Optimizing quantum error correction for superconducting qubit processors

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

OPTIMIZING QUANTUM ERROR CORRECTION FOR SUPERCONDUCTING QUBIT PROCESSORS

Optimizing quantum error correction for superconducting qubit processors

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Monday 29, January 2024 at 15:00 o'clock

by

Boris Mihailov Varbanov

Master of Science in Applied Physics,
Delft University of Technology, the Netherlands,
born in Ruse, Bulgaria.

CONTENTS

SUMMARY

The theory of quantum mechanics describes many phenomena that may initially seem to be counter-intuitive and, in some cases, impossible, given the understanding of classical mechanics that most of us are more intimately familiar with. Following its initial introduction, there was a great deal of debate among scientists regarding the predictions made by this theory. The strange nature of quantum mechanics has led to many memorable quotes and the use of "spooky" to describe some of these predictions. Since its initial introduction, quantum mechanics has been rigorously tested and has proven to be quite a successful theory. Quantum mechanics has found many different applications and has led to the existence of devices and technologies we use daily. Another potential application of quantum mechanics is quantum computation, which Richard Feynman first put forward as an idea in 1982. Quantum computers have the potential to solve specific problems that can be infeasible for even the most powerful (classical) supercomputers and have potential applications in many different areas, such as quantum chemistry, cryptography, and optimization.

However, performing a quantum computation is challenging and requires overcoming the inherent fragility of quantum systems. Storing information in a quantum system requires it to be well isolated from the environment to avoid any unwanted interactions that can corrupt the stored data. Unfortunately, at the same time, we need the ability to control this system, make it interact with other such systems, and ultimately measure it for us to perform an actual computation. This is a universal issue and all of the systems we have so far developed to be used as quantum bits (qubits) have been plagued by noise. Each operation applied to the qubit or even the act of leaving the qubit idling for some time generally leads to an error with a non-negligible probability. The impact of this noise has so far prevented quantum computers from performing any practical computation. While substantial efforts have been made to reduce these physical error rates over the past several years, we are still far from the universal fault-tolerant quantum computers we ultimately strive for.

Fortunately, quantum error correction can help us reach the low error rates necessary for quantum computers to realize their potential applications in the future. This can be achieved by storing the quantum information in a logical qubit instead of a noisy physical one. When using a stabilizer code, which will be the focus of this dissertation, this logical information is distributed over many (noisy) physical qubits, referred to as data qubits. Another set of qubits, the so-called ancilla qubits, is used to perform indirect parity measurements, which do not destroy the stored information but give some information about whether an error has occurred. We then try to interpret this information to identify what errors have happened and correct them, which is done by a classical algorithm referred to as the *decoder*. Increasing the number of physical qubits used to encode the logical qubits allows more physical errors to be detected and corrected. The number of correctable errors is captured by the distance of the code, defined as the

minimum number of physical single-qubit errors that constitute a logical error. One of the critical properties of error correction is the ability to reduce the logical error rate by increasing the code distance, which requires the physical error rates to be below some threshold value.

The valiant experimental effort over the years has led to several recent experiments that implement various error-correcting codes and demonstrate the reduction of the error rates promised by error correction. In particular, these experiments (and the experiments leading up to them) identified several noise sources that had not been explored in sufficient detail and could significantly impact the logical performance of the code. In this dissertation, we explore the impact of the noise encountered in transmon-qubit devices on the performance of error-correcting codes, namely the surface code. Transmon qubits are, in practice, multi-level systems, and only the lowest two energy levels are used for computation. Unfortunately, they are also weakly anharmonic, leading to the applied operations having some probability of exciting the qubit outside of this computational subspace, referred to as a leakage error. We explore the impact of leakage in both simulations and experiments and develop schemes to mitigate it. We also consider other approaches to improve the logical performance or to reduce unwanted interactions.

In Chapter 2, we develop a realistic model of leakage induced by the two-qubit gates between flux-tunable transmon qubits. We show that leaked qubits effectively spread errors on their neighboring qubits, which are then detected by the parity measurements. We show that a Hidden Markov model can detect the increased error rate due to leakage. This enables us to post-select out runs during which any qubit has leaked to restore the code performance.

Unfortunately, post-selection is ultimately not scalable. Instead, it is desirable to have operations that return leaked qubits to the computational subspace. These operations are called leakage-reduction units and convert leakage into a regular error. In Chapter 3, we propose a leakage-reduction scheme, which does not require any overhead in the time needed to perform the parity measurements or an overhead in the quantum hardware. For data qubits, we propose an operation that transfers the leakage to a dedicated readout resonator, where it can quickly decay. This operation is designed to not disturb the computational states, allowing it to be applied unconditionally. For the ancilla qubit, we use the fact that measurements can determine if a qubit is in the leaked state. We then apply a conditional operation to return the qubit to the computational subspace whenever it is measured to be leaked. Using detailed density-matrix simulation, we show that this scheme can be easily implemented to remove qubit leakage from the system, mitigating its impact on the logical performance of the code.

In Chapter 4, we realize the data-qubit leakage reduction unit in an experiment and show it can also be used to remove ancilla-qubit leakage, removing the need for fast conditional operations and readout that distinguishes the leaked states. We show that these operations can remove most of the leaked population in about a hundred nanoseconds while having a negligible impact on the computational subspace. We also demonstrate that these operations decrease the number of observed errors by a two-qubit parity check, showing that the effect of leakage can be mitigated.

Chapter 5 considers an architecture employing two types of superconducting qubits, the transmon qubit and the fluxonium qubit. These qubits have very different frequen-

cies, making it unclear whether these qubits can even interact with each other in the first place. We show that the interactions with the higher-excited states can be utilized to perform operations between them, and we propose two types of gates. In practice, qubit frequencies are targeted with only a certain precision in fabrication. In certain cases, this can lead to unwanted interaction between qubits that increase the physical error rates, referred to as *frequency collisions*. We show that the large detuning between these qubits reduces the frequency of frequency collision, thereby increasing the expected fabrication yield.

In Chapter 6, we realize a distance-two surface code experiment and perform repeated parity measurements to detect and post-select errors, given that it's impossible to correct them when using such a small code. We implement a suite of logical operations for this code, including initialization, measurement, and several single-qubit gates. In the context of error detection, a logical operation is said to be fault-tolerant if the errors produced by each operation are detectable. We show that fault-tolerant variants of operations perform better than non-fault-tolerant ones. We also characterize the impact of various noise sources on the code performance.

In Chapter 7, we look at another small-distance code, in this case, the distance-seven repetition code. We show that increasing the distance weakly suppresses the logical error rate of the code. We investigate the limiting factors behind the observed logical performance by analyzing the correlation between the observed parity measurements and performing simulations using noise models parameterized by the measured physical error rates.

Chapter 8 considers a decoder that can perform the error inference more accurately. In particular, we implement a neural network decoder and investigate how it performs on experimental data from surface code experiments. We show that the accuracy of this decoder approaches what can be achieved by an optimal and computationally inefficient tensor network decoder. Transmon measurement produces analog outcomes. These are then typically converted to binary ones, leading to some information loss. We show how a neural network can also use this analog information to improve the achieved logical performance further.

We have investigated the impact of non-conventional errors in simulation and in several experiments, demonstrating the importance of characterizing and mitigating these errors. We expect the methods introduced in this dissertation to lead to lower logical error rates. In the short term, this can aid in demonstrations of the usefulness of error correction. In the long term, addressing such errors is important to ensure the ability to suppress logical error rates to sufficiently low levels. We finish this dissertation with a brief conclusion of each chapter. We also outline several potential challenges that can impact future error-correction experiments, namely how to reduce the larger qubit overhead needed for fault-tolerant computation and several error sources that might become a limiting factor for future error-correction experiments.

SAMENVATTING

De theorie van de kwantummechanica beschrijft veel verschijnselen die in eerste instantie contra-intuïtief en in sommige gevallen zelfs onmogelijk lijken, uitgaande van het begrip van de klassieke mechanica waar de meesten van ons beter bekend mee zijn. Na haar introductie was er veel discussie onder wetenschappers over de voorspellingen die deze theorie deed. De vreemde aard van de kwantummechanica heeft geleid tot gedenkwaardige citaten en zelfs het gebruik van 'spookachtig' om sommige van deze voorspellingen te beschrijven. Sinds haar introductie is de kwantummechanica uitvoerig getest en een behoorlijk succesvolle theorie gebleken. Kwantummechanica heeft veel verschillende toepassingen en heeft geleid tot het bestaan van apparaten en technologieën die we hedendaags gebruiken. Een mogelijke toepassing van de kwantummechanica is de zogenaamde kwantumberekening, een idee dat voor het eerst geopperd werd door Richard Feynman in 1982. Kwantumcomputers kunnen in potentie specifieke problemen oplossen die zelfs voor de krachtigste (klassieke) supercomputers onhaalbaar zijn en hebben potentiële toepassingen op veel verschillende gebieden, zoals kwantumchemie, cryptografie en optimalisatie.

Het uitvoeren van een kwantumberekening is echter een uitdaging en vereist het overkomen van de inherente kwetsbaarheid van kwantumsystemen. Het opslaan van informatie in een kwantumsysteem vereist het systeem goed geïsoleerd is van de omgeving zodat ongewenste interacties die de opgeslagen gegevens kunnen beschadigen voorkomen kunnen worden. Helaas moeten we tegelijkertijd het systeem kunnen controleren, het kunnen laten communiceren met andere soortgelijke systemen, en uiteindelijk het kunnen meten om daadwerkelijk een berekening uit te voeren. Dit is een universeel probleem en alle systemen die we tot nu toe hebben ontwikkeld om als kwantumbits (qubits) te gebruiken, worden geplaagd door de aanwezigheid van ruis. Elke bewerking die wordt toegepast op de qubit en zelfs het inactief laten van de qubit gedurende enige tijd leidt over het algemeen tot een fout met een niet te verwaarlozen waarschijnlijkheid. De gevolgen van deze ruis hebben er tot dusver voor gezorgd dat kwantumcomputers geen enkele praktische berekening uit konden voeren. Hoewel er de afgelopen jaren aanzienlijke inspanningen zijn geleverd om deze fysieke fouten te verminderen, zijn we nog steeds ver verwijderd van de universele fouttolerante kwantumcomputers waar we naar streven.

Gelukkig kan kwantumfoutcorrectie ons helpen de lage foutwaarschijnlijkheden te behalen die nodig zijn voor het realiseren van de potentiële toepassingen van kwantumcomputers. Dit kan worden bereikt door de kwantuminformatie op te slaan in een zogeheten logische qubit in plaats van in een ruisachtige fysieke qubit. Bij het gebruik van een zogeheten stabilizer code, waar dit proefschrift zich op zal richten, wordt deze logische informatie verdeeld over vele (ruisachtige) fysieke qubits, de zogenaamde dataqubits. Een andere set qubits, de zogenaamde ancilla qubits, wordt gebruikt om indirecte pariteitsmetingen uit te voeren, die de opgeslagen informatie niet vernietigen maar

wel informatie opleveren over of er een fout is opgetreden. Vervolgens proberen we deze informatie te interpreteren om te identificeren welke fouten zijn opgetreden en corrigeren we ze. Dit wordt gedaan door een klassiek algoritme dat de textitdecoder wordt genoemd. Door het vergroten van het aantal gebruikte fysieke qubits dat gebruikt wordt om de logische qubits te coderen, kunnen we meer fysieke fouten detecteren en corrigeren. Het aantal corrigeerbare fouten wordt vastgelegd door de afstand van de code, gedefinieerd als het minimale aantal fysieke single-qubit-fouten dat een logische fout vormt. Eén van de essentiele eigenschappen van foutcorrectie is het vermogen om het aantal logische fouten te verminderen door het vergroten van de codeafstand, waarvoor de fysieke foutwaarschijnlijkheden beneden een bepaalde drempelwaarde moeten liggen.

De moedige experimentele inspanningen door de jaren heen hebben geleid tot verschillende recente experimenten die verschillende foutcorrectiecodes implementeren en reductie van de foutenwaarschijnlijkheden aantonen die door foutcorrectie worden beloofd. In het bijzonder identificeerden deze experimenten (en de experimenten die daaraan voorafgingen) verschillende ruisbronnen die nog niet voldoende waren onderzocht en mogelijk een aanzienlijke invloed kunnen hebben op de logische prestaties van de code. In dit proefschrift onderzoeken we de impact van de ruis die wordt aangetroffen in zogeheten transmon-qubit-apparaten op de prestaties van foutcorrectiecodes, namelijk de zogenaamde surface code. Transmon qubits zijn in de praktijk systemen met meerdere niveaus, en alleen de laagste twee energieniveaus worden gebruikt voor berekeningen. Helaas zijn ze weinig anharmonisch, wat met enige waarschijnlijkheid leidt tot het exciteren van de qubit naar buiten zijn rekendeelruimte wanneer een bewerking wordt toegepast, ook wel een leakage fout genoemd. We onderzoeken de impact van leakage in zowel simulaties als experimenten en ontwikkelen methoden om deze te matigen. Wij bekijken ook andere methoden om de logische prestaties te verbeteren of ongewenste interacties te verminderen.

In hoofdstuk 2 ontwikkelen we een realistisch model van leakage veroorzaakt door de twee-qubit gates tussen flux-afstembare transmon qubits. We laten zien dat gelekte qubits in essentie fouten verspreiden naar aangrenzende qubits, die vervolgens worden gedetecteerd door de pariteitsmetingen. We laten zien dat een Hidden Markov-model het verhoogde foutenpercentage als gevolg van leakage kan detecteren. Hierdoor kunnen we runs achteraf selecteren waarin een qubit is gelekt om de codeprestaties te herstellen.

Helaas is deze na-selectie uiteindelijk niet schaalbaar. In plaats daarvan is het wenselijk om Bewerkingen uit te voeren die gelekte qubits terugbrengen naar de rekendeelruimte. Deze bewerkingen worden leakage-reduction units genoemd en zetten leakage om in een reguliere fout. In hoofdstuk 3 stellen we een leakage-reduction methode voor, waarvoor geen overhead nodig is in duur van de pariteitsmetingen, noch in de kwantumhardware. Voor dataqubits stellen we een bewerking voor die de leakage overdraagt naar een aangewezen uitleesresonator, waar het snel kan vervallen. Deze operatie is ontworpen om de computationele toestanden niet te verstoren, waardoor het onvoorwaardelijk kan worden toegepast. Voor de ancilla qubit maken we gebruik van het feit dat metingen kunnen bepalen of een qubit in de leaked toestand zit. Vervolgens passen we een voorwaardelijke bewerking toe om de qubit terug te brengen naar de rekendeelruimte

wanneer gemeten is dat deze leaked is. Met behulp van gedetailleerde dichtheidsmatrixsimulatie laten we zien dat dit schema eenvoudig kan worden geïmplementeerd om qubit-leakage uit het systeem te verwijderen, waardoor de impact ervan op de logische prestaties van de code wordt beperkt.

In hoofdstuk 4 realiseren we de data-qubit leakage-reduction unit in een experiment en laten we zien dat het ook kan worden gebruikt om ancilla-qubit leakage te verwijderen, waardoor de noodzaak voor snelle voorwaardelijke bewerkingen en een uitlezing die de leaked toestanden kan onderscheiden weggenomen wordt. Wij laten zien dat deze operaties het grootste deel van de leaked populatie in ongeveer honderd nanoseconden kunnen verwijderen, terwijl ze een verwaarloosbare impact hebben op de rekendeelruimte. We laten ook zien dat deze bewerkingen het aantal waargenomen fouten verminderen door een pariteitscontrole van twee qubits, wat aantoont dat het effect van leakage kan worden beperkt.

In hoofdstuk 5 beschouwen we apparatuur die gebruik maakt van twee soorten supergeleidende qubits, namelijk de transmon-qubit en de fluxonium-qubit. Deze qubits hebben zeer verschillende frequenties, waardoor het onduidelijk is of deze qubits überhaupt met elkaar kunnen interacteren. We laten zien dat de interacties met de hoger geëxciteerde toestanden kunnen worden benut om operaties tussen hen uit te voeren, en we stellen twee soorten gates voor. In de praktijk worden de frequenties van qubits met slechts een bepaalde precisie gerealiseerd. In bepaalde gevallen kan dit leiden tot ongewenste interactie tussen qubits waardoor de fysieke foutwaarschijnlijkheden toenemen, *frequentie botsingen* genoemd. We laten zien dat de grote ontstemming tussen deze qubits de frequentie van frequentiebotsingen kan verminderen, waardoor de verwachte fabricageopbrengst toeneemt.

In hoofdstuk 6 voeren we een afstand-twee surface code experiment uit en voeren we herhaalde pariteitsmetingen uit om fouten te detecteren en achteraf te selecteren, aangezien het onmogelijk is om ze te corrigeren bij het gebruik van zo'n kleine code. We implementeren een reeks logische bewerkingen voor deze code, waaronder initialisatie, meting en verschillende single-qubit gates. In de context van foutdetectie wordt van een logische bewerking gezegd dat deze fouttolerant is als de fouten die door elke bewerking worden geproduceerd, detecteerbaar zijn. We laten zien dat fouttolerante varianten van operaties beter presteren dan niet-fouttolerante varianten. Ook karakteriseren we de impact van verschillende ruisbronnen op de codeprestaties.

In hoofdstuk 7 kijken we naar een andere code voor kleine afstanden, in dit geval de afstand-zeven repetition code. We laten zien dat het vergroten van de afstand de logische foutwaarschijnlijkheid van de code zwak onderdrukt. We onderzoeken de beperkende factoren achter de waargenomen logische prestaties door de correlatie tussen de waargenomen pariteitsmetingen te analyseren en door het uitvoeren van simulaties met behulp van ruismodellen die zijn geparametriseerd door de gemeten fysieke foutwaarschijnlijkheden.

Hoofdstuk 8 behandelt een decoder die de foutinferentie nauwkeuriger kan uitvoeren. In het bijzonder implementeren we een neurale-netwerk decoder en onderzoeken we hoe deze presteert op experimentele gegevens uit surface code experimenten. We laten zien dat de nauwkeurigheid van deze decoder de nauwkeurigheid die kan worden bereikt door een optimale en computationeel inefficiënte tensor-netwerk decoder be-

nadert. Transmon-meting levert analoge uitkomsten op. Deze worden vervolgens door-gaans omgezet naar binaire uitkomsten, wat tot enig informatieverlies leidt. We laten zien hoe een neuraal netwerk deze analoge informatie ook kan gebruiken om de be-haalde logische prestaties verder te verbeteren.

We hebben de impact van niet-conventionele fouten in simulatie en in verschillende experimenten onderzocht, waarbij we het belang aantonen van het karakteriseren en be-perken van deze fouten. We verwachten dat de methoden die in dit proefschrift worden geïntroduceerd, zullen leiden tot lagere logische foutwaarschijnlijkheden. Op de korte termijn kan dit helpen bij het aantonen van het nut van foutcorrectie. Op de lange ter-mijn is het aanpakken van dergelijke fouten belangrijk om ervoor te zorgen dat de logi-sche foutwaarschijnlijkheden tot voldoende lage niveaus kunnen worden teruggebracht. We sluiten dit proefschrift af met een korte conclusie van elk hoofdstuk. We schetsen ook verschillende potentiële uitdagingen die van invloed kunnen zijn op toekomstige foutcorrectie-experimenten, namelijk hoe de grotere qubit-overhead die nodig is voor fouttolerante berekeningen kan worden verminderd en verschillende foutbronnen die een beperkende factor kunnen worden voor toekomstige foutcorrectie-experimenten.

1

INTRODUCTION

1.1. QUANTUM COMPUTING

1.1.1. A BRIEF HISTORY OF QUANTUM COMPUTERS

The field of quantum computing traces its roots back to the 1980s, with Richard Feynman being commonly cited as the first person to propose using quantum mechanics to perform simulations and computations that would otherwise be hard for classical computers [1]. Shortly after, some of the first quantum algorithms that demonstrated the potential of quantum computers were proposed. David Deutsch and Richard Jozsa proposed a quantum algorithm in 1992 [2] for determining if a given boolean function mapping a bitstring to a single bit is either constant or balanced. While this was just a simple toy problem, this algorithm was the first to show the potential speedup quantum computers can provide over classical algorithms when solving specific problems. However, the field of quantum computation truly gained momentum with Peter Shor's groundbreaking proposal in 1994 [3, 4] for a quantum algorithm that could factor large numbers exponentially faster than any classical computer. Previously, the problem of factoring large numbers was considered very difficult. It was so difficult that it was used as the basis of a key-encryption scheme proposed by Ron Rivest, Adi Shamir, and Leonard Adleman, more commonly known as RSA encryption [5]. This scheme quickly became widely used for secure data transmission online. Therefore, Shor's algorithm posed a significant threat to the protocols and applications utilizing this encryption scheme and, as a result, sparked widespread interest in the potential of quantum computing. Of course, this also led computer scientists to start developing public-key encryption algorithms that do not rely on number factorization or other problems that are (currently) known to be efficiently solvable by quantum computers. Alternatively, quantum encryption schemes that rely on quantum entanglement have also been proposed [6, 7]. Regardless of whether Shor's algorithm will eventually turn out to be the killer application of quantum computing or not, it ultimately drew a lot of attention to the field and significantly accelerated its development. Since these initial algorithms, many other potential applications of quantum computers have been proposed in areas such as quantum sim-

1

ulation [8] and quantum chemistry [9–12].

It was almost immediately realized that the noise experienced by quantum systems (also called qubits) can be an obstacle to performing any practical quantum computation and that a quantum computer needs to be robust to these errors. Inspired by (classical) error correction, Peter Shor introduced the first quantum error correction (QEC) scheme [13] and later showed that quantum computation could, in principle, be made fault-tolerant [14]. Soon after, the theorem of fault tolerance was established [15–18], showing that QEC can be used to achieve an arbitrarily low error rate at the expense of an increase in the number of physical qubits required. This bolstered the confidence researchers had in the feasibility of quantum computers, paving the way for the first experimental realization of a quantum computer. Nowadays, QEC is considered the most promising approach to realizing large-scale fault-tolerant quantum computers, and many current experiments are aiming to utilize error correction to suppress the error rates [19–33].

Realizing error correction in experiments has proven to be rather challenging. In this dissertation, we will focus on experiments using superconducting-qubit processors. On the one hand, the physical error rates observed in even state-of-the-art experiments are relatively high, often above or near the threshold required for error correction to lead to a reduction of the logical error rate [20, 32, 34–36]. On the other hand, fabricating devices with enough qubits to implement even small-distance codes was also a significant hurdle. This was partially due to the limited precision with which the device parameters (such as the qubit frequencies) could be targeted in fabrication, which could significantly impact the operational error rate and ultimately reduce the device yield [37–41]. Therefore, these initial demonstrations required improvements in the design and fabrication of quantum devices, as well as the calibration of the various operations used in these experiments. However, as devices became larger and error rates lower, it was realized that the physical noise in these devices was much more complex than what was typically considered in textbook QEC theory. Namely, performing multiple operations in parallel increases the overall error rate due to various forms of crosstalk [20, 32, 42, 43]. Other types of errors present in physical systems, such as qubit leakage outside of the computational subspace, were also found to have a significant impact on the code performance [44–47]. Finally, more potential noise sources are discovered as physical error rates continuously improve. One such example is the impact of ionizing radiation from cosmic rays [48–50], which leads to bursts of correlated error across many qubits of the device. Since error-correcting codes typically do not consider such correlated errors, this is another important error source to consider in future experiments. Building a large-scale error-corrected quantum computer will require scaling up to even larger devices and further lowering the physical error rates, meaning that these non-conventional errors must be eventually addressed for fault-tolerant computers to become a reality.

1.1.2. QUANTUM BITS

In classical computing, the basic unit of information is the bit, representing a logical unit with only two possible states, most commonly labeled as 0 (or off) and 1 (or on). Similarly, the unit of information used in a quantum computer is (most often) a quantum two-level system called a quantum bit or simply a qubit. The states that a qubit can

take can be described in terms of a pair of basis states, which are commonly labeled $|0\rangle$ and $|1\rangle$. While classical bits can either be in 0 or 1, qubits can be $|0\rangle$, $|1\rangle$ or any state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ that is a linear combination (or *superposition*) of these two basis states, where $\alpha, \beta \in \mathbb{C}$ are complex amplitudes. The probabilities of observing the qubit in $|0\rangle$ or $|1\rangle$ upon measurement are given by $|\alpha|^2$ and $|\beta|^2$, respectively. Of course, since one must observe one or the other upon measurement, these probabilities must be normalized, such that $|\alpha|^2 + |\beta|^2 = 1$. Therefore, the qubit states are unit vectors that live in a two-dimensional Hilbert space. The two basis states are vectors, typically defined as

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad\qquad |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

also known as the *computational* basis states. We can also rewrite any superposition state as $|\psi\rangle = \cos(\theta/2)|0\rangle + e^{i\phi}\sin(\theta/2)|1\rangle$, where θ and ϕ are the polar and azimuthal angles that define the direction of a unit vector in a three-dimensional space. This allows quantum states to be visualized as vectors lying on the surface of the *Bloch sphere*.

The real power of quantum information starts to become apparent only when we consider multiple qubits. For example, suppose we have a system of 2 qubits, which has four basis states $|00\rangle$, $|01\rangle$, $|10\rangle$, and $|11\rangle$. Here, $|00\rangle = |0\rangle \otimes |0\rangle$, where $\otimes$ is the tensor product operator. Similar as before, this system can be in any superposition state involving these basis states $|\psi\rangle = c_{00}|00\rangle + c_{01}|01\rangle + c_{10}|10\rangle + c_{11}|11\rangle$, where $c_{ij} \in \mathbb{C}$ are the complex amplitudes, under the constraint that $\sum_{i,j}|c_{ij}|^2 = 1$. Therefore, a system with n qubits can create superposition states involving 2^n basis states and amplitudes. This enormous computational space, when combined with other quantum phenomena, such as *quantum entanglement* and *quantum interference*, enables quantum computers to solve problems that are otherwise infeasible for classical computers to solve.

Quantum entanglement is perhaps the most defining feature of quantum computers and also plays an essential role in QEC. Let us return to a system of two qubits and consider the state $|\psi\rangle = (|00\rangle + |01\rangle)\sqrt{(2)}$. This state can be expressed as $|\psi\rangle = |i\rangle \otimes |j\rangle$, where $|i\rangle = |0\rangle$ is the state of the first qubit, while $|j\rangle = (|0\rangle + |1\rangle)/\sqrt{2} = |+\rangle$ is the state of the second one. States that can be written in this form are called *product states* or *separable states*. Next, consider instead the state $|\psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$, which cannot be written as a product of two single-qubit states, making it an *entangled state*. Measuring these two qubits will reveal that the two outcomes are perfectly *correlated*. Of course, if we consider the state $|00\rangle$, the resulting measurement outcomes would also be correlated. However, the correlations exhibited by entangled states are a much more powerful and useful resource that can be exploited in quantum computation. It is also worth noting that creating an entangled state using only single-qubit operations is impossible. Instead, the two qubits need to interact with each other to create these non-trivial states.

So far, we have expressed the state of a qubit using the state vector formalism. However, there is a different state representation, called the *density matrix*, which is more convenient for describing quantum states that are not completely known. For example, consider the case where a qubit is in the (*pure*) state $|\psi_i\rangle$ with probability p_i, where i indexes each of these states and probabilities. In other words, the state of the qubit is

1

described by the *ensemble* of states $\{p_i, |\psi_i\rangle\}$, which can be expressed as

$$\rho = \sum_i p_i |\psi_i\rangle \langle \psi_i|,$$

with ρ being the density matrix. Furthermore, ρ must be positive semi-definite, Hermitian, and of unit trace, i.e., $\text{Tr}(\rho) = 1$. Such a mixture of pure states is called a *mixed* state. For comparison, the density matrix for a qubit in a pure state is in the form of $\rho = |\psi\rangle \langle \psi|$.

1.1.3. QUANTUM OPERATIONS

Qubits are the units of information used in a quantum computer. However, to perform quantum computations, we need the ability to perform various operations on these qubits. In particular, quantum operations include initializing the qubit in a given initial state, measuring the qubit, and performing an n-qubit quantum gate. Implementing a many-qubit interaction with high fidelity is very challenging in most platforms. Luckily, having access only to single-qubit and two-qubit gates enables universal quantum computation since gates involving three or more qubits can be decomposed into single-qubit and two-qubit gates. In this section, we will introduce the most common operations used in quantum computing.

An n-qubit quantum gate can be described by a $2^n \times 2^n$ unitary operator U, for which it holds that $UU^\dagger = U^\dagger U = I$, where $U^\dagger$ denotes the Hermitian conjugate of U. Applying an operation U to a system initially is a state ρ maps the system to another state $\Lambda(\rho) = U\rho U^\dagger$. A unitary operation describes the evolution of a *closed* system. In practice, every quantum system is *open*, meaning that, unfortunately, each system is coupled to and interacting with some environment. It is possible to consider the unitary evolution of a larger system, including both the environment and our original system, and perform a partial trace over the environment to obtain the dynamics of the system of interest. Under the assumption that the system and bath are initially decoupled, this leads to the so-called *Kraus operator* representation.

More specifically, a general quantum process is a linear, completely positive map of density operators $\rho \rightarrow \Lambda(\rho)$ that take a system initially in the state ρ to another state $\Lambda(\rho)$. For any physical process, this map needs to be completely positive and trade-preserving. Such an operation can be expressed in the form

$$\Lambda(\rho) = \sum_i K_i \rho K_i^\dagger,$$

where K_i are the Kraus operators and for which it holds that $\sum_i K_i^\dagger K_i = I$, where I is the identity. This representation is particularly useful when describing noise processes that lead to errors.

Before going through the list of commonly used operations, we will first introduce the Pauli matrices

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad X = \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad Y = \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad Z = \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \qquad (1.1)$$

The n-qubit *Pauli group* $\mathscr{P}_n$ is defined as $\mathscr{P}_n = \{I, X, Y, Z\}^{\otimes n} \times \{\pm 1, \pm i\}$. Each element of $\mathscr{P}_n$ is therefore of the form $\lambda P_1 \otimes P_2 \otimes \ldots \otimes P_n$, i.e. the tensor product of n Pauli matrices P_i

and an overall phase $\lambda \in \{\pm 1, \pm i\}$. Note that each Pauli matrix is Hermitian and unitary, with $P_i^2 = I$ for $P_i \in \{I, X, Y, Z\}$. Taking the product of two Pauli matrices leads to another Pauli matrix along with some phase prefactor i (or $-i$ depending on the order of the matrices), that is,

$$XY = iZ, \qquad ZX = iY, \qquad YZ = iX.$$

Pauli matrices either commute or anti-commute with each other. In particular, the Pauli matrices, excluding the identity, anti-commute with each other

$$\{X, Y\} = 0, \qquad \{Y, Z\} = 0, \qquad \{Z, X\} = 0.$$

Of course, this means that these matrices do not commute with each other, and in particular, they lead to the following commutation relations

$$[X, Y] = 2iZ, \qquad [Y, Z] = 2iX, \qquad [Z, X] = 2iY.$$

Each Pauli matrix commutes with both the identity operator and with itself.

The Pauli matrices form an orthogonal basis for any $2 \otimes 2$ Hermitian. In particular, we can decompose any density matrix ρ as a linear combination of the Pauli matrices

$$\rho = \frac{1}{2}(I + r_X X + r_Y Y + r_Z Z), \tag{1.2}$$

where $r_X, r_Y, r_Z \in \mathbb{R}$ and $\sum_i r_i^2 \leq 1$. Therefore, a density matrix can be represented by the Pauli vector $\mathbf{r} = (r_X, r_Y, r_Z)$.

SINGLE-QUBIT GATES
Each single-qubit gate can be visualized as a rotation of the Bloch sphere around some axis. The first three commonly-seen single-qubit gates are given by X, Y, and Z unitary operators, corresponding to π-rotations around the $\hat{x}$, $\hat{y}$, and $\hat{z}$ axes of the Bloch sphere, respectively. Another set of commonly used gates are the Hadamard (H), S, and T gates, defined as the operators

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad S = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\frac{\pi}{2}} \end{pmatrix} = \sqrt{Z}, \qquad T = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\frac{\pi}{4}} \end{pmatrix} = \sqrt{S}. \tag{1.3}$$

The Hadamard gate can be visualized as a π-rotation around the $(\hat{x} + \hat{z})/\sqrt{2}$ axis of the Bloch sphere. This can also be seen as performing a change of basis, switching the $\hat{x}$ and $\hat{z}$ axes of the Bloch sphere. The S and T gates are types of phase-shift gates of the form $R_Z(\varphi) = \text{diag}(1, e^{i\varphi})$, where φ corresponds to the angle of rotation around the $\hat{z}$ axis. Therefore, $S = R_Z(\pi/2)$ and $T = R_Z(\pi/4)$, respectively. Similarly, the Z gate is also a phase-shift gate with $Z = R_Z(\pi)$. Periods of qubit idling are equivalent to applying an I gate. Unfortunately, this is only the case in an ideal world. In practice, the qubits experience decoherence during idling periods, which we will discuss later in Sec. 1.2.2.

1

The most commonly used two-qubit gates are the controlled-not (CNOT) and controlled-phase (CZ) gates, given by the unitary operators

$$\text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad \text{CZ} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{1.4}$$

The pair of qubits that these gates are applied to are called the *control* and *target* qubits. The CNOT applies an X gate on the target qubit whenever the control qubit is in state $|1\rangle$. The CZ gates applied a Z gate to the target qubit conditioned on the control qubit being in state $|1\rangle$. More generally, we can introduce the controlled-phase rotation gate of the form $\text{diag}(1,1,1,e^{i\varphi})$, where φ is the conditional phase picked up by the target qubit depending on the state of the control qubit, with the CZ gate corresponding to φ being an odd multiple of π.

The most general kind of measurement is a positive operator-valued measurement, or a POVM for short. Such a measurement is described by a set of measurement operators $\{M_i\}$, where i refers to the observed measurement outcome. These operators must satisfy the completeness relation $\sum_i M_i^\dagger M_i = I$. Consider a qubit in a state ρ. The probability of observing outcome i is given by $p_i = \text{Tr}\left(M_i^\dagger M_i \rho\right)$. The state of the system ρ' immediately after a measurement with outcome i is $\rho' = M_i \rho M_i^\dagger / p_i$. For a projective measurement, the measurement operators $M_i = \Pi_i$, where Π_i is an orthogonal projector, i.e. $\Pi_i^2 = \Pi_i$ and $\Pi_i^\dagger = \Pi_i$. In this case, the probability of observing outcome i reduces to $p_i = \text{Tr}\left(\Pi_i \rho\right)$, with $\sum_i \Pi_i = I$.

1.1.4. NOISE AND ERROR CHANNELS

As previously mentioned, quantum systems are never truly isolated, and the interactions with their environment lead to qubit decoherence. Furthermore, quantum operations are realized only with a limited fidelity. To model the errors resulting from these noise processes, theorists typically consider error channels that apply Pauli errors with some probability each. A commonly considered noise model is that of *independent depolarizing* noise, where each qubit is affected by the single-qubit depolarizing channel

$$\rho \mapsto \mathscr{E}(\rho) = (1-p)\rho + \frac{p}{3}\left(X\rho X + Y\rho Y + Z\rho Z\right). \tag{1.5}$$

In this case, the channel applies an X, Y or a Z error each with probability $p/3$. Otherwise, the state remains unchanged with probability $1-p$. The choice that each of these Pauli errors occurs with the same probability is an assumption typically made to reduce the number of parameters describing the noise model. Alternatively, one can consider a more general Pauli-noise channel of the general form

$$\rho \mapsto \mathscr{E}(\rho) = p_I\rho + p_X X\rho X + p_Y Y\rho Y + p_Z Z\rho Z, \tag{1.6}$$

with the constraints that $0 \le p_i \le 1$ and $\sum_i p_i = 1$. There are a few channels commonly considered in the literature that correspond to special subcases of this general channel. The bit-flip channel applies an X error with probability p, corresponding to

$$\rho \mapsto \mathcal{E}(\rho) = (1-p)\rho + pX\rho X. \tag{1.7}$$

The phase-flip channel replaces the X error with a Z error. Another relevant error channel in the context of the surface code is the independent bit-flip and phase-flip error channel, which independently applies an X or a Z error with probability p each, leading to a probability of a Y error that is $O(p^2)$.

When modeling noise, the choice of where the error channels are inserted plays a significant role. Simpler models typically consider a set of independent single-qubit channels affecting a subset of the qubits before a set of otherwise ideal operations. A more realistic approach is to consider errors during each idling period and as a result of each applied operation, commonly referred to as *circuit-level* noise models. In such a model, a single-qubit noise channel (typically a depolarizing channel) is applied after each single-qubit gate or idling period. Naturally, two-qubit operations are also considered noisy. It is essential to consider two-qubit error channels after two-qubit gates. Typically, models consider the two-qubit depolarizing channel

$$\mathcal{E}(\rho) = (1-p)\rho + \frac{p}{15} \sum_{P_i \in \{I,X,Y,Z\}^{\otimes 2}/II} P_i \rho P_i. \tag{1.8}$$

However, similar to the single-qubit case, this channel can be generalized to a general two-qubit Pauli noise channel. Such noise channels can be used when the noise in a given system is biased towards a specific type of error, i.e., when some Pauli errors occur with a higher probability than others. Finally, in a circuit-level noise model, single-qubit noise channels are also applied following qubit initialization and before measurement operations. Depending on the noise model, these are typically either depolarizing channels or either bit-flip or phase-flip channels, depending on the basis in which the qubit was prepared or measured.

1.2. Superconducting quantum processors

Several experimental qubit platforms are currently being explored, with superconducting qubits and trapped-ion qubits perhaps being the most promising platforms nowadays. In this dissertation, we focus on superconducting qubits, particularly on the superconducting *transmon* qubit [51]. In this section, we will give a brief summary of the design and operation of this qubit.

1.2.1. The transmon qubit

One of the simplest and most fundamental circuits used in superconducting-qubit processors is the linear LC-resonator, often referred to simply as a resonator. This system consists of a capacitor with capacitance C in parallel with a linear inductor with an inductance L, with the corresponding circuit shown in Fig. 1.1**a**. The energy stored in this system oscillates between the electrical energy (analogous to the "kinetic energy" of a

1
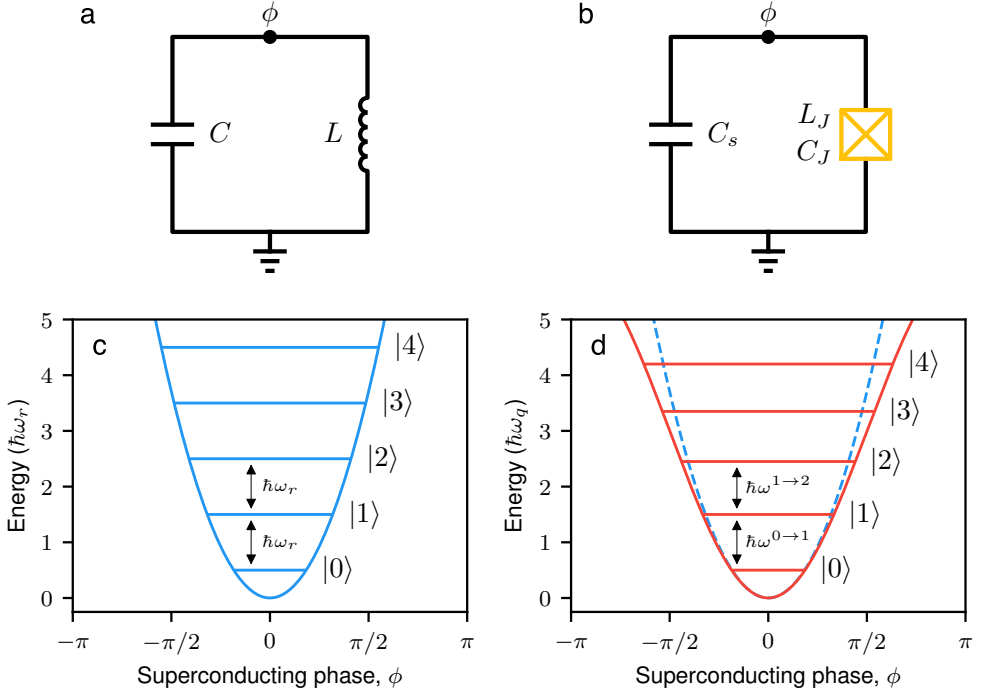
Figure 1.1: **a** Circuit diagram for a linear resonator consisting of a capacitance C in parallel with an inductance L. **b** Circuit diagram for a transmon qubit, consisting of a shunting capacitance C_s in parallel with a Josephson junction (yellow), characterized by a nonlinear inductance L_J and a self-capacitance C_J. **c** The energy potential of the resonator as a function of the superconducting phase ϕ. The energy levels of the resonator are equidistant, with an energy separation of $\hbar\omega_r$. **d** The energy potential of the transmon qubit (solid red line). The anharmonicity introduced by the Josephson junction transforms the quadratic potential of the resonator (dashed blue lines) into a cosine potential, resulting in non-equidistant energy levels. In particular, $\omega^{0\to1} = \omega_q$, while $\omega^{1\to2} = \omega_q + \alpha$. Since $\omega^{0\to1} \neq \omega^{1\to2}$, the states $|0\rangle$ and $|1\rangle$ can be used to encode a qubit.

system) stored in the capacitor and the magnetic energy stored in the inductor (analogous to the "potential energy"). The Hamiltonian for this system can be expressed as

$$H_r = \frac{Q^2}{2C} + \frac{\Phi^2}{2L}, \tag{1.9}$$

where Q and Φ are the charge and flux operators, which can be interpreted as the charge on the capacitor and the flux threading the inductor. These operators satisfy the commutation relation $[\Phi, Q] = i\hbar I$, where $\hbar$ is the reduced Planck's constant $h/(2\pi)$. By introducing the reduced (dimensionless) charge $q \equiv Q/(2e)$ and flux $\phi \equiv 2\pi\Phi/\Phi_0$ operators, we can rewrite the Hamiltonian as

$$H_r = 4E_C q^2 + \frac{1}{2}E_L\phi^2, \tag{1.10}$$

where $E_C = e^2/(2C)$ is the charging energy, $E_L = (\Phi_0/2\pi)^2/L$ is the inductive energy, $\Phi_0 = h/(2e)$ is the magnetic flux quantum, and e is the electron charge.

Following standard circuit quantization theory [52, 53], we can introduce the standard creation $a^\dagger$ and annihilation a operators, which allows us to express the resonator Hamiltonian in the form

$$H_r = \hbar\omega_r(a^\dagger a + 1/2), \tag{1.11}$$

where $\omega_r = \sqrt{8E_L E_C}/\hbar = 1/\sqrt{LC}$ is the resonant frequency of the resonator. These operators are related to the charge and flux operators by the relations

$$q = i q_{\text{zpf}}(a^\dagger - a), \qquad\qquad \phi = \phi_{\text{zpf}}(a^\dagger + a),$$

where $q_{\text{zpf}} = [E_L/(32E_C)]^{1/4}$ and $\phi_{\text{zpf}} = (2E_C/E_L)^{1/4}$ are the magnitudes of zero-point fluctuations of the charge and flux, respectively. The eigenstates $|k\rangle$ of this Hamiltonian satisfy $a^\dagger a|k\rangle = k|k\rangle$ for $k = 0,1,2,\ldots$, with each eigenstate having a corresponding eigenenergy of $E_k = \hbar\omega_r(k+1/2)$, or equivalently a frequency $\omega_k = E_k/\hbar$. In other words, there are infinitely many eigenstates, each separated by an energy $E_{k+1} - E_k = \hbar\omega_r$, where ω_r is the resonant frequency of the resonator, with $\omega_r/(2\pi)$ typically in the $3-9$ GHz range (though it can also be as high as 20 GHz, for example). This energy potential is illustrated in Fig. 1.1**c**.

Let us consider whether we can use such a resonator as a qubit, for example, by using the two lowest-energy states $|0\rangle$ and $|1\rangle$. While these two states indeed define a computational subspace, the equidistant level spacing makes driving specific transitions impossible without exciting the qubit to $|2\rangle$, $|3\rangle$, and other higher-excited states. Some degree of non-linearity is therefore required to define a qubit that we can actually use. Such a non-linearity is also commonly referred to as anharmonicity. Let us define $\omega^{i\to j} = \omega_j - \omega_i = (E_j - E_i)/\hbar$ as the transition frequency between state $|i\rangle$ and $|j\rangle$. Therefore, we require the transition frequency $\omega^{0\to 1}$ and $\omega^{1\to 2}$ to be sufficiently different to avoid driving any of the population outside the computational subspace defined by the lowest two energy levels, which is more commonly referred to as *leakage*.

Fortunately, we can introduce such an anharmonicity using a Josephson junction [54, 55], a non-linear circuit element that is at the heart of many superconducting qubits. A Josephson junction consists of two pieces of superconductor coupled by a thin insulating barrier. In particular, the Josephson junction behaves as a non-linear inductor, which can be derived by considering the Josephson relations

$$I = I_c \sin(\phi), \qquad\qquad V = \frac{\hbar}{2e}\frac{d\phi}{dt},$$

where I_c is the critical current of the junction and ϕ is the superconducting phase difference across the junction. We can then introduce the Josephson energy $E_J = \Phi_0 I_c/(2\pi)$ and the Josephson inductance $L_J = \Phi_0^2/(4\pi^2 E_J)$. Finally, by replacing the linear inductor of an LC resonator circuit with a Josephson junction (see Fig. 1.1**b**), we then obtain the Hamiltonian [53]

$$H_{\text{tr}} = 4E_C q^2 - E_J \cos(\phi), \tag{1.12}$$

with ϕ now being an operator. The Josephson junction typically has some self-capacitance C_J, while the parallel capacitor has a capacitance C_s, typically referred to as the *shunting*

capacitance. The total capacitance is then given by $C_\Sigma = C_s + C_J$, which also defines the charging energy in this case $E_C = e^2/(2C_\Sigma)$. Including this non-linear element makes the energy levels of this system no longer equivalent, making it possible for specific energy transitions to be addressed, which is illustrated in Fig. 1.1**d**. A qubit defined by such a circuit is commonly known as a transmon qubit [51, 56]. The transmon is typically designed to be in the limit of $E_J \gg E_c$, which allows for charge noise to be exponentially suppressed. This is typically achieved by using a larger shunting capacitance C_s, such that $C_s \gg C_J$ (typically $E_J/E_C \geq 30$). However, in this regime, the circuit is only weakly anharmonic. We can define the qubit transition frequency as $\omega_q = \omega^{0\to1} = (\sqrt{8E_JE_C} - E_C)/\hbar$, with $\omega_q/(2\pi)$ typically being the range between $3 - 7$ GHz. We can also define the anharmonicity $\alpha = \omega^{1\to2} - \omega^{0\to1} = -E_C/\hbar$, with $\alpha/(2\pi)$ typically chosen to be in the range $200 - 400$ MHz. Performing a power series expansion of the potential term of Eq. (1.12), keeping the terms up to fourth order, introducing the creation and annihilation operators, and performing a rotating-wave approximation allows us to rewrite the transmon Hamiltonian as that of a Duffing oscillator

$$H_{\mathrm{tr}} \approx \hbar\omega_q a^\dagger a + \hbar\frac{\alpha}{2}a^\dagger a^\dagger a a. \tag{1.13}$$

Since $|\alpha| \ll \omega_q$, this more clearly illustrates that the transmon can be seen as a weakly-anharmonic resonator. Unfortunately, the weak anharmonicity means each operation applied to the transmon will have some probability of exciting the qubit to the higher-excited states.

The transmon frequency can be made flux-tunable by replacing the Josephson junction with a superconducting quantum interference device (SQUID) consisting of a pair of Josephson junctions connected in a loop [51]. Tunability is achieved by considering the external flux Φ_e threading the SQUID loop. The Hamiltonian is then given by

$$H_{\mathrm{tr}} = 4E_c q^2 - E_{J1}\cos(\phi) - E_{J2}\cos(\phi - \phi_e), \tag{1.14}$$

where $\phi_e = 2\pi\Phi_e/\Phi_0$ is the reduced external flux, while E_{J1} and E_{J2} are the Josephson energies of two junctions. By performing the change of variables $\phi \to \phi - \phi_e/2$ and by using some trigonometric identities, the expression for the Hamiltonian can be rewritten as [51, 53]

$$H_{\mathrm{tr}} = 4E_c q^2 - E_{J\Sigma}(\phi_e)\cos(\phi - \varphi_e), \tag{1.15}$$

where the angle φ_e is defined by $\tan\varphi_e = d\tan(\phi_e/2)$, $d = |E_{J2} - E_{J1}|/(E_{J1} + E_{J2})$ is the junction asymmetry, and

$$E_{J\Sigma}(\phi_e) := (E_{J1} + E_{J2})\sqrt{\cos^2\left(\frac{\phi_e}{2}\right) + d^2\sin^2\left(\frac{\phi_e}{2}\right)}. \tag{1.16}$$

For a constant magnetic flux, φ_e can be ignored by a change of variable [51], which brings Eq. (1.15) in the form of Eq. (1.12). We see that $E_{J\Sigma}(\phi_e)$ plays the role of an effective and tunable E_J. When the two junctions are symmetric i.e. when $E_{J1} = E_{J2} = E_J$ such that $d = 0$, the expression $E_{J\Sigma}(\phi_e)$ further simplifies to $E_{J\Sigma}(\phi_e) = 2E_J|\cos(\phi_e/2)|$.

Introducing the SQUID loop makes the qubit transition frequency ω_q flux-tunable, i.e., $\omega_q \rightarrow \omega_q(\phi_e) \propto \sqrt{|\cos(\phi_e/2)|}$.

Flux tunability allows for great flexibility when operating a transmon qubit. For example, transmon qubits can be fluxed away to avoid interactions with two-level systems that lead to lower coherence times [57]. Qubits can also be fluxed away to avoid any unwanted interactions when executing operations in parallel [58, 59]. Finally, conditional operations can be performed by flux-pulsing qubits close to interaction points [60] or by parametrically driving the interaction using a flux modulation [61]. However, this flexibility comes at the price of an increased sensitivity to flux noise. Transmons qubits are typically operated at a frequency sweet spot [56], where the qubit is first-order insensitive to flux noise. The most commonly used sweet spot is where the transmon frequency ω_q is at its maximum, typically at $\phi_e = 0$. Therefore, flux-pulsing a qubit to a lower frequency reduces the qubit dephasing time due to an increased flux-noise sensitivity, which we will discuss in the following section.

1.2.2. QUBIT DECOHERENCE

The interactions between the qubit and its environment lead to the decoherence of the quantum state. For transmon qubits, the dominant sources of decoherence are the energy relaxation, characterized by a timescale T_1, and the dephasing, characterized by a timescale T_2. We will assume that the dynamics of the system are Markovian, which enables us to describe the time evolution of the system using a *Lindblad* master equation

$$\dot{\rho} = \frac{d\rho}{dt} = -\frac{i}{\hbar}[H, \rho] + \sum_k L_k \rho L_k^\dagger - \frac{1}{2}\left\{L_k^\dagger L_k, \rho\right\}, \tag{1.17}$$

where H is the Hamiltonian of the system, ρ is the density matrix describing the state, and $\{L_k\}$ are the Lindblad operators (also called the quantum jump operators). The first term describes the coherent evolution of the system, while the second term, involving the Lindblad operators, models the decoherence experienced by the qubit due to the environment. For the sake of simplicity, we will first treat the transmon qubit as a two-level system to investigate the impact of decoherence on the computational subspace and then extend the results to the higher-excited states. In particular, the transmon Hamiltonian is then given by $H = -\hbar\omega_q \sigma_z/2$.

Thermal relaxation typically refers to the tendency of a transmon qubit prepared in the excited state $|1\rangle$ to relax back to the ground state $|0\rangle$, corresponding to the process of losing energy to the environment. Generalizing this, we can consider the energy exchange with an environment in thermal equilibrium at some finite temperature T. This process can be modeled by the Lindblad operators $L_\downarrow = \sqrt{\gamma_\downarrow}\sigma_-$ and $L_\uparrow = \sqrt{\gamma_\uparrow}\sigma_+$, where $\sigma_- = |0\rangle\langle 1|$ and $\sigma_+ = |1\rangle\langle 0|$. Here, $L_\downarrow$ models energy relaxation, while $L_\uparrow$ corresponds to energy excitation. The relaxation $\gamma_\downarrow$ and excitation $\gamma_\uparrow$ are related to the relaxation time T_1 via

$$\frac{1}{T_1} = \gamma_1 = \gamma_\downarrow + \gamma_\uparrow. \tag{1.18}$$

1

These rates also obey a detailed balance condition

$$\frac{\gamma_\uparrow}{\gamma_\downarrow} = e^{-\frac{\hbar\omega_q}{k_B T}}, \qquad (1.19)$$

where k_B is the Boltzmann constant. These operators are also commonly expressed in the form

$$\gamma_\downarrow = \gamma\left[1 + \bar{n}(\omega_q)\right], \qquad (1.20)$$

$$\gamma_\uparrow = \gamma\bar{n}(\omega_q), \qquad (1.21)$$

where $\bar{n}(\omega_q)$ is the thermal photon number of the environment, given by

$$\bar{n}(\omega_q) = \frac{1}{e^{\frac{\hbar\omega_q}{k_B T}} - 1}. \qquad (1.22)$$

Transmon qubits are relatively high-frequency qubits (in the range $3 - 7$ GHz) and are typically operated at low temperatures (about 20 mK). For example, a transmon qubit with a frequency of $\omega_q/(2\pi) = 6$ GHz and a temperature of $T = 20$ mK is expected to have a $\bar{n}(\omega_q) \approx 6 \times 10^{-7}$. In other words, in the limit of $\lim_{T\to 0} \bar{n}(\omega_q) = 0$ the relaxation rate $\gamma_\downarrow$ dominates over the excitation rate $\gamma_\uparrow$ and $\gamma \simeq \gamma_1$ in this case. Because of this, most error models neglect the impact of excitation and only model the relaxation that the qubit experiences. In practice, $\bar{n}(\omega_q)$ is not typically negligible, which leads to the *residual excitation* of the qubit [62–65]. The other major noise source that qubits experience is dephasing, which can be attributed to a dispersive coupling to the environment or fluctuations of the external flux field that controls the qubit frequency. This process is modeled by the Lindblad operator $L_2 = \sqrt{\gamma_\phi/2}\sigma_z$, where γ_ϕ is the pure dephasing rate. This is related to the decoherence time T_2 (sometimes referred to as the dephasing time) via

$$\frac{1}{T_2} = \gamma_2 = \frac{\gamma_1}{2} + \gamma_\phi, \qquad (1.23)$$

with γ_2 being the decoherence rate, which includes contributions from both the pure dephasing and the relaxation. The time $T_\phi = 1/\gamma_\phi$ is referred to as the pure-dephasing time. Combining the effects of relaxation and assuming that $\bar{n}(\omega_q) = 0$, an initial density matrix $\rho(t = 0)$ evolves as

$$\rho(t = 0) = \begin{pmatrix} 1 - \alpha & \beta \\ \beta^* & \alpha \end{pmatrix} \mapsto \rho(t) = \begin{pmatrix} 1 - \alpha e^{-t\gamma_1} & \beta e^{i\omega_q t} e^{-t\gamma_2} \\ \beta^* e^{-i\omega_q t} e^{-t\gamma_2} & \alpha e^{-t\gamma_1} \end{pmatrix}. \qquad (1.24)$$

So far, we considered the transmon to be an ideal two-level system, which, at this point, we know that it isn't. In particular, let us also consider the second-excited state in our model. The transmon is weakly anharmonic, which allows approximating the relaxation experienced by the higher-excited states with the Lindblad operator $L_\downarrow = \sqrt{\gamma_\downarrow}a$, with a being the annihilation operator. Similarly, the excitation can be modeled with the operator $L_\uparrow = \sqrt{\gamma_\uparrow}a^\dagger$. The dephasing experienced by the higher-excited states is typically modeled by considering the operator $L_2 = \sqrt{2\gamma_\phi}a^\dagger a$ [53]. However, in Ref. [66], it

was observed that the scaling of the dephasing rate with the flux-noise sensitivity (see below) predicted by this model does not match what was observed in the experiment. A different set of operators was instead proposed, which are given in Chapter 2, more specifically in Sec. 2.4.2.

Away from the frequency sweet spot, the transmon qubit becomes first-order sensitive to flux noise, which increases the pure dephasing rate. Flux noise has a power spectral density

$$S(f) \sim A/f,$$

where f is a frequency and $\sqrt{A}$ is a constant that is typically in the range $1 - 10\,\mu\Phi_0$. The sensitivity away from the sweet spot is given by

$$D_\phi = \frac{1}{2\pi} \left| \frac{\partial \omega_q}{d\Phi_e} \right|, \tag{1.25}$$

with Φ_e being the external flux threading the SQUID loop. The high-frequency components of the noise (relative to the time the qubit spends away from the sweet spot) lead to an increase in γ_ϕ, given by [67]

$$\gamma_\phi = 2\pi \sqrt{\ln 2} \sqrt{A} D_\phi. \tag{1.26}$$

The value of $\sqrt{A}$ can be extracted experimentally by fluxing the qubit away and measuring the increase in the dephasing time. Note that the dephasing rate might not increase linearly with the sensitivity if the qubit is fluxed far away from the sweet spot since second-order effects might become important. If the qubit is fluxed to only a few different frequencies, the dephasing rate at those points can also be directly measured and included in the error model. The low-frequency components of flux noise can be modeled as quasi-static over the duration of an operation or an individual run of an experiment but fluctuating across repetitions of this operation/experiment. In particular, the quasi-static components of the flux noise can be modeled as a random shift to the external flux field [66, 68]. Fortunately, these slow components can often be mitigated using dynamical decoupling methods [69, 70] or be compensated for in calibration (depending on the time scale that they actually change over).

We note that T_2 refers to the intrinsic decoherence time of a qubit. In practice, two different decoherence times are measured in experiments, T_2^* and T_2^E. If the decoherence time is measured using a standard Ramsey experiment, it refers to T_2^*. Since there is no dynamical decoupling used in this experiment, one typically measures a time $T_2^* \leq T_2$ that might fluctuate over some time. Instead, if the decoherence time is measured using an Echo experiment, it refers to T_2^E. The echo pulse is then expected to remove the impact of these noise components, leading to $T_2^E \geq T_2^*$ and perhaps closer to the value of T_2. The exact improvement in the dephasing time from using echo pulses depends on the power spectrum of the noise. However, since an improvement is typically observed, most experiments employ operations with built-in echo effects [66] and further include dynamical decoupling during periods of qubit idling or between operations [20, 30]. Because of this, we normally consider the measured T_2^E for our dephasing models.

Finally, for a more detailed overview of the physical noise sources encountered in superconducting circuits, we refer the reader to Refs. [71, 72].

1.2.3. NOISE MODELS IN THIS DISSERTATION

In this section, we briefly comment on our approach to modeling errors in the simulations performed in this dissertation. In particular, we perform both density-matrix simulations and Clifford simulations, which are simulations of circuits involving only Clifford operations and measurements in the Pauli basis, making them efficient to simulate using the Gottesman-Knill theorem [73, 74].

When performing Clifford simulation, we are generally restricted in the noise we can consider, and typically, only Pauli error channels are included in such simulations. In this work, we either consider a standard circuit-level depolarizing noise model or attempt to parameterize the Pauli noise model using the physical error rates to construct a more physically realistic model. In particular, we consider the relaxation and dephasing that the qubit experiences during idling and Pauli-twirl the resulting channel to make it compatible with the simulation, see Chapter 7. For operations, we consider depolarizing channels parameterized by the measured error rates. In the case of the measurements, we model the assignment error rate and the quantum non-demolition probability by introducing bit-flip error channels before and after an otherwise ideal measurement operation, see Sec. 7.4.2.

For the density-matrix simulations, we consider more physically motivated error models. However, in many cases, we want to investigate the behavior of some specific error (namely, leakage) or to capture the most significant error sources impacting an experiment. Therefore, we make several simplifying assumptions and neglect certain physical noise sources. This also reduces the simulation's complexity and the time required to perform the numerical scans. In particular, we don't typically consider the slow components of the flux noise (or other noise sources) since we assume they are echoed out. The impact of the fast components is included by considering the relaxation and dephasing experienced by the qubits (including the increased dephasing away from the sweet spot) during periods of idling or operations. In particular, we normally consider the ideal operation and symmetrically introduce amplitude-phase damping operations for half of the duration. This is described in more detail in Chapter 2, see also Ref. [68]. Although we typically only consider the decoherence occurring during readout, the measurement error model used in Chapter 6 is instead directly parameterized by the assignment error rates benchmarked in the experiment. Finally, we consider leakage due to the two-qubit controlled-phase gates, following the model introduced in Chapter 2.

Generally, we have neglected several non-conventional error sources that might have a considerable impact, particularly residual ZZ crosstalk or other types of crosstalk, and the impact that two-level-system defects can have on the operational error rates or coherence times [43, 57]. Finally, we do not consider leakage in our Clifford simulations, even though a stochastic leakage model can still be simulated efficiently and capture the impact of this error [20].

1.2.4. OPERATING A TRANSMON QUBIT

INITIALIZATION AND RESET

The most straightforward way to initialize a transmon qubit is simply by leaving the qubit idle for several times longer than the relaxation time T_1. For transmon qubits, this typically leads to the qubit being in the ground state with high probability. This is more

commonly known as passive initialization. However, there is still some probability of the qubit being excited, widely referred to as a residual excitation [62–65]. Furthermore, idling for an extended period can become inefficient for qubits with higher T_1 [63]. If a high-fidelity, quantum non-demolition (QND) readout is available, the qubit can be initialized by measurement and post-selection on the qubit being in the ground state [63]. This is sometimes referred to as a *heralded* initialization. To avoid any post-selection, a conditional operation can be used to transfer the population in $|1\rangle$ to $|0\rangle$ whenever the qubit is measured to be excited [75, 76]. However, the latency between the measurement and the conditional operation should be low to maximize the initialization fidelity. Finally, initialization may be performed using an unconditional reset operation [47, 77–79]. Resetting the state of a qubit is one of the slower operations, with a typical duration being in the range $100 - 500$ ns range. These operations reset a qubit to the ground state with a typical fidelity of around 99% or higher.

READOUT

Transmon measurement is typically based on dispersive readout [80–82]. We consider a transmon qubit capacitively coupled to a dedicated readout resonator. In the dispersive regime, when the qubit frequency is far detuned from the resonator frequency (relative to the coupling strength), there is a state-dependent shift of the resonator frequency. This can be used to infer the qubit's state by probing the resonator. The Hamiltonian of this system is given by [71]

$$H \approx \hbar\omega_r a^\dagger a + \hbar\omega_q b^\dagger b + \hbar\frac{\alpha}{2}b^\dagger b^\dagger bb + \hbar g(ab^\dagger + a^\dagger b), \tag{1.27}$$

where a ($a^\dagger$) and b ($b^\dagger$) are the annihilation (creation) operators for the resonator and transmon qubit, respectively, ω_r and ω_q are the resonator and qubit frequencies, α is the transmon anharmonicity, and g is the coupling strength. Here, we have approximated the transmon Hamiltonian as that of a Duffing oscillator and taken a rotating-wave approximation. To further simplify the discussion, it is useful to restrict this Hamiltonian to the computational levels by taking $b^\dagger \to \sigma_+$ and $b \to \sigma_-$, which leads to the infamous Jaynes-Cumming Hamiltonian [51, 83]

$$H_{\mathrm{JC}} = \hbar\omega_r a^\dagger a + \hbar\frac{\omega_q}{2}\sigma_z + \hbar g(a\sigma_+ + a^\dagger \sigma_-). \tag{1.28}$$

This Hamiltonian is exactly solvable and can accurately describe many interactions and processes where the interactions with the higher-excited states do not play a significant role. The dispersive regime is obtained when the qubit-resonator detuning $\Delta = \omega_q - \omega_r$ is much larger than the coupling strength g, i.e., when $\left|g/\Delta\right| \ll 1$. In this regime, there is no direct exchange interaction between the transmon qubit and the readout resonator. This means that we can accurately approximate the Hamiltonian to second order using perturbation theory in terms of g/Δ, resulting in [53, 80, 81]

$$H_{\mathrm{disp}} \approx \hbar\omega_r a^\dagger a + \hbar\frac{(\omega_q + \chi)}{2}\sigma_z + \hbar\chi\sigma_z a^\dagger a, \tag{1.29}$$

where $\chi = g^2/\Delta$ is the qubit-state dependent shift of the resonator frequency, commonly known as the *dispersive shift*. If the qubit is in $|0\rangle$, the frequency of the resonator is $\omega_r - \chi$.

1

Alternatively, if the qubit is in $|1\rangle$, the frequency is shifted to $\omega_r + \chi$. Measurement can be performed by applying a microwave tone to the resonator at some probe frequency. This populates the resonator with photons, resulting in a state-dependent coherent state. The output field of the resonator is typically measured using heterodyne detection, extracting the "in-phase" (I) and "quadrature" (Q) components of the signal. These signal components are combined and integrated over time to produce the final analog outcome in the IQ plane, which contains information about the qubit's state. This value can then be converted to a binary outcome by using a threshold that maximizes the measurement fidelity, which is determined during calibration.

In the dispersive regime, the interactions between the qubit and the resonators are virtual processes. However, these processes can involve the higher-excited states, which we have ignored so far but are important to consider. Accounting for these interactions [53] leads to a dispersive shift of $\chi = g^2 \alpha / (\Delta(\Delta + \alpha))$. The qubit frequency is shifted to $\tilde{\omega}_q = \omega_q + g^2/\Delta$, and the resonator frequency is shifted to $\tilde{\omega}_r = \omega_r - g^2/(\Delta + \alpha)$, which together lead to the Hamiltonian

$$H_{\text{disp}} \approx \hbar\tilde{\omega}_r a^\dagger a + \hbar\frac{\tilde{\omega}_q}{2}\sigma_z + \hbar\chi\sigma_z a^\dagger a. \tag{1.30}$$

Since σ_z commutes with H_{disp}, dispersive readout constitutes a QND measurement, meaning that any subsequent measurement should lead to the same outcome as the initial measurement. For the dispersive approximation to be accurate, the number of photons n in the resonator should be well below the critical photon number n_{crit}, i.e., $n \ll n_{\text{crit}}$, where $n_{\text{crit}} = \Delta^2/(4g^2)$ [53]. We refer the reader to Refs. [53, 71] for a more detailed discussion on the readout and the detection of the outgoing resonator field.

The readout of transmon qubits is one of the slowest operations and typically exhibits a higher error rate. In most cases, the readout is done in around 500 ns [20, 29, 30, 32], though there have been demonstrations of optimized setups that can measure a qubit in about 100 ns or less [84–87]. The readout fidelity is typically around 99% [20, 29, 30, 32, 84, 86, 88], with some measurements approaching 99.9% [85, 87]. As with other operations, the readout can also induce leakage [89–91]. Finally, we note that there are methods for improving the readout performance by using a flux pulse [87] or a microwave pulse [92, 93] that are typically not employed in most current experiments.

SINGLE-QUBIT GATES

Single-qubit gates are typically performed using a microwave drive. Typically, each transmon qubit has a dedicated microwave-drive line (see Ref. [59]) that is capacitively coupled to the qubit and a time-dependent voltage $V_d(t)$ used to perform rotations. This introduces a drive term to the Hamiltonian of the form $H_d = i\hbar\xi_d(t)\left(a^\dagger - a\right)$, with ξ_d being proportional to the voltage $V_d(t)$. In particular, choosing the form of the applied voltage such that $\xi_d(t) = 2\varepsilon_d(t)\cos\left(\omega_d t + \phi_d\right)$ results in

$$H_d = i\hbar\varepsilon_d(t)\left[e^{i(\omega_d t + \phi_d)} + e^{-i(\omega_d t + \phi_d)}\right]\left(a^\dagger - a\right), \tag{1.31}$$

where ω_d is the drive frequency, ϕ_d is the phase of the drive, $\varepsilon_d(t)$ is the time-dependent drive amplitude. Such a microwave drive can be used to perform rotations around any

axis in the equatorial plane of the Bloch sphere. More specifically, the axis of rotation is determined by the phase ϕ_d, while the angle of rotation is determined by the duration t and amplitude $\varepsilon_d(t)$.

Since transmon qubits are weakly anharmonic, these microwave pulses can lead to leakage outside the computational subspace due to a non-zero spectral overlap with the ω_{12} transition frequency. Note that this can also lead to phase errors as the drive may lead to a repulsion of $|1\rangle$ and $|2\rangle$. This results in a shift of ω_{01}, leading to the accumulation of a relative phase between these states. To mitigate this, the so-called DRAG (Derivative Reduction Adiabatic Gate) [94, 95] pulse-shaping technique can minimize both types of errors. This technique has enabled single-qubit gates to be routinely realized with high fidelity (99.9% or higher [35, 96]), with typical gate times being about 20 ns.

Phase gates, or more generally, rotations around the $\hat{z}$ axis of any angle, can be performed "in software" by updating the phase of the microwave drive [71, 97]. This method is preferred over realizing these rotations by flux-pulsing the qubit away from the frequency sweet spot for a fixed duration. This is because performing these virtual gates does not require additional microwave or flux pulses, meaning that they can be applied instantly (by being realized together with other pulses) and with no error rate.

TWO-QUBIT GATES

The interactions between pairs of transmon qubits are enabled by introducing a coupling between them. While this can be a direct capacitive coupling, typically, the coupling is mediated by a resonator instead [80, 81, 98]. This allows the transmon qubits to be placed further apart on the device. The resonator mediating this interaction is typically called the *coupling bus resonator*. Here, we will focus on such a coupling.

Let us consider a system of two transmon qubits, each coupled to the same bus resonator. The Hamiltonian corresponds to this system can be written as

$$H \approx \sum_{i=1}^{2} \left[\hbar\omega_{qi} b_i^\dagger b_i + \hbar\frac{\alpha_i}{2} b_i^\dagger b_i^\dagger b_i b_i \right] + \hbar\omega_r a^\dagger a + \sum_{i=1}^{2} \hbar g_i (a b_i^\dagger + a^\dagger b_i), \tag{1.32}$$

where ω_{qi} and α_i are the frequency and anharmonicity of each transmon qubit, with $i \in \{1, 2\}$. The frequency of the resonator is ω_r, g_i is the coupling between each transmon with the resonator. Finally, a is the annihilation operator for the resonator, and b_i is the annihilation operator for each transmon. Here, we have once again modeled each transmon as a Duffing oscillator and performed a rotating-wave approximation. Next, we consider the dispersive regime, where both qubits are far detuned from the resonator, i.e., when $|g_i/\Delta_i| \ll 1$ with $\Delta_i = \omega_{qi} - \omega_r$ for $i \in \{1, 2\}$. In this regime, the two transmon qubits can interact by only virtually populating the resonator [53, 80, 81, 98].

The Hamiltonian can be rewritten by using the approximate dispersive transformation and an expansion to second order in g_i/Δ_i, leading to [53]

$$\begin{aligned}\tilde{H} \approx & \sum_{i=1}^{2} \left[\hbar\tilde{\omega}_{qi} b_i^\dagger b_i + \hbar\frac{\alpha_i}{2} b_i^\dagger b_i^\dagger b_i b_i \right] + \hbar\tilde{\omega}_r a^\dagger a + \hbar J \left(b_1^\dagger b_2 + b_1 b_2^\dagger \right) \\ & + \sum_{i=1}^{2} \hbar\chi_i a^\dagger a b_i^\dagger b_i + \sum_{i \neq j} \hbar\Xi_{ij} b_i^\dagger b_i \left(b_i^\dagger b_j + b_j^\dagger b_i \right), \end{aligned} \tag{1.33}$$

1

where the frequencies $\tilde{\omega}_{qi}$ and $\tilde{\omega}_r$ include Lamb shifts, $\chi_i \simeq 2\alpha_i g_i^2/\Delta_i^2$ is a cross-Kerr coupling strength between each qubit and the resonator, while $\Xi_{ij} = -\alpha_i g_i g_j/(2\Delta_i\Delta_j)$ is the strength of a coupling interaction that depends on the qubit excitation number and that is small and can therefore be ignored. Importantly, J is the resonator-induced coupling strength between the two qubits that enables their interaction, which is given by

$$J = \frac{g_1 g_2}{2}\left(\frac{1}{\Delta_1} + \frac{1}{\Delta_2}\right). \tag{1.34}$$

If the resonator is in the ground state, then the cross-Kerr coupling and the resonator Hamiltonian terms can be ignored, leading $\tilde{H}$ to be in the same form as the Hamiltonian that can be obtained when considering a direct capacitive coupling, see Refs. [53, 71]. We can expand the coupling term in the form

$$\hbar J\left(b_1^\dagger b_2 + b_1 b_2^\dagger\right) = \hbar\left[J\left(|01\rangle\langle10| + \text{h.c.}\right) + \sqrt{2}J\left(|11\rangle\langle02| + \text{h.c.}\right) + \ldots\right], \tag{1.35}$$

where $|ij\rangle = |i\rangle \otimes |j\rangle$ denotes the product state of the first and second transmon qubits, respectively. Expressing the coupling term in this form reveals exchange interactions in the different excitation manifolds that scale with (approximately) a factor of $\sqrt{2}$ with the excitation number. Therefore, interactions involving the higher-excited states are expected to lead to shorter two-qubit gate times. Note that the frequencies of the two coupled transmon qubits are chosen such that when they are at their respective frequency sweet spots, this interaction is significantly suppressed. However, there is still some residual coupling, which is an important error source that we discuss in more detail later in this section.

Next, we consider both of these transmon qubits to be flux-tunable and explore the frequency spectrum of this system as a function of the applied flux. In particular, we will consider using the interactions between the computational and higher-excited states to perform a two-qubit controlled-phase gate [60], which is also the two-qubit gate we will consider throughout this dissertation. We assume that $\omega_{q2} > \omega_{q1}$ such that an external flux is applied only to the second transmon to bring its frequency down to an interaction point while the first transmon remains at its frequency sweet spot.

When the two transmon qubits are coupled to each other, the exchange interactions result in an avoided crossing between the states $|11\rangle$ and $|02\rangle$ at the interaction point where $\omega_2 = \omega_{q1} - \alpha_2$. Without a coupling, these states would be degenerate at this point. However, the coupling mediated by the resonator lifts this degeneracy and shifts the frequency ω_1 of the state $|11\rangle$ (relative to the single-excitation states) by an amount $\zeta = \omega_{11} - \omega_{01} - \omega_{10} + \omega_{00}$, typically referred to as the ZZ coupling strength. This frequency shift can be exploited to realize a conditional-phase gate [60, 99].

Flux-pulsing the higher-frequency qubit down from its maximum frequency at the sweet spot to the interaction point $\omega_{\text{int}} = \omega_{q1} - \alpha_2$, i.e., to the $|11\rangle \leftrightarrow |02\rangle$ avoided crossing, for some fixed time leads to an accumulation a conditional phase. More specifically, restricting ourselves to the computational subspace, this interaction leads to a phase-

gate of the form

$$U_{\text{CZ}} = \begin{pmatrix} e^{i\phi_{00}} & 0 & 0 & 0 \\ 0 & e^{i\phi_{01}} & 0 & 0 \\ 0 & 0 & e^{i\phi_{10}} & 0 \\ 0 & 0 & 0 & e^{i\phi_{11}} \end{pmatrix},$$

where $\phi_{ij} = \int dt \omega_{ij}(t)$ represents the dynamical phase accumulated by $|ij\rangle$ over the frequency excursion, with ω_{ij} being the frequency of this state. Using single-qubit phase (Z) rotations, this gate can be brought to the form of

$$U_{\text{CZ}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{i\phi_{2Q}} \end{pmatrix},$$

where $\phi_{2Q} = \phi_{11} - \phi_{01} - \phi_{10} + \phi_{00} = \int dt \zeta(t)$ is the acquired two-qubit conditional phase. Control over ϕ_{2Q} enables the realization of any controlled-phase gate. In particular, for $\phi_{2Q} = (2m+1)\pi$ this leads to a CZ gate (see Eq. (1.4)), with m being an integer. An accurate expression for ζ is given in Ref. [60].

Given that these gates rely on an interaction between the computational and higher-excited states, it is vital to consider the leakage induced when calibrating this gate. Perhaps the most successful approach for realizing this gate uses slower ("baseband") flux pulses that bring the higher-frequency qubit to the interaction point. These flux pulses can be adiabatic with respect to the avoided crossing, in which case the eigenstate $|11\rangle$ at the qubit sweet spot is allowed to slowly evolve into the instantaneous eigenstate $|\overline{11}\rangle = (|11\rangle + |02\rangle)/\sqrt{2}$ of the Hamiltonian at the avoided crossing. At the avoided crossing, the frequency of this state is shifted, allowing it to pick up the conditional phase. The flux pulse is then again slowly turned off, adiabatically returning the state $|\overline{11}\rangle$ back to $|11\rangle$. Such a trajectory aims to minimize the leakage to $|02\rangle$ and can be further optimized using a fast-adiabatic approach [100]. An alternative approach is to use a non-adiabatic flux pulse, which quickly brings $|11\rangle$ to the avoided crossing [101]. There, $|11\rangle$ is allowed to rotate to $|02\rangle$ and back, picking up a conditional phase of $\phi_{2Q} = \pi$ at the end of the full rotation. The goal, in this case, is to calibrate the gate duration such that the population returns to the computational subspace at the end of the gate. In practice, calibrations of this gate aim to minimize the gate duration and generally fall in between these two approaches. When using bi-polar flux pulses, leakage interference can also be exploited to minimize the leakage induced by this gate [66, 102]. In particular, Ref. [102] outlines the possible strategies to minimize leakage when using relatively fast flux pulses. In Chapter 2, we develop a parameterized error model for the leakage induced by this gate and further discuss the interactions between leaked and computational qubits.

Another challenge when performing such gates are the distortions of the flux pulse applied to the higher-frequency qubit. These can be partially resolved by using pre-distortion corrections and by using bi-polar flux pulses that exploit the symmetry around the frequency sweet spot [66, 102, 103]. Flux-pulsed CZ gates are routinely implemented in $20-60$ ns and can achieve high fidelities (around 99.9% and possibly higher) and low

1

leakage rates (roughly 0.1%) [43, 66, 102, 104]. However, in larger devices, the gate fidelities are often closer to 99% [20, 30, 32]. Apart from the possible crosstalk problems, the performance of these gates can be impacted by the presence of two-level-system defects that can become resonant with one of the qubits at some point during the frequency excursion, leading to lower fidelities and higher leakage [43].

When both qubits are at their sweet spots and sufficiently detuned from each other, the ZZ interaction strength is significantly suppressed. However, there is still some residual coupling left, with $\zeta/2\pi$ typically being ~ 100 kHz, but it can be as high as ~ 1 MHz [29]. This is commonly called the residual ZZ coupling or crosstalk. Note that this coupling is a function of the qubit frequencies. Pulsing either of the two transmon qubits away to avoid an interaction can lead to an increase or a decrease in this coupling. This residual ZZ coupling can lead to coherent and correlated ZZ errors during qubit idling and spectator errors [105, 106]. Replacing the coupling resonator with a tunable coupler [107–115] enables the suppression of the residual ZZ crosstalk, but this comes at the price of more complicated gate calibration and the introduction of additional control lines. In fact, when performing multiple operations simultaneously, it is challenging to cancel out the residual crosstalk fully. However, crosstalk errors can also be a significant error source, especially in the context of quantum error correction.

The controlled-phase gate is not the only two-qubit gate possible with the coupling considered in this section. For an overview of the possible two-qubit gates, we refer the reader to Ref. [71].

1.2.5. OTHER SUPERCONDUCTING QUBITS

While the transmon qubit is arguably the most successful superconducting qubit, many other qubits can be defined using superconducting circuits. One particularly promising candidate is the fluxonium qubit [116], which is designed to overcome some of the limitations of the transmon design. The circuit of the fluxonium is similar to the transmon, where in addition to the capacitance and Josephson junction in parallel, it also features a large shunting inductance (see Fig. 5.1 for a schematic of the circuit). This leads to the Hamiltonian

$$H = 4E_C q^2 - E_J \cos\left(\phi - \phi_e\right) + \frac{1}{2}E_L \phi^2, \tag{1.36}$$

where the symbols and operators were introduced when discussing the harmonic resonator and transmon Hamiltonians earlier in this chapter. The large shunting inductance leads to an insensitivity to charge noise without sacrificing the anharmonicity like the transmon qubit. These large inductances can be realized using an array of Josephson junctions [117, 118] or materials like granular aluminum [119] with high kinetic inductance. The large inductance allows fluxonium qubits to be highly anharmonic, which is one aspect in which this can be seen as an improvement over the transmon design. The large inductance also suppresses the sensitivity to flux noise, while a large capacitance further localizes the lowest energy states, increasing the coherence times. The fluxonium parameters are typically chosen such that $2 \leq E_J/E_C, E_J/E_L \leq 10$ [120]. This leads to a low qubit transition frequency, typically below 1 GHz, and to remarkably high coherence times, which have surpassed the millisecond barrier [121, 122]. However, this

comes with the price that operations are generally not as straightforward to implement. We discuss the fluxonium qubit in more detail in Chapter 5.

We refer the reader to Ref. [53] for a review of superconducting circuits and circuit quantum electrodynamics and Ref. [71] for a review of superconducting qubits and their operation.

1.3. QUANTUM ERROR CORRECTION

Quantum systems are inherently susceptible to noise due to the decoherence resulting from interactions with their environment or the finite fidelity with which qubit operations can be realized. Even the current state-of-the-art quantum processors typically exhibit error rates that are significantly higher than the low error rates necessary to perform practical quantum computation [10, 123]. Lowering the physical error rates has proven to be a formidable engineering challenge, and it is unlikely that an inherently protected qubit exhibiting sufficiently low error rates will emerge over the coming years.

Once the concept of quantum computation started to gain ground and the first potentially useful algorithms began to emerge, researchers soon realized that the fragility of quantum systems is a fundamental problem and that for a quantum computer to ever exist in practice, some robustness to noise will most likely be necessary. Similar concerns were also raised when classical processors were still being developed, which ultimately led to introducing the theory of (classical) error correction [124]. Error correction schemes generally revolve around introducing some redundancy and encoding the information to enable the detection and subsequent correction of the errors that have occurred.

Translating the concepts of error correction from the classical setting to the quantum one turned out to be non-trivial. This can be mainly attributed to the following reasons:

- The no-cloning theorem [125–127] states that it is impossible to copy the quantum information in the qubits. This contrasts with the classical case, where a bit can be copied when introducing redundancy.

- Qubit measurements lead to the collapse of the quantum state, meaning it is impossible to directly observe the state of a qubit encoding the quantum information. For comparison, the state of a classical bit can be measured.

- Quantum information is continuous. Classical bits of information can only take one of two discrete states for all intents and purposes, and an error typically flips the state of the bit. Qubits, on the other hand, use continuous phases and amplitudes to encode information. This also implies that small shifts in these continuous variables can accumulate over time and are not trivial to detect.

These obstacles made it seem that error correction would not be possible in the quantum case. This changed when researchers realized that quantum information could be redundantly encoded in highly entangled states involving many physical qubits. Furthermore, it was realized that it was possible to measure the parity of subsets of these without disturbing the encoded state. Importantly, these parity measurements can detect whether errors have occurred on the encoding qubits. These measurements also

1

lead to the discretization of continuous errors. This ultimately led to Peter Shor propos-ing the first QEC code [13].

1.3.1. STABILIZER CODES

Many error-correcting codes, including the code that Peter Shor first proposed [13], fall under the category of stabilizer codes [128]. The surface code [129–131], which is the focus of this dissertation, is another example of a stabilizer code. Generally, stabilizer codes are attractive due to the relative simplicity of their formalism, which makes it sim-pler to characterize their distance and define their logical operators.

An $[[n,k,d]]$ stabilizer code encodes k logical qubits into a system of n physical qubits in a 2^n-dimensional Hilbert space $\mathcal{H}$. These physical qubits are typically called the *data* qubits. The code is defined by a stabilizer group $\mathcal{S}$, which is an Abelian subgroup of the n-qubit Pauli $\mathcal{P}_n$ group such that $-I \notin \mathcal{S}$. The stabilizer group $\mathcal{S}$ can be charac-terized by $n-k$ independent generator $S_1, S_2, \ldots, S_{n-k}$, i.e. $\mathcal{S} = \langle S_1, S_2, \ldots, S_{n-k} \rangle$. These generators define a subspace $\mathcal{H}_L \subseteq \mathcal{H}$, referred to as the *code space* or the logical sub-space, which is spanned by the states $|\psi\rangle$ (also called *code words*) that are stabilized by the elements of $\mathcal{S}$, i.e.

$$\mathcal{H}_L = \left\{ |\psi\rangle \in \mathcal{H} \mid s|\psi\rangle = |\psi\rangle \, \forall s \in \mathcal{S} \right\}.$$

For simplicity, the generators are often just referred to as the stabilizers of the code. One can always find a pair of logical operators for each encoded logical qubit. These logical operators (in the Pauli group) commute with each stabilizer but are not in $\mathcal{S}$. In other words these operators belong to the centralizer of $\mathcal{S}$, defined as $\mathcal{C}(\mathcal{S}) = \{P \in \mathcal{P}_n \mid sP = Ps \forall s \in \mathcal{S}\}$. Therefore the set of logical operators is given by $C(\mathcal{S})/\mathcal{S}$. In this disserta-tion, we will commonly denote a logical operator as P_L, where P is some Pauli operator. The distance d of the code is then defined as $d = \min_{P_L \in \mathcal{C}(\mathcal{S})/\mathcal{S}} |P_L|$, i.e., the minimum weight of any logical operator. Since the logical operators commute with the stabilizers, it should be clear that an error that is in $\mathcal{C}(\mathcal{S})/\mathcal{S}$ (or in other words, it happens to be a logical operator) will not be detectable by the stabilizer checks. Therefore, the code distance d is related to the maximum weight of an error correctable by a given code. For a more thorough description of the stabilizer formalism, we refer the user to Ref. [132].

There is some freedom in choosing these generators, and they are typically chosen to make their measurement in experiments easier, which involves considering the weight of each of these generators and their locality given a specific device layout and connec-tivity. These measurements are usually done with the help of additional qubits, called the *ancilla* qubits that are generally interspersed with the data qubits encoding the log-ical information. The measurements are typically performed using several two-qubit gates to map the data-qubit parity to the state of the ancilla qubit, after which the ancilla qubit is measured and ideally reset to the ground state in preparation for the following round of error correction.

THE SURFACE CODE

One of the most promising and popular stabilizer codes is the surface code [130, 131], which is derived from the toric code that Alexei Kitaev initially proposed [129]. One of the main advantages of the surface code over other codes is that it can be implemented
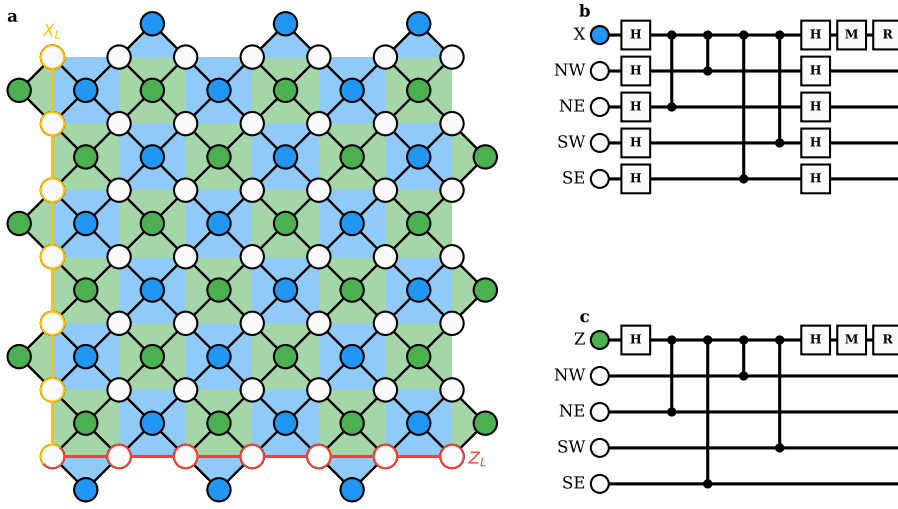
Figure 1.2: **a** Schematic depicting the fabric of the surface code, in this case, for a distance-seven surface code patch. The data qubits are shown as white circles. The ancilla qubits performing the X-type (blue plaquettes) and Z-type (green plaquettes) stabilizer measurements are shown as blue and green circles, respectively. Representatives of the X_L and Z_L logical operators are shown in yellow and red, respectively. The standard circuits measuring the X-type and Z-type stabilizers are shown in **b** and **c**, respectively. The data qubits are labeled according to their position relative to the ancilla qubit performing the measurement (NW=northwest, NE=northeast, SW=southwest, SE=southeast). The single-qubit operations labeled H, M, and R correspond to Hadamard gates, measurement, and qubit reset, respectively.

on a planar layout and requires only nearest-neighbor coupling between the physical qubits. This makes the surface code relatively straightforward to realize in an experiment. In addition to the surface code's modest requirements on the quantum hardware, the surface code is highly resilient to physical error, as exhibited by the high thresholds (defined in the following subsection) achieved for realistic error models [130, 133–135]. These characteristics make the surface code an ideal choice for near-term demonstrations of quantum memory experiments [20, 32]. The good logical performance that the surface code can achieve also makes it a promising candidate for future fault-tolerant architectures [130, 136].

The fabric of the rotated surface code is depicted in Fig. 1.2**a** and consists of data qubits and ancilla qubits arranged in a square 2D array. A distance d rotated surface code is composed of $n = d^2$ data qubits and $d^2 - 1$ ancilla qubits. The stabilizers of the surface code are typically defined as the X-type and Z-type plaquette operators of the form $XXXX$ and $ZZZZ$, respectively (or XX and ZZ at the boundary of the code). The X_L and Z_L logical operators of the surface code correspond to chain-like products of X and Z operators connecting the vertical or horizontal boundaries of the code, see Fig. 1.2**a** for examples of these logical operators. The X-type stabilizers are used to detect phase-flip (Z) errors, while the Z-type stabilizers are used to identify bit-flip (X) errors. Since

1

$Y \sim XZ$, these stabilizers can also detect Y errors occurring on the data qubits.

Typically, a dedicated ancilla qubit measures each stabilizer, commonly called the X-type and Z-type ancilla qubits. These ancilla qubits are positioned in the center of each plaquette and are, therefore, coupled to two or four data qubits each. The circuits used to perform these measurements are illustrated in Fig. 1.2**b,c**. These circuits use the controlled-phase gate as the two-qubit gate of choice, which is typical for flux-tunable transmon-qubit processors. The order of the two-qubit gates in each circuit is chosen to minimize the logical error rates [59, 137]. This order also allows for the stabilizers to be measured in parallel, reducing the overall time it takes to perform these measurements. However, executing these measurements in parallel is impossible in some architectures (namely, architectures not employing tunable couplers such as Ref. [20]) since performing the two-qubit gates in parallel also drives unwanted interactions between qubits [59]. In these cases, the two types of stabilizers can either be measured sequentially or in a *pipelined* fashion [59], where the measurement of the X-type stabilizers begins while the Z-type ancilla qubits are still being measured or vice versa. This pipelined approach uses the fact that qubit readout is typically much slower than performing single-qubit or two-qubit gates to reduce the duration of an error-correction cycle. Finally, while ancilla qubits are typically reset to the ground state after each measurement [47, 77–79], this is not required for measuring the stabilizers of the surface code, and many experiments that do not have access to a fast and high-fidelity reset do not include this operation [32].

The stabilizer measurement outcomes $m_{a,r}$ measured by ancilla qubit a at QEC round r are typically referred to as the syndromes. While these contain information about the errors that have occurred, they are generally harder for a decoder to process. For example, consider a single X error on one of the data qubits in the bulk of the code before several rounds of otherwise ideal stabilizer measurements are done. This error will anti-commute with the two neighboring Z-type stabilizers, flipping their measurement outcomes. However, unless corrected, this error will also flip all subsequent outcomes (in the absence of any other error). This means that a single error flips the outcome of multiple syndrome outcomes. For this reason, decoders typically consider the syndrome defects $d_{a,r} = m_{a,r} \oplus m_{a,r-1}$, where $\oplus$ is the binary addition operator. This can effectively be considered as taking the time derivative of the syndromes. The syndrome defects isolate an error's location in space and time, and in particular, a detectable error leads to one or more $d_{a,r} = 1$. For example, the X error that we previously considered will now only flip the syndrome defects measured by the two neighboring Z-type ancilla qubits in the following round. The syndrome outcomes required for calculating the defects at the first round of QEC can be inferred from the prepared initial state (or assumed to be trivial if, for example, they are not well defined). Finally, the above equations hold if the ancilla qubits are reset after each QEC round. However, if these reset operations are omitted from the circuit, the defects are instead calculated using $d_{a,r} = m_{a,r} \oplus m_{a,r-2}$. We note that the concept of the syndrome defects can be generalized to that of *detectors* and *detection events* [123]. In particular, in error-correction experiments, it is generally possible to identify sets of measurement outcomes, referred to as the detectors, that should ideally have a deterministic parity. Errors will then lead to observing a different parity than the expected one, referred to as a detection event. For more information, we refer the reader to Refs. [123, 138].

1.3.2. FAULT-TOLERANCE

QEC is currently the most promising approach to achieving fault-tolerant quantum computation by suppressing the logical error rates to a sufficiently low level. However, QEC might seem counter-intuitive at first. In particular, we have already emphasized that it introduces an enormous qubit overhead, depending on the code distance needed to achieve a sufficiently low error rate. However, repeated stabilizer measurements also introduce many single-qubit gates, two-qubit gates, and measurements. Each of these physical operations is itself noisy and can induce errors, which can propagate through the circuit and either directly lead to a logical error or to a detectable one that could become a logical error when incorrectly decoded. Therefore, it is not immediately apparent whether measuring the stabilizers will ultimately allow for removing more errors than it introduces. Similar considerations should also be made for the logical operations performed during the computation, which are much more involved than any single physical operation.

The theory of fault tolerance [132, 139] considers this problem and provides design principles for constructing circuits that can tolerate the errors induced by the noisy operations employed in these circuits. Of course, whether a circuit is fault tolerant is a function of the gates and the order in which they are performed, the code, the noise model considered, and the decoder. Because of this, defining if an operation is fault-tolerant is often challenging. Instead, fault-tolerance principles should be considered a design philosophy when constructing circuits or algorithms. In particular, one should ensure that single faults do not lead to logical errors (especially for codes with $d > 3$), as that would mean the logical error rate will be comparable to the physical error rate. More generally, circuits should be designed to minimize their logical error rate. This typically involves considering each fault location in the circuit and ensuring that as many faults as possible are correctable. In an analogy to the code distance, the circuit distance d_c refers to the minimum number of faults that lead to an undetectable logical error [140]. By definition, $d_c \le d$ and the circuit should ideally be distance-preserving, such that $d_c = d$.

Transversality becomes a useful tool for designing such circuits when considering logical operations. A transversal logical single-qubit gate is realized by performing a physical operation on each data qubit (or a subset of these qubits) in parallel. Similarly, a transversal logical two-qubit gate is realized by doing two-qubit gates between pairs of data qubits (that are in separate code blocks) in parallel. Transversal gates are typically much easier to implement in practice. A fault in such a gate is guaranteed to propagate to, at most, a single error in each code block, meaning that errors remain localized. Since the physical gates are ideally executed in parallel, performing transversal gates takes as long as doing a single physical operation. Therefore, transversal gates are generally expected to achieve lower logical error rates than non-transversal ones. Unfortunately, a no-go theorem states that quantum codes that can detect any single-qubit error do not allow for a universal set of transversal gates [141]. It was also shown in Ref [142] that for 2D stabilizer codes, only Clifford gates can be performed in constant depth using only local gates.

So far, we considered the number of faults in a circuit that can lead to a logical error. However, it is also essential to ensure that the logical error rate is suppressed when increasing the code distance d (or equivalent, the number of encoding qubits n). When

the physical error rates are high, we expect that the operations introduced by QEC will induce more errors than they can correct, such that increasing the distance d will likely increase the logical error rate. On the other hand, when the physical error rates are low, and error events are sparse, we expect to be able to correct more errors than those introduced by the stabilizer measurements, such that the logical error rate goes down with the code distance. The threshold is the physical error rate corresponding to the crossover between these two regimes. Note that this threshold is a property of the noise model, the code, and the decoder.

To simplify the discussion, let us assume that each fault occurs with the same probability p and let the threshold error probability be p_{th}. We will also consider the case where we store the logical information and perform repeated rounds of QEC. Let $\varepsilon_L(d)$ be the logical error probability per round of QEC for a fixed p. Then, for $p \ll p_{th}$ it should hold that

$$\lim_{d \to \infty} \varepsilon_L(d) = 0.$$

Note that for p near the threshold, there might be an initial suppression of the logical error rate for some lower code distances, followed by an increase for larger distances due to finite-size effects [20]. In the regime of asymptotically low physical error rates, we only need to consider the logical error rates due to the minimum-weight error chains. For a surface code of an odd distance d, these are chains of weight $(d+1)/2$. In this regime, the logical error rate is expected to scale with p according to [130]

$$\varepsilon_L(d) \sim A(d) \, p^{(d+1)/2}, \tag{1.37}$$

where $A(d)$ is the entropic contribution to the error rate, i.e., it corresponds to the number of low-weight error chains that lead to a logical error. This entropic factor is an important contribution to the error rate. For example, the rotated version of the surface code uses fewer physical qubits than the unrotated one for the same distance d. Naively, one might assume that the rotated code will have the same or lower logical error rate. However, it has been shown that the rotated code has a much larger entropic contribution to the error rate, which can result in it performing worse than the unrotated version for certain error rates and code distances [143].

Thresholds are not an ideal metric for comparing the performance of different codes, as they not only depend on the decoder but can also have a strong dependence on the assumptions made in the error model. Furthermore, the threshold is hard to estimate experimentally, as it would require varying the physical error rates (which are all different in the first place). Instead, the ability to suppress the logical error rate is typically characterized by increasing the code distance and performing a fit to

$$\varepsilon_L(d) = C/\Lambda^{(d+1)/2}, \tag{1.38}$$

where Λ is the error-suppression factor and C is some fitting constant [19, 20, 30]. The factor Λ attempts to capture how far below the threshold you are, with $\Lambda \approx 1$ corresponding to being near the threshold. It should also be clear that Λ is a function of the physical error rates and circuit used. The goal of many current experiments is to demonstrate $\Lambda \gg 1$, which is desirable as a higher value of Λ will reduce the qubit overhead required

to reach sufficiently low logical error rates. Finally, note that while the suppression of ε_L is typically expressed in terms of the code distance d, a more suitable choice is the circuit distance d_c [140].

1.3.3. ERROR CORRECTION IN PRACTICE

Realizing fault tolerance in practice is considerably more challenging. While circuit-level depolarizing noise models are fairly realistic, the physical noise is often much more complex. For example, increasing the scale of a processor may lead to a reduction in the coherence times [144]. Performing more operations in parallel may also increase the physical error rates due to the crosstalk [20, 43]. Although there have been some promising demonstrations of logical error suppression [20, 30], unfortunately, there is no guarantee that QEC will be sufficient to enable fault-tolerant computation, even if each operation is designed to be theoretically fault-tolerant. Therefore, it is vital to follow the principles behind fault tolerance and to characterize the limiting factors behind the achieved logical error rates in the experiment.

When describing the formalism of stabilizer codes, we assumed that each physical qubit is a two-level system. Since transmon qubits are weakly anharmonic, operations can induce leakage outside the computational subspace. Leakage is an example of a non-conventional error that falls outside of stabilizer formalism, meaning that a stabilizer code is not guaranteed to detect and correct such errors. This makes leakage a potentially very threatening error that can increase the logical error rates or possibly even prevent the suppression of the logical error rate by increasing the code distance [45]. In Chapter 2, we develop a realistic model for leakage in flux-tunable transmon-qubit processors and investigate the impact of leakage on the logical performance of a surface code. We observe that leakage can lead to significantly higher error rates. In Chapters 3 and 4, we propose and experimentally implement operations to remove leakage outside the computational subspace. We expect these operations can lead to lower logical error rates and ensure the ability of these codes to suppress the logical error rate to sufficiently low levels.

1.3.4. BEYOND THE SURFACE CODE

The performance of the surface code has been extensively explored for depolarizing noise models, where each Pauli error occurs with the same probability. However, in many experimental platforms, the noise experienced by the qubits is instead biased toward one type of error. For example, flux-tunable transmon qubits typically exhibit approximately the same relaxation and dephasing rates when operated at the sweet spot, where the qubit is insensitive to flux noise. However, transmon qubits may be fluxed away to interact with other qubits [60] or avoid any unwanted interactions [59], increasing the dephasing rate experienced by the qubit.

Several modifications to the standard surface code to deal with biased noise have been proposed in the past [145–147]. For physical error rates that are well below the threshold, rectangular surface codes with optimized aspect ratios have also been shown to achieve good logical performance in the presence of biased noise [148]. The $XZZX$ surface code has been shown to have remarkably high thresholds under biased noise models [149–151]. The $XZZX$ surface code is a variant of the surface code, where the

1

stabilizers of the code are of the form $XZZX$ instead of the standard $XXXX$ or $ZZZZ$ stabilizers used in the conventional surface code. The excellent performance of the $XZZX$ surface code has motivated proposals for architectures that are engineered to exhibit a high noise bias [150, 152–154] *and* that have access to a bias-preserving CNOT gate [152], allowing them to maintain the noise bias during the measurement of the stabilizers.

While the surface code is generally promising, it has several downsides that might be particularly relevant when considering future fault-tolerant architectures. This code has a low encoding rate, requiring significantly more physical qubits for each logical one, especially compared to certain quantum low-density parity check (LDPC) codes [140, 155–158]. Therefore, architectures based on the surface code will need a significant qubit overhead to realize. Engineering such an extensive system can be challenging, as it might require a very high fabrication accuracy, developing dilution refrigerators that can accommodate these large processors, and possibly interconnecting multiple such refrigerators.

It is also essential to consider the ease with which logical operations can be performed using a given code. The surface code allows X, Z, and CNOT logical gates to be performed transversally [159]. Hadamard gates are nearly transversal but require a subsequent rotation of the code patch. However, other codes, such as two-dimensional color code [160, 161], allow all logical Clifford operations to be done transversally [162, 163]. Therefore, it is possible for the surface code to ultimately perform worse than other codes when performing a computation despite them having lower thresholds than the surface code [164].

1.4. DECODERS

The measured syndromes, and by extension, the syndrome defects, contain information about the errors that have occurred during the computation. It is then the task of a decoder to process this information to infer the most probable correction, hopefully restoring the encoded logical information. Most decoders also require information about the physical error rates to determine the most likely correction or error consistent with the measured syndromes.

One of the essential characteristics of a decoder is the accuracy with which it can predict the necessary corrections, as this directly impacts the logical performance of the code. A more accurate decoder results in higher thresholds, reducing the qubit overhead required to reach a sufficiently low logical error rate. However, the speed with which the decoder can process the syndromes is also essential for ensuring scalability. In particular, a decoder must be able to decode the syndromes in *real-time*, i.e., the rate with which the decoder processes the syndromes should be equal to or lower than the rate with which they are generated. Otherwise, the decoding may lead to an exponential slowdown of the computation due to an increasing backlog volume of syndromes that need to be processed, more commonly known as the *backlog problem* [132, 165, 166]. How fast a decoder needs to be depends on the QEC cycle duration, which can differ significantly depending on the physical platform. For transmon-qubit processors, a round of QEC can typically be performed in ∼ 1 μs [165]. This also happens to be the tightest constraint imposed by any platform, leading to this time being often used when bench-

marking these decoders. Unfortunately, there is a general trade-off between the speed and the accuracy of the decoder. Therefore, the goal is to design a decoder that is as accurate as possible while still being fast enough for real-time decoding.

In practice, the reliance of a decoder on an error model can also be a practical problem. Constructing such a model in the first place requires an accurate characterization of the operational error rates, typically using tomographic reconstruction (e.g., process [167, 168] or gate-set [169–171] tomography) or partial characterization protocols (e.g., randomized benchmarking [172–178]). However, each of these techniques has certain downsides. For example, tomographic protocols struggle to scale past a few qubits and might fail to capture the impact of crosstalk when performing operations in parallel. While randomized benchmarking protocols are more scalable, they do not provide detailed information about the noise. Therefore, it is desirable for the decoder or some protocol to infer the physical noise from the measured syndromes directly [30, 179–185]. However, these models still rely on certain assumptions and might not capture certain physical errors, such as leakage. Unfortunately, the error rates fluctuate over time in real physical systems [43, 57], which can further impact the accuracy of the decoder. Hence, another desirable feature is the ability to adapt to such fluctuations as the computation is being performed, the feasibility of which naturally depends on the timescale of these fluctuations [180].

1.4.1. MAXIMUM-LIKELIHOOD DECODERS

A maximum-likelihood (ML) decoder finds the most likely correction given the observed syndromes and a physical error model (typically only considering Pauli errors), making it an optimal decoder by definition. In particular, this decoder considers all possible errors consistent with the observed syndrome and divides them into several sets depending on whether they correspond to a certain logical error. It then calculates the probability of each set occurring and chooses the correction corresponding to the most probable one. In particular, let us consider a stabilizer code that encodes a single logical qubit in n data qubits. For any Pauli operator $P \in \mathscr{P}_n$, we can define the (left) coset of $\mathscr{S}$ that contains P, that is $P\mathscr{S} \equiv \{Ps : s \in \mathscr{S}\}$. Let $\mathbf{s}$ be the vector of observed syndrome outcomes and let E be some representative Pauli operator that is consistent with $\mathbf{s}$. The set of all Pauli operators consistent with this syndrome is then given by the coset $E\mathscr{C}(\mathscr{S})$, which can further be partitioned as [186]

$$E\mathscr{C}(\mathscr{S}) = C_I \cup C_X \cup C_Y \cup C_Z,$$

where $C_P = EP_L\mathscr{S}$ with P_L being one of the logical operators of the code. The optimal correction is, therefore, any Pauli operator belonging to the most likely coset of this partition. This can be expressed as

$$C^* = \underset{C \in \{C_I, C_X, C_Y, C_Z\}}{\operatorname{argmax}} P(C),$$

where C^* is the most likely coset, with the most likely correction being any $c \in C^*$. The probability of the coset C_P is given by $P(C_P) = \sum_{s \in \mathscr{S}} P(EP_L s)$, with $P \in \{I, X, Y, Z\}$. The ML decoder achieves higher accuracy than decoders that find the most likely error given the observed syndrome because it considers the degeneracy of errors. In particular, it

is possible that the most probable coset does not correspond to the one containing the most probable error.

Unfortunately, the general problem of ML decoder is inefficient to solve and has been proven to be #P-complete [187]. However, certain codes under some specific noise models allow for an efficient implementation of ML decoding [186]. For more general noise models, tensor network (TN) contraction can be used to solve an approximation of ML decoding, resulting in a nearly optimal decoder [20, 148, 186]. In particular, TN decoders are most commonly seen and often referred to as ML decoders instead. While technically more computationally efficient than general ML decoding, TN decoders are still very computationally expensive, making them too slow for real-time decoding. As a matter of fact, TN decoders are so computationally expensive that they are typically only considered for small-distance to medium-distance codes [20, 148].

1.4.2. PERFECT-MATCHING DECODERS

The minimum-weight perfect matching (MWPM) decoder [131, 188] is perhaps the most popular decoding algorithm for the surface code. Historically, some of the first explorations of the performance of this code considered independent bit-flip and phase-flip error models [131], where these errors could occur on each qubit before an otherwise ideal round of stabilizer measurement. Under such a model, each error leads to at most two syndrome defects measured either by the X-type or the Z-type stabilizers. This led to the simple and intuitive idea of pairwise matching these defects based on the likelihood of the shortest error chain leading to them. In other words, finding the most likely pairing is equivalent to finding the most likely errors consistent with the measured syndromes, which can be expressed as:

$$E^* = \operatorname*{argmax}_{E \in \mathscr{P}_n} P(E \mid \mathbf{s}).$$

Unlike the ML decoder, MWPM does not consider any error degeneracy. Errors on any data qubits on the code's boundary (see Fig. 1.2) can lead to a single syndrome defect. While this might first seem incompatible with pairwise matching, MWPM decodes can deal with such an error by introducing a virtual defect when solving the matching problem.

The problem of minimum-weight perfect matching is well-known in graph theory and is solvable in polynomial time using the Blossom algorithm [189], making the MWPM a theoretically fast and efficient decoder. In particular, the worst-case complexity of the MWPM decoder is $O(N^3 \log(N))$, where N is the number of nodes in the matching graph, corresponding to the number of observed non-trivial syndrome defects (and potentially a virtual boundary defect) [190]. Unfortunately, this complexity can potentially make the decoder too slow for real-time decoding, especially in the case of superconducting-qubit processors and when decoding large-distance surface codes. However, several implementations of this algorithm have been shown to run with close to linear complexity and possibly in constant average time [190–193].

In addition to the computational efficiency it boats, MWPM decoders lead to generally good logical performance. In the case of independent bit-flip and phase-flip errors, the code-capacity threshold was found to be very high (approximately 10.3% [133]), approaching the threshold obtained using an ML decoder (about 10.9% [131]). For more

realistic error models, such as phenomenological or circuit-level noise models considering depolarizing errors, MWPM decoders continue to demonstrate good logical performance [130, 133–135] relative to other decoders. Note that some of the results above are for the unrotated surface code, which performs similarly to the rotated version. Ultimately, the exact values of the thresholds are not too important since they ultimately depend on the error model considered. Importantly, the MWPM decoder achieves thresholds in the range $0.5 - 1.1$% for circuit-level noise models [135], which are values that seem feasible to achieve in experiments.

Despite their popularity, MWPM decoders have several downsides that ultimately limit the accuracy that they can achieve. In particular, this decoder does not consider the degeneracy of errors. Most implementations of MWPM also assume that the bit-flip and phase-flip errors are uncorrelated. That is, if X and Z errors are happening with a probability of $O(p)$ each, it is assumed that Y errors are happening with a probability of $O(p^2)$. However, this isn't the case when considering the depolarizing noise model. More importantly, the noise observed in physical transmon-qubit processors also does not follow this assumption. Ultimately, this contributes to MWPM decoders being less accurate than ML decoders or other decoding algorithms that consider Y errors [20, 148, 186, 194, 195], which we further discuss in Chapter 8.

Another practical downside of MWPM is that it requires a physical error model, specifically the probability of each independent error mechanism leading to one or two syndrome defects. While this can be extracted from a physical noise model, obtaining such a model for experimental devices is not a trivial task (see the discussion in Sec. 1.4). There are algorithms for estimating these probabilities directly from the measured syndrome defects [30, 180, 184], bypassing the need for benchmarking operations. However, these algorithms are also based on some assumptions about the errors. In particular, errors such as crosstalk or leakage that can lead to multiple non-trivial syndrome defects violate the typical assumptions behind these algorithms and potentially lead to inaccurate estimates [184]. We discuss this in greater detail in Chapter 7.

1.4.3. BEYOND PERFECT-MATCHING DECODERS

Naturally, many other decoding algorithms have been proposed that improve the speed or accuracy achieved by MWPM decoders. The Union-find (UF) decoder [196, 197] is an example of another computationally efficient and fast algorithm that outperforms MWPM in terms of speed. In particular, the UF decoder has a worst-case runtime that is nearly linear in the number of syndrome defects [196, 197]. However, UF decoders lead to slightly worse logical performance, such as the lower thresholds achieved under independent bit-flip and phase-flip noise (around 9.9% for the surface code [196]). Efficient hardware architectures for this decoder based on Field Programmable Gate Arrays (FPGAs) or Application Specific Integrated Circuits (ASICs) have been either proposed [198] or implemented [199]. Other clustering-based decoders similar to the UF decoder have also been realized and implemented on FPGA and ASIC [200]. In either case, these implementations have demonstrated runtimes enabling real-time decoding in future experiments (up to considerable code distances).

As previously mentioned, one of the more significant issues with the standard MWPM decoder is its inability to deal with Y errors. In particular, Y errors lead to three or four

1

non-trivial syndrome defects, which are, therefore, incompatible with pairwise matching. More generally, X or Z errors can be represented as edges in a graphical representation of the physical error model. On the other hand, Y errors correspond to hyperedges instead, transforming the graph into a hypergraph. Unfortunately, the MWPM algorithm can only be executed on graphs, forcing it to ignore the correlations between the X-type and the Z-type syndrome defects that Y errors lead to, ultimately leading to an increase in the logical error rate [148].

There have been modifications to MWPM based on a two-pass correlation strategy to better deal with these errors at the expense of increased decoder runtime [20, 194]. However, several other decoding algorithms have also been shown to achieve higher accuracy than the MWPM decoder while still being efficient enough to be considered a possible candidate for real-time decoding. For example, decoders based on *belief-propagation* (BP) have shown that they can achieve good logical performance when applied to the surface code [20, 148, 201]. These decoders use an iterative *message-passing* algorithm to estimate the marginal probability of each error mechanism occurring based on the observed syndrome defects and a hypergraph, where each hyperedge represents an error mechanism. The estimated marginal probabilities are then used to infer the most likely error that has occurred. Unfortunately, the BP algorithm is not guaranteed to converge to a solution when applied to the quantum setting due to the degeneracy of the errors [201]. Instead, these decoders employ a secondary algorithm that uses the estimated marginal error probabilities to find the solution whenever BP fails to converge. One notable example of such an algorithm is ordered-statistics decoding (OSD), leading to the BP-OSD decoder [140, 156, 202]. Alternatively, the marginal probabilities can be translated to a set of edge weights, after which MWPM can be used to find the most likely errors, leading to the so-called belief-matching decoder [20, 148, 201, 203]. Neural network (NN) decoders have also achieved significantly higher accuracy than MWPM [195, 204]. These decoders can be trained directly on the measured syndromes, bypassing the need to estimate the physical error rate accurately. We discuss such a decoder in Chapter 8 and note that, in some cases, these decoders can demonstrate an accuracy matching or even exceeding that of an ML decoder [195].

1.4.4. DECODER EXTENSIONS

In the discussion about the decoders so far, we have assumed that the stabilizer measurement outcomes and, by extension, the syndrome defects are binary. In practice, measurements of transmon qubits lead to continuous analog outcomes [53, 71]. It has been shown that the information available in these analog outcomes, commonly referred to as the *soft information*, can be useful to a decoder and lead to lower logical error rates [205, 206]. In particular, there have been extensions of the MWPM decoder [206] and BP decoder [207] using the analog outcomes. In Chapter 8, we incorporate the analog outcome into an ML decoder, again demonstrating an improvement in the logical performance. Ref. [195] also provided the soft information to a transformer-based NN decoder, again showing decreased logical error rates. In addition to the analog outcomes, information about non-conventional errors such as leakage can also improve the accuracy of the decoder. Ref. [195] also demonstrated that providing information about the qubit leakage to their ML decoder improves the logical performance of the

code. Therefore, extending other decoders to utilize this information is an open research question.

1.5. DISSERTATION OUTLINE

In this dissertation, we focus on the challenges encountered when implementing error-correcting codes in practice and how logical performance is affected by the noise encountered in transmon-qubit processors. We focus on the surface code, the most promising approach to achieving fault tolerance at the time of writing. Though we consider several different error sources, we focus on the impact of non-conventional errors, namely on leakage outside of the computational subspace. By performing simulations using realistic error models or directly analyzing experimental data, we characterize the impact and signatures of leakage in the surface code and develop schemes to detect and remove this error. In addition to mitigating leakage, we investigate more accurate decoding algorithms and heterogenous qubit architectures that can reduce the frequency crowding problem.

In Chapter 2, we develop a physically realistic leakage error model for flux-tunable transmon qubits and use density-matrix simulations to study how leakage behaves in a distance-three surface code. We observe that leakage exhibits a stochastic behavior and that a leaked qubit leads to an increase in the number of non-trivial syndrome defects observed by the neighboring stabilizers. This can be interpreted as a signature of leakage, enabling the indirect detection by a set of computationally efficient Hidden Markov models. We show that these models can detect the location and time of leakage events with considerable accuracy, allowing us to post-select out the runs where leakage occurred to restore the logical performance of the code.

While post-selection can be useful in some near-term experiments, leakage-reduction units that bring leaked qubits back to the computational subspace are required to deal with leakage in a scalable way. In Chapter 3, we present a leakage-reduction scheme for the surface code that does not introduce a significant in either hardware or the time required to perform a round of error correction. For data qubits, we propose using a microwave pulse to transfer the population in the second-excited state to a readout resonator, which then quickly decays back to the ground state. We use another microwave pulse that transfers the population between the second- and first-excited states of the ancilla qubits, which we only apply when the qubit is measured to be in the second-excited state. We show that these leakage-reduction units are effective in removing leakage and can reduce the impact of this error on the logical error rate of the code.

In Chapter 4, we realize the data-qubit leakage-reduction unit proposed in Chapter 3 in an experiment, where we demonstrate that this operation can be implemented to remove most of the leaked population while having a minimal impact on the computational subspace. Furthermore, we demonstrate that this operation can be extended to remove the population in the third-excited state and that it can also be applied to the ancilla qubits. We show that incorporating these operations in a repeated weight-2 stabilizer measurement experiment suppresses the build-up of leakage and the associated signatures observed in the measurement outcomes.

We next take a closer look at another superconducting qubit, the fluxonium. Chapter 5 explores how two-qubit gates between a transmon qubit and a fluxonium qubit can

be realized. In particular, we propose a microwave-activated cross-resonance gate and a controlled-phase gate between these qubits that rely on interactions with the higher-excited states. By performing density-matrix simulations, we show that these gates can be implemented with high fidelity and low leakage in a few hundred nanoseconds. The controlled-phase gate is optimal for low-frequency fluxonium qubits, while the cross-resonance gate achieves better performance over a large range of medium-range frequencies. Finally, we show that an architecture employing both types of qubits and using the cross-resonance gate as the two-qubit gate of choice mitigates the problem of frequency crowding and can have significantly higher fabrication yield compared to architectures using only fixed-frequency transmon qubits and employing the same type of gate.

Next, we consider some error-correction experiments implementing small-distance codes. In Chapter 6, we realize a distance-two surface code and perform repeated rounds of error detection to post-select out the detectable errors. We implement a suite of logical operations, including arbitrary state initialization, measurement, and a universal single-qubit gate set. We observe that the fault-tolerant variants of these operations achieve a lower logical error rate than non-fault-tolerant ones. In addition, we explore the impact of several error sources on the observed logical performance using density-matrix simulations.

In Chapter 7, we implement a distance-seven quantum repetition code and demonstrate the ability to suppress the logical error rate of the code when increasing the code distance. However, we observe that going from distance three to distance seven leads to only a slight decrease in the logical error rate. We study the reasons behind this performance by performing circuit-level noise simulations of this experiment and by analyzing the correlations between the observed stabilizer measurement outcomes.

Finally, we consider improving the logical performance of the surface code by using a decoder that can more accurately infer the errors that have occurred given the observed syndrome defects. Neural network decoders are particularly promising as they have been shown to achieve good logical performance without requiring any prior information about the physical error rates. In Chapter 8, we explore the performance of such a decoder on both simulated data using circuit-level noise models and experimental data from recent surface-code experiments. We demonstrate that this decoder can achieve logical error rates approaching those possible with approximate maximum-likelihood decoders. We furthermore show that incorporating the additional information available from the analog measurement outcomes obtained in practice when reading out transmon qubits to the decoder allows it to achieve even lower logical error rates.

We conclude this dissertation by providing a summary of each chapter and providing an outlook on some challenges that can significantly impact the logical performance of error-correcting codes in near-term and long-term experimental realization, or that can help reduce the qubit overhead required to reach the low logical error rates required for performing useful computation.

BIBLIOGRAPHY

[1] R. P. Feynman, *Simulating physics with computers*, International Journal of Theoretical Physics **21**, 467–488 (1982).

[2] D. Deutsch and R. Jozsa, *Rapid solution of problems by quantum computation*, Proc. R. Soc. Lond. A **439**, 553–558 (1992).

[3] P. W. Shor, *Algorithms for quantum computation: discrete logarithms and factoring*, Proceedings 35th Annual Symposium on Foundations of Computer Science, 124–134 (1994).

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

[5] R. L. Rivest, A. Shamir, and L. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Commun. ACM **21**, 120–126 (1978).

[6] C. H. Bennett, F. Bessette, G. Brassard, L. Salvail, and J. Smolin, *Experimental quantum cryptography*, Journal of Cryptology **5**, 3–28 (1992).

[7] C. H. Bennett and G. Brassard, *Quantum cryptography: public key distribution and coin tossing*, Theoretical Computer Science **560**, Theoretical Aspects of Quantum Cryptography – celebrating 30 years of BB84, 7–11 (2014).

[8] S. Lloyd, *Universal quantum simulators*, Science **273**, 1073 (1996).

[9] D. Wecker, M. B. Hastings, N. Wiebe, B. K. Clark, C. Nayak, and M. Troyer, *Solving strongly correlated electron models on a quantum computer*, Phys. Rev. A **92**, 062318 (2015).

[10] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, *Elucidating reaction mechanisms on quantum computers*, Proceedings of the National Academy of Sciences **114**, 7555–7560 (2017).

[11] Y. Cao, J. Romero, J. Olson, M. Degroote, P. Johnson, M. Kieferová, I. Kivlichan, T. Menke, B. Peropadre, N. Sawaya, S. Sim, L. Vies, and A. Aspuru-Guzik, *Quantum chemistry in the age of quantum computing*, ArXiv:1812.09976 (2018).

[12] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, *Quantum computational chemistry*, Rev. Mod. Phys. **92**, 015003 (2020).

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

[14] P. Shor, *Fault-tolerant quantum computation*, in Proceedings of 37th conference on foundations of computer science (1996), pp. 56–65.

[15] A. Y. Kitaev, *Quantum computations: algorithms and error correction*, Russian Mathematical Surveys **52**, 1191–1249 (1997).

1

[16] E. Knill, R. Laflamme, and W. H. Zurek, *Resilient quantum computation*, Science **279**, 342–345 (1998).

[17] D. Gottesman, *Theory of fault-tolerant quantum computation*, Phys. Rev. A **57**, 127–137 (1998).

[18] D. Aharonov and M. Ben-Or, *Fault-tolerant quantum computation with constant error rate*, SIAM Journal on Computing **38**, 1207–1282 (2008).

[19] 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**, 66–69 (2015).

[20] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush, D. Bacon, J. C. Bardin, J. Basso, A. Bengtsson, S. Boixo, G. Bortoli, A. Bourassa, J. Bovaird, L. Brill, M. Broughton, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, J. Cogan, R. Collins, P. Conner, W. Courtney, A. L. Crook, B. Curtin, D. M. Debroy, A. Del Toro Barba, S. Demura, A. Dunsworth, D. Eppens, C. Erickson, L. Faoro, E. Farhi, R. Fatemi, L. Flores Burgos, E. Forati, A. G. Fowler, B. Foxen, W. Giang, C. Gidney, D. Gilboa, M. Giustina, A. Grajales Dau, J. A. Gross, S. Habegger, M. C. Hamilton, M. P. Harrigan, S. D. Harrington, O. Higgott, J. Hilton, M. Hoffmann, S. Hong, T. Huang, A. Huff, W. J. Huggins, L. B. Ioffe, S. V. Isakov, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, P. Juhas, D. Kafri, K. Kechedzhi, J. Kelly, T. Khattar, M. Khezri, M. Kieferová, S. Kim, A. Kitaev, P. V. Klimov, A. R. Klots, A. N. Korotkov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, K.-M. Lau, L. Laws, J. Lee, K. Lee, B. J. Lester, A. Lill, W. Liu, A. Locharla, E. Lucero, F. D. Malone, J. Marshall, O. Martin, J. R. McClean, T. McCourt, M. McEwen, A. Megrant, B. Meurer Costa, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, A. Morvan, E. Mount, W. Mruczkiewicz, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, H. Neven, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, M. Y. Niu, T. E. O'Brien, A. Opremcak, J. Platt, A. Petukhov, R. Potter, L. P. Pryadko, C. Quintana, P. Roushan, N. C. Rubin, N. Saei, D. Sank, K. Sankaragomathi, K. J. Satzinger, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, V. Smelyanskiy, W. C. Smith, G. Sterling, D. Strain, M. Szalay, A. Torres, G. Vidal, B. Villalonga, C. Vollgraff Heidweiller, T. White, C. Xing, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, Y. Zhang, N. Zhu, and G. Q. AI, *Suppressing quantum errors by scaling a surface code logical qubit*, Nature **614**, 676–681 (2023).

[21] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, *Demonstrating multi-round subsystem quantum error correction using matching and maximum likelihood decoders*, Nature Communications **14**, 2852 (2023).

[22] N. Ofek, A. Petrenko, R. Heeres, P. Reinhold, Z. Leghtas, B. Vlastakis, Y. Liu, L. Frunzio, S. M. Girvin, L. Jiang, M. Mirrahimi, M. H. Devoret, and R. J. Schoelkopf, *Extending the lifetime of a quantum bit with error correction in superconducting circuits*, Nature **536**, 441 (2016).

1

[23] A. Grimm, N. E. Frattini, S. Puri, S. O. Mundhada, S. Touzard, M. Mirrahimi, S. M. Girvin, S. Shankar, and M. H. Devoret, *Stabilization and operation of a kerr-cat qubit*, Nature **584**, 205–209 (2020).

[24] P. Campagne-Ibarcq, A. Eickbusch, S. Touzard, E. Zalys-Geller, N. E. Frattini, V. V. Sivak, P. Reinhold, S. Puri, S. Shankar, R. J. Schoelkopf, L. Frunzio, M. Mirrahimi, and M. H. Devoret, *Quantum error correction of a qubit encoded in grid states of an oscillator*, Nature **584**, 368–372 (2020).

[25] V. V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. L. Brock, A. Z. Ding, L. Frunzio, S. M. Girvin, R. J. Schoelkopf, and M. H. Devoret, *Real-time quantum error correction beyond break-even*, Nature **616**, 50–55 (2023).

[26] 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 control of an error-corrected qubit*, Nature **598**, 281–286 (2021).

[27] M. H. Abobeih, Y. Wang, J. Randall, S. J. H. Loenen, C. E. Bradley, M. Markham, D. J. Twitchen, B. M. Terhal, and T. H. Taminiau, *Fault-tolerant operation of a logical qubit in a diamond quantum processor*, Nature **606**, 884–889 (2022).

[28] C. Ryan-Anderson, J. G. Bohnet, K. Lee, D. Gresh, A. Hankin, J. P. Gaebler, D. Francois, A. Chernoguzov, D. Lucchetti, N. C. Brown, T. M. Gatterman, S. K. Halit, K. Gilmore, J. A. Gerber, B. Neyenhuis, D. Hayes, and R. P. Stutz, *Realization of real-time fault-tolerant quantum error correction*, Phys. Rev. X **11**, 041058 (2021).

[29] 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*, Nat. Phys. **18**, 80–86 (2022).

[30] 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**, 383–387 (2021).

[31] 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**, 875–880 (2020).

1

[32] S. Krinner, N. Lacroix, A. Remm, A. Di Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, *Realizing repeated quantum error correction in a distance-three surface code*, Nature **605,** 669–674 (2022).

[33] Y. Zhao, Y. Ye, H.-L. Huang, Y. Zhang, D. Wu, H. Guan, Q. Zhu, Z. Wei, T. He, S. Cao, F. Chen, T.-H. Chung, H. Deng, D. Fan, M. Gong, C. Guo, S. Guo, L. Han, N. Li, S. Li, Y. Li, F. Liang, J. Lin, H. Qian, H. Rong, H. Su, L. Sun, S. Wang, Y. Wu, Y. Xu, C. Ying, J. Yu, C. Zha, K. Zhang, Y.-H. Huo, C.-Y. Lu, C.-Z. Peng, X. Zhu, and J.-W. Pan, *Realization of an error-correcting surface code with superconducting qubits*, Phys. Rev. Lett. **129**, 030501 (2022).

[34] 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).

[35] 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).

[36] M. Kjaergaard, M. Schwartz, J. Braumüller, P. Krantz, J. Wang, S. Gustavsson, and W. Oliver, *Superconducting qubits: current state of play*, Annu. Rev. Condens. Matter Phys. **11**, 369–395 (2020).

[37] J. M. Kreikebaum, K. P. O'Brien, A. Morvan, and I. Siddiqi, *Improving wafer-scale josephson junction resistance variation in superconducting quantum coherent circuits*, Superconductor Science and Technology **33,** 06LT02 (2020).

[38] J. Hertzberg, E. Zhang, S. Rosenblatt, E. Magesan, J. Smolin, J. Yau, V. Adiga, M. Sandberg, B. M., J. M. Chow, and J. S. Orcutt, *Laser-annealing josephson junctions for yielding scaled-up superconducting quantum processors*, npj Quantum Inf. **7,** 1–8 (2021).

[39] E. J. Zhang, S. Srinivasan, N. Sundaresan, D. F. Bogorin, Y. Martin, J. B. Hertzberg, J. Timmerwilke, E. J. Pritchett, J.-B. Yau, C. Wang, W. Landers, E. P. Lewandowski, A. Narasgond, S. Rosenblatt, G. A. Keefe, I. Lauer, M. B. Rothwell, D. T. McClure, O. E. Dial, J. S. Orcutt, M. Brink, and J. M. Chow, *High-performance superconducting quantum processors via laser annealing of transmon qubits*, Science Advances **8**, eabi6690 (2022).

[40] H. Kim, C. Jünger, A. Morvan, E. S. Barnard, W. P. Livingston, M. V. P. Altoé, Y. Kim, C. Song, L. Chen, J. M. Kreikebaum, D. F. Ogletree, D. I. Santiago, and I. Siddiqi, *Effects of laser-annealing on fixed-frequency superconducting qubits*, App. Phys. Lett. **121**, 142601 (2022).

[41] S. Vallés-Sanclemente, S. L. M. van der Meer, M. Finkel, N. Muthusubramanian, M. Beekman, H. Ali, J. F. Marques, C. Zachariadis, H. M. Veen, T. Stavenga, N. Haider, and L. DiCarlo, *Post-fabrication frequency trimming of coplanar-waveguide resonators in circuit QED quantum processors*, Applied Physics Letters **123**, 034004 (2023).

[42] 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].

[43] P. V. Klimov, A. Bengtsson, C. Quintana, A. Bourassa, S. Hong, A. Dunsworth, K. J. Satzinger, W. P. Livingston, V. Sivak, M. Y. Niu, T. I. Andersen, Y. Zhang, D. Chik, Z. Chen, C. Neill, C. Erickson, A. G. Dau, A. Megrant, P. Roushan, A. N. Korotkov, J. Kelly, V. Smelyanskiy, Y. Chen, and H. Neven, *Optimizing quantum gates towards the scale of logical qubits*, 2023, arXiv:2308.02321 [quant-ph].

[44] 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).

[45] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code*, Quantum Info. Comput. **15**, 997–1016 (2015).

[46] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. Di-Carlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Information **6**, 102 (2020).

[47] 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**, 1761 (2021).

[48] 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, S. Gustavsson, J. A. Formaggio, B. A. VanDevender, and W. D. Oliver, *Impact of ionizing radiation on superconducting qubit coherence*, Nature **584**, 551–556 (2020).

[49] C. D. Wilen, S. Abdullah, N. A. Kurinsky, C. Stanford, L. Cardani, G. D'Imperio, C. Tomei, L. Faoro, L. B. Ioffe, C. H. Liu, A. Opremcak, B. G. Christensen, J. L. DuBois, and R. McDermott, *Correlated charge noise and relaxation errors in superconducting qubits*, Nature **594**, 369–373 (2021).

[50] 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.

1

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*, Nature Physics **18**, 107–111 (2022).

[51] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, et al., *Charge-insensitive qubit design derived from the Cooper-pair box*, Phys. Rev. A **76**, 042319 (2007).

[52] U. Vool and M. Devoret, *Introduction to quantum electromagnetic circuits*, International Journal of Circuit Theory and Applications **45**, 897–934 (2017).

[53] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, *Circuit quantum electrodynamics*, Rev. Mod. Phys. **93**, 025005 (2021).

[54] B. D. Josephson, *Possible new effects in superconductive tunnelling*, Phys. Lett. **1**, 251–253 (1962).

[55] B. D. Josephson, *Coupled superconductors*, Rev. Mod. Phys. **36**, 216–220 (1964).

[56] 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).

[57] 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).

[58] 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).

[59] 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. Applied **8**, 034021 (2017).

[60] 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).

[61] 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. Why-land, K. Yadav, W. Zeng, and C. Rigetti, *Parametrically activated entangling gates using transmon qubits*, Phys. Rev. Applied **10**, 034050 (2018).

[62] A. D. Córcoles, J. M. Chow, J. M. Gambetta, C. Rigetti, J. R. Rozen, G. A. Keefe, M. Beth Rothwell, M. B. Ketchen, and M. Steffen, *Protecting superconducting qubits from radiation*, en, App. Phys. Lett. **99**, 181906 (2011).

[63] D. Ristè, J. G. van Leeuwen, H.-S. Ku, K. W. Lehnert, and L. DiCarlo, *Initialization by measurement of a superconducting quantum bit circuit*, Phys. Rev. Lett. **109**, 050507 (2012).

[64] X. Y. Jin, A. Kamal, A. P. Sears, T. Gudmundsen, D. Hover, J. Miloshi, R. Slattery, F. Yan, J. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Thermal and residual excited-state population in a 3d transmon qubit*, Phys. Rev. Lett. **114**, 240501 (2015).

[65] Z. Wang, S. Shankar, Z. Minev, P. Campagne-Ibarcq, A. Narla, and M. Devoret, *Cavity attenuators for superconducting qubits*, Phys. Rev. Appl. **11**, 014031 (2019).

[66] 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).

[67] F. Luthi, T. Stavenga, O. Enzing, A. Bruno, C. Dickel, N. 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).

[68] T. E. O'Brien, B. M. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Information **3** (2017).

[69] J. M. Martinis, S. Nam, J. Aumentado, K. M. Lang, and C. Urbina, *Decoherence of a superconducting qubit due to bias noise*, Phys. Rev. B **67**, 094510 (2003).

[70] J. Bylander, S. Gustavsson, F. Yan, F. Yoshihara, K. Harrabi, G. Fitch, D. G. Cory, Y. Nakamura, J.-S. Tsai, and W. D. Oliver, *Noise spectroscopy through dynamical decoupling with a superconducting flux qubit*, Nat. Phys. **7**, 565 (2011).

[71] 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).

[72] F. Battistel, *Mitigating leakage and noise in superconducting quantum computing* (Delft University of Technology, 2022).

[73] D. Gottesman, *The heisenberg representation of quantum computers*, 1998, arXiv:quant-ph/9807006 [quant-ph].

[74] S. Aaronson and D. Gottesman, *Improved simulation of stabilizer circuits*, Phys. Rev. A **70**, 052328 (2004).

1

1

[75] D. Ristè, C. C. Bultink, K. W. Lehnert, and L. DiCarlo, *Feedback control of a solid-state qubit using high-fidelity projective measurement*, Phys. Rev. Lett. **109**, 240502 (2012).

[76] 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**, 1–7 (2019).

[77] 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).

[78] 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).

[79] 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).

[80] A. Blais, R.-S. Huang, A. Wallraff, S. M. Girvin, and R. J. Schoelkopf, *Cavity quantum electrodynamics for superconducting electrical circuits: an architecture for quantum computation*, Phys. Rev. A **69**, 062320 (2004).

[81] A. Blais, J. Gambetta, A. Wallraff, D. I. Schuster, S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf, *Quantum-information processing with circuit quantum electrodynamics*, Phys. Rev. A **75**, 032329 (2007).

[82] M. D. Reed, L. DiCarlo, B. R. Johnson, L. Sun, D. I. Schuster, L. Frunzio, and R. J. Schoelkopf, *High-fidelity readout in circuit quantum electrodynamics using the Jaynes-Cummings nonlinearity*, Phys. Rev. Lett. **105**, 173601 (2010).

[83] M. Boissonneault, J. M. Gambetta, and A. Blais, *Dispersive regime of circuit qed: photon-dependent qubit dephasing and relaxation rates*, Phys. Rev. A **79**, 013819 (2009).

[84] 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).

[85] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Pechal, M. Mondal, M. Oppliger, C. Eichler, and A. Wallraff, *Rapid High-Fidelity Single-Shot Dispersive Readout of Superconducting Qubits*, Phys. Rev. App. **7**, 054020 (2017).

[86] Y. Sunada, S. Kono, J. Ilves, S. Tamate, T. Sugiyama, Y. Tabuchi, and Y. Nakamura, *Fast readout and reset of a superconducting qubit coupled to a resonator with an intrinsic purcell filter*, Phys. Rev. Appl. **17**, 044016 (2022).

[87] F. Swiadek, R. Shillito, P. Magnard, A. Remm, C. Hellings, N. Lacroix, Q. Ficheux, D. C. Zanuz, G. J. Norris, A. Blais, S. Krinner, and A. Wallraff, *Enhancing dispersive readout of superconducting qubits through dynamic control of the dispersive shift: experiment and theory*, 2023, arXiv:2307.07765 [quant-ph].

[88] 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).

[89] 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).

[90] M. Khezri, A. Opremcak, Z. Chen, A. Bengtsson, T. White, O. Naaman, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, A. L. Crook, B. Curtin, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, B. Foxen, G. Garcia, W. Giang, M. Giustina, R. Gosula, A. G. Dau, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, J. Kelly, S. Kim, P. V. Klimov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J. Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S. Martin, M. McEwen, A. Megrant, X. Mi, K. C. Miao, S. Montazeri, A. Morvan, M. Neeley, C. Neill, A. Nersisyan, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Quintana, C. Rocque, P. Roushan, K. Sankaragomathi, K. J. Satzinger, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, N. Zhu, N. Zobrist, D. Sank, A. Korotkov, Y. Chen, and V. Smelyanskiy, *Measurement-induced state transitions in a superconducting qubit: within the rotating wave approximation*, 2022, arXiv:2212.05097 [quant-ph].

[91] R. Shillito, A. Petrescu, J. Cohen, J. Beall, M. Hauru, M. Ganahl, A. G. Lewis, G. Vidal, and A. Blais, *Dynamics of transmon ionization*, Phys. Rev. Appl. **18**, 034031 (2022).

[92] C. Lledó, R. Dassonneville, A. Moulinas, J. Cohen, R. Shillito, A. Bienfait, B. Huard, and A. Blais, *Cloaking a qubit in a cavity*, 2022, arXiv:2211.05758 [quant-ph].

[93] M. H. Muñoz-Arias, C. Lledó, and A. Blais, *Qubit readouts enabled by qubit cloaking*, 2023, arXiv:2305.00895 [quant-ph].

[94] 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).

[95] J. M. Chow, L. DiCarlo, J. M. Gambetta, F. Motzoi, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, *Optimized driving of superconducting artificial atoms for improved single-qubit gates*, Phys. Rev. A **82**, 040305 (2010).

1

[96] 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).

[97] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, *Efficient Z gates for quantum computing*, Phys. Rev. A **96,** 022330 (2017).

[98] J. Majer, J. M. Chow, J. M. Gambetta, J. Koch, B. R. Johnson, J. A. Schreier, L. Frunzio, D. I. Schuster, A. A. Houck, A. Wallraff, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Coupling superconducting qubits via a cavity bus*, Nature **449,** 443–447 (2007).

[99] 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).

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

[101] 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).

[102] 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).

[103] 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).

[104] 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).

[105] S. Krinner, S. Lazar, A. Remm, C. K. Andersen, N. Lacroix, G. J. Norris, et al., *Benchmarking coherent errors in controlled-phase gates due to spectator qubits*, Phys. Rev. Appl. **14**, 024042 (2020).

[106] 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**, 020318 (2020).

[107] Y. Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, et al., *Qubit architecture with high coherence and fast tunable coupling*, Phys. Rev. Lett. **113**, 220502 (2014).

[108] 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).

[109] P. Mundada, G. Zhang, T. Hazard, and A. Houck, *Suppression of qubit crosstalk in a tunable coupling superconducting circuit*, Phys. Rev. Applied **12**, 054023 (2019).

[110] 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).

[111] 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).

[112] 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).

[113] 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*, Phys. Rev. Lett. **127**, 200502 (2021).

[114] E. A. Sete, A. Q. Chen, R. Manenti, S. Kulshreshtha, and S. Poletto, *Floating tunable coupler for scalable quantum computing architectures*, Phys. Rev. Appl. **15**, 064063 (2021).

[115] J. Stehlik, D. M. Zajac, D. L. Underwood, T. Phung, J. Blair, S. Carnevale, D. Klaus, G. A. Keefe, A. Carniol, M. Kumph, M. Steffen, and O. E. Dial, *Tunable coupling architecture for fixed-frequency transmon superconducting qubits*, Phys. Rev. Lett. **127**, 080505 (2021).

[116] V. E. Manucharyan, J. Koch, L. I. Glazman, and M. H. Devoret, *Fluxonium: single cooper-pair circuit free of charge offsets*, Science **326**, 113–116 (2009).

[117] N. A. Masluk, I. M. Pop, A. Kamal, Z. K. Minev, and M. H. Devoret, *Microwave characterization of Josephson junction arrays: implementing a low loss superinductance*, Phys. Rev. Lett. **109**, 137002 (2012).

1

[118] V. E. Manucharyan, *Superinductance* (Yale University, 2012).

[119] L. Grünhaupt, M. Spiecker, D. Gusenkova, N. Maleeva, S. T. Skacel, I. Takmakov, F. Valenti, P. Winkel, H. Rotzinger, W. Wernsdorfer, A. V. Ustinov, and I. M. Pop, *Granular aluminium as a superconducting material for high-impedance quantum circuits*, Nature Materials **18**, 816–819 (2019).

[120] L. B. Nguyen, G. Koolstra, Y. Kim, A. Morvan, T. Chistolini, S. Singh, K. N. Nesterov, C. Jünger, L. Chen, Z. Pedramrazi, B. K. Mitchell, J. M. Kreikebaum, S. Puri, D. I. Santiago, and I. Siddiqi, *Blueprint for a high-performance fluxonium quantum processor*, PRX Quantum **3**, 037001 (2022).

[121] L. B. Nguyen, Y.-H. Lin, A. Somoroff, R. Mencia, N. Grabon, and V. E. Manucharyan, *High-coherence fluxonium qubit*, Phys. Rev. X **9**, 041041 (2019).

[122] A. Somoroff, Q. Ficheux, R. A. Mencia, H. Xiong, R. Kuzmin, and V. E. Manucharyan, *Millisecond coherence in a superconducting qubit*, Phys. Rev. Lett. **130**, 267001 (2023).

[123] C. Gidney, *Stim: a fast stabilizer circuit simulator*, Quantum **5**, 497 (2021).

[124] C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal **27**, 379–423 (1948).

[125] J. L. Park, *The concept of transition in quantum mechanics*, Foundations of Physics **1**, 23–33 (1970).

[126] W. K. Wootters and W. H. Zurek, *A single quantum cannot be cloned*, Nature **299**, 802–803 (1982).

[127] D. Dieks, *Communication by epr devices*, Physics Letters A **92**, 271–272 (1982).

[128] D. Gottesman, *Stabilizer codes and quantum error correction* (California Institute of Technology, 1997).

[129] A. Kitaev, *Fault-tolerant quantum computation by anyons*, Annals of Physics **303**, 2–30 (2003).

[130] 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).

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

[132] B. M. Terhal, *Quantum error correction for quantum memories*, Rev. Mod. Phys. **87**, 307–346 (2015).

[133] 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).

[134] A. G. Fowler, A. M. Stephens, and P. Groszkowski, *High-threshold universal quantum computation on the surface code*, Phys. Rev. A **80**, 052312 (2009).

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

[136] D. Litinski, *A Game of Surface Codes: Large-Scale Quantum Computing with Lattice Surgery*, Quantum **3**, 128 (2019).

[137] Y. Tomita and K. M. Svore, *Low-distance surface codes under realistic quantum noise*, Phys. Rev. A **90**, 062320 (2014).

[138] M. McEwen, D. Bacon, and C. Gidney, *Relaxing Hardware Requirements for Surface Code Circuits using Time-dynamics*, Quantum **7**, 1172 (2023).

[139] D. Gottesman, *Quantum error correction and fault-tolerance*, 2005, arXiv:quant-ph/0507174 [quant-ph].

[140] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, *High-threshold and low-overhead fault-tolerant quantum memory*, 2023, arXiv:2308.07915 [quant-ph].

[141] B. Eastin and E. Knill, *Restrictions on transversal encoded quantum gate sets*, Phys. Rev. Lett. **102**, 110502 (2009).

[142] S. Bravyi and R. König, *Classification of topologically protected gates for local stabilizer codes*, Phys. Rev. Lett. **110**, 170503 (2013).

[143] M. E. Beverland, B. J. Brown, M. J. Kastoryano, and Q. Marolleau, *The role of entropy in topological quantum error correction*, Journal of Statistical Mechanics: Theory and Experiment **2019**, 073404 (2019).

[144] C. Wang, X. Li, H. Xu, Z. Li, J. Wang, Z. Yang, Z. Mi, X. Liang, T. Su, C. Yang, G. Wang, W. Wang, Y. Li, M. Chen, C. Li, K. Linghu, J. Han, Y. Zhang, Y. Feng, Y. Song, T. Ma, J. Zhang, R. Wang, P. Zhao, W. Liu, G. Xue, Y. Jin, and H. Yu, *Towards practical quantum computers: transmon qubit with a lifetime approaching 0.5 milliseconds*, npj Quantum Information **8**, 3 (2022).

[145] D. K. Tuckett, S. D. Bartlett, and S. T. Flammia, *Ultrahigh error threshold for surface codes with biased noise*, Phys. Rev. Lett. **120**, 050505 (2018).

[146] D. K. Tuckett, A. S. Darmawan, C. T. Chubb, S. Bravyi, S. D. Bartlett, and S. T. Flammia, *Tailoring surface codes for highly biased noise*, Phys. Rev. X **9**, 041031 (2019).

[147] D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, *Fault-tolerant thresholds for the surface code in excess of* 5% *under biased noise*, Phys. Rev. Lett. **124**, 130501 (2020).

[148] O. Higgott, T. C. Bohdanowicz, A. Kubica, S. T. Flammia, and E. T. Campbell, *Improved decoding of circuit noise and fragile boundaries of tailored surface codes*, Phys. Rev. X **13**, 031007 (2023).

[149] J. P. Bonilla Ataides, D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, *The xzzx surface code*, Nature Communications **12**, 2172 (2021).

[150] A. S. Darmawan, B. J. Brown, A. L. Grimsmo, D. K. Tuckett, and S. Puri, *Practical quantum error correction with the xzzx code and kerr-cat qubits*, PRX Quantum **2**, 030345 (2021).

[151] Q. Xu, N. Mannucci, A. Seif, A. Kubica, S. T. Flammia, and L. Jiang, *Tailored xzzx codes for biased noise*, Phys. Rev. Res. **5**, 013035 (2023).

1

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

[153] C. Chamberland, K. Noh, P. Arrangoiz-Arriola, E. T. Campbell, C. T. Hann, J. Iverson, H. Putterman, T. C. Bohdanowicz, S. T. Flammia, A. Keller, G. Refael, J. Preskill, L. Jiang, A. H. Safavi-Naeini, O. Painter, and F. G. Brandão, *Building a fault-tolerant quantum computer using concatenated cat codes*, PRX Quantum **3,** 010329 (2022).

[154] J. Guillaud and M. Mirrahimi, *Repetition cat qubits for fault-tolerant quantum computation*, Phys. Rev. X **9,** 041053 (2019).

[155] A. Leverrier and G. Zémor, *Quantum tanner codes*, in 2022 ieee 63rd annual symposium on foundations of computer science (focs) (2022), pp. 872–883.

[156] P. Panteleev and G. Kalachev, *Asymptotically good quantum and locally testable classical ldpc codes*, in Proceedings of the 54th annual acm sigact symposium on theory of computing, STOC 2022 (2022), pp. 375–388.

[157] O. Higgott and N. P. Breuckmann, *Constructions and performance of hyperbolic and semi-hyperbolic floquet codes*, 2023, arXiv:2308.03750 [quant-ph].

[158] M. A. Tremblay, N. Delfosse, and M. E. Beverland, *Constant-overhead quantum error correction with thin planar connectivity*, Phys. Rev. Lett. **129,** 050504 (2022).

[159] E. T. Campbell, B. M. Terhal, and C. Vuillot, *Roads towards fault-tolerant universal quantum computation*, Nature **549,** 172–179 (2017).

[160] H. Bombin and M. A. Martin-Delgado, *Topological quantum distillation*, Phys. Rev. Lett. **97,** 180501 (2006).

[161] H. Bombin and M. A. Martin-Delgado, *Topological computation without braiding*, Phys. Rev. Lett. **98,** 160502 (2007).

[162] H. Bombín, *Gauge color codes: optimal transversal gates and gauge fixing in topological stabilizer codes*, New Journal of Physics **17,** 083002 (2015).

[163] A. Kubica and M. E. Beverland, *Universal transversal gates with color codes: a simplified approach*, Phys. Rev. A **91,** 032330 (2015).

[164] A. J. Landahl, J. T. Anderson, and P. R. Rice, *Fault-tolerant quantum computing with color codes*, arXiv:1108.5738 (2011).

[165] F. Battistel, C. Chamberland, K. Johar, R. W. J. Overwater, F. Sebastiano, L. Skoric, Y. Ueno, and M. Usman, *Real-time decoding for fault-tolerant quantum computing: progress, challenges and outlook*, Nano Futures **7,** 032003 (2023).

[166] L. Skoric, D. E. Browne, K. M. Barnes, N. I. Gillespie, and E. T. Campbell, *Parallel window decoding enables scalable fault tolerant quantum computation*, 2023, arXiv:2209.08552 [quant-ph].

[167] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information* (Cambridge University Press, Cambridge, 2000).

[168] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, *Self-consistent quantum process tomography*, Phys. Rev. A **87**, 062119 (2013).

[169] R. Blume-Kohout, J. K. Gamble, E. Nielsen, J. Mizrahi, J. D. Sterk, and P. Maunz, *Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit*, arXiv:1310.4492 (2013).

[170] 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*, Nat. Commun. **8,** 14485 (2017).

[171] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout, *Gate Set Tomography*, Quantum **5,** 557 (2021).

[172] 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).

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

[174] J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, M. B. Ketchen, and M. Steffen, *Characterization of addressability by simultaneous randomized benchmarking*, Phys. Rev. Lett. **109,** 240504 (2012).

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

[176] 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).

[177] A. K. Hashagen, S. T. Flammia, D. Gross, and J. J. Wallman, *Real Randomized Benchmarking*, Quantum **2,** 85 (2018).

[178] J. Helsen, X. Xue, L. M. K. Vandersypen, and S. Wehner, *A new class of efficient randomized benchmarking protocols*, npj Quantum Information **5,** 71 (2019).

[179] A. G. Fowler, D. Sank, J. Kelly, R. Barends, and J. M. Martinis, *Scalable extraction of error models from the output of error detection circuits*, arXiv:1405.1454 (2014).

[180] S. Spitz, B. M. Tarasinski, C. Beenakker, and T. O'Brien, *Adaptive weight estimator for quantum error correction in a time-dependent environment*, Advanced Quantum Technologies **1**, 1800012 (2018).

[181] S. T. Flammia and J. J. Wallman, *Efficient estimation of pauli channels*, ACM Transactions on Quantum Computing **1**, 10.1145/3408039 (2020).

[182] R. Harper, S. T. Flammia, and J. J. Wallman, *Efficient learning of quantum noise*, Nature Physics **16**, 1184–1188 (2020).

[183] T. Wagner, H. Kampermann, D. Bruß, and M. Kliesch, *Learning logical pauli noise in quantum error correction*, Phys. Rev. Lett. **130,** 200601 (2023).

1

[184] E. H. Chen, T. J. Yoder, Y. Kim, N. Sundaresan, S. Srinivasan, M. Li, A. D. Córcoles, A. W. Cross, and M. Takita, *Calibrated decoders for experimental quantum error correction*, Phys. Rev. Lett. **128**, 110504 (2022).

[185] R. Harper and S. T. Flammia, *Learning correlated noise in a 39-qubit quantum processor*, PRX Quantum **4**, 040311 (2023).

[186] S. Bravyi, M. Suchara, and A. Vargo, *Efficient algorithms for maximum likelihood decoding in the surface code*, Phys. Rev. A **90**, 032326 (2014).

[187] P. Iyer and D. Poulin, *Hardness of decoding quantum stabilizer codes*, IEEE Transactions on Information Theory **61**, 5209–5223 (2015).

[188] A. G. Fowler, A. C. Whiteside, and L. C. L. Hollenberg, *Towards practical classical processing for the surface code*, Phys. Rev. Lett. **108**, 180501 (2012).

[189] J. Edmonds, *Paths, trees, and flowers*, Canadian Journal of Mathematics **17**, 449–467 (1965).

[190] O. Higgott and C. Gidney, *Sparse blossom: correcting a million errors per core second with minimum-weight matching*, 2023, arXiv:2303.15933 [quant-ph].

[191] A. G. Fowler, *Minimum weight perfect matching of fault-tolerant topological quantum error correction in average o(1) parallel time*, Quantum Info. Comput. **15**, 145–158 (2015).

[192] O. Higgott, *Pymatching: a python package for decoding quantum codes with minimum-weight perfect matching*, ACM Transactions on Quantum Computing **3**, 10.1145/3505637 (2022).

[193] Y. Wu and L. Zhong, *Fusion blossom: fast mwpm decoders for qec*, 2023, arXiv:2305.08307 [quant-ph].

[194] A. G. Fowler, *Optimal complexity correction of correlated errors in the surface code*, arXiv:1310.0863 (2013).

[195] J. Bausch, A. W. Senior, F. J. H. Heras, T. Edlich, A. Davies, M. Newman, C. Jones, K. Satzinger, M. Y. Niu, S. Blackwell, G. Holland, D. Kafri, J. Atalaya, C. Gidney, D. Hassabis, S. Boixo, H. Neven, and P. Kohli, *Learning to decode the surface code with a recurrent, transformer-based neural network*, 2023, arXiv:2310.05900 [quant-ph].

[196] N. Delfosse and N. H. Nickerson, *Almost-linear time decoding algorithm for topological codes*, Quantum **5**, 595 (2021).

[197] S. Huang, M. Newman, and K. R. Brown, *Fault-tolerant weighted union-find decoding on the toric code*, Phys. Rev. A **102**, 012419 (2020).

[198] P. Das, C. A. Pattison, S. Manne, D. M. Carmean, K. M. Svore, M. Qureshi, and N. Delfosse, *Afs: accurate, fast, and scalable error-decoding for fault-tolerant quantum computers*, in 2022 ieee international symposium on high-performance computer architecture (hpca) (2022), pp. 259–273.

[199] N. Liyanage, Y. Wu, A. Deters, and L. Zhong, *Scalable quantum error correction for surface codes using fpga*, 2023, arXiv:2301.08419 [quant-ph].

[200] B. Barber, K. M. Barnes, T. Bialas, O. Buğdaycı, E. T. Campbell, N. I. Gillespie, K. Johar, R. Rajan, A. W. Richardson, L. Skoric, C. Topal, M. L. Turner, and A. B. Ziad, *A real-time, scalable, fast and highly resource efficient decoder for a quantum computer*, 2023, arXiv:2309.05558 [quant-ph].

[201] L. Caune, J. Camps, B. Reid, and E. Campbell, *Belief propagation as a partial decoder*, 2023, arXiv:2306.17142 [quant-ph].

[202] J. Roffe, D. R. White, S. Burton, and E. Campbell, *Decoding across the quantum low-density parity-check code landscape*, Phys. Rev. Res. **2**, 043423 (2020).

[203] B. Criger and I. Ashraf, *Multi-path Summation for Decoding 2D Topological Codes*, Quantum **2**, 102 (2018).

[204] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. J. Beenakker, *Machine-learning-assisted correction of correlated qubit errors in a topological code*, Quantum **2**, 48 (2018).

[205] B. D'Anjou, *Generalized figure of merit for qubit readout*, Phys. Rev. A **103**, 042404 (2021).

[206] C. A. Pattison, M. E. Beverland, M. P. da Silva, and N. Delfosse, *Improved quantum error correction using soft information*, 2021, arXiv:2107.13589 [quant-ph].

[207] N. Raveendran, N. Rengaswamy, A. Pradhan, and B. Vasic, *Soft syndrome decoding of quantum ldpc codes for joint correction of data and syndrome errors*, in 2022 ieee international conference on quantum computing and engineering (qce) (Sept. 2022), pp. 275–281.

2

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. In this chapter, 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.

2.1. 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 the exploration of 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 chapter, 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 mea-

surements 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].

2.2. RESULTS

2.2.1. 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 [10] of the controlled-phase gate (CZ), considered as the native two-qubit gate in a transmon-based Surface-17, with experimentally targeted parameters (see Tab. 2.1 and Tab. 2.2). 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. 2.1 **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. 2.1 **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. 2.1 **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 correc-

2
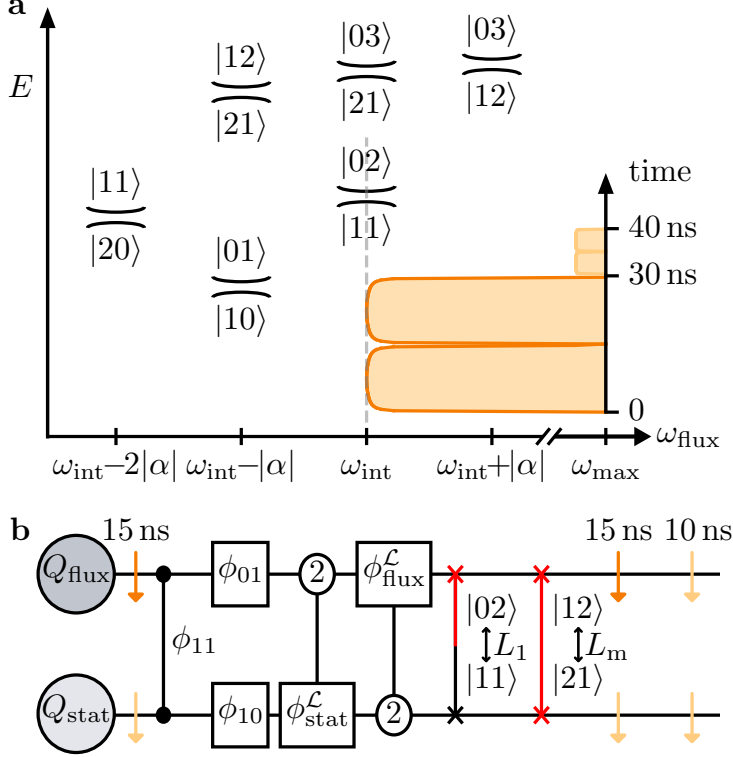
Figure 2.1: **CZ error model for two transmon qubits.** Schematic of the relevant interactions and the CZ error model for two transmons, a higher frequency one Q_{flux} and a lower frequency one Q_{stat}. The inset of **a** shows the frequency excursion taken by Q_{flux} from its sweetspot frequency 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 $\phi_{\text{flux}}^{\mathcal{L}}$ and $\phi_{\text{stat}}^{\mathcal{L}}$, 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 Sec. 2.2.1 for precise definitions). Relaxation and decoherence, indicated by the orange arrows, are taken into account as well.

tions. The conditional phase is defined as $\phi_{CZ} = \phi_{11} - \phi_{01} - \phi_{10} + \phi_{00}$, which for an ideal CZ is $\phi_{CZ} = \pi$. In this chapter, we assume $\phi_{00} = \phi_{01} = \phi_{10} = 0$ and $\phi_{CZ} = \phi_{11} = \pi$. 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 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 $\phi_{stat}^{\mathscr{L}} := \phi_{02} - \phi_{12}$. Similarly, if Q_{stat} is leaked, then Q_{flux} acquires a leakage conditional phase $\phi_{flux}^{\mathscr{L}} := \phi_{20} - \phi_{21}$. These rotations are generally non-trivial, i.e., $\phi_{stat}^{\mathscr{L}} \neq \pi$ and $\phi_{flux}^{\mathscr{L}} \neq 0$, due to the interactions in the 3-excitation manifold which cause a shift in the energy of $|12\rangle$ and $|21\rangle$ (see Sec. 2.5.1). If the only interaction leading to non-trivial $\phi_{stat}^{\mathscr{L}}$, $\phi_{flux}^{\mathscr{L}}$ 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 $\phi_{stat}^{\mathscr{L}} = \pi - \phi_{flux}^{\mathscr{L}}$.

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 $|11\rangle\langle02|$ and $|02\rangle\langle11|$ do not affect the results presented in this chapter, so we set them to 0 for computational efficiency (see Sec. 2.4.2). 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 Fig. 2.9 and Sec. 2.5.1.

The described operations are implemented as instantaneous in the *quantumsim* package (introduced in [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. 2.1 **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 chapter are summarized in Sec. 2.4.2. In particular, unless otherwise stated, L_1 is set to 0.125% and $\phi_{flux}^{\mathscr{L}}$ and $\phi_{stat}^{\mathscr{L}}$ 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 $\phi_{flux}^{\mathscr{L}}$ and $\phi_{stat}^{\mathscr{L}}$ have not been characterized in experiment, we instead choose to randomize 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 Sec. 2.5.1). 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.

2.2.2. EFFECT OF LEAKAGE ON THE CODE PERFORMANCE

We implement density-matrix simulations [48] to study the effect of leakage in Surface-17 (Fig. 2.2). We follow the frequency arrangement and operation scheduling proposed in [47], which employs three qubit frequencies for the surface-code lattice, arranged as shown in Fig. 2.2 **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. 2.1). Based on the leakage model in Sec. 2.2.1, 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]. The physical error rate of a single transmon qubit under the effect of decoherence is defined in [28].

Fig. 2.2 **b** shows that the logical error rate ε_{L} is sharply pushed above the memory break-even point (the physical error rate) by the 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. 2.2 **c,d**. While not included in these simulations, we do not expect the inclusion of leakage mobility or the possibility of further leaking to $|3\rangle$ to have a considerable effect on the logical performance (see Sec. 2.5.7).

2.2.3. 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 $p_{\text{DM}}^{\mathscr{L}}(Q) = \mathbb{P}(Q \in \mathscr{L}) = \langle 2|\rho_Q|2\rangle$ 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, $p_{\text{DM}}^{\mathscr{L}}(Q)$ 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 Tab. 2.1). We refer to this sharp increase of $p_{\text{DM}}^{\mathscr{L}}(Q)$ 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. 2.3 **a**. This is the typical behavior of leakage, as shown in Fig. 2.3 **b** by the bi-modal density distribution of the probabilities $p_{\text{DM}}^{\mathscr{L}}(Q)$ 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-
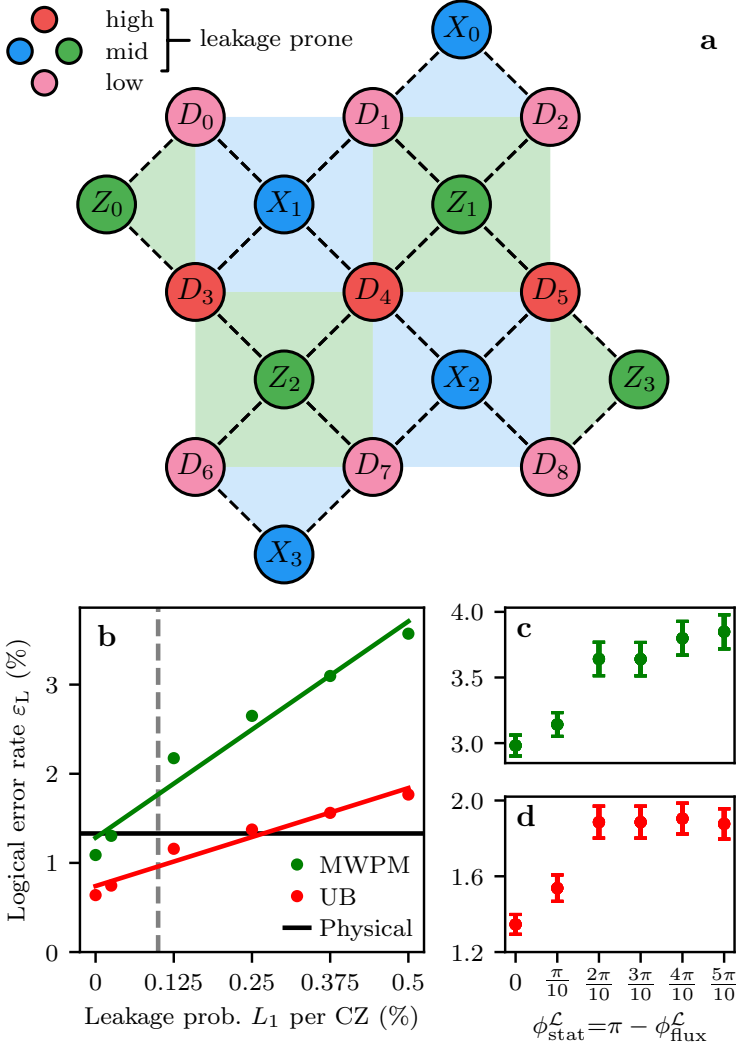
Figure 2.2: **The effect of leakage on the performance of Surface-17. a** Schematic overview of the Surface-17 layout [47]. Pink (resp. red) circles with D labels represent low- (high-) frequency data qubits, while blue (resp. green) circles with X (Z) labels represent ancilla qubits of intermediate 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 [10]. Logical error rate ε_L for MWPM (**c**) and upper bound (UB) (**d**) as a function of the leakage conditional phases $\phi_{\mathrm{flux}}^{\mathscr{L}}$ and $\phi_{\mathrm{stat}}^{\mathscr{L}}$ (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).
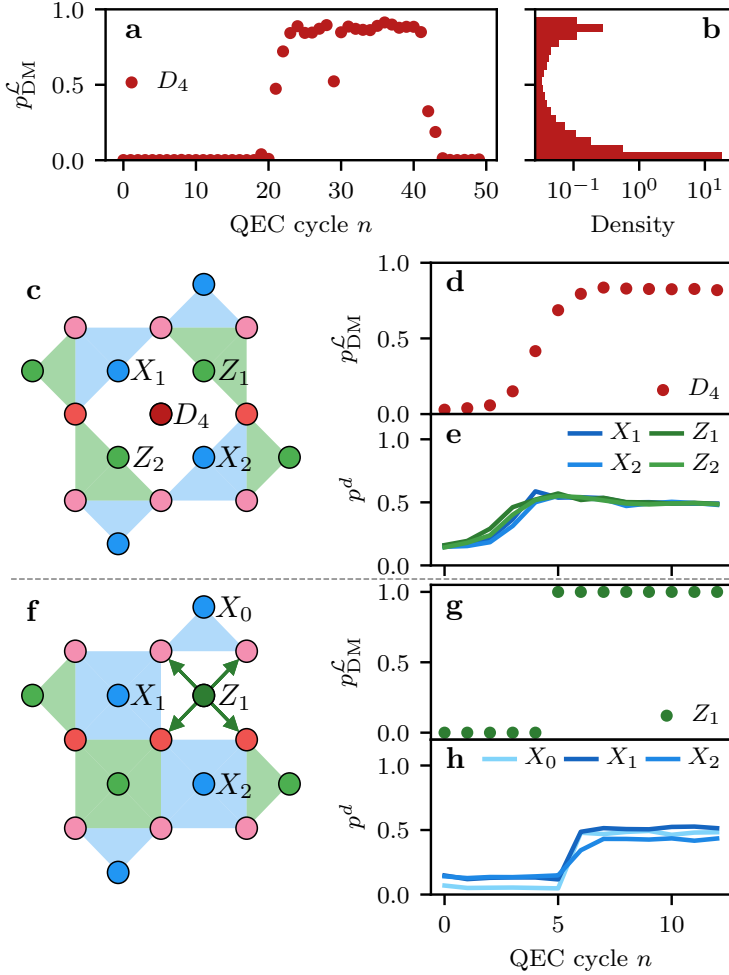
Figure 2.3: **Projection and signatures of qubit leakage. a-b** Projection of data-qubit leakage. **a** Example realization of a data-qubit leakage event, extracted from the density-matrix simulations. **b** Density histogram of all data-qubit leakage probabilities over 20 bins, extracted over 4×10^4 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 Sec. 2.5.6). **d** The average projection of the leakage probability $p_{\mathrm{DM}}^{\mathcal{L}}$ of D_4 over all events, where this probability is first below and then above a threshold of $p_{\mathrm{th}}^{\mathcal{L}} = 0.5$ 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 $p_{\mathrm{th}}^{\mathcal{L}}$. **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.

action effect of the repetitive stabilizer measurements (see Fig. 2.10 and Sec. 2.5.2). Given this projective behavior, we identify individual events by introducing a threshold $p_{\text{th}}^{\mathscr{L}}(Q)$, above which a qubit is considered as leaked. Here we focus on leakage on D_4, sketched in Fig. 2.3 **c**. Based on a threshold $p_{\text{th}}^{\mathscr{L}}(D_4) = 0.5$, we select leakage events and extract the average dynamics shown in Fig. 2.3 **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. Each of the high-frequency data qubits equilibrates towards a steady-state population (extracted by averaging $p_{\text{DM}}^{\mathscr{L}}(Q)$ over all runs without selecting individual events) after many QEC cycles (see Fig. 2.11 and Sec. 2.5.3).

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. 2.3 **e**. As $p_{\text{DM}}^{\mathscr{L}}(D_4)$ 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. 2.3 **c** and as observed in [21]. This anti-commutation leads to some of the increase in ε_{L} for the MWPM and UB decoders in Fig. 2.2 **b**. Furthermore, we attribute the observed sharp projection of leakage (see Fig. 2.3 **d**) to a back-action effect of the measurements of the neighboring stabilizers, whose outcomes are nearly randomized when the qubit is leaked (see Sec. 2.5.6 and Sec. 2.5.2). 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, $p_{\text{DM}}^{\mathscr{L}}$ is defined as the leakage probability after the ancilla projection during measurement. Since in the simulations ancilla qubits are fully projected, $p_{\text{DM}}^{\mathscr{L}}(Q) = 0, 1$ for an ancilla qubit Q, allowing to directly obtain the individual leakage events, as shown in Fig. 2.3 **g**. We note that an ancilla qubit remains leaked for 17 QEC cycles on average for the considered parameters (given in Tab. 2.1). We extract p^d during the selected events, as shown in Fig. 2.3 **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. 2.3 **f** and described in Sec. 2.2.1 The angle of rotation is determined by $\phi_{\text{flux}}^{\mathscr{L}}$ or $\phi_{\text{stat}}^{\mathscr{L}}$, depending on whether the leaked ancilla qubit takes the roles of Q_{stat} or Q_{flux}, respectively (see Sec. 2.4.1). 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. 2.3 **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. 2.2 **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.

2.2.4. HIDDEN MARKOV MODELS
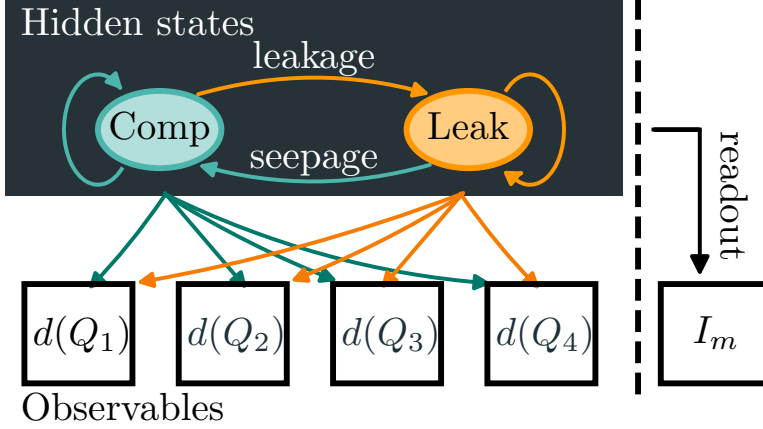
Figure 2.4: **Schematic representation of an HMM for leakage detection.** For both ancilla and data qubits only two hidden states are considered, corresponding to the qubit being either in the computational (teal) or leakage subspace (orange). Transitions between these states occur each QEC cycle, depending 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.

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. 2.4 and Sec. 2.4.3) 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 Sec. 2.2.3) to high $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$. 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 Sec. 2.2.3), 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 Sec. 2.5.5).

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 Sec. 2.4.3). 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 a nearly constant value by the stabilizer anti-commutation, while when an ancilla qubit is leaked, this probability depends on $\phi_{\text{flux}}^{\mathscr{L}}$ and $\phi_{\text{stat}}^{\mathscr{L}}$. 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 Sec. 2.2.6 and Sec. 2.5.5).

2.2.5. DATA-QUBIT LEAKAGE DETECTION
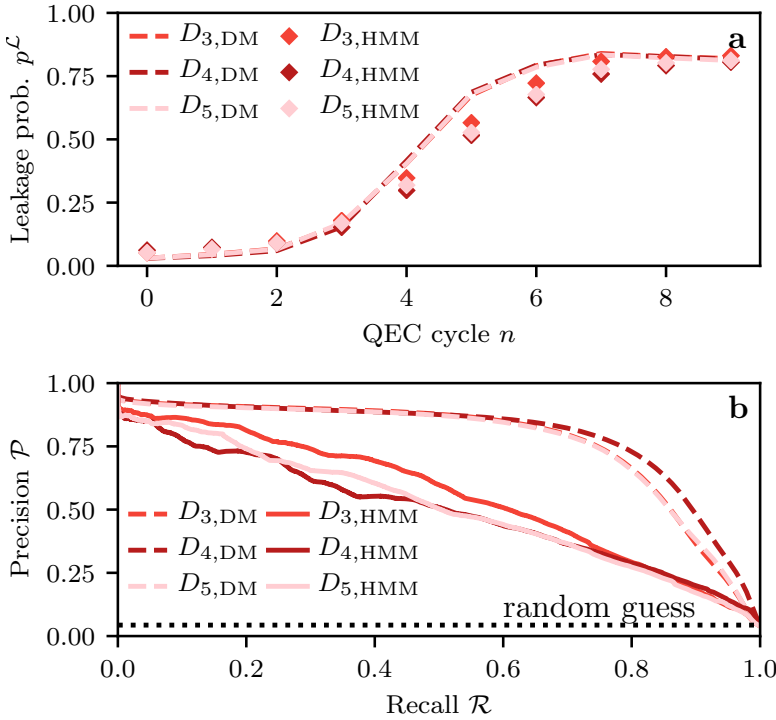
Figure 2.5: **Data-qubit leakage detection. a** Average response in time of the HMMs (diamonds) to leakage, in comparison to the actual leakage probability extracted from the density-matrix simulations (dashed lines). The average is computed by selecting single realizations where $p_{\text{DM}}^{\mathscr{L}}(Q)$ was below a threshold $p_{\text{th}}^{\mathscr{L}} = 0.5$ 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.

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 $p_{\text{HMM}}^{\mathscr{L}}(Q)$ of the HMMs to an event is shown in Fig. 2.5 and compared to the leakage probabilities $p_{\text{DM}}^{\mathscr{L}}(Q)$ extracted from the density-matrix simulation. Events are selected when crossing a threshold $p_{\text{th}}^{\mathscr{L}}$, as described in Sec. 2.2.3, and the response is averaged over these events. For the data-qubit HMMs, the response $p_{\text{HMM}}^{\mathscr{L}}(Q)$ closely follows the probability $p_{\text{DM}}^{\mathscr{L}}(Q)$ 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

$$\mathscr{P}_{\text{HMM}}(Q) = \mathbb{P}\left(Q \in \mathscr{L} \mid p_{\text{HMM}}^{\mathscr{L}}(Q) > p_{\text{th}}^{\mathscr{L}}(Q)\right) \tag{2.1}$$

and can be computed as

$$\mathscr{P}_{\text{HMM}}(Q) = \frac{\sum_i p_{\text{DM}}^{\mathscr{L}}(Q,i)\,\theta\left[p_{\text{HMM}}^{\mathscr{L}}(Q,i) - p_{\text{th}}^{\mathscr{L}}(Q)\right]}{\sum_i \theta\left[p_{\text{HMM}}^{\mathscr{L}}(Q,i) - p_{\text{th}}^{\mathscr{L}}(Q)\right]}, \tag{2.2}$$

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

$$\mathscr{R}_{\text{HMM}}(Q) = \mathbb{P}\left(p_{\text{HMM}}^{\mathscr{L}}(Q) > p_{\text{th}}^{\mathscr{L}}(Q) \mid Q \in \mathscr{L}\right), \tag{2.3}$$

and can be computed as

$$\mathscr{R}_{\text{HMM}}(Q) = \frac{\sum_i p_{\text{DM}}^{\mathscr{L}}(Q,i)\,\theta\left[p_{\text{HMM}}^{\mathscr{L}}(Q,i) - p_{\text{th}}^{\mathscr{L}}(Q)\right]}{\sum_i p_{\text{DM}}^{\mathscr{L}}(Q,i)}. \tag{2.4}$$

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. 2.5 **b**) is a parametric curve obtained by sweeping $p_{\text{th}}^{\mathscr{L}}(Q)$ and plotting the value of $\mathscr{P}$ and $\mathscr{R}$. Since the PR curve is constructed from $p_{\text{HMM}}^{\mathscr{L}}(Q)$ 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

$$\mathscr{O}(Q) = \text{AUC}_{\text{HMM}}(Q) / \text{AUC}_{\text{DM}}(Q), \tag{2.5}$$

where $\text{AUC}_{\text{HMM}}(Q)$ is the area under the PR curve of the HMM and $\text{AUC}_{\text{DM}}(Q)$ is the area for the optimal model that predicts leakage with probability $p_{\text{DM}}^{\mathscr{L}}(Q)$, achieving the best possible $\mathscr{P}_{\text{DM}}$ and $\mathscr{R}_{\text{DM}}$. 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 Fig. 2.12 and Sec. 2.5.4).

2.2.6. ANCILLA-QUBIT LEAKAGE DETECTION
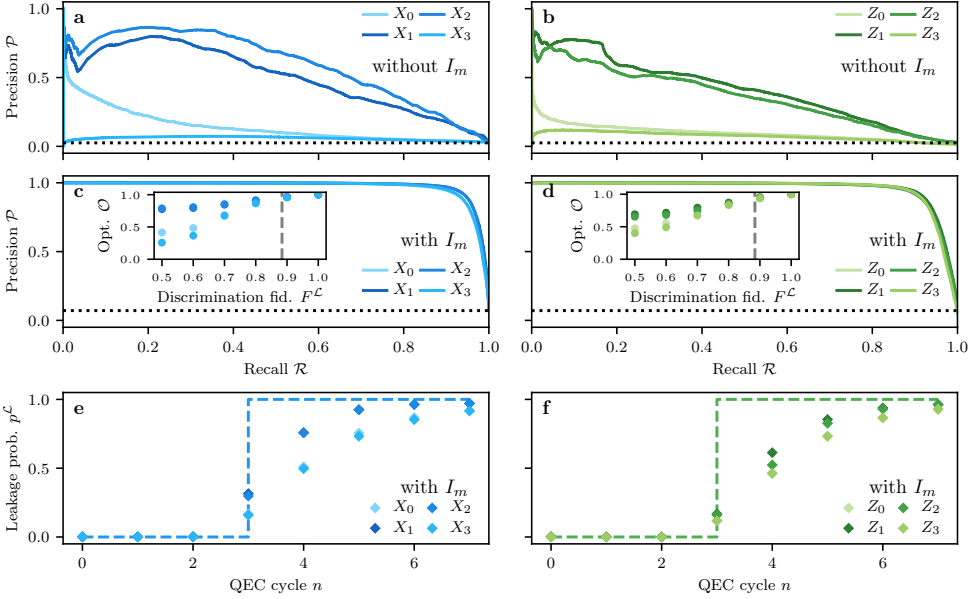
Figure 2.6: **Ancilla-qubit leakage detection. a-d** Precision-recall curves for the ancilla-qubit HMMs over 4×10^4 runs of 50 QEC cycles each. In **a,b** the HMMs rely only on the observed defects on the neighboring stabilizers. In **c-f** the HMMs further get the in-phase component I_m of the analog readout as input, from which $p_m^{\mathcal{L}}$ is extracted. The dotted line corresponds to a random guess classifier for which $\mathscr{P}$ is equal to the fraction of leakage events over all QEC cycles and runs. As ancilla-qubit leakage is directly measured, $\mathscr{P}_{DM} = 1$ for all values of $\mathscr{R}$ (not shown). Insets in **c,d**: the HMM optimality $\mathscr{O}$ 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.

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. 2.6 **a,b**. Given that projective measurements are used in the simulations, $\text{AUC}_{DM}(Q) = 1$ for ancilla qubits. Bulk ancilla qubits have a moderate $\mathscr{O}(Q) \approx 47\%$, 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. 2.6 **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 [47] and Fig. 2.8), 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 $\phi_{\text{stat}}^{\mathcal{L}}$ and $\phi_{\text{flux}}^{\mathcal{L}}$ (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 Sec. 2.4.3). 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 $\phi_{\text{stat}}^{\mathcal{L}}$ and $\phi_{\text{flux}}^{\mathcal{L}}$) 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. 2.2).

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

$$F^{\mathcal{L}} = 1 - \frac{\mathbb{P}(1 \mid 2) + \mathbb{P}(2 \mid 1)}{2}, \tag{2.6}$$

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. 2.13). We focus on the discrimination fidelity as in our simulations relaxation is already accounted for in the measurement outcomes (see Sec. 2.4.2). We extract $F^{\mathcal{L}}$ 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 $F^{\mathcal{L}} = 88.4\%$ (see Sec. 2.5.5).

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

$$p_m^{\mathcal{L}} = \frac{\mathcal{N}_2(I_m)}{\sum_{j \in \{0,1,2\}} \mathcal{N}_j(I_m)}. \tag{2.7}$$

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. 2.6 **c,d**, from which an average $\mathcal{O}(Q) \approx 97\%$ can be extracted for the ancilla-qubit HMMs. The temporal responses of the HMMs to leakage are compared to the leakage probabilities extracted from measurement in Fig. 2.6 **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 $F^{\mathcal{L}} = 88.4\%$ might suggest an even sharper response, this is not the case as the HMM update depends on both the prior $p_{\text{HMM}}^{\mathcal{L}}$ (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 (see Sec. 2.4.3). 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. 2.6 **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 $\phi_{\text{stat}}^{\mathscr{L}}$ and $\phi_{\text{flux}}^{\mathscr{L}}$.

We explore $\mathscr{O}(Q)$ as a function of $F^{\mathscr{L}}$, as shown in the inset of Fig. 2.6 **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 Sec. 2.5.5). At low $F^{\mathscr{L}}$ ($\lesssim 60\%$) 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 $\phi_{\text{stat}}^{\mathscr{L}}$ and $\phi_{\text{flux}}^{\mathscr{L}}$, without compromising the capability of the HMMs to detect leakage. In Sec. 2.5.8 we explore an alternative scheme for increasing the performance of the ancilla-qubit HMMs without using the analog measurements, which comes at the cost of a lower optimality for data-qubit HMMs.

2.2.7. IMPROVING CODE PERFORMANCE VIA POST-SELECTION
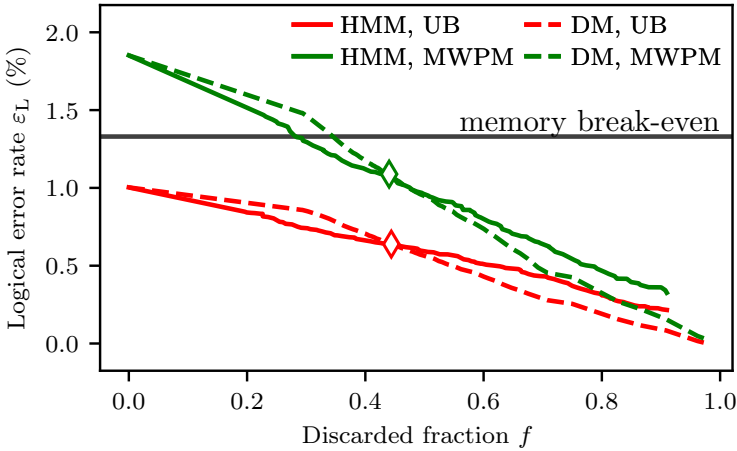
Figure 2.7: **Restoring code performance by post-selecting on leakage detection.** Improvement in the logical error rate ε_{L} via post-selecting on the detection of leakage for a MWPM decoder (green) and the decoder upper bound (red). The post-selection is based on the probabilities predicted by the 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.

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

$$\max_{Q,n} p^{\mathscr{L}}(Q, n) \geq p^{\mathscr{L}}_{\text{th}}(Q).$$

(2.8)

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 Sec. 2.5.4), 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

$$\min_{p^{\mathscr{L}}_{\text{th}}(Q)} \qquad (f, \varepsilon_{\text{L}}),$$

$$\text{subject to} \qquad 0.02 \leq p^{\mathscr{L}}_{\text{th}}(Q) \leq 1,$$

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. 2.7 for both the MWPM and UB decoders. In Fig. 2.7 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 Sec. 2.2.6). 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.

2.3. 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 dedicated 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 Sec. 2.5.6), 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 Sec. 2.2.6. 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 nontrivial 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 [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 in superconducting qubits are the swap scheme introduced in [37], or the use of the readout resonator to reset a leaked data-qubit into the computational subspace, similarly to [55, 56]. 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. 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 ~ 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 Sec. 2.5.6). 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 [51–53]. Given control of the leakage conditional phases, the performance of this decoder can be optimized by setting $\phi_{\text{stat}}^{\mathscr{L}} = \pi$ and $\phi_{\text{flux}}^{\mathscr{L}} = 0$, 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 Fig. 2.14 and Sec. 2.5.6). Given a moderate discrimination fidelity of the leaked state, this is not expected to compromise the detectability of leakage, as discussed in Sec. 2.2.6. 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 chapter 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 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.

2.4. METHODS

2.4.1. SIMULATION PROTOCOL
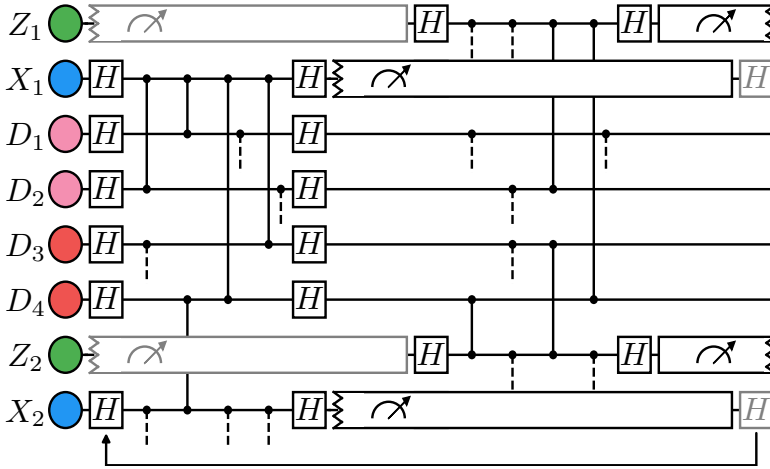
Figure 2.8: **The quantum circuit for a single QEC cycle employed in simulation.** The unit-cell scheduling is defined in [47]. The qubit labels and frequencies correspond to the lattice arrangement shown in Fig. 2.2. Gray elements correspond to operations belonging to the previous or the following QEC 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 Tab. 2.1. The arrow at the bottom indicates the repetition of QEC cycles.

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 Sec. 2.4.2).

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 chapter), with the quantum circuit given in Fig. 2.8. 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. In Sec. 2.2.6 we considered the discrimination of $|2\rangle$ in read-

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 2.1: The parameters for the qubit decoherence times and for the gate, measurement and QEC-cycle durations used in the density-matrix simulations.

out, 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 chapter.

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

$$F_{\text{L}}(n) = \frac{1}{2}\left[1 + (1 - 2\varepsilon_{\text{L}})^{n-n_0}\right], \tag{2.10}$$

where n_0 is a fitting parameter (usually close to 0) [28].

2.4.2. 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

$$\frac{d\rho}{dt} = -i\left[H, \rho\right] + \sum_i L_i \rho L_i^\dagger - \frac{1}{2}\left\{L_i^\dagger L_i, \rho\right\}, \tag{2.11}$$

where H is the transmon Hamiltonian

$$H = \omega a^\dagger a + \frac{\alpha}{2}(a^\dagger)^2 a^2, \tag{2.12}$$

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

$$L_{\text{amp}} = \sqrt{\frac{1}{T_1}}\, a. \tag{2.13}$$

The $|2\rangle$ lifetime is then characterized by a relaxation time $T_1/2$. Dephasing is described by

$$L_{\text{deph,1}} = \sqrt{\frac{8}{9T_\phi}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \tag{2.14}$$

$$L_{\text{deph,2}} = \sqrt{\frac{2}{9T_\phi}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.15}$$

$$L_{\text{deph,3}} = \sqrt{\frac{2}{9T_\phi}} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \tag{2.16}$$

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 $|2\rangle$ [10]. 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 ω_{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 rate at sweetspot), given by $\Gamma_\phi = 2\pi\sqrt{\ln 2A} D_\phi$ [57], while the low-frequency components are not included due to the built-in echo effect of Net-Zero pulses [10]. 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 Tab. 2.1. 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 [10].

We now consider the coherence of leakage in the CZ gates, which in the rotating frame of the qutrit is modeled as the exchanges

$$|11\rangle \mapsto \sqrt{1 - 4L_1}\,|11\rangle + e^{i\phi}\sqrt{4L_1}\,|02\rangle, \tag{2.17}$$

$$|02\rangle \mapsto \sqrt{1 - 4L_1}\,|02\rangle - e^{-i\phi}\sqrt{4L_1}\,|11\rangle, \tag{2.18}$$

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. (2.17) and (2.18) leads to a coherent superposition of computational and leaked states

$$\rho = \left(\begin{array}{c|c} \rho^{\mathscr{C}} & \rho^{\mathrm{coh}} \\ \hline \rho^{\mathrm{coh}} & \rho^{\mathscr{L}} \end{array} \right), \tag{2.19}$$

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 chapter 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 measure-

ments (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.

2.4.3. 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 $B_{o_i[n],s} := \mathbb{P}\left(o_i[n] = o_i \mid s[n] = s\right)$. Inverting this problem, the inference of the posterior state probabilities $p^s[n]$ from the realized observables is possible via

$$p^s[n] = \mathbb{P}\left(s[n] \mid o[n], o[n-1], \ldots, o[1]\right) \tag{2.20}$$

$$= \frac{\mathbb{P}\left(o[n] \mid s[n]\right) p^s_{\text{prior}}[n]}{\mathbb{P}\left(o[n]\right)} \tag{2.21}$$

$$= \frac{\prod_i \mathbb{P}\left(o_i[n] \mid s[n]\right) p^s_{\text{prior}}[n]}{\prod_i \mathbb{P}\left(o_i[n]\right)} \tag{2.22}$$

$$= \frac{\prod_i B_{o_i[n],s} p^s_{\text{prior}}[n]}{\sum_{s'} \prod_i B_{o_i[n],s'} p^{s'}_{\text{prior}}[n]}, \tag{2.23}$$

where $p^s_{\text{prior}}[n]$ is the prior probability

$$p^s_{\text{prior}}[n] = \sum_{s'} A_{s,s'} p^{s'}[n-1]. \tag{2.24}$$

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, $S = \{\mathcal{C}, \mathcal{L}\}$, namely whether the qubit is in the computational ($\mathcal{C}$) or the leakage subspace ($\mathcal{L}$). The transition matrix is parameterized in terms of the leakage and seepage probabilities per QEC cycle. The leakage probability is estimated as $\Gamma_{\mathcal{C} \to \mathcal{L}} \approx N_{\text{flux}} L_1$ (for low L_1), where N_{flux} is the number of CZ gates for which the qubit is fluxed during a QEC cycle and L_1 is the leakage probability per CZ gate. The seepage probability is estimated by $\Gamma_{\mathcal{L} \to \mathcal{C}} \approx N_{\text{flux}} L_2 + \left(1 - e^{\frac{t_c}{T_{1/2}}}\right)$, where t_c is the QEC cycle duration and T_1 the relaxation time (see Tab. 2.1), while L_2 is the seepage contribution from the gate, where $L_2 = 2L_1$ due to the dimensionality ratio between $\mathcal{C}$ and $\mathcal{L}$ for a qubit-qutrit pair [49]. The transition matrix A is then given by

$$A = \begin{pmatrix} 1 - \Gamma_{\mathcal{C} \to \mathcal{L}} & \Gamma_{\mathcal{L} \to \mathcal{C}} \\ \Gamma_{\mathcal{C} \to \mathcal{L}} & 1 - \Gamma_{\mathcal{L} \to \mathcal{C}} \end{pmatrix}. \tag{2.25}$$

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 $p^{\mathscr{L}}_{\text{DM}}(Q_i)$ as observed from the density matrix is above a threshold of $p^{\mathscr{L}}_{\text{th}} = 0.5$. In the case of ancilla-qubit leakage, $B_{d_i[n],\mathscr{L}}$ depends on the values of the leakage conditional phases $\phi^{\mathscr{L}}_{\text{stat}}$ and $\phi^{\mathscr{L}}_{\text{flux}}$. 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 Sec. 2.2.3).

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 Sec. 2.5.5.

2.5. SUPPLEMENTARY MATERIAL

2.5.1. SECOND-ORDER LEAKAGE EFFECTS

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 2.2: 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. 2.2.

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 Sec. 2.2.1. 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
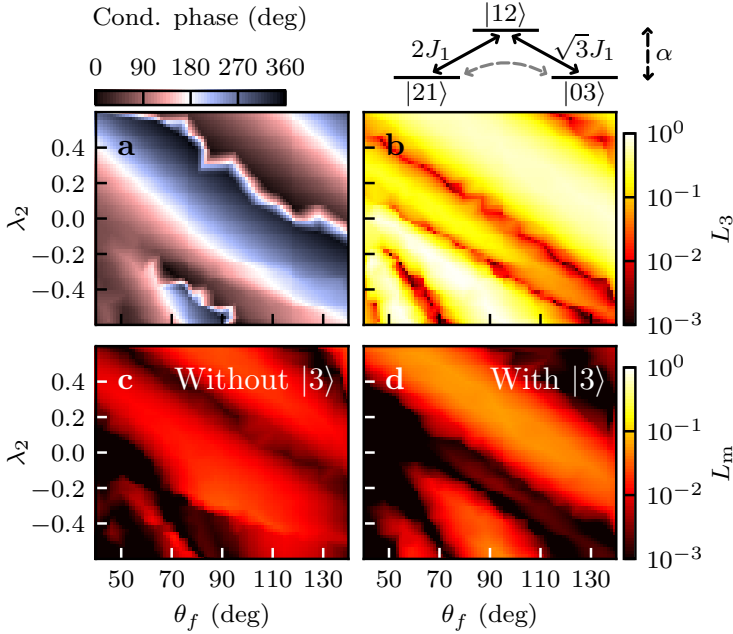
Figure 2.9: Heatmaps obtained from CZ full-trajectory simulations. **a,b,d** include $|3\rangle$ in the Hilbert space of the fluxing qubit, while (**c**) does not. The conditional phase ϕ_{CZ} (**a**), L_1 (**b**) and L_m (**c,d**) are plotted versus the flux-pulse parameters (see [10, 34] for definitions). The interaction point is located at $\theta_f = 90\,°$. 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 [10]

$$
\begin{aligned}
H(t) = {} & \omega_{\text{stat}} a_{\text{stat}}^{\dagger} a_{\text{stat}} + \frac{\alpha_{\text{stat}}}{2} (a_{\text{stat}}^{\dagger})^2 a_{\text{stat}}^2 \\
& + \omega_{\text{flux}}(\Phi(t)) a_{\text{flux}}^{\dagger} a_{\text{flux}} + \frac{\alpha_{\text{flux}}}{2} (a_{\text{flux}}^{\dagger})^2 a_{\text{flux}}^2 \\
& + J_1(\Phi(t)) (a_{\text{stat}} a_{\text{flux}}^{\dagger} + a_{\text{stat}}^{\dagger} a_{\text{flux}}),
\end{aligned}
\tag{2.26}
$$

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. 2.1. We perform full-trajectory simulations (following the same structure as in [10], excluding distortions and quasi-static flux noise) using the parameters reported in Table 1 and Tab. 2.2. 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

$$
L_3 := |\langle 03 | \mathscr{S}_{\text{CZ}}(|12\rangle \langle 12|) | 03 \rangle|^2,
\tag{2.27}
$$

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. 2.9 **b** shows for the high-mid qubit pair, even when $\phi_{\text{CZ}} = \pi$ (see Fig. 2.9 **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 [10], 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

$$
L_{\text{m}} := |\langle 21 | \mathscr{S}_{\text{CZ}}(|12\rangle \langle 12|) | 21 \rangle|^2.
\tag{2.28}
$$

If $|3\rangle$ is not included, L_{m} takes small but non-negligible values, as shown in Fig. 2.9 **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. 2.9, 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.

$$\frac{1}{2\pi}\frac{(2J_1)(\sqrt{3}J_1)}{\alpha} \approx 2.6\,\text{MHz}, \tag{2.29}$$

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.

2.5.2. 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. 2.10 **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 $|\Phi^+\rangle_{Q_1 Q_2} = (|00\rangle + |11\rangle)/\sqrt{2}$, 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

$$|\psi_1\rangle_{Q_1 Q_2 A} = \frac{1}{2\sqrt{2}}\Big(\big[2|00\rangle + (1+a)|11\rangle + b|21\rangle\big]|0\rangle$$
$$+ \big[(1-a)|11\rangle + b|21\rangle\big]|1\rangle\Big), \tag{2.30}$$

with probability $1-p$, where $a = \sqrt{1-4L_1}$ and $b = \sqrt{4L_1}$, or in state

$$|\psi_2\rangle_{Q_1 Q_2 A} = \frac{1}{2\sqrt{2}}\Big(\big[(1-a)|10\rangle + b|20\rangle\big]|0\rangle$$
$$+ \big[(1+a)|10\rangle + 2|01\rangle + b|20\rangle\big]|1\rangle\Big), \tag{2.31}$$

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:

$$\rho|_{|1\rangle} = \frac{1}{2(1-a)+4p(1+a)}\Big($$
$$(1-p)\big[(1-a)^2|11\rangle\langle 11| + b^2|21\rangle\langle 21|\big]$$
$$p\big[(1+a)^2|10\rangle\langle 10| + 4|01\rangle\langle 01| + b^2|20\rangle\langle 20|$$
$$+ 2(1+a)(|10\rangle\langle 01| + |01\rangle\langle 10|)\big]\Big), \tag{2.32}$$
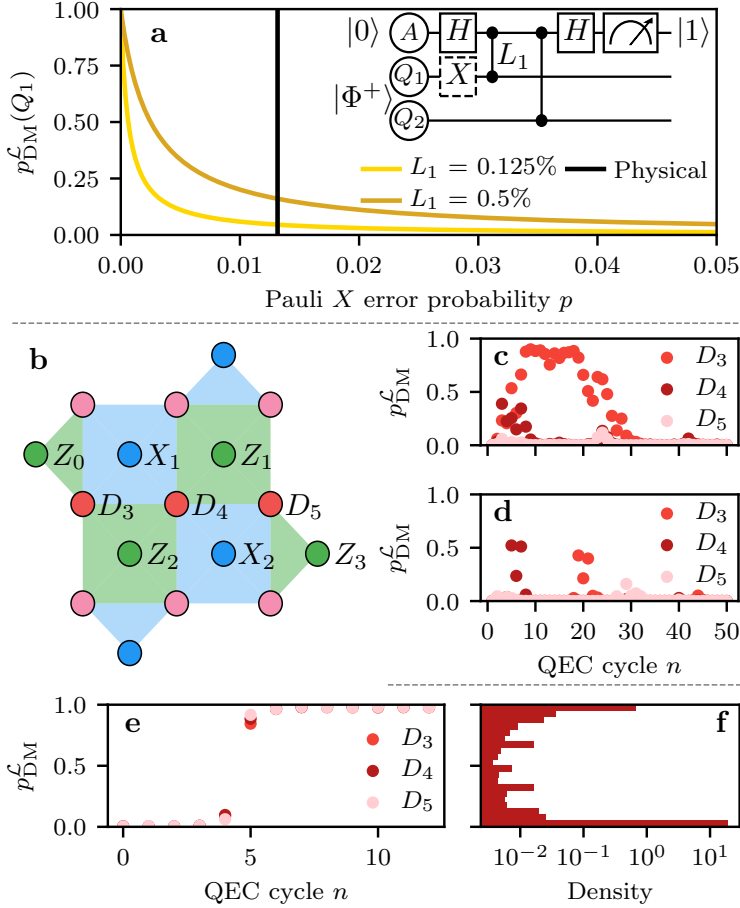
Figure 2.10: Projection of data-qubit leakage. **a** Inset: an ancilla qubit A, initialized in $|0\rangle$, measures ZZ on two data qubits Q_1, Q_2, initialized in the Bell state $|\Phi^+\rangle$, and we assume that the measurement projects A onto $|1\rangle$ (thus resulting in a defect here). All operations 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 $p_{\mathrm{DM}}^{\mathscr{L}}(Q_1)$ 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 $p_{\mathrm{DM}}^{\mathscr{L}}$ for all three high-frequency data qubits. **e** The average projection of the leakage probability $p_{\mathrm{DM}}^{\mathscr{L}}$ 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 $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$ was below a threshold $p_{\mathrm{th}}^{\mathscr{L}} = 0.5$ 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.

where we have set the off-diagonal terms containing a $|2\rangle$ to 0, consistently with the simulations in this chapter (in any case, they do not matter for the present discussion), see Sec. 2.4.2. Tracing out Q_2, the leakage probability of Q_1 is

$$p_{\mathrm{DM}}^{\mathscr{L}}(Q_1) = \frac{4L_1}{2(1 - \sqrt{1 - 4L_1}) + 4p(1 + \sqrt{1 - 4L_1})}, \tag{2.33}$$

where the denominator is just the probability of observing a defect. Thus, the product of this probability and of $p_{\mathrm{DM}}^{\mathscr{L}}(Q_1)$ 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 Sec. 2.5.3 for Surface-17. However, Eq. (2.33), plotted in Fig. 2.10 **a**, shows that $p_{\mathrm{DM}}^{\mathscr{L}}(Q_1)$, 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. 2.10 **a**, $p_{\mathrm{DM}}^{\mathscr{L}}(Q_1) = 4.5\%$, 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 $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$ of any of the data qubits sharply increases. An example of a leakage event of D_3 is shown in Fig. 2.10 **c**, where $p_{\mathrm{DM}}^{\mathscr{L}}(D_3)$ is sharply projected to a high value. However, during the initial projection, $p_{\mathrm{DM}}^{\mathscr{L}}(D_4)$ 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. 2.10 **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 $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$ of both data qubits. A second example of a realization of data-qubit leakage is shown in Fig. 2.10 **d**, where both D_3 and D_4 exhibit sharp and brief projections to $p_{\mathrm{DM}}^{\mathscr{L}}(Q) \approx 0.5$ 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, $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$ 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 $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$ throughout leakage events across individual realizations. The observed bi-modal density distribution shown in Fig. 2.3 **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 Sec. 2.2.3). To make the selection of leakage events (in Fig. 2.3 **d**, 2.5 **a**, 2.14 **b**, **c**, 2.10 **e**, 2.12 **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 chapter. Fig. 2.10 **e** shows that, in the absence of relaxation and decoherence ($T_1 = T_2 = \infty$), the average $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$ 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. 2.3 **d**, in agreement with the expectation based on Fig. 2.10 **a** for $p = 0$ and Eq. (2.33). The density distribution of all $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$ of the three high-frequency data qubits, shown in Fig. 2.10 **f**, while highly bi-modal is still supported on intermediate values between 0 and 1 of $p_{\mathrm{DM}}^{\mathscr{L}}(Q)$, contrarily to what Fig. 2.10 **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. 2.10 **c-d**.

2.5.3. 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 Sec. 2.5.1), 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:

$$
\begin{aligned}
\dot{p}^{\mathscr{C}} &= -p^{\mathscr{C}} \Gamma_{\mathscr{C} \to \mathscr{L}} + p^{\mathscr{L}} \Gamma_{\mathscr{L} \to \mathscr{C}}, \\
\dot{p}^{\mathscr{L}} &= p^{\mathscr{C}} \Gamma_{\mathscr{C} \to \mathscr{L}} - p^{\mathscr{L}} \Gamma_{\mathscr{L} \to \mathscr{C}}.
\end{aligned}
\tag{2.34}
$$

The steady-state condition is $\dot{p}^i = 0$ for $i = \mathscr{C}, \mathscr{L}$, resulting in the steady-state populations p_{ss}^i:

$$
\begin{aligned}
p_{ss}^{\mathscr{C}} &= \frac{\Gamma_{\mathscr{L} \to \mathscr{C}}}{\Gamma_{\mathscr{C} \to \mathscr{L}} + \Gamma_{\mathscr{L} \to \mathscr{C}}}, \\
p_{ss}^{\mathscr{L}} &= \frac{\Gamma_{\mathscr{C} \to \mathscr{L}}}{\Gamma_{\mathscr{C} \to \mathscr{L}} + \Gamma_{\mathscr{L} \to \mathscr{C}}}.
\end{aligned}
\tag{2.35}
$$

Considering the CZ error model in Sec. 2.2.1, for a qubit it approximately holds that

$$
\Gamma_{\mathscr{C} \to \mathscr{L}} \approx N_{\mathrm{flux}} L_1,
\tag{2.36}
$$

$$
\Gamma_{\mathscr{L} \to \mathscr{C}} \approx N_{\mathrm{flux}} L_2 + (1 - e^{-\frac{t_c}{T_1/2}}),
\tag{2.37}
$$

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. (2.37)) 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].
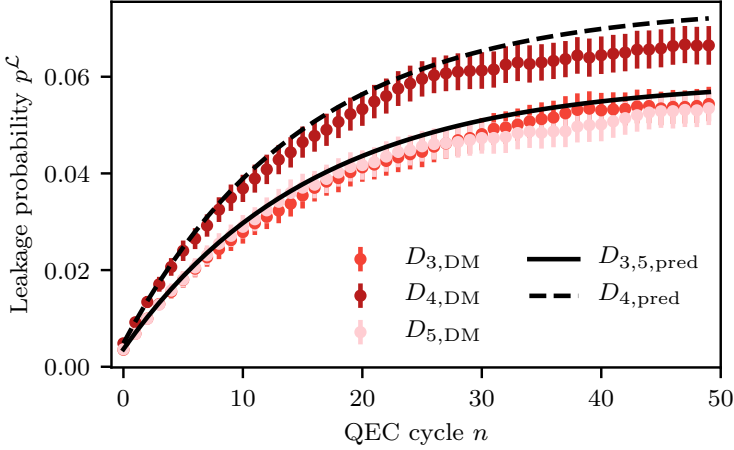
Figure 2.11: Evolution of the average leakage population $p^{\mathscr{L}}$ towards the steady state over 50 QEC cycles for the high-frequency data qubits in Surface-17. The leakage populations extracted from the density-matrix simulation (dots) agree well with the predicted one (black lines). The extracted populations are averaged over 4×10^4 runs. Error bars correspond to 2 standard deviations estimated by bootstrapping.

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 $N_{\mathrm{CZ}} = N_{\mathrm{flux}} = 4$ (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 $p_{ss}^{\mathscr{L}}(D_4) = 7.5\%$ and $p_{ss}^{\mathscr{L}}(D_3) = p_{ss}^{\mathscr{L}}(D_5) = 5.9\%$. Furthermore, Eq. (2.34) can be solved to find that the time evolution of $p^{\mathscr{L}}$ towards the steady state is

$$p^{\mathscr{L}}(n) = \frac{\Gamma_{\mathscr{C} \to \mathscr{L}}}{\Gamma_{\mathscr{C} \to \mathscr{L}} + \Gamma_{\mathscr{L} \to \mathscr{C}}} (1 - e^{-(\Gamma_{\mathscr{C} \to \mathscr{L}} + \Gamma_{\mathscr{L} \to \mathscr{C}})n}), \qquad (2.38)$$

where n is the QEC cycle number, shown in Fig. 2.11 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. 2.11).

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 $|3\rangle$, respec-

tively. The rate equations (Eq. (2.34)) are extended to

$$
\begin{aligned}
\dot{p}^{\mathscr{C}} &= -p^{\mathscr{C}}\Gamma_{\mathscr{C}\to\mathscr{L}_2} + p^{\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{C}}, \\
\dot{p}^{\mathscr{L}_2} &= p^{\mathscr{C}}\Gamma_{\mathscr{C}\to\mathscr{L}_2} - p^{\mathscr{L}_2}(\Gamma_{\mathscr{L}_2\to\mathscr{C}} + \Gamma_{\mathscr{L}_2\to\mathscr{L}_3}) + p^{\mathscr{L}_3}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2}, \\
\dot{p}^{\mathscr{L}_3} &= p^{\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3} - p^{\mathscr{L}_3}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2}.
\end{aligned}
\tag{2.39}
$$

The steady-state populations $\{p_{ss}^i\}$ then become:

$$
\begin{aligned}
p_{ss}^{\mathscr{C}} &= \frac{\Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2}}{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}}, \\
p_{ss}^{\mathscr{L}_2} &= \frac{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2}}{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}}, \\
p_{ss}^{\mathscr{L}_3} &= \frac{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}}{\Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{L}_2\to\mathscr{C}}\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} + \Gamma_{\mathscr{C}\to\mathscr{L}_2}\Gamma_{\mathscr{L}_2\to\mathscr{L}_3}}.
\end{aligned}
\tag{2.40}
$$

In addition to Eqs. (2.36) and (2.37), in this model we have

$$
\Gamma_{\mathscr{L}_2\to\mathscr{L}_3} \approx N_{\text{flux}}L_3/2,
\tag{2.41}
$$

$$
\Gamma_{\mathscr{L}_3\to\mathscr{L}_2} \approx N_{\text{flux}}L_3/2 + (1 - e^{-\frac{t_c}{T_{1/3}}}).
\tag{2.42}
$$

The factor of $1/2$ in Eq. (2.41) 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 $p_{ss}^{\mathscr{L}_2}(D_4) = 7.1\%$, $p_{ss}^{\mathscr{L}_3}(D_4) = 5.1\%$ 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.

2.5.4. 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 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
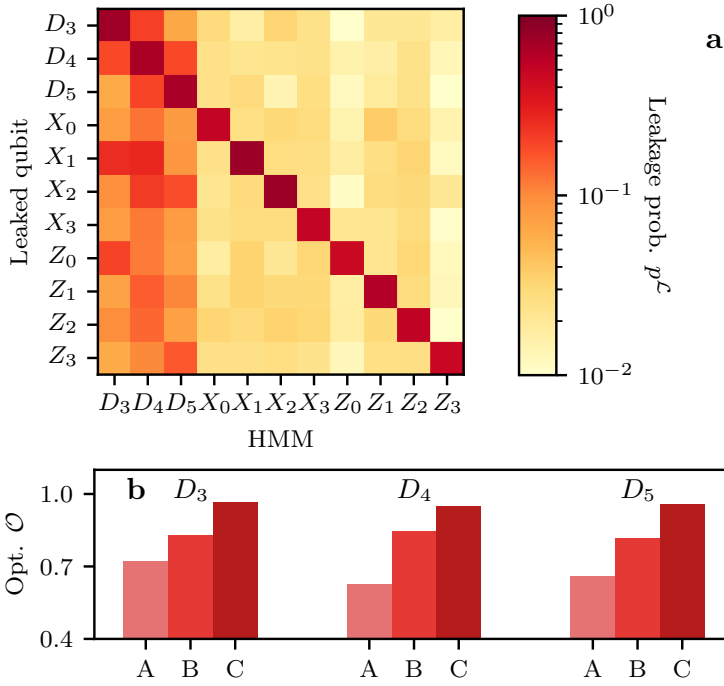
Figure 2.12:   The crosstalk between the HMMs.   **a** Average responses of all HMMs 1 QEC cycle after a given qubit leaks. We select individual realizations where the leakage probability $p^{\mathcal{L}}$ is first below and then above a threshold $p^{\mathcal{L}}_{\text{th}} = 0.5$ for 5 and 8 QEC cycles, 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.

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 $F^{\mathcal{L}}$, see Sec. 2.2.6). 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. 2.12 **b**. The responses of the neighboring HMMs im-

mediately (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. 5 and 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. 2.12). 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 $\mathcal{O}$ (defined in Sec. 2.2.5) of each of the data-qubit HMMs due to the crosstalk, shown in Fig. 2.12 **b**. Post-selecting out runs where ancilla-qubit leakage is detected from the density matrix increases the average $\mathcal{O}$ of the three data-qubit HMMs from $\mathcal{O} \approx 67.0\%$ to $\mathcal{O} \approx 83.3\%$. 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 $\mathcal{O} \approx 95.9\%$. The larger contribution from neighboring data-qubit leakage is consistent with the higher crosstalk (see Fig. 2.12 **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.

2.5.5. 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 $\mathcal{N}_i$ with mean $\vec{\mu}_i$ and standard deviation $\vec{\sigma}_i$ in the IQ plane [17, 58], as exemplified in Fig. 2.13 **a**.

Given an analog measurement of I and Q, the probability of a transmon being in state $|i\rangle$ can be expressed as

$$\mathbb{P}(i \mid I, Q) = \frac{\mathbb{P}(I, Q \mid i)\,\mathbb{P}(i)}{\mathbb{P}(I, Q)}, \tag{2.43}$$

where

$$\mathbb{P}(I, Q) = \sum_{j \in \{0,1,2\}} \mathbb{P}(I, Q \mid j)\,\mathbb{P}(j). \tag{2.44}$$
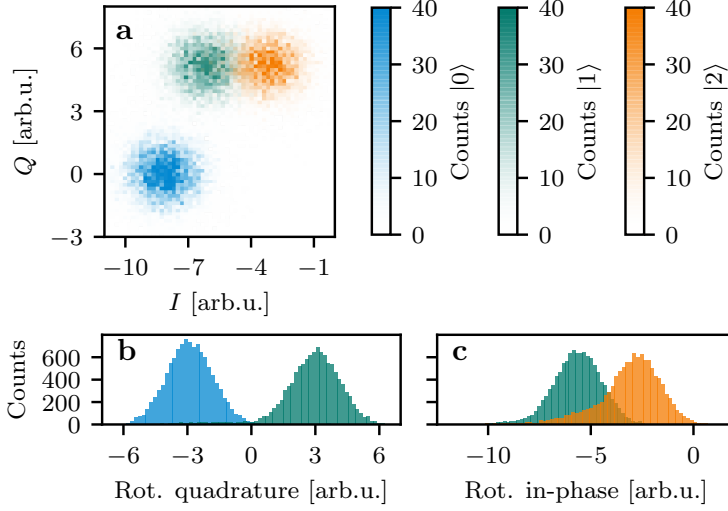
Figure 2.13: The analog measurement of transmons as extracted from experiment. **a** Histograms of the in-phase I and quadrature Q components of the measured readout for a transmon prepared in $|0\rangle$, $|1\rangle$ or $|2\rangle$. **b** The histograms of the responses for the transmon initially prepared 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

$$\mathbb{P}(i \mid I, Q) = \frac{\mathcal{N}_i(I, Q)}{\sum_{j \in \{0,1,2\}} \mathcal{N}_j(I, Q)}. \tag{2.45}$$

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

$$F^{ij} = 1 - \mathbb{P}(j \mid i)\mathbb{P}(i) - \mathbb{P}(i \mid j)\mathbb{P}(j), \tag{2.46}$$

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 Sec. 2.4.2). We assume $\mathbb{P}(i) = \mathbb{P}(j) = \frac{1}{2}$, which leads to

$$F^{ij} = 1 - \frac{\mathbb{P}(j \mid i) + \mathbb{P}(i \mid j)}{2}. \tag{2.47}$$

This can be related to the signal-to-noise ratio $\text{SNR} = \left| \vec{\mu}_i - \vec{\mu}_j \right| / 2\sigma$, assuming symmetric Gaussian distributions, as

$$F^{ij} = 1 - \frac{1}{2} \text{erfc} \left(\frac{\text{SNR}}{\sqrt{2}} \right). \tag{2.48}$$

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. (2.45) as

$$\mathbb{P}(i \mid I) = \frac{\mathcal{N}_i(I)}{\sum_{j \in \{0,1,2\}} \mathcal{N}_j(I)}, \tag{2.49}$$

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. 2.13 **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.

2.5.6. 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. 2.2, 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 Sec. 2.2.1, 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

$$|0\rangle \langle 0|_A \otimes \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}_D$$

$$+ |1\rangle \langle 1|_A \otimes \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -e^{-i\phi_{\text{stat}}^{\mathcal{L}}} \end{pmatrix}_D \tag{2.50}$$

$$=: |0\rangle \langle 0|_A \otimes \tilde{I}_D + |1\rangle \langle 1|_A \otimes \tilde{Z}_D. \tag{2.51}$$
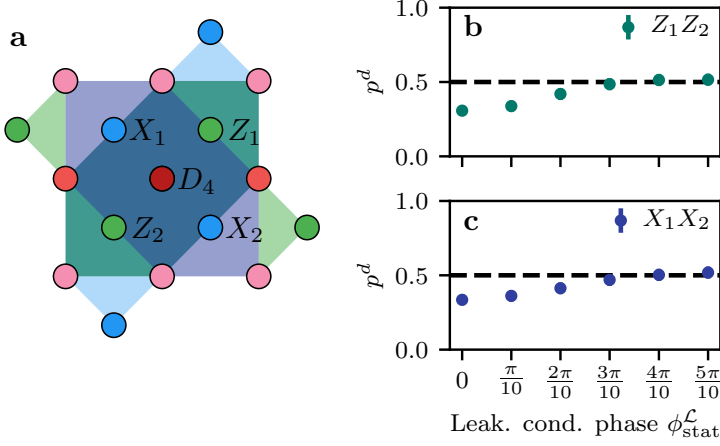
Figure 2.14: The effects of data-qubit leakage on the stabilizers of the code. **a** Sketch of how data-qubit leakage in the bulk (e.g. on D_4) effectively defines weight-3 gauge operators, whose product forms a weight-6 X-type (purple) or Z-type (teal) "supercheck" stabilizer, in addition to the standard 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 $\phi_{\text{stat}}^{\mathscr{L}}$.

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 [10], can be treated as a perturbation to this.

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

$$M_{\pm}^Z = \frac{\tilde{I}_{abcd} \pm \tilde{Z}_{abcd}}{2}, \tag{2.52}$$

where $\tilde{I}_{abcd} := \tilde{I}_a \tilde{I}_b I_c I_d$ and $\tilde{Z}_{abcd} := \tilde{Z}_a \tilde{Z}_b Z_c Z_d$. 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

$$M_{\pm}^X = \frac{\tilde{I}_{abcd} \pm \tilde{X}_{abcd}}{2}, \tag{2.53}$$

where $\tilde{X}_{abcd} := \tilde{X}_a \tilde{X}_b X_c X_d$ and

$$\tilde{X} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -e^{-i\phi_{\text{stat}}^{\mathscr{L}}} \end{pmatrix}, \tag{2.54}$$

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 $M_{\pm}^Z$ and $M_{\pm}^X$ commute, as it holds

$$\left[M_{\pm}^Z, M_{\pm}^X \right] = \frac{1}{4} \left[\tilde{Z}_{abcd}, \tilde{X}_{abcd} \right] \tag{2.55}$$

(and also $\left[M_{\pm}^Z, M_{\pm}^X \right] = - \left[M_{\pm}^Z, M_{\mp}^X \right]$). To compute the commutator we first evaluate

$$\left[\tilde{Z}, \tilde{X} \right] = 2\mathrm{i} \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.56}$$

$$\{ \tilde{Z}, \tilde{X} \} = 2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & e^{-2\mathrm{i}\phi_{\mathrm{stat}}^{\mathscr{L}}} \end{pmatrix}. \tag{2.57}$$

It follows that

$$\{ \tilde{Z}, \tilde{X} \} \Big|_{\mathscr{C}} = \{Z, X\} = 0, \tag{2.58}$$

$$\left[\tilde{Z}, \tilde{X} \right] \Big|_{\mathscr{L}} = 0, \tag{2.59}$$

when restricted to the computational and leakage subspaces, respectively. Notice that, when all data qubits are in the computational subspace, it holds

$$\{ \tilde{Z}_{abcd}, \tilde{X}_{abcd} \} \Big|_{\mathscr{C}} = \{Z_{abcd}, X_{abcd}\} = 0 \tag{2.60}$$

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. (2.59) 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

$$\{ \tilde{Z}_{abcd}, \tilde{X}_{abcd} \} \Big|_{\mathscr{L}_a} = \left\{ -e^{-\mathrm{i}\phi_{\mathrm{stat}}^{\mathscr{L}}} Z_{bcd}, -e^{-\mathrm{i}\phi_{\mathrm{stat}}^{\mathscr{L}}} X_{bcd} \right\}$$

$$= e^{-\mathrm{i}\phi_{\mathrm{stat}}^{\mathscr{L}}} \{Z_{bcd}, X_{bcd}\} = 0. \tag{2.61}$$

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

$$M_{\pm}^Z |_{\mathscr{L}_a} = \frac{I_{bcd} \pm e^{-\mathrm{i}\phi_{\mathrm{stat}}^{\mathscr{L}}} Z_{bcd}}{2} \tag{2.62}$$

and similarly for $M_{\pm}^X |_{\mathscr{L}_a}$.

For $\phi_{\mathrm{stat}}^{\mathscr{L}} = 0, \pi$, $M_{\pm}^X |_{\mathscr{L}_a}$ 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. (2.61)]

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. 2.14 **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 $\phi^{\mathscr{L}}_{\mathrm{stat}} \neq 0, \pi$, 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 $\phi^{\mathscr{L}}_{\mathrm{stat}}$ is randomized across runs (see Fig. 2.3) 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 $\phi^{\mathscr{L}}_{\mathrm{stat}}$ 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. 2.14 **b,c**, where in the case of $\phi^{\mathscr{L}}_{\mathrm{stat}} = 0, \pi$ 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 $\phi^{\mathscr{L}}_{\mathrm{stat}}$ in experiment would be beneficial for decoding in the presence of data-qubit leakage whenever the superchecks are considered.

2.5.7. 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 $L_m \in [0, 1.5\%]$ for a fixed leakage probability $L_1 = 0.125\%$ and randomized leakage-conditional phases $\phi_{\text{stat}}^{\mathscr{L}}$ and $\phi_{\text{flux}}^{\mathscr{L}}$ (see Sec. 2.2.1). 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 Sec. 2.5.1) 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.

2.5.8. 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 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. 2.6 **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 Sec. 2.2.5) to $\mathcal{O}(D) \approx 31.2\%$. While such scheme may be beneficial for the post-selection-based scheme defined in Sec. 2.2.7 (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.

BIBLIOGRAPHY

[1] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. Di-Carlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Information **6**, 102 (2020).

[2] A. D. Córcoles, 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**, 1338–1352 (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, 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. Land-huis, 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 super-conducting processor*, Nature **574**, 505–510 (2019).

[4] J. S. Otterbach, R. Manenti, N. Alidoust, A. Bestwick, M. Block, B. Bloom, S. Cald-well, 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 thresh-old 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*, Phys. Rev. A **93**, 060302(R), 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. App. **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. App. **10**, 034040 (2018).

[18] 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**, 66–69 (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, et al., *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**, 1–7 (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**, 527–531 (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*, Nat. Phys. **16**, 875–880 (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*, Nat. Phys. **14**, 595–600 (2018).

2

[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**, 195–199 (2018).

[27] S. Bravyi, D. Gosset, and R. König, *Quantum advantage with shallow circuits*, Science **362**, 308–311 (2018).

[28] T. E. O'Brien, B. M. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Information **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,** 139–156 (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(R) (2015).

[41] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code*, Quantum Info. Comput. **15,** 997–1016 (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. Applied **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. Spitz, B. M. Tarasinski, C. Beenakker, and T. 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).

2

[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 allmicrowave reset of a superconducting qubit*, Phys. Rev. Lett. **121**, 060502 (2018).

[57] F. Luthi, T. Stavenga, O. Enzing, A. Bruno, C. Dickel, N. 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).

3

A 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). In this chapter, 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.

3.1. INTRODUCTION

Quantum computing has recently achieved 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 [31, 32] 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 [25].

Leakage outside of the computational subspace [12–14, 16, 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 [40]. Furthermore, leakage can last for many QEC cycles [44], making operations on a leaked qubit fail and possibly spread correlated errors through the code [25, 40, 43]. 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 [40, 46, 49]. If the average leakage lifetime $l_{\mathrm{avg}}^{\mathscr{L}}$, that is, the average number of QEC cycles that a qubit stays leaked (after leaking in the first place), fulfills $l_{\mathrm{avg}}^{\mathscr{L}} = \mathscr{O}(1)$ QEC cycles and $l_{\mathrm{avg}}^{\mathscr{L}} \ll d$, then for low-enough error rates a threshold is likely to exist [43] as leakage would have a relatively local effect in space and time. Due to a finite energy-relaxation time, leakage does indeed last for $l_{\mathrm{avg}}^{\mathscr{L}} = \mathscr{O}(1)$ QEC cycles. However, in practice it is important how large $l_{\mathrm{avg}}^{\mathscr{L}}$ is, since if it is low the noise threshold is expected to be higher. Shortening the relaxation time to reduce $l_{\mathrm{avg}}^{\mathscr{L}}$ is not effective as this increases the physical error rate as well.

A leakage-reduction unit (LRU) [39, 41–43, 45, 46, 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 [40] 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 $l_{\mathrm{avg}}^{\mathscr{L}} = \mathscr{O}(1) \ll d$, 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 [41, 45, 46] 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 $|3\rangle$ [42]) to $|0\rangle$ [54–56], or conditionally using real-time feedback [30, 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 [42]), $l_{\mathrm{avg}}^{\mathscr{L}}$ 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 [12–14, 16, 33, 35–37]. Moreover, in the surface code an extra row of qubits is needed to perform all the SWAPs [45], 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 [40, 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, 43, 46] 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, 43, 46, 60–62] that uses measurement information about leakage to better correct leakage-induced correlated

errors.

3.2. 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. 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.

3.2.1. 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

$$H = H_0 + H_c + H_d \tag{3.1}$$

$$H_0 = \delta^r a^\dagger a + \delta^q b^\dagger b + \frac{\alpha}{2}(b^\dagger)^2 b^2 \tag{3.2}$$

$$H_c = g(ab^\dagger + a^\dagger b) \tag{3.3}$$

$$H_d = \frac{\Omega}{2}(e^{i\phi}b + e^{-i\phi}b^\dagger) \tag{3.4}$$

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 work and we fix it to $\phi = 0$.

We can qualitatively understand (see Fig. 3.1(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. (3.1) 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. (3.62)). 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. 3.1(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
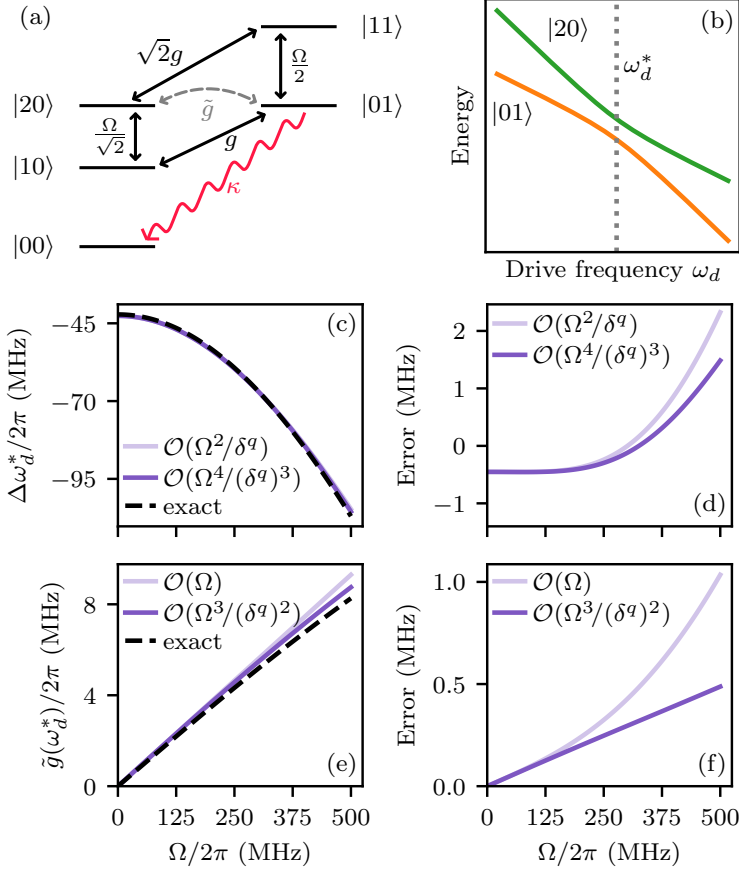
Figure 3.1: Concept of the readout-resonator LRU. (a) The state $|20\rangle$ (with the notation $|\text{transmon, resonator}\rangle$) is connected to $|01\rangle$ by two main paths via either $|11\rangle$ or $|10\rangle$, due to the capacitive coupling g or the transmon-drive amplitude Ω, respectively. 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. (3.1), or by approximate analytical formulas (see Sec. 3.2.1 and Sec. 3.5) for the parameters in Tab. 3.1. The absolute errors with respect to the exact curves are shown in (d),(f) respectively.

Schrieffer-Wolff transformation e^S [63–66]. Let $\{|ij\rangle_D\}$ be the basis of eigenvectors 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 parame-

ter g/Δ, such that $e^{-S_1}|ml\rangle \approx |ml\rangle_D$, we get (see Sec. 3.5)

$$H^D := e^S H e^{-S} \approx e^{S_1} H e^{-S_1} \tag{3.5}$$

$$= H_0^D + H_{d1}^D + H_{d2}^D \tag{3.6}$$

with

$$H_0^D = \left(\delta^r - \sum_{m=0}^{\infty} \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m|\right) a^\dagger a$$

$$+ \sum_{m=1}^{\infty} \left(m\delta^q + \frac{\alpha}{2}m(m-1) + \frac{g^2 m}{\Delta_{m-1}}\right)|m\rangle\langle m| \tag{3.7}$$

$$H_{d1}^D = \frac{\Omega e^{i\phi}}{2} b + \text{h.c.} \tag{3.8}$$

$$H_{d2}^D = \frac{\Omega e^{i\phi}}{2}\left(a \sum_{m=0}^{\infty} \frac{g\Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m|\right.$$

$$+ a^\dagger \sum_{m=0}^{\infty} \frac{g\alpha\sqrt{m+1}\sqrt{m+2}}{\Delta_m \Delta_{m+1}} |m\rangle\langle m+2|\right) + \text{h.c.}, \tag{3.9}$$

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. (3.9) we get a lowest order approximation of $\tilde{g}$:

$$\tilde{g} \approx \frac{\Omega g \alpha}{\sqrt{2}\Delta(\Delta + \alpha)}. \tag{3.10}$$

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. 3.1(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. (3.2). This value is shifted due to the capacitive coupling (as can be seen from Eq. (3.7)), 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 [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 Sec. 3.5. We also use this transformation to compute $\tilde{g}$ up to order $\Omega^3/(\delta^q)^2$. Figs. 3.1(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 Tab. 3.1. We consider 6 energy levels for the transmon and 3 for the resonator as we see that the

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 3.1: Parameters used both in the analysis and Lindblad simulations of the readout-resonator LRU, similar to the experimental ones in [29]. The transmon parameters correspond to the target parameters of a high-frequency data qubit in Sec. 3.3.

exact curves converge for such choice. In Fig. 3.1(c)(d) we see that the two approximations are both pretty good, while in Fig. 3.1(e)(f) we see that Eq. (3.10) 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. 3.1(d).

3.2.2. 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

$$\Omega(t) = \begin{cases} \Omega \sin^2(\pi \frac{t}{2t_{\text{rise}}}) & \text{for} \quad 0 \le t \le t_{\text{rise}} \\ \Omega & \text{for} \quad t_{\text{rise}} \le t \le t_{\text{p}} - t_{\text{rise}} \\ \Omega \sin^2(\pi \frac{t_{\text{p}}-t}{2t_{\text{rise}}}) & \text{for} \quad t_{\text{p}} - t_{\text{rise}} \le t \le t_{\text{p}} \end{cases} \quad (3.11)$$

similarly to [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 [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 Sec. 3.3.1).

We use Lindblad simulations of the transmon-resonator system to optimize over Ω, ω_d
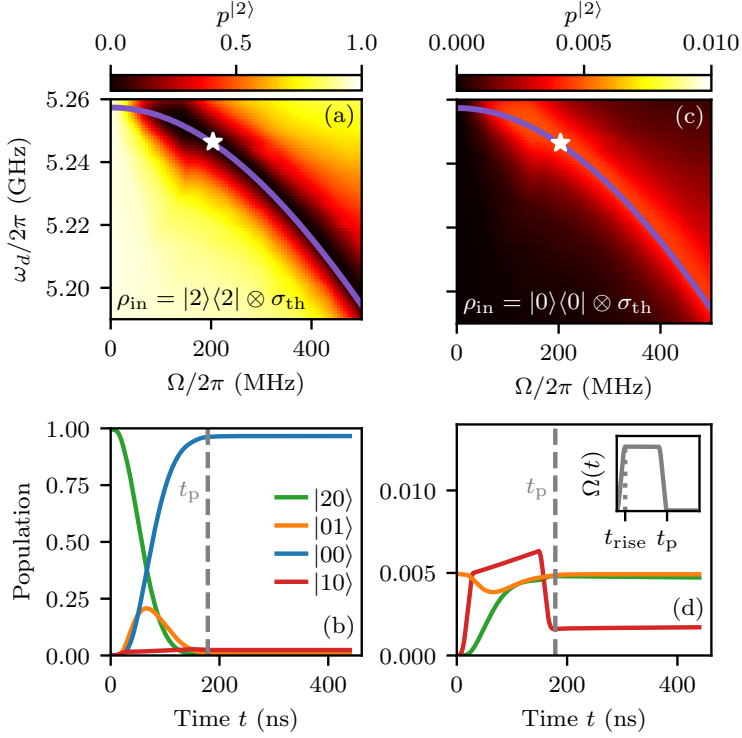
Figure 3.2: Lindblad simulations of the transmon-resonator system for the readout-resonator LRU. In (a),(b) the initial state is $|2\rangle\langle2| \otimes \sigma_{\rm th}$, while in (c),(d) it is $|0\rangle\langle0| \otimes \sigma_{\rm th}$, where $\sigma_{\rm th}$ is the resonator thermal state. (a),(c) Transmon leakage population $p^{|2\rangle} = \langle2|\,{\rm Tr}_r(\rho(T_{\rm slot}))|2\rangle$ at the end of the time slot of $T_{\rm slot} = 440$ ns. For each choice of (Ω, ω_d) we optimize the total pulse duration $t_{\rm p} \leq T_{\rm slot}$ to minimize $p^{|2\rangle}$ given the initial state $|2\rangle\langle2| \otimes \sigma_{\rm th}$, for fixed $t_{\rm rise} = 30$ ns. The white star indicates the chosen operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_{\rm p} = 178.6$ ns) with $p^{|2\rangle}_{\rm op.} \approx 0.5\%$ 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. 3.1(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 $t_{\rm p}$. The inset in (d) shows a schematic of the pulse $\Omega(t)$.

and $t_{\rm p}$. The Lindblad equation is given by

$$\dot{\rho} = -i\left[H^D, \rho\right] + \sum_j \left(K_j \rho K_j^\dagger - \frac{1}{2}\{K_j^\dagger K_j, \rho\}\right) \tag{3.12}$$

with $\{K_j\}$ the quantum jump operators. We express (and solve) this equation in the exact unitarily transformed frame. That is, while in Sec. 3.2.1 we have used a first-order Schrieffer-Wolff transformation e^{S_1} (see Eq. (3.5)), in the numerics we compute the full transformation e^S (see also Eq. (3.5)). 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 Sec. 3.2.1. In other words, the simulations reproduce the dynam-

ics under the Hamiltonian in Eqs. (3.1) to (3.4) without any approximation.

The Hamiltonian parameters are the same as in Sec. 3.2.1 and are reported in Tab. 3.1, 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]

$$\sigma_{\text{th}} \approx \left(1 - \frac{\bar{n}}{1+2\bar{n}}\right)|0\rangle\langle0| + \frac{\bar{n}}{1+2\bar{n}}|1\rangle\langle1| \tag{3.13}$$

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 $\{K_j\}$ are explicitly given by

$$\frac{1}{\sqrt{T_1^r}}a = \sqrt{\kappa}\,a, \quad \sqrt{\frac{\bar{n}}{1+\bar{n}}}\sqrt{\kappa}\,a^\dagger, \quad \sqrt{\frac{2}{T_\phi^r}}a^\dagger a, \tag{3.14}$$

$$\frac{1}{\sqrt{T_1^q}}b, \quad \sqrt{\frac{2}{T_\phi^q}}b^\dagger b, \tag{3.15}$$

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 $e^{-S}ae^S = \sum_{l=0}^1 \sqrt{l+1}|l\rangle_D\langle l+1|_D = a_D$ 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 $|2\rangle\langle2|\otimes\sigma_{\text{th}}$, the leakage population $p^{|2\rangle} = \langle2|\text{Tr}_r(\rho(T_{\text{slot}}))|2\rangle$ at the end of the available time slot is minimized (see Fig. 3.2(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 Tab. 3.1 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 [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 $\pi/2\tilde{g}_{\mathrm{damp}}$ with $\tilde{g}_{\mathrm{damp}} := \sqrt{\tilde{g}^2 - (\kappa/4)^2}\, e^{-\kappa/7\tilde{g}}$ for larger κ. Then for the optimization over t_{p} we use the bounds $t_{\mathrm{p}} - 2t_{\mathrm{rise}} \in [0, 1.1 \times \pi/2\tilde{g}_{\mathrm{damp}}]$ (using the bounded Brent method in *scipy*; see [70] for the code). 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 Sec. 3.6.2).

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 $\Omega/2\pi \in [0, 500\,\mathrm{MHz}]$, $\omega_d/2\pi \in [5.19, 5.26\,\mathrm{GHz}]$) 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. 3.2) 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. 3.2(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 Sec. 3.2.1. 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. 3.2 as the operating point ($\Omega/2\pi \approx 204\,\mathrm{MHz}$, $\omega_d/2\pi \approx 5.2464\,\mathrm{GHz}$, $t_{\mathrm{p}} = 178.6\,\mathrm{ns}$). This point reaches $p^{|2\rangle}_{\mathrm{op.}} \approx 0.5\%$ while affecting the least the coherence within the computational subspace (see Sec. 3.6.1). 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. 3.2(a) we find good $p^{|2\rangle} \lesssim 5\%$ up to $\Omega/2\pi \gtrsim 100\,\mathrm{MHz}$, which could be used to further ease the requirements on the drive (see Sec. 3.3).

The time evolution for a few selected states is shown in Fig. 3.2(b) for the operating point, given the initial state $|2\rangle\langle2| \otimes \sigma_{\mathrm{th}}$. 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 Sec. 3.3) at every QEC cycle, thus does not lead to heating of the resonator with these system parameters (see Sec. 3.4 for a discussion about other parameter regimes).

We now evaluate the effect of the pulse on a non-leaked transmon (see Fig. 3.2(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 pro-

cess by unitarity. Fig. 3.2(c) shows that indeed the induced leakage is greater where $p^{|2\rangle}$ is lower in Fig. 3.2(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. 3.2(c) at the operating point, which is comparable to state-of-the-art CZ leakage rates, see Sec. 3.3.2) 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. 3.2(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 Sec. 3.6.1). 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.

3.3. SURFACE CODE WITH LRUS

3.3.1. LAYOUT AND OPERATION SCHEDULING

We study the distance-3 rotated surface code (see Fig. 3.3(a)), nicknamed Surface-17, in the presence of leakage and LRUs. We follow the frequency and pipelined scheme in [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 [13, 14, 33, 35, 36] 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 [40], 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 [40, 42]) 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. 3.3(b). The X-type and Z-type parity-check units are implemented in an interleaved way, with the CZs for one unit being applied while the other ancilla-qubit type is measured. The duration of each operation is summarized in Sec. 3.7.1, 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 Sec. 3.2, to the high-frequency data qubits. Notice that the optimal pulse selected in Sec. 3.2.2, which was simulated for the target parameters of the high-frequency data qubits, takes about $t_p = 180$ ns and thus easily fits within this time slot (see Sec. 3.4 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
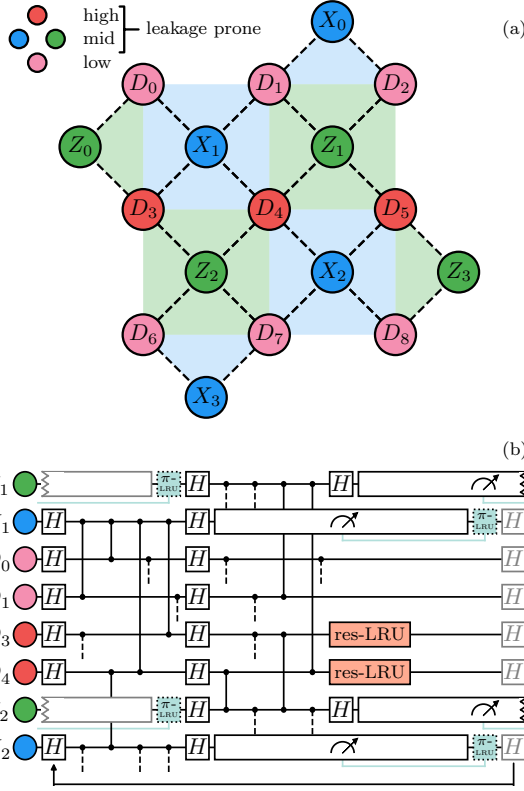
Figure 3.3: (a) Schematic overview of the Surface-17 layout [40, 58]. Pink (resp. red) circles with D labels represent low- (high-) frequency data qubits, while blue (resp. green) circles with X (Z) labels represent ancilla qubits, which have an intermediate frequency. Ancilla qubits and high-frequency 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 [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 Sec. 3.7.1. The arrow at the bottom indicates the repetition of QEC cycles.

(analog) measurement outcome contains information about leakage [40]. 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.

3.3.2. 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 [40]. L_1 is defined as the average leakage from the computational to the leakage subspace [71]. The state of the art is $L_1 \approx 0.1\%$ [13, 14], although the actual L_1 is expected to be higher when operating a multi-transmon processor [32, 72], 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 [9, 34, 38]. The noise parameters used are reported in Sec. 3.7.1. Furthermore, during a CZ with a leaked transmon, the non-leaked transmon acquires a phase called the leakage conditional phase [40]. We select these phases uniformly at random (see Sec. 3.7.3) and, in contrast to [40], 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 Sec. 3.7.3 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 [40] 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 Sec. 3.7.1 for details). Furthermore, in Sec. 3.2.2 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 [71]), with probability $1/2$ each. Since almost all induced leakage comes from $|0\rangle$ (see Sec. 3.2.2), this means that $p_f^{|2\rangle} \approx 0$ for $p_i^{|1\rangle} = 1$ and that $p_f^{|2\rangle} \approx 2L_1^{\text{LRU}}$ 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

$$p_f^{|2\rangle} \approx (1 - R)\, p_i^{|2\rangle} + 2L_1^{\text{LRU}}\, p_i^{|0\rangle} \tag{3.16}$$

for an arbitrary incoming state. Notice that, given these definitions, Fig. 3.2(a),(c) respectively show a heatmap of $1 - R$ and $2L_1^{\text{LRU}}$ for the considered transmon-resonator parameters. In particular, the operating point achieves $R \approx 99.5\%$ and $L_1^{\text{LRU}} \approx 0.25\%$. The achieved leakage reduction can be compared with the one given purely by relaxation during T_{slot}, namely $R_{T_1} = 1 - e^{-T_{\text{slot}}/(T_1/2)} = 2.9\%$, 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 [30, 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 $1 - e^{-t_{\text{extra}}/T_2}$ 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. 3.4 and 3.9).

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

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & p_M(1|1) & 1 - p_M(1|1) \\ 0 & 1 - p_M(2|2) & p_M(2|2) \end{pmatrix}. \tag{3.17}$$

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 chapter) 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.

3.3.3. 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 [40]. 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 [40] 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

$$\Gamma_{\mathscr{C} \to \mathscr{L}} \approx N_{\text{flux}} L_1, \tag{3.18}$$

$$\Gamma_{\mathscr{L} \to \mathscr{C}} \approx N_{\text{flux}} L_2 + (1 - e^{-\frac{t_c}{T_{1/2}}}), \tag{3.19}$$

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 [71]. 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. (3.19) by introducing an extra seepage mechanism. Hence we expect that $\Gamma_{\mathscr{L} \to \mathscr{C}}^{\text{LRU}} \sim \Gamma_{\mathscr{L} \to \mathscr{C}} + R$ for data qubits and $\Gamma_{\mathscr{L} \to \mathscr{C}}^{\text{LRU}} \sim \Gamma_{\mathscr{L} \to \mathscr{C}} + p_M(2|2)$ 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

$$l_{\text{avg}}^{\mathscr{L}} = \sum_{n=1}^{\infty} n \, \mathbb{P}(\text{stay in } \mathscr{L} \text{ for } n \text{ QEC cycles}) \tag{3.20}$$

$$= \sum_{n=1}^{\infty} n (1 - \Gamma_{\mathscr{L} \to \mathscr{C}})^{n-1} \Gamma_{\mathscr{L} \to \mathscr{C}} = \frac{1}{\Gamma_{\mathscr{L} \to \mathscr{C}}}, \tag{3.21}$$

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 number of QEC cycles n is well-approximated by [40]

$$\bar{p}^{\mathscr{L}}(n) = \frac{\Gamma_{\mathscr{C} \to \mathscr{L}}}{\Gamma_{\mathscr{C} \to \mathscr{L}} + \Gamma_{\mathscr{L} \to \mathscr{C}}} (1 - e^{-(\Gamma_{\mathscr{C} \to \mathscr{L}} + \Gamma_{\mathscr{L} \to \mathscr{C}})n}). \tag{3.22}$$

The steady state is the long-time limit and is given by

$$\bar{p}_{\text{ss}}^{\mathscr{L}} = \lim_{n \to \infty} \bar{p}^{\mathscr{L}}(n) = \frac{\Gamma_{\mathscr{C} \to \mathscr{L}}}{\Gamma_{\mathscr{C} \to \mathscr{L}} + \Gamma_{\mathscr{L} \to \mathscr{C}}}. \tag{3.23}$$

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.

Fig. 3.4 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. (3.22) 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

$$l_{\text{avg}}^{\mathscr{L}} = \frac{1}{\Gamma_{\mathscr{L}\to\mathscr{C}}^{\text{LRU}}} \sim \frac{1}{\Gamma_{\mathscr{L}\to\mathscr{C}} + R} \sim \frac{1}{R} \tag{3.24}$$

$$\bar{p}_{\text{ss}}^{\mathscr{L}} = \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}}{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}} + \Gamma_{\mathscr{L}\to\mathscr{C}}^{\text{LRU}}} \sim \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}}{\Gamma_{\mathscr{L}\to\mathscr{C}}^{\text{LRU}}} \sim \frac{\Gamma_{\mathscr{C}\to\mathscr{L}}^{\text{LRU}}}{R} \tag{3.25}$$

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). Figs. 3.4(b),(d) show that this is indeed the case.

Fig. 3.4 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.
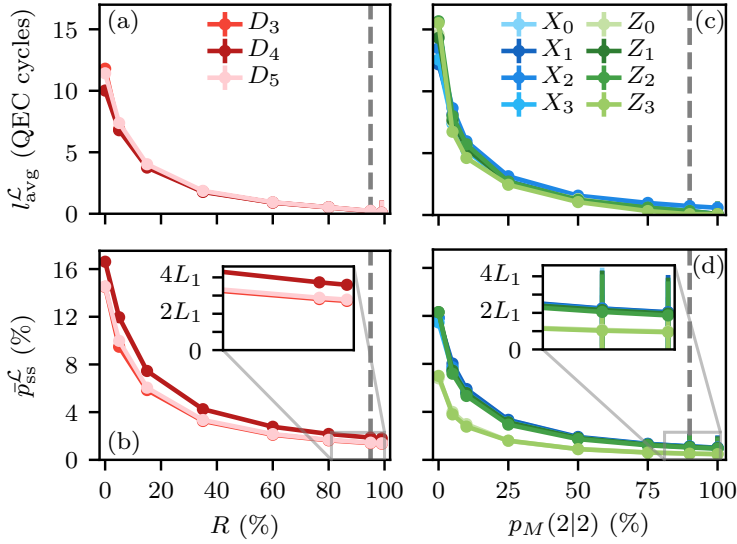
Figure 3.4: Average leakage lifetime $l_{\text{avg}}^{\mathscr{L}}$ [(a),(c)] and leakage steady state $\bar{p}_{\text{ss}}^{\mathscr{L}}$ [(b),(d)] as a function of the leakage-reduction rate R for data qubits [(a),(b)] and as a function of the readout probability $p_M(2|2)$ for ancilla qubits [(c),(d)]. Here we fix the CZ leakage rate to $L_1 = 0.5\%$. The insets in (b),(d) show that $\bar{p}_{\text{ss}}^{\mathscr{L}}$ tends to $\approx N_{\text{flux}}L_1$ ($N_{\text{flux}} = 4$ 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 Sec. 3.3.4. 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.

3.3.4. Logical performance

In the simulations the logical qubit is initialized in $|0\rangle_L$ and the logical fidelity $\mathscr{F}_L(n)$ 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 $|+\rangle_L$ 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 $\mathscr{F}_L(n) = [1 + (1 - 2\varepsilon_L)^{n-n_0}]/2$, 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, 73] and an overview is given in Sec. 3.7.1.

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 Sec. 3.2.2 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 [7, 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 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 Sec. 3.7.2.

Fig. 3.5 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.

3.4. 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 [42] 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 significantly 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
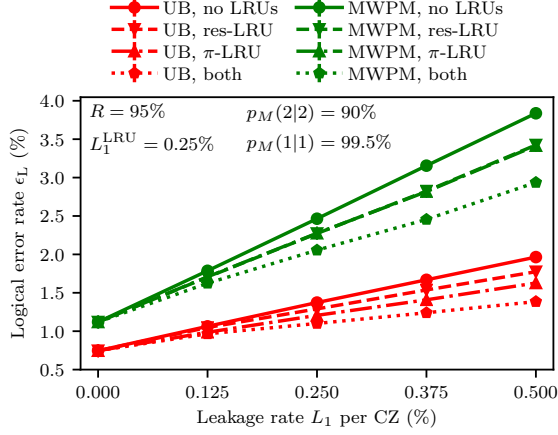
Figure 3.5: Logical error rate ε_{L} per QEC cycle for the upper bound (UB, red) and minimum-weight perfect-matching (MWPM, green) decoders versus the CZ leakage rate L_1, in the cases with: no LRUs, only res-LRU, only π-LRU and both LRUs (the point without leakage at $L_1 = 0$ is always without LRUs 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.

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 [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 [29]. 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 Sec. 3.2.2 (≈ 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 [74]), or they can be mitigated altogether by using tunable couplers [2, 7, 15]. The res-LRU is compatible with tunable-coupler schemes and their possibly different operation scheduling than in [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 Sec. 3.6.3). 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 [40].

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 $T_{\mathrm{slot}} = t_{\mathrm{m}} - 4 t_{\mathrm{CZ}}$, 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 [16]. 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 $T_{\mathrm{slot}} - t_{\mathrm{p}} \sim 16/\kappa$ and $t_{\mathrm{p}} < t_{\mathrm{read}}$. 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. $T_{\mathrm{slot}} - t_{\mathrm{p}} + 8 t_{\mathrm{CZ}} + 2 t_{\mathrm{H}} \geq 4/\kappa$. 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 ($d \gg l_{\mathrm{avg}}^{\mathscr{L}}$). 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, 43, 46]. We expect that the demonstrated MWPM logical error rate can be further lowered by the use of decoders [24, 43, 46, 60–62] that use information about leakage extracted directly or indirectly (e.g. with hidden Markov models [40]) from the measurement outcomes.

3.5. APPROXIMATE TRANSMON-RESONATOR HAMILTONIAN

3.5.1. 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

$$H = H_0 + \epsilon V \tag{3.26}$$

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

$$H' := e^S H e^{-S} \tag{3.27}$$

is block diagonal. In other words, using $\{|\bar{\psi}_n\rangle\}$ to refer to 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

$$S = \sum_{k=1}^{\infty} \epsilon^k S_k \tag{3.28}$$

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

$$H' = e^S H e^{-S} = \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{[S, [S, \dots [S,}_{k \text{ times}} H] \dots]]. \tag{3.29}$$

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

$$[\text{diagonal}, \text{diagonal}] = \text{diagonal}, \tag{3.30}$$

$$[\text{diagonal}, \text{off-diagonal}] = \text{off-diagonal}, \tag{3.31}$$

$$[\text{off-diagonal}, \text{off-diagonal}] = \text{diagonal}. \tag{3.32}$$

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 $O = O_D + O_{OD}$, respectively.

Here we expand H and S up to $k = 3$ in Eq. (3.29), assuming that the 4th-order block off-diagonal term is negligible. We get the following pieces:

$$\text{0th order:} \quad H_0 \tag{3.33}$$

$$\text{1st order:} \quad V + [S_1, H_0] \tag{3.34}$$

$$\text{2nd order:} \quad [S_1, V] + \frac{1}{2}[S_1, [S_1, H_0]] + [S_2, H_0] \tag{3.35}$$

$$\text{3rd order:} \quad [S_2, V] + \frac{1}{2}\Big([S_2, [S_1, H_0]] + [S_1, [S_1, V]] + [S_1, [S_2, H_0]]\Big)$$
$$+ \frac{1}{6}[S_1, [S_1, [S_1, H_0]]] + [S_3, H_0] \tag{3.36}$$

$$\text{4th order:} \quad [S_3, V] + \frac{1}{2}\Big([S_1, [S_3, H_0]] + [S_2, [S_2, H_0]] + [S_3, [S_1, H_0]] + [S_1, [S_2, V]]$$
$$+ [S_2, [S_1, V]]\Big)$$
$$+ \frac{1}{6}\Big([S_1, [S_1, [S_1, V]]] + [S_2, [S_1, [S_1, H_0]]] + [S_1, [S_2, [S_1, H_0]]]$$
$$+ [S_1, [S_1, [S_2, H_0]]]\Big)$$
$$+ \frac{1}{24}[S_1, [S_1, [S_1, [S_1, H_0]]]]. \tag{3.37}$$

Setting the block off-diagonal parts at 1st, 2nd and 3rd order to 0 we get

$$[H_0, S_1] = V \tag{3.38}$$

$$[H_0, S_2] = \frac{1}{2}[S_1, V]_{OD} \tag{3.39}$$

$$[H_0, S_3] = \frac{1}{2}[S_2, V]_{OD} + \frac{1}{3}[S_1, [S_1, V]_D]_{OD}$$
$$+ \frac{1}{12}\big[S_1, [S_1, V]_{OD}\big]_{OD}, \tag{3.40}$$

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

$$H' = H_0 + \frac{\epsilon^2}{2}[S_1, V]_D$$
$$+ \epsilon^3\Big(\frac{1}{2}[S_2, V]_D + \frac{1}{12}\big[S_1, [S_1, V]_{OD}\big]_D\Big)$$
$$+ \epsilon^4\Big(\frac{1}{2}[S_3, V]_D - \frac{1}{24}\big[S_1, [S_1, [S_1, V]_D]_{OD}\big]_D$$
$$- \frac{1}{6}\big[S_2, [S_1, V]_{OD}\big]_D + \frac{1}{12}\big[S_1, [S_2, V]_{OD}\big]_D\Big). \tag{3.41}$$

This expression has been simplified using Eqs. (3.38) to (3.40), together with the fact that e.g. $[S_k, [\ldots, \ldots]_D]_D = 0$ since S_k is block off-diagonal.

3.5.2. 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. (3.1) to (3.4).

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]

$$S_1 = g \sum_{m=1}^{\infty} \frac{\sqrt{m}}{\Delta + \alpha(m-1)} \Big(a|m\rangle\langle m-1| - \text{h.c.} \Big), \tag{3.42}$$

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 Sec. 3.2.1 is then given by

$$H^D \approx e^{S_1} H e^{-S_1} = e^{S_1}(H_0 + H_c)e^{-S_1} + e^{S_1} H_d e^{-S_1} \tag{3.43}$$

with

$$e^{S_1}(H_0 + H_c)e^{-S_1} = H_0 + \frac{1}{2}[S_1, H_c] \tag{3.44}$$

$$\approx \delta^r a^\dagger a + \sum_{m=1}^{\infty} \left(m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}} \right) |m\rangle\langle m|$$

$$- a^\dagger a \sum_{m=0}^{\infty} \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle\langle m| \tag{3.45}$$

$$:= H_0^D \tag{3.46}$$

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. (3.45) 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 chapter 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

$$e^{S_1} H_d e^{-S_1} = H_{d1}^D + H_{d2}^D \tag{3.47}$$

where

$$H_{d1}^D := \frac{\Omega e^{i\phi}}{2} b + \text{h.c.} \tag{3.48}$$

$$H_{d2}^D := \frac{\Omega e^{i\phi}}{2} \left(a \sum_{m=0}^{\infty} \frac{g\Delta_{-1}}{\Delta_m\Delta_{m-1}} |m\rangle\langle m| + a^\dagger \sum_{m=0}^{\infty} \frac{g\alpha\sqrt{m+1}\sqrt{m+2}}{\Delta_m\Delta_{m+1}} |m\rangle\langle m+2| \right) + \text{h.c.} \tag{3.49}$$

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.

3.5.3. SWT OF THE PURE DRIVE HAMILTONIAN

Summarizing, in the unitarily transformed frame the original Hamiltonian H takes (approximately) the form

$$H^D \approx H_0^D + H_{d1}^D + H_{d2}^D, \tag{3.50}$$

where H_0^D is given in Eq. (3.45) and H_{d1}^D, H_{d2}^D are given in Eq. (3.47).

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 Sec. 3.5.1, defining a "double-dressed" Hamiltonian

$$H^{DD} := e^{S'} H^D e^{-S'} \tag{3.51}$$

$$= \underbrace{e^{S'}(H_0^D + H_{d1}^D)e^{-S'}}_{=:H_0^{DD}} + \underbrace{e^{S'} H_{d2}^D e^{-S'}}_{=:H_d^{DD}} \tag{3.52}$$

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 Sec. 3.5.1, to find S_1' we need to solve Eq. (3.38), i.e.

$$[H_0^D, S_1'] = H_{d1}^D \tag{3.53}$$

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

$$\langle ml|S_1'|nk\rangle = \frac{\langle ml|H_{d1}^D|nk\rangle}{E_{ml}^D - E_{nk}^D}, \tag{3.54}$$

where $\{E_{ml}^D\}$ are the eigenenergies of H_0^D, which can be easily inferred from Eq. (3.45). We neglect the dispersive shift since it is proportional to α/Δ. Then

$$\langle ml|S_1'|nk\rangle = \frac{\Omega}{2}\left(-\frac{\sqrt{m+1}\,\delta_{m,n-1}\delta_{l,k}}{\delta^q + \alpha m + \frac{g^2\Delta_{-1}}{\Delta_{m-1}\Delta_m}}e^{i\phi}\right.$$
$$\left. + \frac{\sqrt{m}\,\delta_{m,n+1}\delta_{l,k}}{\delta^q + \alpha(m-1) + \frac{g^2\Delta_{-1}}{\Delta_{m-2}\Delta_{m-1}}}e^{-i\phi}\right), \tag{3.55}$$

where $\delta_{i,j}$ is the Kronecker delta. From this equation one can infer that

$$S_1' = -\frac{\Omega}{2}e^{i\phi}\sum_{m=0}^{\infty}\frac{\sqrt{m+1}}{\delta_m^q}|m\rangle\langle m+1| - \text{h.c.}, \tag{3.56}$$

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. (3.39), i.e.

$$[H_0^D, S_2'] = \frac{1}{2}[S_1', H_{d1}^D]_{\text{OD}} \tag{3.57}$$

with

$$
\begin{aligned}
\left[S_1', H_{d1}^D\right] = & -\frac{\Omega^2}{2} \sum_{m=0}^{\infty} \frac{\tilde{\delta}_m^q}{\delta_m^q \delta_{m-1}^q} |m\rangle \langle m| \\
& -\frac{\Omega^2}{4} \sum_{m=0}^{\infty} \sqrt{m+1}\sqrt{m+2}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right) \\
& (e^{2i\phi}|m\rangle \langle m+2| + \text{h.c.}),
\end{aligned}
\tag{3.58}
$$

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

$$
\begin{aligned}
S_2' = & \frac{\Omega^2}{8} e^{2i\phi} \sum_{m=0}^{\infty} \frac{\sqrt{m+1}\sqrt{m+2}}{\delta_m^q + \delta_{m+1}^q}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right) \\
& |m\rangle \langle m+2| - \text{h.c.}
\end{aligned}
\tag{3.59}
$$

We can then compute S_3' from Eq. (3.40), i.e.

$$
\left[H_0^D, S_3'\right] = \frac{1}{2}\left[S_2', H_{d1}^D\right]_{\text{OD}} + \frac{1}{3}\left[S_1', \left[S_1', H_{d1}^D\right]_{\text{D}}\right]_{\text{OD}} + \frac{1}{12}\left[S_1', \left[S_1', H_{d1}^D\right]_{\text{OD}}\right]_{\text{OD}}.
\tag{3.60}
$$

The result is

$$
\begin{aligned}
S_3' = & \Omega^3 e^{i\phi} \sum_{m=0}^{\infty} |m\rangle \langle m+1| \left(\frac{1}{12}\frac{\sqrt{m+1}}{(\delta_m^q)^3}\left(\frac{\tilde{\delta}_{m+1}^q}{\delta_{m+1}^q}-\frac{\tilde{\delta}_m^q}{\delta_{m-1}^q}\right)\right. \\
& + \frac{1}{96\delta_m^q}\left((m+2)\sqrt{m+1}\frac{\delta_m^q + 4\delta_{m+1}^q}{\delta_{m+1}^q(\delta_m^q + \delta_{m+1}^q)}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right)\right. \\
& \left.\left.-\sqrt{m+1}\,m\frac{4\delta_{m-1}^q + \delta_m^q}{\delta_{m-1}^q(\delta_{m-1}^q + \delta_m^q)}\left(\frac{1}{\delta_{m-1}^q}-\frac{1}{\delta_m^q}\right)\right)\right) - \text{h.c.} \\
& + \frac{\Omega^3}{96} e^{3i\phi} \sum_{m=0}^{\infty} |m\rangle \langle m+3| \frac{\sqrt{m+1}\sqrt{m+2}\sqrt{m+3}}{\delta_m^q + \delta_{m+1}^q + \delta_{m+2}^q} \\
& \left(\frac{3\delta_{m+2}^q - \delta_{m+1}^q - \delta_m^q}{\delta_{m+2}^q(\delta_m^q + \delta_{m+1}^q)}\left(\frac{1}{\delta_m^q}-\frac{1}{\delta_{m+1}^q}\right) - \frac{3\delta_m^q - \delta_{m+1}^q - \delta_{m+2}^q}{\delta_m^q(\delta_{m+1}^q + \delta_{m+2}^q)}\left(\frac{1}{\delta_{m+1}^q}-\frac{1}{\delta_{m+2}^q}\right)\right) - \text{h.c.}
\end{aligned}
\tag{3.61}
$$

We can eventually use Eqs. (3.56), (3.59) and (3.61) together with Eq. (3.41) to ob-

tain H_0^{DD} (defined in Eq. (3.52)):

$$
\begin{aligned}
H_0^{DD} = {}& \delta^r a^\dagger a + \sum_{m=0}^{\infty} |m\rangle \langle m| \Bigg(m\delta^q + \frac{\alpha}{2} m(m-1) + \frac{g^2 m}{\Delta_{m-1}} \\
&- \frac{\Omega^2 \tilde{\delta}_m^q}{4\delta_m^q \delta_{m-1}^q} - \frac{\Omega^4}{32} \left(\frac{m+1}{(\delta_m^q)^3} \left(\frac{\tilde{\delta}_{m+1}^q}{\delta_{m+1}^q} - \frac{\tilde{\delta}_m^q}{\delta_{m-1}^q} \right) - \frac{m}{(\delta_{m-1}^q)^3} \left(\frac{\tilde{\delta}_m^q}{\delta_m^q} - \frac{\tilde{\delta}_{m-1}^q}{\delta_{m-2}^q} \right) \right) \\
&- \frac{\Omega^4}{192} \Bigg(\frac{1}{\delta_m^q} \Bigg((m+2)(m+1) \frac{\delta_m^q + 5\delta_{m+1}^q}{\delta_{m+1}^q (\delta_m^q + \delta_{m+1}^q)} \left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q} \right) \\
&\qquad\qquad - (m+1)m \frac{5\delta_{m-1}^q + \delta_m^q}{\delta_{m-1}^q (\delta_{m-1}^q + \delta_m^q)} \left(\frac{1}{\delta_{m-1}^q} - \frac{1}{\delta_m^q} \right) \Bigg) \\
&\qquad\quad - \frac{1}{\delta_{m-1}^q} \Bigg((m+1)m \frac{\delta_{m-1}^q + 5\delta_m^q}{\delta_m^q (\delta_{m-1}^q + \delta_m^q)} \left(\frac{1}{\delta_{m-1}^q} - \frac{1}{\delta_m^q} \right) \\
&\qquad\qquad - m(m-1) \frac{5\delta_{m-2}^q + \delta_{m-1}^q}{\delta_{m-2}^q (\delta_{m-2}^q + \delta_{m-1}^q)} \left(\frac{1}{\delta_{m-2}^q} - \frac{1}{\delta_{m-1}^q} \right) \Bigg) \Bigg) \\
&+ \frac{\Omega^4}{96} \Bigg(\frac{(m+2)(m+1)}{\delta_m^q + \delta_{m+1}^q} \left(\frac{1}{\delta_m^q} - \frac{1}{\delta_{m+1}^q} \right)^2 - \frac{m(m-1)}{\delta_{m-2}^q + \delta_{m-1}^q} \left(\frac{1}{\delta_{m-2}^q} - \frac{1}{\delta_{m-1}^q} \right)^2 \Bigg) \Bigg) \\
&- a^\dagger a \sum_m \frac{g^2 \Delta_{-1}}{\Delta_m \Delta_{m-1}} |m\rangle \langle m| .
\end{aligned}
\tag{3.62}
$$

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. (3.52)) up to 2nd order in Ω/δ^q is instead given by

$$
H_d^{DD} = H_{d2}^D + [S_1', H_{d2}^D] + [S_2', H_{d2}^D] + \frac{1}{2} [S_1', [S_1', H_{d2}^D]]
\tag{3.63}
$$

$$
=: H_{\text{eff.coupl.}}^{DD} + H_{\text{resid.}}^{DD},
\tag{3.64}
$$

where

$$
\begin{aligned}
H_{\text{eff.coupl.}}^{DD} = {}& e^{i\phi} a^\dagger \sum_{m=0}^{\infty} |m\rangle \langle m+2| \Bigg(\tilde{g}_m \left(1 - \frac{\Omega^2}{8} \left(\frac{m+3}{(\delta_{m+2}^q)^2} + \frac{m+2}{(\delta_{m+1}^q)^2} + \frac{m+1}{(\delta_m^q)^2} + \frac{m}{(\delta_{m-1}^q)^2} \right) \right) \\
&+ \frac{\Omega^2}{4} \left(\frac{\sqrt{m+1}\sqrt{m+3}}{\delta_m^q \delta_{m+2}^q} \tilde{g}_{m+1} + \frac{\sqrt{m}\sqrt{m+2}}{\delta_{m-1}^q \delta_{m+1}^q} \tilde{g}_{m-1} \right) \\
&+ \frac{\Omega^2}{4} \sqrt{m+1}\sqrt{m+2} \left(\frac{g_{m+2}'}{\delta_m^q (\delta_m^q + \delta_{m+1}^q)} - \frac{g_{m+1}'}{\delta_m^q \delta_{m+1}^q} + \frac{g_m'}{\delta_{m+1}^q (\delta_m^q + \delta_{m+1}^q)} \right) \Bigg) + \text{h.c.}
\end{aligned}
\tag{3.65}
$$

with

$$
\tilde{g}_m := \frac{g\alpha\Omega\sqrt{m+1}\sqrt{m+2}}{2\Delta_m \Delta_{m+1}}
\tag{3.66}
$$

$$
g_m' := \frac{g\Omega\Delta_{-1}}{2\Delta_m \Delta_{m-1}},
\tag{3.67}
$$

and

$$
\begin{aligned}
H_{\text{resid.}}^{DD} = {}& (e^{i\phi}a + \text{h.c.}) \sum_{m=0}^{\infty} |m\rangle\langle m| \left(g'_m \left(1 - \frac{\Omega^2}{4}\left(\frac{m+1}{(\delta_m^q)^2} + \frac{m}{(\delta_{m-1}^q)^2} \right) \right) \right. \\
& + \frac{\Omega^2}{4}\left(\frac{m+1}{(\delta_m^q)^2} g'_{m+1} + \frac{m}{(\delta_{m-1}^q)^2} g'_{m-1} \right) \\
& + \frac{\Omega^2}{4}\left(\frac{\sqrt{m+1}\sqrt{m+2}\,\tilde{g}_m}{\delta_m^q(\delta_m^q + \delta_{m+1}^q)} + \frac{\sqrt{m}\sqrt{m+1}\,\tilde{g}_{m-1}}{\delta_m^q \delta_{m-1}^q} \right. \\
& \left.\left. + \frac{\sqrt{m-1}\sqrt{m}\,\tilde{g}_{m-2}}{\delta_{m-1}^q(\delta_{m-2}^q + \delta_{m-1}^q)} \right) \right) \\
& - \frac{\Omega}{2}e^{2i\phi}a \sum_{m=0}^{\infty} |m\rangle\langle m+1| \frac{\sqrt{m+1}}{\delta_m^q}(g'_{m+1} - g'_m) + \text{h.c.} \\
& - \frac{\Omega}{2}a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+1| \left(\frac{\sqrt{m+1}}{\delta_m^q}(g'_{m+1} - g'_m) + \frac{\sqrt{m+2}}{\delta_{m+1}^q}\tilde{g}_m - \frac{\sqrt{m}}{\delta_{m-1}^q}\tilde{g}_{m-1} \right) + \text{h.c.} \\
& + \frac{\Omega^2}{4}e^{3i\phi}a \sum_{m=0}^{\infty} |m\rangle\langle m+2| \sqrt{m+1}\sqrt{m+2}\left(\frac{g'_{m+2}}{\delta_m^q(\delta_m^q + \delta_{m+1}^q)} - \frac{g'_{m+1}}{\delta_m^q \delta_{m+1}^q} \right. \\
& \left. + \frac{g'_m}{\delta_{m+1}^q(\delta_m^q + \delta_{m+1}^q)} \right) + \text{h.c.} \\
& - \frac{\Omega}{2}e^{2i\phi}a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+3| \left(\frac{\sqrt{m+1}}{\delta_m^q}\tilde{g}_{m+1} - \frac{\sqrt{m+3}}{\delta_{m+2}^q}\tilde{g}_m \right) + \text{h.c.} \\
& + \frac{\Omega^2}{4}e^{3i\phi}a^\dagger \sum_{m=0}^{\infty} |m\rangle\langle m+4| \left(\frac{\sqrt{m+1}\sqrt{m+2}\,\tilde{g}_{m+2}}{\delta_m^q(\delta_m^q + \delta_{m+1}^q)} - \frac{\sqrt{m+4}\sqrt{m+1}\,\tilde{g}_{m+1}}{\delta_m^q \delta_{m+3}^q} \right. \\
& \left. + \frac{\sqrt{m+3}\sqrt{m+4}\,\tilde{g}_m}{\delta_{m+3}^q(\delta_{m+3}^q + \delta_{m+2}^q)} \right) + \text{h.c.} \qquad (3.68)
\end{aligned}
$$

All terms in $H_{\text{resid.}}^{DD}$ 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.

3.5.4. ANALYSIS OF THE $|20\rangle \leftrightarrow |01\rangle$ AVOIDED CROSSING

In this section we give the methods used to calculate the curves in Fig. 3.1(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. (3.1). 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 Sec. 3.2.1. 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 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 Sec. 3.2.2 the res-LRU can reach a very high R, for which the effect of this phase is then minimal. Assuming that $H_{\text{resid.}}^{DD}$ in Eq. (3.68) is negligible, an analytical approximation of η is given by

$$\eta(\omega_d) \approx \langle 20|H_0^{DD}(\omega_d)|20\rangle - \langle 01|H_0^{DD}(\omega_d)|01\rangle, \tag{3.69}$$

where we have made the dependence of H_0^{DD} in Eq. (3.62) 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. (3.62) is not analytically solvable, we use the secant method available in *scipy* to find ω_d^* that fulfills this condition in Eq. (3.69). It is then straightforward to compute the (approximate) analytical estimate for the effective coupling as $\tilde{g}(\omega_d^*) = |\langle 01|H_{\text{eff.coupl.}}^{DD}(\omega_d^*)|20\rangle|$ from Eq. (3.65), which is plotted in Fig. 3.1(e).

3.6. FURTHER CHARACTERIZATION OF THE READOUT-RESONATOR LRU

3.6.1. EFFECTIVE T_1 AND T_2 DUE TO THE DRIVE
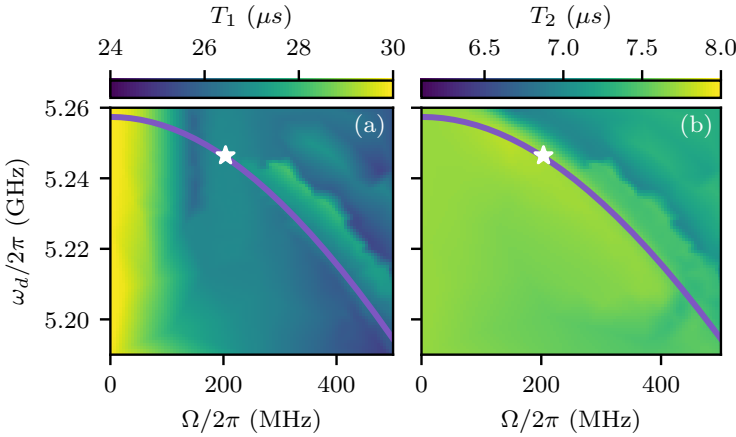
Figure 3.6: Effective T_1 (a) and T_2 (b) which account for the extra decoherence caused by the drive during the time slot $T_{\text{slot}} = 440$ ns. We can see that the variation is small as a function of the drive amplitude compared to the values at $\Omega = 0$. The white star indicates the chosen operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_p = 178.6$ ns, see Sec. 3.2.2). The purple line corresponds to the higher order estimate of the optimal drive frequency ω_d^* as a function of Ω (see Fig. 3.1(c)). The heatmaps are sampled using the *adaptive* package [67].

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 $T_{slot} = 440$ ns, leading to a uniform metric for the whole (ω_d, Ω) landscape. Specifically, to estimate T_1 we prepare the state $|1\rangle\langle 1| \otimes \sigma_{th}$, we simulate the Lindblad equation in Eq. (3.12) 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 $p^{|1\rangle} = e^{-T_{slot}/T_1}$ we then compute T_1 by inverting this formula. To estimate T_2 we prepare $|+\rangle\langle +| \otimes \sigma_{th}$ 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 $\left|\langle 0| \text{Tr}_r(\rho(T_{slot}))|1\rangle\right| = e^{-T_{slot}/T_2}/2$ to get T_2.

Fig. 3.6 shows the resulting effective T_1 and T_2. In Fig. 3.6(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 $\Omega_{cr}/2\pi = 143$ MHz, where the pulses are very long, suggesting that driving slightly into the underdamped regime is favourable. In Fig. 3.6(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 Tab. 3.1). 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. 3.6(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. 3.6(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. 3.6(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. 3.2(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. 3.2(d) one can notice that a non-negligible amount of population ends up in $|10\rangle$ from the initial state $|0\rangle\langle 0| \otimes \sigma_{th}$. 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.
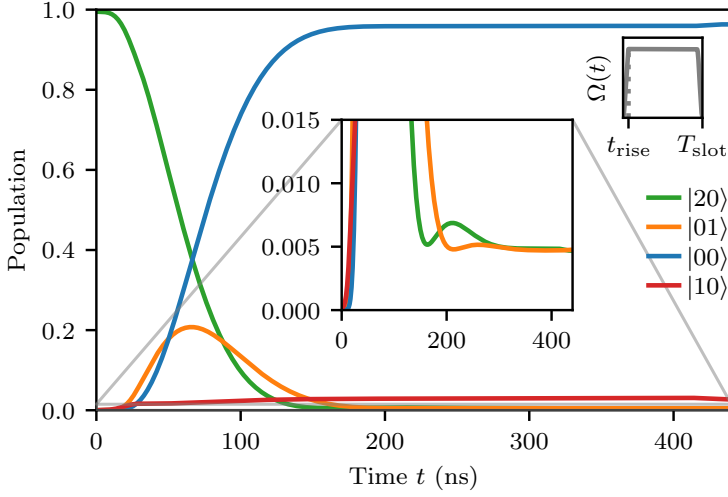
Figure 3.7: Time evolution from the initial state $|2\rangle\langle 2| \otimes \sigma_{\mathrm{th}}$ for $t_{\mathrm{rise}} = 30$ ns and for an otherwise always-on drive during T_{slot}. This is simulated with the same $\Omega/2\pi \approx 204$ MHz and $\omega_d/2\pi \approx 5.2464$ GHz as at the operating point in Fig. 3.2.

3.6.2. LONG-DRIVE LIMIT IN THE UNDERDAMPED REGIME AND ITS DRAW-BACK AS A LRU

In this section we compare the reset schemes in [55, 56] versus [54] in terms of their performance as a LRU in the underdamped regime. The approach of [55, 56], which we have adopted in Sec. 3.2.2, 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 Sec. 3.2.2, this approach allows for a residual leakage population $p_{\mathrm{op.}}^{|2\rangle} \approx 0.5\%$ at the operating point (see Fig. 3.2(a)), given our parameters (see Tab. 3.1). While this already reaches thermal-state levels (here $\bar{n} = 0.5\%$) with the considered system parameters, the approach in [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 [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. Fig. 3.7 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 Sec. 3.2.2). 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 $T_{\mathrm{slot}} = 440$ 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 [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. 3.7 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. 3.2(b). By evaluating T_1 we find $T_1 \approx 23\ \mu$s instead of 27 μs (see Sec. 3.6.1). Furthermore, if one would have to use a $t_{\text{p}} > T_{\text{slot}}$ 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.

3.6.3. SENSITIVITY TO RESIDUAL ZZ CROSSTALK
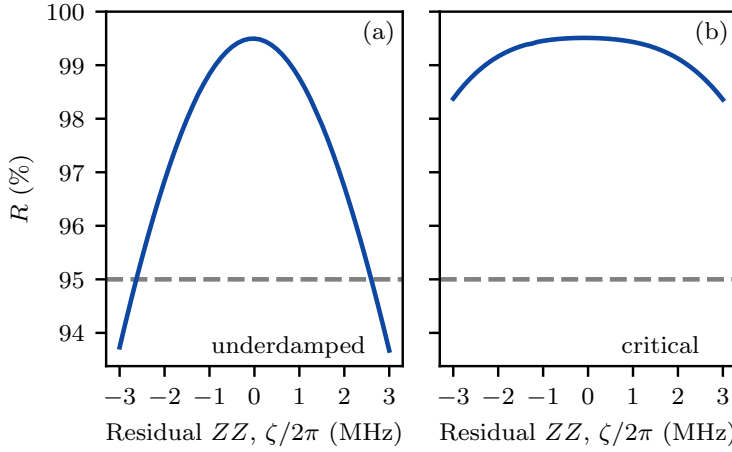
Figure 3.8: Sensitivity of the leakage-reduction rate R of the readout-resonator LRU as a function of the overall residual ZZ coupling ζ. (a) Underdamped regime, specifically at the operating point ($\Omega/2\pi \approx 204$ MHz, $\omega_d/2\pi \approx 5.2464$ GHz, $t_{\text{p}} = 178.6$ ns, see Sec. 3.2.2). (b) Critical regime ($\Omega/2\pi \approx 143$ MHz, $\omega_d/2\pi \approx 5.252$ GHz, $t_{\text{p}} = 440$ ns).

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, 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 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. 3.8 we perform the analysis for the operating point (see Sec. 3.2.2), 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 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 Sec. 3.3.4 and for which the logical error rate was already close to optimal in Surface-17 (see Sec. 3.7.2). 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 $\lesssim 1\%$.

3.7. FURTHER SURFACE-17 CHARACTERIZATION

3.7.1. DETAILS ABOUT THE DENSITY-MATRIX SIMULATIONS

The parameters used in this chapter are reported in Tab. 3.2.

Parameter	Value
Relaxation time T_1	30 μs
Sweetspot pure-dephasing time $T_{\phi,\mathrm{max}}$	60 μs
High-freq. pure-dephasing time at interaction point $T_{\phi,\mathrm{int}}$	8 μs
Mid-freq. pure-dephasing time at interaction point $T_{\phi,\mathrm{int}}$	6 μs
Mid-freq. pure-dephasing time at parking point $T_{\phi,\mathrm{park}}$	8 μs
Low-freq. pure-dephasing time at parking point $T_{\phi,\mathrm{park}}$	9 μs
Single-qubit gate time t_{gate}	20 ns
Two-qubit interaction time t_{int}	30 ns
Single-qubit phase-correction time t_{pc}	10 ns
Readout-resonator LRU time $t_{\mathrm{res\text{-}LRU}}$	100 ns
$|1\rangle \leftrightarrow |2\rangle$ π-pulse time $t_{\pi\text{-}\mathrm{LRU}}$	20 ns
Measurement time t_{m}	580 ns
QEC-cycle time t_{c}	800 ns

Table 3.2: 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].

RES-LRU IN *quantumsim*

A comprehensive review of the density-matrix simulations and the use of the *quantum-sim* package [59] is available at [40, 51]. 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 $t_{\mathrm{res\text{-}LRU}}$. We use the Pauli Transfer Matrix $S_{\mathrm{res\text{-}LRU}} = S_\uparrow S_\downarrow$, where $S_\downarrow$ is the Pauli Transfer Matrix of the superoperator $\mathscr{S}_\downarrow = e^{t_{\mathrm{res\text{-}LRU}} \mathscr{L}_\downarrow}$ and the Lindbladian $\mathscr{L}_\downarrow$ has the quantum
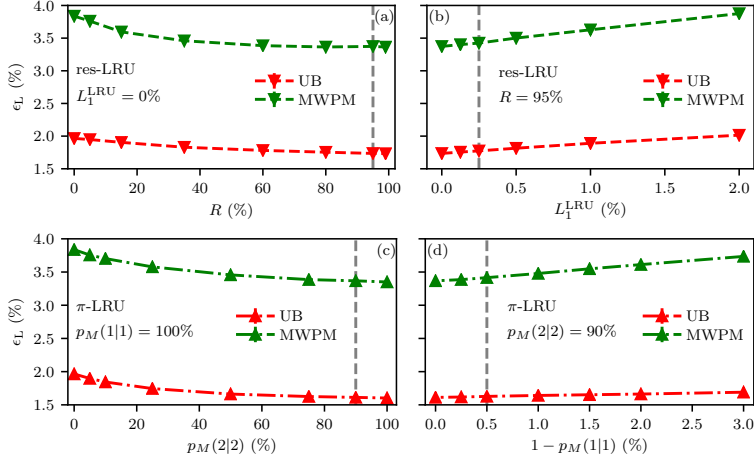
Figure 3.9: Logical error rate ε_{L} per QEC cycle as a function of various LRU parameters. (a),(b) use only the res-LRU, while (c),(d) the π-LRU. We fix $L_1 = 0.5\%$ for all. Vertical dashed lines indicate the values considered in Sec. 3.3.4. These results are extracted from 2×10^4 runs of 20 QEC cycles each per choice of parameters. Error bars are estimated using bootstrapping and are smaller than the symbol size.

jump operator

$$K_\downarrow = \frac{1}{\sqrt{\frac{t_{\text{res-LRU}}}{-\log(1-R_{\text{sim}})}}} |0\rangle \langle 2| \tag{3.70}$$

with R_{sim} to be determined. Besides this, $\mathscr{L}_\downarrow$ has the standard qutrit jump operators for relaxation and dephasing [40]. 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

$$K_\uparrow = \frac{1}{\sqrt{\frac{1}{-\log(1-2L_1^{\text{LRU}})}}} |2\rangle \langle 0| \tag{3.71}$$

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 Sec. 3.3.2. 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 Sec. 3.3.2 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, 73].

UB considers the 32 computational states that differ by a purely X error on top of $|0\rangle_L$ 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 $|0\rangle_L$ and the other with $|1\rangle_L$ 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 $\mathscr{F}_L(n)$ 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 $|0\rangle_L$. 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 [75] 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.

3.7.2. 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. 3.9(a),(c) one can see that this is the case and that the values of ε_L at the parameters used in Sec. 3.3.4 ($R = 95\%$ 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. 3.9(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 Sec. 3.3.4 ($L_1^{\mathrm{LRU}} = 0.25\%$ and $1 - p_M(1|1) = 0.5\%$) rather than at 0, although the difference is small.
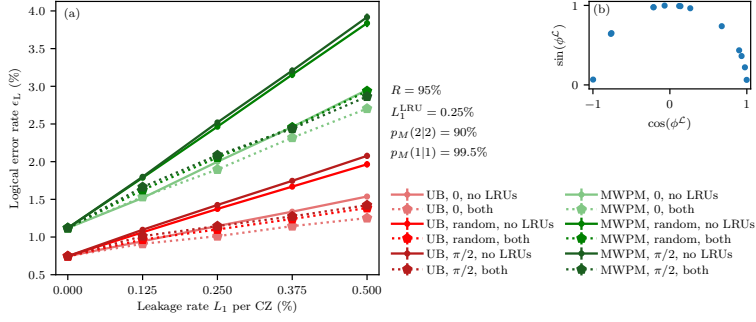
Figure 3.10: Variation of the logical error rate ε_L for different choices of leakage conditional phases $\phi^{\mathscr{L}}$. (a) ε_L per QEC cycle for UB (shades of red) and MWPM (shades of green) versus L_1, in the cases with: no LRUs and both LRUs, each for all $\phi^{\mathscr{L}}$ set to 0, $\pi/2$ 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 $\phi^{\mathscr{L}}$ used across this work. These values are extracted from a uniform distribution in $[0,\pi]$. We have excluded negative values as $\pm\phi^{\mathscr{L}}$ corresponds to the same chance of spreading a Z error under the twirling action of the parity-check measurements.

3.7.3. 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 |low-f. transmon, high-f. transmon⟩. Note that for a controller-phase gate 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⟩ 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 [40] 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⟩ and |1⟩.

For a flux-based CZ with conditional phase π for |11⟩, ideally one should have $\phi^{\mathscr{L}}_{\text{flux}} = 0$ and $\phi^{\mathscr{L}}_{\text{stat}} = \pi$ [40] as |02⟩ acquires a conditional phase equal and opposite to |11⟩. If only |12⟩ and |21⟩ 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 [40]. However, |03⟩ interacts with |12⟩ and |21⟩ and breaks the relationship above, for which we can consider $\phi^{\mathscr{L}}_{\text{flux}}$ and $\phi^{\mathscr{L}}_{\text{stat}}$ as effectively unconstrained. The randomized values used across the main text are reported in Fig. 3.10(b). We use 14 values, of which 3 for $\phi^{\mathscr{L}}_{\text{stat}}$ and 3 for $\phi^{\mathscr{L}}_{\text{flux}}$ when each high-frequency data qubit is leaked or interacts with a leaked ancilla qubit, respectively, and 8 only for $\phi^{\mathscr{L}}_{\text{stat}}$ 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 $\phi^{\mathscr{L}} = 0$, since no Z rotations are spread then. Instead, the worst-case scenario corresponds to all $\phi^{\mathscr{L}} = \pi/2$, since under the twirling effect of the parity-check measurements this corresponds to spreading a Z error with 50% chance. Notice that, if all $\phi^{\mathscr{L}} = \pi$, 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.

Fig. 3.10(a) compares the logical performance for both UB and MWPM in the cases where $\phi^{\mathscr{L}} = 0$, $\phi^{\mathscr{L}} = \pi/2$ and when they are random as in Fig. 3.5 and in the rest of this work. First, one can notice that the performance of random $\phi^{\mathscr{L}}$ is very close to the worst-case scenario ($\phi^{\mathscr{L}} = \pi/2$). 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 $\phi^{\mathscr{L}} = 0$ without implementing LRUs is almost as good (or even better) as using the LRUs when $\phi^{\mathscr{L}}$ 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 $\phi^{\mathscr{L}} = 0$ 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. 3.10(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.

BIBLIOGRAPHY

[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 control of an error-corrected qubit*, Nature **598**, 281–286 (2021).

[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] 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].

[7] 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).

[8] 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 read-out of a trapped-ion quantum bit*, Phys. Rev. Lett. **113**, 220501 (2014).

[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] 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).

[11] 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), 060302 (2016).

[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*, Physical Review A **101** (2020).

[13] 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).

[14] 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).

[15] F. Yan, D. Campbell, P. Krantz, M. Kjaergaard, D. Kim, J. L. Yoder, D. Hover, A. Sears, A. J. Kerman, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Distinguishing coherent and thermal photon noise in a circuit quantum electrodynamical system*, Phys. Rev. Lett. **120**, 260504 (2018).

[16] 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).

[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. App. **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. App. **10**, 034040 (2018).

[20] S. Bravyi, D. Gosset, and R. König, *Quantum advantage with shallow circuits*, Science **362**, 308–311 (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, et al., *Quantum computational advantage using photons*, Science **370**, 1460–1463 (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] M. P. Harrigan, K. J. Sung, M. Neeley, K. J. Satzinger, F. Arute, K. Arya, J. Atalaya, 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, A. Fowler, B. Foxen, C. Gidney, M. Giustina, R. Graff, S. Habegger, A. Ho, S. Hong, T. Huang, L. B. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, C. Jones, D. Kafri, K. Kechedzhi, J. Kelly, S. Kim, P. V. Klimov, A. N. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, M. Lindmark, M. Leib, O. Martin, J. M. Martinis, J. R. McClean, M. McEwen, A. Megrant, X. Mi, M. Mohseni, W. Mruczkiewicz, J. Mutus, O. Naaman, C. Neill, F. Neukart, M. Y. Niu, T. E. O'Brien, B. O'Gorman, E. Ostby, A. Petukhov, H. Putterman, C. Quintana, P. Roushan, N. C. Rubin, D. Sank, A. Skolik, V. Smelyanskiy, D. Strain, M. Streif, M. Szalay, A. Vainsencher, T. White, Z. J. Yao, P. Yeh, A. Zalcman, L. Zhou, H. Neven, D. Bacon, E. Lucero, E. Farhi, and R. Babbush, *Quantum approximate optimization of non-planar graph problems on a planar superconducting processor*, Nature Physics **17**, 332–336 (2021).

[24] 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**, 66–69 (2015).

3

[25] Z. Chen, K. J. Satzinger, J. Atalaya, A. N. Korotkov, A. Dunsworth, D. Sank, C. Quin-
      tana, 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. Court-
      ney, 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. Hug-
      gins, 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. Naa-
      man, 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, *Expo-
      nential suppression of bit or phase errors with cyclic error correction*, Nature **595**,
      383–387 (2021).

[26] D. Ristè, S. Poletto, M. Z. Huang, A. Bruno, V. Vesterinen, O. P. Saira, and L. Di-
      Carlo, *Detecting bit-flip errors in a logical qubit using stabilizer measurements*,
      Nat. Commun. **6**, 6983 (2015).

[27] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, et al., *Demon-
      stration of weight-four parity measurements in the surface code architecture*, Phys.
      Rev. Lett. **117**, 210505 (2016).

[28] 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 reg-
      ister*, Nature **563**, 527–531 (2018).

[29] 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).

[30] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, J. Heinsoo, J.-C. Besse, M. Gabu-
      reac, A. Wallraff, and C. Eichler, *Entanglement stabilization using ancilla-based
      parity detection and real-time feedback in superconducting circuits*, npj Quantum
      Information **5**, 1–7 (2019).

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

[32] 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*,
      Nat. Phys. **18**, 80–86 (2022).

[33] F. W. Strauch, P. R. Johnson, A. J. Dragt, C. J. Lobb, J. R. Anderson, and F. C. Well-
      stood, *Quantum logic gates for coupled superconducting phase qubits*, Phys. Rev.
      Lett. **91**, 167005 (2003).

[34] 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**, 14 (2021).

[35] 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).

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

[37] 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).

[38] 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).

[39] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults*, Quantum Info. Comput. **7**, 139–156 (2007).

[40] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. Di-Carlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Information **6**, 102 (2020).

[41] 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.

[42] 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**, 1761 (2021).

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

[44] 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).

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

[46] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code*, Quantum Info. Comput. **15**, 997–1016 (2015).

[47] 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).

3

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

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

[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. M. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Information **3** (2017).

[52] 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).

[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*, Phys. Rev. Lett. **109**, 240502 (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. Applied **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**, 491–492 (1966).

[64] S. Bravyi, D. P. DiVincenzo, and D. Loss, *Schrieffer–Wolff transformation for quantum many-body systems*, Annals of Physics **326**, 2793–2826 (2011).

[65] E. Magesan and J. M. Gambetta, *Effective hamiltonian models of the cross-resonance gate*, Physical Review A **101**, 052308 (2020).

[66] M. Boissonneault, J. M. Gambetta, and A. Blais, *Dispersive regime of circuit qed: photon-dependent qubit dephasing and relaxation rates*, Phys. Rev. A **79**, 013819 (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, 2007).

[69] S. Haroche and J. Raimond, *Exploring the quantum: atoms, cavities, and photons*, Oxford Graduate Texts (Oxford University Press, 2006).

[70] F. Battistel, B. Varbanov, and B. Terhal, *Further supporting material for "a hardware-efficient leakage-reduction scheme for quantum error correction with superconducting transmon qubits"*, 2021.

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

[72] S. Krinner, S. Lazar, A. Remm, C. K. Andersen, N. Lacroix, G. J. Norris, et al., *Benchmarking coherent errors in controlled-phase gates due to spectator qubits*, Phys. Rev. Appl. **14**, 024042 (2020).

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

[74] J. Hertzberg, E. Zhang, S. Rosenblatt, E. Magesan, J. Smolin, J. Yau, V. Adiga, M. Sandberg, B. M., J. M. Chow, and J. S. Orcutt, *Laser-annealing josephson junctions for yielding scaled-up superconducting quantum processors*, npj Quantum Inf. **7**, 1–8 (2021).

[75] S. Spitz, B. M. Tarasinski, C. Beenakker, and T. O'Brien, *Adaptive weight estimator for quantum error correction in a time-dependent environment*, Advanced Quantum Technologies **1**, 1800012 (2018).

4

ALL-MICROWAVE LEAKAGE REDUCTION UNITS FOR QUANTUM ERROR CORRECTION WITH SUPERCONDUCTING TRANSMON QUBITS

Minimizing leakage from computational states is a challenge when using many-level systems like superconducting quantum circuits as qubits. In this chapter, we realize and extend the quantum-hardware-efficient, all-microwave leakage reduction unit (LRU) for transmons in a circuit QED architecture proposed in Chapter 3 [2]. This LRU effectively reduces leakage in the second- and third-excited transmon states with up to 99% efficacy in 220 ns*, with minimum impact on the qubit subspace. As a first application in the context of quantum error correction, we show how multiple simultaneous LRUs can reduce the error detection rate and suppress leakage buildup within* 1% *in data and ancilla qubits over 50 cycles of a weight-2 stabilizer measurement.*

4.1. INTRODUCTION

Superconducting qubits, such as the transmon [3], are many-level systems in which a qubit is represented by the two lowest-energy states $|g\rangle$ and $|e\rangle$. However, leakage to non-computational states is a risk for all quantum operations, including single-qubit gates [4], two-qubit gates [5–7] and measurement [8, 9]. While the typical probability of leakage per operation may pale in comparison to conventional qubit errors induced by control errors and decoherence [7, 10], unmitigated leakage can build up with increasing circuit depth. A prominent example is multi-round quantum error correction (QEC) with stabilizer codes such as the surface code [11]. In the absence of leakage, such codes successfully discretize all qubit errors into Pauli errors through the measurement of stabilizer operators [12, 13], and these Pauli errors can be detected and corrected (or kept track of) using a decoder. However, leakage errors fall outside the qubit subspace and are not immediately correctable [14–16]. The signature of leakage on the stabilizer syndrome is often not straightforward, hampering the ability to detect and correct it [17, 18]. Additionally, the build-up of leakage over QEC rounds accelerates the destruction of the logical information [10, 19]. Therefore, despite having low probability per operation, methods to reduce leakage must be employed when performing experimental QEC with multi-level systems.

Physical implementations of QEC codes [20–25] use qubits for two distinct functions: Data qubits store the logical information and, together, comprise the encoded logical qubits. Ancilla qubits perform indirect measurement of the stabilizer operators. Handling leakage in ancilla qubits is relatively straightforward as they are measured in every QEC cycle. This allows for the use of reset protocols [19, 26] without the loss of logical information. Leakage events can also be directly detected using three- or higher-level readout [21] and reset using feedback [27, 28]. In contrast, handling data-qubit leakage requires a subtle approach as it cannot be reset nor directly measured without loss of information or added circuit complexity [29–31]. A promising solution is to interleave QEC cycles with operations that induce seepage without disturbing the qubit subspace, known as leakage reduction units (LRUs) [2, 14, 15, 29, 30, 32–35]. An ideal LRU returns leakage back to the qubit subspace, converting it into Pauli errors which can be detected and corrected, while leaving qubit states undisturbed. By converting leakage into conventional errors, LRUs enable a moderately high physical noise threshold, below which the logical error rate decreases exponentially with the code distance [15, 30]. A more powerful operation called 'heralded leakage reduction' would both reduce and herald leakage, leading to a so-called erasure error [36, 37]. Unlike Pauli errors, the exact location of erasures is known, making them easier to correct and leading to higher error thresholds [38–41].

In this chapter, we present the realization and extension of the LRU scheme proposed in Ref. [2]. This is a highly practical scheme requiring only microwave pulses and the quantum hardware typically found in contemporary circuit QED quantum processors: a microwave drive and a readout resonator dispersively coupled to the target transmon (in our case, a readout resonator with dedicated Purcell filter). We show its straightforward calibration and the effective removal of the population in the first two leakage states of the transmon ($|f\rangle$ and $|h\rangle$) with up to > 99% efficacy in 220 ns. Process tomography reveals that the LRU backaction on the qubit subspace is only an AC-Stark shift, which
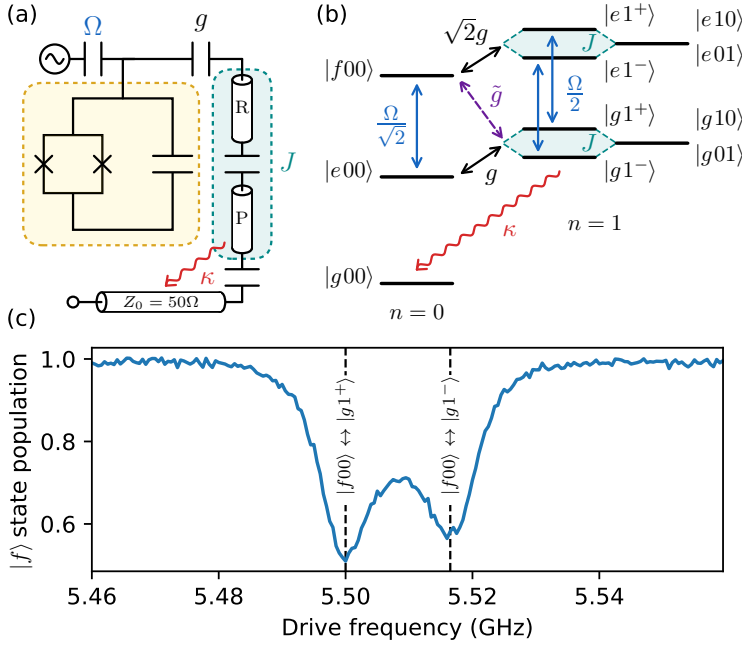
Figure 4.1: **Leakage reduction unit scheme.** (a) Schematic for the driven transmon-resonator system. A transmon (T, yellow) with three lowest-energy levels $|g\rangle$, $|e\rangle$, and $|f\rangle$ is coupled to a readout resonator (R) with strength g. The latter is coupled to a frequency-matched Purcell resonator (P) with strength J. The Purcell resonator also couples to a 50 Ω feedline through which its excitations quickly decay at rate κ. The transmon is driven with a pulse of strength Ω applied to its microwave drive line. (b) Energy level-spectrum of the system. Levels are denoted as $|T, R, P\rangle$, with numbers indicating photons in R and P. As the two resonators are frequency matched, the right-most degenerate states split by $2J$, and g is shared equally among the two hybridized resonator modes $|1^-\rangle$ and $|1^+\rangle$. An effective coupling $\tilde{g}$ arises between $|f00\rangle$ and the two hybridized states $|g1^\pm\rangle$ via $|e00\rangle$ and $|e1^\pm\rangle$. (c) Spectroscopy of the $|f00\rangle \leftrightarrow |g1^\pm\rangle$ transition. Measured transmon population in $|f\rangle$ versus drive frequency, showing dips corresponding to the two transitions assisted by each of the hybridized resonator modes.

can be easily corrected using a Z-axis rotation. As a first application in a QEC setting, we interleave repeated measurements of a weight-2 parity check [18, 28] with simultaneous LRUs on data and ancilla qubits, showing the suppression of leakage and error detection rate buildup.

4.2. Results

Our leakage reduction scheme [Fig. 4.1(a)] consists of a transmon with states $|g\rangle$, $|e\rangle$ and $|f\rangle$, driven by an external drive Ω, coupled to a resonant pair of Purcell and readout resonators [42] with effective dressed states $|00\rangle$ and $|1^\pm\rangle$. The LRU scheme transfers leakage population in the second-excited state of the transmon, $|f\rangle$, to the ground state, $|g\rangle$, via the resonators using a microwave drive. It does so using an effective coupling, $\tilde{g}$, mediated by the transmon-resonator coupling, g, and the drive Ω, which couples states

$|f00\rangle$ and $|g1^{\pm}\rangle$. Driving at the frequency of this transition,

$$\omega_{f00} - \omega_{g1^{\pm}} \approx 2\omega_Q + \alpha - \omega_{RP}, \tag{4.1}$$

transfers population from $|f00\rangle$ to $|g1^{\pm}\rangle$, which in turn quickly decays to $|g00\rangle$ provided the transition rate, $\tilde{g}$, is small compared to κ. Here, ω_Q and α are the transmon qubit transition frequency and anharmonicity, respectively, while ω_{RP} is the resonator mode frequency. In this regime, the drive effectively pumps any leakage in $|f\rangle$ to the computational state $|g\rangle$. We perform spectroscopy of this transition by initializing the transmon in $|f\rangle$ and sweeping the drive around the expected frequency. The results [Fig. 4.1(c)] show two dips in the f-state population corresponding to transitions with the hybrized modes of the matched readout-Purcell resonator pair. The dips are broadened by $\sim \kappa_{eff}/2\pi \approx 8$ MHz, where $\kappa_{eff} = \kappa/2$ is the effective linewidth of the dressed resonator (see Sec. 4.4.1 and Tab. 4.1), making them easy to find. We achieve typical couplings of $\tilde{g}/2\pi \sim 1$ MHz for this transition, see Sec. 4.4.3.

To make use of this scheme for a LRU, we calibrate a pulse that can be used as a circuit-level operation. We use the pulse envelope proposed in Ref. [2]:

$$A(t) = \begin{cases} A\sin^2\left(\pi\frac{t}{2t_r}\right) & \text{for } 0 \leq t \leq t_r, \\ A & \text{for } t_r \leq t \leq t_p - t_r, \\ A\sin^2\left(\pi\frac{t_p - t}{2t_r}\right) & \text{for } t_p - t_r \leq t \leq t_p, \end{cases} \tag{4.2}$$

where A is the amplitude, t_r is the rise and fall time, and t_p is the total duration. We conservatively choose $t_r = 30$ ns to avoid unwanted transitions in the transmon. To measure the fraction of leakage removed, R, we apply the pulse on the transmon prepared in $|f\rangle$ and measure it [Fig. 4.2(a)], correcting for readout error using the measured 3-level assignment fidelity matrix [Fig. 4.2(c)]. To optimize the pulse parameters, we first measure R while sweeping the pulse frequency and A [Fig. 4.2(d)]. A second sweep of t_p and A [Fig. 4.2(e)] shows that $R > 99\%$ can be achieved by increasing either parameters. This value is limited by thermal population in the resonator modes. We estimate values of $P(n = 1) \approx 0.5\%$, see Sec. 4.4.4. Simulation [2] suggests that $R \approx 80\%$ is already sufficient to suppress most of the impact of current leakage rates, which is comfortably achieved over a large region of parameter space. For QEC, a fast operation is desirable to minimize the impact of decoherence. However, one must not excessively drive the transmon, which can cause extra decoherence (see Fig. 6 in Ref. [2]). Considering the factors above, we opt for $t_p = 220$ ns and adjust A such that $R \gtrsim 80\%$. Additionally, we benchmark the repeated action of the LRU and verify that its performance is maintained over repeated applications, thus restricting leakage events to approximately a single cycle (see Fig. 4.6).

With the LRU calibrated, we then benchmark its impact on the qubit subspace using quantum process tomography. The results (Fig. 4.3) show that the qubit incurs a Z-axis rotation. We find that the rotation angle increases linearly with t_p [Fig. 4.3(g)], consistent with a 71(9) kHz AC-Stark shift induced by the LRU drive. This phase error in the qubit subspace can be avoided using decoupling pulses or corrected with a virtual Z gate. Figures 4.3(h) and 4.3(i) show the Pauli transfer matrix (PTM) for the operation before and after applying a virtual Z correction, respectively. From the measured PTM [Fig. 4.3(i)] and enforcing physicality constraints [43], we obtain an average gate fidelity
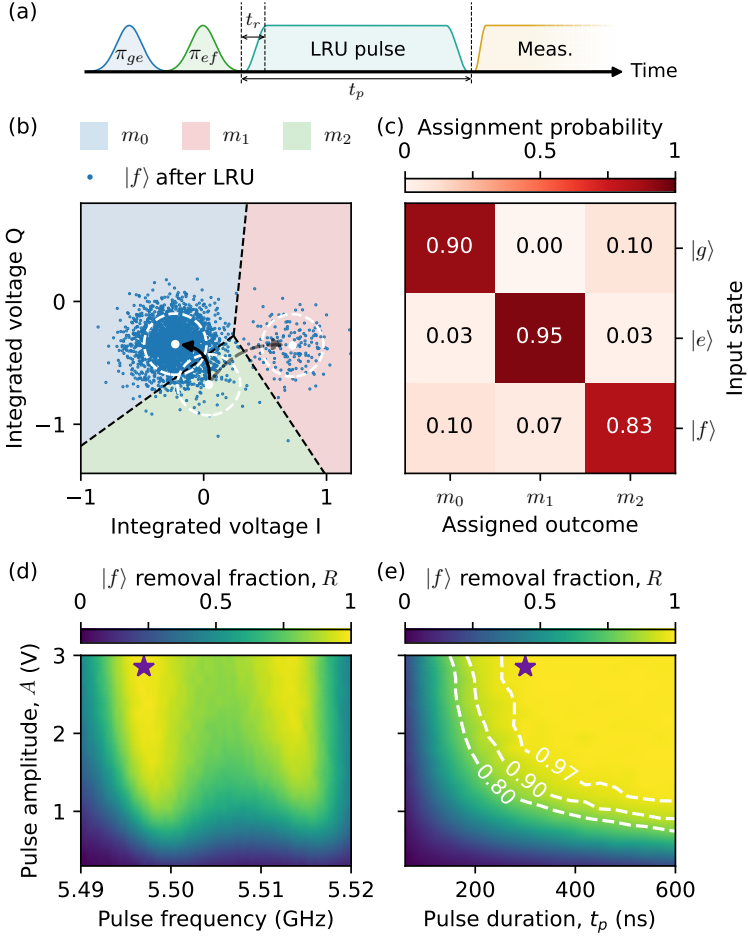
Figure 4.2: **Calibration of the leakage reduction unit pulse.** (a) Pulse sequence used for LRU calibration. (b) Single-shot readout data obtained from the experiment. The blue, red and green areas denote m_0, m_1, and m_2 assignment regions, respectively. The mean (white dots) and 3σ standard deviation (white dashed circles) shown are obtained from Gaussian fits to the three input-state distributions. The blue data shows the first 3×10^3 (from a total of 2^{15}) shots of the experiment described in (a), indicating 99.(3)% $|f\rangle$-state removal fraction. (c) Measured assignment fidelity matrix used for readout correction. (d-e) Extracted $|f\rangle$-state removal fraction versus pulse parameters. Added contours (white dashed curves) indicate 80, 90 and 97% removal fraction. The purple star indicates the pulse parameters used in (b).
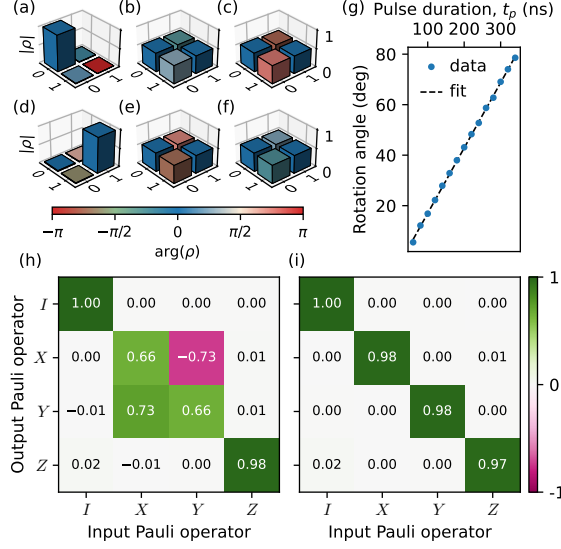
Figure 4.3: **Process tomography of the leakage reduction unit.** (a-f) Measured density matrices after the LRU gate for input states $|0\rangle$, $|+\rangle$, $|+i\rangle$, $|1\rangle$, $|-\rangle$, and $|-i\rangle$, respectively. (g) Z-rotation angle induced on the qubit versus the LRU pulse duration. The linear best fit (black dashed line) indicates an AC-Stark shift of 71(9) kHz. (h-i) Pauli transfer matrix of the LRU with (i) and without (h) virtual phase correction ($t_{\mathrm{p}} = 220$ ns and $R = 84.(7)\%$).

$F_{\mathrm{avg}} = 98.(9)\%$. Compared to the measured 99.(2)% fidelity of idling during the same time ($t_{\mathrm{p}} = 220$ ns), there is evidently no significant error increase.

Finally, we implement the LRU in a QEC scenario by performing repeated stabilizer measurements of a weight-2 X-type parity check [18, 28] using three transmons (Fig. 4.4). We use the transmon in Figs. 1-3, D_1, plus an additional transmon (D_2) as data-qubits together with an ancilla, A. LRUs for D_2 and A are tuned using the same procedure as above. A detailed study of the performance of this parity check and of the impact of simultaneous LRUs is shown in Figs. 4.10 and 4.9. Given their frequency configuration [44], D_1 and A are most vulnerable to leakage during two-qubit controlled-Z (CZ) gates, as shown by the avoided crossings in Fig. 4.4(a). Additional leakage can occur during other operations: in particular, we observe that leakage into states above $|f\rangle$ can occur in A due to measurement-induced transitions [8] (see Fig. 4.14). Therefore, a LRU acting on $|f\rangle$ alone is insufficient for A. To address this, we develop an additional LRU for $|h\rangle$ (h-LRU), the third-excited state of A (see Fig. 4.13). The h-LRU can be employed simultaneously with the f-LRU without additional cost in time or impact on the $|f\rangle$ removal fraction, R. Thus, we simultaneously employ f-LRUs for all three qubits and an h-LRU for A [Fig. 4.4(a)]. To evaluate the impact of the LRUs, we measure the error detection probability (probability of a flip occurring in the measured stabilizer parity) and leakage population of the three transmons over multiple rounds of stabilizer measurement. Without leakage reduction, the error detection probability rises $\sim 8\%$ in 50 rounds. We attribute this feature to leakage build-up [19, 24, 35]. With the LRUs, the rise
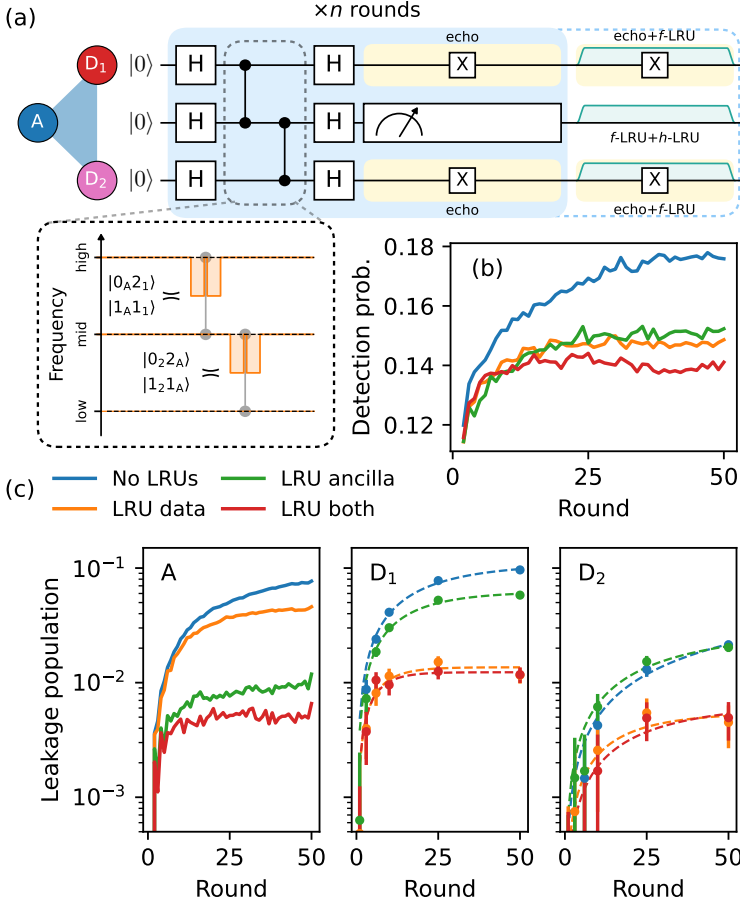
Figure 4.4: **Repeated stabilizer measurement with leakage reduction.** (a) Quantum circuit using ancilla A to measure the X-type parity of data qubits D_1 and D_2. The dashed box shows the frequency arrangement for two-qubit CZ gates. A CZ gate is performed by fluxing the higher-frequency transmon down in frequency to the nearest avoided crossing (orange shaded trajectory). The duration of single-qubit gates, CZ gates, and measurement are 20, 60 and 340 ns, respectively, totalling 500 ns for the parity check (light-blue region). Performing the LRUs extends the circuit by 220 ns (blue-dashed region). Echo pulses on data qubits mitigate phase errors caused by residual ZZ crosstalk and AC-stark shift during the measurement and LRUs (light yellow slots). (b-c) Measured error-detection probability (b) and leakage (c) versus the number of parity-check rounds in four settings. Here, leakage includes any population outside the computational subspace. The No LRUs setting (blue) does not apply any LRUs. LRU data (orange) and LRU ancilla (green) settings apply LRUs exclusively on the data qubits and the ancilla, respectively. The LRU both (red) setting applies LRUs on all qubits.

stabilizes faster (in ~ 10 rounds) to a lower value and is limited to 2%, despite the longer cycle duration (500 versus 720 ns without and with the LRU, respectively). Leakage is overall higher without LRUs, in particular for D_1 and A [Fig. 4.4(c)], which show a steady-state population of $\approx$ 10%. Using leakage reduction, we lower the leakage steady-state population by up to one order of magnitude to $\lesssim$ 1% for all transmons. Additionally, we find that removing leakage on other transmons leads to lower overall leakage, suggesting that leakage is transferred between transmons [17, 35]. This is particularly noticeable in A [Fig. 4.4(c)], where the steady-state leakage is always reduced by adding LRUs on D_1 and D_2.

4.3. DISCUSSION

We have demonstrated and extended the all-microwave LRU for superconducting qubits in circuit QED proposed in Ref. [2]. We have shown how these LRUs can be calibrated using a straightforward procedure to deplete leakage in the second- and third-excited states of the transmon. This scheme could potentially work for even higher transmon states using additional drives. We have verified that the LRU operation has minimal impact in the qubit subspace, provided one can correct for the static AC-Stark shift induced by the drive(s).

   This scheme does not reset the qubit state and is therefore compatible with both data and ancilla qubits in the QEC context. We have showcased the benefit of the LRU in a building-block QEC experiment where LRUs decrease the steady-state leakage population of data and ancilla qubits by up to one order of magnitude (to $\lesssim$ 1%), and thereby reduce the error detection probability of the stabilizer and reaching a faster steady state. We find that the remaining ancilla leakage is dominated by higher states above $|f\rangle$ (see Fig. 4.14) likely caused by the readout [8, 9]. Given the observation leakage transfer between transmons, which can result in higher excited leakage states [35], data qubits can also potentially benefit from h-LRUs. Compared to other LRU approaches [19, 35], we believe this scheme is especially practical as it is all-microwave and very quantum-hardware efficient, requiring only the microwave drive line and dispersively coupled resonator that are already commonly found in the majority of circuit QED quantum processors [21, 22, 24]. Extending this leakage reduction method to larger QEC experiments can be done without further penalty in time as all LRUs can be simultaneously applied. However, we note that when extending the LRU to many qubits, microwave crosstalk should be taken into account in order to avoid driving unwanted transitions. This can be easily avoided by choosing an appropriate resonator-qubit detuning.

4.4. SUPPLEMENTAL MATERIAL

4.4.1. DEVICE

The device used (Fig. 4.5) has 17 flux-tunable transmons arranged in a square lattice with nearest-neighbor connectivity (as required for a distance-3 surface code). Transmons are arranged in three frequency groups as prescribed in the pipelined architecture of Ref. [44]. Each transmon has a dedicated microwave drive line used for single-qubit gates and leakage reduction, and a flux line used for two-qubit gates. Nearest-neighbor transmons have fixed coupling mediated by a dispersively coupled bus resonator. Each
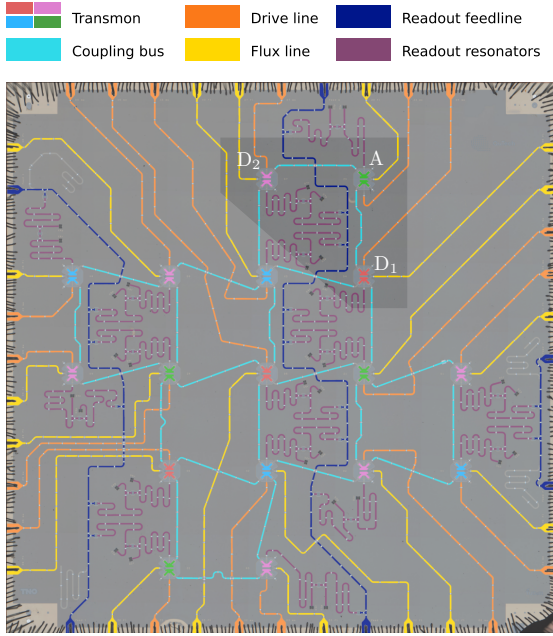
Figure 4.5: **Circuit QED device.** Optical image of the 17-transmon quantum processor, with added falsecolor to highlight different circuit elements. The shaded area indicates the three transmons used in this experiment.

transmon has a dedicated pair of frequency-matched readout and Purcell resonators coupled to one of three feedlines, used for fast multiplexed readout in the architecture of Ref. [42]. Single-qubit gates are realized using standard DRAG pulses [4]. Two-qubit controlled-Z gates are implemented using sudden net-zero flux pulses [7]. Characteristics and performance metrics of the three transmons used in the experiment are shown in Tab. 4.1.

4.4.2. Repeated LRU application

For QEC we require that the LRU performance remains constant over repeated applications. To assess this, we perform repeated rounds of the experiment shown in Fig. 4.6 while idling or using the LRU. In each round apply an e-f rotation with rotation angle θ to induce a leakage rate

$$L_1 = \frac{\sin^2(\theta/2)}{2}, \tag{4.3}$$

and choose θ such that $L_1 = 2\%$. For the purpose of this experiment, we lower the readout amplitude in order to suppress leakage to higher states during measurement [8], The results (Fig. 4.6) show that while idling, leakage in $|f\rangle$ builds up to a steady-state population of about 16%. Using the LRU, it remains constant at $P_f = L_1$ throughout all rounds. This behavior shows that LRU performance is maintained throughout repeated applications and suggests that leakage events are restricted to a single round.

Transmon	D_1	A	D_2
Frequency at sweetspot, $\omega_Q/2\pi$ (GHz)	6.802	6.033	4.788
Anharmonicity, $\alpha/2\pi$ (MHz)	-295	-310	-321
Resonator frequency, $\omega_{R/P}/2\pi$ (GHz)	7.786	7.600	7.105
Purcell res. linewidth, $\kappa/2\pi$ (MHz)	15.5	22.5	12.6
Qubit-res. coupling, $g/2\pi$ (MHz)	172	212	301
f-LRU drive frequency (GHz)	5.498	4.135	2.152
h-LRU drive frequency (GHz)	-	3.496	-
T_1 (μs)	17	26	37
T_2 (μs)	19	22	27
Single-qubit gate error (%)	0.1(0)	0.0(7)	0.0(5)
Two-qubit gate error (%)		1.(1) 1.(9)	
Two-qubit gate leakage (%)		0.3(7) 0.1(1)	
f-LRU removal fraction, R^f (%)	84.(7)	99.(2)	80.(3)
h-LRU removal fraction, R^h (%)	-	96.(1)	-
Operation	Duration (ns)		
Single-qubit gate	20		
Two-qubit gate	60		
Measurement	340		
LRU	220		

Table 4.1: **Device metrics.** Frequencies and coherence times are measured using standard spectroscopy and time-domain measurements [45]. Gate errors are evaluated using randomized benchmarking protocols [46–48].

4.4.3. ESTIMATING EFFECTIVE TRANSITION COUPLING

In order to assess the effective transition coupling, $\tilde{g}^f$, we study the system described in Fig. 1(a). Excluding the drive, the total Hamiltonian of this system is given by

$$H = H_R + H_P + H_Q +$$
$$J(a_R^\dagger a_P + a_R a_P^\dagger) + g(a_R^\dagger b + a_R b^\dagger), \qquad (4.4)$$

where $H_{R/P} = \omega_{R/P} a^\dagger a$ and $H_Q = \omega_Q b^\dagger b + \frac{\alpha}{2} b^\dagger b^\dagger bb$ are the resonators and transmon Hamiltonians, respectively. When the two resonator modes are resonant, i.e., $\omega_R = \omega_P$,

$$H = H_R^+ + H_R^- + H_Q +$$
$$\frac{g}{\sqrt{2}} [(a_+^\dagger b + a_+ b^\dagger) + (a_-^\dagger b + a_- b^\dagger)], \qquad (4.5)$$

where $a_\pm = (a_R \pm a_P)/\sqrt{2}$ and $H_R^\pm = (\omega_R \pm J) a_\pm^\dagger a_\pm$ describe the dressed resonator modes. In this basis, we find that g is shared by the dressed resonator modes with effective
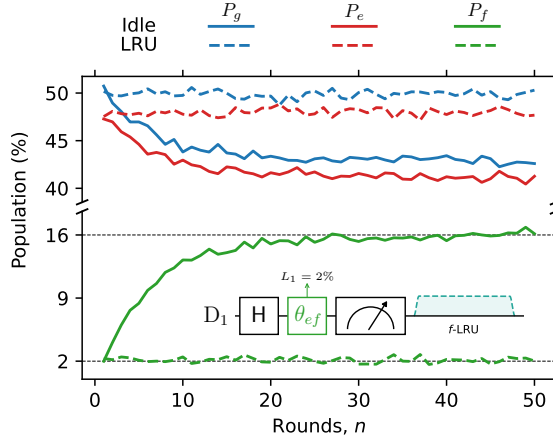
Figure 4.6: **Repeated LRUs on** D_1**.** Computational (top) and leakage (bottom) state population over repeated cycles of the circuit as shown. Transmon D_1 is repeatedly put in superposition, controllably leaked with rate L_1 and measured. Measurement is followed by either idling (solid) or the LRU (dashed). The horizontal lines denote the steady-state leakage with and without LRUs.
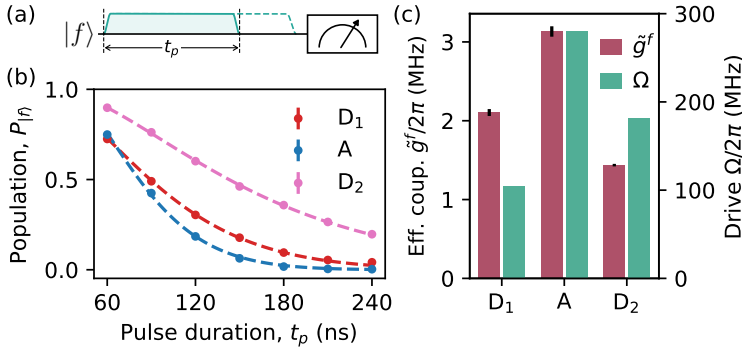
Figure 4.7: **Assessing the** $|f00\rangle \leftrightarrow |g1^{\pm}\rangle$ **transition coupling.** (a) Experimental circuit to measure $P_{|f\rangle}$ versus LRU pulse duration, t_p. After initializing the transmon in $|f\rangle$, we perform a LRU pulse of duration t_p within a constant interval of 240 ns followed by a measurement. (b) Measured $|f\rangle$ state population, $P_{|f\rangle}$, as a function of t_p for each of the three transmons used. The dashed lines show fits to the data used to estimate the effective transition coupling $\tilde{g}^f$. (c) Transition couplings, $\tilde{g}^f$, and corresponding drive strengths, Ω, estimated from fitting.
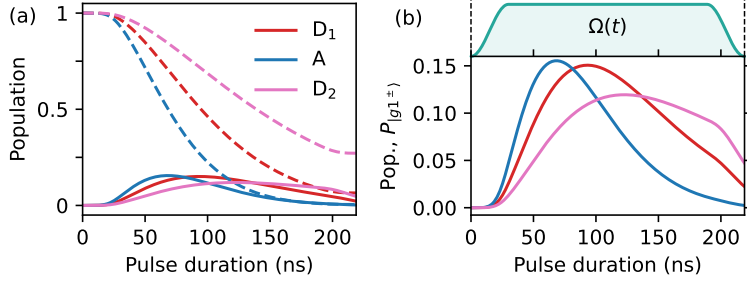
Figure 4.8: **Dynamics of the states $|f00\rangle$ and $|g1^\pm\rangle$ during the LRU**. (a) Population of the state $|f00\rangle$ (dashed) and $|g1^\pm\rangle$ (solid) during the LRU pulse, $\Omega(t)$, obtained by numerically solving Eq. 4.6. (b) Close-up of population in $|g1^\pm\rangle$ and LRU pulse (top) versus time.

strength $g_{\text{eff}} = g/\sqrt{2}$. Because of this, each dresssed resonator mode approximately inherits an effective linewidth $\kappa_{\text{eff}} = \kappa/2$. We then study the dynamics of this system in the three-level manifold $\{|g00\rangle, |g1^\pm\rangle, |f00\rangle\}$ by solving the Lindblad equation

$$\dot{\rho} = -i\left[H, \rho\right] + L\rho L^\dagger - \frac{1}{2}\{L^\dagger L, \rho\}, \tag{4.6}$$

where $L = \sqrt{\kappa_{\text{eff}}}|g00\rangle\langle g1^\pm|$ is the decay operator modeling loss in the resonator mode and

$$H = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \tilde{g}^f \\ 0 & (\tilde{g}^f)^* & 0 \end{bmatrix} \tag{4.7}$$

is the approximate Hamiltonian of the system in the drive frame on resonance with $|f00\rangle \leftrightarrow |g1^\pm\rangle$. In this model, an initial state $|f00\rangle$ evolves as (Eq. S2 of Ref. [26]),

$$P_{|f00\rangle}(t) = e^{-\frac{\kappa_{\text{eff}}}{2}t}\left|\cosh\left(\frac{\zeta}{2}t\right) + \frac{\kappa_{\text{eff}}}{2\zeta}\sinh\left(\frac{\zeta}{2}t\right)\right|^2, \tag{4.8}$$

where, $\zeta = \sqrt{-(2\tilde{g}^f)^2 + (\kappa_{\text{eff}}/2)^2}$. To estimate $\tilde{g}^f$, we measure $P_{|f\rangle}$ as function of pulse duration, t_p, as shown in Fig. 4.7 and fit

$$P_{|f\rangle}(t_p, \kappa_{\text{eff}}, \tilde{g}^f, t_0) = P_{|f00\rangle}(t_p - t_0), \tag{4.9}$$

where t_0 is introduced to account for the finite rise and fall time of the LRU drive. We fix $\kappa_{\text{eff}} \equiv \kappa/2$ (shown in Tab. 4.1) leaving only two parameters to fit $\tilde{g}^f$, and t_0. The results [Fig. 4.7(c)] show $\tilde{g}^f/2\pi$ lies between 1 and 3 MHz. Additionally, we find $t_0 \approx 20$ ns similar to the rise time, $t_r = 30$ ns, used in experiment. From these results, one can also estimate the drive strength Ω used for each transmon using (Eq. A35 of Ref. [2]),

$$\tilde{g}^f = \Omega\frac{g_{\text{eff}}\alpha}{\sqrt{2}\Delta(\Delta + \alpha)}, \tag{4.10}$$

where $\Delta \equiv \omega_Q - \omega_R$ and $g_{eff} \equiv g/\sqrt{2}$. We find drive strengths, $\Omega/2\pi$, lying between 100 and 300 MHz depending on the qubit. This disparity occurs because the drives are synthesized at distinct frequencies (from 2 to 5.5 GHz) and therefore use different RF components and experience different attenuation. We note that the amplitudes used in Fig. 4.7 are slightly different from the ones used in the experiments reported in the main text however they lead to similar removal fractions.
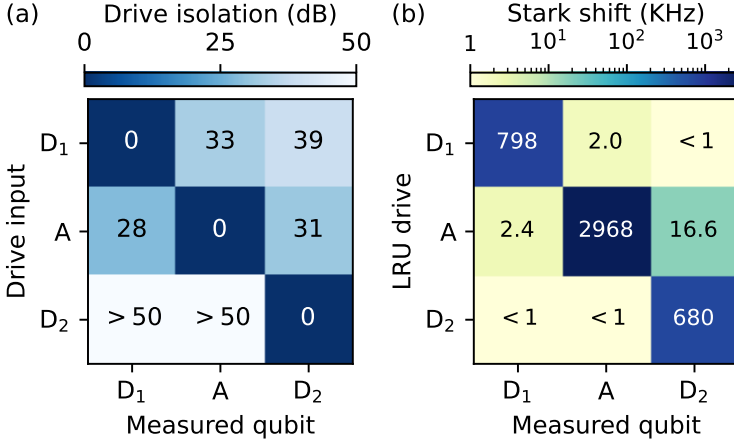
Figure 4.9: **Cross-driving induced AC Stark shift from LRU drives.** (a) Measured drive isolation between the qubits. For each element in the matrix S_{ij}, we measure the isolation of qubit i to a drive Ω_j driven through the drive line of qubit j such that its effective strength is $\Omega_i^{eff} = 10^{-S_{ij}/20}\Omega_j$. (b) Estimated AC Stark shift induced by the LRU drive between qubits based on the drive amplitudes calculated in Fig. 4.7(c). We plot absolute frequency shifts, however, actual estimated values are negative.

4.4.4. POPULATION IN THE READOUT RESONATORS
In this section, we consider the impact of population in the resonator modes. Thermal population in the resonators can converted into leakage during the LRU [2]. As shown in Fig. 2(d) of Ref. [2], when there is no leakage, the system will evolve such that $P_{|f00\rangle} = P_{|g1\pm\rangle}$ in the steady state. It is thus important that the mean photon number, $\bar{n}$, characterising the thermal field in the resonators is low, $\bar{n} \ll 1$. One can estimate an upper bound of $\bar{n}$ assuming that the pure dephasing rate of the transmon, Γ_ϕ, is limited by photon shot noise in the readout resonators using [45], namely

$$\Gamma_\phi = \eta \frac{4\chi^2}{\kappa}\bar{n}, \tag{4.11}$$

where $\eta = \kappa^2/(\kappa^2 + 4\chi^2)$. Using this estimate, we find $P(n = 1) \approx 0.5\%$. Therefore, one can expect at most $P_{|f00\rangle} \approx 0.5\%$ in the steady state, which will limit the removal fraction $R < 99.5\%$. Another important aspect to consider when using the LRU repeatedly, is the amount of population left in $|g1\pm\rangle$ after a leakage event. If significant population is left in the resonator, errors can occur in subsequent operations. We investigate this

by numerically solving Eq. 4.6 taking into account the rise time of the LRU pulse and using the $\tilde{g}$ estimated in Fig. 4.7. Figure 4.8 shows the evolution of state $|g1^\pm\rangle$ for the three transmons. We find that the population in this level never exceeds 15%. Moreover, Fig. 4.8(b) reveals that the leftover population by the end of the pulse is below 5%. In particular, we see that population is significantly suppressed during the pulse fall time t_r. Therefore, we do not expect errors following the LRU to be significant after a leakage event.

4.4.5. DRIVE CROSSTALK

Given the relatively high strength of the drives required to drive the LRUs, we consider the impact of drive crosstalk. In particular, we study the impact of each LRU drive on each qubit. By coupling off-resonantly, cross-drives may shift the qubit frequency which could lead to lower performance of the LRUs when driven simultaneously. To investigate this, we first characterize drive crosstalk between the three transmons used by measuring the drive isolation,

$$S_{ij} = 20\log_{10}(\Omega_j/\Omega_i^{\text{eff}}), \tag{4.12}$$

between a drive, Ω_j, on qubit j and the effective drive, Ω_i^{eff}, felt on qubit i. The results [Fig. 4.9(a)] show that drive isolation is at least 28 dB and crosstalk is most significant when driving via the drive lines of D_1 and A. Using the drive strength estimated in Fig. 4.7, we then calculate the AC Stark shift, δ, induced by an effective drive Ω_{eff} using,

$$\delta = \frac{\alpha\Omega_{\text{eff}}^2}{2\Delta_d(\Delta_d + \alpha)}, \tag{4.13}$$

where $\Delta_d = \omega_{\text{drive}} - \omega_Q$. The results [Fig. 4.9(b)] show that crosstalk induced shifts are at most ≈ 17 kHz. This is much smaller than the typical linewidth of the $|f00\rangle \leftrightarrow |g1^\pm\rangle$ transition ~ 10 MHz [as shown in Fig. 1(c)]. Therefore, one does not expect lower simultaneous performance of LRUs due to cross-drive Stark shifts. One can also verify this experimentally by measuring simultaneous leakage removal fractions, R, and comparing them to individual ones. Performing this experiment, we find the same removal fractions, as expected.

4.4.6. BENCHMARKING THE WEIGHT-2 PARITY CHECK

We benchmark the performance of the weight-2 parity check using three experiments assessing different error types. First, we assess the ability to accurately assign the parity of the data-qubit register by measuring the ancilla outcome for all data-qubit input computational states. The results [Fig. 4.10(a)] show an average parity assignment fidelity of 95.6%. Next, we look at errors occurring on the data qubits when projecting them onto a Bell state using a X-type parity check [Fig. 4.10(b)]. From data-qubit state tomography conditioned on ancilla outcome, we obtain an average Bell-state fidelity of 97.7% (96.9% for $m = +1$ and 98.5% for $m = -1$). For each reported density matrix, we apply readout corrections and enforce physicality constraints via maximum likelihood estimation [43]. Finally, we look at the backaction of two back-to-back parity checks [Fig. 4.10(c)]. Here, we measure the correlation of the two parity outcomes. Ideally, the first parity outcome should be random while the second should be the same as the first.
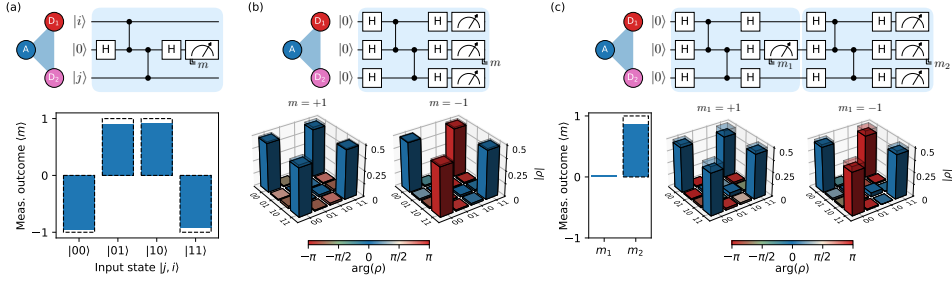
Figure 4.10: **Benchmarking of the weight-2 parity check.** (a) Quantum circuit of the weight-2 X-type parity check and bar plot of the average measured ancilla outcome for the different input computational states of the data-qubit register. Dashed bars show ideal average outcome: $\langle m \rangle = +1\,(-1)$ for even (odd) data-qubit input parity. (b) Generation of Bell states via stabilizer measurement (top) and corresponding data-qubit state tomography (bottom) conditioned on the ancilla outcome. The obtained fidelity to the ideal Bell states (shaded wireframe) is 96.9% and 98.5% for $m = +1$ and $m = -1$, respectively. (c) Repeated stabilizer experiment. (Bottom left) Average measured ancilla outcome $\langle m \rangle$ for each round of stabilizer measurement. Ideally, the first outcome should be random and the second always +1. The measured probability is $P(m_2 = +1) = 90.0\%$. (Bottom right) Reconstructed data-qubit states conditioned on the first ancilla outcome. The obtained Bell-state fidelities are 90.6% and 91.5% for $m_1 = +1$ and $m_1 = -1$, respectively.

Since our ancilla is not reset after measurement, the probability of both parities being correlated is $P(m_2 = +1) = 90.0\%$ [bar plot in Fig. 4.10(c)]. We can also reconstruct the data qubit state after the experiment. Here, we find that the average Bell-state fidelity drops to 91.0% (90.6% for $m_1 = +1$ and 91.5% for $m_1 = -1$). This drop in fidelity is likely due to decoherence from idling during the first ancilla measurement.

4.4.7. MEASUREMENT-INDUCED TRANSITIONS

Previous studies have observed measurement-induced state transitions that can lead to leakage [8, 9]. To evaluate the backaction of ancilla measurement, we model the measurement as a rank 3 tensor $\epsilon_i^{m,j}$ which takes an input state i, declares an outcome m and outputs a state j with normalization condition,

$$\sum_m \sum_j \epsilon_i^{m,j} = 1. \tag{4.14}$$

To find $\epsilon_i^{m,j}$, we perform the experiment in Fig. 4.11(a). For each input state i, the probability distribution of the measured results $P_i(M_0, M_1)$ follows

$$P_i(M_0 = m_k, M_1 = m_\ell) = \sum_s \sum_j \epsilon_i^{m_k, s} \epsilon_s^{m_\ell, j}. \tag{4.15}$$

This system of 27 second-order equations is used to estimate the 27 elements of $\epsilon_i^{m,j}$ through a standard optimization procedure. In this description, the assignment fidelity matrix $M_{i,m}$ [Fig. 4.11(b)] is given by

$$M_{i,m} = \sum_j \epsilon_i^{m,j}. \tag{4.16}$$
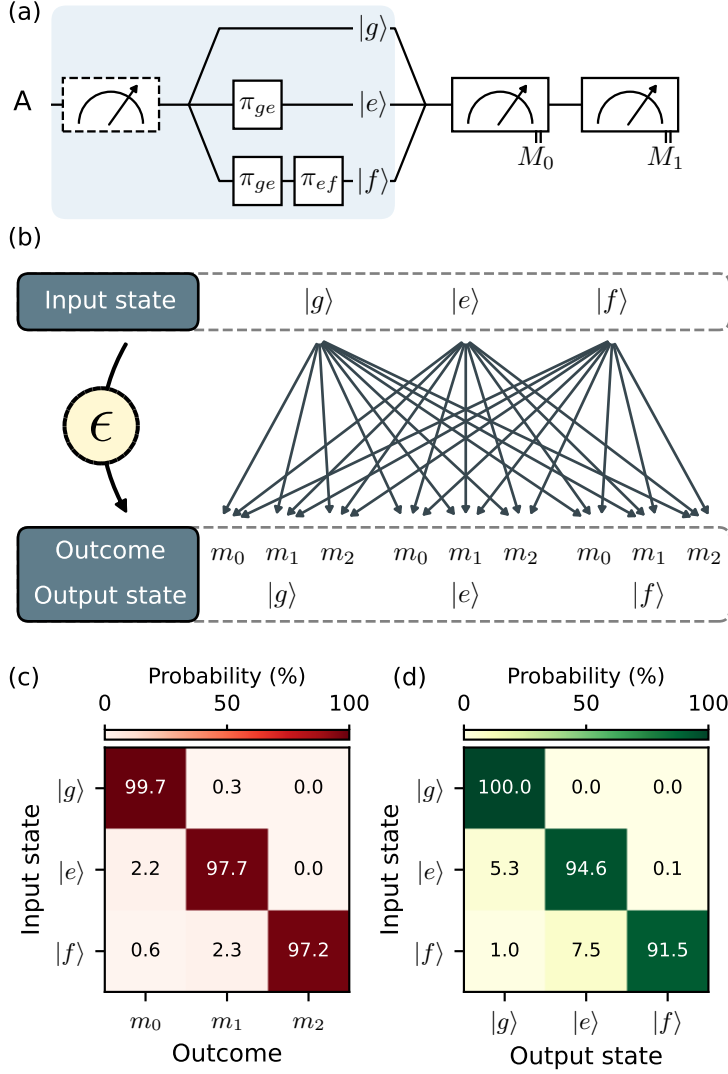
Figure 4.11: **Characterizing measurement-induced transitions.** (a) Quantum circuit used to characterize transmon measurement. A transmon is initialized into states $|g\rangle$, $|e\rangle$ and $|f\rangle$ after a heralding (dashed) measurement (blue panel). Following prepration, two consecutive measurements M_0 and M_1 are performed, yielding three-level outcomes. (b) Illustration of the extracted measurement model. The model is described by a rank 3 tensor $\epsilon_i^{m,j}$, where input states i are connected to measurement outcomes m and output states j. From it, the assignment probability matrix (c) and the QNDness matrix (d) can be extracted.
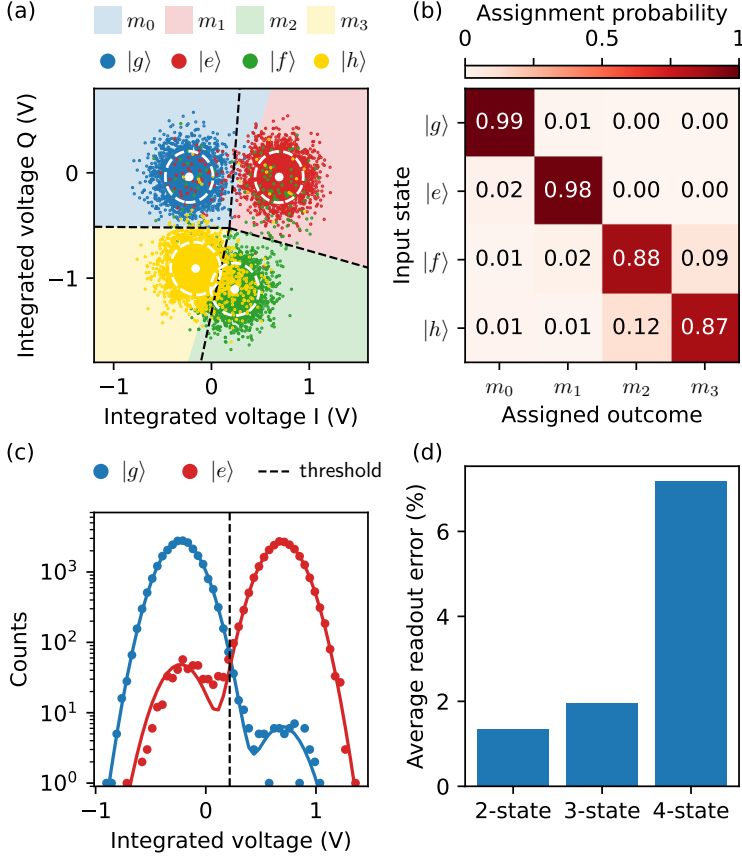
Figure 4.12: **Four state readout.** (a) Single-shot readout data for the four lowest-energy transmon states $|g\rangle$, $|e\rangle$, $|f\rangle$ and $|h\rangle$ of A. Data are plotted for the first 3×10^3 from a total of 2^{15} shots for each input state. The dashed lines show decision boundaries obtained from fitting a linear discrimination classifier to the data. The mean (white dot) and 3σ standard deviation (white dashed circles) shown are obtained from Gaussian fits to each input state distribution. (b) Assignment probability matrix obtained from classification of each state into a quaternary outcome. (c) Histogram of shots for qubit states taken along the projection maximizing the signal-to-noise ratio. (d) The average assignment errors for 2-, 3- and 4-state readout are 1.(3), 1.(9) and 7.(2)%, respectively.

Furthermore, this model allows us to assess the probability of transitions occurring during the measurement. This is given by the QNDness matrix

$$Q_{i,j} = \sum_m \epsilon_i^{m,j}. \tag{4.17}$$

The results [Fig. 4.11(b)] show an average QNDness of 95.4% across all states. The average leakage rate ($(Q_{g,f} + Q_{e,f})/2$) is 0.06%, predominantly occurring for input state $|e\rangle$.

4.4.8. READOUT OF TRANSMON STATES

In order to investigate leakage to higher states in the ancilla, we need to discriminate between the first two leakage states, $|f\rangle$ and $|h\rangle$. To do this without compromising the performance of the parity check, we simultaneously require high readout fidelity for the qubit states $|g\rangle$ and $|e\rangle$. We achieve this for the ancilla for the states $|g\rangle$ through $|h\rangle$ using a single readout pulse. Figure 4.12(a) shows the integrated readout signal for each of the states along with the decision boundaries used to classify the states. Any leakage to even higher states will likely be assigned to $|h\rangle$ since the resonator response at the readout frequency is mostly flat for $|h\rangle$. The average assignment error for the four states is 7.(2)% [Fig. 4.12(b)] while the average qubit readout error is 1.(3)% [Fig. 4.12(c)]. Here, we assume that state preparation errors are small compared to assignment errors.

4.4.9. LEAKAGE REDUCTION FOR HIGHER STATES

Although most common leakage mechanisms usually populate the second-excited state of the transmon, $|f\rangle$, some operations such as the measurement can leak into higher-excited states [8]. We observe the build-up of population in these higher states in the repeated parity-check experiment (Fig. 4). Figure 4.14 shows the fraction of total leakage to these higher states for the ancilla. Therefore, leakage reduction for higher states is necessary for ancillas. Similar to the leakage reduction mechanism that drives $|f\rangle \rightarrow |g\rangle$ [with effective coupling $\tilde{g}^f$ in Fig. 4.13(a)], one can drive $|h\rangle \rightarrow |e\rangle$ (with effective coupling $\tilde{g}^h$ in Fig. 4.13(b)]. This transition can be induced much like the former, with an extra drive at frequency

$$\omega_{h00} - \omega_{e1\pm} \approx 2\omega_Q + 3\alpha - \omega_{R/P}, \tag{4.18}$$

2α below the f-LRU transition. The effective coupling for each LRU is given by (Eq. A35 of Ref. [2]):

$$\tilde{g}^f \approx \Omega \frac{g_{eff}\alpha}{\sqrt{2}\Delta(\Delta + \alpha)} \tag{4.19}$$

and

$$\tilde{g}^h \approx \Omega \frac{\sqrt{3}g_{eff}\alpha}{\sqrt{2}(\Delta + \alpha)(\Delta + 2\alpha)}, \tag{4.20}$$

where $\Delta = \omega_Q - \omega_{R/P}$ and $g_{eff} = g/\sqrt{2}$ (see section S3). As $\tilde{g}^h/\tilde{g}^f = \sqrt{3}\Delta/(\Delta + \alpha) > 1$, one should be able to drive $|h\rangle \rightarrow |e\rangle$ with comparable performance using similar drive amplitudes. We then have two LRU mechanisms, f-LRU and h-LRU, increasing seepage from $|f\rangle$ and $|h\rangle$, respectively. We drive both of these transitions simultaneously using two independent drives. Following the same calibration procedure shown in Fig. 2 for
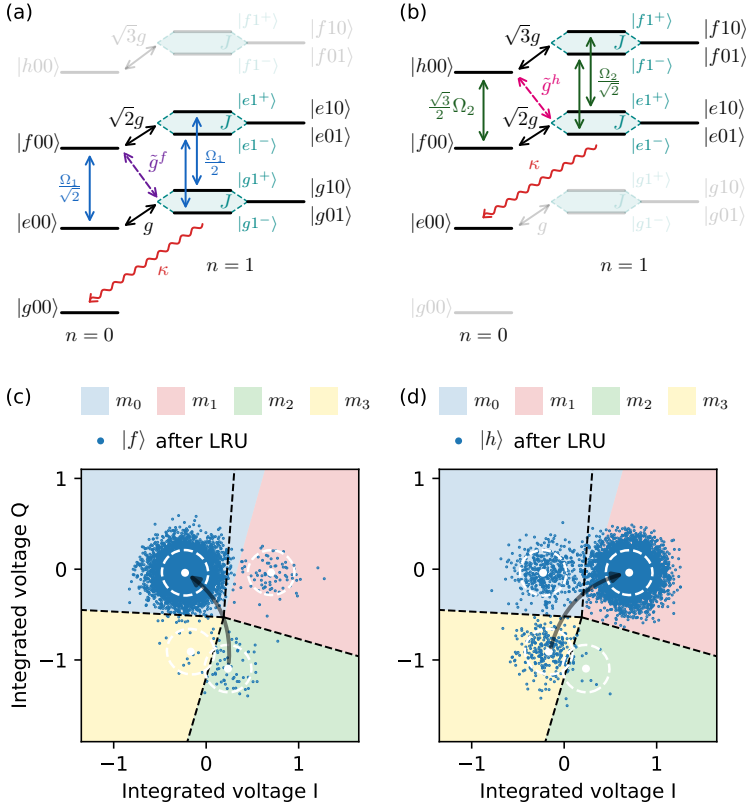
Figure 4.13: **Leakage reduction for $|f\rangle$ and $|h\rangle$.** (a, b) Transmon-resonator system level structure showing the relevant couplings for the f-LRU (a) and h-LRU (b). Each effective coupling, $\tilde{g}^f$ and $\tilde{g}^h$, is mediated by its respective drive Ω_1 and Ω_2 and transmon-resonator coupling g. (c, d) Readout data (2^{13} shots) of leakage states $|f\rangle$ (c) and $|h\rangle$ (d) after applying both LRU pulses simultaneously. The white dots and dashed cicles show the mean and 3σ standard deviation obtained from fitting calibration data for each state.
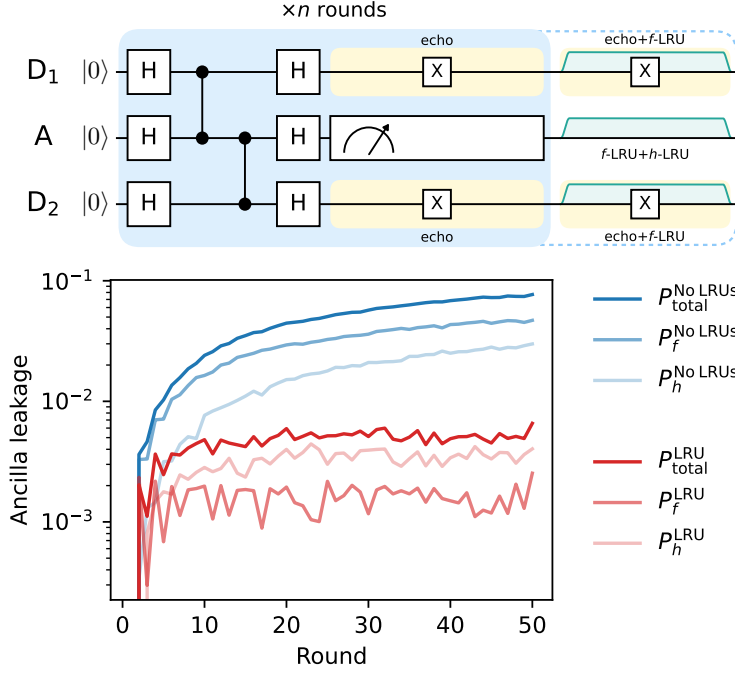
Figure 4.14: **Higher leakage states of the ancilla qubit.** Composition of ancilla leakage during the repeated stabilizer measurement of Figure 4. $P_{\text{total}} = P_f + P_h$ denotes the total leakage population without (blue) and with (red) LRUs.

the f-LRU, we tune up a pulse for the h-LRU. Figures 4.13(c) and 4.13(d) show readout data for states $|f\rangle$ and $|h\rangle$ after performing both LRUs simultaneously. The corresponding removal fraction for each state is $R^f = 99.(2)\%$ and $R^h = 96.(1)\%$ for $t_{\text{p}} = 220$ ns. Using this scheme, we can effectively reduce leakage in both states (Fig. 4.14). In particular, leakage in $|f\rangle$ is effectively kept under 0.2%, while that in $|h\rangle$ sits below 0.4% (red curves in Fig. 4.14). The former shows a flat curve and therefore corresponds to the L_1 of the cycle (similar to Fig. 4). The apparent remaining leakage in $|h\rangle$ could possibly be due to higher-excited states, which are naively assigned as $|h\rangle$ by the readout as they cannot be distinguished. One could potentially address these with additional drives. The general expression for the effective coupling of the transition targeting the mth leakage state (with $m_f = 0$, $m_h = 1$, etc.) is given by (Eq. A35 of Ref. [2])

$$\tilde{g}^m / \Omega \approx \frac{\sqrt{(m+1)(m+2)} g_{\text{eff}} \alpha}{2(\Delta + m\alpha)(\Delta + (m+1)\alpha)}, \qquad (4.21)$$

which increases monotonically with m. Therefore, one could expect comparable performances for LRUs acting on higher states.

BIBLIOGRAPHY

[1] J. F. Marques, H. Ali, B. M. Varbanov, M. Finkel, H. M. Veen, S. L. M. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider, B. M. Terhal, and L. DiCarlo, *All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits*, Phys. Rev. Lett. **130**, 250602 (2023).

[2] 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).

[3] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, et al., *Charge-insensitive qubit design derived from the Cooper-pair box*, Phys. Rev. A **76**, 042319 (2007).

[4] 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).

[5] 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).

[6] 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).

[7] 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).

[8] 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]   M. Khezri, A. Opremcak, Z. Chen, A. Bengtsson, T. White, O. Naaman, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, A. L. Crook, B. Curtin, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, B. Foxen, G. Garcia, W. Giang, M. Giustina, R. Gosula, A. G. Dau, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, J. Kelly, S. Kim, P. V. Klimov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J. Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S. Martin, M. McEwen, A. Megrant, X. Mi, K. C. Miao, S. Montazeri, A. Morvan, M. Neeley, C. Neill, A. Nersisyan, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Quintana, C. Rocque, P. Roushan, K. Sankaragomathi, K. J. Satzinger, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, N. Zhu, N. Zobrist, D. Sank, A. Korotkov, Y. Chen, and V. Smelyanskiy, *Measurement-induced state transitions in a superconducting qubit: within the rotating wave approximation*, 2022, arXiv:2212.05097 [quant-ph].

[10] 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**, 383–387 (2021).

[11] 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).

[12] 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).

[13] M. Takita, A. W. Cross, A. D. Córcoles, J. M. Chow, and J. M. Gambetta, *Experimental demonstration of fault-tolerant state preparation with superconducting qubits*, Phys. Rev. Lett. **119**, 180501 (2017).

[14] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults*, Quantum Info. Comput. **7**, 139–156 (2007).

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

[16] 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).

[17] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. Di-Carlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Information **6**, 102 (2020).

[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**, eaay3050 (2020).

[19] 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. Bush-nell, 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. Smelyan-skiy, 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**, 1761 (2021).

[20] C. Ryan-Anderson, J. G. Bohnet, K. Lee, D. Gresh, A. Hankin, J. P. Gaebler, D. Fran-cois, A. Chernoguzov, D. Lucchetti, N. C. Brown, T. M. Gatterman, S. K. Halit, K. Gilmore, J. A. Gerber, B. Neyenhuis, D. Hayes, and R. P. Stutz, *Realization of Real-Time Fault-Tolerant Quantum Error Correction*, Physical Review X **11**, Publisher: American Physical Society, 041058 (2021).

[21] S. Krinner, N. Lacroix, A. Remm, A. Di Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, *Realizing repeated quantum error correction in a distance-three surface code*, Nature **605**, 669–674 (2022).

[22] Y. Zhao, Y. Ye, H.-L. Huang, Y. Zhang, D. Wu, H. Guan, Q. Zhu, Z. Wei, T. He, S. Cao, F. Chen, T.-H. Chung, H. Deng, D. Fan, M. Gong, C. Guo, S. Guo, L. Han, N. Li, S. Li, Y. Li, F. Liang, J. Lin, H. Qian, H. Rong, H. Su, L. Sun, S. Wang, Y. Wu, Y. Xu, C. Ying, J. Yu, C. Zha, K. Zhang, Y.-H. Huo, C.-Y. Lu, C.-Z. Peng, X. Zhu, and J.-W. Pan, *Realization of an error-correcting surface code with superconducting qubits*, Phys. Rev. Lett. **129**, 030501 (2022).

[23] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, *Demonstrating multi-round subsys-tem quantum error correction using matching and maximum likelihood decoders*, Nature Communications **14**, 2852 (2023).

4

[24] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush, D. Bacon, J. C. Bardin, J. Basso, A. Bengtsson, S. Boixo, G. Bortoli, A. Bourassa, J. Bovaird, L. Brill, M. Broughton, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, J. Cogan, R. Collins, P. Conner, W. Courtney, A. L. Crook, B. Curtin, D. M. Debroy, A. Del Toro Barba, S. Demura, A. Dunsworth, D. Eppens, C. Erickson, L. Faoro, E. Farhi, R. Fatemi, L. Flores Burgos, E. Forati, A. G. Fowler, B. Foxen, W. Giang, C. Gidney, D. Gilboa, M. Giustina, A. Grajales Dau, J. A. Gross, S. Habegger, M. C. Hamilton, M. P. Harrigan, S. D. Harrington, O. Higgott, J. Hilton, M. Hoffmann, S. Hong, T. Huang, A. Huff, W. J. Huggins, L. B. Ioffe, S. V. Isakov, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, P. Juhas, D. Kafri, K. Kechedzhi, J. Kelly, T. Khattar, M. Khezri, M. Kieferová, S. Kim, A. Kitaev, P. V. Klimov, A. R. Klots, A. N. Korotkov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, K.-M. Lau, L. Laws, J. Lee, K. Lee, B. J. Lester, A. Lill, W. Liu, A. Locharla, E. Lucero, F. D. Malone, J. Marshall, O. Martin, J. R. McClean, T. McCourt, M. McEwen, A. Megrant, B. Meurer Costa, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, A. Morvan, E. Mount, W. Mruczkiewicz, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, H. Neven, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, M. Y. Niu, T. E. O'Brien, A. Opremcak, J. Platt, A. Petukhov, R. Potter, L. P. Pryadko, C. Quintana, P. Roushan, N. C. Rubin, N. Saei, D. Sank, K. Sankaragomathi, K. J. Satzinger, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, V. Smelyanskiy, W. C. Smith, G. Sterling, D. Strain, M. Szalay, A. Torres, G. Vidal, B. Villalonga, C. Vollgraff Heidweiller, T. White, C. Xing, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, Y. Zhang, N. Zhu, and G. Q. AI, *Suppressing quantum errors by scaling a surface code logical qubit*, Nature **614**, 676–681 (2023).

[25] L. Postler, S. Heuβen, I. Pogorelov, M. Rispler, T. Feldker, M. Meth, C. D. Marciniak, R. Stricker, M. Ringbauer, R. Blatt, P. Schindler, M. Müller, and T. Monz, *Demonstration of fault-tolerant universal quantum gate operations*, Nature **605**, 675–680 (2022).

[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] D. Ristè, J. G. van Leeuwen, H.-S. Ku, K. W. Lehnert, and L. DiCarlo, *Initialization by measurement of a superconducting quantum bit circuit*, Phys. Rev. Lett. **109**, 050507 (2012).

[28] 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**, 1–7 (2019).

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

[30] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code*, Quantum Info. Comput. **15**, 997–1016 (2015).

[31] M. McEwen, D. Bacon, and C. Gidney, *Relaxing Hardware Requirements for Surface Code Circuits using Time-dynamics*, Quantum **7**, 1172 (2023).

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

[33] 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.

[34] 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).

[35] K. C. Miao, M. McEwen, J. Atalaya, D. Kafri, L. P. Pryadko, A. Bengtsson, A. Opremcak, K. J. Satzinger, Z. Chen, P. V. Klimov, C. Quintana, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, P. Conner, A. L. Crook, B. Curtin, D. M. Debroy, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, A. G. Fowler, B. Foxen, G. Garcia, W. Giang, C. Gidney, M. Giustina, R. Gosula, A. G. Dau, J. A. Gross, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, J. Kelly, S. Kim, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J. Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S. Martin, A. Megrant, X. Mi, S. Montazeri, A. Morvan, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Rocque, P. Roushan, K. Sankaragomathi, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, N. Shutty, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, T. White, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, N. Zhu, N. Zobrist, H. Neven, V. Smelyanskiy, A. Petukhov, A. N. Korotkov, D. Sank, and Y. Chen, *Overcoming leakage in quantum error correction*, Nature Physics **19**, 1780–1786 (2023).

[36] M. Grassl, T. Beth, and T. Pellizzari, *Codes for the quantum erasure channel*, Phys. Rev. A **56**, 33–38 (1997).

[37] C. H. Bennett, D. P. DiVincenzo, and J. A. Smolin, *Capacities of quantum erasure channels*, Phys. Rev. Lett. **78**, 3217–3220 (1997).

[38] T. M. Stace, S. D. Barrett, and A. C. Doherty, *Thresholds for topological codes in the presence of loss*, Phys. Rev. Lett. **102**, 200501 (2009).

[39] S. D. Barrett and T. M. Stace, *Fault tolerant quantum computation with very high threshold for loss errors*, Phys. Rev. Lett. **105**, 200502 (2010).

[40] A. Kubica, A. Haim, Y. Vaknin, F. Brandão, and A. Retzker, *Erasure qubits: overcoming the T_1 limit in superconducting circuits*, 2022, arXiv:2208 . 05461 [quant-ph].

[41] Y. Wu, S. Kolkowitz, S. Puri, and J. D. Thompson, *Erasure conversion for fault-tolerant quantum computing in alkaline earth rydberg atom arrays*, Nature Communications **13**, 4657 (2022).

[42] 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).

[43] 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).

[44] 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. Applied **8**, 034021 (2017).

[45] 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).

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

[47] 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).

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

4

5

MICROWAVE-ACTIVATED GATES BETWEEN A FLUXONIUM AND A TRANSMON QUBIT

We propose and analyze two types of microwave-activated gates between a fluxonium and a transmon qubit, namely a cross-resonance (CR) and a CPHASE gate. The large frequency difference between a transmon and a fluxonium makes the realization of a two-qubit gate challenging. For a medium-frequency fluxonium qubit, the transmon-fluxonium system allows for a cross-resonance effect mediated by the higher levels of the fluxonium over a wide range of transmon frequencies. This allows one to realize the cross-resonance gate by driving the fluxonium at the transmon frequency, mitigating typical problems of the cross-resonance gate in transmon-transmon chips related to frequency targeting and residual ZZ coupling. However, when the fundamental frequency of the fluxonium enters the low-frequency regime below $100\,\mathrm{MHz}$, the cross-resonance effect decreases leading to long gate times. For this range of parameters, a fast microwave CPHASE gate can be implemented using the higher levels of the fluxonium. In both cases, we perform numerical simulations of the gate showing that a gate fidelity above 99% can be obtained with gate times between 100 and $300\,\mathrm{ns}$. Next to a detailed gate analysis, we perform a study of chip yield for a surface code lattice of fluxonia and transmons interacting via the proposed cross-resonance gate. We find a much better yield as compared to a transmon-only architecture with the cross-resonance gate as native two-qubit gate.

5.1. Introduction

The transmon qubit [2, 3] is the most successful superconducting qubit to date, with superconducting chips of around a hundred qubits currently being realized [4–6]. The success of the transmon is due to its resilience to charge noise, the relative simplicity of the circuit and its fabrication, the straightforward control and readout using microwave pulses, and the possibility to couple transmons either via direct capacitances [7] or via bus resonators [8, 9]. Coherence times between 10 and $100\,\mu s$ are routinely reported in two-dimensional transmon chips [6, 10, 11], and even longer T_1 times have been obtained by using different superconducting materials [12, 13]. High-fidelity two-qubit gates have been successfully demonstrated for transmon architectures using several different schemes that rely on either flux pulses [9, 14–18], microwave drives [19–23] or tunable couplers [24–31].

Despite its success, the fact that transmons are essentially slightly anharmonic oscillators is a limiting factor in transmon architectures. Apart from the problem of leakage out of the computational subspace [32], the small anharmonicity of the transmon implies that the transmons must be separated in frequency by, at most, their anharmonicity to enable fast entangling gates. As observed in Ref. [33], this is intuitively due to the fact that when the transmons are far away from each other in frequency, they behave as uncoupled harmonic oscillators. Notice, however, that in this frequency range, the unwanted, spurious ZZ coupling is relatively large, and this limits the performances of the gates [19, 33, 34]. These problems affect fixed, non-tunable coupling architectures such as those based on the cross-resonance (CR) gate [19, 33, 35], giving rise to the problem of frequency collisions [36]. The consequence is a low chip yield when qubit connectivity is as required for the surface code [37], prompting research into optimizing the choice of qubit frequencies [38], improving the accuracy with which qubit frequencies are targeted via laser annealing [36, 39, 40] or pursuing alternative heavy-hexagonal codes that require a lower qubit connectivity [36, 41]. Another issue in transmon chips is the problem of ZZ crosstalk [42] for which some solutions have been discussed [20, 22, 23, 31, 43–45].

The fluxonium qubit [46, 47] is a suitable candidate to go beyond the limitations of a transmon-only architecture. The circuit of the fluxonium is similar to that of the transmon in being composed of a capacitance and a Josephson junction in parallel, but it also features an additional large, shunting inductance. The fluxonium is operated in the regime where the characteristic impedance of the parallel LC-circuit $Z_{LC} = \sqrt{L/C}$ is larger than, say, a few kΩs. To achieve this regime, the high impedance can be realized effectively as an array of hundreds of Josephson junctions [47, 48] or using a material such as granular aluminum [49] or niobium-titanium-nitrate [50, 51]. The inductive shunt provides intrinsic protection against charge noise without the need for a large capacitance, as in transmon qubits. Crucially, this breaks the tradeoff between anharmonicity and charge noise sensitivity that limits the transmon qubit. The large inductance also suppresses the sensitivity to flux noise in the loop formed with the Josephson junction (see Fig. 5.1a), and moreover, the fluxonium is usually operated in the double well configuration, where the qubit frequency is first-order insensitive to flux noise. In this configuration, the fluxonium qubit shows large coherence times [52], which have surpassed the millisecond barrier in 3D devices [53].

The enhanced protection of the fluxonium comes at the price of requiring a more involved scheme for the manipulation of its quantum state. The low fundamental frequency of the fluxonium (below 1 GHz) complicates the execution of single-qubit gates due to the lesser accuracy of the rotating wave approximation compared to the transmon case [54]. In addition, the reduced matrix elements of charge and flux operators that control the strength of the coupling between the computational levels need to be compensated using higher drive power. At very low frequencies of around 10 MHz, new schemes have to be devised for state preparation (reset) and single-qubit gates [55]. On the other hand, the measurement of a fluxonium qubit can have advantages as compared to a transmon qubit. The off-resonant fluxonium-resonator coupling can give rise to relatively large dispersive shifts [54, 56], which enable fast measurement. For granular aluminum-based fluxonium qubits, highly accurate quantum measurements using strong drive power – populating the read-out cavity with a large number of photons – have been reported [57, 58].

Despite this increased complexity, the higher coherence times reached by the fluxonium still pay off in terms of single-qubit gate fidelity [53]. Recently, several two-qubit gate schemes between fluxonia have been proposed and experimentally realized in two-qubit chips [59–65]. A whole architecture for fluxonium qubits has been analyzed in Ref. [54] with two-qubit gates implemented using either the CR gate or the CPHASE gate induced by the differential AC-Stark shift effect [20].

In most superconducting qubit research, the focus is on coupling 'same-type' qubits, i.e., coupling two or more transmon qubits or, alternatively, coupling fluxonia. Exceptions where different types of qubits are coupled are, for example, Refs. [44, 66, 67]. In this chapter, we consider the idea of using chips with heterogeneous qubits. In particular, we analyze how to realize microwave-activated two-qubit gates between capacitively coupled transmons and fluxonia. For fluxonia with medium frequencies, between 0.25-1.0 GHz, we show that the CR gate activated by driving the fluxonium at the transmon frequency is an ideal two-qubit gate candidate. In order to achieve similar gate times, the coupling capacitance needs to be large as compared to the transmon-transmon case [33], but smaller than the fluxonium-fluxonium case [65] due to the transmon being the better antenna. A purely capacitive coupling is easier to engineer than the inductive (combined with a small capacitive) coupling proposed in [54].

By means of a Schrieffer-Wolff analysis [68], we show that the CR effect is mainly mediated by the higher levels of the fluxonium and that it stays large over a wide range of transmon frequencies. Importantly, the frequency of the $|1\rangle - |2\rangle$ and $|0\rangle - |3\rangle$ transitions of the fluxonium should be designed to be relatively far away from the transmon frequency in order to limit residual, static ZZ interactions and leakage in the fluxonium during the gate operation. By means of numerical simulations, which include noise, we show that the CR gate can be realized with leakage below 10^{-4} and gate fidelity above 99% with gate times between 100 and 200 ns. On the other hand, we show that when the fluxonium frequency decreases to around 10 MHz, the CR effect vanishes. In this case, we find that a possible way of implementing a CPHASE gate is to drive to the higher levels of the fluxonium, similar to Ref. [60]. The entangling power of the gate is then due to the coupling-induced hybridization between the bare $|13\rangle_0$ and $|04\rangle_0$ levels of the transmon-fluxonium system, where the subscript denotes that these are the bare states. We argue

that despite the drawback of using the higher levels of the fluxonium, which have coherence times comparable to that of the transmon, the CPHASE gate can be implemented in 100 to 200 ns with arbitrary conditional phases and fidelities above 99%.

This chapter is organized as follows: In Sec. 5.2, we introduce the transmon-fluxonium system. Sec. 5.3 presents the CR gate between a transmon and medium-frequency fluxonium. We provide a comparison with the transmon-transmon CR gate, highlighting the advantage of the transmon-fluxonium case with respect to the frequency crowding problem. We substantiate our understanding by explicit numerical simulations. In Sec. 5.4, we study a low-frequency fluxonium coupled to a transmon and propose a CPHASE gate similar to the one implemented in Ref. [60] between two fluxonia. Also, in this case, we perform a numerical simulation of the gate, showing that, by changing the pulse parameters, CPHASE gates with arbitrary conditional phases can be implemented. Sec. 5.5 presents two possible transmon-fluxonium surface-code-like architectures based on either the CR or the CPHASE gate. We also perform a yield fabrication analysis for the CR gate architecture, similar to those in Ref. [36, 38] for the transmon-transmon case and in Ref. [54] for the fluxonium-fluxonium case. We conclude in Sec. 5.6.

5.2. The Transmon-Fluxonium system

(a)

(b)
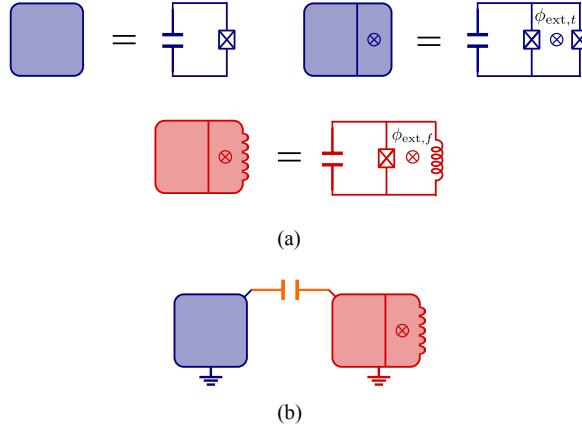
Figure 5.1: (a) Symbolic representation of a fixed-frequency transmon (top left), a flux-tunable transmon (top right) and a fluxonium (bottom). (b) Capacitively coupled (fixed-frequency) transmon and fluxonium qubits.

The basic circuit of the coupled transmon-fluxonium system is shown in Fig. 5.1b. Following standard circuit quantization [69, 70] and directly approximating the transmon as a Duffing oscillator, the Hamiltonian can be written as

$$
\begin{aligned}
H = {} & \hbar \omega_t b^\dagger b + \hbar \frac{\delta_t}{2} b^\dagger b^\dagger b b \\
& + 4 E_{C,f} q_f^2 + \frac{1}{2} E_{L,f} \phi_f^2 - E_{J,f} \cos\left(\phi_f - \phi_{\text{ext},f}\right) \\
& + J_C q_t q_f,
\end{aligned}
\tag{5.1}
$$

with $E_{C,f}, E_{L,f}, E_{J,f}$ the fluxonium charging, inductive and Josephson energy respectively, $\omega_t/2\pi$ the fundamental transmon frequency and $\delta_t/2\pi < 0$ its anharmonicity. The transmon operators b and $b^\dagger$ satisfy commutation relations $[b, b^\dagger] = I$, while the fluxonium (dimensionless) reduced charge q_f and reduced flux ϕ_f operators satisfy $[\phi_f, q_f] = iI$, with I the identity. The transmon charge operator q_t can be expressed in terms of b and $b^\dagger$ as

$$q_t = i \underbrace{\left(\frac{E_{J,t}}{32\hbar|\delta_t|}\right)^{1/4}}_{q_{\text{zpf}}} (b^\dagger - b), \tag{5.2}$$

where the transmon Josephson energy $E_{J,t}$ is related to the qubit energy and anharmonicity by

$$E_{J,t} = \frac{\hbar(\omega_t - \delta_t)^2}{8|\delta_t|}. \tag{5.3}$$

The coefficient q_{zpf} represents the charge zero-point fluctuations of the transmon. In the usual Duffing approximation, the charging energy of the transmon is simply $E_{C,t} = \hbar|\delta_t|$.

Parameter set	Fluxonium						Transmon		
	$\frac{E_{C,f}}{h}$ (GHz)	$\frac{E_{L,f}}{h}$ (GHz)	$\frac{E_{J,f}}{h}$ (GHz)	$\frac{\omega_{f,01}}{2\pi}$ (MHz)	$T_1^{1\to0}$ (μs)	$T_1^{3\to0}$ (μs)	$\frac{\omega_t}{2\pi}$ (GHz)	T_1 (μs)	$\frac{J_C}{h}$ (MHz)
CR	1.0	1.0	4.0	582	126	20	$\in [4.2, 5.8]$	≈ 130	20
CPHASE	1.0	0.5	8.0	30	3700	7	4.37	130	30

Table 5.1: Parameter sets used in the chapter. The transmon always has anharmonicity $\delta_t/2\pi = -300\,\text{MHz}$. The relaxation times correspond to dielectric losses as described in Sec. 5.7.4. The dielectric loss tangent for the transmon is taken to be $\tan\delta_{\text{diel},t} = 3 \times 10^{-7}$ and assumed to be frequency independent. For the fluxonium, similar as in Ref. [52], we take a frequency-dependent dielectric loss tangent $\tan\delta_{\text{diel},f}(\omega) = 3.5 \times 10^{-6}(\omega/\omega_{\text{ref}})^{0.15}$ with $\omega_{\text{ref}}/2\pi = 6.0\,\text{GHz}$. This is needed to take into account the various frequencies that are present in the fluxonium. The temperature of the environment is always assumed to be $T = 20\,\text{mK}$. Additional relaxation and excitation times for other relevant fluxonium transitions are reported in Table 5.3 in Sec. 5.7.4.

In what follows we will assume the fluxonium to be biased with a reduced external flux $\phi_{\text{ext},f} = \pi$ so that it is operated in the double-well potential configuration. We will denote by $|kl\rangle_0$ the bare, uncoupled energy levels of the two qubits, where the first label k identifies the transmon level, while the second label l the fluxonium level. The symbol $|kl\rangle$ denotes the dressed, coupled energy level obtained by adiabatic continuation of the bare level $|kl\rangle_0$ when J_C goes from 0 to a nonzero value. The computational basis is defined as the dressed basis and the projector onto the computational subspace equals $P_c = \sum_{k,l=0}^{1} |kl\rangle\langle kl|$. The projector onto the leakage subspace is then $P_l = I - P_c$. We define $\omega_{f,kl}$ as the transition frequency between the bare fluxonium levels k and l (not to be confused with ω_{kl} in, say, Eq. (5.15) which denote the dressed energy levels of the transmon-fluxonium system). Also, let

$$q_{f,k>l} = \text{Im}(\langle k| q_f |l\rangle_f) = -i\langle k| q_f |l\rangle_f, \tag{5.4}$$

be the imaginary part of the fluxonium matrix element with respect to the bare levels. Note that $\langle k| q_f |k\rangle_f = 0$. In addition, since at $\phi_{\text{ext},f} = \pi$ the fluxonium Hamiltonian, just like the transmon, has a parity symmetry, we also have $\langle k| q_f |k+2m\rangle_f = 0$ for $m \in \mathbb{N}$.
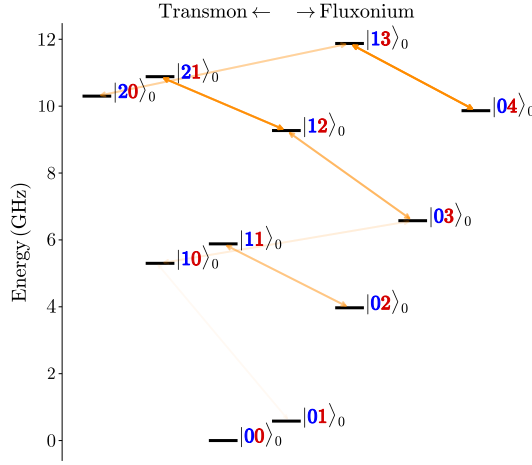
Figure 5.2: An example of a typical energy level diagrams for a capacitively coupled transmon-fluxonium system. The arrows denote the levels that show non-zero matrix elements of the coupling Hamiltonian $J_C q_t q_f$: the darker the color of the arrow the larger the matrix element. The figure corresponds to the parameter set CR in Table 5.1 with $\omega_t/2\pi = 5.3\,\text{GHz}$ and $\phi_{\text{ext},f} = \pi$.

For the parameters listed in Table 5.1, the energy levels $|02\rangle_0$ and $|03\rangle_0$ have frequencies of the same order of magnitude as the $|10\rangle_0$, $|11\rangle_0$ levels as seen in the level diagram in Fig. 5.2. Due to the relatively large matrix elements of the two lowest levels of the fluxonium with the higher levels (see Fig. 5.3), the $q_f q_t$ term in the Hamiltonian directly couples levels $|10\rangle_0 \leftrightarrow |03\rangle_0$ and levels $|11\rangle_0 \leftrightarrow |02\rangle_0$. As we show in Sec. 5.7.1, the coupling to these levels induces a spurious ZZ coupling, but also gives rise to the CR interaction.

The gates considered in this chapter will be activated by a microwave drive on the fluxonium. A drive at carrier frequency ω_d can be modelled by the time-dependent Hamiltonian

$$H_{\text{drive}}(t) = \hbar g(t)\varepsilon_d \cos(\omega_d t + \theta_d) q_f, \tag{5.5}$$

with $0 \le g(t) \le 1$ a dimensionless envelope function and ε_d the maximum drive amplitude that characterizes the drive strength and θ_d the phase of the drive.

5.3. THE CROSS-RESONANCE GATE FOR MEDIUM FREQUENCY FLUXONIA

The CR effect manifests itself when we drive one of the qubits (the control) at the fundamental frequency of the other (the target). The effect arises due to the presence of the coupling term, which enables the drive operator of the control qubit to drive transitions between energy levels of the target qubit. Importantly, the drive on the target qubit depends on the state of the control. Here we take the fluxonium as the control qubit and the
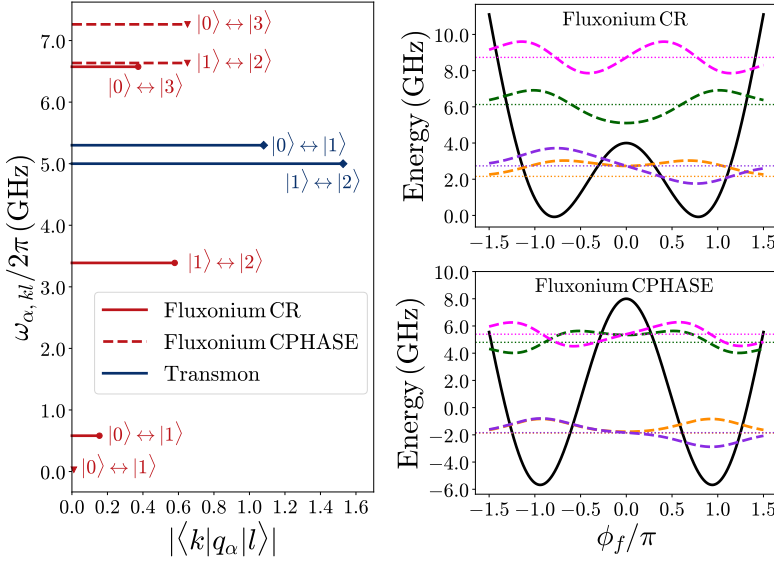
Figure 5.3: (Left) Matrix elements of the fluxonium charge operator ($\alpha = f$) for the fluxonia in Table 5.1 and of the transmon charge operator ($\alpha = t$) with corresponding transition frequency on the y axis. The transmon is taken to have fundamental frequency $\omega_t/2\pi = 5.3\,\mathrm{GHz}$ and anharmonicity $\delta_t/2\pi = -300\,\mathrm{MHz}$. (Right) First 4 eigenfunctions and potential energy for the fluxonia with parameter set CR (top) and CPHASE (bottom) in Table 5.1. The dotted lines represent the energy corresponding to each level.

transmon as the target qubit, so we drive the fluxonium at the fundamental frequency of the transmon. This choice is motivated by the fact that we can have large CR effect leading to a gate time around 100 ns, while the same does not hold if we take the transmon as control and the fluxonium as target. At low drive strengths [33, 71], the CR effect can be simply understood by looking at the matrix elements of the charge operator in the dressed basis. We provide a perturbative analysis of the CR coefficient in Sec. 5.7.1.

Including the envelope function, the fundamental CR Hamiltonian is

$$H_{\mathrm{CR}}(t) = \hbar g(t)\mu_{\mathrm{CR}}X_t Z_f, \tag{5.6}$$

with X_t the transmon Pauli X operator and Z_f the fluxonium Pauli Z operator. As remarked in Sec. 5.7.1, by changing the phase of the drive we can always make μ_{CR} positive. The CR gate with the fluxonium as control and the transmon as target is given by the unitary

$$U_{\mathrm{CR}} = e^{-i\frac{\pi}{4}X_t Z_f}, \tag{5.7}$$

and thus, it is implemented by letting $H_{\mathrm{CR}}(t)$ act for a time t_{gate} such that

$$\frac{\mu_{\mathrm{CR}}}{\hbar} \int_0^{t_{\mathrm{gate}}} dt\, g(t) = \frac{\pi}{4}. \tag{5.8}$$

If $g(t)$ is a constant equal to 1 we obtain the simple formula

$$t_{\mathrm{gate}}^{(\mathrm{id})} = \frac{\hbar\pi}{4\mu_{\mathrm{CR}}}. \tag{5.9}$$
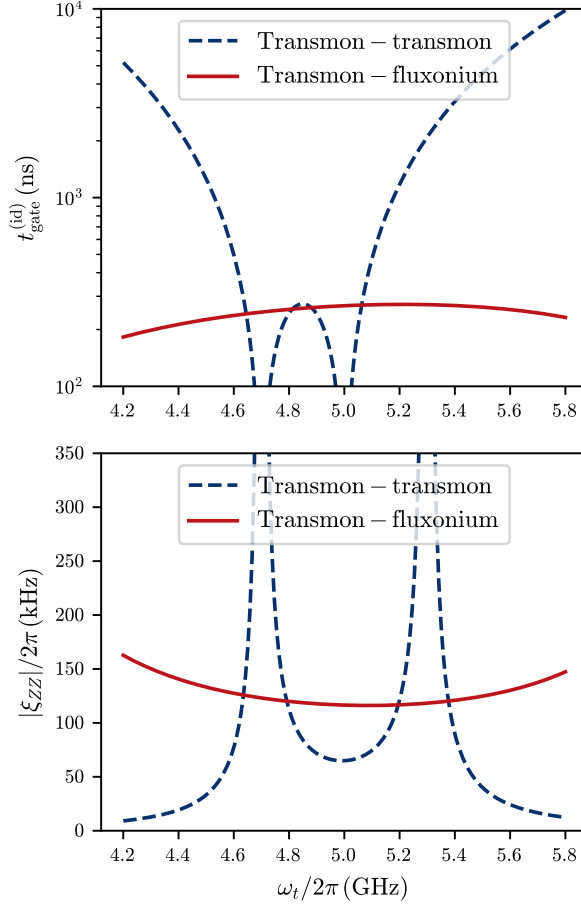
Figure 5.4: Comparison of the gate time (top) and the residual, static ZZ coupling (bottom) for the transmon-transmon and the transmon-fluxonium CR gate. The frequency of the control transmon in the transmon-transmon case is fixed to $\omega_c/2\pi = 5.0\,\text{GHz}$ and both transmons have anharmonicity $\delta_t/2\pi = \delta_c/2\pi = -300\,\text{MHz}$. The coupling between the transmons is set to $J_C/h = 2\,\text{MHz}$. In the transmon-fluxonium case the parameters are taken as in parameter set CR in Table 5.1. In the top figure the drive on the control transmon in the transmon-transmon case is set to $\varepsilon_d/2\pi = 30\,\text{MHz}$, while the drive on the control fluxonium in the transmon-fluxonium case is taken to be $\varepsilon_d/2\pi = 300\,\text{MHz}$. We use Eqs. (5.10) and (5.11) to estimate the ideal gate time in Eq. (5.9). The ZZ coupling ξ_{ZZ} is evaluated via numerical diagonalization using Eq. (5.12). For the transmon-transmon case we see the ZZ coupling blowing up at the resonances $|11\rangle \leftrightarrow |02\rangle$ and $|11\rangle \leftrightarrow |20\rangle$ (not at the resonance $|01\rangle \leftrightarrow |10\rangle$!), while the gate time blows up at the resonances $|01\rangle \leftrightarrow |10\rangle$ and $|11\rangle \leftrightarrow |02\rangle$, leaving a narrow frequency window of opportunity.

Using additional single-qubit gates the CR gate can be turned into a CNOT gate [71].

In Sec. 5.7.1 we derive an approximate formula for the CR coefficient in the transmon-fluxonium case, using a second-order Schrieffer-Wolff transformation, which reads

$$\mu_{\text{CR}} = \frac{J_C q_{\text{zpf}}}{4\hbar} \left[\frac{q_{f,30}^2}{\omega_t - \omega_{f,30}} - \frac{q_{f,21}^2}{\omega_t - \omega_{f,21}} \right] \varepsilon_d. \tag{5.10}$$

This expression has to be compared with the standard expression for the transmon-transmon case with control qubit c and target qubit t [33, 72] :

$$\mu_{\text{CR}}^{(tt)} = -\frac{J_C q_{\text{zpf},t} q_{\text{zpf},c}^2}{\hbar(\omega_c - \omega_t)} \left[\frac{\delta_c}{\delta_c + \omega_c - \omega_t} \right] \frac{\varepsilon_d}{2}, \tag{5.11}$$

where $q_{\text{zpf},t}$ ($q_{\text{zpf},c}$), $\omega_t/2\pi$ ($\omega_c/2\pi$) are the charge zero point fluctuation and the fundamental frequency of the target (control) transmon, respectively, and $\delta_c/2\pi$ the anharmonicity of the control transmon.

We show a comparison between the transmon-transmon and transmon-fluxonium CR gate in Fig. 5.4. Due to the small anharmonicity of the transmons, Eq. (5.11) predicts that the CR coefficient is only large enough to lead to a gate time below 300 ns when the condition $\omega_c + \delta_c \leq \omega_t \leq \omega_c$ is satisfied, see Fig. 5.4 (top). Outside this region the gate time quickly increases to values above 1 μs. Thus, the CR gate for transmon-only systems requires careful frequency engineering and the constraints leads to frequency crowding [36, 38].

In contrast, the transmon-fluxonium CR gate can be activated for a large range of target transmon frequencies and presents more stable CR gate times. While a larger drive strength ε_d is necessary for the transmon-fluxonium case in order to compensate for the smaller charge matrix elements of the fluxonium, we show below, via numerical simulations, that small leakage and high-fidelities can be achieved. In Fig. 5.4 (bottom) we also plot the residual, static ZZ coupling defined as

$$\frac{\xi_{ZZ}}{2\pi} = \frac{E_{11} - E_{10} - E_{01} + E_{00}}{h}, \tag{5.12}$$

with E_{kl} the dressed eigenenergy of level $|kl\rangle$. While the transmon-transmon case achieves a smaller minimum ZZ coupling, the transmon-fluxonium case shows a more stable ξ_{ZZ} without resonant peaks.

In addition, we observe that the average leakage L_1 (see Sec. 5.7.3 for a definition) in the CR gate for the transmon-fluxonium case is generally lower than for the transmon-transmon case for different target transmon frequencies, see Fig. 5.5. In the transmon-fluxonium case, and in the frequency range $4.2 - 5.8$ GHz, we identify two peaks in the leakage at $\omega_t/2\pi \approx 4.41$ GHz and at $\omega_t/2\pi \approx 4.93$ GHz. The former is due to a three-photon transition between the levels $|0\rangle$ and $|5\rangle$ of the fluxonium qubit with frequency $\omega_{f,05}/2\pi = 13.23$ GHz, while the latter is caused by a two-photon transition between levels $|0\rangle$ and $|4\rangle$ of the fluxonium with frequency $\omega_{f,04}/2\pi = 9.86$ GHz. These are frequency collisions that one must avoid in order for the CR gate to achieve high fidelities (see also the discussion in Sec. 5.5). Away from these frequencies the leakage can be as low as 10^{-7} and this happens for a wide target frequency range. In contrast, the transmon-transmon case achieves leakage below 10^{-5} only close to $\omega_t/2\pi = \omega_c/2\pi = 5.0$ GHz, where however a problem of qubit addressability emerges. Moreover, the leakage increases when the two-photon transition between levels $|0\rangle$ and $|2\rangle$ of the control is triggered at $\omega_t/2\pi = 4.85$ GHz and when ω_t matches the $|1\rangle$-$|2\rangle$ transition of the control at $\omega_t/2\pi = 4.7$ GHz.

In Fig. 5.6 we plot gate infidelities in the noiseless and noisy case as a function of the drive strength for the four target transmon frequencies that we consider in the yield analysis in Sec. 5.5. We refer the reader to Sec. 5.7.3 for the definition of gate fidelities

and to Sec. 5.7.4 for details of the dielectric loss error model used for the noisy simulations. We consider a simple piece-wise Gaussian pulse and its echo version, where two piece-wise Gaussian pulses on the fluxonium at frequency $\omega_d/2\pi = \omega_t/2\pi$ and with opposite phase, are interleaved with single-qubit π rotations around the X axis on the fluxonium qubit (as for the transmon-transmon case [73]). We provide more details about the pulses in Sec. 5.7.5. We see that for the chosen target frequency range, the echo pulse generally outperforms the simple Gaussian pulse, especially at low drive strengths. This is because the echo pulse ideally cancels the effect of the unwanted ZZ interaction (besides canceling the X-rotation on the transmon qubit), while the smaller the drive strength the larger the effect of the ZZ interaction is compared to the CR effect [34], which explains the lower fidelities in the simple pulse case. In all cases, we observe that with the echo pulse gate fidelities can be at least 99.5% in the noiseless case and at least 99.3% in the noisy case for all target frequencies. We remark that these were obtained without any pulse optimization and simply by matching the CR condition in Eq. (5.8) and the corresponding one for the echo pulse (see Sec. 5.7.5). Optimal control techniques developed for the CR gate [74] and more detailed techniques to understand the sources of error [75, 76] for the transmon-transmon case can also be applied to the transmon-fluxonium case to achieve higher fidelities.
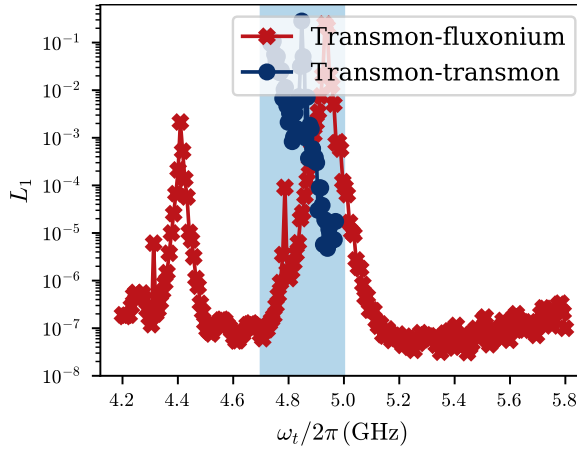
Figure 5.5: Average leakage for a simulated, noiseless CR gate for the transmon-transmon and the transmon-fluxonium case as a function of the target frequency. In both cases parameters are taken as in Fig. 5.4 and we set $\omega_d = \omega_t$. We numerically simulate the gates using a piece-wise Gaussian envelope (see Sec.5.7.5) with rise time $t_{\text{rise}} = 10\,\text{ns}$ and gate time chosen to satisfy the CR condition in Eq. (5.8). We plot the result for the transmon-transmon case only in the relevant, blue-shaded region such that $\omega_c + \delta_c \le \omega_t \le \omega_c$. In the simulations we include 3 bare levels for the transmons and 8 for the fluxonium.

5.4. THE CPHASE GATE FOR LOW FREQUENCY FLUXONIA

If we keep the fluxonium charging energy $E_{C,f}$ fixed and decrease the ratio between $E_{L,f}/E_{J,f}$, the energy barrier between the two lowest minima of the fluxonium potential increases, while the kinetic term stays constant. In this limit, the two lowest eigen-
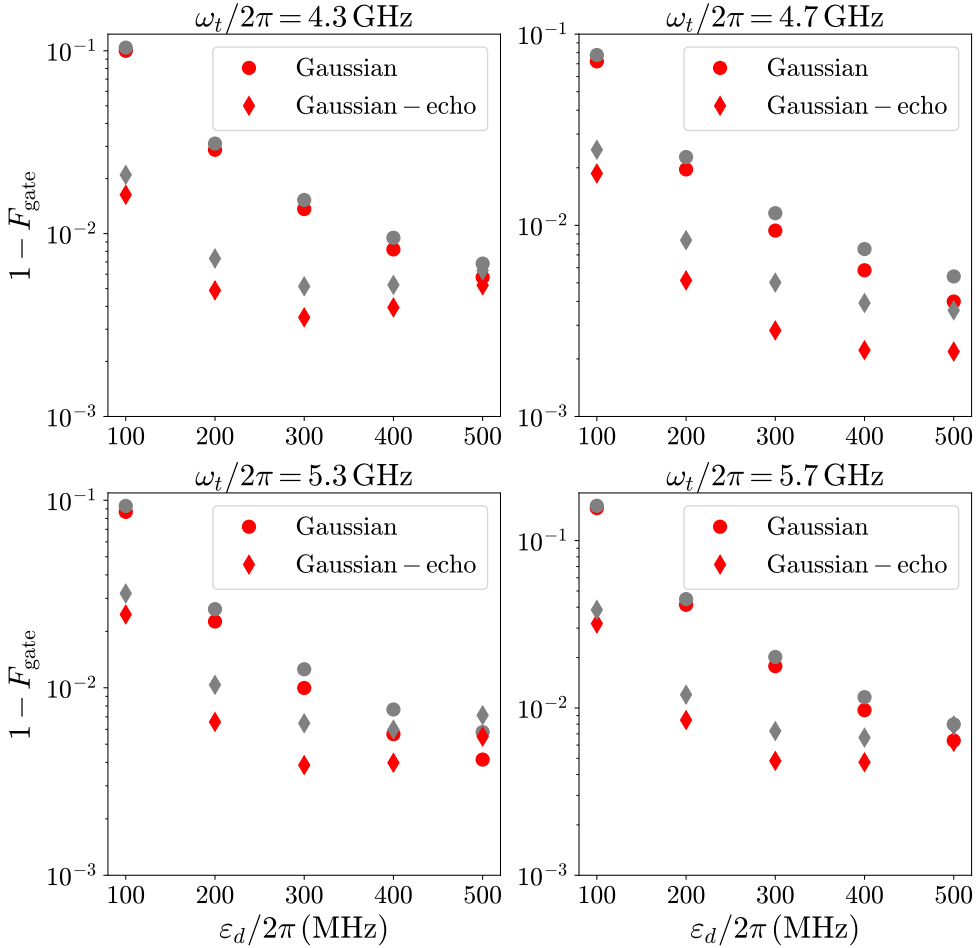
Figure 5.6: Gate infidelity as a function of the drive strength for four different target transmon frequencies in the transmon-fluxonium case. The piece-wise Gaussian envelope are always taken to have rise time $t_{rise} = 10\,$ns. We also plot in gray the infidelities in the presence of dielectric loss. When $\varepsilon_d/2\pi = 300\,$MHz, gate times for the simple Gaussian pulse (no echo) are $t_{gate} = 233, 297, 330, 295\,$ns for $\omega_t/2\pi = 4.3, 4.7, 5.3, 5.7\,$GHz, respectively. In the simulations we include 3 bare levels for the transmons and 6 for the fluxonium.

states of the fluxonium are an even and odd superpositions of two states that are more and more localized in the left and right well of the potential, respectively, see Fig. 5.3(b). As a result, the fundamental frequency of the fluxonium, i.e., the energy splitting between the two lowest levels, decreases, as well as the magnitude of the matrix elements $\langle 0| q_f |1\rangle$, $\langle 0| \phi_f |1\rangle$ of the charge and flux operators between the two lowest levels, see Fig. 5.3(a). This naturally leads to longer coherence times, although the control of the quantum state of the fluxonium becomes more involved [55]. In this scenario, the frequencies of the transmon and of the fluxonium are extremely different, since the fundamental frequency of the fluxonium can be smaller than 100 MHz.

Moreover, the CR scheme that we analyzed in Sec. 5.3 is not applicable in this case because of a vanishing cross-resonance coefficient μ_{CR}. We can understand this by looking at the approximate, analytical formula for μ_{CR} in Eq. (5.10). In the limit of low fluxonium frequency the difference between the frequencies $\omega_{f,30}$ and $\omega_{f,21}$ becomes smaller (see the energy levels in Fig. 5.7 for an example). In order to have a large μ_{CR} the two terms on the right-hand side of Eq. (5.10) need to constructively interfere, i.e., the transmon frequency must be chosen between $\omega_{f,30}$ and $\omega_{f,21}$, leading to a small range of transmon frequencies with a sizable cross-resonance coefficient. Outside this transmon frequency range the μ_{CR} goes to zero also because $|q_{f,21}| \approx |q_{f,30}|$ in the low fluxonium frequency limit.

In this section we show that the scheme proposed in Ref. [60] to realize a CPHASE gate between two fluxonia can be adapted using different fluxonium levels. In Fig. 5.7 we show the energy level diagram for parameter set CPHASE which we will use in this section, see Table 5.1. We see that the system is chosen to have (approximately) a resonance between the transmon frequency and the $|3\rangle - |4\rangle$ transition of the fluxonium, i.e., $\omega_t \approx \omega_{f,34}$. Denoting by E_{kl} the dressed energies of level $|kl\rangle$ and by $E_{kl}^{(0)}$ the bare energy
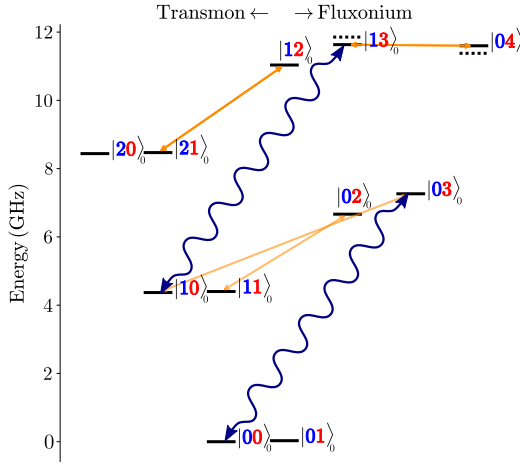
Figure 5.7: Energy level diagram for a low frequency fluxonium coupled to a transmon corresponding to parameter set CPHASE in Table 5.1. The blue wavy lines identify the levels that are driven. The dashed lines above the level $|13\rangle_0$ and below the level $|04\rangle_0$ represent the dressed levels $|13\rangle$ and $|04\rangle$, respectively.

of level $|kl\rangle_0$, we will thus have $E_{13} - E_{04} \neq E_{13}^{(0)} - E_{04}^{(0)}$ due to the coupling term. This energy shift in turn gives rise to a difference between the frequencies associated with the $|00\rangle - |03\rangle$ transition and with the $|10\rangle - |13\rangle$ transition, expressed by a parameter Δ:

$$\frac{\Delta}{2\pi} = \frac{(E_{13} - E_{10}) - (E_{03} - E_{00})}{h} \neq 0. \tag{5.13}$$

For parameter set CPHASE in Table 5.1 we have $\Delta/2\pi = 14.0\,\text{MHz}$. This effect can be used

to implement a microwave-activated CZ gate and more generally a CPHASE gate with an arbitrary phase ϕ given by the general unitary

$$\text{CPHASE}(\phi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{i\phi} \end{pmatrix}. \tag{5.14}$$

As schematically shown in Fig. 5.7, the idea is to apply a drive on the fluxonium qubit with a frequency approximately given by the $|0\rangle - |3\rangle$ transition frequency. More precisely, the drive amplitude and the drive frequency is chosen so that the generalized Rabi frequencies of the $|00\rangle - |03\rangle$ and $|10\rangle - |13\rangle$ transitions are matched to a certain value Ω, which is expressed in the condition

$$\sqrt{\varepsilon_d^2 q_{f,10-13}^2 + (\omega_{13} - \omega_{10} - \omega_d)^2} = \sqrt{\varepsilon_d^2 q_{f,00-03}^2 + (\omega_{13} - \omega_{10} - \omega_d - \Delta)^2} = \Omega, \tag{5.15}$$

where $q_{f,10-13} = |\langle 10| q_f |13\rangle|$ and $q_{f,00-03} = |\langle 00| q_f |03\rangle|$.

This condition leads to drive frequencies that are between the frequencies of the $|00\rangle - |03\rangle$ and $|10\rangle - |13\rangle$ transition, which for parameter set CPHASE in Table Tab. 5.1 is $\omega_d/2\pi \approx 7.26\,\text{GHz}$. Also, the typical drive strength is $\varepsilon_d/2\pi \approx 10\,\text{MHz}$.

Eq. (5.15) guarantees that, assuming $g(t) = 1$, after a time $t_{\text{gate}} = 2\pi/\Omega$ both transitions give rise to a Rabi oscillation which ideally induces no leakage. The nonzero Δ gives rise to a differential phase $\phi \approx \pi\Delta/\Omega$ which is acquired by the state $|11\rangle$, see the detailed analysis of the gate in Sec. 5.7.2. By means of single-qubit Z rotations the implemented gate can be turned into the CPHASE gate in Eq. (5.14).

Crucial for the implementation of the gate is that the system undergoes a Rabi oscillations for both the involved transitions, inducing little leakage. We investigate this property in Fig. 5.8, where the average leakage L_1 (see its definition in Sec. 5.7.3) is plotted as a function of gate time for different target conditional phases. We see that in the coherent case there is a very sharp minimum where the leakage is minimized. Also, the minimum value of the leakage decreases for larger conditional phases, which have larger optimal gate time since $t_{\text{gate}} \approx 2\phi/\Delta$.

However, comparing Figs. 5.5 and 5.8, we notice that the leakage is much higher than that achievable with the CR gate discussed in Sec. 5.3 due to the fact that the CPHASE gate is directly populating non-computational states during the gate. While lower leakage could be potentially achieved with optimized pulses, leakage is definitely more pronounced for this CPHASE gate as compared to the CR gate, assuming the CR gate is operated away from frequency collisions. In addition, in Fig. 5.8 we observe (unsurprisingly) that the presence of noise increases the leakage compared to the noiseless case, and makes the leakage-minima less sharp. Also, the use of higher levels of the fluxonium inevitably exposes the system to the shorter decay times of these levels. Despite these facts, we obtain optimal gate fidelities in the noise free case of $\{99.76\%, 99.87\%, 99.93\%, 99.96\%\}$ and in the noisy case of $\{99.36\%, 99.44\%, 99.51\%, 99.61\%\}$ for the conditional phases $\phi = \{\pi, 5\pi/4, 2\pi/2, 7\pi/4\}$, respectively. The noise free gate fidelity is limited by leakage. In particular, in the noiseless case and when the condition Eq. (5.15) is matched, the state
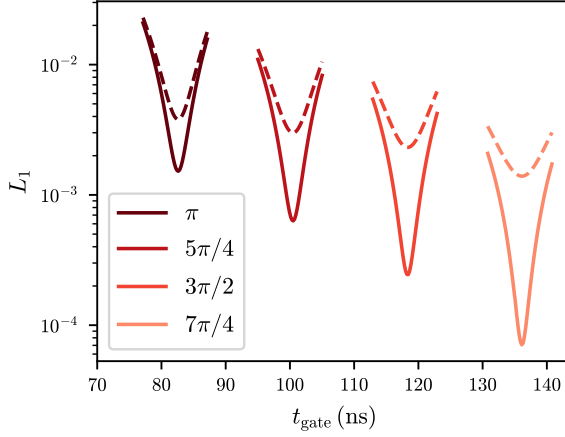
Figure 5.8: Leakage as a function of gate time for different target conditional phases. We use a piece-wise Gaussian envelope $g(t)$ with rise time $t_{\text{rise}} = 10\,\text{ns}$. Drive frequencies and drive strengths are chosen to satisfy Eq. (5.15) with $\Omega = \Delta/(\pi\phi)$ which depends on the target conditional phase ϕ. Minimum leakage (approximately) corresponds to the condition $\int_0^{t_{\text{gate}}} dt\, g(t) = 2\phi/\Delta$. Solid lines represent the coherent results, while the dashed lines represent the noisy ones in the presence of dielectric losses. In the simulations we include 3 bare levels for the transmons and 5 for the fluxonium.

with the highest probability of leakage is $|04\rangle$. This state accounts for approximately 95% of the leakage. Due to the hybridization of the bare levels $|13\rangle_0$ and $|04\rangle_0$, the dressed state $|04\rangle$ also acquires a non-zero fluxonium charge matrix element, and thus the drive can stimulate the $|10\rangle - |04\rangle$ transition, although it is more off-resonant than the $|10\rangle - |13\rangle$ transition. This also explains the decrease of leakage with the increase of the target conditional phase in Fig. 5.8. In fact, the larger the target conditional phase, the larger the ideal gate time and the smaller the drive amplitude, which causes the unwanted $|04\rangle$ to be less populated. The remaining 5% of the leakage is due to the imperfect cancellation of the $|13\rangle$ population. We believe that these two main sources of leakage can be both reduced by further pulse optimization. Finally, we remark that in the noisy case we see an increase in the population of $|13\rangle$ as well as of $|03\rangle$. The latter seems due to a direct relaxation from the $|04\rangle$ state (see Table 5.3 in Sec. 5.7.4 for the relaxation time).

When the drive is turned off, the static, residual ZZ coupling is small and it is evaluated to be $40\,\text{kHz}$ for parameter set CPHASE in Table 5.1 used in this section. As can be seen from Fig. 5.7, the bare levels $|11\rangle_0$ and $|02\rangle_0$ have a non-zero matrix element induced by the capacitive coupling. However, these levels are far detuned in frequency, by more than $2\,\text{GHz}$, and as a result have small level hybridization. The same holds for the pair of bare levels $|10\rangle_0$, $|03\rangle_0$. As shown in Sec. 5.7.1, these transitions are the main cause of ZZ coupling in this system, which is suppressed given the large detuning between the levels.

5.5. ARCHITECTURES BASED ON FLUXONIUM AND TRANSMON

In this section we provide some architectural considerations for multi-qubit transmon-fluxonium chips based on either the CR gate or the CPHASE gate analyzed in the previous sections. For concreteness, we focus on a surface code architecture, where each qubit is directly coupled to at most four neighbors [10, 11, 37]. In this case, either the transmons or the fluxonia play the role of the data or ancilla qubit in the surface code. For both the CR and CPHASE gate, we drive the fluxonium qubit and since up to four transmons couple to the same fluxonium qubit we need to examine how to avoid operation crosstalk. We then explore the expected yield for the architecture based on the CR gate, for which we expect frequency collisions to impact the fidelity of operations.

5.5.1. FREQUENCY ALLOCATION AND OPERATIONS

For the CR gate we can make use of the wide range of transmon frequencies over which the gate can be implemented. A possible frequency setup is shown in Fig. 5.9. In this case the architecture is fully microwave, without the need of frequency tunable transmons or tunable couplers, thus, flux control is only needed for static biasing of the fluxonia.

The fixed-frequency transmons are well-separated in frequency by at least 400 MHz. With these parameters the gate times, assuming $J_C = 20$ MHz and $\varepsilon_d/2\pi = 300$ MHz, would be roughly between 200 and 350 ns and the residual ZZ coupling between 100 and 150 kHz. In this range of frequencies the ZZ coupling is quite stable and does not have any sharp peaks (see Fig. 5.4).

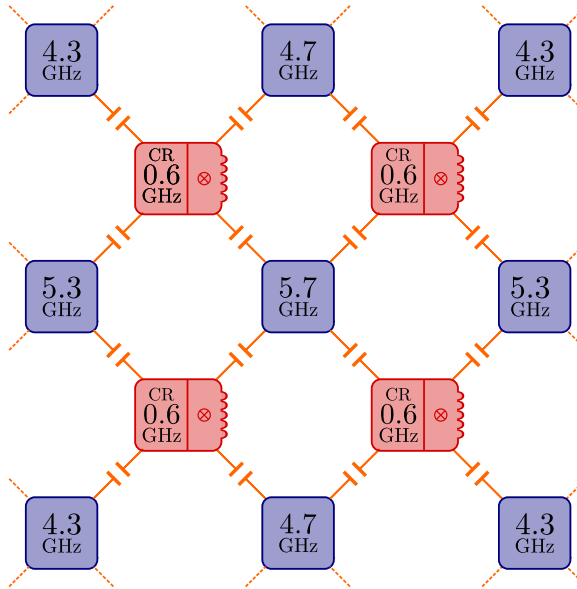Figure 5.9: Surface code archtitecture for fixed frequency transmons (blue) and medium frequency fluxonia (red) based on the CR gate. We choose transmons at four different fundamental frequencies $\omega_t/2\pi = \{4.3, 4.7, 5.3, 5.7\}$ GHz and anharmonicity $\delta_t/2\pi = -0.3$ GHz, while all the fluxonia have target parameters as in parameter set CR in Table 5.1 with fundamental frequency $\omega_{f,01}/2\pi \approx 0.6$ GHz.

In Fig. 5.10 we show a fundamental unit cell for a transmon-fluxonium chip based on the CPHASE gate described in Sec. 5.4. In this case we require frequency-tunable transmons, since we need the ability to selectively tune the transmons to a frequency that matches approximately that of the $|3\rangle - |4\rangle$ transition of the fluxonium. This is an inevitable consequence of the fact that the gate relies on a single resonance. The transmons can all be placed approximately at the same frequency taken to be 4.8 GHz in the example, while the fluxonium has target parameters equal to parameter set CPHASE in Table 5.1. Thus, in order to activate the gate, we first apply a static flux to the desired transmon, in order to bring its frequency close to $\omega_{f,34}/2\pi = 4.3$ GHz. Then we apply a microwave tone to the fluxonium as detailed in Sec. 5.4. Typical gate times depend on the conditional phase and are estimated between 50 and 150 ns (not including the time to flux-tune the qubit back), with static ZZ of 60 kHz [1]. Thus after the gate is completed, the transmon is brought back to the sweet spot at its normal frequency. In an all flux-onium multi-qubit chip where the CPHASE gate is implemented using the scheme of Ref. [60], one would need to flux the fluxonia away from their flux sweet spots, i.e., the double-well configuration, in order to implement the gate selectively. This could lead to complications because additional fluxonium transitions are activated when the system is moved out of the double-well configuration. In contrast, in our scheme the fluxonia always remain at their flux sweet spots.

We remark that our procedure to selectively activate the CPHASE gate is different than the scheme used for the flux-activated CPHASE gate between two transmons [14] in multi-qubit architectures [37]. There the flux pulse is used to implement the gate, while in addition, a neighbor qubit is parked at a different frequency to avoid a frequency collision during the gate operation. In our case, the qubits that are not involved in the gate are left untouched and the gate is activated by the microwave pulse on the fluxonium. A disadvantage is that moving the transmon away from a flux sweet spot triggers $1/f$ flux noise during the gate operation, which impacts leakage and the gate fidelity. This problem can be mitigated using Net-Zero flux pulses [14]. The fluxonium instead always remains at its sweet spot. Another disadvantage is that while a transmon and a fluxonium are interacting via the CPHASE gate, the other transmons coupled to the fluxonium need to remain at their sweet spots to avoid cross-driving. Therefore, these transmons cannot simultaneously interact with other fluxonia during the activation of the CPHASE gate, limiting the number of gates that can be executed in parallel in this architecture.

5.5.2. Frequency collisions and chip yield study

The transmon-fluxonium architecture based on the CR gate that we propose employs only fixed-frequency qubits. Thus, the impact of frequency crowding and the possible frequency collisions on the chip yield is an important consideration when describing the scalability of the system. In contrast, transmons are flux-tunable in the architecture using the CPHASE gate and we do not expect frequency collisions to be an issue in that case. Therefore we focus on exploring the problem of frequency crowding only in the fixed-frequency architecture. In this section, we discuss the fluctuations in the parameters of both fluxonia and transmons that we expect due to fabrication imprecision.

---

[1] By increasing the coupling, the gate times can be further reduced, at the price of a higher static ZZ interaction.
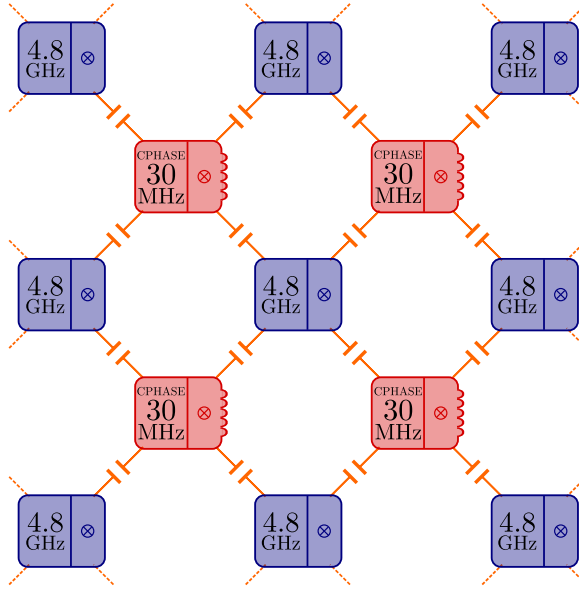
Figure 5.10: Surface code architecture for flux-tunable transmons (blue) and low frequency fluxonia (red) based on the CPHASE gate. All transmons are chosen at the same sweet spot fundamental frequency $\omega_t/2\pi = 4.8\,\mathrm{GHz}$ and anharmonicity $\delta_t/2\pi = -0.3\,\mathrm{GHz}$, while the fluxonia and the J_C parameter are taken as in parameter set CPHASE in Table 5.1.

We outline a set of frequency collisions that we expect to degrade the device performance. To identify the dominant collisions we numerically explore the CR gate performance as a function of the target transmon frequencies ω_t and extract the regions of increased leakage or crosstalk. We associate each region to a specific transition that is driven and that corresponds to a collision condition. We then define a window around each collision, inside of which the increase in crosstalk or leakage is expected to significantly degrade the gate fidelity. We perform a detailed scan in the vicinity of each collision to extract the corresponding bound by requiring that the resulting error is below a given threshold, which we specify below. For collisions involving a spectator qubit we simulate the gate using the full three-qubit system instead. Given these bounds, we simulate the zero-collision yield over a range of variation in the tunnel junction resistance, which determines the qubit frequencies, for surface code lattices up to distance 7 and different drive amplitudes.

The transition frequencies of the transmon are determined by $E_{C,t}$ and $E_{J,t}$, while for the fluxonium the frequencies are a function of $E_{C,f}$, $E_{L,f}$, and $E_{J,f}$. We do not assume any variation in the targeted $E_{C,t}$ or $E_{C,f}$ as the shunting capacitance can be consistently reproduced [36, 54, 77]. On the other hand, it is hard to fabricate the Josephson junction reliably, thus leading to large fluctuations in the critical current I_c, which can be related to the tunnel barrier resistance at room temperature R via the Ambegaokar-Baratoff formula $I_c = \frac{\pi\Delta}{2eR}$, where Δ is the superconducting energy gap. Since R is readily measurable experimentally, we define the variation due to the fabrication of the Josephson junction

in terms of the standard deviation σ_R of the resistance. Given that the Josephson energy can be expressed as $E_J = \frac{\hbar I_c}{2e}$ we expect a variation in $E_{J,f}$ of $\frac{\sigma_{E_{J,f}}}{E_{J,f}} \approx \frac{\sigma_R}{R}$ for the fluxonium qubit. For $E_{L,f}$, we consider the superinductance to be realized via a Josephson-junction array consisting of approximately $N \approx 100$ junctions. Therefore, independent fluctuations in the Josephson energy of each junction would lead to a variation in the inductive energy of $\frac{\sigma_{E_{L,f}}}{E_{L,f}} = \frac{\sigma_R}{\sqrt{N}R}$, using the approximation that the effective inductance due to the array is $L_{\text{eff}} = N L_J$ [47]. When the fluctuations are too large, this simple approximation may break down and one expects spectral shifts due to the coupling of the fluxonium mode with other modes of the array [78].

The transmon frequency ω_t is approximately given by $\omega_t \approx \sqrt{8 E_{C,t} E_{J,t}}/\hbar$ and we thus expect a deviation in the transmon frequency of $\frac{\sigma_{\omega_t}}{\omega_t} = \frac{\sigma_R}{2R}$. Given that $E_{C,t} \approx \hbar|\delta_t|$ we do not expect any variation in δ_t. A resistance variation as low as $\sigma_R/R \approx 2\%$ has been previously reported [77] and a variation of $\sigma_R/R \approx 0.5\%$ has been obtained after the use of laser annealing [36].

Type	Frequency Collision	Bounds		
		$\varepsilon_d/2\pi = 100\,\text{MHz}$	$\varepsilon_d/2\pi = 300\,\text{MHz}$	$\varepsilon_d/2\pi = 500\,\text{MHz}$
1	$\omega_t = \omega_{f,12}$ or $\omega_t = \omega_{f,03}$	$\pm 100\,\text{MHz}$	$\pm 100\,\text{MHz}$	$\pm 100\,\text{MHz}$
2	$\omega_t < \omega_{f,12}$ or $\omega_t > \omega_{f,03}$	–	–	–
3	$2\omega_t = \omega_{f,04}$	$\pm 15\,\text{MHz}$	$\pm 40\,\text{MHz}$	$\pm 60\,\text{MHz}$
4	$2\omega_t = \omega_{f,15}$	$\pm 5\,\text{MHz}$	$\pm 40\,\text{MHz}$	$\pm 50\,\text{MHz}$
5	$2\omega_t = \omega_t + \delta_t + \omega_{f,03}$	$\pm 9\,\text{MHz}$		
6	$3\omega_t = \omega_{f,05}$	–	$\pm 17\,\text{MHz}$	$\pm 35\,\text{MHz}$
7	$\omega_t = \omega_s$	$\pm 5\,\text{MHz}$	$\pm 15\,\text{MHz}$	$\pm 15\,\text{MHz}$
8	$\omega_t = \omega_s + \delta_s$	$\pm 7\,\text{MHz}$	$\pm 20\,\text{MHz}$	$\pm 20\,\text{MHz}$
9	$\omega_t + \omega_s = \omega_{f,04}$	$\pm 10\,\text{MHz}$	$\pm 25\,\text{MHz}$	$\pm 50\,\text{MHz}$

Table 5.2: Frequency collisions and chosen bounds (on $\omega/2\pi$) on forbidden windows around the collision at three different drive amplitudes ε_d for the fixed-frequency transmon-fluxonium architecture employing the CR gate. The bounds on the forbidden windows for each collision are estimated from numerical simulations with the exception of collision type 7, which is taken to be a similar ratio relative to collision type 8 as the ratio between collisions types 5 and 6 reported in [36]. Collision types 1-6 involve a fluxonium acting as the control qubit and a transmon at a frequency ω_t acting as the target qubit. Collision types 7-9 further consider a spectator transmon at a frequency ω_s that is coupled to the fluxonium. The other transmon and fluxonium parameters are given in Ref. [36].

Frequency collisions generally lead to an increase in the infidelity or time-duration of the targeted two-qubit gate due to an increase in leakage and crosstalk. Based on our (noiseless) numerical simulations, we have identified the 9 most likely frequency collisions for our architecture, listed in Tab. 5.2. Each collision involves either only the control fluxonium and target transmon or it further involves a spectator transmon, coupled to the fluxonium. The impact of each collision can be understood as follows: Type 1 collisions lead to a high residual ZZ coupling between the transmon and the fluxonium qubit. This has to be avoided as it gives strong ZZ crosstalk when we do not want to couple the qubits. We define the bound around this collision such that $\xi_{ZZ}/2\pi \leq 1\,\text{MHz}$. Collision type 2 defines conditions leading to a relatively long CR gate time $t_{\text{gate}} \gg 100\,\text{ns}$, as it breaks the condition in Eq. (5.29). Type 3 (resp. type 4) collisions are two-photon

transitions due to the drive at frequency $\omega_d = \omega_t$ which lead to the fluxonium leaking from $|0\rangle$ to $|4\rangle$ (resp. $|1\rangle$ to $|5\rangle$). Collision type 5 represents the drive causing the transmon to leak from $|1\rangle$ to $|2\rangle$ while the fluxonium leaks from $|0\rangle$ to $|3\rangle$. Collision type 6 leads to the fluxonium leaking from $|0\rangle$ to $|5\rangle$ via a three-photon transition due to the drive at frequency ω_t. To bound each of these collisions, we require that the average leakage L_1 to satisfy $L_1 \leq 10^{-3}$. The next collisions consider the impact of driving a CR gate between a fluxonium and a target transmon on a spectator transmon at frequency ω_s. Collisions type 7 leads to the drive of the CR gate addressing the neighboring spectator transmon, while type 8 collisions instead cause the spectator transmon to leak from $|1\rangle$ to $|2\rangle$. Collision type 9 describes a transition involving both the target and spectator transmon and leading to the fluxonium leaking from $|0\rangle$ to $|4\rangle$. To extract the bounds around collision type 7, we prepare the spectator transmon in $|0\rangle$ and the control and target qubits in the state $P_c/2$ and require that the population in $|1\rangle$ of the spectator qubit after the gate to be below 10^{-3}. Finally, to bound collisions type 8 and 9, we require again that the resulting average leakage from the computational states is sufficiently small, that is, $L_1 \leq 10^{-3}$. We note that collision types 4 and 5 are separated by only $\approx 50\,\text{MHz}$ for the targeted fluxonium parameters. A sufficiently strong drive ($\varepsilon_d/2\pi \geq 300\,\text{MHz}$) leads to a (drive-strength dependent) detuning of the qubit frequencies, which can lead to the frequencies of these two collisions to shift closer to each other. In such a case, we instead place a single bound around collision type 4, which also includes collision type 5, and ensures that the resulting leakage $L_1 \leq 10^{-3}$ outside of this window.

To simulate the expected zero-collision yield, we consider a transmon-fluxonium surface code lattice of distance d and take the fluxonia to be the ancilla qubits and the transmons to be the data qubits. We sample a ω_t for each transmon and $E_{L,f}$ and $E_{J,f}$ for each fluxonium drawing from a Gaussian distribution characterized by a standard deviation determined by σ_R (as described above) and centered around the targeted parameter value. We then evaluate the transition frequencies of each fluxonium via numerical diagonalization and check if any collisions have occurred across the lattice. We perform 6000 repetitions of this process for each lattice and drive amplitude. In Fig. 5.11 we show the results for lattices of distance $d = 3, 5, 7$ and for drive amplitudes $\varepsilon_d/2\pi = 100, 300, 500\,\text{MHz}$ (with the corresponding bounds given in Tab. 5.2) as a function of σ_R/R. We observe that for a resistance variation of $\sigma_R/R = 2\%$, we expect all lattices up to $d = 7$ to be producible with a yield $\gtrsim 10\%$ when $\varepsilon_d/2\pi = 100\,\text{MHz}$. When the drive amplitude is increased to $\varepsilon_d/2\pi \geq 300\,\text{MHz}$, the $d = 7$ yield lattice drops to $\gtrsim 1\%$. In the case of a strong drive of $\varepsilon_d/2\pi = 500\,\text{MHz}$, the yield for the $d = 5$ and $d = 7$ lattices drops to be above 1% and 0.1% respectively. If we instead consider the resistance variation of $\sigma_R/R \approx 0.5\%$ achieved following laser annealing, we see that any lattice up to $d = 7$ can be fabricated with a yield $> 99\%$ for each drive amplitude considered here. In Ref. [36], a transmon-transmon architecture utilizing a CR gate of duration $200-400\,\text{ns}$ is explored. At the same resistance variation of $\sigma_R/R \approx 0.5\%$, there is no transmon-transmon surface code lattice achieving a yield $> 10\%$, while for a heavy-hexagon code lattice, only code distances $d = 3, 5$ lead to a yield $> 10\%$. The high yield demonstrates that the frequency crowding issue is greatly mitigated in a transmon-fluxonium architecture by the large detuning between the qubit frequencies. In Sec. 5.7.6, we show the average number of collisions observed for a $d = 3$ lattice at $\varepsilon_d/2\pi = 300\,\text{MHz}$, demonstrat-
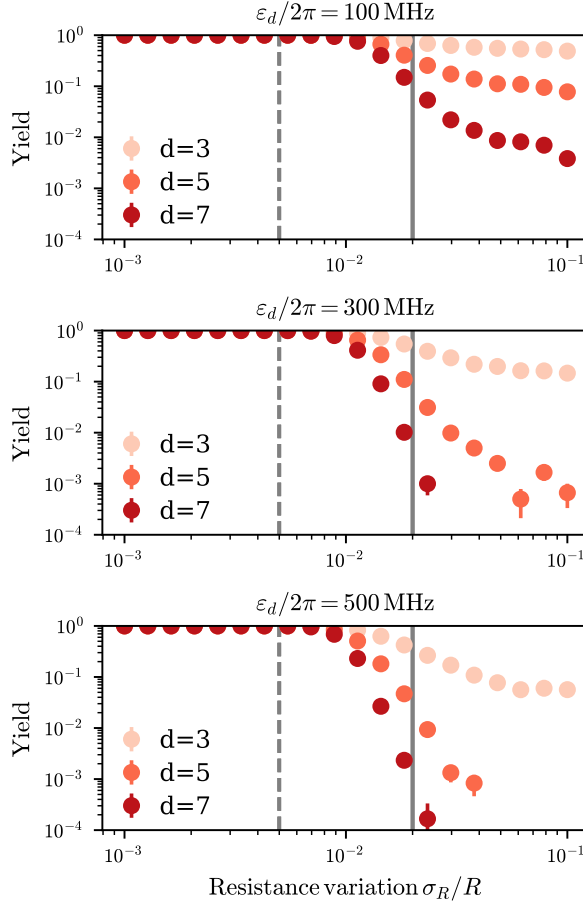
Figure 5.11: Zero-collision yield as a function of the tunnel barrier resistance dispersion for a fixed-frequency transmon-fluxonium surface code lattice of distance $d = 3,5,7$ (light, medium and dark red respectively) and for drive amplitudes $\varepsilon_d/2\pi = 100\,\mathrm{MHz}$ (top), $\varepsilon_d/2\pi = 300\,\mathrm{MHz}$ (middle) and $\varepsilon_d/2\pi = 500\,\mathrm{MHz}$ (bottom). The solid-gray line shows the state-of-the-art resistance variation measured, while the dashed-gray line shows the achieved variation following laser annealing. The yield is extracted over 6000 resamples of the lattice parameters.

ing that the loss in yield is dominated by frequency collisions involving next-nearest-neighboring transmons. We expect a more optimal assignment of the transmon frequencies to further increase the zero-collision yield. Furthermore, a factor of 2 improvement in the resistance variation achieved from fabrication would enable a $\gtrsim 10\%$ yield on transmon-fluxonium lattices up to distance 7, mitigating the need for post-fabrication adjustments.

5.6. Conclusions

In this chapter we have studied two-qubit gates between transmons and fluxonia to be used in a multi-qubit architecture. Despite the typical large fundamental frequency difference between transmons and fluxonia, two-qubit gates are still possible thanks to the direct or indirect use of the higher levels of the fluxonium. We have analyzed two different microwave-activated gates: the CR gate and the CPHASE gate. The CR gate is suited for medium frequency fluxonia and, compared to its transmon-transmon counterpart, it can be implemented over a wider range of transmon frequencies, mitigating the frequency crowding problem. For low frequency fluxonia, the CR effect decreases and therefore we have studied a different scheme that implements a CPHASE gate using the third level of the fluxonium. While this gate is more prone to leakage, one can get arbitrary conditional phases with gate times around 100 to 200 ns and have small residual ZZ coupling. We have also provided some architectural considerations for a surface-code-like architecture where each qubit is coupled to up to four neighbors. In case the architecture is based on the CR gate, it can be fully microwave-activated, while some flux control on the transmons is needed for the CPHASE case. We have shown that the fixed-frequency architecture based on the CR gate between transmons and fluxonia greatly mitigates the problem of frequency crowding. We show that a tunnel barrier resistance variation achieved by laser-annealing enables a yield of near unity for surface codes up to distance 7 and possibly higher: this is a yield which is considerably higher as compared to fixed-frequency transmon-transmon architectures using the CR gate. It would be interesting to also make a multi-qubit chip yield comparison between our transmon-fluxonium architecture and a fluxonium-fluxonium architecture as in [54].

5.7. Supplementary Information

5.7.1. Schrieffer-Wolff and CR gate analysis

In this section we perform a perturbative analysis of the coupled transmon-fluxonium system based on the Schrieffer-Wolff transformation, following closely Ref. [33, 68, 72]. We first execute the analysis without the drive and obtain perturbative formulas for the frequency shifts of the levels as well as for the residual ZZ coupling. In our analysis, we include the first 3 levels of the transmon and the first 4 levels of the fluxonium. Within this subspace the Hamiltonian in Eq. (5.1) reads

$$H = H_0 + V, \tag{5.16}$$

with the Hamiltonians H_0 and V which, in the bare basis, are

$$\frac{H_0}{\hbar} = \omega_t |1\rangle \langle 1|_t + (2\omega_t + \delta_t) |2\rangle \langle 2|_t + \sum_{k=1}^{3} \omega_{f,k} |k\rangle \langle k|_f, \tag{5.17}$$

and

$$V \stackrel{\text{RWA}}{=} J_C q_{\text{zpf}} \left(\sigma_{01}^t + \sqrt{2} \sigma_{12}^t \right) \times \left(q_{f,10} \sigma_{10}^f + q_{f,30} \sigma_{30}^f + q_{f,21} \sigma_{21}^f + q_{f,32} \sigma_{32}^f \right) + \text{h.c.} \tag{5.18}$$

Here $|k\rangle \langle k|_t = |k\rangle \langle k|_0 \otimes I_{0f}$ and $|k\rangle \langle k|_f = I_{0t} \otimes |k\rangle \langle k|_0$, with I_{0t} and I_{0f} the identity on the transmon and fluxonium, respectively. Analogously, $\sigma_{kl}^t = |k\rangle \langle l|_0 \otimes I_{0f}$ and $\sigma_{kl}^f =$

$I_{0t} \otimes |k\rangle \langle l|_0$, while the $q_{f,kl}$ are given in Eq. (5.4) in the main text. Furthermore, the $\omega_{f,k}/2\pi$ are the frequencies associated with the fluxonium levels, i.e., with respect to the ground state of the fluxonium, while $\omega_t/2\pi$ and $\delta_t/2\pi$ the fundamental frequency and the anharmonicity of the transmon, respectively. So in this section, for notational simplicity, $\omega_{f,0}$ is set as 0 (while the explicit dependence is given in expressions in the main text and Eq. (5.29)).

In Eq. (5.18) we performed a rotating wave approximation (RWA) neglecting terms $\sigma_{kl}^t \sigma_{k'l'}^f$ with $k > k', l > l'$, and their Hermitian conjugate.

In order to proceed with the Schrieffer-Wolff analysis let us define the relevant projectors in the bare basis

$$P_0 = \sum_{k,l=0}^{1} |kl\rangle \langle kl|_0, \tag{5.19a}$$

$$Q_0 = I - P_0, \tag{5.19b}$$

and in the dressed basis

$$P = \sum_{k,l=0}^{1} |kl\rangle \langle kl|, \tag{5.20a}$$

$$Q = I - P. \tag{5.20b}$$

The Schrieffer-Wolff transformation is defined as the unitary U that transforms the projectors in the bare basis to those in the dressed basis:

$$U^\dagger P_0 U = P, \quad U^\dagger Q_0 U = Q. \tag{5.21}$$

The unitary U exists and is unique if and only if $\|P - P_0\| < 1$ and in this case it is given by [68]

$$U = \sqrt{(I - 2P_0)(I - 2P)}. \tag{5.22}$$

In addition, U can be written as $U = \exp(S)$ with S being anti-Hermitian, $S = -S^\dagger$, and block-off-diagonal with respect to P_0, i.e., $P_0 S P_0 = (I - P_0)S(I - P_0) = 0$. The goal of the Schrieffer-Wolff transformation is to obtain an effective Hamiltonian H_{eff} that has the same spectrum as PHP, that is, the projection of H onto the subspace associated with P, i.e., the computational subspace in our case. The effective Hamiltonian is given by

$$H_{\text{eff}} = P_0 U H U^\dagger P_0. \tag{5.23}$$

Eq. (5.22) provides a numerical method to obtain the Schrieffer-Wolff unitary and thus also the effective Hamiltonian [79]. However, the standard use of the Schrieffer-Wolff is to find an analytical, perturbative expansion of the effective Hamiltonian. The norm of the coupling operator $\|V\|$ quantifies the strength of the coupling, and thus plays the role of the coupling parameter. Following Ref. [68], the second-order expansion of the effective Hamiltonian is given by

$$H_{\text{eff}}^{(2)} = P_0(H_0 + V)P_0 + \frac{1}{2}P_0[S_1, V_{\text{od}}]P_0, \tag{5.24}$$

where $V_{\text{od}} = P_0 V Q_0 + Q_0 V P_0$ is the off-diagonal part of V. In our case V is fully off-diagonal, i.e. $V_{od} = V$, and the anti-Hermitian operator S_1 reads

$$S_1 = \frac{J_C q_{\text{zpf}} q_{f,30}}{\hbar(\omega_{f,3} - \omega_t)} \sigma_{01}^t \sigma_{30}^f + \frac{J_C q_{\text{zpf}} q_{f,21}}{\hbar[(\omega_{f,2} - \omega_{f,1}) - \omega_t]} \sigma_{01}^t \sigma_{21}^f$$
$$+ \frac{\sqrt{2} J_C q_{\text{zpf}} q_{f,10}}{\hbar[\omega_{f,1} - (\omega_t + \delta_t)]} \sigma_{12}^t \sigma_{10}^f - \text{h.c.}. \tag{5.25}$$

This gives the effective Hamiltonian

$$\frac{H_{\text{eff}}^{(2)}}{\hbar} = \omega_t |1\rangle \langle 1|_t + \omega_{f,1} |1\rangle \langle 1|_f + J_C q_{\text{zpf}} q_{f,10} [\sigma_{01}^t \sigma_{10}^f + \text{h.c.}]$$
$$+ \zeta_{10} |10\rangle \langle 10|_0 + \zeta_{11} |11\rangle \langle 11|_0, \tag{5.26}$$

where we define the coefficients

$$\hbar \zeta_{10} = -\frac{J_C^2 q_{\text{zpf}}^2 q_{f,30}^2}{\omega_{f,3} - \omega_t}, \tag{5.27a}$$

$$\hbar \zeta_{11} = J_C^2 q_{\text{zpf}}^2 \left[\frac{2 q_{f,10}^2}{\omega_{f,1} - (\omega_t + \delta_t)} - \frac{q_{f,21}^2}{(\omega_{f,2} - \omega_{f,1}) - \omega_t} \right], \tag{5.27b}$$

We note that the first contribution in ζ_{11} is due to the $|11\rangle_0 - |20\rangle_0$ transition. It is quite small as $q_{f,10}$ is small (see Fig. 5.3), and the levels are fairly off-resonant (see Fig. 5.2). The second term is dominant and due to the (more resonant) $|11\rangle_0 - |02\rangle_0$ transition with larger $q_{f,21}$ in Fig. 5.2. The sign of the contributions is opposite as $|20\rangle_0$ is higher than $|11\rangle_0$ while $|02\rangle_0$ is lower. The coefficient ζ_{10} is due to the $|10\rangle_0 - |03\rangle_0$ transition.

From Eq. (5.26) we see that we still have the exchange coupling term proportional to J_C that couples levels $|10\rangle_0$ and $|01\rangle_0$. This could be further removed with a second Schrieffer-Wolff transformation that would give two effective Hamiltonians: one for the subspace $\{|00\rangle_0, |01\rangle_0\}$ and one for the subspace $\{|10\rangle_0, |11\rangle_0\}$. However, we remark that the second-order corrections due to this flip-flop term are very small in our typical setup because $|q_{f,10}| \ll 1$ and, in addition, the transmon and fluxonium frequency always differ by at least $3.5\,\text{GHz}$. Thus, we can simply neglect the effect of this term. Within this approximation, the ZZ coupling in second-order Schrieffer-Wolff is given by

$$\hbar \xi_{ZZ}^{(\text{SW})} = \langle 11| H_{\text{eff}}^{(2)} |11\rangle_0 + \langle 00| H_{\text{eff}}^{(2)} |00\rangle_0 - \langle 01| H_{\text{eff}}^{(2)} |01\rangle_0 - \langle 10| H_{\text{eff}}^{(2)} |10\rangle_0$$
$$= \hbar(\zeta_{11} - \zeta_{10}). \tag{5.28}$$

Thus the ZZ coupling is enhanced in strength when the signs of ζ_{11} and ζ_{10} are opposite. Neglecting the first term in Eq. (5.27b), this is achieved when the transmon frequency is chosen as

$$\omega_{f,2} - \omega_{f,1} < \omega_t < \omega_{f,3} - \omega_{f,0}. \tag{5.29}$$

In a multi-qubit architecture where all qubits are capacitively coupled, the SW transformation to the dressed, computational basis, will not only slightly entangle nearest-neighbor capacitively-coupled qubits but also entangle non-nearest neighbors. That is,

iS_1 will be a 2-local many-qubit Hamiltonian with non-commuting 2-local terms each representing the nearest-neighbor qubit couplings. This implies that the computational qubits, on which we also apply single-qubit gates and which we measure, are represented by two-level subspaces which are partially multi-qubit entangled. As a consequence, a drive on one qubit in the bare basis, will be transferred not only to its nearest-neighbor qubits, but also, more weakly, to non-nearest neighbor qubits.

DRIVE AND CR COEFFICIENT

In the presence of a drive on the fluxonium the Hamiltonian gets an additional term given in Eq. (5.5). We can use the previous analysis to get an expression for the CR effect by simply applying the Schrieffer-Wolff transformation to the drive Hamiltonian [33, 72, 80] so we see the effect of the drive in the dressed, computational, basis.

In what follows we assume, for simplicity, that the envelope function is a constant, i.e., $g(t) = 1$, and set $\theta_d = \pi/2$, but we will comment on what happens when we change θ_d.

After obtaining the effective Hamiltonian in Eq. (5.26), we switch to a rotating (or interacting) frame at the drive frequency for both qubits defined by the reference Hamiltonian $H_{\text{ref}}/\hbar = \omega_d(|1\rangle\langle1|_t + |1\rangle\langle1|_f)$, for the purpose of analysis. In general, for a Hamiltonian H, moving to rotating frame set by $U_{\text{ref}} = e^{-iH_{\text{ref}}t/\hbar}$ results in the dynamics being given by a Hamiltonian $\tilde{H}$ given by

$$\tilde{H}(t) = U_{\text{ref}}^\dagger H U_{\text{ref}} + i\left(\frac{d}{dt}U_{\text{ref}}^\dagger\right)U_{\text{ref}} \tag{5.30}$$

$$= e^{iH_{\text{ref}}t/\hbar}He^{-iH_{\text{ref}}t/\hbar} - H_{\text{ref}}. \tag{5.31}$$

We then approximate and calculate

$$\begin{aligned}\frac{\tilde{H}_{\text{drive,eff}}(t)}{\hbar} &= e^{iH_{\text{ref}}t/\hbar}P_0 e^{S_1}\frac{H_{\text{drive}}(t)}{\hbar}e^{-S_1}P_0 e^{-iH_{\text{ref}}/\hbar t}\\ &\approx e^{iH_{\text{ref}}/\hbar t}P_0\frac{(H_{\text{drive}}(t) + [S_1, H_{\text{drive}}])}{\hbar}P_0 e^{-iH_{\text{ref}}/\hbar t}\\ &\overset{\text{RWA}}{\approx} \mu_{X_f}X_f + \mu_{\text{CR}}X_t Z_f + \mu_{X_t}X_t.\end{aligned} \tag{5.32}$$

Here the first term is simply due to $H_{\text{drive}}(t)$ and $[S_1, H_{\text{drive}}]$ gives the other two terms, i.e. the (fluxonium-controlled) rotation on the transmon qubit. Here $X_f (X_t)$ is the Pauli X operator acting on the fluxonium (transmon), and Z_f the Pauli Z acting on the fluxonium.

The coefficients within this approximation read

$$\hbar\mu_{X_f} = \frac{1}{2}q_{f,10}\varepsilon_d \tag{5.33a}$$

and

$$\hbar\mu_{X_t} = \frac{J_C q_{\text{zpf}}}{4}\left[\frac{q_{f,30}^2}{\omega_t - \omega_{f,3}} + \frac{q_{f,21}^2}{\omega_t - (\omega_{f,2} - \omega_{f,1})}\right]\varepsilon_d, \tag{5.33b}$$

while the CR coefficient is

$$\hbar\mu_{CR} = \frac{J_C q_{\text{zpf}}}{4}\left[\frac{q_{f,30}^2}{\omega_t - \omega_{f,3}} - \frac{q_{f,21}^2}{\omega_t - (\omega_{f,2} - \omega_{f,1})}\right]\varepsilon_d. \qquad (5.34)$$

As for the strength of the CR coefficient μ_{CR}, we can observe the following. Similar as for the ZZ coupling, the largest coefficient μ_{CR} is obtained when the two terms in Eq. (5.34) add constructively, i.e. we choose the transmon frequency according to Eq. (5.29). Naturally, the more entangling the Schrieffer-Wolff unitary e^{S_1} to the dressed basis is, the more the drive on the bare transmon qubit becomes transferred to a coupling term and this entangling power of e^{S_1} with S_1 in Eq. (5.25) is proportional to J_C. Additionally, our perturbative formula Eq. (5.34) predicts a linear increase of μ_{CR} with the drive amplitude ε_d. As for the transmon-transmon case, a more refined analysis that includes the drive in the perturbation would predict a saturation of the cross-resonance coefficient with the drive amplitude [33]. By adapting this analysis to our case, we verified that for the parameters used in this chapter the linear approximation Eq. (5.34) is quite accurate and reproduces the "exact" result with error below 5% for all the considered drive amplitudes.

Notice that if μ_{CR} is negative for $\theta_d = \pi/2$, we can always change its sign by taking $\theta_d = 3\pi/2$ instead in Eq. (5.5). Thus, μ_{CR} can always be assumed to be positive. By choosing a different phase θ_d of the drive, say $\theta_d = 0$, one can go through the math behind Eq. (5.32) and observe that Pauli $X_t \to Y_t$ and $X_f \to Y_f$ as one may expect.

In the rotating frame set by H_{ref}, Eq. (5.26) equals

$$\frac{\tilde{H}_{\text{eff}}^{(2)}}{2\hbar} \approx (\omega_t - \omega_d)|1\rangle\langle 1|_t + (\omega_{f,1} - \omega_d)|1\rangle\langle 1|_f + \zeta_{10}|10\rangle\langle 10| + \zeta_{11}|11\rangle\langle 11|, \qquad (5.35)$$

neglecting the flip-flop coupling in the computational subspace. Hence, we see that if ω_d is chosen as ω_t, one drives the fluxonium at the frequency of the transmon qubit, $\tilde{H}_{\text{eff}}^{(2)}$ contains no single-qubit Z_t and the effect of the $\mu_{CR} X_t Z_f$ and $\mu_{X_t} X_t$ terms in Eq. (5.32) is maximal. If instead the drive frequency and the transmon frequency are sufficiently different, the (fluxonium-controlled) X_t-rotation is very small as compared to Z_t, and hence would induce at most some renormalization of the transmon frequency and the ZZ-coupling.

In order to completely understand the dynamics, we can re-evaluate this Hamiltonian in the standard computational rotating frame where the precession of each (dressed) qubit at its eigenfrequency is undone. For the transmon qubit, H_{ref} already selects this frame (as $\omega_d = \omega_t$), but for the fluxonium qubit we can undo the first rotation frame and use the first computational one for the fluxonium by picking a new $H_{\text{ref}}'/\hbar = (\omega_{1,f} - \omega_d)|1\rangle\langle 1|_f$. This ensures that $\tilde{H}_{\text{eff}}^{(2)}$ has no more single-qubit Z_t or Z_f terms, and note it does not affect the relevant $X_t Z_f$ term. At the same time it introduces a time-dependence $e^{\pm 2i(\omega_{1,f} - \omega_d)t}$ in the single-qubit term $\mu_{X_f} X_f$ in Eq. (5.32). Both due to the smallness of $q_{f,10}$ as well as the difference in frequency between the transmon and the fluxonium (Fig. 5.3), the effect of this off-resonant term is thus very small.

In Fig. 5.12 we compare the results obtained with the perturbative Schrieffer-Wolff analysis and the exact numerical values for the ZZ coupling in Eq. (5.28). We see good
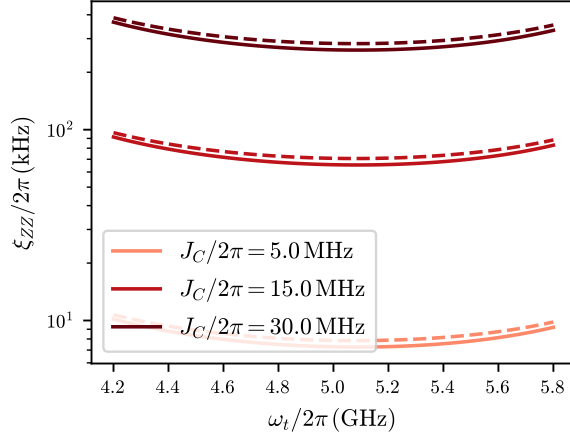
Figure 5.12: Residual ZZ coupling as a function of the transmon frequency at zero drive strength. The solid lines represent the exact numerical value, while the dashed line shows the result obtained from the second-order Schrieffer-Wolff transformation in Eq. (5.28). We use the fluxonium parameters of parameter set CR in Table 5.1.

agreements between the numerics and the results from the perturbative Schrieffer-Wolff analysis. We notice that the ZZ coupling is relatively constant over a wide range of frequencies. This is due to the fact that the transmon frequency is always in between the transition frequencies of the $|1\rangle - |2\rangle$ and the $|0\rangle - |3\rangle$ transition of the fluxonium, as in Eq. (5.29).

We conclude this section by commenting on the role of the coupling parameter J_C on the gate implementation. The unwanted ZZ coupling coefficient depends quadratically on J_C, while the cross-resonance coefficient scales linearly with J_C. This suggests that smaller J_C should decrease the error associated with the coherent ZZ interaction. However, smaller J_C also means longer gate times, which we would like to limit in order to have small errors from decoherence processes. We remark that the transmon-transmon implementation of the cross-resonance gate faces exactly the same trade-off.

5.7.2. CPHASE gate analysis

In this section we provide a simplified analysis of the CPHASE gate, restricting ourselves to the computational subspace plus the higher (dressed) levels $|03\rangle$ and $|13\rangle$ to understand the idea behind the gate. We will work in the dressed computational basis $|kl\rangle$ which can be obtained from the bare basis $|kl\rangle_0$ by a Schrieffer-Wolff transformation which is discussed in Sec. 5.7.1. This transformation gives a ZZ coupling between the qubits (and a very off-resonant flip-flop interaction which we neglect here) which is used for the CPHASE gate.

We assume that the drive $H_{\text{drive}}(t)$ in Eq. (5.5) has $g(t) = 1$ and the phase is chosen as $\theta_d = \pi/2$. We solely focus on the drive enacting the transitions $|00\rangle \leftrightarrow |03\rangle$ and $|10\rangle \leftrightarrow |13\rangle$

where $|kl\rangle$ are dressed energy levels, so we write

$$\frac{H_{\text{drive}}(t)}{\hbar} \approx \frac{\varepsilon_d q_{f,00-03}}{2} (|00\rangle \langle 03| (-e^{i\omega_d t} + e^{-i\omega_d t}) + \text{h.c.})$$
$$+ \frac{\varepsilon_d q_{f,10-13}}{2} (|10\rangle \langle 13| (-e^{i\omega_d t} + e^{-i\omega_d t}) + \text{h.c.}), \quad (5.36)$$

with $q_{f,kl-mn} = |\langle kl| q_f |mn\rangle|$. The rest of the Hamiltonian of the transmon-fluxonium system, restricted to this six-dimensional subspace equals

$$\frac{H|_6}{\hbar} = \omega_{01} P_{01} + \omega_{11} P_{11} + \omega_{00} P_{00} + \omega_{03} P_{03} + \omega_{10} P_{10} + \omega_{13} P_{13},$$

where $P_{ij} = |ij\rangle \langle ij|$. Due to the ZZ coupling, Δ defined in Eq. (5.13) is unequal to zero and the entangling rate of the gate is (roughly) determined by Δ since driving an uncoupled fluxonium qubit could not generate a CPHASE entangling gate. Going to a rotating (interaction) frame with reference Hamiltonian $H_{\text{ref}}/\hbar = \omega_d (P_{03} + P_{13})$ (and neglecting fast-rotating terms depending on $e^{\pm 2\omega_d t}$) gives a time-independent Hamiltonian

$$\frac{\tilde{H}}{\hbar} = \omega_{01} P_{01} + \omega_{11} P_{11}$$
$$+ \omega_{00} P_{00} + (\omega_{03} - \omega_d) P_{03} - \frac{\varepsilon_d q_{f,00-03}}{2} (|00\rangle \langle 03| + \text{h.c.}) \quad (5.37)$$
$$+ \omega_{10} P_{10} + (\omega_{13} - \omega_d) P_{13} - \frac{\varepsilon_d q_{f,10-13}}{2} (|10\rangle \langle 13| + \text{h.c.}).$$

Thus we see that one is driving Rabi oscillations in two effective qubit subsystems, namely the two-level subsystem $|00\rangle - |03\rangle$ and the two-level subsystem $|10\rangle - |13\rangle$. For a qubit Hamiltonian $H_{\text{qubit}}/\hbar = \begin{pmatrix} \alpha & \gamma/2 \\ \gamma/2 & \beta \end{pmatrix}$, one can use that

$$U(t) = e^{-i H_{\text{qubit}} t/\hbar} = e^{-i \text{Tr}(H_{\text{qubit}}) t/(2\hbar)} e^{-i\theta \hat{n} \cdot \vec{\sigma}/2}, \quad (5.38)$$

with angle $\theta = t \sqrt{(\alpha - \beta)^2 + \gamma^2}$. A full Rabi oscillation, which induces no leakage, occurs for $\theta = 2\pi$, so that $e^{-i\pi \hat{n} \cdot \vec{\sigma}} = -I$, $t_{\text{gate}} = \frac{2\pi}{\sqrt{(\alpha - \beta)^2 + \gamma^2}}$ and $U(t_{\text{gate}}) = -e^{-i(\alpha+\beta) t_{\text{gate}}/2} I$. Applying this to the simultaneous Rabi oscillations in the two subspaces, we see that one needs to fulfill the condition in Eq. (5.15) to get a full Rabi oscillation in both qubit subspaces. For this $t_{\text{gate}} = 2\pi/\Omega$, the phases picked up by the computational states are

$$\begin{aligned} \phi_{00} &= \pi - (\omega_{00} + \omega_{03} - \omega_d) t_{\text{gate}}/2, \\ \phi_{01} &= -\omega_{01} t_{\text{gate}} \\ \phi_{10} &= \pi - (\omega_{10} + \omega_{13} - \omega_d) t_{\text{gate}}/2 \\ \phi_{11} &= -\omega_{11} t_{\text{gate}}. \end{aligned} \quad (5.39)$$

A CPHASE gate can be brought to the form in Eq. (5.14) by single-qubit Z gates with $\phi = \phi_{11} - \phi_{10} - \phi_{01} + \phi_{00} \approx \frac{t_{\text{gate}}}{2} (\omega_{13} - \omega_{10} - \omega_{03} + \omega_{00}) = \frac{t_{\text{gate}}}{2} \Delta = \frac{\pi\Delta}{\Omega}$ where we have neglected the effect of the ZZ coupling in the computational subspace. Hence $\phi \approx \pi\Delta/\Omega$.

A given targeted phase ϕ thus leads to a targeted Ω (which sets the gate time t_{gate}), and the targeted Ω is used to solve for a drive-frequency ω_d and a drive-power ε_d which satisfies Eq. (5.15). The frequency ω_d is chosen to be close to $\omega_{03} - \omega_{00}$ and $\omega_{13} - \omega_{10}$, say midway between those transitions. The spectrum of the fluxonium and transmon qubit should be such that this choice of ω_d avoids it being close to other fluxonium transitions, such as $|01\rangle \leftrightarrow |04\rangle$.

We note that changing the phase θ_d of the drive has no effect on the gate as it simply changes the Rabi driving to be around an axis in the XY-plane instead of around the X-axis. We also note that if the transmon qubit is flux-tunable, one can vary Δ (letting it range from, say, negative to positive), and hence get a varying phase at a fixed t_{gate}.

5.7.3. Fidelity and leakage definitions

We take the definitions of gate fidelity and leakage from Ref. [81]. We report them here for completeness with some notational adaptation. Let P_c denote the projector onto the computational subspace encoding n qubits and $d_c = 2^n$ be its dimension. Let $\mathscr{U}$ be the quantum operation associated with a target unitary U that we want to implement within the computational subspace and that acts as the identity on the leakage subspace, i.e. $\mathscr{U}(\rho) = U\rho U^\dagger$ and $\mathscr{U}^\dagger$ is its inverse. Let $\mathscr{E}$ be the quantum operation we actually apply to the system. The average gate fidelity within the computational subspace is given by

$$F_{\text{gate}} = \int d\psi_c \langle\psi_c| U^\dagger \mathscr{E}(|\psi_c\rangle\langle\psi_c|)U |\psi_c\rangle, \tag{5.40}$$

where $d\psi_c$ denotes the Haar measure over states in the computational subspace. The process or entanglement fidelity in the computational subspace equals

$$F_{\text{ent}} = \langle\Psi_c| I \otimes \mathscr{U}^\dagger \circ \mathscr{E}(|\Psi_c\rangle\langle\Psi_c|) |\Psi_c\rangle, \tag{5.41}$$

where $|\Psi_c\rangle = \frac{1}{\sqrt{d_c}}\sum_{i=1}^{d_c} |i,i\rangle$ is the maximally-entangled state in the computational subspace. The average leakage is defined as

$$L_1 = 1 - \frac{1}{d_c}\text{Tr}[P_c \mathscr{U}^\dagger \circ \mathscr{E}(P_c)]. \tag{5.42}$$

For any trace-preserving channel $\mathscr{S}$ on a d_c-dimensional system we have the relation $F = \frac{d_c F_{\text{ent}} + 1}{d_c + 1}$ where F is the fidelity. Here, the effective channel on the computational space is $\mathscr{S}(\rho) = P_c[\mathscr{U}^\dagger \circ \mathscr{E}(P_c \rho P_c)]P_c$ which is not trace-preserving but trace-decreasing and $\text{Tr}\mathscr{S}(I/d_c) = 1 - L_1$. Incorporating this trace-decreasing property into the standard derivation [82] of the relation between process fidelity and gate fidelity gives

$$F_{\text{gate}} = \frac{d_c F_{\text{ent}} + 1 - L_1}{d_c + 1}. \tag{5.43}$$

Thus to compute the gate fidelity, one computes the entanglement or process fidelity F_{ent} and L_1. In turn, F_{ent} can be re-expressed, using $|\Psi_c\rangle\langle\Psi_c| = \frac{1}{d_c}\sum_\mu P_\mu \otimes P_\mu$ with (normalized) Pauli matrices P_μ ($\text{Tr}\,P_\mu P_\nu = \delta_{\mu\nu}$) as

$$F_{\text{ent}} = \frac{1}{d_c^2} \sum_{\mu=1}^{d_c^2} \text{Tr}(U P_\mu U^\dagger \mathscr{E}(P_\mu)). \tag{5.44}$$

So one evaluates $\mathcal{E}$ on the Pauli eigenstates of P_μ and takes the expectation value with the appropriate observable $UP_\mu U^\dagger$ etc. to compute F_{ent}. L_1 can be computed similarly, resulting through Eq. (5.43) in the evaluation of F_{gate}.

Parameter set	Fluxonium				
	$T_1^{0 \to 1}$ (μs)	$T_1^{2 \to 1}$ (μs)	$T_1^{4 \to 1}$ (μs)	$T_1^{3 \to 2}$ (μs)	$T_1^{4 \to 3}$ (μs)
CR	510	9	60	8	4
CPHASE	3976	7	90	81	4

Table 5.3: Relaxation and excitation times for other relevant fluxonium transitions for the CR and CPHASE parameter set in Table 5.1. The dielectric loss tangent and the temperature of the environment are taken as described in the caption of Table 5.1.

5.7.4. ERROR MODEL
In the main part of the chapter we have shown results of simulations under coherent and noisy evolutions. In this section we detail our noise model. As noise source we considered only relaxation due to dielectric losses. We do not include the effect of $1/f$ flux noise in the fluxonium since it is always assumed to be biased at $\phi_{\text{ext},f} = \pi$, which is a flux sweet spot. Clearly in a multi-qubit architecture with tunable transmons that could be biased away from the flux-insensitive point this source of error would play a role, similarly to the CPHASE gate in transmon-transmon architectures [14]. We do not include pure dephasing mechanisms, since the model of dephasing highly depends on the experimental setup one considers. Thus, while this noise source should be included when modeling an experiment, we left it out from our analysis.

Errors are assumed to be Markovian and modelled via a Lindblad master equation of the following form [83]

$$\frac{d\rho}{dt} = \frac{1}{i\hbar}[H(t), \rho] + \sum_k \mathscr{D}[L_k](\rho), \tag{5.45}$$

where ρ is the density matrix for the system, $H(t)$ is a general time-dependent Hamiltonian and $\mathscr{D}$ is the Lindblad dissipator

$$\mathscr{D}[L_k](\rho) = L_k \rho L_k^\dagger - \frac{1}{2} L_k^\dagger L_k \rho - \frac{1}{2} \rho L_k^\dagger L_k, \tag{5.46}$$

with L_k the so-called jump operators. In the following we specify the form of the jump operators modeling dielectric loss.

For both transmon as well as fluxonium qubits, dielectric loss can be modelled by adding a real part to the admittance (in the frequency domain) of the shunting capacitance [52]. More precisely, this admittance is assumed to be of the following form

$$Y_C(\omega) = \frac{\omega C}{Q_{\text{diel}}} + i\omega C, \tag{5.47}$$

with Q_{diel} the quality factor related to the dielectric loss tangent, namely $Q_{\text{diel}} = 1/\tan\delta_{\text{diel}}$. The dielectric loss tangent can in turn be frequency-dependent [52] and we consider this

in the case of the fluxonium, so $Q_{\text{diel}} = Q_{\text{diel}}(\omega)$ (see caption of Table 5.1 in the main text). The following discussion applies to both transmon and fluxonium. Let $|k\rangle$ and $|l\rangle$ be a pair of (bare) energy levels with transition frequency $\omega_{kl} = (E_k - E_l)/\hbar$ with E_k and E_l the energies associated with the levels and $k > l$. The decay rate from level $|k\rangle$ to $|l\rangle$ at temperature $T = 0$ reads

$$\gamma_{kl} = \frac{\Phi_0^2}{\hbar 2\pi^2} |\langle k|\phi|l\rangle|^2 \omega_{kl} \text{Re}\left[Y_C(\omega_{kl})\right], \tag{5.48}$$

with $\Phi_0 = h/2e$ the superconducting flux quantum and ϕ the reduced (dimensionless) flux operator of the system. At finite temperature $T > 0$, this is replaced by a relaxation rate

$$\gamma_{kl}^{\downarrow} = \gamma_{kl}[1 + \bar{n}(\omega_{kl})], \tag{5.49}$$

and an excitation rate

$$\gamma_{kl}^{\uparrow} = \gamma_{kl}\bar{n}(\omega_{kl}), \tag{5.50}$$

with average photon number

$$\bar{n}(\omega) = \frac{1}{e^{\beta\hbar\omega} - 1}, \tag{5.51}$$

where $\beta = 1/k_b T$. The relaxation times from level k to level l with $k > l$ reported in Tab. 5.1 are

$$T_1^{k \rightarrow l} = \frac{1}{\gamma_{kl}^{\downarrow}}, \tag{5.52}$$

while the excitation times are defined as

$$T_1^{l \rightarrow k} = \frac{1}{\gamma_{kl}^{\uparrow}}. \tag{5.53}$$

Thus, for any pair of energy levels we take two jump operators $L_{kl}^{\downarrow} = \sqrt{\gamma_{kl}^{\downarrow}} |l\rangle\langle k|$ and $L_{kl}^{\uparrow} = \sqrt{\gamma_{kl}^{\uparrow}} |k\rangle\langle l|$, which model relaxation and excitation between the two levels. While for transmon qubits the excitation rate between the first two levels can be neglected at the typical operating temperatures of few mK, this is not the case for the fluxonium. For low frequency fluxonia, the excitation rate $\gamma_{01}^{\uparrow}$ can be comparable to the relaxation rate. Table 5.3 shows some relevant relaxation and excitation times for the fluxonia we considered in this chapter.

5.7.5. DETAILS OF THE MICROWAVE PULSE AND ECHO
In the simulations in the main text we use a piece-wise Gaussian envelope which is defined as

$$g(t) = \begin{cases} \frac{1}{1 - e^{-t_{\text{rise}}^2/2\sigma^2}}\left[e^{-(t - t_{\text{rise}})^2/2\sigma^2} - e^{-t_{\text{rise}}^2/2\sigma^2}\right], & 0 \leq t < t_{\text{rise}} \\ 1, & t_{\text{rise}} \leq t < t_{\text{pulse}} - t_{\text{rise}} \\ \frac{1}{1 - e^{-t_{\text{rise}}^2/2\sigma^2}}\left[e^{-(t - (t_{\text{pulse}} - t_{\text{rise}}))^2/2\sigma^2} - e^{-t_{\text{rise}}^2/2\sigma^2}\right], & t_{\text{pulse}} - t_{\text{rise}} \leq t \leq t_{\text{pulse}}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.54}$$

using $\sigma = t_{\mathrm{rise}}/\sqrt{2\pi}$ and t_{pulse} the total pulse duration.

In Sec. 5.7.1 we have shown that, as in the transmon-transmon case, the CR effect comes with an additional X drive on the transmon. This term commutes with the CR term $X_t Z_f$ and gives rise to an unwanted X-rotation on the transmon during the gate. After each simulated gate, we undo this X rotation on the transmon by applying its (noiseless) inverse quantum operation. While an X drive is also present on the fluxonium, the large detuning between the drive frequency and the fluxonium frequency makes the effect of this term negligible.

In addition, we also consider an echo pulse similar to Refs. [19, 34, 73] with the goal to cancel the ZZ coupling and the X-rotation on the transmon qubit during the gate. Let $U_{GP}(t_{\mathrm{pulse}}, \varepsilon_d, \omega)$ be the time evolution operator when a Gaussian pulse as in Eq. (5.54) with total pulse time t_{pulse}, amplitude ε_d and frequency ω is applied to the control fluxonium. The echo pulse consists of two pulses with opposite sign of the drive amplitude applied on the fluxonium at chosen frequency $\omega = \omega_t$ of the target transmon, interleaved with single-qubit π rotations around the X-axis on the control, fluxonium, qubit $R_X^{(f)}(\pi)$. This gives the time evolution operator

$$U_{\mathrm{echo}} = R_X^{(f)}(\pi) U_{GP}(t_{\mathrm{pulse}}, -\varepsilon_d, \omega_t) \times R_X^{(f)}(\pi) U_{GP}(t_{\mathrm{pulse}}, \varepsilon_d, \omega_t). \tag{5.55}$$

In order to implement a CR gate the pulse time of each Gaussian pulse is chosen such that

$$\frac{|\mu_{\mathrm{CR}}|}{\hbar} \int_0^{t_{\mathrm{pulse}}} dt\, g(t) = \frac{\pi}{8}, \tag{5.56}$$

so applying essentially half the CR gate. Note that $\varepsilon_d \to -\varepsilon_d$ changes the sign $\mu_{CR} \to -\mu_{CR}$ in Eq. (5.34) and conjugation by π-pulses causes $Z_f \to -Z_f$, so that U_{echo} implements the CR gate. When we apply U_{echo}, both terms $\mu_{X_f} X_f$ and $\mu_{X_t} X_t$ in (5.32) cancel due to $\varepsilon_d \to -\varepsilon_d$.

In our simulation we take the single-qubit π-rotations $R_X^{(f)}(\pi)$ to be perfect. As discussed in Ref. [34], the echo pulse ideally cancels the effect of the ZZ coupling and the additional X-rotations but can also introduce some other unwanted terms in the effective time evolution operator although the overall effect is positive. In all cases, we always undo the accumulated single-qubit phases via virtual Z gates.

5.7.6. AVERAGE NUMBER OF COLLISIONS

In Sec. 5.5 we obtained the zero-collision device yield based on the frequency collisions and bounds outlined in Tab. 5.2. In this section we explore the average number of collisions for each type that have occurred for a $d = 3$ surface code lattice using a drive amplitude of $\varepsilon_d/2\pi = 300\,\mathrm{MHz}$, shown in Fig. 5.13. We observe that most collisions involve a spectator transmon, specifically, these are collisions of type 8 or 9. Together, these collisions account for most of the reduction in the zero-collision yield observed in Fig. 5.11. Collision type 8 results in the excitation of a spectator transmon from $|1\rangle$ to $|2\rangle$ during a CR gate. Given the target frequencies in Fig. 5.9, there exist pairs of transmons ($\omega_t/2\pi$ and $\omega_s/2\pi$ at 4.3 GHz and 4.7 GHz, or 5.3 GHz and 5.7 GHz, respectively) whose frequencies are ideally 100 MHz away from this collision. Collision type 9 results
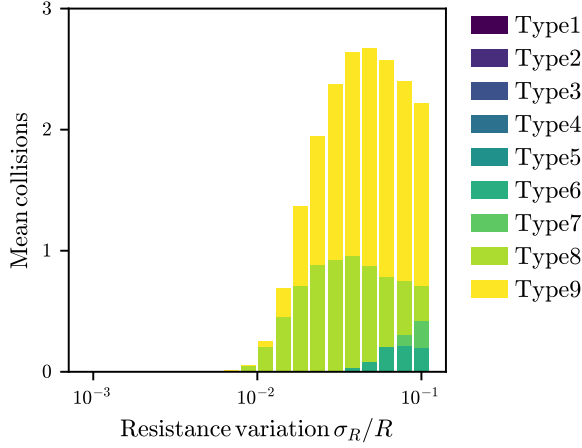
Figure 5.13: Average number of collisions of each type as a function of the resistance variation σ_R/R for a fixed-frequency transmon-fluxonium $d = 3$ surface code lattice and drive amplitudes $\varepsilon_d/2\pi = 300$ MHz. The number of collisions are collected over 6000 resamples of the lattice parameters. Collision type 5 is disabled due to the proximity of the transition to that of type 4. Instead the window around collision type 4 accounts for collision 5 as well.

in the fluxonium qubit leaking from $|0\rangle$ to $|4\rangle$, corresponding to a transition frequency of $\omega_{f,04}/2\pi = 9.86$ GHz for the target fluxonium parameters in Tab. 5.1. In this case, there are pairs of transmon frequencies ($\omega_t/2\pi$ and $\omega_s/2\pi$ at 4.3 GHz and 5.7 GHz, or 4.7 GHz and 5.3 GHz, respectively) the sum of which is 140 MHz away from this transition. In either case, the variation in the tunnel resistance of $\sigma_R/R = 2\%$ translates to a variation in ω_t and ω_s of about 1% each, which translates to a standard deviation of approximately 50 MHz, leading to the onset of these types of collisions. Collision type 6, which is the next most dominant collision, does not involve any spectator transmons and leads to the $|0\rangle$ to $|5\rangle$ transition on the fluxonium that happens at a frequency $\omega_{f,05}/2\pi = 13.23$ GHz. For the ideal transmon frequencies, this collision is ideally about 110 MHz away from the transmon at frequency $\omega_t = 4.3$ GHz and detuned by 300 MHz or more from any other transmon. The relatively large number of collisions of type 8 or 9 compared to any other types indicates that the target frequency allocation of the transmons is the main limiting factor behind the current yield.

BIBLIOGRAPHY

[1] A. Ciani, B. M. Varbanov, N. Jolly, C. K. Andersen, and B. M. Terhal, *Microwave-activated gates between a fluxonium and a transmon qubit*, Phys. Rev. Res. **4**, 043127 (2022).

[2] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, et al., *Charge-insensitive qubit design derived from the Cooper-pair box*, Phys. Rev. A **76**, 042319 (2007).

[3] 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).

[4] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, M. Gong, C. Guo, C. Guo, S. Guo, L. Han, L. Hong, H.-L. Huang, Y.-H. Huo, L. Li, N. Li, S. Li, Y. Li, F. Liang, C. Lin, J. Lin, H. Qian, D. Qiao, H. Rong, H. Su, L. Sun, L. Wang, S. Wang, D. Wu, Y. Xu, K. Yan, W. Yang, Y. Yang, Y. Ye, J. Yin, C. Ying, J. Yu, C. Zha, C. Zhang, H. Zhang, K. Zhang, Y. Zhang, H. Zhao, Y. Zhao, L. Zhou, Q. Zhu, C.-Y. Lu, C.-Z. Peng, X. Zhu, and J.-W. Pan, *Strong quantum computational advantage using a superconducting quantum processor*, Phys. Rev. Lett. **127**, 180501 (2021).

[5] 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).

[6] *IBM Quantum*, (2021) https://quantum-computing.ibm.com/.

[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] J. Majer, J. M. Chow, J. M. Gambetta, J. Koch, B. R. Johnson, J. A. Schreier, L. Frunzio, D. I. Schuster, A. A. Houck, A. Wallraff, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Coupling superconducting qubits via a cavity bus*, Nature **449,** 443–447 (2007).

[9] 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).

[10] 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,** 875–880 (2020).

[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*, Nat. Phys. **18,** 80–86 (2022).

[12] A. P. M. Place, L. V. H. Rodgers, P. Mundada, B. M. Smitham, M. Fitzpatrick, Z. Leng, A. Premkumar, J. Bryon, A. Vrajitoarea, S. Sussman, G. Cheng, T. Madhavan, H. K. Babla, X. H. Le, Y. Gang, B. Jäck, A. Gyenis, N. Yao, R. J. Cava, N. P. de Leon, and A. A. Houck, *New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds*, Nature Communications **12,** 1779 (2021).

[13] C. Wang, X. Li, H. Xu, Z. Li, J. Wang, Z. Yang, Z. Mi, X. Liang, T. Su, C. Yang, G. Wang, W. Wang, Y. Li, M. Chen, C. Li, K. Linghu, J. Han, Y. Zhang, Y. Feng, Y. Song, T. Ma, J. Zhang, R. Wang, P. Zhao, W. Liu, G. Xue, Y. Jin, and H. Yu, *Towards practical quantum computers: transmon qubit with a lifetime approaching 0.5 milliseconds*, npj Quantum Information **8,** 3 (2022).

[14] 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).

[15] 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).

[16] 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).

[17] 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).

[18] 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).

[19] 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), 060302 (2016).

[20] 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*, Phys. Rev. Lett. **127,** 200502 (2021).

[21] S. Krinner, P. Kurpiers, B. Royer, P. Magnard, I. Tsitsilin, J.-C. Besse, A. Remm, A. Blais, and A. Wallraff, *Demonstration of an all-microwave controlled-phase gate between far-detuned qubits*, Phys. Rev. Applied **14,** 044039 (2020).

[22] K. X. Wei, E. Magesan, I. Lauer, S. Srinivasan, D. F. Bogorin, S. Carnevale, G. A. Keefe, Y. Kim, D. Klaus, W. Landers, N. Sundaresan, C. Wang, E. J. Zhang, M. Steffen, O. E. Dial, D. C. McKay, and A. Kandala, *Hamiltonian engineering with multicolor drives for fast entangling gates and quantum crosstalk cancellation*, Phys. Rev. Lett. **129,** 060501 (2022).

[23] A. Kandala, K. X. Wei, S. Srinivasan, E. Magesan, S. Carnevale, G. A. Keefe, D. Klaus, O. Dial, and D. C. McKay, *Demonstration of a high-fidelity cnot gate for fixed-frequency transmons with engineered ZZ suppression*, Phys. Rev. Lett. **127,** 130501 (2021).

[24] M. R. Geller, E. Donate, Y. Chen, M. T. Fang, N. Leung, C. Neill, P. Roushan, and J. M. Martinis, *Tunable coupler for superconducting Xmon qubits: perturbative nonlinear model*, Phys. Rev. A **92,** 012320 (2015).

5

[25] F. Yan, P. Krantz, Y. Sung, M. Kjaergaard, D. L. Campbell, T. P. Orlando, S. Gustavs-son, and W. D. Oliver, *Tunable coupling scheme for implementing high-fidelity two-qubit gates*, Phys. Rev. App. **10**, 054062 (2018).

[26] 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. Marti-nis (Google AI Quantum), *Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms*, Phys. Rev. Lett. **125**, 120504 (2020).

[27] X. Li, T. Cai, H. Yan, Z. Wang, X. Pan, Y. Ma, W. Cai, J. Han, Z. Hua, X. Han, Y. Wu, H. Zhang, H. Wang, Y. Song, L. Duan, and L. Sun, *Tunable coupler for realizing a controlled-phase gate with dynamically decoupled regime in a superconducting circuit*, Phys. Rev. Applied **14**, 024070 (2020).

[28] M. Roth, M. Ganzhorn, N. Moll, S. Filipp, G. Salis, and S. Schmidt, *Analysis of a parametrically driven exchange-type gate and a two-photon excitation gate be-tween superconducting qubits*, Phys. Rev. A **96**, 062323 (2017).

[29] D. C. McKay, S. Filipp, A. Mezzacapo, E. Magesan, J. M. Chow, and J. M. Gambetta, *Universal gate for fixed-frequency qubits via a tunable bus*, Phys. Rev. Applied **6**, 064007 (2016).

[30] 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).

[31] 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).

[32] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. Di-Carlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Information **6**, 102 (2020).

[33] E. Magesan and J. M. Gambetta, *Effective hamiltonian models of the cross-resonance gate*, Physical Review A **101**, 052308 (2020).

[34] M. Malekakhlagh, E. Magesan, and D. C. McKay, *First-principles analysis of cross-resonance gate operation*, Phys. Rev. A **102**, 042605 (2020).

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

[36] J. Hertzberg, E. Zhang, S. Rosenblatt, E. Magesan, J. Smolin, J. Yau, V. Adiga, M. Sandberg, B. M., J. M. Chow, and J. S. Orcutt, *Laser-annealing josephson junctions for yielding scaled-up superconducting quantum processors*, npj Quantum Inf. **7**, 1–8 (2021).

[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. Applied **8**, 034021 (2017).

[38] A. Morvan, L. Chen, J. M. Larson, D. I. Santiago, and I. Siddiqi, *Optimizing frequency allocation for fixed-frequency superconducting quantum processors*, Phys. Rev. Res. **4**, 023079 (2022).

[39] N. Muthusubramanian, A. Bruno, B. Tarasinski, A. Fognini, R. Hagen, and L. DiCarlo, *Local trimming of transmon qubit frequency by laser annealing of Josephson junctions*, in Aps march meeting abstracts, Vol. 2019, APS Meeting Abstracts (Jan. 2019), B29.015.

[40] E. J. Zhang, S. Srinivasan, N. Sundaresan, D. F. Bogorin, Y. Martin, J. B. Hertzberg, J. Timmerwilke, E. J. Pritchett, J.-B. Yau, C. Wang, W. Landers, E. P. Lewandowski, A. Narasgond, S. Rosenblatt, G. A. Keefe, I. Lauer, M. B. Rothwell, D. T. McClure, O. E. Dial, J. S. Orcutt, M. Brink, and J. M. Chow, *High-performance superconducting quantum processors via laser annealing of transmon qubits*, Science Advances **8**, eabi6690 (2022).

[41] 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*, Phys. Rev. X **10**, 011022 (2020).

[42] S. Krinner, S. Lazar, A. Remm, C. K. Andersen, N. Lacroix, G. J. Norris, et al., *Benchmarking coherent errors in controlled-phase gates due to spectator qubits*, Phys. Rev. Appl. **14**, 024042 (2020).

[43] P. Mundada, G. Zhang, T. Hazard, and A. Houck, *Suppression of qubit crosstalk in a tunable coupling superconducting circuit*, Phys. Rev. Applied **12**, 054023 (2019).

[44] 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).

[45] P. Zhao, D. Lan, P. Xu, G. Xue, M. Blank, X. Tan, H. Yu, and Y. Yu, *Suppression of static ZZ interaction in an all-transmon quantum processor*, Phys. Rev. Applied **16**, 024037 (2021).

[46] V. E. Manucharyan, J. Koch, L. I. Glazman, and M. H. Devoret, *Fluxonium: single cooper-pair circuit free of charge offsets*, Science **326**, 113–116 (2009).

[47] V. E. Manucharyan, *Superinductance* (Yale University, 2012).

[48] N. A. Masluk, I. M. Pop, A. Kamal, Z. K. Minev, and M. H. Devoret, *Microwave characterization of Josephson junction arrays: implementing a low loss superinductance*, Phys. Rev. Lett. **109**, 137002 (2012).

5

[49] L. Grünhaupt, M. Spiecker, D. Gusenkova, N. Maleeva, S. T. Skacel, I. Takmakov, F. Valenti, P. Winkel, H. Rotzinger, W. Wernsdorfer, A. V. Ustinov, and I. M. Pop, *Granular aluminium as a superconducting material for high-impedance quantum circuits*, Nature Materials **18**, 816–819 (2019).

[50] T. M. Hazard, A. Gyenis, A. Di Paolo, A. T. Asfaw, S. A. Lyon, A. Blais, and A. A. Houck, *Nanowire superinductance fluxonium qubit*, Phys. Rev. Lett. **122**, 010504 (2019).

[51] M. Pita-Vidal, A. Bargerbos, C.-K. Yang, D. J. van Woerkom, W. Pfaff, N. Haider, P. Krogstrup, L. P. Kouwenhoven, G. de Lange, and A. Kou, *Gate-tunable field-compatible fluxonium*, Phys. Rev. Applied **14**, 064038 (2020).

[52] L. B. Nguyen, Y.-H. Lin, A. Somoroff, R. Mencia, N. Grabon, and V. E. Manucharyan, *High-coherence fluxonium qubit*, Phys. Rev. X **9**, 041041 (2019).

[53] A. Somoroff, Q. Ficheux, R. A. Mencia, H. Xiong, R. Kuzmin, and V. E. Manucharyan, *Millisecond coherence in a superconducting qubit*, Phys. Rev. Lett. **130**, 267001 (2023).

[54] L. B. Nguyen, G. Koolstra, Y. Kim, A. Morvan, T. Chistolini, S. Singh, K. N. Nesterov, C. Jünger, L. Chen, Z. Pedramrazi, B. K. Mitchell, J. M. Kreikebaum, S. Puri, D. I. Santiago, and I. Siddiqi, *Blueprint for a high-performance fluxonium quantum processor*, PRX Quantum **3**, 037001 (2022).

[55] H. Zhang, S. Chakram, T. Roy, N. Earnest, Y. Lu, Z. Huang, D. K. Weiss, J. Koch, and D. I. Schuster, *Universal fast-flux control of a coherent, low-frequency qubit*, Phys. Rev. X **11**, 011010 (2021).

[56] G. Zhu, D. G. Ferguson, V. E. Manucharyan, and J. Koch, *Circuit QED with fluxonium qubits: theory of the dispersive regime*, Phys. Rev. B **87**, 024510 (2013).

[57] D. Gusenkova, M. Spiecker, R. Gebauer, M. Willsch, D. Willsch, F. Valenti, N. Karcher, L. Grünhaupt, I. Takmakov, P. Winkel, et al., *Quantum nondemolition dispersive readout of a superconducting artificial atom using large photon numbers*, Phys. Rev. App. **15**, 064030 (2021).

[58] I. Takmakov, P. Winkel, F. Foroughi, L. Planat, D. Gusenkova, M. Spiecker, D. Rieger, L. Grünhaupt, A. Ustinov, W. Wernsdorfer, I. Pop, and N. Roch, *Minimizing the discrimination time for quantum states of an artificial atom*, Phys. Rev. Applied **15**, 064029 (2021).

[59] K. N. Nesterov, I. V. Pechenezhskiy, C. Wang, V. E. Manucharyan, and M. G. Vavilov, *Microwave-activated controlled-Z gate for fixed-frequency fluxonium qubits*, Phys. Rev. A **98**, 030301 (2018).

[60] 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).

[61] 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*, Phys. Rev. Res. **4**, 023040 (2022).

[62] Y. Chen, K. N. Nesterov, V. E. Manucharyan, and M. G. Vavilov, *Fast flux entangling gate for fluxonium circuits*, Phys. Rev. Appl. **18**, 034027 (2022).

[63] K. N. Nesterov, Q. Ficheux, V. E. Manucharyan, and M. G. Vavilov, *Proposal for entangling gates on fluxonium qubits via a two-photon transition*, PRX Quantum **2**, 020345 (2021).

[64] F. Bao, H. Deng, D. Ding, R. Gao, X. Gao, C. Huang, X. Jiang, H.-S. Ku, Z. Li, X. Ma, X. Ni, J. Qin, Z. Song, H. Sun, C. Tang, T. Wang, F. Wu, T. Xia, W. Yu, F. Zhang, G. Zhang, X. Zhang, J. Zhou, X. Zhu, Y. Shi, J. Chen, H.-H. Zhao, and C. Deng, *Fluxonium: an alternative qubit platform for high-fidelity operations*, Phys. Rev. Lett. **129**, 010502 (2022).

[65] E. Dogan, D. Rosenstock, L. Le Guevel, H. Xiong, R. A. Mencia, A. Somoroff, K. N. Nesterov, M. G. Vavilov, V. E. Manucharyan, and C. Wang, *Two-fluxonium cross-resonance gate*, Phys. Rev. Appl. **20**, 024011 (2023).

[66] A. Noguchi, A. Osada, S. Masuda, S. Kono, K. Heya, S. P. Wolski, H. Takahashi, T. Sugiyama, D. Lachance-Quirion, and Y. Nakamura, *Fast parametric two-qubit gates with suppressed residual interaction using the second-order nonlinearity of a cubic transmon*, Phys. Rev. A **102**, 062408 (2020).

[67] A. Maiani, M. Kjaergaard, and C. Schrade, *Entangling transmons with low-frequency protected superconducting qubits*, PRX Quantum **3**, 030329 (2022).

[68] S. Bravyi, D. P. DiVincenzo, and D. Loss, *Schrieffer–Wolff transformation for quantum many-body systems*, Annals of Physics **326**, 2793–2826 (2011).

[69] U. Vool and M. Devoret, *Introduction to quantum electromagnetic circuits*, International Journal of Circuit Theory and Applications **45**, 897–934 (2017).

[70] S. Rasmussen, K. Christensen, S. Pedersen, L. Kristensen, T. Bækkegaard, N. Loft, and N. Zinner, *Superconducting circuit companion—an introduction with worked examples*, PRX Quantum **2**, 040204 (2021).

[71] 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).

[72] D. DiVincenzo, ed., *Quantum information processing: lecture notes of the 44th IFF spring school 2013* (Forschungszentrum Juelich, Zentralbibliothek, 2013).

[73] 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).

[74] 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, 042348 (2018).

[75] 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).

[76] M. Malekakhlagh and E. Magesan, *Mitigating off-resonant error in the cross-resonance gate*, Phys. Rev. A **105**, 012602 (2022).

5

[77] J. M. Kreikebaum, K. P. O'Brien, A. Morvan, and I. Siddiqi, *Improving wafer-scale josephson junction resistance variation in superconducting quantum coherent circuits*, Superconductor Science and Technology **33**, 06LT02 (2020).

[78] D. G. Ferguson, A. A. Houck, and J. Koch, *Symmetries and collective excitations in large superconducting circuits*, Phys. Rev. X **3**, 011003 (2013).

[79] G. Consani and P. A. Warburton, *Effective Hamiltonians for interacting superconducting qubits: local basis reduction and the Schrieffer–Wolff transformation*, New Journal of Physics **22**, 053040 (2020).

[80] J. Pommerening, *Multiqubit coupling dynamics and the cross-resonance gate* (RWTH Aachen University, 2017).

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

[82] M. A. Nielsen, *A simple formula for the average gate fidelity of a quantum dynamical operation*, Physics Letters A **303**, 249–252 (2002).

[83] H. P. Breuer and F. Petruccione, *The theory of open quantum systems* (Oxford University Press, Oxford, 2007).

5

6

LOGICAL-QUBIT OPERATIONS IN AN ERROR-DETECTING SURFACE CODE

Future fault-tolerant quantum computers will require storing and processing quantum data in logical qubits. In this chapter, we realize a suite of logical operations on a distance-two surface code qubit built from seven physical qubits and stabilized using repeated error detection cycles. Logical operations include initialization into arbitrary states, measurement in the cardinal bases of the Bloch sphere, and a universal set of single-qubit gates. For each type of operation, we observe higher performance for fault-tolerant variants over non-fault-tolerant variants, and quantify the difference. In particular, we demonstrate process tomography of logical gates, using the notion of a logical Pauli transfer matrix. This integration of high-fidelity logical operations with a scalable scheme for repeated stabilization is a milestone on the road to quantum error correction with higher-distance superconducting surface codes.

6.1. INTRODUCTION

Two key capabilities will distinguish an error-corrected quantum computer from present-day noisy intermediate-scale quantum (NISQ) processors [2]. First, it will initialize, transform, and measure quantum information encoded in logical qubits rather than physical qubits. A logical qubit is a highly entangled two-dimensional subspace in the larger Hilbert space of many more physical qubits. Second, it will use repetitive quantum parity checks to discretize, signal, and (with aid of a decoder) correct errors occurring in the constituent physical qubits without destroying the encoded information [3]. Provided the incidence of physical errors is below a code-specific threshold and the quantum circuits for logical operations and stabilization are fault-tolerant, the logical error rate can be exponentially suppressed by increasing the distance (redundancy) of the quantum error correction (QEC) code employed [4]. At present, the exponential suppression for specific physical qubit errors (bit-flip or phase-flip) has been experimentaly demonstrated [5–8] for repetition codes [9–11].

Leading experimental quantum platforms have taken key steps towards implementing QEC codes protecting logical qubits from general physical qubit errors. In particular, trapped-ion systems have demonstrated logical-level initialization, gates and measurements for single logical qubits in the Calderbank-Shor-Steane [12] and Bacon-Shor [13] codes. Most recently, entangling operations between two logical qubits have been demonstrated in the surface code using lattice surgery [14]. However, except for smaller-scale experiments using two ion species [15], trapped-ion experiments in QEC have so far been limited to a single round of stabilization.

In parallel, taking advantage of highly-non-demolition measurement in circuit quantum electrodynamics [16], superconducting circuits have taken key strides in repetitive stabilization of two-qubit entanglement [17, 18] and logical qubits. Quantum memories based on 3D-cavity logical qubits in cat [19, 20] and Gottesman-Kitaev-Preskill [21] codes have crossed the memory break-even point. Meanwhile, monolithic architectures have focused on logical qubit stabilization in a surface code realized with a 2D lattice of transmon qubits. Currently, the surface code [22] is the most attractive QEC code for solid-state implementations owing to its practical nearest-neighbor-only connectivity requirement and high error threshold. Recent experiments [6, 23] have demonstrated repetitive stabilization by post-selection in a surface code which, owing to its small size, is capable of quantum error detection but not correction. In particular, Ref. [23] has demonstrated the preparation of logical cardinal states and logical measurement in two cardinal bases. In this chapter, we go beyond previous work by demonstrating a complete suite of logical-qubit operations for this small (distance-2) surface code while preserving multi-round stabilization. Our logical operations include initialization anywhere on the logical Bloch sphere (with significant improvement over previously reported fidelities), measurement in all cardinal bases, and a universal set of single-qubit logical gates. For each type of operation, we quantify the increased performance of fault-tolerant variants over non-fault-tolerant ones. We use a logical Pauli transfer matrix to describe a logical gate, analogous to the procedure commonly used to describe gates on physical qubits [24]. Finally, we perform logical state stabilization by means of repeated error detection where we compare the performance of two scalable, fault-tolerant stabilizer measurement schemes compatible with our quantum hardware architecture [25].

The distance-2 surface code (Fig. 6.1a) uses four data qubits (D_1 through D_4) to encode one logical qubit, whose two-dimensional codespace is the even-parity (i.e., eigenvalue +1) subspace of the stabilizer set

$$\mathscr{S} = \{Z_{D1}Z_{D3}, X_{D1}X_{D2}X_{D3}X_{D4}, Z_{D2}Z_{D4}\}. \tag{6.1}$$

This codespace has logical Pauli operators

$$Z_L = Z_{D1}Z_{D2}, \; Z_{D3}Z_{D4}, \; Z_{D1}Z_{D4}, \text{ and } Z_{D2}Z_{D3}, \tag{6.2}$$

$$X_L = X_{D1}X_{D3} \text{ and } X_{D2}X_{D4}, \tag{6.3}$$

that anti-commute with each other and commute with $\mathscr{S}$, and logical computational basis

$$|0_L\rangle = \frac{1}{\sqrt{2}}\left(|0000\rangle + |1111\rangle\right), \tag{6.4}$$

$$|1_L\rangle = \frac{1}{\sqrt{2}}\left(|0101\rangle + |1010\rangle\right). \tag{6.5}$$

Measuring the stabilizers using three ancilla qubits (A_1, A_2 and A_3 in Fig. 6.1a) allows detection of all physical errors that change the outcome of one or more stabilizers to $m = -1$. This list includes all errors on any one single qubit. However, no error syndrome is unique to a specific physical error. For instance, a phase flip in any one data qubit triggers the same syndrome: $m_{A2} = -1$. Consequently, this code cannot be used to correct such errors. We thus perform state stabilization by post-selecting runs in which no error is detected by the stabilizer measurements in any cycle. In this error-detection context, an operation is fault-tolerant if any single-fault produces a non-trivial syndrome and can therefore be post-selected out [26] (see Sec. 6.5.4 and Sec. 6.5.6).

6.2. RESULTS

6.2.1. STABILIZER MEASUREMENTS
Achieving high performance in a code hinges on performing projective quantum parity (stabilizer) measurements with high assignment fidelity, meaning one can accurately discriminate parity, and low additional backaction such that the state of the qubits after the measurement is properly projected onto the parity subspace. We implement each of the stabilizers in $\mathscr{S}$ using a standard indirect-measurement scheme [27, 28] with a dedicated ancilla. We benchmark the accuracy of each parity measurement by preparing the data-qubits in a computational state and measuring the probability of ancilla outcome $m_A = -1$. As a fidelity metric, we calculate the average probability to correctly assign the parity $Z_{D1}Z_{D3}$, $Z_{D1}Z_{D2}Z_{D3}Z_{D4}$ and $Z_{D1}Z_{D3}$, finding 94.2%, 86.1% and 97.2%, respectively (see Fig. 6.6).

6.2.2. LOGICAL STATE INITIALIZATION USING STABILIZER MEASUREMENTS
A practical means to quantify the backaction of stabilizer measurements is using them to initialize logical states. As proposed in Ref. [23], we can prepare arbitrary logical states
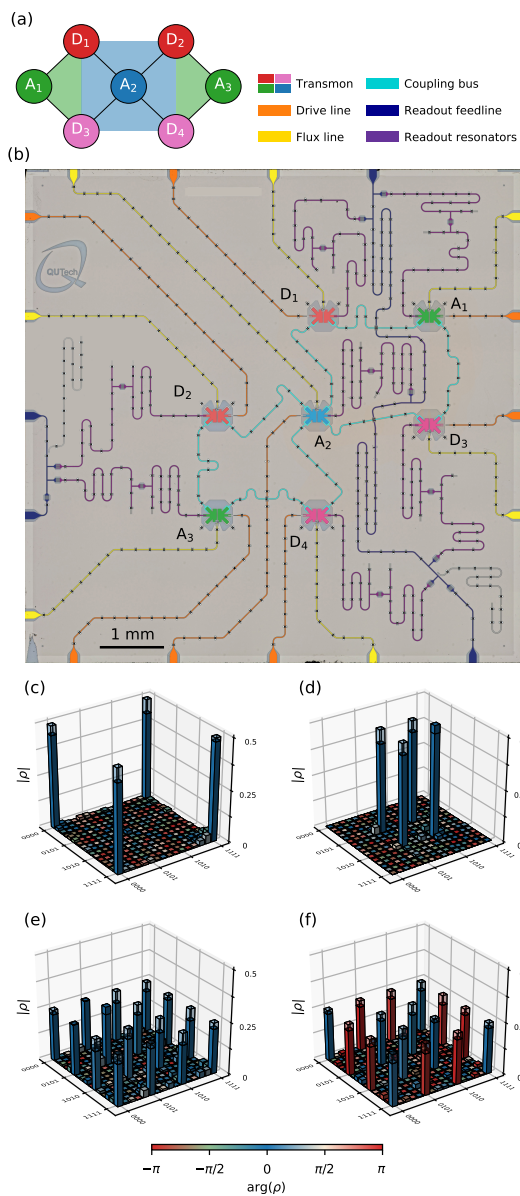
Figure 6.1: **Surface-7 quantum processor and initialization of logical cardinal states.** (a) Distance-two surface code. (b) Optical image of the quantum hardware with added false-color to emphasize different circuit elements. (c-f) Estimated physical density matrices, ρ, after targeting the preparation of the logical cardinal states $|0_L\rangle$ (c), $|1_L\rangle$ (d), $|+_L\rangle$ (e) and $|-_L\rangle$ (f). Each state is measured after preparing the data qubits in $|0000\rangle$, $|1010\rangle$, $|++++\rangle$ and $|++--\rangle$, respectively. The ideal target state density matrix is shown in the shaded wireframe.

by first initializing the data-qubit register in the product state

$$|\psi\rangle = (C_{\theta/2}|0\rangle + S_{\theta/2}|1\rangle)|0\rangle \left(C_{\theta/2}|0\rangle + S_{\theta/2}e^{i\phi}|1\rangle \right)|0\rangle \tag{6.6}$$

using single-qubit rotations R_y^θ on D_1 and R_ϕ^θ on D_3 acting on $|0000\rangle$ ($C_\alpha = \cos\alpha$ and $S_\alpha = \sin\alpha$). A follow-up round of stabilizer measurements ideally projects the four-qubit state onto the logical state

$$|\psi_L\rangle = \left(C_{\theta/2}^2|0_L\rangle + S_{\theta/2}^2 e^{i\phi}|1_L\rangle \right) / \sqrt{C_{\theta/2}^4 + S_{\theta/2}^4} \tag{6.7}$$

with probability

$$P = \frac{1}{2}\left(C_{\theta/2}^4 + S_{\theta/2}^4 \right). \tag{6.8}$$

We use this procedure to target initialization of the logical cardinal states $|0_L\rangle$, $|1_L\rangle$, $|+_L\rangle = \left(|0_L\rangle + |1_L\rangle\right)/\sqrt{2}$, and $|-_L\rangle = \left(|0_L\rangle - |1_L\rangle\right)/\sqrt{2}$. For the first two states, the procedure is fault-tolerant according to the definition above. We characterize the produced states using full four-qubit state tomography including readout calibration and maximum-likelihood estimation (MLE) (Fig. 6.1c-f). The fidelity F_{4Q} to the ideal four-qubit target states is $90.0 \pm 0.3\%$, $92.9 \pm 0.2\%$, $77.3 \pm 0.5\%$, and $77.1 \pm 0.5\%$, respectively. For each state, we can extract a logical fidelity F_L by further projecting the obtained four-qubit density matrix onto the codespace [23], finding $99.83 \pm 0.08\%$, $99.97 \pm 0.04\%$, $96.82 \pm 0.55\%$, and $95.54 \pm 0.55\%$, respectively (see Sec. 6.4.2). This sharp increase from F_{4Q} to F_L demonstrates that the vast majority of errors introduced by the parity check are weight-1 and detectable. A simple modification makes the initialization of $|+_L\rangle$ ($|-_L\rangle$) also fault-tolerant: initialize the data-qubit register in a different product state, namely $|++++\rangle$ ($|++--\rangle$), before performing the stabilizer measurements. With this modification, F_{4Q} increases to $85.4 \pm 0.3\%$ ($84.6 \pm 0.3\%$) and F_L to $99.78 \pm 0.09\%$ ($99.64 \pm 0.17\%$), matching the performance achieved when targetting $|0_L\rangle$ and $|1_L\rangle$.

6.2.3. Logical measurement of arbitrary states

A key feature of a code is the ability to measure logical operators. In the surface code, we can measure X_L (Z_L) fault-tolerantly, albeit destructively, by simultaneously measuring all data qubits in the X (Z) basis to obtain a string of data-qubit outcomes (each $+1$ or -1). The value assigned to the logical operator is the computed product of data-qubit outcomes as prescribed by Eq. 6.3 (6.2). Additionally, the outcome string is used to compute a value for the stabilizer(s) $X_{D1}X_{D2}X_{D3}X_{D4}$ ($Z_{D1}Z_{D3}$ and $Z_{D2}Z_{D4}$), enabling a final step of error detection (Fig. 6.2a). Measurement of $Y_L = +iX_LZ_L = Y_{D1}Z_{D2}X_{D3}$ is not fault-tolerant. However, we lower the logical assignment error by also measuring D_4 in the Z basis to compute a value for $Z_{D2}Z_{D4}$ and thereby detect bit-flip errors in D_2 and D_4.

We demonstrate Z_L, X_L and Y_L measurements on logical states prepared on two orthogonal planes of the logical Bloch sphere. Setting $\theta = \pi/2$ and sweeping ϕ, we ideally prepare logical states on the equator (Fig. 6.2d)

$$|\psi_L\rangle = \left(|0_L\rangle + e^{i\phi}|1_L\rangle\right) / \sqrt{2}. \tag{6.9}$$
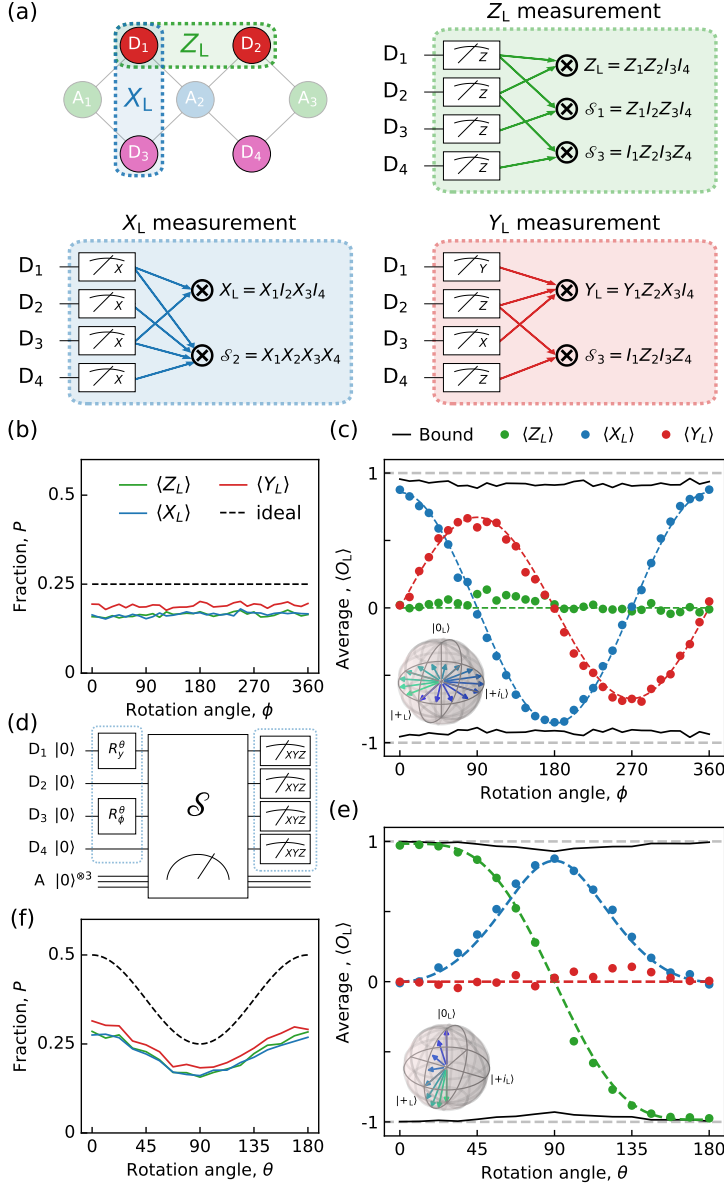
Figure 6.2: **Arbitrary logical-state initialization and measurement in the logical cardinal bases.** (a) Assembly of data-qubit measurements used to evaluate logical operators Z_L, X_L and Y_L with additional error detection. (d) Initialization of logical states using the procedure described in Eq. 6.6. (c, e) Z_L, X_L and Y_L logical measurement results as a function of the gate angles ϕ (c) and θ (e). The colored dashed curves show a fit of the analytical prediction based on Eqs. 6.9 and 6.11 to the data and the dark curve denotes a bound based on the measured F_L of each state. (b, f) Total fraction P of post-selected data as a function of the input angle for each logical measurement. The dashed curve shows the ideal fraction given by Eq. 6.8.

We measure the produced states in the Z_L, X_L and Y_L bases and obtain experimental averages $\langle Z_L \rangle$, $\langle X_L \rangle$ and $\langle Y_L \rangle$. As expected, we observe sinusoidal oscillations in $\langle X_L \rangle$ and $\langle Y_L \rangle$ and near-zero $\langle Z_L \rangle$. The reduced range of the $\langle Y_L \rangle$ oscillation evidences the non-fault-tolerant nature of Y_L measurement. A second manifestation is the higher fraction P of post-selected data for Y_L (Fig. 6.2b). To quantify the logical assignment fidelity F_L^R with correction for initialization error, it is tempting to apply the formula

$$\frac{\langle O_L \rangle_{\max} - \langle O_L \rangle_{\min}}{2} = (2F_L^R - 1)(2F_L - 1), \ O \in \{X, Y\} \tag{6.10}$$

inspired by the standard method to quantify readout fidelity of physical qubits from Rabi oscillations with limited initialization fidelity (described in Sec. 6.5.7). This method suggests $F_L^R = 95.8\%$ for X_L and 87.5% for Y_L. However, this method is not accurate for a logical qubit because not all input states outside the codespace are rejected by the limited set of stabilizer checks computable from the data-qubit outcome string and, moreover, detectable initialization errors can become undetectable when compounded with data-qubit readout errors. An accurate method to extract F_L^R based on the measured 16×16 data-qubit assignment probability matrix (detailed in Sec. 6.5.7) gives $F_L^R = 98.7\%$ for X_L and 91.4% for Y_L.

Setting $\phi = 0$ and sweeping θ, we then prepare logical states on the X_L-Z_L plane (Fig. 6.2e), ideally

$$|\psi_L\rangle = \left(C_{\theta/2}^2 |0_L\rangle + S_{\theta/2}^2 |1_L\rangle \right) / \sqrt{C_{\theta/2}^4 + S_{\theta/2}^4}. \tag{6.11}$$

Note that due to the changing overlap of the initial product state with the codespace, P is now a function of θ (Eq. 6.8). The approximate extraction method based on the range of $\langle Z_L \rangle$ suggests $F_L^R = 99.4\%$, while the accurate method gives 99.8%. Note that while both are fault-tolerant, the Z_L measurement has higher fidelity than the X_L measurement as the former is only vulnerable to vertical double bit-flip errors while the latter is vulnerable to both horizontal and diagonal double phase-flip errors.

6.2.4. LOGICAL GATES

Finally, we demonstrate a suite of gates enabling universal logical-qubit control (Fig. 6.3). Full control of the logical qubit requires a gate set comprising Clifford and non-Clifford logical gates. Some Clifford gates, like $R_{Z_L}^\pi$ and $R_{X_L}^\pi$ (where $R_{O_L}^\theta = e^{-i\theta O_L/2}$), can be implemented transversally and therefore fault-tolerantly (Fig. 6.3d). We perform arbitrary rotations (generally non-fault-tolerant) about the Z_L axis using the standard gate-by-measurement circuit [29] shown in Fig. 6.3a. In our case, the ancilla is physical (A_2), while the qubit transformed is our logical qubit. The rotation angle θ is set by the initial ancilla state $|A_\theta\rangle = (|0\rangle + e^{i\theta}|1\rangle)/\sqrt{2}$. Since we cannot do binary-controlled Z_L rotations, we simply post-select runs in which the measurement outcome is $m_{A2} = +1$. However, we note that these gates can be performed deterministicaly using repeat-until-success [30]. Choosing $\theta = \pi/4$ implements the non-Clifford $T_L = R_{Z_L}^{\pi/4}$ gate. A similar circuit (Fig. 6.3b) can be used to perform arbitrary rotations around the X_L axis. We compile both circuits using our hardware-native gateset (Figs. 6.3c,d). To assess logical-gate performance, we perform logical process tomography using the procedure illustrated in Fig. 6.3e for T_L. First, we initialize into each of the six logical cardinal
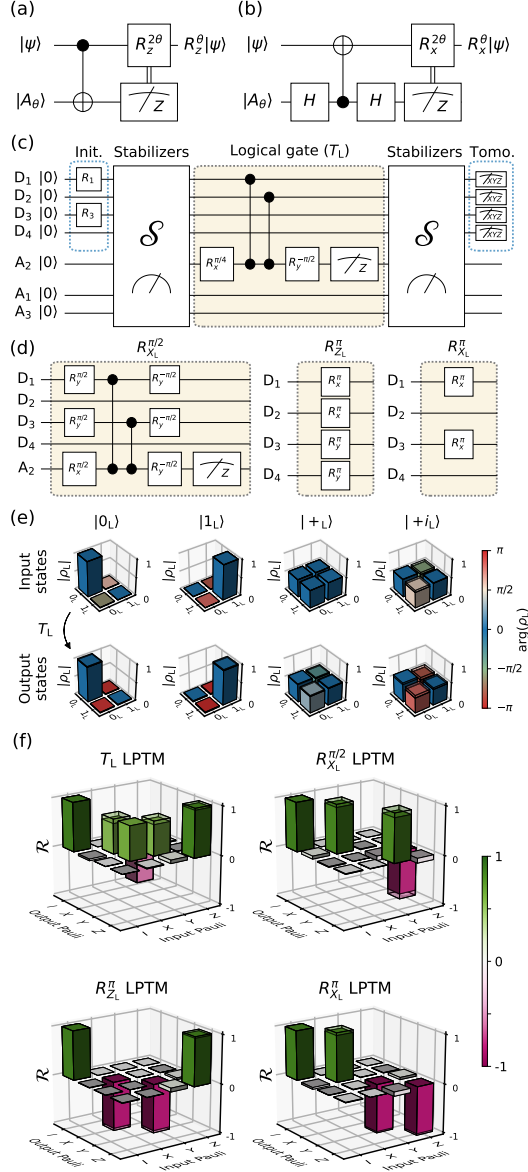
Figure 6.3: **Logical gates and their characterization.** (a, b) General gate-by-measurement schemes realizing arbitrary rotations around the Z (a) and X (b) axis of the Bloch sphere. (c) Process tomography experiment of the T_L gate. Input cardinal logical states are initialized using the method of Fig. 6.2. Output states are measured following a second round of stabilizer measurements. (d) Logical $R_{X_L}^{\pi/2}$, $R_{Z_L}^{\pi}$ and $R_{X_L}^{\pi}$ gates compiled using our hardware-native gateset. (e) Logical state tomography of input and output states of the T_L gate. These logical density matrices are obtained by performing four-qubit tomography of the data qubits and then projecting onto the codespace. (f) Extracted (solid) and ideal (wireframe) logical Pauli transfer matrices.

states $\{|0_L\rangle, |1_L\rangle, |+_L\rangle, |-_L\rangle, |+i_L\rangle, |-i_L\rangle\}$. We characterize each actual input state by four-qubit state tomography and project to the codespace to obtain a logical density matrix. Next, we similarly characterize each output state produced by the logical gate and a second round of stabilizer measurements to detect errors occurred in the gate (full data in Fig. 6.7). Using this over-complete set of input-output logical-state pairs, combined with MLE (see Sec. 6.4.3), we extract a logical Pauli transfer matrix (LPTM). The resulting LPTMs for the non-fault-tolerant T_L and $R_{X_L}^{\pi/2}$ gates as well as the fault-tolerant $R_{Z_L}^{\pi}$ and $R_{X_L}^{\pi}$ are shown in Fig. 6.3e. From the LPTMs, we extract average logical gate fidelities F_L^G (Eq. 6.19) 97.3%, 95.6%, 97.9%, and 98.1%, respectively.

6.2.5. PIPELINED VERSUS PARALLEL STABILIZER MEASUREMENTS
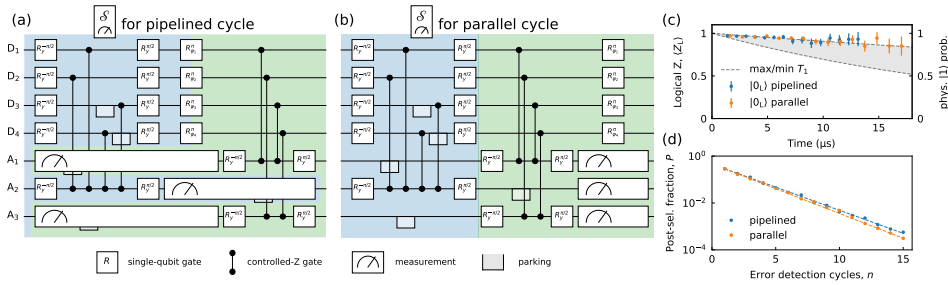
Figure 6.4: **Repetitive error detection using pipelined and parallel stabilizer measurement schemes.** (a, b) Gate sequences used to implement the pipelined (a) and parallel (b) stabilizer measurement schemes. Gate duration is 20 ns for single-qubit gates, 60 ns for controlled-Z (CZ) gates and parking [17, 25], and 540 ns for ancilla readout. The order of CZs in the $X_{D1}X_{D2}X_{D3}X_{D4}$ stabilizer (blue shaded region) prevents the propagation of ancilla errors into logical qubit errors [26]. The total cycle duration for the pipelined (parallel) scheme is 840 ns (1000 ns). (c) Estimated Z_L expectation value, $\langle Z_L \rangle$, measured for the $|0_L\rangle$ state versus the duration of the experiment using the pipelined (blue) and the parallel (orange) schemes. We also plot the excited-state probability (right axis) set by the maximum and minimum physical qubit T_1. (d) Post-selected fraction of data versus the number of error detection cycles n for the pipelined (blue) and parallel (orange) scheme.

A scalable control scheme is fundamental to realize surface codes with large code distance. To this end, we now compare the performance of two schemes suitable for the quantum hardware architecture proposed in Ref. [25]. These schemes are scalable in the sense that their cycle duration remains independent of code distance. The pipelined scheme interleaves the coherent operations and ancilla readout steps associated with stabilizer measurements of type X and Z by performing the coherent operations of X (Z) type stabilizers during the readout of Z (X) type stabilizers (Fig. 6.4a). The parallel scheme performs all ancilla readouts simultaneously (Fig. 6.4b). The pipelined cycle scheme duration is shorter than the parallel scheme by 16% which can potentially increase the performance of the code. This only occurs if the interleaved readout of ancillas does not result in increased measurement-induced dephasing between them. To compare their performance, we initialize and stabilize $|0_L\rangle$ for up to $n = 15$ cycles. We perform refocusing pulses ($R_{\varphi_i}^{\pi}$) on the data qubits to correct for coherent errors during the measurement of ancilla qubits. We also separately calibrate the equatorial rotation

axis of this gate for each scheme to extract the best performance. At each n, we take data back-to-back for the two schemes in order to minimize the effect of parameter drift, repeating each experiment up to 256×10^3 times. Figure 6.4c shows the Z_L measurement outcome averaged over the post-selected runs. We extract the error-detection rate γ from the n-dependence of the fraction of post-selected data P (Fig. 6.4d) using the procedure described in Sec. 6.4.4. We observe that the error rate is slightly lower for the pipelined scheme ($\gamma_{pip} \sim 45\%$), most likely due to the shorter duration of the cycle. This superiority is consistent across different input logical states (see Fig. 6.8) with an average ratio $\gamma_{pip}/\gamma_{par} \sim 97\%$.

6.3. DISCUSSION

We have demonstrated a suite of logical-level initialization, gate and measurement operations in a distance-2 superconducting surface code undergoing repetitive stabilizer measurements. For each type of logical operation, we have quantified the increased performance of fault-tolerant variants over non-fault-tolerant variants. Table 6.1 summarizes all the results. We can initialize the logical qubit to any point on the logical Bloch sphere, with logical fidelity surpassing Ref. [23]. In addition to characterizing initialized states using full four-qubit tomography, we also demonstrate logical measurements in all logical cardinal bases. Finally, we demonstrate a universal single-qubit set of logical gates by performing logical process tomography, using the concept of a logical-level Pauli transfer matrix. As expected, the fidelity of the fault-tolerant gates is higher than the non-fault-tolerant ones. However, one would expect a sharper difference given the typical error rates of the operations envolved. We believe this could be due to errors introduced by the stabilizer measurements which might be dominant over the errors of the logical gate itself.

With a view towards implementing higher-distance surface codes using our quantum-hardware architecture [25], we have compared the performance of two scalable stabilization schemes: the pipelined and parallel measurement schemes. In this comparison, two main factors compete. On one hand, the shorter cycle time favors pipelining. On the other, the pipelining introduces extra dephasing on ancilla qubits of one type during readout of the other. The performance of both schemes is comparable, but slightly higher for the pipelined scheme. From detailed density-matrix simulations discussed in Sec. 6.5.8, we further understand that conventional qubit errors such as energy relaxation, dephasing and readout assignment error alone do not fully account for the net error-detection rate observed in the experiment (see Fig. 6.14 and also not for the P reduction in Figs. 6.2b,f; see Fig. 6.15). We believe that the dominant error source is instead leakage to higher transmon states incurred during CZ gates. Our data (Fig. 6.13) shows that the error detection scheme successfully post-selects leakage errors in both the ancilla and data qubits. Learning to identify these non-qubit errors and to correct them without post-selection is the subject of ongoing research [31–33] and an outstanding challenge in the quest for quantum fault-tolerance with higher-distance superconducting surface codes [34], which to this date have yet to be implemented with repeated error correction.

Logical operation		Characteristic	Logical fidelity metric	value (%)
Init.	$|0_L\rangle$	FT	F_L	99.83
	$|1_L\rangle$	FT		99.97
	$|+_L\rangle$	Non-FT/FT		96.82/99.78
	$|-_L\rangle$	Non-FT/FT		95.54/99.64
Meas.	Z_L	FT	F_L^R	99.8
	X_L	FT		98.7
	Y_L	Non-FT		91.4
Gate	$R_{X_L}^\pi$	FT	F_L^G	97.9
	$R_{Z_L}^\pi$	FT		98.1
	$R_{X_L}^{\pi/2}$	Non-FT		95.6
	T_L	Non-FT		97.3

Table 6.1: **Summary of logical initialization, measurement, and gate operations and their performance.** Fault-tolerant operations are labelled FT and non-fault tolerant ones Non-FT. Quoted F_L^R values are those extracted with the accurate method described in Sec. 6.5.7.

6.4. METHODS

6.4.1. DEVICE

We use a superconducting circuit-QED processor (Fig. 6.1b) featuring the quantum hardware architecture proposed in Ref. [25]. Seven flux-tunable transmons are arranged in three frequency groups: a high-frequency group for D_1 and D_2; a middle-frequency group for A_1, A_2 and A_3; and a low-frequency group for D_3 and D_4. Similar to the device in Ref. [23], each transmon is transversely coupled to its nearest neighbors using a coupling bus resonator dedicated to each pair. This simplest and minimal connectivity minimizes multi-qubit crosstalk. Also, every transmon has a dedicated flux line for two-qubit gating, and a dispersively coupled readout resonator with Purcell filter enabling frequency-multiplexed readout [18, 35] using two feedlines. In contrast to Ref. [23], every transmon has a dedicated microwave drive line for single-qubit gating, avoiding the need to drive any via a feedline and thus reducing driving crosstalk.

All transmons are flux biased to their maximal frequency (i.e., flux sweetspot [36]), where measured qubit relaxation (T_1) and dephasing (T_2) times lie in the range 27—102 μs and 55—117 μs, respectively. Detailed information on the implementation and performance of single- and two-qubit gates for this same device can be found in Ref. [37]. Device characteristics are also summarized in Table S1.

The device was fabricated on a high-resistivity intrinsic Si<100> wafer that was first descummed using UV-ozone cleaner and stripped of native oxides using buffered oxide etch solution (BOE 7:1). The wafer was subjected to vapor of hexamethyldisilazane (HMDS) at 150°C and sputtered with 200 nm of niobium titanium nitride (NbTiN). Post dicing into smaller dies, a layer of hydrogen silsesquioxane (HSQ) was spun and baked at 300°C to serve as an inorganic sacrificial mask for wet etching of NbTiN. This layer was

removed post base-patterning steps. The quantum plane was defined using electron-beam (e-beam) lithography of a high-contrast, positive-tone resist spun on top of the NbTiN-HSQ stack. Post development, the exposed region was first dry etched using SF6/O2 mixture and then wet etched to remove any residual metal. Dolan-bridge-style Al/AlOx/Al Josephson junctions were then fabricated using standard double-angle e-beam evaporation. Airbridges and crossovers were added using a two-step process. The first step involved patterning galvanic contact using e-beam resist (~ 6 μm thick) subjected to reflow. In the second step, the airbridges and crossovers were patterned with e-beam evaporated Al (450 nm thick). Finally, the device underwent dicing, resist lift-off and Al wirebonding to a printed circuit board.

6.4.2. STATE TOMOGRAPHY

To perform state tomography on the prepared logical states, we measure the $4^4 - 1$ expectation values of data-qubit Pauli observables, $p_i = \langle \sigma_i \rangle, \sigma_i \in \{I, X, Y, Z\}^{\otimes 4}$ (except $I^{\otimes 4}$). Interleaved with this measurement we also characterize the measurement POVM used to correct for readout errors in p_i. These are then used to construct the density matrix

$$\rho = \sum_{i=0}^{4^4-1} \frac{p_i \sigma_i}{2^4} \tag{6.12}$$

with $p_0 = 1$, corresponding to $\sigma_0 = I^{\otimes 4}$. Due to statistical uncertainty in the measurement, the constructed state, ρ, might lack the physicality characteristic of a density matrix, that is, $\text{Tr}(\rho) = 1$ and $\rho \geq 0$. Specifically, ρ might not satisfy the latter constraint, while the former is automatically satisfied by $p_0 = 1$. To enforce these constraints, we use a maximum-likelihood method [24] to find the physical density matrix, ρ_{ph}, that is closest to the measured state, where closeness is defined in terms of best matching the measurement results. We thus minimize the cost function $\sum_{i=0}^{4^4-1} |p_i - \text{Tr}(\rho_{\text{ph}}\sigma_i)|^2$, subject to $\text{Tr}(\rho_{\text{ph}}) = 1$ and $\rho_{\text{ph}} \geq 0$. We find the optimal $\rho_{\text{ph}}^{\text{opt}}$ using the convex-optimization package *cvxpy* via *cvx-fit* in Qiskit [38]. The fidelity to a target pure state, $|\psi\rangle$, is then computed as

$$F = \langle \psi | \rho_{\text{ph}}^{\text{opt}} | \psi \rangle . \tag{6.13}$$

One can further project ρ_{ph} onto the codespace to obtain a logical state ρ_{L} using

$$\rho_{\text{L}} = \frac{1}{2} \sum_i \frac{\text{Tr}(\rho_{\text{ph}}\sigma_i^{\text{L}})}{\text{Tr}(\rho_{\text{ph}}I_{\text{L}})} \sigma_i^{\text{L}} , \ \sigma_i^{\text{L}} \in \{I_{\text{L}}, X_{\text{L}}, Y_{\text{L}}, Z_{\text{L}}\} \tag{6.14}$$

where I_{L} is the projector onto the codespace. Here, we can compute the logical fidelity F_{L} using Eq. 6.13.

6.4.3. PROCESS TOMOGRAPHY IN THE CODESPACE

A general single-qubit gate can be described [24] by a Pauli transfer matrix (PTM) $\mathscr{R}$ that maps an input state described by $p_i = \langle \sigma_i \rangle, \sigma_i \in \{I, X, Y, Z\}$, with $p_0 = 1$, to an output state p':

$$p'_j = \sum_i \mathscr{R}_{ij} p_i . \tag{6.15}$$

To construct the PTM $\mathcal{R}$ in the codespace, we use the overcomplete set of input states, $\{|0_L\rangle, |1_L\rangle, |+_L\rangle, |-_L\rangle, |+i_L\rangle, |-i_L\rangle\}$, and their corresponding output states and perform linear inversion. The input and output logical states are characterized using state tomography of the data qubits to find the four-qubit state ρ, which is then projected to the codespace using:

$$p_i^L = \frac{\text{Tr}(\rho\sigma_i^L)}{\text{Tr}(\rho I_L)} , \ \sigma_i^L \in \{I_L, X_L, Y_L, Z_L\}, \tag{6.16}$$

We find that all the measured logical states already satisfy the constraints of a physical density matrix. This is likely to happen as one-qubit states that are not very pure usually lie within the Bloch sphere even within the uncertainty in the measurement. The constructed LPTM, however, might not satisfy the constraints of a physical quantum channel, that is, trace preservation and complete positivity (TPCP). These are better expressed by switching from the PTM representation to the Choi representation. The Choi state $\rho^{\mathcal{R}}$ can be computed as

$$\rho^{\mathcal{R}} = \frac{1}{4} \sum_{i,j} \mathcal{R}_{ij} \sigma_j^T \otimes \sigma_i, \tag{6.17}$$

where the first tensor-product factor corresponds to an auxiliary subsystem. The TPCP constraints are $\text{Tr}(\rho_{ph}^{\mathcal{R}}) = 1$, $\rho_{ph}^{\mathcal{R}} \geq 0$ and $\text{Tr}_1(\rho_{ph}^{\mathcal{R}}) = 1/2$, where Tr_1 is the partial trace over the auxiliary subsystem. In other words, $\rho_{ph}^{\mathcal{R}}$ is a density matrix satisfying an extra constraint. We then find the optimal $\rho_{ph}^{\mathcal{R},\text{opt}}$ using the same convex-optimization methods as for state tomography and adding this extra constraint [24, 39]. We compute the corresponding LPTM via

$$(\mathcal{R}_{ph}^{\text{opt}})_{ij} = \text{Tr}(\rho_{ph}^{\mathcal{R},\text{opt}} \sigma_j^T \otimes \sigma_i). \tag{6.18}$$

and the average logical gate fidelity using

$$F_L^G = \frac{\text{Tr}(\mathcal{R}_{\text{ideal}}^\dagger \mathcal{R}_{ph}^{\text{opt}}) + 2}{6}, \tag{6.19}$$

where $\mathcal{R}_{\text{ideal}}$ is the LPTM of the ideal target gate.

6.4.4. Extraction of error-detection rate
The fraction of post-selected data P in the repetitive error detection experiment (Fig. 6.4b) decays exponentially with the number of cycles n. This is consistent with a constant error-detection rate per cycle γ. We extract this rate by fitting the function

$$P(n) = A(1-\gamma)^n. \tag{6.20}$$

6.5. Supplemental Material
6.5.1. Device characteristics

Qubit	D_1	D_2	D_3	D_4	A_1	A_2	A_3
Qubit transition frequency at sweetspot, $\omega_q/2\pi$ (GHz)	6.433	6.253	4.535	4.561	5.770	5.881	5.785
Transmon anharmonicity, $\alpha/2\pi$ (MHz)	-280	—	-320	—	-290	-285	—
Readout frequency, $\omega_r/2\pi$ (GHz)	7.493	7.384	6.913	6.645	7.226	7.058	7.101
Relaxation time, T_1 (μs)	27	44	32	102	38	58	43
Ramsey dephasing time, T_2^* (μs)	44	55	51	103	55	60	52
Echo dephasing time, T_2 (μs)	59	70	55	117	69	79	73
Best multiplexed readout fidelity, F_{RO}, (%)	98.6	98.9	96.0	96.5	98.6	94.2	98.9
Single-qubit gate fidelity, F_{SQ}, (%)	99.95	99.86	99.83	99.98	99.95	99.91	99.95

Table 6.2: **Summary of frequency, coherence and readout parameters of the seven transmons.** Coherence times are obtained using standard time-domain measurements [40]. Note that temporal fluctuations of several μs are typical for these values. The multiplexed readout fidelity, F_{RO}, is the average assignment fidelity [41] extracted from single-shot readout histograms after mitigating residual excitation using initialization by measurement and post-selection [42, 43].
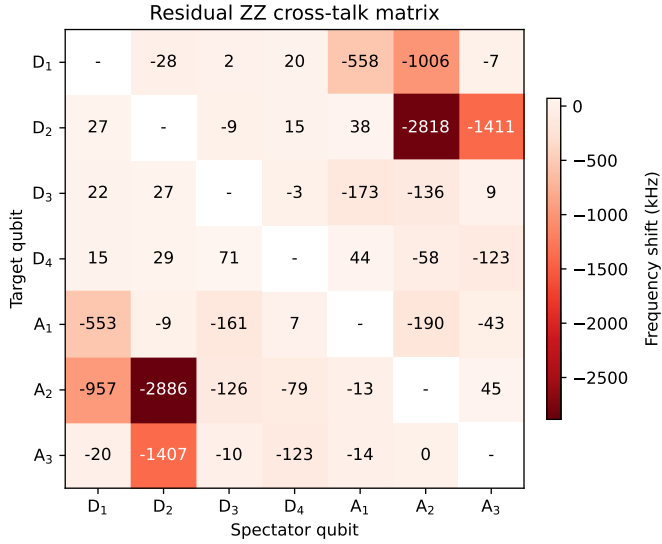
Figure 6.5: **Residual ZZ-coupling matrix.** Measured residual ZZ coupling between all transmon pairs at the bias point (their simultaneous flux sweetspot [36]). Each matrix element denotes the frequency shift that the target qubit experiences due to the spectator qubit being in the excited state, $|1\rangle$. The procedure used for this measurement is similar to the one described in Ref. [44].

6.5.2. PARITY-CHECK PERFORMANCE
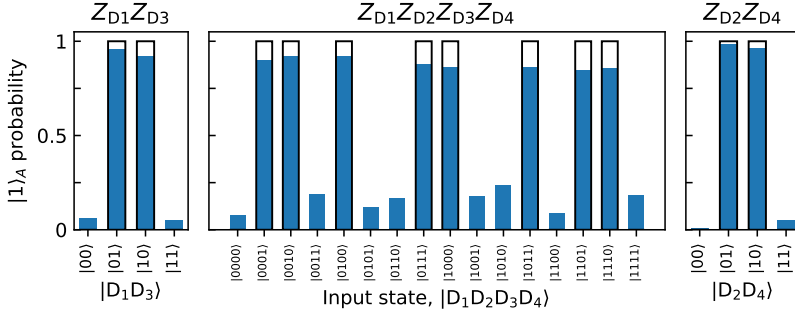
Figure 6.6: **Characterization of the assignment fidelity of Z-type parity checks.** (a) $Z_{D1}Z_{D3}$, (b) *$Z_{D1}Z_{D2}Z_{D3}Z_{D4}$, and (c) $Z_{D2}Z_{D4}$ parity checks implemented using A_1, A_2, and A_3, respectively. Each parity check is benchmarked by preparing the relevant data qubits in a computational state and then measuring the probability of ancilla outcome $m_{Ai} = -1$. Measured (ideal) probabilities are shown as solid blue bars (black wireframe). From the measured probabilities we extract average assignment fidelities 94.2%, 86.1% and 97.2%, respectively. *This parity check implements the $X_{D1}X_{D2}X_{D3}X_{D4}$ stabilizer measurement with the addition of single-qubit gates on data qubits to perform a change of basis.

6

6.5.3. PROCESS TOMOGRAPHY
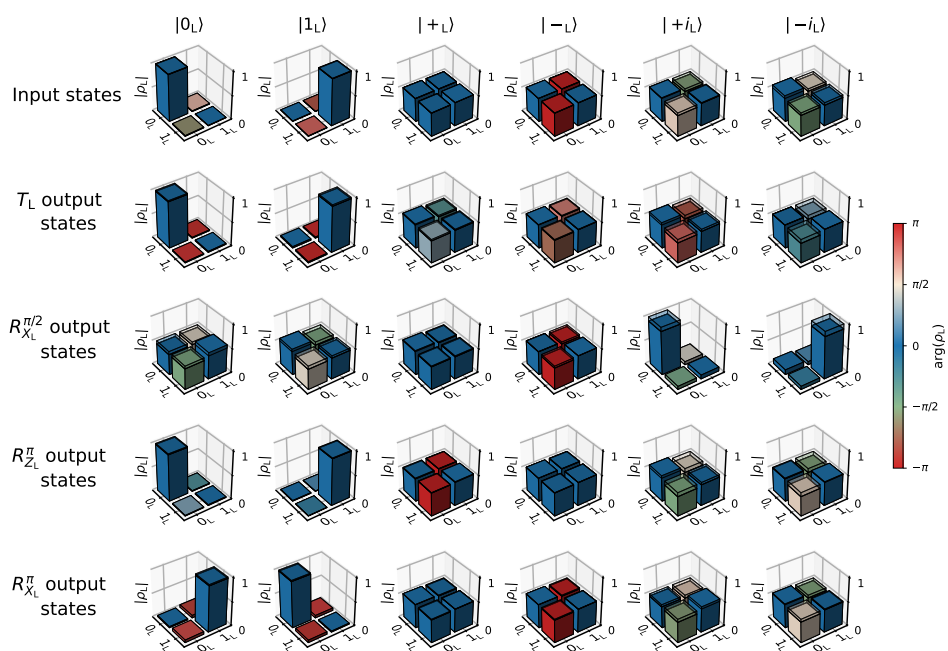
Figure 6.7: **Full set of logical states measured in the logical process tomography procedure.** Measured input and output logical states for each logical gate. Each state is measured using the procedure described in Sec. 6.4.3.

6.5.4. Logical state stabilization
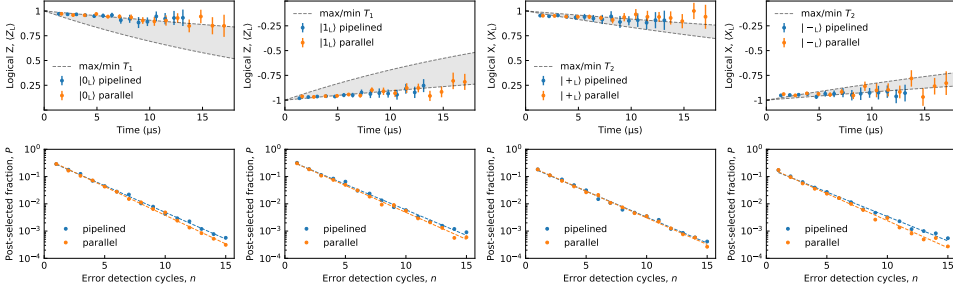
Figure 6.8: **Stabilization of logical cardinal states by repetitive error detection using the pipelined and parallel schemes.** From left to right, the stabilized logical states are $|0_L\rangle$, $|1_L\rangle$, $|+_L\rangle$ and $|-_L\rangle$. For each logical state, the top panel shows the evolution of the relevant logical operator as a function of number of cycles, n, plotted versus wall-clock time. Error bars are estimated based on the statistical uncertainty given by $P(n)$. The shaded area indicates the range of physical qubit T_1 values (a and b) and T_2 values (c and d) plotted on the right-axis. Each bottom panel shows the corresponding post-selected fraction of data, $P(n)$.
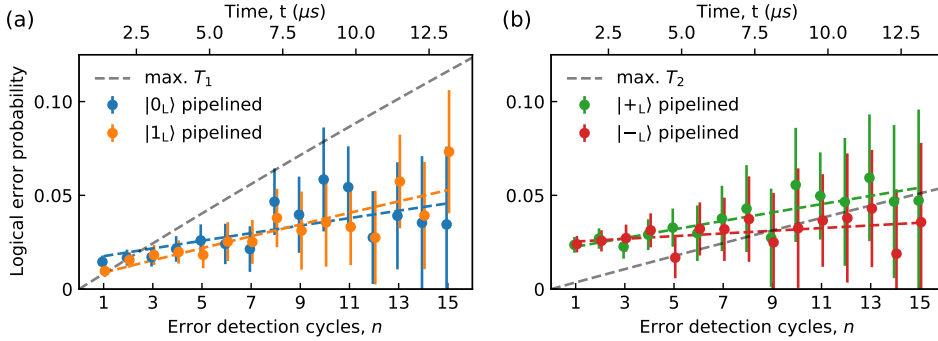
Figure 6.9: **Logical error probability versus number of error detection cycles.** Logical error probability after n cycles of error detection for states $|0_L\rangle$, $|1_L\rangle$ (a) and $|+_L\rangle$, $|-_L\rangle$ (b) measured using the pipelined scheme. For comparison, the grey dashed curves in (a) and (b) correspond to the physical error probability of the best T_1 and T_2 respectively. The logical error rate per round of detection, extracted by fitting the data (colored dashed lines), is 0.43%, 0.67%, 0.49% and 0.15% respectively.

6.5.5. Logical error rate

Here, we study the error rate of the logical qubit and compare it to that of a physical qubit. The probability for a logical error on an eigenstate of O_L after n cycles is given by

$$P_{\text{error}}^{\text{L}} = \frac{1 - |\langle O_L \rangle(n)|}{2}. \tag{6.21}$$

For eigenstates of Z_L we compare the measured error rate to the error experienced due to T_1 on a physical qubit (Fig. 6.9a). In a physical qubit in the excited state, $|1\rangle$, this error

is given by

$$P_{\text{error}} = 1 - e^{-t/T_1}. \tag{6.22}$$

For eigenstates of X_{L} (Fig. 6.9b) we now consider the error experienced due to T_2. This error for physical qubits whose state lies on the equator of the Bloch sphere is

$$P_{\text{error}} = \frac{1 - e^{-t/T_2}}{2}. \tag{6.23}$$

We find that the logical error rates for all states, corresponding to the slopes of the colored dashed curves in Fig. 6.9, are lower than the corresponding best physical error rates.

6.5.6. Fault tolerance of logical operations

Fault tolerance of an operation

We begin by elaborating the definition of a fault-tolerant logical operation. We consider a single fault occurring during the circuit implementing the logical operation, where a fault can refer to any single-qubit Pauli error following a single-qubit gate or an idling step, or any two-qubit Pauli error following a two-qubit gate or a measurement error. Furthermore, a single fault can also refer to any single-qubit error on the input state of the logical operation. Thus we consider the performance of the logical operation either when there is a single error at input *or* a fault in the logical operation. In the context of error detection, the logical operation such as state initialization or gate execution, is fault-tolerant if any such fault either produces a non-trivial syndrome (in case the circuit involves the measurement of the stabilizers) and is thus post-selected out or leads to an outgoing state that is either the desired logical state or any logical state together with a detectable error. This implies that if the logical operation is followed up by a fictitious and ideal measurement of the stabilizers, the detectable error would lead to a non-trivial syndrome and be post-selected out, ensuring that the outgoing state could only be the desired logical state. For a fault-tolerant logical measurement we require that the logical measurement outcome is correct, i.e. if it is applied to a logical state with single error *or* a fault happens during the logical measurement, we either post-select or get the correct outcome.

Logical state initialization using stabilizer measurements

We perform fault-tolerant initialization of the logical cardinal states $|0_{\text{L}}\rangle$, $|1_{\text{L}}\rangle$, $|+_{\text{L}}\rangle$ and $|-_{\text{L}}\rangle$. We focus on the initialization of $|0_{\text{L}}\rangle$ and $|1_{\text{L}}\rangle$ and then extend these arguments to $|+_{\text{L}}\rangle$ and $|-_{\text{L}}\rangle$.

To prepare $|0_{\text{L}}\rangle$ ($|1_{\text{L}}\rangle$) the data-qubit register is prepared in the state $|0000\rangle$ ($|1010\rangle$) and a single round of stabilizer measurements is performed. Consider any single-qubit error occurring during the initialization of the data qubits. Any such error will be detected by the following stabilizer measurement and post-selected out.

Then, consider a fault occurring during the stabilizer measurement, implemented by the circuits shown in Fig. 6.4. Any single-qubit error on the data qubits following an idling step either leads to non-trivial syndrome (if it occurs before the two-qubit gate) or constitutes a detectable error (if it occurs after the two-qubit gate). A single-qubit error following any of the single-qubit gates will similarly either produce non-trivial syndrome or lead to an error that is either detectable or an element in $\mathscr{S}$.

A measurement error on ancilla qubits A_1 or A_3 will always produce a non-trivial syndrome (since the input state is an eigenstate of the measured Z type stabilizers) and will thus be post-selected out. However, a measurement error on A_2 can still lead to trivial syndrome (as the input state is not an eigenstate of the X-type stabilizer). This will result in the preparation of the desired logical state together with a detectable phase-flip error on any of the qubits.

Any two-qubit Pauli error after each CZ gate involved in the measurement of the Z-type stabilizers will lead to the preparation of the desired logical state together with an error that is either in $\mathscr{S}$ (for example in the case when a bit-flip error occurs on the ancilla qubit and phase-flip error on the data-qubit following the first CZ gate of the circuits) or one that is detectable. The same statement holds for any two-qubit error after each of the CZ gates involved in the X type stabilizer check. Here the order of the gates is crucial to ensure that any two-qubit error after the second CZ gate of the circuit is detectable [26].

The fault-tolerant preparation of $|+_L\rangle$ ($|-_L\rangle$) involves initializing the data-qubit register in the state $|++++\rangle$ ($|++--\rangle$) instead. The arguments for the fault-tolerance of this operation follow closely the ones presented for $|0_L\rangle$ ($|1_L\rangle$) with the only difference being that a single measurement error on ancilla qubits A_1 or A_3 can now lead to a trivial outcome and an outgoing state that involves a detectable (bit-flip) error, while a measurement error on A_2 will instead always lead to a non-trivial syndrome.

When initializing arbitrary logical states by preparing the data qubit register in the state given by Eq. 6, the procedure is fault-tolerant only when preparing $|0_L\rangle$ (corresponding to $\theta = 0$ and $\phi = 0$) or $|1_L\rangle$ (corresponding to $\theta = \pi$ and $\phi = 0$), which are discussed above. When preparing $|+_L\rangle$ (corresponding to $\theta = \pi/2$ and $\phi = 0$) or $|-_L\rangle$ (corresponding to $\theta = \pi/2$ and $\phi = \pi$), the input states are $|+0+0\rangle$ and $|+0-0\rangle$ respectively. In these cases a single fault (for example a phase-flip error on qubit D_3 on the input state) is not detectable by the stabilizer measurement and instead leads to the initialization of (the opposite states) $|-_L\rangle$ and $|+_L\rangle$, respectively. The preparation of any other state on the equator of the Bloch sphere is not fault-tolerant either, following the same reasoning: an under- or overrotation of ϕ will directly translate to an error at the logical level.

FAULT-TOLERANT LOGICAL MEASUREMENTS

We now consider the fault-tolerance of the logical measurement, which is performed following the procedure described in Sec. 6.2.3 (see Fig. 6.2). The only fault to consider in these circuits is a measurement error on one of the data qubits. When measuring X_L or Z_L, any such error will result in a non-trivial syndrome (given the assumption that the input state is in the codespace) and the logical measurement outcome is post-selected out. When the fault is instead a single-qubit error on the input state, bringing this state outside of the codespace, the fault-free logical measurement will either detect this error or this error will not have an affect on the logical measurement outcome.

For the non-fault-tolerant measurement of Y_L only the value for the $Z_{D2}Z_{D4}$ stabilizer can be computed and used to detect errors on D_2 and D_4. Thus a single fault (for example a measurement error on either D_1 or D_3) can lead to an incorrect logical measurement outcome, making this operation non-fault-tolerant.

6

TRANSVERSAL LOGICAL GATES AND NON-FAULT-TOLERANT GATE INJECTION

The logical gates $R_{X_L}^\pi$ and $R_{Z_L}^\pi$ (shown in Fig. 6.3d) are clearly fault-tolerant as any single-qubit error following any of the single-qubit gates involved in the circuits is detectable. At the same time the transversal execution of these gates ensures that no single qubit error on the input state can spread to two or more qubits, ensuring that any such fault is detectable. These fault-tolerant properties do not hold when we consider the T_L and $R_{X_L}^{\pi/2}$ logical gates implemented by the gate-by-measurement circuits shown in Fig. 6.3b (and Fig. 6.3d). For example a bit-flip error on A_2 following the first single-qubit gate of the circuit will result in a logical error. More generally, any under- or over-rotation in the rotation angle θ used in preparing the ancilla qubit in $|A_\theta\rangle$ translates to a different rotation at the logical level than desired.

6.5.7. QUANTIFYING THE LOGICAL ASSIGNMENT FIDELITY

We start this section reviewing how the readout fidelity F^R of a physical qubit is standardly quantified from the contrast of a Rabi oscillation when the input states $\rho_\pm$ closest to the eigenstates $|\Psi_\pm\rangle$ of the measured observable O have limited fidelity F (assumed equal for both). Evidently, we want F^R to quantify the performance of readout only, independent of errors in the input state. To this end, consider the probability flow diagram of Fig. 6.10. We define F^R as the average probability of proper assignment for perfect input states $|\Psi_\pm\rangle$. Therefore, $F^R = 1 - (\epsilon_+ + \epsilon_-)/2$, where $\epsilon_\pm$ is the probability of wrongly assigning outcome ∓ 1 for input state $|\Psi_\pm\rangle$. The positive and negative extremes of the Rabi oscillation are

$$\langle O \rangle_{\max} = F\bar{\epsilon}_+ + \bar{F}\epsilon_- - F\epsilon_+ - \bar{F}\bar{\epsilon}_-. \tag{6.24}$$

$$\langle O \rangle_{\min} = -F\bar{\epsilon}_- - \bar{F}\epsilon_+ + F\epsilon_- + \bar{F}\bar{\epsilon}_+, \tag{6.25}$$

where $\bar{F} = 1 - F$ and $\bar{\epsilon}_\pm = 1 - \epsilon_\pm$. Combining these expressions and simplifying terms, it follows that the contrast of the Rabi oscillation (defined as half the peak-to-peak range), is

$$\frac{\langle O \rangle_{\max} - \langle O \rangle_{\min}}{2} = (2F - 1)(2F^R - 1). \tag{6.26}$$

Turning over to the logical qubit, we similarly define the logical readout fidelity F_L^R as the average probability of proper assignment for perfectly prepared logical states $|\Psi_{L\pm}\rangle$ i.e., the logical states that are eigenstates of the measured observable O_L with eigenvalue ± 1. To this end, it is tempting to apply the above equation to the oscillations in Fig. 6.2, simply substituting $F \rightarrow F_L$ and $F^R \rightarrow F_L^R$. However, this approach is not accurate. This is because the probability flow diagram for the logical qubit, shown in Fig. 6.11, is more complex. The quantities we seek to determine are those inside the dashed box, describing logical readout on perfect input logical states: $p_{R\pm}$ is the probability that the experimental logical measurement on $|\Psi_{\pm L}\rangle$ is rejected (R), which occurs whenever the data-qubit outcome string produces a value of -1 on at least one of the stabilizers computable from the string; $\epsilon_{L\pm}$ is the probability of wrongly assigning logical outcome ∓ 1 for $|\Psi_{L\pm}\rangle$, conditioned on no rejection. Using these definitions, $F_L^R = 1 - (\epsilon_{L+} + \epsilon_{L-})/2$. Outside the dashed box, $\rho_\pm$ is the experimental input state closest to $|\Psi_{L\pm}\rangle$. This imperfect input state has probability $p_{L\pm}$ of being in the codespace and its projection onto the

Figure 6.10: **Probability flow diagram for physical qubit readout**. Please see text for the definition of all variables shown. The characterization of physical qubit readout robust to initialization errors determines the probabilities within the dashed box.

codespace, $\rho_{L\pm}$, has fidelity $F_{L\pm}$ to $|\Psi_{L\pm}\rangle$. Finally, $\rho_{\perp\pm}$ is the projection of $\rho_\pm$ outside the codespace.

We now discuss the more accurate method used to quantify F_L^R that does not rely on Eq. 6.26. We first consider the transformation of $|\Psi_{L\pm}\rangle$ by the pre-rotations that we perform when measuring in each cardinal logical basis, $O_L \in \{Z_L, X_L, Y_L\}$. For Z_L there are no measurement pre-rotations, so the states are

$$|0_L\rangle = \frac{1}{\sqrt{2}} (|0000\rangle + |1111\rangle),\tag{6.27}$$

$$|1_L\rangle = \frac{1}{\sqrt{2}} (|0101\rangle + |1010\rangle).\tag{6.28}$$

For X_L, the transformed states are

$$R_{1y}^{-\pi/2} R_{2y}^{-\pi/2} R_{3y}^{-\pi/2} R_{4y}^{-\pi/2} |+_L\rangle = \frac{1}{2} (|0000\rangle + |0101\rangle + |1010\rangle + |1111\rangle),\tag{6.29}$$

$$R_{1y}^{-\pi/2} R_{2y}^{-\pi/2} R_{3y}^{-\pi/2} R_{4y}^{-\pi/2} |-_L\rangle = \frac{1}{2} (|0011\rangle + |0110\rangle + |1001\rangle + |1100\rangle).\tag{6.30}$$

Finally, for Y_L, these are

$$R_{1x}^{\pi/2} R_{3y}^{-\pi/2} |+i_L\rangle = \frac{1}{2} (|0000\rangle - i|0111\rangle + i|1010\rangle + |1101\rangle),\tag{6.31}$$

$$R_{1x}^{\pi/2} R_{3y}^{-\pi/2} |-i_L\rangle = \frac{1}{2} (-|0010\rangle - i|0101\rangle - i|1000\rangle + |1111\rangle).\tag{6.32}$$

The above expressions make clear, for each logical cardinal basis, which data-qubit outcome strings are rejected and which ones are accepted with declared logical outcome $+1$ or -1. For completeness, all cases are detailed in Table 6.3.
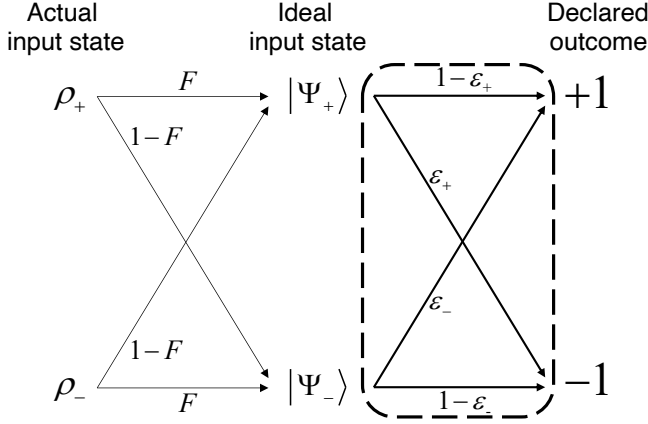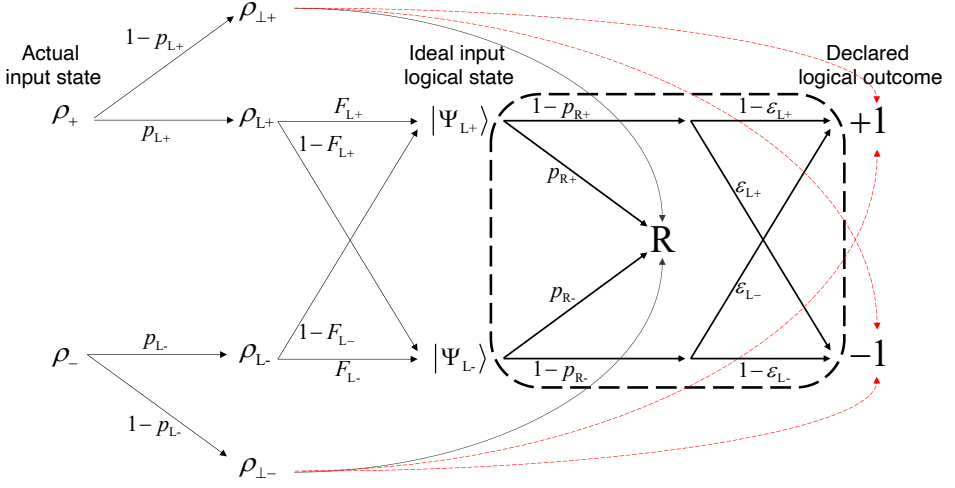
Figure 6.11: **Probability flow diagram for logical readout.** Please see text for the definition of all variables shown. The characterization of logical qubit readout robust to initialization errors determines the probabilities within the dashed box. The probability for states outside the codespace to not be rejected (red dashed curves) is the primary reason why using Eq. 6.26 with the substitutions $F \to F_L$ and $F^R \to F_L^R$ does not yield an accurate estimate of F_L^R.

Data-qubit outcome	Logical assignment			Data-qubit outcome	Logical assignment		
$(m_{D1}, m_{D2}, m_{D3}, m_{D4})$	Z_L	X_L	Y_L	$(m_{D1}, m_{D2}, m_{D3}, m_{D4})$	Z_L	X_L	Y_L
$(+1,+1,+1,+1)$	$+1$	$+1$	$+1$	$(-1,+1,+1,+1)$	R $(\mathscr{S}_1)$	R $(\mathscr{S}_2)$	-1
$(+1,+1,+1,-1)$	R $(\mathscr{S}_3)$	R $(\mathscr{S}_2)$	R $(\mathscr{S}_3)$	$(-1,+1,+1,-1)$	R $(\mathscr{S}_1,\mathscr{S}_3)$	-1	R $(\mathscr{S}_3)$
$(+1,+1,-1,+1)$	R $(\mathscr{S}_1)$	R $(\mathscr{S}_2)$	-1	$(-1,+1,-1,+1)$	-1	$+1$	$+1$
$(+1,+1,-1,-1)$	R $(\mathscr{S}_1,\mathscr{S}_3)$	-1	R $(\mathscr{S}_3)$	$(-1,+1,-1,-1)$	R $(\mathscr{S}_3)$	R $(\mathscr{S}_2)$	R $(\mathscr{S}_3)$
$(+1,-1,+1,+1)$	R $(\mathscr{S}_3)$	R $(\mathscr{S}_2)$	R $(\mathscr{S}_3)$	$(-1,-1,+1,+1)$	R $(\mathscr{S}_1,\mathscr{S}_3)$	-1	R $(\mathscr{S}_3)$
$(+1,-1,+1,-1)$	-1	$+1$	-1	$(-1,-1,+1,-1)$	R $(\mathscr{S}_1)$	R $(\mathscr{S}_2)$	$+1$
$(+1,-1,-1,+1)$	R $(\mathscr{S}_1,\mathscr{S}_3)$	-1	R $(\mathscr{S}_3)$	$(-1,-1,-1,+1)$	R $(\mathscr{S}_3)$	R $(\mathscr{S}_2)$	R $(\mathscr{S}_3)$
$(+1,-1,-1,-1)$	R $(\mathscr{S}_1)$	R $(\mathscr{S}_2)$	$+1$	$(-1,-1,-1,-1)$	$+1$	$+1$	-1

Table 6.3: **Logical assignments from data-qubit measurement outcomes**. When measuring in the specified logical cardinal basis (as shown in Fig. 6.2a), the final string of data-qubit outcomes is rejected (R) whenever at least one of the computable stabilizers has value -1 (indicated within parentheses). The computable stabilizers are $\mathscr{S}_1$ and $\mathscr{S}_3$ when measuring Z_L, $\mathscr{S}_2$ when measuring X_L, and $\mathscr{S}_3$ when measuring Y_L. When the data-qubit outcome string is accepted, the logical assignment (+1 or -1) is given by the appropriate product of data-qubit outcomes: $m_{D1} \times m_{D2}$ for Z_L, $m_{D1} \times m_{D3}$ for X_L, and $m_{D1} \times m_{D2} \times m_{D3}$ for Y_L.

The key experimental input needed to proceed is the data-qubit assignment probability matrix A, shown in Fig. 6.12. Each element of this 16×16 matrix gives the experimental probability of measuring a string of data outcomes $(m_{D1}, m_{D2}, m_{D3}, m_{D4})$, $m_{Di} \in \{-1, 1\}$ (varying across rows) when performing simultaneous readout of the data qubits having prepared them in physical computational state $|n_{D1} n_{D2} n_{D3} n_{D4}\rangle$, $n_{Di} \in \{0, 1\}$ (varying across columns). These computational states are prepared by applying the needed parallel combination of R_{ix}^{π} pulses on data qubits starting with all qubits (including ancillas) initialized in $|0\rangle$.
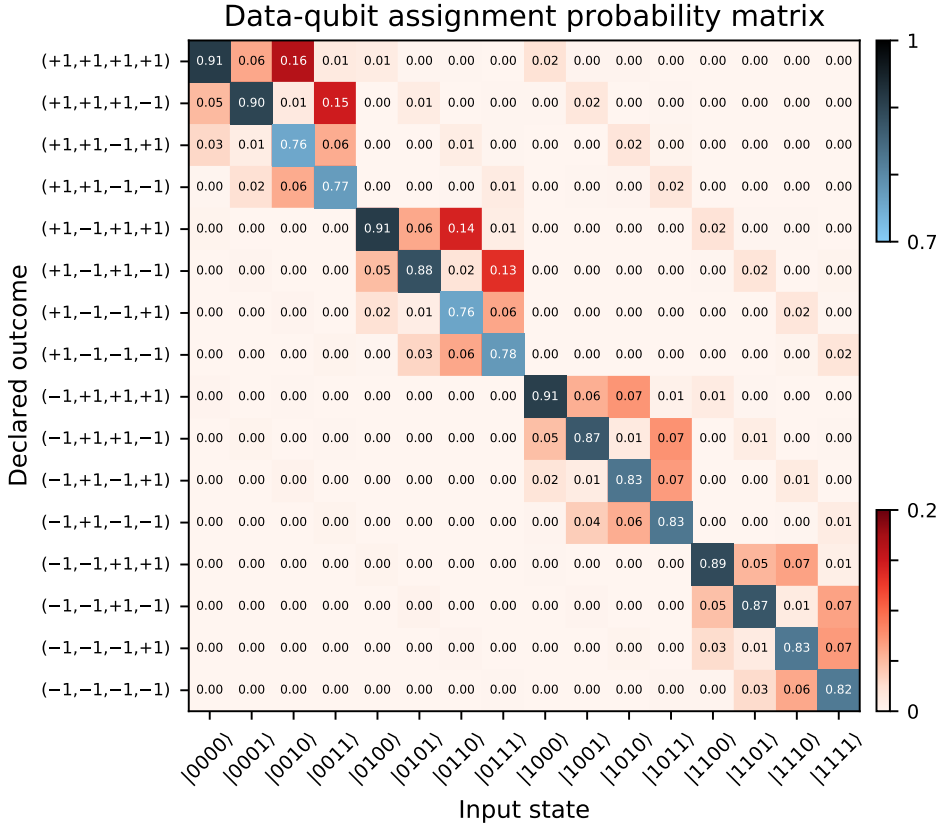
Figure 6.12: **Experimental data-qubit assignment probability matrix.** Each element of A gives the experimental probability of measuring outcome string $(m_{D1}, m_{D2}, m_{D3}, m_{D4})$ (varying across rows) when performing simultaneous measurement of the data qubits prepared in $|n_{D1} n_{D2} n_{D3} n_{D4}\rangle$, $n_{Di} \in \{0, 1\}$ (varying across columns).

With A in hand, it is straightforward to compute the probabilities for all strings of data-qubit outcomes for each choice of O_L and $|\Psi_{L\pm}\rangle$. This is given by $A\vec{p}$, where $\vec{p}$ is vector (size 16) whose elements are the probabilities (in the physical data-qubit computational basis) of the corresponding state in Eqs. 6.27-6.32. For example, $\vec{p} = (1/2, 0, \ldots, 0, 1/2)^T$ for Z_L and $|0_L\rangle$, and $\vec{p} = (1/4, 0, 0, 0, 0, 1/4, 0, 0, 0, 0, 1/4, 0, 0, 0, 0, 1/4)^T$ for X_L and $|+_L\rangle$. From

Logical measurement basis O_L	Z_L	X_L	Y_L
$|\Psi_{\text{L}+}\rangle$	$|0_\text{L}\rangle$	$|+_\text{L}\rangle$	$|+i_\text{L}\rangle$
$|\Psi_{\text{L}-}\rangle$	$|1_\text{L}\rangle$	$|-_\text{L}\rangle$	$|-i_\text{L}\rangle$
$p_{\text{R}+}$	0.184	0.164	0.078
$p_{\text{R}-}$	0.170	0.118	0.073
$\epsilon_{\text{L}+}$	0.003	0.016	0.089
$\epsilon_{\text{L}-}$	0.002	0.010	0.083
F_L^R	0.998	0.987	0.914

Table 6.4: **Quantified performance of logical measurement.** Final results of the analysis performed to quantify logical measurement in the logical cardinal bases without corruption from initialization errors. See Fig. 6.11 for reference. The extracted logical readout fidelities are those quoted in the main text.

$A\vec{p}$ and the rejection and logical assignment rules in Table 6.3, it is straightforward to compute all the probabilities within the dashed box of Fig. 6.11. The final results are presented in Table 6.4. The key assumption behind this analysis is that errors induced by single-qubit gates (both during preparation of the physical data-qubit computational states needed for determination of A and the measurement pre-rotations when performing logical measurement in X_L and Y_L) are small compared to the errors induced by data-qubit readout. This assumption is safe given the performance metrics summarized in Table 6.2.

6.5.8. NUMERICAL ANALYSIS

LEAKAGE IN EXPERIMENT

We observe a clear signature of leakage accumulation with the increasing number of error-detection cycles in the single-shot readout histograms obtained at the end of each experiment. In Fig. 6.13 we show examples of this accumulation for D_2, D_3 and A_3 at cycles $n = 1$, $n = 8$ and $n = 15$. For dispersive readout, a transmon in state $|2\rangle$ induces a different frequency shift in the readout resonator compared to state $|0\rangle$ or $|1\rangle$. The increased number of data points at $n = 8$ and $n = 15$ shown in Fig. 6.13, following a Gaussian distribution with a mean and standard deviation different from those observed at $n = 1$ is thus a clear manifestation of leakage to the higher-excited states (mostly to $|2\rangle$). We believe that the dominant source of leakage in our processor are the CZ gates [37, 45]. However, the leakage rate L_1 for each gate has not been experimentally characterized, e.g., by performing leakage-modified randomized benchmarking experiments [46, 47]. This is because our CZ tune-up procedure is performed in a parity-check block unit. This maximizes the performance of the parity-check but makes the gate unfit for randomized benchmarking protocols. We can estimate the population $p^{\mathscr{L}}(n)$ in the leakage subspace $\mathscr{L}$ at cycle n from the single-shot readout histograms. We perform a fit of a triple Gaussian model to the histograms from which we extract the voltage that allows for the best discrimination of $|2\rangle$ from $|1\rangle$ and $|0\rangle$. The leaked population $p^{\mathscr{L}}(n)$ is then given by the fraction of shots declared as $|2\rangle$ over the total number of shots. Assuming that leakage is only induced by the CZ gates (on the transmon being fluxed to perform
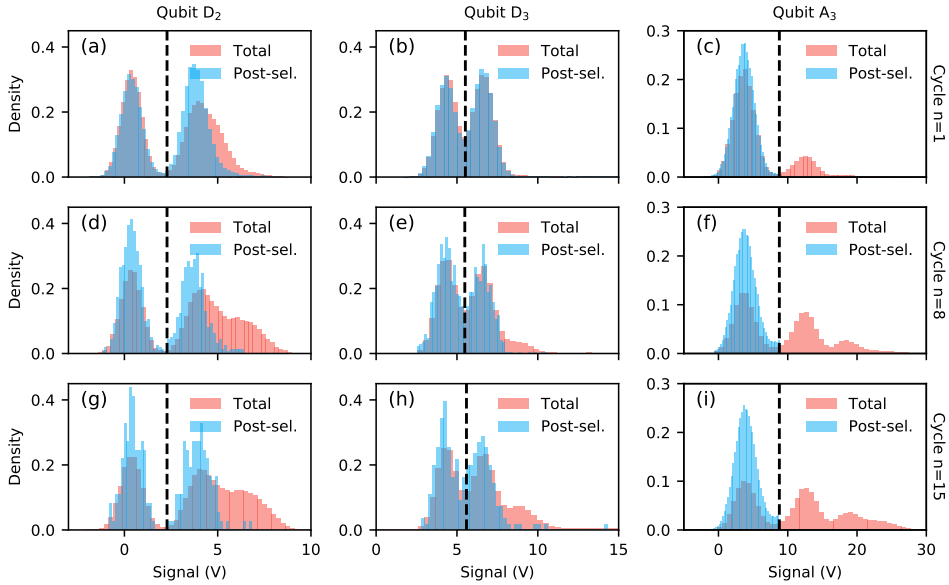
Figure 6.13: **Signature of transmon leakage in experimental data.** Single-shot readout histograms obtained at cycle n over all shots (red) and the post-selected shots based on detecting no error in any cycles up to n (blue) for D_2 (left), D_3 (middle) and A_3 (right) and at cycle $n = 1$ (top row), $n = 8$ (middle row) and $n = 15$ (bottom row). The dashed black lines indicate the thresholds used to discriminate $|0\rangle$ from $|1\rangle$.

the gate) and that each CZ gate has the same leakage rate L_1, we can use the Markovian model presented in Ref. [31] to estimate the L_1 value leading to the observed population $p^{\mathscr{L}}(n)$. This analysis gives a L_1 estimate in the approximate range $1 - 4\%$ for most transmons. However, we do not consider these estimates to be accurate due to the low fidelity with which $|2\rangle$ can be distinguished from $|1\rangle$ and instead treat L_1 as a free parameter in our simulations (see below).

The histograms of the post-selected shots in Fig. 6.13 demonstrate that post-selection rejects runs where leakage on those transmons occurred. Thus, while leakage may considerably impact the error-detection rate in the experiment [31], we do not expect it to significantly affect the fidelity of the logical initialization, and gates.

DENSITY-MATRIX SIMULATIONS

We perform numerical density-matrix simulations using the *quantumsim* package [48] to study the impact of the expected error sources on the performance of the code. We focus on repetitive error detection using the pipelined scheme and with the logical qubit initialized in $|0_L\rangle$. In Fig. 6.14a, we show the post-selected fraction $P(n)$ as a function of the number n of error-detection cycles for a series of models. Model 0 is a no-error model, which we take as the starting point of the comparison. Model 1 adds amplitude and phase damping experienced by the transmon. Model 2 adds the increased dephasing away from the sweetspot arising from flux noise. Model 3 adds residual qubit excitation and readout (SPAM) errors. Finally, Model 4 adds crosstalk due to the residual
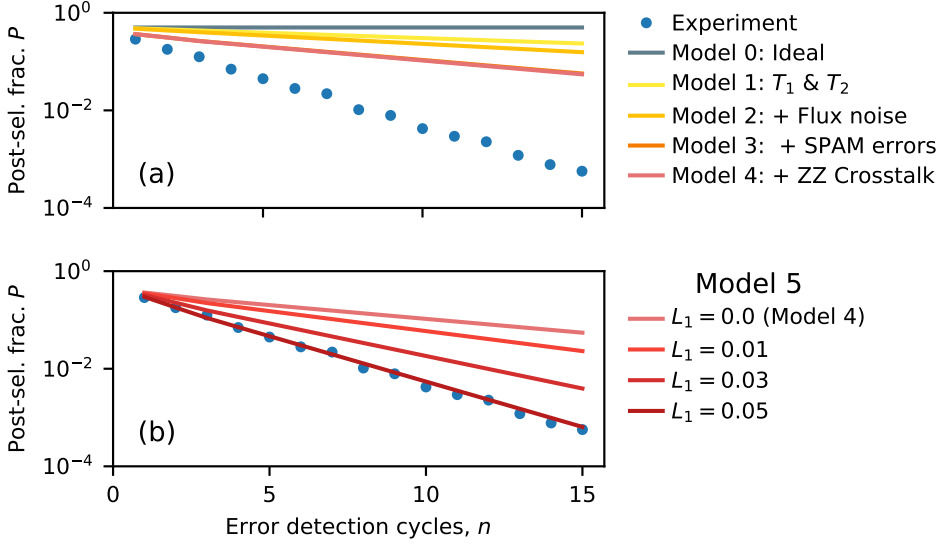
Figure 6.14: **Simulation of error-detection rate.** Post-selected fraction P as a function of the number n of error-detection cycles for $|0_L\rangle$. The experimental P (blue dots) is compared to numerical simulation under various models (solid curves). (a) Simulated P obtained by incremental addition of error sources starting from the no-error (Model 0, gray); qubit relaxation and dephasing (Model 1, yellow); extra dephasing due to flux noise away from the sweetspot (Model 2, amber); state preparation and measurement errors (Model 3, orange); and crosstalk due to residual ZZ interactions (Model 4, red). (b) Simulated P for Model 5 adding CZ gate leakage with 4 different values of L_1, the leakage per CZ gate, assumed equal for all CZ gates.

ZZ coupling during the coherent operations of the stabilizer measurement circuits. The details of each model and their input parameters drawn from experiment are detailed below. We find that the dominant contributors to the error-detection rate are SPAM errors and decoherence. However, we also observe that the noise sources included through Model 4 clearly fail to quantitatively capture the decay of the post-selected fraction observed in experiment.

We believe that an important factor behind the observed discrepancy is the presence of leakage, as suggested by the single-shot readout histograms in Fig. 6.13. We consider the leakage per CZ gate L_1 as a free parameter and assume the same value for all CZ gates. We simulate the post-selected fraction for a range of L_1 values, shown in Fig. 6.14b. We observe that $L_1 \approx 5\%$ produces a good match with experiment, suggesting that leakage significantly impacts the error-detection rate observed. We perform a similar analysis now considering the logical measurement of Z_L experiment depicted in Fig. 6.2f which also finds similar agreement with experimental data (Fig. 6.15). This value of L_1 is significantly higher than achieved in Ref. [37], which used the same device. However, note that in this earlier experiment CZ gates were characterized while keeping all other qubits in $|0\rangle$. Spectator transmons with residual ZZ coupling to either of the transmons involved in a CZ gate can increase L_1 when not in $|0\rangle$ (which is certainly the case in the present experiment). Note that leakage may also be further induced by the measurement [49],
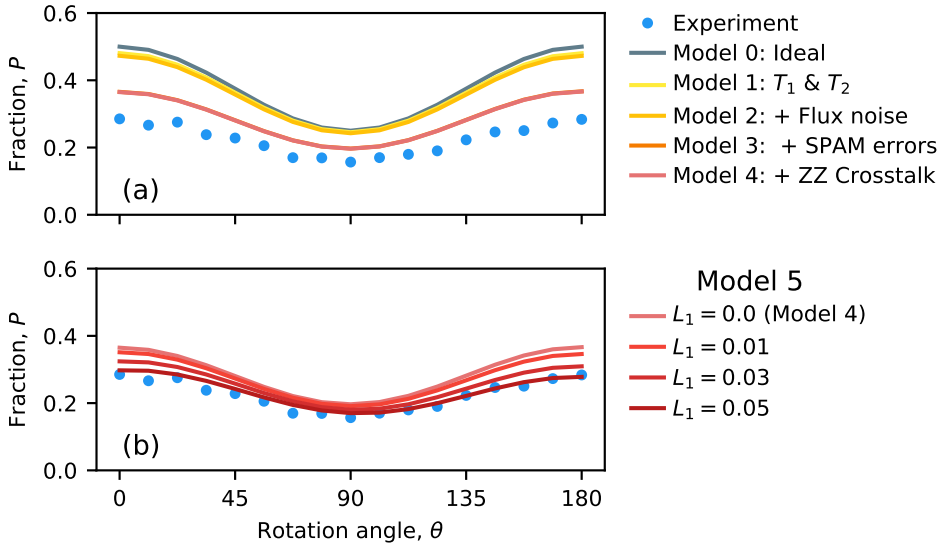
Figure 6.15: **Simulated post-selected fraction.** Post-selected fraction P of Fig. 6.2f for the Z_L measurement with the same error models used in Fig. 6.14.

an effect that we do not consider in our simulation. However, the assumption that all CZ gates have the same L_1, the approximations used in our models, and other error sources that we have not considered here may lead to an overestimation of the true L_1.

Leakage is an important error source to consider in quantum error correction experiments of larger distance codes, requiring either post-selection based on detection [31] or the use of leakage reduction units [33]. We leave the detailed investigation of the exact leakage rates in our experiment and the mechanisms leading to them to future work.

6.5.9. ERROR MODELS
Lastly, we detail the error models used in the numerical simulations in Fig. 6.14.

MODEL 1
We take into account transmons decoherence by including an amplitude-damping channel parameterized by the measured relaxation time T_1 and a phase-damping channel parameterized by the pure-dephasing time at the sweetspot

$$\frac{1}{T_\phi^{\text{max}}} = \frac{1}{T_2} - \frac{1}{2T_1},$$

where T_2 is the measured echo dephasing time (see Table 6.2). The qutrit Kraus operators defining these channels are detailed in Ref. [31] and we similarly introduce these channels during idling periods and symmetrically around each single-qubit or two-qubit gate (each period lasting half the duration of the gate).

MODEL 2

We consider the pure-dephasing rate $1/T_\phi = 2\pi\sqrt{\ln 2A}D_\phi + 1/T_\phi^{\max}$ away from the sweetspot due to the fast-frequency components of the $1/f$ flux noise, where D_ϕ is the flux sensitivity at a given qubit frequency and A is the scaling parameter for the flux-noise spectral density. We use a $\sqrt{A} \approx 3\ \mu\Phi_0$, the average of the extracted $\sqrt{A}$ values for D_3, A_1 and A_2 obtained by fitting the measured decrease of T_2 as a function of the applied flux bias, following the model described above. This allows us to estimate the dephasing time at the CZ interaction and parking frequencies, which then parameterize the applied amplitude-phase damping channel inserted during those operations [31]. We neglect the slow-frequency components of the flux noise due to the use of sudden Net Zero pulses, which echo out this noise to first order [37, 45].

MODEL 3

We further include state-preparation and measurement errors. We consider residual qubit excitations, where instead of the transmon being initialized in $|0\rangle$ at the start of the experiment, it is instead excited to $|1\rangle$ with probability p_e. We extract p_e for each transmon from a double-Gaussian fit to the histogram of the single-shot readout voltages with the transmon nominally initialized in $|0\rangle$ [43]. We model measurement errors via the POVM operators $M_i = \sum_{j=0}^{2} \sqrt{P(i|j)}|j\rangle\langle j|$ for $i \in 0,1,2$ being the measurement outcome, while $P(i|j)$ is the probability of measuring the qubit in state $|i\rangle$ when having prepared state $|j\rangle$. We extract the probability $P(Q = |i\rangle) = \text{Tr}\left(M_i^\dagger M_i \rho\right)$ of measuring qubit Q in state $|i\rangle$ from simulation, where ρ is the density matrix, while application of the POVM transforms $\rho \to M_i \rho M_i^\dagger / P(Q = |i\rangle)$. In our simulations we condition on the detection of no error and thus we calculate $P(Q = |0\rangle)$ and then apply M_0 to the state ρ. We obtain $P(0|j)$ for $j \in 0,1$ from the experimental assignment fidelity matrix [35] (where a heralded initialization protocol was used to prepare the qubits in $|0\rangle$ [42]) and we assume $P(0|2) = 0$, consistent with the observed histograms in Fig. 6.13. At the end of each experiment with n error-detection cycles we calculate the probability P_n^f of obtaining trivial syndromes from the final measurements of the data qubits (see Sec. 6.2.5). From this and from the probability $P_n(A_i = |0\rangle)$ of measuring ancilla A_i in $|0\rangle$ at cycle n, we calculate the post-selected fraction of experiments defined as $P(n) = P_n^f \prod_n \prod_{i=1}^{3} P_n(A_i = |0\rangle)$.

MODEL 4

We consider the crosstalk due to residual ZZ interactions between transmons. The CZ gates involved in a parity check are jointly calibrated to minimize phase errors for the whole check as one block (see Fig. 6.6). Instead of modeling this crosstalk as an always-on interaction and taking into account the details of the check calibration, we instead capture the net effect of this noise by including it as single-qubit and two-qubit phase errors in each CZ gate. This assumes that the crosstalk only occurs between transmons that are directly coupled, which the measured frequency shifts observed in Fig. 6.5 validate. We characterize the phases picked up during the CZ gates using $k \times 2^{k-1}$ Ramsey experiments for a check involving a total of k transmons (including the ancilla). In each experiment, we perform a Ramsey experiment on one transmon labelled Q_k. Q_k is initialized in a maximal superposition using a $R_x^{-\pi/2}$ pulse, while the remaining $k-1$

transmons are prepared in each of the 2^{k-1} computational states $|l\rangle$. Following this initialization, the parity check is performed, followed by a rotation of $R_\phi^{-\pi/2}$ (while the other transmons are rotated back to $|0\rangle$) and by a measurement of Q_k. By varying the axis of rotation ϕ, we extract the phase $\phi_{\text{Ram}}^k (l)$ picked up by Q_k with the remaining transmons in state $|l\rangle$. We perform this procedure for each of the k transmons of the check, resulting in a total of $k \times 2^{k-1}$ measured phases, which are arranged in a column vector $\vec{\phi}^{\text{Ram}}$. We parameterize each CZ gate used in the parity check by a matrix $\text{diag}\left(1, e^{i\phi_{01}}, e^{i\phi_{10}}, e^{i\phi_{11}}\right)$. The column vector $\vec{\phi}^{\text{CZ}}$ then contains all of the phases parameterizing each of the $k-1$ CZ gates involved in the parity checks, with $k = 3$ for the $Z_{\text{D1}}Z_{\text{D3}}$ and $Z_{\text{D2}}Z_{\text{D4}}$ checks and $k = 5$ for the $Z_{\text{D1}}Z_{\text{D2}}Z_{\text{D3}}Z_{\text{D4}}$ check. We can express each of the measured phases in the Ramsey experiment as a linear combination of the acquired phases as a result of the CZ interactions between transmons, i.e., $\vec{\phi}^{\text{Ram}} = A\vec{\phi}^{\text{CZ}}$, where the matrix A encodes the linear dependence. Given the measured $\vec{\phi}^{\text{Ram}}$ we perform an optimization to find the closest $\vec{\phi}^{\text{CZ}}$ as given by

$$\min_{\vec{\phi}^{\text{CZ}}} \quad \sum_i \left(\sum_j A_{ij}\vec{\phi}_j^{\text{CZ}} - \vec{\phi}_i^{\text{Ram}}\right)^2,$$

$$\text{subject to} \quad 0 \leq \vec{\phi}_j^{\text{CZ}} < 2\pi.$$

The optimal $\vec{\phi}^{\text{CZ}}$ then captures the net effect of the ZZ crosstalk during the parity checks, which we include in the simulation. We do not model phase errors accrued during the ancilla readout, since in our simulation we condition on each ancilla being measured in $|0\rangle$.

MODEL 5
We model leakage due to CZ gates following the model and numerical implementation presented in Ref. [31]. Here, we do not consider the phases picked up when non-leaked transmons interact with leaked ones (the leakage-conditional phases [31]) and we set them to their ideal values. We also neglect higher-order leakage effects, such as excitation to higher-excited states or leakage mobility. Thus, we only consider the exchange of population between $|11\rangle$ and $|02\rangle$ given by $4L_1$, except for the CZ between A_1 and D_3, where the population is instead exchanged with $|20\rangle$ as we use the $|11\rangle$-$|20\rangle$ avoided crossing for this gate [37].

There remain several relevant error sources beyond those included in our numerical simulation. For example, we do not include dephasing of data or other ancilla qubits induced by ancilla measurement, which we expect to be a relevant error source for comparing the performance of the pipelined and parallel schemes. Also, we only consider the net effect of crosstalk due to residual ZZ interactions during coherent operations of the parity-check circuits, which we include via errors in the single-qubit and two-qubit phases in the CZ gates. Thus, we do not capture the crosstalk present whenever an ancilla is projected to state $|1\rangle$ by the readout but declared to be in $|0\rangle$ instead. Furthermore, as ZZ crosstalk does not commute with the amplitude damping included during the execution of the circuit, we are not capturing the increased phase error rate that this leads to.

6

BIBLIOGRAPHY

[1] 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*, Nat. Phys. **18**, 80–86 (2022).

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

[3] B. M. Terhal, *Quantum error correction for quantum memories*, Rev. Mod. Phys. **87**, 307–346 (2015).

[4] J. M. Martinis, *Qubit metrology for building a fault-tolerant quantum computer*, npj Quantum Inf. **1**, 15005 (2015).

[5] 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**, 66–69 (2015).

[6] 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**, 383–387 (2021).

[7] R. Lescanne, M. Villiers, T. Peronnin, A. Sarlette, M. Delbecq, B. Huard, T. Kontos, M. Mirrahimi, and Z. Leghtas, *Exponential suppression of bit-flips in a qubit encoded in an oscillator*, Nature Physics **16**, 509–513 (2020).

[8] A. Grimm, N. E. Frattini, S. Puri, S. O. Mundhada, S. Touzard, M. Mirrahimi, S. M. Girvin, S. Shankar, and M. H. Devoret, *Stabilization and operation of a kerr-cat qubit*, Nature **584**, 205–209 (2020).

[9] D. Ristè, S. Poletto, M. Z. Huang, A. Bruno, V. Vesterinen, O. P. Saira, and L. Di-Carlo, *Detecting bit-flip errors in a logical qubit using stabilizer measurements*, Nat. Commun. **6**, 6983 (2015).

[10] J. Cramer, N. Kalb, M. A. Rol, B. Hensen, M. S. Blok, M. Markham, D. J. Twitchen, R. Hanson, and T. H. Taminiau, *Repeated quantum error correction on a continuously encoded qubit by real-time feedback*, Nature Communications **7**, 11526 (2016).

[11] 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).

[12] D. Nigg, M. Müller, E. A. Martinez, P. Schindler, M. Hennrich, T. Monz, M. A. Martin-Delgado, and R. Blatt, *Quantum computations on a topologically encoded qubit*, Science **345**, 302–305 (2014).

[13] 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 control of an error-corrected qubit*, Nature **598**, 281–286 (2021).

[14] A. Erhard, H. Poulsen Nautrup, M. Meth, L. Postler, R. Stricker, M. Stadler, V. Negnevitsky, M. Ringbauer, P. Schindler, H. J. Briegel, R. Blatt, N. Friis, and T. Monz, *Entangling logical qubits with lattice surgery*, Nature **589**, 220–224 (2021).

[15] 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**, 527–531 (2018).

[16] A. Blais, R.-S. Huang, A. Wallraff, S. M. Girvin, and R. J. Schoelkopf, *Cavity quantum electrodynamics for superconducting electrical circuits: an architecture for quantum computation*, Phys. Rev. A **69**, 062320 (2004).

[17] 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**, 1–7 (2019).

[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**, eaay3050 (2020).

[19] N. Ofek, A. Petrenko, R. Heeres, P. Reinhold, Z. Leghtas, B. Vlastakis, Y. Liu, L. Frunzio, S. M. Girvin, L. Jiang, M. Mirrahimi, M. H. Devoret, and R. J. Schoelkopf, *Extending the lifetime of a quantum bit with error correction in superconducting circuits*, Nature **536**, 441 (2016).

[20] L. Hu, Y. Ma, W. Cai, X. Mu, Y. Xu, W. Wang, Y. Wu, H. Wang, Y. P. Song, C. L. Zou, S. M. Girvin, L.-M. Duan, and L. Sun, *Quantum error correction and universal gate set operation on a binomial bosonic logical qubit*, Nature Physics **15**, 503–508 (2019).

6

[21] P. Campagne-Ibarcq, A. Eickbusch, S. Touzard, E. Zalys-Geller, N. E. Frattini, V. V. Sivak, P. Reinhold, S. Puri, S. Shankar, R. J. Schoelkopf, L. Frunzio, M. Mirrahimi, and M. H. Devoret, *Quantum error correction of a qubit encoded in grid states of an oscillator*, Nature **584**, 368–372 (2020).

[22] 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).

[23] 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**, 875–880 (2020).

[24] 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).

[25] 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. Applied **8**, 034021 (2017).

[26] Y. Tomita and K. M. Svore, *Low-distance surface codes under realistic quantum noise*, Phys. Rev. A **90**, 062320 (2014).

[27] O.-P. Saira, J. P. Groen, J. Cramer, M. Meretska, G. de Lange, and L. DiCarlo, *Entanglement genesis by ancilla-based parity measurement in 2D circuit QED*, Phys. Rev. Lett. **112**, 070502 (2014).

[28] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, et al., *Demonstration of weight-four parity measurements in the surface code architecture*, Phys. Rev. Lett. **117**, 210505 (2016).

[29] P. Aliferis, D. Gottesman, and J. Preskill, *Quantum accuracy threshold for concatenated distance-3 codes*, Quantum Info. Comput. **6**, 97–165 (2006).

[30] A. Paetznick and K. M. Svore, *Repeat-until-success: non-deterministic decomposition of single-qubit unitaries*, Quantum Information & Computation **14**, 1277–1301 (2014).

[31] 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**, 102 (2020).

[32] 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**, 1761 (2021).

6

[33] 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).

[34] K. J. Satzinger, Y.-J. Liu, A. Smith, C. Knapp, M. Newman, C. Jones, Z. Chen, C. Quintana, X. Mi, A. Dunsworth, C. Gidney, I. Aleiner, F. Arute, K. Arya, J. Atalaya, R. Babbush, J. C. Bardin, R. Barends, J. Basso, A. Bengtsson, A. Bilmes, M. Broughton, B. B. Buckley, D. A. Buell, B. Burkett, N. Bushnell, B. Chiaro, R. Collins, W. Courtney, S. Demura, A. R. Derk, D. Eppens, C. Erickson, L. Faoro, E. Farhi, A. G. Fowler, B. Foxen, M. Giustina, A. Greene, J. A. Gross, M. P. Harrigan, S. D. Harrington, J. Hilton, S. Hong, T. Huang, W. J. Huggins, L. B. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, T. Khattar, S. Kim, P. V. Klimov, A. N. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, A. Locharla, E. Lucero, O. Martin, J. R. McClean, M. McEwen, K. C. Miao, M. Mohseni, S. Montazeri, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, T. E. O'Brien, A. Opremcak, B. Pató, A. Petukhov, N. C. Rubin, D. Sank, V. Shvarts, D. Strain, M. Szalay, B. Villalonga, T. C. White, Z. Yao, P. Yeh, J. Yoo, A. Zalcman, H. Neven, S. Boixo, A. Megrant, Y. Chen, J. Kelly, V. Smelyanskiy, A. Kitaev, M. Knap, F. Pollmann, and P. Roushan, *Realizing topologically ordered states on a quantum processor*, Science **374**, 1237–1241 (2021).

[35] 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).

[36] 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).

[37] 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).

[38] G. Aleksandrowicz, T. Alexander, P. Barkoutsos, L. Bello, Y. Ben-Haim, D. Bucher, F. J. Cabrera-Hernández, J. Carballo-Franquis, A. Chen, C.-F. Chen, J. M. Chow, A. D. Córcoles-Gonzales, A. J. Cross, A. Cross, J. Cruz-Benito, C. Culver, S. D. L. P. González, E. D. L. Torre, D. Ding, E. Dumitrescu, I. Duran, P. Eendebak, M. Everitt, I. F. Sertage, A. Frisch, A. Fuhrer, J. Gambetta, B. G. Gago, J. Gomez-Mosquera, D. Greenberg, I. Hamamura, V. Havlicek, J. Hellmers, Ł. Herok, H. Horii, S. Hu, T. Imamichi, T. Itoko, A. Javadi-Abhari, N. Kanazawa, A. Karazeev, K. Krsulich, P. Liu, Y. Luh, Y. Maeng, M. Marques, F. J. Martín-Fernández, D. T. McClure, D. McKay, S. Meesala, A. Mezzacapo, N. Moll, D. M. Rodríguez, G. Nannicini, P. Nation, P. Ollitrault, L. J. O'Riordan, H. Paik, J. Pérez, A. Phan, M. Pistoia, V. Prutyanov, M. Reuter, J. Rice, A. R. Davila, R. H. P. Rudy, M. Ryu, N. Sathaye, C. Schnabel, E. Schoute, K. Setia, Y. Shi, A. Silva, Y. Siraichi, S. Sivarajah, J. A. Smolin, M. Soeken, H. Takahashi, I. Tavernelli, C. Taylor, P. Taylour, K. Trabing, M. Treinish, W. Turner,

D. Vogt-Lee, C. Vuillot, J. A. Wildstrom, J. Wilson, E. Winston, C. Wood, S. Wood, S. Wörner, I. Y. Akhalwaya, and C. Zoufal, *Qiskit: An Open-source Framework for Quantum Computing*, Feb. 2019.

[39] J. de Jong, *Implementation of a fault-tolerant swap operation on the ibm 5-qubit device* (Technical University of Delft, 2019).

[40] 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).

[41] 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).

[42] D. Ristè, J. G. van Leeuwen, H.-S. Ku, K. W. Lehnert, and L. DiCarlo, *Initialization by measurement of a superconducting quantum bit circuit*, Phys. Rev. Lett. **109**, 050507 (2012).

[43] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Pechal, M. Mondal, M. Oppliger, C. Eichler, and A. Wallraff, *Rapid High-Fidelity Single-Shot Dispersive Readout of Superconducting Qubits*, Phys. Rev. App. **7**, 054020 (2017).

[44] R. Sagastizabal, S. P. Premaratne, B. A. Klaver, M. A. Rol, V. Negîrneac, M. S. Moreira, X. Zou, S. Johri, N. Muthusubramanian, M. Beekman, C. Zachariadis, V. P. Ostroukh, N. Haider, A. Bruno, A. Y. Matsuura, and L. DiCarlo, *Variational preparation of finite-temperature states on a quantum computer*, npj Quantum Inf. **7**, 130 (2021).

[45] 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).

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

[47] 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).

[48] T. E. O'Brien, B. M. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Information **3** (2017).

[49] 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).

6

7

ERROR SUPPRESSION WITH A TRANSMON-BASED REPETITION CODE

Quantum error correction enables the suppression of the logical error rate by increasing the code distance d, an essential ingredient for realizing fault-tolerant computation. In this chapter, we implement a d = 7 repetition code experiment and perform up to 15 rounds of error correction. We show that by increasing the distance from d = 3 to d = 7, we are able to decrease the logical error rate per round, albeit at a low rate. We perform simulations to show that the logical performance is mainly limited by the qubit relaxation times and measurement error rates. By analyzing the correlations between the measured syndrome defects, we characterize the errors observed in the experiment and observe signatures of crosstalk and leakage errors, which lead to an increase in the defect rate throughout the experiment and a higher logical error rate.

7.1. INTRODUCTION

Quantum error correction (QEC) is considered essential for the practical realization of fault-tolerant quantum computation [1–9], which is required for any potentially useful application of a quantum computer [10–12]. Two-dimensional stabilizer codes [13], such as the surface codes [10, 14], have emerged as one of the promising frameworks for implementing QEC on current quantum processors [15–29]. In this framework, the quantum information is encoded into a logical qubit consisting of many physical qubits, referred to as the data qubits. Ancilla qubits perform repeated measurements of the parity of their respective neighboring data qubits. These projective measurements ideally discretize the decoherence experienced by the physical qubits and provide information about any errors that may have occurred without disturbing the encoded information [9]. A classical decoder then attempts to infer the most likely physical or logical corrections based on the observed measurement outcomes. Ultimately, the logical error rate is a function of the physical error rates, the choice of stabilizer code, and the decoding algorithm used to infer the corrections. Minimum-weight perfect matching (MWPM) decoding has emerged as a promising decoding method for several codes, such as the repetition code and the surface code, due to the overall good logical performance it can achieve and its computational efficiency [30–34]. Provided that the physical error rate is below a certain threshold, the logical error rate can be exponentially suppressed by increasing the code distance [1, 5, 8, 10].

The repetition code is the simplest stabilizer code that holds fundamental importance and illustrates the principles behind error correction. Implementing a distance-d repetition code requires only $2d-1$ physical qubits arranged in a linear chain, consisting of d data qubits and $d-1$ ancilla qubits interspersed between them, with each one measuring the parity of the two neighboring data qubits. This makes large-distance repetition code experiments substantially easier to implement on near-term hardware, where only a limited number of qubits are available. The high threshold of around 3% for a circuit-level depolarizing noise model [35] that this code exhibits allows for good logical performance to be achieved even when the physical error rates are relatively high, as is typically the case in near-term experiments. The simplicity of the repetition code comes at the expense that it can correct either only bit-flip or phase-flip errors, making it inadequate for realizing fault tolerance. Despite this, repetition code implementations can still be used to benchmark the ability of a quantum processor to perform repeated parity-check measurements, especially as this will also explore the impact of error sources not typically considered by textbook error corrections, such as leakage, crosstalk or other non-Markovian noise [27, 36]. So far, many small-distance repetition code experiments have been realized in various platforms to demonstrate the ability to perform parity measurements and correct errors [22, 37–43]. On the other hand, repetition codes of distances as large as $d = 30$ have been realized in both trapped-ion [21] and superconducting-qubit processors [15, 19, 27, 36, 44], demonstrating the ability to suppress errors by scaling the code distance. Such experiments are often a direct precursor to more complex QEC experiments using stabilizer codes that can correct both bit-flip and phase-flip errors, such as the surface code [17, 19].

In this chapter, we implement a distance-7 repetition code in a 17-qubit transmon processor and perform repeated parity measurements. We use an MWPM decoder to

process the collected measurements and correct the bit-flip errors that have occurred during the experiment. By subsampling the measurement outcomes collected in the experiment, we extract the performance of three distance-5 and five distance-3 codes possible within our linear chain layout. Averaging the logical performance of each of the distance-d codes, we show that the logical error rate is exponentially suppressed with the code distance d, albeit with a low suppression factor due to the high physical error rates. We infer these error rates directly from the correlations between the experimentally observed syndrome measurement outcomes. We observe a significant discrepancy between the estimated error probabilities and those predicted by a circuit-level Pauli-noise model based on the qubit relaxation and dephasing times, the single-qubit and two-qubit gate error rates measured by randomized benchmarking experiments, the assignment error probability, and the probability of qubit relaxation or excitation during readout. In particular, we find significant probabilities of errors that lead to pairs of defects in non-consecutive QEC rounds or measured by ancilla qubits that do not share any neighbors. These errors are a signature of non-conventional errors, such as leakage, that are not considered in our error model and further degrade the code's logical performance.

7.2. RESULTS

7.2.1. MEMORY EXPERIMENTS USING THE REPETITION CODE
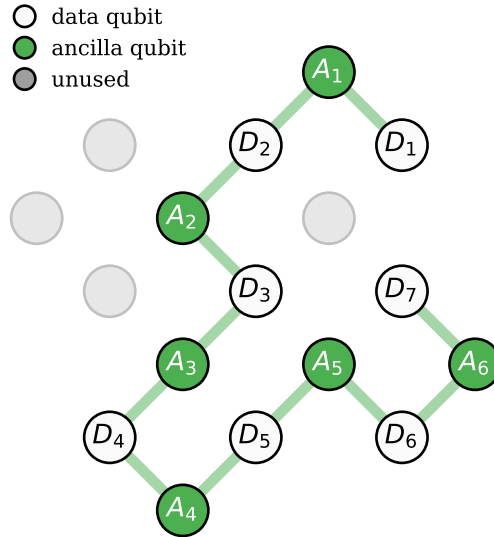
Figure 7.1: Schematic of a distance-7 repetition code embedded into a 17-transmon processor. The logical information is encoded into seven data qubits (white circles, labeled D_i). There are six ancilla qubits (green circles, labeled A_i) used to measure the parity of the neighboring data qubits, indicated by the green lines. Unused transmon qubits are shown as gray circles.

We implement a distance-7 repetition code embedded into a 17-transmon supercon-

ducting processor, with the layout of the code shown in Fig. 7.1. The logical information is encoded into a chain of 7 data qubits, which are labeled as D_i for $i = 1, 2, \ldots, 7$. In addition, interspersed between the data qubits are six ancilla qubits, labeled as A_i for $i = 1, 2, \ldots, 6$, each measuring the $Z_i Z_{i+1}$ parity operator on the neighboring data qubits. Additional information about the design of the processors is presented in Sec. 7.4.1. The logical states of the code are typically defined by the total parity of all data qubits and are typically defined as $|0\rangle_L = |0000000\rangle$ and $|1\rangle_L = |1111111\rangle$. Repeatedly measuring the parity operators allows bit-flip (X) errors on the data qubits to be detected and corrected.
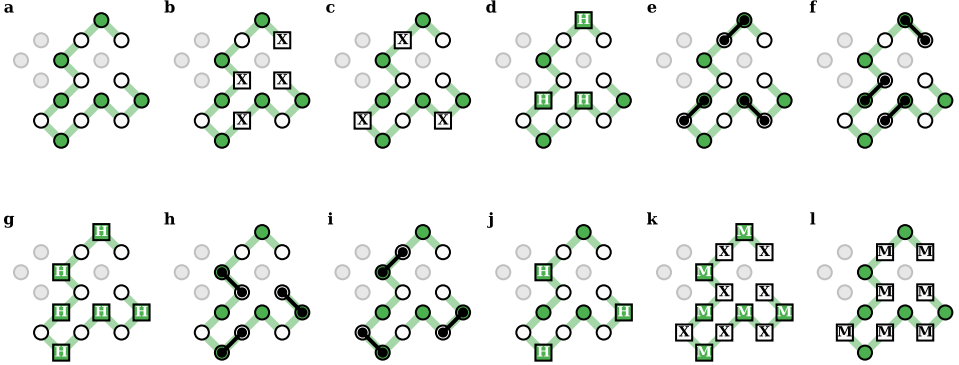
Figure 7.2: Schematic of the circuits used for the repetition-code memory experiments. At the start of each run, all the qubits are first prepared in their ground state (**a**) using heralded initialization. Several X gates are then applied to some of the data qubits in order to prepare either the $|0\rangle_L = |1010101\rangle$ (**b**) or $|1\rangle_L = |0101010\rangle$ (**c**) logical states. Each QEC cycle begins by putting some of the ancilla qubits into a superposition using three H (Hadamard) gates (**d**), after which a pair of CZ (controlled-phase) gates map the parity of the neighboring data qubit to the state of the corresponding ancilla qubit (**e** and **f**). In step (**g**), a set of H gates rotates these ancilla qubits back, while the remaining ancilla qubits that have been idling so far are instead put into a superposition. The parity of the neighboring qubits is similarly mapped to the state of these ancilla qubits using two CZ gates (**h** and **i**), after which these qubits are also rotated back by a set of H gates (**j**). The cycle ends with each ancilla qubit being measured simultaneously (**k**). At the same time, an X gate is applied to each data qubit halfway through the measurement period to dynamically decouple these qubits. This QEC round (**d-k**) is repeated several times. At the end of each experiment, all data qubits are measured on the Z-basis simultaneously, which also measures the value of the Z_L logical operator. For the definition of the qubits, see this Fig. 7.1.

In practice, however, the energy relaxation that transmon qubits experience during operations or periods of idling leads to a much higher probability of the qubit in the excited state $|1\rangle$ decaying to $|0\rangle$ as opposed to a qubit in the ground state $|0\rangle$ getting excited to $|1\rangle$. This results in the typically-chosen $|1\rangle_L$ to have a much higher logical error rate than $|0\rangle_L$ and for $|1\rangle_L$ to naturally tend to decay to $|0\rangle_L$ after many rounds of QEC. To symmetrize the logical error rates of the two logical states, we instead define the logical states as $|0\rangle_L = |1010101\rangle$ and $|1\rangle_L = |0101010\rangle$, such that each state has approximately the same number of excited qubits. Furthermore, at the end of each cycle, we apply an X gate on each data qubit to prevent data qubits from simply decaying to $|0\rangle$ after many QEC rounds. In each memory experiment, we first prepare one of the two logical states, after which we perform several rounds of parity check measurements. At the end of each experiment, we measure all data qubits, which allows us to extract a final set of parity outcomes that can be used in decoding and determining the total parity of the final

state, which we refer to as the logical observable. The full circuits that realize these experiments are shown in Fig. 7.2, while the device itself is described in Sec. 7.4.1. The logical states we have chosen will lead to odd parity measurement outcomes in the absence of errors, leading to the state of the ancilla qubits being flipped every QEC round. In comparison, even-parity states lead to the ancilla qubits typically remaining in their initially prepared ground states over the first several rounds. As a result, the ancilla qubits have a higher error rate due to the asymmetric nature of energy relaxation. To compensate for this, we include a virtual π-phase correction with the final Hadamard gates applied to each ancilla qubit, flipping the measured parity outcomes and reducing the error rates.

7.2.2. UNDERSTANDING THE DEVICE PERFORMANCE FROM THE SYNDROME DEFECTS
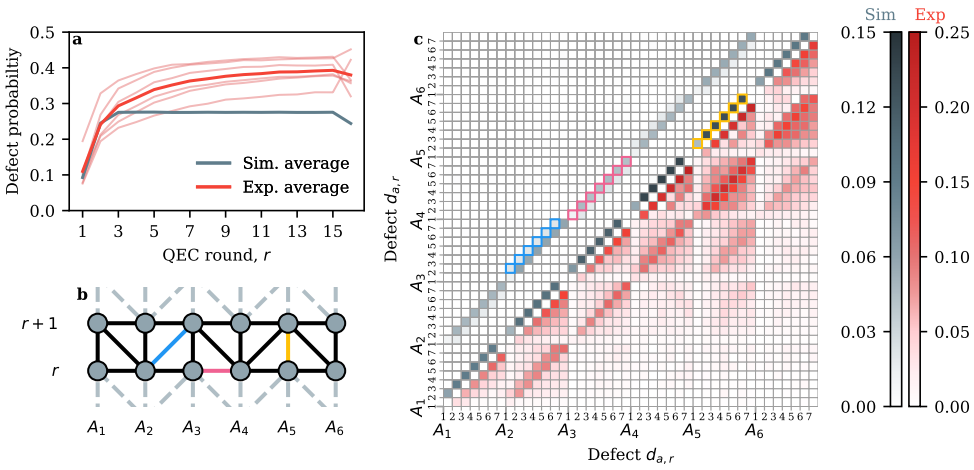
Figure 7.3: Syndrome defect analysis. **a** The probability of observing a non-trivial defect as a function of the number of QEC rounds for the $d = 7$ code, evaluated on simulated data (gray) and experimental data (red). The last QEC round $r = 16$ corresponds to syndrome defects calculated using the final data qubit measurement outcomes. The defect probability for each ancilla qubit extracted from the experimental data is also shown (light red). **b** Schematic of the decoding graph shown for $r = 2$ rounds. Each node (light gray point) in this graph corresponds to a syndrome defect being measured at some round r by ancilla qubit $a = A_1, \ldots, A_6$. Error events that lead to a pair of non-trivial defects are represented as edges (black lines) between the corresponding nodes. An example of a time-like (yellow line), a space-like (pink line), and a diagonal (blue line) error are highlighted. Edges connecting the shown nodes to nodes correspond to defects measured in the previous or the following QEC rounds are shown as dashed light gray lines. **c** The error probabilities p_m associated with each possible edge in the graph over the first $r = 7$ QEC rounds of the experiment, inferred from either simulated data (gray, upper triangle) or experimental data (red, lower triangle). The axes correspond to the measured defects $d_{a,r}$, with the minor ticks marking the rounds r while major ticks mark the ancilla qubits a. The probabilities associated with the error type examples shown in **b** are also highlighted in the upper triangle of this matrix.

The information about which errors have occurred is contained in the measurement outcomes $m_{a,r}$ of each ancilla qubit a at QEC round $r = 1, 2, \ldots, N_r$. However, these are typically first pre-processed to the syndrome defects, which contain the same amount of information but are more directly useful for decoding. In particular, non-trivial syn-

drome defects (sometimes referred to simply as defects) correspond to the points where the parity measurements change between consecutive QEC rounds, indicating the occurrence of an error.

One of the typical assumptions of textbook QEC is that the ancilla qubits are reset following each measurement. In this case, the measurement outcomes $m_{a,r}$ directly correspond to measurements of the stabilizers, also referred to as the syndromes $s_{a,r}$. However, in our experiment and many others, the ancilla qubits are not reset after the readout due to the difficulty of implementing a fast and high-fidelity qubit reset protocol, like the ones proposed in [45–47]. In this case, the syndromes are instead given by $s_{a,r} = m_{a,r-1} \oplus m_{a,r}$. The data qubit measurement outcomes $\{m_d\}$ obtained at the end of entire experiment can be used to extract a final set of syndromes $s_{a,N_r+1} = m_d \oplus m_{d+1}$, where d and $d+1$ correspond to the two data qubits for which ancilla a is measuring the parity. The syndrome defects are then given by the differences in the observed syndromes in consecutive rounds, i.e., $d_{a,r} = s_{a,r-1} \oplus s_{a,r}$. The values of $s_{a,0}$ correspond to the expected parity measurement outcomes given the prepared data qubit state, which is typically chosen such that $s_{a,0} = 0$. Similarly, $m_{a,0} = 0$ is taken when calculating $s_{a,1}$ when the ancilla qubits are not reset.

The physical error rates directly determine the probability of observing a non-trivial defect at each QEC round. Some other standard assumptions behind QEC experiments are that physical errors can be decomposed into a set of Pauli errors and that the corresponding error rates do not change over time. Under these assumptions, the non-trivial defect probability is expected to be some constant value at each QEC round, except for the first two rounds and the final one. The defects in the last round are calculated using the data-qubit measurement outcomes, while the defects in the first two rounds depend on the initial data-qubit state.

Note that when the ancilla qubits are reset, the average defect probability is instead expected to reach a constant value after only the first QEC round. In Fig. 7.3**a**, we show the average defect probabilities in our experiment over $r = 1, 2, \ldots, 16$ rounds of QEC. Aside from the expected rate changes in the first few and final rounds, we observe a relative increase in the probabilities of approximately 30% over the 13 QEC rounds, towards a steady-state value of approximately 0.4. This increase can be due to qubit leakage outside the computational subspace and the increase in the average leakage population over the course of the experiment. A similar increase in the defect probability has also been observed in other QEC experiments [15, 17, 19, 27, 48]. A leaked qubit can remain outside the computational subspace for several QEC rounds before decaying back. During this time, the leaked qubit can spread correlated errors, increasing the defect probability on its neighboring ancilla qubits [47–50]. Notably, a single leakage event can give rise to many correlated non-trivial defects and, therefore violate the assumption that errors lead to at most two defects that MWPM decoders rely on, which, in turn, can lead to a severe impact on the logical performance of the code [47, 49].

To further investigate the presence of such correlated errors in our data, we next consider the correlations between the defects observed in the measured defects, following closely the analysis done in [27]. We consider a circuit-level noise model, where each operation, preparation, or idling period is followed by any of the possible Pauli errors (sometimes referred to as a *fault* on the involved qubit(s) with some probability.

Note that this also includes two-qubit Pauli errors following two-qubit gates. Given this model, it can be shown that for the repetition code, any single error leads to either one (if the error involves a data qubit on the boundary of the code) or two non-trivial defects $d_{a,r} = 1$. These errors can be classified into several groups depending on which defects they lead to. Errors involving only an ancilla qubit will lead to a pair of non-trivial defects measured by the ancilla qubit in two consecutive QEC rounds, i.e., a $d_{a,r} = 1$ together with a $d_{a,r+1} = 1$, which we refer to as a *time-like* error. If ancilla qubits were reset after each measurement, these time-like errors would also include errors after the reset operations and classical readout errors, where the declared measurement outcome does not correspond to the measured state. However, when the ancilla qubits are not reset, the classical readout errors instead lead to defects that are two QEC rounds apart, namely $d_{a,r} = 1$ and $d_{a,r+2} = 1$. On the other hand, errors involving only a data qubit typically lead to a pair of non-trivial defects measured by the two neighboring ancilla qubit a and a' in the same QEC round, i.e., $d_{a,r} = 1$ together with a $d_{a',r} = 1$, which we refer to as *space-like* errors. Exceptions to this rule are certain errors involving a data qubit on the code's boundary that lead only to a single non-trivial defect measured by the only neighboring ancilla qubit of that data qubit, which we refer to as a *boundary* error. Other exceptions are the errors involving data qubit in the bulk of the code that happen between the two controlled-phase (CZ) gates used to map the parity of the data qubit when measuring the checks involving that qubit. These errors instead lead to non-trivial defects measured by the neighboring ancilla qubits in consecutive rounds, $d_{a,r} = 1$ and a $d_{a',r+1} = 1$, which we refer to as *diagonal* errors. Errors following each CZ gate in our stabilizer measurement circuit can lead to a space-like, time-like, boundary-like, or diagonal error.

These single errors can be represented as edges in an undirected graph, where the nodes correspond to the possible syndrome defects measured in each experiment. We refer to this graph as the *decoding graph*, and it is illustrated for our repetition code experiment in Fig. 7.3**b**. For a precise definition of this graph, see Sec. 7.4.3. Note that several faults in the circuit may lead to the same pair of non-trivial defects and correspond to the same edge in this graph. Each edge h_m can then be associated with a total probability p_m for the errors that lead to the defects at the endpoints of this edge, where m is the index of this edge. The fact that the possible errors lead to, at most, two defects is important in the context of decoding. Namely, it enables mapping the problem of inferring the most likely error consistent with the observed syndromes to the problem of minimum-weight perfect matching on a graph [30–32].

The total physical error probabilities can be directly inferred from the average correlations between the observed defects in each repetition of the experiment, as derived in [27, 28, 51]. Given the error model discussed above, the probability p_m associated with an edge $h_m = \{v_i, v_j\}$, where the nodes correspond to defects d_i and d_j, is given by

$$p_m = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\langle d_i d_j \rangle - \langle d_i \rangle \langle d_j \rangle}{1 - 2\langle d_i \rangle - 2\langle d_j \rangle + 4\langle d_i d_j \rangle}}, \tag{7.1}$$

assuming that $p_m < 0.5$ and where each syndrome defect is labeled by a single index $i = (a, r)$ and $j = (a', r')$. For a full derivation of this equation, see Sec. 7.4.4. This equation

can be used to estimate the error probability associated with any edge except for those corresponding to boundary errors. Instead, the probability of these is estimated using

$$p_m = \frac{\langle d_i \rangle - q_m^i}{1 - 2q_m^i},$$ (7.2)

where q_m^i is the total probability of observing an odd number of errors that together lead to $d_i = 1$, excluding the boundary error associated with h_m itself. Calculating q_m^i requires first obtaining the probabilities of all non-boundary errors that lead to a defect on d_i as well. For a full derivation of these equations, please see Sec. 7.4.5.

In Fig. 7.3c, we show the matrix of estimated error probabilities between any two syndrome defects from both the experimental data (bottom triangle) and from simulated data (upper triangle) using the circuit-level noise models used, see Sec. 7.4.2. Note that the decoding graph is undirected, meaning that all the edge probabilities p_m extracted from each dataset are fully contained in the corresponding triangle. In the simulation, given our error model, we observe the exact types of errors that we expected in the first place. These probabilities associated with these errors appear along diagonal lines in the matrix, as illustrated in Fig. 7.3c. Note that this matrix does not include any of the boundary error probabilities, which we discuss in more detail in the following section. We can see that the most prominent errors, according to the constructed error model, are the time-like errors corresponding to ancilla-qubit and measurement errors, see Tab. 7.1 and Tab. 7.2 for the exact parameters used in our simulations. The probabilities estimated from the experimental data reach much higher values by the end of the experiment, consistent with the observed increase in the average defect probabilities over the QEC rounds. While time-like errors remain the most prominent type, we observe that many errors that are not expected from the circuit-level noise model are being estimated to occur with a relatively high probability. Some of these include errors leading to non-trivial defects on non-adjacent ancilla qubits, suggesting the presence of crosstalk errors. On the other hand, we also observe errors leading to non-trivial defects measured by the same ancilla qubit or by adjacent qubits that occur more than two QEC rounds apart, which can be attributed to leakage. The defects that a leaked data qubit can propagate depend on the phase error propagated from the interaction with the ancilla qubits in the computational subspace, referred to as the leakage-conditional phase errors in [49]. Let us first consider the case when the controlled-phase gate acts as the identity whenever one of the interacting qubits is leaked, corresponding to no leakage-conditional errors being propagated. We also assume that single-qubit gates act trivially on any leaked qubits, which is typically the case in experiments. In this case, the X gates applied to the data qubits at the end of each round will flip the parity measured between the leaked and the computational data qubits, leading to a defect at every round until the leaked qubit seeps back into the computational subspace. If the interaction with the leaked data qubit leads to non-trivial leakage-conditional errors, then a defect is propagated with a probability depending on the phase error. In the case of a leaked ancilla qubit, these leakage-conditional phases do not lead to defects as they remain undetected by the repetition code. Instead, the propagated defects depend on what outcomes the measurements of leaked ancilla qubits lead to. Leaked qubits typically remain outside the computational subspace for several rounds, during which time

they can generate multiple defects. This, combined with the increase of the average leakage population of the qubits towards a steady-state population over the first several rounds of the experiment, can also explain the observed increase of the average defect probability observed in Fig. 7.3**a**.

Errors leading to three or more non-trivial defects can be represented as *hyperedges* in a *decoding hypergraph*, each again associated with a corresponding probability. We discuss this generalization of the decoding graph in Sec. 7.4.3 and Sec. 7.4.6. Such errors can also lead to the inaccurate estimation of the expected edge probabilities and, in some cases, even non-physical (negative) probability estimates. In particular, we have found that this to be the case for the estimation of the boundary edges, which rely on the accurate inference of the probabilities associated with all other edges, see Eq. (7.2) and Sec. 7.4.5. Specifically, we find that many boundary edge probabilities are estimated to be negative when using the experimental dataset, see Sec. 7.4.5 for an explanation and additional details.

Assuming that one can identify a set of expected hyperedges, their probabilities can also be inferred from the measured syndrome defects and accounted for when estimating the edge probabilities, leading to more accurate estimates [28]. However, we were not able to correct the estimated edge probabilities due to the complexity of identifying the set of hyperedges required to model the observed defect correlations. Therefore, we opted not to use these probabilities for the weights of the MWPM decoder and instead extracted these weights from the circuit-level noise model that we use in our simulations, which is described in Sec. 7.4.2 and which is based on the measured coherence times and operational error rates.

7.2.3. Logical performance

Next, we characterize the logical performance of the code and demonstrate an exponential suppression of the logical error rate with increasing code distance. After applying the corrections obtained from the MWPM decoder, we compare the corrected logical qubit parity to the initially prepared state to determine if a logical error occurred. Each experiment performing r rounds of QEC is then repeated many times to calculate the logical error probability $p_L(r)$ at round r. Repeating this procedure for $r = 0, 1, 2, \ldots, 15$ enables us to extract the logical error rate per round ε_L by fitting the decay of $p_L(r)$ over the QEC rounds, using

$$p_L(r) = \frac{1}{2}\left[1 - (1 - 2\varepsilon_L)^{r - r_0}\right],$$

where r_0 is a fitting constant [52]. We perform this series of experiments only for the $d = 7$ repetition code. To extract the performance of the $d = 3$ or $d = 5$ code, we instead follow the procedure used in [15, 27] and subsample the measured defects into a set of smaller datasets corresponding to the possible codes that can be defined in our linear layout, see Fig. 7.1. This leads to five and three datasets for the possible $d = 3$ and $d = 5$ codes, respectively.

The logical performance of the $d = 3, 5, 7$ codes extracted from either simulated or experimental data are shown in Fig. 7.4**a** and Fig. 7.4**b**, respectively. The logical error rates obtained by fitting the decay of average p_L over the subsampled datasets are shown
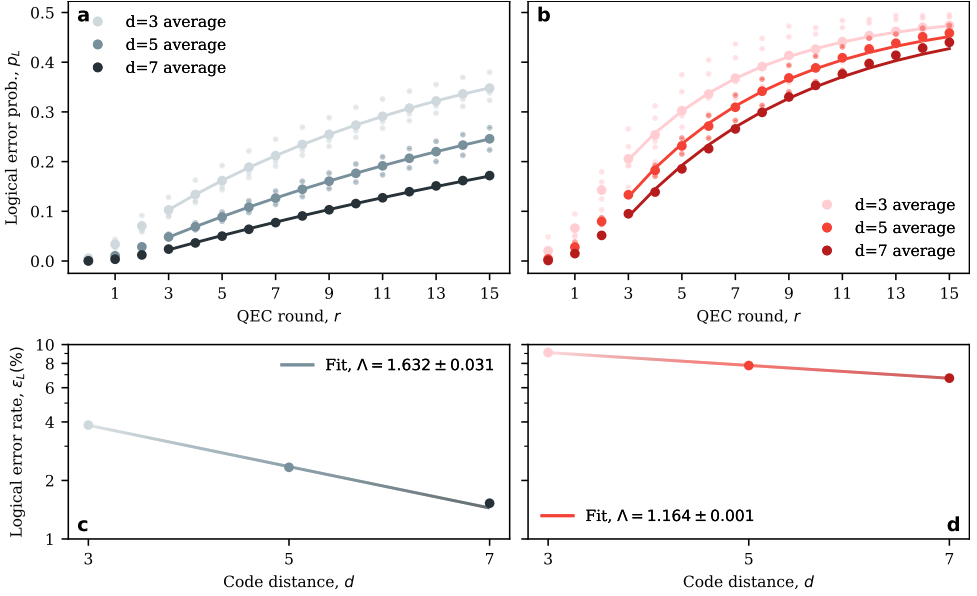
Figure 7.4: Logical performance. **a** The probability of a logical error p_L as a function of the number of QEC rounds r extracted from simulated data for the $d = 3$ (light gray), $d = 5$ (gray), and $d = 7$ (dark gray) repetition codes. In contrast, **b** shows the probabilities extracted from experimental data for the $d = 3$ (light red), $d = 5$ red, and $d = 5$ (dark red) codes. The logical performance of the $d = 3, 5$ codes is obtained by subsampling the $d = 7$ dataset. The small transparent dots in these subfigures show the probabilities of each subsampled code individually, while the large dots show their averages. The solid lines show the fits to the average error probabilities from $r = 3$ used to extract the logical error rates per round, plotted in **c** and **d**. The fits start from round $r = 3$ to avoid any time-boundary effects. **c, d** The logical error rate ε_L per round versus the code distance for the simulated (**c**, gray) and experimental (**d**, red) data, demonstrating suppression of the error rate. The solid lines show the fits used to extract the error suppression factors.

in Fig. 7.4**c** and Fig. 7.4**d**, respectively. Assuming that the physical error rate is below the threshold, ε_L is exponentially suppressed with increasing code distance d, modeled by

$$\varepsilon_L(d) = C / \Lambda^{(d+1)/2},$$

where Λ is the suppression factor and C is a fitting constant [27]. The fits to the extracted ε_L and the obtained factors Λ are shown in Fig. 7.4**c** and Fig. 7.4**d**. While we observe a suppression of ε_L with increasing distance d in both the simulation and experiment, the suppression factors achieved are moderate in either case and lower than those achieved in other repetition code experiments [15, 19, 27]. This performance is largely a result of the relatively high operational error rates, especially in the case of the readout. Further lowering these error rates remains paramount for achieving better logical performance for the repetition code and, ultimately, the surface code. The discrepancy in the logical performance between the simulation and the experiment is likely due to the impact that non-conventional errors, particularly leakage, have on the decoding and logical error rate. While this conclusion is motivated by the strong signatures of leakage observed in the error probabilities inferred from the measured syndrome defects, a more detailed

error model will be required to determine the precise impact of leakage and whether it can explain the observed discrepancy. However, it is worth noting that other repetition-code experiments have observed similar signatures and discrepancies, attributing them to leakage or crosstalk errors [19, 27]. We discuss the possible improvements and extensions to our error model in Sec. 7.3.

7.3. Discussion

We have demonstrated a repetition code experiment performing up to 15 rounds of stabilizer measurements. We have also shown that our system can suppress the logical error rate by increasing the distance of the code despite the moderate logical performance of the code. However, implementing a surface code memory experiment is much more challenging and will require several challenges to be addressed. In particular, the high physical error rates lead to the overall poor logical performance observed in this experiment. The low relaxation and dephasing times, together with the qubit readout and the two-qubit gate error rates, are the dominant error sources in this experiment, see Tab. 7.1 and Tab. 7.2. Achieving lower operational error rates will be necessary for any break-even quantum memory demonstration and, eventually, scaling up the code size to reduce the error rate.

On the other hand, the measured syndrome defects analysis revealed signatures of leakage and crosstalk errors. These non-conventional errors, in particular, can severely impact the logical performance of code and interfere with estimating the weights used by the MWPM decoder directly from the experimental data [49]. Addressing these errors will also be essential for any future implementation. Several leakage reduction schemes have been proposed and already integrated into some QEC experiments [47, 48, 50, 53], where they have been shown to significantly improve the system's stability, lead to lower defect probabilities, and possibly lower logical error rates. Addressing crosstalk errors might require the inclusion of tunable couplers between the transmon qubits or a more precise optimization of the qubit and pulse frequencies in the case of microwave crosstalk. However, further research and characterization are needed to address the crosstalk issues in the device. Finally, as we scale up to larger systems, we expect two-level system defects also to become another important source of errors that must be considered in greater detail.

Next, we discuss possible improvements to the error models we used in our simulations that might lead to a better match between simulation and experiment. We have implemented an approximate Pauli-noise model (described in detail in Sec. 7.4.2) that enables us to use a Pauli-frame simulation (otherwise referred to as stabilizer simulation), making the simulations scalable to large code distances. Similar models have shown to be able to achieve an overall good match with experimental results [19, 27], demonstrating that such models can still accurately capture the impact of the various physical error sources to a large degree. In our model, we consider a Pauli-twirled approximation of the amplitude-phase damping channel during periods of qubit idling. However, we only include depolarizing channels after single-qubit and two-qubit gates, parameterized by the operational error rate extracted from randomized benchmarking experiments. A possible improvement to our model would be to model the impact of decoherence throughout the operation by including a pair of twirled amplitude-phase

damping channels for half the gate time around the ideal operation, similar to the error model presented in [49]. As these channels typically do not fully capture the measured error rate in randomized benchmarking experiments, we would likely also need to consider a depolarizing channel after the second twirled amplitude-phase damping channel, such that the noisy operation has the same fidelity as the one measured.

A more accurate model of the readout would consider a pair of bitflip channels before and after the ideal projective measurement, approximating the relaxation or excitation that the qubit experiences during the readout, as well as a probability to flip the measurement outcome. In the model we consider here, we assume that there is no process that leads to classical declaration errors. These three probabilities can be extracted from measurement butterfly characterizations, see [50] for a description of this procedure. Finally, an error source we did not consider in this work but that can easily be included is the residual excitations when the qubits are prepared at the start of each experiment, which can be modeled as bitflip channels following the initial preparation of the qubits.

The models described above consider only the errors within the qubit computational subspace. However, leakage to non-computational states is typically a significant error source in current experiments, especially considering the impact it can have on the logical error rate of a quantum code. An incoherent leakage model could capture the effect of qubit leakage with a reasonable degree of accuracy by considering a stochastic leakage process following different operations and tracking whether each qubit is leaked in the simulation, as demonstrated in [19]. In such a model, qubits have a certain probability of leaking after each operation. In the case of a leakage event, the qubit is flagged as leaked, and the gates applied to this leaked qubit either act trivially or induce errors. In particular, the phase errors due to the interactions between qubits in the computational and leaked subspaces can be modeled as a probability of Z error being applied to the computational qubit, conditioned on the other qubit being leaked. A stabilizer simulation, including such a stochastic leakage model, can maintain a simulation cost scaling quadratically with the number of qubits and linearly with the number of QEC rounds [19].

Estimating the error probabilities from the measured syndrome defects helps avoid the need for complex error models and makes matching-based decoders, such as MWPM decoders or belief-matching decoders [54], a lot more flexible. However, errors that lead to more than two defects directly impact the accuracy of the estimated probabilities, leading to unphysically small or negative probabilities in some instances. Resolving this typically involves either introducing a minimal value that these probabilities can take based on some error model [19] or explicitly considering a subset of these errors in the estimation procedure [28]. One problem is that estimating the probability of the errors leading to any possible combination of four or more defects directly from the data is a very computationally expensive task, mainly due to the large number of combinations needed to be considered, see Sec. 7.4.6 for a further discussion.

Leakage-reduction operations can reduce the leakage lifetime to about a single cycle, limiting the number of defects these errors can lead to. Therefore, these operations will likely be necessary to extend the estimation procedure to include leakage errors in this model while keeping the computational complexity feasible. However, another issue is that these probabilities might not be immediately helpful for a decoder since there is

no clear, logical correction associated with these non-conventional errors. Neural network decoders [55–65] can instead be used to directly infer the most likely logical corrections from the measured syndrome defects, generalizing the process of estimating these weights and then providing them to a matching decoder.

7.4. SUPPLEMENTAL INFORMATION

7.4.1. DEVICE

The device follows the frequency arrangement scheme proposed in [66], which employs three frequency groups. Data qubits D_3 and D_7 are part of the high-frequency group, while the remaining data qubits and the ancilla qubits are in the low-frequency and mid-frequency groups, respectively. Each transmon qubit is capacitively coupled to its nearest-neighboring qubits via a coupling bus resonator, even if they are unused in this experiment. Control of the transmon qubits is achieved via individual microwave-drive lines used for single-qubit gates and dedicated flux lines for two-qubit gates. Each qubit is also dispersively coupled to a Purcell-filtered readout resonator, enabling the dispersive readout of these qubits.

7.4.2. ERROR MODELS

Qubit	Relaxation, time T_1 (μs)	Dephasing, time T_2 (μs)	Single-qubit gate error, ε_g (%)	Assignment, error p_a (%)	QND prob., p_{QND} (%)
D1	18.16	14.40	0.21	5.63	-
D2	11.27	14.26	0.12	5.02	-
D3	11.51	10.68	0.04	1.35	-
D4	13.97	7.18	0.06	2.81	-
D5	14.46	12.20	0.08	11.82	-
D6	10.24	15.77	0.30	2.75	-
D7	5.86	9.74	0.17	2.16	-
A1	7.06	11.12	0.17	3.02	93.70
A2	8.72	11.48	0.16	1.82	92.39
A3	5.11	4.47	0.46	1.98	92.90
A4	9.90	6.59	0.14	3.10	88.40
A5	12.42	13.41	0.25	1.56	89.59
A6	10.61	18.04	0.11	3.81	92.65

Table 7.1: A summary of the measured coherence times, single-qubit gate error rates, and readout error rates for each qubit used in the experiment. The coherence times are measured using standard time-domain measurements [67], with the dephasing time measured using a Hahn echo experiment. The gate error is extracted using randomized benchmarking experiments [68–70]. The readout parameters include the assignment error [71] and quantum non-demolition probability [28].

The error model we consider here makes several simplifying assumptions about the noise and ultimately considers Pauli-error channels before or after operations. For such an error model, the circuits we consider here can be simulated by propagating the Pauli frame through the circuit, making these simulations computationally efficient and, therefore, scalable. Of course, the model's approximations make it less accurate in its predictions compared to a density-matrix simulation, for example. However, similar approximate models have been shown to be capable of capturing the dominant error sources in some experiments with reasonable accuracy [19, 27]. In particular, we use *stim* [72] to perform the simulations done in this chapter and *pymatching* [33] to perform MWPM decoding. Using these packages also allows us to directly extract the weights used by the MWPM-decoder from the constructed circuits using the error model described above. More specifically, each edge $h_m = \{v_i, v_j\}$ in the decoding graph is assigned a weight $w_m = \log\big((1 - p_m)/p_m\big)$, where p_m is the probability of the error event leading to defects $d_i = 1$ and $d_j = 1$ that is associated with that edge.

Interacting qubits	Two-qubit gate error ε_g (%)
A1, D1	2.18
A1, D2	1.35
A2, D2	1.12
A2, D3	2.55
A3, D3	1.92
A3, D4	2.02
A4, D4	1.34
A4, D5	1.62
A5, D5	1.94
A5, D6	1.07
A6, D6	0.93
A6, D7	2.50

Table 7.2: The controlled-phase error rates measured in the experiment using randomized benchmarking [68–70, 73]

We model the decoherence experienced by transmon qubits during periods of idling by including a Pauli-twirled amplitude-phase damping channel after each such operation. Whenever a transmon qubit idles for a time t, it undergoes an amplitude-damping channel parameterized by the relaxation time T_1 and described by the Pauli-transfer ma-

trix (PTM)

$$R_{\mathrm{amp}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{1-\gamma} & 0 & 0 \\ 0 & 0 & \sqrt{1-\gamma} & 0 \\ \gamma & 0 & 0 & 1-\gamma \end{pmatrix},$$

where $\gamma = 1 - e^{-t/T_1}$. At the same time, the qubit also undergoes dephasing, which is parameterized by the pure-dephasing time T_ϕ and described by the transfer matrix

$$R_{\mathrm{deph}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-\lambda & 0 & 0 \\ 0 & 0 & 1-\lambda & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where $\lambda = 1 - e^{-t/T_\phi}$. The pure-dephasing time T_ϕ is related to the dephasing time T_2 by

$$\frac{1}{T_\phi} = \frac{1}{T_2} - \frac{1}{2T_1}.$$

The transfer matrix corresponding to the amplitude-phase damping channel is then given by $R = R_{\mathrm{amp}} R_{\mathrm{deph}}$. We then apply the Pauli-twirling approximation over the Pauli basis to R to obtain the twirled transfer matrix $\tilde{R}$, which is diagonal. The twirled single-qubit process then corresponds to a Pauli-noise channel that applies an X, Y, or Z with probabilities p_X, p_X and p_X, respectively. For the amplitude-phase channel defined above, one finds

$$p_X = p_Y = \frac{\gamma}{4},$$

$$p_Z = \frac{1}{2} - \frac{\gamma}{4} - \frac{(1-\lambda)\sqrt{1-\gamma}}{2}.$$

Single-qubit gates are followed by a depolarizing channel, which, with a probability $p_g/3$, applies an error X, Y and Z. This probability is related to the gate error ε_g measured from randomized benchmarking (also referred to as the gate infidelity) by $p_g = (2^n + 1)\varepsilon_g/2^n$, n is the number of qubits involved in the gate, with $n = 1$ in this case [68, 69]. Similarly, controlled-phase gates are followed, with probability p_g, by an error drawn uniformly from the 15 elements in the set $\{I, X, Y, Z\}^{\otimes 2} \setminus II$. Here p_g is again determined by its relation with the infidelity ε_g with $n = 2$.

In our model, each measurement is preceded by a symmetric bitflip channel that leads to an X error with probability p_a, where p_a is the average assignment error probability of the readout. To estimate this probability, we prepare each qubit in a state s and then measure it. By repeating this experiment and preparing $|0\rangle$ and $|1\rangle$ an equal number of times, we can estimate

$$p_a = \frac{P(m=0 \mid s=1) + P(m=1 \mid s=0)}{2},$$

with m being the observed measurement outcome. In our model, each measurement operation can also be followed by an X error with probability p_q, which is derived from an experimentally estimated quantum non-demolition (QND) probability p_{QND}. To estimate this probability, we follow the experiment outlined in [28] and prepare each qubit in the state $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$ after which we perform two consecutive measurements that produce outcomes m_1 and m_2, respectively. In this case, the first outcome should be ideally random, while the second outcome should be the same as the first in the absence of errors. In practice, the measurement can have some back-action on the state of the ancilla, which can be captured as

$$p_{QND} = \frac{P(m_2 = 0 \mid m_1 = 0) + P(m_2 = 1 \mid m_1 = 1)}{2},$$

see Tab. 7.1 for the measured values. We assume that the initial state is prepared perfectly, that each measurement is projective, and that the measurement outcome always corresponds to the state of the qubit. In this case, the QND probability captures the impact of both the assignment error in the second measurement as well as any back-action on the qubit from the first measurement that changes its state, but that does not affect m_1. Here, we model this error by a bitflip channel after each measurement that flips the state with probability p_q. Therefore

$$p_{QND} = p_a p_q + (1 - p_a)(1 - p_q),$$

which can be rearranged to give

$$p_q = \frac{1 - p_{QND} - p_a}{1 - 2p_a},$$

where $p_a < 1/2$ given that one has the freedom to swap the declared outcomes for each state otherwise. All of the experimentally measured coherence times and operational error rates used in our model are listed in Tab. 7.1 and Tab. 7.2.

7.4.3. The decoding graph

Here we define the *decoding* hypergraph $G = (V, H)$ that represents the effective physical error model, where V is the set of nodes and H is the set of hyperedges connecting two or more nodes (even though typically error models consider only edges in this graph). First, we define a single index $i = (a, r)$ to label each possible syndrome defect $d_{a,r}$ by ancilla qubit a in QEC round r, such that there are a total of $N = N_a N_r$ defects, with N_a being the number of ancilla qubits and N_r the total number of QEC rounds. Thus, each syndrome defect will be represented as a node v_i in the hypergraph. Note that these nodes are included in this hypergraph regardless of the specific non-trivial or trivial defects measured in each experiment. Each Pauli error that occurs at some location in the circuit and leads to one or more non-trivial syndrome defects will be associated with a hyperedge $h \in H$ in the hypergraph.

Specific errors will lead to only a single non-trivial defect, such as some errors on data qubits on the boundary of a code. We refer to the set of nodes corresponding to these syndrome defects as the boundary nodes V_{bound}. In contrast, the nodes corresponding to

syndrome defects that always occur together with one or more other defects, including the boundary defects, are referred to as the bulk nodes V_{bulk}. We will also introduce a single "image" node v_0 and a corresponding defect d_0 that is not actually being measured in any experiment. Note that this is not a unique choice, and one can also introduce such an image node for each boundary node in the hypergraph. The errors leading to only a single non-trivial syndrome defect will be represented as edges between v_0 and the boundary nodes $v_i \in V_{\text{bound}}$. The full set of nodes included in the hypergraph is then $V = V_{\text{bulk}} \cup V_{\text{bound}} \cup \{v_0\}$, with $|V| = N + 1$.

The hyperedge set H of the hypergraph G is defined as follows. If a single error occurring at some location in the circuit leads to two or more non-trivial syndrome defects $d_i, \ldots, d_j$, then there is a corresponding hyperedge $h_m = \{v_i, \ldots, v_j\}$ in the hypergraph between the relevant nodes, where m is the index of the edge. Not that these edges do not include the image node v_0. Instead, if an error leads only to a single non-trivial defect d_i, then there is an edge $h_m = \{v_i, v_0\}$ between the corresponding boundary node and the image node. We refer to the latter set of edges as the *boundary* edges. It is possible for multiple errors occurring at different locations in the circuit to lead to the same defects (thus, these errors are indistinguishable from each other). We associate each hyperedge h_m with a single independent error event that happens with a probability p_m and leads to the non-trivial defects corresponding to the nodes connected by the hyperedge. Therefore, this error event probability combines the probabilities of the errors in the circuit that lead to the same set of non-trivial defects.

7.4.4. Estimating the bulk edge probabilities

We first consider the standard case, where G is just a graph, meaning that every $h_m \in H$ is simply an edge, i.e., such that $h_m = \{v_i, v_j\}$. We discuss the extension to a hypergraph in Sec. 7.4.6 and discuss some of the difficulties that come when working with hyperedges at the end of that section. Let us introduce a random binary variable e_m associated with each edge h_m that has possible outcomes $y_m \in \{0, 1\}$ and that corresponds to whether the corresponding error event occurred or not. These hidden variables influence the observed non-trivial defects d_i, which we also treat as a random variable with possible outcome $x_i \in \{0, 1\}$. To simplify the notation, we define the random defect vector $\mathbf{d} = (d_0, d_1, d_2, \ldots, d_N)$ with possible outcomes $\mathbf{x} = (x_0, x_1, x_2, \ldots, x_N)$. Similarly, define the random error event vector $\mathbf{e} = (e_1, e_2, \ldots, e_M)$ with corresponding outcomes $\mathbf{y} \in \Omega$, where Ω is the outcome sample space and $M = |H|$ is the size of the graph. We can now consider the probability distribution

$$
\begin{aligned}
p(\mathbf{x}) &= P(\mathbf{d} = \mathbf{x}) \\
&= \sum_{\mathbf{y} \in \Omega} P(\mathbf{d} = \mathbf{x} \mid \mathbf{e} = \mathbf{y}) P(\mathbf{e} = \mathbf{y}) \\
&= \sum_{\mathbf{y} \in \Omega} p(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}).
\end{aligned}
\tag{7.3}
$$

We assume that the error events are independently distributed such that

$$
p(\mathbf{y}) = \prod_{m=1}^{M} p(y_m) = \prod_{m=1}^{M} p_m^{y_m} \left(1 - p_m\right)^{1 - y_m},
\tag{7.4}
$$

where, we used that error event $P(e_m = 1) = p_m$, while $P(e_m = 0) = 1 - p_m$.

Next, consider a subset of the nodes $V_S \subseteq V$ and their corresponding possible defect outcome $\mathbf{x}_S$. We can then define the set of edges $H_S = \{h \in H \mid h \subseteq N[V_S]\}$ together with their corresponding event outcomes $\mathbf{y}_S$, where $N[V_S]$ denotes the closed neighborhood of V_S. It then follows that $p(\mathbf{x}_S \mid \mathbf{y}) = p(\mathbf{x}_S \mid \mathbf{y}_S)$, which simply says that the defects $\mathbf{d}_S$ are only affected by the edges in H that involve the nodes V_S, namely the edges in H_S. In particular, if we consider the subset V_S to be the pair of vertices connected via an edge h_m, then H_S will contain h_m and any other edge that is incident to any of these two nodes. For later convenience, we define:

Definition 1. *Let us introduce the edge set $A_m = \{h \in H \mid h \subseteq N[h_m] \text{ and } h \neq h_m\}$, i.e., A_m is the set of edges which are adjacent to edge h_m, excluding h_m itself. For an edge $h_m = \{v_i, v_j\}$, we also define $A_m^i \subseteq A_m$ as the set of edges which are incident to node v_i and similarly the subset A_m^j for those incident to node v_j, such that $A_m = A_m^i \cup A_m^j$.*

Now, we shall consider a specific edge h_m and let us denote $\mathbf{y}_m^i \in \Omega_m^i$ to be the event outcomes corresponding to the edges in A_m^i, and analogously for the edges in A_m^j. The error model that we consider here implies that

$$p(x_i, x_j \mid y_m, \mathbf{y}_m^i, \mathbf{y}_m^j) = \delta_{x_i, y_m \oplus_l (\mathbf{y}_m^i)_l} \delta_{x_j, y_m \oplus_l (\mathbf{y}_m^j)_l}. \tag{7.5}$$

This expresses that observing a defect outcome $d_i = x_i$ and $d_j = x_j$ is *deterministically* determined by the error event outcomes corresponding to edge h_m as well as the other edges that are incident to v_i and v_j. Particularly, observing a non-trivial defect $d_i = 1$ requires an odd number of events that lead to a defect at d_i to have occurred.

Next, we consider the experimentally observable outcomes and how to infer the edge probabilities from these statistics. Namely, we have access to the expectation values

$$\begin{aligned}
\langle d_i d_j \rangle &= \sum_{x_i, x_j} p(x_i, x_j) x_i x_j \\
&= P(d_i = 1, d_j = 1) \\
&= \sum_{\mathbf{y} \in \Omega} P(d_i = 1, d_j = 1 \mid \mathbf{e} = \mathbf{y}) P(\mathbf{e} = \mathbf{y}),
\end{aligned} \tag{7.6}$$

for any pair of defects d_i and d_j. Similarly, we have

$$\begin{aligned}
\langle d_i \rangle &= \sum_{x_i} p(x_i) x_i = P(d_i = 1) \\
&= \sum_{\mathbf{y} \in \Omega} P(d_i = 1 \mid \mathbf{e} = \mathbf{y}) P(\mathbf{e} = \mathbf{y}),
\end{aligned} \tag{7.7}$$

where $p(x_i, x_j)$ and $p(x_i)$ are, of course, the marginal distributions of Eq. (7.3). Note that we do not have access to the expectation value for $\langle d_i d_0 \rangle$, where d_0 is the defect corresponding to the image node v_0. Furthermore, we also do not have access to $\langle d_0 \rangle$, which ultimately complicates the estimation of the probabilities for the boundary edges, which are treated separately, see Sec. 7.4.5.

Considering Eq. (7.6) we can see that the edges that influence d_i and d_j are only the edge h_m and the edges in A_m. The defects in A_m influence either d_i or d_j, but not both. Splitting A_m into the two subsets A_m^i and A_m^j, we can use Eq. (7.5) to obtain

$$\langle d_i d_j \rangle = p_m \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (\mathbf{y}_m^i)_l = 0}} p(\mathbf{y}_m^i) \sum_{\substack{\mathbf{y}_m^j \in \Omega_A^j \\ \oplus_l (\mathbf{y}_m^j)_l = 0}} p(\mathbf{y}_m^j)$$
$$+ (1 - p_m) \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (\mathbf{y}_m^i)_l = 1}} p(\mathbf{y}_m^i) \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^j \\ \oplus_l (\mathbf{y}_m^j)_l = 1}} p(\mathbf{y}_m^j). \tag{7.8}$$

We can analogously derive the equations for the expectation value $\langle d_i \rangle$, given by

$$\langle d_i \rangle = p_m \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (\mathbf{y}_m^i)_l = 0}} p(\mathbf{y}_m^i) + (1 - p_m) \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (\mathbf{y}_m^i)_l = 1}} p(\mathbf{y}_m^i). \tag{7.9}$$

These expressions hold for any edge h_m in the decoding graph. Using Eq. (7.4) we can define the probability

$$q_m^i \equiv \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (\mathbf{y}_m^i)_l = 1}} p(\mathbf{y}_m^i)$$
$$= \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (\mathbf{y}_m^i)_l = 1}} \prod_l p_l^{y_l} (1 - p_l)^{1 - y_l}, \tag{7.10}$$

where y_l is a shorthand for $(\mathbf{y}_m^i)_l$. Thus q_m^i is the probability of observing $d_i = 1$ due to an odd number of errors which each trigger d_i, excluding the error event e_m that corresponds to the edge h_m that we are specifically considering. Since either an even or an odd number of errors must have occurred, we can also define the complementary probability.

$$1 - q_m^i = \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (\mathbf{y}_m^i)_l = 0}} p(\mathbf{y}_m^i)$$
$$= \sum_{\substack{\mathbf{y}_m^i \in \Omega_m^i \\ \oplus_l (y_m^i)_l = 0}} \prod_l p_l^{y_l} (1 - p_l)^{1 - y_l}, \tag{7.11}$$

corresponding to an even number of events happening and (thus not triggering d_i) (note that this also includes the case where none of these error events occurred, i.e., $\mathbf{y}_m^i = \mathbf{0}$). Using these two new definitions, we can rewrite Eq. (7.8) and Eq. (7.9) as

$$\langle d_i d_j \rangle = p_m \left(1 - q_m^i \right) \left(1 - q_m^j \right) + (1 - p_m) \, q_m^i q_m^j, \tag{7.12}$$

and

$$\langle d_i \rangle = p_m \left(1 - q_m^i\right) + \left(1 - p_m\right) q_m^i. \tag{7.13}$$

This leads to a system of three equations for $\langle d_i d_j \rangle$, $\langle d_i \rangle$, $\langle d_j \rangle$ which involve three variables, p_m, q_m^i, and q_m^j. More generally, such a system of equations can be constructed for every edge h_m, with the exception of the boundary edges. The system of equations is straightforward to solve analytically, resulting in an expression for the probability p_m associated with edge h_m (which is not a boundary edge)

$$p_m = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\langle d_i d_j \rangle - \langle d_i \rangle \langle d_j \rangle}{1 - 2\langle d_i \rangle - 2\langle d_j \rangle + 4\langle d_i d_j \rangle}}. \tag{7.14}$$

We note that this procedure was also derived in [27, 51].

If a pair of defect variables d_i and d_j are uncorrelated, then the evaluation of Eq. (7.14) leads to $p_m = 0$. If d_i and d_j are instead correlated, we expect to observe a $p_m > 0$. However, if these variables are instead anti-correlated, then evaluating the above expression becomes a problem as the term under the square root is negative.

As observed in our experiment or in [27] one finds a $p_m > 0$ for more edges than expected from a circuit-level Pauli-noise error model. Although this analysis allows us to estimate the probability of the error leading to any pair of defects, using these probabilities in decoding requires also to assign a correction corresponding to each edge in the decoding graph. This correction is not immediately obvious based on the analysis presented so far. A common strategy is to only consider the edges expected from a circuit-level noise model and assign them the corresponding corrections [19, 28]. Another strategy would be to assign the minimum-weight correction.

A different complication is that some errors lead to more than two non-trivial defects. For many codes, these can be certain Pauli errors happening at some location in the circuit. For example, the Y errors when running a surface code memory experiment lead to three or four non-trivial defect outcomes, as they affect both X-and Z-parity check outcomes. However, they can also be due to the presence of non-conventional errors, namely crosstalk and leakage errors. These errors are associated with hyperedges in a hypergraph, and the presence of these hyperedges can lead to an inaccurate estimation of the edge probabilities if not accounted for, as demonstrated in [28]. In certain cases, this can even lead to the estimation of negative probabilities, which is the case in our experiment. The probabilities of these hyperedges can also be estimated from the data. However, it requires solving a larger system of equations that scales with the number of nodes that the hyperedge connects, see Sec. 7.4.6. Practically, this means that estimating the probability of very large hyperedges will be too computationally expensive. A procedure for calculating these hyperedge probabilities and for correcting the estimated edge probabilities was also given in [19, 28]. Since solving a larger system of equations quickly becomes too computationally expensive, the probabilities for hyperedges of a certain size are typically not calculated for all possible node combinations.

7.4.5. ESTIMATING THE BOUNDARY EDGE PROBABILITIES

Now let us now consider an edge $h_m = \{d_i, d_0\}$ between a defect d_i and the image defect d_0, for which the error probability cannot be estimated using Eq. (7.14). As a reminder, the main issue when dealing with these edges is that one does not have access to $\langle d_i d_0 \rangle$ or $\langle d_0 \rangle$ in order to construct the system of equations that can be solved to obtain the edge probability. Instead, we can only estimate $\langle d_i \rangle$, for which Eq. (7.13) holds. Rearranging that equation yields

$$p_m = \frac{\langle d_i \rangle - q_m^i}{1 - 2q_m^i},\tag{7.15}$$

with the expression for q_m^i being given in Eq. (7.10). In particular, q_m^i can be determined from the probabilities p_l of all edges h_l incident to node v_i, excluding h_m itself. Note that in our definition, there is only a single image node v_0, meaning that v_i can have only one edge connecting it to v_0. In turn, this means that we can calculate q_m^i after estimating p_l for all the non-boundary edges in the graph, and this will yield the probability p_m of the boundary edge h_m that we are after.

In practice, we find that estimating the boundary edge probabilities using this method can be prone to errors. In the first place, both $\langle d_i \rangle$ and $\langle d_i d_j \rangle$ are estimated with some sampling noise, which translates to some error in the estimates for the non-boundary edge probabilities p_l. Unless we consider a certain set of edges expected from some error model, we can generally estimate an edge probability p_m between *any possible* pair of nodes. In that case, each v_i will have $n - 1$ edges connecting to the other possible nodes, excluding v_0. Since estimating q_m^i requires summing over all of these edges, the errors in the edge probabilities are propagated to the estimate of the boundary edge probability. Furthermore, errors leading to more than two defects lead to further errors in the edge estimates, which we discuss in more detail in Sec. 7.4.6. In our experiment, this leads to all boundary edge probabilities being estimated to be negative. Considering only the edges corresponding to the errors expected from a circuit-level depolarizing noise reduces the number of unphysical boundary edge probabilities, as discussed in [74]. However, even in that case, we still estimate approximately half of the boundary edges to be negative, which we associate primarily with the non-conventional errors present in the experiment instead of the sampling noise due to the large number of runs used for these estimates.

7.4.6. HYPEREDGES

It is clear that excluding the hyperedges and the errors they correspond to can considerably impact the estimated edge probabilities, especially for the boundary edges, as is the case in our experiment. Therefore, it is necessary to extend the above analysis to include any hyperedges in the decoding graph corresponding to non-conventional errors not considered in our error model to get accurate edge estimates.

Here, we first consider hyperedges of size three and outline a procedure for estimating their corresponding edge probabilities from the experimentally measured syndrome defects. One can then use these hyperedge probabilities to obtain more accurate estimates for the edge probabilities, namely using the procedure outline in [28], which is
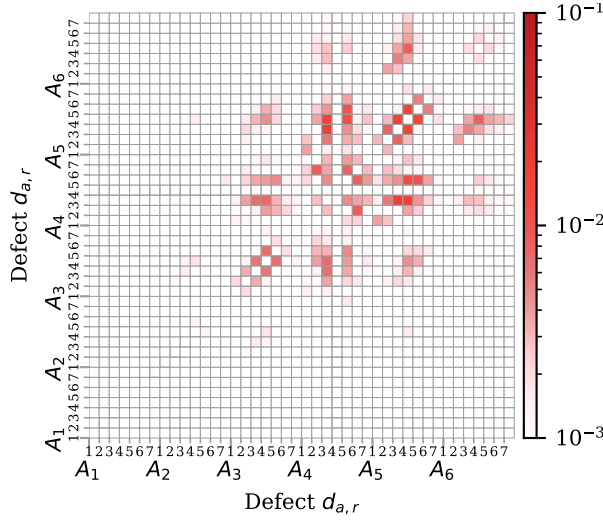
Figure 7.5: The discrepancies $\mathscr{D}_{i,j,k}$ extracted from experimental data that involve the defect d_i measured by ancilla qubit A_4 at QEC round 4 i.e., $i = (A_4, 4)$, and all other pairs of defects d_j and d_k. Here, $\left\langle d_i d_j d_k \right\rangle_{\text{edge}}$ corresponds to the expectation value inferred from the measured syndrome defects assuming that all errors lead either to a single defect or a pair of defects. A positive discrepancy $\mathscr{D}_{i,j,k} > 0$ indicates that there are errors triggering these three (and possibly more) defects. Note that the matrix is symmetric along the main anti-diagonal.

7

based on a clustering algorithm. Furthermore, we propose a procedure for determining what hyperedges should be included when there isn't a set of expected hyperedges given some error model. While this offers a general approach to including hyperedges based on the correlations between the measured syndrome defects, it quickly becomes too computationally expensive for hyperedges of size four or more.

Before introducing any hyperedges to the graph, let us first consider the expectation value $\left\langle d_i d_j d_k \right\rangle$ that solely the edges in the graph lead to, which we will refer to as $\left\langle d_i d_j d_k \right\rangle_{\text{edge}}$ for clarity. The possible edges between the nodes corresponding to these defects are $h_m = \{v_i, v_j\}$, $h_n = \{v_j, v_k\}$, $h_o = \{v_i, v_k\}$. We then define the probability $q_{m,n}^i$ for a non-trivial defect $d_i = 1$ being observed due to the possible error events corresponding to all edges incident to node v_i with the exception of h_m and h_n. This is simply an extension of the definition of q_l^i given in Eq. (7.10), which can also be expressed as

$$q_{m,n}^i = \frac{q_m^i - p_n}{1 - 2p_n},\tag{7.16}$$

where p_n is the probability associated with edge h_n. This can also be expressed in a similar fashion to Eq. (7.10), by instead considering an incident set of edges $A_{m,n}^i$ to node v_i which excludes both h_m and h_n and the corresponding possible outcomes $\mathbf{y}_{m,n}^i \in \Omega_{m,n}^i$. To estimate $\left\langle d_i d_j d_k \right\rangle_{\text{edge}}$ in particular, we need to consider the probabilities $q_{m,o}^i$, $q_{m,n}^j$ and $q_{n,o}^k$ of observing a non-trivial value on each of these three defects, excluding

the errors associated with the two edges connecting the corresponding node to any of the remaining nodes. We can then write

$$
\begin{aligned}
\langle d_i d_j d_k \rangle_{\text{edge}} = {} & p_m q_{n,o}^k \left(1 - p_n\right) \left(1 - p_o\right) \left(1 - q_{m,o}^i\right) \left(1 - q_{m,n}^j\right) \\
& + p_n q_{m,o}^i \left(1 - p_m\right) \left(1 - p_o\right) \left(1 - q_{n,o}^k\right) \left(1 - q_{m,n}^j\right) \\
& + p_o q_{m,n}^j \left(1 - p_m\right) \left(1 - p_n\right) \left(1 - q_{m,o}^i\right) \left(1 - q_{n,o}^k\right) \\
& + p_m p_n q_{m,n}^j \left(1 - p_o\right) \left(1 - q_{m,o}^i\right) \left(1 - q_{n,o}^k\right) \\
& + p_m p_o q_{m,o}^i \left(1 - p_n\right) \left(1 - q_{m,n}^j\right) \left(1 - q_{n,o}^k\right) \\
& + p_n p_o q_{n,o}^k \left(1 - p_m\right) \left(1 - q_{m,n}^j\right) \left(1 - q_{m,o}^i\right) \\
& + q_{m,o}^i q_{m,n}^j q_{n,o}^k \left(1 - p_m\right) \left(1 - p_n\right) \left(1 - p_o\right) \\
& + p_m p_n p_o q_{m,n}^j q_{m,o}^i q_{n,o}^k,
\end{aligned}
$$

which we have obtained by going over all combinations of errors leading to three non-trivial defects, analogously to Eq. (7.8), since $\langle d_i d_j d_k \rangle = P\left(d_i = 1, d_j = 1, d_k = 1\right)$. We then introduce $\mathscr{D}_{i,j,k} = \left| \langle d_i d_j d_k \rangle - \langle d_i d_j d_k \rangle_{\text{edge}} \right|$ as the *discrepancy* between the observed $\langle d_i d_j d_k \rangle$ that is calculated from the measured syndrome defects and the predicted $\langle d_i d_j d_k \rangle_{\text{edge}}$ for a model that does not consider a hyperedge between nodes v_i, v_j, and v_k. Calculating $\langle d_i d_j d_k \rangle_{\text{edge}}$ involves solving a system of six equations for each of the expectation values $\langle d_i d_j \rangle$, $\langle d_i d_k \rangle$, $\langle d_j d_k \rangle$, $\langle d_i \rangle$, $\langle d_j \rangle$, and $\langle d_k \rangle$ that involves six variables, which we solve numerically using a least-squares algorithm. The expressions for each of these expectation values can be derived analogously to the equation for $\langle d_i d_j d_k \rangle_{\text{edge}}$ by considering the possible combinations that would lead to only one or two defects. If the hypothesis that hyperedges do indeed not exist is indeed true, then one would observe $\mathscr{D}_{i,j,k} \approx 0$. In turn, finding a $\mathscr{D}_{i,j,k} > 0$ suggests that there is a hyperedge connecting at least these three nodes and possibly even more nodes beyond that. We show an example of the discrepancies extracted from experimental data that involve defect d_i measured by ancilla qubit A_4 at round $r = 4$, i.e., $i = (A_4, 4)$, and any other pair of defects in Fig. 7.5. For comparison, the same discrepancies extracted from the simulated data for the same number of shots are all $\mathscr{D}_{i,j,k} \lesssim 10^{-3}$ (not shown in the figure), unless we specifically consider some crosstalk error which, say, with some probability, induces a pair of correlated X errors on two data qubits that do not share a neighbor. In particular, if one of those data qubits is on the boundary of the code, this error would lead to exactly three defects. In that case, we only observe a significant discrepancy exactly between those defects, demonstrating that this technique can, in principle, accurately detect the presence of errors associated with hyperedges. In contrast, we see multiple $\mathscr{D}_{i,j,k} > 10^{-3}$ extracted from the experimental data, with the largest values being $\approx 2 \times 10^{-2}$, indicating the presence of many errors that lead to *at least* three non-trivial defects in our experiment. These hyperedges would mainly involve the nodes corresponding to defects measured by A_4 at other rounds $r \neq 4$ or to the defects measured

by the two ancilla qubits A_3 and A_4. However, we also observe discrepancies involving the defects measured by the next-nearest neighboring ancilla qubit A_6, corresponding to crosstalk errors. We note that we observe a similar number of significant discrepancies when considering defects d_i other than the one considered in this example.

Now, let us assume that the hyperedge $h_l = \{v_i, v_j, v_k\}$ corresponding to some error that triggers the defects d_i, d_j, d_k exists in the hypergraph and occurs with a probability p_l. Including this edge in our analysis and considering the possible error event combinations that can lead to three non-trivial defects, we find

$$
\begin{aligned}
\langle d_i d_j d_k \rangle = {} & p_l \left(1 - p_m\right)\left(1 - p_n\right)\left(1 - p_0\right)\left(1 - q^i_{m,o,l}\right)\left(1 - q^j_{m,n,l}\right)\left(1 - q^k_{n,o,l}\right) \\
& + p_m q^k_{n,o,l}\left(1 - p_l\right)\left(1 - p_n\right)\left(1 - p_o\right)\left(1 - q^i_{m,o,l}\right)\left(1 - q^j_{m,n,l}\right) \\
& + p_n q^i_{m,o,l}\left(1 - p_l\right)\left(1 - p_m\right)\left(1 - p_o\right)\left(1 - q^k_{n,o,l}\right)\left(1 - q^j_{m,n,l}\right) \\
& + p_o q^j_{m,n,l}\left(1 - p_l\right)\left(1 - p_m\right)\left(1 - p_n\right)\left(1 - q^i_{m,o,l}\right)\left(1 - q^k_{n,o,l}\right) \\
& + p_m p_n q^j_{m,n,l}\left(1 - p_l\right)\left(1 - p_o\right)\left(1 - q^i_{m,o,l}\right)\left(1 - q^k_{n,o,l}\right) \\
& + p_m p_o q^i_{m,o,l}\left(1 - p_l\right)\left(1 - p_n\right)\left(1 - q^j_{m,n,l}\right)\left(1 - q^k_{n,o,l}\right) \\
& + p_n p_o q^k_{n,o,l}\left(1 - p_l\right)\left(1 - p_m\right)\left(1 - q^j_{m,n,l}\right)\left(1 - q^i_{m,o,l}\right) \\
& + q^i_{m,o,l} q^j_{m,n,l} q^k_{n,o,l}\left(1 - p_l\right)\left(1 - p_m\right)\left(1 - p_n\right)\left(1 - p_o\right) \\
& + p_l p_m q^i_{m,o,l} q^j_{m,n,l}\left(1 - p_n\right)\left(1 - p_o\right)\left(1 - q^k_{n,o,l}\right) \\
& + p_l p_n q^j_{m,n,l} q^k_{n,o,l}\left(1 - p_m\right)\left(1 - p_o\right)\left(1 - q^i_{m,o,l}\right) \\
& + p_l p_o q^i_{m,o,l} q^k_{n,o,l}\left(1 - p_m\right)\left(1 - p_n\right)\left(1 - q^j_{m,n,l}\right) \\
& + p_l p_m p_n p_o\left(1 - q^i_{m,o,l}\right)\left(1 - q^j_{m,n,l}\right)\left(1 - q^k_{n,o,l}\right) \\
& + p_l p_m p_n q^i_{m,o,l} q^k_{n,o,l}\left(1 - p_o\right)\left(1 - q^j_{m,n,l}\right) \\
& + p_l p_m p_o q^j_{m,n,l} q^k_{n,o,l}\left(1 - p_n\right)\left(1 - q^i_{m,o,l}\right) \\
& + p_l p_n p_o q^i_{m,o,l} q^j_{m,n,l}\left(1 - p_m\right)\left(1 - q^k_{n,o,l}\right) \\
& + p_m p_n p_o q^j_{m,n,l} q^i_{m,o,l} q^k_{n,o,l}\left(1 - p_l\right),
\end{aligned}
$$

where we have once again extended the definition of $q^i_{m,n}$ to now exclude three edges, leading to $q^i_{m,n,l}$. Calculating the hyperedge probabilities involves numerically solving the system of seven equations for each of the seven expectation values $\langle d_i d_j d_k \rangle$, $\langle d_i d_j \rangle$, $\langle d_i d_k \rangle$, $\langle d_j d_k \rangle$, $\langle d_i \rangle$, $\langle d_j \rangle$, and $\langle d_k \rangle$ that are expressed in terms of seven variables. When the decoding graph has multiple hyperedges, the $q^i_{m,n,l}$ obtained by solving this system of equations will also include the probabilities of the hyperedges that trigger d_i together with any other defects, excluding d_j and d_k. Similarly, the estimated p_m will not correspond to the error event probability associated with the edge h_m, but will account for

hyperedges that lead to both $d_i = 1$ and $d_j = 1$ along with any other defects, excluding d_k. Therefore, solving this system of equations can be used to obtain the probability p_l associated with the hyperedge h_l. We can estimate all the size-3 hyperedge probabilities by considering the possible triples of defects, which can then be accounted for when solving for the edge probabilities, following the procedure outlined in [28].

It is also possible to observe $\mathcal{D}_{i,j,k} > 0$ when there is instead a hyperedge of size four or more that connects nodes v_i, v_j and v_k. However, a caveat to considering larger hyperedges is that a system of $2^s - 1$ equations has to be solved to estimate the hyperedge probability, where s is the size of the hyperedge. Considering all possible hyperedges of size s in the decoding graph leads to a total of $\binom{N}{s}$ hyperedge probabilities to be estimated. This quickly becomes too computationally expensive for hyperedges of size $s \geq 4$ when considering experiments executing more than just a few rounds of error correction. Another caveat of this approach is that some errors might not deterministically lead to certain non-trivial defects. For example, a leaked qubit can remain outside the computational subspace for several QEC rounds and lead to potentially many non-trivial defects measured by the neighboring ancilla qubits, each happening with some probability. This means that leakage errors are associated with a set of possible defects, each occurring with some probability. Therefore, extending this analysis to estimate the probability of leakage errors is more complex than considering all possible hyperedges up to some given size. In particular, keeping this analysis computationally feasible would require including leakage-reduction operations that limit the lifetime of leakage events to only a few rounds, thereby limiting the number of correlated non-trivial defects that a single leakage event can lead to.

7

BIBLIOGRAPHY

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

[2] A. R. Calderbank and P. W. Shor, *Good quantum error-correcting codes exist*, Phys. Rev. A **54**, 1098–1105 (1996).

[3] A. M. Steane, *Error correcting codes in quantum theory*, Phys. Rev. Lett. **77**, 793 (1996).

[4] S. B. Bravyi and A. Y. Kitaev, *Quantum codes on a lattice with boundary*, arXiv:quant-ph/9811052 (1998).

[5] E. Knill, R. Laflamme, and W. H. Zurek, *Resilient quantum computation*, Science **279**, 342–345 (1998).

[6] R. Raussendorf and J. Harrington, *Fault-tolerant quantum computation with high threshold in two dimensions*, Phys. Rev. Lett. **98**, 190504 (2007).

[7] D. Aharonov and M. Ben-Or, *Fault-tolerant quantum computation with constant error rate*, SIAM Journal on Computing **38**, 1207–1282 (2008).

[8] D. Gottesman, *Fault-tolerant quantum computation with constant overhead*, Quantum Info. Comput. **14**, 1338–1372 (2014).

[9] B. M. Terhal, *Quantum error correction for quantum memories*, Rev. Mod. Phys. **87**, 307–346 (2015).

[10] 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).

[11] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, *Elucidating reaction mechanisms on quantum computers*, Proceedings of the National Academy of Sciences **114**, 7555–7560 (2017).

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

[13] D. Gottesman, *Stabilizer codes and quantum error correction* (California Institute of Technology, 1997).

[14] A. Kitaev, *Fault-tolerant quantum computation by anyons*, Annals of Physics **303**, 2–30 (2003).

[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**, 66–69 (2015).

[16] M. H. Abobeih, Y. Wang, J. Randall, S. J. H. Loenen, C. E. Bradley, M. Markham, D. J. Twitchen, B. M. Terhal, and T. H. Taminiau, *Fault-tolerant operation of a logical qubit in a diamond quantum processor*, Nature **606**, 884–889 (2022).

[17] S. Krinner, N. Lacroix, A. Remm, A. Di Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, *Realizing repeated quantum error correction in a distance-three surface code*, Nature **605**, 669–674 (2022).

[18] Y. Zhao, Y. Ye, H.-L. Huang, Y. Zhang, D. Wu, H. Guan, Q. Zhu, Z. Wei, T. He, S. Cao, F. Chen, T.-H. Chung, H. Deng, D. Fan, M. Gong, C. Guo, S. Guo, L. Han, N. Li, S. Li, Y. Li, F. Liang, J. Lin, H. Qian, H. Rong, H. Su, L. Sun, S. Wang, Y. Wu, Y. Xu, C. Ying, J. Yu, C. Zha, K. Zhang, Y.-H. Huo, C.-Y. Lu, C.-Z. Peng, X. Zhu, and J.-W. Pan, *Realization of an error-correcting surface code with superconducting qubits*, Phys. Rev. Lett. **129**, 030501 (2022).

[19] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush, D. Bacon, J. C. Bardin, J. Basso, A. Bengtsson, S. Boixo, G. Bortoli, A. Bourassa, A. Bovaird, L. Brill, M. Broughton, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, J. Cogan, R. Collins, P. Conner, W. Courtney, A. L. Crook, B. Curtin, D. M. Debroy, A. Del Toro Barba, S. Demura, A. Dunsworth, D. Eppens, C. Erickson, L. Faoro, E. Farhi, R. Fatemi, L. Flores Burgos, E. Forati, A. G. Fowler, B. Foxen, W. Giang, C. Gidney, D. Gilboa, M. Giustina, A. Grajales Dau, J. A. Gross, S. Habegger, M. C. Hamilton, M. P. Harrigan, S. D. Harrington, O. Higgott, J. Hilton, M. Hoffmann, S. Hong, T. Huang, A. Huff, W. J. Huggins, L. B. Ioffe, S. V. Isakov, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, P. Juhas, D. Kafri, K. Kechedzhi, J. Kelly, T. Khattar, M. Khezri, M. Kieferová, S. Kim, A. Kitaev, P. V. Klimov, A. R. Klots, A. N. Korotkov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, K.-M. Lau, L. Laws, J. Lee, K. Lee, B. J. Lester, A. Lill, W. Liu, A. Locharla, E. Lucero, F. D. Malone, J. Marshall, O. Martin, J. R. McClean, T. McCourt, M. McEwen, A. Megrant, B. Meurer Costa, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, A. Morvan, E. Mount, W. Mruczkiewicz, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, H. Neven, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, M. Y. Niu, T. E. O'Brien, A. Opremcak, J. Platt, A. Petukhov, R. Potter, L. P. Pryadko, C. Quintana, P. Roushan, N. C. Rubin, N. Saei, D. Sank, K. Sankaragomathi, K. J. Satzinger, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, V. Smelyanskiy, W. C. Smith, G. Sterling, D. Strain, M. Szalay, A. Torres, G. Vidal, B. Villalonga, C. Vollgraff Heidweiller, T. White, C. Xing, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, Y. Zhang, N. Zhu, and G. Q. AI, *Suppressing quantum errors by scaling a surface code logical qubit*, Nature **614**, 676–681 (2023).

[20] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, *Demonstrating multi-round subsystem quantum error correction using matching and maximum likelihood decoders*, Nature Communications **14**, 2852 (2023).

[21] S. A. Moses, C. H. Baldwin, M. S. Allman, R. Ancona, L. Ascarrunz, C. Barnes, J. Bartolotta, B. Bjork, P. Blanchard, M. Bohn, J. G. Bohnet, N. C. Brown, N. Q. Burdick, W. C. Burton, S. L. Campbell, J. P. C. I. au2, C. Carron, J. Chambers, J. W. Chan, Y. H. Chen, A. Chernoguzov, E. Chertkov, J. Colina, J. P. Curtis, R. Daniel, M. DeCross, D. Deen, C. Delaney, J. M. Dreiling, C. T. Ertsgaard, J. Esposito, B. Estey, M. Fabrikant, C. Figgatt, C. Foltz, M. Foss-Feig, D. Francois, J. P. Gaebler, T. M. Gatterman, C. N. Gilbreth, J. Giles, E. Glynn, A. Hall, A. M. Hankin, A. Hansen, D. Hayes, B. Higashi, I. M. Hoffman, B. Horning, J. J. Hout, R. Jacobs, J. Johansen, L. Jones, J. Karcz, T. Klein, P. Lauria, P. Lee, D. Liefer, C. Lytle, S. T. Lu, D. Lucchetti, A. Malm, M. Matheny, B. Mathewson, K. Mayer, D. B. Miller, M. Mills, B. Neyenhuis, L. Nugent, S. Olson, J. Parks, G. N. Price, Z. Price, M. Pugh, A. Ransford, A. P. Reed, C. Roman, M. Rowe, C. Ryan-Anderson, S. Sanders, J. Sedlacek, P. Shevchuk, P. Siegfried, T. Skripka, B. Spaun, R. T. Sprenkle, R. P. Stutz, M. Swallows, R. I. Tobey, A. Tran, T. Tran, E. Vogt, C. Volin, J. Walker, A. M. Zolot, and J. M. Pino, *A race track trapped-ion quantum processor*, 2023, arXiv:2305.03828 [quant-ph].

[22] 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).

[23] M. Takita, A. D. Córcoles, E. Magesan, B. Abdo, M. Brink, A. Cross, et al., *Demonstration of weight-four parity measurements in the surface code architecture*, Phys. Rev. Lett. **117**, 210505 (2016).

[24] 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**, 527–531 (2018).

[25] 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).

[26] C. Ryan-Anderson, J. G. Bohnet, K. Lee, D. Gresh, A. Hankin, J. P. Gaebler, D. Francois, A. Chernoguzov, D. Lucchetti, N. C. Brown, T. M. Gatterman, S. K. Halit, K. Gilmore, J. A. Gerber, B. Neyenhuis, D. Hayes, and R. P. Stutz, *Realization of real-time fault-tolerant quantum error correction*, Phys. Rev. X **11**, 041058 (2021).

[27] 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,

7

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,** 383–387 (2021).

[28] E. H. Chen, T. J. Yoder, Y. Kim, N. Sundaresan, S. Srinivasan, M. Li, A. D. Córcoles, A. W. Cross, and M. Takita, *Calibrated decoders for experimental quantum error correction*, Phys. Rev. Lett. **128,** 110504 (2022).

[29] 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*, Nat. Phys. **18**, 80–86 (2022).

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

[31] A. G. Fowler, A. C. Whiteside, and L. C. L. Hollenberg, *Towards practical classical processing for the surface code*, Phys. Rev. Lett. **108,** 180501 (2012).

[32] A. G. Fowler, *Minimum weight perfect matching of fault-tolerant topological quantum error correction in average o(1) parallel time*, Quantum Info. Comput. **15,** 145–158 (2015).

[33] O. Higgott and C. Gidney, *Sparse blossom: correcting a million errors per core second with minimum-weight matching*, 2023, arXiv:2303.15933 [quant-ph].

[34] Y. Wu and L. Zhong, *Fusion blossom: fast mwpm decoders for qec*, 2023, arXiv:2305.08307 [quant-ph].

[35] M. Rispler, P. Cerfontaine, V. Langrock, and B. M. Terhal, *Towards a realistic gaas-spin qubit device for a classical error-corrected quantum memory*, Phys. Rev. A **102**, 022416 (2020).

[36] J. R. Wootton, *Benchmarking near-term devices with quantum error correction*, Quantum Science and Technology **5**, 044004 (2020).

[37] D. G. Cory, M. D. Price, W. Maas, E. Knill, R. Laflamme, W. H. Zurek, T. F. Havel, and S. S. Somaroo, *Experimental quantum error correction*, Phys. Rev. Lett. **81,** 2152 (1998).

[38] O. Moussa, J. Baugh, C. A. Ryan, and R. Laflamme, *Demonstration of sufficient control for two rounds of quantum error correction in a solid state ensemble quantum information processor*, Phys. Rev. Lett. **107,** 160501 (2011).

[39] P. Schindler, J. T. Barreiro, T. Monz, V. Nebendahl, D. Nigg, M. Chwalla, M. Hennrich, and R. Blatt, *Experimental repetitive quantum error correction*, Science **332,** 1059–1061 (2011).

[40] J. Zhang, M.-H. Yung, R. Laflamme, A. Aspuru-Guzik, and J. Baugh, *Digital quantum simulation of the statistical mechanics of a frustrated magnet*, Nat. Commun. **3,** 880 (2012).

[41] M. D. Reed, L. DiCarlo, S. E. Nigg, L. Sun, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, *Realization of three-qubit quantum error correction with superconducting circuits*, Nature **482,** 382–385 (2012).

7

[42] G. Waldherr, Y. Wang, S. Zaiser, M. Jamali, T. Schulte-Herbrüggen, H. Abe, T. Ohshima, J. Isoya, J. F. Du, P. Neumann, and J. Wrachtrup, *Quantum error correction in a solid-state hybrid spin register*, Nature **506**, 204–207 (2014).

[43] J. Cramer, N. Kalb, M. A. Rol, B. Hensen, M. S. Blok, M. Markham, D. J. Twitchen, R. Hanson, and T. H. Taminiau, *Repeated quantum error correction on a continuously encoded qubit by real-time feedback*, Nature Communications **7**, 11526 (2016).

[44] J. R. Wootton and D. Loss, *Repetition code of 15 qubits*, Phys. Rev. A **97**, 052313 (2018).

[45] 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).

[46] 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).

[47] 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**, 1761 (2021).

[48] K. C. Miao, M. McEwen, J. Atalaya, D. Kafri, L. P. Pryadko, A. Bengtsson, A. Opremcak, K. J. Satzinger, Z. Chen, P. V. Klimov, C. Quintana, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, P. Conner, A. L. Crook, B. Curtin, D. M. Debroy, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, A. G. Fowler, B. Foxen, G. Garcia, W. Giang, C. Gidney, M. Giustina, R. Gosula, A. G. Dau, J. A. Gross, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, J. Kelly, S. Kim, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J. Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S. Martin, A. Megrant, X. Mi, S. Montazeri, A. Morvan, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Rocque, P. Roushan, K. Sankaragomathi, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, N. Shutty, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, T. White, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, N. Zhu, N. Zobrist, H. Neven, V. Smelyanskiy, A. Petukhov, A. N. Korotkov, D. Sank, and Y. Chen, *Overcoming leakage in quantum error correction*, Nature Physics **19**, 1780–1786 (2023).

7

[49] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. Di-Carlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Information **6**, 102 (2020).

[50] J. F. Marques, H. Ali, B. M. Varbanov, M. Finkel, H. M. Veen, S. L. M. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider, B. M. Ter-hal, and L. DiCarlo, *All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits*, Phys. Rev. Lett. **130**, 250602 (2023).

[51] S. Spitz, B. M. Tarasinski, C. Beenakker, and T. O'Brien, *Adaptive weight estimator for quantum error correction in a time-dependent environment*, Advanced Quantum Technologies **1**, 1800012 (2018).

[52] T. E. O'Brien, B. M. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Information **3** (2017).

[53] 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).

[54] O. Higgott, T. C. Bohdanowicz, A. Kubica, S. T. Flammia, and E. T. Campbell, *Improved decoding of circuit noise and fragile boundaries of tailored surface codes*, Phys. Rev. X **13**, 031007 (2023).

[55] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. J. Beenakker, *Machine-learning-assisted correction of correlated qubit errors in a topological code*, Quantum **2**, 48 (2018).

[56] M. Lange, P. Havström, B. Srivastava, V. Bergentall, K. Hammar, O. Heuts, E. van Nieuwenburg, and M. Granath, *Data-driven decoding of quantum error correcting codes using graph neural networks*, 2023, arXiv:2307.01241 [quant-ph].

[57] B. M. Varbanov, M. Serra-Peralta, D. Byfield, and B. M. Terhal, *Neural network decoder for near-term surface-code experiments*, 2023, arXiv:2307.03280 [quant-ph].

[58] C. Chamberland and P. Ronagh, *Deep neural decoders for near term fault-tolerant experiments*, Quantum Science and Technology **3**, 044002 (2018).

[59] P. Baireuther, M. D. Caio, B. Criger, C. W. J. Beenakker, and T. E. O'Brien, *Neural network decoder for topological color codes with circuit level noise*, New Journal of Physics **21**, 013003 (2019).

[60] S. Varsamopoulos, K. Bertels, and C. Almudever, *Comparing neural network based decoders for the surface code*, IEEE Transactions on Computers **69**, 300–311 (2020).

[61] R. Sweke, M. S. Kesselring, E. P. L. van Nieuwenburg, and J. Eisert, *Reinforcement learning decoders for fault-tolerant quantum computation*, Machine Learning: Science and Technology **2**, 025005 (2020).

[62] K. Meinerz, C.-Y. Park, and S. Trebst, *Scalable neural decoder for topological surface codes*, Phys. Rev. Lett. **128**, 080505 (2022).

[63] Y. Ueno, M. Kondo, M. Tanaka, Y. Suzuki, and Y. Tabuchi, *Neo-qec: neural network enhanced online superconducting decoder for surface codes*, 2022, arXiv:2208.05758 [quant-ph].

[64] C. Chamberland, L. Goncalves, P. Sivarajah, E. Peterson, and S. Grimberg, *Techniques for combining fast local decoders with global decoders under circuit-level noise*, Quantum Science and Technology **8**, 045011 (2023).

[65] M. Zhang, X. Ren, G. Xi, Z. Zhang, Q. Yu, F. Liu, H. Zhang, S. Zhang, and Y.-C. Zheng, *A scalable, fast and programmable neural decoder for fault-tolerant quantum computation using surface codes*, 2023, arXiv:2305.15767 [quant-ph].

[66] 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. Applied **8**, 034021 (2017).

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

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

[70] 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).

[71] 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).

[72] C. Gidney, *Stim: a fast stabilizer circuit simulator*, Quantum **5,** 497 (2021).

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

[74] J. Mulder, *Estimation of error probabilities for the quantum repetition code* (Delft University of Technology, 2022).

7

8

NEURAL NETWORK DECODER FOR NEAR-TERM SURFACE-CODE EXPERIMENTS

Neural-network decoders can achieve a lower logical error rate compared to conventional decoders, like minimum-weight perfect matching, when decoding the surface code. Furthermore, these decoders require no prior information about the physical error rates, making them highly adaptable. In this chapter, we investigate the performance of such a decoder using both simulated and experimental data obtained from a transmon-qubit processor, focusing on small-distance surface codes. We first show that the neural network typically outperforms the matching decoder due to better handling errors leading to multiple correlated syndrome defects, such as Y errors. When applied to the experimental data of Ref. [2], the neural network decoder achieves logical error rates approximately 25% lower than minimum-weight perfect matching, approaching the performance of a maximum-likelihood decoder. To demonstrate the flexibility of this decoder, we incorporate the soft information available in the analog readout of transmon qubits and evaluate the performance of this decoder in simulation using a symmetric Gaussian-noise model. Considering the soft information leads to an approximately 10% lower logical error rate, depending on the probability of a measurement error. The good logical performance, flexibility, and computational efficiency make neural network decoders well-suited for near-term demonstrations of quantum memories.

8.1. Introduction

Quantum computers are anticipated to outperform classical computers in solving specific problems, such as integer factorization [3] and quantum simulation [4]. However, for a quantum computer to perform any meaningful computation, it has to be able to execute millions of operations, requiring error rates per operation lower than 10^{-10} [5, 6]. Despite a valiant experimental effort aimed at enhancing operational performance, state-of-the-art processors typically exhibit error rates per operation around 10^{-3} [7–16], which is far from what is needed to perform any useful computation.

Fortunately, quantum error correction (QEC) provides a means to reduce the error rates, albeit at the cost of additional overhead in the required physical qubits [17–20]. Two-dimensional stabilizer codes [21], such as the surface codes [22], have emerged as a prominent approach to realizing fault-tolerant computation due to their modest connectivity requirements and high tolerance to errors [23–25]. These codes encode the logical information into an array of physical qubits, referred to as data qubits. Ancilla qubits are used to repeatedly measure parities of sets of neighboring data qubits. Changes between consecutive measurement outcomes, which are typically referred to as syndrome defects, indicate that errors have occurred. A classical decoder processes this information and aims at inferring the most likely correction.

The increased number of available qubits [2, 26–28] and the higher fidelities of physical operations [7–16, 29–34] in modern processors have enabled several experiments employing small-distance codes to demonstrate the capacity to detect and correct errors [2, 28, 35–47]. In a recent milestone experiment, the error rate per QEC round of a surface-code logical qubit was reduced by increasing the code distance [2], demonstrating the fundamental suppression achieved by QEC.

The performance of the decoder directly influences the performance of a QEC code. Minimum-weight perfect matching (MWPM) is a good decoding algorithm for the surface code, which is computationally efficient and, therefore, scalable [23, 48–51]. Its good performance is ensured under the assumption that the errors occurring in the experiment can be modeled as independent X and Z errors [23]. This leads to the MWPM decoder performing worse than decoders based on belief propagation [52–55] or a (more computationally-expensive) approximate maximum-likelihood decoder based on tensor-network (TN) contraction [56, 57]. A more practical concern is that a decoder relies on a physical error model to accurately infer the most likely correction. Typically, this requires constructing an approximate model and a series of benchmarking experiments to extract the physical error rates. While there are methods to estimate the physical error rates based on the measured defects [2, 44, 58, 59], they typically ignore non-conventional errors like crosstalk or leakage. The presence of these errors can impact both the accuracy with which the physical error rates are estimated from the data and the performance of the decoder itself [59].

An alternative approach to decoding is based on using neural networks (NN) to infer the most likely correction given a set of measured defects [60–80]. These decoders do not require any prior information about the error model and therefore alleviate the need to construct any error model, making them highly adaptable. This flexibility comes at the cost of requiring a significant amount of data for training the network and optimizing the hyper-parameters to ensure that the optimal performance of the decoder

is reached during training. Despite the potential issues during the training, it has been shown that they can match and generally exceed the performance of MWPM decoders, in several cases achieving near-optimal performance [75, 77]. Depending on the NN architecture employed, these decoders can be scalable and run in real time [68–70, 72, 79]. While decoders based on recurrent NNs are more computationally expensive, they enable the decoding of experiments performing a variable number of stabilizer measurement rounds [62, 75, 77], making them well-suited for decoding near-term memory [75] and stability experiments [81].

In this chapter, we assess the performance of a neural-network decoder using both simulated and experimental data. Our work goes beyond [75] and previous NN decoding works in applying and partially training a NN decoder for the first time on data from a surface-code experiment [2], thus capturing realistic performance and showing the versatility of NN decoders. In addition, we go beyond [75] in training the NN decoder for a distance-7 surface code and extract its exponential error suppression factor on simulated data. Thirdly, we show that our NN decoder can be trained with (simulated) soft measurement data and get a performance enhancement.

We begin by simulating the performance of a $d = 3$ surface code using a circuit-level noise model to show that the NN decoder outperforms MWPM by learning to deal with Y errors, as previous studies have suggested [75].

Next, we investigate the performance of the NN decoder when applied to data from a recent surface code experiment [2]. Due to the limited volume of available experimental data, we train the NN decoder on simulated data generated using an error model based on the measured physical error rates. However, we evaluate the decoder's performance on simulated and experimental data. The NN decoder significantly outperforms MWPM when decoding simulated data and achieves a lower logical error rate for the $d = 5$ code than the constituent $d = 3$ codes. When evaluated on experimental data, the NN decoder achieves a performance approaching that of a tensor-network decoder, which approximates a maximum-likelihood decoder. However, contrary to the finding in [2], the logical error rate observed in the $d = 5$ experiment is higher than the average of each of the $d = 3$ experiments, which we attribute to either a sub-optimal choice of hyperparameters or the mismatch between the simulated data that the decoder was trained on and the experimental data.

To further explore the performance of NNs, we consider the continuous information available in the measurement outcomes of transmon qubits [82, 83], typically referred to as soft information [84]. By calculating the defect probabilities given the soft outcomes and providing them to the neural network during training and evaluation, we demonstrate that the soft decoder can achieve an approximately 10% lower logical error rate if the measurement error probability is sufficiently high.

8.2. Background

8.2.1. The surface code

A (rotated) surface code encodes a single logical qubit into a two-dimensional array of $n = d \times d$ physical qubits, referred to as data qubits, where d is the distance of the code. The logical state of the qubit is determined by the stabilizers of the code, which
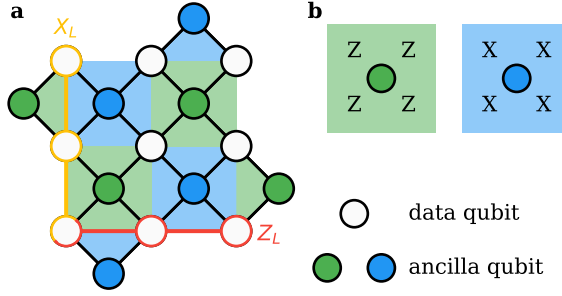
Figure 8.1: **a** Schematic of a distance $d = 3$ surface-code logical qubit, where 9 data qubits (white circles) store the logical information. 8 ancilla qubits (blue and green circles) are used to measure the Z-type (green plaquettes) and X-type (blue plaquettes) stabilizers of the code. Examples of the X_L (yellow) and Z_L (red) logical operators of the code. **b** Illustration of the Z-type plaquette (left, green) and X-type (right, blue) plaquette corresponding to the $ZZZZ$ and $XXXX$ stabilizer operators measured by each ancilla qubit.

are the weight-four or weight-two X-type (blue plaquettes) or Z-type (green plaquettes) Pauli operators, see Fig. 8.1. In addition to the stabilizers, the code is given by a pair of anti-commuting logical operators, X_L and Z_L, which commute with the code stabilizers. The stabilizers are typically measured indirectly with the help of $n-1$ ancilla qubits. To perform this measurement, each ancilla coherently interacts with its neighboring data qubits in a specific order [85], after which the ancilla qubit is measured and reset. The stabilizer measurement outcomes are typically referred to as the syndromes and hold information about the errors that have occurred. The full circuits used to perform these measurements are shown in Fig. 8.8. In particular, we use the circuits used in [2], which feature several echo gates used for dynamical decoupling in the experiment, see Sec. 8.5.1 for additional details.

To characterize the performance of the code, we perform a series of logical memory experiments. In each experiment, the physical qubits are prepared in an eigenstate of either the X_L (resp. Z_L) logical operator, after which $N-1$ rounds of stabilizer measurements are executed. The experiment is concluded by reading out each data qubit in the X (resp. Z basis), which also performs a logical X_L (resp. Z_L) measurement. The goal of each experiment is to maintain the logical state for as many QEC rounds as possible by using error correction, see Sec. 8.5.1 for more details.

The information about errors is contained in the stabilizer measurement outcome $m_{r,a}$ of ancilla a at round r. The final data qubit measurements can also be used to infer a final set of outcomes $m_{r=N,a}$ for either the X-type or Z-type stabilizers. The defects $d_{r,a} = m_{r,a} \oplus m_{r-1,a}$ isolate the changes in $m_{r,a}$ such that an error is signaled by an observation of one or more $d_{r,a} = 1$. The choice of initial state and the dynamical decoupling gates can also flip some of the measured $m_{r,a}$, which is accounted for when calculating $d_{r,a}$. A decoder processes the observed $d_{r,a}$ to infer a correction for the measured logical observable. By repeating each experiment many times, we extract the probability of a logical error $p_L(r)$ at QEC round r, from which we calculate the logical fidelity $F_L(r) = 1 - 2p_L(r)$, which decays exponentially with the number of executed QEC rounds. We model this decay as $F_L(r) = (1 - 2\varepsilon_L)^{r-r_0}$, where ε_L is the logical error

rate per QEC round and r_0 is a fitting constant. When fitting the decay of $F_L(r)$ to extract ε_L, we start the fit at $r = 3$ to avoid any time-boundary effects that might impact this estimate.

8.2.2. ERROR MODELS

To explore the performance of the NN decoder, we perform simulations using circuit-level Pauli-noise models. For most of our simulations, we consider a depolarizing circuit-level noise, which is defined as

1. After each single-qubit gate or idling period, with a probability $p/3$, we apply an error drawn from $\{X, Y, Z\}$.

2. After each two-qubit gate, with a probability $p/15$, we apply an error drawn from $\{I, X, Y, Z\}^{\otimes 2} \setminus \{II\}$.

3. With a probability p, we apply an X error before each measurement.

4. With a probability p, we apply an X error after each reset operation or after the qubits are first prepared at the start of an experiment.

In some of our simulations, we consider noise models that are biased to have a higher or a lower probability of applying Y errors. To construct this model, we define a Y-bias factor η and modify the standard depolarizing circuit-level noise model, as follows:

1. After each single-qubit gate or idling period, there is a probability $\eta p/(\eta + 2)$ to apply a Y error and a probability $p/(\eta + 2)$ to apply an X or a Z error.

2. After each two-qubit gate, there is a probability $\eta p/(7\eta + 8)$ of applying an error drawn from $\mathscr{P}_B = \{IY, XY, YI, YX, YY, YZ, ZY\}$ and a probability $p/(7\eta + 8)$ of applying an error drawn from $\{I, X, Y, Z\}^{\otimes 2} \setminus (\mathscr{P}_B \cup \{II\})$.

This biased error model is a generalization of the depolarizing model. In particular, choosing $\eta = 1$ makes this noise model equivalent to the depolarizing one. On the other hand, when $\eta = 0$, the model leads to only X or Z errors applied after operations. In the other limiting case, as $\eta \to \infty$, the model applies only Y errors after idling periods and gates. Given that the error probability is the same across all operations of the same type, we will refer to these error models as uniform circuit-level noise models.

Finally, we also perform simulations of the recent experiment conducted by Google Quantum AI, using the error model which they provided together with the experimental data [2]. This is once again a circuit-level Pauli-noise model similar to the ones presented above, but the probability of a depolarizing error after each operation is based on the measured physical error rates. We will refer to this model as the experimental circuit-level noise model.

We use *stim* [86] to perform the stabilizer simulations. We have written a wrapper package that helps with constructing the circuit for each experiment, which is available in [87]. We use *pymatching* [50] for the MWPM decoding. The weights used in the MWPM decoder are directly extracted from the sampled circuit using the built-in integration between *stim* and *pymatching*

8

8.2.3. NEURAL NETWORK ARCHITECTURE

Here we describe the NN architecture that we employ in this chapter, which nearly exactly follows the one proposed in [75, 77]. Many NN decoders studied previously are based on feed-forward or convolutional NN architecture. These decoders can generally decode experiments running a fixed number of QEC rounds. Decoders based on recurrent NN architectures, on the other hand, can learn the temporal correlations in the data, allowing them to directly process experiments performing a variable number of QEC rounds. We have used the *TensorFlow* library [88] to implement the NN architecture, with the source code of the decoder available in [89], the parameters used for each training are listed in Tab. 8.1, while the scripts that perform the training are available upon request.

The NN architecture takes as input the defects $d_{a,r}$ with $r = 1, 2, \ldots, N$. The decoder solves a binary classification problem and determines whether a correction of the logical observable is required based on the observed defects. In practice, the architecture is based on a two-headed network that makes two predictions p_{main} and p_{aux}, which are used to improve the training of the network, see Fig. 8.2. To train a decoder, a series of memory experiments are performed. Since the logical qubit is prepared in a known logical state and measured at the end of each experiment, it is possible to extract the actual value $p_{\mathrm{true}} \in \{0, 1\}$ of whether a correction is required or not. In particular, the cost function I that the network attempts to minimize during training is the weighted sum of the binary cross-entropies between each prediction and p_{true}, expressed as

$$I = H(p_{\mathrm{main}}, p_{\mathrm{true}}) + w_a H(p_{\mathrm{aux}}, p_{\mathrm{true}}),$$

where w_a is a weight that is typically chosen as $w_a = 0.5$ in our runs, while

$$H(p_i, p_j) = -p_i \log p_j - (1 - p_i) \log(1 - p_j)$$

is the binary cross-entropy function. The choice behind this loss function is elaborated below.
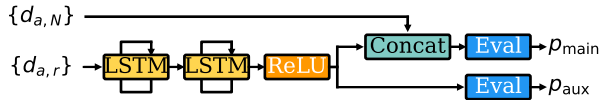

Figure 8.2: Schematic of the recurrent NN architecture used in this chapter, following the design proposed in [77]. The inputs to the network are the set of defects $\{d_{a,r}\}$, which are calculated from the measurement outcomes of each ancilla qubit a at QEC round $r = 1, 2, \ldots, N-1$, and the final defects $\{d_{a,N}\}$, which are inferred from data qubit measurements. The time-invariant input $\{d_{a,r}\}$ is provided to the recurrent part of the network, consisting of two stacked LSTM layers (yellow rectangles) and a ReLU activation layer (orange rectangle). The recurrent output is then passed to the two heads of the decoder, which consist of an evaluation layer (blue rectangle) that predict a probability of a logical error. The lower head takes as input only the recurrent output and outputs a probability p_{aux}. The upper head, on the other hand, combines (teal rectangle) the recurrent output with $\{d_{a,N}\}$ and outputs a probability p_{main}. Arrows indicate the flow of information through the network.

Fig. 8.2 schematically illustrates the architecture of the recurrent network. The recurrent body of the neural network consists of two stacked long short-term memory (LSTM) layers. Each LSTM layer is defined by a pair of internal memory states: a short-term

memory, referred to as the hidden state, and a long-term memory, referred to as the cell state. Here, we use the same internal states size N_L for both LSTM layers [90, 91], with $N_L = 64, 96, 128$ for surface codes of distance $d = 3, 5, 7$, unless otherwise specified. The LSTM layers receive the defects for each QEC round as input, calculated from both the X-type and the Z-type stabilizer measurement outcomes. The first LSTM layer outputs a hidden state for each QEC round, which is then provided as input to the second LSTM layer, which outputs only its final hidden state. A rectified linear unit (ReLU) activation function is applied to the output of the second LSTM layer before being passed along to each of the two heads of the network.

The heads of the network are feed-forward evaluation networks consisting of a single hidden layer of size N_L using the ReLU activation function and an output layer using the sigmoid activation function, which maps the hidden layer output to a probability used for binary classification. The output of the recurrent part of the network is directly passed to the lower head of the network, which uses this information to predict a probability p_{aux} of a logical error. The upper head also considers the defects inferred from the data qubit measurements, which are combined with the recurrent output and provided as input. Therefore, unlike the lower head, the upper one uses the full information about the errors that have occurred when making its prediction p_{main} of whether a logical error occurred. Both p_{main} and p_{aux} are used when training the network, which helps the neural network to generalize to handle longer input sequences. However, only p_{main} is used when evaluating the performance of the decoder. We provide additional details about the training procedure in Sec. 8.5.2 and list the hyper-parameters of the network in Tab. 8.1.

8.3. RESULTS

8.3.1. PERFORMANCE ON CIRCUIT-LEVEL NOISE SIMULATIONS

We first demonstrate that the NN decoder can achieve a lower logical error rate than the MWPM decoder by learning error correlations between the defects, which are otherwise ignored by the MWPM decoder. We consider the Y-biased circuit-level noise model described previously, parameterized by the bias η towards Y errors and a probability $p = 0.001$ of inserting an error after each operation. We use this noise model to simulate the performance of a $d = 3$ surface-code quantum memory experiment in the Z-basis, initially preparing either $|0\rangle^{\otimes n}$ or $|1\rangle^{\otimes n}$. To train the NN decoder, we generated datasets of $r = 1, 5, \ldots, 37$ QEC rounds, sampling 5×10^5 shots for each round and initial state. When evaluating the decoder's performance, we simulate the code performance over $r = 10, 30, \ldots, 290$ QEC rounds and sample 2×10^4 shots instead.

To benchmark the logical performance, we calculate the logical fidelity F_L at the end of each experiment. Averaging F_L over each initial state, we fit the exponential decay of F_L with the number of QEC rounds to extract the logical error rate per round ε_L. Fig. 8.3 shows that the NN decoder maintains a constant ε_L when evaluated on datasets going up to 300 QEC rounds, demonstrating the ability of the decoder to generalize to significantly longer sequences than those used for training. On the other hand, the NN decoder achieves about 20% lower ε_L compared to the MWPM decoder. We then evaluate the trained NN decoder on simulated data using $\eta \in \{0, 0.5, 1, 2, 10, 100\}$ and keep all
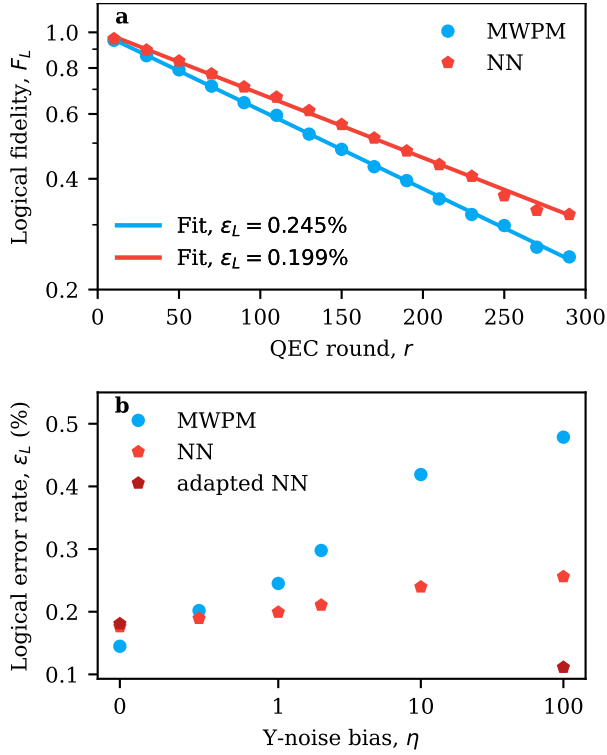
Figure 8.3: **a** Logical fidelity F_L as a function of the number of QEC rounds r for the MWPM (blue) and the NN decoders (red) using a uniform circuit-level depolarizing noise model. Each data point is averaged over 4×10^4 shots. Solid lines show the fits to the data used to extract the logical error rate per round ε_L. **b** The logical error rate ε_L as a function of the bias η towards Y errors for the MWPM decoder (blue) and a NN decoder trained on simulated data using depolarizing noise (red), corresponding to $\eta = 1$. The performance of an adapted NN decoder at a bias of $\eta = 0$ or $\eta = 100$ is shown in dark red. Each point is extracted from a fit of the decay of the logical fidelity over 300 QEC rounds. The error bar is smaller than the marker size. The error bars are smaller than the marker sizes.

other parameters the same without training any new neural networks, with the resulting error rates shown in Fig. 8.3**b**. At $\eta = 0$, corresponding to an error model leading to X and Z errors, the NN decoder displays a higher ε_L than the MWPM decoder. For $\eta \geq 0.5$, the NN decoder instead demonstrates a lower logical error, with the relative reduction increasing with the bias. This demonstrates that the NN decoder can achieve a lower logical error rate by learning the correlations between the defects caused by Y errors, consistent with the results presented in [75]. The NN decoder can achieve an even lower logical error rate at a bias of $\eta = 100$ by being trained on a dataset generated using this bias (referred to as the adapted NN decoder in Fig. 8.3). On the other hand, training a model for $\eta = 0$ does not lead to any improvement in ε_L of the NN decoder, showing that the MWPM decoder is more optimal in this setting.
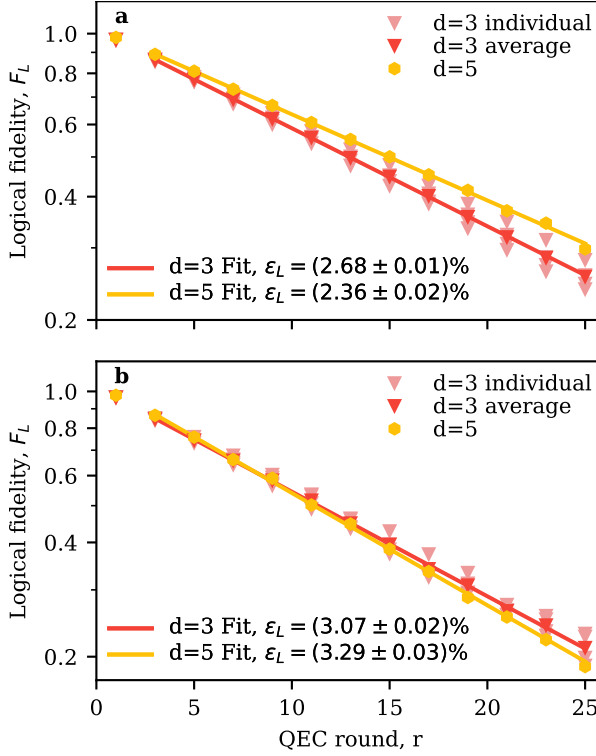
Figure 8.4: Logical fidelity F_L as a function of the number of QEC rounds r for the NN decoder evaluated on simulated data (shown in **a**) and on experimental data (shown in **b**). The average performance of the $d = 3$ surface code (red triangle), which is the average of the performance of each of the four constituent codes (bright red triangles), is compared to the $d = 5$ code (orange hexagons). Each data point is averaged over 5×10^4 shots for both experiment and simulation. Solid lines show the fits to the data used to extract the logical error rate per round ε_L. The error bars are smaller than the marker sizes.

8.3.2. Performance on experimental data

Next, we evaluate the performance of the NN decoder on experimental data available from the recent experiment executed by Google Quantum AI [2], where a 72-qubit quantum processor was used to implement a $d = 5$ surface code as well as the four $d = 3$ surface codes which use a subset of the qubits of the larger code. The stabilizer measurement circuits used in that experiment are the same as those shown in Fig. 8.8. For each distance-d surface code, the data qubits are prepared in several random bitstrings, followed by $r = 25$ rounds of stabilizer measurement, followed by a logical measurement, with experiments performed in both the X-basis and Z-basis. The experiment demonstrated that the $d = 5$ surface code achieves a lower ε_L compared to the average of the four constituent $d = 3$ patches when using a tensor-network (TN) decoder, an approximation to a maximum-likelihood decoder.

We find that training a NN decoder to achieve good logical performance requires a large number of shots (approximately 10^7 in total or more) obtained from experiments

preparing different initial states and running a different number of rounds. As the amount of experimental data is too small to train the NN decoder (the total number of shots being 6.5×10^5), we instead opt to simulate the experiments using the Pauli error model based on the measured error rates of each operation, available in [2]. Keeping the same number of rounds and prepared state, we generate a total of 2×10^7 shots for training the decoder for each $d = 3$ experiment and 6×10^7 to train the decoder for the $d = 5$ experiment, see Tab. 8.1. While we train the network on simulated data, we still evaluate the decoder performance on both simulated and the experimental data, with the results shown in Fig. 8.4**a** and Fig. 8.4**b** respectively. Both the training and evaluation data consist of $r = 1, 3, \dots, 25$ rounds of QEC and consider the same initial states. When evaluating the NN decoder on simulated data, we observe that the $d = 5$ code achieves a lower ε_L compared to the average of the $d = 3$ codes, see Fig. 8.4**a**. Evaluating the decoder on the experimental data leads to an approximately 15% (40%) higher ε_L for the $d = 3$ ($d = 5$) code, demonstrating that the approximate error model used in simulation fails to fully capture the errors in the experiment. Furthermore, we observe that the $d = 5$ has a higher ε_L instead, see Fig. 8.4**a**, contrary to what was demonstrated in [2] using a tensor-network decoder.

In order to put the performance of the NN decoder in perspective, in Fig. 8.5, we compare the logical performance of the NN decoder to the performance of several other decoders that were also implemented in [2]. We perform this comparison both on simulated (see Fig. 8.5**a**) and experimental (see Fig. 8.5**b**) data. We find that the NN decoder consistently outperforms the standard MWPM decoder in either case. On the experimental dataset, the NN decoder performs equivalent to the TN decoder when decoding the $d = 3$ surface codes. However, when decoding the $d = 5$ surface code experiment, the NN decoder displays a higher ε_L than the TN decoder and the computationally efficient belief-matching (BM) decoder [54]. When evaluated on simulated data, the NN and BM decoders exhibit similar error rates, with the NN decoder again demonstrating better performance when decoding the $d = 3$ code but worse when dealing with the $d = 5$ code. The BM decoder we use for the simulated data is described in [55] and uses the belief propagation implemented in [93]. The higher error rate of the NN decoder for the $d = 5$ code in both simulation and experiment can be related to the difficulty of optimizing the performance of the substantially larger NN model used (see Tab. 8.1 for the model hyper-parameters). However, the discrepancy in the experiment can also be attributed to a mismatch between the simulated data used for training (based on an approximate error model) and the experimental data used for evaluation. Compared to the $d = 3$ surface code data, the accumulation of qubit leakage can cause the $d = 5$ performance to degrade faster over the QEC rounds [2]. We expect that training on experimental data and a better hyper-parameter optimization to enable a NN performance comparable to state-of-the-art decoders like BM and TN while offering additional flexibility to the details of the noise model. Compared to the TN decoder, both NN and BM can achieve similar logical performance while remaining significantly faster, and if their implementation is optimized, they can potentially be used to decode experiments in real time.
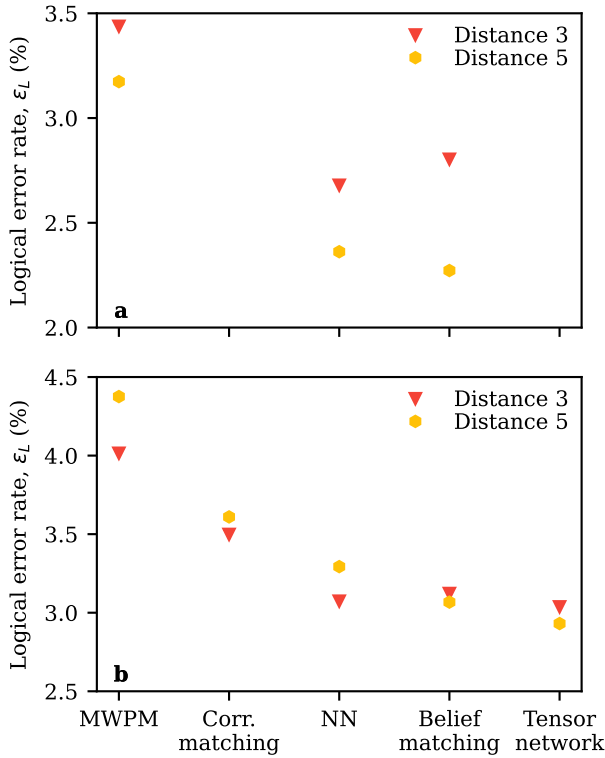
Figure 8.5: The logical error rate per round ε_L for the $d = 3$ (red triangle) and $d = 5$ (orange hexagon) for several decoder implementations applied to either simulated data (shown in **a**) or experimental data (shown in **b**). These correspond (from left to right) to minimum-weight perfect matching (MWPM), a correlated modification of MWPM (Corr. MWPM) [92], our neural network (NN) decoder, belief matching (BM) [54], and a tensor network (TN) decoder, which approximates maximum-likelihood decoding. We did not run the corr. MWPM or TN decoder on the simulated data so fewer data points appear in **a**. All logical error rates on the experimental data, except for the NN decoder, are taken from [2]. The error bars are smaller than the marker sizes.

8.3.3. LOGICAL ERROR RATE SUPPRESSION

An exponential suppression of the logical error rate, assuming that the physical error rates are below 'threshold', is vital for realizing a fault-tolerant quantum computer. We explore the error suppression achieved when using the NN decoder. We characterize the logical performance of $d = 3, 5, 7$ surface codes simulated using a uniform depolarizing circuit-level noise model with an error probability of $p = 0.1\%$, close to the state-of-the-art physical error rates achieved in the experiment. To train the NN decoder, we use data generated using this error probability. We find that also training using a higher probability of $p = 0.2\%$ leads to a significantly lower logical error rate for the $d = 7$ code. Furthermore, we evaluate the performance of the NN decoder on data simulated using $p = 0.05\%$, which is an example of the physical error rate needed to achieve practical sub-threshold scaling of the error rate. For each distance d and error probability p, we perform simulations of memory experiments in the Z-basis with varying numbers of
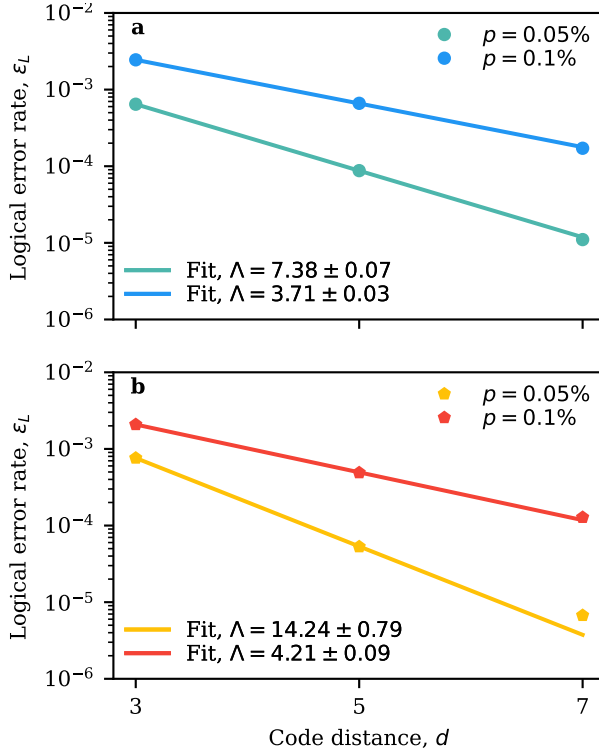
Figure 8.6: The logical error rate per round ε_L for surface codes of distance $d = 3, 5, 7$ for an MWPM decoder, shown in **a**, and our NN decoder, shown in **b**. This is evaluated on datasets using a uniform depolarizing circuit-level noise model with error probabilities of $p = 0.1\%$ (blue for the MWPM, red for the NN decoder) and $p = 0.05\%$ (teal for the MWPM, orange for the NN decoder). Solid lines show the fits to the data used to extract the logical error suppression factor Λ. Each data point is extracted from a fit to the F_L as a function of QEC rounds. The logical fidelities are extracted over 10^5 shots. The error bars are smaller than the marker sizes.

8

QEC rounds, going up to 600 rounds for the $d = 7$ code with an error rate of $p = 0.05\%$ to extract the logical error per round ε_L. The logical error rates obtained when using an MWPM decoder are shown in Fig. 8.6**a**, while those achieved by the NN decoder are shown in Fig. 8.6**b**. If the physical error rate is below threshold, ε_L is expected to decay exponentially with the code distance d, following $\varepsilon_L(d) = C/\Lambda^{(d+1)/2}$, where Λ is the suppression factor and C is a fitting constant. The data shows an apparent exponential suppression of the error rates by either decoder for the considered error rates, which we fit to extract the suppression factor Λ, shown in Fig. 8.6. In either case, the NN decoder achieves better logical performance compared to the MWPM decoder. While for $p = 0.1\%$, the NN decoder achieves an approximately 13% higher Λ, for $p = 0.05\%$, the more accurate NN decoder leads roughly twice as high Λ. The higher suppression factors Λ obtained from using better decoders significantly reduces the code distance required to achieve algorithmically-relevant logical error rates. For example, for an error rate of $p = 0.05\%$, realizing $\varepsilon_L \approx 10^{-10}$ would require a $d = 19$ surface code when using

the MWPM decoder and $d = 15$ when using the NN decoder, corresponding to roughly 40% less physical qubits required. However, whether the NN can continue to exhibit similar performance when decoding higher distance codes remains to be demonstrated.

8.3.4. Decoding with soft information

Measurements of physical qubits generally produce a continuous signal that is subsequently converted into declared binary outcomes by classical processing and thresholding. For example, transmon qubits are dispersively coupled to a dedicated readout resonator, which itself is connected to a readout feedline. Readout is performed by applying a microwave pulse to the feedline, populating the readout resonator. Due to a state-dependent shift of the resonator frequency, the outgoing signal is phase-shifted depending on whether the qubit is in the state $|0\rangle$ or $|1\rangle$. This leads to a change in the real and imaginary components of the outgoing signal, which is experimentally measured. This two-dimensional output can be transformed into a single continuous real variable and converted to a binary outcome by applying some threshold calibrated using a separate experiment [33, 82, 83].

While binary variables are convenient to work with and store, continuous measurement outcomes hold much more information about the state of the qubit, referred to as soft information. It has been demonstrated that an MWPM-based decoder which considers the soft information of the individual measurements when decoding, offers higher thresholds and lower logical error rates than a hard decoder, which only considers the binary outcomes [84]. To demonstrate the flexibility of machine-learning decoders, we consider providing the soft information available from readout when training and evaluating the NN decoder.

In our simulations, measurements project the qubit into either $|0\rangle$ or $|1\rangle$. A measurement outcome $m_{r,q} = i$ of qubit q at round r corresponds to the ancilla qubit being in $|i\rangle$ directly after the measurement. Given $m_{r,q} = i$, we model the soft outcome $\tilde{m}_{r,q} \in \mathbb{R}$ to follow a Gaussian distribution $\mathcal{N}_i$ with mean μ_i and standard deviation σ. The soft outcome $\tilde{m}_{r,q}$ can then be converted to a binary outcome $\bar{m}_{r,q}$ by introducing a threshold t, such that

$$\bar{m}_{r,a} = \begin{cases} 0 & \text{if } \tilde{m}_{r,a} \le t, \\ 1 & \text{otherwise.} \end{cases}$$

For the symmetric Gaussian distributions that we consider, this process leads to an assignment error probability $P(\bar{m}_{r,q} = 0 \mid m_{r,q} = 1) = P(\bar{m}_{r,q} = 1 \mid m_{r,q} = 0) = p_m$. This assignment error is *added* to the errors considered in our circuit-level noise models, specifically the X error before each measurement that happens with a probability p. The assignment error probability can related to the signal-to-noise ratio $\text{SNR} = |\mu_0 - \mu_1|/2\sigma$ as $p_m = \frac{1}{2}\text{erfc}\left(\frac{\text{SNR}}{\sqrt{2}}\right)$. We fix $\mu_0 = -1$ and $\mu_1 = 1$ such that a given probability p_m fixes the standard deviation σ of the two distributions.

The most straightforward approach to incorporating the soft information into the NN decoder is to directly provide the soft measurement outcomes $\tilde{m}_{r,q}$ as input during training and evaluation. However, we find that doing this leads to an overall poor logical performance. Instead, we estimate the probability of a defect $P(d_{r,a} = 1 \mid \tilde{m}_{r,a}, \tilde{m}_{r-1,a})$, given the soft measurement outcomes of an ancilla qubit a in consecutive QEC rounds
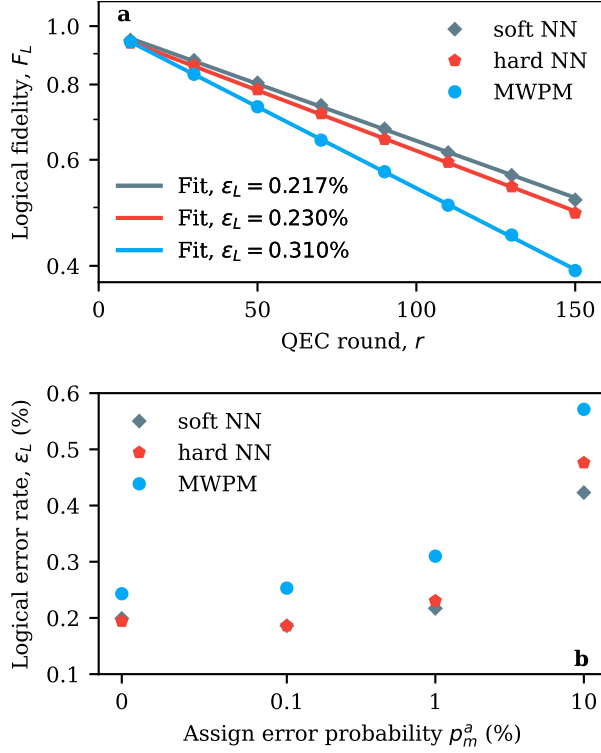
Figure 8.7: **a** The logical fidelity F_L as a function of the number of QEC rounds r for an MWPM decoder (blue circles), a soft NN decoder (gray diamonds) that uses the probability of observing defects and a hard NN decoder (red pentagons) that uses the defects obtained from the hard measurement outcomes. This performance is estimated on simulated data using a uniform depolarizing circuit-level noise model with an error probability $p = 0.1\%$. The soft outcome distributions are such that ancilla and data qubits have a probability of assignment errors of $p_m^a = 1\%$ and $p_m^d = 0.1\%$, respectively. Solid lines show the fits to the data used to extract the logical error rate per round ε_L. Each data is averaged over 10^5 shots. **b** The extracted logical error rate ε_L for each of the three decoders as a function of the ancilla qubit assignment error probability p_m^a, keeping $p_m^d = 0.1\%$ and $p = 0.1\%$. The error bars are smaller than the marker sizes.

and provide this as input to the decoder. Given a soft outcome $\tilde{m}_{r,q}$, the probability of the measured qubit 'having being in the state' $|i\rangle$ can be expressed as

$$P(i \mid \tilde{m}_{r,q}) = \frac{P(\tilde{m}_{r,q} \mid i)P(i)}{\sum_{j \in \{1,2\}} P(\tilde{m}_{r,q} \mid j)P(j)}.$$

The soft outcomes follow a Gaussian distribution, that is, $P(\tilde{m}_{r,q} \mid i) = \mathcal{N}_i(\tilde{m}_{r,q})$. Finally, we make the simplifying assumption that the prior state probabilities $P(i) = P(j) = \frac{1}{2}$,

---

We note that we discovered a software bug affecting the performance of the soft NN and hard NN decoders, shown in Fig. 8.7. We have since corrected this bug and are currently re-evaluating the performance of the decoders at the time of writing. We believe the resulting logical error rates achieved by these decoders are similar, but not identical, to the ones reported in this chapter.

such that

$$P(i \mid \tilde{m}_{r,q}) = \frac{\mathcal{N}_i(\tilde{m}_{r,q})}{\sum_{j\in\{1,2\}} \mathcal{N}_j(\tilde{m}_{r,q})}.$$

The probability of observing a defect can then be expressed as

$$P(d_{r,a} = 1 \mid \tilde{m}_{r,a}, \tilde{m}_{r-1,a}) =$$
$$1 - \sum_{i\in\{0,1\}} P(i \mid \tilde{m}_{r,a})P(i \mid \tilde{m}_{r-1,a}).$$

The expression for the defect probability inferred from using the soft (final) data qubit measurement outcomes can be derived similarly.

To explore the performance of the soft NN decoder, we simulate the $d = 3$ surface-code memory experiment using a circuit-level noise model with an error rate per operation of $p = 0.1\%$. We consider two separate assignment error probabilities p_m^a and p_m^d for ancilla qubit and data qubit measurements. We motivate this choice by the fact that data qubits remain idling while the ancilla qubits are being measured. A shorter measurement time can reduce the decoherence experienced by the data qubits but will typically lead to a higher p_m^a. The data qubit measurements at the end of the experiment, on the other hand, can be optimized to minimize p_m^d. Therefore, we focus on how a soft decoder can help with decoding when p_m^a is higher, similar to the discussion in [84]. We train the NN decoder using datasets of $r = 1, 5, \ldots, 37$ QEC rounds, sampling 5×10^5 shots for each round and initial logical state. When evaluating the performance, we use simulate $r = 10, 30, \ldots, 150$ QEC rounds, sampling 5×10^4 shots instead.

The results for $p_m^a = 1\%$ are shown in Fig. 8.7**a**. The hard NN decoder achieves an approximately 20% lower logical error rate compared to an MWPM decoder, consistent with the results shown in Fig. 8.3. In comparison, the soft NN decoder leads to an approximately 30% lower logical error rate instead, demonstrating the ability of the decoder to adapt to the provided soft information. In Fig. 8.7**b** the logical error rate ε_L of the three decoders is shown for $p_m^a \in \{0, 0.1\%, 1\%, 10\%\}$, where both NN decoders are trained at the corresponding p_m^a. For low p_m^a, the performance of the soft NN decoder is essentially equivalent to the hard NN decoder, with a moderate reduction in ε_L achieved for $p_m^a \geq 1\%$.

It is possible that the probability of defects is not the optimal way to provide the soft information to the decoder. One downside of this representation is that for a high assignment error probability $p_m^a \geq 20\%$, the probability of observing a defect is close to 50%, which also impacts the training and leads the soft NN decoder to exhibit a higher logical error rate compared to the hard one (not shown in Fig. 8.7). Optimizing the performance of the soft NN decoder and comparing it to alternative approaches, namely the soft MWPM decoder proposed in [84], remains an open question.

8.4. DISCUSSION

We now discuss in more detail the performance of the NN decoder on the experimental data. Unfortunately, we only use simulated data to train the NN decoder throughout this work. These simulations use approximate Pauli-noise models that account for the most significant error mechanisms in the experiment, such as decoherence and readout errors. However, they do not include several important error sources present in the

actual experiments, such as leakage, crosstalk, and stray interactions. The exclusion of these error mechanisms leads to the Pauli-noise models underpredicting the logical error rate compared to the rates observed in the experiment, as observed in Fig. 8.4. Furthermore, it was shown that the $d = 5$ code is more sensitive to errors like leakage and crosstalk, which can lead to a more significant deviation relative to simulations of the $d = 3$ codes [2]. Despite using these approximate models for training, when evaluating the NN decoder on experimental data, we observe that it outperforms MWPM and can achieve logical error rates comparable to those obtained using maximum-likelihood decoding, which is approximated by the TN decoder. The TN decoder requires information about the error probabilities, what defects they lead to, and their corresponding corrections, which can be encoded into a hypergraph, where the nodes correspond to defects and the hyperedges represent errors. Importantly, this hypergraph also does not explicitly include hyperedges corresponding to non-conventional errors, such as leakage or crosstalk. We expect that training on experimental data and optimizing the hyperparameters of the network will enable it to match the performance of the TN decoder closely and potentially exceed it by learning about errors not included in the hypergraph.

Despite the large volume of training data required to achieve good performance, we don't expect that generating sufficient experimental data for training will be an issue. Assuming that the QEC round duration is 1 μs and that it takes 200 ns to reset all qubits between subsequent runs, we estimate that it would take approximately three minutes to generate the datasets with 10^7 shots running $r = 1, 5, \ldots, 37$ rounds of QEC that were used for training the $d = 3, 5, 7$ surface codes, see Tab. 8.1.

The soft NN decoder used in this chapter achieves only a moderate performance increase. A direct comparison of this decoder with the soft MWPM decoder [84] will be useful to put this performance into perspective. It is possible that using the defect probabilities as the decoder input is not an optimal choice. An alternative approach to incorporating the soft information into the decoder is to estimate the likelihood of an assignment error $L_{r,a} = \mathcal{N}_{\neg i}(\tilde{m}_{r,a}) / \mathcal{N}_i(\tilde{m}_{r,a})$ given a soft outcome $\tilde{m}_{r,a}$ that leads to a hard outcome of $\bar{m}_{r,a} = i$, which is used by the soft MWPM decoder proposed in [84]. The likelihoods $L_{r,a}$ can then be provided as input to the NN decoder together with the binary defects $d_{r,a}$ that were measured. In addition to the representation of the input data, it is an open question whether using a soft NN decoder will be useful in practice, where assignment error rates are typically low. Specifically, it would be interesting to see if using a soft NN decoder will enable using a shorter measurement time that might lead to a higher assignment error rate but maximize the logical performance overall, as discussed in [84]. The symmetric Gaussian distributions of the continuous measurement outcomes we consider here are only very simple approximations of the distributions seen in experiments, and in our modeling we could adapt these. In particular, the relaxation that the qubit experiences during the readout leads to an asymmetry between the distributions and a generally higher probability of an assignment error when the qubit was prepared in $|1\rangle$. Furthermore, the continuous outcomes observed in the experiment can also contain information about leakage [29, 94, 95] or correlations with other measurements. Therefore, it will be essential to investigate and optimize the performance of the soft decoders using experimental data.

Finally, we outline some possible directions for future research necessary to use these

decoders for decoding large-distance experiments. Decoders based on feedforward and convolutional architectures have been shown to achieve low-latency decoding, making them a possible candidate for being used in real time [68–70, 72]. On the other hand, recurrent networks generally have a larger number of parameters and carry out more complex operations when processing the data. However, recurrent NN decoders have been shown to achieve higher accuracy and be more easily trainable than other architectures, especially when considering realistic noise models [62]. Therefore, whether hardware implementations of recurrent NN decoders can be used for real-time decoding is an open question. In addition to the latency, the scalability of NN decoders is an open question. Decoding higher-distance codes will require larger neural networks and larger training datasets, which will most likely be more challenging to train, given that approaches based on machine learning generally struggle when the dimension of the input becomes very large. Practically, one might be interested in whether the NN decoder can be trained and used to decode some finite code distance, which is expected to lead to algorithmically-relevant logical error rates given the processor's performance. Alternatively, there exist approaches that enable scalable NN decoders. These are typically based on convolutional neural networks that learn to infer and correct the physical errors that have occurred while a secondary global decoder handles any possibly remaining errors [68, 70], but a purely convolutional NN method has been explored as well [79]. The recurrent NN decoder used in this chapter is not scalable, and adapting it to work with larger code distances and using it to decode through logical operations is another open research venue.

8.5. SUPPLEMENTAL MATERIAL

8.5.1. QUANTUM MEMORY EXPERIMENTS

To characterize the logical performance of a surface code, we look at its ability to maintain an initial logical state as a function of the number of QEC rounds, commonly referred to as a quantum memory experiment. The circuits used to perform these experiments are illustrated in Fig. 8.8 and follow the ones used in the recent $d = 5$ surface code experiment done by Google Quantum AI [2]. Removing some of the Hadamard gates when compiling the stabilizer measurement circuits leads to each ancilla qubit measuring the $ZXXZ$ operator instead of the standard $XXXX$ and $ZZZZ$ stabilizers of the surface code. Implementing this $ZXXZ$ variant of the surface code symmetrizes the logical error rates between experiments done in the logical X-basis or Z-basis [2]. Despite this modification, we use notations associated with the traditional stabilizers measured by the surface code.

Each experiment begins by preparing a given logical state, performed by the circuits in Fig. 8.8 **a-d**. The data qubits are first initialized in the ground state and then prepared in either $|0\rangle$ or $|1\rangle$ by a layer of conditional X gates. A subset of the data qubits is then rotated and transforms the initial state into an eigenstate of the X- or Z-type stabilizers. The parity of the initial bitstring state determines whether $|0\rangle_L$ or $|1\rangle_L$ ($|+\rangle_L$ or $|-\rangle_L$) is prepared if the experiment is done in the Z-basis (X-basis). In simulation, we prepare either $|0\rangle^{\otimes n}$ or $|1\rangle^{\otimes n}$ when using uniform circuit-level noise models. In the experiment, several random bitstring states are used in order to symmetrize the impact of amplitude
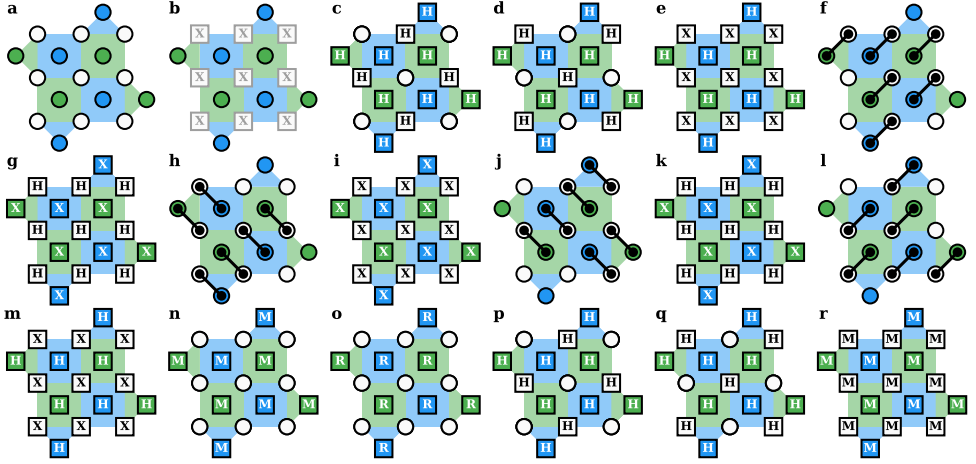
Figure 8.8: Schematic of the circuits used in the quantum memory experiments for a $d = 3$ surface code. **a-d** are used to initialize the logical state at the start of each experiment. The qubits are first prepared in the ground state (**a**), after which a set of conditional X gates (gray) are used to prepare the data qubits in a bit-string state (**b**). Afterward, a set of H (Hadamard) gates transform this into an eigenstate of the X-type (**c**) or Z-type (**d**) stabilizers. **e-o** show the circuits used to measure the stabilizers. The ancilla qubits are first placed in a superposition by a set of H gates (**c** or **d** in the first round, **e** otherwise). The parity of the neighboring data qubits is then mapped using four CZ gates (**f**, **h**, **j**, and **l**). The order of the gates used to measure the X- and Z-type stabilizers are chosen to avoid any "hook" errors propagating to a logical error. Two layers of H gates are applied to the data qubits (**g** and **k**) to measure the parity in the X-basis. In the middle of this sequence, X gates are applied to all qubits (**i**) for dynamical decoupling. Finally, the ancilla qubits are rotated back (**m**) using a set of H gates, measured (**n**, denoted by M) and reset (**o**, denoted by R). Several X gates are applied to the data qubit throughout this sequence for dynamical decoupling. In the final round, all data qubits are measured **p-r**, which is also a logical measurement. Some of the data qubits are rotated depending on whether the experiment is done in the X (**p**) or Z (**q**) logical basis. This step replaces **m** in the final round. Afterward, all qubits are measured simultaneously (**r**), replacing **n** in the final round. Data qubits are denoted with white circles, while ancilla qubits are illustrated as blue and green circles. For the definition of the plaquettes, see Fig. 8.1. The circuits we run follow the ones used in [2].

8

damping [2].

    The prepared logical state is then maintained over a total of $r \in \{1, 2, \ldots, N - 1\}$ QEC rounds, with the circuit given by Fig. 8.8 **e-o**. The first QEC round then projects this initial state into a simultaneous eigenstate of both the X- or Z-type stabilizers. Each cycle involves a series of four interactions between each ancilla qubit and its neighboring data qubits, which map the X or Z parity onto the state of the ancilla qubit. The order in which these two-qubit operations are executed is carefully chosen to minimize the impact of errors occurring during the execution of the circuit [85]. At the end of each QEC round, all of the ancilla qubits are measured and reset. The stabilizer measurement circuits also contain several X gates on either the data or ancilla qubits, which dynamically decouple the qubits in the experiment [2]. Naturally, these gates do not improve the logical performance for the simulations using approximate Pauli-error models that we consider here. In the final QEC round, the data qubits rotated during the state preparation are rotated back and measured in the Z-basis together with the ancilla qubits, illustrated in Fig. 8.8 **p-r**. The data qubit measurement outcomes are then used to cal-

culate the value of the X_L or Z_L logical observable as well as to infer a final set of X- or Z-type stabilizer measurement outcomes.

8.5.2. DECODER TRAINING AND EVALUATION

Distance	Shots	Rounds	Dim. N_L	Learning rate	Batch size	Dropout rate
Experimental circuit-level noise						
3	2×10^7	[1, 25, 2]	64	5×10^{-4}	64	5%
5	6×10^7	[1, 25, 2]	253	5×10^{-4}	256	5%
Uniform circuit-level noise						
3	10^7	[1, 37, 4]	64	10^{-3}	256	20%
5	10^7	[1, 37, 4]	96	10^{-3}	256	20%
7	10^7	[1, 37, 4]	128	10^{-3}	256	20%

Table 8.1: The hyper-parameters used for training the NN decoders. Different parameters are used for simulations based on the uniform circuit-level noise model and the experimental circuit-level noise, which models the experiments done in [2]. The internal state size of the network layers N_L is chosen to scale with the code distance d. The QEC round parameters $[i, j, k]$ for each dataset refer to performing experiments starting with i QEC rounds and going up to j rounds in steps of k. The total number of shots used for training is given, which is equally divided over the QEC rounds and prepared states (not shown in the table). The learning rate, batch size, and dropout rate are the hyper-parameters we tune to help the network to train.

Here we provide additional details about how we train the NN decoder and the hyper-parameters we use. We use the Adam optimizer typically with a learning rate of 10^{-3} or 5×10^{-4} for training. In addition, we apply dropout after the hidden layer of the feed-forward network of each head and, in some cases, after the second LSTM layer with a dropout rate of either 20% or 5% to avoid over-fitting and assist with the generalization of the network. We use a batch size of 256 or 64, which we found to lead to a smoother minimization of the loss. After each training epoch, we evaluate the loss of the network on a separate dataset that considers the same number of QEC rounds and prepared states as the training dataset but samples fewer shots for each experiment. After each epoch, we save the networks' weights if a lower loss has been achieved. Furthermore, we use early stopping to end the training if the loss has not decreased over the last 20 epochs to reduce the time it takes to train each model. We have observed that not using early-stopping and leaving the training to continue does not typically lead the network to reach a lower loss eventually. For some datasets, we lower the learning rate after the initial training has stopped early and train the network once more to achieve better performance. The hyper-parameters we have used for training each network and the parameters of the training datasets used are presented in Tab. 8.1.

The NN architecture we employ in this chapter uses two stacked LSTM layers to process the recurrent input [77]. We observe poor logical performance for a $d = 3$ surface code when using only a single LSTM layer. On the other hand, we see no significant improvement in the logical error rate when using four layers instead, motivating the choice

to use only two. This network architecture also performs well when decoding $d = 5$ and $d = 7$ surface code experiments. However, we expect that a deeper recurrent network might improve the logical error rates when decoding larger-distance codes or when training on and decoding experimental data. We have also practically observed that training the NN decoder for larger distances is more challenging, especially if the physical error rates are small. Training the neural network on a dataset with a higher physical error rate (in addition to data using the same error rate as the evaluation dataset) can also improve the performance of the decoder, as we also discussed in Sec. 8.3.3.

The training of our neural networks was performed on the DelftBlue supercomputing cluster [96] and was carried out on an NVIDIA Tesla V100S GPU. Once trained, the decoder takes approximately 0.7 seconds per QEC round for a $d = 3$ surface code (corresponding to an internal state size of $N_L = 64$) using a batch size of 50000 shots on an Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz. For a $d = 5$ surface code ($N_L = 96$), it takes about 0.8 seconds per round, while for a $d = 7$ surface code ($N_L = 128$), it takes about 1.1 seconds per round, using the same batch size of 50000 shots. We note that using smaller batch sizes leads to a higher overall runtime due to parallelism when the network processes the inputs. Therefore, larger batch sizes are preferable as long as they fit into the memory. Each runtime was extracted by decoding simulated datasets running $r = 10, 30, \ldots, 290$ rounds of QEC and averaging the runtime per QEC round over all the datasets.

8

BIBLIOGRAPHY

[1] B. M. Varbanov, M. Serra-Peralta, D. Byfield, and B. M. Terhal, *Neural network decoder for near-term surface-code experiments*, 2023, arXiv:2307.03280 [quant-ph].

[2] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush, D. Bacon, J. C. Bardin, J. Basso, A. Bengtsson, S. Boixo, G. Bortoli, A. Bourassa, J. Bovaird, L. Brill, M. Broughton, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, J. Cogan, R. Collins, P. Conner, W. Courtney, A. L. Crook, B. Curtin, D. M. Debroy, A. Del Toro Barba, S. Demura, A. Dunsworth, D. Eppens, C. Erickson, L. Faoro, E. Farhi, R. Fatemi, L. Flores Burgos, E. Forati, A. G. Fowler, B. Foxen, W. Giang, C. Gidney, D. Gilboa, M. Giustina, A. Grajales Dau, J. A. Gross, S. Habegger, M. C. Hamilton, M. P. Harrigan, S. D. Harrington, O. Higgott, J. Hilton, M. Hoffmann, S. Hong, T. Huang, A. Huff, W. J. Huggins, L. B. Ioffe, S. V. Isakov, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, P. Juhas, D. Kafri, K. Kechedzhi, J. Kelly, T. Khattar, M. Khezri, M. Kieferová, S. Kim, A. Kitaev, P. V. Klimov, A. R. Klots, A. N. Korotkov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, K.-M. Lau, L. Laws, J. Lee, K. Lee, B. J. Lester, A. Lill, W. Liu, A. Locharla, E. Lucero, F. D. Malone, J. Marshall, O. Martin, J. R. McClean, T. McCourt, M. McEwen, A. Megrant, B. Meurer Costa, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, A. Morvan, E. Mount, W. Mruczkiewicz, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, H. Neven, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, M. Y. Niu, T. E. O'Brien, A. Opremcak, J. Platt, A. Petukhov, R. Potter, L. P. Pryadko, C. Quintana, P. Roushan, N. C. Rubin, N. Saei, D. Sank, K. Sankaragomathi, K. J. Satzinger, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, V. Smelyanskiy, W. C. Smith, G. Sterling, D. Strain, M. Szalay, A. Torres, G. Vidal, B. Villalonga, C. Vollgraff Heidweiller, T. White, C. Xing, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, Y. Zhang, N. Zhu, and G. Q. AI, *Suppressing quantum errors by scaling a surface code logical qubit*, Nature **614**, 676–681 (2023).

[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] S. Lloyd, *Universal quantum simulators*, Science **273**, 1073 (1996).

[5] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, *Elucidating reaction mechanisms on quantum computers*, Proceedings of the National Academy of Sciences **114**, 7555–7560 (2017).

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

[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] W. Huang, C. H. Yang, K. W. Chan, T. Tanttu, B. Hensen, R. C. C. Leon, M. A. Fogarty, J. C. C. Hwang, F. E. Hudson, K. M. Itoh, A. Morello, A. Laucht, and A. S. Dzurak, *Fidelity benchmarks for two-qubit gates in silicon*, Nature **569**, 532–536 (2019).

[9] 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).

[10] 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).

[11] 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).

[12] 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).

[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, 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).

[14] 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).

[15] 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).

[16] 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).

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

[18] E. Knill, R. Laflamme, and W. H. Zurek, *Resilient quantum computation*, Science **279**, 342–345 (1998).

[19] D. Aharonov and M. Ben-Or, *Fault-tolerant quantum computation with constant error rate*, SIAM Journal on Computing **38**, 1207–1282 (2008).

[20] D. Gottesman, *Fault-tolerant quantum computation with constant overhead*, Quantum Info. Comput. **14**, 1338–1372 (2014).

[21] D. Gottesman, *Stabilizer codes and quantum error correction* (California Institute of Technology, 1997).

[22] A. Kitaev, *Fault-tolerant quantum computation by anyons*, Annals of Physics **303**, 2–30 (2003).

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

[24] 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).

[25] R. Raussendorf and J. Harrington, *Fault-tolerant quantum computation with high threshold in two dimensions*, Phys. Rev. Lett. **98**, 190504 (2007).

[26] A. D. Córcoles, 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**, 1338–1352 (2020).

8

[27] 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).

[28] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, *Demonstrating multi-round subsystem quantum error correction using matching and maximum likelihood decoders*, Nature Communications **14**, 2852 (2023).

[29] 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).

[30] J. F. Marques, H. Ali, B. M. Varbanov, M. Finkel, H. M. Veen, S. L. M. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider, B. M. Terhal, and L. DiCarlo, *All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits*, Phys. Rev. Lett. **130**, 250602 (2023).

[31] 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**, 1761 (2021).

[32] K. C. Miao, M. McEwen, J. Atalaya, D. Kafri, L. P. Pryadko, A. Bengtsson, A. Opremcak, K. J. Satzinger, Z. Chen, P. V. Klimov, C. Quintana, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, P. Conner, A. L. Crook, B. Curtin, D. M. Debroy, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, A. G. Fowler, B. Foxen, G. Garcia, W. Giang, C. Gidney, M. Giustina, R. Gosula, A. G. Dau, J. A. Gross, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, J. Kelly, S. Kim, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J.

Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S. Martin, A. Megrant, X. Mi, S. Montazeri, A. Morvan, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Rocque, P. Roushan, K. Sankaragomathi, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, N. Shutty, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, T. White, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, N. Zhu, N. Zobrist, H. Neven, V. Smelyanskiy, A. Petukhov, A. N. Korotkov, D. Sank, and Y. Chen, *Overcoming leakage in quantum error correction*, Nature Physics **19**, 1780–1786 (2023).

[33] 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).

[34] 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. App. **6**, 034008 (2016).

[35] 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**, 66–69 (2015).

[36] N. Ofek, A. Petrenko, R. Heeres, P. Reinhold, Z. Leghtas, B. Vlastakis, Y. Liu, L. Frunzio, S. M. Girvin, L. Jiang, M. Mirrahimi, M. H. Devoret, and R. J. Schoelkopf, *Extending the lifetime of a quantum bit with error correction in superconducting circuits*, Nature **536**, 441 (2016).

[37] A. Grimm, N. E. Frattini, S. Puri, S. O. Mundhada, S. Touzard, M. Mirrahimi, S. M. Girvin, S. Shankar, and M. H. Devoret, *Stabilization and operation of a kerr-cat qubit*, Nature **584**, 205–209 (2020).

[38] P. Campagne-Ibarcq, A. Eickbusch, S. Touzard, E. Zalys-Geller, N. E. Frattini, V. V. Sivak, P. Reinhold, S. Puri, S. Shankar, R. J. Schoelkopf, L. Frunzio, M. Mirrahimi, and M. H. Devoret, *Quantum error correction of a qubit encoded in grid states of an oscillator*, Nature **584**, 368–372 (2020).

[39] V. V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. L. Brock, A. Z. Ding, L. Frunzio, S. M. Girvin, R. J. Schoelkopf, and M. H. Devoret, *Real-time quantum error correction beyond break-even*, Nature **616**, 50–55 (2023).

[40] 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 control of an error-corrected qubit*, Nature **598**, 281–286 (2021).

[41] M. H. Abobeih, Y. Wang, J. Randall, S. J. H. Loenen, C. E. Bradley, M. Markham, D. J. Twitchen, B. M. Terhal, and T. H. Taminiau, *Fault-tolerant operation of a logical qubit in a diamond quantum processor*, Nature **606**, 884–889 (2022).

8

[42] C. Ryan-Anderson, J. G. Bohnet, K. Lee, D. Gresh, A. Hankin, J. P. Gaebler, D. Francois, A. Chernoguzov, D. Lucchetti, N. C. Brown, T. M. Gatterman, S. K. Halit, K. Gilmore, J. A. Gerber, B. Neyenhuis, D. Hayes, and R. P. Stutz, *Realization of real-time fault-tolerant quantum error correction*, Phys. Rev. X **11**, 041058 (2021).

[43] 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*, Nat. Phys. **18**, 80–86 (2022).

[44] 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**, 383–387 (2021).

[45] 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**, 875–880 (2020).

[46] S. Krinner, N. Lacroix, A. Remm, A. Di Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, *Realizing repeated quantum error correction in a distance-three surface code*, Nature **605**, 669–674 (2022).

[47] Y. Zhao, Y. Ye, H.-L. Huang, Y. Zhang, D. Wu, H. Guan, Q. Zhu, Z. Wei, T. He, S. Cao, F. Chen, T.-H. Chung, H. Deng, D. Fan, M. Gong, C. Guo, S. Guo, L. Han, N. Li, S. Li, Y. Li, F. Liang, J. Lin, H. Qian, H. Rong, H. Su, L. Sun, S. Wang, Y. Wu, Y. Xu, C. Ying, J. Yu, C. Zha, K. Zhang, Y.-H. Huo, C.-Y. Lu, C.-Z. Peng, X. Zhu, and J.-W. Pan, *Realization of an error-correcting surface code with superconducting qubits*, Phys. Rev. Lett. **129**, 030501 (2022).

[48] A. G. Fowler, A. C. Whiteside, and L. C. L. Hollenberg, *Towards practical classical processing for the surface code*, Phys. Rev. Lett. **108**, 180501 (2012).

[49] A. G. Fowler, *Minimum weight perfect matching of fault-tolerant topological quantum error correction in average o(1) parallel time*, Quantum Info. Comput. **15**, 145–158 (2015).

8

[50] O. Higgott and C. Gidney, *Sparse blossom: correcting a million errors per core second with minimum-weight matching*, 2023, arXiv:2303.15933 [quant-ph].

[51] Y. Wu and L. Zhong, *Fusion blossom: fast mwpm decoders for qec*, 2023, arXiv:2305.08307 [quant-ph].

[52] J. Roffe, D. R. White, S. Burton, and E. Campbell, *Decoding across the quantum low-density parity-check code landscape*, Phys. Rev. Res. **2**, 043423 (2020).

[53] B. Criger and I. Ashraf, *Multi-path Summation for Decoding 2D Topological Codes*, Quantum **2**, 102 (2018).

[54] O. Higgott, T. C. Bohdanowicz, A. Kubica, S. T. Flammia, and E. T. Campbell, *Improved decoding of circuit noise and fragile boundaries of tailored surface codes*, Phys. Rev. X **13**, 031007 (2023).

[55] L. Caune, J. Camps, B. Reid, and E. Campbell, *Belief propagation as a partial decoder*, 2023, arXiv:2306.17142 [quant-ph].

[56] S. Bravyi, M. Suchara, and A. Vargo, *Efficient algorithms for maximum likelihood decoding in the surface code*, Phys. Rev. A **90**, 032326 (2014).

[57] C. T. Chubb and S. T. Flammia, *Statistical mechanical models for quantum codes with correlated noise*, Annales de l'Institut Henri Poincare D **8**, 269–321 (2021).

[58] S. Spitz, B. M. Tarasinski, C. Beenakker, and T. O'Brien, *Adaptive weight estimator for quantum error correction in a time-dependent environment*, Advanced Quantum Technologies **1**, 1800012 (2018).

[59] E. H. Chen, T. J. Yoder, Y. Kim, N. Sundaresan, S. Srinivasan, M. Li, A. D. Córcoles, A. W. Cross, and M. Takita, *Calibrated decoders for experimental quantum error correction*, Phys. Rev. Lett. **128**, 110504 (2022).

[60] G. Torlai and R. G. Melko, *Neural decoder for topological codes*, Phys. Rev. Lett. **119**, 030501 (2017).

[61] M. Sheth, S. Z. Jafarzadeh, and V. Gheorghiu, *Neural ensemble decoding for topological quantum error-correcting codes*, Phys. Rev. A **101**, 032338 (2020).

[62] S. Varsamopoulos, K. Bertels, and C. Almudever, *Comparing neural network based decoders for the surface code*, IEEE Transactions on Computers **69**, 300–311 (2020).

[63] S. Varsamopoulos, K. Bertels, and C. G. Almudever, *Decoding surface code with a distributed neural network–based decoder*, Quantum Machine Intelligence **2**, 3 (2020).

[64] D. Fitzek, M. Eliasson, A. F. Kockum, and M. Granath, *Deep q-learning decoder for depolarizing noise on the toric code*, Phys. Rev. Res. **2**, 023230 (2020).

[65] R. Sweke, M. S. Kesselring, E. P. L. van Nieuwenburg, and J. Eisert, *Reinforcement learning decoders for fault-tolerant quantum computation*, Machine Learning: Science and Technology **2**, 025005 (2020).

[66] K. Meinerz, C.-Y. Park, and S. Trebst, *Scalable neural decoder for topological surface codes*, Phys. Rev. Lett. **128**, 080505 (2022).

8

[67] Y. Ueno, M. Kondo, M. Tanaka, Y. Suzuki, and Y. Tabuchi, *Neo-qec: neural network enhanced online superconducting decoder for surface codes*, 2022, arXiv:2208.05758 [quant-ph].

[68] C. Chamberland, L. Goncalves, P. Sivarajah, E. Peterson, and S. Grimberg, *Techniques for combining fast local decoders with global decoders under circuit-level noise*, Quantum Science and Technology **8**, 045011 (2023).

[69] R. W. J. Overwater, M. Babaie, and F. Sebastiano, *Neural-network decoders for quantum error correction using surface codes: a space exploration of the hardware cost-performance tradeoffs*, IEEE Transactions on Quantum Engineering **3**, 1–19 (2022).

[70] S. Gicev, L. C. L. Hollenberg, and M. Usman, *A scalable and fast artificial neural network syndrome decoder for surface codes*, Quantum **7**, 1058 (2023).

[71] S. Krastanov and L. Jiang, *Deep neural network probabilistic decoder for stabilizer codes*, Scientific Reports **7**, 11003 (2017).

[72] M. Zhang, X. Ren, G. Xi, Z. Zhang, Q. Yu, F. Liu, H. Zhang, S. Zhang, and Y.-C. Zheng, *A scalable, fast and programmable neural decoder for fault-tolerant quantum computation using surface codes*, 2023, arXiv:2305.15767 [quant-ph].

[73] E. Egorov, R. Bondesan, and M. Welling, *The end: an equivariant neural decoder for quantum error correction*, 2023, arXiv:2304.07362 [quant-ph].

[74] S. Varsamopoulos, B. Criger, and K. Bertels, *Decoding small surface codes with feedforward neural networks*, Quantum Science and Technology **3**, 015004 (2017).

[75] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. J. Beenakker, *Machine-learning-assisted correction of correlated qubit errors in a topological code*, Quantum **2**, 48 (2018).

[76] C. Chamberland and P. Ronagh, *Deep neural decoders for near term fault-tolerant experiments*, Quantum Science and Technology **3**, 044002 (2018).

[77] P. Baireuther, M. D. Caio, B. Criger, C. W. J. Beenakker, and T. E. O'Brien, *Neural network decoder for topological color codes with circuit level noise*, New Journal of Physics **21**, 013003 (2019).

[78] P. Andreasson, J. Johansson, S. Liljestrand, and M. Granath, *Quantum error correction for the toric code using deep reinforcement learning*, Quantum **3**, 183 (2019).

[79] X. Ni, *Neural Network Decoders for Large-Distance 2D Toric Codes*, Quantum **4**, 310 (2020).

[80] T. Wagner, H. Kampermann, and D. Bruß, *Symmetries for a high-level neural decoder on the toric code*, Phys. Rev. A **102**, 042411 (2020).

[81] C. Gidney, *Stability Experiments: The Overlooked Dual of Memory Experiments*, Quantum **6**, 786 (2022).

[82] 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).

[83] A. Blais, A. L. Grimsmo, S. M. Girvin, and A. Wallraff, *Circuit quantum electrodynamics*, Rev. Mod. Phys. **93,** 025005 (2021).

[84] C. A. Pattison, M. E. Beverland, M. P. da Silva, and N. Delfosse, *Improved quantum error correction using soft information*, 2021, arXiv:2107.13589 [quant-ph].

[85] Y. Tomita and K. M. Svore, *Low-distance surface codes under realistic quantum noise*, Phys. Rev. A **90**, 062320 (2014).

[86] C. Gidney, *Stim: a fast stabilizer circuit simulator*, Quantum **5,** 497 (2021).

[87] B. M. Varbanov and M. Serra-Peralta, *surface-sim*, version 1, July 2023.

[88] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, *Tensorflow: large-scale machine learning on heterogeneous distributed systems*, 2015.

[89] B. M. Varbanov and M. Serra-Peralta, *qrennd*, version 1, July 2023.

[90] S. Hochreiter and J. Schmidhuber, *Long Short-Term Memory*, Neural Computation **9,** 1735–1780 (1997).

[91] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature **521,** 436–444 (2015).

[92] A. G. Fowler, *Optimal complexity correction of correlated errors in the surface code*, arXiv:1310.0863 (2013).

[93] J. Roffe, *LDPC: Python tools for low density parity check codes*, version 0.1.0, Jan. 2022.

[94] 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).

[95] M. Khezri, A. Opremcak, Z. Chen, A. Bengtsson, T. White, O. Naaman, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, A. L. Crook, B. Curtin, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, B. Foxen, G. Garcia, W. Giang, M. Giustina, R. Gosula, A. G. Dau, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, J. Kelly, S. Kim, P. V. Klimov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J. Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S.

8

Martin, M. McEwen, A. Megrant, X. Mi, K. C. Miao, S. Montazeri, A. Morvan, M. Neeley, C. Neill, A. Nersisyan, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Quintana, C. Rocque, P. Roushan, K. Sankaragomathi, K. J. Satzinger, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, N. Zhu, N. Zobrist, D. Sank, A. Korotkov, Y. Chen, and V. Smelyanskiy, *Measurement-induced state transitions in a superconducting qubit: within the rotating wave approximation*, 2022, arXiv:2212.05097 [quant-ph].

[96]   D. H. P. C. C. (DHPC), *DelftBlue Supercomputer (Phase 1)*, https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1, 2022.

8

9

CONCLUSION

9.1. SUMMARY AND DISCUSSION

This dissertation focuses on some of the challenges in implementing error correction using the surface code in superconducting processors and how to characterize them and potentially resolve them. We have explored these issues using realistic error models and detailed simulations or by analyzing the performance of experiments implementing small stabilizer codes.

- In Chapter 2, we demonstrated that leakage can significantly impact the logical performance of a stabilizer code. We observed that leakage on both data and ancilla qubits behaves stochastically and increases the defect rate observed on the neighboring stabilizers. Data-qubit leakage effectively decreases the code distance, combined with the corruption of the neighboring stabilizers due to the anti-commutation of the stabilizers involving this data qubit. Ancilla-qubit leakage inhibits the parity measurements typically performed by the ancilla qubit and can instead propagate errors on the neighboring data qubits. The higher defect rates can also be interpreted as a signature of leakage, enabling the indirect detection of leakage. We show that a set of independent Hidden Markov models can accurately detect leakage events and ultimately allow restoring the code performance using post-selection. In practice, ancilla-qubit leakage can often be detected accurately from the readout signal, at least when the measurement is calibrated to distinguish the second-excited state. The signal produced by data-qubit leakage generally depends on the code chosen and how the individual operations interact with the leaked data-qubit. In the surface code, data-qubit leakage leads to the anti-commutation of the neighboring stabilizers if the Hadamard gates act as the identity of the leaked qubit. A leaked data qubit in the bulk of the surface code will, therefore, lead to an increase in the defect rate observed on the four neighboring ancilla qubits, constituting a strong signal that an HMM can detect. In the repetition code, which measures only a single type of stabilizer, data-qubit leakage will not lead to anti-commutation. The signatures of this leakage event

311

will generally depend on the leakage-conditional phases. If echoing pulses are applied at the end of each cycle, these can also lead to a strong signal, similar to what was observed in Ref. [1]. While data-qubit leakage detection can be helpful in some near-term experiments, post-selection is not a scalable approach to dealing with leakage. Instead, a scalable error correction scheme will require a dedicated leakage-reduction mechanism.

- In Chapter 3, we proposed a leakage reduction scheme using separate operations for data qubits and ancilla qubits that can enable scalable computation. In the case of data qubits, the population in $|2\rangle$ is transferred to a readout resonator using a microwave pulse, where it can quickly decay. For ancilla-qubits, a conditional microwave pulse moves the population in $|2\rangle$ to $|1\rangle$ whenever the qubit is measured to be in $|2\rangle$. We showed that even when implemented with limited fidelity, these two operations can reduce the lifetime of leakage events to a single QEC round and substantially reduce the logical error rate.

  In Chapter 4, we realize one of these leakage reduction operations in a three-qubit parity-check experiment. In particular, we implement the microwave pulse that transfers the leakage to the readout resonator and apply it to both data and ancilla qubits. We find that these LRUs can be implemented in about 200 ns with a fidelity exceeding 80% and as high as 99% while having a negligible impact on the computational subspace (provided we corrected for the phase shifts induced by the strong drive). In the case of the ancilla qubits, this has the benefit of no longer requiring fast classical feedback, which is challenging to realize in practice and necessary for our initial proposal. Finally, we extend the LRU to simultaneously reduce the population in both $|2\rangle$ and $|3\rangle$. These operations lead to steady-state leakage populations below 1% on all three qubits and remove the increase in the defect probability associated with leakage. The ancilla-qubit LRU increases the duration of the QEC round, which also increases the logical error rate. Therefore, either optimizing the effective coupling strength to reduce the duration of this operation or implementing sufficiently low-latency feedback to enable the initially proposed π-LRU is expected to improve the logical performance. Applying these LRUs to large codes will also require optimizing the qubit and resonator frequencies to avoid any possible microwave crosstalk. Therefore, these LRUs can also reduce the expected device yield when using fixed-frequency transmon qubits.

- In Chapter 5, we considered a heterogeneous qubit architecture that employed transmon and fluxonium qubits. We showed that fast, high-fidelity, and low-leakage two-qubit gates could be realized between these two qubits. We have proposed a microwave-activated cross-resonance gate for fluxonium qubits with frequencies ranging from approximately 250 MHz to about 1 GHz. The strength of this cross-resonance interaction vanishes when the frequency of the fluxonium is below this range. In that case, we show that a flux-pulsed controller-phase gate can instead be realized using interactions involving higher-excited states. Using the cross-resonance gate, we demonstrated that this homogeneous architecture can significantly reduce the frequency crowding problem compared to homogeneous qubit architectures using fixed-frequency transmon architectures. However, it is

unclear whether our architecture offers any advantage regarding the expected device yield compared to an architecture employing only fluxonium qubits. More generally, a heterogeneous architecture can be a particularly natural choice for realizing error-correcting codes, where ancilla qubits should be designed to enable a fast readout. In contrast, data qubits can be designed to have a slower readout if this ultimately allows them to have higher coherence times or other better single-qubit or two-qubit gates.

- In Chapter 6, we realized a suite of logical operations in an error-detecting distance-two surface code. In particular, we demonstrate the ability to initialize arbitrary initial states, to measure in the cardinal bases of the Bloch sphere, and to perform a universal set of two-qubit gates. We characterized these logical gates using process tomography on the logical level and constructed the logical Pauli transfer matrices corresponding to each operation. Logical process tomography will be essential for the characterization of the logical operations in near-term experiments, where codes can not perform logical randomized benchmarking. We also observe that the fault-tolerant operations achieve higher fidelity, demonstrating the benefits of error detection and correction. For the logical non-Clifford gate, we implemented T_L using a gate-by-measurement scheme. Due to the small distance of the code, we were able to use a single ancilla qubit to implement this operation, avoiding the need for qubit swapping. However, implementing non-Clifford gates becomes much more challenging for larger-distance codes, with one of the more promising approaches being magic-state distillation protocols.

- In Chapter 7, we implemented a repetition code experiment and demonstrated an exponential suppression of the logical error rate with increasing code distance. However, we achieve this with a low error suppression factor, suggesting that the physical error rates are only slightly below the threshold. Achieving error suppression or memory-break-even performance with the surface code will likely require several improvements, including lower readout and two-qubit error rate and the inclusion of leakage reduction operations discussed in this thesis. We also outlined some difficulties in estimating the physical error probabilities from the observed syndrome defects, namely when these errors lead to three or more non-trivial defects. In principle, these errors must be explicitly accounted for to achieve accurate probability estimates and avoid any potential increase in the logical error rates if the decoder is calibrated using these probabilities. In practice, however, the impact of excluding these errors from the estimation procedure depends on the likelihood of these errors and the number of defects they can lead to. For example, experiments in which the leakage rates are relatively low and in which leakage reduction operations are employed such that the lifetime of leakage events is limited to around a single QEC cycle will likely be able to ignore the presence of these errors when inferring the conventional error rates.

- In Chapter 8, we explored the performance of a recurrent neural network when decoding surface code experiments, using both experimental data and simulated data based on a circuit-level noise model with physical error rates in the range of

what is achievable in near-term experiments. We found that this decoder can effectively deal with errors leading to more than two non-trivial syndrome defects, leading it to outperform perfect-matching decoders and achieve error rates approach maximum-likelihood decoders. Furthermore, these decoders are adaptive to the physical noise and can be modified to use the soft information in the readout. A practical problem when using these decoders is that the size of the networks generally scales with the code distance. Training these larger models becomes more challenging and can lead to sub-optimal performance. It might be necessary to develop a procedure for optimizing the model hyperparameters or to consider possible modifications to the architecture that can enable easier training. For example, one can extend the model to predict the defects in the next QEC round and modify the loss function accordingly. It also needs to be determined whether these decoders can be used for real-time decoding, especially considering the increase in the runtime with the size of the recurrent layers. Finally, it can be helpful to investigate how well such a decoder can adapt to non-conventional noise sources, such as leakage or crosstalk, present in experimental data.

9.2. OUTLOOK

Throughout this dissertation, we have outlined several challenges encountered in realizing error correction experiments, mainly focusing on the problem of leakage. Over the past few years, there has been significant progress in improving the quantum hardware, leading to modern superconducting processors featuring anywhere from around 50 to more than 100 qubits [2–5]. At the same time, significant progress has been made towards lowering the gate [6–15] and measurement [16–22] error rates. In the case of transmon qubits, the dispersive readout is slower than the other operations and typically exhibits higher error rates. However, there are some proposals based on *qubit cloaking* that can lead to lower error rates and faster readout times [23, 24]. Furthermore, several schemes have been proposed to mitigate non-conventional errors in these processors, such as crosstalk [7, 25–32] and leakage [33–41]. This has led to several impressive experiments demonstrating the ability of error correction to suppress the logical error rate [5, 42]. I believe there are several challenges that will be crucial to overcome in the next generation of experiments and, ultimately, to realize scalable fault-tolerant computation.

- **Improving the qubit coherence times.** Quantum error correction can suppress the logical error rates by increasing the code distance, assuming that the physical error rates are below the threshold. This suppression comes at the expense of a significant overhead in the required qubits. This overhead is typically captured by the encoding rate k/n, which is the number of encoded logical qubits k over the number of required physical qubits n. For example, the surface code, arguably the most promising stabilizer code, encodes only a single logical qubit ($k = 1$) using $n = d^2$ physical data qubits, where d is the code distance. This poor encoding rate means that fault-tolerant algorithms based on the surface code will likely require millions of qubits.

  This qubit overhead imposes considerable engineering challenges when considering the possible errors in fabrication, the heating that the control pulses induce

in the refrigerator, and the challenge of either engineering larger dilution refrigerators or interconnecting multiple smaller ones (or possibly both). Therefore, it is vital to reduce the number of physical qubits required for fault-tolerant computation as much as possible.

One of the ways to achieve this is by increasing the qubit coherence times and, therefore, lowering the physical error rates, assuming that the applied operations generally are coherence-limited. The lower physical error rates, in turn, lead to a larger error suppression factor, enabling a lower-distance code to achieve a sufficiently low logical error rate. While this does not improve the encoding rate of the code, it ultimately reduces the number of physical qubits necessary for fault-tolerant computation.

The typical relaxation times that transmons are typically on the order of 20 μs to 50 μs [43]. However, there has been significant progress in understanding the different loss mechanisms and mitigating them. Using lower-loss materials has led to a substantial increase in the transmon relaxation times. Specifically, using sapphire substrates and tantalum films as the base superconductor has been shown to reduce the bulk dielectric and surface losses, respectively. This has led to transmons qubits with relaxation times as high as 500 μs and average values on the order of 100 μs [44–46].

These high coherence times are typically achieved in few-qubit devices. In contrast, the relaxation times measured in a 50-qubit processor made using the same fabrication methods were about an order of magnitude lower, with an average of $T_1 \approx 30$ μs, partially attributed to decoherence caused by the control lines [45]. Therefore, increasing the number of qubits available while maintaining high coherence times also requires careful thermal engineering.

Finally, while a $T_1 \approx 500$ μs is an impressive number, the design of the transmon qubit has several downsides, including a low anharmonicity and a sensitivity to flux noise. The fluxonium qubit is an alternative design that addresses some of these limitations [47]. There have been fluxonium qubits with a relaxation time $T_1 \gtrsim 1$ ms [48, 49]. Exploring the noise sources limiting these coherence times and how they can be mitigated is worthwhile, given that the fluxonium qubit may be a superior choice for superconducting processors.

- **Two-level-system defects and relaxation time fluctuations.** In addition to increasing the qubit-coherence times, it is also essential to consider their stability throughout the experiment. In particular, the transmon relaxation times have been shown to exhibit significant fluctuations on a timescale of several minutes to a few hours [50–52]. These fluctuations are typically attributed to two-level systems (TLS) that couple electrically to the transmon and lead to additional loss. The TLS frequencies can fluctuate in time, leading to fluctuations in the relaxation time. These relaxation time fluctuations can impact the performance of the logical qubit, especially if the decoder is not re-calibrated frequently enough to account for the changes in the physical error rates.

Besides impacting the coherence times at the operational point, TLS can also lead

9

to significantly higher gate error rates. Flux-tunable transmon qubits can be fluxed away from the operating point to perform a controlled-phase gate with a neighboring qubit. TLS that are close in frequency to the interaction point or that interact with the qubit while it is being fluxed to that point will lead to higher two-qubit gate error rates [32]. During dispersive readout, the transmon-qubit frequency is broadened and shifted, which can lead to strong interactions with TLS that would only be weakly coupled to the transmon qubit at the operating point, leading to a typical decrease of T_1 during readout [53]. Since this effect is largely due to the broadening of the qubit frequency during the readout, this effect can also be expected to occur for other superconducting qubits. Furthermore, the drift in TLS frequencies would require the operations to be re-calibrated [32, 51, 54]. Therefore, lowering the density of TLS will be critical, especially for larger transmon processors.

- **Ionizing radiation.** It has been proposed that gamma rays emitted by radioactive impurities in building materials such as concrete and high-energy cosmic rays ionize the substrate of the processor, generating phonons that can ballistically travel throughout the substrate for several millimeters [55–57]. These phonons can ultimately reach the superconductor at the surface of the substrate, where they can break Cooper pairs and lead to a burst of quasiparticles. This leads to a correlated decrease in the relaxation times of potentially many qubits on the device, which has been recently observed in experiments [56–58]. It has also been observed that this can lead to the frequencies of multiple TLS in the vicinity of the radiation impact to shift in frequency, which can, in some instances, also impact the relaxation times of the qubits [54]. Interestingly, that same experiment did not observe a decrease in the relaxation time due to the generated quasiparticles, the reason for which has yet to be entirely determined. Regardless, ionizing radiation can simultaneously impact many qubits, leading to correlated errors that are hard for error-correcting codes to handle [57]. Such an event occurs once every 10 seconds on average [57], meaning that this is an important error source that will eventually impact large-scale fault-tolerant computation. Even though this error might not be devastating for a very large-distance code [59], it still affects the logical performance of the code, increasing the code size required to achieve a sufficiently low logical error rate throughout the entire computation. Therefore, it might be necessary to introduce additional shielding [60] or quasiparticle traps [61–65] to mitigate this noise on the hardware level instead.

- **Stabilizer codes that lead to lower overhead.** Apart from lowering the physical error rates, the significant qubit overhead required for fault-tolerant computation can be addressed more fundamentally by considering error-correcting codes that can achieve better logical performance or a higher encoding rate than the surface code. In most cases, these codes will also require adapting the quantum hardware to enable their experimental implementation.

Recent work has shown that a significant improvement in the logical performance of the code can be achieved by some tailored codes in the case of biased physical noise [66–72]. Maintaining the bias during the parity-check measurements also

requires a bias-preserving CNOT gate [73]. Typically, most superconducting qubits either do not exhibit a bias in their noise or have an interaction that enables this bias-preserving gate. However, both biased noise and the required gates can be engineered in superconducting cat qubits [70, 73–75]. Another example of noise that can enable a better logical performance is that of erasure errors, which can also be engineered in superconducting or ion-qubits [76–82].

Another, more promising, direction towards reducing the qubit overhead is using codes with a higher encoding rate. In particular, codes with a constant encoding rate can lead to a constant overhead [83, 84]. Low-density parity-check (LDPC) codes are particularly promising, as they limit the number of parity checks each qubit participates in and the number of qubits each check acts on [85]. These are experimentally desirable properties, as high-weight parity measurements are hard to measure indirectly with an ancilla qubit, while a lower number of checks that each qubit participates in leads to shorter measurement circuits. More importantly, families of *good* LDPC codes have been discovered that offer both a constant encoding rate k/n and a distance d scaling linearly with the number of qubits n [86, 87]. Note that this definition of the encoding rate does not consider the additional ancilla qubits needed to perform the parity checks.

However, several challenging problems still need to be solved to realize an LDPC code in an experiment. Firstly, the logical performance of these codes remains to be explored in detail, especially under realistic circuit-level noise models. More importantly, LDPC codes require some non-local and possibly high-degree qubit connectivity, which is challenging to implement in planar layouts. Regardless, a few recent results demonstrate LDPC codes that achieve higher encoding rate thresholds comparable to the surface code for circuit-level noise while requiring connectivity compatible with bi-layer layouts [88–90].

It is, therefore, of considerable interest to investigate the advantages of other codes and their feasibility. For LDPC codes, a few open problems include finding code families with better encoding rates or distance scaling, developing real-time decoders that achieve good logical performance, as well as performing logical operations with these codes. Developing the required connectivity or qubit engineering to implement these more optimal codes is also an outstanding design and fabrication challenge. Constructing models that consider the expected crosstalk for these new connectivity layouts can also be useful to evaluate how practical a certain LDPC code is to realize in practice.

- **Simulations and error modelling.** There has been a considerable effort in developing accurate physical error models that can accurately predict the logical performance observed in experiments, with some simulations achieving an impressive agreement with the measured results [5]. The performance of the small-distance codes realized in recent experiments can still be numerically explored with significant accuracy using density matrix simulations [91, 92]. Therefore, this is an excellent opportunity to gain an in-depth understanding of the impact of various error sources on the logical performance of these codes. Achieving a good agreement between the defect rates or logical error rates predicted by such simulations and

9

the experimental observations when employing leakage reduction units can give new insights into the impact of leakage in this setting and whether the decoder can be adapted to handle this error better. For example, in some recent distance-three surface code experiments performing only a single type of stabilizer measurements (referred to as Surface-13), it was observed that data-qubit LRUs lead to lower defect rates but higher logical error rates (unpublished). It is not immediately obvious why this is the case and whether it is due to some specific errors induced by these LRUs or due to suboptimal decoding.

On the other hand, the number of qubits available in current processors is quickly progressing outside of the range of what can be explored using density-matrix simulations. Therefore, developing stochastic and scalable error models that can still make relatively accurate predictions on the code performance and the expected thresholds is very important. Recently, this effort has been extended to include stochastic models of non-conventional errors, namely leakage and crosstalk [5]. Other important error sources that can significantly impact the code performance include TLS interactions and ionizing radiation, the impact of which was described above. A straightforward way to capture the effect of these error sources is to create a model that captures the increase in the qubit relaxation rates that these error sources generally lead to. In particular, TLS can impact the relaxation rates at the operating point, during the readout, or during operations that involve moving the qubit to a lower frequency. However, this might not accurately capture the impact of strongly coupled TLS, which can sometimes interact coherently with a qubit, potentially leading to non-Markovian qubit dynamics. Therefore, a better understanding of the impact of TLS on the operations and the overall impact of TLS and ionizing radiation on the code performance can lead to better mitigation strategies for these errors.

- **Accurate real-time decoders.** Real-time decoding is particularly challenging for superconducting-qubit processors due to the fast gate durations, leading to an error correction round typically taking on the order of 1 μs [93]. Notably, this window also includes the time it takes to process the readout signal and send the observed binary outcome to the decoder, leaving even less time for a decoder to process the measured syndromes. Unfortunately, real-time decoding is necessary for scalable fault-tolerant computation to avoid the possibility of exponential increase in the computation time, known as the *backlog problem* [94, 95]. There have been a few real-time decoder implementations, typically based on the minimum-weight perfect-matching [96] on union-find algorithms [97] that run on field programmable gate arrays or application-specific integrated circuits, which can deal with codes of large distances. However, these algorithms are based on certain approximations, such as the assumption that the physical errors can be modeled as independent X and Z errors [98]. This typically is not the case in reality, which results in higher logical error rates when compared to more complex decoders, such as belief-matching [5, 99] or neural network decoders [100–102]. More generally, decoders should be able to deal with errors that lead to more than two non-trivial defects that may arise from Y errors, for example. These more accurate

algorithms can substantially decrease the qubit overhead, depending on the details of the physical error model. Another feature that might be important in the future is the ability of the decoder to adapt to fluctuations in the noise, which can be achieved by retraining a neural network or via the adaptive estimation of the error rates from the measured data [103]. Therefore, it is essential to investigate and optimize the runtime of these more accurate and flexible decoders and determine whether they can be used for real-time correction.

Naturally, these are only some challenges that must be solved to build a fault-tolerant computer. Exploring alternative qubit platforms with better noise resilience or more flexible connectivity is extremely valuable. Furthermore, engineering interactions that enable lower two-qubit and readout error rates remain essential for reducing the overhead required by error corrections. Another key challenge is developing more accurate fabrication or post-fabrication correction methods that will allow precise targeting of the design parameters. Finally, none of these issues consider the applications of these computers, which will require optimized algorithms that outperform their classical counterparts. Ultimately, implementing error correction in the first place seems a daunting challenge for current experiments. I hope this dissertation has made a small contribution towards making fault-tolerant computation a reality in the future.

9

BIBLIOGRAPHY

[1] 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).

[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] A. D. Córcoles, 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**, 1338–1352 (2020).

[4] E. Pelofske, A. Bärtschi, and S. Eidenbenz, *Quantum volume in practice: what users can expect from NISQ devices*, IEEE Transactions on Quantum Engineering **3**, 1–19 (2022).

[5] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush, D. Bacon, J. C. Bardin, J. Basso, A. Bengtsson, S. Boixo, G. Bortoli, A. Bourassa, J. Bovaird, L. Brill, M. Broughton, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, J. Cogan, R. Collins, P. Conner, W. Courtney, A. L. Crook, B. Curtin, D. M. Debroy, A. Del Toro Barba, S. Demura, A. Dunsworth, D. Eppens, C. Erickson, L. Faoro, E. Farhi, R. Fatemi, L. Flores Burgos, E. Forati, A. G. Fowler, B. Foxen, W. Giang, C. Gidney, D. Gilboa, M. Giustina, A. Grajales Dau, J. A. Gross, S. Habegger, M. C. Hamilton, M. P. Harrigan, S. D. Harrington, O. Higgott, J. Hilton, M. Hoffmann, S. Hong, T. Huang, A. Huff, W. J. Huggins, L. B. Ioffe, S. V. Isakov, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, P. Juhas, D. Kafri, K. Kechedzhi, J. Kelly, T. Khattar, M. Khezri, M. Kieferová, S. Kim, A. Kitaev, P. V. Klimov, A. R. Klots, A. N. Korotkov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, K.-M. Lau, L. Laws, J. Lee, K. Lee, B. J.

Lester, A. Lill, W. Liu, A. Locharla, E. Lucero, F. D. Malone, J. Marshall, O. Martin, J. R. McClean, T. McCourt, M. McEwen, A. Megrant, B. Meurer Costa, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, A. Morvan, E. Mount, W. Mruczkiewicz, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, H. Neven, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, M. Y. Niu, T. E. O'Brien, A. Opremcak, J. Platt, A. Petukhov, R. Potter, L. P. Pryadko, C. Quintana, P. Roushan, N. C. Rubin, N. Saei, D. Sank, K. Sankaragomathi, K. J. Satzinger, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skruzny, V. Smelyanskiy, W. C. Smith, G. Sterling, D. Strain, M. Szalay, A. Torres, G. Vidal, B. Villalonga, C. Vollgraff Heidweiller, T. White, C. Xing, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, Y. Zhang, N. Zhu, and G. Q. AI, *Suppressing quantum errors by scaling a surface code logical qubit*, Nature **614**, 676–681 (2023).

[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]   L. Ding, M. Hays, Y. Sung, B. Kannan, J. An, A. Di Paolo, A. H. Karamlou, T. M. Hazard, K. Azar, D. K. Kim, B. M. Niedzielski, A. Melville, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, J. A. Grover, K. Serniak, and W. D. Oliver, *High-fidelity, frequency-flexible two-qubit fluxonium gates with a transmon coupler*, Phys. Rev. X **13**, 031035 (2023).

[8]   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), 060302 (2016).

[9]   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).

[10] 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).

[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] 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).

[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, 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).

[14] 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*, Phys. Rev. Res. **4**, 023040 (2022).

[15] E. Dogan, D. Rosenstock, L. Le Guevel, H. Xiong, R. A. Mencia, A. Somoroff, K. N. Nesterov, M. G. Vavilov, V. E. Manucharyan, and C. Wang, *Two-fluxonium cross-resonance gate*, Phys. Rev. Appl. **20**, 024011 (2023).

[16] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Pechal, M. Mondal, M. Oppliger, C. Eichler, and A. Wallraff, *Rapid High-Fidelity Single-Shot Dispersive Readout of Superconducting Qubits*, Phys. Rev. App. **7**, 054020 (2017).

[17] 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. App. **6**, 034008 (2016).

[18] 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).

[19] Y. Sunada, S. Kono, J. Ilves, S. Tamate, T. Sugiyama, Y. Tabuchi, and Y. Nakamura, *Fast readout and reset of a superconducting qubit coupled to a resonator with an intrinsic purcell filter*, Phys. Rev. Appl. **17**, 044016 (2022).

[20] A. Bengtsson, A. Opremcak, M. Khezri, D. Sank, A. Bourassa, K. J. Satzinger, S. Hong, C. Erickson, B. J. Lester, K. C. Miao, A. N. Korotkov, J. Kelly, Z. Chen, and P. V. Klimov, *Model-based optimization of superconducting qubit readout*, 2023, arXiv:2308.02079 [quant-ph].

[21] F. Swiadek, R. Shillito, P. Magnard, A. Remm, C. Hellings, N. Lacroix, Q. Ficheux, D. C. Zanuz, G. J. Norris, A. Blais, S. Krinner, and A. Wallraff, *Enhancing dispersive readout of superconducting qubits through dynamic control of the dispersive shift: experiment and theory*, 2023, arXiv:2307.07765 [quant-ph].

9

[22] L. Chen, H.-X. Li, Y. Lu, C. W. Warren, C. J. Križan, S. Kosen, M. Rommel, S. Ahmed, A. Osman, J. Biznárová, A. Fadavi Roudsari, B. Lienhard, M. Caputo, K. Grigoras, L. Grönberg, J. Govenius, A. F. Kockum, P. Delsing, J. Bylander, and G. Tancredi, *Transmon qubit readout fidelity at the threshold for quantum error correction without a quantum-limited amplifier*, npj Quantum Information **9**, 26 (2023).

[23] C. Lledó, R. Dassonneville, A. Moulinas, J. Cohen, R. Shillito, A. Bienfait, B. Huard, and A. Blais, *Cloaking a qubit in a cavity*, 2022, arXiv:2211.05758 [quant-ph].

[24] M. H. Muñoz-Arias, C. Lledó, and A. Blais, *Qubit readouts enabled by qubit cloaking*, 2023, arXiv:2305.00895 [quant-ph].

[25] 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).

[26] 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).

[27] 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*, Phys. Rev. Lett. **127**, 200502 (2021).

[28] E. A. Sete, A. Q. Chen, R. Manenti, S. Kulshreshtha, and S. Poletto, *Floating tunable coupler for scalable quantum computing architectures*, Phys. Rev. Appl. **15**, 064063 (2021).

[29] J. Stehlik, D. M. Zajac, D. L. Underwood, T. Phung, J. Blair, S. Carnevale, D. Klaus, G. A. Keefe, A. Carniol, M. Kumph, M. Steffen, and O. E. Dial, *Tunable coupling architecture for fixed-frequency transmon superconducting qubits*, Phys. Rev. Lett. **127**, 080505 (2021).

[30] K. X. Wei, E. Magesan, I. Lauer, S. Srinivasan, D. F. Bogorin, S. Carnevale, G. A. Keefe, Y. Kim, D. Klaus, W. Landers, N. Sundaresan, C. Wang, E. J. Zhang, M. Steffen, O. E. Dial, D. C. McKay, and A. Kandala, *Hamiltonian engineering with multicolor drives for fast entangling gates and quantum crosstalk cancellation*, Phys. Rev. Lett. **129**, 060501 (2022).

[31] 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].

[32] P. V. Klimov, A. Bengtsson, C. Quintana, A. Bourassa, S. Hong, A. Dunsworth, K. J. Satzinger, W. P. Livingston, V. Sivak, M. Y. Niu, T. I. Andersen, Y. Zhang, D. Chik, Z. Chen, C. Neill, C. Erickson, A. G. Dau, A. Megrant, P. Roushan, A. N. Korotkov, J. Kelly, V. Smelyanskiy, Y. Chen, and H. Neven, *Optimizing quantum gates towards the scale of logical qubits*, 2023, arXiv:2308.02321 [quant-ph].

[33] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults*, Quantum Info. Comput. **7,** 139–156 (2007).

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

[35] M. Suchara, A. W. Cross, and J. M. Gambetta, *Leakage suppression in the toric code*, Quantum Info. Comput. **15,** 997–1016 (2015).

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

[37] 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).

[38] 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,** 1761 (2021).

[39] K. C. Miao, M. McEwen, J. Atalaya, D. Kafri, L. P. Pryadko, A. Bengtsson, A. Opremcak, K. J. Satzinger, Z. Chen, P. V. Klimov, C. Quintana, R. Acharya, K. Anderson, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. C. Bardin, A. Bourassa, J. Bovaird, L. Brill, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, J. Campero, B. Chiaro, R. Collins, P. Conner, A. L. Crook, B. Curtin, D. M. Debroy, S. Demura, A. Dunsworth, C. Erickson, R. Fatemi, V. S. Ferreira, L. F. Burgos, E. Forati, A. G. Fowler, B. Foxen, G. Garcia, W. Giang, C. Gidney, M. Giustina, R. Gosula, A. G. Dau, J. A. Gross, M. C. Hamilton, S. D. Harrington, P. Heu, J. Hilton, M. R. Hoffmann, S. Hong, T. Huang, A. Huff, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, J. Kelly, S. Kim, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, L. Laws, K. Lee, B. J. Lester, A. T. Lill, W. Liu, A. Locharla, E. Lucero, S. Martin, A. Megrant, X. Mi, S. Montazeri, A. Morvan, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, R. Potter, C. Rocque, P. Roushan, K. Sankaragomathi, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, N. Shutty, V. Shvarts, J. Skruzny, W. C. Smith, G. Sterling, M. Szalay, D. Thor, A. Torres, T. White, B. W. K. Woo, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, N. Zhu, N. Zobrist, H. Neven, V. Smelyanskiy, A. Petukhov, A. N. Korotkov, D. Sank, and Y. Chen, *Overcoming leakage in quantum error correction*, Nature Physics **19,** 1780–1786 (2023).

[40] J. F. Marques, H. Ali, B. M. Varbanov, M. Finkel, H. M. Veen, S. L. M. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider, B. M. Terhal, and L. DiCarlo, *All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits*, Phys. Rev. Lett. **130,** 250602 (2023).

9

[41] N. Lacroix, L. Hofele, A. Remm, O. Benhayoune-Khadraoui, A. McDonald, R. Shillito, S. Lazar, C. Hellings, F. Swiadek, D. Colao-Zanuz, A. Flasby, M. B. Panah, M. Kerschbaum, G. J. Norris, A. Blais, A. Wallraff, and S. Krinner, *Fast flux-activated leakage reduction for superconducting quantum circuits*, 2023, arXiv:2309.07060 [quant-ph].

[42] 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**, 383–387 (2021).

[43] M. Kjaergaard, M. Schwartz, J. Braumüller, P. Krantz, J. Wang, S. Gustavsson, and W. Oliver, *Superconducting qubits: current state of play*, Annu. Rev. Condens. Matter Phys. **11**, 369–395 (2020).

[44] A. P. M. Place, L. V. H. Rodgers, P. Mundada, B. M. Smitham, M. Fitzpatrick, Z. Leng, A. Premkumar, J. Bryon, A. Vrajitoarea, S. Sussman, G. Cheng, T. Madhavan, H. K. Babla, X. H. Le, Y. Gang, B. Jäck, A. Gyenis, N. Yao, R. J. Cava, N. P. de Leon, and A. A. Houck, *New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds*, Nature Communications **12**, 1779 (2021).

[45] C. Wang, X. Li, H. Xu, Z. Li, J. Wang, Z. Yang, Z. Mi, X. Liang, T. Su, C. Yang, G. Wang, W. Wang, Y. Li, M. Chen, C. Li, K. Linghu, J. Han, Y. Zhang, Y. Feng, Y. Song, T. Ma, J. Zhang, R. Wang, P. Zhao, W. Liu, G. Xue, Y. Jin, and H. Yu, *Towards practical quantum computers: transmon qubit with a lifetime approaching 0.5 milliseconds*, npj Quantum Information **8**, 3 (2022).

[46] S. Ganjam, Y. Wang, Y. Lu, A. Banerjee, C. U. Lei, L. Krayzman, K. Kisslinger, C. Zhou, R. Li, Y. Jia, M. Liu, L. Frunzio, and R. J. Schoelkopf, *Surpassing millisecond coherence times in on-chip superconducting quantum memories by optimizing materials, processes, and circuit design*, 2023, arXiv:2308.15539 [quant-ph].

[47] V. E. Manucharyan, J. Koch, L. I. Glazman, and M. H. Devoret, *Fluxonium: single cooper-pair circuit free of charge offsets*, Science **326**, 113–116 (2009).

[48] L. B. Nguyen, Y.-H. Lin, A. Somoroff, R. Mencia, N. Grabon, and V. E. Manucharyan, *High-coherence fluxonium qubit*, Phys. Rev. X **9**, 041041 (2019).

[49] A. Somoroff, Q. Ficheux, R. A. Mencia, H. Xiong, R. Kuzmin, and V. E. Manucharyan, *Millisecond coherence in a superconducting qubit*, Phys. Rev. Lett. **130**, 267001 (2023).

[50] C. Müller, J. Lisenfeld, A. Shnirman, and S. Poletto, *Interacting two-level defects as sources of fluctuating high-frequency noise in superconducting circuits*, Phys. Rev. B **92**, 035442 (2015).

[51] 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).

[52] J. J. Burnett, A. Bengtsson, M. Scigliuzzo, D. Niepce, M. Kudra, P. Delsing, and J. Bylander, *Decoherence benchmarking of superconducting qubits*, npj Quantum Information **5**, 54 (2019).

[53] T. Thorbeck, Z. Xiao, A. Kamal, and L. C. G. Govia, *Readout-induced suppression and enhancement of superconducting qubit lifetimes*, 2023, arXiv:2305.10508 [quant-ph].

[54] T. Thorbeck, A. Eddins, I. Lauer, D. T. McClure, and M. Carroll, *Two-level-system dynamics in a superconducting qubit due to background ionizing radiation*, PRX Quantum **4**, 020356 (2023).

[55] L. Grünhaupt, N. Maleeva, S. T. Skacel, M. Calvo, F. Levy-Bertrand, A. V. Ustinov, H. Rotzinger, A. Monfardini, G. Catelani, and I. M. Pop, *Loss mechanisms and quasiparticle dynamics in superconducting microwave resonators made of thin-film granular aluminum*, Phys. Rev. Lett. **121**, 117001 (2018).

[56] 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, S. Gustavsson, J. A. Formaggio, B. A. VanDevender, and W. D. Oliver, *Impact of ionizing radiation on superconducting qubit coherence*, Nature **584**, 551–556 (2020).

[57] 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*, Nature Physics **18**, 107–111 (2022).

[58] C. D. Wilen, S. Abdullah, N. A. Kurinsky, C. Stanford, L. Cardani, G. D'Imperio, C. Tomei, L. Faoro, L. B. Ioffe, C. H. Liu, A. Opremcak, B. G. Christensen, J. L. DuBois, and R. McDermott, *Correlated charge noise and relaxation errors in superconducting qubits*, Nature **594**, 369–373 (2021).

9

[59] A. Strikis, S. C. Benjamin, and B. J. Brown, *Quantum computing is scalable on a planar array of qubits with fabrication defects*, Phys. Rev. Appl. **19**, 064081 (2023).

[60] L. Cardani, F. Valenti, N. Casali, G. Catelani, T. Charpentier, M. Clemenza, I. Colantoni, A. Cruciani, G. D'Imperio, L. Gironi, L. Grünhaupt, D. Gusenkova, F. Henriques, M. Lagoin, M. Martinez, G. Pettinari, C. Rusconi, O. Sander, C. Tomei, A. V. Ustinov, M. Weber, W. Wernsdorfer, M. Vignati, S. Pirro, and I. M. Pop, *Reducing the impact of radioactivity on quantum circuits in a deep-underground facility*, Nature Communications **12**, 2733 (2021).

[61] J. M. Martinis, *Saving superconducting quantum processors from decay and correlated errors generated by gamma and cosmic rays*, npj Quantum Information **7**, 90 (2021).

[62] R.-P. Riwar, A. Hosseinkhani, L. D. Burkhart, Y. Y. Gao, R. J. Schoelkopf, L. I. Glazman, and G. Catelani, *Normal-metal quasiparticle traps for superconducting qubits*, Phys. Rev. B **94**, 104516 (2016).

[63] U. Patel, I. V. Pechenezhskiy, B. L. T. Plourde, M. G. Vavilov, and R. McDermott, *Phonon-mediated quasiparticle poisoning of superconducting microwave resonators*, Phys. Rev. B **96**, 220501 (2017).

[64] F. Henriques, F. Valenti, T. Charpentier, M. Lagoin, C. Gouriou, M. Martínez, L. Cardani, M. Vignati, L. Grünhaupt, D. Gusenkova, J. Ferrero, S. T. Skacel, W. Wernsdorfer, A. V. Ustinov, G. Catelani, O. Sander, and I. M. Pop, *Phonon traps reduce the quasiparticle density in superconducting circuits*, Applied Physics Letters **115**, 212601 (2019).

[65] V. Iaia, J. Ku, A. Ballard, C. P. Larson, E. Yelton, C. H. Liu, S. Patel, R. McDermott, and B. L. T. Plourde, *Phonon downconversion to suppress correlated errors in superconducting qubits*, Nature Communications **13**, 6425 (2022).

[66] D. K. Tuckett, S. D. Bartlett, and S. T. Flammia, *Ultrahigh error threshold for surface codes with biased noise*, Phys. Rev. Lett. **120**, 050505 (2018).

[67] D. K. Tuckett, A. S. Darmawan, C. T. Chubb, S. Bravyi, S. D. Bartlett, and S. T. Flammia, *Tailoring surface codes for highly biased noise*, Phys. Rev. X **9**, 041031 (2019).

[68] D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, *Fault-tolerant thresholds for the surface code in excess of 5% under biased noise*, Phys. Rev. Lett. **124**, 130501 (2020).

[69] J. P. Bonilla Ataides, D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, *The xzzx surface code*, Nature Communications **12**, 2172 (2021).

[70] A. S. Darmawan, B. J. Brown, A. L. Grimsmo, D. K. Tuckett, and S. Puri, *Practical quantum error correction with the xzzx code and kerr-cat qubits*, PRX Quantum **2**, 030345 (2021).

[71] B. Srivastava, A. Frisk Kockum, and M. Granath, *The XYZ^2 hexagonal stabilizer code*, Quantum **6**, 698 (2022).

[72] Q. Xu, N. Mannucci, A. Seif, A. Kubica, S. T. Flammia, and L. Jiang, *Tailored xzzx codes for biased noise*, Phys. Rev. Res. **5**, 013035 (2023).

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

[74] C. Chamberland, K. Noh, P. Arrangoiz-Arriola, E. T. Campbell, C. T. Hann, J. Iverson, H. Putterman, T. C. Bohdanowicz, S. T. Flammia, A. Keller, G. Refael, J. Preskill, L. Jiang, A. H. Safavi-Naeini, O. Painter, and F. G. Brandão, *Building a fault-tolerant quantum computer using concatenated cat codes*, PRX Quantum **3**, 010329 (2022).

[75] J. Guillaud and M. Mirrahimi, *Repetition cat qubits for fault-tolerant quantum computation*, Phys. Rev. X **9**, 041053 (2019).

[76] T. M. Stace, S. D. Barrett, and A. C. Doherty, *Thresholds for topological codes in the presence of loss*, Phys. Rev. Lett. **102**, 200501 (2009).

[77] S. D. Barrett and T. M. Stace, *Fault tolerant quantum computation with very high threshold for loss errors*, Phys. Rev. Lett. **105**, 200502 (2010).

[78] Y. Wu, S. Kolkowitz, S. Puri, and J. D. Thompson, *Erasure conversion for fault-tolerant quantum computing in alkaline earth rydberg atom arrays*, Nature Communications **13**, 4657 (2022).

[79] K. S. Chou, T. Shemma, H. McCarrick, T.-C. Chien, J. D. Teoh, P. Winkel, A. Anderson, J. Chen, J. Curtis, S. J. de Graaf, J. W. O. Garmon, B. Gudlewski, W. D. Kalfus, T. Keen, N. Khedkar, C. U. Lei, G. Liu, P. Lu, Y. Lu, A. Maiti, L. Mastalli-Kelly, N. Mehta, S. O. Mundhada, A. Narla, T. Noh, T. Tsunoda, S. H. Xue, J. O. Yuan, L. Frunzio, J. Aumentado, S. Puri, S. M. Girvin, J. a. S. Harvey Moseley, and R. J. Schoelkopf, *Demonstrating a superconducting dual-rail cavity qubit with erasure-detected logical measurements*, 2023, arXiv:2307.03169 [quant-ph].

[80] K. Sahay, J. Jin, J. Claes, J. D. Thompson, and S. Puri, *High threshold codes for neutral atom qubits with biased erasure errors*, 2023, arXiv:2302.03063 [quant-ph].

[81] M. Kang, W. C. Campbell, and K. R. Brown, *Quantum error correction with metastable states of trapped ions using erasure conversion*, PRX Quantum **4**, 020358 (2023).

[82] A. Kubica, A. Haim, Y. Vaknin, F. Brandão, and A. Retzker, *Erasure qubits: overcoming the T_1 limit in superconducting circuits*, 2022, arXiv:2208.05461 [quant-ph].

[83] D. Gottesman, *Fault-tolerant quantum computation with constant overhead*, Quantum Info. Comput. **14**, 1338–1372 (2014).

[84] M. A. Tremblay, N. Delfosse, and M. E. Beverland, *Constant-overhead quantum error correction with thin planar connectivity*, Phys. Rev. Lett. **129**, 050504 (2022).

[85] N. P. Breuckmann and J. N. Eberhardt, *Quantum low-density parity-check codes*, PRX Quantum **2**, 040101 (2021).

[86] A. Leverrier and G. Zémor, *Quantum tanner codes*, in 2022 ieee 63rd annual symposium on foundations of computer science (focs) (2022), pp. 872–883.

9

[87] P. Panteleev and G. Kalachev, *Asymptotically good quantum and locally testable classical ldpc codes*, in Proceedings of the 54th annual acm sigact symposium on theory of computing, STOC 2022 (2022), pp. 375–388.

[88] N. P. Breuckmann and B. M. Terhal, *Constructions and noise threshold of hyperbolic surface codes*, IEEE Transactions on Information Theory **62**, 3731–3744 (2016).

[89] O. Higgott and N. P. Breuckmann, *Constructions and performance of hyperbolic and semi-hyperbolic floquet codes*, 2023, arXiv:2308.03750 [quant-ph].

[90] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, *High-threshold and low-overhead fault-tolerant quantum memory*, 2023, arXiv:2308.07915 [quant-ph].

[91] T. E. O'Brien, B. M. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Information **3** (2017).

[92] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. Di-Carlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Information **6**, 102 (2020).

[93] F. Battistel, C. Chamberland, K. Johar, R. W. J. Overwater, F. Sebastiano, L. Skoric, Y. Ueno, and M. Usman, *Real-time decoding for fault-tolerant quantum computing: progress, challenges and outlook*, Nano Futures **7**, 032003 (2023).

[94] B. M. Terhal, *Quantum error correction for quantum memories*, Rev. Mod. Phys. **87**, 307–346 (2015).

[95] L. Skoric, D. E. Browne, K. M. Barnes, N. I. Gillespie, and E. T. Campbell, *Parallel window decoding enables scalable fault tolerant quantum computation*, 2023, arXiv:2209.08552 [quant-ph].

[96] B. Barber, K. M. Barnes, T. Bialas, O. Buğdaycı, E. T. Campbell, N. I. Gillespie, K. Johar, R. Rajan, A. W. Richardson, L. Skoric, C. Topal, M. L. Turner, and A. B. Ziad, *A real-time, scalable, fast and highly resource efficient decoder for a quantum computer*, 2023, arXiv:2309.05558 [quant-ph].

[97] N. Liyanage, Y. Wu, A. Deters, and L. Zhong, *Scalable quantum error correction for surface codes using fpga*, 2023, arXiv:2301.08419 [quant-ph].

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

[99] O. Higgott, T. C. Bohdanowicz, A. Kubica, S. T. Flammia, and E. T. Campbell, *Improved decoding of circuit noise and fragile boundaries of tailored surface codes*, Phys. Rev. X **13**, 031007 (2023).

[100] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. J. Beenakker, *Machine-learning-assisted correction of correlated qubit errors in a topological code*, Quantum **2**, 48 (2018).

[101] P. Baireuther, M. D. Caio, B. Criger, C. W. J. Beenakker, and T. E. O'Brien, *Neural network decoder for topological color codes with circuit level noise*, New Journal of Physics **21**, 013003 (2019).

9

[102]   B. M. Varbanov, M. Serra-Peralta, D. Byfield, and B. M. Terhal,   *Neural network decoder for near-term surface-code experiments*, 2023, arXiv:2307.03280 [quant-ph].

[103]   S. Spitz, B. M. Tarasinski, C. Beenakker, and T. O'Brien, *Adaptive weight estimator for quantum error correction in a time-dependent environment,* Advanced Quantum Technologies **1**, 1800012 (2018).

9

ACKNOWLEDGEMENTS

Writing this dissertation has allowed me to reflect on my work over the past few years. I am fortunate to have worked with so many amazing people. I want to use the last few pages of my dissertation to express my gratitude to everyone who has helped me in one way or another during this journey.

I couldn't have asked for a better supervisor than you, **Barbara**. Working with you has been very rewarding, and I have learned so much over these past years. I am amazed by your depth of knowledge and your seemingly endless energy. You have shown me that research is also about having fun and exploring whatever you are curious about. I am very grateful for the freedom you gave me when choosing the projects I worked on and the support you offered me whenever I was stuck on something.

I must also thank my other promoter, **Leo**. I am very grateful that I could work so closely together with your lab's experimental efforts. The potential impact that my research could have was always a big motivation for me. I also appreciate your attention to detail and scientific integrity.

I would also like to thank all the members of my defense committee for the time they dedicated to reading my dissertation, providing feedback, and participating in my defense ceremony.

I am also very grateful to **Francesco** for being my paranymph and comrade in the battle to detect and mitigate leakage. Thank you for the many discussions we had while we were trying to understand leakage better and for showing me how valuable a ruler is when making plots. I am happy that we have stayed in touch since you graduated, and I hope we continue to keep in touch in the future. I would also like to thank **Marc** - working with you on neural network decoding was fun. I was very impressed with your work ethic, your numerical skills, and your appreciation for writing clean code. It is great that you decided to join the group and will continue working on error correction experiments. **Alessandro**, it was great to collaborate with you on the transmon-fluxonium project. I had a lot of fun exploring the fluxonium qubit and figuring out how to realize two-qubit gates with a transmon qubit. I also think that we ended up developing a pretty nice simulation package. **Maarten**, it was nice to hear you talk about the problems you were working on, even if I did not always understand everything. Also, thank you once more for your help with translating my summary into Dutch. **Mac**, it's great that we once again have someone in the group working on bosonic codes, though I also think the work you are doing with time-dynamic circuits is fascinating. **Yaroslav**, it was always impressive to me how many ideas you had and how many projects you were working on simultaneously. You also had a great sense of humor, and it was always fun to chat with you. **Yang**, even though we were working with two very different systems, it was great to have somebody to whom I could complain about the difficulties of error correction experiments. **Jasper**, you did an excellent job investigating the various problems one can encounter when estimating the error rates from the syndrome defects, and it was nice

BIBLIOGRAPHY

working with you. I also want to thank all the previous members of the group with whom I have had the pleasure of discussing various topics: **Minh**, **Zherui**, **Marios**, **Joel**, **Matteo**, **Xiaotong**, **Daniel**, and **Manuel**.

I want to thank **Jorge** and **Hany** for being great friends and collaborating with me on several projects. Thank you for all the time you've spent explaining to me how the experiments were set up and how the various operations were calibrated. Even though it was often challenging, I had a lot of fun when we were trying to figure out what was limiting the experimental results. I also want to thank **Brian** for introducing me to *quantumsim* and the problem of qubit leakage and how to simulate it. I also want to thank **Tom** for all the help and discussions on decoding the surface code and for developing the method we used to detect leakage. I am grateful to **Slava** for his help with developing and maintaining *quantumsim*. I want to thank **Adriaan**, **Niels**, **Ramiro**, and **Thijs** for all the discussions when I first started learning about transmon qubits and how the various operations were realized in practice. I also want to thank **Berend**, **Matvey**, **Sean**, **Nandini**, **Santi**, **Tim**, and **Miguel** for all the conversations they have had with me.

I would like to thank **David**, **Joonas**, **Ophelia**, and **Earl** from Riverlane for discussing decoding with me. A special mention to **David** - it was great to collaborate with you on the neural network decoding paper. I want to thank **Christian** for collaborating with us on the transmon-fluxonium project. I am grateful to the management staff at QuTech and, in particular, **Jenny** and **Marja** for making my life significantly easier.

Of course, I can't forget all of the fantastic friends I made in Delft and with whom I have shared many unforgettable moments. We have had many great trips together and fun evenings watching movies or playing games. We often stayed up later than we should have, going either for a game of darts or a trip to the "bee house" at three in the morning. **Christian**, it was terrific that we shared so many hobbies and interests. It was great to discuss and share music with you, and I also had a blast going to concerts and music festivals with you. But apart from that, I could talk to you about cooking and video games. Of course, you also have a great sense of humor, and I appreciate all the silly puns and dad jokes you've shared over the years. **Helene**, you are hilarious, and I loved hanging out with you. Of course, it was nice that you were also into obscure music. I appreciate all the country, metal, and pop songs you shared with me. It is remarkable how creative you are – making cheese, growing vegetables, painting, sewing. And, of course, I can't forget about "theory coffee," which was great and perhaps why we often spent more time talking than we should have. **Ivan**, thank you for being my paranymph. You are perhaps the hardest-working person I know, even though we often complain about how busy we are. It felt like you were always willing to lend a helping hand - thanks for helping me move apartments and for all the other times you've helped me out with something. Whether we were working out at the gym, walking back home from work, or watching TV, I always had a great time hanging out with you. **Alina**, **Marta**, **Hugo**, **Mattias**, **Rasa**, **Eugenio**, **Chien**, **Santi**, you are all awesome people, and I am happy we have shared so many great moments together. **Guille**, thank you for discussing machine learning with me and for all the jokes you've shared with me.

Last but not least, I would like to thank my family. Without your unconditional support, I would not be the man I am today. I love you and hope to support you just as much as you have supported me.

CURRICULUM VITÆ

Boris Mihailov VARBANOV

10-09-1994   Born in Ruse, Bulgaria.

EDUCATION

2009–2013   Diploma for Secondary Education

First English Language School, Sofia, Bulgaria

2013–2016   Bachelor of Science in Physics

Royal Holloway University of London, Egham, United Kingdom

BSc. Thesis: Characterization of the helimagnetic insulator Cu_2OSeO_3 using X-ray diffraction and magnetic susceptibility

Supervisor: Dr. P. Niklowitz and Dr. J. Poulten

2017–2019   Master of Science in Applied Physics

Delft University of Technology, Delft, The Netherlands

Internship: Simulation and design of SQUID-array based Josephson Parametric Amplifiers

Supervisors: Dr. T. Last and R. Vollmer

MSc. Thesis: Characterization, detection, and mitigation of leakage in simulations of a superconducting-qubit implementation of Surface-17

Supervisors: Prof. dr. L. DiCarlo and Dr. B. Tarasinski

2019–2024   Ph.D. in Applied Physics

Delft University of Technology, Delft, The Netherlands

Thesis: Optimizing Quantum Error Correction for Superconducting-qubit Processors

Promotors: Prof. dr. B.M. Terhal and Prof. dr. L. DiCarlo

LIST OF PUBLICATIONS

7. **B.M. Varbanov**, M. Serra-Peralta, D. Byfield, B.M. Terhal *Neural network decoder for near-term surface-code experiments*, arXiv: 2307.03280 (2023).

6. J.F. Marques, H. Ali, **B.M. Varbanov**, M. Finkel, H.M. Veen, S.L.M. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider, B.M. Terhal, L. DiCarlo *All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits*, Phys. Rev. Lett. **130**, 250602 (2023).

5. A. Ciani, **B.M. Varbanov**, N. Jolly, C.K. Andersen, and B.M. Terhal, *Microwave-activated gates between a fluxonium and a transmon qubit*, Phys. Rev. Research **4**, 043127 (2022).

4. 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*, Nat. Phys. **18**, 80–86 (2022).

3. 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).

2. **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).

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