 when the full system is taken into account). Here with the given κ we find that this occurs in a time comparable to T_{slot} only from about $\Omega = \Omega_{\text{cr}}$, so for $\Omega \leq \Omega_{\text{cr}}$ we drive for the entire T_{slot}. For $\Omega > \Omega_{\text{cr}}$ the optimization has many local minima as a function of t_{p}, corresponding to the minima of the $|20\rangle \leftrightarrow |01\rangle$ oscillations induced by the drive. Here we choose to target the first minimum as in Refs. [55, 56] since it is the fastest approach. For a sudden pulse this minimum would occur around $\pi/2\tilde{g}$ for sufficiently small κ, whereas we find heuristically that a good initial guess for the optimization is $\pi/2\tilde{g}_{\text{damp}}$ with $\tilde{g}_{\text{damp}} := \sqrt{\tilde{g}^2 - (\kappa/4)^2}\, e^{-\kappa/7\tilde{g}}$ for larger κ. Then for the optimization over t_{p} we use the bounds $t_{\text{p}} - 2t_{\text{rise}} \in [0, 1.1 \times \pi/2\tilde{g}_{\text{damp}}]$ (using the bounded Brent method in *scipy*; we provide the code at `https://doi.org/10.4121/14762052`). While using a longer t_{p} in the underdamped regime (possibly even greater than the allotted T_{slot}) would eventually lead to an even lower leakage population [54], it is not necessarily desirable as a longer t_{p} may mean that the disturbance to a non-leaked transmon is greater as well (see Section 9.6.2).

While the procedure above optimizes t_{p} given a certain pair (Ω, ω_d), we use the package *adaptive* [67] to choose the next pair to sample and we iterate this process. This package searches a given parameter space (here $\Omega/2\pi \in [0, 500\text{ MHz}]$, $\omega_d/2\pi \in [5.19, 5.26\text{ GHz}]$) in a finer way where the given cost function changes faster. Here we use $(\log p^{|2\rangle})^2$ as the cost function since it changes faster where $p^{|2\rangle}$ is small, allowing us to get both a high-resolution heatmap (see Fig. 9.2) and a good first estimation of the $p^{|2\rangle}$ minima in a single run. Then we run a local optimization with tight bounds around some of these candidate points for fine tuning.

In Fig. 9.2(a) one can observe a band with low $p^{|2\rangle}$ as desired. This band occurs at drive frequencies slightly above $\omega_d^*(\Omega)$, which one would expect to be optimal based on Section 9.2.1. We attribute this to the fact that a significant share of the time is taken by the rise and fall of the pulse, where $\Omega(t)$ is smaller than the maximum. We find that one can choose a broad range of Ωs to achieve a $p^{|2\rangle} \gtrsim 0.5\%$, from 130 MHz (slightly below the critical point) to deep in the underdamped regime. However, other considerations apply, namely, on the high end using a very high Ω poses strong experimental requirements on the drive, while on the low end the pulse takes much longer and it is not a priori given that driving at the critical point would be best. Actually, notice that driving at the critical point with good performance is possible only due to the relatively high T_{slot} for the given κ. In the following we choose the point marked by a star in Fig. 9.2 as the operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_{\text{p}} = 178.6$ ns). This point reaches $p_{\text{op.}}^{|2\rangle} \approx 0.5\%$ while affecting the least the coherence within the computational subspace (see Section 9.6.1). We attribute the fact that this minimum does not reach 0 to re-heating from $|00\rangle$ to $|01\rangle$, as well as transmon decoherence (resonator pure dephasing would contribute as well but here $T_\phi^r = \infty$) and interactions with higher energy levels. We note that in Fig. 9.2(a) we find good $p^{|2\rangle} \lesssim 5\%$ up to $\Omega/2\pi \gtrsim 100$ MHz, which could be used to further ease the requirements on the drive (see Section 9.3).

The time evolution for a few selected states is shown in Fig. 9.2(b) for the operating point, given the initial state $|2\rangle\langle 2| \otimes \sigma_{\text{th}}$. The first few ns make $|20\rangle$ rotate into $|01\rangle$, while the latter decays relatively fast to $|00\rangle$ due to the large relaxation rate κ of the readout resonator. Already after ≈ 220 ns the remaining $|01\rangle$ population has practically returned to the thermal state. The repetition of the pulse, such as in the surface code (see Section 9.3) at every QEC cycle, thus does not lead to heating of the resonator with these system parameters (see Section 9.4 for a discussion about other parameter regimes).

We now evaluate the effect of the pulse on a non-leaked transmon (see Fig. 9.2(c),(d)). There should ideally be no effect, except for an acquired single-qubit phase which can easily be determined and corrected by either a real or virtual Z rotation. First, if the transmon is in $|0\rangle$ and there is some thermal population in the resonator, part of the state is supported on $|01\rangle$, which rotates into $|20\rangle$ in the same way as the opposite process by unitarity. Figure 9.2(c) shows that indeed the induced leakage is greater where $p^{|2\rangle}$ is lower in Fig. 9.2(a). However, due to the low $\bar{n} = 0.005$, the induced leakage is also overall low ($p^{|2\rangle} \approx 0.48\%$ in Fig. 9.2(c) at the operating point, which is comparable to state-of-the-art CZ leakage rates, see Section 9.3.2) and can be made even lower by engineering colder resonators. If the initial state is $|1\rangle\langle 1| \otimes \sigma_{\text{th}}$ there is little induced leakage ($p^{|2\rangle} \approx 0.02\%$ at the operating point and $p^{|2\rangle} \lesssim 0.04\%$ across the whole landscape) as the drive is off-resonant with transitions from this state. Second, the pulse might affect the coherence times of the transmon by driving transitions within or outside the computational subspace (and back), as the small but non-negligible transitory population in $|10\rangle$ in Fig. 9.2(b),(d) seems to suggest. However, we find that both the effective T_1^q and T_2^q are only marginally affected as a function of Ω (see Section 9.6.1). This is because stronger pulses cause a somewhat stronger disturbance to the qubit, but they are shorter so that in total the effect is small.

9.3. SURFACE CODE WITH LRUS

9.3.1. LAYOUT AND OPERATION SCHEDULING

We study the distance-3 rotated surface code (see Fig. 9.3(a)), nicknamed Surface-17, in the presence of leakage and LRUs. We follow the frequency and pipelined scheme in Ref. [58], in which the 9 data qubits are subdivided into 3 high- and 6 low-frequency ones. The 4 X and the 4 Z ancilla qubits have an intermediate frequency. We consider the flux-pulse implementation of the CZs [10, 11, 33–35] for tunable-frequency transmons, in which the transmon with the greater frequency is lowered towards the other one with a flux pulse. With this technique fluxed transmons are prone to leakage. This means that the high-frequency data qubits and all the ancilla qubits can leak. As shown in Section 8.4, leakage can last for many QEC cycles and be quite detrimental to the logical performance of the code. Here we address these issues with the res-LRU for high-frequency data qubits and with the π-LRU for ancilla qubits, as described below. If due to a different implementation of the CZs (or due to leakage mobility [49]; see Section 8.11.6) also the low-frequency data qubits can leak, one can apply the res-LRU to them as well but we do not explore this here.

The circuit executed for each QEC cycle is shown in Fig. 9.3(b). The X-type and Z-type parity-check units are implemented in an interleaved way, with the CZs for one unit being

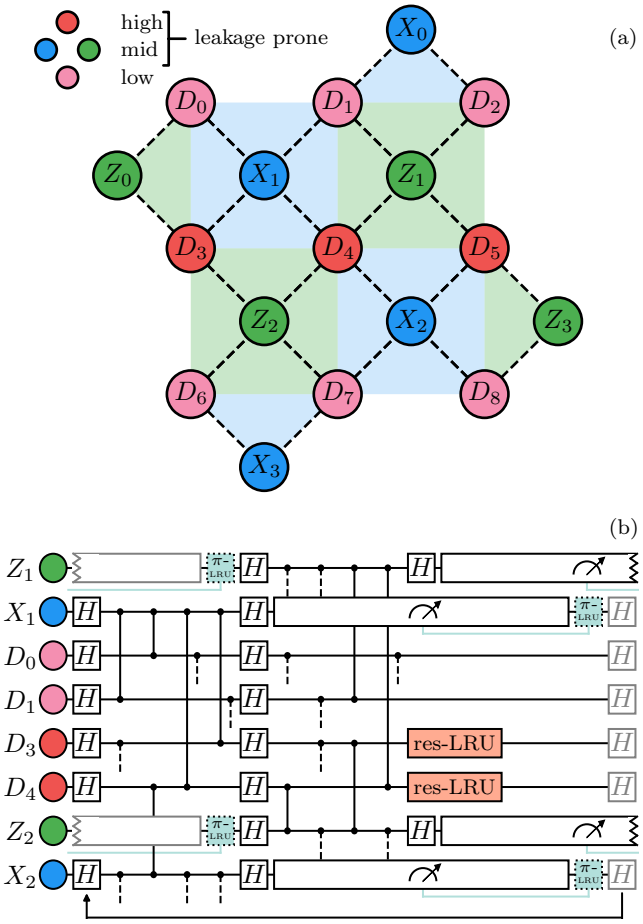

Figure 9.3: (a) Schematic overview of the Surface-17 layout [58]. Pink (resp. red) circles with D labels represent low- (high-) frequency data qubits, while blue (resp. green) circles with X (Z) labels represent ancilla qubits, which have an intermediate frequency. Ancilla qubits and high-frequency data qubits are prone to leakage during the CZ gates. (b) The quantum circuit for a single QEC cycle employed in simulation, for the unit-cell scheduling defined in Ref. [58], in which we insert the LRUs. The res-LRUs (orange) are applied unconditionally on the high-frequency data qubits after the CZs, while the π-LRUs (teal) are applied on the ancilla qubits depending on the measurement outcome. Gray elements correspond to operations belonging to the previous or the following QEC cycle. The duration of each operation is given in Section 9.7.1. The arrow at the bottom indicates the repetition of QEC cycles.

applied while the other ancilla-qubit type is measured. The duration of each operation is summarized in Section 9.7.1, with a total QEC-cycle duration of 800 ns. The data qubits are idling for a considerable amount of time, namely $T_{\text{slot}} = 440$ ns, while the ancilla qubits are measured. We choose this time slot as the ideal place to apply the res-LRUs, introduced in Section 9.2, to the high-frequency data qubits. Notice that the optimal pulse selected in Section 9.2.2, which was simulated for the target parameters of the high-frequency data qubits, takes about $t_{\text{p}} = 180$ ns and thus easily fits within this time slot (see Section 9.4 for a discussion about other parameter regimes).

For the ancilla qubits there is no available time slot to apply the res-LRU. A possibility would be to make the QEC-cycle time longer by inserting these LRUs when the measurement is completed. However, this approach would lower the logical error rate of the code by a non-negligible amount. On the other hand, ancilla qubits are measured and the (analog) measurement outcome contains information about leakage (see Section 8.11.1). We choose to use a different type of LRU altogether which uses this information. Specifically, we consider a $|1\rangle \leftrightarrow |2\rangle$ π pulse, conditioned on the measurement outcome reporting a $|2\rangle$. Below we discuss further details of the implementation of this π-LRU.

9.3.2. IMPLEMENTATION OF THE LRUS IN THE DENSITY-MATRIX SIMULATIONS

We use density-matrix simulations [51] using the open-source package *quantumsim* [59] to study Surface-17 with res-LRUs and π-LRUs. We include relaxation and dephasing (T_1 and T_2), as well as flux-dependent T_2 and leakage rate L_1 during the CZs, following the same error model as in Sections 8.10.1 and 8.10.2. L_1 is defined as the average leakage from the computational to the leakage subspace [70]. The state of the art is $L_1 \approx 0.1\%$ (see Sections 6.8 and 6.18), although the actual L_1 is expected to be higher when operating a multi-transmon processor [31, 71], thus here we consider up to $L_1 = 0.5\%$. We assume that single-qubit gates do not induce any leakage as their leakage rates are typically negligible compared to the CZs [6, 37, 38]. The noise parameters used are reported in Section 9.7.1. Furthermore, during a CZ with a leaked transmon, the non-leaked transmon acquires a phase called the leakage conditional phase (see Section 8.2). We select these phases uniformly at random (see Section 9.7.3) and, in contrast to Chapter 8, we then keep them fixed for every Surface-17 simulation in this chapter. This makes it easier to recognize trends as a function of the LRU parameters. In Section 9.7.3 we discuss the variability of the logical error rate depending on the leakage conditional phases. We do not consider further leakage from $|2\rangle$ to $|3\rangle$ in subsequent CZ gates (see Section 8.11.6) as we expect it to be negligible when LRUs make $|2\rangle$ short-lived.

RES-LRU FOR DATA QUBITS

In the simulations, leakage-prone transmons are modeled as 3-level systems and non-leakage-prone ones as 2-level systems, leading to an already computationally expensive size for the density matrix. As a consequence, we do not include the readout resonator explicitly in these simulations. The resonator is initially in the ground state and is returned to it at the end of the time slot, approximately. We can thus trace the resonator out and model the res-LRU on the transmon qubit as an incoherent $|2\rangle \mapsto |0\rangle$ relaxation (see Section 9.7.1 for details). Furthermore, in Section 9.2.2 we have observed that the

res-LRU can also cause a non-leaked transmon to partially leak, so we include that as an incoherent $|0\rangle \mapsto |2\rangle$ excitation.

Calling $p_i^{|j\rangle}, p_f^{|j\rangle}$ the populations before and after the res-LRU, we define the leakage-reduction rate $0 \leq R \leq 1$ as $R = 1 - p_f^{|2\rangle}$ conditioned on an initially fully leaked transmon, i.e. for $p_i^{|2\rangle} = 1$. Furthermore, we define the average res-LRU leakage rate L_1^{LRU} as the average of the induced leakage starting from either $|0\rangle$ or $|1\rangle$ (consistently with the definition for CZ [70]), with probability 1/2 each. Since almost all induced leakage comes from $|0\rangle$ (see Section 9.2.2), this means that $p_f^{|2\rangle} \approx 0$ for $p_i^{|1\rangle} = 1$ and that $p_f^{|2\rangle} \approx 2L_1^{\mathrm{LRU}}$ for $p_i^{|0\rangle} = 1$ (neglecting relaxation effects as the used $T_1 = 30\ \mu$s is relatively long). Combining these two definitions one gets the expression

$$p_f^{|2\rangle} \approx (1 - R)\, p_i^{|2\rangle} + 2L_1^{\mathrm{LRU}}\, p_i^{|0\rangle} \tag{9.16}$$

for an arbitrary incoming state. Notice that, given these definitions, Fig. 9.2(a),(c) respectively show a heatmap of $1 - R$ and $2L_1^{\mathrm{LRU}}$ for the considered transmon-resonator parameters. In particular, the operating point achieves $R \approx 99.5\%$ and $L_1^{\mathrm{LRU}} \approx 0.25\%$. The achieved leakage reduction can be compared with the one given purely by relaxation during T_{slot}, namely $R_{T_1} = 1 - e^{-T_{\mathrm{slot}}/(T_1/2)} = 2.9\%$, which shows that the LRU provides a much stronger additional seepage channel.

π-LRU FOR ANCILLA QUBITS

The dispersive readout of a transmon qubit is in general performed by sending a pulse to the readout resonator, integrating the reflected signal to obtain a point in the IQ plane and depleting the photons in the resonator (either passively by relaxation or actively with another pulse) [17–19]. The measured point is compared to one or more thresholds to declare the measurement outcome. These thresholds are determined as to optimally separate the distributions for the different outcomes, which have a Gaussian(-like) form. Here we assume that the distribution for $|2\rangle$ is sufficiently separated from $|0\rangle$ and $|1\rangle$ [17]. This is generally expected to be possible thanks to the different dispersive shift. Then one uses three thresholds in the IQ plane to distinguish between $|0\rangle$, $|1\rangle$ and $|2\rangle$ (or two if $|2\rangle$ is well-separated from e.g. $|0\rangle$). We also assume that an outcome can be declared during photon depletion, thus enabling real-time conditional feedback. This is challenging to perform in 200-300 ns in experiment due to the classical-postprocessing requirements, but it has been previously achieved [29, 57]. We can then apply the π-LRU right at the end of the depletion time. The $|1\rangle \leftrightarrow |2\rangle$ π pulse is expected to be implementable as a simple pulse in the same way and time as single-qubit gates (20 ns) and with comparable, coherence-limited fidelity.

If conditional feedback is not possible in the allotted time, one can either increase the QEC-cycle duration (at the cost of extra decoherence for all qubits, scaling as $1 - e^{-t_{\mathrm{extra}}/T_2}$ per qubit per QEC cycle) or postpone the conditional gate to the next QEC cycle. In the latter case, one source of error corresponds to the ancilla qubit already seeping before the application of the π-LRU, which then causes it to leak instead. The probability of this error is already low and is expected to become even lower with longer T_1s and lower-leakage CZs. The other errors are the Z rotations (depending on the leakage conditional phases) that the leaked ancilla qubit spreads for at least 1 extra QEC cycle, as well as the

fact that the parity-check stays disabled. We do not simulate these variants and we expect a relatively low logical-performance loss, corresponding to an average leakage lifetime of about 2 QEC cycles (see Figs. 9.4 and 9.9).

Readout-declaration errors are expected to affect the performance of the π-LRU. On one hand, an incorrect declaration of $|1\rangle$ as a $|2\rangle$ makes the π pulse induce leakage. On the other hand, declaring a $|2\rangle$ as a $|1\rangle$ would lead to leakage not being corrected and lasting for at least one extra QEC cycle. We define the readout matrix M with entries $M_{ij} =: p_M(i|j)$ being the probability that the actual state $|j\rangle$ resulting from the projective measurement is declared as an $|i\rangle$. In the simulations we use

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & p_M(1|1) & 1-p_M(1|1) \\ 0 & 1-p_M(2|2) & p_M(2|2) \end{pmatrix}. \tag{9.17}$$

In particular, this means that we do not consider declaration errors within the computational subspace. While that would change the value of the logical error rate since the error syndrome gets corrupted, it is not relevant for evaluating the performance of the π-LRU since a $|0\rangle$ mistaken for a $|1\rangle$ or vice-versa does not trigger the π-LRU anyway. Furthermore, we assume that a $|0\rangle$ cannot be mistaken as a $|2\rangle$ since their readout signals are often much more separated than the signals of $|1\rangle$ and $|2\rangle$. Note that if a $|0\rangle$ (rather than a $|1\rangle$, as we assume in this work) could be mistakenly declared as a $|2\rangle$, then a $|1\rangle \leftrightarrow |2\rangle$ π pulse does not induce leakage, so here we consider the worst-case scenario for the π-LRU.

9.3.3. AVERAGE LEAKAGE LIFETIME AND STEADY STATE

Once a qubit leaks, it tends to remain leaked for a significant amount of time, up to 10-15 QEC cycles on average (see Section 8.4). Starting from an initial state with no leakage, the probability that a qubit is in the leaked state tends towards a steady state within a few QEC cycles. It was shown in Section 8.11.8 that this evolution is well captured by a classical Markov process with leakage (resp. seepage) rate $\Gamma_{\mathscr{C}\to\mathscr{L}}$ ($\Gamma_{\mathscr{L}\to\mathscr{C}}$) per QEC cycle, where $\mathscr{C}$ (resp. $\mathscr{L}$) is the computational (leakage) subspace. Note that here $\mathscr{L}$ is 1-dimensional, corresponding to $|2\rangle$. In our error model, without accounting for LRUs, these rates are approximately given by

$$\Gamma_{\mathscr{C}\to\mathscr{L}} \approx N_{\text{flux}} L_1, \tag{9.18}$$

$$\Gamma_{\mathscr{L}\to\mathscr{C}} \approx N_{\text{flux}} L_2 + (1 - e^{-\frac{t_c}{T_{1/2}}}), \tag{9.19}$$

where N_{flux} is in how many CZ gates the transmon is fluxed during a QEC cycle, t_c is the duration of a QEC cycle and L_1 (resp. L_2) is the average leakage (seepage) probability of a CZ [70]. Thus the two native mechanisms that generate seepage are the CZs themselves and relaxation.

The major effect of a LRU is to effectively increase $\Gamma_{\mathscr{L}\to\mathscr{C}}$ in Eq. (9.19) by introducing an extra seepage mechanism. Hence we expect that $\Gamma^{\text{LRU}}_{\mathscr{L}\to\mathscr{C}} \sim \Gamma_{\mathscr{L}\to\mathscr{C}} + R$ for data qubits and $\Gamma^{\text{LRU}}_{\mathscr{L}\to\mathscr{C}} \sim \Gamma_{\mathscr{L}\to\mathscr{C}} + p_M(2|2)$ for ancilla qubits, preventing leakage from accumulating and lasting long for large R or $p_M(2|2)$.

The average leakage lifetime $l_{\text{avg}}^{\mathscr{L}}$ is the average duration of leakage and for a Markov process it is calculated as

$$l_{\text{avg}}^{\mathscr{L}} = \sum_{n=1}^{\infty} n\,\mathbb{P}(\text{stay in } \mathscr{L} \text{ for } n \text{ QEC cycles}) \tag{9.20}$$

$$= \sum_{n=1}^{\infty} n(1-\Gamma_{\mathscr{L}\to\mathscr{C}})^{n-1}\Gamma_{\mathscr{L}\to\mathscr{C}} = \frac{1}{\Gamma_{\mathscr{L}\to\mathscr{C}}}, \tag{9.21}$$

thus assuming that the qubit starts in $\mathscr{L}$. The evolution of the leakage probability $\bar{p}^{\mathscr{L}}(n)$, averaged over surface-code runs, as a function of the QEC-cycle number n is well-approximated by (see Section 8.11.8)

$$\bar{p}^{\mathscr{L}}(n) = \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}}{\Gamma_{\mathscr{C}\to\mathscr{L}}+\Gamma_{\mathscr{L}\to\mathscr{C}}}(1-e^{-(\Gamma_{\mathscr{C}\to\mathscr{L}}+\Gamma_{\mathscr{L}\to\mathscr{C}})n}). \tag{9.22}$$

The steady state is the long-time limit and is given by

$$\bar{p}_{\text{ss}}^{\mathscr{L}} = \lim_{n\to\infty}\bar{p}^{\mathscr{L}}(n) = \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}}{\Gamma_{\mathscr{C}\to\mathscr{L}}+\Gamma_{\mathscr{L}\to\mathscr{C}}}. \tag{9.23}$$

For ancilla qubits $\bar{p}^{\mathscr{L}}(n)$ can be computed directly from the "true" measurement outcomes (i.e. without declaration errors on top). For data qubits it can be computed from the density matrix. Specifically, for data qubits we evaluate $\bar{p}^{\mathscr{L}}(n)$ right after the CZs.

Figure 9.4 shows $l_{\text{avg}}^{\mathscr{L}}$ and $\bar{p}_{\text{ss}}^{\mathscr{L}}$ extracted from the Surface-17 simulations by fitting $\bar{p}^{\mathscr{L}}(n)$ to Eq. (9.22) for each qubit. We can indeed observe that these quantities drop substantially for both data and ancilla qubits. The decays follow an inverse proportionality as e.g. for data qubits

$$l_{\text{avg}}^{\mathscr{L}} = \frac{1}{\Gamma_{\mathscr{L}\to\mathscr{C}}^{\text{LRU}}} \sim \frac{1}{\Gamma_{\mathscr{L}\to\mathscr{C}}+R} \sim \frac{1}{R} \tag{9.24}$$

$$\bar{p}_{\text{ss}}^{\mathscr{L}} = \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}}{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}+\Gamma_{\mathscr{L}\to\mathscr{C}}^{\text{LRU}}} \sim \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}}{\Gamma_{\mathscr{L}\to\mathscr{C}}^{\text{LRU}}} \sim \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}}{R} \tag{9.25}$$

for sufficiently large R and small $\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}$. For ancilla qubits we expect, similarly, a $1/p_M(2|2)$ dependence. The lifetime drops from values $\gtrsim 10$ to ≈ 1, which is the minimum value it can achieve (some points drop below 1 within error bars as it is difficult for the fit to estimate such a short lifetime). As of course the LRUs do not prevent leakage from occurring during the CZs in the first place, one cannot expect the steady state to reach 0 even for a perfect LRU ($R = 1$), but rather $\bar{p}_{\text{ss}}^{\mathscr{L}} \sim \Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}} \approx N_{\text{flux}}L_1$ ($+L_1^{\text{LRU}}$ if the LRU can mistakenly induce leakage). Figures 9.4(b),(d) show that this is indeed the case.

Figure 9.4 also demonstrates that both $l_{\text{avg}}^{\mathscr{L}}$ and $\bar{p}_{\text{ss}}^{\mathscr{L}}$ get close to their minimum values already for R, $p_M(2|2) \gtrsim 80\%$. This suggests that res-LRU and π-LRU may not necessarily need to be perfect to provide a good logical performance in Surface-17. This means that one could use e.g. a weaker pulse to implement the res-LRU or that the readout of $|2\rangle$ may not need to be particularly optimized in practice.

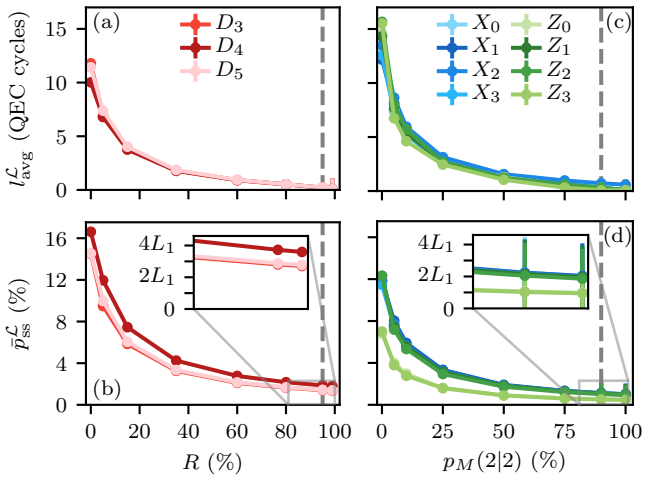

Figure 9.4: Average leakage lifetime $l^{\mathscr{L}}_{\mathrm{avg}}$ [(a),(c)] and leakage steady state $\bar{p}^{\mathscr{L}}_{\mathrm{ss}}$ [(b),(d)] as a function of the leakage-reduction rate R for data qubits [(a),(b)] and as a function of the readout probability $p_M(2|2)$ for ancilla qubits [(c),(d)]. Here we fix the CZ leakage rate to $L_1 = 0.5\%$. The insets in (b),(d) show that $\bar{p}^{\mathscr{L}}_{\mathrm{ss}}$ tends to $\approx N_{\mathrm{flux}} L_1$ ($N_{\mathrm{flux}} = 4$ for D_4, 3 for D_3, D_5, 1 for Z_0, Z_3 and 2 for the remaining ancilla qubits). The vertical dashed lines correspond to the values used in Section 9.3.4. These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are mostly smaller than the symbol size.

9.3.4. LOGICAL PERFORMANCE

In the simulations the logical qubit is initialized in $|0\rangle_{\mathrm{L}}$ and the logical fidelity $\mathscr{F}_{\mathrm{L}}(n)$ is computed at the end of each QEC cycle as the probability that the decoder correctly determines whether a logical error has occurred or not. We do not perform a similar analysis with initial state $|+\rangle_{\mathrm{L}}$ or other states as the density-matrix simulations are computationally expensive and we expect a similar performance. The logical error rate ε_{L} per QEC cycle can be extracted by fitting $\mathscr{F}_{\mathrm{L}}(n) = [1 + (1 - 2\varepsilon_{\mathrm{L}})^{n-n_0}]/2$, where n_0 is a fitting parameter (usually close to 0) [51]. We evaluate ε_{L} for the upper bound decoder (UB) which uses the complete density-matrix information to infer a logical error, and for the minimum-weight perfect-matching decoder (MWPM). Detailed information about these decoders can be found in [51, 72] and an overview is given in Section 9.7.1.

By mapping a leaked qubit back to the computational subspace, a LRU does not fully remove a leakage error but can at most convert it into a regular (Pauli) error. Hence, it is not to be expected that ε_{L} in the presence of leakage can be restored to the value at $L_1 = 0$. We consider realistic parameters for the LRUs. Specifically, we use $R = 95\%$, $L_1^{\mathrm{LRU}} = 0.25\%$, $p_M(2|2) = 90\%$ and $p_M(1|1) = 99.5\%$. We have shown in Section 9.2.2 that the first two parameters can be attained with realistic parameters for the transmon-readout system, while the last two are close to be achievable in experiment [15, 54]. In particular, while the operating point has $R = 99.5\%$, we conservatively choose $R = 95\%$ here. Notice that $p_M(1|1) = 99.5\%$ is quite high. We argue that the state of the art can be squeezed as the threshold to distinguish between $|1\rangle$ and $|2\rangle$ in the IQ plane could be

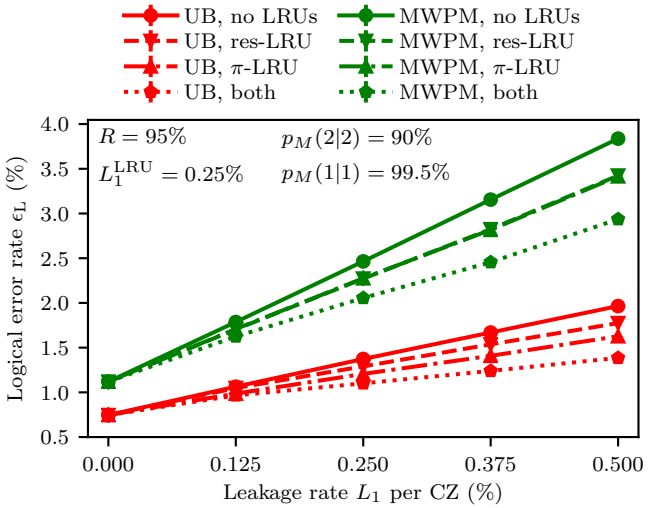

Figure 9.5: Logical error rate ε_L per QEC cycle for the upper bound (UB, red) and minimum-weight perfect-matching (MWPM, green) decoders versus the CZ leakage rate L_1, in the cases with: no LRUs, only res-LRU, only π-LRU and both LRUs (the point without leakage at $L_1 = 0$ is always without LRUs as well). These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are smaller than the symbol size.

moved towards $|2\rangle$, rather than placing it in the middle as is common practice. In this way one would slightly reduce $p_M(2|2)$ in favor of $p_M(1|1)$ if $p_M(1|1)$ is not high enough. A broader study of the logical performance as a function of the LRU parameters can be found in Section 9.7.2.

Figure 9.5 shows the reduction in ε_L as a function of the CZ leakage rate L_1 when LRUs with the given parameters are employed. Using only the res-LRU or the π-LRU lowers $\varepsilon_L^{\mathrm{MWPM}}$ by basically the same amount, while $\varepsilon_L^{\mathrm{UB}}$ is lower for the π-LRU than for the res-LRU. We attribute this to the fact that UB directly uses the information in the density matrix, while MWPM relies on the measured syndrome, thus being more susceptible to ancilla-qubit leakage. When both LRUs are used, we see that ε_L is reduced by an amount which is close to the sum of the reductions when only one kind of LRU is used. As expected, ε_L is not restored to the value at $L_1 = 0$, but the reduction is overall significant and can reach up to 30% for both MWPM and UB compared to the case without LRUs.

9.4. DISCUSSION

In this chapter we have introduced a leakage-reduction scheme using res-LRUs and π-LRUs which does not require any additional hardware or a longer QEC cycle. Furthermore, while the scheme in Ref. [49] is applicable only to ancilla qubits, our combination of res-LRU for data qubits and π-LRU for ancilla qubits enables to significantly reduce leakage in the whole transmon processor. We have shown with detailed simulations using realistic parameters that the reset scheme in [54–56] can be adapted to be a LRU without signifi-

cantly affecting the states in the computational subspace, allowing to unconditionally apply the res-LRU in the surface code. The use of the res-LRU for data qubits, as well as the use of the π-LRU for ancilla qubits, leads to a substantial reduction of the average leakage lifetime and leakage steady state, preventing leakage from lasting more than ≈ 1 QEC cycles on average, even when the LRUs are imperfect and can introduce leakage themselves. Using full density-matrix simulations of Surface-17 we have demonstrated that this leads to a significant reduction of the logical error rate for both the UB and MWPM decoders.

Regarding the practical implementation of the res-LRU, the required drive amplitude is relatively strong, similarly to the one used in the experiments in [54–56]. It is thus important that the microwave crosstalk is minimized by careful engineering of the drive lines. Furthermore, in a multi-transmon processor it is relevant that the drive frequency does not accidentally match any two-qubit or neighboring single-qubit transitions. E.g., in the original scheme in Ref. [58] that we followed, the target frequencies are 6.7, 6.0 and 4.9 GHz for high-, mid- and low-frequency qubits, respectively, and 7.8 GHz for the readout resonator [28]. In particular, the mid-frequency qubits (the ancilla qubits) are parked around 5.4-5.5 GHz during measurement, with their $|1\rangle \leftrightarrow |2\rangle$ transition around 5.1-5.2 GHz. This is close to the optimal drive frequency found in Section 9.2.2 (≈ 5.25 GHz), which can lead to an indirect ancilla-qubit drive mediated by the bus resonator, albeit weaker. The difficulty of precise frequency targeting in fabrication can further lead to undesired frequency collisions. These issues can be alleviated by choosing slightly different transmon/resonator frequencies and anharmonicities to make the drive more off-resonant with that transition (combined with better frequency targeting [73]), or they can be mitigated altogether by using tunable couplers [2, 12, 15]. The res-LRU is compatible with tunable-coupler schemes and their possibly different operation scheduling than in Ref. [58], as well as potentially applicable to superconducting qubits which use a resonator for dispersive readout other than the transmon. Tunable couplers would also be advantageous to fully protect the res-LRU performance from residual ZZ crosstalk, even though we find that a cumulative ZZ interaction up to ~ 2 MHz can be tolerated with fixed couplers (see Section 9.6.3). Beside this, if the low-frequency data qubits can leak depending on the implementation of the CZ, the res-LRU can be applied to them in the same time slot as the high-frequency ones. If the thermal population in the readout resonator is relatively high in a given experiment, the effect of a correspondingly high L_1^{LRU} can potentially be mitigated by applying res-LRU conditionally on the detection of leakage by a set of hidden Markov models (see Section 8.5).

Regarding the viability of inserting the res-LRU in the surface-code time scheduling, the necessary condition is that $t_{\mathrm{p}} \leq T_{\mathrm{slot}}$. We can express T_{slot} as $T_{\mathrm{slot}} = t_{\mathrm{m}} - 4t_{\mathrm{CZ}}$, where t_{m} is the measurement time for the ancilla qubits. Slower CZs might make T_{slot} too short, although CZs even faster than 40 ns (as assumed here) have been realized in 15 ns [13]. The measurement time can be further broken down into readout-pulse time and photon-depletion time, $t_{\mathrm{m}} = t_{\mathrm{read}} + t_{\mathrm{depl}}$. Both of these would be reduced by a larger κ, however, assuming that the κ's of ancilla- and data-qubit resonators are comparable, t_{p} would be reduced as well. Even if we keep t_{p} and t_{CZ} fixed to the values in this chapter, we get $t_{\mathrm{m}} \geq 340$ ns, which is significantly lower than $t_{\mathrm{m}} = 580$ ns as considered here. A desirable, additional condition to the necessary one is that $T_{\mathrm{slot}} - t_{\mathrm{p}} \geq 4/\kappa$, i.e. that there is enough leftover time in T_{slot} to allow for the data-qubit resonator to return the thermal

state, where we estimate that 4 decay constants would suffice (together with the fact that the resonator was already relaxing during t_p). Assuming similar depletion time for data- and ancilla-qubit resonators, this roughly means that the res-LRU is easily applicable if t_p is smaller or similar to t_{read}. Note that in this chapter we have $T_{slot} - t_p \sim 16/\kappa$ and $t_p < t_{read}$. If the additional condition above is not satisfied, one could demand that at least the resonator has returned to the thermal state before the res-LRU in the following QEC cycle, i.e. $T_{slot} - t_p + 8t_{CZ} + 2t_H \geq 4/\kappa$. In this case the disadvantage would be that the presence of a fraction of a photon in the resonator would cause additional data-qubit dephasing especially during the first few CZs. As the extra photon is present only when the qubit was previously leaked, we expect this disadvantage to be small as long as the overall leakage rate is small. If even the relaxed additional condition is violated, on top of the additional dephasing the resonator would also heat up, effectively leading to a higher L_1^{LRU} in the QEC cycle(s) following the one in which the qubit leaked. As also this effect scales with L_1, we expect that it would not be an issue as long as κ is not very low (allowing for at most 1 extra QEC cycle to thermalize we get $\kappa/2\pi \geq 1$ MHz). Otherwise, leakage would not really be removed from the system but would be largely moved back and forth from the transmon to the resonator.

The demonstrated reduction in the average leakage lifetime and in the logical error rate is expected to lead to a higher noise threshold for the surface code in the presence of leakage, compared to the case without LRUs. Furthermore, for error rates below threshold (both regular and leakage) we believe that the logical error rate would be exponentially suppressed with increasing code distance when employing LRUs. Without LRUs this might hold only when the code distance is sufficiently larger than the average leakage lifetime ($d \gg l_{avg}^{\mathscr{L}}$). For smaller distances the relatively long correlated error chains induced by leakage might lead to a sub-exponential scaling. To study the noise threshold and sub-threshold behavior it is necessary to implement simulations of large code sizes which use a simplified error model, such as a stochastic error model for leakage and Pauli errors [24, 40, 43]. We expect that the demonstrated MWPM logical error rate can be further lowered by the use of decoders [24, 40, 43, 60–62] that use information about leakage extracted directly or indirectly (e.g. with hidden Markov models; see Section 8.5) from the measurement outcomes.

The data underlying this chapter, as well as the code to analyze it, are available at https://doi.org/10.4121/c.5320331. The code used to generate the data is available upon request to the corresponding author.

9.5. APPROXIMATE TRANSMON-RESONATOR HAMILTONIAN

9.5.1. SCHRIEFFER-WOLFF TRANSFORMATION

In this section we explain the concept of the Schrieffer-Wolff transformation (SWT) [63–65] and derive the equations that we use in the following sections.

Consider a Hamiltonian

$$H = H_0 + \epsilon V \tag{9.26}$$

expressed in a certain basis $\{|\psi_n\rangle\}$, where H_0 is block diagonal with respect to this basis and the perturbation V can be taken as block off-diagonal without loss of generality (block-

diagonal terms can be included in the definition of H_0). Furthermore, we assume $||V|| = \mathcal{O}(1)$ and $\epsilon \ll \Delta_{ij}$, where we set Δ_{ij} as the minimum energy separation between blocks i and j.

The SWT corresponds to finding an anti-hermitian matrix S such that

$$H' := e^S H e^{-S} \tag{9.27}$$

is block diagonal. In other words, calling $\{|\bar{\psi}_n\rangle\}$ the basis of eigenstates of H, $e^S = \sum_n |\psi_n\rangle\langle\bar{\psi}_n|$. The matrix S can be expanded in a series

$$S = \sum_{k=1}^{\infty} \epsilon^k S_k \tag{9.28}$$

where each S_k is block off-diagonal. If $\epsilon \ll \Delta_{ij}$ one can expect the first order (S_1) to provide a good approximation, otherwise one needs to consider higher orders depending on ϵ (although the series does not always converge for extensive systems [64]). Using the Baker-Campbell-Hausdorff formula one gets

$$H' = e^S H e^{-S} = \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{[S,[S,\dots[S}_{k \text{ times}},H]\dots]]. \tag{9.29}$$

The procedure for the SWT is to group terms of the same order in ϵ in this formula and set the block off-diagonal part of H' to 0, thus getting equations for $\{S_k\}$, in the usual case with *two* blocks [64]. One uses the relationships

$$[\text{diagonal}, \text{diagonal}] = \text{diagonal}, \tag{9.30}$$
$$[\text{diagonal}, \text{off-diagonal}] = \text{off-diagonal}, \tag{9.31}$$
$$[\text{off-diagonal}, \text{off-diagonal}] = \text{diagonal}. \tag{9.32}$$

However, the last line only holds for the case with two blocks. In the following we consider the generalization of the SWT to the case with an arbitrary number of blocks [65]. We use the notation O_{D} and O_{OD} for the block diagonal and off-diagonal parts of an operator $O = O_{\text{D}} + O_{\text{OD}}$, respectively.

Here we expand H and S up to $k = 3$ in Eq. (9.29), assuming that the 4th-order block

off-diagonal term is negligible. We get the following pieces:

$$\text{0th order}: \quad H_0 \tag{9.33}$$

$$\text{1st order}: \quad V + [S_1, H_0] \tag{9.34}$$

$$\text{2nd order}: \quad [S_1, V] + \frac{1}{2}[S_1, [S_1, H_0]] + [S_2, H_0] \tag{9.35}$$

$$\begin{aligned}\text{3rd order}: \quad & [S_2, V] + \frac{1}{2}\Big([S_2, [S_1, H_0]] + [S_1, [S_1, V]] + [S_1, [S_2, H_0]]\Big) \\ & + \frac{1}{6}[S_1, [S_1, [S_1, H_0]]] + [S_3, H_0]\end{aligned} \tag{9.36}$$

$$\begin{aligned}\text{4th order}: \quad & [S_3, V] + \frac{1}{2}\Big([S_1, [S_3, H_0]] + [S_2, [S_2, H_0]] + [S_3, [S_1, H_0]] + [S_1, [S_2, V]] \\ & \qquad + [S_2, [S_1, V]]\Big) \\ & + \frac{1}{6}\Big([S_1, [S_1, [S_1, V]]] + [S_2, [S_1, [S_1, H_0]]] + [S_1, [S_2, [S_1, H_0]]] \\ & \quad + [S_1, [S_1, [S_2, H_0]]]\Big) \\ & + \frac{1}{24}[S_1, [S_1, [S_1, [S_1, H_0]]]]\,.\end{aligned} \tag{9.37}$$

Setting the block off-diagonal parts at 1st, 2nd and 3rd order to 0 we get

$$[H_0, S_1] = V \tag{9.38}$$

$$[H_0, S_2] = \frac{1}{2}[S_1, V]_{\mathrm{OD}} \tag{9.39}$$

$$[H_0, S_3] = \frac{1}{2}[S_2, V]_{\mathrm{OD}} + \frac{1}{3}[S_1, [S_1, V]_{\mathrm{D}}]_{\mathrm{OD}} + \frac{1}{12}\big[S_1, [S_1, V]_{\mathrm{OD}}\big]_{\mathrm{OD}}\,, \tag{9.40}$$

where we have used the first equation to simplify the following ones. These equations can be solved iteratively for S_k (given knowledge of the eigenstates of H_0). The Hamiltonian H' is then block diagonal up to 4th order and is explicitly given by

$$\begin{aligned}H' = H_0 & + \frac{\epsilon^2}{2}[S_1, V]_{\mathrm{D}} \\ & + \epsilon^3\Big(\frac{1}{2}[S_2, V]_{\mathrm{D}} + \frac{1}{12}\big[S_1, [S_1, V]_{\mathrm{OD}}\big]_{\mathrm{D}}\Big) \\ & + \epsilon^4\Big(\frac{1}{2}[S_3, V]_{\mathrm{D}} - \frac{1}{24}\big[S_1, [S_1, [S_1, V]_{\mathrm{D}}]_{\mathrm{OD}}\big]_{\mathrm{D}} \\ & \qquad - \frac{1}{6}\big[S_2, [S_1, V]_{\mathrm{OD}}\big]_{\mathrm{D}} + \frac{1}{12}\big[S_1, [S_2, V]_{\mathrm{OD}}\big]_{\mathrm{D}}\Big).\end{aligned} \tag{9.41}$$

This expression has been simplified using Eqs. (9.38) to (9.40), together with the fact that e.g. $[S_k, [\ldots, \ldots]_{\mathrm{D}}]_{\mathrm{D}} = 0$ since S_k is block off-diagonal.

9.5.2. SWT OF THE CAPACITIVE COUPLING

We consider the Hamiltonian $H = H_0 + H_c + H_d$ of a driven transmon capacitively coupled to a resonator, as given in Eqs. (9.1) to (9.4).

The SWT of H_c up to 1st order in the perturbation parameter $\epsilon = g/\Delta$, where $\Delta = \omega_q - \omega_r$, is implemented using the matrix [66]

$$S_1 = g \sum_{m=1}^{\infty} \frac{\sqrt{m}}{\Delta + \alpha(m-1)} \Big(a\,|m\rangle\langle m-1| - \text{h.c.} \Big), \tag{9.42}$$

where $\{|m\rangle\}$ are transmon states and where we have absorbed ϵ in the definition of S_1. The Hamiltonian in the unitarily transformed frame as defined in Section 9.2.1 is then given by

$$H^D \approx e^{S_1} H e^{-S_1} = e^{S_1}(H_0 + H_c)e^{-S_1} + e^{S_1} H_d\, e^{-S_1} \tag{9.43}$$

with

$$\begin{aligned} e^{S_1}(H_0 + H_c)e^{-S_1} &= H_0 + \frac{1}{2}[S_1, H_c] && (9.44)\\ &\approx \delta^r a^\dagger a + \sum_{m=1}^{\infty} \Big(m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}} \Big) |m\rangle\langle m| \\ &\quad - a^\dagger a \sum_{m=0}^{\infty} \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| && (9.45)\\ &:= H_0^D && (9.46) \end{aligned}$$

where we define $\Delta_m = \Delta + \alpha m = \Delta - |\alpha| m$ as $\alpha < 0$ for transmons. The second term above contains a Stark shift of the transmon frequency and the last term is the state-dependent dispersive shift. The approximation in Eq. (9.45) is due to the fact that we have ignored a double-excitation exchange term coming from $[S_1, H_c]$, since it is proportional to $g\alpha/(\Delta_m \Delta_{m-1})$. This is negligible for low anharmonicity and, secondly, for $\omega_r > \omega_q$ as then $\Delta < 0$ and $|\Delta_m|$ increases with m. If instead $\omega_r < \omega_q$, $\Delta > 0$ and $|\Delta_m|$ decreases with m, so even if the approximation is good for the two lowest levels, there can be some higher level which does not sit well within the dispersive regime. However, in this work we consider a system with $\omega_r > \omega_q$, hence we do not need to take this into account.

The drive Hamiltonian in the unitarily transformed frame takes the form

$$e^{S_1} H_d\, e^{-S_1} = H_{d1}^D + H_{d2}^D, \tag{9.47}$$

where

$$\begin{aligned} H_{d1}^D &:= \frac{\Omega e^{i\phi}}{2} b + \text{h.c.} && (9.48)\\ H_{d2}^D &:= \frac{\Omega e^{i\phi}}{2} \Big(a \sum_{m=0}^{\infty} \frac{g\Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| + a^\dagger \sum_{m=0}^{\infty} \frac{g\alpha\sqrt{m+1}\sqrt{m+2}}{\Delta_m \Delta_{m+1}} |m\rangle\langle m+2| \Big) + \text{h.c.} && (9.49) \end{aligned}$$

The last term contains a 1st-order approximation in g/Δ of the $|20\rangle \leftrightarrow |01\rangle$ effective coupling $\tilde{g}$, which is linear in Ω. However, the "pure" drive term H_{d1}^D can be quite strong, so we need to evaluate how it affects $\tilde{g}$ and the rest of the Hamiltonian.

9.5.3. SWT OF THE PURE DRIVE HAMILTONIAN

Summarizing, in the unitarily transformed frame the original Hamiltonian H takes (approximately) the form

$$H^D \approx H_0^D + H_{d1}^D + H_{d2}^D, \tag{9.50}$$

where H_0^D is given in Eq. (9.45) and H_{d1}^D, H_{d2}^D are given in Eqs. (9.48) and (9.49), respectively.

We now want to find an additional SWT transformation $S' = S_1' + S_2' + S_3'$, with H_{d1}^D taking the role of V in Section 9.5.1, defining a "double-dressed" Hamiltonian

$$H^{DD} := e^{S'} H^D e^{-S'} \tag{9.51}$$

$$= \underbrace{e^{S'}(H_0^D + H_{d1}^D)e^{-S'}}_{=:H_0^{DD}} + \underbrace{e^{S'} H_{d2}^D e^{-S'}}_{=:H_d^{DD}} \tag{9.52}$$

such that H_0^{DD} is fully diagonal up to 3rd order in the perturbation parameter $\epsilon = \Omega/\delta^q$. Then H_d^{DD} gives the couplings within the manifold of interest ($|20\rangle, |01\rangle$) and outside of it. We absorb ϵ^k in the definition of S_k' so it does not explicitly appear below.

Following Section 9.5.1, to find S_1' we need to solve Eq. (9.38), i.e.

$$[H_0^D, S_1'] = H_{d1}^D \tag{9.53}$$

in this specific case. Bracketing it with the eigenstates $\{|ml\rangle\}$ of H_0^D, with the notation $|\text{transmon}, \text{resonator}\rangle$, we get the matrix elements of S_1' as

$$\langle ml|S_1'|nk\rangle = \frac{\langle ml|H_{d1}^D|nk\rangle}{E_{ml}^D - E_{nk}^D}, \tag{9.54}$$

where $\{E_{ml}^D\}$ are the eigenenergies of H_0^D, which can be easily inferred from Eq. (9.45). We neglect the dispersive shift since it is proportional to α/Δ. Then

$$\langle ml|S_1'|nk\rangle = \frac{\Omega}{2}\left(-\frac{\sqrt{m+1}\delta_{m,n-1}\delta_{l,k}}{\delta^q + \alpha m + \frac{g^2\Delta_{-1}}{\Delta_{m-1}\Delta_m}} e^{i\phi} + \frac{\sqrt{m}\delta_{m,n+1}\delta_{l,k}}{\delta^q + \alpha(m-1) + \frac{g^2\Delta_{-1}}{\Delta_{m-2}\Delta_{m-1}}} e^{-i\phi}\right), \tag{9.55}$$

where $\delta_{i,j}$ is the Kronecker delta. From this equation one can infer that

$$S_1' = -\frac{\Omega}{2} e^{i\phi} \sum_{m=0}^{\infty} \frac{\sqrt{m+1}}{\delta_m^q} |m\rangle\langle m+1| - \text{h.c.}, \tag{9.56}$$

where we have defined $\delta_m^q = \delta^q + \alpha m + \frac{g^2\Delta_{-1}}{\Delta_{m-1}\Delta_m}$.

Having derived S_1', we can compute S_2' from Eq. (9.39), i.e.

$$[H_0^D, S_2'] = \frac{1}{2}[S_1', H_{d1}^D]_{\text{OD}} \tag{9.57}$$

with

$$[S_1', H_{d1}^D] = -\frac{\Omega^2}{2}\sum_{m=0}^{\infty}\frac{\tilde{\delta}_m^q}{\delta_m^q\delta_{m-1}^q}|m\rangle\langle m| - \frac{\Omega^2}{4}\sum_{m=0}^{\infty}\sqrt{m+1}\sqrt{m+2}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right)(e^{2i\phi}|m\rangle\langle m+2| + \text{h.c.}), \tag{9.58}$$

where $\tilde{\delta}_m^q = \delta^q - \alpha + \frac{g^2\Delta_{-1}\Delta_{3m}}{\Delta_m\Delta_{m-1}\Delta_{m-2}}$. Clearly the first term is the diagonal part while the second term is the off-diagonal one. With a similar procedure as the one used for S_1', it follows that

$$S_2' = \frac{\Omega^2}{8}e^{2i\phi}\sum_{m=0}^{\infty}\frac{\sqrt{m+1}\sqrt{m+2}}{\delta_m^q+\delta_{m+1}^q}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right)|m\rangle\langle m+2| - \text{h.c.} \tag{9.59}$$

We can then compute S_3' from Eq. (9.40), i.e.

$$[H_0^D, S_3'] = \frac{1}{2}[S_2', H_{d1}^D]_{\text{OD}} + \frac{1}{3}[S_1', [S_1', H_{d1}^D]_{\text{D}}]_{\text{OD}} + \frac{1}{12}[S_1', [S_1', H_{d1}^D]_{\text{OD}}]_{\text{OD}}. \tag{9.60}$$

The result is

$$\begin{aligned} S_3' = \Omega^3 e^{i\phi}\sum_{m=0}^{\infty}|m\rangle\langle m+1|\Bigg(&\frac{1}{12}\frac{\sqrt{m+1}}{(\delta_m^q)^3}\left(\frac{\tilde{\delta}_{m+1}^q}{\delta_{m+1}^q}-\frac{\tilde{\delta}_m^q}{\delta_{m-1}^q}\right) \\ &+\frac{1}{96\delta_m^q}\Bigg((m+2)\sqrt{m+1}\frac{\delta_m^q+4\delta_{m+1}^q}{\delta_{m+1}^q(\delta_m^q+\delta_{m+1}^q)}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right) \\ &-\sqrt{m+1}m\frac{4\delta_{m-1}^q+\delta_m^q}{\delta_{m-1}^q(\delta_{m-1}^q+\delta_m^q)}\left(\frac{1}{\delta_{m-1}^q}-\frac{1}{\delta_m^q}\right)\Bigg)\Bigg) - \text{h.c.} \\ +\frac{\Omega^3}{96}e^{3i\phi}\sum_{m=0}^{\infty}&|m\rangle\langle m+3|\frac{\sqrt{m+1}\sqrt{m+2}\sqrt{m+3}}{\delta_m^q+\delta_{m+1}^q+\delta_{m+2}^q} \\ &\left(\frac{3\delta_{m+2}^q-\delta_{m+1}^q-\delta_m^q}{\delta_{m+2}^q(\delta_m^q+\delta_{m+1}^q)}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right)-\frac{3\delta_m^q-\delta_{m+1}^q-\delta_{m+2}^q}{\delta_m^q(\delta_{m+1}^q+\delta_{m+2}^q)}\left(\frac{1}{\delta_{m+1}^q}-\frac{1}{\delta_{m+2}^q}\right)\right) - \text{h.c.} \end{aligned} \tag{9.61}$$

We can eventually use Eqs. (9.56), (9.59) and (9.61) together with Eq. (9.41) to ob-

tain H_0^{DD} (defined in Eq. (9.52)):

$$
\begin{aligned}
H_0^{DD} = \delta^r a^\dagger a + \sum_{m=0}^{\infty} |m\rangle\langle m| \Bigg(& m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}} \\
& - \frac{\Omega^2 \tilde{\delta}_m^q}{4\delta_m^q \delta_{m-1}^q} - \frac{\Omega^4}{32}\Bigg(\frac{m+1}{(\delta_m^q)^3}\Big(\frac{\tilde{\delta}_{m+1}^q}{\delta_{m+1}^q} - \frac{\tilde{\delta}_m^q}{\delta_{m-1}^q}\Big) - \frac{m}{(\delta_{m-1}^q)^3}\Big(\frac{\tilde{\delta}_m^q}{\delta_m^q} - \frac{\tilde{\delta}_{m-1}^q}{\delta_{m-2}^q}\Big)\Bigg) \\
& - \frac{\Omega^4}{192}\Bigg(\frac{1}{\delta_m^q}\Bigg((m+2)(m+1)\frac{\delta_m^q + 5\delta_{m+1}^q}{\delta_{m+1}^q(\delta_m^q + \delta_{m+1}^q)}\Big(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q}\Big) \\
& \qquad - (m+1)m\frac{5\delta_{m-1}^q + \delta_m^q}{\delta_{m-1}^q(\delta_{m-1}^q + \delta_m^q)}\Big(\frac{1}{\delta_{m-1}^q} - \frac{1}{\delta_m^q}\Big)\Bigg) \\
& \qquad - \frac{1}{\delta_{m-1}^q}\Bigg((m+1)m\frac{\delta_{m-1}^q + 5\delta_m^q}{\delta_m^q(\delta_{m-1}^q + \delta_m^q)}\Big(\frac{1}{\delta_{m-1}^q} - \frac{1}{\delta_m^q}\Big) \\
& \qquad\qquad - m(m-1)\frac{5\delta_{m-2}^q + \delta_{m-1}^q}{\delta_{m-2}^q(\delta_{m-2}^q + \delta_{m-1}^q)}\Big(\frac{1}{\delta_{m-2}^q} - \frac{1}{\delta_{m-1}^q}\Big)\Bigg)\Bigg) \\
& + \frac{\Omega^4}{96}\Bigg(\frac{(m+2)(m+1)}{\delta_m^q + \delta_{m+1}^q}\Big(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q}\Big)^2 - \frac{m(m-1)}{\delta_{m-2}^q + \delta_{m-1}^q}\Big(\frac{1}{\delta_{m-2}^q} - \frac{1}{\delta_{m-1}^q}\Big)^2\Bigg)\Bigg) \\
& - a^\dagger a \sum_m \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| .
\end{aligned}
\tag{9.62}
$$

We note that this expression implicitly contains all cross terms between the perturbative parameters g/Δ and Ω/δ^q up to the chosen orders. The approximate coupling Hamiltonian H_d^{DD} (defined in Eq. (9.52)) up to 2nd order in Ω/δ^q is instead given by

$$
\begin{aligned}
H_d^{DD} &= H_{d2}^D + [S_1', H_{d2}^D] + [S_2', H_{d2}^D] + \frac{1}{2}[S_1', [S_1', H_{d2}^D]] && (9.63)\\
&=: H_{\text{eff.coupl.}}^{DD} + H_{\text{resid.}}^{DD}, && (9.64)
\end{aligned}
$$

where

$$
\begin{aligned}
H_{\text{eff.coupl.}}^{DD} = e^{i\phi} a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+2| \Bigg(& \tilde{g}_m\Bigg(1 - \frac{\Omega^2}{8}\Bigg(\frac{m+3}{(\delta_{m+2}^q)^2} + \frac{m+2}{(\delta_{m+1}^q)^2} + \frac{m+1}{(\delta_m^q)^2} + \frac{m}{(\delta_{m-1}^q)^2}\Bigg)\Bigg) \\
& + \frac{\Omega^2}{4}\Bigg(\frac{\sqrt{m+1}\sqrt{m+3}}{\delta_m^q \delta_{m+2}^q}\tilde{g}_{m+1} + \frac{\sqrt{m}\sqrt{m+2}}{\delta_{m-1}^q \delta_{m+1}^q}\tilde{g}_{m-1}\Bigg) \\
& + \frac{\Omega^2}{4}\sqrt{m+1}\sqrt{m+2}\Bigg(\frac{g'_{m+2}}{\delta_m^q(\delta_m^q + \delta_{m+1}^q)} - \frac{g'_{m+1}}{\delta_m^q \delta_{m+1}^q} + \frac{g'_m}{\delta_{m+1}^q(\delta_m^q + \delta_{m+1}^q)}\Bigg)\Bigg) + \text{h.c.}
\end{aligned}
\tag{9.65}
$$

with

$$
\begin{aligned}
\tilde{g}_m &:= \frac{g\alpha\Omega\sqrt{m+1}\sqrt{m+2}}{2\Delta_m \Delta_{m+1}} && (9.66)\\
g'_m &:= \frac{g\Omega\Delta_{-1}}{2\Delta_m \Delta_{m-1}}, && (9.67)
\end{aligned}
$$

and

$$
\begin{aligned}
H^{DD}_{\text{resid.}} = (e^{i\phi}a + \text{h.c.}) \sum_{m=0}^{\infty} |m\rangle\langle m| \Bigg(& g'_m \Big(1 - \frac{\Omega^2}{4}\Big(\frac{m+1}{(\delta^q_m)^2} + \frac{m}{(\delta^q_{m-1})^2}\Big)\Big) \\
& + \frac{\Omega^2}{4}\Big(\frac{m+1}{(\delta^q_m)^2} g'_{m+1} + \frac{m}{(\delta^q_{m-1})^2} g'_{m-1}\Big) \\
& + \frac{\Omega^2}{4}\Big(\frac{\sqrt{m+1}\sqrt{m+2}\tilde{g}_m}{\delta^q_m(\delta^q_m + \delta^q_{m+1})} + \frac{\sqrt{m}\sqrt{m+1}\tilde{g}_{m-1}}{\delta^q_m \delta^q_{m-1}} \\
& \quad + \frac{\sqrt{m-1}\sqrt{m}\tilde{g}_{m-2}}{\delta^q_{m-1}(\delta^q_{m-2} + \delta^q_{m-1})}\Big)\Bigg) \\
& - \frac{\Omega}{2} e^{2i\phi} a \sum_{m=0}^{\infty} |m\rangle\langle m+1| \frac{\sqrt{m+1}}{\delta^q_m}(g'_{m+1} - g'_m) + \text{h.c.} \\
& - \frac{\Omega}{2} a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+1| \Big(\frac{\sqrt{m+1}}{\delta^q_m}(g'_{m+1} - g'_m) + \frac{\sqrt{m+2}}{\delta^q_{m+1}}\tilde{g}_m - \frac{\sqrt{m}}{\delta^q_{m-1}}\tilde{g}_{m-1}\Big) + \text{h.c.} \\
& + \frac{\Omega^2}{4} e^{3i\phi} a \sum_{m=0}^{\infty} |m\rangle\langle m+2| \sqrt{m+1}\sqrt{m+2}\Big(\frac{g'_{m+2}}{\delta^q_m(\delta^q_m + \delta^q_{m+1})} - \frac{g'_{m+1}}{\delta^q_m \delta^q_{m+1}} \\
& \quad + \frac{g'_m}{\delta^q_{m+1}(\delta^q_m + \delta^q_{m+1})}\Big) + \text{h.c.} \\
& - \frac{\Omega}{2} e^{2i\phi} a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+3| \Big(\frac{\sqrt{m+1}}{\delta^q_m}\tilde{g}_{m+1} - \frac{\sqrt{m+3}}{\delta^q_{m+2}}\tilde{g}_m\Big) + \text{h.c.} \\
& + \frac{\Omega^2}{4} e^{3i\phi} a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+4| \Big(\frac{\sqrt{m+1}\sqrt{m+2}\tilde{g}_{m+2}}{\delta^q_m(\delta^q_m + \delta^q_{m+1})} - \frac{\sqrt{m+4}\sqrt{m+1}\tilde{g}_{m+1}}{\delta^q_m \delta^q_{m+3}} \\
& \quad + \frac{\sqrt{m+3}\sqrt{m+4}\tilde{g}_m}{\delta^q_{m+3}(\delta^q_{m+3} + \delta^q_{m+2})}\Big) + \text{h.c.}
\end{aligned} \tag{9.68}
$$

All terms in $H^{DD}_{\text{resid.}}$ are relatively small and off-resonant with the $|20\rangle \leftrightarrow |01\rangle$ transition so we expect them to have a small effect and we do not proceed with higher orders of SWTs.

9.5.4. ANALYSIS OF THE $|20\rangle \leftrightarrow |01\rangle$ AVOIDED CROSSING

In this section we give the methods used to calculate the curves in Fig. 9.1(c),(e).

We define ω_d^* as the drive frequency corresponding to the center of the $|20\rangle \leftrightarrow |01\rangle$ avoided crossing of the full Hamiltonian H as given in Eq. (9.1). Then the exact value of the effective $|20\rangle \leftrightarrow |01\rangle$ coupling $\tilde{g}$ is given by half the energy separation at that point. The avoided crossing can be found numerically by exact diagonalization as a function of ω_d.

In the subspace $\mathscr{S} = \text{span}\{|20\rangle, |01\rangle\}$ we can write H as $H|_{\mathscr{S}} \equiv -\eta(\omega_d)Z/2 + \tilde{g}(\omega_d)[\cos(\phi)X + \sin(\phi)Y] = -\eta(\omega_d)Z/2 + \tilde{g}(\omega_d)X$ for $\phi = 0$ as in Section 9.2.1. As we want to implement a $|20\rangle \leftrightarrow |01\rangle$ π rotation, we notice that the choice of ϕ, i.e. the choice of rotation axis in the equator of the Bloch sphere, is irrelevant. We have also ignored a term proportional to the identity I, which gives a phase difference with respect to states outside of $\mathscr{S}$, in particular

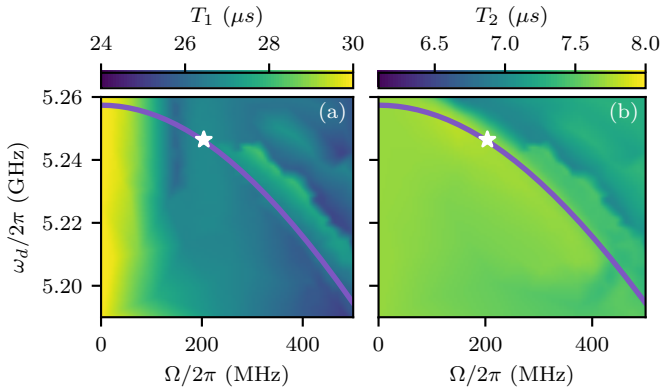

Figure 9.6: Effective T_1 (a) and T_2 (b) which account for the extra decoherence caused by the drive during the time slot $T_{\text{slot}} = 440$ ns. We can see that the variation is small as a function of the drive amplitude compared to the values at $\Omega = 0$. The white star indicates the chosen operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_p = 178.6$ ns, see Section 9.2.2). The purple line corresponds to the higher order estimate of the optimal drive frequency ω_d^* as a function of Ω (see Fig. 9.1(c)). The heatmaps are sampled using the *adaptive* package [67].

between the computational and leakage subspaces of the transmon. However, this phase is irrelevant if $|20\rangle$ is swapped entirely onto $|01\rangle$ since the latter decays and dephases fast, thus suppressing any phase coherence. As demonstrated in Section 9.2.2 the res-LRU can reach a very high R, for which the effect of this phase is then minimal. Assuming that $H^{DD}_{\text{resid.}}$ in Eq. (9.68) is negligible, an analytical approximation of η is given by

$$\eta(\omega_d) \approx \langle 20|H_0^{DD}(\omega_d)|20\rangle - \langle 01|H_0^{DD}(\omega_d)|01\rangle\,, \tag{9.69}$$

where we have made the dependence of H_0^{DD} in Eq. (9.62) on ω_d explicit. This holds since then H_0^{DD} accounts for all the Stark shifts of $|20\rangle$ and $|01\rangle$ due to the capacitive coupling and the drive (up to the given orders). The center of the avoided crossing is found by imposing the condition $\eta(\omega_d) = 0$. As the explicit expression that can be extracted from Eq. (9.62) is not analytically solvable, we use the secant method available in *scipy* to find ω_d^* that fulfills this condition in Eq. (9.69). It is then straightforward to compute the (approximate) analytical estimate for the effective coupling as $\tilde{g}(\omega_d^*) = |\langle 01|H^{DD}_{\text{eff.coupl.}}(\omega_d^*)|20\rangle|$ from Eq. (9.65), which is plotted in Fig. 9.1(e).

9.6. FURTHER CHARACTERIZATION OF THE READOUT-RESONATOR LRU

9.6.1. EFFECTIVE T_1 AND T_2 DUE TO THE DRIVE

In this section we discuss the effects of the readout-resonator LRU within the computational subspace when applied to a non-leaked transmon. As pulses at different (ω_d, Ω) points have a different duration t_p, it would not be fair to report an effective T_1 and T_2 during t_p. That is, stronger pulses potentially produce lower T_1 and T_2, but they also take less time to implement the LRU. However, the overall disturbance to the qubit is

a combination of these two factors. We thus report an effective T_1 and T_2 during the whole time slot of $T_{\text{slot}} = 440$ ns, leading to a uniform metric for the whole (ω_d, Ω) landscape. Specifically, to estimate T_1 we prepare the state $|1\rangle\langle 1| \otimes \sigma_{\text{th}}$, we simulate the Lindblad equation in Eq. (9.12) and we evaluate the remaining population $p^{|1\rangle}$ in $|1\rangle$ at the end of the time slot after tracing out the resonator. Assuming that $p^{|1\rangle} = e^{-T_{\text{slot}}/T_1}$ we then compute T_1 by inverting this formula. To estimate T_2 we prepare $|+\rangle\langle +| \otimes \sigma_{\text{th}}$ and we evaluate the decay of the off-diagonal transmon matrix element $|0\rangle\langle 1|$ as this is directly available in simulation (rather than simulating a full Ramsey experiment). We then invert $\left|\langle 0|\,\text{Tr}_r(\rho(T_{\text{slot}}))|1\rangle\right| = e^{-T_{\text{slot}}/T_2}/2$ to get T_2.

Figure 9.6 shows the resulting effective T_1 and T_2. In Fig. 9.6(a) one can see that T_1 decreases by at most 15% as a function of Ω, showing that a short t_p mostly counterbalances the effect of a strong Ω. In particular, $T_1 \approx 27.1$ μs at the operating point. On the other hand, one can notice that T_1 dips around $\Omega_{\text{cr}}/2\pi = 143$ MHz, where the pulses are very long, suggesting that driving slightly into the underdamped regime is favorable. In Fig. 9.6(b) one can see that the value of T_2 is about 7.7 μs at $\Omega = 0$, i.e. when no pulse is applied. This has to be contrasted with the input T_2 parameter of 30 μs inserted in the Lindblad equation (see Table 9.1). We assume that that implicitly accounts for dephasing caused by flux noise only. Photon-shot noise from the resonator is a further dephasing source which is explicitly included in these simulations. The combination of flux and photon-shot noise leads to the actual effective T_2 reported in Fig. 9.6(b). We note that if $\bar{n} = 0$ then the effective T_2 at $\Omega = 0$ would exactly match the input of 30 μs. While the effective T_2 can be restored from 7.7 μs to 30 μs with colder resonators or by engineering different system parameters altogether, the important information from Fig. 9.6(b) is that T_2 barely changes as a function of Ω. Combined with the similar result for T_1, this means that the drive causes only a marginal effect within the computational subspace. Notice that in the region where the readout-resonator LRU is most effective (just above the purple line in Fig. 9.6(b)), T_2 is even slightly higher than at $\Omega = 0$ (7.9 versus 7.7 μs). We attribute this to the fact that the pulse temporarily reduces the excited-state population in the resonator (see Fig. 9.2(d)). In this way photon-shot noise is reduced until the resonator re-thermalizes, however at the cost of some leakage of the transmon.

In Fig. 9.2(d) one can notice that a non-negligible amount of population ends up in $|10\rangle$ from the initial state $|0\rangle\langle 0| \otimes \sigma_{\text{th}}$. This corresponds to an excitation rate $T_1^{\uparrow} \approx 256$ μs at the operating point. We backtrack this source of error to a combination of the drive and the jump operator $a^{\dagger}$, corresponding to the drive inducing a transmon excitation rate based on the resonator excitation rate. However, as here $T_1^{\uparrow} \gg \max\{T_1, T_2\}$, it is not a limiting factor and we have not included it in the Surface-17 simulations.

9.6.2. LONG-DRIVE LIMIT IN THE UNDERDAMPED REGIME AND ITS DRAWBACKS AS A LRU

In this section we compare the reset schemes in Refs. [55, 56] versus Ref. [54] in terms of their performance as a LRU in the underdamped regime. The approach of Refs. [55, 56], which we have adopted in Section 9.2.2, aims at swapping $|20\rangle$ and $|01\rangle$ by targeting the first minimum of the oscillations induced by the drive (switching the drive off afterwards). As shown in Section 9.2.2, this approach allows for a residual leakage population $p_{\text{op.}}^{|2\rangle} \approx 0.5\%$ at the operating point (see Fig. 9.2(a)), given our parameters (see Table 9.1). While

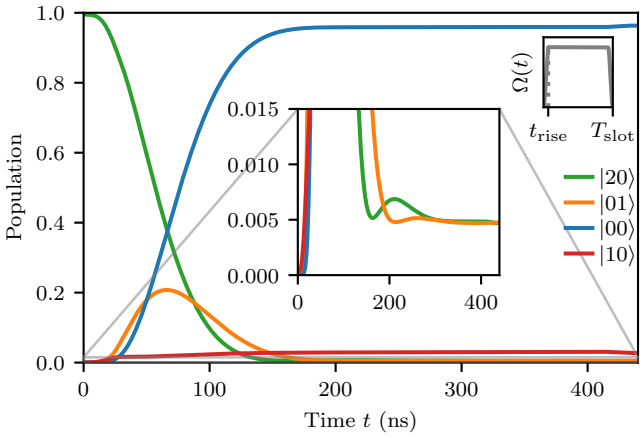

Figure 9.7: Time evolution from the initial state $|2\rangle\langle 2| \otimes \sigma_{\text{th}}$ for $t_{\text{rise}} = 30$ ns and for an otherwise always-on drive during T_{slot}. This is simulated with the same $\Omega/2\pi \approx 204$ MHz and $\omega_d/2\pi \approx 5.2464$ GHz as at the operating point in Fig. 9.2.

this already reaches thermal-state levels (here $\bar{n} = 0.5\%$) with the considered system parameters, the approach in Ref. [54] could be used in general to achieve an even lower or similar $p^{|2\rangle}$ (in particular for lower κ's).

The approach in Ref. [54] keeps the drive on for a much longer period of time (at least one more oscillation) allowing both the populations in $|20\rangle$ and $|01\rangle$ to decay to almost 0, modulo thermal excitations. Figure 9.7 shows that it is indeed possible to suppress these populations to thermal-state levels, where we use the same (Ω, ω_d) as at the operating point (see Section 9.2.2). However, we see that for the operating point there is almost no gain by using this approach. Furthermore, this approach costs much more time and could exceed $T_{\text{slot}} = 440$ ns if κ is not as high as assumed here. In particular, in that case the first few minima after the first one could be slightly higher, due to transmon decoherence, and one would need to wait even longer to overcome this effect.

Another disadvantage of the approach in Ref. [54] is that the disturbance to the qubit is stronger as the drive is kept on for a longer period of time. E.g., in Fig. 9.7 one can see that $|00\rangle$ and $|10\rangle$ reach an equilibrium thanks to the drive (even in the presence of relaxation), where the population in $|10\rangle$ is higher than in Fig. 9.2(b). By evaluating T_1 we find $T_1 \approx 23$ μs instead of 27 μs (see Section 9.6.1). Furthermore, if one would have to use a $t_p > T_{\text{slot}}$ when κ is lower than here, then the QEC cycle would get longer, affecting the coherence of all qubits, not only of the high-frequency data qubits to which the res-LRU is applied.

9.6.3. SENSITIVITY TO RESIDUAL ZZ CROSSTALK

In a multi-transmon chip, each transmon is coupled to one or more neighbors. In general, if the coupling is not tunable there can be some residual ZZ crosstalk (see Section 3.2.7), i.e. a shift of the transmon frequency by an amount ζ based on whether each neighboring transmon is in $|1\rangle$ instead of $|0\rangle$. In this section we study the effect of this ZZ coupling

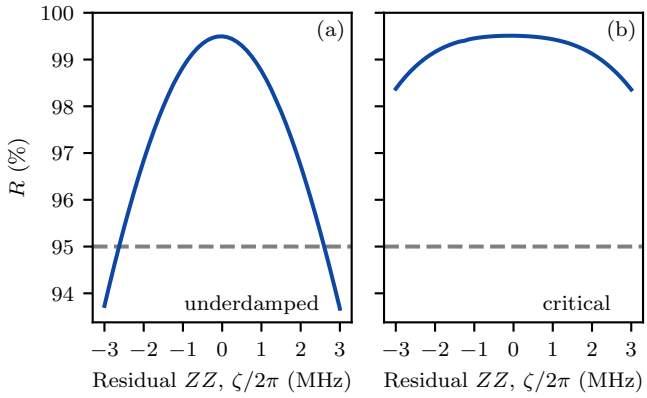

Figure 9.8: Sensitivity of the leakage-reduction rate R of the readout-resonator LRU as a function of the overall residual ZZ coupling ζ. (a) Underdamped regime, specifically at the operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_p = 178.6$ ns, see Section 9.2.2). (b) Critical regime ($\Omega/2\pi \approx 143$ MHz, $\omega_d/2\pi \approx 5.252$ GHz, $t_p = 440$ ns).

on the readout-resonator LRU, which we assume being tuned up when all neighbors are in $|0\rangle$. We do not include neighboring transmons in our simulations, so we mimic it by shifting the transmon frequency (while keeping the drive parameters fixed).

In Fig. 9.8 we perform the analysis for the operating point (see Section 9.2.2), which resides in the underdamped regime, and for the critical point. In both cases the leakage-reduction rate R scales seemingly quadratically. In the underdamped regime the pulse targets the first minimum of the damped Rabi oscillations, so it is more sensitive to a variation in frequency than in the critical regime. However, we can observe that for $|\zeta|/2\pi \lesssim 2$ MHz (note that this is the cumulative ZZ coupling over all neighbors) R stays above 95%, which is the conservative value we have used in Section 9.3.4 and for which the logical error rate was already close to optimal in Surface-17 (see Section 9.7.2). Regarding other performance parameters of the LRU, we find that L_1^{LRU} scales in the same relative way as R by unitarity, whereas T_1, T_2 and $T_1^{\uparrow}$ vary by $\lesssim 1\%$.

9.7. FURTHER SURFACE-17 CHARACTERIZATION

9.7.1. DETAILS ABOUT THE DENSITY-MATRIX SIMULATIONS

The parameters used in this work are reported in Table 9.2.

RES-LRU IN *quantumsim*

A comprehensive review of the density-matrix simulations and the use of the *quantumsim* package [59] is available in Refs. [47, 51] (see also Section 3.3.2). In this section we explain the specific implementation of the newly introduced res-LRU, expressed in the Pauli Transfer Matrix formalism.

We construct a "phenomenological" Lindblad model with input parameters R, L_1^{LRU} and $t_{\text{res-LRU}}$. We use the Pauli Transfer Matrix $S_{\text{res-LRU}} = S_{\uparrow} S_{\downarrow}$, where $S_{\downarrow}$ is the Pauli Transfer

Parameter	Value
Relaxation time T_1	30 μs
Sweetspot pure-dephasing time $T_{\phi,\max}$	60 μs
High-freq. pure-dephasing time at interaction point $T_{\phi,\text{int}}$	8 μs
Mid-freq. pure-dephasing time at interaction point $T_{\phi,\text{int}}$	6 μs
Mid-freq. pure-dephasing time at parking point $T_{\phi,\text{park}}$	8 μs
Low-freq. pure-dephasing time at parking point $T_{\phi,\text{park}}$	9 μs
Single-qubit gate time t_{single}	20 ns
Two-qubit interaction time t_{int}	30 ns
Single-qubit phase-correction time t_{cor}	10 ns
Readout-resonator LRU time $t_{\text{res-LRU}}$	100 ns
$\|1\rangle \leftrightarrow \|2\rangle$ π-pulse time $t_{\pi\text{-LRU}}$	20 ns
Measurement time t_{m}	580 ns
QEC-cycle time t_{c}	800 ns

Table 9.2: The parameters for the qubit coherence times and for the gate, LRU, measurement and QEC-cycle durations used in the density-matrix simulations. The interaction point corresponds to the frequency to which a transmon is fluxed to implement a CZ, whereas the parking point to the frequency at which the ancilla qubits are parked during measurement [58].

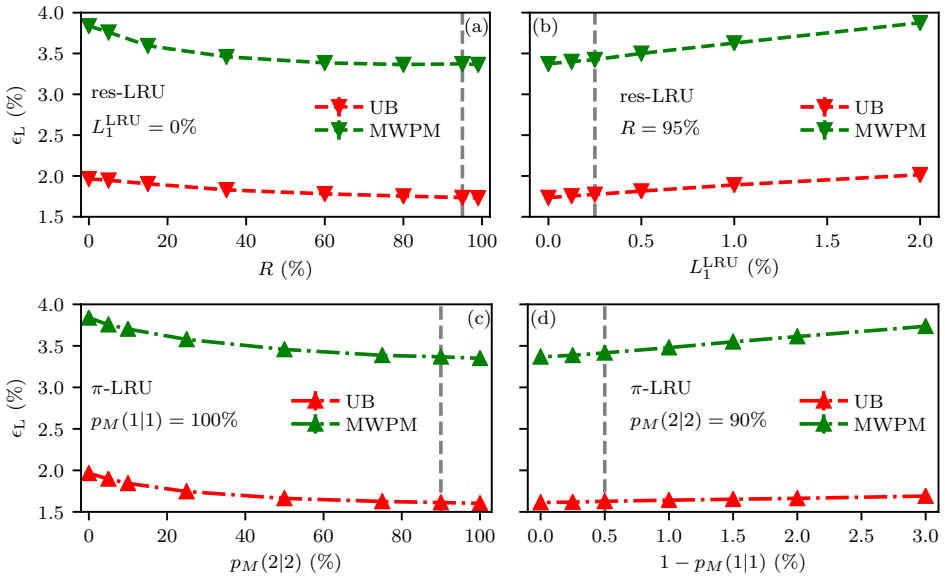

Figure 9.9: Logical error rate ε_L per QEC cycle as a function of various LRU parameters. (a),(b) use only the res-LRU, while (c),(d) the π-LRU. We fix $L_1 = 0.5\%$ for all. Vertical dashed lines indicate the values considered in Section 9.3.4. These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are smaller than the symbol size.

Matrix of the superoperator $\mathscr{S}_\downarrow = e^{t_{\text{res-LRU}}\mathscr{L}_\downarrow}$ and the Lindbladian $\mathscr{L}_\downarrow$ has the quantum jump operator

$$L_\downarrow = \frac{1}{\sqrt{\frac{t_{\text{res-LRU}}}{-\log(1-R_{\text{sim}})}}} |0\rangle\langle 2| \tag{9.70}$$

with R_{sim} to be determined. Besides this, $\mathscr{L}_\downarrow$ has the standard qutrit jump operators for relaxation and dephasing [10, 47]. On the other hand, $S_\uparrow$ is the Pauli Transfer Matrix of the superoperator $\mathscr{S}_\uparrow = e^{\mathscr{L}_\uparrow}$ and the Lindbladian $\mathscr{L}_\uparrow$ has a single jump operator

$$L_\uparrow = \frac{1}{\sqrt{\frac{1}{-\log(1-2L_1^{\text{LRU}})}}} |2\rangle\langle 0| \tag{9.71}$$

since relaxation and dephasing during $t_{\text{res-LRU}}$ are already accounted for by $S_\downarrow$. In this way, calling $p_i^{|j\rangle}, p_f^{|j\rangle}$ the populations before and after the res-LRU, if we apply $S_{\text{res-LRU}}$ on a non-leaked transmon we get $p_f^{|2\rangle} = 2L_1^{\text{LRU}} p_i^{|0\rangle}$, consistently with Section 9.3.2. Instead, if we apply $S_{\text{res-LRU}}$ to a leaked transmon ($p_i^{|2\rangle} = 1$) we get $p_f^{|2\rangle} \approx 1 - R_{\text{sim}} + 2L_1^{\text{LRU}}$. By fixing $R_{\text{sim}} = R + 2L_1^{\text{LRU}}$ we match the definition of R in Section 9.3.2 as well. The approximation is very good for large R and low L_1^{LRU}, which is precisely the interesting regime for res-LRU that we have explored.

DECODING

In this section we provide additional information on the UB and MWPM decoders [51, 72].

UB considers the 32 computational states that differ by a purely X error on top of $|0\rangle_{\text{L}}$ and that are independent (i.e. they cannot be obtained from each other by multiplication with an X-type stabilizer). At the end of each QEC cycle n, each possible final Z syndrome is compatible with a pair of these states, where one can be associated with $|0\rangle_{\text{L}}$ and the other with $|1\rangle_{\text{L}}$ as they differ by the application of any representation of X_{L}. The largest overlap of these two states with the diagonal of the density matrix at QEC cycle n corresponds to the maximum probability of correctly guessing whether a X_{L} error has occurred or not upon performing a logical measurement of Z_{L}. The latter is assumed to be performed by measuring all data qubits in the $\{|0\rangle, |1\rangle, |2\rangle\}$ basis and computing the overall parity. To compute the parity we assume that a $|2\rangle$ is declared as a $|1\rangle$ since decoders usually do not use information about leakage (and since measurements often declare $|2\rangle$ as a $|1\rangle$ rather than as a $|0\rangle$). Then UB computes $\mathscr{F}_{\text{L}}(n)$ by weighing this probability with the chance of measuring the given final Z syndrome (conditioned on the density matrix) and by summing over all possible syndromes. In other words, UB always finds the correction that maximizes the likelihood of the logical measurement returning the initial state, here $|0\rangle_{\text{L}}$. As UB uses information generally hidden in the density matrix, it gives an upper bound to the performance of any realistic decoder, which can at most use the syndrome information extracted via the ancilla qubits.

MWPM tries to approximate the most likely correction by finding the lowest weight correction, which is a good approximation when physical error rates are relatively low. As the ancilla qubits can be faulty, the decoding graph is three dimensional. In particular,

we allow for space-like edges corresponding to data-qubit errors, time-like edges corresponding to ancilla-qubit errors and spacetime-like edges corresponding to data-qubit errors occurring in the middle of the parity-check circuit. The weights are extracted with the adaptive algorithm in Ref. [74] from a simulation (10^5 runs of 20 QEC cycles each) without leakage and an otherwise identical error model. Similarly to UB, for decoding we assume that a $|2\rangle$ is declared as a $|1\rangle$ since the standard MWPM does not account for leakage.

9.7.2. LOGICAL ERROR RATE AS A FUNCTION OF THE LRU PARAMETERS

We study the variation in the logical error rate ε_L per QEC cycle as a function of the performance parameters of the LRUs. Here we fix $L_1 = 0.5\%$ as it is easier to visualize variations in ε_L with a relatively large L_1. The leakage-reduction rate R and the readout probability $p_M(2|2)$ play similar roles for the res-LRU and π-LRU, respectively. In Fig. 9.9(a),(c) one can see that this is the case and that the values of ε_L at the parameters used in Section 9.3.4 ($R = 95\%$ and $p_M(2|2) = 90\%$) are very close to their best values (at least for this system size). This shows that the advantages of a larger R or $p_M(2|2)$ are marginal. We attribute this to the fact that leakage is exponentially suppressed with an already quite large exponent. Furthermore, the parameters L_1^{LRU} and $1 - p_M(1|1) = p_M(2|1)$, regulating the induced leakage, play similar roles as well, as Fig. 9.9(b),(d) show. We see that ε_L is more sensitive to L_1^{LRU} and $1 - p_M(1|1)$ compared to R and $p_M(2|2)$. In particular we see that ε_L is slightly larger at the parameters used in Section 9.3.4 ($L_1^{\text{LRU}} = 0.25\%$ and $1 - p_M(1|1) = 0.5\%$) rather than at 0, although the difference is small.

9.7.3. EFFECT OF THE LEAKAGE CONDITIONAL PHASES ON THE LOGICAL ERROR RATE

As defined in the main text the leakage conditional phases are the phases that a non-leaked transmon acquires when interacting with a leaked one during a CZ. Here we denote them as $\phi^{\mathscr{L}}_{\text{flux}}$ and $\phi^{\mathscr{L}}_{\text{stat}}$ depending on whether the lower or the higher frequency transmon of the pair is leaked, respectively, and we use $\phi^{\mathscr{L}}$ to indicate either of them. Furthermore, in this section we use the notation $|\text{low-f. transmon}, \text{high-f. transmon}\rangle$. Note that for a CZ between two qutrits in principle there are 9 phases $(\phi_{00}, \phi_{01}, \phi_{10}, \phi_{11}, \phi_{02}, \phi_{20}, \phi_{21}, \phi_{12}, \phi_{22})$, where the first 4 are fixed to $0, 0, 0, \pi$, respectively. Of the 5 phases containing a $|2\rangle$ we consider only two of them here, i.e. $\phi^{\mathscr{L}}_{\text{stat}} = \phi_{02} - \phi_{12}$ and $\phi^{\mathscr{L}}_{\text{flux}} = \phi_{20} - \phi_{21}$ as defined above. This is because in our leakage model (see Section 8.10.2) we set to 0 the coherence between the computational and leakage subspace of each qutrit, motivated by the fact that leakage is projected relatively fast and that the stabilizer measurements ideally prevent any interference effect. This means that the individual phases are global phases, whereas their difference cannot be gauged away when the non-leaked qubit is in a superposition of $|0\rangle$ and $|1\rangle$.

For a flux-based CZ with conditional phase π for $|11\rangle$, ideally one should have $\phi^{\mathscr{L}}_{\text{flux}} = 0$ and $\phi^{\mathscr{L}}_{\text{stat}} = \pi$ as $|02\rangle$ acquires a conditional phase equal and opposite to $|11\rangle$ (see also Section 8.2). If only $|12\rangle$ and $|21\rangle$ are coupled in the 3-excitation manifold, it holds $\phi^{\mathscr{L}}_{\text{stat}} = \pi - \phi^{\mathscr{L}}_{\text{flux}}$. The strength of the repulsion times the CZ duration gives e.g. $\phi^{\mathscr{L}}_{\text{flux}} \sim \pi/4$ for the parameters in Table 8.2. However, $|03\rangle$ interacts with $|12\rangle$ and $|21\rangle$ and breaks the

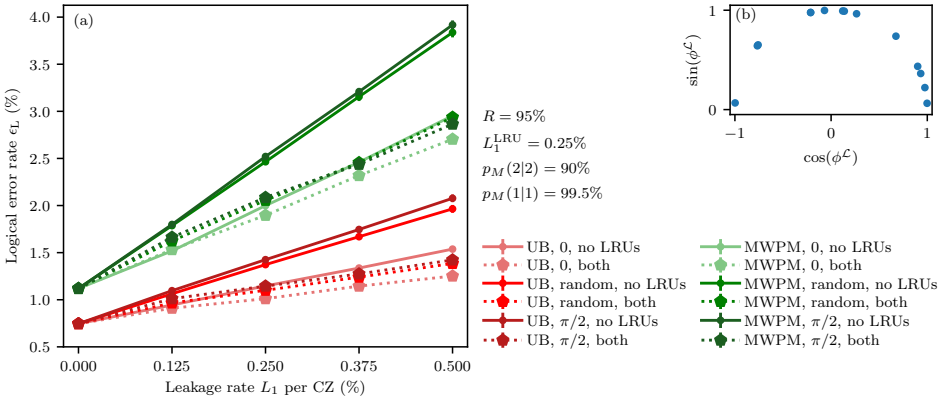

Figure 9.10: Variation of the logical error rate ε_L for different choices of leakage conditional phases $\phi^{\mathcal{L}}$. (a) ε_L per QEC cycle for UB (shades of red) and MWPM (shades of green) versus L_1, in the cases with: no LRUs and both LRUs, each for all $\phi^{\mathcal{L}}$ set to 0, $\pi/2$ or uniformly random in $[0,\pi]$. These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are mostly smaller than the symbol size. (b) The random values for $\phi^{\mathcal{L}}$ used across this chapter. These values are extracted from a uniform distribution in $[0,\pi]$. We have excluded negative values as $\pm\phi^{\mathcal{L}}$ corresponds to the same chance of spreading a Z error under the twirling action of the parity-check measurements.

relationship above, for which we can consider $\phi^{\mathcal{L}}_{\text{flux}}$ and $\phi^{\mathcal{L}}_{\text{stat}}$ as effectively unconstrained. The randomized values used across the main text are reported in Fig. 9.10(b). We use 14 values, of which 3 for $\phi^{\mathcal{L}}_{\text{stat}}$ and 3 for $\phi^{\mathcal{L}}_{\text{flux}}$ when each high-frequency data qubit is leaked or interacts with a leaked ancilla qubit, respectively, and 8 only for $\phi^{\mathcal{L}}_{\text{stat}}$ when each ancilla qubit is leaked and interacts with a low-frequency data qubit (as low-frequency data qubits cannot leak themselves).

In this section we study the dependence of the logical error rate ε_L on the leakage conditional phases, without discussing how one would engineer the system to tune them to certain values. The best-case scenario to minimize ε_L is to set all $\phi^{\mathcal{L}} = 0$, since no Z rotations are spread then. Instead, the worst-case scenario corresponds to all $\phi^{\mathcal{L}} = \pi/2$, since under the twirling effect of the parity-check measurements this corresponds to spreading a Z error with 50% chance. Notice that, if all $\phi^{\mathcal{L}} = \pi$, overall the spread errors amount to a stabilizer (except in the QEC cycle in which leakage occurs), so it is close to the best-case scenario.

Figure 9.10(a) compares the logical performance for both UB and MWPM in the cases where $\phi^{\mathcal{L}} = 0$, $\phi^{\mathcal{L}} = \pi/2$ and when they are random as in Fig. 9.5 and in the rest of this chapter. First, one can notice that the performance of random $\phi^{\mathcal{L}}$ is very close to the worst-case scenario ($\phi^{\mathcal{L}} = \pi/2$). This is due to the fact that it is not necessary to spread an error on every qubit with 50% chance each to cause a logical error with high probability. Second, one can see that just tuning all $\phi^{\mathcal{L}} = 0$ without implementing LRUs is almost as good (or even better) as using the LRUs when $\phi^{\mathcal{L}}$ are random. We attribute this to the fact that one of the major effects of the LRUs is to prevent correlated errors being spread by a leaked qubit for many QEC cycles. Tuning $\phi^{\mathcal{L}} = 0$ achieves this as well, but it still does not address the fact that the code distance is effectively reduced if a data qubit

stays leaked and that the full stabilizer information is not accessible as long as an ancilla qubit is leaked. Indeed, using LRUs even when $\phi^{\mathscr{L}} = 0$ always allows for a lower logical error rate (see Fig. 9.10(a)). Furthermore, the reduction in distance and the corruption of the stabilizer information suggest that a threshold would still likely be low without using LRUs.

REFERENCES

[1] F. Battistel, B. Varbanov, and B. Terhal, *Hardware-efficient leakage-reduction scheme for quantum error correction with superconducting transmon qubits,* PRX Quantum **2**, 030314 (2021).

[2] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, *Quantum supremacy using a programmable superconducting processor,* Nature **574**, 505–510 (2019).

[3] P. Jurcevic, A. Javadi-Abhari, L. S. Bishop, I. Lauer, D. F. Bogorin, M. Brink, L. Capelluto, O. Günlük, T. Itoko, N. Kanazawa, A. Kandala, G. A. Keefe, K. Krsulich, W. Landers, E. P. Lewandowski, D. T. McClure, G. Nannicini, A. Narasgond, H. M. Nayfeh, E. Pritchett, M. B. Rothwell, S. Srinivasan, N. Sundaresan, C. Wang, K. X. Wei, C. J. Wood, J.-B. Yau, E. J. Zhang, O. E. Dial, J. M. Chow, and J. M. Gambetta, *Demonstration of quantum volume 64 on a superconducting quantum computing system,* Quantum Science and Technology **6**, 025020 (2021).

[4] L. Egan, D. M. Debroy, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Newman, M. Li, K. R. Brown, M. Cetina, and C. Monroe, *Fault-tolerant operation of a quantum error-correction code,* (2020), arXiv:2009.11482 [quant-ph] .

[5] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo, *Restless tuneup of high-fidelity qubit gates,* Phys. Rev. Applied **7**, 041001 (2017).

[6] Z. Chen, J. Kelly, C. Quintana, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Lucero, E. Jeffrey, A. Megrant, J. Mutus, M. Neeley, C. Neill, P. J. J. O'Malley, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, T. C. White, A. N. Korotkov, and J. M. Martinis, *Measuring and suppressing quantum state leakage in a superconducting qubit,* Phys. Rev. Lett. **116**, 020501 (2016).

[7] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and J. M. Martinis, *Superconducting quantum circuits at the surface code threshold for fault tolerance.* Nature **508**, 500 (2014).

[8] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, *Procedure for systematically tuning up cross-talk in the cross-resonance gate,* Physical Review A **93**, 060302 (2016).

[9] S. S. Hong, A. T. Papageorge, P. Sivarajah, G. Crossman, N. Didier, A. M. Polloreno, E. A. Sete, S. W. Turkowski, M. P. da Silva, and B. R. Johnson, *Demonstration of a parametrically activated entangling gate protected from flux noise,* Physical Review A **101** (2020).

[10] M. A. Rol, F. Battistel, F. K. Malinowski, C. C. Bultink, B. M. Tarasinski, R. Vollmer, N. Haider, N. Muthusubramanian, A. Bruno, B. M. Terhal, and L. DiCarlo, *Fast, high-fidelity conditional-phase gate exploiting leakage interference in weakly anharmonic superconducting qubits,* Phys. Rev. Lett. **123**, 120502 (2019).

[11] V. Negîrneac, H. Ali, N. Muthusubramanian, F. Battistel, R. Sagastizabal, M. S. Moreira, J. F. Marques, W. J. Vlothuizen, M. Beekman, C. Zachariadis, N. Haider, A. Bruno, and L. DiCarlo, *High-fidelity controlled-Z gate with maximal intermediate leakage operating at the speed limit in a superconducting quantum processor,* Phys. Rev. Lett. **126**, 220502 (2021).

[12] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Tunable Coupling Scheme for Implementing High-Fidelity Two-Qubit Gates,* Physical Review Applied **10**, 054062 (2018).

[13] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Z. Chen, K. Satzinger, R. Barends, F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, S. Boixo, D. Buell, B. Burkett, Y. Chen, R. Collins, E. Farhi, A. Fowler, C. Gidney, M. Giustina, R. Graff, M. Harrigan, T. Huang, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, P. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, E. Lucero, J. McClean, M. McEwen, X. Mi, M. Mohseni, J. Y. Mutus, O. Naaman, M. Neeley, M. Niu, A. Petukhov, C. Quintana, N. Rubin, D. Sank, V. Smelyanskiy, A. Vainsencher, T. C. White, Z. Yao, P. Yeh, A. Zalcman, H. Neven, J. M. Martinis, and Google AI Quantum, *Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms,* Physical Review Letters **125** (2020).

[14] M. Kjaergaard, M. E. Schwartz, A. Greene, G. O. Samach, A. Bengtsson, M. O'Keeffe, C. M. McNally, J. Braumüller, D. K. Kim, P. Krantz, M. Marvian, A. Melville, B. M. Niedzielski, Y. Sung, R. Winik, J. Yoder, D. Rosenberg, K. Obenland, S. Lloyd, T. P. Orlando, I. Marvian, S. Gustavsson, and W. D. Oliver, *Programming a quantum computer with quantum instructions,* (2020), arXiv:2001.08838 [quant-ph] .

[15] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L.

Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler,* Phys. Rev. X **11**, 021058 (2021).

[16] T. P. Harty, D. T. C. Allcock, C. J. Ballance, L. Guidoni, H. A. Janacek, N. M. Linke, D. N. Stacey, and D. M. Lucas, *High-fidelity preparation, gates, memory, and readout of a trapped-ion quantum bit,* Phys. Rev. Lett. **113**, 220501 (2014).

[17] E. Jeffrey, D. Sank, J. Y. Mutus, T. C. White, J. Kelly, R. Barends, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. Megrant, P. J. J. O'Malley, C. Neill, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, *Fast accurate state measurement with superconducting qubits,* Phys. Rev. Lett. **112**, 190504 (2014).

[18] C. C. Bultink, M. A. Rol, T. E. O'Brien, X. Fu, B. C. S. Dikken, C. Dickel, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, R. N. Schouten, and L. DiCarlo, *Active resonator reset in the nonlinear dispersive regime of circuit QED,* Phys. Rev. Appl. **6**, 034008 (2016).

[19] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, and C. Eichler, *Rapid high-fidelity multiplexed readout of superconducting qubits,* Phys. Rev. Appl. **10**, 034040 (2018).

[20] S. Bravyi, D. Gosset, and R. König, *Quantum advantage with shallow circuits,* Science **362**, 308 (2018).

[21] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, P. Hu, X.-Y. Yang, W.-J. Zhang, H. Li, Y. Li, X. Jiang, L. Gan, G. Yang, L. You, Z. Wang, L. Li, N.-L. Liu, C.-Y. Lu, and J.-W. Pan, *Quantum computational advantage using photons,* Science **370**, 1460 (2020).

[22] R. Babbush, J. R. McClean, M. Newman, C. Gidney, S. Boixo, and H. Neven, *Focus beyond quadratic speedups for error-corrected quantum advantage,* PRX Quantum **2**, 010103 (2021).

[23] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, S. Boixo, M. Broughton, B. B. Buckley, D. A. Buell, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, S. Demura, A. Dunsworth, D. Eppens, E. Farhi, A. Fowler, B. Foxen, C. Gidney, M. Giustina, R. Graff, S. Habegger, M. P. Harrigan, A. Ho, S. Hong, T. Huang, W. J. Huggins, L. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, C. Jones, D. Kafri, K. Kechedzhi, J. Kelly, S. Kim, P. V. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, M. Lindmark, E. Lucero, O. Martin, J. M. Martinis, J. R. McClean, M. McEwen, A. Megrant, X. Mi, M. Mohseni, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, H. Neven, M. Yuezhen Niu, T. E. O'Brien, E. Ostby, A. Petukhov, H. Putterman, C. Quintana, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, D. Strain, K. J. Sung, M. Szalay, T. Y. Takeshita, A. Vainsencher, T. White, N. Wiebe, Z. J. Yao, P. Yeh, and A. Zalcman, *Hartree-Fock on a superconducting qubit quantum computer,* Science **369**, 1084–1089 (2020).

[24] 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. Hoi, C. Neill, P. J. J.

O'Malley, C. Quintana, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, *State preservation by repetitive error detection in a superconducting quantum circuit,* Nature **519**, 66 (2015).

[25] D. Ristè, S. Poletto, M. Z. Huang, A. Bruno, V. Vesterinen, O. P. Saira, and L. DiCarlo, *Detecting bit-flip errors in a logical qubit using stabilizer measurements,* Nat. Commun. **6**, 6983 (2015).

[26] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, J. M. Chow, and J. M. Gambetta, *Demonstration of weight-four parity measurements in the surface code architecture,* Phys. Rev. Lett. **117**, 210505 (2016).

[27] V. Negnevitsky, M. Marinelli, K. K. Mehta, H.-Y. Lo, C. Flühmann, and J. P. Home, *Repeated multi-qubit readout and feedback with a mixed-species trapped-ion register,* Nature **563**, 527 (2018).

[28] C. C. Bultink, T. E. O'Brien, R. Vollmer, N. Muthusubramanian, M. W. Beekman, M. A. Rol, X. Fu, B. Tarasinski, V. Ostroukh, B. Varbanov, A. Bruno, and L. DiCarlo, *Protecting quantum entanglement from leakage and qubit errors via repetitive parity measurements,* Science Advances **6**, eaay3050 (2020).

[29] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, J. Heinsoo, J.-C. Besse, M. Gabureac, A. Wallraff, and C. Eichler, *Entanglement stabilization using ancilla-based parity detection and real-time feedback in superconducting circuits,* npj Quantum Information **5** (2019), 10.1038/s41534-019-0185-4.

[30] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, N. Lacroix, G. J. Norris, M. Gabureac, C. Eichler, and A. Wallraff, *Repeated quantum error detection in a surface code,* Nature Physics **16**, 875–880 (2020).

[31] J. F. Marques, B. M. Varbanov, M. S. Moreira, H. Ali, N. Muthusubramanian, C. Zachariadis, F. Battistel, M. Beekman, N. Haider, W. Vlothuizen, A. Bruno, B. M. Terhal, and L. DiCarlo, *Logical-qubit operations in an error-detecting surface code,* Nature Physics (2021).

[32] Z. Chen, K. J. Satzinger, J. Atalaya, A. N. Korotkov, A. Dunsworth, D. Sank, C. Quintana, M. McEwen, R. Barends, P. V. Klimov, S. Hong, C. Jones, A. Petukhov, D. Kafri, S. Demura, B. Burkett, C. Gidney, A. G. Fowler, A. Paler, H. Putterman, I. Aleiner, F. Arute, K. Arya, R. Babbush, J. C. Bardin, A. Bengtsson, A. Bourassa, M. Broughton, B. B. Buckley, D. A. Buell, N. Bushnell, B. Chiaro, R. Collins, W. Courtney, A. R. Derk, D. Eppens, C. Erickson, E. Farhi, B. Foxen, M. Giustina, A. Greene, J. A. Gross, M. P. Harrigan, S. D. Harrington, J. Hilton, A. Ho, T. Huang, W. J. Huggins, L. B. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, K. Kechedzhi, S. Kim, A. Kitaev, F. Kostritsa, D. Landhuis, P. Laptev, E. Lucero, O. Martin, J. R. McClean, T. McCourt, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Newman, M. Y. Niu, T. E. O'Brien, A. Opremcak, E. Ostby, B. Pató, N. Redd, P. Roushan, N. C. Rubin, V. Shvarts, D. Strain, M. Szalay, M. D. Trevithick, B. Villalonga, T. White, Z. J. Yao, P. Yeh, J. Yoo, A. Zalcman, H. Neven, S. Boixo, V. Smelyanskiy, Y. Chen, A. Megrant,

J. Kelly, and G. Q. AI, *Exponential suppression of bit or phase errors with cyclic error correction,* Nature **595**, 383 (2021).

[33] F. W. Strauch, P. R. Johnson, A. J. Dragt, C. J. Lobb, J. R. Anderson, and F. C. Wellstood, *Quantum logic gates for coupled superconducting phase qubits,* Phys. Rev. Lett. **91**, 167005 (2003).

[34] L. DiCarlo, J. M. Chow, J. M. Gambetta, L. S. Bishop, B. R. Johnson, D. I. Schuster, J. Majer, A. Blais, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, *Demonstration of two-qubit algorithms with a superconducting quantum processor,* Nature **460**, 240 (2009).

[35] J. M. Martinis and M. R. Geller, *Fast adiabatic qubit gates using only σ_z control,* Phys. Rev. A **90**, 022307 (2014).

[36] V. Tripathi, M. Khezri, and A. N. Korotkov, *Operation and intrinsic error budget of a two-qubit cross-resonance gate,* Phys. Rev. A **100**, 012301 (2019).

[37] A. P. Babu, J. Tuorila, and T. Ala-Nissila, *State leakage during fast decay and control of a superconducting transmon qubit,* npj Quantum Information **7** (2021), 10.1038/s41534-020-00357-z.

[38] M. Werninghaus, D. J. Egger, F. Roy, S. Machnes, F. K. Wilhelm, and S. Filipp, *Leakage reduction in fast superconducting qubit gates via optimal control,* npj Quantum Information **7** (2021), 10.1038/s41534-020-00346-2.

[39] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults,* Quantum Info. Comput. **7**, 139 (2007).

[40] A. G. Fowler, *Coping with qubit leakage in topological codes,* Phys. Rev. A **88**, 042308 (2013).

[41] J. Ghosh, A. G. Fowler, J. M. Martinis, and M. R. Geller, *Understanding the effects of leakage in superconducting quantum-error-detection circuits,* Phys. Rev. A **88**, 062329 (2013).

[42] J. Ghosh and A. G. Fowler, *Leakage-resilient approach to fault-tolerant quantum computing with superconducting elements,* Phys. Rev. A **91**, 020302 (2015).

[43] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code,* Quantum Info. Comput. **15**, 997 (2015).

[44] N. C. Brown and K. R. Brown, *Comparing zeeman qubits to hyperfine qubits in the context of the surface code: $^{174}Yb^+$ and $^{171}Yb^+$,* Phys. Rev. A **97**, 052301 (2018).

[45] N. C. Brown, M. Newman, and K. R. Brown, *Handling leakage with subsystem codes,* New Journal of Physics **21**, 073055 (2019).

[46] N. C. Brown and K. R. Brown, *Leakage mitigation for quantum error correction using a mixed qubit scheme,* Phys. Rev. A **100**, 032325 (2019).

[47] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. DiCarlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code,* npj Quantum Information **6** (2020), 10.1038/s41534-020-00330-w.

[48] N. C. Brown, A. Cross, and K. R. Brown, *Critical faults of leakage errors on the surface code,* in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)* (2020) pp. 286–294.

[49] M. McEwen, D. Kafri, Z. Chen, J. Atalaya, K. J. Satzinger, C. Quintana, P. V. Klimov, D. Sank, C. Gidney, A. G. Fowler, F. Arute, K. Arya, B. Buckley, B. Burkett, N. Bushnell, B. Chiaro, R. Collins, S. Demura, A. Dunsworth, C. Erickson, B. Foxen, M. Giustina, T. Huang, S. Hong, E. Jeffrey, S. Kim, K. Kechedzhi, F. Kostritsa, P. Laptev, A. Megrant, X. Mi, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Niu, A. Paler, N. Redd, P. Roushan, T. C. White, J. Yao, P. Yeh, A. Zalcman, Y. Chen, V. N. Smelyanskiy, J. M. Martinis, H. Neven, J. Kelly, A. N. Korotkov, A. G. Petukhov, and R. Barends, *Removing leakage-induced correlated errors in superconducting quantum error correction,* Nature Communications **12** (2021), 10.1038/s41467-021-21982-y.

[50] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation,* Phys. Rev. A **86**, 032324 (2012).

[51] T. E. O'Brien, B. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise,* npj Quantum Information **3** (2017), 10.1038/s41534-017-0039-x.

[52] D. Hayes, D. Stack, B. Bjork, A. Potter, C. Baldwin, and R. Stutz, *Eliminating leakage errors in hyperfine qubits,* Physical Review Letters **124** (2020).

[53] V. Langrock and D. P. DiVincenzo, *A reset-if-leaked procedure for encoded spin qubits,* (2020), arXiv:2012.09517 [quant-ph] .

[54] P. Magnard, P. Kurpiers, B. Royer, T. Walter, J.-C. Besse, S. Gasparinetti, M. Pechal, J. Heinsoo, S. Storz, A. Blais, and A. Wallraff, *Fast and unconditional all-microwave reset of a superconducting qubit,* Phys. Rev. Lett. **121**, 060502 (2018).

[55] S. Zeytinoğlu, M. Pechal, S. Berger, A. A. Abdumalikov, A. Wallraff, and S. Filipp, *Microwave-induced amplitude- and phase-tunable qubit-resonator coupling in circuit quantum electrodynamics,* Physical Review A **91** (2015).

[56] D. Egger, M. Werninghaus, M. Ganzhorn, G. Salis, A. Fuhrer, P. Müller, and S. Filipp, *Pulsed reset protocol for fixed-frequency superconducting qubits,* Phys. Rev. Applied **10**, 044030 (2018).

[57] D. Ristè, C. C. Bultink, K. W. Lehnert, and L. DiCarlo, *Feedback control of a solid-state qubit using high-fidelity projective measurement,* Physical Review Letters **109** (2012).

[58] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, *Scalable quantum circuit and control for a superconducting surface code,* Phys. Rev. Appl. **8**, 034021 (2017).

[59] The quantumsim package can be found at https://quantumsim.gitlab.io/.

[60] T. M. Stace and S. D. Barrett, *Error correction and degeneracy in surface codes suffering loss,* Phys. Rev. A **81**, 022317 (2010).

[61] S. Nagayama, A. G. Fowler, D. Horsman, S. J. Devitt, and R. V. Meter, *Surface code error correction on a defective lattice,* New Journal of Physics **19**, 023050 (2017).

[62] J. M. Auger, H. Anwar, M. Gimeno-Segovia, T. M. Stace, and D. E. Browne, *Fault-tolerance thresholds for the surface code with fabrication errors,* Phys. Rev. A **96**, 042316 (2017).

[63] J. R. Schrieffer and P. A. Wolff, *Relation between the Anderson and Kondo Hamiltonians,* Phys. Rev. **149**, 491 (1966).

[64] S. Bravyi, D. P. DiVincenzo, and D. Loss, *Schrieffer – Wolff transformation for quantum many-body systems,* Annals of Physics **326**, 2793 (2011).

[65] E. Magesan and J. M. Gambetta, *Effective hamiltonian models of the cross-resonance gate,* Physical Review A **101** (2020).

[66] M. Boissonneault, J. M. Gambetta, and A. Blais, *Dispersive regime of circuit QED: Photon-dependent qubit dephasing and relaxation rates,* Physical Review A **79** (2009).

[67] B. Nijholt, J. Weston, J. Hoofwijk, and A. Akhmerov, *Adaptive: parallel active learning of mathematical functions,* (2019).

[68] H. P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press, Oxford, 2002).

[69] S. Haroche and J. Raimond, *Exploring the Quantum: Atoms, Cavities, and Photons*, Oxford Graduate Texts (Oxford University Press, 2006).

[70] C. J. Wood and J. M. Gambetta, *Quantification and characterization of leakage errors,* Phys. Rev. A **97**, 032306 (2018).

[71] S. Krinner, S. Lazar, A. Remm, C. Andersen, N. Lacroix, G. Norris, C. Hellings, M. Gabureac, C. Eichler, and A. Wallraff, *Benchmarking coherent errors in controlled-phase gates due to spectator qubits,* Phys. Rev. Applied **14**, 024042 (2020).

[72] T. E. O'Brien, B. M. Varbanov, and S. T. Spitz, *qgarden,* (2019).

[73] J. B. Hertzberg, E. J. Zhang, S. Rosenblatt, E. Magesan, J. A. Smolin, J.-B. Yau, V. P. Adiga, M. Sandberg, M. Brink, J. M. Chow, and J. S. Orcutt, *Laser-annealing Josephson junctions for yielding scaled-up superconducting quantum processors,* npj Quantum Information **7**, 129 (2021).

[74] S. T. Spitz, B. Tarasinski, C. W. J. Beenakker, and T. E. O'Brien, *Adaptive weight estimator for quantum error correction in a time-dependent environment,* Advanced Quantum Technologies **1**, 1800012 (2018).

10

CONCLUSION

10.1. SUMMARY AND DISCUSSION

In Section 1.1 we have identified lowering the error rates, scaling up, and keeping error rates low while scaling up as the three encompassing challenges for quantum computing. Focusing on superconducting quantum computing, in this thesis we have contributed towards the first challenge, while keeping the other challenges in perspective for the techniques that we introduced.

- In Chapter 6 we have demonstrated a high-fidelity, low-leakage controlled-phase gate (Net Zero). This two-qubit gate can serve as a building block in a fully programmable quantum computer since it is history independent, thanks to its insensitivity to long-timescale distortions. The Sudden Net Zero variant allows for easy tuneup and conditional-phase tunability. The former is important for automatized routines that are being implemented to tune up increasingly large processors [1, 2], the latter can allow for the compilation of quantum algorithms with a lower gate count [3]. Numerical simulations matched to experiment show that the gate performance can be further improved by lowering flux-noise levels, as well as lowering relaxation rates.

- In Chapter 7 we have discussed a gate-diagnostic method (Spectral Tomography) that can be used to collect detailed information about the errors affecting a gate. In particular, this method is resistant to state-preparation and measurement errors thanks to the prescription of repeating the gate multiple times, which magnifies gate errors while the other errors remain the same. The collected information can be then used to address the most relevant noise sources identified. As the method is non-scalable, it can reasonably be applied only to single- and two-qubit gates. However, it is straightforwardly applicable to single- and two-qubit *logical* gates, making it a tool to characterize logical noise channels even in large-scale processors.

– In Chapters 8 and 9 we have studied leakage in relation to quantum error correction and improving the logical fidelity by either detecting leakage or removing it via leakage-reduction units (LRUs).

  In Chapter 8 we have first shown that even a small amount of remaining leakage from the controlled-phase gate can have a profound impact on the logical performance, taking the distance-3 rotated surface code as a case study in our numerical simulations. We could attribute this to the fact that data-qubit leakage effectively reduces the code distance and ancilla-qubit leakage corrupts the syndrome and leads to a spread of correlated errors. To mitigate the effect of leakage on the code, we developed Hidden Markov Models (HMMs) that detect leakage using the characteristic syndromes that are likely the result of a qubit being leaked. In particular, we highlighted the importance of measuring the $|2\rangle$ state, even with limited fidelity, to be able to accurately detect ancilla-qubit leakage. By post-selecting out the surface-code runs where the HMMs detect leakage, we could restore the logical error rate below the memory break-even point.

  While leakage detection can be used for the targeted application of LRUs or leakage-aware decoding [4–9], post-selection is non-scalable. By bringing a leaked qubit back to the computational subspace, LRUs are instead a scalable approach.

  In Chapter 9 we proposed a scheme with two separate LRUs for data and ancilla qubits. The res-LRU for data qubits uses the readout resonator, already available on chip, as an energy sink onto which excitations are swapped from a leaked transmon. The π-LRU for ancilla qubits is a conditional operation that brings $|2\rangle$ back to $|1\rangle$ based on the detection of a $|2\rangle$. We showed that the use of these LRUs significantly restores the logical performance even when the LRUs are implemented with limited fidelity. The experimental requirements are a strong microwave drive for the res-LRU and fast conditional feedback for the π-LRU. Implementing the former with good performance will require a careful choice of qubit and resonator frequencies (as well as tunable couplers potentially); implementing the latter will require low-latency control electronics capable of applying conditional operations based on the readout of either $|0\rangle$, $|1\rangle$ and $|2\rangle$ (in particular, this requires that operations can be conditioned on more than one bit, as this can represent only two states).

  To study the threshold of the surface code in the presence of LRUs in a realistic setting, as an extension of this work, it will be necessary to implement efficient simulations on the line of the ones described in Section 5.1.3 (see also Refs. [4–6]).

10

10.2. OUTLOOK

Let us now turn our attention towards some of the specific problems within the domain of the three main challenges outlined in Section 1.1 (and recalled above). I think that the following problems are the most critical and could eventually constitute a roadblock towards building a useful superconducting quantum computer.

- **Crosstalk and tunable couplers.** As discussed in Section 3.2.7, crosstalk encompasses many different aspects. Here I focus on residual ZZ coupling. I believe that tunable couplers are indispensable for mitigating this kind of crosstalk. Being able

to isolate at will a qubit from other qubits and treating it as an atomic unit is key to systematically scale up a quantum processor. Furthermore, tunable coupling allows for a much more flexible choice of qubit frequencies and operation scheduling (versus all the requirements in e.g. Ref. [10], determined by a fixed-coupling architecture). Many designs for tunable couplers have shown promising performance and high on-off ratios [11–16]. However, the problem of the current generation of tunable couplers is that, in a large processor with many qubits and tunable couplers between them, the tuning of one coupler can affect the Hamiltonian of the circuit and in turn affect the optimal tuning of the other. This makes it hard to achieve the zero-coupling condition across all pairs simultaneously [17]. A possible avenue is to use more than one control knob (usually the flux) per tunable coupler, even though this is experimentally challenging since it is already hard to introduce a single knob per tunable coupler (on top of all the control lines for the qubits).

Next to this, it is important to raise the accuracy and yield in fabrication [18], to ensure that the fabricated qubits and couplers achieve their target parameters with good-enough precision. Otherwise, qubits and couplers might not work as well as intended.

- **Real-time decoding and feedback.** From a theoretical point of view, quantum error correction has been mostly studied by simulating a code and later passing the syndrome data to a decoder to identify the most likely correction. In this way the decoding algorithm can run for as long as needed. However, eventually the decoder needs to run in real time, collecting syndrome information every QEC cycle and providing a correction (or a Pauli-frame update) before the next logical (non-Pauli) gate is applied. Note that simply the syndrome information amounts to 100 Mb/s for a distance-10 surface code for a QEC-cycle time of 500 ns. Furthermore, while decoding can be done per e.g. surface-code patch, interpatch communication is required to perform logical two-qubit gates (e.g. the CNOT via lattice surgery [19, 20]). Decoding cannot even use all the available time since signals need to travel up and down the fridge, although this takes at most a few tens of nanoseconds at the speed of light (30 cm/ns). As a consequence, low latency, an efficient decoding algorithm and a fast implementation of such an algorithm are all needed.

  While many efficient decoding algorithms have been developed [21–23], convincing implementations are still lacking. Preliminary studies were conducted in Refs. [24, 25] for the minimum-weight perfect-matching decoder (MWPM). Recently, concrete micro-architectures have been proposed for the repetition code [26] and the surface code [27] in the case of a lookup-table decoder, and for the Union-Find decoder [28], mostly aimed at the surface code. Furthermore, a lookup-table decoder and MWPM have been implemented in an FPGA in the DiCarlo lab [29], with the potential to run in real time for at most the distance-3 surface code. Further demonstrations on actual hardware are needed, with the potential to scale up to higher-distance codes.

- **Relaxation errors in large devices and scalability of the chip.** In few-qubit ($<$ 10) devices, transmons regularly reach relaxation times (T_1) on the order of 50-100 μs [30], and relaxation times beyond 100 μs are not so uncommon. However,

in a 50+ qubit device like Google's Sycamore [17], on average $T_1 = 15\ \mu$s, which is a major issue (although this decrease in T_1 does not seem to occur in IBM's processors [31], which use fixed-frequency transmons). Clearly, it is fundamental to bring the T_1 values (and also T_2 values) found in small chips to the bigger ones, otherwise the overall performance may get worse.

Apart for T_1, which captures an average behavior, cosmic rays (see Section 3.2.3) have been shown to be a devastating error source [32] causing chip-wide correlated errors [33]. These cannot be dealt with error correction (unless the chip is really large and the correlation length is smaller than the system size, or the system is modular with multiple chips). Since these events occur every 10 seconds on average [33], this severely restricts the extent of a quantum computation. Shielding methods, a more modular design and, possibly, moving fridges to underground facilities [34] might become necessary.

Furthermore, with the growing number of qubits in quantum processors, space and heat load in a dilution refrigerator will become a limiting factor. This can be partially compensated by better cable technology, multi-layer chips or a bigger dilution refrigerator. Eventually, it might be necessary to connect multiple dilution refrigerators with some kind of quantum links [35].

These are not the only problems that need to be addressed. In particular, optimizing quantum algorithms to require less resources, as well as developing new qubits with better isolation and insensitivity to errors [36], might be key steps. Furthermore, training new students at the graduate and undergraduate levels is of paramount importance since the quantum-computing field is evolving at a fast pace, with a high demand for qualified people. Finally, lowering physical error rates as much as possible with control or hardware techniques will of course continue to be beneficial to reduce the overhead for quantum error correction. This thesis has contributed to this effort, hoping that quantum computers will fulfill their promise of solving useful computational tasks.

REFERENCES

[1] J. Kelly, P. O'Malley, M. Neeley, H. Neven, and J. M. Martinis, *Physical qubit calibration on a directed acyclic graph,* (2018), arXiv:1803.03226 [quant-ph] .

[2] P. V. Klimov, J. Kelly, J. M. Martinis, and H. Neven, *The snake optimizer for learning quantum processor control parameters,* (2020), arXiv:2006.04594 [quant-ph] .

[3] N. Lacroix, C. Hellings, C. K. Andersen, A. Di Paolo, A. Remm, S. Lazar, S. Krinner, G. J. Norris, M. Gabureac, J. Heinsoo, and et al., *Improving the performance of deep quantum optimization algorithms with continuous gate sets,* PRX Quantum **1** (2020), 10.1103/prxquantum.1.020304.

[4] A. G. Fowler, *Coping with qubit leakage in topological codes,* Phys. Rev. A **88**, 042308 (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. Hoi, C. Neill, P. J. J.

O'Malley, C. Quintana, P. Roushan, A. Vainsencher, J. Wenner, A. N. Cleland, and J. M. Martinis, *State preservation by repetitive error detection in a superconducting quantum circuit,* Nature **519**, 66 (2015).

[6] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code,* Quantum Info. Comput. **15**, 997 (2015).

[7] T. M. Stace and S. D. Barrett, *Error correction and degeneracy in surface codes suffering loss,* Phys. Rev. A **81**, 022317 (2010).

[8] S. Nagayama, A. G. Fowler, D. Horsman, S. J. Devitt, and R. V. Meter, *Surface code error correction on a defective lattice,* New Journal of Physics **19**, 023050 (2017).

[9] J. M. Auger, H. Anwar, M. Gimeno-Segovia, T. M. Stace, and D. E. Browne, *Fault-tolerance thresholds for the surface code with fabrication errors,* Phys. Rev. A **96**, 042316 (2017).

[10] R. Versluis, S. Poletto, N. Khammassi, B. Tarasinski, N. Haider, D. J. Michalak, A. Bruno, K. Bertels, and L. DiCarlo, *Scalable quantum circuit and control for a superconducting surface code,* Phys. Rev. Appl. **8**, 034021 (2017).

[11] Y. Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, B. Campbell, Z. Chen, B. Chiaro, and et al., *Qubit architecture with high coherence and fast tunable coupling,* Physical Review Letters **113** (2014), 10.1103/physrevlett.113.220502.

[12] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Tunable Coupling Scheme for Implementing High-Fidelity Two-Qubit Gates,* Physical Review Applied **10**, 054062 (2018).

[13] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler,* Phys. Rev. X **11**, 021058 (2021).

[14] B. K. Mitchell, R. K. Naik, A. Morvan, A. Hashim, J. M. Kreikebaum, B. Marinelli, W. Lavrijsen, K. Nowrouzi, D. I. Santiago, and I. Siddiqi, *Hardware-efficient microwave-activated tunable coupling between superconducting qubits,* (2021), arXiv:2105.05384 [quant-ph] .

[15] E. A. Sete, A. Q. Chen, R. Manenti, S. Kulshreshtha, and S. Poletto, *Floating tunable coupler for scalable quantum computing architectures,* Physical Review Applied **15** (2021), 10.1103/physrevapplied.15.064063.

[16] J. Stehlik, D. Zajac, D. Underwood, T. Phung, J. Blair, S. Carnevale, D. Klaus, G. Keefe, A. Carniol, M. Kumph, and et al., *Tunable coupling architecture for fixed-frequency transmon superconducting qubits,* Physical Review Letters **127** (2021), 10.1103/physrevlett.127.080505.

[17] Z. Chen, K. J. Satzinger, J. Atalaya, A. N. Korotkov, A. Dunsworth, D. Sank, C. Quintana, M. McEwen, R. Barends, P. V. Klimov, S. Hong, C. Jones, A. Petukhov, D. Kafri, S. Demura, B. Burkett, C. Gidney, A. G. Fowler, A. Paler, H. Putterman, I. Aleiner, F. Arute, K. Arya, R. Babbush, J. C. Bardin, A. Bengtsson, A. Bourassa, M. Broughton, B. B. Buckley, D. A. Buell, N. Bushnell, B. Chiaro, R. Collins, W. Courtney, A. R. Derk, D. Eppens, C. Erickson, E. Farhi, B. Foxen, M. Giustina, A. Greene, J. A. Gross, M. P. Harrigan, S. D. Harrington, J. Hilton, A. Ho, T. Huang, W. J. Huggins, L. B. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, K. Kechedzhi, S. Kim, A. Kitaev, F. Kostritsa, D. Landhuis, P. Laptev, E. Lucero, O. Martin, J. R. McClean, T. McCourt, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Newman, M. Y. Niu, T. E. O'Brien, A. Opremcak, E. Ostby, B. Pató, N. Redd, P. Roushan, N. C. Rubin, V. Shvarts, D. Strain, M. Szalay, M. D. Trevithick, B. Villalonga, T. White, Z. J. Yao, P. Yeh, J. Yoo, A. Zalcman, H. Neven, S. Boixo, V. Smelyanskiy, Y. Chen, A. Megrant, J. Kelly, and G. Q. AI, *Exponential suppression of bit or phase errors with cyclic error correction,* Nature **595**, 383 (2021).

[18] J. B. Hertzberg, E. J. Zhang, S. Rosenblatt, E. Magesan, J. A. Smolin, J.-B. Yau, V. P. Adiga, M. Sandberg, M. Brink, J. M. Chow, and J. S. Orcutt, *Laser-annealing Josephson junctions for yielding scaled-up superconducting quantum processors,* npj Quantum Information **7**, 129 (2021).

[19] C. Horsman, A. G. Fowler, S. Devitt, and R. V. Meter, *Surface code quantum computing by lattice surgery,* New Journal of Physics **14**, 123011 (2012).

[20] A. Erhard, H. Poulsen Nautrup, M. Meth, L. Postler, R. Stricker, M. Stadler, V. Negnevitsky, M. Ringbauer, P. Schindler, H. J. Briegel, and et al., *Entangling logical qubits with lattice surgery,* Nature **589**, 220–224 (2021).

[21] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, *Topological quantum memory,* Journal of Mathematical Physics **43** (2002), 10.1063/1.1499754.

[22] N. Delfosse and N. H. Nickerson, *Almost-linear time decoding algorithm for topological codes,* (2017), arXiv:1709.06218 [quant-ph] .

[23] X. Ni, *Neural network decoders for large-distance 2D toric codes,* Quantum **4**, 310 (2020).

[24] A. G. Fowler, A. C. Whiteside, and L. C. L. Hollenberg, *Towards practical classical processing for the surface code,* Physical Review Letters **108** (2012), 10.1103/physrevlett.108.180501.

[25] A. G. Fowler, A. C. Whiteside, and L. C. L. Hollenberg, *Towards practical classical processing for the surface code: Timing analysis,* Physical Review A **86** (2012), 10.1103/physreva.86.042313.

[26] D. Ristè, L. C. G. Govia, B. Donovan, S. D. Fallek, W. D. Kalfus, M. Brink, N. T. Bronn, and T. A. Ohki, *Real-time processing of stabilizer measurements in a bit-flip code,* npj Quantum Information **6**, 71 (2020).

[27] P. Das, A. Locharla, and C. Jones, *Lilliput: A lightweight low-latency lookup-table based decoder for near-term quantum error correction,* (2021), arXiv:2108.06569 [quant-ph] .

[28] P. Das, C. A. Pattison, S. Manne, D. Carmean, K. Svore, M. Qureshi, and N. Delfosse, *A scalable decoder micro-architecture for fault-tolerant quantum computing,* (2020), arXiv:2001.06598 [quant-ph] .

[29] M. Moreira, B. M. Tarasinski, J. Gloudemans, V. P. Ostroukh, W. J. Vlothuizen, and L. DiCarlo, *Real-time quantum error correction for a surface-code logical qubit,* APS March Meeting (2020).

[30] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, *Superconducting qubits: Current state of play,* Annual Review of Condensed Matter Physics **11**, 369 (2020).

[31] P. Jurcevic, A. Javadi-Abhari, L. S. Bishop, I. Lauer, D. F. Bogorin, M. Brink, L. Capelluto, O. Günlük, T. Itoko, N. Kanazawa, A. Kandala, G. A. Keefe, K. Krsulich, W. Landers, E. P. Lewandowski, D. T. McClure, G. Nannicini, A. Narasgond, H. M. Nayfeh, E. Pritchett, M. B. Rothwell, S. Srinivasan, N. Sundaresan, C. Wang, K. X. Wei, C. J. Wood, J.-B. Yau, E. J. Zhang, O. E. Dial, J. M. Chow, and J. M. Gambetta, *Demonstration of quantum volume 64 on a superconducting quantum computing system,* (2020), arXiv:2008.08571 [quant-ph] .

[32] A. P. Vepsäläinen, A. H. Karamlou, J. L. Orrell, A. S. Dogra, B. Loer, F. Vasconcelos, D. K. Kim, A. J. Melville, B. M. Niedzielski, J. L. Yoder, and et al., *Impact of ionizing radiation on superconducting qubit coherence,* Nature **584**, 551–556 (2020).

[33] M. McEwen, L. Faoro, K. Arya, A. Dunsworth, T. Huang, S. Kim, B. Burkett, A. Fowler, F. Arute, J. C. Bardin, A. Bengtsson, A. Bilmes, B. B. Buckley, N. Bushnell, Z. Chen, R. Collins, S. Demura, A. R. Derk, C. Erickson, M. Giustina, S. D. Harrington, S. Hong, E. Jeffrey, J. Kelly, P. V. Klimov, F. Kostritsa, P. Laptev, A. Locharla, X. Mi, K. C. Miao, S. Montazeri, J. Mutus, O. Naaman, M. Neeley, C. Neill, A. Opremcak, C. Quintana, N. Redd, P. Roushan, D. Sank, K. J. Satzinger, V. Shvarts, T. White, Z. J. Yao, P. Yeh, J. Yoo, Y. Chen, V. Smelyanskiy, J. M. Martinis, H. Neven, A. Megrant, L. Ioffe, and R. Barends, *Resolving catastrophic error bursts from cosmic rays in large arrays of superconducting qubits,* (2021), arXiv:2104.05219 [quant-ph] .

[34] L. Cardani, F. Valenti, N. Casali, G. Catelani, T. Charpentier, M. Clemenza, I. Colantoni, A. Cruciani, G. D'Imperio, L. Gironi, and et al., *Reducing the impact of radioactivity on quantum circuits in a deep-underground facility,* Nature Communications **12** (2021), 10.1038/s41467-021-23032-z.

[35] P. Magnard, S. Storz, P. Kurpiers, J. Schär, F. Marxer, J. Lütolf, T. Walter, J.-C. Besse, M. Gabureac, K. Reuer, and et al., *Microwave quantum link between superconducting circuits housed in spatially separated cryogenic systems,* Physical Review Letters **125** (2020), 10.1103/physrevlett.125.260502.

[36] A. Gyenis, A. D. Paolo, J. Koch, A. Blais, A. A. Houck, and D. I. Schuster, *Moving beyond the transmon: Noise-protected superconducting quantum circuits,* (2021), arXiv:2106.10296 [quant-ph] .

ACKNOWLEDGEMENTS

A PhD is a collection of many micro-skills and interactions. This thesis would have not been possible without the support of multiple people, both scientifically and at a personal level.

My thanks go first to **Barbara**. I remember that the day after I sent an inquiry for a position with you, I told about it to Robert and I was surprised to learn that you had already spoken to him early in the morning. Since then, I've always been impressed by your availability and by the fact that you seem to have the energy of three people. I also appreciate your curiosity and readiness to learn and discover new things. From you I have learned how to do research and I will take these lessons with me in the years to come. Thanks for how you take care of the people in your group, trying to make sure that they feel good and supporting them in their steps.

Leo, thanks for the collaboration between us and the rest of your group. I wanted my work to have a practical impact and the two-qubit-gate and surface-code projects have given me the possibility to do so. I appreciate your high standards and scientific integrity.

Thanks to the committee members for reading my thesis, providing comments on it and being present at my defence.

Thanks to my paranymph **Boris** for all our multiple-hour-long conversations about leakage, the state of the field and any other random thing. Writing the leakage-detection paper together was a lot of fun and I always appreciate your numerical skills. We will forever remember being stranded in Denver once the March Meeting was canceled and wanting to leave the US asap.

Thanks to my paranymph **Alessandro** for TAing together multiple times about circuit quantization and for all the time we have spent together (and for the arrosticini!). The trip to New York was a blast. Thanks for introducing me to the gym in the early days. Good luck in Germany and I hope I'll see you soon.

Daniel and **Christoph**, we met when you were still in Aachen. I felt welcomed and you made me like Barbara's group. Thanks to the other members and visitors of the group for the various conversations we had: **Joel**, **Ben**, **Chris**, **Xiaotong**, **Yang**, **Jonathan**, **Jarn**, **Manuel**, **Matteo**, **Maarten**, **Olivia** and **Marios**. A special mention to **Joey**: it was nice to work together on the small project for your Master.

Adriaan, sitting next to each other in the lab was a great experience. Thanks for teaching all the things about the experiment to a theorist. Finishing a paper during the March Meeting was stressful but it was nice to do it together. From you I learned many good research and writing methods, as well as how to make good figures and presentations.

Victor and **Hany**, thanks for the collaboration on the follow-up paper. Also thanks to the other people in the DiCarlo lab with whom I had the pleasure to interact: **Niels** (looking forward to our next adventure together!), **Ramiro**, **Alessandro**, **Nadia**, **Nandini**, **Filip**,

Slava, **Thijs**, **Jorge**, **Miguel** and **René**. **Tom**, **Brian**, **Xavi**, the multi-qubit meetings with you and some of the people above were great and I learned a lot.

Thanks to my flatmate **Filip** for sharing many adventures together and for all the open conversations. Thanks also for introducing me to many friends in Stephanie's group: **Kaushik**, **Tim**, **Kenneth**, **Valentina**, **Marc**, **Glaucia** and more. **Jonas**, it was cool to write a paper together and to learn from you.

Jake, **Antariksha**, **Mohsen**, **Xiao**, **Sjaak**, **Maximilian**, **Lingling**, **Mohamed** and many other people in QuTech, it was nice to meet you and enjoy some time together. Also thanks to **Jenny** and **Marja** for assisting our group in the organizational part.

Thanks to my flatmate **Rishabh** and my friends **Indushree**, **Eveline**, **Emilia** and **Sophie**, as well as to all the other friends that are dear and close to me here, in Munich and Italy. Thanks to my girlfriend's flatmate **Julia** for accepting me in their home for months. A special thanks to my crew mates during the lockdown.

Last but not the least, thanks to my family for the time together and for all the support during all my life. Un grazie speciale alla **Nonna Pasqua** perché la renderà felice. Thanks to my girlfriend **Jane** for being an amazing person and for being next to me, as well as for the patience to wait that I was done with writing this thesis.

CURRICULUM VITÆ

Francesco BATTISTEL

05-01-1993 Born in Pordenone, Italy.

EDUCATION

2007–2012 High School
Liceo Scientifico Sperimentale “G. A. Pujati”, Sacile, Italy

2012-2015 Bachelor in Physics
Università degli Studi di Trieste, Trieste, Italy
Thesis: Black Hole Entropy and Thermodynamics
Supervisor: Prof.dr. S. Ansoldi

2015-2017 Elite Master Program in Theoretical and Mathematical Physics
Ludwig-Maximilians Universität & Technische Universität München, Munich, Germany
Thesis: General Quantum Error Correction for MERA Codes
Supervisor: Prof.dr. R. König

2017-2021 PhD in Applied Physics
Technische Universiteit Delft, Delft, Netherlands
Thesis: Mitigating Leakage and Noise in Superconducting Quantum Computing
Promotors: Prof.dr. B.M. Terhal
Prof.dr. L. DiCarlo

AWARDS

2012 Olympics of Chemistry
5th in the Italian National Final category B

2012-2015 Collegio per le Scienze “Luciano Fonda”
Scholarship awarded through competitive exams

List of Publications

7. A. Ciani, **F. Battistel**, D. DiVincenzo, B.M. Terhal, *Lecture Notes on Circuit-QED*, in preparation (2022).

6. J.F. Marques, B.M. Varbanov, M.S. Moreira, H. Ali, N. Muthusubramanian, C. Zachariadis, **F. Battistel**, M. Beekman, N. Haider, W. Vlothuizen, A. Bruno, B.M. Terhal, L. DiCarlo, *Logical-qubit operations in an error-detecting surface code*, Nature Physics (2021).

5. **F. Battistel**, B.M. Varbanov, B.M. Terhal, *Hardware-efficient leakage-reduction scheme for quantum error correction with superconducting transmon qubits*, PRX Quantum **2**, 030314 (2021).

4. V. Negîrneac, H. Ali, N. Muthusubramanian, **F. Battistel**, R. Sagastizabal, M.S. Moreira, J.F. Marques, W. J. Vlothuizen, M. Beekman, N. Haider, A. Bruno, L. DiCarlo, *High-fidelity controlled-Z gate with maximal intermediate leakage operating at the speed limit in a superconducting quantum processor*, Phys. Rev. Lett. **126**, 220502 (2021).

3. B.M. Varbanov, **F. Battistel**, B.M. Tarasinski, V.P. Ostroukh, T.E. O'Brien, L. DiCarlo, B.M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Inf **6**, 102 (2020).

2. J. Helsen, **F. Battistel**, B.M. Terhal, *Spectral Quantum Tomography*, npj Quantum Inf **5**, 74 (2019).

1. M.A. Rol, **F. Battistel**, F.K. Malinowski, C.C. Bultink, B.M. Tarasinski, R. Vollmer, N. Haider, N. Muthusubramanian, A. Bruno, B.M. Terhal, and L. DiCarlo, *Fast, High-Fidelity Conditional-Phase Gate Exploiting Leakage Interference in Weakly Anharmonic Superconducting Qubits*, Phys. Rev. Lett. **123**, 120502 (2019).