Building a fault-tolerant quantum computer using concatenated cat codes

Christopher Chamberland,[1, 2] Kyungjoo Noh,[1] Patricio Arrangoiz-Arriola,[1, *] Earl T. Campbell,[1, *] Connor T. Hann,[1, 3, *] Joseph Iverson,[1, *] Harald Putterman,[1, *] Thomas C. Bohdanowicz,[1, 2] Steven T. Flammia,[1] Andrew Keller,[1] Gil Refael,[1, 2] John Preskill,[1, 2] Liang Jiang,[1, 4] Amir H. Safavi-Naeini,[1, 5] Oskar Painter,[1, 2] and Fernando G.S.L. Brandão[1, 2]

[1] *AWS Center for Quantum Computing, Pasadena, CA 91125, USA*
[2] *IQIM, California Institute of Technology, Pasadena, CA 91125, USA*
[3] *Department of Physics, Yale University, New Haven, CT 06511, USA*
[4] *Pritzker School of Molecular Engineering, The University of Chicago, Illinois 60637, USA*
[5] *Department of Applied Physics and Ginzton Laboratory, Stanford University, Stanford, CA 94305, USA*

We present a comprehensive architectural analysis for a fault-tolerant quantum computer based on cat codes concatenated with outer quantum error-correcting codes. For the physical hardware, we propose a system of acoustic resonators coupled to superconducting circuits with a two-dimensional layout. Using estimated near-term physical parameters for electro-acoustic systems, we perform a detailed error analysis of measurements and gates, including CNOT and Toffoli gates. Having built a realistic noise model, we numerically simulate quantum error correction when the outer code is either a repetition code or a thin rectangular surface code. Our next step toward universal fault-tolerant quantum computation is a protocol for fault-tolerant Toffoli magic state preparation that significantly improves upon the fidelity of physical Toffoli gates at very low qubit cost. To achieve even lower overheads, we devise a new magic-state distillation protocol for Toffoli states. Combining these results together, we obtain realistic full-resource estimates of the physical error rates and overheads needed to run useful fault-tolerant quantum algorithms. We find that with around 1,000 superconducting circuit components, one could construct a fault-tolerant quantum computer that can run circuits which are intractable for classical supercomputers. Hardware with 32,000 superconducting circuit components, in turn, could simulate the Hubbard model in a regime beyond the reach of classical computing.

CONTENTS

I. Introduction 2
A. Overview of main results 3

II. Hardware implementation and stabilization schemes 5
A. Overview of cat codes and driven-dissipative stabilization 5
B. Physical implementation of buffer and storage resonators 7
C. Wiring and layout 7
D. Calculation of the loss parameter κ_1/κ_2 8
E. Multiplexed stabilization 10
F. Crosstalk 10

III. Gates and Measurements 12
A. X Gate 12
B. CNOT 12
C. Toffoli 14
D. Z Rotation 16
E. CZ Rotation 16
F. Physical implementation of the gates 16
G. X measurement 18
H. Z measurement 20

IV. Logical failure rates for quantum memory 21
A. Repetition code logical failure rates 22
B. Rotated surface code logical failure rates 24
C. Surface code logical failure rates in the presence of crosstalk errors 25

V. Computation by lattice surgery and timelike errors 27

VI. Toffoli distillation: Bottom-up scheme 28

VII. Top-Down Scheme for higher fidelity Toffoli gates 31

VIII. Overhead Estimates 34

IX. Conclusion 35

Acknowledgments 36

References 36

A. Engineering two-phonon dissipation with piezoelectric nanostructures 39
1. Implementation of the required Josephson nonlinearity 39
2. Calculation of nonlinear interaction rate g_2 40
3. Classical filter theory and derivation of dissipation rates 41
4. Filter design 45
a. General properties 45

[*] These authors contributed equally

b. Optimizing the filter 47
5. Optimization of the dimensionless loss κ_1/κ_2 49

B. Multiplexed stabilization and crosstalk 49
1. Effective operator formalism 49
2. Simultaneous stabilization of multiple cat qubits with a single ATS 50
3. Sources of crosstalk 52
Type I: stochastic errors induced by effective dissipators 53
Type II: stochastic errors induced by effective Hamiltonians 53
Type III: coherent errors induced by effective Hamiltonians 54
4. Crosstalk mitigation: filtering 54
Suppression of Type I errors 54
Suppression of Type II errors 55
5. Crosstalk mitigation: mode frequency optimization 57

C. Shifted Fock basis 60

D. Perturbative analysis of the Z error rates of the cat qubit gates 62
1. Idling 62
2. Z rotations 62
3. CZ rotations 63
4. CNOT 64
5. Toffoli 67

E. Simulations of gate error rates 69
1. CNOT 69
2. Toffoli 72
3. Z and CZ 73

F. Physical implementation of cat qubit gates 73

G. Measurement 77
1. X-basis measurement with transmon 77
2. X-basis measurement without transmon 79
3. Z-basis measurement 81

H. STOP algorithm 83

I. Stabilizer operations with the repetition code 86
1. Computational basis states 86
2. Implementation of logical Clifford gates 87

J. Growing encoded data qubits to larger code distances with the repetition code 88

K. Toffoli simulation twirling approximation 89

L. Fitting procedure for memory and lattice surgery 90

M. Edge weights and decoding graphs for the repetition and surface codes 92
1. Repetition code decoding graphs 92
2. Surface code decoding graphs 94
3. Adding edges for dealing with correlated errors 99
4. Decoding time-like errors 100

N. Toffoli state distillation (TDTOF) 101
1. Prior state of the art 101
2. Transversality proofs 101
3. Trading space and time 104
4. Implementing Conjugated-CCZ gates 105
a. Injection with CNOT gates and single qubit measurements 106
b. Injection using lattice surgery 107
5. Error propagation and detection 107
a. Truncation errors 110
b. Generic noise 110
6. Noise tailoring through Clifford symmetries 111
7. Factory layout and scheduling 112
8. Clifford noise 115

I. INTRODUCTION

Building a fault-tolerant quantum computer is one of the great scientific and engineering challenges of the 21st century. A successful quantum computing architecture must meet many conflicting demands: it must have a threshold error rate that is achievable by hardware on a large scale, a convenient physical layout, and low overhead for fault-tolerant algorithms. All proposed quantum architectures require tradeoffs among these objectives. For example, the most popular proposed architecture, the surface code [1], has a convenient 2D physical layout and relatively high threshold error rates, but the overhead for running useful algorithms remains daunting [2–6], even after years of optimization.

Recent work has shown that qubits with highly *biased* noise are a promising route to fault tolerance [7–10], at least when gates that preserve the noise bias can be easily implemented in the architecture [11–14]. One possible route to realizing such qubits is via two-component cat code [15–17], a bosonic qubit encoded in an oscillator mode [18–20], subjected to engineered two-photon dissipation [15, 21, 22] or engineered Kerr nonlinearity [16, 23, 24]. The engineered interaction heavily suppresses population transfer between the two constituent coherent states of the cat qubit, causing an effective noise bias towards phase-flip errors on the encoded logical qubits [15, 16]. Furthermore, bias-preserving CNOT and Toffoli (TOF) gates can be performed for these cat codes [12, 13]. Therefore, concatenating the (inner) cat code with another (outer) quantum error-correcting code can be done with great efficiency by tailoring the outer code to suppress the dominant phase-flip errors. We call these coding schemes *concatenated cat codes*. This idea has been explored previously for the case where the outer code is a repetition code [13, 14], and experiments suggest that strong suppression of biased errors is possible with this approach [25].

In this paper, we give a full-stack analysis of a fault-tolerant quantum architecture based on cat codes concatenated with outer quantum error-correcting codes. We propose a blueprint for a possible practical implementation based on hybrid electro-acoustic systems consisting of acoustic resonators coupled to superconducting circuits. These systems are a promising platform for realizing concatenated cat codes due to their small footprint [26], potential for ultra-high coherence times [27], and easy integration with superconducting circuits for control and read-out [28].

We give a comprehensive error analysis of this approach that provides a detailed picture of the physically achievable error rates for gates and measurements based on estimated parameters for coupling strengths and phonon loss and dephasing rates. Using these data, we then explicitly analyze quantum error correction when the outer code is either a repetition code or a thin rectangular surface code. Then we show how to build a fault-tolerant quantum computer in our architecture, combining lattice surgery and magic state distillation for Toffoli states. Finally, we provide an estimate of overhead as a function of physical error rates required to run fault-tolerant quantum algorithms.

Our analysis can be broadly classified into three categories: 1) a hardware proposal; 2) a physical-layer analysis of gate and measurements errors; and 3) a logical-level analysis of memory and computation failure rates. More specifically, in Section II we describe our hardware proposal for using phononic bandgap resonators and superconducting circuits to store and process quantum information at the physical level. Then in Section III we give a complete analysis of gate and measurement errors for phononic qubits using realistic noise parameters that we expect from the hardware proposal. In Sections IV, VI, and VII we give a gate-level analysis of universal fault-tolerant quantum computation that looks at logical error rates across a physically relevant parameter regime.

While the main purpose of this paper is to integrate this analysis, many of our results are independently interesting. For example, other architectures that encode a qubit into an oscillator using a cat code can benefit from our analysis of gate errors. Our gate-level analysis of fault-tolerant quantum computation introduces new ideas that will be more broadly useful for any architecture that uses Toffoli magic state distillation or lattice surgery. And our novel frequency-multiplexed stabilization scheme can be leveraged to improve hardware efficiency in any system that uses engineered dissipation and provides the required connectivity to implement the outer codes.

A. Overview of main results

In Section II we describe our hardware proposal for using phononic-crystal-defect resonators (PCDRs), of the type reported in Ref. [28], as the storage elements. These are periodically patterned suspended nanostructures that support localized acoustic resonances in the gigahertz range. They are fabricated from a piezoelectric material such as $\mathrm{LiNbO_3}$, which allows us to couple these resonances to superconducting circuits with nearly the same strength as ordinary electromagnetic cavities.

A key parameter in our proposal is the dimensionless loss κ_1/κ_2, where κ_1 is the single-phonon loss rate (per time) and κ_2 is the engineered two-phonon dissipation rate (per time) stabilizing the cat-code subspace. Calculating accurate predictions for this parameter is crucial for estimating the performance of the higher levels in the stack—such as the outer error-correcting codes—and the feasibility of the architecture.

This loss κ_1/κ_2 compactly summarizes a variety of physical processes. The two-phonon dissipation rate κ_2 is an engineered quantity that we can calculate from first principles. By contrast, the single-phonon loss rate κ_1 involves several dissipation channels, only some of which are directly controllable. Those decay processes that are intrinsic (i.e., not directly controllable) are due to a rich variety of mechanisms such as decay into ensembles of "two-level system" defects or quasi-particles in the superconductors [29], and are difficult to quantify from first principles. Our best reference for these loss rates is experimental data.

Because of these difficulties, we estimate κ_1/κ_2 using a hybrid approach. We first calculate κ_2 using a semi-classical description of the underlying superconducting circuits and then infer the value of κ_1 that is necessary to reach the regime $\kappa_1/\kappa_2 \sim 10^{-5}$, as we show this is the desired regime for running useful algorithms. We note that our architecture can tolerate larger values of κ_1/κ_2 at the cost of using more resources. For instance, if $\kappa_1/\kappa_2 = 2 \times 10^{-5}$, the overhead requirements are still competitive with other architectural proposals.

Following a recent demonstration [25], we propose implementing the two-phonon dissipation by engineering an interaction through which the storage mode exchanges excitations with an ancillary "buffer" mode *in pairs*. This buffer is strongly coupled to a reservoir, so these excitations rapidly decay into the bath. We model this reservoir by an arbitrary admittance function $Y(\omega)$, which makes our analysis quite general. In particular, we compute κ_2 when this reservoir is a multi-pole bandpass filter connected to a $50\,\Omega$ waveguide. The filter allows us to control the density of states of the reservoir, causing it to vanish at all frequencies except those within the filter passband. This is useful not only to protect the storage mode from radiative (or Purcell) decay, but it is also a crucial requirement to suppress correlated phase-flip errors while stabilizing multiple storage modes simultaneously with the same buffer mode. We explicitly design a filter and optimize it to obtain the largest possible value of κ_2. Interestingly, we find that this maximal value of κ_2 is determined solely by the filter bandwidth.

Given the bandwidth limitations imposed by the need to multiplex our stabilization (see below), we find $\kappa_2/2\pi \approx 500\,\mathrm{kHz}$. This imposes the requirement that the intrinsic relaxation time of the storage modes be at least $T_1 \approx 32\,\mathrm{ms}$ and $\approx 16\,\mathrm{ms}$ in order to reach $\kappa_1/\kappa_2 \sim 10^{-5}$ and $\sim 2 \times 10^{-5}$, respectively. While the required T_1 is rather high, it seems within reach in the near future. Indeed, single-crystal designs on silicon have been recently realized with $T_1 \approx 1.5\,\mathrm{s}$ [27]. Unfortunately these devices cannot be easily coupled to superconducting circuits. However we believe they offer insight into the loss processes on nanomechanical resonators and suggest similar levels of coherence with piezoelectric devices might be attainable.

The engineered dissipation needed to stabilize each cat code is provided by coupling each phononic resonator to nonlinear circuit elements. Specifically, we follow the approach of Ref. [25], where the nonlinearity is provided by a circuit element variant of a superconducting quantum interference device (SQUID) called an Asymmetrically Threaded SQUID (ATS). While Ref. [25] demonstrated an ATS can be used to stabilize a single mode into a cat code, our hardware layout necessitates that each ATS couple to and stabilize multiple resonators simultaneously. We present a simple scheme for this multiplexed stabilization, and provide a detailed analysis of the crosstalk that arises from coupling multiple modes to the same ATS. Moreover, we show that by employing a bandpass filter and carefully optimizing the phonon-mode frequencies, we are able to largely suppress the dominant sources of crosstalk in our system.

In Section III, we then analyze the errors in our gates and measurements. To do this, we introduce a method that we call the shifted Fock basis method. This method allows us to efficiently perform a perturbative analysis of the dominant Z error rates of the cat-qubit gates and improve the efficiency of numerical simulation of large cat qubits compared to the usual Fock basis method. The shifted Fock basis method allows us to compute the Z error rates of various cat-qubit gates using a small Hilbert space dimension that is independent of the average excitation number $|\alpha|^2$ of the cat qubit.

Using this method, we go on to show that the optimal Z error rates (per gate) of the cat qubit gates at the optimal gate time scale as $\sqrt{\kappa_1/\kappa_2}$. The optimal Z error rates of the CNOT and TOF gates are in fact independent of the size of the cat qubit, whereas those of Z and CZ rotations decrease linearly in $1/|\alpha|$.

We also study the effects of bosonic dephasing on various cat-qubit gates. We numerically find that although the Z error rates of the Z and CZ rotations are not at all affected by the dephasing, those of CNOT and TOF gates are adversely affected by the dephasing. This is surprising given that dephasing does not change the parity of the cat qubit. We provide a perturbative analysis to explain this unexpected behavior and attribute the enhanced Z error rates of the CNOT and TOF gates to the fact that the stabilizing jump operators for the target cat qubits are not static and instead rotate conditioned on the state of the control qubits. Our perturbative analysis agrees well with our numerical results, and they predict that the optimal Z error rates of the CNOT and TOF gates scale as $\sqrt{\kappa_\phi/\kappa_2}$, where κ_ϕ is the dephasing rate. That is, the effects of dephasing on the CNOT and TOF gates are comparable to the effects of phonon loss and thus should not be ignored unlike in the case of Z and CZ rotations.

We then develop and analyze schemes for readout in both the X and Z bases, enabling fast and hardware-efficient stabilizer measurement. For X-basis readout we propose two methods. The first uses an additional readout mode in every unit cell of our layout which we measure using a transmon. By performing repeated quantum non-demolition (QND) bosonic parity measurements in parallel with the gates of the subsequent error correction cycle, we can suppress the infidelity mechanisms associated with the transmon while having minimal impact on cycle time. We have simulated the X measurement scheme achieving average readout error probabilities of less than $2 * 10^{-3}$ for $\kappa_1/\kappa_2 < 10^{-4}$. The second scheme utilizes the ATS itself for X-basis read-out, avoiding the need of an extra transmon and allowing for fewer modes per ATS. We analyze this scheme which gives average readout error probabilities of roughly $4 * 10^{-3}$ for $\kappa_1/\kappa_2 < 10^{-4}$.

In addition, we propose a method for high-fidelity and fast Z-basis readout using only the storage and buffer modes; the resulting error probability decays exponentially as a function of $|\alpha|^2$, and is only weakly sensitive to phonon loss and dephasing. In this scheme excitations are swapped to the buffer mode where they leak to the transmission line and are detected via a homodyne measurement. We derive the signal-to-noise ratio and infidelity for this readout scheme, which we validate with simulations.

With a clear understanding of gate and measurement error rates, we proceed in Section IV to analyze the logical failure rates for a quantum memory based on concatenation into two codes: a repetition code and a thin rectangular surface code. We compute logical Z failure rates for both the repetition codes and the surface code. In the case of the surface code, we compute explicit leading order failure rates for logical X errors as a function of the Z-distance of the code. Our thresholds are computed using a full circuit-level simulation and a minimum-weight perfect matching (MWPW) decoder.

Using the thin surface code, we consider lattice surgery as a means of performing logical Clifford operations in Section V. By extending our full circuit-level simulation to model timelike errors during lattice surgery, we obtain logical error probabilities for Clifford operations.

To simulate universal quantum computation with Toffoli gates, we introduce in Section VI a new protocol to fault-tolerantly prepare TOF magic states encoded in the repetition code. Due to the fault-tolerant properties of our protocol, all gates required in our circuits can be implemented at the physical level. Hence we refer to such an approach as a bottom-up approach for preparing TOF magic states. The main insight is that a TOF state can be prepared by measuring a single Clifford observable, which can be achieved using a sequence of physical CNOT and TOF gates. To ensure fault-tolerance, this Clifford measurement has to be repeated a fixed number of times, but due to suppressed bit-flip noise the state does not significantly decohere during this measurement process. Using the full circuit-level noise model of Section III and assuming $\kappa_1/\kappa_2 = 2 \times 10^{-5}$, we show that TOF magic states can be prepared with total logical Z failure rates as low as 6×10^{-6}, which is several orders of magnitude lower than what could be achieved using non-fault-tolerant methods to prepare TOF states. Furthermore, the noise on the prepared TOF state is dominated by one specific Pauli error, which is a feature we can further exploit.

In Section VII, we show how TOF magic states probabilistically prepared using our bottom-up approach can be injected in a new magic state distillation scheme. This protocol distills 2 higher fidelity TOF states from 8 lower fidelity TOF states with high success probability. For generic noise, the protocol achieves quadratic error reduction. In the relevant case where a single Pauli error dominates, we can achieve cubic error reduction. The protocol is compiled down to architecture-level lattice surgery operations performed at the encoded level using repetition and surface codes. As such, we refer to such an approach as being top-down. Our top-down approach allows us to distill TOF magic states with low enough logical error rates for use in quantum algorithms of practical interest. Further, we note that given the low error rates achieved using our bottom-up approach, only one round of distillation is required in our top-down approach to prepare TOF states with the desired logical error rates.

Finally, in Section VIII we analyze the overhead required for running quantum algorithms in our architecture. We find that running quantum circuits on 100 qubits with Clifford gates and up to 1,000 Toffoli gates would require 1,000-2,000 ATSs. Such circuits are comfortably beyond the reach of classical simulability. The estimate is based on achieving a ratio of $\kappa_1/\kappa_2 = 10^{-5}$, corresponding to a CNOT error probability of 0.3%. It also assumes that the cat code can correct all X errors, meaning one can concatenate it with a repetition code. The number of hardware components required is compatible with next generation cryogenic dilution refrigerators, making our proposal a promising route for early implementations of fault-tolerant quantum computation.

We also consider how our architecture performs for the task of estimating the ground state energy density of the Hubbard model. For a parameter regime that is very challenging for classical computers, we estimate that our architecture could be implemented using 32,000 ATS components and executed in 49 minutes per run. Our analysis assumes an overall CNOT error probability of 0.3%; with better gates, further reductions in resources would be possible. In this analysis we do not assume the cat code can correct all X errors and use a thin strip surface code as the outer code.

Notably, for this problem the magic-state factory uses at most 7.7% of the total resources and is never a bottleneck on algorithm execution time. This low factory overhead is due to a combination of factors. First, the bottom-up procedure gives initial TOF states with a cost that is not much more than a physical TOF but orders of magnitude lower error rates. Second, at the required TOF error rate it suffices to implement the top-down protocol using a mixture of repetition codes and surface codes, which dramatically reduces the factory footprint. In contrast, the best performing T state factories (in architectures without biased noise) rely completely on surface codes and either require multiple rounds of distillation to achieve the same error suppression [30–32] or only produce 1 T state at a time so that 8 rounds are needed to realize 2 TOF gates [33].

II. HARDWARE IMPLEMENTATION AND STABILIZATION SCHEMES

In our proposal, the lowest-level protection from errors occurs directly at the hardware level and is based on the idea of autonomous quantum error correction [34], where rather than correcting errors at the “software level”, one instead engineers a system whose unitary evolution and dissipation is sufficient to protect the encoded information from Markovian errors. One can think of this process as the continuous analog of the standard, discrete QEC cycle consisting of syndrome measurements and correcting unitaries. The value of AQEC is that it eliminates the need for active measurements and classical feedback.

Historically, proposals for the implementation of autonomous QEC have been formulated in the language of coherent feedback control [35] or reservoir engineering [36], where the evolution is described via a stochastic master equation or a Lindblad master equation, respectively. Here we specifically adopt a bosonic autonomous QEC technique that more neatly fits into the latter category. It was first introduced by Mirrahimi *et al.* in 2014 [15] and demonstrated for individual qubits in recent experiments [21, 25]. We summarize the most relevant pieces here for convenience.

A. Overview of cat codes and driven-dissipative stabilization

The basic idea is to encode a qubit in a two-dimensional subspace $S = \text{span}\{|-\alpha\rangle, |+\alpha\rangle\}$ of a harmonic oscillator, spanned by the two quasi-orthogonal coherent states $|\pm\alpha\rangle$ [37, 38]. The qubit states can be defined in the X basis as the following two-component Schrödinger cat states:

$$|\pm\rangle = \mathcal{N}_\pm(|\alpha\rangle \pm |-\alpha\rangle). \qquad (1)$$

These states are eigenstates of the parity operator $\hat{P} = \exp(i\pi\hat{a}^\dagger\hat{a})$ with eigenvalues ±1, and $\mathcal{N}_\pm =$

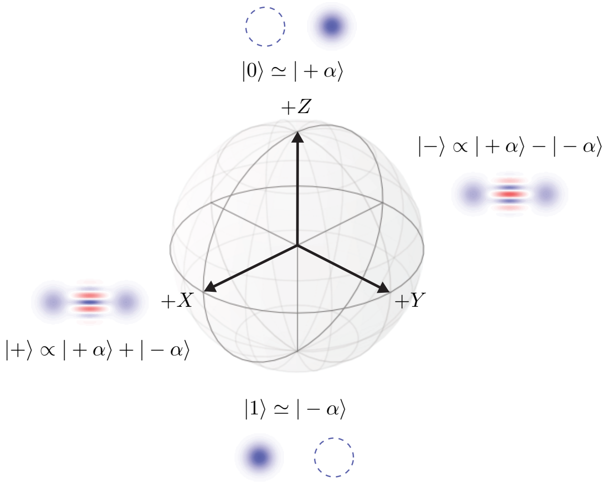

$|0\rangle \simeq |+\alpha\rangle$
$+Z$
$|-\rangle \propto |+\alpha\rangle - |-\alpha\rangle$
$+X$
$+Y$
$|+\rangle \propto |+\alpha\rangle + |-\alpha\rangle$
$|1\rangle \simeq |-\alpha\rangle$

FIG. 1. Bloch sphere of the cat qubit. The codewords $|0\rangle$, $|1\rangle$ and the $|\pm\rangle$ states are indicated on the Z and X axes, respectively, along with their Wigner function representations (shown for $\alpha = 2$).

$1/\sqrt{2(1 \pm e^{-|2\alpha|^2})}$. The codewords of this code are

$$|0\rangle = |+\alpha\rangle + \mathcal{O}(e^{-2|\alpha|^2})|-\alpha\rangle \tag{2}$$

$$|1\rangle = |-\alpha\rangle + \mathcal{O}(e^{-2|\alpha|^2})|+\alpha\rangle. \tag{3}$$

Note that $|0\rangle \approx |+\alpha\rangle$ and $|1\rangle \approx |-\alpha\rangle$ is a very good approximation for $|\alpha|^2 \gg 1$, as will typically be assumed throughout this paper.

The usual error channels that affect real oscillators, such as energy relaxation and dephasing, will eventually corrupt the information encoded in this manner. To protect against these common errors, one can engineer an artificial coupling to a bath such that the oscillator only emits and absorbs excitations to and from this bath *in pairs*. Such dynamics can be modeled by a Lindblad master equation of the form

$$\frac{d\hat\rho(t)}{dt} = \kappa_2\mathcal{D}[\hat a^2 - \alpha^2]\hat\rho(t) + \kappa_1\mathcal{D}[\hat a]\hat\rho(t) + \kappa_\phi\mathcal{D}[\hat a^\dagger \hat a]\hat\rho(t) \tag{4}$$

where $\mathcal{D}[\hat L]\hat\rho := \hat L\hat\rho\hat L^\dagger - \frac{1}{2}(\hat L^\dagger\hat L\hat\rho + \hat\rho\hat L^\dagger\hat L)$, κ_1 is the usual single-photon (or phonon) dissipation rate, κ_ϕ is the pure dephasing rate, and κ_2 is a two-photon (or two-phonon) dissipation rate. In the case where $\kappa_1 = \kappa_\phi = 0$, any linear combination of the codewords $|0\rangle$, $|1\rangle$ is a steady state of Eq. (4). This is straightforward to see, as any state for which $\hat a^2|\psi\rangle = \alpha^2|\psi\rangle$ is stationary under this master equation, and this includes both the even- and odd-parity cats. Any arbitrary pure initial state eventually evolves to a state in the S manifold; in particular, states that are near the fixed points $|\pm\alpha\rangle$ exponentially decay back into the code space at a rate $\kappa_{\text{conf}} = 2|\alpha|^2\kappa_2$, which is typically called the confinement rate. For finite κ_1, κ_ϕ, this description of the dynamics no longer holds true exactly. In particular, the stationary solutions of Eq. (4) are no longer pure states. However, if $\kappa_{\text{conf}} > \kappa_{\text{err}}$, where κ_{err} is the effective error rate then the codewords are still metastable states.

The key feature of this code is that, above the threshold $\kappa_{\text{conf}} > \kappa_{\text{err}}$, the bit-flip rate (or rate of X-type errors) $\Gamma_{0\leftrightarrow 1}$ decays exponentially with the "code distance" $|\alpha|^2$ as

$$\Gamma_{0\leftrightarrow 1} \sim |\alpha|^2 e^{-2|\alpha|^2}\kappa_{\text{err}}, \tag{5}$$

whereas the phase-flip rate (or rate of Z-type errors) $\Gamma_{+\leftrightarrow -}$ increases linearly as

$$\Gamma_{+\leftrightarrow -} \sim |\alpha|^2\kappa_{\text{err}}. \tag{6}$$

For sufficiently small values of the dimensionless loss parameter $\kappa_{\text{err}}/\kappa_2$, and sufficiently large $|\alpha|^2$, this translates to a large noise bias, i.e. a large discrepancy between the X and Z error rates. As alluded to earlier, this bias is a key feature of our proposal and will be exploited to our advantage to achieve fault tolerance with lower overheads.

The driven-dissipative dynamics of Eq. (4) can be physically realized by using a cleverly designed nonlinear element to couple the storage mode $\hat a$ to an engineered environment, or reservoir. Following Refs. [21, 25], the idea is to generate a nonlinear interaction of the form $g_2^*\hat a^2\hat b^\dagger + \text{h.c.}$ between the storage mode and an ancillary mode $\hat b$, which here we refer to as the "buffer mode" in keeping with existing terminology. The buffer mode is in turn strongly coupled to a bath — it is designed to have a large energy relaxation rate κ_b so that it rapidly and irreversibly buffers the photons it contains into the environment. If $\kappa_b \gg g_2$, the $\hat b$ mode is in the vacuum state $|\hat b^\dagger\hat b = 0\rangle$ most of the time, and its excited states can be adiabatically eliminated from the Hamiltonian [21, 39]. In this picture, there exists an effective Markovian description of the $\hat a$ mode dynamics where the $\hat b$ mode has been traced out into the environment, and where the emission of excitations via $g_2^*\hat a^2\hat b^\dagger$ can be accurately modeled as a dissipative process acting on the $\hat a$ mode alone. Usually the bath is not thermally populated at the energy scales of interest, so to stimulate the absorption process $g_2\hat a^{\dagger 2}\hat b$ a linear drive $\epsilon_d^*\hat b e^{-i\omega_d t} + \text{h.c.}$ on the buffer mode is added to supply the required energy. With this drive tuned perfectly on resonance ($\omega_d = \omega_b$), the evolution of the combined system is described by

$$\frac{d\hat\rho(t)}{dt} = -i[g_2^*(\hat a^2 - \alpha^2)\hat b^\dagger + \text{h.c.}, \hat\rho(t)] + \kappa_b\mathcal{D}[\hat b]\hat\rho(t) + \kappa_1\mathcal{D}[\hat a]\hat\rho(t), \tag{7}$$

where $\alpha^2 := -\epsilon_d/g_2^*$. After adiabatically eliminating the $\hat b$ mode, this master equation becomes Eq. (4), with $\kappa_2 = 4|g_2|^2/\kappa_b$.

B. Physical implementation of buffer and storage resonators

To realize the dynamics described by Eq. (7) in practice, previous experiments have relied on Josephson junctions — either in the form of a transmon qubit [21] or an "asymmetrically-threaded SQUID" [25] — as the source of nonlinearity. Other variations of the nonlinear elements exist, for instance the "SNAIL" [24, 40], but here we focus the discussion on junctions and the ATS. The potential energy of an ordinary junction has the form $\cos(\hat{\phi})$, where $\hat{\phi}$ is the superconducting phase difference across it, whereas the energy of an ATS has the form $\sin(\hat{\phi})$. Either of these two forms is nonlinear in $\hat{\phi}$, and because $\hat{\phi} = \varphi_a \hat{a} + \varphi_b \hat{b} + \text{h.c.}$, all terms of higher than quadratic order in $\hat{\phi}$ generate nonlinear couplings between the modes, provided that the required energy is injected with pumps tuned to the appropriate frequencies. It is important to emphasize that the operators $\hat{a}$, $\hat{b}$ in this sum are the *normal modes* of the combined storage and buffer resonators. Because these resonators are typically far-detuned there is little mixing between them, so $\hat{a}$ is "storage-like" and $\hat{b}$ is "buffer-like". The vacuum fluctuation amplitudes φ_a, φ_b quantify the contribution of these normal modes to the total phase difference $\hat{\phi}$ seen by the nonlinear circuit element.

The desired interaction $g_2^* \hat{a}^2 \hat{b}^\dagger + \text{h.c.}$ can be resonantly activated by pumping the system with a tone at frequency $\omega_p = 2\omega_a - \omega_b$. This pump provides the missing energy in the conversion process — crudely speaking a pump photon combines with a buffer photon to create two storage phonons, and vice versa. The specific way that this energy is injected depends on the hardware, but this interaction is possible with both the single-junction and ATS implementations. However, a single junction (or single SQUID) — despite being a simpler device — has a key disadvantage: the $\cos(\hat{\phi})$ potential only has even powers in its series expansion, and so the junction also generates undesired cross-Kerr couplings such as $\hat{a}^\dagger \hat{a} \hat{b}^\dagger \hat{b}$ in addition to the desired interaction. These couplings produce frequency shifts in the storage mode which depend on the number of photons $\hat{b}^\dagger \hat{b}$ in the buffer, and so any fluctuations in the latter will dephase the encoded cat qubit. Indeed, this has been observed to be a limiting factor in previous demonstrations [21, 25]. This problem can be circumvented by instead relying on the $\sin(\hat{\phi})$ potential of the ATS. For these reasons we adopt the ATS as the main nonlinear element in our proposal. For further details on this device, the pump implementation, and the calculation of g_2, see Appendix A and Ref. [25].

For the storage oscillator, the two cited works have used either superconducting 3D microwave cavities [21] or on-chip coplanar-waveguide (CPW) resonators [25], and recent theoretical proposals have focused on similar implementations [13, 14]. Here we study the possibility of using nanomechanical resonators instead, and tailor our calculations specifically to the case of one-dimensional phononic-crystal-defect resonators (PCDRs) made of lithium niobate, a strongly piezoelectric single-crystal material. These devices support resonances at gigahertz frequencies, with mode shapes localized inside a volume $< 1\,\mu\text{m}^3$ of a suspended nanostructure. They have been coupled to transmon qubits in recent experiments [28, 41] and may offer a number of advantages over electromagnetic resonators.

First, a PCDR is a micron-scale nanostructured device, with an on-chip footprint (area) that is at least three orders of magnitude smaller than that of planar superconducting resonators, including lumped-element structures. This is not a significant advantage today, with the largest quantum computers only having a few dozen physical qubits, but it may become important in the future.

A second consideration is that, unlike electromagnetic resonators, appropriately designed acoustic devices do not experience direct crosstalk (unwanted couplings) because acoustic waves do not propagate through vacuum. They can still couple through the circuitry that mediates interactions between them, but this can be mitigated with approaches such as filtering and a carefully chosen connectivity, both of which are important features of our proposal.

The third and most important consideration is that there is recent experimental evidence that phononic-crystal-based devices can have very long coherence times as a result of the high degree of confinement of their modes and the quality of their materials. For example, similar designs fabricated from single-crystal silicon and operating at a frequency of 5 GHz have been shown to have energy relaxation and pure dephasing times of $T_1 \approx 1.5\,\text{s}$ and $T_\phi \approx 130\,\mu\text{s}$, respectively [27]. These devices cannot be easily coupled to superconducting circuits, but they offer insight into the decoherence mechanisms affecting nanomechanical resonators and suggest a roadmap for achieving similar levels of coherence with piezoelectric devices. For example, similar studies with lithium niobate PCDRs are already under way [29], and although their coherence times are currently limited to $\sim 1\,\mu\text{s}$, there is no reason to believe that these numbers could not approach those of the silicon devices after sufficient advances in materials and surface science.

C. Wiring and layout

We now describe a way to combine all of these building blocks to build a two-dimensional grid of cat qubits that form the basis for an outer code, such as the repetition code or the surface code. First, following Ref. [25] we form a buffer resonator with frequency ω_b by shunting an ATS with a capacitor. A bandpass filter with bandwidth $4J$ centered at ω_b is then connected to the output port of the buffer, and an open $50\,\Omega$ waveguide (which can be accurately modeled as a $50\,\Omega$ resistive termination) is connected to the output of the filter. This filter configuration stands in contrast to the implementation in Ref. [25],

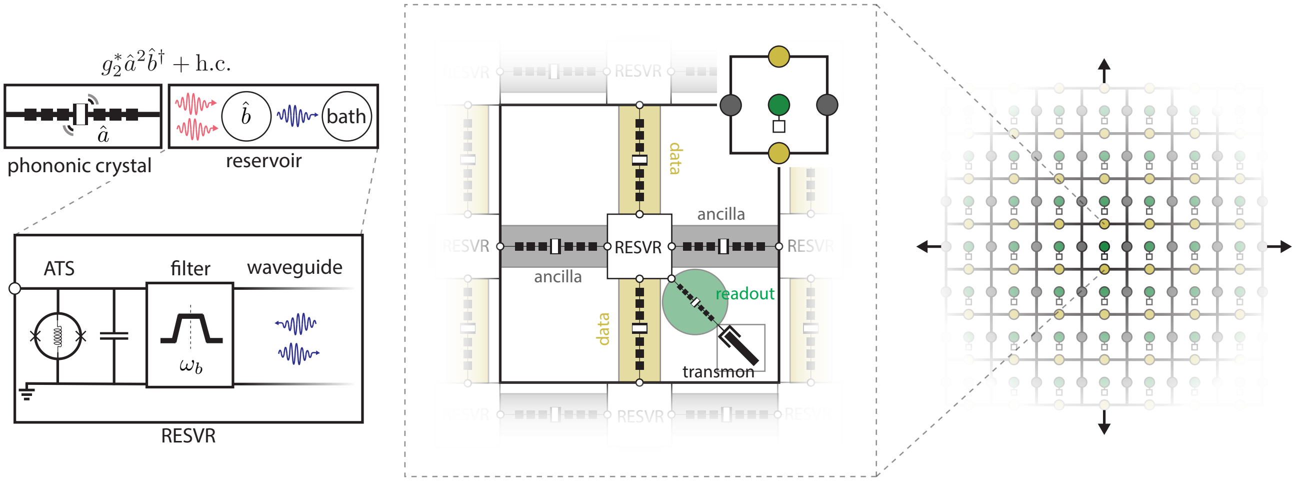


FIG. 2. Hardware implementation of the repetition- and surface-cat codes. In the 2D grid on the right, yellow circles represent data qubits where the logical information is encoded, and gray circles represent ancilla qubits which are used to measure the stabilizers and extract error syndromes. Both data and ancilla qubits are encoded as Schrödinger cat states of localized acoustic modes of phononic-crystal-defect resonators (PCDRs), and are stabilized through a driven-dissipative two-phonon interaction with an engineered reservoir. This stabilization strongly biases the noise, suppressing X errors and increasing Z errors. The white square in each plaquette represents the reservoir, which is implemented with a capacitively-shunted ATS (the "buffer" resonator), a bandpass filter, and an open $50\,\Omega$ waveguide. This circuit is shown inside the "RESVR" box in the left panel and has a single non-grounded terminal, marked with a white circle on the edge of the box. All circles surrounding each RESVR box in the layout diagram in the center panel represent this one physical terminal. The green circle in the center of each plaquette represents an additional acoustic mode used to measure the cat qubits in the X basis with the aid of a transmon, which is represented by a white square. Altogether, five PCDRs are connected to each reservoir: four as active qubits, and one for readout.

where a bandstop filter was used to protect the storage mode from radiatively decaying into the waveguide. In our proposal the bandpass filter also serves this role, but it also plays a more fundamental role as a means of suppressing crosstalk mechanisms that arise as a result of our frequency-multiplexed scheme to stabilize (and perform gates between) multiple modes with a single ATS. From this point on, we refer to the combination of the buffer, filter, and $50\,\Omega$ environment as the "reservoir".

We arrange multiple reservoirs in a two-dimensional grid, as shown in Fig. 2, and connect neighboring reservoirs with a PCDR using each of the two terminals of the resonator. Four of these resonators provide the connectivity between reservoirs and are located above, below, to the left, and to the right of each reservoir. These four resonators serve as data and ancilla qubits in either the repetition or the surface code. In addition, one more resonator coupled to each reservoir serves the purpose of an ancillary readout mode which is used to measure the cat qubits in the X basis with the aid of a ordinary transmon. Alternatively, it is possible to omit this resonator altogether and perform the X readout directly via the buffer — see Section III G for further details.

There are two important considerations that motivate this architecture. The first is that present PCDR designs only have two available terminals, so each of them can be connected to at most two different reservoir circuits. This is simply a design choice — it may be possible to add more terminals without a significant degradation of performance, and this would enable other variations of the 2D layout. The second consideration comes from our analysis of correlated errors in the frequency-multiplexed stabilization scheme, which we overview below and provide details of in Appendix B. Our results show that the correlated error rates scale rapidly with the number of modes connected to a ATS, and the error rates that come with choosing five modes per ATS are the largest that can be tolerated by the outer error-correcting codes.

D. Calculation of the loss parameter κ_1/κ_2

The dimensionless loss κ_1/κ_2 is a crucial parameter: it sets the error rates of the gates, as well as the error rates during idling, state preparation, and measurement. It is therefore important to calculate accurate predictions for this parameter (under different assumptions for the intrinsic losses), especially when considering acoustic devices as the storage elements, for which the relevant stabilization schemes have not yet been demonstrated. We remark that the pure dephasing rate κ_ϕ is also important, especially in a regime where $\kappa_\phi \gg \kappa_1$. Here we focus on calculating the dimensionless loss κ_1/κ_2, keeping in mind that this sets the scale for most relevant error rates while the specific value of κ_ϕ/κ_1 sets the prefactors. For further details see Table I and Table II.

The details of this calculation are presented in Appendix A. The key result is that the two-phonon dissipation rate κ_2 scales linearly with the filter bandwidth $4J$ and inversely with the mean phonon number $|\alpha|^2$ (the "distance" of the cat code):

$$\kappa_2 \approx 4\eta^2 J/|\alpha|^2. \tag{8}$$

Here $\eta \approx 1/5$ is a small parameter relating to the adiabaticity constraints. The single-phonon dissipation rate $\kappa_1 = \kappa_{1,\text{int}} + \kappa_{1,\text{pur}}$, on the other hand, contains two main contributions: the intrinsic decay rate $\kappa_{1,\text{int}}$ of the *bare* storage mode (for example due to two-level systems and quasiparticles) and the Purcell decay rate $\kappa_{1,\text{pur}}$ (due to the mixing of the bare buffer and storage modes). Of these two contributions, the former is largely an empirical quantity that depends on the quality of the materials and the fabrication process, whereas the latter can be calculated and mitigated, because it depends on the way the storage and buffer resonators are coupled. In particular, $\kappa_{1,\text{pur}}$ contains a contribution from direct radiative decay into the reservoir (which is negligibly small because the storage mode frequency ω_a is far outside of the filter passband) and a contribution $\sim (g/\delta)^2\kappa_{b,\text{int}}$ coming from the intrinsic decay of the bare buffer resonator, which the filter cannot protect against. Here g is the *linear* coupling rate between the storage and buffer, $\delta = \omega_b - \omega_a$ is the detuning, and $\kappa_{b,\text{int}}$ is the intrinsic decay rate of the bare buffer resonator. This contribution is important when $\kappa_{1,\text{int}}$ is orders of magnitude smaller than $\kappa_{b,\text{int}}$. Note also that only $\kappa_{b,\text{int}}$ enters this expression as opposed to $\kappa_b (\gg \kappa_{b,\text{int}})$, because the filter prevents the $\hat{a}$ mode from directly emitting photons into the waveguide.

A second key result of our analysis is that $\kappa_{1,\text{pur}}$ can be strongly suppressed by using a buffer resonator with a large characteristic impedance Z_b. The idea is to detune the buffer frequency far away from the storage frequencies until $\kappa_{1,\text{pur}} \sim (g/\delta)^2\kappa_{b,\text{int}}$ is suppressed to a value comparable to or smaller than $\kappa_{1,\text{int}}$. This comes at the cost of reducing the nonlinear interaction rate g_2, which also scales with the detuning as $g_2 \sim 1/\delta^2$. But one can offset this penalty by increasing Z_b, because $g_2 \sim Z_b^{5/2}$. We show that under certain assumptions of $\kappa_{b,\text{int}}$ and $\kappa_{1,\text{int}}$, there is a range of experimentally-feasible impedance values (on the order of a few kΩ) for which one can access a regime where

$$\kappa_1/\kappa_2 \approx \kappa_{1,\text{int}}|\alpha|^2/4\eta^2 J. \tag{9}$$

This is a useful result, as it addresses the problems that arise when coupling a highly coherent, linear storage element to a much lossier superconducting circuit. We show in Fig. 3 a plot of this simple expression for κ_1/κ_2 as a function of $\kappa_{1,\text{int}}$ and for different values of the filter bandwidth $4J$, marking the largest bandwidth $4J/2\pi \approx 100\,\text{MHz}$ that is allowable by our analysis in the case where 5 modes (2 data, 2 ancillas, 1 readout) are coupled to each reservoir. We also assume $|\alpha|^2 = 8$, which is large enough to result in good performance of the outer codes. With this bandwidth the intrinsic loss of the storage resonators must be $\kappa_{1,\text{int}}/2\pi \approx 5\,\text{Hz}$ — equivalent to an energy relaxation time $T_1 \approx 32\,\text{ms}$ — in order to reach the regime $\kappa_1/\kappa_2 = 10^{-5}$ required to run useful algorithms with a competitive resource overhead, as we show in Section VIII. This is a challenging target, but as discussed in Section II B, it is certainly possible that this level of coherence will become accessible in the not-so-distant future. We remark that there are ways to increase this bandwidth. For example, by performing X readout with the buffer itself, only 4 modes have to be coupled to each reservoir, and this allows us to increase the bandwidth to $4J/2\pi \approx 180\,\text{MHz}$, lowering the T_1 requirement by roughly a factor of 2. Future approaches may further reduce the number of modes to 2, by increasing the number of terminals of each PCDR from 2 to 4, and this would allow a more drastic increase in $4J$. Finally, we remark that while we have focused this analysis on the single-phonon loss rate κ_1, the pure dephasing rate κ_ϕ is also relevant as discussed throughout this proposal.

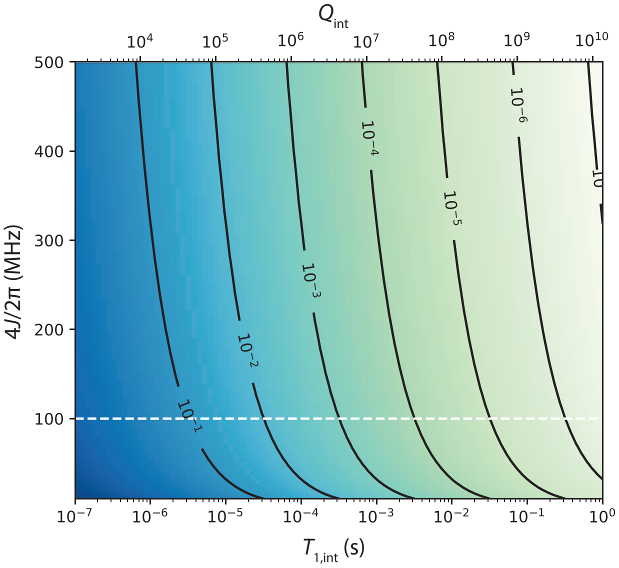


FIG. 3. Dimensionless loss κ_1/κ_2, as given by Eq. (9), as a function of filter bandwidth $4J$ and the intrinsic energy relaxation time $T_{1,\text{int}} = 1/\kappa_{1,\text{int}}$, assuming fixed $\omega_a/2\pi = 2.16\,\text{GHz}$ and $|\alpha|^2 = 8$. We label the corresponding quality factor $Q_\text{int} = \omega_a/\kappa_{1,\text{int}}$ on the upper horizontal axis, and mark the bandwidth value used in our proposal with the white dashed line. At this bandwidth, the relaxation times required to reach $\kappa_1/\kappa_2 = 10^{-4}$ and 10^{-5} are $T_{1,\text{int}} \approx 3\,\text{ms}$ and $32\,\text{ms}$, respectively. Future innovations in the multiplexed stabilization scheme may allow for larger bandwidths, which would relax these requirements proportionally.

It is important to note that the value $\kappa_2 \approx 4\eta^2 J/|\alpha|^2 \sim 2\pi \times 500\,\text{kHz}$ that we derive in this analysis, while theoretically possible, would require substantially (about

30 times) larger values of g_2 than those previously reported [25]. Because $\alpha^2 = -\epsilon_d/g_2^*$ (see Section II A), this would require a larger drive amplitude on the buffer mode in order to maintain a fixed α, which may cause unforeseen problems such as instabilities [42] or the excitation of spurious transitions [43, 44]. Increasing the power-handling capacity of nonlinear circuits such as the ATS is an area of active research, with promising advances such as the use of inductive shunts to suppress instabilities [45].

E. Multiplexed stabilization

In our architecture, each reservoir is responsible for stabilizing multiple storage modes simultaneously. This multimode stabilization can be implemented via a simple extension of the single-mode stabilization scheme demonstrated in Ref. [25]. The main idea is to use *frequency-division multiplexing* to stabilize different modes independently. Here, multiplexing refers to the fact that different regions of the filter passband are allocated to the stabilization of different modes. When the bandwidth allocated to each stabilization process is sufficiently large, multiple modes can be stabilized simultaneously and independently, as we now show.

To stabilize the n-th mode coupled to a given reservoir, we apply a pump frequency $\omega_p^{(n)} = 2\omega_a - \omega_b + \Delta_n$, and drive the buffer mode at frequency $\omega_d^{(n)} = \omega_b - \Delta_n$, where Δ_n denotes a detuning. Analogously to the single-mode stabilization case, due to the nonlinear mixing of the ATS these pumps and drives give rise to an interaction Hamiltonian of the form

$$\hat{H}/\hbar = \sum_n g_2 \left(\hat{a}_n^2 - \alpha^2\right) \hat{b}^\dagger e^{i\Delta_n t} + \text{H.c.} \tag{10}$$

See Appendix B for a derivation of Equation (10) as well as Equations (11) and (12) below. Note that the sum does not run over all modes coupled to the ATS, but rather only over the modes stabilized by that ATS. In our architecture, though five modes couple to each ATS, only two must be stabilized simultaneously, so the sum contains only two terms. By adiabatically eliminating the lossy buffer mode, one obtains an effective master equation describing the evolution of the storage modes

$$\frac{d\hat{\rho}}{dt} \approx \mathcal{D}\left[\sum_n \sqrt{\kappa_{2,n}}(\hat{a}_n^2 - \alpha^2)e^{i\Delta_n t}\right]\hat{\rho}(t), \tag{11}$$

where $\kappa_{2,n} \approx 4|g_2|^2/\kappa_b$ if the corresponding detuning falls inside the filter passband ($|\Delta_n| < 2J$), and $\kappa_{2,n} \approx 0$ otherwise, see Appendix A. If the detunings are chosen such that $|\Delta_n - \Delta_m| \gg 4|\alpha|^2\kappa_2$ for all $m \neq n$, then Equation (11) can be approximated by

$$\frac{d\hat{\rho}}{dt} \approx \sum_n \kappa_{2,n}\mathcal{D}\left[\hat{a}_n^2 - \alpha^2\right]\hat{\rho}(t), \tag{12}$$

which is obtained by neglecting the fast-rotating terms in (11) via a rotating-wave approximation. The dynamics (12) stabilize cat states in different modes independently and simultaneously. Thus, by simply applying additional pumps and drives with appropriately chosen detunings, multiple modes can be simultaneously stabilized by a single ATS.

The efficacy of this multiplexed stabilization scheme can be understood intuitively by considering the frequencies of photons that leak from the buffer mode to the filtered bath. In the case of $\Delta_n = 0$, a pump applied at frequency $2\omega_a - \omega_b$ facilitates the conversion of two phonons of frequency ω_a to a single photon of frequency ω_b. As a result, photons that leak from the buffer to the bath have frequency ω_b. If instead the pump is detuned by an amount $\Delta_n \neq 0$, it follows from energy conservation that the corresponding emitted photons have frequency $\omega_b + \Delta_n$. When the differences in these emitted photon frequencies, $\Delta_n - \Delta_m$, are chosen to be much larger than the emitted photon linewidths, $4|\alpha|^2\kappa_2$ (see Appendix C), emitted photons associated with different storage modes are spectrally resolvable by the environment. Therefore, when the stabilization of mode n causes a photon to leak to the environment, there is no back-action on modes $m \neq n$. These ideas are illustrated pictorially in Figure 4(a). The figure emphasizes an important additional point: the emitted photon frequencies must lie inside the filter bandwidth, lest the engineered dissipation be suppressed by the filter.

F. Crosstalk

In acting as a nonlinear mixing element, the ATS not only mediates the desired ($g_2\hat{a}_n^2\hat{b}^\dagger + \text{H.c.}$) interactions, but it also mediates spurious interactions between different storage modes. We now describe how such interactions can give rise to crosstalk among the cat qubits, and subsequently how this crosstalk can be mitigated through a combination of filtering and phonon-mode frequency optimization.

While most spurious interactions mediated by the ATS are far detuned and can be safely neglected in the rotating-wave approximation, there are others which cannot be neglected. Most concerning among these are interactions of the form

$$g_2\hat{a}_j\hat{a}_k\hat{b}^\dagger e^{i\delta_{ijk}t} + \text{H.c.}, \tag{13}$$

for $j \neq k$, where $\delta_{ijk} = \omega_p^{(i)} - \omega_j - \omega_k + \omega_b$. This interaction converts two phonons from different modes, j and k, into a single buffer mode photon, facilitated by the pump that stabilizes mode i. These interactions cannot be neglected in general because they have the same coupling strength as the desired interactions (10), and they can potentially be resonant or near-resonant, depending on the frequencies of the phonon modes involved.

There are three different mechanisms through which

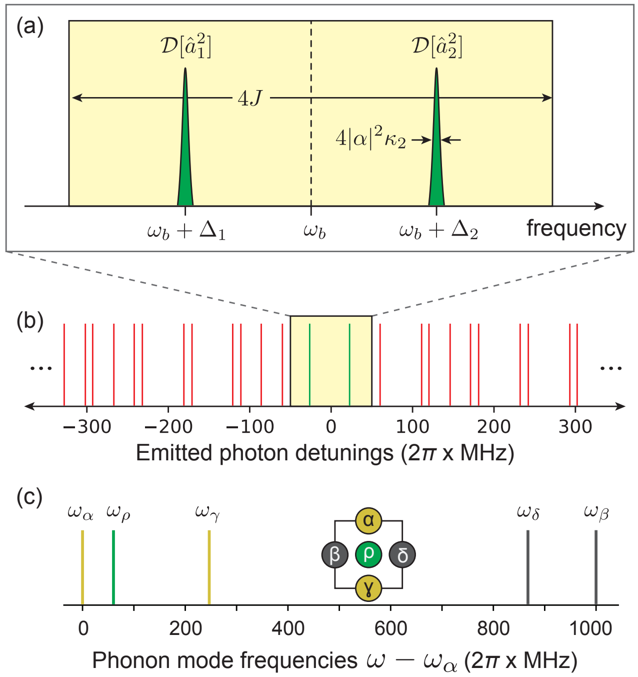


FIG. 4. Multiplexed stabilization and crosstalk mitigation. (a) Frequency multiplexing. Because the desired couplings ($g_2\hat{a}_n^2\hat{b}^\dagger e^{i\Delta_i t}$ + H.c.) are detuned by different amounts, photons lost to the environment via the buffer have different frequencies. When the corresponding emitted photons (green lines) are spectrally well resolved, $|\Delta_n-\Delta_m|\gg 4|\alpha|^2\kappa_2$, the modes are stabilized independently. Dissipation associated with photon emissions at frequencies inside the filter passband (yellow box) is strong, while dissipation associated with emission at frequencies outside the passband is suppressed. (b),(c) Crosstalk suppression. Red lines in (b) denote photon emission frequencies associated with various correlated errors, calculated for the specific phonon mode frequencies plotted in (c). The mode frequencies are deliberately chosen so that *all* emissions associated with correlated errors occur at frequencies outside the filter passband (no red lines fall in the yellow box). In other words, Equations (18) and (19) are simultaneously satisfied for any choices of the indices that lead to nontrivial errors in the cat qubits. See Appendix B for further details.

the interactions (13) can induce crosstalk among the cat qubits. These mechanisms are described in detail in Appendix B, and we summarize them here. First, analogously to how the desired interactions (10) lead to two-phonon losses, the undesired interactions (13) lead to correlated, single-phonon losses

$$\kappa_{\text{eff}}\mathcal{D}[\hat{a}_j\hat{a}_k]\to\kappa_{\text{eff}}|\alpha|^4 D[\hat{Z}_j\hat{Z}_k] \tag{14}$$

where the rate κ_{eff} will be discussed shortly. The arrow denotes projection onto the code space, illustrating that these correlated losses manifest as *stochastic*, correlated phase errors in the cat qubits.

Second, the interplay between different interactions of the form (13) gives rise to new effective dynamics [39, 46, 47] generated by Hamiltonians of the form

$$\hat{H}_{\text{eff}} =\chi\hat{a}_i^\dagger\hat{a}_j^\dagger\hat{a}_m\hat{a}_n e^{i(\delta_{\ell mn}-\delta_{ijk})t}+\text{H.c.}, \tag{15}$$

$$\to\chi|\alpha|^4\hat{Z}_i\hat{Z}_j\hat{Z}_k\hat{Z}_l e^{i(\delta_{\ell mn}-\delta_{ijk})t}+\text{H.c.}, \tag{16}$$

where the coupling rate χ is defined in Appendix B. The projection onto the code space in the second line reveals that $\hat{H}_{\text{eff}}$ can induce undesired, *coherent* evolution within the code space.

Third, $\hat{H}_{\text{eff}}$ can also evolve the system out of the code space, changing the phonon-number parity of one or more modes in the process. Though the engineered dissipation subsequently returns the system to the code space, it does not correct changes to the phonon-number parity. The net result is that $\hat{H}_{\text{eff}}$ also induces *stochastic*, correlated phase errors in the cat qubits,

$$\gamma_{\text{eff}}\mathcal{D}[\hat{Z}_i\hat{Z}_j\hat{Z}_k\hat{Z}_\ell], \tag{17}$$

where the rate γ_{eff} will be discussed shortly.

Remarkably, all of these types of crosstalk can be suppressed through a combination of filtering and phonon-mode frequency optimization. In Appendix B, we show that both $\kappa_{\text{eff}}\approx 0$ and $\gamma_{\text{eff}}\approx 0$, provided

$$|\delta_{ijk}|>2J, \tag{18}$$

$$|\delta_{ijk}-\delta_{\ell mn}|>2J, \tag{19}$$

respectively. This suppression can be understood as follows. The decoherence associated with κ_{eff} and γ_{eff} results from the emission of photons at frequencies $\omega_b+\delta_{ijk}$ and $\omega_b\pm(\delta_{ijk}-\delta_{\ell mn})$, respectively. When the frequencies of these emitted photons lie outside the filter passband, their emission (and the associated decoherence) is suppressed. Crucially, we can arrange for all such errors to be suppressed *simultaneously* by carefully choosing the frequencies of the phonon modes, as shown in Figure 4(b,c). The configuration of mode frequencies in Figure 4(c) was found via a numerical optimization procedure described in Appendix B. The optimization also accounts for the undersired coherent evolution (16): the detunings $\delta_{ijk}-\delta_{\ell mn}$ are maximized so that $\hat{H}_{\text{eff}}$ is rapidly rotating and its damaging effects are mitigated (this suppression is quantified in Appendix B). Additionally, we note that in Figure 4(b) all emitted photon frequencies associated with crosstalk lie at least 10 MHz outside of the filter passband. As a result, the crosstalk suppression is robust to variations in the phonon mode frequencies of the same order. Larger variations in the phonon mode frequencies can be accommodated by reducing the filter bandwidth.

We have demonstrated that crosstalk can be largely suppressed within the five-mode unit cells of our architecture. It is tempting to consider whether more modes could be added to each unit cell to improve hardware efficiency or connectivity, but we find that crosstalk is a limiting factor in this regard. As more modes are added, the number of undesired terms (13) grows combinatorially, increasing the total number of constraints, Equations (18) and (19). At the same time, the filter bandwidth must be increased to accommodate the stabilization of additional modes, making each constraint more challenging to satisfy. Thus, it rapidly becomes difficult or impossible to satisfy all constraints, and crosstalk can become significant. We have accordingly chosen five modes per unit cell because this is the maximum number consistent with our 2D square grid layout for which all crosstalk constraints can be satisfied. While frequency crowding and bandwidth constraints are characteristic of multimode architectures generally [48–50], resonators with additional terminals, or tunable couplers [51, 52], could be employed in future designs to further suppress crosstalk and increase the number of modes per unit cell.

III. GATES AND MEASUREMENTS

In this section, we discuss the gates and measurements of the cat qubits. We first discuss the implementation of the X gate via a rotating two-phonon dissipation; this will be helpful for understanding the CNOT and Toffoli gates. We then review the fundamentals of the bias-preserving CNOT and Toffoli gates acting on cat qubits [13] and present several new analytical and numerical results. In particular, we explicitly characterize the extra geometric phase (Z or CZ rotations) which must be taken into account in the implementation of the CNOT and Toffoli gates if the average excitation number $|\alpha|^2$ is not an even integer. Moreover, we introduce the shifted Fock basis method and demonstrate that it is useful for the perturbative analysis of the Z error rates of various cat-qubit gates. We then illustrate that the shifted Fock basis method also allows more efficient numerical simulation of large cat qubits (up to $|\alpha|^2 = 10$) than the usual Fock basis method. The numerical results on gate error rates are summarized in Table II and detailed descriptions of the methods are given in Appendices C to E. We also discuss physical implementation of the gates in a system consisting of acoustic modes and ATSs.

We discuss two schemes for X-basis readout. In the first we achieve high fidelity readout with a small impact to the length of an error correction cycle. This is achieved using an additional readout mode interrogated by a transmon in parallel to the next error correction cycle. The second X-basis readout scheme does not use a transmon and instead uses deflation in conjunction with a coupling mediated by ATS to a buffer mode to achieve high fidelity readout. We also describe a fast Z-basis readout scheme which uses a coupling between the storage mode and buffer mediated by the ATS. This achieves exponentially improved error rates with $|\alpha|^2$. Combining these X-basis and Z-basis readout schemes allow for hardware efficient stabilizer measurement. More detailed analysis can be found in Appendix G.

A. X Gate

The X gate interchanges the cat-code computational basis states $|0\rangle$ and $|1\rangle$. For large values of α these cat-code states are approximately equal to the coherent states $|\alpha\rangle$ and $|-\alpha\rangle$, so the X gate acts by rotating the coherent states by π in the complex plane. This rotation can be realized by modulating the phase of the drive on the cavity so that the two-phonon dissipation rotates by π over a time T; then the code state evolves according to

$$\frac{d\hat{\rho}(t)}{dt} = \kappa_2 \mathcal{D}[\hat{a}^2 - \alpha^2 e^{2i\frac{\pi}{T}t}]\hat{\rho}(t). \tag{20}$$

This gives an adiabatic implementation of the X gate. Furthermore, we can apply a compensating Hamiltonian given by

$$\hat{H}_X = -\frac{\pi}{T}\hat{a}^\dagger\hat{a}, \tag{21}$$

so that code state rotates along with the fixed point of the dissipator. With this compensating Hamiltonian the gate need not be adiabatic and will succeed for any T. When the X gate is corrupted by phonon loss, gain, or by dephasing, the logical error rates during the X gate are identical to the noise during idle. This is because in the rotating frame of the compensating Hamiltonian $\hat{H}_X$, the noise and the dissipator are identical to the case of idle. The error rates for idle are summarized in Table I.

B. CNOT

We can realize the bias-preserving CNOT gate from [13] using an ATS coupled to a pair of acoustic modes. The CNOT gate rotates the cat-code states of the target cavity just as for the X gate, except that now the rotation is conditioned on the state of the control cavity. Cavity 1 will be the control and cavity 2 the target. A time dependent dissipator that realizes this rotation is given by the Lindblad jump operator

$$\hat{L}_2(t) = \hat{a}_2^2 - \alpha^2 + \frac{\alpha}{2}(e^{2i\frac{\pi}{T}t} - 1)(\hat{a}_1 - \alpha). \tag{22}$$

When cavity 1 is in the $|1\rangle$ cat-code state, which is approximately equal to the $|-\alpha\rangle$ coherent state, the corresponding dissipator reduces approximately to the rotating dissipator for the X gate on the second cavity. On the other hand when cavity 1 is in the $|0\rangle$ cat state, the operator L_2 reduces to the usual time-independent Lindblad operator. The control cavity is always stabilized by the usual time-independent Lindblad operator:

$$\hat{L}_1 = \hat{a}_1^2 - \alpha^2. \tag{23}$$

When a cat-code state $\hat{\rho}(t)$ evolves according to

$$\frac{d\hat{\rho}(t)}{dt} = \kappa_2\mathcal{D}[\hat{L}_1](\hat{\rho}) + \kappa_2\mathcal{D}[\hat{L}_2(t)]\hat{\rho}(t), \tag{24}$$

the encoded state undergoes a CNOT gate (up to an extra Z rotation on the control qubit; see below), assuming the gate time T is long compared to the stabilization rate $\kappa_2\alpha^2$. This gate preserves the bias in the noise because the two cat-code states remain distantly separated during the conditional rotation. Just as for the X gate the CNOT gate can be performed much faster with the help of a compensating Hamiltonian. In this case the compensating Hamiltonian has the form:

$$\hat{H}_{\text{CNOT}} = \frac{\pi}{4\alpha T}(\hat{a}_1 + \hat{a}_1^\dagger - 2\alpha)(\hat{a}_2^\dagger\hat{a}_2 - \alpha^2). \tag{25}$$

This Hamiltonian rotates the state of cavity 2 conditioned on the state of cavity 1, so that the two-cavity system remains in the subspace stabilized by the dissipator $\mathcal{D}[\hat{L}_1]$ and the rotating dissipator $\mathcal{D}[\hat{L}_2(t)]$.

The dissipators $\mathcal{D}[\hat{L}_1]$ and $\mathcal{D}[\hat{L}_2(t)]$ combined with the compensating Hamiltonian $\hat{H}_{\text{CNOT}}$ in Eq. (25) implement a gate

$$CX' \equiv \hat{Z}_1(-\pi\alpha^2)\cdot \text{CNOT}_{1\to 2}, \tag{26}$$

in the $T \gg 1/(\kappa_2\alpha^2)$ limit, which differs from the desired CNOT gate $\text{CNOT}_{1\to 2}$ by an extra Z rotation on the control qubit $\hat{Z}_1(-\pi\alpha^2)$ (see Appendix D for more details). Here, $\hat{Z}(\theta)$ is defined as $\hat{Z}(\theta) \equiv \exp[i\theta|1\rangle\langle 1|]$ and $|1\rangle$ is a computational basis state, the -1 eigenstate of the Pauli Z operator. The extra Z rotation is trivial if the average excitation number α^2 is an even integer. We also remark that the same extra Z rotation persists even if we use an ideal compensating Hamiltonian $-(\pi/T)|-\alpha\rangle\langle-\alpha|_1(\hat{a}_2^\dagger\hat{a}_2-\alpha^2)$. However, the extra Z rotation is not present if another variant of ideal compensating Hamiltonian $-(\pi/T)|-\alpha\rangle\langle-\alpha|_1\hat{a}_2^\dagger\hat{a}_2$ is used.

Note that the compensating Hamiltonian in Eq. (25) is only an approximation of an ideal compensating Hamiltonian, e.g., $-(\pi/T)|-\alpha\rangle\langle-\alpha|_1(\hat{a}_2^\dagger\hat{a}_2-\alpha^2)$. Hence, there is a residual non-adiabatic error that scales like $1/T$, where T is the gate time. Phonon loss, gain, and dephasing noise during the CNOT gate give rise to a Z error rate on both cavities that is proportional to T. The balance between the non-adiabatic errors and the noise gives rise to an optimal gate time that maximizes the fidelity.

In Ref. [13], it was noticed that the residual non-adiabatic error scales as $c/(\kappa_2\alpha^2 T)$ and found that the constant coefficient is given by $c \simeq 1/(2\pi)$ via a numerical fit. In Appendix D, we provide a first-principle perturbative analysis of the Z error rates of the CNOT gate by using the shifted Fock basis as a main tool. The key idea of the shifted Fock basis is to use the displaced Fock states $\hat{D}(\pm\alpha)|\hat{n}=n\rangle$ as the (unorthonormalized) basis states, where $\hat{n}=\hat{a}^\dagger\hat{a}$ is the mode occupation number. In particular, for the perturbative analysis of the Z error rates, it suffices to consider only the ground state manifold consisting of the coherent states $\hat{D}(\pm\alpha)|\hat{n}=0\rangle = |\pm\alpha\rangle$ and the first excited state manifold consisting of the displaced single-phonon Fock states $\hat{D}(\pm\alpha)|\hat{n}=1\rangle$. See Appendix C for a detailed description of the shifted Fock basis, including orthonormalization and matrix elements of the annihilation operator $\hat{a}$ in the shifted Fock basis. By taking the ground and the first excited state manifolds in the shifted Fock basis and using perturbation theory, we find that the Z error rates (per gate) of the implemented CX' gate are given by

$$\begin{aligned}\bar{p}_{Z_1} &= \kappa_1\alpha^2 T + \frac{\pi^2}{64\kappa_2\alpha^2 T},\\ \bar{p}_{Z_2} &= \bar{p}_{Z_1Z_2} = \frac{1}{2}\kappa_1\alpha^2 T.\end{aligned} \tag{27}$$

Here, κ_1 is the single-phonon loss rate (per time) and we assumed no dephasing and gain for the moment. We use $\bar{p}$ for error rates predicted by the perturbation theory and p for numerical results. Note that the coefficient $\pi^2/64 = 0.154$ in the non-adiabatic error term is close to the coefficient $1/(2\pi) = 0.159$ which was found earlier via a numerical fit [13]. Hence, the optimal gate time that minimizes the total gate infidelity is given by

$$\bar{T}^\star_{CX'} = \frac{\pi}{8\alpha^2\sqrt{2\kappa_1\kappa_2}}, \tag{28}$$

and at the optimal gate time, the Z error rates are given by

$$\bar{p}^\star_{Z_1} = 6\bar{p}^\star_{Z_2} = 6\bar{p}^\star_{Z_1Z_2} = \frac{3\pi}{8}\sqrt{\frac{\kappa_1}{2\kappa_2}} = 0.833\sqrt{\frac{\kappa_1}{\kappa_2}}. \tag{29}$$

These agree well with the numerical results (see Table II)

$$p^\star_{Z_1} = 6.067p^\star_{Z_2} = 6.067p^\star_{Z_1Z_2} = 0.91\sqrt{\frac{\kappa_1}{\kappa_2}}, \tag{30}$$

within a relative error of 10% (see Appendix D for the reasons for the discrepancy). Note that the perturbation theory predicts that the optimal Z error rates of the CX' gate (or the CNOT gate for even $|\alpha|^2$) are independent of the size of the cat code $|\alpha|^2$.

We simulated the CNOT gate using the effective dissipators and Hamiltonian acting on two cavities. Our method was to use the shifted Fock basis as described in Appendix C to find the optimal gate time and perform tomography at the optimal gate. This allowed us to compute all of the two-qubit Pauli error rates. The shifted Fock basis approach allowed us to compute the Z error rates with a small Hilbert space dimension which does not depend on α. In the standard Fock basis the required Hilbert space dimension increases rapidly with α. In contrast to the Z error rates, to accurately resolve the full set of Pauli error rates, a large dimension that increases with α is required even for the shifted Fock basis. However, even in this case the shifted Fock basis approach requires a smaller Hilbert space dimension and is still several times faster than the standard Fock basis. Our code was written in Python using the QuTIP package to solve the master equation including the disspators and Hamiltonian terms. We ran the simulations using AWS EC2 C5.18xlarge instances with 72 virtual CPUs, and the total time required for the CNOT simulations was about 150 hours. We considered four noise models: first pure phonon loss at rate κ_1 and then phonon loss and phonon gain with $n_{th} = 1/100$ and with dephasing at three rates, $\kappa_\phi = 1$, 2.5, and 10 times κ_1. The error rates for the CNOT gate at optimal gate time for each of the four noise models are in Table II. Our numerical results for Z error rates and the optimal gate time are in good agreement with the perturbative calculations based on the shifted Fock basis in Appendix D. We confirmed that the non-Z error rates are exponentially small in $|\alpha|^2$, and we observed that the 12 non-Z Pauli error rates, e.g. X_1, Y_1Z_2 or Y_1X_2, fall into two classes. Our perturbation theory calculations do not extend to the exponentially small error rates. However, we observe numerically that half these error rates scale with $\sqrt{\kappa_1/\kappa_2}$, while the other half scale like κ_1/κ_2. A more detailed discussion of our numerical results can be found in Appendix E.

While idling, bosonic dephasing does not induce any additional Z errors since dephasing preserves the excitation number parity. Thus, one might be tempted to conclude that dephasing only affects non-Z-type error rates of the CNOT gate and leaves the Z error rates unchanged. However, surprisingly, we numerically find that this is not the case. In particular, as shown in Table II, we observe that the optimal gate time decreases noticeably and the total optimal Z error rate (per gate) of the CNOT gate increases as the dephasing rate (per time) κ_ϕ increases.

In Appendix D, we show that the enhanced Z error rates of the CNOT gate due to dephasing are attributed to the fact that the target stabilization operator $\hat{L}_2(t)$ is not static and instead rotates conditioned on the state of the control mode. More specifically, dephasing in each mode causes direct population transfer from the ground state manifold of a cat qubit to its first excited state manifold. While such a heating itself does not cause a phase-flip error since dephasing preserves the excitation number parity, the rotating target stabilization operator $\hat{L}_2$ does cause a Z error on the control qubit while it brings the excited states of the target mode back to the ground state manifold. In the general case with a non-zero thermal population $n_{\rm th}$ and dephasing rate κ_ϕ, our perturbation theory predicts that the Z error rates of the CNOT gate (CX' gate to be more precise) are given by

$$\begin{aligned}\bar{p}_{Z_1} &= \kappa_1(1+2n_{\rm th})\alpha^2 T + \frac{1}{2}\kappa_\phi\alpha^2 T + \frac{\pi^2}{64\kappa_2\alpha^2 T},\\ \bar{p}_{Z_2} &= \bar{p}_{Z_1Z_2} = \frac{1}{2}\kappa_1(1+2n_{\rm th})\alpha^2 T.\end{aligned} \tag{31}$$

Note that dephasing adds $\kappa_\phi\alpha^2 T/2$ to p_{Z_1}. As a result, even in the lossless case (i.e., $\kappa_1 = 0$), dephasing still limits the performance of the CNOT gate since the optimal Z_1 error rate scales as $p^\star_{Z_1} \propto \sqrt{\kappa_\phi/\kappa_2}$ although $p^\star_{Z_2}$ and $p^\star_{Z_1Z_2}$ vanish. For instance, for $n_{\rm th} = 0.01$ and $\kappa_\phi = 10\kappa_1$, our perturbation theory result yields

$$\bar{p}^\star_{Z_1} = 25.6\bar{p}^\star_{Z_2} = 25.6\bar{p}^\star_{Z_1Z_2} = 1.93\sqrt{\frac{\kappa_1}{\kappa_2}}, \tag{32}$$

which agrees well with the numerical result

$$p^\star_{Z_1} = 27.1p^\star_{Z_2} = 27.1p^\star_{Z_1Z_2} = 2.14\sqrt{\frac{\kappa_1}{\kappa_2}}, \tag{33}$$

up to a relative error of 10%. See Appendix D for more details.

C. Toffoli

The bias-preserving Toffoli or CCX gate is directly analogous to the CNOT gate. The two control cavities are stabilized by the usual jump operator $\hat{L}_1 = \hat{a}_1^2 - \alpha^2$ and $\hat{L}_2 = \hat{a}_2^2 - \alpha^2$, while the third cavity is stabilized by a jump operator that couples the three cavities and rotates the third conditioned on the state of the two controls,

$$\hat{L}_3(t) = \hat{a}_3^2 - \alpha^2 - \frac{1}{4}(e^{2i\frac{\pi}{T}t} - 1)(\hat{a}_1 - \alpha)(\hat{a}_2 - \alpha). \tag{34}$$

When both cavities 1 and 2 are in the $|1\rangle \simeq |-\alpha\rangle$ cat-code state, this jump operator reduces to approximately $\hat{a}_3^2 - \alpha^2 e^{2i\frac{\pi}{T}t}$, which is the rotating jump operator that realizes the X gate on the third cavity. When one of the control cavities is in the $|0\rangle \simeq |\alpha\rangle$ cat-code state, the jump operator is approximately equal to the usual $\hat{a}_3^2 - \alpha^2$ jump operator that stabilizes the cat-code states. In this way the jump operators $\hat{L}_1$, $\hat{L}_2$, and $\hat{L}_3(t)$ implement the Toffoli gate (up to a controlled-Z rotation on the two control qubits). Also like the CNOT gate we can apply a Hamiltonian to drive the desired evolution and perform the gate much faster while cancelling part of the non-adiabatic errors. For the Toffoli gate this Hamiltonian is given by

$$\hat{H}_{\rm TOF} = -\frac{\pi}{8\alpha^2 T}((\hat{a}_1 - \alpha)(\hat{a}_2^\dagger - \alpha) + \text{h.c.})(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2). \tag{35}$$

This Hamiltonian is the natural extension of Eq. (25). It does not cancel all non-adiabatic noise, and like the CNOT in the presence of noise, the trade-off between non-adiabatic errors and noise from loss or dephasing gives rise to an optimal gate time for each value α and the noise parameters.

Similarly as in the case of the CNOT gate, we emphasize that the dissipators $\mathcal{D}[\hat{L}_1]$, $\mathcal{D}[\hat{L}_2]$, and $\mathcal{D}[\hat{L}_3(t)]$ combined with the compensating Hamiltonian $\hat{H}_{\rm TOF}$ in Eq. (35)

realize a gate

$$CCX' \equiv CZ_{1,2}(-\pi\alpha^2)\cdot \mathrm{TOF}_{1,2\to 3}, \tag{36}$$

which differs from the desired Toffoli gate $\mathrm{TOF}_{1,2\to 3}$ by a CZ rotation on the two control qubits (see Appendix D for more details). Here, $CZ(\theta)$ is defined as $CZ(\theta)\equiv \exp[i\theta|11\rangle\langle 11|]$ and $|11\rangle$ is the simulatenous -1 eigenstate of the Pauli Z operators $\hat{Z}_1$ and $\hat{Z}_2$. The extra CZ rotation persists even with an ideal compensating Hamiltonian $-(\pi/T)|{-\alpha},-\alpha\rangle\langle -\alpha,-\alpha|_{1,2}(\hat{a}_3^\dagger\hat{a}_3-\alpha^2)$ but is not present with another variant of ideal compensating Hamiltonian $-(\pi/T)|-\alpha,-\alpha\rangle\langle-\alpha,-\alpha|_{1,2}\hat{a}_3^\dagger\hat{a}_3$. Note that the extra CZ rotation $CZ_{1,2}(-\pi\alpha^2)$ is trivial if $|\alpha|^2$ is an even integer.

We simulated the Toffoli gate subject to phonon loss, gain, and dephasing at different rates by solving the master equation given by the Hamiltonian $\hat{H}_{\mathrm{TOF}}$, the dissipator on each cavity, and the Lindblad operators for the noise. These simulations were carried out using AWS EC2 c5.18xlarge instances and took about 170 hours running on instances with 72 virtual CPUs. Because we simulated three cavities for the Toffoli gate, we were able to resolve only the dominant Z-type error rates and not the other Pauli error rates that are exponentially small in α^2. These simulations used the shifted Fock basis approach. With this method we are able to use a Hilbert space dimension of 8 for each of the three cavities and simulate all of the Z Pauli error rates with high precision. The numerical results for the optimal gate time and the 7 Z-type Pauli error rates are summarized in Table II. Our simulations match our perturbation theory calculations for the Z error rates. Similar to the case of the CNOT gate, dephasing noise increases the Z error rates and shortens the optimal gate time. The dominant Z error on the control qubits 1 and 2 increases from $0.58\sqrt{\kappa_1/\kappa_2}$ to $0.91\sqrt{\kappa_1/\kappa_2}$ as the dephasing rate increases from 0 to $10\kappa_1$. Dephasing noise primarily affects the Z_1, Z_2 and Z_1Z_2 error rates. Many of the other Pauli Z error rates decrease because of the reduction in optimal gate time. Also, with dephasing noise in addition to loss, the optimal gate time for the Toffoli gate differs from the optimal gate time for the CNOT gate. With large dephasing $\kappa_\phi = 10\kappa_1$ the optimal gate time for the Toffoli gate is about 1.18 times the optimal gate time for CNOT. For simplicity, we have chosen to always operate the Toffoli gate using a gate time equal to the CNOT optimal gate time. This has a small effect on the total fidelity of the Toffoli gate and on the relative size of the different Pauli Z error probabilities.

In the presence of phonon loss, thermal population n_{th}, and dephasing, our perturbation theory yields the following Z error rates of the CCX' gate, or the Toffoli gate for even $|\alpha|^2$ (see Appendix D):

$$\begin{aligned}
\bar{p}_{Z_1} &= \bar{p}_{Z_2} = \kappa_1(1+2n_{\mathrm{th}})\alpha^2 T + \frac{1}{8}\kappa_\phi\alpha^2 T + \frac{\pi^2}{128\kappa_2\alpha^2 T},\\
\bar{p}_{Z_3} &= \frac{5}{8}\kappa_1(1+2n_{\mathrm{th}})\alpha^2 T,\\
\bar{p}_{Z_1Z_2} &= \frac{1}{8}\kappa_\phi\alpha^2 T + \frac{\pi^2}{128\kappa_2\alpha^2 T},\\
\bar{p}_{Z_1Z_3} &= \bar{p}_{Z_2Z_3} = \bar{p}_{Z_1Z_2Z_3} = \frac{1}{8}\kappa_1(1+2n_{\mathrm{th}})\alpha^2 T.
\end{aligned} \tag{37}$$

In the loss-only case ($n_{\mathrm{th}} = \kappa_\phi = 0$), the optimal gate time that minimizes the total gate infidelity is given by $\bar{T}^\star_{CCX'} = (\pi/(8\alpha^2\sqrt{2\kappa_1\kappa_2}))$ which is identical to the optimal gate time of the CX' gate (or the CNOT gate for even $|\alpha|^2$) predicted by the perturbation theory. At the optimal gate time, the Z error rates (per gate) are given by

$$\begin{aligned}
\bar{p}^\star_{Z_1} &= \bar{p}^\star_{Z_2} = 3.2\bar{p}^\star_{Z_3} = 2\bar{p}^\star_{Z_1Z_2}\\
&= 16\bar{p}^\star_{Z_1Z_3} = 16\bar{p}^\star_{Z_2Z_3} = 16\bar{p}^\star_{Z_1Z_2Z_3}\\
&= \frac{\pi}{4}\sqrt{\frac{\kappa_1}{2\kappa_2}} = 0.555\sqrt{\frac{\kappa_1}{\kappa_2}}.
\end{aligned} \tag{38}$$

which agree well with the numerical results (see Table II)

$$\begin{aligned}
p^\star_{Z_1} &= p^\star_{Z_2} = 3.05p^\star_{Z_3} = 1.81p^\star_{Z_1Z_2}\\
&= 14.9p^\star_{Z_1Z_3} = 14.9p^\star_{Z_2Z_3} = 14.9p^\star_{Z_1Z_2Z_3} = 0.58\sqrt{\frac{\kappa_1}{\kappa_2}},
\end{aligned} \tag{39}$$

up to a relative error of 5%. Thus, similarly as in the case of the CNOT gate, the perturbation theory predicts that the optimal Z error rates of the CCX' gate (or the Toffoli gate for even $|\alpha|^2$) are independent of the size of the cat code $|\alpha|^2$.

For $\kappa_\phi = 10\kappa_1$ and $n_{\mathrm{th}} = 0.01$, we find that the perturbation theory predicts the optimal gate time for the Toffoli gate is 1.25 times that of the CNOT gate by comparing Eqs. (31) and (37). Also, at the optimal gate time of the CNOT gate (predicted by the perturbation theory), the Z error rates of the Toffoli gate are predicted to be

$$\begin{aligned}
\bar{p}_{Z_1} &= \bar{p}_{Z_2} = 9.08\bar{p}_{Z_3} = 1.21\bar{p}_{Z_1Z_2}\\
&= 45.4\bar{p}_{Z_1Z_3} = 45.4\bar{p}_{Z_2Z_3} = 45.4\bar{p}_{Z_1Z_2Z_3} = 0.857\sqrt{\frac{\kappa_1}{\kappa_2}},
\end{aligned} \tag{40}$$

and agree well with the numerical result in Table II

$$\begin{aligned}
p_{Z_1} &= p_{Z_2} = 9.29p_{Z_3} = 1.08p_{Z_1Z_2}\\
&= 45.5p_{Z_1Z_3} = 45.5p_{Z_2Z_3} = 45.5p_{Z_1Z_2Z_3} = 0.91\sqrt{\frac{\kappa_1}{\kappa_2}}.
\end{aligned} \tag{41}$$

D. Z Rotation

Recall that Z and CZ rotations are needed to complete the CNOT and Toffoli gates if the average excitation number $|\alpha|^2$ is not an even integer. While we only consider even $|\alpha|^2$ in the rest of the paper, we discuss the Z and CZ rotations for completeness. A Z-axis rotation by an arbitrary angle θ can be realized using a static jump operator $\hat{L} = \hat{a}^2 - \alpha^2$ and a linear drive Hamiltonian

$$\hat{H}_Z = \epsilon_Z(\hat{a} + \hat{a}^\dagger). \tag{42}$$

Applying this Hamiltonian to a cat-code state for a time T will perform a Z rotation

$$\hat{Z}(\theta) = \exp[i\theta|1\rangle\langle 1|] = |0\rangle\langle 0| + e^{i\theta}|1\rangle\langle 1| \tag{43}$$

by angle $\theta = 4\epsilon_Z\alpha T$. That is, the Hamiltonian gives a relative phase between the $|0\rangle$ and $|1\rangle$ cat-code states. The bias in the noise is preserved under this gate because the cat-code states remain distantly separated.

By using the perturbation theory based on the shifted Fock basis, we find that the dominant Z error rate of the Z rotation $\hat{Z}(\theta)$ is given by

$$\bar{p}_Z = \kappa_1\alpha^2 T + \frac{\theta}{16\kappa_2\alpha^4 T}, \tag{44}$$

where T is the gate time (see Appendix D for more details). Hence, the optimal gate time is given by

$$\bar{T}^{\star}_{Z(\theta)} = \frac{|\theta|}{4\alpha^3\sqrt{\kappa_1\kappa_2}}, \tag{45}$$

and the optimal Z error rate is given by

$$\bar{p}^{\star}_Z = \frac{|\theta|}{2\alpha}\sqrt{\frac{\kappa_1}{\kappa_2}}. \tag{46}$$

For the Z gate (i.e., $\theta = \pi$), for instance, the perturbation theory predicts $\bar{p}^{\star}_Z = (\pi/2)\sqrt{\kappa_1/\kappa_2}/\alpha = 1.57\sqrt{\kappa_1/\kappa_2}/\alpha$ which agrees well with the numerical result $p^{\star}_Z = 1.63\sqrt{\kappa_1/\kappa_2}/\alpha$ (see Appendix E). Note that unlike in the case of the CNOT and the Toffoli gates, the optimal Z error rate decreases as the size $|\alpha|^2$ of the cat code increases. Also, since the jump operator $\hat{a}^2 - \alpha^2$ is static, dephasing does not cause any additional Z errors. In the presence of gain, the non-zero thermal population n_{th} simply replaces κ_1 by $\kappa_1(1 + 2n_{\text{th}})$.

E. CZ Rotation

A bias-preserving ZZ rotation can be realized using static jump operators $\hat{L}_1 = \hat{a}_1^2 - \alpha^2$, $\hat{L}_2 = \hat{a}_2^2 - \alpha^2$, and a beam-splitter Hamiltonian

$$\hat{H}_{ZZ} = \epsilon_{ZZ}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger). \tag{47}$$

The beam-splitter Hamiltonian gives a phase shift depending on the value of $\hat{Z}_1\hat{Z}_2$ in the cat-code basis. This Hamiltonian can be supplemented by linear drives on each cavity to produce a CZ rotation. More specifically, the following Hamiltonian

$$\begin{aligned}\hat{H}_{CZ} = \epsilon_{ZZ}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger) - \epsilon_{ZZ}\alpha(\hat{a}_1 + \hat{a}_1^\dagger)\\ - \epsilon_{ZZ}\alpha(\hat{a}_2 + \hat{a}_2^\dagger)\end{aligned} \tag{48}$$

realizes a CZ rotation

$$\begin{aligned}CZ(\theta) &= \exp[i\theta|11\rangle\langle 11|]\\ &= (\hat{I} - |11\rangle\langle 11|) + e^{i\theta}|11\rangle\langle 11|\end{aligned} \tag{49}$$

by angle $\theta = -8\epsilon_{ZZ}\alpha^2 T$. The perturbation theory predicts that the Z error rates of the CZ rotation $CZ(\theta)$ are given by

$$\begin{aligned}\bar{p}_{Z_1} &= \bar{p}_{Z_2} = \kappa_1\alpha^2 T + \frac{\theta^2}{64\kappa_2\alpha^4 T},\\ \bar{p}_{Z_1Z_2} &= \frac{\theta^2}{32\kappa_2\alpha^4 T},\end{aligned} \tag{50}$$

and thus the total gate error rate is minimized when the gate time T is given by

$$\bar{T}^{\star}_{CZ(\theta)} = \frac{|\theta|}{4\alpha^3\sqrt{\kappa_1\kappa_2}}. \tag{51}$$

At this optimal gate time, the Z error rates are given by

$$\bar{p}^{\star}_{Z_1} = \bar{p}^{\star}_{Z_2} = 1.5\bar{p}^{\star}_{Z_1Z_2} = \frac{3|\theta|}{8\alpha}\sqrt{\frac{\kappa_1}{\kappa_2}}. \tag{52}$$

For the CZ gate (i.e., $\theta = \pi$), for example, the perturbation theory yields $\bar{p}^{\star}_{Z_1} = \bar{p}^{\star}_{Z_2} = 1.5\bar{p}^{\star}_{Z_1Z_2} = (3\pi/(8\sqrt{2}))\sqrt{\kappa_1/\kappa_2}/\alpha = 0.833\sqrt{\kappa_1/\kappa_2}/\alpha$, which agrees well with the numerical result $p^{\star}_{Z_1} = p^{\star}_{Z_2} = 1.48p^{\star}_{Z_1Z_2} = 0.83\sqrt{\kappa_1/\kappa_2}/\alpha$. Similarly as in the case of the Z rotations, the optimal Z error rates decrease as the size of the cat code increases. Also, since the dissipators are static, dephasing does not cause additional Z errors.

F. Physical implementation of the gates

Here, we discuss physical realization of the cat qubit gates. Note that engineering static two-phonon dissipations in a multiplexed setting has been extensively discussed in Appendix B. Also implementation of the rotating dissipators for the CNOT and Toffoli gates are discussed in detail in Ref. [13]. We thus focus on engineering Hamiltonian interactions needed to implement the cat-qubit gates. In particular, we discuss realization of linear drive on a phononic mode, compensating Hamiltonians for the CNOT and Toffoli gates in the multiplexed setting (for a more comprehensive discussion, see

Appendix F). That is, we consider the Hamiltonian of a system consisting of multiple phononic modes $\hat{a}_k$ coupled to a shared ATS mode $\hat{b}$:

$$\hat{H}_{\text{rot}} = -2E_J\epsilon_p(t)\sin\Big(\sum_{k=1}^{N}\varphi_k\hat{a}_k e^{-i\omega_k t} + \text{h.c.} \\ + \varphi_b\hat{b}e^{-i\omega_b t} + \text{h.c.}\Big), \qquad (53)$$

where φ_k and φ_b quantify zero-point fluctuations of the modes $\hat{a}_k$ and $\hat{b}$, respectively. Note that we used the rotating frame where each mode rotates with its own frequency.

First, a linear drive on a phononic mode, say $\hat{a}_k$, can be readily realized by using a pump $\epsilon_p(t) = \epsilon_p\cos(\omega_p t)$ and choosing the pump frequency ω_p to be the frequency of the mode we want to drive, that is, $\omega_p = \omega_k$. Then, by taking only the leading order linear term in the sine potential (i.e., $\sin(\hat{x}) \simeq \hat{x}$), we get the desired linear drive

$$\hat{H}_{\text{rot}} = -E_J\epsilon_p\varphi_k(\hat{a}_k + \hat{a}_k^\dagger) + \hat{H}', \qquad (54)$$

where $\hat{H}'$ contains fast-oscillating terms such as $-E_J\epsilon_p(\varphi_l\hat{a}_l e^{-i(\omega_l-\omega_k)t} + \text{h.c.})$ with $l \neq k$ and $-E_J\epsilon_p(\varphi_b\hat{b}e^{-i(\omega_b-\omega_k)t} + \text{h.c.})$ as well as other terms that rotate even faster, e.g., $-E_J\epsilon_p\varphi_k(\hat{a}_k e^{-2i\omega_k t} + \text{h.c.})$. Since the frequency differences between different modes are on the order of 100MHz but $|\epsilon_Z|/(2\pi)$ is typically much smaller than 100MHz (see Appendix F), the fast-oscillating terms can be ignored by using a rotating wave approximation (RWA). Note that the subleading cubic term in the sine potential is also neglected here. In particular, terms like $\hat{a}_k^\dagger\hat{a}_k^2$ rotate at the same frequency as the desired linear term $\hat{a}_k$. However, the coupling strength of these terms are smaller than that of the linear term by a factor of φ_k^2. To avoid driving unwanted higher order terms, one may alternatively drive the phononic mode directly, at the expense of increased hardware compleixty, instead of using the pump $\epsilon_p(t)$ at the ATS node.

Let us move on to the implementation of the compensating Hamiltonian for the CNOT gate in Eq. (25). Without loss of generality, we focus on the CNOT gate between the modes $\hat{a}_1$ (control) and $\hat{a}_2$ (target). Note that $\hat{H}_{\text{CNOT}}$ consists of an optomechanical coupling $(\pi/(4\alpha T))(\hat{a}_1 + \hat{a}_1^\dagger)\hat{a}_2^\dagger\hat{a}_2$ between two phononic modes, a linear drive on the control mode $-(\pi\alpha/(4T))(\hat{a}_1 + \hat{a}_1^\dagger)$, which is discussed above, and a selective frequency shift of the target mode $-(\pi/(2T))\hat{a}_2^\dagger\hat{a}_2$. To realize the optomechanical coupling, one might be tempted to directly drive the cubic term $\hat{a}_1\hat{a}_2^\dagger\hat{a}_2 + \text{h.c.}$ in the sine potential via a pump $\epsilon_p(t) = \epsilon_p\cos(\omega_p t)$. However, the direct driving scheme is not suitable for a couple of reasons: since the term $\hat{a}_1\hat{a}_2^\dagger\hat{a}_2$ rotates with frequency ω_1, the required pump frequency is given by $\omega_p = \omega_1$ which is the same pump frequency reserved to engineer a linear drive on the $\hat{a}_1$ mode. Moreover, the term $\hat{a}_1\hat{a}_2^\dagger\hat{a}_2$ rotates at the same frequency as those of undesired cubic terms such as $\hat{a}_1\hat{a}_3^\dagger\hat{a}_3$, $\hat{a}_1\hat{a}_4^\dagger\hat{a}_4$, and also $\hat{a}_1^\dagger\hat{a}_1^2$. Hence, even if the linear drive is realized by directly driving the phononic mode $\hat{a}_1$, one cannot selectively drive the desired optomechanical coupling by using the pump frequency $\omega_p = \omega_1$ due to the frequency collision with the other cubic terms.

To circumvent the above frequency-collision issue, we propose to realize the optomechanical coupling by off-resonantly driving the term $(\hat{a}_1 + \lambda)\hat{a}_2\hat{b}^\dagger$. In particular, we use the fact that a time-dependent Hamiltonian $\hat{H} = \chi\hat{A}\hat{b}^\dagger e^{-i\Delta t}$ yields an effective Hamiltonian $\hat{H}_{\text{eff}} = (\chi^2/\Delta)\hat{A}^\dagger\hat{A}$ upon time-averaging assuming that the population of the $\hat{b}$ mode is small (i.e., $\hat{b}^\dagger\hat{b} \ll 1$) and the detuning Δ is sufficiently large (see Appendix F for more details). Hence, given a Hamiltonian $\hat{H} = \chi(\hat{a}_1 + \lambda)\hat{a}_2\hat{b}^\dagger e^{-i\Delta t} + \text{h.c.}$, we get

$$\hat{H}_{\text{eff}} = \frac{\chi^2\lambda}{\Delta}\Big(\hat{a}_1 + \hat{a}_1^\dagger + \lambda + \frac{1}{\lambda}\hat{a}_1^\dagger\hat{a}_1\Big)\hat{a}_2^\dagger\hat{a}_2. \qquad (55)$$

In particular, by choosing $\lambda = -2\alpha$, we can realize the optomechanical coupling as well as the selective frequency shift of the $\hat{a}_2$ mode, i.e., $\hat{H}_{\text{eff}} \propto (\hat{a}_1 + \hat{a}_1^\dagger - 2\alpha)\hat{a}_2^\dagger\hat{a}_2$ up to an undesired cross-Kerr term $-\hat{a}_1^\dagger\hat{a}_1\hat{a}_2^\dagger\hat{a}_2/(2\alpha)$. In this scheme, we have the desired selectivity because the term $(\hat{a}_1 + \lambda)\hat{a}_2\hat{b}^\dagger$ is detuned from other undesired terms such as $(\hat{a}_1 + \lambda)\hat{a}_k\hat{b}^\dagger$ with $k \geq 3$ by a frequency difference $\omega_2 - \omega_k$. Thus, the unwanted optomechanical coupling $(\hat{a}_1 + \hat{a}_1^\dagger)\hat{a}_k^\dagger\hat{a}_k$ can be suppressed by a suitable choice of the detuning Δ. We remark that the unwanted cross-Kerr term $\hat{a}_1^\dagger\hat{a}_1\hat{a}_2^\dagger\hat{a}_2$ can in principle be compensated by off-resonantly driving another cubic term $\hat{a}_1\hat{a}_2\hat{b}^\dagger$ with a different detuning $\Delta' \neq \Delta$.

Lastly, let us consider the compensating Hamiltonian $\hat{H}_{\text{TOF}}$ for the Toffoli gate in Eq. (35). $\hat{H}_{\text{TOF}}$ is explicitly given by

$$\hat{H}_{\text{TOF}} = -\frac{\pi}{8\alpha^2 T}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2) \\ + \frac{\pi}{8\alpha T}(\hat{a}_1 + \hat{a}_1^\dagger - \alpha)(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2) \\ + \frac{\pi}{8\alpha T}(\hat{a}_2 + \hat{a}_2^\dagger - \alpha)(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2). \qquad (56)$$

Note that the terms in the second and the third lines are in the same form as the compensating Hamiltonian for the CNOT gate. Thus, they can be realized in a similar way as described above. The terms in the first line contain a beam-splitter interaction $(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)$ as well as a quartic term $(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)\hat{a}_3^\dagger\hat{a}_3$. Realization of the beam-splitter interaction is discussed in Appendix F. Since the sine potential has an odd parity, it is not possible to drive the quartic term directly. The quartic term can nevertheless be realized by off-resonantly driving the term $(\hat{a}_1 + \hat{a}_2)\hat{a}_3\hat{b}^\dagger$: given $\hat{H} = \chi(\hat{a}_1 + \hat{a}_2)\hat{a}_3\hat{b}^\dagger e^{-i\delta t} + \text{h.c.}$, we get

$$\hat{H}_{\text{eff}} = \frac{\chi^2}{\Delta}(\hat{a}_1^\dagger \hat{a}_2 + \hat{a}_1 \hat{a}_2^\dagger)\hat{a}_3^\dagger \hat{a}_3 + \frac{\chi^2}{\Delta}(\hat{a}_1^\dagger \hat{a}_1 + \hat{a}_2 \hat{a}_2^\dagger)\hat{a}_3^\dagger \hat{a}_3, \tag{57}$$

i.e., the desired quartic interaction and unwanted cross-Kerr interactions between a control and the target modes. Similarly as in the case of the CNOT gate, the undesired cross-Kerr terms, which are as strong as the desired quartic term, can in principle be cancelled by off-resonantly driving the terms $\hat{a}_1 \hat{a}_3 \hat{b}^\dagger$ and $\hat{a}_2 \hat{a}_3 \hat{b}^\dagger$ with detunings Δ_1 and Δ_2 which are different from each other and also from Δ.

G. X measurement

X-basis readout entails determining the parity of a bosonic mode. Specifically this is readout in the basis of even and odd cat states, i.e., $|\pm\rangle \propto |\alpha\rangle \pm |-\alpha\rangle$. In the following we discuss two schemes for X-basis readout. The first uses an additional readout phononic mode upon which we repeatedly perform parity measurements with a transmon qubit in parallel with the gates of the next error correction cycle. Infidelities corresponding to this scheme are what we use for most of the error correction simulations. Second we discuss a measurement scheme that does not rely on a transmon and can achieve competitive fidelities and times with the first scheme in the context of error correction.

In the first measurement scheme we use an additional phononic readout mode in every unit cell that we do not stabilize with two phonon dissipation. This readout mode is interrogated by a transmon qubit in parallel with the gates of the next error correction cycle. This allows us to achieve high measurement fidelity and minimal idling time for the data qubits. The layout used is pictured in Fig. 2. Note that in addition, for purposes of reading out the transmon it will be coupled to a readout resonator coupled to a transmission line.

Here we outline the steps for the readout of an ancilla qubit in the X basis. The first step in the measurement procedure is the deflation of the ancilla mode ($\hat{a}_1$) which maps the even parity cat state to the Fock state $|\hat{n} = 0\rangle$ and the odd cat state to the Fock state $|\hat{n} = 1\rangle$ [24]. With evolution under two-phonon dissipation given by

$$\frac{d\hat{\rho}(t)}{dt} = \kappa_2 \mathcal{D}[\hat{a}_1^2 - \alpha(t)^2]\hat{\rho}(t) \tag{58}$$

deflation is achieved by varying $\alpha(t)$ from an initial $\alpha(t_{\text{initial}}) = \alpha_0$ to $\alpha(t_{\text{final}}) = \alpha_1 < \alpha_0$. For our readout scheme we use $\alpha_1 = 0$ so the final states of the ancilla mode are the $|\hat{n} = 0\rangle$ and $|\hat{n} = 1\rangle$ Fock states. For readout we do not need to maintain the full phase coherence between the even and odd parity states, so we may rapidly take $\alpha(t)$ from α_0 to α_1 and wait for a few $1/\kappa_2$. The purpose of the deflation is to reduce susceptibility to single phonon loss events which change the parity of the cat qubit. Note that after this deflation we turn off the two dissipation.

Subsequent to the deflation the ancilla mode and readout mode ($\hat{a}_2$) evolve under a beamsplitter Hamiltonian (derived starting with Eq. (F7))

$$\hat{H} = g(\hat{a}_1^\dagger \hat{a}_2 + \hat{a}_2^\dagger \hat{a}_1) \tag{59}$$

which performs a SWAP gate between the ancilla mode and readout mode in a time $\pi/2g$ (the swapped state is rotated by 90 degrees). In the above equation $g = E_J \epsilon_p \beta \varphi_{a1} \varphi_{a2} \varphi_b$. Since this depends twice on the zero point fluctations of storage modes the strength without the drive factor β would be similar to g_2. In order to maintain a low population of the buffer mode necessary in the derivation of the Hamiltonian we take $\beta \ll 1$. We assume a strength for the coupling of $g/2\pi = 1$ MHz.

After the SWAP the excitation is in the readout mode, we perform repeated QND (quantum non-demolition) measurements [53–55] and take a majority vote to get our final measurement outcome. In our case the individual measurements are standard QND bosonic parity measurements which are performed using a dispersive coupling between the readout mode and a transmon qubit ($\hat{q}$) of the form $\hat{H}_{\text{dispersive}} = \chi \hat{a_2}^\dagger \hat{a_2} \hat{q}^\dagger \hat{q}$. Evolution under the Hamiltonian for a time π/χ gives the controlled parity gate

$$U = I \otimes |g\rangle\langle g| + e^{i\hat{a}_2^\dagger \hat{a}_2 \pi}|e\rangle\langle e|. \tag{60}$$

which can be used in conjunction with simple transmon state preparation and measurement to realize parity measurements of the readout mode [53].

While this repeated parity measurement is taking place, the CNOT gates of the next error correction cycle can occur in parallel. This enables us to reach high readout fidelity without affecting the length of an error correction cycle. Specifically for the case of the repetition code or surface code we are able to complete 3 or 5 parity measurements respectively during the error correction gates of the next cycle.

We simulated this measurement scheme with varying κ_1/κ_2, gain, and κ_ϕ for the phononic modes, including the effects of transmon measurement error to get a rough sense of the expected measurement fidelities. The effects of gain and dephasing contribute predominantly during the deflation and SWAP gate but in general do not have a large effect on the overall readout fidelity. The misassignment probabilities and measurement times can be found in Table III for select values of κ_ϕ and κ_1/κ_2. As expected the effects of the transmon infidelity mechanisms become suppressed as we increase the number of measurements included in the majority vote. The error, especially for large κ_1/κ_2, is larger for the odd parity state since it is susceptible to single phonon loss for a longer duration.

For most of the error correction simulations we have used the X-basis measurement scheme just outlined. In

Idle	$\kappa_\phi = 0$, $n_{th} = 0$	$\kappa_\phi = \kappa_1$, $n_{th} = 1/100$	$\kappa_\phi = 2.5\kappa_1$, $n_{th} = 1/100$	$\kappa_\phi = 10\kappa_1$, $n_{th} = 1/100$
Z	$\kappa_1\alpha^2 T$	$1.02\kappa_1\alpha^2 T$	$1.02\kappa_1\alpha^2 T$	$1.02\kappa_1\alpha^2 T$
Y	$\kappa_1\alpha^2 T\exp(-4\alpha^2)$	$\ll p_X$	$\ll p_X$	$\ll p_X$
X	$\ll p_Y$	$\kappa_1\alpha^2 T\exp(-2\alpha^2)$	$2.5\kappa_1\alpha^2 T\exp(-2\alpha^2)$	$10\kappa_1\alpha^2 T\exp(-2\alpha^2)$
$\|0\rangle$ Prep				
Time	$0.1 * (\kappa_2\alpha^2)^{-1}$	$0.1 * (\kappa_2\alpha^2)^{-1}$	$0.1 * (\kappa_2\alpha^2)^{-1}$	$0.1 * (\kappa_2\alpha^2)^{-1}$
$X + Y$	$0.39\exp(-4\alpha^2)$	$0.39\exp(-4\alpha^2)$	$0.39\exp(-4\alpha^2)$	$0.39\exp(-4\alpha^2)$
$\|+\rangle$ Prep				
Time	$10 * (\kappa_2\alpha^2)^{-1}$	$10 * (\kappa_2\alpha^2)^{-1}$	$10 * (\kappa_2\alpha^2)^{-1}$	$10 * (\kappa_2\alpha^2)^{-1}$
$X + Y$	$7.5\kappa_1/\kappa_2$	$7.5\kappa_1/\kappa_2$	$7.5\kappa_1/\kappa_2$	$7.5\kappa_1/\kappa_2$

TABLE I. Table of error rates for idle and state preparation in the cat code. For the $|+\rangle$ ($|0\rangle$) state preparation, we initialize the system to the vaccum state $|\hat{n} = 0\rangle$ (the coherent state $|\alpha\rangle$), turn on the engineered two-phonon dissipation, and wait until the system relaxes to the $|+\rangle$ ($|0\rangle$) state.

CNOT	$\kappa_\phi = 0$, $n_{th} = 0$	$\kappa_\phi = \kappa_1$, $n_{th} = 1/100$	$\kappa_\phi = 2.5\kappa_1$, $n_{th} = 1/100$	$\kappa_\phi = 10\kappa_1$, $n_{th} = 1/100$	Scaling
Optimal Gate Time	$0.31\|\alpha\|^{-2}$	$0.27\|\alpha\|^{-2}$	$0.24\|\alpha\|^{-2}$	$0.16\|\alpha\|^{-2}$	$(\kappa_1\kappa_2)^{-\frac{1}{2}}$
Z_1	0.91	1.10	1.33	2.14	$\sqrt{\kappa_1/\kappa_2}$
Z_2, Z_1Z_2	0.15	0.14	0.12	0.079	$\sqrt{\kappa_1/\kappa_2}$
X_1, X_2, X_1X_2, Y_1, Y_1Y_2, Z_1Y_2	$0.93\exp(-2\|\alpha\|^2)$	$1.07\exp(-2\|\alpha\|^2)$	$1.28\exp(-2\|\alpha\|^2)$	$2.01\exp(-2\|\alpha\|^2)$	$\sqrt{\kappa_1/\kappa_2}$
Y_2, Y_1Y_2, X_1Y_2, X_1Z_2, Y_1Z_2, Z_1Y_2	$0.28\exp(-2\|\alpha\|^2)$	$0.29\exp(-2\|\alpha\|^2)$	$0.30\exp(-2\|\alpha\|^2)$	$0.28\exp(-2\|\alpha\|^2)$	(κ_1/κ_2)
Toffoli					
At CNOT optimal Time					
$Z_1 = Z_2$	0.58	0.62	0.68	0.91	$\sqrt{\kappa_1/\kappa_2}$
Z_3	0.19	0.17	0.15	0.098	$\sqrt{\kappa_1/\kappa_2}$
Z_1Z_2	0.32	0.41	0.50	0.84	$\sqrt{\kappa_1/\kappa_2}$
$Z_1Z_3 = Z_2Z_3$	0.039	0.035	0.031	0.020	$\sqrt{\kappa_1/\kappa_2}$
$Z_1Z_2Z_3$	0.039	0.034	0.030	0.020	$\sqrt{\kappa_1/\kappa_2}$

TABLE II. Table of gate error rates for CNOT and Toffoli. The κ_1 dependence for each error rate is shown in the rightmost column, and the error rate is the product of this term and the corresponding function of alpha to the left. The second column is for a noise model with only phonon loss. The next three columns include phonon gain with $n_{th} = 1/100$ and dephasing noise at various rates.

experiments it is advantageous to avoid using an additional transmon for every unit cell since this adds significant spatial and control-line overhead. In addition, the requirement of having 5 modes per ATS limits us due to crosstalk considerations. As is discussed in Appendix G in more detail, using deflation and a coupling of the form

$$\hat{H} = ig\hat{a}^\dagger\hat{a}(\hat{b}^\dagger - \hat{b}) \tag{61}$$

we can perform X-basis readout without a transmon (longitudinal readout [56]). Above $\hat{b}$ corresponds to a buffer mode and $\hat{a}$ corresponds to the storage mode we aim to readout. By deflating the cat qubit to the $|\hat{n} = 0\rangle$, $|\hat{n} = 1\rangle$ manifold and evolving under the above Hamiltonian the parity of the cat qubit can be determined by homodyne measurement of the $\hat{b}$ mode.

We discuss how to achieve this Hamiltonian from the ATS potential and the signal-to-noise ratio (SNR) for the measurement subsequent to the deflation in Appendix G. The SNR for the measurement distinguishing $|\hat{n} = 0\rangle$ from $|\hat{n} = 1\rangle$ is given by

$$\mathrm{SNR}(\tau) = \frac{4g_r}{\kappa_b\sqrt{2\kappa_b t}}(-2 + 2e^{-\frac{\kappa_b t}{2}} + \kappa_b t). \tag{62}$$

Using a $\hat{b}$ mode with $\kappa_b/2\pi = 20$ MHz, a coupling of

$\frac{\kappa_1}{\kappa_2}$ Initial Parity	10^{-5} Even	10^{-5} Odd	10^{-4} Even	10^{-4} Odd	10^{-3} Even	10^{-3} Odd	T_{affected}	T_{total}
$\kappa_\phi = 0$								
1 Parity Meas.	$1.0 * 10^{-2}$	$1.0 * 10^{-2}$	$1.0 * 10^{-2}$	$1.1 * 10^{-2}$	$1.2 * 10^{-2}$	$1.6 * 10^{-2}$	550 ns	.95 μs
3 Parity Meas.	$3.1 * 10^{-4}$	$4.0 * 10^{-4}$	$4.6 * 10^{-4}$	$1.5 * 10^{-3}$	$1.9 * 10^{-3}$	$1.2 * 10^{-2}$	550 ns	2.15 μs
5 Parity Meas.	$2.5 * 10^{-5}$	$1.8 * 10^{-4}$	$1.7 * 10^{-4}$	$1.8 * 10^{-3}$	$1.6 * 10^{-3}$	$1.8 * 10^{-2}$	550 ns	3.35 μs
$\kappa_\phi = \kappa_1^*$								
1 Parity Meas.	$1.0 * 10^{-2}$	$1.0 * 10^{-2}$	$1.0 * 10^{-2}$	$1.1 * 10^{-2}$	$1.2 * 10^{-2}$	$1.7 * 10^{-2}$	550 ns	.95 μs
3 Parity Meas.	$3.2 * 10^{-4}$	$4.0 * 10^{-4}$	$4.8 * 10^{-4}$	$1.6 * 10^{-3}$	$2.1 * 10^{-3}$	$1.3 * 10^{-2}$	550 ns	2.15 μs
5 Parity Meas.	$2.6 * 10^{-5}$	$1.8 * 10^{-4}$	$2.0 * 10^{-4}$	$1.9 * 10^{-3}$	$1.8 * 10^{-3}$	$1.9 * 10^{-2}$	550 ns	3.35 μs
ATS Readout								
	$3.5 * 10^{-3}$	$3.0 * 10^{-2}$	$3.6 * 10^{-3}$	$3.5 * 10^{-3}$	$5.1 * 10^{-3}$	$7.7 * 10^{-3}$	800 ns	800 ns

TABLE III. Table containing X-basis measurement error rates for various values of κ_1/κ_2 with fixed $|\alpha|^2 = 8$. The noise parameters marked with * also include $n_{th} = 0.01$. For measurement error rates corresponding to more than one parity measurement, we used majority voting in obtaining the final result. For the error correction simulations performed in Section IV, we used the average measurement error rate between even and odd parities. Such an approach is justified since in an experiment, one could randomly choose whether the ancilla qubit is initialized in the even or odd cat state. Additionally we use the number of measurements that gives us the optimal infidelity in error correction simulations. In the bottom row (labeled "ATS Readout") we added error probabilities for the second X readout scheme (where only four modes are coupled to an ATS) for the case $\kappa_\phi = 0$. More details on the assumptions made to obtain the results for both schemes can be found in Appendix G in addition to plots including data points for all values of κ_1/κ_2 used.

$g/2\pi = 6$ MHz, and assuming a quantum efficiency of .5, we can achieve average readout error probabilities of roughly $3 * 10^{-3} - 4 * 10^{-3}$ in a time of $T_{\text{deflation}} + T_{\text{measure}} = 4/\kappa_2 + 400$ ns for $\kappa_1/\kappa_2 < 10^{-4}$. Error probabilities for select κ_1/κ_2 are given in Table III. We expect to be able to achieve this κ_b using a multimode buffer. These parameters give $\langle \hat{b}^\dagger \hat{b} \rangle < 1$ as is required for the implementation of the readout Hamiltonian. In future work we will confirm numerically the feasibility of these parameters in the context of the effective Hamiltonian theory. Use of lower readout efficiencies and couplings can be compensated by changing the duration of the readout and the coupling strength in addition to optimizing the readout window function. The error due to the deflation is also discussed in Appendix G.

H. *Z* measurement

With Z-basis measurement the goal is to distinguish $|0\rangle$ and $|1\rangle$ which are approximately the coherent states $|\alpha\rangle$ and $|-\alpha\rangle$. In the following we describe how high fidelity and fast readout can be achieved without using a transmon. Specifically, while the buffer and storage mode undergo a swapping interaction we perform a homodyne measurement of a transmission line coupled to the buffer mode. The swapping interaction transfers excitations from the storage mode to the buffer mode where they rapidly leak to the transmission line and are measured. For this readout we use a Hamiltonian of the form

$$\hat{H} = g(\hat{a}^\dagger \hat{b} + \hat{b}^\dagger \hat{a}) \tag{63}$$

where $\hat{a}$ is the phononic ancilla mode we are trying to read out and $\hat{b}$ is the buffer mode. Here $g = E_J \epsilon_p \beta \varphi_a \varphi_b^2$ where β is the strength of a drive on the $\hat{b}$ mode and ϵ_p the strength of a pump. To realize this form of the beam-splitter interaction we drive the two terms $\hat{a}^\dagger \hat{b}^2$ and $\hat{b}$ off-resonantly. In particular, we use a pump $\epsilon_p(t) = \epsilon_p \cos(\omega_p t)$ with frequency $\omega_p = 2\omega_b - \omega_a + \Delta$ to off-resonantly drive the term $\hat{a}\hat{b}^{\dagger 2}$ and directly drive the $\hat{b}$ mode via

$$\hat{H}_d = \epsilon_d (\hat{b}^\dagger e^{-i\omega_d t} + \text{h.c.}) \tag{64}$$

with a drive frequency $\omega_d = \omega_b + \Delta$ to off-resonantly drive the term $\hat{b}$. Then, by taking up to the third order terms in the sine potential we get the Hamiltonian in the rotating frame of all of the modes.

$$\begin{aligned} \hat{H}_{\text{rot}} = &\frac{1}{2} E_J \epsilon_p \varphi_a \varphi_b^2 \hat{a}^\dagger \hat{b}^2 e^{i\Delta t} + \text{h.c.} \\ &+ \epsilon_d \hat{b}^\dagger e^{-i\Delta t} + \text{h.c.} + \hat{H}' \end{aligned} \tag{65}$$

where $\hat{H}'$ contains rapidly rotating terms. Now let $\chi_1 \equiv E_J \epsilon_p \varphi_a \varphi_b^2 / 2$. Then, neglecting $\hat{H}'$ and constants, the average Hamiltonian theory yields [46, 47]

$$\begin{aligned} \hat{H}_{\text{eff}} &= \frac{1}{\Delta} [\chi_1 \hat{a}^\dagger \hat{b}^2 + \epsilon_d \hat{b}, \chi_1 \hat{a} \hat{b}^{\dagger 2} + \epsilon_d \hat{b}^\dagger] \\ &= \frac{1}{\Delta} \Big[\chi_1^2 [2(1 + 2\hat{b}^\dagger \hat{b}) \hat{a}^\dagger \hat{a} - \hat{b}^{\dagger 2} \hat{b}^2] + 2\chi_1 \epsilon_d (\hat{a}^\dagger \hat{b} + \hat{a} \hat{b}^\dagger) \Big] \\ &\xrightarrow{\hat{b}^\dagger \hat{b} \ll 1} g(\hat{a}^\dagger \hat{b} + \hat{a} \hat{b}^\dagger) + g_b \hat{a}^\dagger \hat{a}. \end{aligned} \tag{66}$$

The coupling constant is given by $g = E_J \epsilon_p \beta \varphi_a \varphi_b^2$ where $\beta = \epsilon_d/\Delta$ and there is an energy shift. The strength of the coupling is on the order of g_2 since it depends twice on $\varphi_b > \varphi_a$ and $\beta < 1$ to ensure $\hat{b}^\dagger \hat{b} < 1$. Note also that the assumption $\hat{b}^\dagger \hat{b} \ll 1$ can be relaxed since the parasitic terms can be somewhat accounted and compensated for. Evolution under this Hamiltonian swaps excitations from the storage mode to the buffer mode where they rapidly leak to the transmission line with a rate κ_b. By a homodyne measurement of the output of the transmission line we can determine which coherent state the storage mode was in. Now we outline a computation of the SNR for this readout scheme. The SNR is defined as

$$\mathrm{SNR}^2 = \frac{|\langle \hat{M} \rangle_\alpha - \langle \hat{M} \rangle_{-\alpha}|^2}{\langle \hat{M}^2_{N(\alpha)} \rangle + \langle \hat{M}^2_{N(-\alpha)} \rangle} \tag{67}$$

where $\hat{M}(\tau) = \sqrt{\kappa_b} \int_0^\tau dt [\hat{b}^\dagger_{out}(t) e^{i\phi_h} + \hat{b}_{out}(t) e^{-i\phi_h}]$ is the operator corresponding to the homodyne detection with angle ϕ_h [56, 57]. Specifically the signal for the measurement is $|\langle \hat{M} \rangle_\alpha - \langle \hat{M} \rangle_{-\alpha}|$ and the noise associated with the measurement is $|\langle \hat{M}^2_{N(\alpha)} \rangle + \langle \hat{M}^2_{N(-\alpha)} \rangle|^{1/2}$ where $\hat{M}_N = \hat{M} - \langle \hat{M} \rangle$. $\hat{b}_{out}$ is defined by the standard input-output theory relations to be $\hat{b}_{out} = \hat{b}_{in} + \sqrt{\kappa_b} \hat{b}$. We take the input field ($\hat{b}_{in}$) to be vacuum though technically there is a weak drive to realize the readout Hamiltonian (this could be replaced by a pump at ω_b). The coupled equations for the storage and buffer mode are

$$\begin{aligned} \dot{\hat{a}} &= -i[\hat{a}, \hat{H}_r] = -ig_r \hat{b}, \\ \dot{\hat{b}} &= -i[\hat{b}, \hat{H}_r] - \frac{\kappa_b}{2} \hat{b} - \sqrt{\kappa_b} \hat{b}_{in} = \\ &- ig_r \hat{a} - \frac{\kappa_b}{2} \hat{b} - \sqrt{\kappa_b} \hat{b}_{in}. \end{aligned} \tag{68}$$

Solving these equations and substituting as is done in Appendix G we find that the SNR with the optimal homodyne angle is approximately

$$\mathrm{SNR}(\tau) = \alpha \sqrt{8 \kappa_b} \frac{\left[1 - e^{-\kappa_b \tau/4} \left[\cosh \frac{\beta\tau}{4} + \frac{\kappa_b}{\beta} \sinh \frac{\beta\tau}{4}\right]\right]}{g\sqrt{\tau}} \tag{69}$$

where $\beta = \sqrt{\kappa_b^2 - (4g)^2}$. Given this expression one can easily find numerically which measurement time gives the minimum infidelity. The measurement SNR scales as α which means there is an exponential improvement of the measurement fidelity with $|\alpha|^2$. This measurement process is not quantum non-demolition (QND) and at long times the measurement SNR goes as $1/\sqrt{\tau}$. This is expected since the storage mode is empty and there is no more signal at long times so we are integrating noise. From the measurement SNR, the measurement separation error is computed as

$$\epsilon_{\mathrm{sep}}(\tau) = \frac{1}{2} \mathrm{Erfc}(\frac{\mathrm{SNR}(\tau)}{2}) \tag{70}$$

[58]. As is discussed in Appendix G single photon loss lowers the effective α^2 with some probability leading to a worse separation error. This effect is subleading. Additionally the manifestation of dephasing is approximately as an enhanced single phonon loss rate $\kappa_a \longrightarrow \kappa_a + \kappa_\phi$. We have also simulated the stochastic master equation corresponding to this measurement scheme and confirmed that it agrees with our analytics. We used one conservative fit of the measurement separation error vs. $|\alpha|^2$ for all of the cases considered in error correction simulations.

In the error correction simulations the time and error relations used for Z measurement were

$$\begin{aligned} \epsilon(|\alpha|^2) &= e^{-1.5 - .9|\alpha|^2} \\ \mathrm{T}_{\mathrm{measure}} &= 150 \text{ ns} \end{aligned} \tag{71}$$

where ϵ is the probability to readout the incorrect state. To get these numbers we used a $g/2\pi = 4$ MHz, $\kappa_b/2\pi = 20$ MHz, and a quantum efficiency of 0.5. Note that in the final parameters κ_b was higher but this does not have a large effect on the conservative readout fidelities we used and the time remains less than that of the X measurement up to $\kappa_b/2\pi$ of roughly 100 MHz. We can also envision using a multimode buffer to achieve a lower κ_b. This readout scheme can also be optimized further by tuning the integration window of the readout. In the future, we plan to confirm that we can loosen the assumptions for the Z-basis measurement with a minor effect on the surface code logical failure rate. This is expected given the robustness of the surface code to higher X-basis measurement errors we observed.

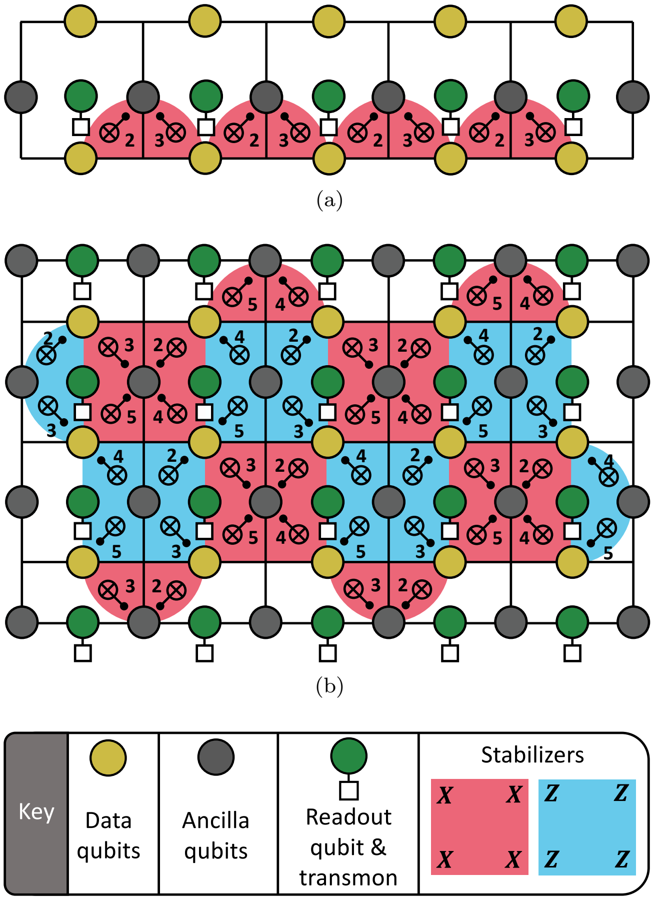
We note that potentially better measurement fidelity can be achieved by using an interaction Hamiltonian

$$\hat{H} = g \hat{X}_a \hat{P}_b = ig(\hat{a}^\dagger + \hat{a})(\hat{b}^\dagger - \hat{b}) \tag{72}$$

which allows for QND readout of the bosonic mode's position $\hat{X}_a$. This Hamiltonian can easily be achieved by adding additional pumps and drives.

IV. LOGICAL FAILURE RATES FOR QUANTUM MEMORY

The two codes that we use in our architecture for implementing quantum algorithms are the repetition code and the rotated surface code [59]. The stabilizer measurement circuits for such codes are given in Fig. 5. As described in Sections VI and VII, the repetition code is used for preparing $|\mathrm{TOF}\rangle$ magic states which will allow us to implement logical Toffoli gates. However, the repetition code alone is insufficient for universal quantum computation since, without the ability to correct at least one bit-flip error, the logical X-failure rates would be



(a)

(b)

(c)

FIG. 5. (a) Circuit illustration of a $d = 5$ repetition code embedded in our ATS layout. As explained in Fig. 2, the yellow vertices correspond to the data qubits and the gray vertices to the ancillas, whereas green circles are readout modes and the white squares are transmon qubits, which we use for X-basis measurements (see Appendix G and Fig. 40 for more details). The pink semi-circles are used to illustrate the X_iX_{i+1} stabilizer of the repetition code. We also label each CNOT gate by the corresponding time step in which it is applied. (b) Circuit illustration of a $d_x = 3$ by $d_z = 5$ thin rotated surface code. The pink and blue plaquettes correspond to the X and Z-type stabilizers respectively, with the numbers indicating the time steps in which the CNOT gates are applied. (c) Key illustrating the different components of the repetition and surface code lattices.

too high during the implementation of most quantum algorithms of interest for reasonable values of α^2 (see Fig. 7). As such, apart from the preparation of $|\text{TOF}\rangle$ states (which will be converted to $|\text{TOF}\rangle$ states encoded in the surface code using lattice surgery), all logical gates of quantum algorithms are performed in a $d_x = 3$ by d_z rotated surface code lattice. Here d_x and d_z denote the minimum weight of the X- and Z-type logical operators of the rotated surface code. We fix $d_x = 3$ since as will be seen, we will only need to correct one bit-flip error at the surface code level to get the desired logical X failure rates for the implementation of quantum algorithms of practical interest such as those considered in Section VIII.

In this section, we provide logical Z failure rates for the repetition code and rotated surface code in the context of quantum memories using a minimum-weight perfect matching (MWPM) decoding algorithm with weighted edges described in Appendix M and the noise model described in Section III. We also provide general logical X failure rate polynomials of the rotated surface code as a function of the d_z distance.

A. Repetition code logical failure rates

The logical Z failure rates of the repetition code for distance $3 \leq d \leq 19$ are provided in Fig. 6. All results were obtained from a Monte-Carlo simulation based on the circuit level noise model where each gate, state-preparation, idling qubits and measurements fail with probabilities given in Tables I to III.

In error correction there are two settings of interest: where the logical information needs to be stored for some fixed period of time; and where there is flexibility to adapt the number of rounds before proceeding to the next stage of the computation. Here we introduce the `STOP` algorithm, which is an adaptive policy for decoding how many rounds to repeat the syndrome measurements. In the limit of large code distances, `STOP` terminates (with high probability) in the same number of rounds as an algorithm using fixed d rounds. For smaller code distances and low noise regimes, `STOP` provides an advantage over a fixed round decoder as it requires $(d + 1)/2$ rounds. Full details for the implementation of the `STOP` algorithm are provided in Appendix H. We now give two important remarks:

Remark one: Consider first the setting where the logical information is stored for a fixed period of time. The standard approach that is followed in the literature when obtaining numerical results for decoding such codes is to perform d rounds of noisy syndrome measurements followed by one round of perfect syndrome measurement (where no additional errors are introduced). Errors are then corrected using the full syndrome history. The round of perfect syndrome measurement is added to ensure that the final error after correction is either in the stabilizer group or corresponds to a logical operator (i.e. we must ensure that we project to the code-space to declare success or failure). Furthermore, if the error syndrome was decoded based *only* on d noisy syndrome measurement rounds (i.e. without the round of perfect error correction), a single measurement error occurring in the dth round could result in a logical failure (a fact that is often not fully appreciated). However for many models of universal quantum computation, the data qubits are measured directly as part of the quantum algorithm or during the implementation of state injection for performing non-Clifford gates (see Refs. [60–62] and Fig. 43). As illustrated in Fig. 43 of Appendix H, the direct measure-

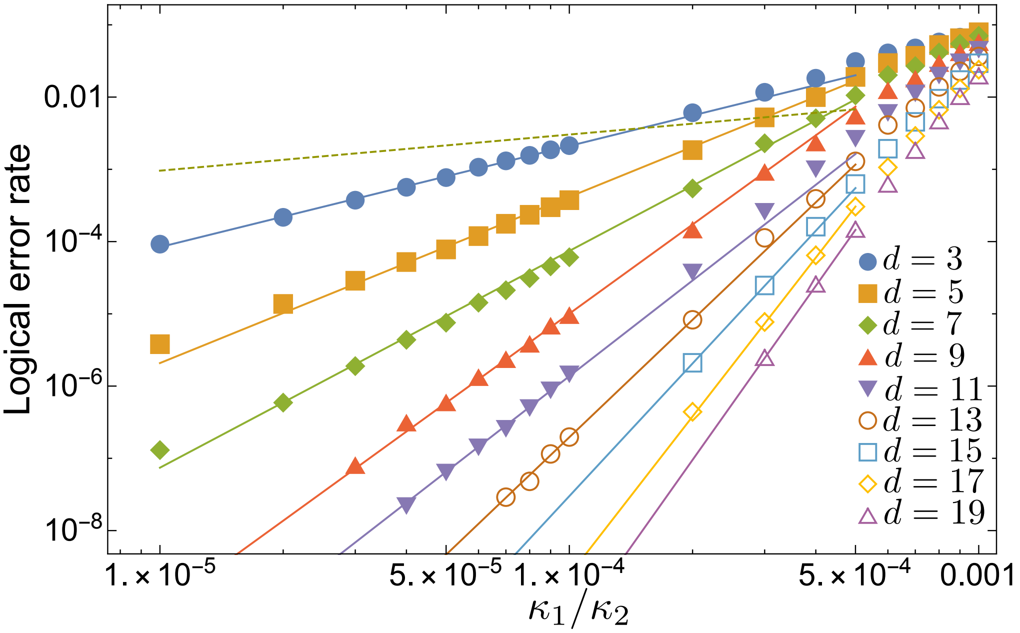


FIG. 6. Logical Z failure rates for the repetition code, for a variety of values of the code distance d. We use the circuit-level noise model described in Section III with $\kappa_\phi = 0$ and $n_{th} = 0$. The X-basis measurement error rates are obtained from Table III with three parity measurements. The number of syndrome measurement rounds r for each distance is obtained using the STOP algorithm described in Appendix H. The dashed green line is used as a stand-in for comparison with the logical memory error rates and corresponds to the function $0.3025\sqrt{\kappa_1/\kappa_2}$ which is a quarter of the total Z failure rate of a physical CNOT gate (see Table II).

ment of the data qubits can be viewed as a round of perfect error correction since measurement errors in such a process are equivalent to data qubit errors occurring immediately prior to the measurement of the data. However for our purposes, the repetition code will be used during the preparation of $|\mathrm{TOF}\rangle$ magic states where the circuits used in the preparation protocol contain non-Clifford gate locations (see Fig. 15). Prior to the application of these non-Clifford gates, errors on the encoded code-blocks need to be corrected without having access to a round of perfect syndrome measurements (since the data qubits cannot be measured directly prior to applying the non-Clifford gates). Hence, it is important to have a decoder which is robust to measurement errors occurring in the last round when rounds of perfect syndrome measurements cannot effectively be applied in the hardware. A solution is that instead of repeating the syndrome measurement d times, one can repeat the syndrome r times where r is computed using the STOP algorithm mentioned above. Note that in this case, r is not fixed but instead is a function of the observed syndrome history. For all logical Z failure rates plotted in Fig. 6, the simulations were performed using the STOP algorithm for determining when to stop measuring the error syndrome. To ensure projection onto the codespace, we add 1 round of ideal syndrome measurement after the last round given by the STOP algorithm and implement MWPM over the full syndrome history.

Remark two: The x-axis in Fig. 6 is plotted as a function of κ_1/κ_2. It is important to note that some components of the hardware fail with probabilities proportional to κ_1/κ_2 whereas other components (such as the CNOT gates) fails with probabilities proportional to $\sqrt{\kappa_1/\kappa_2}$ (see Tables I and II). In particular, for regimes where $\kappa_1/\kappa_2 \sim 10^{-5}$, the noise is dominated by CNOT gates, whereas in the regime where $\kappa_1/\kappa_2 \sim 10^{-3}$, some idling locations have noise rates comparable with the CNOT failure rates, hence changing the slope of the logical failure rate curves. To be clear, in our simulations we took into account all different types of idling locations; for this reason, and also because we use the STOP algorithm for determining the number of syndrome measurement rounds instead of repeating a fixed d times, our numerics should not be directly compared with previous works such as in [14]. Note further that for comparisons with other works (such as in [14]), the x-axis of our plots would need to be re-scaled as a function of $\sqrt{\kappa_1/\kappa_2}$.

Given two strips of neighboring repetition codes, a logical CNOT gate can be implemented transversally between the two strips, and the failure probability of such a gate is approximately four times the values showed in Fig. 6. One possible interesting quantum error-correction experiment would be to demonstrate a logical CNOT gate with lower failure probability compared to a physical CNOT gate. As such, in Fig. 6, we also plotted a dashed green curve which corresponds to the function $0.3025\sqrt{\kappa_1/\kappa_2}$ which is a quarter of the total Z failure rate of a CNOT gate (see Table II). As can be seen, for $\kappa_1/\kappa_2 < 4.5 \times 10^{-4}$, the probability of failure of a CNOT gate encoded in a

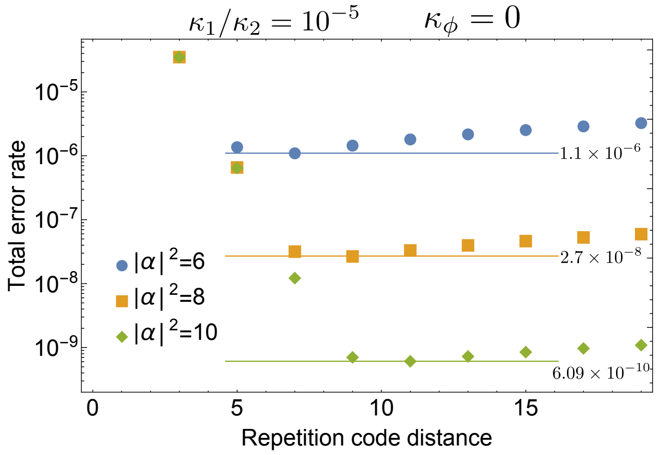


FIG. 7. Total logical failure rate per code cycle for various repetition code distances and values of $|\alpha|^2$ with fixed $\kappa_1/\kappa_2 = 10^{-5}$, $\kappa_\phi = 0$ and $n_{th} = 0$. The logical X and Y failure rates were computed analytically (to leading order) using the noise model presented in Section III while taking into account all malignant and benign fault locations. For $|\alpha|^2 = 8$, the lowest achievable total logical error rate is 2.7×10^{-8} per code cycle using $d = 9$.

$d = 7$ repetition code is lower than that of a physical CNOT gate. From the hardware analysis, we find that $\kappa_2/(2\pi) = 500$kHz (or $\kappa_2 = 3.14 \times 10^6 s^{-1}$) is achievable for $|\alpha|^2 = 8$ (see Section II and Appendix A 5 for more details). In this case, $\kappa_1/\kappa_2 = 10^{-4}$ corresponds to a lifetime of 3ms. From Fig. 6, a logical CNOT gate implemented transversally with two $d = 9$ repetition code strips fails with probability 3.7×10^{-5} which would correspond to the highest CNOT fidelities achieved to date. Furthermore, we find numerically that the general polynomial describing the logical Z failure rate of a distance d repetition code for d rounds of syndrome measurements is given by

$$p_L^{(Z)}(d) = 0.014d \left(770\frac{\kappa_1}{\kappa_2}\right)^{0.41d}. \tag{73}$$

The justifications for the chosen scaling of $p_L^{(Z)}(d)$ and the scaling of the logical failure rates for the rotated surface code in Section IV B are given in Appendix L.

Lastly, in Fig. 7 we compute the total logical failure rate per code cycle (which include contributions from logical X and Y failures) of the repetition code for distances in the range $3 \leq d \leq 19$ with fixed $\kappa_1/\kappa_2 = 10^{-5}$, $\kappa_\phi = 0$ and $n_{th} = 0$. For $|\alpha|^2 = 8$, it can be seen that above $d = 9$, contributions from bit-flip errors are the dominant factor in the total logical failure rate. As such, going to larger repetition code distances results in higher logical failure rates. Such features demonstrate the importance of taking into account contributions from bit-flip errors, even though they are exponentially suppressed. Further, such results demonstrate that the logical X error rate when implementing a logical Toffoli gate using the piecewise fault-tolerant construction of [13, 14] would be too high for the algorithms considered in Section VIII.

B. Rotated surface code logical failure rates

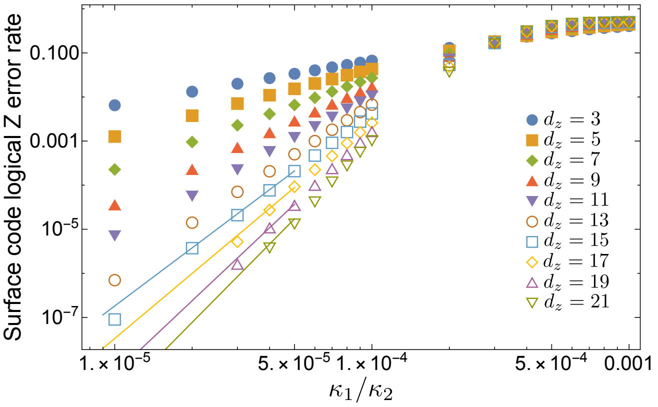


FIG. 8. Logical Z failure rates for the rotated surface code with $d_x = 3$ and varying d_z. We use the circuit-level noise model described in Section III with $\kappa_\phi = 0$ and $n_{th} = 0$. The X-basis measurement error rates are obtained from Table III with five parity measurements. We point out that $\kappa_1/\kappa_2 = 10^{-5}$, $\kappa_1/\kappa_2 = 10^{-4}$ and $\kappa_1/\kappa_2 = 10^{-3}$ correspond to CNOT failure rates of 3.8×10^{-3}, 1.2×10^{-2} and 3.8×10^{-2}. The simulations were done by performing d_z rounds of noisy syndrome measurements followed by one round of perfect syndrome measurement.

Using the circuit-level noise model described in Tables I to III and Eq. (71), the logical Z failure rates for the rotated surface code with $d_x = 3$ and varying d_z are given in Fig. 8. Note that the logical X operator has minimum support on d_x qubits along each column of the lattice. The logical Z operator has minimum support on d_z qubits along each row of the lattice. Contrary to our repetition code simulation methodology, the simulation results were obtained by performing d_z rounds of noisy syndrome measurements followed by one round of perfect syndrome measurement in order to guarantee projection onto the code-space. Throughout this paper, we use the surface code with only a fixed number of error correction rounds. Furthermore, in our proposal we never perform non-Clifford gates directly on surface-code patches, rather non-Clifford gates are always achieved by gate injection of a magic state. As such, all simulation are performed for fixed d rounds followed by 1 ideal round to project onto the codespace.

As can be seen from Fig. 8, in order to obtain low logical Z failure rates without requiring a very large d_z distance (say $d_z > 40$), it is required that $\kappa_1/\kappa_2 \leq 5 \times 10^{-5}$. Put another way, the total CNOT gate Z failure rate should be less than 7.6×10^{-3} to achieve very low logical failure rates with reasonably small surface code distances.

Comparing the results of Figs. 6 and 8, one sees that the surface code significantly under-performs the repetition code. This is mainly due to the fact that a distance-d repetition code requires a total of $2d-1$ data and ancilla qubits compared to the rotated surface code which requires $2d_xd_z-1$ data and ancilla qubits. Further, the surface code requires weight-four stabilizer measurements compared to weight-two stabilizers for the repetition code and thus the syndrome measurement circuit is deeper.

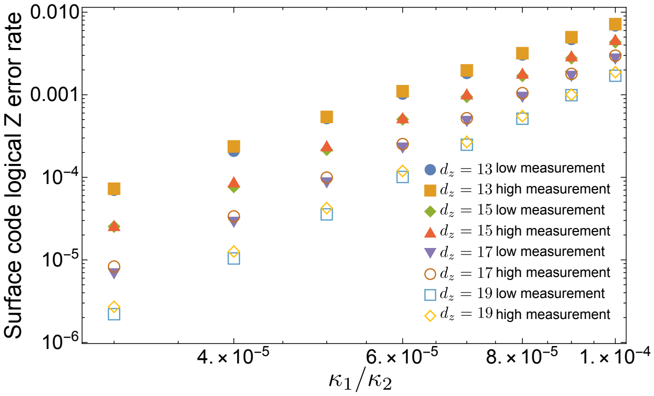


FIG. 9. Logical Z error rates of the thin rotated surface code with five modes coupled to an ATS (which includes a transmon qubit and an additional readout mode in each unit cell and corresponds to the data in Fig. 8) and four modes coupled to an ATS (which excludes the transmon qubit and performs direct X-basis measurements). We labeled measurements with the transmon qubit as *low measurement* and the direct X-basis measurement as *high measurement*. For the direct X-basis measurement, the measurement error rate is fixed at 2×10^{-3} for all values of κ_1/κ_2. Measurement error rates with the transmon qubit were obtained from Table III with five parity measurements. The simulations were performed by setting $\kappa_\phi=0$ and $n_{th}=0$.

The logical Z and X failure rate polynomials for fixed $d_x=3$ and arbitrary d_z distances (with d_z rounds of stabilizer measurements) were found numerically to be given by

$$p_L^{(Z)}(d_z)=0.028d_z\left(3559\frac{\kappa_1}{\kappa_2}\right)^{0.292d_z}, \tag{74}$$

$$p_L^{(X)}(d_z)=3449d_z^2e^{-4|\alpha|^2}\left(\frac{\kappa_1}{\kappa_2}\right). \tag{75}$$

See Appendix L for further details on the fitting procedure and additional results on errors during lattice surgery.

For the algorithms considered in Section VIII, we require $p_L^{(X)}(d_z)\leq10^{-10}$. As can be seen from Eq. (75), if $|\alpha|^2=6$, $\kappa_1/\kappa_2=10^{-5}$ (which requires $d_z=31$), the logical X error rate is approximately 1.3×10^{-9} which is an order of magnitude worse than the minimum requirements. However, setting $|\alpha|^2=8$, we obtain $p_L^{(X)}(d_z)=4.2\times10^{-13}$. We conclude that for the algorithms considered in Section VIII, we require $|\alpha|^2\geq8$.

In Fig. 9, we plot the logical Z error rate of the thin rotated surface code where five modes are coupled to an ATS (which uses a transmon qubit and additional readout mode to implement X-basis measurement) in addition to the case where only four modes are coupled to each ATS (where the X-basis measurement is implemented directly through a buffer mode). For the case of four modes coupled to the ATS (labeled *high measurement*), we fixed the measurement error rate to 2×10^{-3}. Both schemes are discussed in Section III G. For the case with five modes coupled to the ATS, we used the same data as shown in Fig. 8. As can be seen, even though the measurement error rate can be more than an order of magnitude larger when only four modes are coupled to the ATS, the logical Z error rates increase by a small amount in the low κ_1/κ_2 regime. The reason the logical failure rate is not greatly affected by the large increase in measurement failure rates is that CNOT failures are the dominant source of noise. As such, we do not expect the overhead results of Section VIII to increase when using an architecture with four modes coupled to an ATS given that the same code distances can be used for implementing the algorithms of interest.

C. Surface code logical failure rates in the presence of crosstalk errors

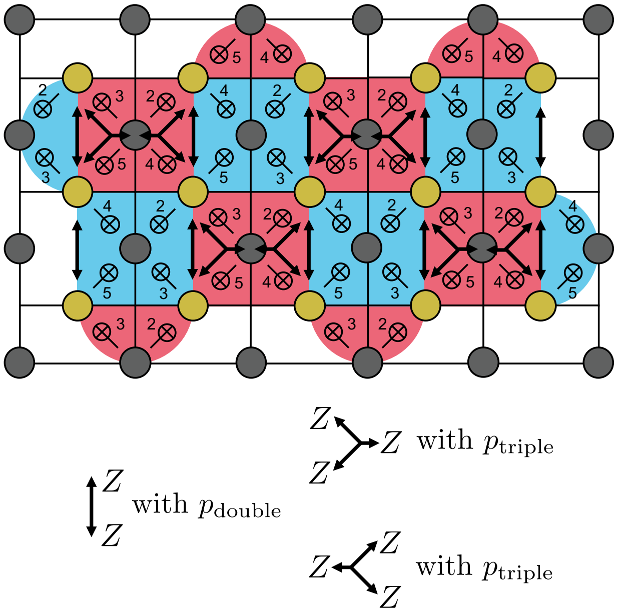
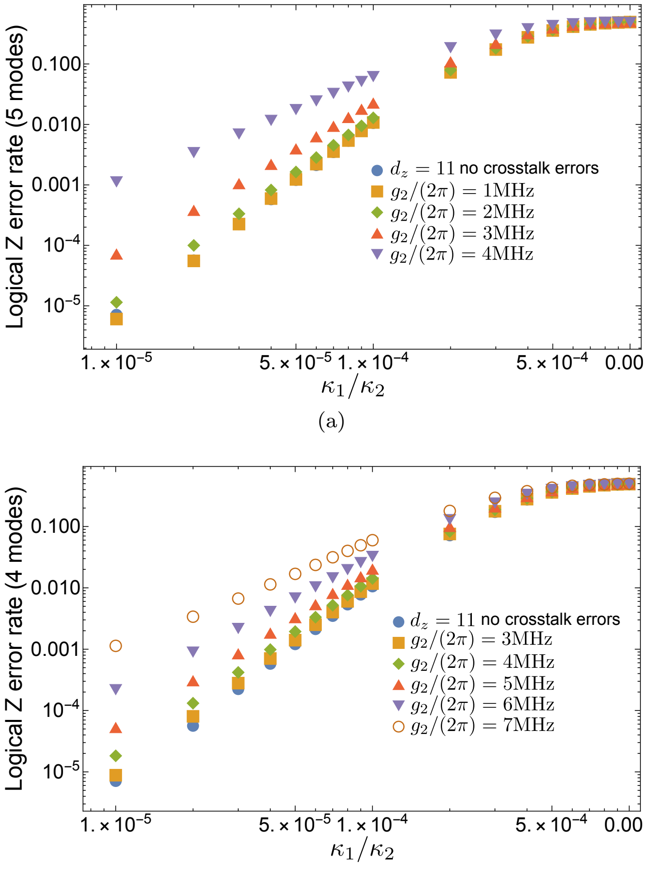


FIG. 10. Crosstalk errors due to multiplexed stabilization. Phononic modes that are connected via a shared ATS experience correlated Pauli Z errors due to micro-oscillation (see Appendix B 5). Every pair of two data qubits that are shared by the same ATS (hence aligned vertically) undergoes a correlated Z error with a probability p_{double}. Also, every triple of two data qubits and an ancilla qubit that measures an X-type surface-code stabilizer undergoes a correlated Z error with a probability p_{triple}, where the Z error on the ancilla qubit manifests as a flipped outcome of the corresponding X-type stabilizer.



(a)

(b)

FIG. 11. (a) Logical Z failure rates for a $d_x = 3$ and $d_z = 11$ thin surface code in the presence of the residual crosstalk errors (given in Eq. (76)) arising from the coherent micro-oscillations and the circuit-level noise model of Section III with $\kappa_\phi = 0$ and $n_{th} = 0$. The X-basis measurement error rates are obtained from Table III with five parity measurements. We compute the logical Z error rates for different values of g_2 shown in the legend, and compare such results to the case where the crosstalk errors are not present. Each ATS has five modes coupled to it as in Fig. 2. (b) Same noise model as in (a) except that the crosstalk error rates are given in Eq. (77) (i.e., where measurements are performed using an ATS via an optomechanical coupling) .

Recall that in our architecture proposal, each ATS stabilizes multiple phononic modes. Since the ATS mediates various spurious interactions as well as desired interactions, phononic modes that are connected by the same ATS undergo crosstalk errors. While stochastic crosstalk errors can be strongly suppressed by filtering and careful choice of the frequencies of the phononic modes (see Appendix B 4), coherent micro-oscillation errors cannot be eliminated by the filters (see Appendix B 5). In particular, such residual crosstalk errors result in two non-trivial noise processes: every pair of data qubits that are connected by a shared ATS (hence aligned vertically) experiences a correlated Z error with probability p_{double} and every triple of two data qubits and an ancilla qubit that measures an X stabilizer of the surface code experiences a correlated Z error with probability p_{double}. In particular, the Z error on the ancilla qubit is realized in the form of a flipped measurement outcome of the corresponding X-type stabilizer (see Fig. 10).

In Appendix B 5, we optimize the frequencies of the five phononic modes coupled to a shared ATS to minimize p_{double} and p_{double}, assuming that the maximum frequency difference between different phononic modes is $2\pi\times 1\text{GHz}$. With the optimal choice of phononic mode frequencies, we find that the correlated error rates p_{double} and p_{double} are given by

$$\begin{aligned} p_{\text{double}} &= 1.829\times 10^{-8}|\alpha|^8\Big(\frac{g_2/(2\pi)}{1\text{MHz}}\Big)^4, \\ p_{\text{triple}} &= 5.205\times 10^{-10}|\alpha|^8\Big(\frac{g_2/(2\pi)}{1\text{MHz}}\Big)^4. \end{aligned} \tag{76}$$

Here, g_2 is the strength of the desired interaction $\hat{a}^2\hat{b}^\dagger$ needed for the engineered two-phonon dissipation. See Appendix B 5 for more details on why p_{double} and p_{double} scale as g_2^4. Note that p_{triple} is 35 times smaller than p_{double}. For $g_2/(2\pi) = 1\text{MHz}$ and $|\alpha|^2 = 8$, p_{double} is given by $p_{\text{double}} = 7.5\times 10^{-5}$, which is negligible compared to the total error rate of the physical CNOT gate between two cat qubits. However, since p_{double} scales as $p_{\text{double}} \propto g_2^4$, it increases rapidly as we use larger coupling strength. For instance, p_{double} is given by $p_{\text{double}} = 1.2\times 10^{-3}$ at $g_2/(2\pi) = 2\text{MHz}$ and $p_{\text{double}} = 1.9\times 10^{-2}$ at $g_2/(2\pi) = 4\text{MHz}$.

If the X readout of cat qubits is performed directly by using an ATS via an optomechanical coupling, the fifth readout mode (green mode in Fig. 2) is not needed and thus we can work with an architecture where each ATS is coupled to four phononic modes. In this case, we find that there is a frequency arrangement that yields (see Appendix B 5)

$$\begin{aligned} p_{\text{double}} &= 1.218\times 10^{-9}|\alpha|^8\Big(\frac{g_2/(2\pi)}{1\text{MHz}}\Big)^4, \\ p_{\text{triple}} &= 3.866\times 10^{-10}|\alpha|^8\Big(\frac{g_2/(2\pi)}{1\text{MHz}}\Big)^4. \end{aligned} \tag{77}$$

In particular, p_{double} is reduced by an order of magnitude by removing the fifth readout phononic mode. Thus, in the 4-modes-per-ATS setting, we have $p_{\text{double}} = 0.003$ at $g_2/(2\pi) = 5\text{MHz}$ and $|\alpha|^2 = 8$.

In Fig. 11a we provide logical Z failure rates of the thin surface code under the presence of the crosstalk errors described above for various values of g_2, where five modes are coupled to each ATS as in Fig. 2. We note that in the presence of crosstalk errors with probabilities p_{double} and p_{triple} (which are given in Eq. (76)), extra edges need to be added to the matching graphs of the surface code. Details of the modified graphs in addition to the edge weight calculations are provided in Appendix M 3. In Fig. 11b, we consider logical Z errors in the presence of crosstalk where the X readout of the cat qubits is performed directly using an optomechanical coupling (so that the crosstalk error probabilities are given in Eq. (77)).

As can be seen from Fig. 11a, when $g_2/(2\pi) = 1$MHz, the effects from crosstalk errors are negligible (the logical error rate curves with and without crosstalk almost perfectly overlap). When $g_2/(2\pi) = 2$MHz, the effects are very small. However, if $g_2/(2\pi) \geq 3$MHz, the difference between logical Z error rates of the surface code with and without crosstalk errors is large enough such that one would need to use larger code distances to achieve the target logical failure rates for the algorithms considered in Section VIII. Hence, to maintain the overhead results obtained in Section VIII, it would be preferable to use values of $g_2/(2\pi) \leq 2$MHz since in such a case, effects from crosstalk errors are very small.

Lastly, when only 4 modes are coupled to the ATS, the results from Fig. 11b indicate that $g_2/(2\pi)$ can go up to 4MHz before effects from crosstalk become large enough such that one would need to use larger surface codes to achieve logical failure rates for the algorithms considered in Section VIII. As shown in Appendix A 5, the maximum achievable g_2 is fundamentally limited by $4J/(10\alpha)$ due to the filter design and validity of adibatic elimination. Here, $4J$ is the filter bandwidth which is given by $2\pi \times 100$MHz for 5 modes per ATS and $2\pi \times 180$MHz for 4 modes per ATS under the optimal choice of phononic mode frequencies. Hence, the maximum achievable $g_2/(2\pi)$ set by the filter design is given by 3.53MHz and (6.36MHz) for the setting with 5 and (4) modes per ATS. On the other hand, the crosstalk errors limit $g_2/(2\pi)$ to be bounded below 2MHz (4MHz). Thus, the crosstalk errors are currently the most limiting factor and need to be further suppressed. See Section II for a discussion of possible ways to further minimize the crosstalk errors.

V. COMPUTATION BY LATTICE SURGERY AND TIMELIKE ERRORS

In both repetition and surface codes, the logical CNOT gate is transversal, which means $\text{CNOT}_L = \text{CNOT}^{\otimes n}$. Therefore, a logical CNOT can be fault-tolerantly implemented whenever the hardware supports physical CNOTs between corresponding qubits in the code blocks. For the repetition code, we can realize a transversal CNOT gate in a 2D layout between two repetition codes. However, for the surface code, a logical CNOT cannot be realized in a way that is both transversal and uses physical CNOT gates in a 2D hardware geometry. A well known solution is to use lattice-surgery between blocks of surface codes [63–66]. The simplest example of lattice surgery realizes a logical $X_L \otimes X_L$ or $Z_L \otimes Z_L$ measurement between two surface code patches separated by a distance ℓ. The two code blocks are merged into a single code block for d_m rounds of surface code stabilizer measurement and then split apart. We illustrate this in Fig. 12 with more fine-grained details in Fig. 50 of Appendix L.

During lattice surgery, certain types of logical errors can occur resulting in the wrong measurement outcome of multi-qubit logical Pauli operators. We call these timelike errors since in the spacetime picture they correspond to strings of errors in the time direction (see Fig. 12). As shown in Appendix L, such logical failure modes are exponentially suppressed by increasing d_m, which comes at the price of increasing the execution time for this logical operation. A seemingly natural choice is to set $d_m = d_x = d_z$, but since our noise model is highly biased this leads to an asymmetry in the optimal choices. We discuss timelike errors in more detail in Appendix L and present simulation results showing that for our noise model, the rate of timelike errors is comparable (even slightly lower) than logical Z error rates. A detailed decoding scheme used for such simulations is described in Appendix M 4.

Lattice surgery measurements combined with logical $|0\rangle$ and $|+\rangle$ preparations, and logical single-qubit X and Z measurements, can be used to perform logical CNOT, Hadamard and CZ gates [63]. Furthermore, the two codeblock lattice surgery sketched in Fig. 12 can be generalized to act on multiple codeblocks to enable measurements of any tensor product of Z and X operators. By making use of lattice twists and dislocations, any logical multi-qubit Pauli operator can be measured by lattice surgery [67].

However, all these operations are either Clifford group gates or Pauli measurements, so some non-Clifford operation is required to complete a universal gate set. The model of Pauli-based computation [68] shows that it is possible to perform universal quantum computation using just Pauli measurements and access to suitable magic states and performing *magic state injection.* We denote the magic state for simulating Toffoli gates as

$$|\text{TOF}\rangle = \frac{1}{2} \sum_{a,b \in \mathbb{F}_2} |a\rangle|b\rangle|a \wedge b\rangle, \tag{78}$$

where $a \wedge b$ is the AND of bits a and b. The $|\text{TOF}\rangle$ state is stabilized by the Abelian group $\mathcal{S}_{\text{TOF}} = \langle g_A, g_B, g_C \rangle$ where

$$g_A = X_A \text{CNOT}_{B,C}, \tag{79}$$
$$g_B = X_B \text{CNOT}_{A,C}, \tag{80}$$
$$g_C = Z_C CZ_{A,B}. \tag{81}$$

To simplify the notation used in Section VII, we label the three qubits involved in a Toffoli gate by A, B and C instead of 1, 2 and 3. Given one copy of a $|\text{TOF}\rangle$ state, magic state injection is performed using the circuit in Fig. 13 to realize a logical Toffoli gate. Notice that the circuit requires a Clifford correction $g_A^a g_B^b g_C^b$ for the binary measurement outcome (a, b, c) of the single qubit Pauli measurements.

In a purely Pauli-based computation, rather than using lattice surgery to simulate the CNOT circuit for magic state injection, the CNOTs can be completely eliminated using the circuit identities shown in Fig. 13. Furthermore, the Clifford corrections and Clifford gates in an algorithm do not necessarily need to be performed. Rather we can

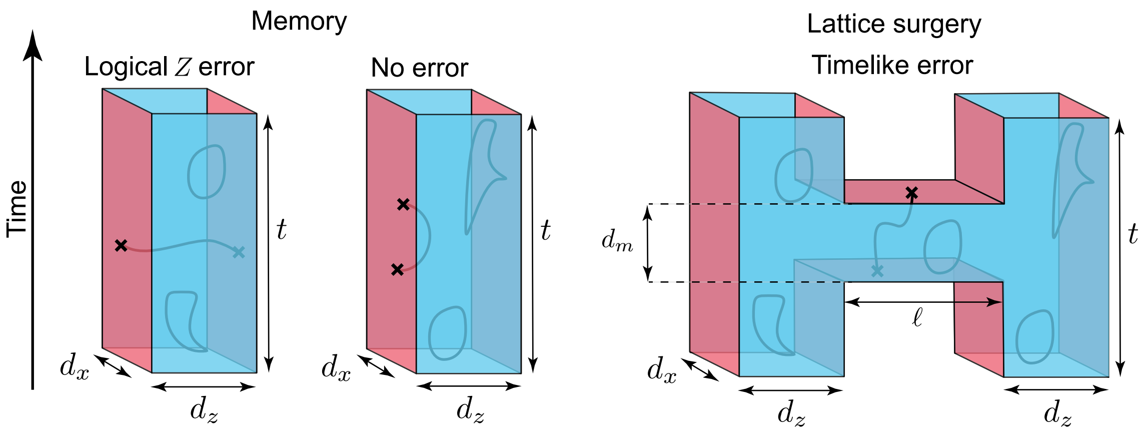


FIG. 12. A spacetime diagram of memory and lattice surgery processes using a thin, rotated surface code with boundaries. Pink (the left and right sides) represents boundaries where Z strings can terminate. Blue (the fore and rear sides) represents boundaries where X strings can terminate. We show examples of Z strings: when traveling in a spatial direction they represent physical Z errors and when traveling in the vertical time direction, they represent errors on X stabilizer measurements. We only show Z strings that are closed loops or terminate on suitable boundaries. These can be regarded as the final Z strings (after matching) including physical/measurement errors combined with recovery operations. In the case of memory, a logical Z error occurs whenever a Z string propagates between two topologically disconnected red boundaries. When performing lattice surgery to measure the $X_{L1} \otimes X_{L2}$ logical operator between two patches, an additional failure mechanism is possible. If a Z string propagates between two red boundaries disconnected in the time direction then we have a timelike Z-error. Computationally, this flips the outcome of the $X_{L1} \otimes X_{L2}$ measurement. Such processes are exponentially suppressed by increasing the measurement distance d_m.

keep a record of the accumulated Clifford gates so far into a Clifford *frame* (see for instance Ref. [61]). When we need to measure a Pauli P, we instead measure the Pauli $CPC^\dagger$ whenever the Clifford frame records C. In such a Pauli-based computational model, Clifford gates do not contribute to an algorithms runtime. Rather the runtime is determined by two factors: how fast we can prepare high fidelity TOF states; and how fast they can be injected into the algorithm. The rate of injection depends on how much routing space between qubits is budgeted for in the device. Using a fast data access structure [66], it is known that lattice surgery can perform a single arbitrary multi-qubit Pauli operator with approximately $\sim 2\times$ overhead in routing costs. Such a space overhead cost is pessimistic since not all qubits need to be involved in every lattice surgery operation, so considerable compression is possible. Ref. [69] assumed a $\sim 1.5\times$ overhead suffices and Refs. [2, 6, 70] assumed this cost could be made negligible. In our later analysis of overheads, we assume a $\sim 1.3\times$ routing overhead cost suffices to maintain this pace of injection, though more study is needed to better quantify this important trade-off.

One can also inject at a considerably faster pace than sequentially injecting magic states, up-to the limit of time-optimal quantum computation [71], though this approach incurs significantly higher routing overhead costs and is not practical for modest size quantum computers. In the next two sections, we consider the pace and fidelity with which we can prepare TOF magic states. In what follows, we use $|\mathrm{TOF}\rangle$ and TOF interchangeably when refering to the state in Eq. (78).

VI. TOFFOLI DISTILLATION: BOTTOM-UP SCHEME

In magic state distillation schemes, the goal is to distill magic states with circuits that require only stabilizer operations [30, 31, 33]. The circuits used to distill such magic states are typically not fault-tolerant to all Clifford gate errors and thus must be implemented using a sufficiently large error-correcting code. Recently, with the advent of flag qubits and redundant ancilla encoding, scalable approaches to fault-tolerantly preparing magic states have been devised such that all stabilizer operations can be implemented directly at the physical level [72, 73]. We refer to such methods as a bottom-up approach to preparing magic states.

In this section, we provide a protocol to fault-tolerantly prepare TOF magic states encoded in the repetition code using a bottom-up approach (herein `BUTOF`). In Section VII, we show how the scheme presented in this section can be supplemented by using a top-down approach to prepare TOF states with the very high fidelities required to implement the algorithms considered in Section VIII.

We now describe how to fault-tolerantly prepare the $|\mathrm{TOF}\rangle$ state. First, note that the state $|\psi_1\rangle = \frac{1}{\sqrt{2}}(|100\rangle + |111\rangle)$ is stabilized by g_B and g_C. Such a state can straightforwardly be prepared using the circuit in the dashed blue box of Fig. 14. In what follows, physical Toffoli gates will

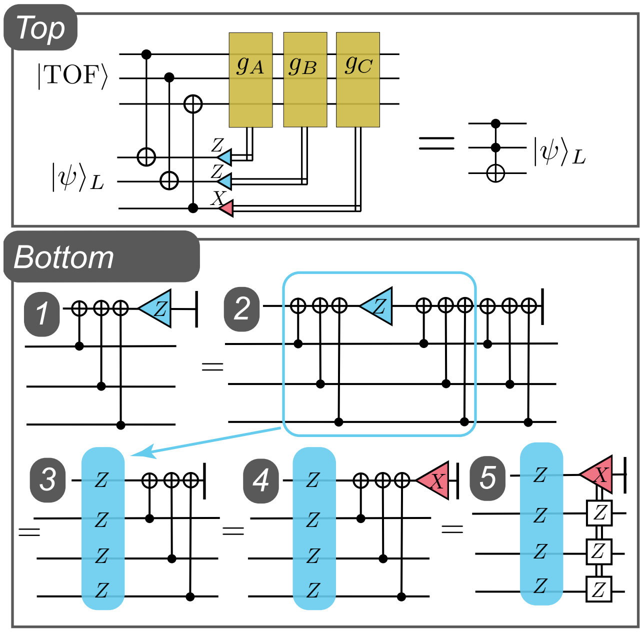


FIG. 13. (*Top*): Circuit for implementing a Toffoli gate given a $|\mathrm{TOF}\rangle$ resource state. The Clifford corrections depend on the three measurement outcomes and are given in Eqs. (79) to (81). All qubits and gates are implemented at the logical level. (*Bottom*): Five equivalent circuits showing how to convert from a CNOT followed by measurements with m CNOT gates into a Pauli-based computation that can be realized by lattice surgery. Circuit 1 to 2: we insert the identity. Circuit 2 to 3: we have replaced the highlighted box with a multi-qubit $Z^{\otimes m+1}$ measurement. Circuit 3 to 4: we add a single qubit X measurement before we discard the qubits. Circuit 4 to 5: we use the X measurement to replace the CNOT gates with classically controlled Z gates. A similar identity holds with the CNOT direction reversed and the roles of X and Z interchanged. Applying the identities of the bottom figure in the $m = 1$ case to the top figure yields a Pauli-based magic state injection procedure. We make use of the the $m > 1$ case in Section VII and Appendix N 4 b.

need to be applied between ancilla qubits and $|\psi_1\rangle$ prior to measuring the data. As such it is very important that the states $|0\rangle_L$ and $|1\rangle_L$ in the circuit of Fig. 14 (which are encoded in the repetition code) be prepared using the STOP algorithm since otherwise measurement errors in the last ancilla measurement round could lead to logical failures [74]. An alternative to avoid using the STOP algorithm would be to prepare $|0\rangle_L$ and $|1\rangle_L$ using post selection. However, such an approach would reduce the acceptance probability of our scheme (see below) thus increasing its space-time overhead cost. Once $|+\rangle_L = |+\rangle^{\otimes n}$, $|1\rangle_L$ and $|0\rangle_L$ have been prepared, the CNOT gate in the dashed blue box of Fig. 14 is applied transversally.

Now, given a copy of $|\psi_1\rangle$, we can prepare $|\mathrm{TOF}\rangle$ by measuring g_A using the circuit in the dashed red box of Fig. 14 resulting in the state $|\psi\rangle_{\mathrm{out}}$. If the measurement outcome is $+1$, then $|\psi\rangle_{\mathrm{out}} = |\mathrm{TOF}\rangle$, and if it is -1, then $|\psi\rangle_{\mathrm{out}} = Z_A|\mathrm{TOF}\rangle$. Hence we apply a Z_A correction given a -1 measurement outcome. Note that neither error detection nor error correction is applied to any of the data blocks at this stage. The reason is that it is not necessary for ensuring the fault-tolerance of our scheme. Further, we found numerically that adding error correction at this stage results in higher logical failure rates when preparing $|\mathrm{TOF}\rangle$. Furthermore, adding unnecessary error detection units would lower the acceptance probability of our scheme. We provide a more detailed implementation of the controlled-g_A gate in Fig. 15 below.

A measurement error on the ancilla results in a logical Z_A failure and so the measurement of g_A needs to be repeated (similar repetitions are needed for the preparation of logical computational basis states, see Appendix I). This can be done using the STOP algorithm. However, due to the increasing circuit depth with increasing repetition code distance in addition to the high cost of the controlled-g_A gate, such a scheme does not have a threshold and results in relatively high logical failure rates. As in Refs. [72, 73], an alternative approach is to use an error detection scheme by repeating the measurement of g_A exactly $(d-1)/2$ times for a distance d repetition code. In between each measurement of g_A, one round of error detection is applied to the data qubits by measuring the stabilizers of the repetition code (see Fig. 14). If any of the measurement outcomes are non-trivial, the BUTOF protocol is aborted and reinitialized. In Fig. 15a, we provide an example of the two-dimensional layout and sequence of operations for measuring g_A which is compatible with our ATS architecture for a distance-5 repetition code. To realize the protocol with local operations, we replace the $|+\rangle$ ancilla in Fig. 14 with 5 qubits that we prepare in a GHZ state. Subsequently, the required Toffoli and CNOT gates are applied, followed by a disentangling of the GHZ states and measurement of the $|+\rangle$ state ancilla. The equivalent circuit implementing the g_A measurement for a $d = 5$ repetition code is shown in Fig. 15b.

As a remark, we point out that in general, it is possible to use one fewer ancilla in the circuit of Fig. 15a with a lattice that is no longer translationally invariant with respect to yellow and gray vertices. However, such a layout could not straightforwardly be used with our lattice surgery implementation of Appendix N.

In Fig. 16a, we provide the total Z failure probability of our BUTOF protocol for various repetition code distances ranging from $d = 3$ to $d = 9$. We note that given the increasing circuit depth of BUTOF with the repetition code distance d, our scheme does not have a threshold even though it is fault-tolerant. Further, as can be seen from Fig. 16b, the acceptance probability for preparing such states (i.e. the probability that all measurement outcomes in Fig. 14 are trivial) decreases exponentially with increasing code distances. Hence, large repetition code distances should be avoided. However in the regime where $\kappa_1/\kappa_2 \approx 10^{-5}$, we can still obtain $|\mathrm{TOF}\rangle$ states with total failure probabilities on the order of 6×10^{-6}, which is orders of magnitude better than the failure probabilities that would be obtained by preparing $|\mathrm{TOF}\rangle$ states using

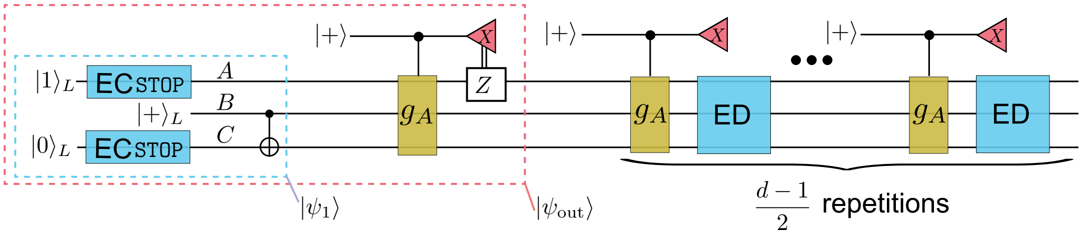


FIG. 14. Circuit for our entire BUTOF protocol. The first step (shown in the dashed blue box) consists of preparing the state $|\psi_1\rangle = \frac{1}{\sqrt{2}}((|100\rangle + |111\rangle)$. The preparation of the states $|0\rangle_L$ and $|1\rangle_L$ are described in Appendix I1. The next step consists of measuring $g_A = X_A \mathrm{CNOT}_{B,C}$. If the measurement outcome on the ancilla is -1, a Z_A correction is applied to the output state. Note that at this stage, error correction is not applied to the data block. The first two steps are enclosed within the dashed red box. We label the output state of the first two steps as $|\psi_{\rm out}\rangle$. Lastly, the measurement of g_A is repeated $(d-1)/2$ times for a distance d repetition code. The ED blocks correspond to one round of stabilizer measurements of the repetition code. If any of the measurement outcomes of ED or ancillas are non-trivial, the protocol is aborted and begins anew.

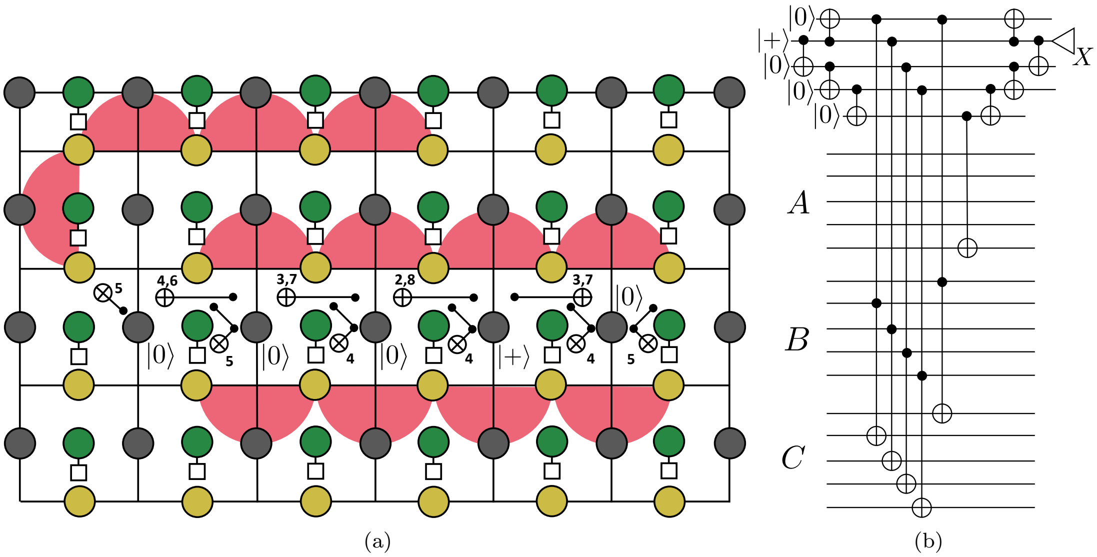
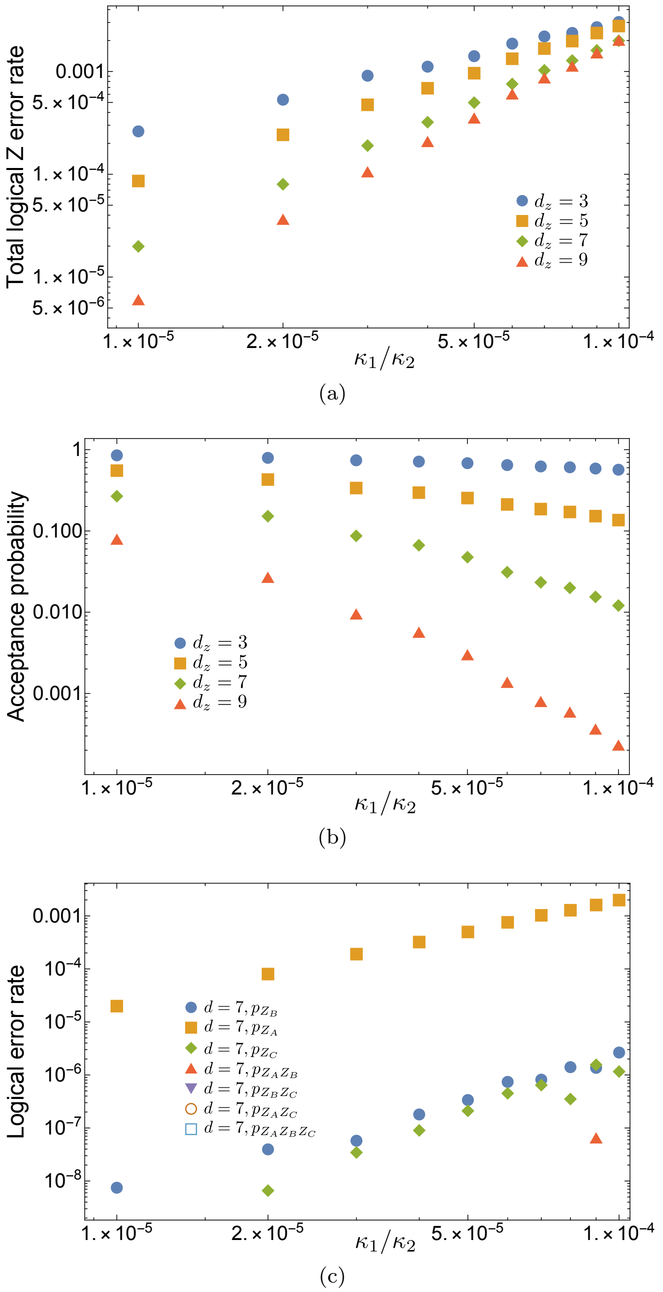


FIG. 15. (a) Implementation of the g_A measurement (for a distance $d = 5$ repetition code) compatible with our ATS layout and lattice surgery implementation for universal quantum computation described in Section VII. All operations are performed respecting the connectivity constraints of the ATS's and use the fewest possible ancilla qubits for preparing the GHZ state necessary for the fault-tolerant measurement of g_A. (b) Equivalent circuit for the implementation of (a).

non-fault-tolerant methods. This drastically reduces the overhead requirements of the top-down approach of Section VII. Also, as can be seen from Fig. 16c, logical Z errors are highly concentrated on block A. The reason is that while the error detection units on each block can detect up to $d-1$ physical Z errors, $(d-1)/2$ measurement errors on the GHZ ancilla will lead to a logical Z error on block A.

We note that the GHZ circuit in Fig. 15b, which is used to measure g_A, is not fault-tolerant to X or Y errors [75]. However, since we are assuming that X and Y errors are exponentially suppressed, flag qubits for detecting X-type error propagation are unnecessary as long as X or Y error rates multiplied by the total number of fault locations are below the target levels for algorithms of interest. Indeed as is shown in Sections VII and VIII and for the parameters chosen in this work, X error rates are low enough such that the desired failure rates can be achieved for implementing the quantum algorithms with over a million Toffoli gates (see Table V).

Lastly, we note that simulating the circuit in Fig. 15 can be challenging given the presence of physical Toffoli gates. In Appendix K, we provide a method for performing a near exact simulation of such circuits (the simulation is exact if there are less than d Z-type errors on block C prior to applying the physical Toffoli gates). Also, when using the STOP algorithm to simulate the preparation of $|0\rangle_L$ and $|1\rangle_L$ prior to applying the physical Toffoli gates, we do not add one round of perfect error correction (since projecting to the codespace is not necessary at this stage). Residual errors at the output of the preparation of $|0\rangle_L$ and $|1\rangle_L$ using the STOP algorithm are propagated to the



(a)

(b)

(c)

FIG. 16. (a) Total logical Z failure rate for preparing a $|\mathrm{TOF}\rangle$ state using the fault-tolerant BUTOF protocol described in this section. (b) Acceptance probabilities for preparing $|\mathrm{TOF}\rangle$ states using the fault-tolerant protocols described in this section. (c) Decomposition of the logical Z errors for a $d = 7$ $|\mathrm{TOF}\rangle$ state prepared using the fault-tolerant protocol described in this section. As can be seen, from all seven possible combinations of logical Z errors, a logical Z error on block A is more likely by several orders of magnitude. All numerical simulations were performed by setting $n_{th} = 0$, $\kappa_\phi = 0$ and using the circuit level noise model described in Section III.

next stage of the protocol.

VII. TOP-DOWN SCHEME FOR HIGHER FIDELITY TOFFOLI GATES

The lowest infidelity we reported for the BUTOF protocol was 6×10^{-6}, which is insufficient for quantum algorithms using over a million Toffoli gates. Furthermore, our BUTOF protocol was realized using logical qubits encoded in a repetition code that does not protect against bit-flip errors. While bit-flip errors are exponentially suppressed in $|\alpha|^2$, a large scale quantum computer will have a very large number of potential fault locations so some protection against bit flips will be needed for reasonable values of $|\alpha|^2$. To prepare TOF states at very high fidelity – in a manner that tolerates some bit-flip errors – we propose a magic-state distillation protocol that utilizes the output of BUTOF. We use thin surface code qubits wherever a potential bit-flip would lead to an error on the output TOF state. We call this the top-down Toffoli protocol (herein TDTOF) because it assumes access to high-fidelity encoded Clifford gates, so we are attacking the problem with a view from the top of the stack. We show that using 1 round of BUTOF (see Section VI) concatenated with 1 round of TDTOF, achieves high enough fidelties to implement some quantum algorithms of interest. Alone, BUTOF struggles to reach the necessary fidelities, while concatenating multiple rounds of TDTOF, without prior BUTOF, would have a higher resource overhead. The marriage of BUTOF and TDTOF leverages the benefits of both.

Magic-state distillation protocols convert some input magic states into less noisy output magic states. To exploit BUTOF, we need a protocol that converts TOF states to higher fidelity TOF states. We know of no proposed protocols for this task that are surface code compatible and protect against generic noise, other than those which simply piggyback on known T-state distillation results (see Appendix N 1 for further discussion). Note that Toffoli and CCZ differ by only a Hadamard conjugation, so are essentially the same, and here it is convenient to work in terms of CCZ gates.

Our starting point for design of TDTOF is identifying a trio of n-qubit codes each encoding k logical qubits, which have a transversal, logical control-control-Z (CCZ) gate. That is, we consider a physical gate

$$\mathrm{CCZ}^{\otimes n} = \prod_{j=1}^{n} \mathrm{CCZ}_j, \tag{82}$$

where CCZ_j acts on j^{th} qubit in each of the 3 codeblocks. Then we need codes where this performs a logical $\mathrm{CCZ}^{\otimes k}$ in parallel across the $3k$ logical qubits.

For our trio of codes, each block encodes $k = 2$ logical qubits into $n = 8$ physical qubits and can detect an error on any single qubit. They all have one X-stabilizer $X^{\otimes 8}$

but different logical X operators

$$\begin{aligned}
X_{L1A} &= (X\otimes X\otimes X\otimes \mathbb{1}\otimes \mathbb{1}\otimes X\otimes \mathbb{1}\otimes \mathbb{1})_A, \quad (83)\\
X_{L2A} &= (X\otimes X\otimes \mathbb{1}\otimes \mathbb{1}\otimes X\otimes X\otimes \mathbb{1}\otimes \mathbb{1})_A,\\
X_{L1B} &= (X\otimes X\otimes \mathbb{1}\otimes \mathbb{1}\otimes X\otimes X\otimes \mathbb{1}\otimes \mathbb{1})_B,\\
X_{L2B} &= (X\otimes \mathbb{1}\otimes X\otimes \mathbb{1}\otimes X\otimes \mathbb{1}\otimes X\otimes \mathbb{1})_B,\\
X_{L1C} &= (X\otimes \mathbb{1}\otimes X\otimes \mathbb{1}\otimes X\otimes \mathbb{1}\otimes X\otimes \mathbb{1})_C,\\
X_{L2C} &= (X\otimes X\otimes X\otimes \mathbb{1}\otimes \mathbb{1}\otimes X\otimes \mathbb{1}\otimes \mathbb{1})_C,
\end{aligned}$$

where the index $\{A, B, C\}$ labels the three different codeblocks and the numerical index labels the 2 logical qubits in this code block. We write $(\ldots)_{D=A,B,C}$ to emphasize that the operator acts non-trivially on codeblock D and trivially on other codeblocks. While the code blocks share the same X-stabilizer, they will have different Z-stabilizers as a consequence of having different logical X operators. From just the X-stabilizer and logical operator information, we can use the same proof technique as Vasmer and Browne used for 3D surface codes [76] to show that these codes are CCZ transversal. The key element of the proof is a lemma relating tranversality to the support of logical X operators and X stabilizers (see Appendix N 2 for details). The lemma requires that for $j = 1, 2$, the operators X_{LjA}, X_{LjB} and X_{LjC} share support on an odd number of qubit indices. Furthermore, we need that for any other choice of three X operators (either logical X or X stabilizer) with one selected from each code block, they must share support on an even number of qubit indices. It is easy to verify the operators provided above have this property.

A standard recipe for magic state distillation protocols goes as follows [77]: prepare $|+\rangle_L^{\otimes 3k}$ encoded in the relevant codes; perform imperfect CCZ$^{\otimes n}$ gates by injection from noisy TOF states; measure the X-stabilizers and post-select on "+1" outcomes; and decode. This would require $3n = 24$ qubits (encoded in some underlying error-correction code) plus workspace for Cliffords and routing. However, one can make a space-time tradeoff [66, 78, 79] so that the full 24-qubit code is never prepared; rather we work with 9 qubits that we herein call the *factory qubits*. We label these 9 factory qubits with (j, D) where $D \in \{A, B, C\}$ denotes the codeblock and $j \in \{1, 2, 3\}$ specifies the qubit within the codeblock. To achieve the space-time tradeoff, we can define an encoding Clifford V such that for $D \in \{A, B, C\}$ we have

$$\begin{aligned}
VX_{1,D}V^\dagger &= X_{L1D}, \quad (84)\\
VX_{2,D}V^\dagger &= X_{L2D},\\
VX_{3,D}V^\dagger &= (X^{\otimes 8})_D.
\end{aligned}$$

Instead of encoding $|+\rangle_L^{\otimes 3k}$ and performing CCZ$^{\otimes n}$, we prepare $|+\rangle^{\otimes 9}$ and perform $V(\text{CCZ}^{\otimes n})V^\dagger$. At the end of the protocol, instead of measuring the X-stabilizers we need only measure the 3 check qubits labelled $X_{3,D}$.

It is important that $V(\text{CCZ}^{\otimes n})V^\dagger$ acts non-trivially on only the 9 qubits identified and error correction properties of the protocol are unaffected (see App. N 3 or Refs. [66, 78, 79] for details). In a Pauli-based computation, each noisy gate $V\text{CCZ}_jV^\dagger$ can be realized using a single noisy TOF state (produced by BUTOF) followed by a sequence of multi-qubit Pauli measurements implemented through lattice surgery (recall Section V and see also Appendix N 4). An explicit factory layout is given in Fig. 60 of Appendix N 7 that provides ample routing space for lattice surgery to be executed rapidly, with 4 magic state injections happening in parallel.

To describe the fault-tolerance properties of TDTOF, let us first assume the underlying memory and lattice surgery operations are implemented perfectly. Since the protocol is based on a trio of codes that can detect a single error, we can detect any fault affecting a single noisy TOF state. Even if an error affects multiple qubits within a single TOF state (e.g. a $Z \otimes Z \otimes Z$ error) we still call it a single fault-location error because it leads to no more than 1 error in each codeblock, so is detectable. Therefore, if the noisy TOF states have infidelity ϵ, then after postselection the protocol TOF outputs will have infidelity $O(\epsilon^2)$. In Appendix N 5, we show exactly how the output fidelity depends on the noise model of the input TOF states. As a toy example, in Appendix N 5 we show that for depolarizing noise the output infidelity is $1.878\epsilon^2 + O(\epsilon^3)$ per TOF state output. However, we saw in Fig. 16 that BUTOF outputs states with errors heavily dominated by $Z \otimes \mathbb{1} \otimes \mathbb{1}$. Let us consider the case when the TOF states are generated by BUTOF with $d_{BU} = 7$ and $\kappa_1/\kappa_2 = 10^{-5}$ and $\kappa_\phi = 0$, and we refer to this throughout as our benchmark example. Assuming an ideal implementation of TDTOF (without any further optimisation to the noise profile) gives an output error of $8 * 10^{-10} \sim 2\epsilon^2$, so the noise correlations slightly degrade performance relative to a depolarizing noise model with the same total error.

However, we can take advantage of this noise bias by tailoring the protocol, which we achieve using Clifford symmetries. An error-free TOF state has certain Clifford symmetries C such that $C|\text{TOF}\rangle = |\text{TOF}\rangle$. For example, the group of Clifford symmetries include $\{g_A, g_B, g_C\}$ of Eqs. (79) to (81). However, under these symmetries the noise model changes as an error E maps to $CEC^\dagger$. Therefore, we can apply a different Clifford symmetry to each of the 8 input TOF states. For any choice of Clifford symmetries, we keep the promise that $\epsilon \to O(\epsilon^2)$ but improved performance can be obtained by tailoring the Clifford symmetries to the noise profile. We prove in Appendix N 6 the existence of a set of Clifford symmetries with the following property: given an initial noise model dominated by a Pauli $Z \otimes \mathbb{1} \otimes \mathbb{1}$ error occurring with probability $\epsilon_1 = p_{Z_A}$ and rarer Z errors occurring with total probability $\epsilon_2 = \sum_{E \neq Z_A} p_E$, then the tailored protocol outputs TOF states with infidelity $O(\epsilon_1^3) + O(\epsilon_1\epsilon_2) + O(\epsilon_2^2)$. Furthermore, performing the Clifford symmetries adds a mere 2 CNOT gates to the protocol's gate complexity because most of the Clifford symmetries can be chosen as permutations of qubit labels. Having accounted for both space-time tradeoffs and noise tailoring, the full final protocol is described in Table XI of Appendix N 7.

If we return to the previously discussed benchmark example, then $\epsilon_1 = 2 * 10^{-5}$ and $\epsilon_2 = 7.5 * 10^{-9}$ so $\epsilon_2 \ll \epsilon_1$ and we expect an improvement from noise tailoring. Assuming an ideal implement of noise tailored TDTOF, then we have an output error of $1.2*10^{-12}$ that is dominated by a contribution of $\sim 8\epsilon_1\epsilon_2$. However, the protocol will not be implemented ideally. The protocol is realized with each qubit encoded into a memory: either a repetition code or a thin surface code. We can exponentially suppress memory and lattice surgery errors by increasing the code distance, though this comes at increased resource cost. The tuning of these code distances is one of the most important aspects of optimal factory design. Following a similar approach from prior work on code distance tuning [79], we present our analysis of Clifford noise in Appendix N 8.

infidelity ϵ_{TD}	# ATS	P_{ACC} (%)	T_{TD} Time (μs) per TOF	d_{BU}
$3.3 * 10^{-9}$	2646	97	485	7
$6.4 * 10^{-9}$	2016	83	455	7
$3.1 * 10^{-8}$	1680	99	362	5
$5.2 * 10^{-8}$	1596	99	340	5
$2.5 * 10^{-7}$	1470	99	318	5
$8.9 * 10^{-7}$	1386	98	318	5
$1.4 * 10^{-6}$	1302	94	311	5
$7.1 * 10^{-6}$	1176	93	289	5

TABLE IV. Resource costs for TDTOF generation of TOF states using as input 8 noisy state produced by BUTOF using distance d_{BU}. The protocol outputs 2 TOF states with infidelity $2\epsilon_{\mathrm{TD}}$ and success probability P_{ACC} rounded to nearest integer. We give the expected runtime per Toffoli as T_{TD}. The whole factory (including BUTOF modules) has a footprint given in terms of the number of ATS components; we need to triple this number to obtain the number of PCDRs (qubits). Further details provided in Tables XIII and XIV. Hardware parameters assumed: $\kappa_1/\kappa_2 = 10^{-5}$, $|\alpha^2| = 8$ and $\kappa_\phi = 0$.

Accounting for the full circuit level noise model described in Section III and the results of Sections IV and VI, we optimize the various factory settings and code distances to minimize the resource requirements for a particular target output infidelity. We focus on the regime where more than 1 bit-flip error is unlikely to occur within the factory, and so work with input magic states encoded in repetition codes and the 9 factory qubits encoded in a thin surface code with $d_x = 3$. For this regime, a comprehensive set of results is provided by Tables XIII and XIV of Appendix N 8. Here we present a selected sample of results in Table IV for target Toffoli fidelities 10^{-6} to $\sim 3.3 * 10^{-9}$. This provides a very low overhead route for reliable implementation of quantum algorithms with up to 10^8 Toffoli gates, as we discuss in the next section. Our approach can be adapted to reach even lower Toffoli error rates, but first we discuss the limiting factors influencing the results of Table IV.

size L	TOF gates N_{TOF}	T-gates N_T	# ATS	RT mins	fac %
		$u/\tau = 4$			
8	$1.8 * 10^5$	$1.5 * 10^6$	$3.2 * 10^4$	19	7.7
10	$1.8 * 10^5$	$1.2 * 10^6$	$4.7 * 10^4$	16	5.2
12	$1.9 * 10^5$	$1.1 * 10^6$	$6.6 * 10^4$	16	3.7
14	$1.9 * 10^5$	$9.5 * 10^5$	$8.9 * 10^4$	15	2.8
16	$1.9 * 10^5$	$8.8 * 10^5$	$1.1 * 10^5$	14	2.1
18	$1.9 * 10^5$	$8.4 * 10^5$	$1.4 * 10^5$	14	1.7
20	$2.0 * 10^5$	$7.8 * 10^5$	$1.8 * 10^5$	13	1.4
22	$1.9 * 10^5$	$8.3 * 10^5$	$2.1 * 10^5$	14	1.1
24	$1.9 * 10^5$	$8.0 * 10^5$	$2.5 * 10^5$	13	1.0
26	$1.9 * 10^5$	$8.4 * 10^5$	$3.2 * 10^5$	14	0.8
28	$2.0 * 10^5$	$8.1 * 10^5$	$3.7 * 10^5$	13	0.7
30	$2.0 * 10^5$	$8.3 * 10^5$	$4.2 * 10^5$	14	0.6
32	$2.0 * 10^5$	$8.3 * 10^5$	$4.8 * 10^5$	15	0.5
		$u/\tau = 8$			
8	$4.3 * 10^5$	$3.8 * 10^6$	$3.2 * 10^4$	49	7.7
10	$4.4 * 10^5$	$3.0 * 10^6$	$4.7 * 10^4$	42	5.2
12	$4.6 * 10^5$	$2.7 * 10^6$	$6.6 * 10^4$	39	3.7
14	$4.6 * 10^5$	$2.3 * 10^6$	$8.9 * 10^4$	36	2.8
16	$4.6 * 10^5$	$2.2 * 10^6$	$1.2 * 10^5$	34	2.0
18	$4.6 * 10^5$	$2.0 * 10^6$	$1.6 * 10^5$	33	1.6
20	$4.7 * 10^5$	$1.9 * 10^6$	$1.9 * 10^5$	34	1.3
22	$4.6 * 10^5$	$2.0 * 10^6$	$2.3 * 10^5$	35	1.1
24	$4.7 * 10^5$	$1.9 * 10^6$	$2.7 * 10^5$	35	0.9
26	$4.6 * 10^5$	$2.0 * 10^6$	$3.2 * 10^5$	35	0.8
28	$4.6 * 10^5$	$2.0 * 10^6$	$3.7 * 10^5$	35	0.7
30	$4.7 * 10^5$	$2.0 * 10^6$	$4.2 * 10^5$	35	0.6
32	$4.7 * 10^5$	$2.0 * 10^6$	$4.8 * 10^5$	35	0.5

TABLE V. Resource estimates for performing phase estimation upto multiplicative (extensive) error of 5% and algorithm success probability of over 90% for the $L \times L$ Hubbard model with parameters $u/T = 4$ and $u/T = 8$. We use the Trotterization scheme and analysis of Ref. [80] to count the required Toffoli and T gates. Using catalysis, 1 TOF state can perform 2 T-gates. Column #ATS refers to the total number of ATS component used, we need to triple this number to obtain the number of PCDRs (qubits). The total #ATS count includes: $2L^2$ logical qubits to represent the Hubbard model fermions; ancilla qubits for phase estimation, ancilla-assisted circuit synthesis [81], Hamming weight phasing and catalysis [82]; and the ATS space for 1 TDTOF factory (%fac counts the percentage of this contribution rounded up to nearest integer); and we also include a generous +30% space overhead for routing and lattice surgery costs. Notice that the gate complexity and runtime is roughly constant as a function of system size since for an extensive error estimate the absolute error decreases with system size and this outweights other factors. For an additive (intensive) error energy estimation, the gate complexity will grow with system size [6, 80]. Hardware parameters assumed: $\kappa_1/\kappa_2 = 10^{-5}$, $|\alpha^2| = 8$ and $\kappa_\phi = 0$.

There are many contributing sources of error, but we can understand this $3.3 * 10^{-9}$ limit by considering the process of bit-flip errors on noisy input TOF states encoded in repetition codes. For hardware parameters $\kappa_1/\kappa_2 = 10^{-5}$, $\kappa_\phi = 0$ and $|\alpha^2| = 8$, the lowest total error rate is $\delta := 2.7 * 10^{-8}$ per repetition code cycle (recall Fig. 7). If a single TOF state is stored for r repetition code cycles, then roughly the accumulated error on qubits B and C is δr and this gives an additional contribution of $2\delta r$ to the non-dominate noise contributions ϵ_2. Returning to our benchmark example, the output error is (roughly) lower bounded by

$$\begin{aligned} C\epsilon_1\epsilon_2' &\sim C\epsilon_1\left(\epsilon_2 + 2\delta r\right) \\ &= C(2*10^{-5})\left(7.5*10^{-9} + 2*2.7*10^{-8}r\right). \end{aligned} \tag{85}$$

where C is some constant that depends on the exact details of the noise profile and we have discussed examples where $C \sim 2$ and $C \sim 8$. Furthermore, for the larger factory examples in Table IV repetition codes could be in storage for as long as $r \sim 200$ repetition code cycles. Together this approximate accounting indicates (with $r = 200$ and $C = 5$) that we should not expect output infidelities lower than $\sim 1.1 * 10^{-9}$. Of course, this is a rough estimation of one error source, just to provide the reader with some intuition. Rather, for a precise accounting of all error sources the lowest observed infidelity was $3.3 * 10^{-9}$. However, there are several straightforward routes to reaching even lower infidelities. By converting immediately after BUTOF from repetition code to thin surface code we would reduce the time exposed to bit-flip errors. For our benchmark example, encoding directly into surface codes should enable us to get much closer to $1.2 * 10^{-12}$ infidelity (the ideal Clifford limit for noise tailored TDTOF). Ultimately, arbitrarily high fidelities can be reached by concatenating TDTOF, though resource costs jump substantially with each level of concatenation. Alternatively, better fidelities could be reached it hardware parameters could be improved by either further suppressing bit-flips by increasing $|\alpha|^2$ or decreasing κ_1/κ_2.

Let us compare to the factory of Gidney and Fowler [83] that concatenates T state distillation with a protocol that distills a single TOF state from a supply T states. Using a square surface code distance d, the factory requires $12d \times 6d$ qubits and takes $5.5d$ surface code cycles. They assume a superconducting transmon architecture with $p_{SC} = 10^{-3}$ CNOT gate infidelity that can execute one cycle of surface code error correction in $1\mu s$. For example algorithms with $\sim 1 - 100$ million Toffoli gates, they considered a $d = 31$ surface codes which gives a $6.9 * 10^4$ qubit footprint generating 1 TOF state every $170\mu s$. This is a considerably larger size than our factory mainly because we exploit BUTOF, thin surface codes and where possible we use repetition codes. Note that Table IV assumed hardware parameters leading to surface code cycles of $5.5\mu s$ rather than $1.1\mu s$, so while our factory typically needs far fewer surface code cycles per TOF state, our slower physical gate times mean that the overall factory runtime (per TOF state) is slower by about a factor 2.

VIII. OVERHEAD ESTIMATES

Here we consider how our architecture could be used to fault-tolerantly implement a quantum algorithm beyond the reach of classical computers. Throughout this section we assume hardware parameters $|\alpha|^2 = 8$ and $\kappa_1/\kappa_2 = 10^{-5}$. Using a Pauli-based computation the complexity is mainly determined by the number of qubits and Toffoli gates required for the algorithm. Classical simulations of 100 qubit circuits are substantially beyond the reach of current classical methods unless they have low depth or are near-Clifford circuits. The best known classical simulation algorithm of near-Clifford circuits [84] for an n-qubit circuit with N_{TOF} total Toffoli gate count has a runtime $O(\mathrm{poly}(n,\tau)2^{0.83N_{\mathrm{TOF}}})$. For $N_{\mathrm{TOF}} = 1000$, the exponential component of the runtime is comfortably in the classically intractable regime. Let us consider a $n = 100$ and $N_{\mathrm{TOF}} = 1000$ computation. Using $d_{BU} = 9$ for BUTOF, we could achieve $6 * 10^{-6}$ error per TOF and so a total of 0.6% algorithm error from TOF gates. For memory using the repetition code, the lowest achievable logical error rate is $2.7 * 10^{-8}$ using a $d_{\mathrm{rep}} = 9$ repetition code (recall Fig. 7). With $\tau d_{\mathrm{rep}} = 9000$ repetition code cycles and $n = 100$ logical qubits, the total error probability is $\sim 2.4\%$. The combined memory and Toffoli error is then $\sim 3\%$. Such a computation would be possible with 900 ATS components for memory and several hundred ATS components to parallelize BUTOF. Some additional resources will be needed for routing and performing Clifford operations, so the entire device would require between 1 and 2 thousand ATS components depending on routing and Clifford requirements. Combining it with the ATS-based X-measurement scheme which does not use a transmon, we believe our scheme gives a promising route for early implementations of fault-tolerant quantum computing.

While a thousand Toffoli gate computation would be feasible in such a purely repetition code architecture, for anything larger we will need bitflip protection to reduce memory logical error rates. Unfortunately, there are no known algorithms where it is believed that a thousand Toffoli gates gives a quantum advantage for a useful problem of commercial or societal value. The smallest such quantum algorithms include simulations of the Hubbard model [6, 80] that has a Hamiltonian

$$H = u\sum_i a_{i,\uparrow}^\dagger a_{i,\uparrow} a_{i,\downarrow}^\dagger a_{i,\downarrow} + T \sum_{i,j\in N(i)} (a_{i,\uparrow}^\dagger a_{j,\uparrow} + a_{i,\downarrow}^\dagger a_{j,\downarrow}), \tag{86}$$

where $N(i)$ indicates the set of neighboring lattice sites to i with respect to an $L \times L$ square lattice with periodic boundary conditions. The variables u and T are important physical parameters that govern the phase transitions of the system. We consider $u/T = 4$ to enable easier comparison with Ref. [6]. However, classical simulation is most difficult in the regime near $u/T = 8$ [85] and so we also consider this choice. Estimating the ground state energy upto multiplicative error typically requires over 1 million Toffoli gates and over 100 logical qubits [6]. While this is not feasible using just repetition encoding, only very little bit-flip protection suffices so we can use a $d_x = 3$ thin surface code as our primary storage and the `TDTOF` protocol for Toffoli states. We present resource estimates for this problem in Table V, with technical details in the caption. We give separately the number of logical TOF gates (N_{TOF}) and logical T gates (N_T) required by the algorithm. We can catalyze 1 TOF state into 2 T states [82], so that the algorithm consumes a total of

$$N_{\mathrm{TOT-TOF}} = N_{\mathrm{TOF}} + (N_T/2) \tag{87}$$

TOF states.

Caveats in architectural comparisons.- Our results for $u/\tau = 4$ in Table V can be compared with the transmon architecture resource estimates of Table I of [6], though subject to several caveats that we list first. Direct comparisons are difficult because the noise models are very different. Transmons architecture are typically considered with CNOT error probabilities of $p_{SC} = 10^{-3}$ and $p_{SC} = 10^{-4}$. Though for our hardware parameters the CNOT infidelity is 3.8×10^{-3} and to perform a CNOT with infidelity of 10^{-4} we would need $\kappa_1/\kappa_2 = 10^{-8}$ (see caption of Fig. 8 for further discussion). So although we benefit greatly from bit-flip suppression due to catcodes, our current projections for Z error rates are far less optimistic than typically assumed for transmon qubits. Furthermore, transmon architecture resource estimates are based on a toy depolarizing noise model, whereas our noise model has been derived from first principles modeling of the hardware. An additional important caveat is that we exploit the Hubbard model simulation analysis of [80], which provides a 5.5× reduction in gate count for $L = 8$ and a larger improvement for larger L (compared to Ref. [6]). These gate count reductions lead to a comparable reduction in runtime, which we account for in our discussion below. However, modest reductions in gate counts have a very mild influence on qubit requirements and furthermore this is mitigated since we present results for a higher total algorithm success probability (90% instead of 70%). Overall, it is therefore meaningful to directly compare qubit tallies of our Table V with Table I of [6].

Qubit cost discussion: Compared to a superconducting transmon qubit architecture [6] with CNOT infidelity p_{SC}: when $p_{SC} = 10^{-3}$ we need $\sim 3\times$ fewer qubits; and when $p_{SC} = 10^{-4}$ we use a comparable number of qubits. If we could achieve lower κ_1/κ_2 then there would be additional resource savings. Extrapolating our surface code simulations, we estimate that $\kappa_1/\kappa_2 = 10^{-6}$ would lead to an extra 1.7× qubit reduction and $\kappa_1/\kappa_2 = 10^{-8}$ would lead to an extra 3.1× qubit reduction. One loose assumption in our qubit counting is that we include a 1.3× overhead to account for routing and lattice surgery costs (see discussion of Section V) whereas we do not know what overhead was assumed in Ref. [6].

Runtime discussion: The total runtime of our architecture is also competitive, with 19-49 minutes for a classically challenging task. There are two important factors in the runtime analysis: the time it takes to prepare τ TOF states, which is $T_a = \tau T_{TD}$ with example T_{TD} values in Table IV; the time required to inject magic states via lattice surgery $T_b = (3N_{\mathrm{TOF}} + N_T)d_m T_{\mathrm{surf}}$ where T_{surf} is the time per surface code cycle and d_m is the number of surface code cycles per lattice surgery operation (recall Fig. 12). We take the runtime to be $RT = \max[T_a, T_b]$. We say the architecture is Clifford bottlenecked if $RT = T_b$ and magic-state bottlenecked if $RT = T_a$. Note that our estimate of T_a assumes that we can only inject 1 magic state at a time; since faster injection rates could incur higher routing or Clifford gate costs. For our hardware and factory design, we are Clifford bottlenecked as the `TDTOF` factory is producing Toffoli states slightly faster than they can be transported into the main algorithm. In contrast, estimates for superconducting transmon architectures [6] have assumed a single factory leading to them being magic-state bottlenecked (with the algorithm often idle and waiting for the factory). Let us consider the instance with $u/\tau = 4$ and $L = 8$ for which we estimate a runtime of 19 minutes. For a transmon architecture with $p_{SC} = 10^{-3}$, one obtains an estimate of 3 minute runtime by reducing the results of Ref. [6] of by a factor 5.5 to account for recent algorithmic improvements [80]. A similar runtime estimate (2.6 minutes) is obtained for the transmon architecture by assuming it generates 1 TOF state per 170μs using the factory of Ref. [83]. Overall, the transmon archiecture is executing about 7× faster that is mostly attributable to faster execution of each surface code cycle.

IX. CONCLUSION

In this paper we presented a comprehensive analysis of an architecture for a fault-tolerant quantum computer. At the lowest level, it is based on hybrid acoustic-electro devices to implement a stabilized cat code with highly biased noise, dominated by dephasing. This cat code is then concatenated with an outer code that focuses mostly on correcting the dephasing errors. Our estimated overheads for performing fault-tolerant quantum algorithms showcase the promise of this approach. There are several interesting directions for future work to improve on our current proposal.

On the hardware side, we would like to explore ways to increase the value of κ_2, which would allow us to achieve the desired ratio of κ_1/κ_2 with a less stringent constraint on $T_1 = 1/\kappa_1$ of the acoustic oscillators. Currently the value of κ_2 is upper bounded by the cross-talk error and the bandwidth of the filter. We believe similar set ups with tunable couplers, multiport resonators, and multiple dump modes are promising for increasing substantially the attainable value of κ_2. Higher κ_2 would also give faster gates, allowing for larger quantum advantage over classical computing.

As was shown in this work, the magic state factory only accounts for approximately 7% of the total resource overhead requirements. The other 93% of the overhead requirements is largely dominated by the performance of the thin rotated surface code. Recently, an $XZZX$-type surface code which takes advantage of the noise bias for phase-flip errors was introduced and shown to have better thresholds compared to the rotated surface code [10]. An interesting avenue for future work would be to consider the implementation of the $XZZX$ surface code (or other topological codes which take advantage of the noise bias) in our architecture to determine if further reductions in overhead costs can be achieved. Further, one could use compass codes [86–89] which potentially require fewer resources compared to surface codes given the low-weight gauge operator measurements. However, details for implementing such codes in a lattice surgery scheme such as the one presented in this work remain to be addressed.

We have considered a standard model of Pauli-based computation with Pauli operators measured by lattice surgery in order to inject magic states. This approach comes with an additional qubit cost for data access and routing, and the choice of routing solution also leads to a lower bound on runtime execution. In previous resource analyses these considerations were not especially important because algorithms were bottlenecked by the pace at which they could produce magic states. In contrast, this emerged as a bottleneck in our architecture and so more careful optimization of routing costs and speed of magic state injection is crucial. Indeed, a rapid runtime is especially important in an architecture where bitflips are rare because it is desirable to execute the algorithm fast enough that we can avoid needing a higher X distance code.

ACKNOWLEDGMENTS

We thank Qian Xu for helping with the displaced Fock basis calculation and Alex Retzker for discussions. C.C. thanks Yunong Shi and Pierre-Yves Aquilanti for their help in setting up the AWS clusters where most of the error correction simulations were performed. We thank all the members of the AWS Center of Quantum Computing for our collaboration on building more powerful quantum technologies. We thank Richard Moulds, Nadia Carlsten, Eric Kessler, and all the members of the Amazon Braket and Quantum Solutions Lab teams. We thank Simone Severini for creating an environment where this research was possible in the first place. We thank Bill Vass, James Hamilton and Charlie Bell for their support and guidance throughout this project.

[1] S. B. Bravyi and A. Y. Kitaev, Quantum codes on a lattice with boundary (1998), arXiv:quant-ph/9811052.

[2] J. O'Gorman and E. T. Campbell, Physical Review A **95**, 032338 (2017), arXiv:1605.07197.

[3] M. Motta, E. Ye, J. R. McClean, Z. Li, A. J. Minnich, R. Babbush, and G. K.-L. Chan, Low rank representations for quantum simulation of electronic structure (2018), arXiv:1808.02625.

[4] C. Gidney and M. Ekerå, How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits (2019), arXiv:1905.09749 [quant-ph].

[5] E. Campbell, A. Khurana, and A. Montanaro, Quantum **3**, 167 (2019), arXiv:1810.05582.

[6] I. D. Kivlichan, C. Gidney, D. W. Berry, N. Wiebe, J. McClean, W. Sun, Z. Jiang, N. Rubin, A. Fowler, A. Aspuru-Guzik, and et al., Quantum **4**, 296 (2020), arXiv:1902.10673.

[7] D. K. Tuckett, S. D. Bartlett, and S. T. Flammia, Phys. Rev. Lett. **120**, 050505 (2018), arXiv:1708.08474.

[8] D. K. Tuckett, A. S. Darmawan, C. T. Chubb, S. Bravyi, S. D. Bartlett, and S. T. Flammia, Phys. Rev. X **9**, 041031 (2019), arXiv:1812.08186.

[9] D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, Phys. Rev. Lett. **124**, 130501 (2020), arXiv:1907.02554.

[10] J. P. Bonilla Ataides, D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, The XZZX surface code (2020), arXiv:2009.07851 [quant-ph].

[11] P. Aliferis and J. Preskill, Phys. Rev. A **78**, 052331 (2008), arXiv:0710.1301.

[12] S. Puri, L. St-Jean, J. A. Gross, A. Grimm, N. E. Frattini, P. S. Iyer, A. Krishna, S. Touzard, L. Jiang, A. Blais, and et al., Science Advances **6**, eaay5901 (2020), arXiv:1905.00450.

[13] J. Guillaud and M. Mirrahimi, Phys. Rev. X **9**, 041053 (2019), arXiv:1904.09474.

[14] J. Guillaud and M. Mirrahimi, Error Rates and Resource Overheads of Repetition Cat Qubits (2020), arXiv:2009.10756 [quant-ph].

[15] M. Mirrahimi, Z. Leghtas, V. V. Albert, S. Touzard, R. J. Schoelkopf, L. Jiang, and M. H. Devoret, New Journal of Physics **16**, 045014 (2014).

[16] S. Puri, S. Boutin, and A. Blais, npj Quantum Inf. **3**, 18 (2017).

[17] J. Cohen, *Autonomous quantum error correction with superconducting qubits*, Ph.D. thesis, Université Paris sciences et lettres (2017).

[18] V. V. Albert, K. Noh, K. Duivenvoorden, D. J. Young, R. T. Brierley, P. Reinhold, C. Vuillot, L. Li, C. Shen, S. M. Girvin, B. M. Terhal, and L. Jiang, Phys. Rev. A **97**, 032346 (2018).

[19] A. Joshi, K. Noh, and Y. Y. Gao, arXiv e-prints , arXiv:2008.13471 (2020), arXiv:2008.13471 [quant-ph].

[20] W. Cai, Y. Ma, W. Wang, C. L. Zou, and L. Sun, arXiv e-prints , arXiv:2010.08699 (2020), arXiv:2010.08699 [quant-ph].

[21] Z. Leghtas, S. Touzard, I. M. Pop, A. Kou, B. Vlastakis, A. Petrenko, K. M. Sliwa, A. Narla, S. Shankar,

M. J. Hatridge, M. Reagor, L. Frunzio, R. J. Schoelkopf, M. Mirrahimi, and M. H. Devoret, Science **347**, 853 (2015), 1412.4633.

[22] S. Touzard, A. Grimm, Z. Leghtas, S. Mundhada, P. Reinhold, C. Axline, M. Reagor, K. Chou, J. Blumoff, K. Sliwa, and et al., Physical Review X **8**, 021005 (2018), arXiv:1705.02401.

[23] S. Puri, A. Grimm, P. Campagne-Ibarcq, A. Eickbusch, K. Noh, G. Roberts, L. Jiang, M. Mirrahimi, M. H. Devoret, and S. M. Girvin, Phys. Rev. X **9**, 041009 (2019).

[24] A. Grimm, N. E. Frattini, S. Puri, S. O. Mundhada, S. Touzard, M. Mirrahimi, S. M. Girvin, S. Shankar, and M. H. Devoret, Nature **584**, 205 (2020), arXiv:1907.12131.

[25] R. Lescanne, M. Villiers, T. Peronnin, A. Sarlette, M. Delbecq, B. Huard, T. Kontos, M. Mirrahimi, and Z. Leghtas, Nature Physics **16**, 509 (2020).

[26] P. Arrangoiz-Arriola and A. H. Safavi-Naeini, Phys. Rev. A **94**, 063864 (2016).

[27] G. S. MacCabe, H. Ren, J. Luo, J. D. Cohen, H. Zhou, A. Sipahigil, M. Mirhosseini, and O. Painter, Science **370**, 840 (2020), https://science.sciencemag.org/content/370/6518/840.full.pdf.

[28] P. Arrangoiz-Arriola, E. A. Wollack, Z. Wang, M. Pechal, W. Jiang, T. P. McKenna, J. D. Witmer, R. Van Laer, and A. H. Safavi-Naeini, Nature **571**, 537 (2019).

[29] E. A. Wollack, A. Y. Cleland, P. Arrangoiz-Arriola, T. P. McKenna, R. G. Gruenke, R. N. Patel, W. Jiang, C. J. Sarabalis, and A. H. Safavi-Naeini, Loss channels affecting lithium niobate phononic crystal resonators at cryogenic temperature (2020), arXiv:2010.01025 [physics.app-ph].

[30] S. Bravyi and J. Haah, Phys. Rev. A **86**, 052329 (2012).

[31] A. M. Meier, B. Eastin, and E. Knill, Quant. Inf. and Comp. **13**, 195 (2013).

[32] E. T. Campbell and M. Howard, Quantum **2**, 56 (2018).

[33] S. Bravyi and A. Kitaev, Phys. Rev. A **71**, 022316 (2005).

[34] J. P. Paz and W. H. Zurek, Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **454**, 355 (1998).

[35] C. Ahn, A. C. Doherty, and A. J. Landahl, Phys. Rev. A **65**, 042301 (2002).

[36] M. Sarovar and G. J. Milburn, Phys. Rev. A **72**, 012306 (2005).

[37] P. T. Cochrane, G. J. Milburn, and W. J. Munro, Phys. Rev. A **59**, 2631 (1999).

[38] H. Jeong and M. S. Kim, Phys. Rev. A **65**, 042305 (2002).

[39] F. Reiter and A. S. Sørensen, Phys. Rev. A **85**, 032111 (2012).

[40] N. E. Frattini, U. Vool, S. Shankar, A. Narla, K. M. Sliwa, and M. H. Devoret, Applied Physics Letters **110**, 222603 (2017), https://doi.org/10.1063/1.4984142.

[41] M. Mirhosseini, A. Sipahigil, M. Kalaee, and O. Painter, Quantum transduction of optical photons from a superconducting qubit (2020), arXiv:2004.04838 [quant-ph].

[42] R. Lescanne, L. Verney, Q. Ficheux, M. H. Devoret, B. Huard, M. Mirrahimi, and Z. Leghtas, Phys. Rev. Applied **11**, 014030 (2019).

[43] 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, Phys. Rev. Lett. **117**, 190503 (2016).

[44] Y. Zhang, B. J. Lester, Y. Y. Gao, L. Jiang, R. J. Schoelkopf, and S. M. Girvin, Phys. Rev. A **99**, 012314 (2019).

[45] L. Verney, R. Lescanne, M. H. Devoret, Z. Leghtas, and M. Mirrahimi, Phys. Rev. Applied **11**, 024003 (2019).

[46] D. F. V. James and J. Jerke, Can. J. Phys. **85**, 625 (2007).

[47] O. Gamel and D. F. V. James, Phys. Rev. A **82**, 052106 (2010).

[48] R. K. Naik, N. Leung, S. Chakram, P. Groszkowski, Y. Lu, N. Earnest, D. C. McKay, J. Koch, and D. I. Schuster, Nat. Commun. **8**, 1904 (2017).

[49] M. Pechal, P. Arrangoiz-Arriola, and A. H. Safavi-Naeini, Quantum Sci. Technol. **4**, 015006 (2018).

[50] C. T. Hann, C.-L. Zou, Y. Zhang, Y. Chu, R. J. Schoelkopf, S. M. Girvin, and L. Jiang, Phys. Rev. Lett. **123**, 250501 (2019).

[51] P. Mundada, G. Zhang, T. Hazard, and A. Houck, Phys. Rev. Appl. **12**, 10.1103/PhysRevApplied.12.054023 (2019).

[52] Y. Chen, C. Neill, P. Roushan, N. Leung, M. Fang, R. Barends, J. Kelly, B. Campbell, Z. Chen, B. Chiaro, A. Dunsworth, E. Jeffrey, A. Megrant, J. Y. Mutus, P. J. J. O'Malley, C. M. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, M. R. Geller, A. N. Cleland, and J. M. Martinis, Phys. Rev. Lett. **113**, 220502 (2014).

[53] L. Sun, A. Petrenko, Z. Leghtas, B. Vlastakis, G. Kirchmair, K. Sliwa, A. Narla, M. Hatridge, S. Shankar, J. Blumoff, *et al.*, Nature **511**, 444 (2014).

[54] C. T. Hann, S. S. Elder, C. S. Wang, K. Chou, R. J. Schoelkopf, and L. Jiang, Phys. Rev. A **98**, 022305 (2018).

[55] S. S. Elder, C. S. Wang, P. Reinhold, C. T. Hann, K. S. Chou, B. J. Lester, S. Rosenblum, L. Frunzio, L. Jiang, and R. J. Schoelkopf, Phys. Rev. X **10**, 011001 (2020).

[56] N. Didier, J. Bourassa, and A. Blais, Phys. Rev. Lett. **115**, 203601 (2015).

[57] A. A. Clerk, M. H. Devoret, S. M. Girvin, F. Marquardt, and R. J. Schoelkopf, Rev. Mod. Phys. **82**, 1155 (2010).

[58] P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, Applied Physics Reviews **6**, 021318 (2019).

[59] Y. Tomita and K. M. Svore, Phys. Rev. A **90**, 062320 (2014).

[60] D. P. DiVincenzo and P. Aliferis, Phys. Rev. Lett. **98**, 020501 (2007).

[61] C. Chamberland, P. Iyer, and D. Poulin, Quantum **2**, 43 (2018).

[62] J. Kelly, R. Barends, A. G. Fowler, A. Megrant, E. Jeffrey, T. C. White, D. Sank, J. Y. Mutus, B. Campbell, Y. Chen, *et al.*, Nature **519**, 66 (2015).

[63] C. Horsman, A. G. Fowler, S. Devitt, and R. V. Meter, New Journal of Physics **14**, 123011 (2012).

[64] A. J. Landahl and C. Ryan-Anderson, arXiv preprint arXiv:1407.5103 (2014).

[65] D. Litinski and F. v. Oppen, Quantum **2**, 62 (2018).

[66] D. Litinski, Quantum **3**, 128 (2019).

[67] D. Litinski and F. von Oppen, Quantum **2**, 62 (2018).

[68] S. Bravyi, G. Smith, and J. A. Smolin, Physical Review X **6**, 021043 (2016).

[69] A. Paler and A. G. Fowler, arXiv preprint arXiv:1906.07994 (2019).

[70] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Phys. Rev. A **86**, 032324 (2012).

[71] A. G. Fowler, arXiv preprint arXiv:1210.4626 (2012).

[72] C. Chamberland and A. W. Cross, Quantum **3**, 143 (2019).

[73] C. Chamberland and K. Noh, npj Quantum Information **6**, 91 (2020).

[74] Recently, it was also shown in Ref. [14] that when implementing a logical Toffoli gate using a piece-wise fault-tolerant approach, one can track the CZ errors that arise when Z errors propagate through the target qubits of the physical Toffoli gates (see Appendix K) and correct all errors at the final output of piece-wise circuit (instead of in between each blocks of physical Toffoli gates). Such an approach could potentially be used in our bottom-up $|\text{TOF}\rangle$ state preparation scheme, allowing us to avoid using the STOP algorithm to prepare $|0\rangle_L$ and $|1\rangle_L$. However we leave such an analysis to future work.

[75] This should be compared to the circuits used in [73] for preparing $|H\rangle$ type magic states which are fault-tolerant to all types of Pauli noise given that a depolarizing circuit level noise model was assumed.

[76] M. Vasmer and D. E. Browne, arXiv preprint arXiv:1801.04255 (2018).

[77] S. Bravyi and A. Kitaev, Phys. Rev. A **71**, 022316 (2005).

[78] J. Haah and M. B. Hastings, Quantum **2**, 71 (2018).

[79] D. Litinski, Quantum **3**, 205 (2019).

[80] E. Campbell (2020), in preparation.

[81] N. Wiebe and C. Granade, Phys. Rev. Lett. **117**, 010503 (2016).

[82] C. Gidney, Quantum **2**, 74 (2018).

[83] C. Gidney and A. G. Fowler, Quantum **3**, 135 (2019).

[84] S. Bravyi, D. Browne, P. Calpin, E. Campbell, D. Gosset, and M. Howard, Quantum **3**, 181 (2019).

[85] B.-X. Zheng, C.-M. Chung, P. Corboz, G. Ehlers, M.-P. Qin, R. M. Noack, H. Shi, S. R. White, S. Zhang, and G. K.-L. Chan, Science **358**, 1155 (2017).

[86] M. Li, D. Miller, M. Newman, Y. Wu, and K. R. Brown, Phys. Rev. X **9**, 021041 (2019).

[87] C. Chamberland, G. Zhu, T. J. Yoder, J. B. Hertzberg, and A. W. Cross, Phys. Rev. X **10**, 011022 (2020).

[88] D. M. Debroy, M. Li, S. Huang, and K. R. Brown, Quantum Science and Technology **5**, 034002 (2020).

[89] S. Huang and K. R. Brown, Phys. Rev. A **101**, 042312 (2020).

[90] S. E. Nigg, H. Paik, B. Vlastakis, G. Kirchmair, S. Shankar, L. Frunzio, M. H. Devoret, R. J. Schoelkopf, and S. M. Girvin, Phys. Rev. Lett. **108**, 240502 (2012).

[91] M. Pechal and A. H. Safavi-Naeini, Phys. Rev. A **96**, 042305 (2017).

[92] J. M. Kreikebaum, K. P. O'Brien, A. Morvan, and I. Siddiqi, Superconductor Science and Technology **33**, 06LT02 (2020).

[93] V. S. Ferreira, J. Banker, A. Sipahigil, M. H. Matheny, A. J. Keller, E. Kim, M. Mirhosseini, and O. Painter, Collapse and revival of an artificial atom coupled to a structured photonic reservoir (2020), arXiv:2001.03240 [quant-ph].

[94] R. Azouit, A. Sarlette, and P. Rouchon, Adiabatic elimination for open quantum systems with effective lindblad master equations (2016), arXiv:1603.04630 [quant-ph].

[95] R. Azouit, F. Chittaro, A. Sarlette, and P. Rouchon, Quantum Sci. Technol. **2**, 044011 (2017).

[96] E. A. Sete, J. M. Martinis, and A. N. Korotkov, Phys. Rev. A **92**, 012325 (2015).

[97] D. F. James and J. Jerke, Canadian Journal of Physics **85**, 625 (2007), https://doi.org/10.1139/p07-060.

[98] 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, Phys. Rev. Applied **10**, 034040 (2018).

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

[100] W. Pfaff, C. Axline, L. Burkhart, *et al.*, Nature Physics **13**, 882 (2017).

[101] D. Gottesman, Proceedings of Symposia in Applied Mathematics **68**, 13 (2010).

[102] C. Chamberland and M. E. Beverland, Quantum **2**, 53 (2018).

[103] P. Aliferis, D. Gottesman, and J. Preskill, Quantum Info. Comput. **6**, 97 (2006).

[104] J. Edmonds, Canadian Journal of mathematics **17**, 449 (1965).

[105] N. Delfosse and N. H. Nickerson, arXiv e-prints , arXiv:1709.06218 (2017), arXiv:1709.06218.

[106] There are cases where families of error correcting codes have thresholds when using the STOP algorithm to decode with the syndrome from the last round. One such example includes concatenated codes using the methods of [103].

[107] P. Brooks and J. Preskill, Physical Review A **87**, 032310 (2013).

[108] J. T. Anderson, G. Duclos-Cianci, and D. Poulin, Phys. Rev. Lett. **113**, 080501 (2014).

[109] C. Vuillot, L. Lao, B. Criger, C. G. Almudéver, K. Bertels, and B. M. Terhal, New Journal of Physics **21**, 033028 (2019).

[110] A. J. Landahl and C. Ryan-Anderson, arXiv e-prints , arXiv:1407.5103 (2014), arXiv:1407.5103.

[111] A. G. Fowler and C. Gidney, arXiv:1808.06709 (2018).

[112] D. Gottesman, Proceedings, XXII International Colloquium on Group Theoretical Methods in Physics , 32 (1999).

[113] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, J. Math. Phys. **43**, 4452 (2002).

[114] A. A. Kovalev and L. P. Pryadko, arXiv preprint arXiv:1208.2317 (2012).

[115] R. Raussendorf and J. Harrington, Physical review letters **98**, 190504 (2007).

[116] D. S. Wang, A. G. Fowler, and L. C. L. Hollenberg, Phys. Rev. A **83**, 020302 (2011).

[117] A. G. Fowler, Phys. Rev. Lett. **109**, 180502 (2012).

[118] C. Chamberland, A. Kubica, T. J. Yoder, and G. Zhu, New Journal of Physics **22**, 023019 (2020).

[119] P. W. Shor, in *Proceedings of the 37th Annual Symposium on Foundations of Computer Science*, FOCS '96 (IEEE Computer Society, USA, 1996) p. 56.

[120] E. Dennis, Phys. Rev. A **63**, 052314 (2001).

[121] C. Jones, Phys. Rev. A **87**, 042305 (2013).

[122] A. G. Fowler, arXiv:1310.0863 (2013).

[123] M. B. Hastings and J. Haah, Physical review letters **120**, 050504 (2018).

[124] C. Jones, Phys. Rev. A. **87**, 022328 (2013).
[125] B. Eastin, Phys. Rev. A. **87**, 032321 (2013).
[126] E. T. Campbell and J. O'Gorman, Quantum Science and Technology **1**, 015007 (2016).
[127] E. T. Campbell and M. Howard, Physical Review A **95**, 022316 (2017).
[128] E. T. Campbell and M. Howard, Physical review letters **118**, 060501 (2017).
[129] J. Haah, M. B. Hastings, D. Poulin, and D. Wecker, Quantum **1**, 31 (2017).
[130] A. Paetznick and B. W. Reichardt, Physical review letters **111**, 090505 (2013).
[131] A. Kubica, B. Yoshida, and F. Pastawski, New Journal of Physics **17**, 083026 (2015).
[132] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes* (North Holland, 1988).

Appendix A: Engineering two-phonon dissipation with piezoelectric nanostructures

In this Appendix we calculate the dimensionless loss parameter κ_1/κ_2 — the ratio of the single-phonon and two-phonon dissipation rates — and show how to minimize it to the lowest level allowable by the intrinsic loss of the hardware and the crosstalk constraints derived in Appendix B. This Appendix is divided into four parts. First, in Appendix A 1 we revisit an existing method to engineer two-photon (or in this case two-phonon) dissipation using an asymmetrically-threaded SQUID (ATS) device. [25]. Next we show in Appendix A 2 how to calculate the interaction rate g_2 when the storage resonator is an arbitrary piezoelectric nanostructure, and explicitly calculate g_2 for the specific case of a one-dimensional phononic-crystal-defect resonator (PCDR) [28]. Then in Appendix A 3 we derive, using a classical description of the underlying superconducting circuits, a general expression for κ_2 when a bandpass filter is placed in between the output port of the buffer resonator and the external $50\,\Omega$ environment and show how to design the filter to optimize κ_2. We include a filter in our analysis because filtering the output — or engineering the density of states of the system's reservoir — is crucial to the multiplexed stabilization protocol described in Appendix B. Finally, in Appendix A 5 we show that the loss κ_1/κ_2 can be minimized by utilizing a high-impedance buffer resonator and calculate a lower bound for this loss.

1. Implementation of the required Josephson nonlinearity

In Section II in the main text, we described at a high level how the two-phonon dissipation can be generated by engineering a nonlinear interaction $g_2^*\hat{a}^2\hat{b}^\dagger + \text{h.c.}$ between the storage mode $\hat{a}$ and a very lossy "buffer" mode $\hat{b}$. Here we describe in detail how this interaction can be engineered and calculate estimates of g_2 specifically for the hardware in this proposal. Following the method introduced in Ref. [25], we propose implementing the required nonlinearity using an asymmetrically-threaded SQUID ("ATS") device, which consists of an ordinary superconducting quantum interference device (SQUID) that is split in the middle by a linear inductor — see Fig. 17. We reproduce some of the results of Ref. [25] here for convenience.

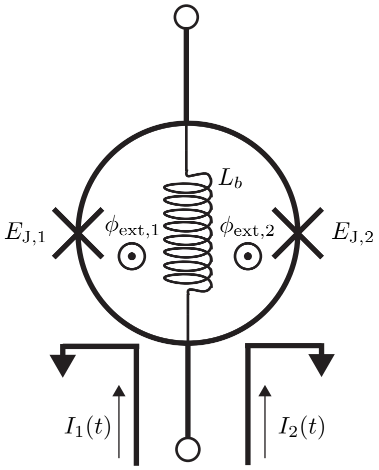


FIG. 17. Schematic diagram of an ATS. Two junctions with Josephson energies $E_{J,1}$, $E_{J,2}$ are connected in parallel, forming a SQUID. The SQUID loop in turn is 'split' in the middle by a linear inductor with inductance L_b, effectively forming two loops on either side of the inductor. The magnetic fluxes $\phi_{\text{ext},1}$ and $\phi_{\text{ext},2}$ threading the left and right loops, respectively, are controlled via externally applied, time-dependent currents $I_1(t)$, $I_2(t)$ that are buffered to ground in the vicinity of the loops using on-chip fluxlines.

In its most general form, the ATS potential is given by

$$U(\hat{\phi}) = \frac{1}{2}E_{L,b}\hat{\phi}^2 - 2E_J\cos(\phi_\Sigma)\cos(\hat{\phi}+\phi_\Delta) + 2\Delta E_J\sin(\phi_\Sigma)\sin(\hat{\phi}+\phi_\Delta), \quad \text{(A1)}$$

where $\hat{\phi}$ is the phase difference across the ATS, $\phi_\Sigma := (\phi_{\text{ext},1}+\phi_{\text{ext},2})/2$, $\phi_\Delta := (\phi_{\text{ext},1}-\phi_{\text{ext},2})/2$, and $\phi_{\text{ext},1}$ ($\phi_{\text{ext},2}$) is the magnetic flux threading the left (right) loop, in units of the reduced flux quantum $\Phi_0 = \hbar/2e$. Here $E_{L,b} = \Phi_0^2/L_b$, $E_J = (E_{J,1}+E_{J,2})/2$, and $\Delta E_J = (E_{J,1}-E_{J,2})/2$ is the junction asymmetry. This ATS potential can be further simplified by tuning ϕ_Σ and ϕ_Δ with two separate fluxlines, setting them to

$$\phi_\Sigma = \pi/2 + \epsilon_p(t), \quad \text{(A2)}$$

$$\phi_\Delta = \pi/2 \quad \text{(A3)}$$

where $\epsilon_p(t) = \epsilon_{p,0}\cos(\omega_p t)$ is a small ac component added on top of the dc bias. At this bias point, and assuming that $|\epsilon_p(t)| \ll 1$, Eq. (A1) reduces to

$$U(\hat{\phi}) = \frac{1}{2}E_{L,b}\hat{\phi}^2 - 2E_J\epsilon_p(t)\sin(\hat{\phi}) + 2\Delta E_J\cos(\hat{\phi}). \quad \text{(A4)}$$

2. Calculation of nonlinear interaction rate g_2

To make further progress, it is necessary to represent the nanomechanical element as an equivalent circuit that accurately captures its linear response. This can be done straightforwardly using the method of Foster synthesis, provided we know the admittance $Y_m(\omega)$ seen from the terminals of the mechanical resonator. This admittance can be accurately computed using modern FEM solvers. For further details on the piezoelectrics simulations, see Ref. [26].

The equivalent circuit (or "Foster network") is shown in Fig. 18(a) and in its simplest form consists of a 'dc capacitance' in series with an LC block, with an additional resistor (not shown) inserted to include the effects of loss in the resonator. We note that this "lossy Foster" method is not exact but is accurate enough for our purposes provided that losses are sufficiently small [90]. The linear part of the buffer resonator (including the inductor that splits the ATS) can also be represented as an LC block. In this representation the buffer and storage resonators are two linear circuits with a linear coupling and can therefore be diagonalized by a simple transformation of coordinates. The resulting "storage-like" ($\hat{a}$) and "buffer-like" ($\hat{b}$) eigenmodes both contribute to the total phase difference across the ATS, $\hat{\phi} = \varphi_a(\hat{a} + \hat{a}^\dagger) + \varphi_b(\hat{b} + \hat{b}^\dagger)$. These modes therefore mix via the ATS potential, which we redefine as $U(\hat{\phi}) \mapsto U(\hat{\phi}) - E_{L,b}\hat{\phi}^2/2$ because we already absorbed the inductor into the linear network. The vacuum fluctuation amplitudes of each mode mode are given by

$$\varphi_{k,j} = \sqrt{\frac{\hbar}{2\omega_k}}(C^{-1/2}U)_{jk}, \quad \text{(A5)}$$

where C is the Maxwell capacitance matrix of the circuit, U is the orthogonal matrix that diagonalizes $C^{-1/2}L^{-1}C^{-1/2}$, and L^{-1} is the inverse inductance matrix [91]. The index $k \in \{a, b\}$ labels the mode and j labels the node in question. Note that generally we omit the j index in our notation because the node of interest is clear from context (it is the one where the ATS is located).

The way in which the ATS mixes the modes is now explicitly clear: the third-order term in the Taylor series expansion of the $\sin(\hat{\phi})$ function in Eq. (A4) contains terms of the form $\hat{a}^2\hat{b}^\dagger$ + h.c., which is precisely the required coupling. This is the key reason for using an ATS as opposed to an ordinary junction, which has a potential $\sim \cos(\hat{\phi})$. Note also that a finite junction asymmetry $|\Delta E_J| > 1$ partially eliminates the benefit of using an ATS, as this introduces additional self- and cross-Kerr terms. For the remainder of this analysis we assume we are operating in the ideal case $\Delta E_J = 0$, noting that with state-of-the-art fabrication one can reliably achieve $\Delta E_J/E_J \sim 10^{-2}$ [92].

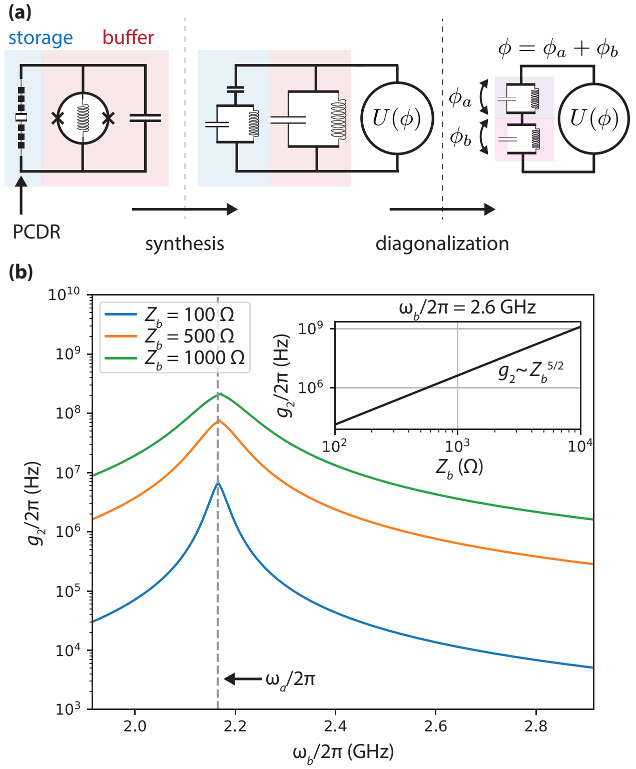


FIG. 18. Calculation of g_2. (a) Schematic summary of our method for calculating g_2. A PCDR, connected in parallel to a buffer resonator that is formed by shunting an ATS with a capacitance C_b, is synthesized as a simple Foster network with the same admittance function $Y_m(\omega)$ as the original piezoelectric structure. The Foster network consists of a parallel combination of an inductance L_a and a capacitance C_a, in series with a 'coupling capacitance' C_g. In turn, the linear components of the buffer resonator L_b and C_b are lumped together with the mechanical Foster circuit, leaving the nonlinear part of the ATS potential as an additional circuit element that we label by $U(\phi)$ in the diagram. The linear network is then diagonalized and the vacuum fluctuation amplitudes φ_a and φ_b of the storage-like and buffer-like eigenmodes are used to calculate g_2. (b) Dependence of g_2 on the buffer resonator frequency ω_b and impedance Z_b. The g_2 curves peak at the storage mode frequency ω_a where the modes are maximally hybridized. Inset: g_2 plotted as a function of Z_b for a fixed ω_b, showing the 5/2 power law dependence.

In order to select the desired terms one must set the pump frequency to $\omega_p = 2\omega_a - \omega_b$ [25]. This brings the term $g_2^*\hat{a}^2\hat{b}^\dagger$+h.c. into resonance and allows us to drop the other terms using a rotating-wave approximation (RWA). The coupling rate is given by $g_2 = (E_J/\hbar)\epsilon_{p,0}\varphi_a^2\varphi_b/2$. Additionally, a linear drive $\epsilon_d^*\hat{b}$+h.c. at frequency $\omega_d = \omega_b$ is added to supply the required energy for the two-phonon drive.

We now explicitly calculate g_2 *assuming that the storage resonator is a one-dimensional lithium niobate phononic-crystal-defect resonator (PCDR)* as reported in Ref. [28]. We use for its Foster network parameters the values $C_g = 0.385\,\mathrm{fF}$, $C_a = 1.682\,\mathrm{fF}$, and $L_a = 2.614\,\mu\mathrm{H}$, which in previous work have produced accurate estimates of the linear coupling rate between the phononic mode and other electrical circuits [28, 29]. These parameters set $\omega_a/2\pi \approx 2.17\,\mathrm{GHz}$ as the storage mode frequency, which will remain fixed for the reaminder of this Appendix. We further take $E_J/h = 90\,\mathrm{GHz}$ and $\epsilon_{p,0} = \pi/80$ as representative values that are experimentally realistic. [25]. We note that Ref. [25] did not explicitly report a value for $\epsilon_{p,0}$, but we inferred it by reproducing their reported value of g_2. In some instances we will set $\epsilon_{p,0}$ to an even smaller value, which we will indicate accordingly. In Fig. 18(b) we show g_2 plotted as a function of the buffer mode's frequency $\omega_b \approx 1/\sqrt{L_b C_b}$ for three different values of the impedance $Z_b = \sqrt{L_b/C_b}$. The two parameters ω_b and Z_b completely specify the properties of the buffer resonator for the purposes of this work. One salient feature is that g_2 scales as

$$g_2 \sim Z_b^{5/2}, \tag{A6}$$

which is due to the fact that $\varphi_b \sim \sqrt{Z_b}$ and $\varphi_a \sim Z_b$. This rapid scaling will prove useful later on, when we explore how to configure the system to minimize the dimensionless loss κ_1/κ_2.

3. Classical filter theory and derivation of dissipation rates

The above calculation of g_2 is only half the story, since we are ultimately interested in making accurate predictions of κ_1/κ_2. Indeed $\kappa_2 = 4g_2^2/\kappa_b$ in the simple two-mode model with the pump tuned perfectly on resonance $\omega_p = 2\omega_a - \omega_b$. However, as we show in Appendix B, in order to stabilize multiple modes with a single ATS (which is necessary to achieve the required connectivity for the surface code), it is a critical requirement to utilize a bandpass filter between the buffer resonator and the open $50\,\Omega$ port in order to protect the storage modes from radiative (Purcell) decay and to suppress unwanted correlated decay processes — see Fig. 19(a) for a sketch of the device. We therefore need a more general expression for the two-phonon dissipation rate κ_2 in the case where the bath that the b mode couples to is described by a general admittance function $Y(\omega)$. We begin with the Hamiltonian of the closed system comprising the storage mode a and the buffer mode b, neglecting dissipation:

$$H = \frac{1}{2}q^T C^{-1} q + \frac{1}{2}\Phi^T L^{-1}\Phi - 2E_J\epsilon_p(t)\sin(\phi_2), \tag{A7}$$

$q = (q_1, q_2)^T$, $\Phi = (\Phi_1, \Phi_2)^T$, $\Phi_j = \int dt V_j(t)$ is the node flux at node j (with the voltage V_j defined with respect to the ground node), and

$$C = \begin{pmatrix} C_a + C_g & -C_g \\ -C_g & C_b + C_g \end{pmatrix}, \quad L^{-1} = \begin{pmatrix} L_a^{-1} & 0 \\ 0 & L_b^{-1} \end{pmatrix}. \tag{A8}$$

We are also using the notation $\phi_j := \Phi_j/\Phi_0$ for the dimensionless flux, where $\Phi_0 = \hbar/2e$ is the reduced flux quantum. The equations of motion (EOMs) are

$$\begin{aligned} \dot{\Phi} &= \partial_q H = C^{-1}q, \\ \dot{q} &= -\partial_\Phi H = -L^{-1}\Phi + 2I_J\epsilon_p(t)\cos\phi_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \end{aligned} \tag{A9}$$

where we defined $I_J := E_J/\Phi_0$. Note that the charge EOM in Eq. (A9) is simply Kirchhoff's current law (KCL). To include the effect of the external admittance $Y(\omega)$, which describes both the filter and the $50\,\Omega$ output line, we add an additional source of current $I_s(t)$ flowing into node 2:

$$I_s(t) = \int_{-\infty}^{\infty} d\omega Y(\omega)\dot{\Phi}_{F,2}(\omega)e^{i\omega t} \tag{A10}$$

$$= \int_{-\infty}^{\infty} d\omega Y(\omega)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} dt'\dot{\Phi}_2(t')e^{-i\omega t'}\right]e^{i\omega t} \tag{A11}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} dt'\int_{-\infty}^{\infty} d\omega Y(\omega)\dot{\Phi}_2(t')e^{i\omega(t-t')}, \tag{A12}$$

where $\dot{\Phi}_{F,2}(\omega)$ is the Fourier transform of the voltage $\dot{\Phi}_2(t)$. Combining the EOMs Eq. (A9) and adding the source term, we find

$$C\ddot{\Phi}(t) + L^{-1}\Phi(t) = F(t)\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{A13}$$

where $F(t)$ is defined as

$$F(t) \equiv 2I_J\epsilon_p(t)\cos\phi_2(t) - \frac{1}{2\pi}\int dt'\int d\omega Y(\omega)\dot{\Phi}_2(t')e^{i\omega(t-t')}. \tag{A14}$$

Here both integrals run from $-\infty$ to $+\infty$. We will use this convention for the remainder of this section for notational simplicity, unless otherwise stated. Let $\Phi' = C^{1/2}\Phi$. Then Eq. (A13) becomes

$$\ddot{\Phi}'(t) + C^{-1/2}L^{-1}C^{-1/2}\Phi'(t) = F(t)C^{-1/2}\begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{A15}$$

We now diagonalize the matrix $C^{-1/2}L^{-1}C^{-1/2}$ as

$$C^{-1/2}L^{-1}C^{-1/2} = U\Omega^2 U^T, \tag{A16}$$

where $\Omega = \mathrm{diag}(\omega_a, \omega_b)$ is a diagonal matrix containing the normal mode frequencies and U is an orthogonal

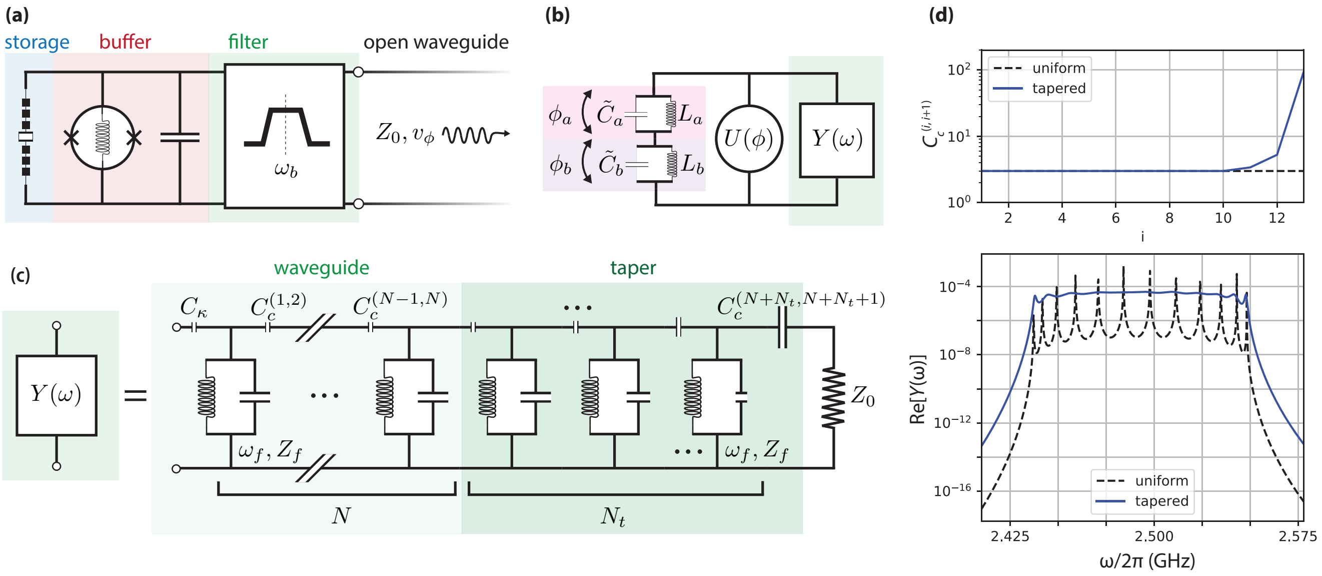


FIG. 19. Filter design. (a) Schematic of the filtering setup. A bandpass filter centered at frequency ω_b is placed in between the output of the buffer resonator and an open waveguide with characteristic impedance Z_0 and phase velocity v_ϕ. Photons that are transmitted through the filter enter the semi-infinite waveguide and are irreversibly lost. (b) Circuit diagram showing the normal modes a and b and their connection to the filter described by an admittance function $Y(\omega)$. (c) Detailed circuit diagram for the filter structure, which consists of a main waveguide section with N "unit cells" followed by a taper section with N_t cells, terminated at the end with a load resistance Z_0 that accurately models the infinite waveguide at the output port. Every cell of the filter has frequency ω_f and impedance Z_f, and neighboring cells are coupled capacitively with capacitances $C_c^{(i,i+1)}$. The coupling capacitance C_κ between the buffer resonator and the first filter cell is defined separately for generality. (d) Top: coupling capacitances plotted as a function of cell index i for tapered ($(N, N_t) = (10, 3)$) and uniform ($(N, N_t) = (13, 0)$) filters. The tapered structure, found automatically by a Nelder-Mead optimizer, is characterized by a rapid increase in C_c near the end of the structure. Bottom: typical filter response, here shown as the real part of $Y(\omega)$ for tapered and uniform filters. The response of the uniform structure shows multiple sharp peaks, each corresponding to a standing-wave resonance of the structure, whereas the tapered response is relatively flat throughout the filter passband. In effect, the taper allows propagating waves to be transmitted to the external waveguide over a broad bandwidth.

matrix. The normal modes are

$$\Phi'' = U^T\Phi' = U^T C^{-1/2}\Phi = (\Phi_1'', \Phi_2'')^T. \tag{A17}$$

In terms of Φ'', the flux EOM Eq. (A15) is given by

$$\ddot{\Phi}''(t) + \Omega^2\Phi''(t) = F(t)U^T C^{-1/2}\begin{pmatrix}0\\1\end{pmatrix} \tag{A18}$$

$$= F(t)\begin{pmatrix}(U^T C^{-1/2})_{12}\\(U^T C^{-1/2})_{22}\end{pmatrix} \tag{A19}$$

$$= F(t)\begin{pmatrix}(C^{-1/2}U)_{21}\\(C^{-1/2}U)_{22}\end{pmatrix}, \tag{A20}$$

where we have used the fact that C (and therefore $C^{-1/2}$) is symmetric. If we define $\Phi_a := (C^{-1/2}U)_{21}\Phi_1''$ and $\Phi_b := (C^{-1/2}U)_{22}\Phi_2''$, Eq. (A18) can be written more neatly as

$$\tilde{C}_a\ddot{\Phi}_a + \tilde{C}_a\omega_a^2\Phi_a = \tilde{C}_b\ddot{\Phi}_b + \tilde{C}_b\omega_b^2\Phi_b = F(t), \tag{A21}$$

where

$$\tilde{C}_a := (C^{-1/2}U)_{21}^{-2}, \quad \tilde{C}_b := (C^{-1/2}U)_{22}^{-2} \tag{A22}$$

are the effective capacitances of the a and b normal modes. Eq. (A18) is KCL for a different network — one where two LC stages, one for each of the normal modes, are placed in series with each other. The series combination is in turn connected to the filtered environment $Y(\omega)$ and the ATS potential $U(\Phi)$ (see Fig. 19(b)). Note that this diagonalization procedure is completely equivalent to synthesizing a Foster network representing the coupled storage and buffer resonators, for example as done in black-box quantization [90].

Note that $\Phi = C^{-1/2}U\Phi''$, and in particular

$$\Phi_2 = (C^{-1/2}U)_{21}\Phi_1'' + (C^{-1/2}U)_{22}\Phi_2'' \tag{A23}$$

$$= \Phi_a + \Phi_b. \tag{A24}$$

In terms of these normal mode amplitudes, $F(t)$ is given by

$$F(t) = 2I_J\epsilon_p(t)\cos\left[\phi_a(t) + \phi_b(t)\right] - \frac{1}{2\pi}\int dt'\int d\omega Y(\omega)\left[\dot{\Phi}_a(t') + \dot{\Phi}_b(t')\right]e^{i\omega(t-t')}. \tag{A25}$$

We now define the following dimensionless, time-varying mode amplitudes:

$$a(t) := \frac{1}{\sqrt{2\hbar}}\left[\sqrt{\tilde{C}_a\omega_a}\Phi_a(t) + i\frac{1}{\sqrt{\tilde{C}_a\omega_a}}\tilde{C}_a\dot{\Phi}_a(t)\right],$$
$$b(t) := \frac{1}{\sqrt{2\hbar}}\left[\sqrt{\tilde{C}_b\omega_b}\Phi_b(t) + i\frac{1}{\sqrt{\tilde{C}_b\omega_b}}\tilde{C}_b\dot{\Phi}_b(t)\right]. \tag{A26}$$

Defining $\varphi_j = \Phi_0^{-1}\sqrt{\hbar/2\omega_j\tilde{C}_j}$, where $j \in \{a, b\}$, we have

$$\phi_a = \varphi_a(a + a^\dagger), \;\; \phi_b = \varphi_b(b + b^\dagger). \tag{A27}$$

Here the † symbol indicates complex conjugation. *We identify φ_j as the amplitude of the vacuum fluctuations of the phase at node 2 due to mode j.*

It is straightforward to show that the EOMs of these "annihilation variables" are

$$\dot{a}(t) = -i\omega_a a(t) + i(\Phi_0/\hbar)\varphi_a F(t),$$
$$\dot{b}(t) = -i\omega_b b(t) + i(\Phi_0/\hbar)\varphi_b F(t). \tag{A28}$$

In terms of a and b, the source term $F(t)$ is given by

$$F(t) = 2I_J\epsilon_p(t)\cos\left[\varphi_a(a(t) + a^\dagger(t)) + \varphi_b(b(t) + b^\dagger(t))\right]$$
$$- \frac{1}{2\pi}\int dt'\int d\omega Y(\omega)\left[\frac{i\hbar}{2\tilde{C}_a\Phi_0\varphi_a}(a^\dagger(t') - a(t')) + \frac{i\hbar}{2\tilde{C}_b\Phi_0\varphi_b}(b^\dagger(t') - b(t'))\right]e^{i\omega(t-t')}. \tag{A29}$$

We now invoke the rotating wave approximation (RWA) and neglect terms that are fast-rotating, namely $a^\dagger(t')$ and $b^\dagger(t')$ in both EOMs and $a(t')$ and $b(t')$ in the EOMs for b and a, respectively. This is well-justified in the regime where ω_a, ω_b, and $|\omega_a - \omega_b|$ are all much larger than the dissipation rates $\mathrm{Re}[Y]/2\tilde{C}_j$, $j \in \{a, b\}$. We will see shortly that indeed these quantities emerge as dissipation rates from our analysis, so this assumption is self-consistent. The EOMs Eq. (A28) then become

$$\dot{a}(t) = -i\omega_a a(t) - \frac{1}{2\pi}\int dt'\int d\omega\frac{Y(\omega)}{2\tilde{C}_a}a(t')e^{i\omega(t-t')} + 2i(E_J/\hbar)\epsilon_p(t)\varphi_a\cos\left[\varphi_a(a(t) + a^\dagger(t)) + \varphi_b(b(t) + b^\dagger(t))\right]$$
$$\dot{b}(t) = -i\omega_b b(t) - \frac{1}{2\pi}\int dt'\int d\omega\frac{Y(\omega)}{2\tilde{C}_b}b(t')e^{i\omega(t-t')} + 2i(E_J/\hbar)\epsilon_p(t)\varphi_b\cos\left[\varphi_a(a(t) + a^\dagger(t)) + \varphi_b(b(t) + b^\dagger(t))\right]. \tag{A30}$$

We now go to an "interaction frame" (or rotating frame) defined by the transformations

$$a(t) \mapsto a(t)e^{i\omega_a t}, \tag{A31}$$
$$b(t) \mapsto b(t)e^{i(\omega_b+\Delta)t}, \tag{A32}$$

and explicitly add the flux pump

$$\epsilon_p(t) = \epsilon_{p,0}\cos\omega_p t, \;\; \omega_p = 2\omega_a - \omega_b - \Delta, \tag{A33}$$

which was introduced in Appendix A 1. We have also added a detuning Δ to keep the analysis general and also because finite Δ is a key requirement for multiplexed stabilization — see Appendix B. Expanding the cosine term to second order and keeping only the resonant terms, we find:

$$\begin{aligned}\dot{a}(t) &= -\frac{1}{2\pi}\int dt'\int d\omega\frac{Y(\omega)}{2\tilde{C}_a}a(t')e^{i(\omega+\omega_a)(t-t')} + 2ig_2a^\dagger(t)b(t),\\ \dot{b}(t) &= i\Delta b(t) - \frac{1}{2\pi}\int dt'\int d\omega\frac{Y(\omega)}{2\tilde{C}_b}b(t')e^{i(\omega+\omega_b+\Delta)(t-t')} + ig_2a^2(t),\end{aligned} \tag{A34}$$

where $g_2 := (E_J/\hbar)\epsilon_{p,0}\varphi_a^2\varphi_b/2$.

The EOMs Eq. (A34) do not have simple solutions in general because they are non-local in time. However, we can drastically simplify them — and re-cast them into a form that is time-local — under a specific regime of interest, which we describe next. First, note that

$$\int dt'\int d\omega Y(\omega)b(t')e^{i(\omega+\delta)(t-t')} = \int dt' Y_T(t-t')b(t')e^{i\delta(t-t')}, \tag{A35}$$

where δ equals either ω_a or $\omega_b+\Delta$ depending on which EOM we are referring to, and $Y_T(t)$ is the Fourier transform of the admittance function $Y(\omega)$:

$$Y_T(t) := \int d\omega Y(\omega)e^{i\omega t}. \tag{A36}$$

Now suppose for illustration that $Y(\omega)$ is a simple function

$$Y(\omega) = \begin{cases} Y_0 & |\omega| \leq 2J \\ 0 & |\omega| > 2J, \end{cases} \tag{A37}$$

which describes an "ideal" filter with bandwidth J. We note this is not a physical admittance function and we are using this simply as an example — in particular, it doesn't satisfy certain basic properties such as causality. Its Fourier transform is

$$Y_T(t) = (2Y_0J)\frac{\sin(2Jt)}{Jt}, \tag{A38}$$

so $|Y_T(t-t')e^{i\delta t}| = |Y_T(t-t')|$ is localized in the range defined by $J|t-t'| \sim 1$. Therefore, assuming $b(t')$ evolves much more slowly compared to the timescale $1/J$, the following approximation holds:

$$\int dt' Y_T(t-t')e^{i\delta(t-t')}b(t') \tag{A39}$$

$$\approx \int dt' Y_T(t-t')e^{i\delta(t-t')}b(t) \tag{A40}$$

$$= \int dt' Y_T(t')e^{i\delta t'}b(t) \tag{A41}$$

$$= 2\pi Y^*(\delta)b(t), \tag{A42}$$

where in the last line we used $Y(-\delta) = Y^*(\delta)$. We shall verify shortly that this slowness assumption is self-consistent. For now, this approximation transforms the EOMs Eq. (A34) to the following form:

$$\begin{aligned}\dot{a}(t) &= -\frac{\kappa_1}{2}a(t) + 2ig_2a^\dagger(t)b(t),\\ \dot{b}(t) &= \left[i\tilde{\Delta} - \frac{\kappa_{b,\mathrm{eff}}(\Delta)}{2}\right]b(t) + ig_2a^2(t) + \epsilon_d.\end{aligned} \tag{A43}$$

Here $\kappa_1 := \mathrm{Re}\left[Y^*(\omega_a)\right]/\tilde{C}_a$ and $\kappa_{b,\mathrm{eff}}(\Delta) := \mathrm{Re}\left[Y^*(\omega_b+\Delta)\right]/\tilde{C}_b$ are the effective *linear* dissipation rates of the a and b modes, respectively. We have also added an additional drive term ϵ_d (which rotates at frequency $\omega_b+\Delta$ in the lab frame and therefore here it is static), and defined $\tilde{\Delta} := \Delta - \mathrm{Im}\left[Y^*(\omega_b+\Delta)\right]/2\tilde{C}_b$, which now includes the frequency shift of the b mode due to its coupling to the filter. Note we have also neglected the corresponding shift $-\mathrm{Im}\left[Y^*(\omega_a)\right]/2\tilde{C}_a$ of the a mode, since this is negligibly small for the purposes of this analysis.

Let us now find an effective description of the a mode alone, valid in a regime where the linear dissipation rate $\kappa_{b,\mathrm{eff}}$ is large (in a sense that will be made rigorous shortly). This procedure is the classical analogue of the formal adiabatic elimination procedure used in Appendix B 1. Let us assume that $\dot{b}(t) = 0$, i.e. the b mode is evolving sufficiently slowly that the time derivative can be neglected. Then Eq. (A43) becomes

$$b(t) = \frac{ig_2a^2(t) + \epsilon_d}{-i\tilde{\Delta} + \kappa_{b,\mathrm{eff}}(\Delta)/2}, \tag{A44}$$

and

$$\dot{a}(t) = -\frac{\kappa_1}{2}a(t) - \kappa_2a^\dagger(t)a^2(t) + \alpha_da^\dagger(t), \tag{A45}$$

where

$$\kappa_2(\Delta) := \mathrm{Re}\left[\frac{4g_2^2}{-2i\tilde{\Delta}+\kappa_{b,\mathrm{eff}}(\Delta)}\right] \tag{A46}$$

$$= \frac{4g_2^2}{4\tilde{\Delta}^2+\kappa_{b,\mathrm{eff}}^2(\Delta)}\kappa_{b,\mathrm{eff}}(\Delta), \tag{A47}$$

and $\alpha_d := 2ig_2\epsilon_d[-i\tilde{\Delta}+\kappa_b/2]^{-1}$. As a final step, let us linearize the EOMs around the static solutions $a = \pm\alpha$ given by setting $\dot{a}(t) = 0$. Assuming $2\kappa_2|\alpha|^2 \gg \kappa_1$, the solutions are $\alpha = \pm\sqrt{\epsilon_d/g_2}$. Defining $d := a - \alpha$ as the "fluctuations" around these fixed points, the linearized equation of motion for $d(t)$ becomes

$$\dot{d}(t) = -\frac{\kappa_1}{2}d(t) - 2\kappa_2|\alpha|^2 d(t) \approx -\kappa_{\mathrm{conf}}d(t), \tag{A48}$$

where $\kappa_{\mathrm{conf}} := 2|\alpha|^2\kappa_2$. We call this rate the confinement rate in keeping with existing terminology [25]. Applying this linearization to Eq. (A44), we find

$$b(t) = \frac{2ig_2\alpha}{-i\tilde{\Delta}+\kappa_{b,\mathrm{eff}}(\Delta)/2}d(t) + const. \tag{A49}$$

The rate κ_2 we previously defined is now manifestly the two-phonon dissipation rate that we wanted to find, as it sets the rate κ_{conf} at which fluctuations away from the fixed points $a = \pm\alpha$ decay back into the "code space". It reduces to the familiar form $\kappa_2 = 4g_2^2/\kappa_{b,\mathrm{eff}}$ in the case of a perfectly resonant pump $\tilde{\Delta} = 0$, and to the form $\kappa_2 = (g_2/\tilde{\Delta})^2\kappa_{b,\mathrm{eff}}$ in the far off-resonant limit $|\tilde{\Delta}| \gg \kappa_{b,\mathrm{eff}}$. This latter form is indeed equivalent to the expressions for κ_2 derived in Appendix B, where the filter is modeled as a linear chain of oscillators with nearest-neighbor linear couplings. Here, the function $\kappa_{b,\mathrm{eff}}(\Delta)$ contains all the information about the filtered environment, capturing effects such as the exponential suppression of κ_2 when Δ lies outside of the filter passband. Finally, we note that the straightforward linearization procedure above is the classical analogue of the shifted Fock basis technique described in Appendix C.

Let us go back and re-examine the two main assumptions that we have made so far: 1) that $b(t')$ evolves much more slowly compared to the filter response timescale $1/J$, and 2) the adiabatic assumption that $\dot{b}(t) = 0$ in Eq. (A43).

First, by inspecting equation Eq. (A48) we can extract the effective timescale of the dynamics of d mode. We see that d evolves on a timescale $1/|\alpha|^2\kappa_2$ (assuming $|\alpha|^2\kappa_2 \gg \kappa_1$, which is the regime we are interested in). Therefore, from the solution for $b(t)$ in Eq. (A49) we infer that the b mode also evolves on this timescale. The slowness assumption that led to Eq. (A43) is therefore self-consistent as long as $|\alpha|^2\kappa_2 \ll J$. Furthermore, even though we used a 'toy model' for $Y(\omega)$ to illustrate the required hierarchy of timescales, we verified numerically using the simulations in Appendix A 4 that this exact logic remains valid even when $Y(\omega)$ describes a real, appropriately designed filter.

Second, under which conditions is the adiabatic elimination $\dot{b}(t) = 0$ valid? The solution for $b(t)$ in Eq. (A49), obtained by assuming $\dot{b}(t) = 0$, evolves on the same timescale $1/|\alpha|^2\kappa_2$ as $d(t)$. Therefore the adiabatic elimination step is self-consistent so long as $|\alpha|^2\kappa_2 \ll \kappa_b$, because $1/\kappa_b$ is the timescale in which $b(t)$, as described by the full EOM Eq. (A43), converges to its steady state. Since $\kappa_2(\Delta) \leq \kappa_2(0)$, this condition is equivalent to $2|\alpha|g_2 \ll \kappa_b$:

$$|\alpha|^2\kappa_2 \ll \kappa_b \iff |\alpha|^2\kappa_2(0) = 4|\alpha|^2 g_2^2/\kappa_b \ll \kappa_b$$
$$\iff 2|\alpha|g_2 \ll \kappa_b. \tag{A50}$$

For the purposes of this work we shall assume that $2|\alpha|g_2 = \eta\kappa_b$ is sufficient, for some small number $\eta < 1$. Using time-domain master equation simulations (not shown) we have verified that using $\eta = 1/5$ is sufficient to stabilize the storage mode.

4. Filter design

Here we turn to the problem of filter design. What should we use as the physical embodiment of the filtered environment described by $Y(\omega)$? We can start by outlining some general design principles based on the preceding analysis. First, recall that the effective dissipation rate of the b mode is $\kappa_{b,\mathrm{eff}}(\Delta) = \mathrm{Re}\left[Y^*(\omega_b+\Delta)\right]/\tilde{C}_b$, and second, note that the two-phonon dissipation rate is given by Eq. (A46), which we repeat here for convenience: $\kappa_2(\Delta) = 4g_2^2\kappa_{b,\mathrm{eff}}(\Delta)\left[4\tilde{\Delta}^2+\kappa_{b,\mathrm{eff}}^2(\Delta)\right]^{-1}$. As discussed in Appendix B, different values of Δ are required to stabilize multiple modes with a single ATS — one value for each mode. Therefore, the function $\kappa_2(\Delta)$ should be constant — and as large as possible — over a certain band of frequencies $B = [\omega_b - \Delta_{\max}, \omega_b + \Delta_{\max}]$. In effect, there should be a finite density of states that the b mode can radiate into within this band. Outside of this band, however, the density of states should vanish in order to suppress correlated phase-flip errors resulting from the multiplexed stabilization (see Appendix B). These requirements translate to a simple design principle: the function $\mathrm{Re}\left[Y(\omega)\right]$ should ideally be a constant in the band $\omega \in B$, and zero elsewhere, much like in the toy model discussed in Appendix A 3 where we took $\Delta_{\max} = 2J$. This is akin to a resistor that only absorbs radiation at certain frequencies.

a. *General properties*

One of the simplest possible networks with these properties is a linear chain of N LC resonators with capacitive couplings, as shown in Fig. 19(c). This "metamaterial

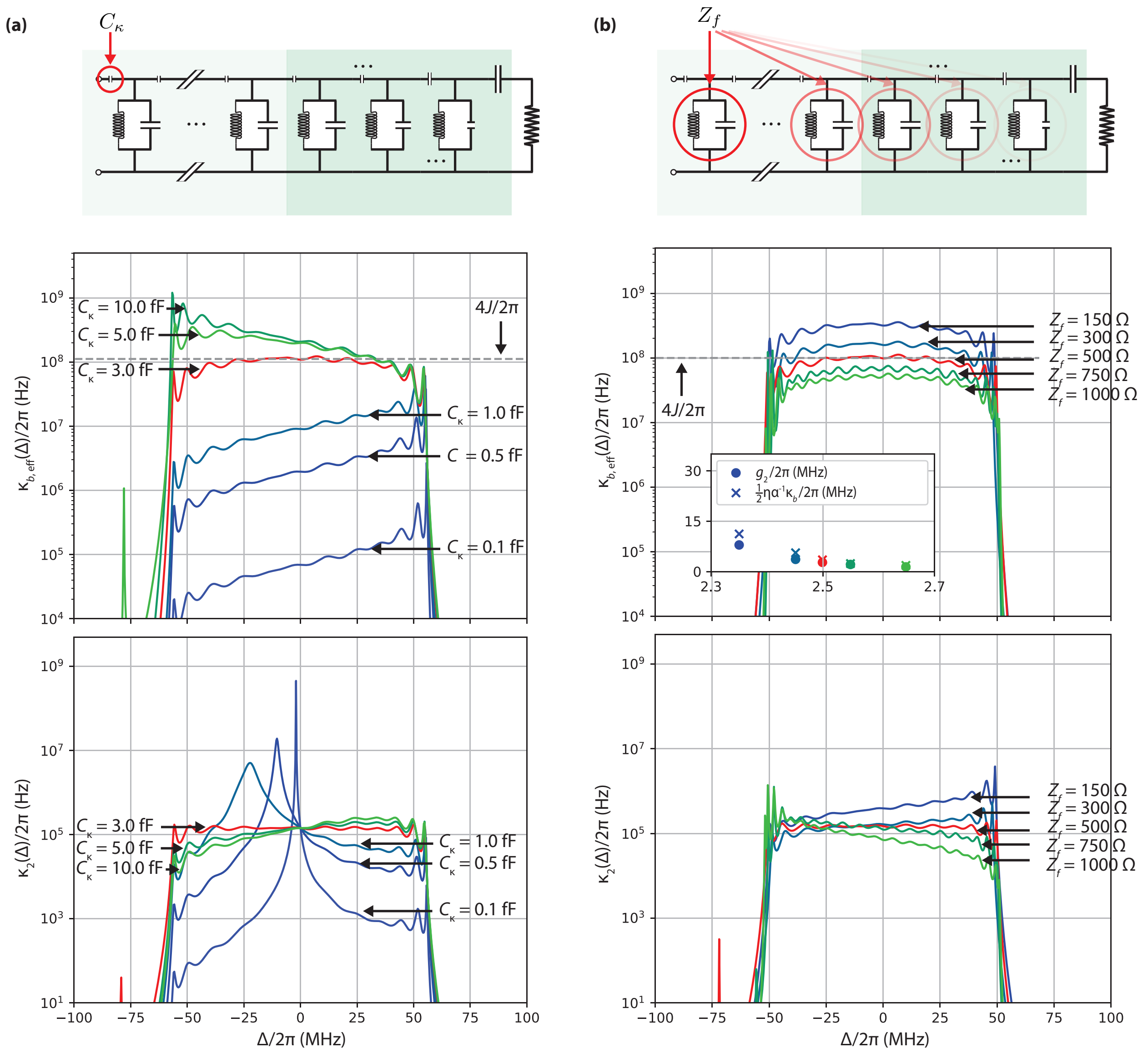


FIG. 20. Exploring the filter design space. (a) Dependence of $\kappa_{b,\mathrm{eff}}(\Delta)$ and $\kappa_2(\Delta)$ on the coupling capacitance C_κ between the b mode and the first filter resonator. Here we fix $\epsilon_{p,0} = 0.015$, $\omega_f/2\pi = 2.55\,\mathrm{GHz}$, $Z_f = 500\,\Omega$, $\omega_b = \omega_f - 2J$, $Z_b = 1\,\mathrm{k}\Omega$, $C_c = 3.0\,\mathrm{fF}$, and $(N, N_t) = (10, 3)$. As C_κ increases, we observe two regimes: an 'undercoupled' regime $C_\kappa \ll C_c$ characterized by a sharply peaked $\kappa_2(\Delta)$, where the narrow b mode filters the dissipation process, and an 'overcoupled' regime $C_\kappa \gg C_c$ where κ_2 saturates and becomes asymmetric. In this latter regime the b mode strongly hybridizes with the first filter cell. For large enough C_κ, their normal mode frequencies shift outside of the filter passband, forming bound resonances that are visible as sharp peaks to the left of the passband in some of the curves. The optimal value is $C_\kappa = C_c = 3.0\,\mathrm{fF}$, where $\kappa_2(\Delta)$ is maximized and flat, is shown in red. Note that at this optimal coupling, $\kappa_{b,\mathrm{eff}} = 4J$ (gray dashed line). Note also that the adiabatic condition $g_2 < \eta\kappa_b/2\alpha$ is not respected for several of the plots shown, as g_2 is fixed. (b) Dependence of $\kappa_{b,\mathrm{eff}}(\Delta)$ and $\kappa_2(\Delta)$ on the characteristic filter impedance Z_f. In order to keep the filter bandwidth $4J$ constant, increasing Z_f requires decreasing C_c, and to keep $g_2 < \eta\kappa_b/2\alpha$ (adiabatic threshold), increasing Z_f requires increasing ω_f (which decreases g_2 due to the larger detuning between the a and b modes). The values used for the plotted curves are $\omega_f/2\pi = 2.4$, 2.5, 2.55, 2.6, $2.7\,\mathrm{GHz}$, $C_c = 10$, 4.5, 2.7, 1.7, $1.2\,\mathrm{fF}$, $\omega_b = \omega_f - 2J$, and $(N, N_t) = (10, 3)$, $(10, 3)$, $(10, 3)$, $(14, 6)$, $(14, 6)$. Larger values of Z_f required larger N_t to compensate for the larger impedance mismatch to the $50\,\Omega$ line. We also fix $Z_b = 1\,\mathrm{k}\Omega$ here. The optimal value is $Z_f = Z_b/2 = 500\,\Omega$, which produces a flat $\kappa_2(\Delta)$ curve (shown in red). Also note that at this optimal value, $\kappa_{b,\mathrm{eff}} = 4J$ (gray dashed line). Inset: g_2 and $\eta\kappa_b/2\alpha$ corresponding to each of the simulations for the different values of Z_f, plotted as a function of ω_b, showing the adiabatic constraint $g_2 < \eta\kappa_b/2\alpha$ is satisfied (here $\alpha = \sqrt{8}$ and $\eta = 1/5$).

waveguide" has a well-defined band with dispersion [93]

$$\omega(k) = \omega_f + 2J[\cos(\pi k/N) - 1],\ \ k \in \{0, ..., N-1\}. \tag{A51}$$

Here J is the coupling rate between neighboring resonators and is approximately given by

$$J \approx \frac{\omega_f}{2}\frac{C_c}{C_f + 2C_c}, \tag{A52}$$

where ω_f is the resonance frequency of each LC block, C_c is the coupling capacitance, and C_f is the shunt capacitance. This rate is directly tied to the filter bandwidth,

$$(\text{bandwidth}) = 4J, \tag{A53}$$

and is controllable via C_c. Note also that we usually specify the frequency ω_f and impedance Z_f of each LC block of the filter, which together with C_c uniquely specify the shunt inductance $L_f = Z_f/\omega_f$ and shunt capacitance $C_f = 1/\omega_f Z_f$. Usually $C_f \gg C_c$, so

$$J \sim \frac{1}{2}\omega_f^2 C_c Z_f. \tag{A54}$$

This means that for fixed values of ω_f and C_c, the filter bandwidth is directly proportional to Z_f. This formula will be useful shortly.

Normally the N filter modes with dispersion relation Eq. (A51) are standing waves that span the entire waveguide. These modes would therefore hybridize with the b mode, effectively forming a "multimode buffer" with $N+1$ sharp resonances that the a mode interacts with via the ATS. This is not the behavior we are interested in. Instead of a structure supporting standing resonances, we ideally seek a waveguide that is perfectly transparent to photons with frequency $\omega \in B$ and perfectly reflective otherwise. One way to achieve this is to add a small number of additional resonators at the end of the waveguide and rapidly ramp up the values of the coupling capacitances $C_c^{(i,i+1)}$ between neighboring cells (see Fig. 19(d)). We refer to this region as the 'taper' in keeping with existing terminology [93]. The shunt capacitances are also adjusted in order to keep the frequency of each cell constant across the filter, including the taper cells. The effect of the taper is to significantly broaden the resonances of the structure so that the entire B band is filled by their overlapping lineshapes, or alternatively, it allows the waves that propagate along the waveguide to be transmitted to the outside $50\,\Omega$ environment with negligible reflections.

We show in Fig. 19(d) the typical response of such a filter. The taper parameters (coupling capacitances and shunt capacitances) have been chosen to minimize the cost function $C = -\sum_{\omega\in B}\log\text{Re}[Y(\omega)]$, producing a relatively flat response over the band of interest B. We note that this choice of cost function is only a design heuristic that approximately produces the desired response.

b. *Optimizing the filter*

Given fixed properties ω_a, ω_b, Z_b, etc. of the coupled storage-buffer system, what is the optimal choice of filter parameters? By now it should be self-evident what we mean by "optimal": those which maximize the two-phonon dissipation rate $\kappa_2(\Delta)$ across the filter band $\{\omega_b + \Delta \in B\}$ and make it as flat (constant) as possible within B. There are many parameters that describe the filter: C_κ, C_c, ω_f, Z_f, N (the number of "unit cells"), N_t (the number of "taper cells"), and the set of coupling capacitances $\{C_c^{(i,i+1)}\}$ in the taper region. For fixed values of these first six parameters, the set $\{C_c^{(i,i+1)}\}$ is automatically optimized using the method described in the preceding paragraph, leaving six free parameters. What we show next is how to choose these parameters in order to optimize the function of interest $\kappa_2(\Delta)$ while simultaneously respecting the following constraints:

1. $4J/2\pi = 100\,\text{MHz}$
2. $\omega_b = \omega_f - 2J$
3. $g_2 < \eta\kappa_b/2\alpha$

Constraint (1) is to ensure that photons created as a result of correlated decay of multiple storage modes during multiplexed stabilization have frequencies $\omega_{\text{corr. decay}} \notin B$ outside of the passband. This prevents these photons from radiating into the environment and suppresses correlated phase-flip errors. The value $4J/2\pi = 100\,\text{MHz}$ is approximately the largest possible bandwidth the filter can have while still satisfying this requirement — for further detail see Appendix B. Constraint (2) sets the b mode frequency exactly in the middle of the passband, making the functions $\kappa_{b,\text{eff}}(\Delta)$ and $\kappa_2(\Delta)$ symmetric. This is not absolutely necessary but is rather a matter of convenience. Constraint (3) is to ensure that the system is in a regime where adiabatic elimination is valid, as found at the end of Appendix A 3. Here we fix $\alpha = \sqrt{8}$ and $\eta = 1/5$. Finally, we comment on what are reasonable values for N and N_t. The number of taper cells N_t depends on Z_f and $Z_0 (= 50\,\Omega)$, with N_t needing to be larger the farther Z_f deviates from Z_0. This agrees with the intuition that the taper is acting as an impedance-matching network. Once Z_f and N_t are chosen, we observe numerically that it is sufficient to choose a number of unit cells $N \gtrsim N_t/2$. Anything larger than this is unnecessary and does not change the results — the waveguide being longer does not affect the dissipation rates we are interested in calculating.

In Fig. 20(a) we show the effect of varying the capacitance C_κ, which sets the strength of coupling between the b mode and the first resonator in the filter waveguide. Here $C_c = 3.0\,\text{fF}$ is fixed, as well as $Z_f = 500\,\Omega$. We observe two "regimes": a weak-coupling regime defined by $C_\kappa \ll C_c$, where κ_b is small and $\kappa_2(\Delta)$ is sharply peaked near $\Delta = 0$. This peak indicates that the b mode is filtering the conversion process $g_2^* a^2 b^\dagger + \text{h.c.}$, only allowing

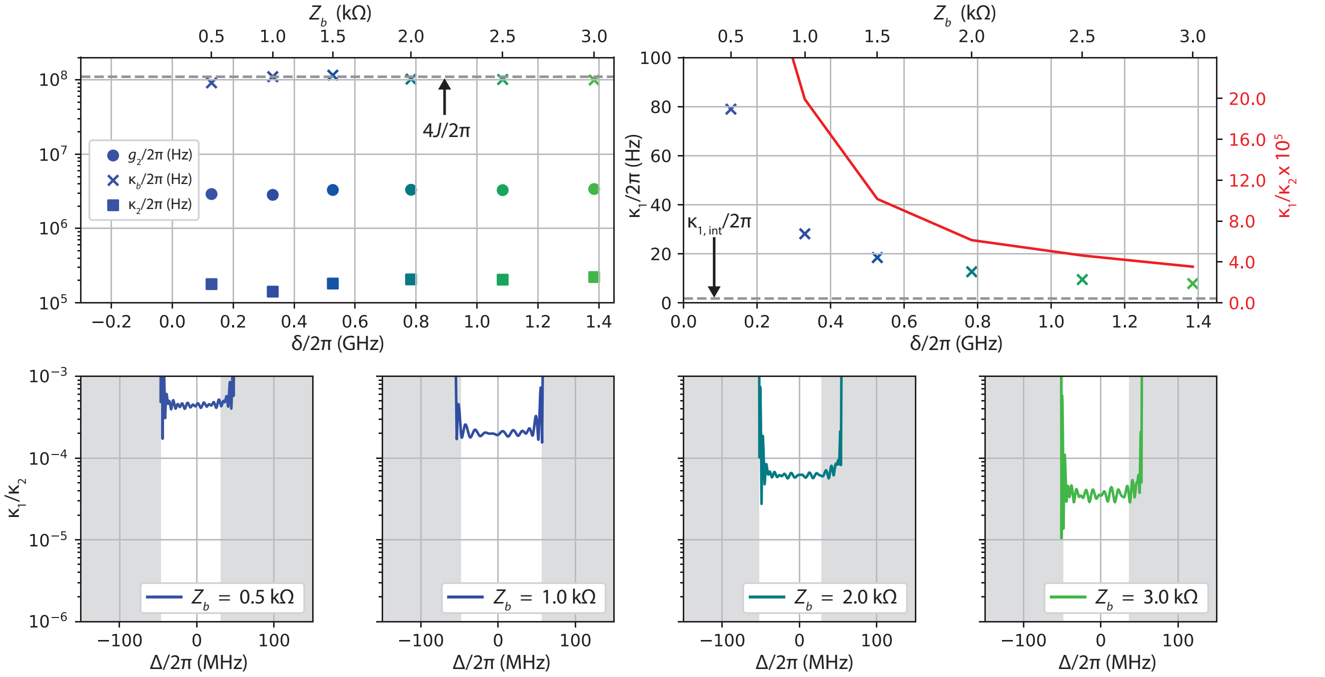


FIG. 21. Behavior of κ_1/κ_2 in the large Z_b limit. Top left panel: g_2, κ_b and κ_2 plotted as a function of Z_b. The latter two rates are averages over the middle of filter passband, $\omega \in [\omega_b - J, \omega_b + J]$. The detunings $\delta = \omega_b - \omega_a$ corresponding to each similuation are also indicated on the horizontal axis. Each result is obtained by optimizing the filter for each value of Z_b, following the procedure outlined in Appendix A 4. We observe that all of these rates remain practically constant, an in particular $\kappa_b = 4J$. Top right panel: single-phonon relaxation rate κ_1 plotted as a function of Z_b and δ. This relaxation rate $\kappa_1 = \kappa_{1,\text{int}} + \kappa_{1,\text{purcell}}$ includes two contributions: the intrinsic loss $\kappa_{1,\text{int}}$ of the resonator, which here we assume has a fixed intrinsic quality factor $Q_{a,\text{int}} = \omega_a/\kappa_{1,\text{int}} = 10^9$, and the Purcell loss $\kappa_{1,\text{purcell}}$ due to its coupling to the buffer resonator. This latter rate has a contribution due to radiation into the waveguide (which is vanishingly small due to the strong filter suppression), and an important contribution $\sim (g/\delta)^2\kappa_{b,\text{int}}$ due to the intrinsic decay of the buffer resonator itself, which we assume has $Q_{b,\text{int}} = 10^6$. This loss channel is *not* suppressed by the filter. However, it can be mitigated by increasing the detuning δ. Indeed, at large values of δ, κ_1 asymptotes to $\kappa_{1,\text{int}}$ (gray dashed line). We also show the loss parameter κ_1/κ_2 plotted in red, where κ_2 is averaged over the filter band, which also asymptotes to a lower bound as Z_b and δ become large. Bottom panels: loss spectra $\kappa_1/\kappa_2(\Delta)$ shown for a few selected values of Z_b. The gray shading indicates the regions where the adiabatic condition $g_2 < \eta\kappa_b/2\alpha$ is *not* satisfied. These regions roughly correspond to the frequencies outside of the passband. Here $\alpha = \sqrt{8}$ and $\eta = 1/5$.

the emission of photons with frequencies inside its narrow bandwidth. Conversely, in the "overcoupled" regime $C_\kappa \gg C_c$ the b mode decays rapidly, but interestingly $\kappa_2(\Delta)$ saturates and becomes asymmetric about $\Delta = 0$. The optimal coupling $C_\kappa = C_c$, in between these two regimes, is where $\kappa_2(\Delta)$ is nearly perfectly symmetric and flat, and practically saturated. We remark that although κ_b saturates to a value of around $\sim 4J$, it is possible to set C_κ to a small enough value so that κ_b is much smaller than this saturation value, assuming a flat $\kappa_2(\Delta)$ were not needed (which is not the case here).

Next, we show in Fig. 20(b) the effect of varying the characteristic filter impedance Z_f. Because of Eq. (A54), in order to keep J constant as Z_f is varied we must adjust C_c as well. Furthermore, we observe that κ_b decreases with Z_f, and so in order to respect the adiabatic threshold $g_2 < \eta\kappa_b/2\alpha$ introduced earlier we reduce g_2 by increasing $\omega_b(= \omega_f - 2J)$ to further detune the a and b modes. The key observation is that the optimal value of Z_f, for which $\kappa_2(\Delta)$ is flat, is precisely $Z_f = Z_b/2$. This is true regardless of the chosen value of Z_b.

Together, these observations constrain $C_\kappa = C_c$ and $Z_f = Z_b/2$, and through Eq. (A54), C_c is constrained to the value $C_c \approx 4J/\omega_f^2 Z_b$. Once a value of Z_b is chosen, the only remaining free parameter is ω_f. But as we did in the preceding exercise, in what follows we will again use $\omega_f(= \omega_b + 2J)$ to fine-tune g_2 in order to satisfy the adiabaticity constraint. Therefore, with this design methodology, *all of the (optimal) filter parameters are dictated by the properties of the storage and buffer resonators, with the exception of N and $N_t \sim N/2$.*

5. Optimization of the dimensionless loss κ_1/κ_2

We finally address the problem of optimizing the loss parameter κ_1/κ_2. For this we turn our attention back to Eq. (A6), which we repeat here: $g_2 \sim Z_b^{5/2}$. Since $\kappa_2 \sim g_2^2 \sim Z_b^5$, the obvious question is, can we exploit this scaling to maximize κ_2? The answer is yes, but surprisingly this is not because of the obvious reason one would expect. In fact, as Z_b increases, all of the filter parameters must be adjusted accordingly as described in Appendix A 4. We observe numerically that as this procedure is carefully repeated with different values of Z_b, *the dissipation rate κ_2 remains practically constant and is independent of Z_b*. A semi-quantitative explanation is as follows: 1) because $C_\kappa = C_c$ and $Z_f = Z_b/2$ as found in the preceding section, the "b" resonator is hardly distinguishable from any other resonator in the main section of the filter waveguide. Its effective decay rate is therefore $\kappa_{b,\text{eff}} \sim J$, because the hopping rate J is the rate that determines how quickly an excitation is transferred to the filter and out of the b mode. Indeed, we observe numerically that this decay precisely matches the filter bandwidth, $\kappa_{b,\text{eff}} = 4J$, as we increase Z_b while re-optimizing all of the filter parameters every time Z_b changes. 2) Since $g_2 = \eta\kappa_b/2\alpha = 2\eta J/\alpha$ (to satisfy the adiabaticity constraint), $\kappa_{2,\max} = 4g_2^2/\kappa_b \approx 4\eta^2 J/\alpha^2 \approx 4J/25\alpha^2$. *Therefore, κ_2 only depends on the filter bandwidth, which is upper-bounded by the crosstalk analysis of Appendix B, and the mean phonon number $|\alpha|^2$*. This result has important implications for our proposal and, as we will see shortly, imposes a lower bound on the phonon relaxation rate κ_1 required to reach the low values of κ_1/κ_2 that are necessary for our architecture.

Even though κ_2 depends solely on J and $|\alpha|^2$, there is still something to be gained by increasing Z_b. In Fig. 21 we show the "loss spectrum" $\kappa_1/\kappa_2(\Delta)$ for different values of Z_b. We observe that this loss does indeed decrease as Z_b increases, but only relatively slowly and eventually asymptotes to a fixed value. This is because as Z_b increases, g_2 increases as well, so the optimization procedure pushes ω_b further away from $\omega_a = \omega_b + \delta$ to compensate and keep g_2 below the adiabatic threshold $\eta\kappa_b/2\alpha$. In doing so, the Purcell decay $\sim (g/\delta)^2\kappa_{b,\text{int}}$ that originates from the hybridization of the buffer and storage modes (here g is the linear coupling between them) decreases as well. Note that only the intrinsic loss $\kappa_{b,\text{int}}$ of the buffer resonator enters this formula, because the radiative contribution is strongly suppressed since ω_a lies far outside the filter passband. Nevertheless this intrinsic contribution is still important, because in this proposal we operate under the assumption that the intrinsic decay rate $\kappa_{b,\text{int}}$ of the buffer mode (which is a superconducting circuit that suffers from several loss channels including two-level systems, quasiparticles, etc.) is at least two orders of magnitude larger than that of the storage mode, $\kappa_{1,\text{int}}$. In the limit $g/\delta \ll 1$, this contribution becomes negligibly small, and the phonon relaxation rate is purely intrinsic: $\kappa_1 \approx \kappa_{1,\text{int}}$. This causes the loss κ_1/κ_2 to asymptote to

$$\kappa_1/\kappa_2 \xrightarrow[Z_b\to\infty]{} \kappa_{1,\text{int}}|\alpha|^2/4\eta^2 J. \qquad \text{(A55)}$$

This is of course only a theoretical exercise: one cannot build a device with arbitrarily large Z_b, and ω_b cannot be arbitrarily large. However, as we show in Fig. 21, there is a feasible — and perhaps even practical — range of values of Z_b with which we could begin to approach the limiting value of loss in Eq. (A55), depending on what assumptions we make for the intrinsic losses of the buffer and storage modes. These limiting values are plotted in Fig. 3 in the main text as a function of $\kappa_{1,\text{int}}$ and for different filter bandwidths. It is important to emphasize that it maybe be possible to increase J beyond its presently constrained value $4J/2\pi = 100\,\text{MHz}$ through further innovations in the stabilization protocols, or by reducing the number of resonators coupled to each ATS. This is why we plot these curves for different bandwidths.

Appendix B: Multiplexed stabilization and crosstalk

In this Appendix, we show how multiple storage resonators coupled to a common ATS can be stabilized simultaneously. Coupling to a common ATS leads to crosstalk, and we discuss how this crosstalk can be quantified and mitigated. The main result of this Appendix is that the predominant sources of crosstalk can be effectively mitigated when up to five modes are coupled to a common ATS, so that the five-mode unit cells of our architecture are largely free of crosstalk.

In Appendix B 1, we begin by reviewing the effective operator formalism described in Ref. [39], which is the main tool we employ to analyze the dynamics of these multimode systems. In Appendix B 2, we describe our procedure for stabilizing multiple modes with a single ATS, and in Appendix B 3 we discuss the resulting sources of crosstalk. Finally, in Appendices B 4 and B 5 we show how this crosstalk can be effectively mitigated through a combination of filtering and phonon mode frequency optimization. Throughout this appendix, we take $\hbar = 1$ to simplify notation.

1. Effective operator formalism

In this Appendix, we frequently employ adiabatic elimination as a tool to extract the effective dynamics of an open quantum system within some stable subspace. The purpose of this subsection is to describe the effective operator formalism that we employ in order to perform this adiabatic elimination. While adiabatic elimination has been described in a variety of prior works (see, e.g., [39, 94, 95]), we privilege the treatment in Ref. [39] due to its simplicity and ease of application. We briefly review the relevant results.

Consider an open quantum system evolving according to the master equation

$$\dot{\hat{\rho}} = -i[\hat{H}, \hat{\rho}] + \sum_i \mathcal{D}[\hat{L}_i](\hat{\rho}), \tag{B1}$$

with Hamiltonian $\hat{H}$, jump operators $\hat{L}_i$, and where $\mathcal{D}[\hat{L}](\hat{\rho}) = \hat{L}\hat{\rho}\hat{L}^\dagger - \frac{1}{2}\left(\hat{L}^\dagger\hat{L}\hat{\rho} + \hat{\rho}\hat{L}^\dagger\hat{L}\right)$. We suppose that the system can be divided into two subspaces: a stable ground subspace, and a rapidly-decaying excited subspace, defined by the projectors $\hat{P}_g$ and $\hat{P}_e$, respectively. The Hamiltonian can be written in block form with respect to these subspaces as

$$\hat{H} = \begin{pmatrix} \hat{H}_g & \hat{V}_- \\ \hat{V}_+ & \hat{H}_e \end{pmatrix} \tag{B2}$$

where $\hat{H}_{g,e} = \hat{P}_{g,e}\hat{H}\hat{P}_{g,e}$, and $\hat{V}_{+,-} = \hat{P}_{e,g}\hat{H}\hat{P}_{g,e}$. We also suppose that the jump operators take the system from the excited to the ground subspace, i.e., $\hat{L}_i = \hat{P}_g\hat{L}_i\hat{P}_e$, and we define the non-Hermitian Hamiltonian

$$\hat{H}_{\mathrm{NH}} = \hat{H}_e - \frac{i}{2}\sum_i \hat{L}_i^\dagger\hat{L}_i. \tag{B3}$$

$\hat{H}_{\mathrm{NH}}$ describes the evolution within the excited subspace; unitary evolution is generated by $\hat{H}_e$, while the remaining term describes the non-unitary, deterministic "no jump" evolution induced by the dissipators $\mathcal{D}[\hat{L}_i]$.

The authors of Ref. [39] consider the case where the evolution between the subspaces induced by $\hat{V}_{+,-}$ is perturbatively weak relative to the evolution induced by $\hat{H}_0 \equiv \hat{H}_g + \hat{H}_{\mathrm{NH}}$. Because the excited subspace is barely populated due to the rapid decays, the dynamics of the system are well-approximated by those within the ground subspace, governed by the effective master equation

$$\dot{\hat{\rho}} = -i[\hat{H}_{\mathrm{eff}}, \hat{\rho}] + \sum_i \mathcal{D}[\hat{L}_{\mathrm{eff},i}](\hat{\rho}), \tag{B4}$$

where

$$\hat{H}_{\mathrm{eff}} = -\frac{1}{2}\hat{V}_-\left[\hat{H}_{\mathrm{NH}}^{-1} + \left(\hat{H}_{\mathrm{NH}}^{-1}\right)^\dagger\right]\hat{V}_+ + \hat{H}_g, \tag{B5}$$

and

$$\hat{L}_{\mathrm{eff},i} = \hat{L}_i\hat{H}_{\mathrm{NH}}^{-1}\hat{V}_+. \tag{B6}$$

These expressions apply for time-independent Hamiltonians. However, we will also be interested in situations where the perturbations $\hat{V}_{+,-}$ are time-dependent and take the form

$$\hat{V}_+(t) = \sum_n \hat{V}_{+,n}e^{i\delta_n t}, \tag{B7}$$

$$\hat{V}_-(t) = \sum_n \hat{V}_{-,n}e^{-i\delta_n t}. \tag{B8}$$

In this case, the effective Hamiltonian and jump operators are given by

$$\begin{aligned}\hat{H}_{\mathrm{eff}} &= \hat{H}_g \\ &- \frac{1}{2}\sum_{m,n}\hat{V}_{-,n}\left[\hat{H}_{\mathrm{NH},m}^{-1} + \left(\hat{H}_{\mathrm{NH},n}^{-1}\right)^\dagger\right]\hat{V}_{+,m}e^{i(\delta_m-\delta_n)t},\end{aligned} \tag{B9}$$

and

$$\hat{L}_{\mathrm{eff},i} = \hat{L}_i\sum_n \hat{H}_{\mathrm{NH},n}^{-1}\hat{V}_{+,n}e^{i\delta_n t}, \tag{B10}$$

where $\hat{H}_{\mathrm{NH},n} = \hat{H}_{\mathrm{NH}} + \delta_n$.

2. Simultaneous stabilization of multiple cat qubits with a single ATS

We consider a collection of N storage modes mutually coupled to a common reservoir. For the moment, we take reservoir to be a capacitively-shunted ATS (buffer resonator) with a large decay rate. The Hamiltonian of the system is

$$\begin{aligned}\hat{H} &= \hat{H}_d + \omega_b\hat{b}^\dagger\hat{b} + \sum_{n=1}^N \omega_n\hat{a}_n^\dagger\hat{a}_n \\ &- 2E_J\epsilon_p(t)\sin\left(\hat{\phi}_b + \sum_{n=1}^N \hat{\phi}_n\right),\end{aligned} \tag{B11}$$

where $\hat{H}_d$ is a driving term (defined below), $\hat{a}_n$ ($\hat{b}$) is the annihilation operator for the n-th storage mode (buffer mode) with frequency ω_n (ω_b), and $\hat{\phi}_n = \varphi_n(\hat{a}_n + \hat{a}_n^\dagger)$ is the phase across the ATS due to mode n, with vacuum fluctuation amplitudes φ_n. To stabilize multiple storage modes simultaneously, we apply separate pump and drive tones for each mode. Explicitly,

$$\epsilon_p(t) = \sum_n \epsilon_p^{(n)}\cos\left(\omega_p^{(n)}t\right), \tag{B12}$$

and

$$\hat{H}_d = \sum_n\left(\epsilon_d^{(n)}\hat{b}\,e^{i\omega_d^{(n)}t} + \mathrm{H.c.}\right). \tag{B13}$$

We choose the frequencies of the n-th pump and drive tones, respectively, as

$$\omega_p^{(n)} = 2\omega_n - \omega_b + \Delta_n, \tag{B14}$$

$$\omega_d^{(n)} = \omega_b - \Delta_n, \tag{B15}$$

where Δ_n denote detunings whose importance will be made clear shortly. Note that, in the architecture proposed in the main text, only a subset of the modes coupled to a given reservoir are stabilized by that reservoir. Accordingly, only the corresponding subset of the drives and pumps above need actually be applied.

To proceed, we expand the sine to third order and move to the frame where each mode rotates at its respective frequency. The resultant Hamiltonian is

$$\begin{aligned}\hat{H} \approx & \sum_n \left(\epsilon_d^{(n)} \hat{b}\, e^{-i\Delta_n t} + \text{H.c.}\right) \\ & - 2E_J\epsilon_p(t)\left[\varphi_b \hat{b}\, e^{-i\omega_b t} + \sum_n \varphi_n \hat{a}_n\, e^{-i\omega_n t} + \text{H.c.}\right] \\ & + \frac{E_J}{3}\epsilon_p(t)\left[\varphi_b \hat{b}\, e^{-i\omega_b t} + \sum_n \varphi_n \hat{a}_n\, e^{-i\omega_n t} + \text{H.c.}\right]^3 \end{aligned} \tag{B16}$$

This Hamiltonian contains terms that lead to the required two-photon dissipators for each storage mode,

$$\sum_n \left[g_{2,n}\left(\hat{a}_n^2 - \alpha_n^2\right)\hat{b}^\dagger e^{i\Delta_n t} + \text{H.c.}\right], \tag{B17}$$

with

$$g_{2,n} = E_J \epsilon_p^{(n)} \varphi_n^2 \varphi_b / 2, \tag{B18}$$

$$\alpha_n^2 = -\left(\epsilon_d^{(n)}\right)^* / g_{2,n}. \tag{B19}$$

However, the Hamiltonian (B16) contains numerous other terms. While many of these other terms are fast-rotating and can be neglected in the rotating wave approximation (RWA), others can have non-trivial effects. For example, the interplay between the terms in the second and third lines of (B16) gives rise to effective frequency shifts (a.c. Stark shifts) of the buffer and storage modes, which modify the resonance conditions (B14) and (B15). One can calculate the magnitudes of these shifts (and hence compensate for them) by applying the effective operator approach of Refs. [46, 47], in which case the Stark shifts are given by the coefficients of the $\hat{b}^\dagger\hat{b}$ and $\hat{a}^\dagger\hat{a}$ terms that arise in the effective Hamiltonian. Alternatively, the shifts can be calculated by moving to a displaced frame with respect to the linear terms on the second line of (B16), as is done in Ref. [25]. The Hamiltonian (B16) also contains terms which lead to crosstalk, but we defer the discussion of these terms to the next section. For now, we keep only the desired terms (B17).

We proceed by adiabatically eliminating the lossy buffer mode $\hat{b}$, following the approach described in Appendix B 1. Specifically, we designate the the ground subspace as the subspace where the buffer mode is in the vacuum state, and the excited subspace as the subspace where the buffer mode contains at least one excitation. We find that the effective dynamics of the storage modes within the ground subspace are described by the master equation

$$\dot{\hat{\rho}} = -i[\hat{H}_{\text{eff}}, \hat{\rho}] + \mathcal{D}\left[\sum_n \frac{g_{2,n}}{\Delta_n - i\kappa_b/2}\left(\hat{a}_n^2 - \alpha_n^2\right) e^{i\Delta_n t}\right](\hat{\rho}), \tag{B20}$$

where

$$\begin{aligned}\hat{H}_{\text{eff}} = -\frac{1}{2}\sum_{m,n} \Bigg\{ & g_{2,n}^* g_{2,m} (\hat{a}_n^2 - \alpha_n^2)^\dagger (\hat{a}_m^2 - \alpha_m^2) \\ & \times \left[\frac{1}{\Delta_m - i\kappa_b/2} + \frac{1}{\Delta_n + i\kappa_b/2}\right] e^{i(\Delta_m - \Delta_n)t}\Bigg\}. \end{aligned} \tag{B21}$$

To understand these dynamics, let us first consider the simple case where $\Delta_n = 0$. The above master equation reduces to

$$\dot{\hat{\rho}} = \kappa_2 D\left[\sum_n \left(\hat{a}_n^2 - \alpha_n^2\right)\right](\hat{\rho}), \tag{B22}$$

where $\kappa_2 = 4|g_2|^2/\kappa_b$. Any product of coherent states

$$|\beta_1\rangle \otimes |\beta_2\rangle \otimes \ldots \otimes |\beta_N\rangle \tag{B23}$$

that satisfies $\sum_n \beta_n^2 = \sum_n \alpha_n^2$ is a steady state of (B22). The subspace of steady states includes states in the code space, for which $\beta_n^2 = \alpha_n^2$, but it also includes states outside of the code space. Because a strictly larger space is stabilized, when noise pushes the system outside of the code space, the stabilization is not guaranteed to return the system to the code space. The coherent dissipation in Eq. (B22) is thus not sufficient for our purposes.

Consider instead the case where the detunings are chosen to be distinct, satisfying $|\Delta_n - \Delta_m| \gg 4|\alpha|^2\kappa_2$. In this limit, we can drop the now fast-rotating cross terms in the dissipator in Eq. (B20), and the effective master equation becomes

$$\dot{\hat{\rho}} = \sum_n \kappa_{2,n} D\left[\hat{a}_n^2 - \alpha_n^2\right](\hat{\rho}), \tag{B24}$$

where

$$\kappa_{2,n} = \frac{\kappa_b |g_{2,n}|^2}{\Delta_n^2 + \kappa_b^2/4}. \tag{B25}$$

The incoherent dissipator Eq. (B24) stabilizes cat states in each mode, as desired. Thus, by simply detuning the pumps and drives used to stabilize each mode, multiple modes can be stabilized simultaneously and independently by a single ATS.

Two remarks about the approximation of Equation (B22) by Equation (B24) are necessary. First, the condition $|\Delta_n - \Delta_m| \gg 4|\alpha|^2\kappa_2$ can be derived by expressing the operators in Equation (B22) in the displaced Fock basis (Appendix C). Roughly speaking, the condition dictates that $|\Delta_n - \Delta_m|$ be much larger than the

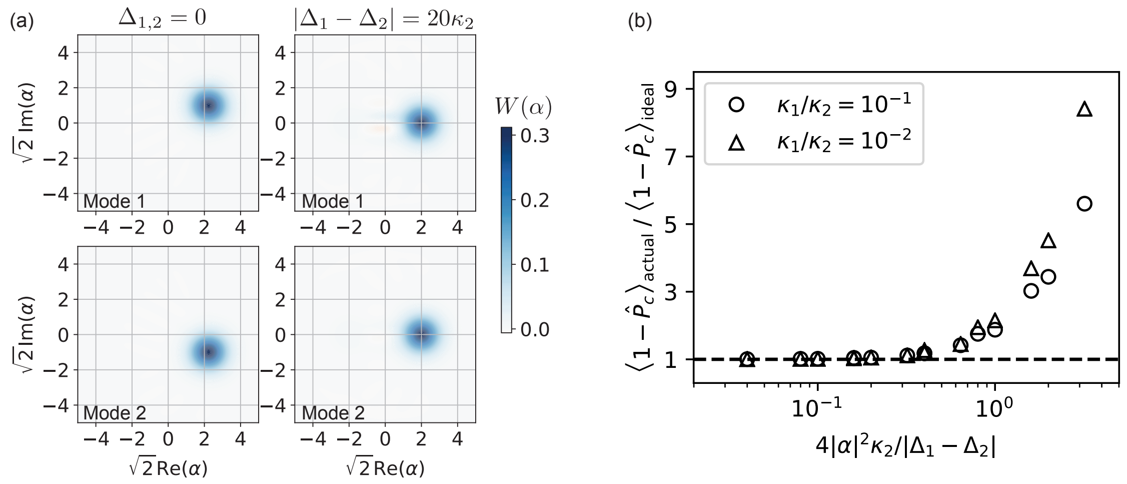


FIG. 22. Multiplexed stabilization. (a) Comparison of stabilization for $\Delta_n = 0$ and $|\Delta_n - \Delta_m| \gg 4|\alpha|^2\kappa_2$. Wigner plots are shown of two storage modes after evolution under the master equation $\dot{\hat{\rho}} = -i[\hat{H}, \hat{\rho}] + \kappa_b \mathcal{D}[\hat{b}]$, with $\hat{H}$ given by (B17). The storage modes are initialized in a product state $|\beta_1\rangle|\beta_2\rangle$ that does not lie in the code space but which is a steady state of (B22). Thus, when $\Delta_n = 0$ (left plots), the evolution is (approximately) trivial. The left two plots thus also serve as Wigner plots of the initial state $|\beta_1\rangle|\beta_2\rangle$. However, when $|\Delta_1 - \Delta_2| \gg 4|\alpha|^2\kappa_2$ (right plots), the system evolves to the code space, defined here by $\alpha = \sqrt{2}$. (b) Validity of approximating Equation (B22) by Equation (B24). Master equations (B22,B24) are simulated (with decoherence added to each mode via the dissipators $\kappa_1\mathcal{D}[\hat{a}]$ and $\kappa_1\mathcal{D}[\hat{a}^\dagger\hat{a}]$), and the expectation value of $1 - \hat{P}_c$ is computed once the system reaches its steady state. Here $\hat{P}_c$ denotes the projector onto the cat code space, and the subscripts "actual" and "ideal" denote expectation with respect to the steady states of (B22) and (B24), respectively. The ratio of expectations, plotted on the vertical axis, quantifies the relative increase in population outside the code space. A ratio ~ 1 indicates the approximation works well. Parameters are chosen from the ranges $|\alpha|^2 \in [1, 4]$ and $|\Delta_1 - \Delta_2|/\kappa_2 \in [5, 100]$.

rate at which photons are lost from the stabilized modes. Second, we have neglected $\hat{H}_{\mathrm{eff}}$; the rotating terms in $\hat{H}_{\mathrm{eff}}$ can be dropped in the RWA in the considered limit, and the non-rotating terms provide an additional source of stabilization [16] that we neglect for simplicity. It is also worth noting that the two-photon dissipation rate, $\kappa_{2,n}$, decreases monotonically with Δ_n. To avoid significant suppression of this engineered dissipation, one can choose $\Delta_n \lesssim \kappa_b$ so that $\kappa_{2,n}$ remains comparable to κ_2, or alternatively one can exploit the filtering procedure described in Appendix A and further analyzed in Appendix B 4 which enables strong effective dissipation even for $\Delta_n > \kappa_b$.

We demonstrate our scheme for multiplexed stabilization numerically in Fig. 22. Through master equation simulations, we observe good stabilization for $|\Delta_1 - \Delta_2| \gg 4|\alpha|^2\kappa_2$, but not $\Delta_{1,2} = 0$, as expected. Moreover, we also quantify the validity of approximating Equation (B22) by Equation (B24). Strictly speaking, the approximation is valid only in the regime $|\Delta_n - \Delta_m| \gg 4|\alpha|^2\kappa_2$, but we find that even for $|\Delta_n - \Delta_m| \sim 4|\alpha|^2\kappa_2$ the stabilization works reasonably well, by which we mean that the population that leaks out of the code space is comparable for the two dissipators (B22) and (B24), see Figure 22(b). The approximation breaks down beyond this point, and accounting for the additional terms in Equation (B22) becomes increasingly important.

We conclude this subsection by providing some physical intuition as to why detuning the pumps and drives allows one to stabilize multiple cat qubits simultaneously. When $\Delta_n = 0$, photons lost from different storage modes via the buffer cannot be distinguished by the environment. As a result, we obtain a single coherent dissipator $\hat{L} \propto \sum_n(\hat{a}_n^2 - \alpha_n^2)$. When distinct detunings are chosen for each mode, however, photons lost from different modes via the buffer are emitted at different frequencies. When these photons are spectrally resolvable, the environment can distinguish them, resulting in a collection of independent, incoherent dissipators $\hat{L}_n \propto (\hat{a}_n^2 - \alpha_n^2)$ instead. The emitted photon linewidth is $4|\alpha|^2\kappa_2$, which can be seen by expressing $\kappa_2\mathcal{D}[\hat{a}^2 - \alpha^2]$ in the displaced Fock basis (Appendix C). Thus, the emitted photons are well-resolved when $|\Delta_n - \Delta_m| \gg 4|\alpha|^2\kappa_2$, which is the same condition assumed in the derivation of (B24). We illustrate this idea pictorially in Figure 4(a) of the main text.

3. Sources of crosstalk

In this subsection we describe how undesired terms in the Hamiltonian (B16) lead to crosstalk among modes coupled to the same ATS. In particular, we show that these undesired terms lead to effective dissipators and effective Hamiltonians that can cause correlated phase errors in the cat qubits.

The predominant sources of crosstalk are undesired terms in the Hamiltonian (B16) of the form

$$g_2\,\hat{a}_i\hat{a}_j\hat{b}^\dagger e^{i\delta_{ijk}t} + \text{H.c.},\tag{B26}$$

where

$$\delta_{ijk} = \omega_k^{(p)} - \omega_i - \omega_j + \omega_b,\tag{B27}$$

and we have neglected the dependence of g_2 on the indices i, j for simplicity. In contrast to the other undesired terms in (B16), these terms have the potential to induce large crosstalk errors because they both (i) have coupling strengths comparable to the desired terms (B17), and (ii) can be resonant or near-resonant. In particular, the undesired term is resonant ($\delta_{ijk} = 0$) for $2\omega_k + \Delta_k = \omega_i + \omega_j$. This resonance condition can be satisfied, for example, when the storage modes have near uniformly-spaced frequencies.

These unwanted terms may not be exactly resonant in practice, but we cannot generally guarantee that they will be rotating fast enough to be neglected in the RWA either. In contrast, all other undesired terms in (B16) are detuned by at least $\min_n|\omega_n - \omega_b|$, which is on the order of $\sim 2\pi \times 1\,\text{GHz}$ for the parameters considered in this work. We therefore focus on crosstalk errors induced by the terms (B26).

The terms (B26) can lead to three different types of correlated errors:

- Type I: Stochastic errors induced by effective dissipators
- Type II: Stochastic errors induced by effective Hamiltonians
- Type III: Coherent errors induced by effective Hamiltonians

We describe each type of error in turn. Without mitigation (see Appendices B 4 and B 5), these correlated phase errors could be a significant impediment to performing high-fidelity operations.

Type I: stochastic errors induced by effective dissipators

The terms (B26) can lead to correlated photon losses at rates comparable to κ_2, resulting in significant correlated phase errors in the cat qubits. These deleterious effects manifest when one adiabatically eliminates the buffer mode. Explicitly, we apply the effective operator formalism described in Subsecton B 1 to the operators

$$\hat{H}^{(1)} = g_2\,\hat{a}_i\hat{a}_j\hat{b}^\dagger e^{i\delta_{ijk}t} + \text{H.c.},\tag{B28}$$

$$\hat{L}^{(1)} = \sqrt{\kappa_b}\,\hat{b}\tag{B29}$$

and obtain the effective operators

$$\hat{H}^{(1)}_{\text{eff}} = -\frac{|g_2|^2\delta_{ijk}}{\delta_{ijk}^2 + \kappa_b^2/4}(\hat{a}_i\hat{a}_j)^\dagger(\hat{a}_i\hat{a}_j) + \text{H.c.},\tag{B30}$$

$$\hat{L}^{(1)}_{\text{eff}} = \frac{g_2\sqrt{\kappa_b}}{\delta_{ijk} - i\kappa_b/2}\hat{a}_i\hat{a}_j e^{i\delta_{ijk}t}.\tag{B31}$$

The effective Hamiltonian preserves phonon-number parity and thus does not induce phase flips. The effective jump operator $\hat{L}_{\text{eff}}$ describes correlated single-phonon losses in modes i and j at a rate

$$\kappa_{\text{eff}} = \frac{\kappa_b|g_2|^2}{\delta_{ijk}^2 + \kappa_b^2/4}\tag{B32}$$

which is comparable to κ_2 for $\delta_{ijk} \lesssim \kappa_b$. These correlated single photon losses induce correlated phase flips in the cat qubits, which can be seen by projecting $\hat{L}_{\text{eff}}$ into the code space,

$$\hat{L}^{(1)}_{\text{eff}} \to \sqrt{\kappa_{\text{eff}}}\,\alpha^2\hat{Z}_i\hat{Z}_j e^{i\delta_{ijk}t}.\tag{B33}$$

Type II: stochastic errors induced by effective Hamiltonians

The interplay between different terms of the form (B26) can lead to further correlated errors. As an example, consider the operators

$$\hat{H}^{(2)} = g_2\,\hat{a}_i\hat{a}_j\hat{b}^\dagger e^{i\delta_{ijk}t} + g_2\,\hat{a}_\ell\hat{a}_m\hat{b}^\dagger e^{i\delta_{\ell mn}t} + \text{H.c.},\tag{B34}$$

$$\hat{L}^{(2)} = \sqrt{\kappa_b}\,\hat{b}.\tag{B35}$$

Adiabatically eliminating the buffer mode yields,

$$\hat{H}^{(2)}_{\text{eff}} = \Big[\chi(\hat{a}_i\hat{a}_j)^\dagger(\hat{a}_\ell\hat{a}_m)e^{i(\delta_{\ell mn}-\delta_{ijk})t} + \text{H.c.}\Big] + \ldots,\tag{B36}$$

$$\begin{aligned}\hat{L}^{(2)}_{\text{eff}} = &\frac{g_2\sqrt{\kappa_b}}{\delta_{ijk} - i\kappa_b/2}\hat{a}_i\hat{a}_j e^{i\delta_{ijk}t}\\ &+ \frac{g_2\sqrt{\kappa_b}}{\delta_{\ell mn} - i\kappa_b/2}\hat{a}_\ell\hat{a}_m e^{i\delta_{\ell mn}t}.\end{aligned}\tag{B37}$$

where

$$\chi = -\frac{|g_2|^2}{2}\left[\frac{1}{\delta_{ijk} - i\kappa_b/2} + \frac{1}{\delta_{\ell mn} + i\kappa_b/2}\right]$$

and "..." denotes additional terms in the effective Hamiltonian that preserve phonon-number parity. Note that the effective dissipator $\hat{L}^{(2)}_{\text{eff}}$ leads to Type I correlated phase errors. Indeed, for sufficiently large $|\delta_{ijk} - \delta_{\ell mn}|$, the action of $\hat{L}^{(2)}_{\text{eff}}$ can be approximated by replacing it with two independent dissipators of the form (B31).

What is different about this example is that the effective Hamiltonian $\hat{H}^{(2)}_{\text{eff}}$ contains terms $\propto (\hat{a}_i\hat{a}_j)^\dagger(\hat{a}_\ell\hat{a}_m)$ that generally do not preserve phonon-number parity. Such terms can unitarily evolve the system out of the code space, changing the parity in the process. In turn, the engineered dissipation returns the system to the code space, but it does so without changing the parity. Therefore, the net effect of such excursions out of the code space and back is to induce *stochastic* parity-flips in the storage modes, which manifest as correlated phase errors on the cat qubits. The errors are stochastic even though the evolution generated by $\hat{H}^{(2)}_{\text{eff}}$ is unitary because the stabilization itself is stochastic. Specifically, the errors are of the form $\mathcal{D}[\hat{Z}_i\hat{Z}_j\hat{Z}_\ell\hat{Z}_m]$, which one can show by adiabatically eliminating the excited states of the storage modes (see Appendix C).

Type III: coherent errors induced by effective Hamiltonians

The parity-non-preserving effective Hamiltonian $\hat{H}^{(2)}_{\text{eff}}$ also induces non-trivial coherent evolution within the code space. This can be seen by projecting $\hat{H}^{(2)}_{\text{eff}}$ into the code space

$$\hat{H}^{(2)}_{\text{eff}} \to (|\alpha|^4\chi\hat{Z}_i\hat{Z}_j\hat{Z}_\ell\hat{Z}_m e^{i(\delta_{\ell mn}-\delta_{ijk})t} + \text{H.c.}). \quad \text{(B38)}$$

This undesired evolution does not decohere the system but can nevertheless degrade the fidelity of operations. See further discussion in Appendix B 5.

4. Crosstalk mitigation: filtering

In this subsection, we show how Type I and Type II crosstalk errors can be suppressed by placing a bandpass filter at the output port of the buffer mode (see Appendix A 3 for additional discussion of filtering). The purpose of the filter is to allow photons of only certain frequencies to leak out of the buffer, such that the desired engineered dissipation remains strong but spurious dissipative processes are suppressed. A crucial requirement of this approach is that the desired dissipative processes be spectrally resolvable from the undesired ones, and we show that adequate spectral resolution is achievable in the next section (Appendix B 5).

We begin by providing a quantum mechanical model of a bandpass filter [93, 96]. While a detailed classical analysis of the filter is given in Appendix A 3, here we employ a complementary quantum model. The quantum model not only allows us to study the filter's effects numerically via master equation simulations, but it is also sufficiently simple so as to enable a straightforward analytical treatment via the effective operator formalism described in Appendix B 1.

Motivated by the filter designs described in Appendix A 4, we employ a tight-binding model where the filter consists of a linear chain of M bosonic modes with annihilation operators $\hat{c}_i$, and each with the same frequency ω_b. Modes in the chain are resonantly coupled to their nearest neighbors with strength J. The first mode in the chain couples to the buffer mode $\hat{b}$, which is no longer coupled directly to the open $50\,\Omega$ waveguide. Instead, the M-th mode is now the one which couples strongly to the waveguide, such that its single-photon loss rate is given by κ_c. The buffer-filter system is described by the Hamiltonian (in the rotating frame)

$$\hat{H}_{\text{buffer+filter}} = J(\hat{c}_1^\dagger\hat{b} + \hat{c}_1\hat{b}^\dagger) + \sum_{i=1}^{M-1} J(\hat{c}_{i+1}^\dagger\hat{c}_i + \hat{c}_{i+1}\hat{c}_i^\dagger), \quad \text{(B39)}$$

together with the dissipator $\kappa_c\mathcal{D}[\hat{c}_M]$. We show below that these additional modes act as a bandpass filter, with center frequency ω_b and bandwidth $4J$, and they suppresses the emission of photons with frequencies outside of this passband.

Suppression of Type I errors

To illustrate the suppression of Type I errors, we consider the operators

$$\hat{H}^{(3)} = \left(g_2\,\hat{a}_i\hat{a}_j\hat{b}^\dagger e^{i\delta_{ijk}t} + \text{H.c.}\right) + \hat{H}_{\text{buffer+filter}}, \quad \text{(B40)}$$

$$\hat{L}^{(3)} = \sqrt{\kappa_c}\,\hat{c}_M \quad \text{(B41)}$$

where the first term in $\hat{H}^{(3)}$ is the same as the unwanted term $\hat{H}^{(1)}$ from Appendix B 3. We adiabatically eliminate *both the buffer and filter modes* in order to obtain an effective dynamics for only the storage modes. We note that adiabatically eliminating the buffer and filter modes together is not fundamentally different from adiabatically eliminating the buffer; both calculations are straightforward applications of the methods in Subsection B 1. We obtain the effective dissipator

$$\hat{L}^{(3)}_{\text{eff}} = \sqrt{\kappa_{\text{eff}}(M)}\,\hat{a}_i\hat{a}_j e^{i\delta_{ijk}t} \quad \text{(B42)}$$

where the rates for the first few values of M are

$$\kappa_{\text{eff}}(0) = \frac{\kappa_c|g_2|^2}{\delta_{ijk}^2 + \kappa_c^2/4} \approx \kappa_c\frac{|g_2|^2}{\delta_{ijk}^2} \quad \text{(B43)}$$

$$\kappa_{\text{eff}}(1) = \frac{\kappa_c|g_2|^2J^2}{(J^2-\delta_{ijk}^2)^2 + \delta_{ijk}^2\kappa_c^2/4} \approx \kappa_{\text{eff}}(0)\left(\frac{J}{\delta_{ijk}}\right)^2 \quad \text{(B44)}$$

$$\kappa_{\text{eff}}(2) = \frac{\kappa_c|g_2|^2J^4}{(2J^2\delta_{ijk}-\delta_{ijk}^3)^2 + (J^2-\delta_{ijk}^2)^2\kappa_c^2/4} \approx \kappa_{\text{eff}}(0)\left(\frac{J}{\delta_{ijk}}\right)^4, \quad \text{(B45)}$$

where the approximations assume that $\delta_{ijk} \gg J, \kappa_c$. In this regime, $\kappa_{\text{eff}}(M)$ is exponentially suppressed with increasing M via the factor $(J/\delta_{ijk})^{2M}$.

We plot these rates as a function of δ_{ijk} in Figure 23(a), where the exponential suppression of the decoherence rates

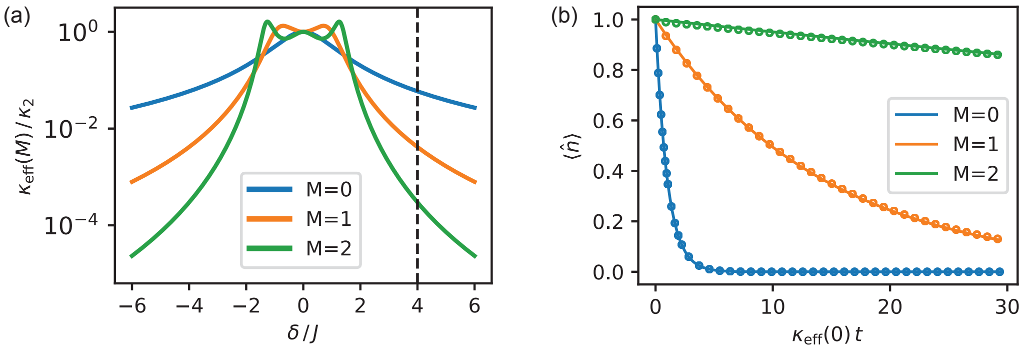


FIG. 23. Suppression of Type I errors. (a) Plots of $\kappa_{\rm eff}(M)$ as a function of the detuning, δ, of the unwanted term. (b) Master equation simulations. The system is initialized with a single excitation in the storage mode and evolved according to the dynamics $\dot{\hat\rho} = -i[(g_2\hat a\hat b^\dagger e^{i\delta t} + \text{H.c.}) + \hat H_{\text{buffer+filter}}, \hat\rho] + \mathcal{D}[\hat L^{(3)}](\hat\rho)$. These dynamics are analogous to those generated by $\hat H^{(3)}$ and $\hat L^{(3)}$; in both cases the unwanted term induces losses at rates $\kappa_{\rm eff}(M)$. Simulation results are indicated by open circles, and the analytical expressions for $\kappa_{\rm eff}(M)$ are plotted as solid lines. Parameters: $\alpha = \sqrt{2}, \kappa_c/g_2 = 10, J/g_2 = 5$. For (b), $\delta = 4J$, as indicated by the dashed line in (a).

outside the filter band is evident. Figure 23(a) should be understood as analogous to Fig. 20 in Appendix A, though we emphasize that here the rates are derived from a fully quantum model of the filter. We also remark that unlike in Appendix A, where the emphasis was on detailed classical filter design, here we do not taper the filter. This explains the "ripples" in $\kappa_{\rm eff}$ within the filter passband. Figure 23(b) shows the results of analogous master equation simulations; good quantitative agreement with the analytical expressions is observed. Thus we conclude that Type I errors are indeed suppressed by the filter, provided $|\delta_{ijk}| > 2J$.

Suppression of Type II errors

To illustrate the suppression of Type II errors, we construct a simple toy model that both captures the relevant physics and is easy to study numerically. Consider the operators

$$\hat H^{(4)} = \left(g\,\hat a\hat b^\dagger e^{i\delta_1 t} + g\,\hat b^\dagger e^{i\delta_2 t} + \text{H.c.}\right) + \left[g_2(\hat a^2-\alpha^2)\hat b^\dagger + \text{H.c.}\right] + \hat H_{\text{buffer+filter}} \qquad \text{(B46)}$$

$$\hat L^{(4)} = \sqrt{\kappa_c}\,\hat c_M. \qquad \text{(B47)}$$

where $\hat a$ is the annihilation operator for the single storage mode that we consider in this model. In this toy model, the first line of $\hat H^{(4)}$ should be understood as analogous to $\hat H^{(2)}$. Indeed we obtain the former from the latter by replacing $\hat a_i\hat a_j \to \hat a$ and $\hat a_\ell \hat a_m \to 1$.

Adiabatically eliminating the buffer and filter modes yields the effective operators

$$\hat H^{(4)}_{\rm eff} = \left[\chi_{\rm eff}(M)\,\hat a\,e^{i(\delta_1-\delta_2)t} + \text{H.c.}\right] + \ldots \qquad \text{(B48)}$$

$$\hat L^{(4)}_{\rm eff} = \sqrt{\kappa^{(\delta_1)}_{\rm eff}(M)}\,\hat a\,e^{i\delta_1 t} + \sqrt{\kappa^{(0)}_{\rm eff}(M)}(\hat a^2-\alpha^2). \qquad \text{(B49)}$$

Here, "…" denotes a parity-preserving term ($\propto \hat a^\dagger\hat a$) that we neglect, $\kappa^{(\delta)}_{\rm eff}(M)$ denotes the effective loss rate [Eqs. (B43) to (B45)] with the replacement $\delta_{ijk}\to\delta$, and

$$\chi_{\rm eff}(M) \approx -\frac{|g|^2}{2}\left(\frac{1}{\delta_1}+\frac{1}{\delta_2}\right) \qquad \text{(B50)}$$

is independent of M in the limit $\delta_{1,2} \gg J, \kappa_b$. The first term in $\hat L^{(4)}_{\rm eff}$ gives rise to the Type I errors that are suppressed by the filter, as already discussed. Our present interest is the Type II errors induced by the interplay of $\hat H^{(4)}_{\rm eff}$, the stabilization, and the filter.

Unfortunately, the effective operators $\hat H^{(4)}_{\rm eff}$ and $\hat L^{(4)}_{\rm eff}$ do not properly capture this interplay. In particular, it follows from energy conservation that Type II errors induced by $\hat H^{(4)}_{\rm eff}$ result in photon emissions at frequency $\omega_b + \delta_2 - \delta_1$. Intuitively, such emissions should be exponentially suppressed when this frequency lies outside the filter band. However, this suppression is not apparent in the operators $\hat H^{(4)}_{\rm eff}, \hat L^{(4)}_{\rm eff}$ because, in the course of deriving $\hat H^{(4)}_{\rm eff}$, we already eliminated the filter. After adiabatic elimination the only vestige of the filter is the term $\sqrt{\kappa^{(0)}_{\rm eff}(M)}(\hat a^2-\alpha^2)$, which embodies the behavior of the filter at frequency ω_b, *but not at frequency* $\omega_b+\delta_2-\delta_1$. As such, proceeding to calculate the Type II error rate from these operators is not valid, and an alternate approach is required.

In order to properly capture the subtle interplay between the effective Hamiltonian, the stabilization, and

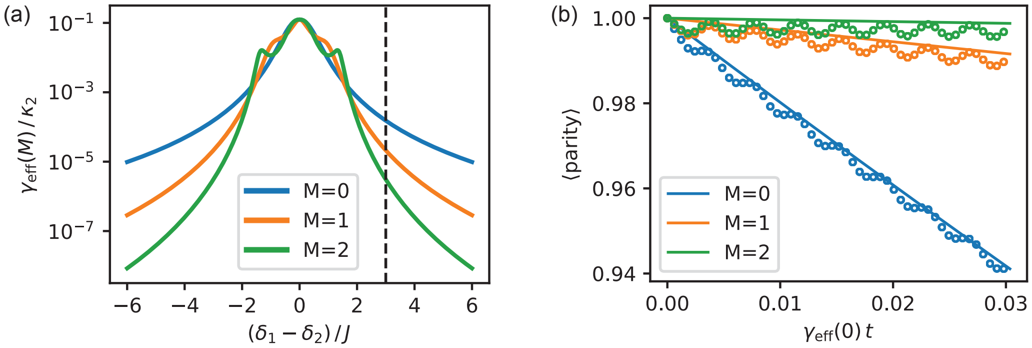


FIG. 24. Suppression of Type II errors. (a) Plots of $\gamma_{\rm eff}(M)$ as a function of the detuning, $\delta_1-\delta_2$, of the effective Hamiltonian. (b) Master equation simulations. The storage mode is initialized in the even parity cat state and evolved according to the dynamics $\dot{\hat\rho} = -i[\bar{\hat H}^{(4)},\hat\rho] + \mathcal{D}[\hat L^{(4)}](\hat\rho)$. Simulation results are indicated by open circles, and the analytical expressions for $\gamma_{\rm eff}(M)$ are plotted as solid lines. Parameters: $\alpha=\sqrt{2}, \kappa_c/g_2 = 10, J/g_2 = 5$. Rather than specify values for g and $\delta_{1,2}$, we simply fix $\chi_{\rm eff}(M)/g_2 = 0.2$. For (b), $\delta = 3J$, as indicated by the dashed line in (a).

filter, we defer adiabatic elimination and instead begin by calculating an effective Hamiltonian that describes the time-averaged dynamics generated by $\hat H^{(4)}$. We restrict our attention to a regime where the terms on the first line of Equation (B46) are rapidly rotating, so that evolution generated by $\hat H^{(4)}$ is well approximated by its time average. We calculate the time-averaged effective Hamiltonian $\bar{\hat H}^{(4)}$ following the approach described in Refs. [46, 47],

$$\bar{\hat H}^{(4)} = \left[g_2(\hat a^2-\alpha^2)\hat b^\dagger + \text{H.c.}\right] + \hat H_{\text{buffer+filter}} - \frac{|g|^2}{2}\left(\frac{1}{\delta_1}+\frac{1}{\delta_2}\right)\left(2\hat b^\dagger \hat b+1\right)\left(\hat a e^{i(\delta_1-\delta_2)t}+\text{H.c.}\right) \tag{B51}$$

where we have neglected a parity-preserving term ($\propto \hat a^\dagger \hat a$), and terms rotating at the fast frequencies $\delta_{1,2}$. Notice that

$$\bar{\hat H}^{(4)} \approx \left[g_2(\hat a^2-\alpha^2)\hat b^\dagger + \text{H.c.}\right] + \hat H_{\text{buffer+filter}} + \hat H^{(4)}_{\text{eff}}, \tag{B52}$$

where the approximation is obtained by preemptively replacing $\hat b^\dagger \hat b$ with its expected value of 0. Doing so reveals that $\hat H^{(4)}_{\rm eff}$ can be understood as arising from the time-averaged dynamics of the the unwanted terms in $\hat H^{(4)}$ in the limit of large $\delta_{1,2}$. In effect, time averaging provides a way of introducing $\hat H^{(4)}_{\rm eff}$ into the dynamics without having to eliminate the filter, thereby allowing us to study the interplay of the filter and effective Hamiltoninan.

We proceed by taking the operators $\bar{\hat H}^{(4)}$ and $\hat L^{(4)}$ and adiabatically eliminating the buffer, the filter, and all excited states of the storage mode, i.e. all states that do not lie in the code space. Adiabatically eliminating the storage mode excited states is valid in the regime where the engineered dissipation is strong relative to couplings that excite the storage mode ($\hat H^{(4)}_{\rm eff}$ in this case), such that these excited states are barely populated. See Appendix C for further details. We obtain

$$\hat{\bar H}^{(4)}_{\rm eff} = \chi_{\rm eff}(M)\,\alpha \hat Z\, e^{i(\delta_1-\delta_2)t} + \text{H.c.}, \tag{B53}$$

$$\hat{\bar L}^{(4)}_{\rm eff} = \sqrt{\gamma_{\rm eff}(M)}\hat Z. \tag{B54}$$

The rates for the first few values of M are

$$\gamma_{\rm eff}(0) = \frac{4\kappa_c|2g_2\alpha\,\chi_{\rm eff}(0)|^2}{4\left(|2g_2\alpha|^2-\delta_{12}^2\right)^2+\delta_{12}^2\kappa_c^2}, \tag{B55}$$

$$\gamma_{\rm eff}(1) = \frac{4J^2\kappa_c|2g_2\alpha\,\chi_{\rm eff}(1)|^2}{4\delta_{12}^2\left(J^2+|2g_2\alpha|^2-\delta_{12}^2\right)^2+\left(|2g_2\alpha|^2-\delta_{12}^2\right)^2\kappa_c^2} \approx \gamma_{\rm eff}(0)\left(\frac{J}{\delta_{12}}\right)^2, \tag{B56}$$

$$\gamma_{\rm eff}(2) = \frac{4J^4\kappa_c|2g_2\alpha\,\chi_{\rm eff}(2)|^2}{4\left(|2g_2\alpha|^2(J-\delta_{12})(J+\delta_{12})+\delta_{12}^4-2J^2\delta_{12}^2\right)^2+\delta_{12}^2\left(|2g_2\alpha|^2+J^2-\delta_{12}^2\right)^2\kappa_c^2} \approx \gamma_{\rm eff}(0)\left(\frac{J}{\delta_{12}}\right)^4, \tag{B57}$$

where we have used the shorthand $\delta_{12}\equiv\delta_1-\delta_2$ to simplify the expressions, and the approximations are obtained in the in the limit of large $|\delta_1 - \delta_2|$. In this limit, we find that the phase flip rate is exponentially suppressed by the filter,

$$\gamma_{\text{eff}}(M) \approx \gamma_{\text{eff}}(0)\left(\frac{J}{\delta_1 - \delta_2}\right)^{2M}, \qquad \text{(B58)}$$

as expected.

We plot the rates $\gamma_{\text{eff}}(M)$ as a function of $\delta_1 - \delta_2$ in Figure 24(a), where the exponential suppression of the decoherence rates outside the filter band is again evident. Figure 24(b) shows the results of corresponding master equation simulations. Good quantitative agreement with the analytical expressions is observed. (Note that the small parity oscillations in the simulation results are Type III errors—coherent micro-oscillations due to evolution generated by the effective Hamiltonian within the code space. These errors are not suppressed by the filter.) Thus we find that Type II errors are also suppressed by the filter, provided the effective Hamiltonian detuning lies outside the filter passband.

5. Crosstalk mitigation: mode frequency optimization

We have shown that stochastic correlated phase errors (Types I and II) can be suppressed by a filter if the corresponding emitted photons have frequencies outside the filter passband. We now show that it is possible to suppress *all* such errors simultaneously by carefully choosing the frequencies of the phonon modes. In doing so, the effects of Type III errors can also be simultaneously minimized. Importantly, the phonon mode frequencies are chosen to be compatible with error correction in the surface code, and we begin this section by describing how the surface code architecture constrains the choice of phonon mode frequencies.

We consider the surface-code architecture and optimize the phonon mode frequencies such that they are compatible with the surface-code stabilizer measurement. To understand the constraints imposed by the implementation of the surface code, recall that each ATS is coupled to five phononic modes in our proposal (see Fig. 2). Among the five modes, four modes (two data and two ancilla modes for the surface code) are stabilized in the cat-code manifold by an ATS. Another mode (readout mode) is dedicated to measuring cat qubits in the X basis and is not stabilized by any ATS. Since every data or ancilla mode couples to two ATSs, each ATS is only responsible for stabilizing two of the five phononic modes to which it couples. Thus, for each given ATS, we must determine which two phononic modes should be stabilized.

An important consideration in deciding which phononic modes should be stabilized by a given ATS is that each ATS is used to realize four CNOT gates (performed in four different time steps) to measure the stabilizers of the surface code. While a CNOT gate is being performed, the target mode of the CNOT gate is stabilized by a rotating jump operator $\hat{L}_2(t) = \hat{a}_2^2 - \alpha^2 + (\alpha/2)(\exp[2i\pi t/T] - 1)(\hat{a}_1 - \alpha)$ that acts non-trivially both on the target mode $(\hat{a}_2)$ and the control mode $(\hat{a}_1)$. Thus, while a CNOT gate is being performed, the target mode must be stabilized by the ATS that also couples to the control mode.

In Fig. 25 we show how these stabilization constraints can be satisfied. In the top panel of the figure, we show four (out of six, state preparation and measurement not show) time steps of the surface-code stabilizer measurement. During each time step, different CNOT gates between data and ancilla cat qubits are applied. We label data modes as α and γ and ancilla modes as β and δ. Ancilla modes labelled as β (δ) are used to measure the X-type (Z-type) stabilizers of the surface code. We use black arrows to indicate which phononic modes are stabilized by each ATS at each time step; each phononic mode at the tip of a black arrow is stabilized by the ATS at the arrow's tail. Importantly, every target mode of a CNOT gate is stabilized by an ATS that also couples to the corresponding control mode at all time steps. Note, however, that a given ATS stabilizes different modes at different time steps, as summarized in the bottom panel of Fig. 25. In particular, there are two stabilization configurations: in configuration 1 (2) modes α, β (γ, δ) are stabilized by the given ATS, and the remaining modes γ, δ (α, β) are stabilized by some other neighboring ATSs.

Now, our goal is to choose the frequencies of the phonon modes and detunings of the pumps in order to minimize crosstalk. In order to ensure that the choice of mode frequencies is compatible with the surface-code stabilizer measurement, we assign modes with the same label in Fig. 25 to have the same frequency. Thus, there are only five mode frequencies that must be chosen: the frequencies $\omega_\alpha, \omega_\beta, \omega_\gamma, \omega_\delta$ corresponding to the four labels in Fig. 25, plus the frequency of the readout mode (not shown in Fig. 25), which we take to be the same in each unit cell and denote by ω_ρ. Similarly, there are four pump detunings, $\Delta_\alpha, \Delta_\beta, \Delta_\gamma, \Delta_\delta$, that must be chosen. Here, as above, Δ_i denotes the detuning of the pump (and buffer drive) used to stabilize mode i. In the following, we construct a cost function C that quantifies crosstalk as a function of these nine parameters (five mode frequencies and four pump detunings). Numerically minimizing C allows us to find the choices of the frequencies and detunings that minimize crosstalk.

First, C should be large if any emitted photons associated with Type I and II errors lie inside the filter's bandwidth $4J$. We thus take $C = 1$ if any of the following conditions are met for either of the two stabilization configurations shown in Fig. 25:

- $|\delta_{ijk}| < 2J$ (Type I errors not suppressed)
- $|\delta_{ijk} - \delta_{\ell mn}| < 2J$ (Type II errors not suppressed)
- $|\delta_{iii}| > 2J$ (desired dissipation suppressed)

In other words, we set $C = 1$ if any Type I or II errors are not suppressed by the filter, or if any of the desired

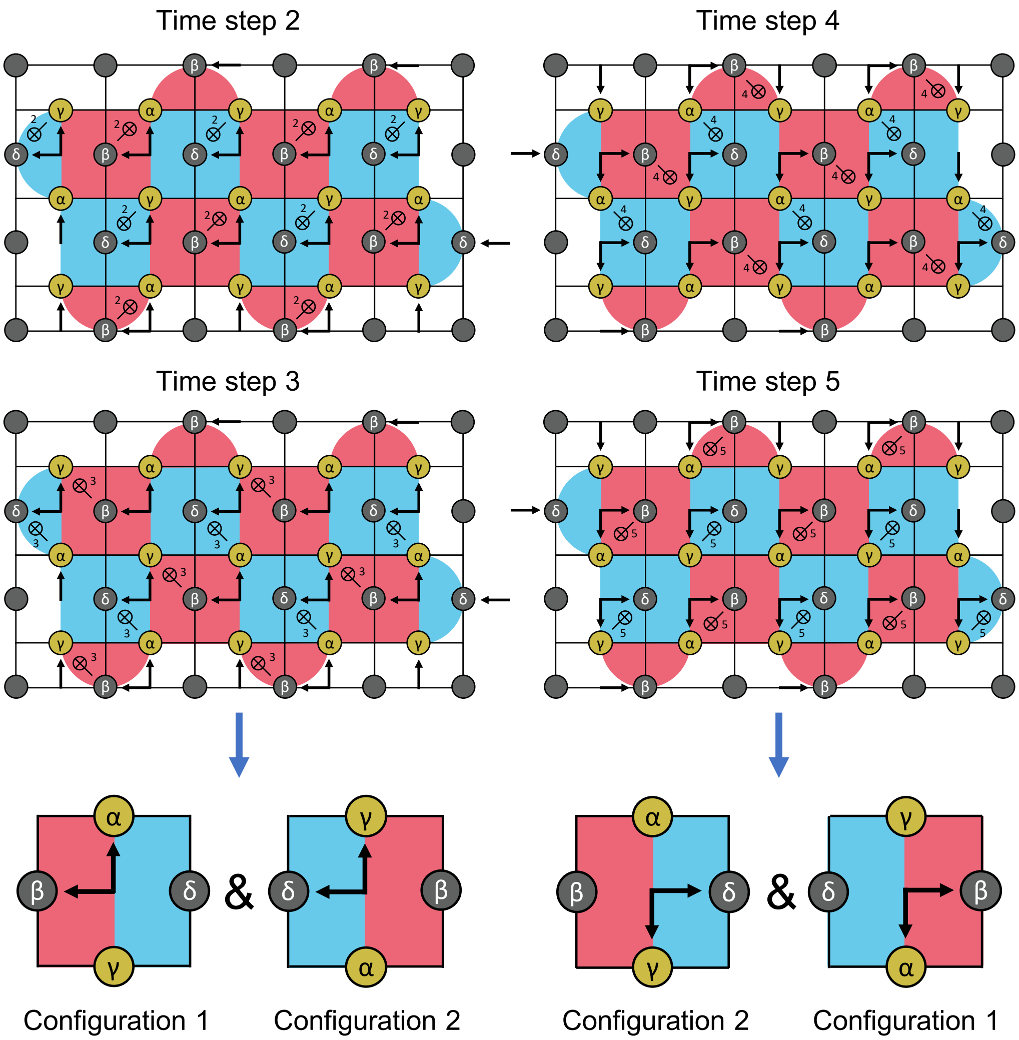


FIG. 25. Cat-qubit stabilization in the surface-code architecture. Each ATS is coupled to two data modes α, γ and two ancilla modes β, δ. In practice, ATSs are also coupled to a fifth readout mode (not shown here because it is not stabilized by any ATS). Each ATS is responsible for performing four CNOT gates (at different time steps) and stabilizing two phononic modes in the cat-code manifold during each time step. In the top panel, we show configurations of the cat-qubit stabilization which respect the constraint discussed in the main text: at each time step, a CNOT's target mode must be stabilized by an ATS that also couples to its control mode. Each phononic mode, pointed by a black arrow, is stabilized by an ATS where the black arrow originates from. In the bottom panel, we show two stabilization configurations in the perspective of each host ATS. In configuration 1 (2), modes α, β (γ, δ) are stabilized by the host ATS and the remaining modes γ, δ (α, β) are stabilized by some other neighboring ATSs.

engineered dissipation is suppressed by the filter. We emphasize that these conditions must be checked for both stabilization configurations in Fig. 25; checking both configurations is necessary in order to ensure that Type I and II crosstalk is suppressed by the filter at *all* time steps.

Second, C should be large if the coherent Type III errors have significant damaging effects, and we now quantify these effects in the context of the surface code. Recall that these errors are generated by effective Hamiltonian terms of the form (B38), which we repeat for convenience,

$$|\alpha|^4\chi\hat{Z}_i\hat{Z}_j\hat{Z}_\ell\hat{Z}_m e^{i(\delta_{\ell mn}-\delta_{ijk})t} + \text{H.c.}. \tag{B59}$$

When these terms are rapidly rotating, i.e., when $|\alpha^4\chi| \ll |\delta_{ijk} - \delta_{\ell mn}|$, their effects are suppressed. Indeed, these terms effectively induce detuned Rabi oscillations between states of different parity, and the magnitude of these oscillations is small in the far-detuned limit. To quantify this suppression, note that these micro-oscillation errors remain coherent during gates but can be converted to incoherent, correlated $\hat{Z}$ errors when the X-type stabilizers are measured. The probability $p_{ijk\ell mn}$ of inducing a correlated phase error upon a such a measurement scales quadratically in the ratio of the coupling strength and

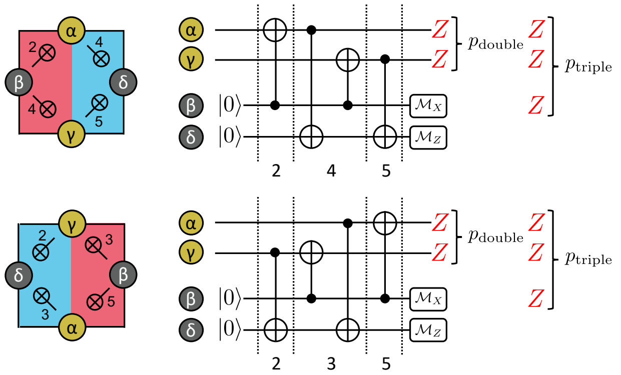


FIG. 26. Type III crosstalk errors in the surface-code architecture. We define $p_{\rm double}$ as the probability of getting a Type III error $\propto \hat{Z}_\alpha\hat{Z}_\gamma\hat{I}_\beta$, and $p_{\rm triple}$ as the probability of getting a Type III error $\propto \hat{Z}_\alpha\hat{Z}_\gamma\hat{Z}_\beta$.

detuning,

$$p_{ijk\ell mn} = \left(\frac{|\alpha^4\chi|}{\delta_{ijk}-\delta_{\ell mn}}\right)^2. \tag{B60}$$

Among the various Type III errors, we focus on those that induce phase errors in both of the data modes α and γ since such errors are specific to our architecture and not taken into account in the standard surface-code analysis. In particular, we define $p_{\rm double}$ as the total probability at least one Type III error $\propto \hat{Z}_\alpha\hat{Z}_\gamma\hat{I}_\beta$, and $p_{\rm triple}$ as the total probability of at least one Type III error $\propto \hat{Z}_\alpha\hat{Z}_\gamma\hat{Z}_\beta$. Explicitly,

$$p_{\rm double} = \sum_{\{ijk\ell mn\}\in\mathcal{D}} p_{ijk\ell mn}, \tag{B61}$$

$$p_{\rm triple} = \sum_{\{ijk\ell mn\}\in\mathcal{T}} p_{ijk\ell mn}, \tag{B62}$$

where $\mathcal{D}$ and $\mathcal{T}$ denote sets of indices that give rise to errors $\propto \hat{Z}_\alpha\hat{Z}_\gamma\hat{I}_\beta$ and $\propto \hat{Z}_\alpha\hat{Z}_\gamma\hat{Z}_\beta$, respectively, see Fig. 26. Note that the $\hat{Z}$ error on the ancilla mode β manifests as a flipped X-basis measurement outcome. On the other hand, $\hat{Z}$ errors on the other ancilla mode δ do not flip the measurement outcomes. This is because the mode δ is measured in the Z basis, and Z-basis measurements commute with $\hat{Z}$ errors.

We incorporate these Type III errors into the cost function as follows. We take $C = 1$ if Type I or II errors are not suppressed by the filter (see aforementioned conditions on the δ_{ijk}), and otherwise we take

$$C = \frac{1}{2}\left(p^{(1)}_{\rm double} + p^{(1)}_{\rm triple} + p^{(2)}_{\rm double} + p^{(2)}_{\rm triple}\right), \tag{B63}$$

where $p^{(i)}_{\rm double}$ and $p^{(i)}_{\rm triple}$ denote the values of $p_{\rm double}$ and $p_{\rm triple}$ for the i-th stabilization configuration. Equation (B63) thus represents the average probability of a Type III error occurring during one time step. Costs $C \ll 1$ are thus only achieved when both the probability of Type III errors is small, and all Type I and II errors are suppressed by the filter.

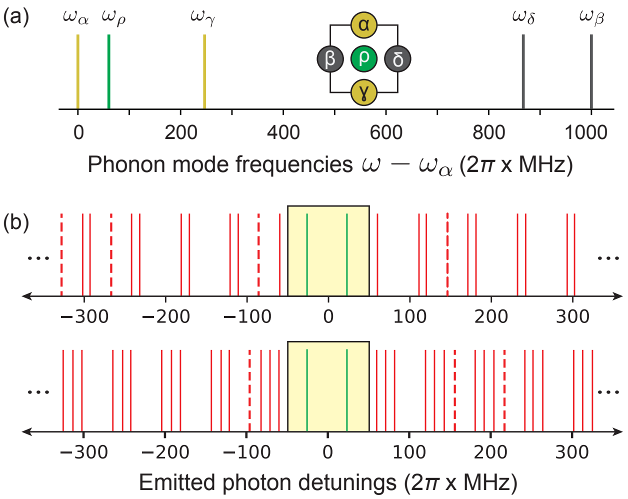


FIG. 27. Optimized mode frequencies. (a) Plot of the optimized frequencies of the five phonon modes. (b) Emitted photon detunings. Red dashed (solid) lines indicate photons emitted via parity-non-preserving Type I (Type II) processes. The yellow box covers the region $[-50, 50]$ $(2\pi \times \text{MHz})$, representing a bandpass filter with center frequency ω_b and a $4J = 2\pi \times 100$ MHz passband. The fact that no lines lie inside the yellow box indicates that all Type I and II processes are sufficiently far detuned so as to be suppressed by the filter. The top (bottom) plot in (b) is for the case where modes α and β (γ and δ) are stabilized simultaneously.

Having defined the cost function C, we perform a numerical search for the values of the mode frequencies and pump detunings which minimize the cost. In performing this optimization, we place two additional restrictions on allowed frequencies and detunings. First, we restrict the mode frequencies to lie within a 1 GHz bandwidth. This is done because the modes are supported by phononic-crystal-defect resonators (PCDRs), and as such all mode frequencies must lie within the phononic bandgap, or at least within the union of two separate bandgaps each associated with different PCDRs. These bandgaps are typically not more than 500 MHz wide for the devices we consider [28]. Second, we restrict the values of the detunings to $\Delta = \pm J$. This is done to maximize use of the filter bandwidth; emitted photons are detuned from one another by $2J$ and from the nearest band edge by J, see Fig. 4(a). Additionally, because the filter bandwidth $4J$ restricts the maximum achievable κ_2, we take J to be as large as possible while still allowing for $C \ll 1$.

The optimization results are listed in Table VI and illustrated in Figure 27. We perform optimization both for the usual case of five modes per ATS, as well as the case of four modes per ATS with the readout mode omitted (this omission is justified if the X readout is performed

# modes	$4J$	$\omega_\alpha, \omega_\beta, \omega_\gamma, \omega_\delta, \omega_\rho$	$\frac{1}{2}(p^{(1)}_{\text{double}} + p^{(2)}_{\text{double}})$	$\frac{1}{2}(p^{(1)}_{\text{triple}} + p^{(2)}_{\text{triple}})$	C
4	180	0, 1000, 798, 101, -	$1.22 \times 10^{-9} \left[\frac{\lvert\alpha\rvert^2 g_2}{2\pi\text{MHz}}\right]^4$	$3.87 \times 10^{-10} \left[\frac{\lvert\alpha\rvert^2 g_2}{2\pi\text{MHz}}\right]^4$	$1.60 \times 10^{-9} \left[\frac{\lvert\alpha\rvert^2 g_2}{2\pi\text{MHz}}\right]^4$
5	100	0, 1000, 242, 879, 61	$1.83 \times 10^{-8} \left[\frac{\lvert\alpha\rvert^2 g_2}{2\pi\text{MHz}}\right]^4$	$5.20 \times 10^{-10} \left[\frac{\lvert\alpha\rvert^2 g_2}{2\pi\text{MHz}}\right]^4$	$1.88 \times 10^{-8} \left[\frac{\lvert\alpha\rvert^2 g_2}{2\pi\text{MHz}}\right]^4$

TABLE VI. Frequency optimization results. The parameters $4J$ and ω are given in units of $2\pi\times$ MHz. The Type III error probabilities and the cost C are expressed in terms of α and g_2. For realistic choices of $|\alpha| = \sqrt{8}$ and $g_2/2\pi = 2$ MHz, the cost function evaluates to $C = 1.05 \times 10^{-4}$ and $C = 1.23 \times 10^{-3}$ for the four- and five-mode configurations respectively. We fix $-\Delta_\alpha = \Delta_\beta = -\Delta_\gamma = \Delta_\delta = J$.

directly using the ATS, see Appendix G). For the optimal configurations, *all* Type I and Type II errors are simultaneously suppressed by the filter. Note also that all emitted photon frequencies associated with Type I or II errors lie at least 10 MHz outside the filter passband. As a result, the optimized configuration is robust to deviations in the mode frequencies of the same order, and larger deviations can be tolerated by decreasing the filter bandwidth. Moreover, for realistic values of $|\alpha|$ and g_2, we have $C \ll 1$, indicating that Type III errors are strongly suppressed. Therefore, all dominant sources of crosstalk are strongly suppressed.

Appendix C: Shifted Fock basis

Simulating a large cat qubit (with large $|\alpha|^2 \gg 1$) by using the usual Fock basis becomes quickly inefficient. Here, we introduce a shifted Fock basis method which can describe large cat states in a more efficient way (i.e., using a smaller Hilbert space dimension) than the usual Fock basis. Specifically, we will explain how to construct the annihilation operator $\hat{a}$ in the shifted Fock basis.

Recall that a cat state is composed of two coherent state components $|\pm\alpha\rangle$ which can be understood as displaced vacuum states $\hat{D}(\pm\alpha)|\hat{n} = 0\rangle$, where $\hat{D}(\alpha) \equiv \exp[\alpha\hat{a}^\dagger - \alpha^*\hat{a}]$ is the displacement operator. In the shifted Fock basis, we use $2d$ displaced Fock states $\hat{D}(\pm\alpha)|\hat{n} = n\rangle$ as basis states, where $n \in \{0, \cdots, d-1\}$. Note that while displaced Fock states in each $\pm\alpha$ branch are orthonormalized, displaced Fock states in different branches are not necessarily orthogonal to each other. We thus need to orthonormalize the displaced Fock states.

We first define the non-orthonormalized basis states as follows:

$$|\phi_{n,\pm}\rangle \equiv \frac{1}{\sqrt{2}}\big[\hat{D}(\alpha) \pm (-1)^n \hat{D}(-\alpha)\big]|\hat{n} = n\rangle, \tag{C1}$$

where $|\phi_{n,+}\rangle$ and $|\phi_{n,-}\rangle$ have even and odd excitation number parity, respectively. Note that we grouped the non-orthonormalized states into the even and odd branches instead of the $\pm\alpha$ branches. As a result, in the ground state manifold ($n = 0$), the normalized basis states $|\phi_{0,\pm}\rangle$ are equivalent to the complementary basis states of the cat qubit $|\pm\rangle$, not the computational basis states $|0/1\rangle$, i.e.,

$$|\pm\rangle \propto |\phi_{0,\pm}\rangle = \frac{1}{\sqrt{2}}(|\alpha\rangle \pm |-\alpha\rangle). \tag{C2}$$

We use the even/odd branching convention so that any two basis states in different branches are orthogonal to each other and hence the orthonormalization can be done separately in each parity sector. Note that

$$\Phi^\pm_{m,n} \equiv \langle\phi_{m,\pm}|\phi_{n,\pm}\rangle = \delta_{m,n} \pm (-1)^m D_{m,n}(2\alpha), \tag{C3}$$

where $D_{m,n}(\alpha) \equiv \langle\hat{n} = m|\hat{D}(\alpha)|\hat{n} = n\rangle$ are the matrix elements of the displacement operator $\hat{D}(\alpha)$ in the usual Fock basis:

$$\begin{aligned} D_{m,n}(\alpha) = e^{-\frac{|\alpha|^2}{2}} \sqrt{\frac{\min(m,n)!}{\max(m,n)!}} L^{(|m-n|)}_{\min(m,n)}(|\alpha|^2) \\ \times \begin{cases} \alpha^{m-n} & m \geq n \\ (-\alpha^*)^{n-m} & m < n \end{cases}. \end{aligned} \tag{C4}$$

Here, $L^{(\alpha)}_n(x)$ is the generalized Laguerre polynomial. Since $|D_{m,n}(2\alpha)| = \mathcal{O}(|\alpha|^{m+n}e^{-2|\alpha|^2})$, $D_{m,n}(2\alpha)$ is negligible if $m + n \ll |\alpha|^2$. In this regime, the basis states $|\phi_{n,\pm}\rangle$ are almost orthonormal. For the purpose of estimating the phase-flip (or Z) error rates within a small multiplicative error, it is often permissible to neglect the non-orthogonality of the states $|\phi_{n,\pm}\rangle$. However, this is generally not the case if we want to evaluate the Z error rates with a very high precision or if we want to estimate the bit-flip (or X) error rates because the bit flip error rates decrease exponentially in $|\alpha|^2$. In these cases, taking into account the non-orthogonality of the states $|\phi_{n,\pm}\rangle$ is essential.

We orthonormalize the basis states $|\phi_{n,\pm}\rangle$ by applying the Gram-Schmidt orthonormalization procedure. Specifically, given the non-orthonormalized basis states $|\phi_{n,\pm}\rangle$, we construct d orthonormalized basis states in each parity sector starting from the ground state $|\phi_{0,\pm}\rangle$:

$$|\psi_{n,\pm}\rangle = \sum_{m=0}^{d-1} c^\pm_{m,n}|\phi_{m,\pm}\rangle. \tag{C5}$$

The coefficients $c^\pm_{m,n}$ ($0 \leq m, n \leq d-1$) are determined inductively. In the base case ($k=0$),

$$c^{\pm}_{0,0}=\frac{1}{\sqrt{\Phi^{\pm}_{0,0}}},\quad c^{\pm}_{m,0}=0\ \text{ for all }\ 1\leq m\leq d-1,\quad \text{(C6)}$$

and thus the logical $|\pm\rangle$ states of the cat qubit are given by

$$|\pm\rangle\equiv|\psi_{0,\pm}\rangle=\frac{1}{\sqrt{\Phi^{\pm}_{0,0}}}|\phi_{0,\pm}\rangle=\frac{|\alpha\rangle\pm|-\alpha\rangle}{\sqrt{2(1\pm e^{-2|\alpha|^2})}}.\quad \text{(C7)}$$

In the general case with $1\leq k\leq d-1$, suppose we are given with $c^{\pm}_{mn}$ for all $0\leq m\leq d-1$ and $0\leq n\leq k-1$. Thus, at this point, the first k columns of $c^{\pm}$ are known. Let $c^{\pm}_{:,0:k-1}$ be the $d\times k$ matrix which is obtained by taking the first k columns of the matrix $c^{\pm}$. Given $c^{\pm}_{:,0:k-1}$, we assign the $k+1^{\text{th}}$ column of $c^{\pm}$ as follows.

$$c^{\pm}_{m,k}=-\frac{(c^{\pm}_{:,0:k-1}(c^{\pm}_{:,0:k-1})^{\dagger}\Phi^{\pm})_{m,k}}{\sqrt{\Phi^{\pm}_{k,k}-((\Phi^{\pm})^{\dagger}c^{\pm}_{:,0:k-1}(c^{\pm}_{:,0:k-1})^{\dagger}\Phi^{\pm})_{k,k}}},\quad \text{(C8)}$$

for $0\leq m\leq k-1$,

$$c^{\pm}_{k,k}=\frac{1}{\sqrt{\Phi^{\pm}_{k,k}-((\Phi^{\pm})^{\dagger}c^{\pm}_{:,0:k-1}(c^{\pm}_{:,0:k-1})^{\dagger}\Phi^{\pm})_{k,k}}},\quad \text{(C9)}$$

and $c^{\pm}_{m,k}=0$ for all $m>k$.

Having constructed the $2d$ orthonormalized shifted Fock basis states $|\psi_{n,\pm}\rangle$, we now need to find the matrix elements of an operator $\hat{O}$ (e.g., $\hat{O}=\hat{a}$) in the orthonormalized basis. Let $|\phi_n\rangle=|\phi_{n,+}\rangle$ and $|\phi_{n+d}\rangle=|\phi_{n,-}\rangle$ for $n\in\{0,\cdots,d-1\}$ and also define $|\psi_n\rangle$ and $|\psi_{n+d}\rangle$ similarly. Suppose that the operator $\hat{O}$ transforms the non-orthonormalized basis states $|\phi_n\rangle$ as follows

$$\hat{O}|\phi_n\rangle=\sum_{m=0}^{2d-1}O_{m,n}|\phi_m\rangle.\quad \text{(C10)}$$

We call $O_{m,n}$ the matrix elements of the operator $\hat{O}$ in the non-orthonormalized basis $|\phi_n\rangle$. Then, in the orthonormalized basis, the matrix elements of the operator $\hat{O}$ are given by

$$O'_{m,n}\equiv\langle\psi_m|\hat{O}|\psi_n\rangle=(c^{\dagger}\Phi Oc)_{m,n},\quad \text{(C11)}$$

where Φ and c are $2d\times 2d$ matrices which are defined as

$$\Phi=\begin{bmatrix}\Phi^{+}&0\\0&\Phi^{-}\end{bmatrix},\quad c=\begin{bmatrix}c^{+}&0\\0&c^{-}\end{bmatrix}.\quad \text{(C12)}$$

The matrix elements of the $d\times d$ matrices $\Phi^{\pm}$ and $c^{\pm}$ are given in Eqs. (C3), (C6), (C8) and (C9).

Consider the annihilation operator $\hat{O}=\hat{a}$ and note that it transforms the non-orthonormalized basis states $|\phi_{n,\pm}\rangle$ as follows:

$$\hat{a}|\phi_{n,\pm}\rangle=\sqrt{n}|\phi_{n-1,\mp}\rangle+\alpha|\phi_{n,\mp}\rangle.\quad \text{(C13)}$$

Note that the annihilation operator $\hat{a}$ flips the $\pm$ parity to the $\mp$ parity. Thus, in the non-orthonormalized basis, the matrix elements of the annihilation operator are given by

$$\begin{bmatrix}0&\hat{b}+\alpha\\\hat{b}+\alpha&0\end{bmatrix}=\hat{X}\otimes(\hat{b}+\alpha),\quad \text{(C14)}$$

where $\hat{X}$ is the Pauli X operator and $\hat{b}$ is the truncated annhilation operator of size $d\times d$. Then, the matrix elements of the annihilaton operator in the orthonormalized basis $|\psi_{n,\pm}\rangle$ can be obtained via the transformation given in Eq. (C11).

Recall that $|\psi_{n,\pm}\rangle$ are complementary basis states. To find the matrix elements of an operator in the computational basis states, we should conjugate the matrix by the Hadamard operator $\hat{H}$. Thus, in the orthonormalized computational basis, the annihilation operator is given by

$$\begin{aligned}\hat{a}&\equiv(\hat{H}\otimes\hat{I})\cdot c^{\dagger}\Phi(\hat{X}\otimes(\hat{b}+\alpha))c\cdot(\hat{H}\otimes\hat{I})\\&\xrightarrow{|\alpha|^2\gg d}\hat{Z}\otimes(\hat{b}+\alpha).\end{aligned}\quad \text{(C15)}$$

The approximate expression $\hat{a}\simeq\hat{Z}\otimes(\hat{b}+\alpha)$ is useful for analyzing the Z error rates of large cat qubits (with $|\alpha|\gg 1$) in the perturbative regime where the cat qubit states may sometimes be excited to the first excited state manifold ($n=1$) but quickly decay back to the ground state manifold ($n=0$). In particular, the engineered two-phonon dissipator $\kappa_2\mathcal{D}[\hat{a}^2-\alpha^2]$ is given by

$$\kappa_2\mathcal{D}[\hat{I}\otimes(\hat{b}^2+2\alpha\hat{b})]\simeq 4\kappa_2\alpha^2\mathcal{D}[\hat{I}\otimes\hat{b}]\quad \text{(C16)}$$

by using the approximate expression $\hat{a}\simeq\hat{Z}\otimes(\hat{b}+\alpha)$ and disregarding higher than second excited states (i.e., $\hat{b}^2=0$). Hence, the linewidth of the engineered two-phonon dissipation is approximately given by $4\kappa_2\alpha^2$, which is twice the confinement rate $\kappa_{\text{conf}}=2\kappa_2\alpha^2$ [25]. The factor of 2 difference is simply due to the fact that quadrature operators decay on average with a rate $\kappa/2$ (corresponding to the confinement rate $\kappa_{\text{conf}}=2\kappa_2\alpha^2$) under the excitation loss rate κ (corresponding to $4\kappa_2\alpha^2$). For numerical simulations (Appendix E), we thoroughly take into account the orthonormalization and use the orthonormalized shifted Fock basis obtained by the Gram-Schmidt process. We lastly remark that the parity operator $e^{i\hat{\pi}\hat{a}^{\dagger}\hat{a}}$ is exactly given by $\hat{X}\otimes\hat{I}$ in the shifted Fock basis (with the orthonormalization accounted for) because of the way we define the basis states, i.e., $|\psi_{n,+}\rangle$ ($|\psi_{n,-}\rangle$) has an even (odd) excitation number parity.

Appendix D: Perturbative analysis of the Z error rates of the cat qubit gates

Here, we analyze the Z error rates of the cat qubit gates (idling, Z rotations, CZ rotations, CNOT, and Toffoli) by using the shifted Fock basis (Appendix C) and adiabatic elimination or effective operator formalism (Appendix B 1).

1. Idling

Consider an idling single cat qubit which is stabilized by the two-phonon dissipation $\kappa_2\mathcal{D}[\hat{a}^2-\alpha^2]$ and is subject to single-phonon loss $\kappa_1\mathcal{D}[\hat{a}]$:

$$\frac{d\hat{\rho}(t)}{dt} = \kappa_2\mathcal{D}[\hat{a}^2-\alpha^2]\hat{\rho}(t)+\kappa_1\mathcal{D}[\hat{a}]\hat{\rho}(t). \tag{D1}$$

Assuming $|\alpha|\gg 1$, the above master equation is given by

$$\begin{aligned}\frac{d\hat{\rho}(t)}{dt} &= \kappa_2\mathcal{D}[\hat{I}\otimes(\hat{b}^2+2\alpha\hat{b})]\hat{\rho}(t)\\ &\quad+\kappa_1\mathcal{D}[\hat{Z}\otimes(\hat{b}+\alpha)]\hat{\rho}(t)\end{aligned} \tag{D2}$$

in the shifted Fock basis, where we used the mapping $\hat{a}\to\hat{Z}\otimes(\hat{b}+\alpha)$. Suppose that the system is initially in the cat qubit manifold, i.e., $\hat{\rho}(0)=\hat{\rho}_g(0)\otimes|0\rangle'\langle 0|'$, where $\hat{\rho}_g(0)$ is a density operator of size 2×2 and $|0\rangle'\equiv|\hat{b}^\dagger\hat{b}=0\rangle$ (not to be confused with the computational basis state $|0\rangle$). When the system is idling, the states are never excited to the excited state manifold and thus $\hat{\rho}(t)=\hat{\rho}_g(t)\otimes|0\rangle'\langle 0|'$. Projecting the master equation in Eq. (D2) to the ground state manifold, we find

$$\frac{d\hat{\rho}_g(t)}{dt} = \kappa_1\alpha^2\mathcal{D}[\hat{Z}]\hat{\rho}_g(t), \tag{D3}$$

and hence

$$\hat{\rho}_g(T)\simeq(1-\bar{p}_Z)\hat{\rho}_g(t)+\bar{p}_Z\hat{Z}\hat{\rho}_g(t)\hat{Z}, \tag{D4}$$

provided that the idling Z error rate (per gate) $\bar{p}_Z\equiv\kappa_1\alpha^2 T$ is small (i.e., $\bar{p}_Z\ll 1$) where T is the idling time. Note that we used the notation $\bar{p}$ with a bar to indicate that the presented expression is obtained via a perturbative analysis. We use p without bar to refer to numerical results.

2. Z rotations

Assume that α is real and positive. To implement a Z rotation $Z(\theta)\equiv\exp[i\theta|1\rangle\langle 1|]$ on a cat qubit (where $|1\rangle\simeq|-\alpha\rangle$ is a cat-code computational basis state), we need to apply a linear drive $\epsilon_Z(\hat{a}+\hat{a}^\dagger)$:

$$\begin{aligned}\frac{d\hat{\rho}(t)}{dt} &= \kappa_2\mathcal{D}[\hat{a}^2-\alpha^2]\hat{\rho}(t)+\kappa_1\mathcal{D}[\hat{a}]\hat{\rho}(t)\\ &\quad-i[\epsilon_Z(\hat{a}+\hat{a}^\dagger),\hat{\rho}(t)].\end{aligned} \tag{D5}$$

In the shifted Fock basis, the linear drive $\epsilon_Z(\hat{a}+\hat{a}^\dagger)$ is given by $\epsilon_Z\hat{Z}\otimes(\hat{b}+\hat{b}^\dagger+2\alpha)$. Thus, in the ground state manifold, it induces a Z rotation via the term $2\epsilon_Z\alpha\hat{Z}$. At the same time, the term $\epsilon_Z\hat{Z}\otimes\hat{b}^\dagger$ excites the cat qubit to its first excited state, which then quickly decays back to the ground state manifold due to the engineered dissipaton $\kappa_2\mathcal{D}[\hat{a}^2-\alpha^2]\leftrightarrow\kappa_2\mathcal{D}[\hat{I}\otimes(\hat{b}^2+2\alpha\hat{b})]$. Thus, to capture the first order effects, we only consider the ground state manifold and the first excited state manifold ($n=0,1$), hence ignoring $\hat{b}^2$ in $\kappa_2\mathcal{D}[\hat{I}\otimes(\hat{b}^2+2\alpha\hat{b})]$. Also, assuming $\kappa_2\gg\kappa_1$, we ignore the intrinsic decay due to the single photon loss in the excited state manifold, i.e., $\kappa_1\mathcal{D}[\hat{a}]\leftrightarrow\kappa_1\mathcal{D}[\hat{Z}\otimes(\hat{b}+\alpha)]\simeq\kappa_1\alpha^2\mathcal{D}[\hat{Z}\otimes\hat{I}]$, where we used $\mathcal{D}[c\hat{A}]=|c|^2\mathcal{D}[\hat{A}]$. Then, the master equation is given by

$$\begin{aligned}\frac{d\hat{\rho}(t)}{dt} &= 4\kappa_2\alpha^2\mathcal{D}[\hat{I}\otimes\hat{b}]\hat{\rho}(t)+\kappa_1\alpha^2\mathcal{D}[\hat{Z}\otimes\hat{I}]\hat{\rho}(t)\\ &\quad-2i\alpha\epsilon_Z[\hat{Z}\otimes\hat{I},\hat{\rho}(t)]-i[\epsilon_Z\hat{Z}\otimes(\hat{b}+\hat{b}^\dagger),\hat{\rho}(t)]\end{aligned} \tag{D6}$$

The second term on the right-hand side of this master equation describes a Z error acting on the encoded cat qubit due to single-photon loss, occurring at the rate (per time) $\kappa_1\alpha^2$. The third term rotates the cat qubit about the Z axis. The fourth term excites the cat qubit from its ground-state manifold to its first-excited-state manifold, with a coupling strength $g=\epsilon_Z$, and at the same time inflicts a Z error on the cat qubit. This excitation decays back to the cat code ground-state manifold with a decay rate $\kappa=4\kappa_2\alpha^2$ due to the engineered dissipation described by the first term. Assuming $\kappa\gg g$ the creation and decay of this excitation results in an additional Z error in the ground state manifold with an effective error rate (per time) $4g^2/\kappa=\epsilon_Z^2/(\kappa_2\alpha^2)$, augmenting the Z error rate due to single-photon loss. The effective master equation therefore becomes

$$\frac{d\hat{\rho}_g(t)}{dt} = \Big(\kappa_1\alpha^2+\frac{\epsilon_Z^2}{\kappa_2\alpha^2}\Big)\mathcal{D}[\hat{Z}]\hat{\rho}_g(t)-i[2\epsilon_Z\alpha\hat{Z},\hat{\rho}_g(t)], \tag{D7}$$

where we have used the subscript g to indicate that $\hat{\rho}_g(t)$ is the density operator in the ground-state manifold of the cat state.

Given this effective master equation, we can analyze the effective Hamiltonian and the effective phase-flip error separately because they commute with each other. The effective Hamiltonian $\hat{H}_{\text{eff}}=2\epsilon_Z\alpha\hat{Z}$ induces a Z rotation $\hat{Z}(\theta)$ with $\theta=4\epsilon_Z\alpha T$ after the gate time T, i.e., $\epsilon_Z=\theta/(4\alpha T)$. Then, the Z error rate (per gate) due to the effective phase-flip is given by

$$\bar{p}_Z = \kappa_1\alpha^2 T + \frac{\epsilon_Z^2}{\kappa_2\alpha^2}T = \kappa_1\alpha^2 T + \frac{\theta^2}{16\kappa_2\alpha^4 T}, \tag{D8}$$

provided that $\bar{p}_Z \ll 1$. This Z error rate is minimized at the optimal gate time

$$\bar{T}^{\star}_{Z(\theta)} = \frac{|\theta|}{4\alpha^3\sqrt{\kappa_1\kappa_2}}, \tag{D9}$$

and the corresponding optimal Z error rate is given by

$$\bar{p}^{\star}_Z = \frac{|\theta|}{2\alpha}\sqrt{\frac{\kappa_1}{\kappa_2}}. \tag{D10}$$

3. CZ rotations

A ZZ interaction between two cat qubits can be implemented by using a beam-splitter coupling $\epsilon_{ZZ}(\hat{a}_1\hat{a}_2^\dagger + \hat{a}_1^\dagger\hat{a}_2)$, which is given by $2\epsilon_{ZZ}\alpha^2\hat{Z}_1\hat{Z}_2$ in the ground state manifold of the cat qubits. To implement a controlled Z rotation $CZ(\theta) \equiv \exp[i\theta|11\rangle\langle 11|]$, we should add single-qubit Z rotations so that only the state $|11\rangle$ accumulates a non-trivial phase. More specifically, we need

$$\begin{aligned}\hat{H} = \epsilon_{ZZ}(\hat{a}_1\hat{a}_2^\dagger + \hat{a}_1^\dagger\hat{a}_2) - \epsilon_{ZZ}\alpha(\hat{a}_1 + \hat{a}_1^\dagger)\\ - \epsilon_{ZZ}\alpha(\hat{a}_2 + \hat{a}_2^\dagger),\end{aligned} \tag{D11}$$

and the master equation is given by

$$\begin{aligned}\frac{d\hat{\rho}(t)}{dt} &= \kappa_2\Big[\mathcal{D}[\hat{a}_1^2 - \alpha^2] + \mathcal{D}[\hat{a}_2^2 - \alpha^2]\Big]\hat{\rho}(t)\\ &+ \kappa_1\Big[\mathcal{D}[\hat{a}_1] + \mathcal{D}[\hat{a}_2]\Big]\hat{\rho}(t) - i[\hat{H}, \hat{\rho}(t)].\end{aligned} \tag{D12}$$

Similarly as in the case of single-qubit Z rotations, the engineered dissipation induces a strong decay from the first excited state manifold to the cat qubit manifold with a decay rate (per time) $\kappa = 4\kappa_2\alpha^2$. Also, the single-phonon loss causes local phase-flip errors in each cat qubit manifold with an error rate $\kappa_1\alpha^2$. In the shifted Fock basis, the Hamiltonian $\hat{H}$ is given by

$$\begin{aligned}\hat{H} &= 2\epsilon_{ZZ}\alpha^2(\hat{Z}_1\hat{Z}_2 - \hat{Z}_1 - \hat{Z}_2)\otimes\hat{I}\\ &+ \epsilon_{ZZ}\alpha(\hat{Z}_1\hat{Z}_2 - \hat{Z}_1)\otimes(\hat{b}_1 + \hat{b}_1^\dagger)\\ &+ \epsilon_{ZZ}\alpha(\hat{Z}_1\hat{Z}_2 - \hat{Z}_2)\otimes(\hat{b}_2 + \hat{b}_2^\dagger)\\ &+ \epsilon_{ZZ}\hat{Z}_1\hat{Z}_2\otimes(\hat{b}_1\hat{b}_2^\dagger + \hat{b}_1^\dagger\hat{b}_2).\end{aligned} \tag{D13}$$

The first term generates an effective Hamiltonian $\hat{H}_{\text{eff}} = 8\epsilon_{ZZ}\alpha^2|11\rangle\langle 11|$ in the ground state manifold. Due to the second (third) term, the first (second) cat qubit is excited to its first excited state manifold with a coupling strength $g = \epsilon_{ZZ}\alpha$ while the encoded cat qubits are subjected to a $\hat{Z}_1\hat{Z}_2 - \hat{Z}_1$ ($\hat{Z}_1\hat{Z}_2 - \hat{Z}_2$) error. The excited state decays back to the ground-state manifold at the rate $\kappa = 4\kappa_2\alpha^2$ due to the engineered dissipation; as a result the cat qubits experience effective $\hat{Z}_1\hat{Z}_2 - \hat{Z}_1$ and $\hat{Z}_1\hat{Z}_2 - \hat{Z}_2$ errors, each with rate (per time) $4g^2/\kappa = \epsilon_{ZZ}^2/\kappa_2$. Note that the last term in the effective Hamiltonian can in principle induce excitation exchange between the two modes but we may neglect this effect because the excited states decay very quickly back to the ground state manifold (i.e., $\epsilon_{ZZ} \ll 4\kappa_2\alpha^2$ which is indeed the case in the parameter regime we focus on). Putting all this together, we find the following effective master equation in the ground-state manifold of two cat qubits:

$$\begin{aligned}\frac{d\hat{\rho}_g(t)}{dt} &= \kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1] + \mathcal{D}[\hat{Z}_2]\Big]\hat{\rho}_g(t)\\ &+ \frac{\epsilon_{ZZ}^2}{\kappa_2}\Big[\mathcal{D}[\hat{Z}_1\hat{Z}_2 - \hat{Z}_1] + \mathcal{D}[\hat{Z}_1\hat{Z}_2 - \hat{Z}_2]\Big]\hat{\rho}_g(t)\\ &- i[8\epsilon_{ZZ}\alpha^2|11\rangle\langle 11|, \hat{\rho}_g(t)].\end{aligned} \tag{D14}$$

The effective Hamiltonian (which commutes with the Z-type effective jump operators) generates a CZ rotation $CZ(\theta)$ with $\theta = -8\epsilon_{ZZ}\alpha^2 T$ where T is the gate time. Hence, $\epsilon_{ZZ} = -\theta/(8\alpha^2 T)$. The remaining effective jump operators induce an error channel

$$\begin{aligned}\mathcal{N}_{CZ(\theta)}(\hat{\rho}) &\simeq \hat{\rho} + \kappa_1\alpha^2 T\Big[\mathcal{D}[\hat{Z}_1] + \mathcal{D}[\hat{Z}_2]\Big]\hat{\rho}\\ &+ \frac{\theta^2}{64\kappa_2\alpha^4 T}\Big[\mathcal{D}[\hat{Z}_1\hat{Z}_2 - \hat{Z}_1]\\ &+ \mathcal{D}[\hat{Z}_1\hat{Z}_2 - \hat{Z}_2]\Big]\hat{\rho},\end{aligned} \tag{D15}$$

provided that the error rates (per gate) $\kappa_1\alpha^2 T$ and $\theta^2/(64\kappa_2\alpha^4 T)$ are much smaller than unity. Ignoring the off-diagonal terms like $\hat{Z}_1\hat{Z}_2\hat{\rho}\hat{Z}_1$, we get Pauli Z error rates

$$\begin{aligned}\bar{p}_{Z_1} &= \bar{p}_{Z_2} = \kappa_1\alpha^2 T + \frac{\theta^2}{64\kappa_2\alpha^4 T},\\ \bar{p}_{Z_1Z_2} &= \frac{\theta^2}{32\kappa_2\alpha^4 T}.\end{aligned} \tag{D16}$$

The total gate infidelity $1 - \bar{p}_{Z_1} - \bar{p}_{Z_2} - \bar{p}_{Z_1Z_2}$ is minimized at the optimal gate time

$$\bar{T}^{\star}_{CZ(\theta)} = \frac{|\theta|}{4\alpha^3\sqrt{2\kappa_1\kappa_2}}, \tag{D17}$$

and the Z error rates (per gate) at this optimal gate time are given by

$$\bar{p}^{\star}_{Z_1} = \bar{p}^{\star}_{Z_2} = \frac{3}{2}p^{\star}_{Z_1Z_2} = \frac{3|\theta|}{8\alpha}\sqrt{\frac{\kappa_1}{2\kappa_2}}. \tag{D18}$$

Note that the optimal Z error rates for Z and CZ rotations decrease as α increases. Below, we show that this is not the case for the CNOT and Toffoli gates.

4. CNOT

The CNOT gate between two cat qubits can be realized by

$$\frac{d\hat{\rho}(t)}{dt} = \kappa_2\Big[\mathcal{D}[\hat{a}_1^2-\alpha^2]+\mathcal{D}[\hat{L}_2(t)]\Big]\hat{\rho}(t) + \kappa_1\Big[\mathcal{D}[\hat{a}_1]+\mathcal{D}[\hat{a}_2]\Big]\hat{\rho}(t) - i[\hat{H},\hat{\rho}(t)], \tag{D19}$$

where $\hat{a}_1$ and $\hat{a}_2$ are the annihilation operators of the control and the target modes, respectively, and $\hat{L}_2(t)$ and $\hat{H}$ are given by

$$\begin{aligned}\hat{L}_2(t) &= \hat{a}_2^2-\alpha^2+\frac{\alpha}{2}(e^{2i\frac{\pi}{T}t}-1)(\hat{a}_1-\alpha),\\ \hat{H} &= \frac{\pi}{4\alpha T}(\hat{a}_1+\hat{a}_1^\dagger-2\alpha)(\hat{a}_2^\dagger\hat{a}_2-\alpha^2).\end{aligned} \tag{D20}$$

How and whether this master equation can be physically implemented is discussed in Appendix F. Here, we focus on analyzing the effective Z error rates on the cat qubits under this master equation.

Note that the time-dependent engineered jump operator $\hat{L}_2(t)$ stabilizes the target mode in the $|\pm\alpha\rangle$ (or $|\pm\alpha e^{i\frac{\pi}{T}t}\rangle$) manifold if the control cat qubit is in the $|0\rangle\simeq|\alpha\rangle$ (or $|1\rangle\simeq|-\alpha\rangle$) state. As a result, the target cat qubit is rotated by 180° at time $t=T$ only if the control qubit is in the $|1\rangle$ state. That is, an $\hat{X}$ gate is applied to the target cat qubit (i.e., $|\pm\alpha\rangle\rightarrow|\mp\alpha\rangle$) conditioned on the control cat qubit being in the $|1\rangle$ state, hence the desired CNOT gate. Note that for this conditional stabilization to work, the engineered jump operator $\hat{L}_2$ should be modulated adiabatically (i.e., $T\gg 1/(\kappa_2\alpha^2)$) such that the target mode does not leak out of the $|\pm\alpha e^{i\frac{\pi}{T}t}\rangle$ manifold if the control qubit is in the $|1\rangle$ state. Adverse effects due to the non-adiabaticity can be partially (but not fully) compensated for by the compensating Hamiltonian $\hat{H}$. See more on this below.

To analyze this master equation, we first use a hybrid basis where the control and the target modes are described by the shifted and usual Fock basis, respectively. In the hybrid basis, assuming $|\alpha|\gg 1$ and using an approximate expression $\hat{a}\simeq\hat{Z}\otimes(\hat{b}+\alpha)$, the compensating Hamiltonian is given by

$$\hat{H} = -\frac{\pi}{T}|1\rangle\langle 1|_1\otimes(\hat{a}_2^\dagger\hat{a}_2-\alpha^2) + \frac{\pi}{4\alpha T}\hat{Z}_1\otimes(\hat{b}_1+\hat{b}_1^\dagger)(\hat{a}_2^\dagger\hat{a}_2-\alpha^2). \tag{D21}$$

Since we are using the shifted Fock basis for the control mode and the usual Fock basis for the target mode at this point, $\hat{b}_1$ is a $d\times d$ matrix whereas $\hat{a}_2$ is a $2d\times 2d$ matrix, where d is defined in Appendix C.

Note that the first term in Eq. (D21), which is a desired term, rotates the target mode conditioned on the control mode being in the $|1\rangle$ state branch. Hence, this term actively brings the target mode to the $|\pm\alpha e^{i\frac{\pi}{T}t}\rangle$ manifold (if the control qubit is in the $|1\rangle$ state) and thus makes it unnecessary for the system to adiabatically relax under the engineered jump operator $\hat{L}_2$. In particular, conditioned on the control qubit being in the $|1\rangle$ state, this term makes the target mode rotate by 180° at $t=T$, implementing an X gate (i.e., $|\pm\alpha\rangle\rightarrow|\mp\alpha\rangle$) to the target cat qubit. While the first term compensates for the adverse effects of the non-adiabaticity, the second term induces an undesirable back-action to the control mode which, as we show below, turns out to be a significant error source for the CNOT gate. Intuitively, the reason why the second term is detrimental is because the cat states in the target mode are not eigenstates of the excitation number operator $\hat{a}_2^\dagger\hat{a}_2$ and rather follow a Poissonian-like distribution with mean excitation number α^2. Due to such fluctuations in the excitation number of the target mode, the undesired second term makes the control mode leak out of its ground state manifold and at the same time causes a Z error on the control qubit space. How this undesired term degrades the CNOT gate fidelity can be best described in a rotating frame and in the full shifted Fock basis which we describe below.

Now, we go to a rotating frame with respect to the desired compensating Hamiltonian

$$\hat{H}' \equiv -\frac{\pi}{T}|1\rangle\langle 1|_1\otimes(\hat{a}_2^\dagger\hat{a}_2-\alpha^2), \tag{D22}$$

that is, we consider the time evolution of $\hat{\rho}_I(t)\equiv e^{i\hat{H}'t}\hat{\rho}(t)e^{-i\hat{H}'t}$ which should ideally be idling. In the rotating frame (assuming $|\alpha|\gg 1$), the annihilation operator of the control mode $\hat{Z}_1\otimes(\hat{b}_1+\alpha)$ is unchanged since $\hat{Z}_1$ commutes with $|1\rangle\langle 1|_1$ in $\hat{H}'$ (this is not the case when the orthonormalization is taken into account as there are exponentially small time-dependent corrections to $\hat{a}_1$ in the rotating frame). On the other hand, $\hat{a}_2$ is transformed as

$$\begin{aligned}\hat{a}_2 &\rightarrow e^{i\hat{H}'t}\hat{a}_2e^{-i\hat{H}'t}\\ &= |0\rangle\langle 0|_1\otimes\hat{a}_2+|1\rangle\langle 1|_1\otimes\hat{a}_2e^{i\frac{\pi}{T}t} = \hat{Z}_1\Big(\frac{\pi}{T}t\Big)\otimes\hat{a}_2,\end{aligned} \tag{D23}$$

where we define $\hat{Z}_k(\theta)$ as $\hat{Z}_k(\theta)\equiv\exp[i\theta|1\rangle\langle 1|_k]$. Having moved to the rotating frame, we finally use the shifted Fock basis for the target mode and replace $\hat{a}_2$ by $\hat{Z}_2\otimes(\hat{b}_2+\alpha)$.

In the rotating frame (and in the full shifted Fock basis), the master equation is given by

$$\begin{aligned}\frac{d\hat{\rho}_I(t)}{dt} &= \kappa_2\Big[\mathcal{D}[\hat{I}_{1,2}\otimes(\hat{b}_1^2+2\alpha\hat{b}_1)]+\mathcal{D}[\hat{L}_2'(t)]\Big]\hat{\rho}_I(t)\\ &\quad+\kappa_1\Big[\mathcal{D}[\hat{Z}_1\otimes(\hat{b}_1+\alpha)]\\ &\qquad+\mathcal{D}[\hat{Z}_1\Big(\frac{\pi}{T}t\Big)\hat{Z}_2\otimes(\hat{b}_2+\alpha)]\Big]\hat{\rho}_I(t)\\ &\quad-i\Big[\frac{\pi}{4\alpha T}\hat{Z}_1\otimes(\hat{b}_1+\hat{b}_1^\dagger)(\hat{b}_2^\dagger\hat{b}_2+\alpha(\hat{b}_2+\hat{b}_2^\dagger)),\hat{\rho}_I(t)\Big],\end{aligned} \tag{D24}$$

where the jump operator $\hat{L}'_2(t) \equiv e^{i\hat{H}'t}\hat{L}_2(t)e^{-i\hat{H}'t}$ in the rotating frame is given by

$$\hat{L}'_2(t) = \hat{Z}_1\Big(\frac{2\pi}{T}t\Big) \otimes (\hat{b}_2^2 + 2\alpha\hat{b}_2) + \frac{\alpha}{2}(e^{2i\frac{\pi}{T}t} - 1)\hat{Z}_1 \otimes \hat{b}_1. \tag{D25}$$

Similarly as in the case of Z and CZ rotations, we only consider the first excited state in each mode ($\hat{b}_1^2 = \hat{b}_2^2 = 0$) and ignore weak internal couplings and dissipations within the excited state manifold assuming that the engineered dissipation rate κ_2 dominates. Lastly, we ignore the second term in the jump operator $\hat{L}_2(t)$ to not complicate the analysis and convey the main idea more easily. This approximation can have a minor quantitative impact as the second term in $\hat{L}_2(t)$ is only four times weaker than the first term in the worst case ($t = T/2$). However, the key qualitative features (e.g., scaling) are not affected by this simplification.

With the above simplifications, the master equation is given by

$$\begin{aligned}\frac{d\hat{\rho}_I(t)}{dt} &= 4\kappa_2\alpha^2\Big[\mathcal{D}[\hat{I}_{1,2} \otimes \hat{b}_1] + \mathcal{D}[\hat{Z}_1\Big(\frac{2\pi}{T}t\Big) \otimes \hat{b}_2]\Big]\hat{\rho}_I(t)\\ &\quad + \kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1 \otimes \hat{I}] + \mathcal{D}[\hat{Z}_1\Big(\frac{\pi}{T}t\Big)\hat{Z}_2 \otimes \hat{I}]\Big]\hat{\rho}_I(t)\\ &\quad - i\Big[\frac{\pi}{4T}\hat{Z}_1 \otimes (\hat{b}_1\hat{b}_2 + \hat{b}_1^\dagger\hat{b}_2^\dagger), \hat{\rho}_I(t)\Big].\end{aligned} \tag{D26}$$

Note that the undesired term in the compensating Hamiltonian $\hat{H} - \hat{H}' = \frac{\pi}{4T}\hat{Z}_1 \otimes (\hat{b}_1\hat{b}_2 + \hat{b}_1^\dagger\hat{b}_2^\dagger)$ jointly excites both the control and the target modes with a coupling strength $g = \pi/(4T)$ and at the same time causes a $\hat{Z}_1$ error on the control qubit. The excited state $|11\rangle'$ (defined as $|\hat{b}_1^\dagger\hat{b}_1 = 1\rangle \otimes |\hat{b}_2^\dagger\hat{b}_2 = 1\rangle$, not to be confused with the computational basis state $|11\rangle$) eventually decays back to the code space through either $|11\rangle' \to |01\rangle' \to |00\rangle'$ or $|11\rangle' \to |10\rangle' \to |00\rangle'$ with a total decay rate (per time) $\kappa = 8\kappa_2\alpha^2$. Note that whichever way the excited state decays, the decay is accompanied by a Z rotation on the control mode, i.e., $\hat{Z}_1(\frac{2\pi}{T}t)$. Thus, after adiabatically eliminating the excited states, we get an effective jump operator $\hat{Z}_1\hat{Z}_1(\frac{2\pi}{T}t)$ with a decay rate (per time) $4g^2/\kappa = \pi^2/(32\kappa_2\alpha^2T^2)$ in the ground state manifold. Thus, we have the following effective master equation.

$$\begin{aligned}\frac{d\hat{\rho}_{I,g}(t)}{dt} &= \kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1] + \mathcal{D}[\hat{Z}_1\Big(\frac{\pi}{T}t\Big)\hat{Z}_2]\Big]\hat{\rho}_{I,g}(t)\\ &\quad + \frac{\pi^2}{32\kappa_2\alpha^2T^2}\mathcal{D}\Big[\hat{Z}_1\hat{Z}_1\Big(\frac{2\pi}{T}t\Big)\Big]\hat{\rho}_{I,g}(t),\end{aligned} \tag{D27}$$

where the dissipators in the first line are due to the single photon loss projected to the ground state manifold. By integrating and ignoring higher order terms, we find

$$\begin{aligned}\hat{\rho}_{I,g}(T) \simeq \hat{\rho}_g(0) + \int_0^T dt\Big(&\kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1] + \mathcal{D}[\hat{Z}_1\Big(\frac{\pi}{T}t\Big)\hat{Z}_2]\Big]\\ &+ \frac{\pi^2}{32\kappa_2\alpha^2T^2}\mathcal{D}\Big[\hat{Z}_1\hat{Z}_1\Big(\frac{2\pi}{T}t\Big)\Big]\Big)\hat{\rho}_g(0)\end{aligned} \tag{D28}$$

at the gate time T.

To go back to the original frame (i.e., $\hat{\rho}(T) = e^{-i\hat{H}'T}\hat{\rho}_I(T)e^{i\hat{H}'T}$), note that $e^{-i\hat{H}'T}$ is given by

$$e^{-i\hat{H}'T} = |0\rangle\langle 0|_1 \otimes \hat{I} + |1\rangle\langle 1|_1 \otimes e^{i\pi\hat{a}_2^\dagger\hat{a}_2}e^{-i\pi\alpha^2} \tag{D29}$$

in the hybrid basis. In the shifted Fock basis, $e^{i\pi\hat{a}^\dagger\hat{a}}$ is exactly given by $\hat{X} \otimes \hat{I}$ and thus we have

$$e^{-i\hat{H}'T} = (\hat{Z}_1(-\pi\alpha^2) \cdot \text{CNOT}_{1\to 2}) \otimes \hat{I} \tag{D30}$$

in the full shifted Fock basis. Thus, projecting $e^{-i\hat{H}'T}$ to the ground state manifold of the cat qubits, we find $\hat{\rho}_g(T) = CX'\hat{\rho}_{I,g}(T)CX'^\dagger$ where $CX' \equiv \hat{Z}_1(-\pi\alpha^2) \cdot \text{CNOT}_{1\to 2}$. Therefore, we can understand $\hat{\rho}_g(T)$ as a state that results from applying a unitary operation CX' to the input state $\hat{\rho}_g(0)$ which is then corrupted by an error channel

$$\begin{aligned}\mathcal{N}_{CX'}(\hat{\rho}) \simeq \hat{\rho} + \int_0^T dt\Big(&\kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1] + \mathcal{D}[\hat{Z}_1\hat{Z}_1\Big(\frac{\pi}{T}t\Big)\hat{Z}_2]\Big]\\ &+ \frac{\pi^2}{32\kappa_2\alpha^2T^2}\mathcal{D}\Big[\hat{Z}_1\hat{Z}_1\Big(\frac{2\pi}{T}t\Big)\Big]\Big)\hat{\rho},\end{aligned} \tag{D31}$$

where we used the fact that $\hat{Z}_2$ is transformed via $\text{CNOT}_{1\to 2}$ into $\hat{Z}_1\hat{Z}_2$. Performing the integration explicitly and ignoring off-diagonal terms similarly as in the analysis of the controlled Z rotations, we find that the Z error rates (per gate) of the CX' gate are given by

$$\begin{aligned}\bar{p}_{Z_1} &= \kappa_1\alpha^2T + \frac{\pi^2}{64\kappa_2\alpha^2T},\\ \bar{p}_{Z_2} &= \bar{p}_{Z_1Z_2} = \frac{1}{2}\kappa_1\alpha^2T.\end{aligned} \tag{D32}$$

Hence, the optimal gate time that minimizes the total gate infidelity is given by

$$\bar{T}^\star_{CX'} = \frac{\pi}{8\alpha^2\sqrt{2\kappa_1\kappa_2}}, \tag{D33}$$

and at the optimal gate time, the Z error rates (per gate) of the CX' gate are given by

$$\bar{p}^\star_{Z_1} = 6\bar{p}^\star_{Z_2} = 6\bar{p}^\star_{Z_1Z_2} = \frac{3\pi}{8}\sqrt{\frac{\kappa_1}{2\kappa_2}} = 0.833\sqrt{\frac{\kappa_1}{\kappa_2}}. \tag{D34}$$

Note that the Z errors (per gate) due to the single photon loss only account for half the total CX' gate error rate at the optimal gate time. The remaining half comes from the Z error due to the undesired term in the compensating Hamiltonian (see the discussion below Eq. (D21)). Numerically, we find that the optimal Z error rates (per gate) of the CNOT gate are given by (see Table II)

$$p^\star_{Z_1} = 6.067 p^\star_{Z_2} = 6.067 p^\star_{Z_1Z_2} = 0.91\sqrt{\frac{\kappa_1}{\kappa_2}}, \tag{D35}$$

which agree well with the perturbative prediction in Eq. (D34) within a relative error of 10%. Note that the quantitative differences are mostly due to the fact that we neglected the second term in Eq. (D25) to make the analysis simpler and also that we only consider the first excited state manifold in each mode.

We emphasize that to really implement the desired $\text{CNOT}_{1\to 2}$ gate, one should apply a Z rotation $\hat{Z}_1(\pi\alpha^2)$ to the control cat qubit to compensate for the extra Z rotation in the CX' gate and such an extra operation will result in additional Z errors (see Appendix D 2). However, if the average excitation number α^2 is an even integer, the extra Z rotation is not needed and thus the Z error rates of the CNOT gate are simply given by the ones in Eq. (D34).

It is often said that bosonic dephasing $\kappa_\phi\mathcal{D}[\hat{a}^\dagger\hat{a}]$ does not cause any Z errors on cat qubits because it preserves the parity. While this is true for idling, Z and CZ rotations, this is not the case for the CNOT and Toffoli gates. To see why this is the case, note that $\kappa_\phi\mathcal{D}[\hat{a}^\dagger\hat{a}]$ is given by

$$\begin{aligned}\kappa_\phi\mathcal{D}[\hat{a}^\dagger\hat{a}] &= \kappa_\phi\mathcal{D}[\hat{I}\otimes(\hat{b}^\dagger+\alpha)(\hat{b}+\alpha)]\\ &= \kappa_\phi\mathcal{D}[\hat{I}\otimes(\hat{b}^\dagger\hat{b}+\alpha(\hat{b}+\hat{b}^\dagger))]\end{aligned} \tag{D36}$$

in the shifted Fock basis, where we assumed $|\alpha|\gg 1$ and used the fact that $\mathcal{D}[\hat{O}+c\hat{I}] = \mathcal{D}[\hat{O}]$ for all hermitian operators $\hat{O}^\dagger = \hat{O}$ and a scalar c. If the cat qubit is in its ground state manifold, $\hat{b}^\dagger\hat{b}+\alpha\hat{b}$ acts trivially and thus the dominant effect due to the dephasing is the heating caused by the term $\alpha\hat{b}^\dagger$, i.e.,

$$\kappa_\phi\mathcal{D}[\hat{a}^\dagger\hat{a}] \simeq \kappa_\phi\alpha^2[\hat{I}\otimes\hat{b}^\dagger]. \tag{D37}$$

Such heating, however, does not induce any Z errors on the qubit space, as indicated by the identity operator in the first slot of the tensor product; this is consistent with the fact that the bosonic dephasing alone cannot change the excitation number parity.

In the case of the CNOT gate, dephasing in each mode independently causes heating, resulting in direct population transfer from the ground state manifold associated with $|00\rangle'$ to the excited states manifolds with $|10\rangle'$ and $|01\rangle'$. As shown in the first line of Eq. (D26), the excited states $|10\rangle'$ and $|01\rangle'$ decay back to the code space via the engineered dissipation $4\kappa_2\alpha^2\mathcal{D}[\hat{I}_{1,2}\otimes\hat{b}_1]$ and $4\kappa_2\alpha^2\mathcal{D}[\hat{Z}_1(\frac{2\pi}{T}t)\otimes\hat{b}_2]$, respectively. While the former engineered dissipation (corresponding to the control mode) is parity preserving, the latter (corresponding to the target mode) induces a Z rotation of the control mode, i.e., $\hat{Z}_1(\frac{2\pi}{T}t)$. This is because the engineered jump operator on the target mode $\hat{L}_2(t)$ rotates conditioned on the state of the control mode. Consequently, while the process $|10\rangle'\to|00\rangle'$ is parity preserving in overall, the other process $|01\rangle'\to|00\rangle'$ induces Z errors on the qubit degree of freedom. More explicitly, the heating followed by the fast relaxation in the target mode induces a new noise process

$$\kappa_\phi\alpha^2\mathcal{D}\Big[\hat{Z}_1\Big(\frac{2\pi}{T}t\Big)\Big]\hat{\rho}_{I,g}(t) \tag{D38}$$

in addition to the noise processes described in the right hand side of Eq. (D27). Integrating over the time window $t\in[0,T]$ and ignoring off-diagonal terms, such a noise process adds an error rate (per gate) $\kappa_\phi\alpha^2T/2$ to p_{Z_1}, i.e.,

$$\begin{aligned}\bar{p}_{Z_1} &= \kappa_1\alpha^2T+\frac{1}{2}\kappa_\phi\alpha^2T+\frac{\pi^2}{64\kappa_2\alpha^2T},\\ \bar{p}_{Z_2} &= \bar{p}_{Z_1Z_2} = \frac{1}{2}\kappa_1\alpha^2T.\end{aligned} \tag{D39}$$

That is, even in the lossless case (i.e., $\kappa_1=0$), the CNOT gate is not free from Z errors and is instead limited by $\bar{p}^\star_{Z_1}\propto\sqrt{\kappa_\phi/\kappa_2}$ at the optimal gate time. In contrast, dephasing does not induce any additional Z errors in the case of idling, Z rotations, and CZ rotations because in these cases the engineered dissipation is always static (i.e., $\kappa_2\mathcal{D}[\hat{a}^2-\alpha^2]$ in the usual Fock basis or approximately $4\kappa_2\alpha^2\mathcal{D}[\hat{I}\otimes\hat{b}]$ in the shifted Fock basis) and thus preserves the parity when it brings the excited states back to the cat code manifold. We also remark that in the presence of non-zero thermal population n_{th}, we simply need to replace κ_1 by $\kappa_1(1+2n_{\text{th}})$.

We reinforce that the above perturbative approach based on an approximate expression $\hat{a}\simeq\hat{Z}\otimes(\hat{b}+\alpha)$ is not capable of capturing non-Z-type errors which decrease exponentially in $|\alpha|^2$. Numerically, however, we simulate the master equation in the shifted Fock basis without making any approximations to capture the exponentially small error rates and get accurate Z error rates. In particular, we use an exact expression of the annihilation operator in the shifted Fock basis (obtained via the procedure described in Appendix C) and perform the frame transformations similarly as in this section (i.e., hybrid basis, rotating frame, and then full shifted Fock basis) but in a way that takes into account exponentially small corrections in $|\alpha|^2$. See Appendix E for numerical results.

5. Toffoli

A Toffoli gate among three cat qubits can be implemented by

$$\begin{aligned}\frac{d\hat{\rho}(t)}{dt} &= \kappa_2\Big[\mathcal{D}[\hat{a}_1^2-\alpha^2]+\mathcal{D}[\hat{a}_2^2-\alpha^2]+\mathcal{D}[\hat{L}_3(t)]\Big]\hat{\rho}(t)\\ &\quad+\kappa_1\Big[\mathcal{D}[\hat{a}_1]+\mathcal{D}[\hat{a}_2]+\mathcal{D}[\hat{a}_3]\Big]\hat{\rho}(t)-i[\hat{H},\hat{\rho}(t)]\end{aligned} \tag{D40}$$

where the engineered dissipation $\hat{L}_3(t)$ and the compensating Hamiltonian $\hat{H}$ are given by

$$\begin{aligned}\hat{L}_3(t) &= \hat{a}_3^2-\alpha^2-\frac{1}{4}(e^{2i\frac{\pi}{T}t}-1)(\hat{a}_1-\alpha)(\hat{a}_2-\alpha),\\ \hat{H} &= -\frac{\pi}{8\alpha^2 T}((\hat{a}_1-\alpha)(\hat{a}_2^\dagger-\alpha)+\text{h.c.})(\hat{a}_3^\dagger\hat{a}_3-\alpha^2).\end{aligned} \tag{D41}$$

Similarly as in the case of the CNOT gate, the time-dependent engineered jump operator $\hat{L}_3(t)$ stabilizes the target mode $\hat{a}_3$ in the $|\pm\alpha e^{i\frac{\pi}{T}t}\rangle$ manifold if the control modes $\hat{a}_1$ and $\hat{a}_2$ are in the "trigger" state $|11\rangle\simeq|{-\alpha},-\alpha\rangle$ or in the usual cat code manifold $|\pm\alpha\rangle$ otherwise. Hence, the target mode is rotated by 180° (i.e., X gate on the cat qubit) at the gate time $t=T$ only if the control qubits are in the trigger state $|11\rangle$, realizing the controlled-controlled-X gate, or the Toffoli gate on the three cat qubits. Moreover, the compensating Hamiltonian $\hat{H}$ mitigates the adverse effects due to the non-adiabaticity by actively bringing the target mode in the desired manifold $|\pm\alpha e^{i\frac{\pi}{T}t}\rangle$ when the control qubits are in the trigger state $|11\rangle\simeq|-\alpha,-\alpha\rangle$.

To analyze the Z error rates of the Toffoli gate perturbatively, we first use to the hybrid basis system where the control modes are described by the shifted Fock basis and the target mode is described by the usual Fock basis. In the hybrid basis, the compensating Hamiltonian is given by

$$\begin{aligned}\hat{H} &= -\frac{\pi}{T}|11\rangle\langle 11|_{1,2}\otimes(\hat{a}_3^\dagger\hat{a}_3-\alpha^2)\\ &\quad-\frac{\pi}{8\alpha T}(\hat{Z}_1-\hat{Z}_1\hat{Z}_2)\otimes(\hat{b}_1+\hat{b}_1^\dagger)(\hat{a}_3^\dagger\hat{a}_3-\alpha^2)\\ &\quad-\frac{\pi}{8\alpha T}(\hat{Z}_2-\hat{Z}_1\hat{Z}_2)\otimes(\hat{b}_2+\hat{b}_2^\dagger)(\hat{a}_3^\dagger\hat{a}_3-\alpha^2)\\ &\quad+\frac{\pi}{8\alpha^2 T}\hat{Z}_1\hat{Z}_2\otimes(\hat{b}_1\hat{b}_2^\dagger+\hat{b}_1^\dagger\hat{b}_2)(\hat{a}_3^\dagger\hat{a}_3-\alpha^2),\end{aligned} \tag{D42}$$

where we used $\hat{a}_k\simeq\hat{Z}_k\otimes(\hat{b}_k+\alpha)$ for $k\in\{1,2\}$. Note that the first term is a desired term that rotates the target mode by 180° over the gate time T only if the two control qubits are in the trigger state. The fourth term acts trivially if the system is in the ground state manifold. The second and the third terms, on the other hand, make the system excited and leak out of the ground state manifold.

Similarly as in the case of the CNOT gate, we go to a rotating frame with respect to the desired compensating Hamiltonian

$$\hat{H}'\equiv-\frac{\pi}{T}|11\rangle\langle 11|_{1,2}\otimes(\hat{a}_3^\dagger\hat{a}_3-\alpha^2), \tag{D43}$$

i.e., $\hat{\rho}_I(t)\equiv e^{i\hat{H}'t}\hat{\rho}(t)e^{-i\hat{H}'t}$. In this frame, the annihilation operators of the control modes $\hat{a}_1$ and $\hat{a}_2$ are unchanged but the annihilation operator of the target mode $\hat{a}_3$ is transformed as

$$\hat{a}_3\to e^{i\hat{H}'t}\hat{a}_3e^{-i\hat{H}'t}=CZ_{1,2}\Big(\frac{\pi}{T}t\Big)\otimes\hat{a}_3, \tag{D44}$$

where $CZ_{1,2}(\theta)\equiv\exp[i\theta|11\rangle\langle 11|_{1,2}]$. Lastly, by using the shifted Fock basis for the target mode as well (i.e., $\hat{a}_3\simeq\hat{Z}_3\otimes(\hat{b}_3+\alpha)$), we find the following equation of motion for $\hat{\rho}_I(t)$:

$$\begin{aligned}\frac{d\hat{\rho}_I(t)}{dt} &= \kappa_2\Big[\mathcal{D}[\hat{I}_{1,2,3}\otimes(\hat{b}_1^2+2\alpha\hat{b}_1)]\\ &\qquad+\mathcal{D}[\hat{I}_{1,2,3}\otimes(\hat{b}_2^2+2\alpha\hat{b}_2)]+\mathcal{D}[\hat{L}_3'(t)]\Big]\hat{\rho}_I(t)\\ &\quad+\kappa_1\Big[\mathcal{D}[\hat{Z}_1\otimes(\hat{b}_1+\alpha)]+\mathcal{D}[\hat{Z}_2\otimes(\hat{b}_2+\alpha)]\\ &\qquad+\mathcal{D}[CZ_{1,2}\Big(\frac{\pi}{T}t\Big)\hat{Z}_3\otimes(\hat{b}_3+\alpha)]\Big]\hat{\rho}_I(t)\\ &\quad-i[\hat{H}-\hat{H}',\hat{\rho}_I(t)].\end{aligned} \tag{D45}$$

Here, $\hat{L}_3'(t)\equiv e^{i\hat{H}'t}\hat{L}_3(t)e^{-i\hat{H}'t}$ is given by

$$\begin{aligned}\hat{L}_3'(t) &= CZ_{1,2}\Big(\frac{2\pi}{T}t\Big)\otimes(\hat{b}_3^2+2\alpha\hat{b}_3)\\ &\quad-\frac{1}{4}(e^{2i\frac{\pi}{T}t}-1)\Big[\hat{Z}_1\hat{Z}_2\otimes\hat{b}_1\hat{b}_2-\alpha(\hat{Z}_1-\hat{Z}_1\hat{Z}_2)\otimes\hat{b}_1\\ &\qquad-\alpha(\hat{Z}_2-\hat{Z}_1\hat{Z}_2)\otimes\hat{b}_2\Big].\end{aligned} \tag{D46}$$

We neglect all the other terms than the first term in the right hand side because they are much smaller than the first term. Also, we only consider the first excited states and set $\hat{b}_1^2=\hat{b}_2^2=\hat{b}_3^2=0$.

In the full shifted Fock basis, $\hat{H}-\hat{H}'$ is given by

$$\begin{aligned}&\hat{H}-\hat{H}'\\ &=-\frac{\pi}{8\alpha T}(\hat{Z}_1-\hat{Z}_1\hat{Z}_2)\otimes(\hat{b}_1+\hat{b}_1^\dagger)(\hat{b}_3^\dagger\hat{b}_3+\alpha(\hat{b}_3+\hat{b}_3^\dagger))\\ &\quad-\frac{\pi}{8\alpha T}(\hat{Z}_2-\hat{Z}_1\hat{Z}_2)\otimes(\hat{b}_2+\hat{b}_2^\dagger)(\hat{b}_3^\dagger\hat{b}_3+\alpha(\hat{b}_3+\hat{b}_3^\dagger))\\ &\quad+\frac{\pi}{8\alpha^2 T}\hat{Z}_1\hat{Z}_2\otimes(\hat{b}_1\hat{b}_2^\dagger+\hat{b}_1^\dagger\hat{b}_2)(\hat{b}_3^\dagger\hat{b}_3+\alpha(\hat{b}_3+\hat{b}_3^\dagger)).\end{aligned} \tag{D47}$$

As explained above, the third term acts trivially on the code space and thus we focus on the first two terms. In particular, we only consider the dominant driving effects due to the first two terms and approximate $\hat{H} - \hat{H}'$ as

$$\hat{H} - \hat{H}' \simeq -\frac{\pi}{8T}(\hat{Z}_1 - \hat{Z}_1\hat{Z}_2) \otimes (\hat{b}_1\hat{b}_3 + \hat{b}_1^\dagger\hat{b}_3^\dagger) - \frac{\pi}{8T}(\hat{Z}_2 - \hat{Z}_1\hat{Z}_2) \otimes (\hat{b}_2\hat{b}_3 + \hat{b}_2^\dagger\hat{b}_3^\dagger). \quad \text{(D48)}$$

Putting everything together, we find the following master equation:

$$\begin{aligned}\frac{d\hat{\rho}_I(t)}{dt} &= 4\kappa_2\alpha^2\Big[\mathcal{D}[\hat{I}_{1,2,3} \otimes \hat{b}_1] + \mathcal{D}[\hat{I}_{1,2,3} \otimes \hat{b}_2] \\ &\quad + \mathcal{D}[CZ_{1,2}\Big(\frac{2\pi}{T}t\Big) \otimes \hat{b}_3]\Big]\hat{\rho}_I(t) \\ &\quad + \kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1 \otimes \hat{I}] + \mathcal{D}[\hat{Z}_2 \otimes \hat{I}] \\ &\quad + \mathcal{D}[CZ_{1,2}\Big(\frac{\pi}{T}t\Big)\hat{Z}_3 \otimes \hat{I}]\Big]\hat{\rho}_I(t) \\ &\quad + i\Big[\frac{\pi}{8T}(\hat{Z}_1 - \hat{Z}_1\hat{Z}_2) \otimes (\hat{b}_1\hat{b}_3 + \hat{b}_1^\dagger\hat{b}_3^\dagger) \\ &\quad + \frac{\pi}{8T}(\hat{Z}_2 - \hat{Z}_1\hat{Z}_2) \otimes (\hat{b}_2\hat{b}_3 + \hat{b}_2^\dagger\hat{b}_3^\dagger), \hat{\rho}_I(t)\Big]. \quad \text{(D49)}\end{aligned}$$

The undesired terms of the compensating Hamiltonian in Eq. (D48) make the system excited to the manifold associated with $|101\rangle'$ ($|011\rangle'$) via $\hat{b}_1^\dagger\hat{b}_3^\dagger$ ($\hat{b}_2^\dagger\hat{b}_3^\dagger$) and at the same time cause an error $\hat{Z}_1 - \hat{Z}_1\hat{Z}_2$ ($\hat{Z}_2 - \hat{Z}_1\hat{Z}_2$) on the qubit space at a rate (per time) $g = \pi/(8T)$. These excited states decay back to the code space via the engineered dissipation. For instance, $|101\rangle'$ decays back to the code space through either $|101\rangle' \to |001\rangle' \to |000\rangle'$ or $|101\rangle' \to |100\rangle' \to |000\rangle'$ with a total decay rate (per time) $\kappa = 8\kappa_2\alpha^2$. In both decay routes, the annihilation of the excitation in the target mode (i.e., $\hat{b}_3$) is accompanied by an additional error $CZ_{1,2}(\frac{2\pi}{T}t)$ on the control qubits. The same is true for the other excited state $|011\rangle'$ which decays back to the code space either via $|011\rangle' \to |001\rangle' \to |000\rangle'$ or $|011\rangle' \to |010\rangle' \to |000\rangle'$. Consequently, by using adiabatic elimination, we find that these driven-dissipative processes induce two independent decay processes with jump operators $(\hat{Z}_1 - \hat{Z}_1\hat{Z}_2)CZ_{1,2}(\frac{2\pi}{T}t)$ and $(\hat{Z}_2 - \hat{Z}_1\hat{Z}_2)CZ_{1,2}(\frac{2\pi}{T}t)$ with an effective decay rate (per time) $4g^2/\kappa = \pi^2/(128\kappa_2\alpha^2T^2)$. Hence, the effective master equation in the ground state manifold is given by

$$\begin{aligned}&\frac{d\hat{\rho}_{I,g}(t)}{dt} \\ &= \kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1] + \mathcal{D}[\hat{Z}_2] + \mathcal{D}\Big[CZ_{1,2}\Big(\frac{\pi}{T}t\Big)\hat{Z}_3\Big]\Big]\hat{\rho}_{I,g}(t) \\ &\quad + \frac{\pi^2}{128\kappa_2\alpha^2T^2}\Big[\mathcal{D}\Big[(\hat{Z}_1 - \hat{Z}_1\hat{Z}_2)CZ_{1,2}\Big(\frac{2\pi}{T}t\Big)\Big] \\ &\quad + \mathcal{D}\Big[(\hat{Z}_2 - \hat{Z}_1\hat{Z}_2)CZ_{1,2}\Big(\frac{2\pi}{T}t\Big)\Big]\Big]\hat{\rho}_{I,g}(t), \quad \text{(D50)}\end{aligned}$$

where $\hat{\rho}_{I,g}(t) \equiv \langle 000|'\hat{\rho}_I|000\rangle'$ is the projected density matrix (of size $2^3 \times 2^3$) to the ground state manifold of the three cat qubits.

To go back to the original frame (i.e., $\hat{\rho}(T) = e^{-i\hat{H}'T}\hat{\rho}_I(T)e^{i\hat{H}'T}$), note that $e^{-i\hat{H}'T}$ is given by

$$\begin{aligned}e^{-i\hat{H}'T} &= (\hat{I}_{1,2,3} - |11\rangle\langle 11|_{1,2}) \otimes \hat{I} \\ &\quad + |11\rangle\langle 11|_{1,2} \otimes e^{i\pi(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2)} \quad \text{(D51)}\end{aligned}$$

in the hybrid basis, and since $e^{i\pi\hat{a}^\dagger\hat{a}}$ is given by $\hat{X} \otimes \hat{I}$ in the shifted Fock basis, we have

$$e^{-i\hat{H}'T} = (CZ_{1,2}(-\pi\alpha^2) \cdot \mathrm{TOF}_{1,2\to 3}) \otimes \hat{I} \quad \text{(D52)}$$

in the full shifted Fock basis, where $\mathrm{TOF}_{1,2\to 3}$ is the desired Toffoli gate. Thus, projecting $e^{-i\hat{H}'T}$ to the ground state manifold, we find $\hat{\rho}_g(T) = CCX'\hat{\rho}_{I,g}(T)CCX'^\dagger$ where $CCX' \equiv CZ_{1,2}(-\pi\alpha^2) \cdot \mathrm{TOF}_{1,2\to 3}$. Therefore, we can understand $\hat{\rho}_g(T)$ as a state that results from applying a unitary operation CCX' to the input state $\hat{\rho}_g(0)$ which is then corrupted by an error channel

$$\begin{aligned}\mathcal{N}_{CCX'}(\hat{\rho}) &\simeq \hat{\rho} + \int_0^T dt\Big(\kappa_1\alpha^2\Big[\mathcal{D}[\hat{Z}_1] \\ &\quad + \mathcal{D}[\hat{Z}_2] + \mathcal{D}\Big[CZ_{1,2}\Big(\frac{\pi}{T}(t+T)\Big)\hat{Z}_3\Big] \\ &\quad + \frac{\pi^2}{128\kappa_2\alpha^2T^2}\Big[\mathcal{D}\Big[(\hat{Z}_1 - \hat{Z}_1\hat{Z}_2)CZ_{1,2}\Big(\frac{2\pi}{T}t\Big)\Big] \\ &\quad + \mathcal{D}\Big[(\hat{Z}_2 - \hat{Z}_1\hat{Z}_2)CZ_{1,2}\Big(\frac{2\pi}{T}t\Big)\Big]\Big]\Big)\hat{\rho}. \quad \text{(D53)}\end{aligned}$$

Here, we used the fact that $\hat{Z}_3$ is transformed via $\mathrm{TOF}_{1,2\to 3}$ into $CZ_{1,2}\hat{Z}_3$. Evaluating the integral explicitly and ignoring off-diagonal Pauli errors, we find the following Z error rates (per gate) of the CCX' gate:

$$\begin{aligned}&\bar{p}_{Z_1} = \bar{p}_{Z_2} = \kappa_1\alpha^2T + \frac{\pi^2}{128\kappa_2\alpha^2T}, \\ &\bar{p}_{Z_3} = \frac{5}{8}\kappa_1\alpha^2T, \quad \bar{p}_{Z_1Z_2} = \frac{\pi^2}{128\kappa_2\alpha^2T}, \\ &\bar{p}_{Z_1Z_3} = \bar{p}_{Z_2Z_3} = \bar{p}_{Z_1Z_2Z_3} = \frac{1}{8}\kappa_1\alpha^2T. \quad \text{(D54)}\end{aligned}$$

Hence, the optimal gate time that minimizes the total gate infidelity is given by

$$\bar{T}^\star_{CCX'} = \frac{\pi}{8\alpha^2\sqrt{2\kappa_1\kappa_2}}, \quad \text{(D55)}$$

and at the optimal gate time, the Z error rates (per gate) of the CCX' gate are given by

$$\begin{aligned}\bar{p}^\star_{Z_1} &= \bar{p}^\star_{Z_2} = 3.2\bar{p}^\star_{Z_3} = 2\bar{p}^\star_{Z_1Z_2} \\ &= 16\bar{p}_{Z_1Z_3} = 16\bar{p}_{Z_2Z_3} = 16\bar{p}_{Z_1Z_2Z_3} \\ &= \frac{\pi}{4}\sqrt{\frac{\kappa_1}{2\kappa_2}} = 0.555\sqrt{\frac{\kappa_1}{\kappa_2}}. \quad \text{(D56)}\end{aligned}$$

Numerically, we find that the optimal Z error rates (per gate) are given by (see Table II)

$$p^\star_{Z_1} = p^\star_{Z_2} = 3.05 p^\star_{Z_3} = 1.81 p^\star_{Z_1 Z_2}$$
$$= 14.9 p_{Z_1 Z_3} = 14.9 p_{Z_2 Z_3} = 14.9 p_{Z_1 Z_2 Z_3} = 0.58\sqrt{\frac{\kappa_1}{\kappa_2}}, \tag{D57}$$

which agree well with the perturbative prediction in Eq. (D56).

Similarly as in the case of the CNOT gate, we remark that the implemented gate CCX' differs from the desired Toffoli gate $\mathrm{TOF}_{1,2\to3}$ by a CZ rotation $CZ_{1,2}(-\pi\alpha^2)$. Thus, unless the average excitation number α^2 is given by an even integer, one should apply $CZ_{1,2}(\pi\alpha^2)$ to compensate for the extra phase shift. Lastly, note that dephasing can induce direct heating in each mode with a heating rate (per time) $\kappa_\phi\alpha^2$ (see Eq. (D37)). The excited states due to the heating decay back to the code space via the engineered dissipation. The engineered jump operators in the control modes are static and thus the excitations in the control modes decay back to the code space in a parity-preserving way, i.e., $4\kappa_2\alpha^2\mathcal{D}[\hat{I}_{1,2,3}\otimes\hat{b}_1]$ and $4\kappa_2\alpha^2\mathcal{D}[\hat{I}_{1,2,3}\otimes\hat{b}_2]$. On the other hand, the engineered jump operator on the target mode is time-dependent and thus the the relaxation of the excitation in the target mode is accompanied by a CZ rotation in the control qubits, i.e., $4\kappa_2\alpha^2\mathcal{D}[CZ_{1,2}(\frac{2\pi}{T}t)\otimes\hat{b}_1]$. Consequently, such a heating-relaxation process in the target mode generates a new noise process

$$\kappa_\phi\alpha^2\mathcal{D}\Big[CZ_{1,2}\Big(\frac{2\pi}{T}t\Big)\Big]\hat{\rho}_{I,g}(t) \tag{D58}$$

in addition to the noise processes described in the right hand side of Eq. (D50) and adds $\kappa_\phi\alpha^2T/8$ to p_{Z_1}, p_{Z_2}, and $p_{Z_1Z_2}$, i.e.,

$$\bar{p}_{Z_1} = \bar{p}_{Z_2} = \kappa_1\alpha^2T + \frac{1}{8}\kappa_\phi\alpha^2T + \frac{\pi^2}{128\kappa_2\alpha^2T},$$
$$\bar{p}_{Z_3} = \frac{5}{8}\kappa_1\alpha^2T, \quad \bar{p}_{Z_1Z_2} = \frac{1}{8}\kappa_\phi\alpha^2T + \frac{\pi^2}{128\kappa_2\alpha^2T},$$
$$\bar{p}_{Z_1Z_3} = \bar{p}_{Z_2Z_3} = \bar{p}_{Z_1Z_2Z_3} = \frac{1}{8}\kappa_1\alpha^2T. \tag{D59}$$

Hence, even in the lossless case (i.e., $\kappa_1 = 0$), the Toffoli gate has non-zero Z error rates which scale as $\bar{p}^*_{Z_1} = \bar{p}^*_{Z_2} = \bar{p}^*_{Z_1Z_2} \propto \sqrt{\kappa_\phi/\kappa_2}$ at the optimal gate time. Lastly, in the presence of non-zero thermal population n_{th}, we simply need to replace κ_1 by $\kappa_1(1+2n_{\mathrm{th}})$.

Appendix E: Simulations of gate error rates

1. CNOT

We simulated the CNOT gate as described in Appendix D 4 using the shifted Fock basis approach on AWS EC2 instances. Our code is written in Python using the QuTiP package. The results presented here took approximately 150 hours to run on an AWS EC2 C5.18xlarge instance with 72 virtual CPUs. To compute the Pauli error rates for the CNOT gate, we use two types of simulation. One set of simulations is aimed at the Z-type Pauli error rates and also determined the optimal gate time. These simulations require only a small dimension in the shifted Fock basis. The second type of simulation uses a much larger Hilbert space dimension to perform full tomography of the CNOT gate at the optimal gate time for relatively small values of the cat-code size α.

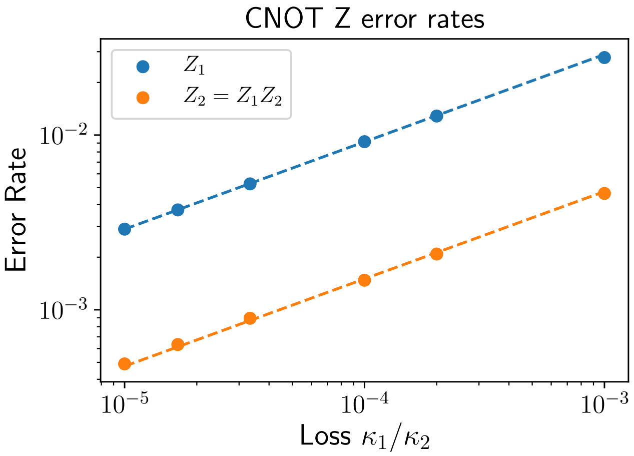


FIG. 28. Log-log plot of Pauli Z-type error rates for the CNOT gate at optimal gate time with mean phonon number $n = 8$ in the presence of pure loss at rate κ_1. The fits are performed over the range κ_1/κ_2 from 10^{-4} to 10^{-5}. The error rates Z_2 and Z_1Z_2 differ by no more than 10^{-5}.

We consider four noise models: first pure phonon loss at a number of different rates. We are most interested in the range of loss (κ_1/κ_2) from 10^{-4} to 10^{-5}. Next, we consider phonon loss at rate κ_1, phonon gain at a rate such that the thermal occupation is given by $n_{th} = 1/100$, and dephasing noise at three different rates $\kappa_\phi = 1$, 2.5, and 10 times κ_1. This value of the thermal occupation number is larger than what we expect in acoustic cavities. We chose $n_{th} = 1/100$ so that we could resolve the contribution of phonon gain on the gate error rates. With $n_{th} = 1/100$ the gate error rates are enhanced by a factor of about 1.01 relative to the error rates with no phonon gain. Dephasing noise is more significant; it increases the dominant error rate, Z error on the control qubit and decreases the optimal gate time.

The Z error rates for the CNOT gate are well-captured by the shifted Fock basis with small dimension, indicating that the Z error rates are dominated by dephasing resulting from the excitation of the cat qubit to the lowest energy excited states. The results plotted in Figs. 28 to 30 were obtained with $d = 7$, or a total Hilbert space dimension of 14. The simulations converge rapidly as the dimension increases. The relative difference in the error

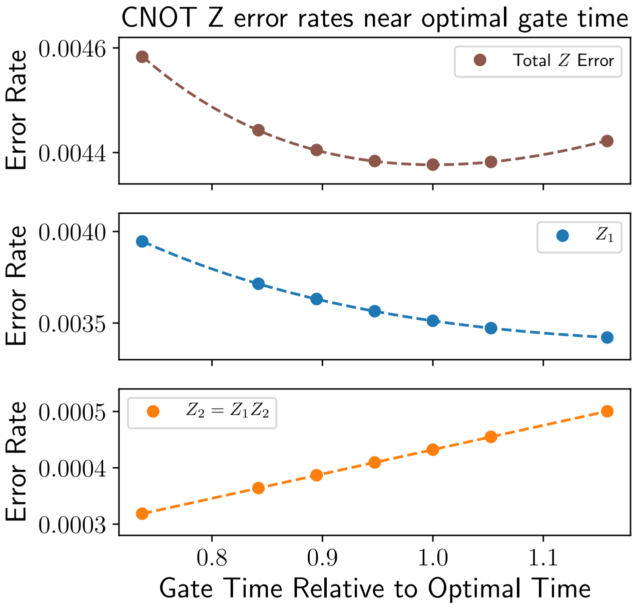


FIG. 29. Plot of Pauli Z-type error rates for the CNOT gate with mean phonon number $n = 10$ at various values of the gate time. The noise model is at rate $\kappa_1 = 10^{-5}\kappa_2$, dephasing at a rate $\kappa_\phi = \kappa_1$, and gain with $n_{th} = 1/100$. The gate time is plotted relative to the optimal gate time for these parameters. The optimal gate time minimizes the total error. The dotted curves are a linear fit for the Z_2 error rate and a sum of a linear term and a $1/T$ term for the Z_1 and total error rates, representing the contributions of loss and non-adiabatic errors.

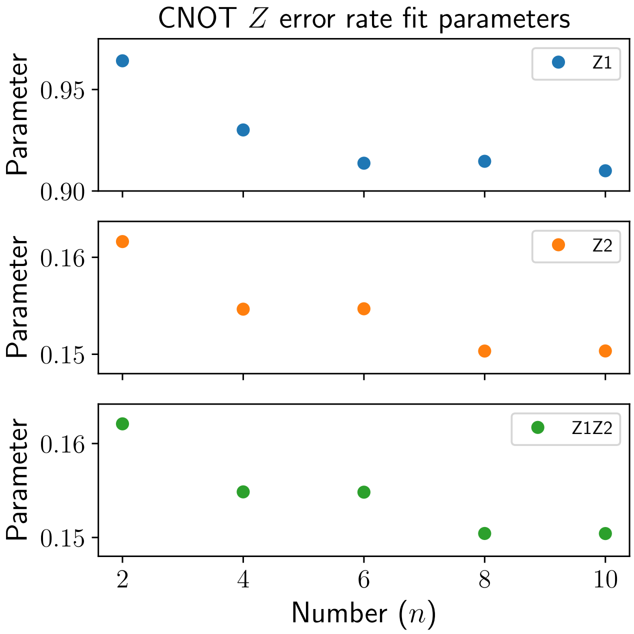


FIG. 30. Plot of the fit parameters of the square root fit as shown in figure 28 for the Pauli Z-type error rates of the CNOT gate for different values of mean phonon number n. The noise model in this figure is pure phonon loss.

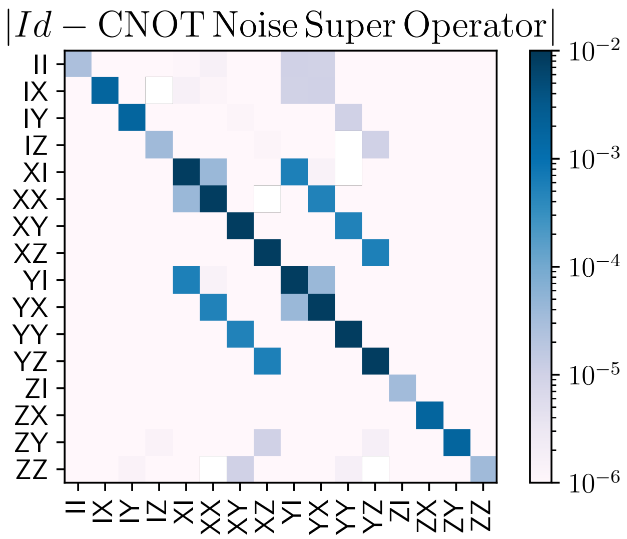


FIG. 31. Plot of identity minus the super operator for the CNOT noise channel in the Pauli basis. The CNOT parameters are $n = 4$, $\kappa_1/\kappa_2 = 10^{-5}$, $n_{th} = 1/100$, and $\kappa_\phi = \kappa_1$. The diagonal components of the matrix are the Pauli infidelities, in other words, one minus the probabilities that the CNOT noise channel maps a given Pauli operator back to itself. The off-diagonal components represent the coherent part of the noise channel. These terms are orders of magnitude smaller than the dominant noise terms. The dominant Z_1 error rate manifests itself as the relatively larger diagonal terms that are sensitive to a Z_1 error, i.e. Pauli operators with X or Y on the first qubit.

rates shown in Fig. 28 between the simulations with $d = 6$ and $d = 7$ is about 10^{-6}, and this gives a bound on how closely these simulations reflect the true error rates in an infinite-dimensional cavity. We call the control cavity 1 and the target 2. As described in Appendix D 4, the non-adiabatic error contribution to the Z_1 error rate of the CNOT gate scales with $1/T$, where T is the gate time, while the error due to single-phonon loss scales with T. As a result of the tradeoff between non-adiabatic error, the optimal gate time scales like $1/\sqrt{T}$. As shown in Fig. 29, around the optimal gate time the Z_1 error rate is decreasing with T, whereas the Z_2 and Z_1Z_2 error rates are increasing. This is because the non-adiabatic errors affect only the control cavity, i.e. Z_1. We find an optimal gate time that differs only slightly from the prediction in Appendix D 4. In Fig. 28 we find the expected square root scaling of the Z error rates with loss rate over a wide range of loss rates. We do observe that the points corresponding to larger values of loss near $\frac{\kappa_1}{\kappa_2} = 10^{-3}$ tend to lie below the square root best-fit curve. For this reason, we perform our fits over the range of loss from 10^{-4} to 10^{-5}, which is our range of interest for our error correction simulations. This leads to slightly larger error rate fit parameters than if we fit over the full range of loss. Fig. 30 shows the dependence of the Z error rate coefficients on the mean phonon number of the cat $n = \alpha^2$.

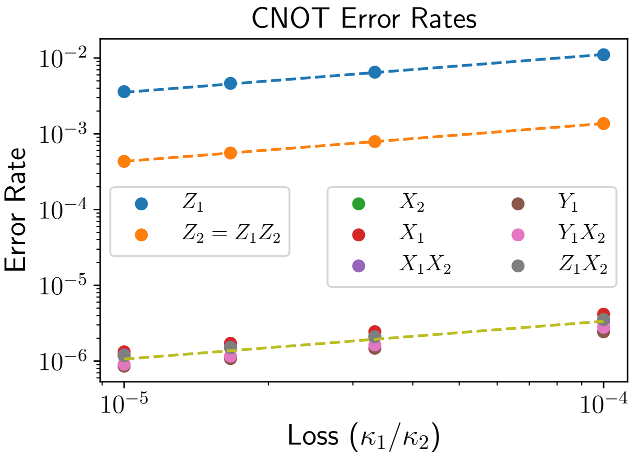


FIG. 32. Log-log plot of the Pauli error rates for the CNOT gate with parameters, $n = 4$, $\kappa_1 = \kappa_\phi$ and $n_{th} = 1/100$. Each of these error rates scale like $\sqrt{\kappa_1/\kappa_2}$.

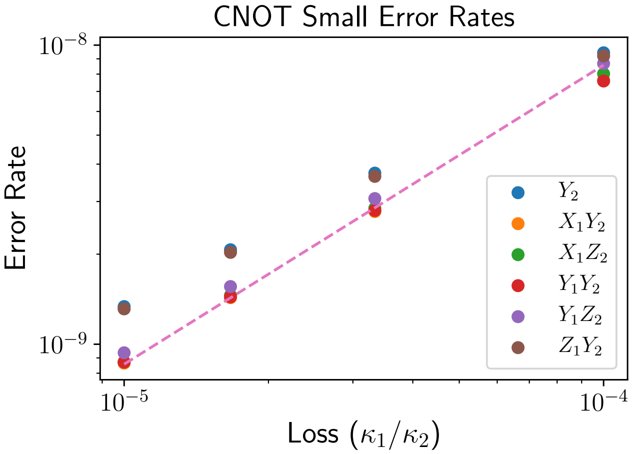


FIG. 33. Log-log plot of the smallest error rates for the CNOT gate with parameters, $n = 4$, $\kappa_1 = \kappa_\phi$ and $n_{th} = 1/100$. These error rates are proportional to κ_1/κ_2 rather than the square root scaling of the other error rates in Fig. 32.

These coefficients come from fits of each error rate to $c\sqrt{\kappa_1/\kappa_2}$ for each value of n. There is variation over the range $n = 2, \ldots 10$, but for $n = 8$ and $n = 10$ the variation is quite small. The values quoted in Table II represent this large-n value.

Once the optimal gate time is found using the Z error rate simulation, we performed tomography for the CNOT gate at several values of loss, dephasing, and n to compute the full noise channel. The noise channel for $n = 4$, $n_{th} = 1/100$, and $\kappa_\phi = \kappa_1 = 10^{-5}$ is illustrated in Fig. 31. The noise channel is largely incoherent with small off-diagonal elements. The diamond distance from identity is equal to about 2.5 times the average infidelity of the channel across all values of α, loss, and dephasing that we simulated. The Hilbert-Schmidt norm of the off-diagonal elements of the super operator in the Pauli basis is 10^{-2}–10^{-3} times the norm of the diagonal elements. Neglecting the off-diagonal components, we are able to read off the full set of 15 two-qubit Pauli error rates. For the values of $n = \alpha^2$ that are not even integers, we must cancel the extra Z_1 rotation by angle $\pi\alpha^2$ that comes with our implementation of the CNOT gate. In practice this would entail additional error, but we do not include the effect of the noisy Z rotation because we are interested in the error intrinsic to the CNOT gate and we expect to operate with even n as much as possible. Besides the dominant Z error rates, each of the other Pauli error rates is exponentially small in α. However, we observe that these exponentially small error rates are divided into two classes—six of them scale like the square root of κ_1/κ_2 just like the Z error rates and the remaining six error rates scale linearly with κ_1/κ_2. The error rates with square root scaling are plotted for one choice of parameters in Fig. 32. The error rates scaling linearly are much smaller and are shown in Fig. 33. A large dimension is required to accurately recover some of the Pauli error rates. As shown in Fig. 34 when $n = 3$ the Pauli errors that involve X or Y acting on the control qubit require a much larger value of d than the other error rates. The difference between the error rates with $d = 8$ and $d = 9$ in this case was as much as 15%. We used a dimension of $d = 9$, 10, and 11 for $n = 3$, 4, and 4.5, respectively, and the total Hilbert space dimension is $2d$ in the shifted Fock basis.

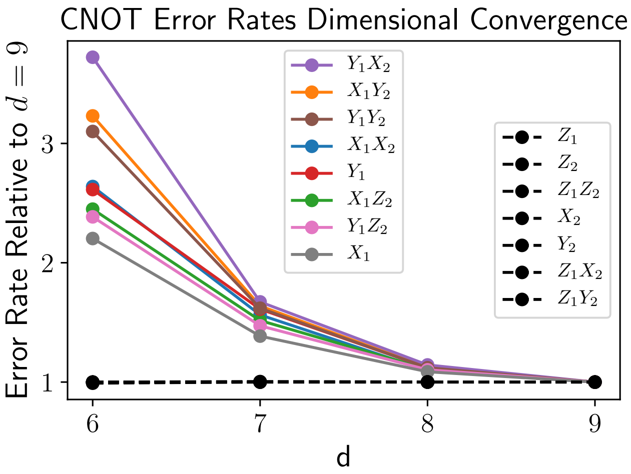


FIG. 34. Plot of the Pauli error rates at a fixed value of the noise parameters and different values of the shifted Fock basis dimension d. Each error rate is scaled by its value at largest value of dimension $d = 9$ to show the convergence as d increases. The parameters are set to $n = 3$, $\kappa_1 = 10^{-5}$, $\kappa_\phi = 0$, and $n_{th} = 0$. One set of Pauli error rates converges rapidly as d increases. This includes all Pauli errors where Z or Id act on the control qubit. Another set of Pauli errors with X or Y acting on the control qubit require much higher Hilbert space dimension to capture accurately. This implies that these error rates include significant contributions from highly excited states.

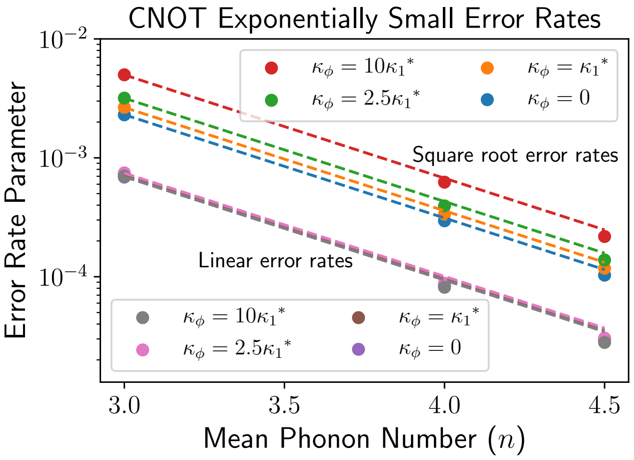


FIG. 35. Log-linear plot showing exponential decay of non-Z error rates as mean phonon number $n = |\alpha|^2$ increases. Four values of dephasing are plotted, $\kappa_\phi = 0$, 1, 2.5, and 10 times κ_1. The asterisk that appears for the non-zero values of dephasing represents that these points include phonon gain at a rate $n_{th} = 1/100$. No gain is present in the $\kappa_\phi = 0$ points. For each set of noise parameters, the upper set of points represents the Pauli error rates from Fig. 32 that scale with $\sqrt{\kappa_1/\kappa_2}$ and exponentially with n. The lower set of points are the Pauli error rates from Fig. 32 that scale linearly with κ_1/κ_2 and exponentially with n. The parameters of the exponential fits can be found in Table II.

We did not go to larger values of n because the required Hilbert space dimension required an unreasonably long time to simulate. Across all values of loss, dephasing, and mean phonon number, the error rates for the largest dimension d that we used and the error rates at $d-1$ differed by several percent. This provides a sense of the difference we expect between the largest dimension we used and the $d \to \infty$ limit. Because of this uncertainty of perhaps several percent in certain of the Pauli error rates and for simplicity, we have chosen to report a single fit for each of the two groups of exponentially small error rates. These include both the small error rates that scale with the square root of loss in Fig. 32 and those that scale linearly in Fig. 33. This is why only a single best fit curve appears over the clusters of small error rates in those plots. We have taken the average within each of the two classes of exponentially small error rates and fit the square root or linear curve to those averages. Correspondingly, in our simulations of error correction we assume that $p_{X_1} = p_{X_2} = p_{X_1X_2} = p_{Y_1} = p_{Y_1X_2} = p_{Z_1X_2}$ and $p_{Y_1} = p_{X_1Y_2} = p_{X_1Z_2} = p_{Y_1Y_2} = p_{Y_1Z_2} = p_{Z_1Y_2}$, and the error probabilities are given by the average fits. Both classes of small error rates exhibit the expected exponential scaling with the mean photon number n of the cat code as shown in Fig. 35.

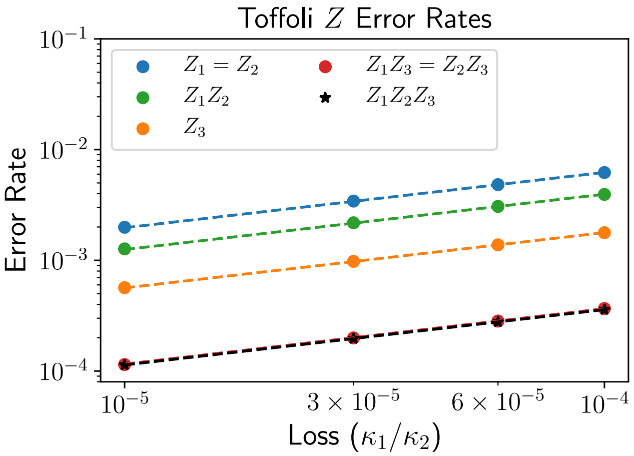


FIG. 36. Log-log plot of the various Z-type error rates for the Toffoli gate at optimal gate time with parameters $n = 8$, $\kappa_\phi = \kappa_1$, and $n_{th} = 1/100$. These error rates were obtained in a shifted Fock basis simulation using $d = 4$ for a total Hilbert space dimension of 8 for each of the three cavities involved in the Toffoli gate. Qubits 1 and 2 are the controls and 3 is the target.

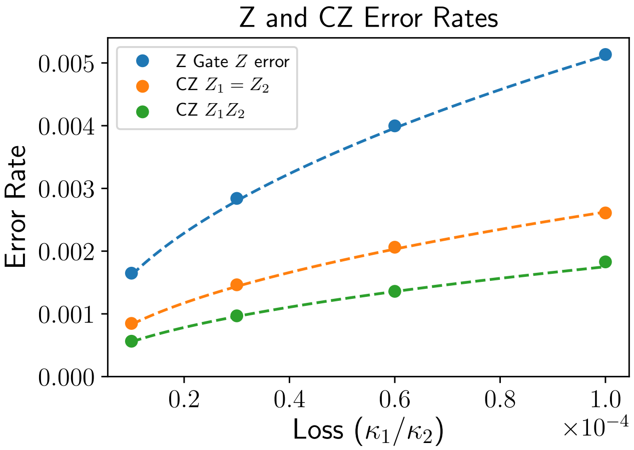


FIG. 37. Plot of the Z error rates for the Z and CZ gates as a function of loss for $n = 10$. The noise model for this plot is pure phonon loss with no gain. Gain will have a small effect of these error rates, while dephasing noise will have only a negligible effect. The dotted curves are best fits in the form $c * \sqrt{\kappa_1/\kappa_2}$.

2. Toffoli

We simulate the Toffoli gate using the shifted Fock basis as we did for the CNOT gate. In this case we solve the master equation for three cavities. This leads to a much larger total Hilbert space dimension, and for this reason we are unable to use the large values of d necessary to resolve all 63 Pauli error probabilities. Instead we focus on the dominant errors, which are the Z-type Pauli errors. These errors do not require a large value of d to calculate

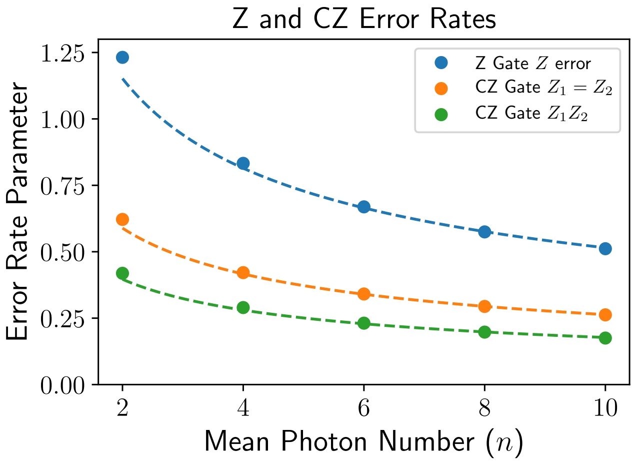


FIG. 38. Plot of the Z error rate parameters from best fits like the ones in Fig. 37 as a function of mean phonon number n. The error rates are the product of these fit parameters and the loss rate $\sqrt{\kappa_1/\kappa_2}$. The dotted best fit curves in this plot are fits to c/α where $\alpha = \sqrt{n}$. For small values of α the scaling differs somewhat from the $1/\alpha$ scaling in the large α limit. For this reason the fits were performed over the range $n = 6$ to 10.

with good precision. We used $d = 4$ for each of the three cavities in these Toffoli simulations, which required a total of about 170 hours running on an AWS EC2 c5.18xlarge instance with 72 virtual CPUs. We simulated the noise channel on a complete set of X eigenstates and averaged over the initial states. We simulated a range of gate times and selected the time that minimizes the total error rate. For loss without gain or dephasing, we found that this gate time matched the optimal gate time for the CNOT gate. With dephasing noise added we found a small difference in optimal gate time. We chose to use the optimal gate times for the CNOT gate throughout. The Toffoli error rates at the true optimal gate time and at the CNOT optimal gate time are shown in Table VII. The difference in the total fidelity of the Toffoli gate is small, however the relative size of individual Pauli Z error rates does differ by several percent when $\kappa_\phi = 10\kappa_1$.

Fig. 36 shows the seven Z-type error probabilities for the Toffoli gate at optimal gate time as a function of the loss rate with $n = 8$, $\kappa_\phi = \kappa_1$ and $n_{th} = 1/100$. We see the expected square root scaling with κ_1/κ_2 for each of the error rates and perform best fits. We simulate Toffoli with $n = 4$, 6, 8, and 10 and for four sets of noise parameters: only phonon loss and then phonon loss, gain and dephasing at three different rates, $\kappa_\phi = 1$, 2.5, and 10 times κ_1. Similar to the CNOT example in Fig. 30, the parameters of the square root fits depend on n but reach a plateau around $n = 8$ or 10. For our error correction simulations we are most interested in values of n in this regime. To produce the numbers in Table VII we have average the values for $n = 8$ and $n = 10$. The relative difference between these two is only order 10^{-2} or less.

3. Z and CZ

To implement our CNOT and Toffoli gates with values of α such that $n = \alpha^2$ is not an even integer, we need to apply an additional Z or CZ rotations on the control cavity or cavities. These rotations can be implemented as described in Section III. The dominant error rates are the Z error rates, and at the optimal gate time the Z error rates scale with $\sqrt{\kappa_1/\kappa_2}$. Unlike the case of CNOT or Toffoli, the error rates decrease with α for the Z and CZ rotations. We simulate the Z error rates for the Z and CZ gates, in other words Z and CZ rotations by angle π. Once again we use the shifted Fock basis as described in Appendix C to simulate the Z error rates using a small Hilbert space dimension. Fig. 37 shows the Z error rates for both the Z and CZ gates when $n = 10$ and the noise model is phonon loss. We fit the error rates to $\sqrt{\kappa_1/\kappa_2}$ for each value of n and for each noise model. Then Fig. 38 shows the scaling of the coefficients of the $\sqrt{\kappa_1/\kappa_2}$ fits as a function of n when the noise model is phonon loss. We fit these curves to $1/\alpha$. The results of the fits that give the Z error rates as functions of α and κ_1/κ_2 are summarized in Table VIII. We also simulated the Z and CZ gates subject to dephasing noise and confirmed that dephasing noise does not contribute significantly to the Z error rates. Including phonon gain with $n_{th} = 1/100$ has a small effect as shown in Table VIII.

Appendix F: Physical implementation of cat qubit gates

Here, we discuss physical realization of the cat qubit gates. Note that engineering static two-phonon dissipations in a multiplexed setting has been extensively discussed in the previous section. Also implementation of the rotating dissipators for the CNOT and Toffoli gates are discussed in detail in Ref. [13]. We thus focus on engineering Hamiltonian interactions needed to implement the cat-qubit gates. In particular, we discuss realization of the linear drive in $\hat{H}_Z$, beam-splitter coupling in $\hat{H}_{CZ}$, selective frequency shift in $\hat{H}_X$, cubic optomechanical coupling in $\hat{H}_{\text{CNOT}}$, and the quartic interaction in $\hat{H}_{\text{TOF}}$ in the stated order.

Recall the Hamiltonian of the system consisting of multiple phononic modes $\hat{a}_k$ coupled to a shared ATS mode $\hat{b}$:

$$\hat{H} = \sum_{k=1}^{N} \omega_k \hat{a}_k^\dagger \hat{a}_k + \omega_b \hat{b}^\dagger \hat{b} - 2E_J \epsilon_p(t) \sin\Big(\sum_{k=1}^{N} \hat{\phi}_k + \hat{\phi}_b \Big). \tag{F1}$$

Here, $\hat{\phi}_k \equiv \varphi_k(\hat{a}_k + \hat{a}_k^\dagger)$ and $\hat{\phi}_b \equiv \varphi_b(\hat{b} + \hat{b}^\dagger)$. Also, φ_k and φ_b quantify zero-point fluctuations of the modes $\hat{a}_k$ and $\hat{b}$. To simplify the discussion, we neglect small frequency shifts due to the pump $\epsilon_p(t)$ for the moment and assume that the frequency of a mode is given by its

Toffoli at optimal gate times	$\kappa_\phi = \kappa_1$ $n_{th} = 1/100$	$\kappa_\phi = 2.5\kappa_1$ $n_{th} = 1/100$	$\kappa_\phi = 10\kappa_1$ $n_{th} = 1/100$	Scaling
Gate Time	$0.28\lvert\alpha\rvert^{-2}$	$0.25\lvert\alpha\rvert^{-2}$	$0.18\lvert\alpha\rvert^{-2}$	$(\kappa_1\kappa_2)^{-\frac{1}{2}}$
$Z_1 = Z_2$	0.62	0.68	0.90	$\sqrt{\kappa_1/\kappa_2}$
Z_3	0.18	0.16	0.12	$\sqrt{\kappa_1\kappa_2}$
Z_1Z_2	0.40	0.48	0.79	$\sqrt{\kappa_1/\kappa_2}$
$Z_1Z_3 = Z_2Z_3$	0.036	0.033	0.024	$\sqrt{\kappa_1/\kappa_2}$
$Z_1Z_2Z_3$	0.035	0.032	0.024	$\sqrt{\kappa_1/\kappa_2}$
Toffoli at CNOT optimal times				
Gate Time	$0.27\lvert\alpha\rvert^{-2}$	$0.24\lvert\alpha\rvert^{-2}$	$0.16\lvert\alpha\rvert^{-2}$	$(\kappa_1\kappa_2)^{-\frac{1}{2}}$
$Z_1 = Z_2$	0.62	0.68	0.91	$\sqrt{\kappa_1/\kappa_2}$
Z_3	0.17	0.15	0.098	$\sqrt{\kappa_1/\kappa_2}$
Z_1Z_2	0.41	0.50	0.84	$\sqrt{\kappa_1/\kappa_2}$
$Z_1Z_3 = Z_2Z_3$	0.035	0.031	0.020	$\sqrt{\kappa_1/\kappa_2}$
$Z_1Z_2Z_3$	0.034	0.030	0.020	$\sqrt{\kappa_1/\kappa_2}$

TABLE VII. Table comparing Toffoli Z Pauli error rates at the optimal gate time and at the optimal gate time for the CNOT gate. The error rates from our numerical simulations were fit to $\sqrt{\kappa_1/\kappa_2}$ to produce the coefficients that appear in the table. Three different values of dephasing are included. The gate times for CNOT and Toffoli match in the case of no dephasing, and the difference between the two increases as the dephasing rate κ_ϕ increases. Qubits 1 and 2 are the controls, and qubit 3 is the target.

Z Gate	Loss, no gain	Loss and gain $n_{th} = 1/100$	Scaling
Opt. Time	0.61	0.61	$(\alpha^3\sqrt{\kappa_1\kappa_2})^{-1}$
Z	1.63	1.64	$\sqrt{\kappa_1/\kappa_2}/\alpha$
CZ Gate			
Opt. Time	0.56	0.56	$(\alpha^3\sqrt{\kappa_1\kappa_2})^{-1}$
$Z_1 = Z_2$	0.83	0.84	$\sqrt{\kappa_1/\kappa_2}/\alpha$
Z_1Z_2	0.56	0.56	$\sqrt{\kappa_1/\kappa_2}/\alpha$

TABLE VIII. Table of Z gate and CZ gate optimal times and Z error rates from numerical simulations. Each error rate or gate time is the product of the coefficient in the second or third columns of the table with the corresponding function from the fourth column. Dephasing noise has a negligible effect on the Z error rates.

bare frequency (in practice, however, the frequency shifts need to be taken into account; see below for the frequency shift due to pump). Then, in the rotating frame where every mode rotates with its own frequency, we have

$$\hat{H}_{\mathrm{rot}} = -2E_J\epsilon_p(t)\sin\Big(\sum_{k=1}^{N}\varphi_k\hat{a}_k e^{-i\omega_k t} + \mathrm{h.c.} + \varphi_b\hat{b}e^{-i\omega_b t} + \mathrm{h.c.}\Big). \tag{F2}$$

Linear drive on a phononic mode, say $\hat{a}_k$, can be readily realized by using a pump $\epsilon_p(t) = \epsilon_p\cos(\omega_p t)$ and choosing the pump frequency ω_p to be the frequency of the mode we want to drive, that is, $\omega_p = \omega_k$. Then, by taking only the leading order linear term in the sine potential (i.e., $\sin(\hat{x}) \simeq \hat{x}$), we get the desired linear drive

$$\hat{H}_{\mathrm{rot}} = -E_J\epsilon_p\varphi_k(\hat{a}_k + \hat{a}_k^\dagger) + \hat{H}', \tag{F3}$$

i.e., $\epsilon_Z = -E_J\epsilon_p\varphi_k$, where $\hat{H}'$ contains fast-oscillating terms such as $-E_J\epsilon_p(\varphi_l\hat{a}_l e^{-i(\omega_l-\omega_k)t} + \mathrm{h.c.})$ with $l \neq k$ and $-E_J\epsilon_p(\varphi_b\hat{b}e^{-i(\omega_b-\omega_k)t} + \mathrm{h.c.})$ as well as other terms that rotate even faster, e.g., $-E_J\epsilon_p\varphi_k(\hat{a}_k e^{-2i\omega_k t} + \mathrm{h.c.})$. Since the frequency differences between different modes are on the order of 100MHz but $|\epsilon_Z|/(2\pi)$ is typically not required to be larger than 1MHz, the fast-oscillating terms can be ignored by using a rotating wave approximation (RWA). For instance, the strength of the linear drive needed for the compensating Hamiltonian for the CNOT gate $\hat{H}_{\mathrm{CNOT}}$ is given by

$$\frac{\pi\alpha}{4T^\star_{\mathrm{CNOT}}} = \frac{\pi\alpha^3}{1.24}\sqrt{\kappa_1\kappa_2} = \begin{cases} 2\pi\times 2.89\mathrm{MHz} & \kappa_1/\kappa_2 = 10^{-3} \\ 2\pi\times 912\mathrm{kHz} & \kappa_1/\kappa_2 = 10^{-4} \\ 2\pi\times 289\mathrm{kHz} & \kappa_1/\kappa_2 = 10^{-5} \end{cases} \tag{F4}$$

at the optimal CNOT gate time $T^{\star}_{\mathrm{CNOT}} = 0.31/(\sqrt{\kappa_1\kappa_2}\alpha^2)$ assuming $\alpha^2 = 8$ and $\kappa_2 = 10^7 s^{-1}$. Note that the subleading cubic term in the sine potential is also neglected here. These unwanted cubic terms are smaller than the desired linear term by a factor of φ_k^2. We remark that to avoid driving unwanted higher order terms, one may alternatively drive the phononic mode directly, at the expense of increased hardware compleixty, instead of using the pump $\epsilon_p(t)$ at the ATS node.

Let us now consider a beam-splitter interaction between two phononic modes, e.g., $\epsilon_{ZZ}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)$, which is needed for implementing a CZ rotation between two cat qubits. It is also used to realize the compensating Hamiltonian for the Toffoli gate $\hat{H}_{\mathrm{TOF}}$ and to realize the SWAP operation for the X readout of a cat qubit. Note that the beam-splitter interaction is quadratic and even. Hence, it cannot be directly driven with a single pump tone since the sine potential has an odd parity. We thus jointly apply one pump tone and another drive tone to off-resonantly drive two odd terms and choose the detunings such that these two odd terms realize a resonant beam-splitter interaction when they are combined together. Since average Hamiltonian theory is useful for the analysis of the above scheme as well as many other schemes we propose below, we briefly state a key result of average Hamiltonian theory [47, 97]: given a time-dependent Hamiltonian

$$\hat{H} = \hat{H}_0 + \sum_n \Big[\hat{V}_n e^{-i\Delta_n t} + \mathrm{h.c.}\Big] \tag{F5}$$

with fast-oscillating time-dependent terms, one gets the following effective Hamiltonian by averaging out fast-oscillating terms

$$\hat{H}_{\mathrm{eff}} = \hat{H}_0 + \frac{1}{2}\sum_{m,n}\Big(\frac{1}{\Delta_m} + \frac{1}{\Delta_n}\Big)[\hat{V}_m^\dagger, \hat{V}_n]e^{i(\Delta_m - \Delta_n)t}. \tag{F6}$$

To realize the beam-splitter interaction $\hat{a}_1^\dagger\hat{a}_2 + \mathrm{h.c.}$, we drive the two terms $\hat{a}_1^\dagger\hat{a}_2\hat{b}^\dagger$ and $\hat{b}$ off-resonantly. In particular, we use a pump $\epsilon_p(t) = \epsilon_p\cos(\omega_p t)$ with a pump frequency $\omega_p = \omega_2 - \omega_1 - \omega_b - \Delta$ to off-resonantly drive the term $\hat{a}_1^\dagger\hat{a}_2\hat{b}^\dagger$ and directly drive the $\hat{b}$ mode via

$$\hat{H}_d = \epsilon_d(\hat{b}^\dagger e^{-i\omega_d t} + \mathrm{h.c.}) \tag{F7}$$

with a drive frequency $\omega_d = \omega_b + \Delta$ to off-resonantly drive the linear term $\hat{b}^\dagger$. Note that the size of the detuning $|\Delta|$ must not be larger than half the filter bandwidth $2J$ so that the drive is not filtered out. Then, by taking up to the third order terms in the sine potential (i.e., $\sin(\hat{x}) \simeq \hat{x} - \hat{x}^3/6$) in Eq. (F2), we find

$$\begin{aligned}\hat{H}_{\mathrm{rot}} &= E_J\epsilon_p\varphi_1\varphi_2\varphi_b\hat{a}_1^\dagger\hat{a}_2\hat{b}^\dagger e^{-i\Delta t} + \mathrm{h.c.}\\ &\quad + \epsilon_d\hat{b}^\dagger e^{-i\Delta t} + \mathrm{h.c.} + \hat{H}',\end{aligned} \tag{F8}$$

where $\hat{H}'$ contains fast-oscillating terms, which we ignore for the moment. Let $\chi_1 \equiv E_J\epsilon_{p,1}\varphi_1\varphi_2\varphi_b$ and $\chi_2 \equiv \epsilon_d$. Then, neglecting $\hat{H}'$, the average Hamiltonian theory yields

$$\begin{aligned}\hat{H}_{\mathrm{eff}} &= \frac{1}{\Delta}[(\chi_1\hat{a}_1\hat{a}_2^\dagger + \chi_2)\hat{b}, (\chi_1\hat{a}_1^\dagger\hat{a}_2 + \chi_2)\hat{b}^\dagger]\\ &= \frac{1}{\Delta}\Big[(\chi_1\hat{a}_1\hat{a}_2^\dagger + \chi_2)(\chi_1\hat{a}_1^\dagger\hat{a}_2 + \chi_2)\\ &\qquad + [(\chi_1\hat{a}_1\hat{a}_2^\dagger + \chi_2), (\chi_1\hat{a}_1^\dagger\hat{a}_2 + \chi_2)]\hat{b}^\dagger\hat{b}\Big]\\ &\xrightarrow{\hat{b}^\dagger\hat{b}\ll 1} \frac{1}{\Delta}(\chi_1\hat{a}_1\hat{a}_2^\dagger + \chi_2)(\chi_1\hat{a}_1^\dagger\hat{a}_2 + \chi_2)\\ &= \frac{\chi_1\chi_2}{\Delta}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger) + \frac{\chi_1^2}{\Delta}(\hat{a}_1^\dagger\hat{a}_1 + 1)\hat{a}_2^\dagger\hat{a}_2.\end{aligned} \tag{F9}$$

Note that we assumed that the population in the $\hat{b}$ mode is negligible (i.e., $\hat{b}^\dagger\hat{b} \ll 1$) and dropped the constant energy shift χ_2^2/Δ in the last line. The first term in the last line is the desired beam-splitter interaction $\epsilon_{ZZ}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)$ with a coupling strength

$$\epsilon_{ZZ} = \frac{\chi_1\chi_2}{\Delta} = E_J\epsilon_{p,1}\varphi_1\varphi_2\varphi_b\beta, \tag{F10}$$

where $\beta \equiv \chi_2/\Delta = \epsilon_d/\Delta$ can be understood as an effective displacement in the $\hat{b}$ mode. For the population of the $\hat{b}$ mode to be negligible, we need $|\beta| \ll 1$. Assuming $\beta = 0.1$ and noting that $E_J\epsilon_{p,1}\varphi_1\varphi_2\varphi_b \sim g_2 \lesssim 2\pi\times 5\mathrm{MHz}$, we find that $\epsilon_{ZZ} \sim 2\pi\times 500\mathrm{kHz}$ is achievable. The strength of the beam-splitter interaction in the compensating Hamiltonian for the Toffoli gate $\hat{H}_{\mathrm{TOF}}$ is given by (see Eq. (F16))

$$\begin{aligned}\frac{\pi}{8T^{\star}_{\mathrm{TOF}}} &= \frac{\pi\alpha^2}{2.48}\sqrt{\kappa_1\kappa_2}\\ &= \begin{cases} 2\pi\times 1.02\mathrm{MHz} & \kappa_1/\kappa_2 = 10^{-3}\\ 2\pi\times 323\mathrm{kHz} & \kappa_1/\kappa_2 = 10^{-4}\\ 2\pi\times 102\mathrm{kHz} & \kappa_1/\kappa_2 = 10^{-5}\end{cases}\end{aligned} \tag{F11}$$

at the optimal Toffoli gate time $T^{\star}_{\mathrm{TOF}} = 0.31/(\sqrt{\kappa_1\kappa_2}\alpha^2)$ assuming $\alpha^2 = 8$ and $\kappa_2 = 10^7 s^{-1}$. We also remark that the second term in the last line of Eq. (F9) gives rise to undesired cross-Kerr interaction and energy shift of the $\hat{a}_2$ mode. The unwanted cross-Kerr interaction $\hat{a}_1^\dagger\hat{a}_1\hat{a}_2^\dagger\hat{a}_2$ can in principle be cancelled by off-resonantly driving the term $\hat{a}_1\hat{a}_2\hat{b}^\dagger$ with a detuning Δ' different from Δ. The frequency shift of the mode $\hat{a}_2$ (i.e., $(\chi_1^2/\Delta)\hat{a}_2^\dagger\hat{a}_2$) can either be incorporated into the frequency matching condition or physically cancelled by off-resonantly driving the term $\hat{a}_2\hat{b}^\dagger$ (see below for more details).

Note that we have so far ignored fast-oscillating terms (i.e., $\hat{H}'$ in Eq. (F8)). These fast-oscillating terms include unwanted cubic terms, e.g., $E_J\epsilon_{p,1}\varphi_2\varphi_3\varphi_b\hat{a}_2^\dagger\hat{a}_3\hat{b}^\dagger e^{i(2\omega_2-\omega_1-\omega_3-\Delta)t} + \mathrm{h.c.}$ which would give rise to an unwanted beam-splitter interaction $\hat{a}_2^\dagger\hat{a}_3 + \mathrm{h.c.}$. If the frequencies of the modes $\hat{a}_1$, $\hat{a}_2$, and $\hat{a}_3$ are equally spaced, $2\omega_2 - \omega_1 - \omega_3$ vanishes and the unwanted term $\hat{a}_2^\dagger\hat{a}_3\hat{b}^\dagger$ interferes with the desired term $\hat{a}_1^\dagger \hat{a}_2 \hat{b}^\dagger$ as they rotate with the same frequency. However, in practice, equal frequency spacing is avoided in the optimization of the frequencies of the phononic modes. Hence, unwanted beam-splitter interactions are far detuned from the desired beam-splitter interaction. We remark that remaining fast-rotating terms in $\hat{H}'$ (different from the above beam-splitter type) are of less concern as their rotating frequencies are farther away from the frequencies of the desired terms.

Let us now move on to the selective frequency shift which is needed, e.g., for removing non-adiabatic errors of the X gate if we were to implement the X gate physically (see Eq. (21)). In practice, the 180° rotation $e^{i\pi\hat{a}^\dagger\hat{a}}$ (or $\hat{a} \to -\hat{a}$) for the X gate can be performed via software by adapting the phases of subsequent drives. However, we still discuss the selective frequency shift because it is conceptually useful for understanding our proposal for implementing the compensating Hamiltonians for the CNOT and Toffoli gates.

We first consider frequency shifts due to a pump $\epsilon_p(t) = \epsilon_p \cos(\omega_p t)$. Note that the terms $\hat{a}_k^\dagger \hat{a}_k \hat{b}^\dagger$ and $\hat{b}^\dagger$ in the sine potential are off-resonantly driven by the pump with the same detuning $\Delta = \omega_p - \omega_b$ and with coupling strengths $E_J \epsilon_p \varphi_k^2 \varphi_b$ and $-E_J \epsilon_p \varphi_b$, respectively. Hence, through the average Hamiltonian theory, we find that the frequency of the $\hat{a}_k$ mode is shifted by

$$\delta\omega_k = -\frac{E_J^2 \epsilon_p^2 \varphi_k^2 \varphi_b^2}{\omega_p - \omega_b}. \tag{F12}$$

Similarly as in the case of beam-splitter interaction, the frequency shift is accompanied by undesirable quartic terms such as self-Kerr $(\hat{a}_k^\dagger)^2 \hat{a}_k^2$ and cross-Kerr $\hat{a}_k^\dagger \hat{a}_k \hat{a}_l^\dagger \hat{a}_l$ nonlinearities. While we have ignored the frequency shifts due to pump in the discussions so far, they need to be carefully taken into account in practice.

Note that the size of frequency shift can be modulated by changing the pump amplitude ϵ_p (i.e., $|\delta\omega_k| \propto \epsilon_p^2$). However, we cannot engineer the frequency shifts due to $\hat{a}_k^\dagger \hat{a}_k \hat{b}^\dagger$ and $\hat{b}^\dagger$ in a mode-selective manner since $\hat{a}_l^\dagger \hat{a}_l \hat{b}^\dagger$ with $l \neq k$ rotates with the same frequency as those of $\hat{a}_k^\dagger \hat{a}_k \hat{b}^\dagger$ and $\hat{b}^\dagger$. In particular, since $\delta\omega_k / \delta\omega_l = \varphi_k^2 / \varphi_l^2$ and the zero-point fluctuations of phononic modes are almost identical, the frequency shifts of the phononic modes $\delta\omega_k$ are approximately independent of the mode index k. Thus, we cannot rely on frequency shifts due to $\hat{a}_k^\dagger \hat{a}_k \hat{b}^\dagger$ and $\hat{b}^\dagger$ to exclusively shift the frequency of a specific mode $\hat{a}_k$.

Selective frequency shift the mode $\hat{a}_k$ can nevertheless be realized by off-resonantly driving the term $\hat{a}_k \hat{b}^\dagger$: if we are given with a Hamiltonian $\hat{H} = \chi \hat{a}_k \hat{b}^\dagger e^{-i\Delta t} + \text{h.c.}$, the average Hamiltonian theory yields (assuming $\hat{b}^\dagger \hat{b} \ll 1$ similarly as in Eq. (F9))

$$\hat{H}_{\text{eff}} = \frac{\chi^2}{\Delta} \hat{a}_k^\dagger \hat{a}_k, \tag{F13}$$

i.e., frequency shift of the mode $\hat{a}_k$. In practice, the pumps used to off-resonantly drive the term $\hat{a}_k \hat{b}^\dagger$ may also drive $\hat{a}_l \hat{b}^\dagger$ with $l \neq k$ which will lead to the frequency shift of another mode $\hat{a}_l$. However, $\hat{a}_l \hat{b}^\dagger$ is detuned from $\hat{a}_k \hat{b}^\dagger$ by $\omega_l - \omega_k$ so the relevant detuning Δ' of the unwanted term $\hat{a}_l \hat{b}^\dagger$ is given by $\Delta' = \Delta + \omega_l - \omega_k$. Hence, the unwanted frequency shift in another mode $\hat{a}_l$ can in principle be suppressed by ensuring $|\Delta'| \gg |\Delta|$.

Building up on the intuitions gained from the discussion of selective frequency shift, we now discuss implementation of the compensating Hamiltonian for the CNOT gate in Eq. (25). Without loss of generality, we focus on the CNOT gate between the modes $\hat{a}_1$ (control) and $\hat{a}_2$ (target). Note that $\hat{H}_{\text{CNOT}}$ consists of an optomechanical coupling $(\pi/(4\alpha T))(\hat{a}_1 + \hat{a}_1^\dagger)\hat{a}_2^\dagger \hat{a}_2$ between two phononic modes, a linear drive on the control mode $-(\pi\alpha/(4T))(\hat{a}_1 + \hat{a}_1^\dagger)$, and a selective frequency shift of the target mode $-(\pi/(2T))\hat{a}_2^\dagger \hat{a}_2$. Similarly as the 180° rotation for the X gate needs not be implemented physically, the selective frequency shift of the target mode can be taken care of via software. That is, instead of using $\hat{H}_{\text{CNOT}}$ in Eq. (25), one may use a different compensating Hamiltonian

$$\hat{H}'_{\text{CNOT}} = \frac{\pi}{4\alpha T}(\hat{a}_1 + \hat{a}_1^\dagger)(\hat{a}_2^\dagger \hat{a}_2 - \alpha^2) \tag{F14}$$

as well as an appropriately modified rotating jump operator $\hat{L}'_2(t)$ such that the cat states $|0\rangle \simeq |\alpha\rangle$ and $|1\rangle \simeq |-\alpha\rangle$ in the target mode are mapped to $|-i\alpha\rangle$ and $|i\alpha\rangle$ if the control mode is in the state $|0\rangle \simeq |\alpha\rangle$, and to $|i\alpha\rangle$ and $|-i\alpha\rangle$ if the control mode is in the trigger state $|1\rangle \simeq |-\alpha\rangle$. Hence, one may simply redefine the cat-code computational basis states of the target mode as $|0\rangle \leftarrow |-i\alpha\rangle$ and $|1\rangle \leftarrow |i\alpha\rangle$ and adjust the phases of subsequent drives accordingly.

Note that the optomechanical coupling and the linear drive on the control mode still need to be implemented physically. Implementation of the linear drive is already discussed above. To realize the optomechanical coupling, one might be tempted to directly drive the cubic term $\hat{a}_1 \hat{a}_2^\dagger \hat{a}_2 + \text{h.c.}$ in the sine potential via a pump $\epsilon_p(t) = \epsilon_p \cos(\omega_p t)$. However, the direct driving scheme is not suitable for a couple of reasons: since the term $\hat{a}_1 \hat{a}_2^\dagger \hat{a}_2$ rotates with frequency ω_1, the required pump frequency is given by $\omega_p = \omega_1$ which is the same pump frequency reserved to engineer a linear drive on the $\hat{a}_1$ mode. Moreover, the term $\hat{a}_1 \hat{a}_2^\dagger \hat{a}_2$ rotates at the same frequency as those of undesired cubic terms such as $\hat{a}_1 \hat{a}_3^\dagger \hat{a}_3$, $\hat{a}_1 \hat{a}_4^\dagger \hat{a}_4$, and also $\hat{a}_1^\dagger \hat{a}_1^2$. Hence, even if the linear drive is realized via a direct driving of the phononic mode, one still cannot selectively drive the desired optomechanical coupling by using the pump frequency $\omega_p = \omega_1$ due to the frequency collision with other unwanted cubic terms. This issue is analogous to the one we had earlier that the selective frequency shift of the $\hat{a}_1$ mode is not possible via the synthesis of two terms $\hat{a}_1^\dagger \hat{a}_1 \hat{b}^\dagger$ and $\hat{b}^\dagger$.

To circumvent the above frequency-collision issue, we propose to realize the optomechanical coupling $(\hat{a}_1 + \hat{a}_1^\dagger)\hat{a}_2^\dagger\hat{a}_2$ by off-resonantly driving the term $(\hat{a}_1 + \lambda)\hat{a}_2\hat{b}^\dagger$. That is, given a Hamiltonian $\hat{H} = \chi(\hat{a}_1 + \lambda)\hat{a}_2\hat{b}^\dagger e^{-i\Delta t} +$ h.c., we get the following effective Hamiltonian through the time averaging

$$\hat{H}_{\text{eff}} = \frac{\chi^2\lambda}{\Delta}\Big(\hat{a}_1 + \hat{a}_1^\dagger + \lambda + \frac{1}{\lambda}\hat{a}_1^\dagger\hat{a}_1\Big)\hat{a}_2^\dagger\hat{a}_2, \tag{F15}$$

where we again assumed that the population of the $\hat{b}$ mode is negligible (i.e., $\hat{b}^\dagger\hat{b} \ll 1$). In particular, by choosing $\lambda = -2\alpha$, we can realize the optomechanical coupling as well as the selective frequency shift of the $\hat{a}_2$ mode, i.e., $\hat{H}_{\text{eff}} \propto (\hat{a}_1 + \hat{a}_1^\dagger - 2\alpha)\hat{a}_2^\dagger\hat{a}_2$ up to an undesired cross-Kerr term $-\hat{a}_1^\dagger\hat{a}_1\hat{a}_2^\dagger\hat{a}_2/(2\alpha)$ (which can in principle be cancelled by off-resonantly driving the term $\hat{a}_1\hat{a}_2\hat{b}^\dagger$). Hence, if we realize $\hat{H}_{\text{CNOT}}$ this way, we need not rely on software to keep track of the phase of the target mode as the phase shift is physically realized. We also remark that the term $(\hat{a}_1 + \lambda)\hat{a}_2\hat{b}^\dagger$ is detuned from other undesired terms such as $(\hat{a}_1 + \lambda)\hat{a}_k\hat{b}^\dagger$ with $k \geq 3$ by a frequency difference $\omega_2 - \omega_k$. Thus, the unwanted optomechanical coupling $(\hat{a}_1 + \hat{a}_1^\dagger)\hat{a}_k^\dagger\hat{a}_k$ can be suppressed by a suitable choice of the detuning Δ similarly as in the case of selective frequency shift.

Note that while the cubic term $\hat{a}_1\hat{a}_2\hat{b}^\dagger$ in $(\hat{a}_1 + \lambda)\hat{a}_2\hat{b}^\dagger$ can be realized by using the sine potential, the other quadratic term $\hat{a}_2\hat{b}^\dagger$ cannot be directly realized from the sine potential which has an odd parity. The quadratic interaction $\hat{a}_2\hat{b}^\dagger$ can in principle be realized by synthesizing (using the average Hamiltonian theory) two odd terms $\hat{a}_2(\hat{b}^\dagger)^2$ and $\hat{b}^\dagger$. To put everything together and get the desired optomechanical coupling, however, the results of average Hamiltonian theory need to be concatenated. In other words, to analyze the full scheme for the desired optomechanical coupling, a higher-order average Hamiltonian theory is needed. We leave it as a future work to thoroughly analyze such a scheme.

Lastly, let us consider the compensating Hamiltonian $\hat{H}_{\text{TOF}}$ for the Toffoli gate in Eq. (35). $\hat{H}_{\text{TOF}}$ is explicitly given by

$$\begin{aligned}\hat{H}_{\text{TOF}} = &-\frac{\pi}{8\alpha^2 T}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2)\\ &+\frac{\pi}{8\alpha T}(\hat{a}_1 + \hat{a}_1^\dagger - \alpha)(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2)\\ &+\frac{\pi}{8\alpha T}(\hat{a}_2 + \hat{a}_2^\dagger - \alpha)(\hat{a}_3^\dagger\hat{a}_3 - \alpha^2).\end{aligned} \tag{F16}$$

Note that the terms in the second and the third lines are in the same form as the compensating Hamiltonian for the CNOT gate. Thus, they can be realized in a similar way as described above. The terms in the first line contain a beam-splitter interaction $(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)$, which we have already discussed above, as well as a quartic term $(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)\hat{a}_3^\dagger\hat{a}_3$. Since the sine potential has an odd parity, it is not possible to drive the quartic term directly. The quartic term can nevertheless be realized by off-resonantly driving the term $(\hat{a}_1 + \hat{a}_2)\hat{a}_3\hat{b}^\dagger$: given $\hat{H} = \chi(\hat{a}_1 + \hat{a}_2)\hat{a}_3\hat{b}^\dagger e^{-i\delta t} +$ h.c., we get

$$\hat{H}_{\text{eff}} = \frac{\chi^2}{\Delta}(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)\hat{a}_3^\dagger\hat{a}_3 + \frac{\chi^2}{\Delta}(\hat{a}_1^\dagger\hat{a}_1 + \hat{a}_2\hat{a}_2^\dagger)\hat{a}_3^\dagger\hat{a}_3, \tag{F17}$$

i.e., the desired quartic interaction and unwanted cross-Kerr interactions between a control and the target modes. The undesired cross-Kerr terms, which are as strong as the desired quartic term, can in principle be cancelled by off-resonantly driving the terms $\hat{a}_1\hat{a}_3\hat{b}^\dagger$ and $\hat{a}_2\hat{a}_3\hat{b}^\dagger$ with detunings Δ_1 and Δ_2 which are different from each other and also from Δ.

The required coupling strength of the quartic interaction $(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1\hat{a}_2^\dagger)\hat{a}_3^\dagger\hat{a}_3$ is given by

$$\begin{aligned}\frac{\pi}{8\alpha^2 T_{\text{TOF}}} &= \frac{\pi}{2.48}\sqrt{\kappa_1\kappa_2}\\ &= \begin{cases} 2\pi \times 128\text{kHz} & \kappa_1/\kappa_2 = 10^{-3}\\ 2\pi \times 40.3\text{kHz} & \kappa_1/\kappa_2 = 10^{-4}\\ 2\pi \times 12.8\text{kHz} & \kappa_1/\kappa_2 = 10^{-5}\end{cases}\end{aligned} \tag{F18}$$

at the optimal Toffoli gate time $T_{\text{TOF}}^\star = 0.31/(\sqrt{\kappa_1\kappa_2}\alpha^2)$ assuming $\kappa_2 = 10^7 s^{-1}$. Note that the coupling strength of the term $\hat{a}_1\hat{2}\hat{b}^\dagger$ and $\hat{a}_1\hat{3}\hat{b}^\dagger$ are comparable to $g_2 \lesssim 2\pi \times 5\text{MHz}$. Incorporating the bosonic enhancement factor due to the average excitation number α^2, we require the detuning Δ to be much larger than $g_2\alpha^2$, e.g., $\Delta = 10g_2\alpha^2$. Then the achievable coupling strength of the quartic interaction is given by $g_2^2/(10g_2\alpha^2) = g_2/(10\alpha^2) \lesssim 2\pi \times 60\text{kHz}$ assuming $\alpha^2 = 8$.

Appendix G: Measurement

In this appendix we discuss measurement schemes for high fidelity readout in both the X and Z basis.

1. X-basis measurement with transmon

Here we discuss in more detail the X-basis readout scheme used to generate the infidelities used in most of the error correction simulations. Note that throughout this appendix, when we refer to measurement infidelities or error probabilities, we are referring to misassignment probabilities

$$\epsilon_s = 1 - \text{P}(s|s) \tag{G1}$$

where $\text{P}(s|s)$ is the probability of reading out the state s given that the cavity was in the state s. In the case of X-basis measurement we are referring to a determination of the parity of a phononic mode or equivalently readout in the basis of even and odd cat states, i.e., $|\pm\rangle \propto |\alpha\rangle \pm |-$

$\alpha\rangle$. To realize such a measurement with minimal impact on the length of an error correction cycle we utilize an additional phononic mode which we refer to as the readout mode. This mode is interrogated by a transmon in parallel with the next error correction cycle. As is pictured in Fig. 2, every unit cell contains this additional readout mode connected to a transmon. Pictured in Fig. 39 is the circuit we use for measuring an X stabilizer which we now walk through in more detail. To perform an X stabilizer measurement first the ancilla qubit $\hat{a}_1$ is entangled with the data qubits. Subsequently we non-adiabatically deflate the ancilla qubit on a timescale of the order $1/\kappa_2$, mapping the even parity state to $|\hat{n} = 0\rangle$ and the odd parity state to $|\hat{n} = 1\rangle$ [24]. Deflation is achieved under evolution with the two-phonon dissipator,

$$\frac{d\hat{\rho}(t)}{dt} = \kappa_2 \mathcal{D}[\hat{a}_1^2 - \alpha(t)^2]\hat{\rho}(t), \tag{G2}$$

by taking $\alpha(t)$ from α_0 to $\alpha_1 < \alpha_0$. In our case we rapidly take $\alpha(t)$ from its initial value to $\alpha_1 = 0$ where we evolve for a time on the order of $1/\kappa_2$. The deflation is not required to be adiabatic since we do not need to maintain phase coherence between the even and odd parity states. The utility of the deflation is that it makes the state of the cavity less susceptible to single phonon loss events which change its parity. After the deflation we perform a SWAP gate between the ancilla qubit and a phononic readout mode transferring the excitation from the ancilla mode to the readout mode. The physical implementation of this Hamiltonian is discussed in Eq. (F7) but here we include a different style of derivation. To realize a SWAP gate we start with the nonlinear part of the ATS Hamiltonian with the buffer mode ($\hat{b}$), the ancilla phononic mode ($\hat{a}_1$) and the phononic readout mode ($\hat{a}_2$).

$$\begin{aligned}\hat{H} = -2E_J\epsilon(t)\sin\big(&\varphi_{\mathrm{a}1}(\hat{a}_1 + \hat{a}_1^\dagger) + \\ &\varphi_{\mathrm{a}2}(\hat{a}_2 + \hat{a}_2^\dagger) + \varphi_{\mathrm{b}}(\hat{b} + \hat{b}^\dagger)\big)\end{aligned} \tag{G3}$$

First a pump

$$\epsilon(t) = \epsilon_p \cos\omega_p t \tag{G4}$$

is applied with $\omega_p = \omega_b + \omega_{a1} - \omega_{a2} + \Delta$. Expanding the sine to third order we realize the Hamiltonian (in the rotating frame of all of the modes)

$$\hat{H}_{rot} = E_J\epsilon\varphi_{\mathrm{a}1}\varphi_{\mathrm{a}2}\varphi_b(\hat{a}_1\hat{b}\hat{a}_2^\dagger e^{i\Delta t} + \mathrm{h.c.}). \tag{G5}$$

Adding a drive on the b mode at frequency $-\Delta$ with effective strength β we realize the Hamiltonian

$$\begin{aligned}\hat{H}_{rot} &= g_r(\hat{a}_1^\dagger\hat{a}_2 + \hat{a}_1^\dagger\hat{a}_2), \\ g_r &= E_J\epsilon_p\beta\varphi_{\mathrm{a}1}\varphi_{\mathrm{a}2}\varphi_b.\end{aligned} \tag{G6}$$

In practice $g_r < g_2$ as it scales quadtratically with the zero point fluctuations of the storage modes and $\beta \ll 1$. Evolution under this Hamiltonian for a duration $\pi/2g_r$ realizes a SWAP gate (there is a rotation of the swapped state by 90 degrees).

An advantage of this readout scheme is that after the SWAP has occurred the next cycle of quantum error correction can continue in parallel with the measurement of the readout mode. Not only does this mean that there is less time where the data qubits are subject to idling error but it also means that we can achieve higher measurement fidelity by repeatedly measuring the readout mode. We note that this simple layout choice could be generally useful in other architectures. In the specific case of this proposal we repeatedly perform QND parity measurements which we majority vote to get our final measurement outcome [53–55]. This allows for a suppression of the measurement error due to the ancilla. CNOT gate times of about $1\mu s$ allow us to conduct 5 parity measurements during the CNOTs of surface code error correction cycle and 3 parity measurement during the CNOTs of a repetition code memory cycle for the realistic numbers we have chosen for the measurement time listed in Fig. 40. In general more advanced methods than majority voting will give higher fidelity [54].

Measurement of readout mode parity is done using a dispersive coupling with a transmon qubit [53] $H = \omega_q\hat{\sigma}_z + (\omega_r - \chi\hat{\sigma}_z)\hat{a}^\dagger\hat{a}$ where $\hat{a}$ corresponds to a bosonic mode and $\hat{\sigma}_z$ corresponds to a transmon qubit. Evolution under this Hamiltonian realizes a unitary of the form $\hat{U} = I \otimes |g\rangle\langle g| + e^{i\hat{a}_2^\dagger\hat{a}_2\pi}|e\rangle\langle e|$ in a time $t = \pi/\chi$ which is a controlled parity gate. As pictured in Fig. 39 placed between an initialization in $|+\rangle$ and measurement of the transmon in the X-basis this realizes a QND measurement of the readout mode parity.

We have performed simple simulations of this measurement scheme to determine the rough infidelities for different single phonon loss rates of the phononic modes κ_1. To start we have conducted master equation simulations of the deflation and swap steps under the influence of single photon loss, gain, and dephasing. Next we have performed master equation evolution punctuated by instantaneous projective measurements to capture the remainder of the measurement process. The idle time is lengthened to capture the effect of cavity decay during measurement. Through an additional parameter we capture the effect of transmon errors such as decay, dephasing, and readout error during the dispersive interaction and readout for every measurement. In other circumstances there can be concern over transmon decay during measurement because it induces dephasing of the cavity. In our case since we are only concerned about measuring the parity this dephasing does not matter since the parity will be unaffected. As a result we are justified in lumping the effect of this transmon decay into our fixed parameter representing the transmon infidelity mechanisms. In any event this effect would be minor given recent advances in transmon coherence. We also note that recent advances in transmon measurement would allow more aggressive transmon measurement fidelities than what we assume [55]. This would allow us to use fewer measurements to achieve the same fidelities we currently expect to achieve.

Here we give an outline of how we sample one measurement sequence. We start by performing master equation evolution under the deflation and SWAP to determine $P(\text{even})$ and $P(\text{odd})$ before the first measurement. Then we sample from these probabilities to determine what state the first measurement will project the readout mode onto. Note that we assume the projection is onto the $|\hat{n}=0\rangle$ and $|\hat{n}=1\rangle$ manifold which is a good approximation given the amount of deflation used. Starting from the state the readout mode is projected onto after the first measurement we perform master equation evolution to roughly include the effects of the single photon loss, gain, and dephasing on the readout mode during the inter measurement period and during measurement. We then repeat this projection and evolution for the remaining number of measurements that are used, giving us one sequence of projections. To include the effect of transmon errors we then add additional randomness associated with a fixed transmon error probability (ϵ_q) giving us the final measurement sequence. We have performed monte carlo sampling of these measurement sequences to determine the measurement infidelities of the majority voting process. A plot of the infidelities are pictured in Fig. 40. The assumed numbers in the simulation are listed in the figure caption.

Defining N (odd) to be the total number of measurements and $k \equiv (N+1)/2$, to leading order the error probability for the majority voting of the repeated measurements for initial even and odd cat states in the case of no gain are

$$\begin{aligned}\epsilon_{\text{even}} &= \epsilon_{(\text{deflate + SWAP})} + \binom{N}{k}\epsilon_q^k(1-\epsilon_q)^{N-k}\\ \epsilon_{\text{odd}} &= \epsilon_{(\text{deflate + SWAP})} + \binom{N}{k}\epsilon_q^k(1-\epsilon_q)^{N-k} + \kappa_1 \mathrm{T}_p\end{aligned} \tag{G7}$$

where T_p is the amount of time after the SWAP and before the kth measurement and $\epsilon_{(\text{deflate + SWAP})}$ is the error from the deflation and SWAP for the given initial state. In the above expressions the first term is the contribution to the error from the deflation and SWAP steps. The second term is due to transmon error where k measurements are incorrect. The last term in the case of an odd initial state is the probability of a T_1 event before the kth measurement which will lead with high probability to all the remaining measurements giving 0. Note that this is the reason that for an odd initial state and larger κ_1/κ_2 values majority voting 5 measurements underperforms majority voting 3 measurements. Using a more advanced procedure than majority voting would mitigate this problem.

2. X-basis measurement without transmon

Here we discuss an alternative X-basis readout scheme which can be done using only the ATS and the buffer mode. Conducting X-basis readout without a transmon is advantageous for device layout since transmons and their readout modes require many control lines and take up a lot of space compared to ATSs and phononic modes. Additionally with fewer modes there are fewer crosstalk terms to deal with. This is achievable using the coupling Hamiltonian

$$\hat{H}_r = i g_r \hat{a}^\dagger\hat{a}(\hat{b}^\dagger - \hat{b}). \tag{G8}$$

Here $\hat{a}$ is the annihilation operator for a storage mode and $\hat{b}$ is the annihilation operator for the buffer mode. Note that this is equivalent to the longitudinal readout discussed with transmons [56]. We can not achieve this Hamiltonian by putting a pump at the b mode frequency since that would also bring the same terms corresponding to the other storage modes on resonance. Now we give an outline for how a Hamiltonian of this form could be realized. We start with the Hamiltonian in the rotating frame of all the modes

$$\begin{aligned}\hat{H}_{\text{rot}} = g_a(\hat{a}\hat{b}^{\dagger 2}e^{-i\Delta t} + \hat{a}^\dagger\hat{b}^2 e^{i\Delta t}) +\\ i g_b(\hat{a}\hat{b}^\dagger e^{-i\Delta t} - \hat{a}^\dagger\hat{b}e^{i\Delta t}).\end{aligned} \tag{G9}$$

Using effective Hamiltonian theory for harmonic terms this corresponds to an effective Hamiltonian [47]

$$\begin{aligned}\hat{H}_{eff} = \frac{g_a^2}{\Delta}[\hat{a}^\dagger\hat{b}^2, \hat{a}\hat{b}^{\dagger 2}] + \frac{g_b^2}{\Delta}[\hat{a}^\dagger\hat{b}, \hat{a}\hat{b}^\dagger] +\\ \frac{i g_a g_b}{\Delta}([\hat{a}^\dagger\hat{b}^2, \hat{a}\hat{b}^\dagger] - [\hat{a}^\dagger\hat{b}, \hat{a}\hat{b}^{\dagger 2}])\\ = \frac{g_a^2}{\Delta}(2\hat{a}^\dagger\hat{a}(1+2\hat{b}^\dagger\hat{b}) - \hat{b}^{\dagger 2}\hat{b}^2) + \frac{g_b^2}{\Delta}(\hat{a}^\dagger\hat{a} - \hat{b}^\dagger\hat{b}) +\\ \frac{i g_a g_b}{\Delta}(2\hat{a}^\dagger\hat{a} - \hat{b}^\dagger\hat{b})(\hat{b} - \hat{b}^\dagger)\\ \xrightarrow{\hat{b}^\dagger\hat{b}\ll 1} \frac{2 i g_a g_b}{\Delta}\hat{a}^\dagger\hat{a}(\hat{b} - \hat{b}^\dagger) + \frac{2g_a^2 + g_b^2}{\Delta}\hat{a}^\dagger\hat{a}\end{aligned} \tag{G10}$$

where we have neglected many of the terms due to the small occupation of the buffer mode. Terms that contain $\hat{a}^\dagger\hat{a}$ simply induce rotations of the storage mode which do not affect the readout. The self-kerr term on the buffer $\hat{b}^{\dagger 2}\hat{b}^2$ and the energy term $\hat{b}^\dagger\hat{b}$ can be neglected because of the negligible occupation of the buffer. Some care needs to be taken with the phases but both parts of the starting Hamiltonian are achievable using simple pumps and drives on the buffer. In the future we plan to numerically confirm the method outlined for off resonantly achieving the desired coupling to understand the necessary limits on Δ and $\langle\hat{b}^\dagger\hat{b}\rangle$ and explore ways to compensate for the undesired terms.

To achieve parity readout using this Hamiltonian we first deflate the storage mode. As before this deflation is

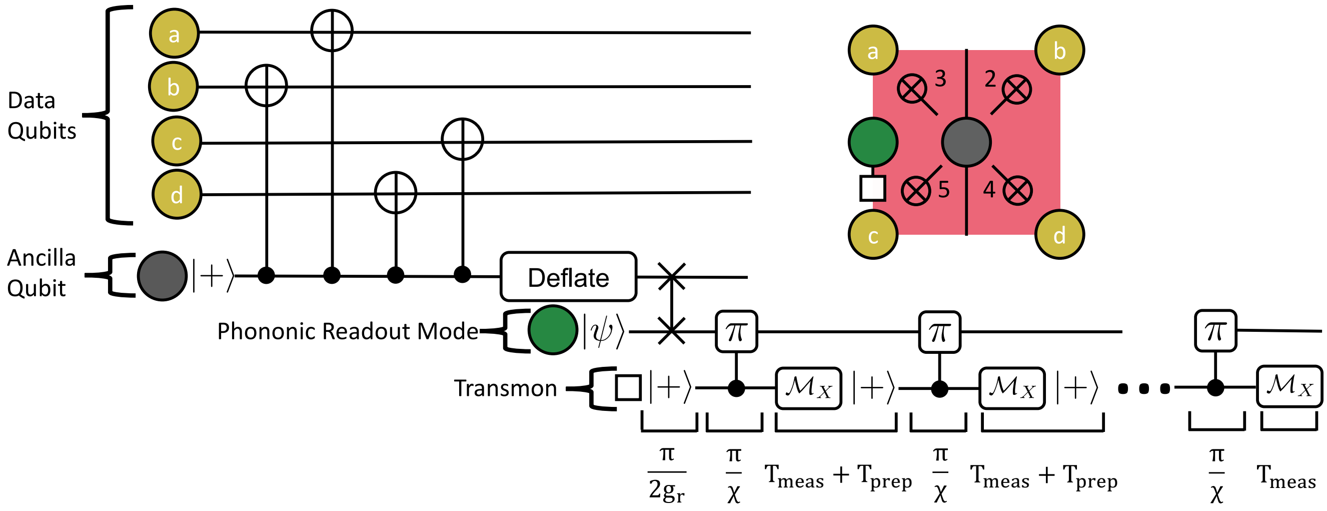


FIG. 39. Circuit used for an X-basis measurement in the context of an X-type stabilizer measurement. The first step consists of entangling the ancilla qubit with the data qubits. Afterwords, the ancilla qubit is deflated followed by a SWAP with a readout mode. Lastly, the readout mode is repeatedly measured using a transmon qubit. The duration's for the parts of the measurement procedure are labeled at the bottom of the figure below each circuit element. While these repeated parity measurements are occurring, the CNOT gates of the next error correction cycle can begin. Also included is a diagram of the physical layout of the stabilizer to give context to the measurement circuit.

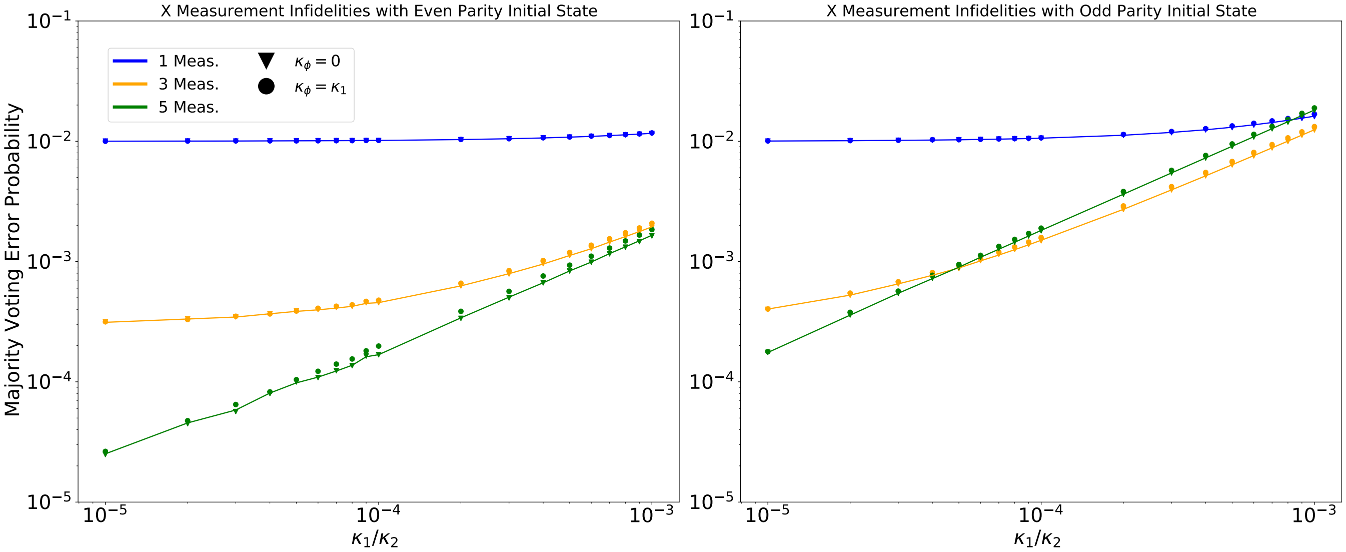


FIG. 40. Simulated error probabilities for X measurement for the cases of both even parity and odd parity initial states. In these simulations we have taken a fixed transmon error probability (ϵ_q) of 99%, $1/\kappa_2 = 100$ ns, deflation time of $3/\kappa_2$, $\mathrm{T_{entangle}} + \mathrm{T_{readout}} + \mathrm{T_{reset}} = 600$ ns, $\alpha^2 = 8$, and $g/2\pi = 1$ MHz [98, 99]. The points with dephasing also include $n_{th} = .01$. We have taken the parity measurement to be QND and fully projective. The dependence on κ_1/κ_2 is stronger for the case of an odd initial state as is expected since this is mapped to $|1\rangle$ after the deflation. The plotted curve is the leading order analytic model for the case of $\kappa_\phi = 0$ Eq. (G7). There is some small numerical imprecision in the deflation simulation for the smallest κ_1/κ_2 but this has a negligible effect on the reported average measurement error used in the error correction simulations.

achieved by abruptly setting the two-phonon dissipator for the ancilla mode equal to $\mathcal{D}[\hat{a}^2]$ and waiting for a timescale comparable to $1/\kappa_2$. Subsequent to this deflation we perform homodyne measurement while evolving under the above Hamiltonian. Now we compute the fidelity of this homodyne readout where we aim to distinguish $|\hat{n}=0\rangle$ from $|\hat{n}=1\rangle$. Here we closely follow the derivation of SNR in [56]. The Langevin equation for the evolution of the buffer mode in the interaction picture is

$$\begin{aligned}\dot{\hat{b}} &= -i[\hat{b},\hat{H}_r] - \frac{\kappa_b}{2}b - \sqrt{\kappa_b}\hat{b}_{in} \\ &= g_r\hat{a}^\dagger\hat{a} - \frac{\kappa_b}{2}\hat{b} - \sqrt{\kappa_b}\hat{b}_{in}\end{aligned} \tag{G11}$$

where $\hat{b}_{in}$ is the input field. In the following calculations we will neglect the single phonon loss of the ancilla mode which in this simple case will add an average readout error probability of roughly $\kappa_a t/4$. We integrate this equation to get the expected value of the buffer mode

$$\langle\hat{b}(t)\rangle = \frac{2g_r}{\kappa_b}\langle\hat{a}^\dagger\hat{a}\rangle(1-e^{-\frac{\kappa_b t}{2}}). \tag{G12}$$

The measurement operator for integration up to time τ and with homodyne angle ϕ_h is defined as [56, 57]

$$\hat{M}(\tau) = \sqrt{\kappa_b}\int_0^\tau dt[\hat{b}^\dagger_{out}(t)e^{i\phi_h} + \hat{b}_{out}(t)e^{-i\phi_h}]. \tag{G13}$$

Evaluating the average of this integral with the optimal phase gives

$$\langle\hat{M}(t)\rangle = \frac{4g_r\langle\hat{a}^\dagger\hat{a}\rangle}{\kappa_b}(-2+2e^{-\frac{\kappa_b t}{2}}+\kappa_b t). \tag{G14}$$

Here we have used the standard input-output condition that $\hat{b}_{out} = \hat{b}_{in} + \sqrt{\kappa_b}\hat{b}$ and the conditions on $\hat{b}_{in}$ that $\langle\hat{b}_{in}\rangle = 0$ and $\langle\hat{b}_{in}(t)\hat{b}^\dagger_{in}(t')\rangle = \delta(t-t')$. There is a drive on $\hat{b}$ to realize the Hamiltonian which we neglect and could be replaced by an appropriate pump. Next we compute the SNR which is defined as

$$\mathrm{SNR}^2 = \frac{|\langle\hat{M}\rangle_1 - \langle\hat{M}\rangle_0|^2}{\langle\hat{M}^2_{N(1)}\rangle + \langle\hat{M}^2_{N(0)}\rangle}. \tag{G15}$$

where $\hat{M}_{N(x)} = \hat{M} - \langle\hat{M}\rangle_x$ so $\langle\hat{M}^2_{N(0)}\rangle = \langle\hat{M}^2_{N(1)}\rangle = \kappa_b t$. Thus the SNR is

$$\mathrm{SNR}(\tau) = \frac{4g_r}{\kappa_b\sqrt{2\kappa_b t}}(-2+2e^{-\frac{\kappa_b t}{2}}+\kappa_b t). \tag{G16}$$

The separation error for this readout, which will be in addition to the effect of the single phonon loss mentioned earlier, will be given by [58]

$$\epsilon_{\mathrm{sep}}(\tau) = \frac{1}{2}\mathrm{Erfc}(\frac{\mathrm{SNR}(\tau)}{2}). \tag{G17}$$

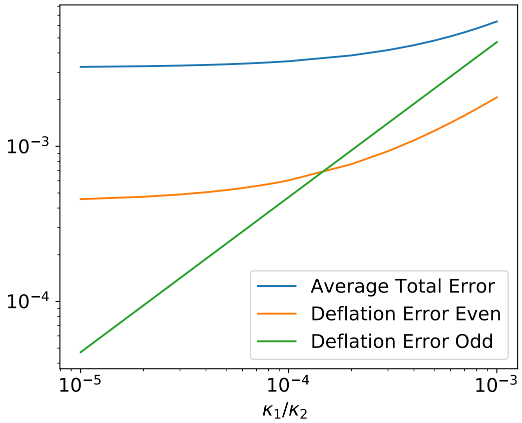


FIG. 41. Plot of deflation and average total error $((P(\text{even}|\text{odd}) + P(\text{odd}|\text{even}))/2)$ for X measurement using optomechanical coupling. In this plot the deflation is for a duration $4/\kappa_2$. Included in the average total error is the separation error of the homodyne readout starting with $|\hat{n}=0\rangle$ and $|\hat{n}=1\rangle$ in addition to the deflation error. The deflation error is defined as $1-\mathrm{P}(0)$ for the even state and $1-\mathrm{P}(1)$ for the odd state. We have taken a quantum efficiency of .5, $\kappa_b/2\pi = 20$ MHz, and $g_r/2\pi = 6$ MHz which correspond to a readout separation error of $3*10^{-3}$ in an integration time 400 ns. We have used this separation error number for all κ_1/κ_2 since the single photon loss is not dominant. This coupling strength ensures $\hat{b}^\dagger\hat{b} < 1$ as is required for self consistency in the derivation of the readout Hamiltonian with effective Hamiltonian theory. We make the assumption that deflation error adds to the readout error.

The deflation will contribute additional error.

For this paper we have taken $\kappa_b/2\pi = 20$ MHz and $g_r/2\pi = 6$ MHz. We expect this κ_b can be achieved using a multimode dump. In the future we will confirm numerically the ability to get this coupling strength in the context of the effective Hamiltonian theory. Plotted in Fig. 41 are the deflation error and average total error for this parameter choice for different values of κ_1/κ_2.

3. *Z*-basis measurement

For Z measurement we use a beamsplitter interaction between the buffer mode and a phononic storage mode

$$\hat{H}_r = g_r(\hat{a}^\dagger\hat{b} + \hat{b}^\dagger\hat{a}) \tag{G18}$$

where $\hat{a}$ is an annihilation operator on a storage mode and $\hat{b}$ is an annihilation operator on the buffer mode. By homodyning the output of the buffer mode we determine the state of the storage mode. We perform this readout scheme with the two-phonon dissipation off. A similar scheme has been realized for Kerr-Cat qubits in [24] while a squeezing drive is on. To achieve this beamsplitter Hamiltonian we use a very similar procedure to the beamplitter used for the SWAP in the X-basis readout. We start from the ATS Hamiltonian with a single storage and buffer mode.

$$\hat{H} = -2E_J\epsilon(t)\sin\left(\varphi_{\mathrm{a}}(\hat{a}+\hat{a}^\dagger) + \varphi_b(\hat{b}+\hat{b}^\dagger)\right) \tag{G19}$$

By applying a pump

$$\epsilon(t) = \epsilon_p\cos(\omega_p t) \tag{G20}$$

at $\omega_p = 2\omega_b - \omega_a + \Delta$ and moving to the rotating frame we get

$$\hat{H}_{rot} = \frac{1}{2}E_J\epsilon\varphi_{\mathrm{a}}\varphi_b^2(\hat{b}^2\hat{a}^\dagger e^{i\Delta t} + \text{h.c.}). \tag{G21}$$

Adding a drive on the b mode at frequency Δ with effective strength β we realize the Hamiltonian

$$\begin{aligned}\hat{H}_r &= g_r(\hat{a}^\dagger\hat{b} + \hat{b}^\dagger\hat{a}),\\ g_r &= E_J\epsilon_p\beta\varphi_{\mathrm{a}}\varphi_b^2.\end{aligned} \tag{G22}$$

The coupling g_r here can be comparable to g_2 since $\varphi_b > \varphi_a$. Note that an equivalent derivation using effective Hamiltonian theory was done in the main text.

The coupled Langevin equations governing the evolution of the storage and buffer modes in the interaction picture are

$$\begin{aligned}\dot{\hat{a}} &= -i[\hat{a},\hat{H}_r] = -ig_r\hat{b},\\ \dot{\hat{b}} &= -i[\hat{b},\hat{H}_r] - \frac{\kappa_b}{2}\hat{b} - \sqrt{\kappa_b}\hat{b}_{in} = -ig_r\hat{a} - \frac{\kappa_b}{2}\hat{b} - \sqrt{\kappa_b}\hat{b}_{in}\end{aligned} \tag{G23}$$

Here κ_b is the single phonon loss rate of the buffer mode and we have neglected the single phonon loss rate of the storage mode since it is far weaker than the relevant scale of κ_b. These equations can be straightforwardly integrated [100] to give

$$\begin{aligned}\hat{a}(t) &= \frac{\hat{a}(0)}{\beta}e^{-\frac{\kappa_b t}{4}}(\beta\cosh\frac{\beta t}{4} + \kappa_b\sinh\frac{\beta t}{4}),\\ \hat{b}(t) &= -i\frac{4g\hat{a}(0)}{\beta}e^{-\kappa_b t/4}\sinh\frac{\beta t}{4}\end{aligned} \tag{G24}$$

where $\beta = \sqrt{\kappa_b^2 - (4g_r)^2}$. Here we have not included the mean zero terms with $\hat{b}_{\mathrm{in}}$ since they are not relevant for computing the signal. The measurement operator with a uniform readout window is defined to be [56, 57]

$$\hat{M}(\tau) = \sqrt{\kappa_b}\int_0^\tau dt[\hat{b}^\dagger_{out}(t)e^{i\phi_h} + \hat{b}_{out}(t)e^{-i\phi_h}] \tag{G25}$$

Using the input-output boundary condition that $\hat{b}_{out} = \hat{b}_{in} + \sqrt{\kappa_b}\hat{b}$. We can determine the average of the measurement operator to be

$$\langle\hat{M}(\tau)\rangle = \frac{2\kappa_b\langle\hat{a}(0)\rangle\sin\phi_h}{g_r} * \left[1 - e^{-\kappa_b\tau/4}\left[\cosh\frac{\beta\tau}{4} + \frac{\kappa_b}{\beta}\sinh\frac{\beta\tau}{4}\right]\right]. \tag{G26}$$

We have also taken the input to be the vacuum with the property $\langle\hat{b}_{in}(t')\hat{b}^\dagger_{in}(t)\rangle = \delta(t-t')$. There is a weak drive on $\hat{b}$ to realize the Hamiltonian which we neglect and could be replaced with a flux pump. From the average of the measurement signal $\langle\hat{M}(\tau)\rangle$ we can determine the measurement SNR using

$$\mathrm{SNR}^2 = \frac{|\langle\hat{M}\rangle_\alpha - \langle\hat{M}\rangle_{-\alpha}|^2}{\langle\hat{M}^2_{N(\alpha)}\rangle + \langle\hat{M}^2_{N(-\alpha)}\rangle}. \tag{G27}$$

The noise terms are approximately given by

$$\langle\hat{M}^2_{N(\pm\alpha)}\rangle = \langle(\hat{M}(\tau) - \langle\hat{M}(\tau)\rangle_{\pm\alpha})\rangle = \kappa_b\tau \tag{G28}$$

which we have confirmed in numerics. This is especially true for longer times ($\tau > 100ns \gg 1/\kappa_b$) in our simulations.

Solving for the SNR and optimizing the phase we get

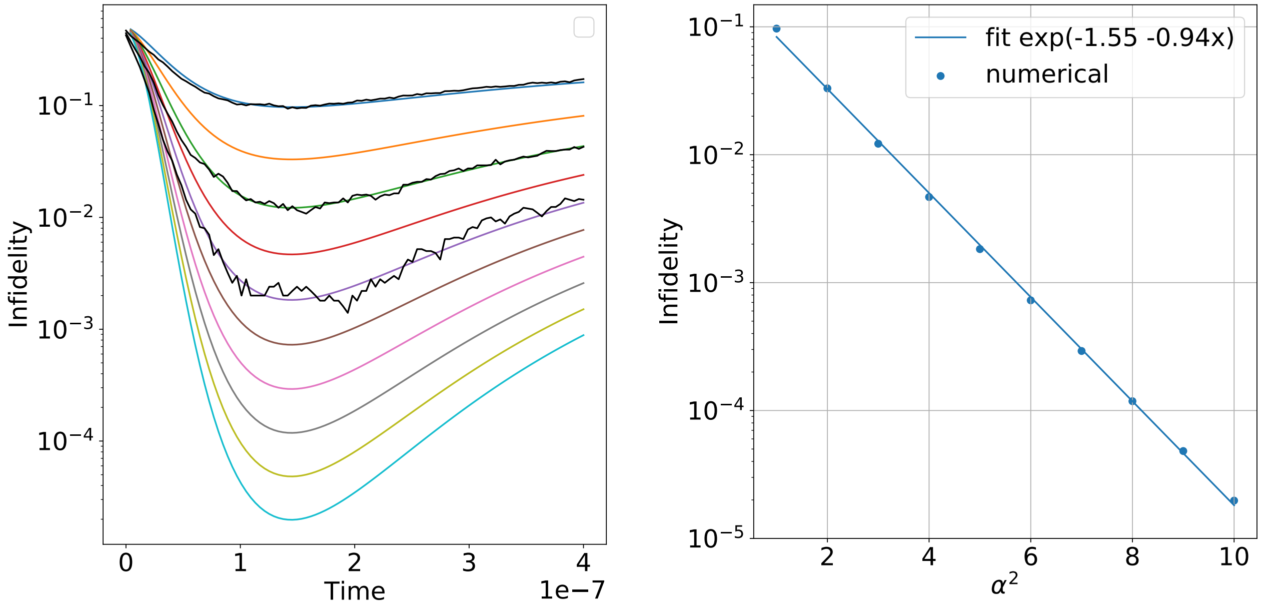
$$\mathrm{SNR}_\alpha(\tau) = \alpha\sqrt{8\kappa_b}\frac{\left[1 - e^{-\kappa_b\tau/4}\left[\cosh\frac{\beta\tau}{4} + \frac{\kappa_b}{\beta}\sinh\frac{\beta\tau}{4}\right]\right]}{g\sqrt{\tau}}. \tag{G29}$$

As is expected since this readout scheme is not QND and doesn't preserve the state of the cavity at long times the readout SNR goes as $1/\sqrt{\tau}$ as we are only integrating noise. The readout separation error will be given by

$$\epsilon_{\mathrm{sep},\alpha}(\tau) = \frac{1}{2}\mathrm{Erfc}(\frac{\mathrm{SNR}_\alpha(\tau)}{2}). \tag{G30}$$

The leading order effect of single photon loss is that loss events lower $\bar{n}$ from $|\alpha|^2$ to $|\alpha|^2 - 1$ decreasing SNR. The error probability associated with this scales as $\kappa_1|\alpha|^2\tau\epsilon_{\mathrm{sep},\sqrt{|\alpha|^2-1}}$. Since the fidelity of readout with $\bar{n} = |\alpha|^2 - 1$ is only different by a small factor than that with $\bar{n} = |\alpha|^2$ and $\kappa_1\tau \ll 1$ this factor is subleading compared to the separation error.

The effect of dephasing on the fidelity can be approximately found given κ_ϕ. First note that with dephasing initial states $|\alpha\rangle$ and $|-\alpha\rangle$ evolve to $|\alpha e^{i\theta}\rangle$ and $|-\alpha e^{i\theta}\rangle$ where θ is distributed as a mean zero normal with standard deviation $\sqrt{\kappa_\phi t}$. Thus we can find the effect of the dephasing on the coherent states projected onto the axis of the homodyne measurement. In other words we can find the effective α for the coherent states subject to



(a) Error vs. Time

(b) Error Fit

FIG. 42. a.) Measurement error probability for Z-basis readout as a function of time. Here we have taken $\kappa_b/2\pi = 20$ MHz, $g_r/2\pi = 4$ MHz, and quantum efficiency of .5. The black lines correspond to infidelities from simulations of the corresponding stochastic master equation (QuTiP) in the interaction picture for a few thousand trajectories for the cases $\alpha^2 = 1, \alpha^2 = 3, \alpha^2 = 5$ (Going to larger α^2 becomes rapidly more difficult since a larger fock space is needed and more trajectories are needed to get resolve the infidelities). The stochastic master equation simulations include $\kappa_1/2\pi = 1$ KHz, $\kappa_\phi = 2.5\kappa_1$, and $n_{th} = .01$ to show their negligible effect. The colored lines correspond to the analytic formula for the separation error Eq. (G30). The analytic curves do not include loss mechanisms. Each color corresponds to an integer value of α^2 from 1 to 10. The simulated and analytic curves agree well. b.) Plot of the minimum infidelities vs. α^2 and the fit line. We used a more conservative relation $\epsilon = e^{-1.5-.9|\alpha|^2}$ for the error correction simulations.

dephasing. Calculating the relevant integral we get

$$\frac{\alpha_0}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} d\theta e^{-\frac{\theta^2}{2\sigma^2}}\cos\theta = \alpha_0 e^{\frac{-\kappa_\phi t}{2}}. \tag{G31}$$

Thus the effect of dephasing with a fixed homodyne axis is to modify the single phonon loss $\kappa_a \longrightarrow \kappa_a + \kappa_\phi$. In general the effect of the single phonon loss and dephasing on the fidelity of the readout is minor.

From these equations we can then determine the fidelity as a function of time for different alpha for our measurement scheme. The chosen parameters for this work are $\kappa_b/2\pi = 20$ MHz and $g/2\pi = 4$ MHz. In the future we will numerically confirm the coupling in the context of the effective Hamiltonian theory. Note also that κ_b can be made larger with the main effect of lengthening the readout time. We have simulated trajectories for this measurement procedure using a stochastic master equation for confirmation. The integrated and classified and measurement results from the stochastic master equation compared to the analytic expression are pictured in Fig. 42. We see exponential suppression of the error probability with increasing α^2. In the future we expect to be able to improve the performance by optimizing the window function for the readout and using the confidence of the measurement result to feed back and improve the matching. These advances in addition to the robustness of the EC to larger measurement errors than those currently assumed would allow us to make looser assumptions.

Appendix H: STOP algorithm

When performing physical non-Clifford operations in between rounds of error correction (EC), in order to maintain the full effective code distance, it is crucial to use a fault-tolerant error correction protocol which satisfies the following definition (taken from [101, 102]):

Definition 1. *Fault-tolerant error correction*

For $t = \lfloor (d-1)/2 \rfloor$, an error correction protocol using a distance-d stabilizer code C is t-fault-tolerant if the following two conditions are satisfied:

1. *For an input codeword with error of weight s_1, if s_2 faults occur during the protocol with $s_1 + s_2 \leq t$, ideally decoding the output state gives the same codeword as ideally decoding the input state.*

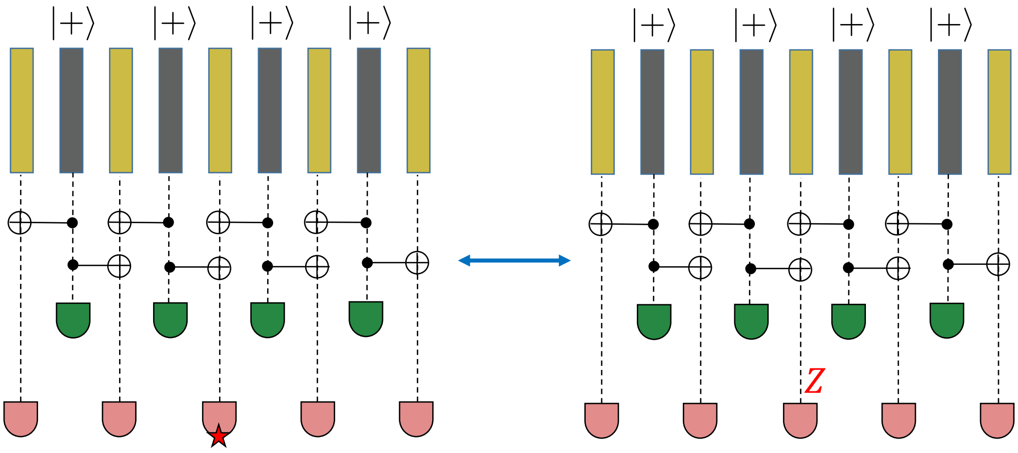


FIG. 43. Distance-5 repetition code, where one round of stabilizer measurements is performed (red measurements) followed by a direct measurement of the data qubits (green measurements). The data qubits are the red rectangles, and the ancilla qubits (prepared in $|+\rangle$) are the yellow rectangles. On the left, a measurement error on the third data qubit occurs during the direct measurement of the data, which is equivalent to having a Z data qubit error immediately before the measurement (shown on the right). Such settings illustrate the importance of the `STOP` algorithm, where one might have to correct errors prior to applying a non-Clifford gate, and a round of perfect error correction (which in practice is achieved by directly measuring the data) cannot be performed. In such settings, a single measurement error during the last round of stabilizer measurements (red measurements in the figure) can lead to a logical failure if the syndrome measurement is repeated a fixed number of times (say d) rather than using the `STOP` algorithm.

2. *For s faults during the protocol with $s \leq t$, no matter how many errors are present in the input state, the output state differs from a codeword by an error of at most weight s.*

Apart from being useful for proving thresholds of fault-tolerant error correction schemes based on code concatenation [103], such a definition of fault-tolerant error correction is also relevant when performing physical non-Clifford operations on encoded qubits *before* directly measuring the data qubits. In particular, if one implements a minimum weight perfect matching (MWPM) decoder (see Ref. [104]) with $\mathcal{O}(d)$ rounds of stabilizer measurements (where d is the code distance of the error correcting code protecting the data), a measurement error in the last round can lead to a logical failure and Definition 1 would not be satisfied. In many fault-tolerant implementations, such a problem can be avoided by implementing non-Clifford operations via magic state injection and stabilizer operations, followed by direct measurement of the data qubits (hence physical non-Clifford gates are never directly applied to the data qubits). An example of direct measurements of the data qubits after performing one round of statbilizer measurements for a $d = 5$ repetition code is given in Fig. 43. By measuring the data qubits, measurement errors can be treated as data qubit errors arising prior to performing the measurement [62]. As such, measuring the data directly acts as a round of perfect error correction.

As was shown in Section VI, post-selection can be avoided when preparing the logical $|0\rangle_L$ and $|1\rangle_L$ states (used to obtain the state $|\psi\rangle_1$) if we have a decoder that is robust to measurement errors in the last syndrome measurement round prior to applying the physical Toffoli gates. For the `BUTOF` protocol, we cannot directly measure the data prior to applying the physical Toffoli gates. Using ideas from Ref. [102], in this section we propose an algorithm which tells us when to terminate the sequence of error syndrome measurements, which we call the `STOP` algorithm, and which satisfies Definition 1 when using the syndrome measurement from the last round to correct errors. Further, in Appendix I, we show how the `STOP` algorithm can be used with magic state injection to perform all stabilizer operations of the repetition code.

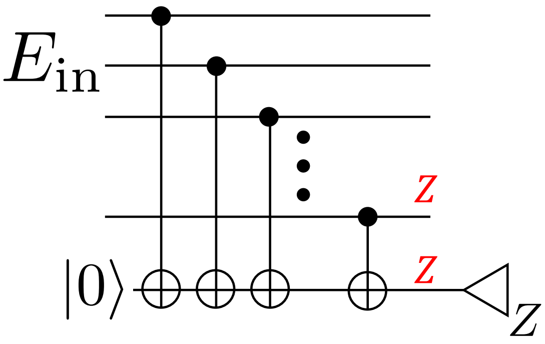


FIG. 44. Example of a single controlled-Z failure resulting in the error $Z_n \otimes Z_{n+1}$ (where Z_{n+1} acts on the ancilla qubit) when measuring the operator $Z^{\otimes n}$. Here n is the number of data qubits. This single fault can cause three consecutive syndrome measurements to yield three distinct outcomes. Here E_{in} is an input error with syndrome $s(E_{\text{in}}) = s_1$.

The goal of the `STOP` algorithm is to track consecutive syndrome outcomes $s_1, s_2, \cdots, s_r$ and to compute the *minimum* number of faults which could have caused this sequence of syndromes. In particular, let n_{diff} be a counter which tracks the minimum number of faults causing changes in syndrome outcomes, and consider the consecutive syndromes s_{k-1}, s_k and s_{k+1}. Given that a single fault can lead to two syndrome changes as in the example below, suppose we obtain different syndromes in rounds k and $k+1$ (so that $s_k \neq s_{k+1}$). In order to decide whether to increment n_{diff} by one, we must first check whether n_{diff} was incremented after measuring the k'th error syndrome. If n_{diff} didn't increase after the k'th round, then we increment n_{diff} by one. Otherwise, n_{diff} remains unchanged.

As an example, suppose a single fault occurs during the second round of stabilizer measurements of an EC protocol adding a weight-one error to the data qubits while also flipping the measurement outcome of one of the stabilizers (in this case $Z^{\otimes n}$ as shown in Fig. 44). Further, suppose the input error to the second round of the EC protocol E_{in} has the error syndrome $s(E_{\text{in}}) = s_1$, and that the error $Z_n E_{\text{in}}$ has error syndrome $s_3 \neq s_1$ (here Z_n is the Z error added to the data qubit arising from the two-qubit gate failure). Since the Z error flipped the measurement outcome of $Z^{\otimes n}$, the syndrome s_2 measured during the second round can differ from both s_1 and s_3.

With the above example in mind, the `STOP` algorithm is given by Algorithm 1. To see why a decoding algorithm based on Algorithm 1 satisfies Definition 1, consider the case where the total number of input errors and faults during the EC is $t = (d-1)/2$ for a distance d error correcting code. If at any time during the EC the same syndrome s_j is measured $t - n_{\text{diff}} + 1$ times in a row, then it must have been the correct syndrome (with very high probability). The reason is that given the value of n_{diff}, which counts the minimum number of faults compatible with the syndrome history since the beginning of the current cycle of error correction, there would need to be more than t faults to cause all $t - n_{\text{diff}} + 1$ consecutive syndromes to be incorrect due to failures resulting in flipped measurement outcomes. As such one could use the syndrome s_j to correct errors and terminate the protocol. Doing so, there could only be $\leq t$ residual leftover errors that went undetected in the last measurement round.

Similarly, if after measuring the $r-1$'th syndrome n_{diff} gets incremented to $n_{\text{diff}} = t$, then we know that at least t faults must have occurred during the EC. As such, by repeating the syndrome measurement one more time (resulting in the syndrome s_r) and using that syndrome to decode, there would need to have been more than t faults for s_r to be the wrong syndrome (due to faults flipping some of the stabilizer measurement outcomes in the last round). Hence using s_r to decode would result in residual errors with weight $v \leq t$ (where, as stated at the beginning of the previous paragraph, the total number of input errors and faults during the EC is t).

Algorithm 1: `STOP` algorithm

```
Result: Final syndrome $s_r$ for $r$ repetitions of the
        syndrome measurement.
initialize: $t = (d-1)/2$; $n_{\text{diff}} = 0$; countSyn = 1;
  SynRep = 1; $n_{\text{diff}}$Increase = 0; test = 0;
while $test = 0$ do
    if $n_{diff} = t$ then
        test = 1
    end
    Measure the error syndrome $s_j$. Store the
      error syndrome $s_{j-1}$ from the previous round
      in synPreviousRound and the current
      syndrome $s_j$ in synCurrentRound.;
    if $countSyn > 1$ then
        if $synPreviousRound = synCurrentRound$
         then
            SynRep = SynRep + 1;
            $n_{\text{diff}}$Increase = 0;
        else
            SynRep = 0;
            if $n_{diff}Increase = 0$ then
                $n_{\text{diff}} = n_{\text{diff}} + 1$;
                $n_{\text{diff}}$Increase = 1;
            else
                $n_{\text{diff}}$Increase = 0;
            end
        end
    end
    if $SynRep = t - n_{diff} + 1$ then
        test = 1;
    end
    countSyn = countSyn + 1;
end
```

Given the above, we conclude that when using Algorithm 1, the sequence of syndrome measurements will terminate if one of the following conditions is satisfied:

1. The syndrome s_r is repeated $t - n_{\text{diff}} + 1$ times in a row.

2. The counter n_{diff} gets incremented to $n_{\text{diff}} = t$. Measure the syndrome one more time resulting in the syndrome s_r. Use s_r to decode.

Decoding will succeed if the total number of input errors and faults during the EC cycle is $\leq t$.

We now provide a few remarks. Firstly, given a particular error correcting code and decoder along with the `STOP` algorithm for repeating the syndrome measurement, one can satisfy Definition 1 by using the last measured syndrome s_r to decode while ignoring the entire syndrome history. Hence in such settings, one can use a simple code-capacity-type decoder to decode with s_r (i.e. a decoder which ignores measurement and space-time correlated errors). As an example, one can decode with the surface code using a MWPM or Union Find decoder (see Ref. [105]) on a two-dimensional graph instead of a three-dimensional graph tracking the entire syndrome history. Doing so could significantly reduce the overall decoding time. In general however, the approach where the STOP algorithm is used to ignore the entire syndrome history apart the last syndrome s_r does not have a threshold [106]. To see this, consider a distance d repetition code and a stochastic noise model where fault locations fail with probability p. After d rounds of repeating the syndrome measurement, there will be approximately pd^2 failures. For large distances d, with high probability the error syndrome will change in every round. Hence the probability of a measurement error in the final round will not depend on the past syndrome history and the decoder will fail to correct the errors with high probability.

On the other hand, tracking the entire syndrome history and using Algorithm 1 to decide when to terminate the rounds of repeated syndrome measurement generally leads to lower failure rates and has a threshold. Indeed, when computing the memory failure rates of the repetition code using Algorithm 1 for deciding when to terminate the syndrome measurements), we found that performing MWPM on the entire syndrome history leads to lower logical failure rates compared to performing MWPM on a one-dimensional graph using only the final syndrome s_r (note that the logical Z failure rates for the repetition code in Fig. 6 were computed by applying MWPM to the full syndrome history of the measured syndromes using the STOP algorithm). As such, the EC protocols used in this work when considering repetition codes implement MWPM on the entire syndrome history in conjunction with Algorithm 1 to decide when to stop measuring the error syndrome.

We conclude this section by providing a lower and upper bound on the maximum number of syndrome measurement repetitions that can be performed using the STOP algorithm. For the case where there are no failures, it is straightforward to see that the syndrome measurement will be repeated $t+1$ times. To find the upper bound, we consider the worst case scenario, where (starting with $n_{\text{diff}} = 0$) there are no failures in the first t rounds of syndrome measurement, so that the same syndrome is repeated t times. However, in round $t+1$, a measurement error occurs and n_{diff} gets incremented to $n_{\text{diff}} = 1$. Now again, suppose there are no failures in the next $t-1$ rounds (so the same syndrome is repeated $t-1$ times) and a measurement error occurs in the t'th round, so that n_{diff} is incremented to $n_{\text{diff}} = 2$. Suppose the same pattern repeats itself until all t faults are exhausted resulting in $n_{\text{diff}} = t$. By the protocol of the STOP algorithm, we must repeat the syndrome measurement one more time. For such a fault pattern, the total number of syndrome measurements s_{tot} is then given by

$$s_{\text{tot}} = \sum_{k=0}^{t-1}(t-k) + t + 1 = \frac{1}{2}(t^2+3t+2) = \binom{t+2}{2}. \tag{H1}$$

For low code distances and low noise rate regimes, the average number of repetitions will approach $t+1$. However for large code distances, with high probability, the syndrome measurement outcome will change every round and thus n_{diff} changes every other round. Thus after $2t$ rounds, $n_{\text{diff}} = t$ and the syndrome must be repeated one more time resulting in a total number of $2t+1 = d$ rounds. It should then be expected that for large code distances, the performance of MWPM decoders based on a fixed d rounds will perform similarly to a decoder which uses the STOP algorithm to terminate while implementing MWPM over the full syndrome history.

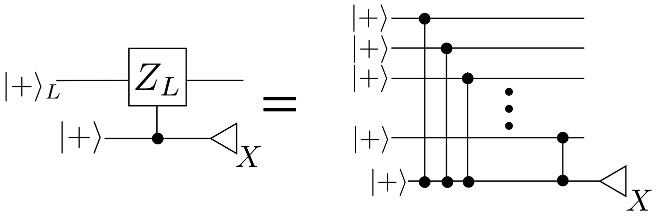


FIG. 45. Circuit for preparing logical computational basis states of the repetition code. A measurement error results in a logical X error applied to the data. Fault-tolerance is achieved by repeating the measurement using the STOP algorithm.

Appendix I: Stabilizer operations with the repetition code

In this section we describe how to do all stabilizer operations with the repetition code. However, the methods we provide apply to any family of Calderbank-Shor-Steane (CSS) codes.

1. Computational basis states

We begin by describing how to prepare the logical computational basis states of the repetition code. Doing so, we provide two schemes for preparing $|0\rangle_L$.

Scheme 1: Using the fact that for an n-qubit repetition code $|+\rangle_L = |+\rangle^{\otimes n}$, preparing $|+\rangle^{\otimes n}$ followed by a logical $Z_L = Z^{\otimes n}$ measurement (see Fig. 45) projects the state to $|0\rangle_L$ given a $+1$ outcome and $|1\rangle_L$ given a -1 outcome. Since a measurement error on the ancilla results in a logical $X_L = X_1$ error applied to the data, fault-tolerance can be achieved by repeating the measurement of Z_L using the STOP algorithm (where the syndrome corresponds to the ancilla measurement outcome) and applying the appropriate X_L correction given the final measurement outcome. For instance, if $|0\rangle_L$ is the desired state and the final measurement outcome at the termination of the STOP algorithm is -1, X_1 would be applied to the data. Lastly, note that only X errors can propagate from the ancilla to the data but these are exponentially suppressed by the cat-qubits.

Scheme 2: Here we present a more conventional approach for preparing the computational basis states which only involves stabilizer measurements (see for instance Refs. [107–109]). Starting with the state $|\psi_1\rangle = |0\rangle^{\otimes n}$

which is a +1 eigenstate of Z_L, measure all stabilizers of the repetition code (each having a random ± 1 outcome) resulting in the state

$$|\psi_2\rangle = \prod_{i=1}^{n-1}\Big(\frac{I \pm X_i X_{i+1}}{2}\Big)|0\rangle^{\otimes n}. \tag{I1}$$

If the measurement outcome of $X_k X_{k+1}$ is -1, the correction $\prod_{j=1}^{k} Z_j$ can be applied to the data to flip the sign back to +1. However given the possibility of measurement errors, the measurement of all stabilizers $\langle X_1X_2, X_2X_3, \cdots, X_{n-1}X_n\rangle$ must be repeated. If physical non-Clifford gates are applied prior to measuring the data, then the STOP algorithm can be used to determine when to stop measuring the syndrome. Subsequently, MWPM is applied to the full syndrome history to correct errors and apply the appropriate Z corrections to fix the code-space given the initial stabilizer measurements. After performing numerical simulations, we found that **scheme 2** achieves lower logical failure rates compared to **scheme 1**. Further, since physical Toffoli gates are applied to the data qubits in order to prepare a $|\text{TOF}\rangle$ magic state (see Section VI) and given the constraints imposed by our ATS architecture (which make performing global Z measurements very challenging using a single ancilla qubit), we always use **scheme 2** along with the STOP algorithm when preparing logical computational basis states.

Lastly, we remark that although the logical component of an uncorrectable error $E^{(z)}Z_L$ (where $E^{(z)}$ is correctable) can always be absorbed by $|0\rangle_L$ resulting in an output state $|\psi_{\text{out}}\rangle = E^{(z)}|0\rangle_L$, it is still important to have a fault-tolerant preparation scheme for $|0\rangle_L$ (and thus to repeat the measurement of all stabilizers enough times). For instance, if a single fault results in a weight-two correctable Z error (assuming $n \geq 5$), a second failure adding one or more data qubit errors during a subsequent part of the computation can combine with the weight-two error resulting in an uncorrectable data qubit error. Hence, such a preparation protocol would not be fault-tolerant up to the full code distance.

2. Implementation of logical Clifford gates

Since the CNOT gate is transversal for the repetition code, we focus on implementing a generating set of single-qubit Clifford operations. Recall that the Clifford group is generated by $\mathcal{P}_n^{(2)} = \langle H_i, S_i, \text{CNOT}_{ij}\rangle$, where

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \; S = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}, \tag{I2}$$

are the Hadamard and phase gate operators. In what follows we show how to implement S and $Q = SHS$ which also forms a generating set for single-qubit Clifford operations. A key to the implementation of such gates will be the injection of the state $|i\rangle = \frac{1}{\sqrt{2}}(|0\rangle + i|1\rangle)$, which is a +1 eigenstate of the Pauli Y operator. The logical state $|i\rangle_L$ can be prepared using **scheme 1** of Appendix I 1 by replacing Z_L with $Y_L = Y_1 Z_2 \cdots Z_n$.

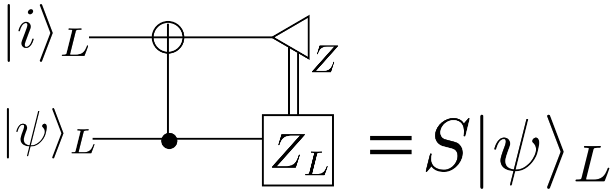


FIG. 46. Circuit for implementing a logical S gate. The circuit requires the preparation of $|i\rangle_L$, and the CNOT gate is transversal. A logical $|Z\rangle_L$ operator is applied when the measurement outcome of the ancilla is -1.

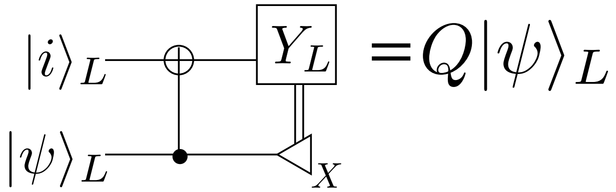


FIG. 47. Circuit for implementing a logical $Q = SHS$ gate. The circuit requires the preparation of $|i\rangle_L$, and the CNOT gate is transversal. A logical $|Y\rangle_L$ operator is applied when the measurement outcome of the ancilla is -1.

In Fig. 46 we provide a circuit for implementing S_L which requires $|i\rangle_L$ as an input state, a transversal CNOT gate, and a logical Z-basis measurement. If a -1 measurement outcome is obtained, we apply a Z_L correction to the data. Note however that a measurement error can result in a logical Z_L being applied incorrectly to the data. As such, to guarantee fault-tolerance, one can repeat the circuit of Fig. 46 and use the STOP algorithm to decide when to terminate. The final measurement outcome is then used to determine if a Z_L correction is necessary. The implementation of S_L can thus be summarized as follows:

S_L **gate implementation:**

1. Implement the circuit in Fig. 46 and let the measurement outcome be s_1.

2. Repeat the circuit in Fig. 46 and use the STOP

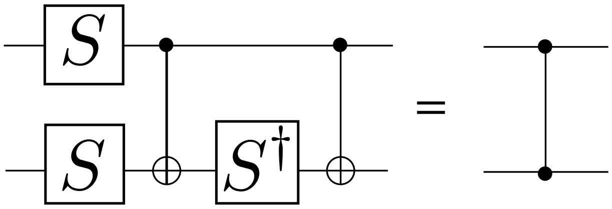


FIG. 48. Efficient circuit for implementing a CZ gate given the higher cost of logical H gates compared to logical S gates.

algorithm to decide when to terminate.

3. If the final measurement outcome $s_r = +1$, do nothing, otherwise apply $Z_L = Z_1 Z_2 \cdots Z_n$ to the data.

The circuit for implementing the logical $Q = SHS$ gate is given in Fig. 47. The circuit consists of an injected $|i\rangle_L$ state, a transversal CNOT gate and a logical X-basis measurement is applied to the input data qubits. If the measurement outcome is -1, Y_L is applied to the data. As with the S_L gate, we repeat the application of the circuit in Fig. 47 to protect against measurement errors. The full implementation of Q_L is given as follows:

Q_L gate implementation:

1. Implement the circuit in Fig. 47 and let the measurement outcome be s_1.

2. Repeat the circuit in Fig. 47 and use the `STOP` algorithm to decide when to terminate.

3. If the final measurement outcome $s_r = +1$, do nothing, otherwise apply $Y_L = Y_1 Z_2 \cdots Z_n$ to the data.

Note that the logical Hadamard gate can be obtained from the S_L and Q_L protocols using the identity $H = S^\dagger SHSS^\dagger = S^\dagger QS^\dagger$. Hence ignoring repetitions of the circuits in Figs. 46 and 47, the implementation of H_L requires three logical CNOT gates, two $|-i\rangle_L$ and one $|i\rangle_L$ state, two logical Z basis measurements, and one logical X basis measurement. Instead of using two logical Hadamard gates and one CNOT gate to obtain a CZ gate, we provide a more efficient circuit in Fig. 48.

Lastly, we point out that since the circuits in Figs. 46 and 47 contain only stabilizer operations and injected $|i\rangle_L$ states, using the `STOP` algorithm to repeat the measurements is not strictly necessary. For instance, one could repeat the measurement a fixed number of times and majority vote instead of using the `STOP` algorithm. However in low noise rate regimes, the `STOP` algorithm can potentially be much more efficient since the average number of repetitions for the measurements can approach $t+1$ where $t = (d-1)/2$.

Appendix J: Growing encoded data qubits to larger code distances with the repetition code

In this section, we provide a simple protocol for growing a state $|\overline{\psi}\rangle_{d_1} = \alpha|0\rangle_{d_1} + \beta|1\rangle_{d_1}$ encoded in a distance d_1 repetition code to a state $|\overline{\psi}\rangle_{d_2} = \alpha|0\rangle_{d_2} + \beta|1\rangle_{d_2}$ encoded in a distance $d_2 > d_1$ repetition code. We emphasize that the protocol presented in this section is applicable to arbitrary states and will be used for growing $|\text{TOF}\rangle$ states prepared using the fault-tolerant methods of Section VI to larger code distances.

Let $\mathcal{S}_{d_1} = \langle X_1X_2, X_2X_3, \cdots, X_{d_1-1}X_{d_1}\rangle$ be the stabilizer group for the distance d_1 repetition code with cardinality $|\mathcal{S}_{d_1}| = d_1 - 1$. Similarly, we define $\mathcal{S}_{d'_1} = \langle X_{d_1+1}X_{d_1+2}, \cdots, X_{d_2-1}X_{d_2}\rangle$ with $|\mathcal{S}_{d'_1}| = d_2 - d_1 - 1$. Furthermore, the stabilizer group for the distance d_2 repetition code is given by $\mathcal{S}_{d_2} = \langle X_1X_2, X_2X_3, \cdots, X_{d_2-1}X_{d_2}\rangle$.

In the remainder of this section, we define $g_i^{(d_1)}$ to be the i'th stabilizer in $\mathcal{S}_{d_1}$ and $g_i^{(d'_1)}$ to be the i'th stabilizer in $\mathcal{S}_{d'_1}$, so that $g_i^{(d_1)} = X_iX_{i+1}$ and $g_i^{(d'_1)} = X_{d_1+i}X_{d_1+i+1}$.

Protocol for growing $|\overline{\psi}\rangle_{d_1}$ to $|\overline{\psi}\rangle_{d_2}$:

1. Prepare the state $|\psi_1\rangle = |0\rangle^{\otimes(d_2-d_1)}$.

2. Measure all stabilizers in $\mathcal{S}_{d'_1}$ resulting in the state $|\psi_2\rangle_{d'_1} = \prod_{i=1}^{d_2-d_1-1}\left(\frac{I \pm g_i^{(d'_1)}}{2}\right)|0\rangle^{\otimes(d_2-d_1)}$.

3. Repeat the measurement of stabilizers in $\mathcal{S}_{d'_1}$ and apply MWPM to the syndrome history to correct errors and project to the code-space. If $g_i^{(d'_1)}$ is measured as -1 in the first round, apply the correction $\prod_{k=d_1+1}^{d_1+i} Z_k$ to the data.

4. Prepare the state $|\psi_3\rangle = |\overline{\psi}\rangle_{d_1} \otimes |\psi_2\rangle_{d'_1}$ and measure $X_{d_1}X_{d_1+1}$.

5. Repeat the measurement of all stabilizers if $\mathcal{S}_{d_2}$ and use MWPM over the syndrome history to correct errors. If in the first round the stabilizer $X_{d_1}X_{d_1+1}$ is measured as -1, apply the correction $\prod_{i=1}^{d_1} Z_i$.

As remark, the corrections stated in step 3 and 5 can be postponed to a later time after the growing protocol is completed. The reason is that one can use the entire syndrome history from each step, in addition to the syndromes measured after the states have merged to apply the appropriate corrections.

The growing scheme involves two blocks, the first being the state $|\overline{\psi}\rangle_{d_1}$ which we want to grow to $|\overline{\psi}\rangle_{d_2}$. The second block involves the set of qubits which are prepared in the state $|\psi_2\rangle_{d'_1}$ and stabilized by $\mathcal{S}_{d'_1}$ (steps 1-3). The key is to measure the boundary operator $X_{d_1}X_{d_1+1}$ between the two blocks, which effectively merges both blocks into the encoded state $|\overline{\psi}\rangle_{d_2}$ and constitutes a simple implementation of lattice surgery [63, 65, 110, 111]. To see this, consider the state prior to step 4:

$$\begin{aligned}
|\psi_3\rangle &= |\overline{\psi}\rangle_{d_1} \otimes |\psi_2\rangle_{d'_1} \\
&= \alpha|0\rangle_{d_1} \otimes |\psi_2\rangle_{d'_1} + \beta|1\rangle_{d_1} \otimes |\psi_2\rangle_{d'_1} \\
&= \alpha \prod_{i=d_1+1}^{d_2-1}\Big(\frac{I + g_i^{(d'_1)}}{2}\Big)|0\rangle_{d_1} \otimes |0\rangle^{\otimes(d_2-d_1)} \\
&+ \beta X_1 \prod_{i=d_1+1}^{d_2-1}\Big(\frac{I + g_i^{(d'_1)}}{2}\Big)|0\rangle_{d_1} \otimes |0\rangle^{\otimes(d_2-d_1)}, \quad \text{(J1)}
\end{aligned}$$

where we used $|1\rangle_{d_1} = X_1|0\rangle_{d_1}$. When measuring $X_{d_1}X_{d_1+1}$ and performing the correction $\prod_{i=1}^{d_1} Z_i$ if the measurement outcome is -1, $|\psi\rangle_3$ is projected to

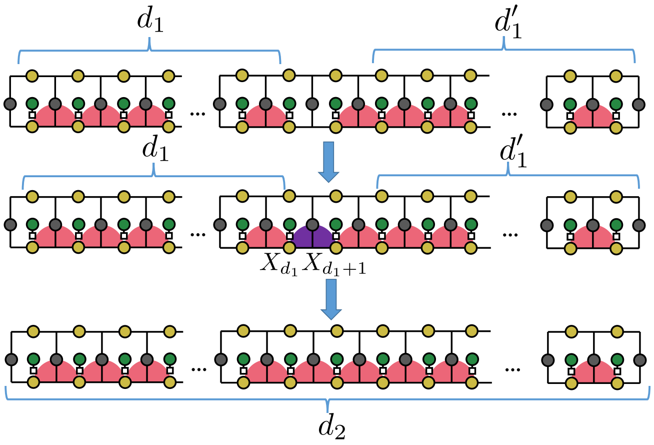


FIG. 49. Diagram illustrating our protocol for growing the state $|\psi\rangle_{d_1}$ to $|\psi\rangle_{d_2}$ with the ATS layout by starting with the two blocks stabilized by $\mathcal{S}_{d_1}$ and $\mathcal{S}_{d_1'}$. The yellow vertices are the data qubits, and the gray vertices correspond to the ancilla qubits used to measure the stabilizers of the repetition code. The measurement of $X_{d_1}X_{d_1+1}$ (with random ±1 outcome) is highlighted by the purple semi-circle. After performing the appropriate corrections, the final block is stabilized by $\mathcal{S}_{d_2}$.

$$\begin{aligned}|\psi\rangle_f &= \alpha \prod_{i=d_1+1}^{d_2-1}\Big(\frac{I+g_i^{(d_1')}}{2}\Big)\Big(\frac{I+X_{d_1}X_{d_1+1}}{2}\Big)\prod_{j=1}^{d_1-1}\Big(\frac{I+g_j^{(d_1)}}{2}\Big)|0\rangle^{\otimes d_2}\\ &+\beta X_1 \prod_{i=d_1+1}^{d_2-1}\Big(\frac{I+g_i^{(d_1')}}{2}\Big)\Big(\frac{I+X_{d_1}X_{d_1+1}}{2}\Big)\prod_{j=1}^{d_1-1}\Big(\frac{I+g_j^{(d_1)}}{2}\Big)|0\rangle^{\otimes d_2}\\ &=\alpha\prod_{i=1}^{d_2-1}\Big(\frac{I+X_iX_{i+1}}{2}\Big)|0\rangle^{\otimes d_2}\\ &+\beta X_1\prod_{i=1}^{d_2-1}\Big(\frac{I+X_iX_{i+1}}{2}\Big)|0\rangle^{\otimes d_2}\\ &=\alpha|0\rangle_{d_2}+\beta X_1|0\rangle_{d_2}\\ &=|\psi\rangle_{d_2}.\end{aligned} \tag{J2}$$

The rounds of repeated stabilizer measurements in steps 3 and 5 are required due to the random outcomes and measurement errors which can occur when performing the appropriate projections. A pictorial representation for the growing scheme is shown in Fig. 49.

Appendix K: Toffoli simulation twirling approximation

To simulate the fault-tolerant preparation of the $|\mathrm{TOF}\rangle$ state taking into account all fault-locations, we implement Monte-Carlo methods using a Gottesman-Knill type simulation [112] to avoid running into scalability issues. However, since the circuit in Fig. 15b contains physical Toffoli gates, some type of approximation is necessary to perform a Gottesman-Knill type simulation with such circuits.

In order to determine the most appropriate type of approximation, writing a Toffoli gate as CCX, we first observe that

$$(\mathrm{CCX})(I\otimes I\otimes Z)|\psi\rangle = (CZ_{A,B}\otimes Z)(\mathrm{CCX})|\psi\rangle, \tag{K1}$$

for some arbitrary state $|\psi\rangle$. In other words, propagating a Z error through the target qubit of the Toffoli gate results in a CZ error on the two control qubits. Recall that we label the three logical qubits by $\{A,B,C\}$.

In what follows, we will consider the transformation of the $|\mathrm{TOF}\rangle$ state with input data qubit Z errors on the third block when using a single $|+\rangle$ ancilla to measure g_A. Note that all conclusions remain unchanged if instead we used the GHZ state of Fig. 15b.

Let A_k be a subset of k qubits and consider $k \geq 1$ data qubit errors on the third block expressed as $E^{(C)} = I \otimes I \otimes \prod_{j \in A_k} Z_j \equiv \prod_{j=1}^{k} Z_j^{(C)}$. We have that

$$|\psi\rangle_{\mathrm{in}} = |+\rangle \prod_{j=1}^{k} Z_j^{(C)} |\mathrm{TOF}\rangle. \tag{K2}$$

After applying g_A and propagating the Z errors through the Toffoli gates, $|\psi\rangle_{\mathrm{in}}$ becomes

$$\begin{aligned}
&\prod_{j=1}^{k} Z_j^{(C)} \Big(|0\rangle|\mathrm{TOF}\rangle + |1\rangle \prod_{j=1}^{k} Z_j^{(A)} |\mathrm{TOF}\rangle \Big) \\
&= \prod_{j=1}^{k} Z_j^{(C)} \Big[|+\rangle \Big(\frac{I + \prod_{j=1}^{k} Z_j^{(A)}}{\sqrt{2}} \Big) |\mathrm{TOF}\rangle \\
&+ |-\rangle \Big(\frac{I - \prod_{j=1}^{k} Z_j^{(A)}}{\sqrt{2}} \Big) \Big],
\end{aligned} \tag{K3}$$

where $\prod_{j=1}^{k} Z_j^{(A)}$ are products of Z errors on the first data block which have identical support with the Z errors on the third block. After measuring the ancilla in the X basis, a ± 1 measurement outcome results in the state $|\psi\rangle_f$ given by

$$|\psi\rangle_f = \prod_{j=1}^{k} Z_j^{(C)} \Big(\frac{I \pm \prod_{j=1}^{k} Z_j^{(A)}}{\sqrt{2}} \Big) |\mathrm{TOF}\rangle. \tag{K4}$$

From Eq. (K4), we see that when performing one round of error detection of the first block A, the error $\Big(\frac{I \pm \prod_{j=1}^{k} Z_j^{(A)}}{\sqrt{2}} \Big)$ will project either to I or $\prod_{j=1}^{k} Z_j^{(A)}$ with 50% probability each unless $\prod_{j=1}^{k} Z_j^{(A)} = Z_L^{(A)}$ in which case the state remains unchanged.

Given the above, when performing our Gottesman-Knill type simulations when measuring g_A, if the input Z errors to the third block are $\prod_{j=1}^{k} Z_j^{(C)}$, we flip the GHZ ancilla measurement outcome with 50% probability and do the following: If $k < d$, we add the Z errors $\prod_{j=1}^{k} Z_j^{(A)}$ to the first block with 50% probability. If $k = d$, we add Z_L to the first block with 100% probability.

Note that such a simulation method is exact when $k < d$ and only introduces a discrepancy when $k = d$. Since such events are rare, our approximation method differs from an exact simulation of the bottom-up $|\mathrm{TOF}\rangle$ state preparation scheme only by a small amount.

Appendix L: Fitting procedure for memory and lattice surgery

Here we extend the discussion of lattice surgery presented in Section V as well as describe and justify the fitting procedures used in our error correction simulations. These fits enable us to reliably extrapolate to larger code sizes than simulated, which is required for our analysis of resource costs for large scale quantum computations (see Section VIII).

In addition, to presenting results for memory errors we also consider lattice surgery errors. Lattice surgery is the primary technique we consider for performing Clifford gates and magic state injection. It is a procedure for measuring multi-qubit logical Pauli operators such as $X_L^{\otimes m}$ with $m \geq 2$. It can be regarded as a code deformation where the m logical qubits are temporally merged into a code of $m-1$ logical qubits, and then split into their constituent m logical qubits. For the simple $m = 2$ case, we illustrate the space-time diagram for this process in Fig. 12 of the main text. Here we present a more detailed in Fig. 50.

An incredibly powerful and beautiful feature of lattice surgery is that decoding via matching naturally extends over this 3D spacetime structure without being interrupted by lattice surgery. However, some care is needed to correctly account for boundaries and assess different failure modes. For a planar surface code, it is well known that one must allow defects to match with the appropriate boundaries in the space direction. When performing lattice surgery, it is also important to match to appropriate boundaries in the time directions.

To understand boundary effects, consider the more detailed explanation of lattice surgery in Fig. 50. The procedure starts and ends with Z basis state preparations and measurements. A bit-flipped single qubit measurement or preparation will yield a pair of Z syndrome defects. That is, the initial and final rounds of Z stabilizer measurements are semi-ideal as they are reconstructed from single-qubit information so that any defects occur in pairs. These short X strings are then easily matched. In contrast, an X syndrome measurement error (at the start/end of lattice surgery) can lead to an isolated defect and is potentially harmful as it flips the outcome of the lattice surgery operation. However, for such an isolated defect near a time boundary, the best explanation is clearly an isolated measurement error. Therefore, we match these defects to red boundaries in the time direction.

As a warm-up to discussing the probability of time-like errors, we first recap the error scaling properties of memory and logical Z-errors. Consider a d_x by d_z surface code patch storing a logical qubit for t surface code cycles. We expect the total logical error probability to scale as $(1 - \exp(-\lambda t))/2$ for some constant rate λ, which for small lambda is approximately $\sim \lambda t/2$. Furthermore, as d_x increases the number of paths across the code increases linearly, so we expect that $\lambda \propto d_x$ and the total Z-logical

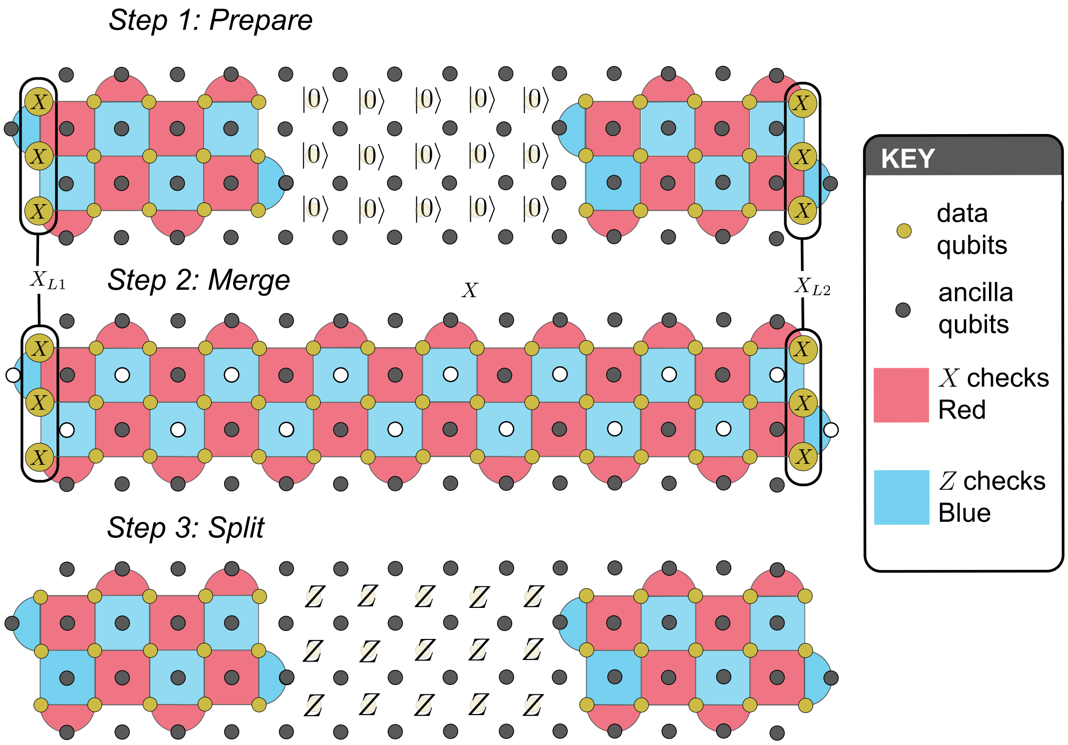


FIG. 50. The three stages of lattice surgery corresponding to cross sections (time slices) of the lattice surgery spacetime diagram in Fig. 12. *Step 1 Prepare:* data qubits between the surface code blocks are prepared in the $|0\rangle$ state. *Step 2 Merge:* start measuring the Z and X stabilizers indicated. The product of the X stabilizers (highlighted with white vertices) yields the outcome $X_{L1}X_{L2}$. However, a measurement error on a white vertex will flip the outcome and so these stabilizer measurements must be repeated d_m times, with d_m chosen sufficiently large to suppress time-like errors to the desired probability. *Step 3 Split:* The qubits between the initial surface code blocks are measured in the Z basis. Note that it is not possible to use X basis measurements to disentangle as this would measure X_{L1} and X_{L2}. If the parity of the single-qubit Z measurements is "-1" then we must apply a Pauli correction X_{L1} (or equivalently X_{L2}) as a correction. Both the measurement of $X_{L1}X_{L2}$ and the estimated Pauli correction but be done fault-tolerantly after having decoded the syndrome. In the case of $X_{L1}X_{L2}$, we choose 1 particular timeslice t_p and make an initial guess by multiplying all the white vertices at time t_p. If the decoder assigns a measurement error to any white vertex at time t_p, then we must account by flipping the $X_{L1}X_{L2}$. If the accumulated physical Z errors before time t_p anticommute with $X_{L1}X_{L2}$ then we flip the outcome. For a similar discussion of lattice surgery see Ref. [111]. Compared to Figs. 2 and 5 we use a similar graphical representation but for simplicity omit the location of the transmon, readout qubit and ATS.

probability to scale as

$$P_Z = d_x t F(d_z, p_1, \dots, p_k), \tag{L1}$$

for some function F of d_z and relevant hardware parameters $(p_1, \dots, p_k)$. Note that $d_x t$ corresponds to the area of the vertical red boundaries in Fig. 12. For fixed parameters $(d_z, p_1, \dots, p_m)$ the value of $F(d_z, p_1, \dots, p_m)$ can be estimated by Monte Carlo simulation and evaluating $P_Z/(d_x t)$. For simulation purposes, standard practice is to assume: at time zero, the system is in a "$+1$" eigenstates of all stabilizers; at time t, the round of stabilizer measurements is ideal. This assumption introduces a finite size effect error into $P_Z/(d_x t)$. This is suppressed by taking t large, and community folklore suggests that $t = \max[d_z, d_x]$ will suffice though one could push higher. The exact form of function F can be quite involved, though we know it will be exponentially suppressed by the relevant distance d_z. Taking our sole experimental parameter to be κ_1/κ_2 we find good fits of the form

$$P_Z = d_x t a_z (b_z \kappa_1/\kappa_2)^{c_z d_z}, \tag{L2}$$

where a_z, b_z, c_z are fitted parameters. For small d_x, there will be a finite size effect so the scaling is not linear in d_x. However, we can still use such a fit when d_x is held constant provided we do not attempt to extrapolate to larger d_x. Note that Eq. (L2) is not necessarily a leading order fit of the classical form $O(p^{(d-1)/2})$. Since the probability of logical failures has a entropic/combinatorial component, it is dominated by errors with a weight much larger than $(d-1)/2$. As such, we do not attempt a leading order fit but rather it is appropriate to fit the scaling exponent c_z.

We present the result of this fitting procedure in Fig. 51 and observe that it works well over the interval $10^{-5} \leq \kappa_1/\kappa_2 \leq 10^{-4}$. At higher values of κ_1/κ_2, higher order contributions to a logical Z failure become important. Similarly, at lower values of κ_1/κ_2, lower order contributions become important. Even if a more sophisticated fitting function of κ_1/κ_2 is assumed, we expect a finite range of applicability since there are other relevant experimental parameters in the noise model.

Similar reasoning can be applied to timelike errors. The relevant boundary has an area ℓd_x where ℓ is the distance between the codeblocks. As with Z-logical errors, the exponential decay of timelike errors follows from a percolation theory analysis [113, 114] of a strings connecting the timelike boundaries. As always in percolation problems, the probability of a percolation event is exponentially suppressed in the distance between the boundaries (whenever below some threshold). The relevant boundaries are separated by a distance d_m, which we call the *measurement distance*, and physically corresponds to the number of repeated rounds of syndrome measurements during the merge step. Therefore, we fit to the ansatz

$$P_M = \ell d_x a_z (b_m \kappa_1/\kappa_2)^{c_m d_m}, \tag{L3}$$

where a_m, b_m, c_m are fitted parameters. To obtain an estimate of P_M we simulate the middle group of qubits in Fig. 50. We wish to isolate the timelike errors and so freeze out Z-logical errors by assuming that the left-most and right-most qubits are ideal and error-free. This is analogous to the assumption of ideal measurements in a memory simulation. Furthermore, since the d_z distance is temporally extended during lattice surgery, such errors will be rare in comparison. Again, this idealization introduces a finite size effect that vanishes as ℓ grows relative to d_m.

We present the result of this fitting procedure for thin surface codes in Fig. 51 and observe that it works well over the interval $10^{-5} \leq \kappa_1/\kappa_2 \leq 5 * 10^{-4}$. We did not collect data for $\kappa_1/\kappa_2 \geq 5 * 10^{-4}$ as we had already identified that the surface code overhead will be prohibitively large in this regime.

To the best of our knowledge, there have not been previous simulations that investigate time-like errors in codes with boundaries and/or using circuit-level noise. For instance, timelike errors were accounted for by Raussendorf and Harrington [115] but using a toy, phenomenological noise model and periodic boundary conditions in both space and time.

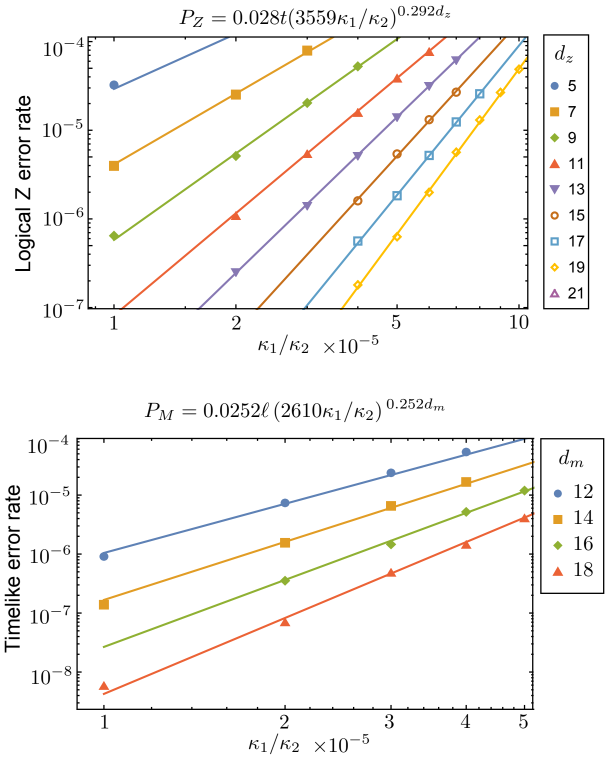


FIG. 51. Fitted results for simulation of $d_x = 3$ surface code for logical Z and timelike errors. We fit according to the anstaz of Eqs. (L2) and (L3). (Top) The logical Z simulations for which we set $t = d_z$ and plot the error probability divided by t. (Bottom) the probability of a timelike error during lattice surgery for which we set $\ell = d_m - 1$. All data points shown are used in fitting. This is a truncated data set eliminating points above 10^{-4} on the error rate axis and eliminating points outside the relevant range of κ_1/κ_2.

Widespread practice is to set $d_m = d_x = d_z$ but there is no *a prior* reason to believe this is optimal. Indeed, just as physical bias in X and Z noise leads to an asymmetry in our choice of d_x and d_z, a realistic noise model will influence the optimal choice of d_m. In later calculations we find that $d_m = d_z - 2$ is the most common optimal choice for the main algorithm. Furthermore, in the design of magic state distillation factories, the time-like errors are not critically important (see Table XII) and so inside the factory d_m can be set much smaller (by about a factor 1/2) than one would otherwise expect.

Appendix M: Edge weights and decoding graphs for the repetition and surface codes

In this section we provide the decoding graphs used to implement MWPM with the repetition and surface codes considered in this paper. We also provide details for computing the edge weights of all edges in a given graph.

1. Repetition code decoding graphs

The circuit for measuring the stabilizer of the $d = 5$ repetition code is shown in Fig. 52a and can straightforwardly be generalized to arbitrary code distances. The corresponding graph for decoding the $d = 5$ repetition

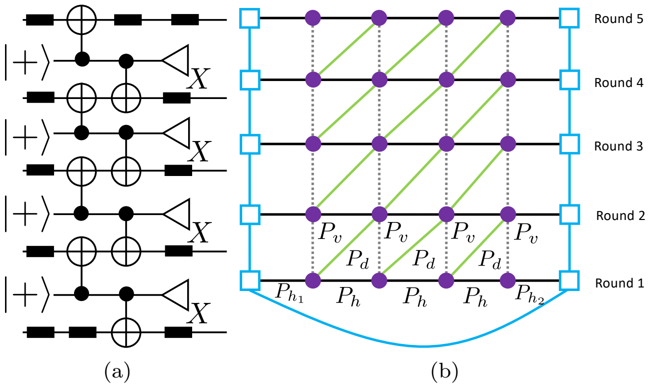


FIG. 52. (a) Circuit for measuring the stabilizers of the $d = 5$ repetition code. The dark rectangular boxes correspond to idling qubit locations. (b) MWPM decoding graph for the $d = 5$ repetition code where the syndrome measurement is repeated five times.

code using MWPM with five rounds of syndrome measurements is shown in Fig. 52b. The purple vertices correspond to the measurement outcome of each ancilla qubit (prepared in $|+\rangle$ and measured in the X-basis), and the horizontal edges correspond to the physical data qubits. A vertex is highlighted if the measurement outcome of the corresponding ancilla is non-trivial. We also add space-like boundary vertices and edges (shown in red). For a given syndrome measurement round (corresponding to a one-dimensional slice of the graph in Fig. 52b), a boundary vertex is highlighted if on odd number of bulk vertices in the corresponding one-dimensional slice are highlighted. To deal with measurement errors, dashed grey vertical edges are added and connect vertices of two one-dimensional graphs. Lastly, cross-diagonal edges (shown in green) are added to deal with space-time correlated errors arising from failures at CNOT gate locations (see below for explicit examples). More details for implementing graph based decoding using MWPM can be found in Refs. [70, 116, 117].

We now describe how to compute the edge weights for each edge type of the graph in Fig. 52b. For a given edge e, we must first compute the probability P_e of all failure events resulting in e being highlighted. The weight w_e for the edge e is then given by $w_e = -\log(P_e)$ (see for instance Refs. [87, 116, 118]). Such a prescription ensures that edges arising from more likely failure events are chosen with higher probability when finding the lowest weight path between two highlighted vertices. In what follows, we will refer to P_e as the edge weight probability for the edge e.

The first and last data qubits in Fig. 52a have an extra idling location compared to all other data qubits, and their edge weight probabilities are labeled P_{h_1} and P_{h_2}, whereas the other data qubits have edge weight probabilities labeled by P_h. The dashed grey vertical edges connecting are labeled P_v and the green space-time correlated egdes are labeled P_d. Next we define P_s to be the probability of a $|+\rangle$ state preparation error, P_m the probability of a measurement error, P_i the probability of an idle error and P_{Z_1}, P_{Z_2} and $P_{Z_1Z_2}$ the probability of a $Z \otimes I$, $I \otimes Z$ and $Z \otimes Z$ CNOT failure (where the first qubit is the control qubit of the CNOT).

We now show how to compute P_h and P_d to leading order (the other edge weight probabilities can be obtained using analogous methods). In the case of a single failure, a bulk horizontal edge (say corresponding to an error on the data qubit q_j) can be highlighted if either a Z error occurs at the idle location during the preparation of $|+\rangle$, a $Z \otimes Z$ failure on the CNOT gate at the second time step with q_j as a target qubit, or an $I \otimes Z$ failure on the CNOT gate on the third time step occuring in the previous syndrome measurement round. Hence we have

$$P_h^{(t_1)} = P_i(1 - P_{Z_1Z_2}) + P_{Z_1Z_2}(1 - P_i), \tag{M1}$$

and

$$\begin{aligned} P_h^{(t>t_1)} &= 2P_i(1 - P_{Z_1Z_2})(1 - P_{Z_2})(1 - P_i) \\ &+ P_{Z_1Z_2}(1 - P_i)^2(1 - P_{Z_2}) \\ &+ P_{Z_2}(1 - P_i)^2(1 - P_{Z_1Z_2}), \end{aligned} \tag{M2}$$

where $P_h^{(t_1)} = P_h$ in the first syndrome measurement round, and $P_h^{(t>t_1)} = P_h$ in all subsequent syndrome measurement rounds.

Now suppose a $Z \otimes Z$ error occurs on a CNOT gate in the third time step of the syndrome measurement round t resulting in a Z data qubit error on qubit q_j while also flipping the measurement outcome of the ancilla a_k. Note that if a Z error had occurred on qubit q_j prior to applying the two CNOT gates, both ancillas a_k and a_{k+1} would be measured non-trivially. Hence in round $t + 1$ (and assuming no other failures), the measurement outcome of a_k will not change whereas the measurement outcome of a_{k+1} will change. To ensure that such an event is treated to leading order, we introduce a green cross-diagonal edge as seen in Fig. 52b. Also note that a $I \otimes Z$ error on a CNOT in the second time step also results in such an edge. Hence we have that

$$P_d = P_{Z_1Z_2}(1 - P_{Z_2}) + P_{Z_2}(1 - P_{Z_1Z_2}). \tag{M3}$$

A similar analysis results in the following expressions for the remaining edge weight probabilities

$$\begin{aligned} P_v = P_m(1 - P_s)(1 - P_{Z_1})^2 + P_s(1 - P_m)(1 - P_{Z_1})^2 \\ + 2P_{Z_1}(1 - P_{Z_1})(1 - P_s)(1 - P_m), \end{aligned} \tag{M4}$$

$$P_{h_1}^{(t_1)} = P_h^{(t_1)}, \tag{M5}$$

$$P_{h_1}^{(t>t_1)} = 3P_i(1-P_i)^2(1-P_{Z_1Z_2})(1-P_{Z_2}) + P_{Z_1Z_2}(1-P_i)^3(1-P_{Z_2}) + P_{Z_2}(1-P_i)^3(1-P_{Z_1Z_2}), \tag{M6}$$

$$P_{h_2}^{(t_1)} = 2P_i(1-P_i)(1-P_{Z_1Z_2}) + P_{Z_1Z_2}(1-P_i)^2, \tag{M7}$$

and

$$P_{h_2}^{(t>t_1)} = P_{h_1}^{(t>t_1)}. \tag{M8}$$

2. Surface code decoding graphs

The two-dimensional graphs for decoding the X and Z stabilizer measurement outcomes of a $d_x = 5$ and $d_z = 7$ surface code, along with their corresponding edge weight probability labels, are shown in Fig. 53. We will show below the edges that need to be added when considering measurement errors and space-time correlated errors arising from CNOT gate failures. However, we first provide edge weight probabilities for the edges of the two-dimensional graphs.

Let $G^{(2D)}_{(d_x)}$ and $G^{(2D)}_{(d_z)}$ be the two-dimensional graphs corresponding to the X and Z stabilizer measurement outcomes. For the graph $G^{(2D)}_{(d_x)}$, we label the bulk edge weight probabilities by $P^{(2D)}_{BLTRX}$ and $P^{(2D)}_{TLBRX}$. All other labels in Fig. 53b are used for boundary edges. Similarly, for the graph $G^{(2D)}_{(d_z)}$, we label the bulk edge weight probabilities by $P^{(2D)}_{BLTRZ}$ and $P^{(2D)}_{TLBRZ}$ with all other labels in Fig. 53c representing boundary edge weight probabilities. In order to simplify the expressions for the edge weight probabilities, we define the following function

$$\Gamma(P_1, P_2, \cdots, P_j; n_1, n_2, \cdots, n_j) \equiv \sum_{k=1}^{j} n_k P_k (1-P_k)^{n_k-1} \prod_{i=1, i\neq k}^{j} (1-P_i)^{n_i}. \tag{M9}$$

In what follows, we define $P^{(P_iP_j)}_{\text{CNOT}}$ to be the probability that a CNOT gate failure results in a two-qubit Pauli error of the form $P_i \otimes P_j$. We also define $P^{(P_i)}_{\text{Id}}$ to be the probability that a single-qubit idling location results in a P_i Pauli error on that qubit. To further simplify the edge weight probability polynomials, we define the following probabilities:

$$P^{(1)}_{ZZCX} = P^{(ZZ)}_{\text{CNOT}} + P^{(ZY)}_{\text{CNOT}} + P^{(YZ)}_{\text{CNOT}} + P^{(YY)}_{\text{CNOT}}, \tag{M10}$$

$$P^{(1)}_{IZCX} = P^{(IZ)}_{\text{CNOT}} + P^{(XZ)}_{\text{CNOT}} + P^{(IY)}_{\text{CNOT}} + P^{(XY)}_{\text{CNOT}}, \tag{M11}$$

$$P^{(1)}_{ZICX} = P^{(ZI)}_{\text{CNOT}} + P^{(ZX)}_{\text{CNOT}} + P^{(ZY)}_{\text{CNOT}} + P^{(ZZ)}_{\text{CNOT}} + P^{(YI)}_{\text{CNOT}} + P^{(YX)}_{\text{CNOT}} + P^{(YZ)}_{\text{CNOT}} + P^{(YY)}_{\text{CNOT}}, \tag{M12}$$

$$P^{(2)}_{IZCX} = P^{(IZ)}_{\text{CNOT}} + P^{(XZ)}_{\text{CNOT}} + P^{(IY)}_{\text{CNOT}} + P^{(XY)}_{\text{CNOT}} + P^{(ZI)}_{\text{CNOT}} + P^{(ZX)}_{\text{CNOT}} + P^{(YI)}_{\text{CNOT}} + P^{(YX)}_{\text{CNOT}}, \tag{M13}$$

$$P^{(3)}_{IZCX} = P^{(IZ)}_{\text{CNOT}} + P^{(IY)}_{\text{CNOT}} + P^{(ZZ)}_{\text{CNOT}} + P^{(ZY)}_{\text{CNOT}} + P^{(XZ)}_{\text{CNOT}} + P^{(XY)}_{\text{CNOT}} + P^{(YZ)}_{\text{CNOT}} + P^{(YY)}_{\text{CNOT}}, \tag{M14}$$

and

$$P^{(2)}_{ZICX} = P^{(ZI)}_{\text{CNOT}} + P^{(YI)}_{\text{CNOT}} + P^{(ZX)}_{\text{CNOT}} + P^{(YX)}_{\text{CNOT}}, \tag{M15}$$

$$P^{(1)}_{d} = P^{(Z)}_{\text{Id}} + P^{(Y)}_{\text{Id}}. \tag{M16}$$

Using Eqs. (M9) to (M16) and the same methods as in Appendix M 1, the leading order edge weight probabilities for the graph $G^{(2D)}_{(d_x)}$ are given by:

$$P^{(2D)}_{BLTRX} = \Gamma(P^{(1)}_{ZZCX}, P^{(1)}_{IZCX}, P^{(1)}_{d}; 1, 1, 1), \tag{M17}$$

$$P^{(2D)}_{TLBRX} = \Gamma(P^{(1)}_{ZZCX}, P^{(1)}_{IZCX}, P^{(1)}_{ZICX}, P^{(2)}_{IZCX}, P^{(1)}_{d}; 2, 2, 1, 1, 1), \tag{M18}$$

$$P_{C1X} = \Gamma(P^{(3)}_{IZCX}, P^{(2)}_{ZICX}, P^{(1)}_{d}; 1, 1, 1), \tag{M19}$$

$$P_{TB2X} = P^{(2D)}_{BLTRX}, \tag{M20}$$

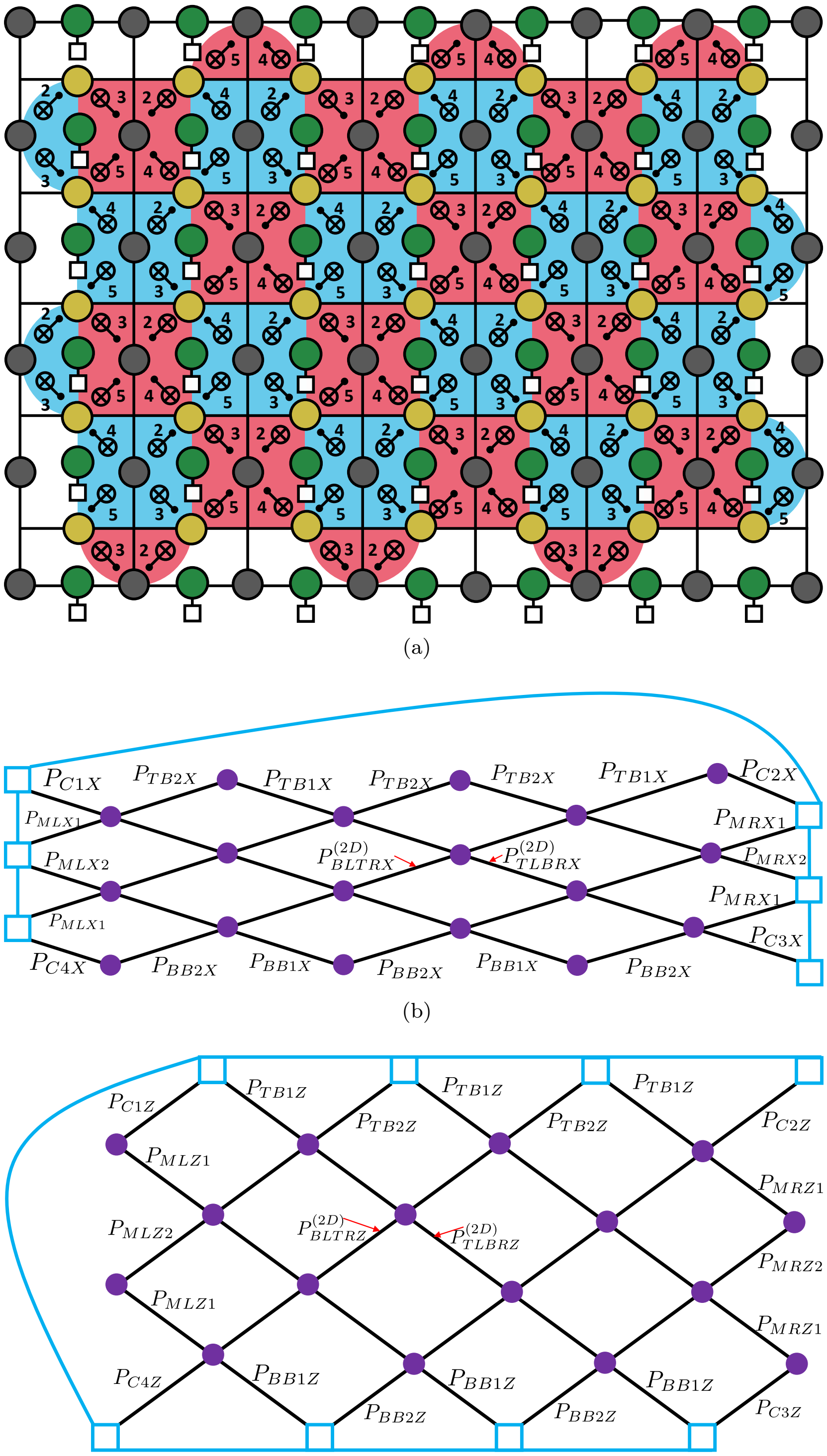
$$P_{TB1X} = \Gamma(P^{(1)}_{ZZCX}, P^{(1)}_{IZCX}, P^{(2)}_{IZCX}, P^{(1)}_{d}; 2, 1, 1, 1), \tag{M21}$$

$$P_{C2X} = \Gamma(P^{(3)}_{IZCX}, P^{(2)}_{IZCX}, P^{(1)}_{ZZCX}, P^{(1)}_{d}; 1, 1, 1, 1), \tag{M22}$$

$$P_{MRX1} = \Gamma(P^{(3)}_{IZCX}, P^{(2)}_{ZICX}, P^{(1)}_{d}; 1, 2, 1), \tag{M23}$$

$$P_{MRX2} = \Gamma(P^{(3)}_{IZCX}, P^{(1)}_{IZCX}, P^{(1)}_{ZICX}, P^{(2)}_{IZCX}, P^{(1)}_{d}; 1, 2, 1, 1, 1), \tag{M24}$$

$$P_{C3X} = \Gamma(P^{(3)}_{IZCX}, P^{(1)}_{IZCX}, P^{(1)}_{ZICX}, P^{(1)}_{d}; 1, 1, 1, 1), \tag{M25}$$



(a)

(b)

(c)

FIG. 53. (a) Surface code lattice with $d_x = 5$ and $d_z = 7$. (b) Graph used for decoding X stabilizer measurement outcomes with both bulk and boundary edge weight probability labels. (c) Graph used for decoding Z stabilizer measurement outcomes with both bulk and boundary edge weight probability labels.

$$P_{BB2X} = P^{(2D)}_{BLTRX}, \tag{M26}$$

$$P_{BB1X} = \Gamma(P^{(1)}_{IZCX}, P^{(1)}_{ZZCX}, P^{(1)}_{ZICX}, P^{(1)}_{d}; 2, 1, 1, 1), \tag{M27}$$

$$P_{C4X} = P_{C3X}, \tag{M28}$$

$$P_{MLX1} = \Gamma(P^{(3)}_{IZCX}, P^{(1)}_{IZCX}, P^{(2)}_{ZICX}, P^{(1)}_{ZICX}, P^{(1)}_{d}; 1, 1, 1, 1, 1), \tag{M29}$$

and

$$P_{MLX2} = P_{MLX1}. \tag{M30}$$

For the graph $G^{(2D)}_{(d_z)}$, we first define the following probabilities:

$$P^{(1)}_{XXCX} = P^{(XX)}_{\text{CNOT}} + P^{(XY)}_{\text{CNOT}} + P^{(YX)}_{\text{CNOT}} + P^{(YY)}_{\text{CNOT}}, \tag{M31}$$

$$P^{(1)}_{XICX} = P^{(XI)}_{\text{CNOT}} + P^{(YI)}_{\text{CNOT}} + P^{(XZ)}_{\text{CNOT}} + P^{(YZ)}_{\text{CNOT}}, \tag{M32}$$

$$P^{(1)}_{IXCX} = P^{(IX)}_{\text{CNOT}} + P^{(ZX)}_{\text{CNOT}} + P^{(IY)}_{\text{CNOT}} + P^{(ZY)}_{\text{CNOT}}, \tag{M33}$$

$$P^{(2)}_{IXCX} = P^{(IX)}_{\text{CNOT}} + P^{(IY)}_{\text{CNOT}} + P^{(ZX)}_{\text{CNOT}} + P^{(ZY)}_{\text{CNOT}} + P^{(XX)}_{\text{CNOT}} + P^{(XY)}_{\text{CNOT}} + P^{(YX)}_{\text{CNOT}} + P^{(YY)}_{\text{CNOT}}, \tag{M34}$$

$$P^{(3)}_{IXCX} = P^{(IX)}_{\text{CNOT}} + P^{(IY)}_{\text{CNOT}} + P^{(ZX)}_{\text{CNOT}} + P^{(ZY)}_{\text{CNOT}} + P^{(XI)}_{\text{CNOT}} + P^{(XZ)}_{\text{CNOT}} + P^{(YI)}_{\text{CNOT}} + P^{(YZ)}_{\text{CNOT}}, \tag{M35}$$

$$P^{(2)}_{XICX} = P^{(XI)}_{\text{CNOT}} + P^{(YI)}_{\text{CNOT}} + P^{(XX)}_{\text{CNOT}} + P^{(YX)}_{\text{CNOT}} + P^{(XZ)}_{\text{CNOT}} + P^{(YZ)}_{\text{CNOT}} + P^{(XY)}_{\text{CNOT}} + P^{(YY)}_{\text{CNOT}}, \tag{M36}$$

and

$$P^{(2)}_{d} = P^{(X)}_{\text{Id}} + P^{(Y)}_{\text{Id}}. \tag{M37}$$

Using Eqs. (M31) to (M37), the leading order edge weight probabilities for the graph $G^{(2D)}_{(d_z)}$ are given by:

$$P^{(2D)}_{BLTRZ} = \Gamma(P^{(1)}_{XXCX}, P^{(1)}_{XICX}, P^{(2)}_{d}; 1, 1, 1), \tag{M38}$$

$$P^{(2D)}_{TLBRZ} = \Gamma(P^{(1)}_{XXCX}, P^{(1)}_{XICX}, P^{(2)}_{IXCX}, P^{(3)}_{IXCX}, P^{(2)}_{d}; 2, 2, 1, 1, 1), \tag{M39}$$

$$P_{C1Z} = \Gamma(P^{(2)}_{XICX}, P^{(1)}_{IXCX}, P^{(2)}_{d}; 1, 1, 1), \tag{M40}$$

$$P_{TB1Z} = \Gamma(P^{(2)}_{XICX}, P^{(1)}_{XICX}, P^{(2)}_{IXCX}, P^{(3)}_{IXCX}, P^{(1)}_{XXCX}, P^{(2)}_{d}; 1, 1, 1, 1, 1, 1), \tag{M41}$$

$$P_{TB2Z} = \Gamma(P^{(2)}_{XICX}, P^{(1)}_{IXCX}, P^{(2)}_{d}; 1, 2, 1), \tag{M42}$$

$$P_{C2Z} = P_{C1Z}, \tag{M43}$$

$$P_{MRZ1} = \Gamma(P^{(1)}_{XXCX}, P^{(1)}_{XICX}, P^{(3)}_{IXCX}, P^{(2)}_{d}; 1, 2, 1, 1), \tag{M44}$$

$$P_{MRZ2} = P^{(2D)}_{BLTRZ}, \tag{M45}$$

$$P_{C3Z} = P_{C1Z}, \tag{M46}$$

$$P_{BB1Z} = \Gamma(P^{(2)}_{XICX}, P^{(1)}_{IXCX}, P^{(2)}_{IXCX}, P^{(1)}_{XICX}, P^{(2)}_{d}; 1, 1, 1, 1, 1), \tag{M47}$$

$$P_{BB2Z} = P_{BB1Z}, \tag{M48}$$

$$P_{C4Z} = \Gamma(P^{(2)}_{XICX}, P^{(1)}_{XICX}, P^{(2)}_{IXCX}, P^{(2)}_{d}; 1, 1, 1, 1), \tag{M49}$$

$$P_{MLZ1} = \Gamma(P^{(1)}_{XXCX}, P^{(1)}_{XICX}, P^{(2)}_{IXCX}, P^{(2)}_{d}; 1, 2, 1, 1), \tag{M50}$$

and

$$P_{MLZ2} = P^{(2D)}_{BLTRZ}. \tag{M51}$$

We now consider the three-dimensional version of the graphs in Figs. 53b and 53c (which we label $G^{(3D)}_{(d_x)}$ and $G^{(3D)}_{(d_z)}$) to deal with measurement errors in addition to space-time correlated errors arising from CNOT gate failures. As an example, consider an $I \otimes Z$ error arising from a CNOT gate failure in the second time-step of an X-type (red) plaquette during the k'th syndrome measurement round. Such a failure adds a Z data-qubit error which propagates through the CNOT in the fifth time-step of the top right red X-type plaquette. Let v_j and v_k be the vertices corresponding to the measurement outcomes of the two ancilla qubits which would detect the Z error. Assuming there were no other failures, only one of the two vertices (say v_j) changes from rounds $k-1$ to round k. In the next syndrome measurement round, both X-

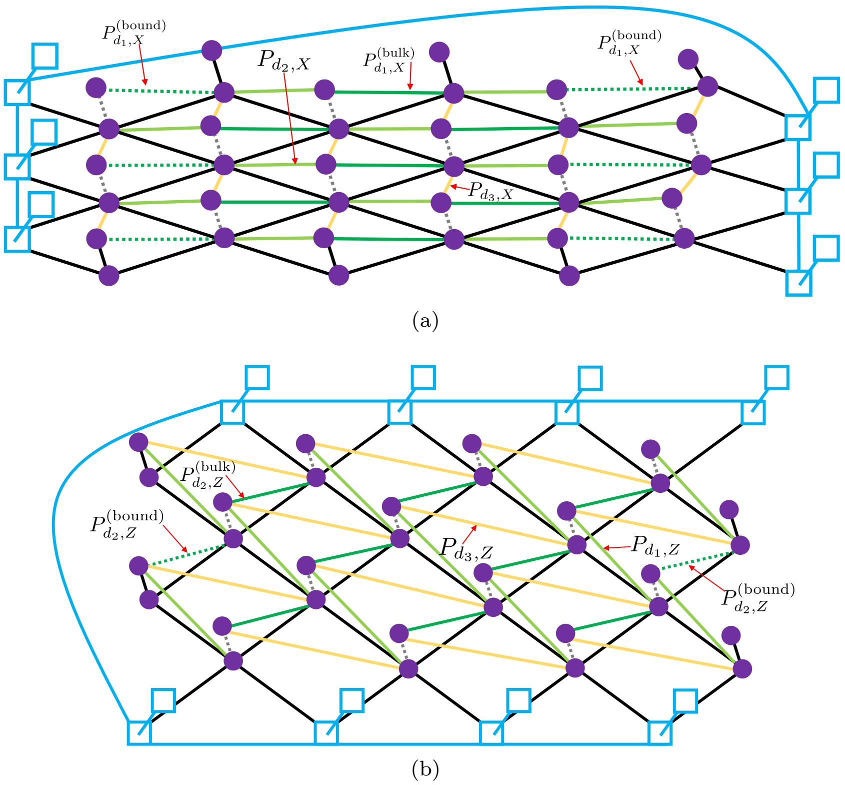

$P_{d_1,X}^{(\text{bound})}$
$P_{d_2,X}$
$P_{d_1,X}^{(\text{bulk})}$
$P_{d_1,X}^{(\text{bound})}$
$P_{d_3,X}$
(a)
$P_{d_2,Z}^{(\text{bulk})}$
$P_{d_2,Z}^{(\text{bound})}$
$P_{d_3,Z}$
$P_{d_1,Z}$
$P_{d_2,Z}^{(\text{bound})}$
(b)

FIG. 54. (a) Graph used for decoding X-type stabilizer measurements which include vertical edges (dashed gray edges) for dealing with measurement errors and space-time correlated edges for correcting errors arising from CNOT gate failures causing two different syndrome measurement outcomes in consecutive rounds. (b) Same as in (a) but for Z-type stabilizer measurements.

type plaquettes will detect the Z data qubit error the ancilla qubits in both X-type plaquettes will be highlighted. Hence only the vertex v_k changes from round k to $k+1$. In order to ensure that the highlighted ancillas arising from failures as in the example considered here can be reached by a single edge when implementing MWPM, the dark green edges in the graph of Fig. 54a (labeled $P_{d_1,X}^{(\text{bulk})}$) are added to the graph of Fig. 53b. The other types of space-time correlated edges are distinguished by their color and associated label (all edges of the same color have identical edge-weight probabilities). Similarly, we add the dashed grey vertical edges in Figs. 54a and 54b connecting identical vertices from two consecutive syndrome measurement rounds to deal with measurement errors. The edge weight probabilities of such edges are labeled P_V^X and P_V^Z. Note that there are also solid dark vertical edges at some of the boundaries of the graphs where weight-two X-type and Z-type stabilizers occur in Fig. 53a. These vertical edges have different edges weight probabilities which are labeled $P_V^{X,\text{bound}}$ and $P_V^{Z,\text{bound}}$.

In order to avoid making the visualization of the three-dimensional graphs too cumbersome, in Figs. 54a and 54b we only included vertices corresponding to the first two syndrome measurement rounds. Further, the two-dimensional edges from the second round were omitted in order to maintain focus on the vertical and space-time correlated edges connecting vertices from two consecutive syndrome measurement rounds.

Let

$$P_{VCX} = P_{\text{CNOT}}^{(ZI)} + P_{\text{CNOT}}^{(ZX)} + P_{\text{CNOT}}^{(YI)} + P_{\text{CNOT}}^{(YX)}, \quad \text{(M52)}$$

and

$$P_{VCZ} = P_{\text{CNOT}}^{(IX)} + P_{\text{CNOT}}^{(IY)} + P_{\text{CNOT}}^{(ZX)} + P_{\text{CNOT}}^{(ZY)}. \quad \text{(M53)}$$

Further, let P_s be the probability of preparing $|-\rangle$ instead of $|+\rangle$ and P_m be the probability that a X-basis measurement outcome is flipped. The edge weight probabilities corresponding to the dashed grey edges in Figs. 54a and 54b (i.e. the vertical edges of $G_{(d_x)}^{(3D)}$ and $G_{(d_z)}^{(3D)}$) are given by

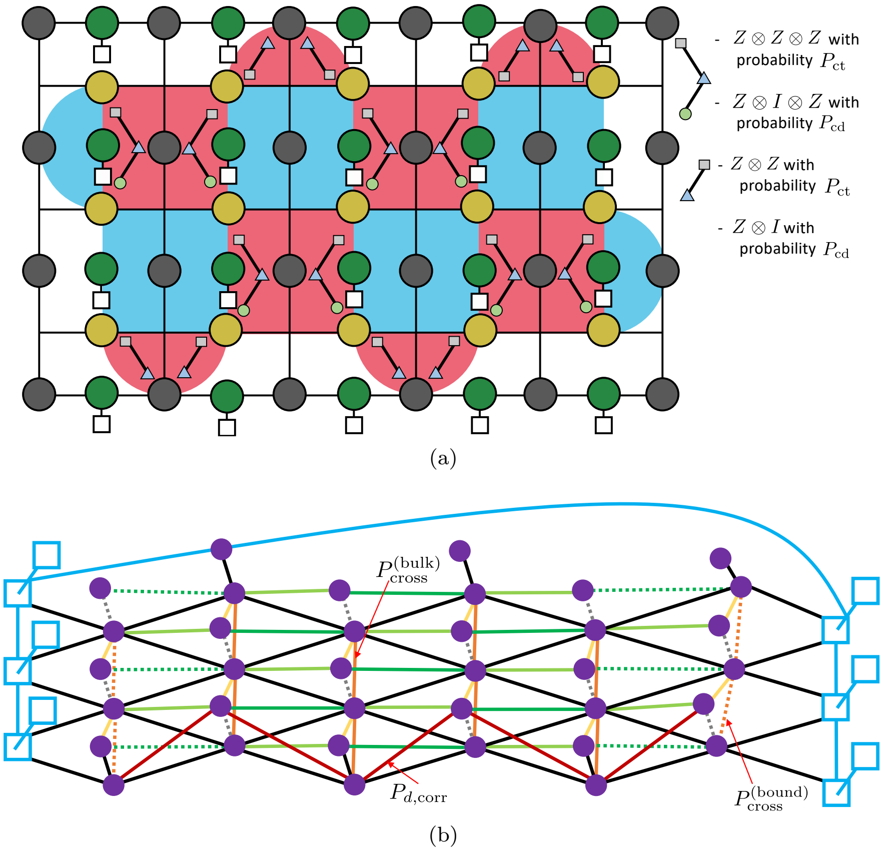
$$P_V^X = \Gamma(P_{VCX}, P_s, P_m; 4, 1, 1), \quad \text{(M54)}$$

$$P_V^{X,\text{bound}} = \Gamma(P_{VCX}, P_s, P_m; 2, 1, 1), \quad \text{(M55)}$$

$$P_V^Z = \Gamma(P_{VCZ}, P_s, P_m; 4, 1, 1), \quad \text{(M56)}$$

and

$$P_V^{Z,\text{bound}} = \Gamma(P_{VCZ}, P_s, P_m; 2, 1, 1). \quad \text{(M57)}$$


- $Z \otimes Z \otimes Z$ with probability P_{ct}
- $Z \otimes I \otimes Z$ with probability P_{cd}
- $Z \otimes Z$ with probability P_{ct}
- $Z \otimes I$ with probability P_{cd}
$P_{\mathrm{cross}}^{(\mathrm{bulk})}$
$P_{d,\mathrm{corr}}$
$P_{\mathrm{cross}}^{(\mathrm{bound})}$

(a)

(b)

FIG. 55. (a) Fictitious identity gates illustrating the possible correlated errors arising before the X-basis measurement of the X-type ancilla qubits. Grey squares correspond to the first qubit, blue triangles to the second qubit and green circles to the third qubit. (b) X-type decoding graph with added edges to correct correlated errors. The edge-weight probabilities of the orange cross-edges are labeled P_{cross}. We also add red edges with edge-weight probabilities labelled $P_{d,\mathrm{corr}}$ at the bottom row of the graph.

Next we consider the edge-weight probabilities for the space-time correlated edges of $G^{(3D)}_{(d_x)}$. The dark green edges labeled by $P^{(\mathrm{bulk})}_{d_1,X}$ have different values at the boundaries (dashed dark green edges in the first and last column of Fig. 54a) and are labeled by $P^{(\mathrm{bound})}_{d_1,X}$. We have that

$$P^{(\mathrm{bulk})}_{d_1,X} = \Gamma(P^{(1)}_{IZCX}, P^{(1)}_{ZZCX}, P^{(2)}_{ZICX}; 1, 1, 2), \tag{M58}$$

and

$$P^{(\mathrm{bound})}_{d_1,X} = \Gamma(P^{(1)}_{IZCX}, P^{(1)}_{ZZCX}, P^{(2)}_{ZICX}; 1, 1, 1). \tag{M59}$$

The edge weight probability $P_{d_2,X}$ (represented by the light green edges in Fig. 54a) is given by

$$P_{d_2,X} = \Gamma(P^{(1)}_{IZCX}, P^{(1)}_{ZZCX}; 1, 1). \tag{M60}$$

Lastly, the edge weight probability $P_{d_3,X}$ (represented by the yellow edges in Fig. 54a) is given by

$$P_{d_3,X} = P_{d_2,X}. \tag{M61}$$

Similarly, for the graph $G^{(3D)}_{(d_z)}$, the edge weight probability $P_{d_1,Z}$ (represented by the light green edges) is given by

$$P_{d_1,Z} = \Gamma(P^{(1)}_{XICX}, P^{(1)}_{XXCX}; 1, 1). \tag{M62}$$

The bulk and boundary edge weight probabilities $P^{(\mathrm{bulk})}_{d_2,Z}$ (dark green edges) and $P^{(\mathrm{bound})}_{d_2,Z}$ (dashed dark green edges) are given by

$$P^{(\mathrm{bulk})}_{d_2,Z} = \Gamma(P^{(1)}_{XICX}, P^{(1)}_{XXCX}, P^{(1)}_{IXCX}; 1, 1, 2), \tag{M63}$$

and

$$P^{(\mathrm{bound})}_{d_2,Z} = \Gamma(P^{(1)}_{XICX}, P^{(1)}_{XXCX}, P^{(1)}_{IXCX}; 1, 1, 1). \tag{M64}$$

Lastly, the edge weight probability $P_{d_3,Z}$ (represented by the yellow edges) is given by

$$P_{d_3,Z} = P_{d_1,Z}. \tag{M65}$$

3. Adding edges for dealing with correlated errors

In this section we provide a modified version of the graph $G^{(3D)}_{(d_x)}$ (described in Appendix M 2) which includes extra edges to deal with two-qubit and three-qubit correlated errors arising from the micro oscillations described in Appendix B 5.

For the purposes of the edge weight analysis, in Fig. 55a, we illustrate fictitious two-qubit and three-qubit gates which act as the identity and which are applied immediately prior to the X-basis measurements of the red plaquettes. The two-qubit correlated errors can be viewed as an $Z \otimes I \otimes Z$-type error at a three-qubit gate location, where the Z errors act on the qubits adjacent to the grey squares and green circles of such gates. Such errors occur with probability P_{cd}. Similarly, the three-qubit correlated errors can be viewed as an $Z \otimes Z \otimes Z$-type error at a three-qubit gate location. Such errors occur with probability P_{ct}. Additionally, there can be correlated errors occurring between the ancilla and data qubits at the top and bottom boundaries of the lattice in Fig. 55a. Hence, we add fictitious two-qubit gate locations at such boundaries as shown in the figure.

In order to incorporate the different types of correlated errors mentioned above into our MWPM decoding protocol, extra edges are added to the graph $G^{(3D)}_{(d_x)}$ as shown in Fig. 55b. The first type of extra edges are two-dimensional cross edges shown in orange that deal with two and three-qubit correlated errors arising at the three-qubit fictitious gate locations of Fig. 55a. The edge-weight probabilities of such edges are labeled $P^{(\text{bulk})}_{\text{cross}}$. Due to boundary effects, we also add dashed orange edges with edge-weight probabilities labeled $P^{(\text{bound})}_{\text{cross}}$. Additionally, extra space-time correlated edges (shown in red) are added at the bottom row of the graph in Fig. 55b with edge weight probabilities labeled by $P_{d,\text{corr}}$. Note that the two-qubit correlated errors arising at the top boundary of Fig. 55a result in space-time correlated edges which are already included in $G^{(3D)}_{(d_x)}$.

In addition to the extra edges added to $G^{(3D)}_{(d_x)}$, the edge-weight probabilities of a subset of the edges already included in $G^{(3D)}_{(d_x)}$ need to be renormalized. The edge-weight probabilities of the added edges in addition to the renormalized edges are given by:

$$P^{(\text{bulk})}_{\text{cross}} = \Gamma(P_{\text{ct}} P_{\text{cd}}; 2, 2), \tag{M66}$$

$$P^{(\text{bound})}_{\text{cross}} = \Gamma(P_{\text{ct}} P_{\text{cd}}; 1, 1), \tag{M67}$$

$$P_{d,\text{corr}} = P_{\text{ct}}, \tag{M68}$$

$$P_{TB2X} = \Gamma(P^{(1)}_{ZZCX}, P^{(1)}_{IZCX}, P^{(1)}_{d}, P_{\text{cd}}; 1, 1, 1, 1), \tag{M69}$$

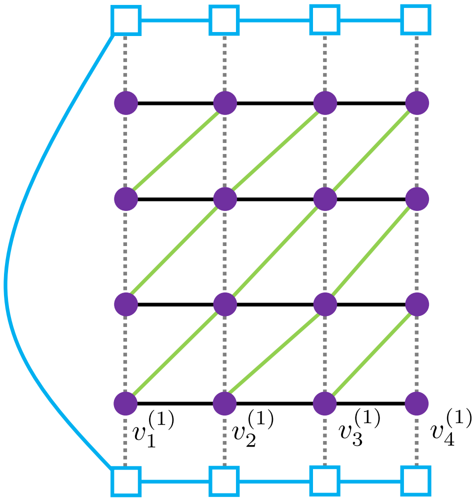


FIG. 56. Example of a decoding graph for correcting timelike errors using a $d = 5$ repetition code with $d_m = 4$. The top and bottom boundary edges (with zero weight) and vertices are shown in blue and are connected by a blue edge with zero weight. As explained in Appendix L, we have removed the left and two-dimensional black edges (which correspond to the left and rightmost qubits) to isolate timelike errors.

$$P_{TB1X} = \Gamma(P^{(1)}_{ZZCX}, P^{(1)}_{IZCX}, P^{(2)}_{IZCX}, P^{(1)}_{d}, P_{\text{cd}}; 2, 1, 1, 1, 1), \tag{M70}$$

$$P_{C2X} = \Gamma(P^{(3)}_{IZCX}, P^{(2)}_{IZCX}, P^{(1)}_{ZZCX}, P^{(1)}_{d}, P_{\text{cd}}, P_{\text{ct}}; 1, 1, 1, 1, 1, 1), \tag{M71}$$

$$P_{BB2X} = \Gamma(P^{(1)}_{ZZCX}, P^{(1)}_{IZCX}, P^{(1)}_{d}, P_{\text{cd}}; 1, 1, 1, 1), \tag{M72}$$

$$P_{BB1X} = \Gamma(P^{(1)}_{IZCX}, P^{(1)}_{ZZCX}, P^{(1)}_{ZICX}, P^{(1)}_{d}, P_{\text{cd}}; 2, 1, 1, 1, 1), \tag{M73}$$

$$P_{C4X} = \Gamma(P^{(3)}_{IZCX}, P^{(1)}_{IZCX}, P^{(1)}_{ZICX}, P^{(1)}_{d}, P_{\text{cd}}, P_{\text{ct}}; 1, 1, 1, 1, 1, 1), \tag{M74}$$

$$P^{X}_{V} = \Gamma(P_{VCX}, P_{s}, P_{m}, P_{\text{ct}}; 4, 1, 1, 1), \tag{M75}$$

For the space-time correlated edges, at the top row of the graph in Fig. 55b, we have

$$P^{(\text{bound,top})}_{d_1,X} = \Gamma(P^{(1)}_{IZCX}, P^{(1)}_{ZZCX}, P^{(2)}_{ZICX}, P_{\text{ct}}; 1, 1, 1, 1), \tag{M76}$$

whereas at the bottom boundary $P^{(\text{bound,bottom})}_{d_1,X}$ is given by Eq. (M59). Similarly, at the top row of Fig. 55b, we

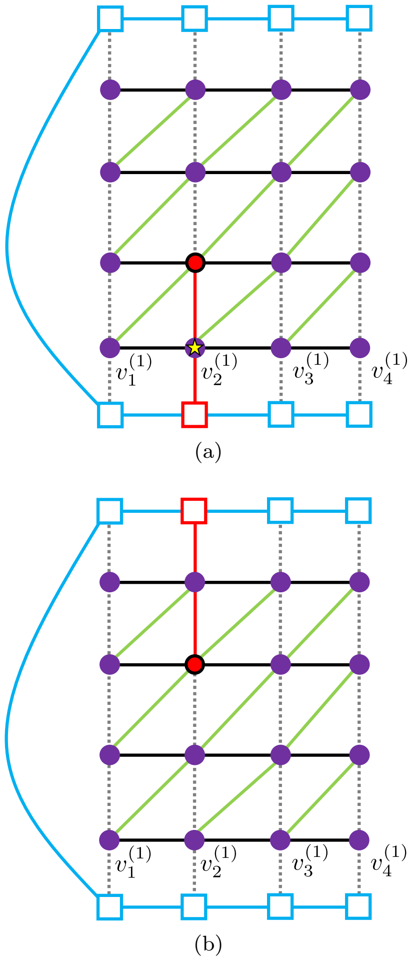


FIG. 57. (a) Implementation of the timelike decoding protocol in the presence of a single measurement error when measuring the stabilizer X_2X_3 during the first round. The minimum weight path matches to the bottom boundary going through the vertex $v_2^{(1)}$ whose outcome is correctly flipped (illustrated by the yellow star). (b) Same as in (a) but with an additional measurement error occurring in the second round when measuring X_2X_3. In this case, the minimum weight path matches to the top boundary and fails to flip the measurement outcome of $v_2^{(1)}$ (which is incorrect given the measurement error in the first round) resulting in a logical failure.

have

$$P_{d_2,X}^{(\mathrm{top})} = \Gamma(P_{IZCX}^{(1)}, P_{ZZCX}^{(1)}, P_{\mathrm{ct}}; 1, 1, 1), \tag{M77}$$

whereas anywhere else in the graph $P_{d_2,X}$ is given by Eq. (M60).

4. Decoding time-like errors

In this section, we show how the decoding graphs in addition to the MWPM decoding protocols need to be modified for correcting timelike errors discussed in Appendix L. Since visualizing three-dimensional graphs can be challenging, we focus on correcting timelike errors in the context of the repetition code, even though timelike errors occur in surface code patches when implementing our lattice surgery schemes. However the main techniques discussed in the context of the repetition can straightforwardly be applied to the rotated surface code.

An example of a decoding graph for timelike errors occurring in a $d = 5$ repetition code with $d_m = 4$ is given in Fig. 56. Note that unlike Fig. 52b, the boundary edges and vertices (shown in blue) are at the top and bottom of the graph since we follow the matching protocol explained in Fig. 12. In particular, we are considering a setting analogous to Fig. 50, where data qubits between two repetition code patches are initially prepared in the $|0\rangle$ state, and the product of the X-stabilizers yields the outcome $X_{L1}X_{L2}$. Note that although the measurement of each X-stabilizer in the first round is random, the parity of the product of all measurement outcomes gives the outcome of $X_{L1}X_{L2}$. Due to possibility of measurement errors, measurements of the X-stabilizers are repeated d_m times. MWPM is then performed over the entire syndrome history in order to determine if measurement errors occurred during the measurement of X-stabilizers in the first round. We thus summarize the decoding protocol as follows:

1. Repeat the measurement of all X-stabilizers d_m times.

2. Implement MWPM using a timelike decoding graph (such as the one in Fig. 56). If there is an odd number of highlighted vertices (purple vertices in Fig. 56), highlight a boundary vertex (the particular choice is irrelevant).

3. Let $v_j^{(1)}$ correspond to the j'th X-stabilizer measurement outcome in the first round (represent by the j'th purple vertex, starting from the left, in the first layer of the graph in Fig. 56). If there are highlighted timelike edges (i.e. vertical edges) incident to $v_j^{(1)}$, flip the measurement outcome of $v_j^{(1)}$.

4. Let $\tilde{v}_j^{(1)}$ correspond to the values of $v_j^{(1)}$ after implementing MWPM and performing the appropriate measurement flips described in the previous step. The outcome v_f of $X_{L1}X_{L2}$ is given by $v_f = \sum_{j=1}^{d-1} \tilde{v}_j^{(1)} \mathrm{mod}(2)$.

In Fig. 57, we provide two examples for the implementation of the timelike decoder. In Fig. 57a, we consider the case where a single measurement error occurs in the first round when measuring the stabilizer X_2X_3. Since the syndrome changes between the first and second round, the second vertex (starting from the left to right) of the second two-dimensional layer is highlighted. A boundary vertex is also highlighted to ensure the total number of highlighted vertices is even. After implementing MWPM, the minimum weight path connecting the two highlighted vertices correctly passes through $v_2^{(1)}$ in the timelike direction. The decoder then flips the measurement outcome of X_2X_3 in the first round resulting in the correct parity for the he outcome of $X_{L1}X_{L2}$. In Fig. 57b, we consider a similar setting but with two consecutive measurement errors of the stabilizer X_2X_3 occurring in the first and second round. In this case, the syndrome only changes between the second and third round resulting in the red highlighted vertex shown in Fig. 57b. After implementing MWPM, the minimum weight path connect to the top boundary and so the measurement outcome of X_2X_3 in the first round is incorrectly left unchanged resulting in a logical failure.

We conclude this section with an important remark. Suppose a measurement error occurs in the first round when measuring the X-type stabilizer $S_j^{(x)}$ of a given code. In order to prevent highlighted timelike edges from being incident to the vertex $v_j^{(1)}$, one requires additional measurement errors such that minimum weight paths are matched in the top timelike portion of the decoding graph (as in Fig. 57b). By increasing d_m to $d_m + 2$, one requires an additional measurement error to guarantee that the minimum weight path is not incident to $v_j^{(1)}$, thus explaining the scaling in Eq. (L3).

Appendix N: Toffoli state distillation (TDTOF)

1. Prior state of the art

Here we give a high-level comparison of how our TDTOF protocol compares to the prior art in terms of magic state conversion rates.

Early protocols for fault-tolerant quantum computation focused on TOF state preparation in concatenated codes [119] or they protected against 1 type of error [120]. However, none of these protocols are suitable for protecting against generic noise in topological (e.g. repetition or surface) codes.

A more modern approach to magic state distillation uses a supply of low fidelity T magic states. There are many protocols for distillation of noisy T-states to purer T-states [30, 31, 33, 121–123]. One can also use T-states as input to protocols that output other types of magic states, including TOF states [124–129]. For instance, there were parallel discoveries of protocols [124, 125] that distill 1 TOF state from 8 noisy T states, which we will write as $8T \rightarrow 1$TOF. This was later generalized using synthillation [127, 128] to a family of protocols $(6k+2)T \rightarrow k$TOF for any integer k. However, in some settings, the supply of noisy TOF states can be prepared with better fidelities than the noisy T states. For instance, in this paper we have shown that in system with highly-biased noise we can use a repetition encoding and the BUTOF protocol to realize TOF state at better fidelities than physical TOF gates, with only a mild additional resource cost.

It has been previously noted [130] that triorthogonal codes enabling $(6k+8)T \rightarrow (2k)T$ state distillation can also be lifted to perform $(6k+8)$TOF $\rightarrow (2k)$TOF. The conversion rate of these protocols is $2k/(6k+8)$, which is poor when k is small (starting at $1/7$ for $k=1$) but improving when k is larger (approaching $1/3$ for $k \rightarrow \infty$). However, the ratio of inputs to outputs is not the sole metric of importance; also crucial is the space-time complexity of the Clifford circuit implementing the distillation protocol. Previous analysis has found that the space-time complexity of Clifford distillation circuits tends to be more favorable for simpler protocols using smaller block sizes [66, 79, 122] and that this effect can outweigh the improvement of conversion rate in the asymptotic regime. In other words, the desiderata for distillation protocols converting $n \rightarrow k$ magic states, are that: the protocol has a good rate, so k/n is large; the protocol is compact so n is as small as possible. These desiderata are in tension since rates tends to improve asymptotically as block size n is increased. A protocol satisfying these desiderata, will likely have a small space-time footprint when compiled down to physical qubits and gates. In this work, we present a 8TOF → 2TOF protocol that protects against any single location fault (of X, Y or Z type), so it has a relatively high conversion rate of $1/4$ without needing to scale to large blocks. In contrast, to achieve the same conversion rate using the ideas of Ref. [130] would require a much larger 32TOF → 8TOF protocol.

2. Transversality proofs

Here we prove that the trio of $[[8,2,2]]$ codes introduced in Section VII have the required CCZ tranversality properties. Recall that CCZ is a 3-qubit gate that adds a "-1" phase to the state $|111\rangle$ and "$+1$" to all other computational basis states. The corresponding magic state $|\mathrm{CCZ}\rangle$ differs from $|\mathrm{TOF}\rangle$ by a single Hadamard gate. For reasons of mathematical elegance, it is simpler to work mostly in terms of $|\mathrm{CCZ}\rangle$ state distillation, but our final description of the distillation protocol will be presented in terms of $|\mathrm{TOF}\rangle$ states.

We say a set of $[[n,k,d]]$ codes is CCZ transversal whenever $\mathrm{CCZ}^{\otimes n}$ performs a logical $\mathrm{CCZ}^{\otimes k}$ gate. Note that if we take three copies of a CCS code that has a transversal T gate (so that $T^{\otimes n} = T_L$ or similar, then it must also be CCZ transversal). This is simply because CCS codes have transversal CNOT gates and we can synthesize CCZ gates from CNOT and T gates. Essentially, this is the observation exploited to construct $(6k+8)$TOF $\rightarrow (2k)$TOF protocols [130]. However, it is possible for a trio of codes to be CCZ transversal, but not be T transversal. To the best of our knowledge this was first shown for the 3D surface codes by showing an equivalence (via unfolding) to 3D colour codes [131]. Later, Vasmer and Brown gave a more direct proof that the 3D surface codes are CCZ transversal [76]. Here, we use similar proof techniques to Vasmer and Brown, though generalized (to $k > 1$) and with a new code construction that code not appear to be a surface code.

We define codes here using slightly different notation from the main text. Given an n-qubit bit string $\mathbf{s} = (s_1, s_2, \ldots, s_n)$, we use $X[\mathbf{s}] := \otimes_j X^{s_j}$. For example, if

$$\mathbf{u} = (1, 1, 1, 0, 0, 1, 0, 0), \tag{N1}$$

then

$$X[\mathbf{u}] = X \otimes X \otimes X \otimes \mathbb{1} \otimes \mathbb{1} \otimes X \otimes \mathbb{1} \otimes \mathbb{1}. \tag{N2}$$

With this notation we can define an $[[n, k, d]]$ CCS code using a binary G-matrix representation as follows.

Let G be a binary matrix that is row-wise linearly independent and partitioned as follows

$$G = \left(\frac{G_1}{G_0}\right), \tag{N3}$$

where G has n columns and G_1 has k rows. Letting m denote the number of rows in G_0, then for a non-trivial ($d \geq 2$) code we know $m \geq 1$. Here, we review the relevant facts for G-matrices, but for additional details and proofs refer the reader to Refs. [78, 127–129]. This allows us to define a CSS code with all-zero logical state

$$|(0, \ldots, 0)\rangle_L = 2^{-m/2} \sum_{\mathbf{u} \in \mathbb{F}_2^m} |\mathbf{u}G_0\rangle. \tag{N4}$$

Note we use bold-font for row vectors. The notation $\mathbf{u}G_0$ represents left multiplication of matrix G_0 by the row vector $\mathbf{u}$, performed modulo 2, which will produce a length n row-vector describing a physical, computational basis state. The set of all $\mathbf{u}G_0$ corresponds to the row-span of G_0 and form a group under addition modulo two.

Furthermore, logical computation basis states can be represented by a k-bit string $\mathbf{x} = (x_1, \ldots, x_k)$ as follows

$$|\mathbf{x}\rangle_L = \frac{1}{\sqrt{|\mathcal{G}_0|}} \sum_{\mathbf{u} \in \mathbb{F}_2^m} |\mathbf{u}G_0 + \mathbf{x}G_1\rangle, \tag{N5}$$

where $\mathbf{x}G_1$ is again obtained by matrix multiplication (modulo 2) and is a constant shift identifying a coset of the group generated by addition (modulo 2) of rows of G_0. We can compress this notation slightly by noting

$$\mathbf{u}G_0 + \mathbf{x}G_1 = (\mathbf{x}, \mathbf{u})G, \tag{N6}$$

where $(\mathbf{x}, \mathbf{u})$ is the row-vector resulting from joining $\mathbf{u}$ and $\mathbf{x}$. Again, note that Eq. (N6) should be read as modulo two and this will be the convention for such expressions throughout the remainder of this appendix.

The j^{th} logical X operator, denoted X_{Lj}, ought to flip the $|\mathbf{0}\rangle_L$ state to $|(\hat{\mathbf{o}}_j)\rangle_L$ state, where $\hat{\mathbf{o}}_j$ is a unit vector with a single "1" entry at the j^{th} location. It is straightforward to verify that $X_{Lj} = X[\hat{\mathbf{o}}_j G_1]$ performs the required flip and that $\hat{\mathbf{o}}_j G_1$ is equal to the j^{th} row of G_1. Therefore, the logical operators of the code are given by the row vectors of G_1. Furthermore, for every $\mathbf{g}$ in the row-span of G_0, the operator $X[\mathbf{g}]$ is an X-stabilizer of the codespace, and this enumerates all the X-stabilizers.

As a final notational preliminary, we will make use of a triple dot product between triples of vectors. If $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are binary vectors of equal length, we define

$$|\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}| = \sum_j a_j b_j c_j \pmod 2, \tag{N7}$$

which we again evaluate modulo 2. It is useful to note that this counts the parity of the number of locations where operators $X[\mathbf{a}]$, $X[\mathbf{b}]$ and $X[\mathbf{c}]$ all act non-trivially.

This G-matrix representation was also used for triorthogonal codes [30] and quasi-triorthogonal codes [127, 128] except we are interested in different transversality properties and so we will require different constraints on the weight of rows in G_0 and G_1. The additional constraints determine the transversality properties that we summarise with the following result, which is a slight generalization (beyond $k = 1$) of the proof techniques used by Vasmer and Browne [76]

Lemma 1. *Let $\{G^A, G^B, G^C\}$ be a trio of G-matrices that represent a trio of $[[n, k, d]]$ codes. Additionally, assume the following triple dot product conditions (recall Eq. N7)*

$$|\hat{\mathbf{o}}_p G^A \wedge \hat{\mathbf{o}}_q G^B \wedge \hat{\mathbf{o}}_r G^C| = \begin{cases} 1 & \textit{if } p = q = r \leq k \\ 0 & \textit{otherwise} \end{cases} \tag{N8}$$

where $\hat{\mathbf{o}}_p$ is a binary unit vector with 1 in location p and 0 everywhere else. Then it follows that a physical CCZ$^{\otimes n}$ realizes a transveral, logical CCZ$^{\otimes k}$.

Let us remark on what Eq. (N8) means in terms of operators. Observe that when $p \leq k$, the operator $X[\hat{\mathbf{o}}_p G^D]$ is the p^{th} logical operator for codeblock $D \in \{A, B, C\}$. Therefore, the condition of Eq. (N8) tells us that the X_{Lp} logical operators must share an odd number of qubit indices where they all act non-trivially. All other combinations of logical operators and stabilizers have an even number of such locations.

Proof. To determine the phase acquired from acting on the codespace with CCZ$^{\otimes n}$, we first ask how this operator acts on an arbitrary computational basis state. Recall that $\text{CCZ}^{\otimes n} = \prod_{j=1}^{n} \text{CCZ}_j$ where CCZ_j acts on qubit j in each block. Given a triple of n-qubit binary vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ we have

$$\text{CCZ}_j |\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle = (-1)^{a_j b_j c_j} |\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle, \tag{N9}$$

and so

$$\text{CCZ}^{\otimes n} |\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle = (-1)^{|\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}|} |\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle, \tag{N10}$$

where $|\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}| = \sum_j a_j b_j c_j$ as we introduced earlier. Next, we ask how this acts on the codespace.

Consider a trio of computational basis states $|\mathbf{x}\rangle_L|\mathbf{y}\rangle_L|\mathbf{z}\rangle_L$ encoded in blocks A, B and C respectively. Using Eqs. (N5) and (N6), we see that

$$\begin{aligned}&|\mathbf{x}\rangle_L|\mathbf{y}\rangle_L|\mathbf{z}\rangle_L \\ &= 2^{-3m/2}\sum_{\mathbf{u},\mathbf{v},\mathbf{w}\in\mathbb{F}_2^m}|(\mathbf{x},\mathbf{u})G^A\rangle|(\mathbf{y},\mathbf{v})G^B\rangle|(\mathbf{z},\mathbf{w})G^C\rangle.\end{aligned} \tag{N11}$$

To determine the phase acquired from acting on $|\mathbf{x}\rangle_L|\mathbf{y}\rangle_L|\mathbf{z}\rangle_L$ with $\mathrm{CCZ}^{\otimes n}$, we consider its action on each term in the superposition using Eq. (N10). Each term acquires a phase

$$\begin{aligned}&\mathrm{CCZ}^{\otimes n}|(\mathbf{x},\mathbf{u})G^A\rangle|(\mathbf{y},\mathbf{v})G^B\rangle|(\mathbf{z},\mathbf{w})G^C\rangle \\ &= (-1)^{\lambda}|(\mathbf{x},\mathbf{u})G^A\rangle|(\mathbf{y},\mathbf{v})G^B\rangle|(\mathbf{z},\mathbf{w})G^C\rangle,\end{aligned} \tag{N12}$$

where the phase exponent is

$$\lambda = |(\mathbf{x},\mathbf{u})G^A \wedge (\mathbf{y},\mathbf{v})G^B \wedge (\mathbf{z},\mathbf{w})G^C|. \tag{N13}$$

Using linearity of the triple dot-product and expanding the vectors in terms of unit-vectors, e.g $(\mathbf{x},\mathbf{u}) = \sum_p \hat{\mathbf{o}}_p(\mathbf{x},\mathbf{u})_p$, we have

$$\lambda = \sum_{p,q,r}(\mathbf{x},\mathbf{u})_p(\mathbf{y},\mathbf{v})_q(\mathbf{z},\mathbf{w})_r|\hat{\mathbf{o}}_pG^A \wedge \hat{\mathbf{o}}_qG^B \wedge \hat{\mathbf{o}}_rG^C|. \tag{N14}$$

Next, using the assumption of Eq. (N8), we see almost all these terms vanish except a few when $p = q = r \leq k$

$$\lambda = \sum_{p\leq k}(\mathbf{u},\mathbf{x})_p(\mathbf{v},\mathbf{y})_p(\mathbf{w},\mathbf{z})_p. \tag{N15}$$

Notice that if $p \leq k$, $(\mathbf{x},\mathbf{u})_p = (\mathbf{x})_p$ since $\mathbf{x}$ is length k. Therefore,

$$\begin{aligned}\lambda &= \sum_{p\leq k}(\mathbf{x})_p(\mathbf{y})_p(\mathbf{z})_p \\ &= |\mathbf{x}\wedge\mathbf{y}\wedge\mathbf{z}|,\end{aligned} \tag{N16}$$

where in the last line we have noted that the summation is exactly the triple dot-product between these vectors. Substituting this back into Eq. (N12) we have

$$\begin{aligned}&\mathrm{CCZ}^{\otimes n}|(\mathbf{x},\mathbf{u})G^A\rangle|(\mathbf{y},\mathbf{v})G^B\rangle|(\mathbf{z},\mathbf{w})G^C\rangle \\ &= (-1)^{|\mathbf{x}\wedge\mathbf{y}\wedge\mathbf{z}|}|(\mathbf{x},\mathbf{u})G^A\rangle|(\mathbf{y},\mathbf{v})G^B\rangle|(\mathbf{z},\mathbf{w})G^C\rangle.\end{aligned} \tag{N17}$$

Since the dependence on $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ has vanished, $\mathrm{CCZ}^{\otimes n}$ acts identically on every term in the superposition comprising the logical computation basis states so we have

$$\mathrm{CCZ}^{\otimes n}|\mathbf{x}\rangle_L|\mathbf{y}\rangle_L|\mathbf{z}\rangle_L = (-1)^{|\mathbf{x}\wedge\mathbf{y}\wedge\mathbf{z}|}|\mathbf{x}\rangle_L|\mathbf{y}\rangle_L|\mathbf{z}\rangle_L. \tag{N18}$$

This is precisely the phase expected from $\mathrm{CCZ}^{\otimes k} = \prod_j \mathrm{CCZ}_{Lj}$ since each CCZ_{Lj} contributes one term $x_jy_jz_j$ to the phase exponent. □

We remark that the above proof closely follows previous work on 3D surface codes [76] but generalised to arbitrary k. This approach could be further extended using a proof technique similar to Refs. [127, 128] to cover cases where: the logical unitary is not $\mathrm{CCZ}^{\otimes k}$ but some other non-Clifford unitary; and/or the full codespace is not necessarily divisible into 3 equal sized blocks. However, this more sophisticated approach is not required for our present purposes.

Rather, we are interested in the special case

Lemma 2. *Consider a trio of G-matrices as follows*

$$G^A = \begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \hline \mathbf{1}\end{pmatrix}, G^B = \begin{pmatrix}\mathbf{u}_2\\ \mathbf{u}_3\\ \hline \mathbf{1}\end{pmatrix}, G^C = \begin{pmatrix}\mathbf{u}_3\\ \mathbf{u}_1\\ \hline \mathbf{1}\end{pmatrix}, \tag{N19}$$

where $\mathbf{1} = (1, 1, \ldots, 1)$. Assume that

1. *$\forall t$: $|\mathbf{u}_t| = \sum_{j=1}^n(\mathbf{u}_t)_j \pmod 2 = 0$;*
2. *$\forall t, t'$: $|\mathbf{u}_t \wedge \mathbf{u}_{t'}| = \sum_{j=1}^n(\mathbf{u}_t)_j(\mathbf{u}_{t'})_j \pmod 2 = 0$;*
3. *$|\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \mathbf{u}_3| = 1$.*

Then the corresponding codes are $[[n, 2, 2]]$ codes with a tranversal logical $CCZ^{\otimes n} = CCZ_L^{\otimes 2}$. For instance, these conditions are met by setting

$$\mathbf{u}_1 = (1, 1, 1, 0, 0, 1, 0, 0), \tag{N20}$$
$$\mathbf{u}_2 = (1, 1, 0, 1, 1, 0, 0, 0), \tag{N21}$$
$$\mathbf{u}_3 = (1, 0, 1, 0, 1, 0, 1, 0), \tag{N22}$$

to produce a trio of $[[8, 2, 2]]$ codes with CCZ transversality as above.

The above lemma provides an example trio of $[[8, 2, 2]]$ codes with the desired transversality property. To be more concrete, by combining Eq. (N19) and Eqs. (N20) to (N22) the trio of codes have G-matrix representation

$$G^A = \left(\begin{array}{cccccccc}1&1&1&0&0&1&0&0\\1&1&0&1&1&0&0&0\\ \hline 1&1&1&1&1&1&1&1\end{array}\right), \tag{N23}$$

$$G^B = \left(\begin{array}{cccccccc}1&1&0&1&1&0&0&0\\1&0&1&0&1&0&1&0\\ \hline 1&1&1&1&1&1&1&1\end{array}\right), \tag{N24}$$

$$G^C = \left(\begin{array}{cccccccc}1&0&1&0&1&0&1&0\\1&1&1&0&0&1&0&0\\ \hline 1&1&1&1&1&1&1&1\end{array}\right), \tag{N25}$$

By translating these matrices into X-stabilizers (all have the $X[\mathbf{1}] = X^{\otimes 8}$ stabilizer) and logical X operators (which differ), we verify that these are the same codes as specified by the operators given the main text (see e.g. Eq. (83)). However, the lemma provides some general conditions under which transversality is satisfied to provide a better insight into the proof technique.

Proof. The proof of Lemma 2 follows quickly from Lemma 1 by simply verifying all the cases. For instance, for $p = q = r$ we have

$$\begin{aligned} |\hat{\mathbf{o}}_1 G^A \wedge \hat{\mathbf{o}}_1 G^B \wedge \hat{\mathbf{o}}_1 G^C| &= |\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \mathbf{u}_3| = 1, \qquad \text{(N26)} \\ |\hat{\mathbf{o}}_2 G^A \wedge \hat{\mathbf{o}}_2 G^B \wedge \hat{\mathbf{o}}_2 G^C| &= |\mathbf{u}_2 \wedge \mathbf{u}_3 \wedge \mathbf{u}_1| = 1, \\ |\hat{\mathbf{o}}_3 G^A \wedge \hat{\mathbf{o}}_3 G^B \wedge \hat{\mathbf{o}}_3 G^C| &= |\mathbf{1} \wedge \mathbf{1} \wedge \mathbf{1}| = 0. \end{aligned}$$

In the second line, we have used that the triple dot product is invariant under permutation of vectors, for instance $|\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}| = |\mathbf{b} \wedge \mathbf{c} \wedge \mathbf{a}|$. The last equality in each line comes from the assumptions in Lemma 2. Since $k = 2$, we see that we indeed get unity when $p = q = r \leq k$ and zero for $p = q = r > 2$ (as required by Eq. (N8)). Let us consider a case when $p, q, r \leq k$ but $p \neq q$, such as

$$\begin{aligned} |\hat{\mathbf{o}}_2 G^A \wedge \hat{\mathbf{o}}_1 G^B \wedge \hat{\mathbf{o}}_1 G^C| &= |\mathbf{u}_2 \wedge \mathbf{u}_2 \wedge \mathbf{u}_3| \qquad \text{(N27)} \\ &= |\mathbf{u}_2 \wedge \mathbf{u}_3| = 0. \end{aligned}$$

We have used the simple identity that in $\mathbb{F}_2$ we have $a^2 b = ab$ and the natural extension to vectors that $|\mathbf{a} \wedge \mathbf{a} \wedge \mathbf{b}| = |\mathbf{a} \wedge \mathbf{b}|$. The last equality comes from the assumptions in Lemma 2 and gives the result required by Eq. (N8). By inspecting Eq. (N19), we find that for any triple of rows (except for the special case when $p = q = r$) from the upper block G_1, two of the selected rows will be equal and so the triple dot product will again give zero, therefore satisfying Eq. (N8).

Next, let us consider a case when one row comes from G_0, for instance $q = 3$ and so

$$\begin{aligned} |\hat{\mathbf{o}}_1 G^A \wedge \hat{\mathbf{o}}_2 G^B \wedge \hat{\mathbf{o}}_3 G^C| &= |\mathbf{u}_1 \wedge \mathbf{u}_3 \wedge \mathbf{1}| \qquad \text{(N28)} \\ &= |\mathbf{u}_1 \wedge \mathbf{u}_3| \\ &= 0. \end{aligned}$$

In the second line, we have use that $a \cdot 1 = a$ extends to vectors so that in general $|\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{1}| = |\mathbf{a} \wedge \mathbf{b}|$. The last line uses assumption 2 of Lemma 2. Indeed, whenever one (or more) of the rows is $\mathbf{1}$, we will be able to deploy assumption 1 (or 2) of Lemma 2. This enumerates all possible cases and confirms that Eq. (N8) always holds, therefore proving the main transversality statement of Lemma 2.

Lastly, that Eqs. (N20) to (N22) satisfy assumptions 1-3 of Lemma 2 is easily verified. For example,

$$\begin{aligned} |\mathbf{u}_1 \wedge \mathbf{u}_2 \wedge \mathbf{u}_3| &= (\mathbf{u}_1)_1 (\mathbf{u}_2)_1 (\mathbf{u}_3)_1 + \sum_{j=2}^{8} (\mathbf{u}_1)_j (\mathbf{u}_2)_j (\mathbf{u}_j)_j \\ &= 1 + \sum_{j=2}^{8} 0 = 1, \end{aligned}$$

where in the first line we split off the $j = 1$ case from the rest of the summation to highlight that this is the only non-zero term. A deeper explanation is provided by noticing that the example vectors $\mathbf{u}_t$ correspond to generators of a Reed-Muller code for which these properties are well-known [132]. □

3. Trading space and time

Here we construct a magic state distillation protocol from the G-matrix representation of Appendix N 2 that minimizes space requirements. The intuition is that one never encodes into the full codespace but rather converts the CCZ$^{\otimes n}$ gate into a product of n conjugated CCZ gates that we can think of as being conjugated by some partial encoding unitary.

In particular, consider some G^D-matrix representing an $[[n, k, d]]$ code and a unitary V^D such that

$$V^D |\mathbf{x}\rangle |\mathbf{u}\rangle |\mathbf{0}\rangle = |(\mathbf{x}, \mathbf{u}) G^D\rangle, \qquad \text{(N29)}$$

where $\mathbf{x}$ is a length k bit-string and $\mathbf{u}$ is length m (recall m the number of rows in G_0^D). Furthermore, it is known that such a unitary V^D can be found that is Clifford and composed solely of CNOT gates [128]. It follows that

$$\begin{aligned} V^D |\mathbf{x}\rangle |+^{\otimes m}\rangle |\mathbf{0}\rangle &= 2^{m/2} \sum_{\mathbf{u} \in \mathbb{F}_2^m} V^D |\mathbf{x}\rangle |\mathbf{u}\rangle |\mathbf{0}\rangle \qquad \text{(N30)} \\ &= 2^{m/2} \sum_{\mathbf{u} \in \mathbb{F}_2^m} |(\mathbf{x}, \mathbf{u}) G^D\rangle \\ &= |\mathbf{x}\rangle_L, \end{aligned}$$

where the second line uses Eq. (N29) and the last line uses Eq. (N5). This confirms that V^D is an encoding unitary for the code associated with G^D. To encode the logical state $|+^{\otimes k}\rangle_L$, we simply use linearity so that

$$V^D |+^{\otimes k}\rangle |+^{\otimes m}\rangle |\mathbf{0}\rangle = |+^{\otimes k}\rangle_L. \qquad \text{(N31)}$$

Given three codeblocks, we can encode simultaneously with $V = (V^A \otimes V^B \otimes V^C)$. The all $|+\rangle$ state encoded across three codeblocks is then

$$V(|+^{\otimes k}\rangle |+^{\otimes m}\rangle |\mathbf{0}\rangle)^{\otimes 3} = |+^{\otimes 3k}\rangle_L. \qquad \text{(N32)}$$

A standard recipe for magic state distillation protocols [30, 33, 127] is to encode into logical $|+\rangle$ states, perform tranversal non-Cliffords as follows

$$\begin{aligned} \mathrm{CCZ}^{\otimes n} V(|+^{\otimes k}\rangle |+^{\otimes m}\rangle |\mathbf{0}\rangle)^{\otimes 3} &= (\mathrm{CCZ} |+^{\otimes 3}\rangle_L)^{\otimes k} \\ &= |\mathrm{CCZ}\rangle_L^{\otimes k}, \qquad \text{(N33)} \end{aligned}$$

which produces k logical CCZ states. Decoding gives

$$\begin{aligned} &(V^\dagger \mathrm{CCZ}^{\otimes n} V)(|+^{\otimes k}\rangle |+^{\otimes m}\rangle |\mathbf{0}\rangle)^{\otimes 3} \qquad \text{(N34)} \\ &= V^\dagger |\mathrm{CCZ}\rangle_L^{\otimes k} \qquad \text{(N35)} \\ &= |\mathrm{CCZ}^{\otimes k}\rangle |+^{\otimes 3m}\rangle |\mathbf{0}^{\otimes 3}\rangle, \end{aligned}$$

where in the last line we have slightly abused qubit order-

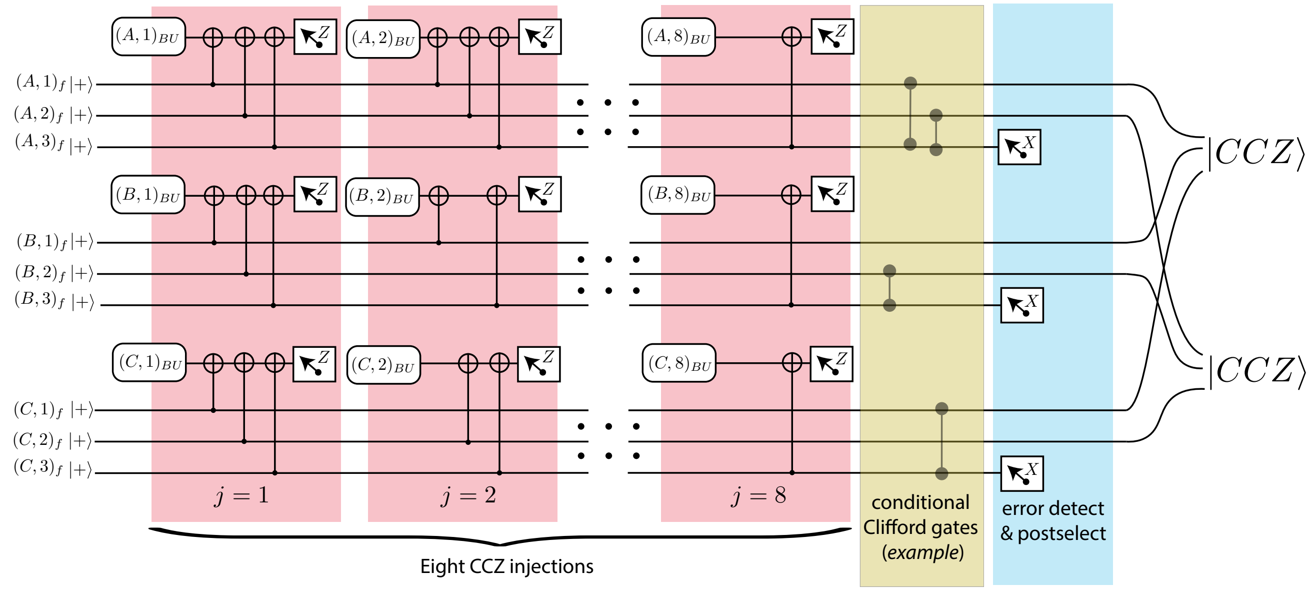


FIG. 58. A magic state distillation protocol for 8CCZ → 2CCZ with the eight CCZ injections performed using Algorithm 2. Qubit labels of form $(D,i)_f$ and $(D,j)_{BU}$ follow notation of Definition 2. For each j, the triple of qubits $(A,1)_{BU}$, $(B,2)_{BU}$ and $(C,3)_{BU}$ are prepared in a noisy CCZ states (e.g. using `BUTOF`) but this preparation is not shown. We show explicitly the CNOT gates for the first 2 steps and the last step, but omit the middle steps for brevity. The full circuit is reproducible using Eqs. (N23) to (N25) to specify the CNOT pattern as outlined in Algorithm 2.

ing to collect together the physical $|\mathrm{CCZ}\rangle$ state output. In the case of a detectable error, at least one of the $|+^{\otimes 3m}\rangle$ qubits will be phase flipped and detected by an X-measurement.

To reduce space overhead, we observe that the $|\mathbf{0}^{\otimes 3}\rangle$ qubits effectively play no role here. Furthermore, the unitary $V^\dagger \mathrm{CCZ}^{\otimes n} V$ acts non-trivially only on the first $3(k+m)$ qubits, so that the $|\mathbf{0}^{\otimes 3}\rangle$ qubits (a total of $3(n-m-k)$ qubits) are truly surplus to requirement. Using our earlier notation $V^\dagger \mathrm{CCZ}^{\otimes n} V = \prod_j V^\dagger \mathrm{CCZ}_j V$ where CCZ_j acts on qubit j of each block, one then has that

$$\begin{aligned}
&(V^\dagger \mathrm{CCZ}_j V)|(\mathbf{x},\mathbf{u})\rangle|(\mathbf{y},\mathbf{v})\rangle|(\mathbf{z},\mathbf{w})\rangle \qquad \text{(N36)}\\
&= (-1)^{[(\mathbf{x},\mathbf{u})G^A]_j[(\mathbf{y},\mathbf{v})G^B]_j(\mathbf{z},\mathbf{w})G^C]_j}|(\mathbf{x},\mathbf{u})\rangle|(\mathbf{y},\mathbf{v})\rangle|(\mathbf{z},\mathbf{w})\rangle\\
&= (-1)^{(\mathbf{x},\mathbf{u})[G^A]_j\cdot(\mathbf{y},\mathbf{v})[G^B]_j\cdot(\mathbf{z},\mathbf{w})[G^C]_j}|(\mathbf{x},\mathbf{u})\rangle|(\mathbf{y},\mathbf{v})\rangle|(\mathbf{z},\mathbf{w})\rangle,
\end{aligned}$$

where $[\ldots]_j$ denotes the j^{th} element of the vector inside or the j^{th} column of a matrix. Notice, we have suppressed the presence of the redundant $|\mathbf{0}\rangle$ qubits.

In Appendix N 4 we describe two concrete implementations of the $V\mathrm{CCZ}V^\dagger$ gates. Of course, it is crucial that the space reduction and $V\mathrm{CCZ}V^\dagger$ implementation does not distort the way errors propagate and that error correction properties are retained, which is proven from first principles in Appendix N 5.

4. Implementing Conjugated-CCZ gates

Here we give explicit implementations for the 8 conjugated-CCZ gates described in Eq. (N36). Any such gate can be realized using a single CCZ magic state and we give further details for two different implementations: the first implementation uses CNOT gates and single qubit measurements (Appendix N 4 a); the second implementation uses only multi-qubit Pauli measurements via lattice surgery (Appendix N 4 b).

Herein, we label qubits as follows.

Definition 2 (Qubit labels)**.** *Consider a magic state distillation protocol using n noisy CCZ magic states and G^D matrices with $k+m$ rows. We label each input magic states by $(D,j)_{BU}$ where $j\in[1,n]$ labels which CCZ state the qubit is part of and $D\in\{A,B,C\}$ distinguishes the 3-qubits within a CCZ state. The BU subscript highlights that these are input noisy state qubits possibly produced by* `BUTOF`*. We also have $3(m+k)$ qubits that we call factory qubits and label $(D,i)_f$ with a subscript f for factory and where $D\in\{A,B,C\}$ and $i\in[1,m+k]$.*

Notice that the qubit labels assume we have made a spacetime tradeoff, so the factory qubits refer to the $3(m+k)$ qubits prepared in a $|+\rangle$ state. The $3(n-m-k)$ qubits described earlier as being in the $|0\rangle$ state are omitted as they are surplus to requirement. For the code of interest, (recall Eqs. (N23) to (N25) there are 9 factory qubits and 24 *BU* qubits, though the *BU*-qubits do not all need to be prepared at the same time and can be encoded in smaller distance codeblocks.

Algorithm 2: A CNOT circuit realizing $V^{\dagger}\mathrm{CCZ}^{\otimes n}V$ as defined in Appendix N 3. Uses a trio of G matrices with n rows and $k+m$ columns. Qubit label convention given in Definition 2.

```
1. For each $j \in \{1, \dots, n\}$
    (a) For each $D \in \{A, B, C\}$ and each $i$ such that $[G^D]_{i,j} = 1$ do a CNOT with control $(D, i)_f$ and target $(D, j)_{BU}$.
    (b) For each $D \in \{A, B, C\}$ measure magic state qubit $(D, j)_{BU}$ in the $Z$ basis and record the outcome as $m_j^D \in \{0, 1\}$.
    (c) For each pair $m_j^D, m_j^{D'} = 1$, apply a $Z$ correction to every qubit $(D'', p)_f$ for which $G^{D''}_{p,j} = 1$.
    (d) For each pair $m_j^D, = 1$, apply a $CZ$ correction to every pair of qubits $(D', p)_f$ and $(D'', q)_f$ for which $G^{D'}_{p,j} = G^{D'}_{q,j} 1$.
```

A circuit using this labeling appears later in Fig. 58.

a. *Injection with CNOT gates and single qubit measurements*

To perform the required sequence of n conjugated-CCZ gates from Eq. (N36), we may implement Algorithm 2. In Items 1b and 1c of Algorithm 2, the indices (D, D', D'') should be read as distinct triples from the set $\{A, B, C\}$. For example, if $D = A$ and $D' = C$ then one infers $D'' = B$. Furthermore, these adaptive Clifford corrections commute with the rest of the circuits and so can all be postponed until later. We illustrate some of the steps in Fig. 58.

Next, we calculate the action of the circuit described by Algorithm 2 for one particular j value. With respect to the factory qubit basis states, we have

$$|\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle = |(\mathbf{x}, \mathbf{u})\rangle|(\mathbf{y}, \mathbf{v})\rangle|(\mathbf{z}, \mathbf{w})\rangle, \tag{N37}$$

where we have broken the state up into 3 blocks corresponding to indices A, B and C. For example, qubit $(A, i)_f$ is in state a_i. Furthermore, a_i equals x_i when $i \leq k$ and u_i when $i > k$. For each $D = \{A, B, C\}$, we implement CNOT gates targeted on the magic state qubit $(D, j)_{BU}$ and controlled on qubits $(D, i)_{BU}$ indicated by $[G^D]_{i,j} = 1$.

Therefore, for $D = A$ the target $(A, j)_{BU}$ qubit is flipped precisely when

$$\sum_i [G^A]_{i,j} a_i = [\mathbf{a}G^A]_j = 1 \pmod 2, \tag{N38}$$

where the summation has been changed to matrix multiplication. Recall $[\mathbf{a}G^A]_j$ just means the j^{th} element of vector $\mathbf{a}G^A$. Similar expressions hold for $D = B, C$. The CCZ magic state is given by

$$|\mathrm{CCZ}\rangle = 2^{-3/2} \sum_{y_D \in \mathbb{F}_2} (-1)^{y_A y_B y_C} |y_A\rangle|y_B\rangle|y_C\rangle. \tag{N39}$$

Ignoring $2^{-3/2}$ for brevity, the CNOTs of Algorithm 2 act as follows on a $|\mathrm{CCZ}\rangle|\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle$ state

$$\sum_{y_D \in \mathbb{F}_2} (-1)^{y_A y_B y_C} |y_A\rangle|y_B\rangle|y_C\rangle|\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle$$
$$\rightarrow \sum_{y_D \in \mathbb{F}_2} (-1)^{y_A y_B y_C} |y'_A\rangle|y'_B\rangle|y'_C\rangle|\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle,$$

with

$$\begin{aligned} y'_A &= y_A + [\mathbf{a}G^A]_j \\ y'_B &= y_B + [\mathbf{b}G^B]_j \\ y'_C &= y_C + [\mathbf{c}G^C]_j. \end{aligned} \tag{N40}$$

We follow these CNOTs by measurement of the BU-qubits in the Z basis, which are afterwards discarded. Assuming measurement outcomes $|m_j^A\rangle|m_j^B\rangle|m_j^C\rangle$ then the only non-vanishing terms have $m_j^D = y'_D$, so

$$y_A = m_j^A + [\mathbf{a}G^A]_j \tag{N41}$$
$$y_B = m_j^B + [\mathbf{b}G^B]_j \tag{N42}$$
$$y_C = m_j^C + [\mathbf{c}G^C]_j. \tag{N43}$$

Discarding the BU-qubits, we get

$$|CZZ\rangle|\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle \rightarrow (-1)^{f(\mathbf{m})}|\mathbf{a}\rangle|\mathbf{b}\rangle|\mathbf{c}\rangle, \tag{N44}$$

where the phase exponent depends on the measurement outcomes $\mathbf{m} = (m_j^A, m_j^B, m_j^C)$ as follows

$$f(\mathbf{m}) = (m_j^A + [\mathbf{a}G^A]_j)(m_j^B + [\mathbf{b}G^B]_j)(m_j^C + [\mathbf{c}G^C]_j). \tag{N45}$$

The value of this phase-exponent was originally $y_A y_B y_C$ but with the substitutions determined by Eqs. (N41) to (N43) we get expression Eq. (N45).

In the case of a $\mathbf{m} = (0, 0, 0)$ projection, we get the phase

$$f(0, 0, 0) = [\mathbf{a}G^A]_j [\mathbf{b}G^B]_j [\mathbf{c}G^C]_j, \tag{N46}$$

so that after switching notation by using Eq. (N37) we get the desired phase in Eq. (N36). However, for non-zero measurement outcomes we have

$$f(\mathbf{m}) = f(0, 0, 0) + g(\mathbf{m}), \tag{N47}$$

where $g(\mathbf{m})$ represents the remaining terms in the expansions of Eq. (N45). We can see that these remaining terms will be quadratic in the variables $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ and so represent Clifford corrections: the quadratic terms correspond to a circuit of CZ gates, the linear terms correspond to a circuit of Z gates, and the constant term gives a global phase.

For example, consider the case when $\mathbf{m} = (1, 0, 0)$ so

$$\begin{aligned} g(\mathbf{m}) &= [\mathbf{b}G^B]_j[\mathbf{c}G^C]_j \\ &= \sum_{p,q} G^B_{p,j} G^C_{q,j} b_p c_q . \end{aligned} \tag{N48}$$

This is corrected by a CZ between qubits $(B, p)_f$ and $(C, q)_f$ for every $\{p, q\}$ such that $G^B_{p,j} = G^C_{q,j} = 1$. This correction precisely matches the rule given in Items 1b and 1c of Algorithm 2. It is straightforward but tedious to verify that the corrections of Algorithm 2 always give the desired phase needed to cancel $(-1)^{g(\mathbf{m})}$.

b. *Injection using lattice surgery*

Here we provide an alternative formulation of the conjugated CCZ injection from that presented in Appendix N 4 a. Instead of using a CNOT circuit, the injection procedure will be described entirely in terms of multi-qubit Pauli operator measurements, as this is the natural set of operations in lattice surgery implementations. We have already discussed the key ideas of this mapping in Appendix L and Section V. Here we wish to allow for the option of performing a lattice surgery operation between a repetition code clock (really just a $d_x = 1$ surface code) and thin surface codes, and we present an example lattice surgery diagram in Fig. 59.

In general, imagine a circuit that performs the following: (i) do n CNOT gates targeted on qubit 0 and controlled on qubits 1 to n; (ii) measure Z_0; (iii) discard qubit zero. This is equivalent to the following measurement driven procedure: (i') measure multi-qubit Pauli $\prod_{j=0}^{n} Z_j$; (ii') measure single-qubit Pauli X_0 and discard; (iii') if second step gives "-1" outcome perform a $\prod_{j=1}^{n} Z_j$ correction. In the bottom diagram of 13, we prove equivalence of these approaches through a series of circuit identities (illustrated for the $n = 3$ case). Applying this equivalence to Algorithm 2 we obtain Algorithm 3.

Note that in a Pauli measurement scheme, we never perform the Clifford corrections. Rather whenever there is a subsequent Pauli measurement P, if the Clifford correction register contains C, we instead measure $CPC^\dagger$. The corrections in Algorithm 3 commute with all the measurements here, and so can be postponed until later.

5. Error propagation and detection

Here we discuss the effect of noisy $|\mathrm{CCZ}\rangle$ states used in the `TDTOF` protocol. For now, we assume all encoded Clifford gates are ideal, but later we will relax this assumption.

Algorithm 3: A Pauli-measurement scheme realizing $V^\dagger \mathrm{CCZ}^{\otimes n} V$ as defined in Appendix N 3. Uses a trio of G^D matrices with n rows and $k + m$ columns. Qubit label convention given in Definition 2.

```
1. For each $j \in \{1, \dots, n\}$
    (a) For each $D \in \{A, B, C\}$: measure a multi-qubit $Z$ operator, with support on $(D, j)_{BU}$ and $(D, i)_f$ for every $i$ such that $[G^D]_{i,j} = 1$ and record the outcome as $\omega^D_j \in \{0, 1\}$.
    (b) For each $\omega^D_j = 1$ , record a $Z$ correction to qubit $(D, i)_f$ for every $i$ such that $[G^D]_{i,j} = 1$.
    (c) For each $D \in \{A, B, C\}$ measure the single-qubit Pauli $(D, j)_{BU}$ in the $X$ basis and record the outcome as $m^D_j \in \{0, 1\}$.
    (d) For each pair $m^D_j, m^{D'}_j = 1$, record a $Z$ correction to every qubit $(D'', p)_f$ for which $G^{D''}_{p,j} = 1$.
    (e) For each pair $m^D_j, = 1$, record a $CZ$ correction to every pair of qubits $(D', p)_f$ and $(D'', q)_f$ for which $G^{D'}_{p,j} = G^{D''}_{q,j} = 1$.
```

To be precise regarding error propagation we introduce the following language

Definition 3. *Given a $|CCZ\rangle$ magic state, we say it has error pattern $\mathbf{e} = (e^A, e^B, e^C) \in \mathbb{F}_2^3$ error if it is in the state*

$$E|CCZ\rangle = Z[\mathbf{e}]|CZZ\rangle \tag{N49}$$

$$= (Z^{e^A} \otimes Z^{e^B} \otimes Z^{e^C})|CCZ\rangle. \tag{N50}$$

Given n such states, for each $j \in [1, n]$ we use $\mathbf{e}_j = (e^A_j, e^B_j, e^C_j)$ to denote the error for the j^{th} $|CCZ\rangle$ state, so that

$$E|CCZ\rangle^{\otimes n} = \bigotimes_{j=1}^{n} \left(Z[\mathbf{e}_j]|CCZ\rangle\right). \tag{N51}$$

We say an error has w fault-locations if there are w non-zero $\mathbf{e}_j$. Furthermore, for $D \in \{A, B, C\}$ we define

$$\mathbf{e}^D = (e^D_1, e^D_2, \dots, e^D_n). \tag{N52}$$

The distinction between our notion of fault-locations and the usual Hamming weight of the concatenated string $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ is important because many methods of preparing a noisy $|\mathrm{CCZ}\rangle$ state will lead to errors such as $Z \otimes Z \otimes Z$

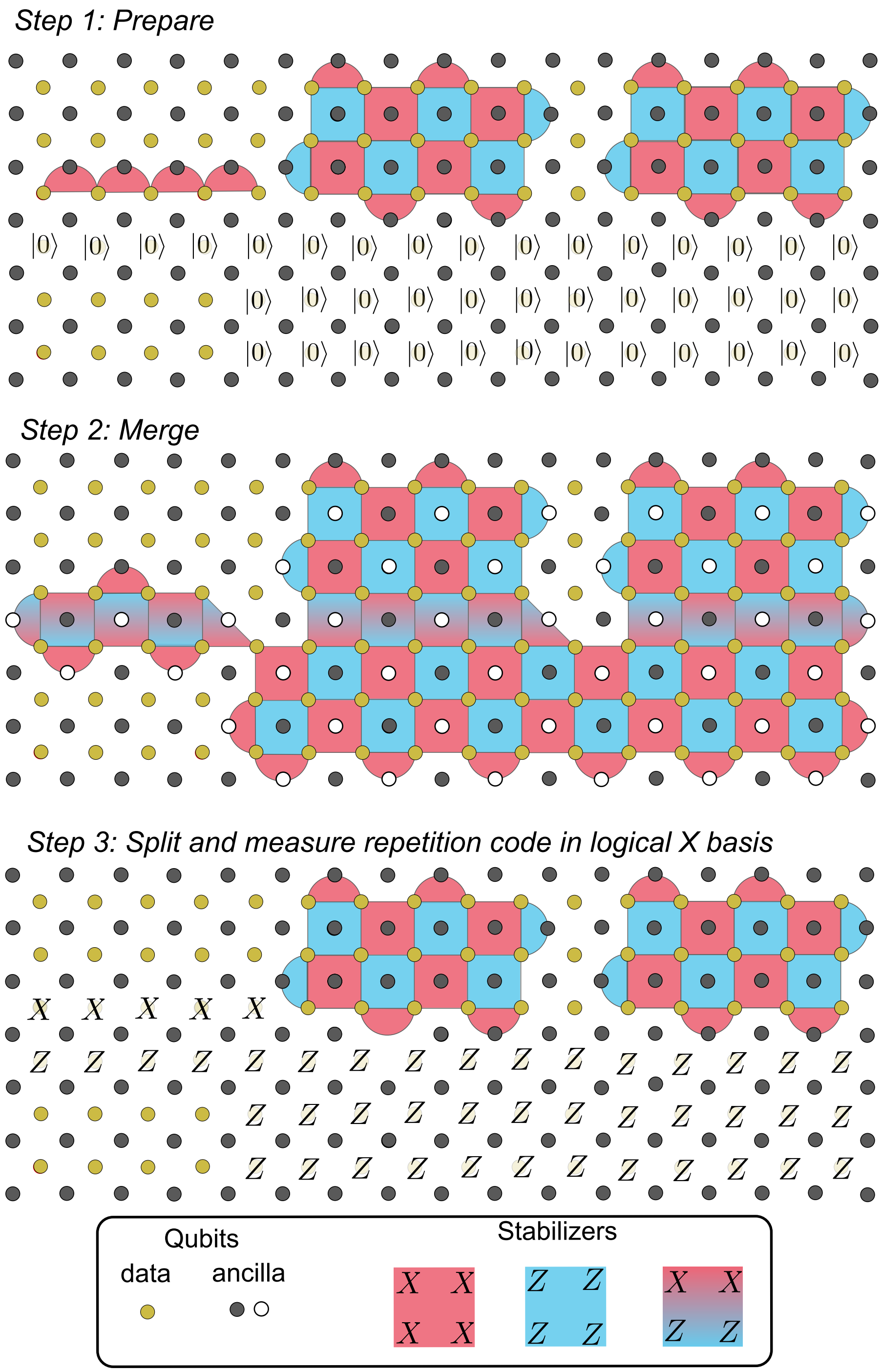


FIG. 59. Lattice surgery to measure a multi-patch Pauli measurement between three codes blocks: one repetition code logical qubit and two thin surface code logical qubits. See Fig. 50 for comparison. This provides the principle building block for the execution of Item 1a. In `TDTOF`, a multi-patch Pauli measurement is always followed by a single-qubit measurement of the repetition code and here we combine this with the third (split) step of lattice surgery. The gradient coloured squares represent where an $X \otimes X \otimes Z \otimes Z$ stabilizer measurement called a dislocation. Multiplying the outcome of the stabilizers labeled with a white dot, gives the outcome of the $Z_L \otimes Z_L \otimes Z_L$ multi-patch Pauli measurement (this is how we obtain the outcomes labeled ω_j^D in Item 1a of Algorithm 3). To ensure fault-tolerance of this measurement outcome, we repeat these stabilizer measurements d_m times as discussed in Appendix M 3. Afterwards all qubits are measured in Z basis, with their product determining the operator $Z_L \otimes \mathbb{1} \otimes \mathbb{1}$ (this is how we obtain the outcomes labeled m_j^D in Item 1a of Algorithm 3). We discuss in Appendix N 8 the effect of errors in lattice surgery due to using finite size code blocks. If we wish to instead measure $X_L \otimes Z_L \otimes Z_L$ then we do not use the dislocation on the repetition code block.

that could have a comparable probability to a single qubit error $Z \otimes \mathbb{1} \otimes \mathbb{1}$. Indeed, we will typically be interested in knowing how many $|\mathrm{CCZ}\rangle$ states are affected by an arbitrary error, though we assume errors are uncorrelated between different $|\mathrm{CCZ}\rangle$ states. Errors propagate as follows

Claim 1 (How errors propagate)**.** *Consider an implementation of Algorithm 3 using noisy CCZ states with Pauli Z error described by* $\{\mathbf{e}^A, \mathbf{e}^B, \mathbf{e}^C\}$ *as in Definition 3. For each* $D \in \{A, B, C\}$*, let*

$$\mathbf{w}^D = \mathbf{e}^D G^D. \tag{N53}$$

The output of Algorithm 3 differs from the ideal case by an error $Z[\mathbf{w}^A] \otimes Z[\mathbf{w}^B] \otimes Z[\mathbf{w}^C]$ *on the factory qubits and where the tensor product represents the three different codeblocks. Identifying the last m qubits of each block as check qubits, we can partition the* $\mathbf{w}^D$ *into two parts as follows*

$$\mathbf{u}^D = \mathbf{e}^D G^D_1, \tag{N54}$$

$$\mathbf{v}^D = \mathbf{e}^D G^D_0. \tag{N55}$$

Claim 1 tells us that Z errors propagate through Algorithm 3 in a manner that is isomorphic to their propagation through error correction codes represented by the corresponding G-matrices.

We can prove Claim 1 by considering how a single Z error on a BU-qubit propagates onto a factory qubit under Algorithm 3. Since an error on a factory qubit propagates to the end of the circuit, they compose independently. Consult the last circuit of Fig. 13 and consider a Z error on the top qubit. It commutes with the multi-qubit Pauli Z measurement but flips the final single qubit X measurement. The outcome for this X measurement decides whether to apply Z to the qubits below. In other words, a Z on the top qubit propagates to all the qubits below. In Algorithm 3, when operating on qubit $(D, j)_{BU}$ we apply the circuit of Fig. 13 to sets of factory qubits identified by $(D, i)_f$ whenever $G^D_{i,j} = 1$. Therefore, a Z error on $(D, j)_{BU}$ occurs whenever $\mathbf{e}^D_j = 1$ and will propagate to every $(D, i)_f$ for which $G^D_{i,j} = 1$. Summing over all $j \in [1, n]$ magic state injections, factory qubit $(D, i)_f$ will have a Z error if

$$\sum_j \mathbf{e}^D_j G^D_{i,j} = \left[\mathbf{e}^D G^D\right]_i = 1 \pmod 2. \tag{N56}$$

Since the i^{th} qubit is Z-flipped according to the i^{th} element of vector $\mathbf{w}^D := \mathbf{e}^D G^D$, this vector describes the Z-error distribution on factory block D. Splitting G^D into its block matrix components G^D_1 and G^D_0 gives Eq. (N54).

We have already described the main components of the magic state distillation routine, but for completeness we recap how they fit together in Algorithm 4. Combining our previous results, we have that

Algorithm 4: A complete magic state distillation routine using a space-time tradeoff and multi-qubit Pauli measurements. It assumes a trio of G^D matrices of size $n \times (m + k)$ representing $[[n, k, d]]$ codes with CCZ transversality as in Lemma 2. We gave suitable G^D matrices in Eqs. (N23) to (N25) for which we have $n = 8$, $k = 2$, $d = 2$ and $m = 1$. Qubit label convention given in Definition 2

1. Prepare $3k$ factory qubits $(D, j)_f$ in the $|+\rangle$ state.
2. Prepare n noisy CCZ magic states (e.g. using `BUTOF`);
3. Perform injections using Algorithm 3.
4. Measure $3m$ check qubits $(D, i)_f$ for all $i = m, \ldots, m + k$ and ACCEPT on $|+\rangle$ for every outcome.

Claim 2 (Distillation)**.** *Consider an implementation of Algorithm 4 using* G^D *matrices of size* $n \times (m + k)$ *satisfying Lemma 2 and using n noisy CCZ states with Pauli Z error described by* $\{\mathbf{e}^A, \mathbf{e}^B, \mathbf{e}^C\}$ *as in Definition 3. The protocol will ACCEPT whenever*

$$\mathbf{v}^D = \mathbf{e}^D G^D_0 = \boldsymbol{0} \tag{N57}$$

for every $D \in \{A, B, C\}$*. Furthermore, provided for every* $D \in \{A, B, C\}$ *we have*

$$\mathbf{u}^D = \mathbf{e}^D G^D_1 = \boldsymbol{0}, \tag{N58}$$

the protocol outputs $|CCZ\rangle^{\otimes k}$*. Furthermore, if the* j^{th} $|CCZ\rangle$ *state has error* $Z[\mathbf{e}_j]$ *with probability* $\mathbb{P}_j(\mathbf{e}_j)$ *then the probability of passing the error detection test is*

$$P_{\text{acc}} = \sum_{\mathbf{e}^D:[\mathbf{e}^D G^D_0 = \mathbf{0}]\forall D} \prod_j \mathbb{P}_j(\mathbf{e}_j), \tag{N59}$$

and the output fidelity is

$$F = \frac{1}{P_{\text{acc}}} \left(\sum_{\mathbf{e}^D:[\mathbf{e}^D G^D_1 = \mathbf{0}]\forall D} \prod_j \mathbb{P}_j(\mathbf{e}_j) \right). \tag{N60}$$

First consider when there are no Z errors. From Appendix N 4 b we see that Algorithm 3 will (when there is no Z noise) apply $\mathrm{CCZ}^{\otimes k}$ to the $3k$ qubits labeled $(D, i)_f$ with $i \leq k$. The check qubits with $i > k$ are unaffected. Therefore, the check qubits should still be in the $|+\rangle$ state and give "+1" in response to an X measurement. This confirms that the protocol acts correctly in the ideal case.

When there are one or more Z errors, Claim 1 shows that the check qubits remain unflipped if and only if

$\mathbf{u}^D = \mathbf{e}^D G_0 = \mathbf{0}$ for all D. Furthermore, if $\mathbf{v}^D = \mathbf{e}^D G_1 = \mathbf{0}$ then Claim 1 tells us whether there are no Z errors propagated onto the factory qubits forming the output $|\mathrm{CCZ}^{\otimes k}\rangle$ state. The formulae for P_{acc} and F follow by simply summing over the probabilities of these events.

For the remainder of this subsection, we consider the special case when $G_0^D = (1, 1, \ldots, 1)$ as we have in Eqs. (N23) to (N25). Then, the state will pass the error detection test whenever

$$\mathbf{v}^D = \mathbf{e}^D G_0^D = \sum_j \mathbf{e}_j^D = \mathbf{0}. \tag{N61}$$

If there are no fault-locations so $\mathbf{e}_j = 0$ for all j, then the protocol will ACCEPT. If there is a single fault location, so a single j for which $\mathbf{e}_j = (e_j^A, e_j^B, e_j^C) \neq (0,0,0)$ then the error must be detected as there is no chance for cancellation. If there are two fault-locations for which $\mathbf{e}_j \neq \mathbf{0}$ and $\mathbf{e}_i \neq \mathbf{0}$ then the errors will go undetected only if they cancel exactly, so $\mathbf{e}_j = \mathbf{e}_i$. Therefore, to leading order

$$P_{\mathrm{acc}} = \prod_{j=1}^{n} \mathbb{P}_j(\mathbf{0}) + \sum_{\{i,j\} \subset [1,n], \mathbf{e} \neq \mathbf{0}} \mathbb{P}_i(\mathbf{e}) \mathbb{P}_j(\mathbf{e}) \prod_{\ell \neq i,j} \mathbb{P}_\ell(\mathbf{0}) + \ldots . \tag{N62}$$

For instance, let us consider an i.i.d depolarizing noise model such that $\mathbb{P}_j(\mathbf{0}) = 1-\epsilon$ and $\mathbb{P}_j(\mathbf{e} \neq \mathbf{0}) = \epsilon/7$. There are 7 types of fault $\mathbf{e} \neq \mathbf{0}$ and 28 pairs of possible locations, making 196 different undetected two fault-location errors, so that

$$P_{\mathrm{acc}} = (1-\epsilon)^8 + 196 \left(\frac{\epsilon}{7}\right)^2 (1-\epsilon)^6 + \ldots . \tag{N63}$$

To leading order, the infidelity $1-F$ is upper bounded by the probability of an undetected two fault-location error,

$$1 - F \leq 196 \left(\frac{\epsilon}{7}\right)^2 (1-\epsilon)^6 + \ldots . \tag{N64}$$

However, some undetected two fault-location errors will not lead to an output error (i.e. when $[\mathbf{e}^D G_1^D = \mathbf{0}] \forall D$). For the G^D matrices of interest (Eqs. (N23) to (N25)), by brute force counting we find that 184 of the undetected 196 two fault-location errors will lead to an error. The 12 harmless faults are listed in Table IX and will return to play an important role in noise tailoring of Appendix N 6. Therefore, we can tighten Eq. (N64) to

$$\begin{aligned} 1 - F &\leq 184 \left(\frac{\epsilon}{7}\right)^2 (1-\epsilon)^6 + \ldots & \text{(N65)} \\ &\sim 3.755\epsilon^2 + O(\epsilon^3). & \text{(N66)} \end{aligned}$$

Therefore, we have quadratic error suppression with quite a small constant factor for depolarizing noise. In the main text, we usually quote the error per TOF state and since the protocol outputs two TOF states, we have $\epsilon_{TD} := \frac{1}{2}(1-F)$. For the depolarizing noise model this leads to:

$$\epsilon_{TD} =\sim 1.878\epsilon^2 + O(\epsilon^3). \tag{N67}$$

a. *Truncation errors*

While we give expressions up to second order, these summations can be easily performed to higher order and any truncation error can be controlled. If we perform calculations up to $t_{\max}$ fault-locations, then the truncation error can be easily upper-bounded by assuming that every error above the cut-off leads to an undetected output error so that we have the rigorous bound

$$1 - F \leq (1 - F_{t_{\max}}) + \sum_{t=t_{\max}+1}^{8} \binom{8}{t} 7^t \left(\frac{\epsilon}{7}\right)^t (1-\epsilon)^{8-t}, \tag{N68}$$

where $(1 - F_{t_{\max}})$ is a estimate counting up to $t_{\max}$ fault-locations and the additional summation is our bound on the truncation error. In all subsequent numerical calculations we have confirmed the possible truncation error is many orders of magnitude smaller than the estimated error. For instance, using $t_{\max} = 3$ then for $\epsilon \leq 10^{-4}$ the truncation error is no more than $3 \cdot 10^{-16}$ and therefore negligible.

In practice, the error distribution from `BUTOF` is far from depolarizing and this is further skewed when we account for Clifford noise (see Appendix N 8). However, truncation error can be estimated of any noise model and controlled in the above manner. Furthermore, one can also tailor the protocol to the noise profile (see Appendix N 6).

b. *Generic noise*

We have show Algorithm 4 tolerates Z error noise. Next, we show it also tolerates X noise on the noisy $|\mathrm{CCZ}\rangle$ states. Abstracting away the details of Claim 2, the protocol maps pure states as follows

$$Z[\mathbf{e}]|\mathrm{CCZ}\rangle^{\otimes n} \rightarrow \det(\mathbf{e}) Z[\nu(\mathbf{e})]|\mathrm{CCZ}\rangle^{\otimes k}, \tag{N69}$$

where $\det(\mathbf{e}) = 0, 1$ depending on whether the error $\mathbf{e}$ is detected or not, and the output error is some function ν of $\mathbf{e}$. Formulae for det and ν can be extracted from Claim 2, but here it is useful to ignore these details. Going to density matrices, we can write

$$\begin{aligned} \rho &:= |\mathrm{CCZ}\rangle\langle\mathrm{CCZ}|^{\otimes n} & \text{(N70)} \\ \sigma &:= |\mathrm{CCZ}\rangle\langle\mathrm{CCZ}|^{\otimes k}. \end{aligned}$$

Because $Z[\mathbf{e}]|\mathrm{CCZ}\rangle^{\otimes n}$ form an orthonormal basis, any input mixed state can be written as

$$\tilde{\rho} := \sum_{\mathbf{e},\mathbf{f}} A_{\mathbf{e},\mathbf{f}} Z[\mathbf{e}] \rho Z[\mathbf{f}]. \tag{N71}$$

If the state suffered stochastic Z noise then it would be diagonal with respect to this basis, so $A_{\mathbf{e},\mathbf{f}} = 0$ whenever $\mathbf{e} \neq \mathbf{f}$. If there are off-diagonal elements $A_{\mathbf{e},\mathbf{f}} \neq 0$ these could be eliminated by applying a random twirl using the Clifford operators that stabilize $|\mathrm{CCZ}\rangle$. However, this would add unnecessary Clifford gates as these off-diagonals are unimportant, as we now show.

By Eq. (N69) and linearity, we have

$$\tilde{\rho} \rightarrow \tilde{\sigma} = \sum_{\mathbf{e},\mathbf{f}} \mathrm{det}(\mathbf{e})\mathrm{det}(\mathbf{f}) A_{\mathbf{e},\mathbf{f}} Z[\nu(\mathbf{e})]\sigma Z[\nu(\mathbf{f})]. \quad \text{(N72)}$$

Because any physical process does not increase the trace of any terms, there is no way for off-diagonal elements (with $\mathbf{e} \neq \mathbf{f}$) to be mapped to on-diagonal elements (with $\nu(\mathbf{e}) \neq \nu(\mathbf{f})$). Since the success probability and fidelity only depend on the output diagonal elements, we conclude that our figures of merit only depend on the diagonal $A_{\mathbf{e},\mathbf{e}}$ elements. In other words, the success probability and output fidelity are unchanged whether or not we twirl the initial state. In all numerics presented, whenever the input magic states suffer a mix of Z and X noise, we have calculated the exact $\tilde{\rho}$ matrix, extracted the diagonal elements and used them to build an equivalent stochastic Z noise model. Consequently, any error on a single $|\mathrm{CCZ}\rangle$ state appears as a stochastic mixture of Z errors at 1 fault location.

After the protocol is complete, we can twirl the output states to ensure that the infidelity matches the trace norm error of the output states. Though again, this twirl is never actually performed but included into the Clifford record to modify Pauli-measurements used to inject the magic state into the algorithm.

6. Noise tailoring through Clifford symmetries

There is some freedom in how injections are scheduled and whether to include certain Clifford gates in `TDTOF` protocol. A $|\mathrm{CCZ}\rangle$ gate is invariant under permutation of qubits A, B and C. More generally, there are Clifford symmetries C such that $C|\mathrm{CCZ}\rangle = |\mathrm{CCZ}\rangle$. A permutation is a sort of Clifford symmetry, but one that can be realized at no further gate count.

As such, we can add Cliffords or freely permute some of the indices in Algorithm 3. In the ideal case, with no errors, these symmetry operations have no effect. However, they can change the noise model. For qubits with depolarizing noise, the noisy state is invariant under all these symmetries. However, for `BUTOF` the output noise model is very asymmetric and highly skewed towards a Z error on qubit A and so applying symmetry operations can change the protocol's performance. Here we explain the idea of noise tailoring through symmetries and find that the change can be dramatic. Indeed, while the protocol usually quadratically suppresses errors, we can tailor the noise for cubic suppression of 1 error type. To make a clean statement we consider a toy noise model.

Claim 3. *Consider a noise model on $|CCZ\rangle$ states such that for every j it experiences error $Z[\mathbf{e}_j]$ (recall Definition 3) with probability*

$$\mathbb{P}_j(\mathbf{e}_j) := \begin{cases} 1-\epsilon_1-\epsilon_2 & \text{if } \mathbf{e}_j = (0,0,0) \\ \epsilon_1 & \text{if } \mathbf{e}_j = (1,0,0) \\ (\epsilon_2/6) & \text{otherwise} \end{cases}, \quad \text{(N73)}$$

where $\epsilon_2 \ll \epsilon_1$. Directly applying Claim 2 leads to an output infidelity of $O(\epsilon_1^2) + O(\epsilon_2^2) + O(\epsilon_1\epsilon_2)$. However, there exists a set of Clifford symmetries $\{C_j\}$ such that $C_j|CCZ\rangle = |CCZ\rangle$ and if applied at the start of the protocol lead to an output infidelity of $O(\epsilon_1^3) + O(\epsilon_1\epsilon_2) + O(\epsilon_2^2)$.

Consider a set of Clifford symmetries such that

$$C_j Z[\mathbf{e}_j] C_j^\dagger = \pm Z[\mathbf{e}_j M_j], \quad \text{(N74)}$$

where M_j is an invertible 3×3 binary matrix and $\mathbf{e}_j M_j$ represents matrix multiplication. The ± phase will depend on $Z[\mathbf{e}_j]$ but is irrelevant to our analysis. For example, if C_j permutes qubits in Hilbert space then M_j represents the permutation of the indices. Then applying C_j to the input magic states generates a new probability distribution for Z errors

$$\mathbb{P}'_j(\mathbf{e}_j M_j) := \mathbb{P}_j(\mathbf{e}_j). \quad \text{(N75)}$$

Using that M must be invertible, we equivalently have

$$\mathbb{P}'_j(\mathbf{e}_j) := \mathbb{P}_j(\mathbf{e}_j M_j^{-1}). \quad \text{(N76)}$$

Only errors with two fault-locations contribution second order contributions to the output infidelity. Recall from Appendix N 5 that for such an error to go undetected, we must have that $\mathbf{e}_i = \mathbf{e}_j =: \mathbf{e} \neq 0$ for some distinct pair $\{i, j\}$. We have introduce the shorthand $\mathbf{e}$ for whatever nonzero error type is under consideration. This occurs with probability

$$\mathbb{P}'_i(\mathbf{e}M_i)\mathbb{P}'_j(\mathbf{e}M_j)\mathbb{P}'_j(\mathbf{0})^6 = \mathbb{P}_i(\mathbf{e}M_i^{-1})\mathbb{P}_j(\mathbf{e}M_j^{-1})\mathbb{P}_j(\mathbf{0})^6. \quad \text{(N77)}$$

This probability is of size $O(\epsilon_1^2)$ if

$$\mathbf{e}M_i^{-1} = \mathbf{e}M_j^{-1} = (1,0,0), \quad \text{(N78)}$$

and otherwise the probability is smaller: either $O(\epsilon_1\epsilon_2)$, $O(\epsilon_2^2)$ or zero. Inverting again, Eq. (N78) can be converted into

$$\mathbf{e} = (1,0,0)M_i = (1,0,0)M_j. \quad \text{(N79)}$$

It follows that to achieve $O(\epsilon_1^3)$ scaling of output infidelity, we require that for every $\{i, j\}$ pair either

(1⋆) $(1,0,0)M_i \neq (1,0,0)M_j$;

(2⋆) *or* if $(1,0,0)M_i = (1,0,0)M_j$ then fault $\mathbf{e} = (1,0,0)M_j$ corresponds to one of the harmless errors listed in Table IX.

Fault type $\mathbf{e}_i = \mathbf{e}_j = \mathbf{e} =$	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$
Fault locations $\{i,j\} =$	$\{1,2\},\{3,6\},\{4,5\},\{7,8\}$	$\{1,5\},\{2,4\},\{3,7\},\{6,8\}$	$\{1,3\},\{2,6\},\{4,8\},\{5,7\}$

TABLE IX. A list of the errors with two fault-locations that are undetected but do not cause a logical fault when executing Algorithm 4 with G^D-matrices as in Eqs. (N23) to (N25). The errors follow the notation of Definition 3. For example, $\mathbf{e} = (1,0,0)$ corresponds to $Z \otimes \mathbb{1} \otimes \mathbb{1}$ and is undetected yet harmless when it acts on BU-qubits $(A,1)_{BU}$ and $(A,2)_{BU}$. This is a direct consequence of $(1,1,0,0,0,0,0,0)G_1^A = (0,0)$ that can be confirmed by inspection of Eq. (N23). Notice that only unit vector $\mathbf{e}$ appears in this list of fault types.

j	1	2	3	4	5	6	7	8
M_j	$\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}$	$\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$	$\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}$	$\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}$	$\begin{pmatrix}1&0&1\\0&0&1\\0&1&0\end{pmatrix}$	$\begin{pmatrix}1&1&0\\1&0&0\\0&0&1\end{pmatrix}$	$\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}$	$\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}$
$(1,0,0)$	$\begin{pmatrix}1&0&0\end{pmatrix}$	$\begin{pmatrix}1&0&0\end{pmatrix}$	$\begin{pmatrix}0&1&0\end{pmatrix}$	$\begin{pmatrix}0&0&1\end{pmatrix}$	$\begin{pmatrix}1&0&1\end{pmatrix}$	$\begin{pmatrix}1&1&0\end{pmatrix}$	$\begin{pmatrix}0&1&0\end{pmatrix}$	$\begin{pmatrix}0&0&1\end{pmatrix}$

TABLE X. A set of transformation matrices M_j that represent Clifford symmetries as defined in Eq. (N74). For every distinct pair of indices $\{i,j\}$ they satisfy either condition $1^\star$ or condition $2^\star$ as stated in the proof of Claim 3. In particular, the only pairs for which condition $1^\star$ does not hold are $\{1,2\}$,$\{3,7\}$ and $\{4,8\}$. However, for these three cases the fault pattern is one of the harmless cases listed in Table IX.

There are 7 different possible values of $(1,0,0)M_j$ but 8 different j indices, so it is (narrowly) not possible to use condition $1^\star$ alone. However, it is possible to find a set of Clifford symmetries such that some pairs $\{i,j\}$ are covered by condition $1^\star$ and some pairs $\{i,j\}$ are covered by condition $2^\star$.

We provide such an $\{M_j\}$ set in Table X, which suffices to prove Claim 3. Furthermore, this set can be implemented especially easily. Consulting Table X we find that the Clifford symmetries for indices $\{1,2,3,4,7,8\}$ all correspond to permutations of indices and so can all be performed in software. The only exceptions are indices $\{5,6\}$ that correspond to a $W = \mathrm{CNOT}_{B,A}X_A$ gate followed by an index permutation. In other words, W bit-flips qubit A if qubit B is in the $|0\rangle$ state. Consequently, the state $|1,1,1\rangle$ is invariant under W and other computational basis states are permuted. Therefore, W is a Clifford symmetry of the $|\mathrm{CCZ}\rangle$ state since all terms except $|1,1,1\rangle$ carry the same amplitude and phase. Since $|\mathrm{TOF}\rangle = H_C|\mathrm{CCZ}\rangle$ and $[W, H_C] = 0$, we know W also stabilizes $|\mathrm{TOF}\rangle$. Furthermore, when conjugating a Z error we have that Eq. (N74) takes the form

$$WZ[\mathbf{e}_j]W^\dagger = W(-1)^{e_j^A} Z[\mathbf{e}_j M_W], \tag{N80}$$

where

$$M_W = \begin{pmatrix}1&1&0\\0&1&0\\0&0&1\end{pmatrix}. \tag{N81}$$

Permuting qubit indices after W corresponds to swapping columns of M_W. We get M_5 of Claim 3 by swapping columns 2 and 3 of M_W. We get M_6 of Table X by swapping columns 1 and 2 of M_W.

The actual noise distribution output from `BUTOF` is not exactly the toy noise model of Table X but it shares the feature that $(Z \otimes \mathbb{1} \otimes \mathbb{1})$ errors dominate. In all numerics presented, we use the Clifford symmetry operations of Claim 3 and Table X but analyzed using the correct `BUTOF` noise model.

Implementing W on repetition encoded qubits A and B is straightforward because W is transversal and the codeblocks are adjacent to each other in the proposed layout.

7. Factory layout and scheduling

We see from Fig. 59 that lattice surgery require some additional workspace to connect the various codeblocks. Fig. 60 presents a 2D layout to realize `TDTOF` using lattice surgery, including all necessary workspace. If bit-flips are sufficiently small, then the factory can be realized completely with repetition codes. In the regime where bit-flips are rare but not completely negligible, we use a mix of repetition codeblocks (for the BU-qubits) and $d_x = 3$ thin surface codes (for the factory qubits) to tolerate a single physical bit-flip anywhere in the factory. Additional bit-flip protection could be achieved by increasing the X distance of all code blocks and/or performing two rounds of `TDTOF`. Here we only describe a single round and primarily focus on the version using $d_x = 3$ surface code blocks.

For now, we assume a supply of TOF states generated from `BUTOF`. Then we can schedule the main `TDTOF` steps as listed in Table XI. The required 8 input $|\mathrm{TOF}\rangle$ are divided into 2 batches of 4. How quickly can a batch of 4 input $|\mathrm{TOF}\rangle$ magic states be injected? Each $|\mathrm{TOF}\rangle$ state comprises 3 qubits, so there are a total of $12 = 3 \times 4$ multi-patch Pauli measurements needed per batch. These can be partly parallelised. Fig. 60 shows 4 horizontal empty regions that we will call access corridors labeled $\{A, AB, BC, C\}$. This allows us to perform 4 multi-patch

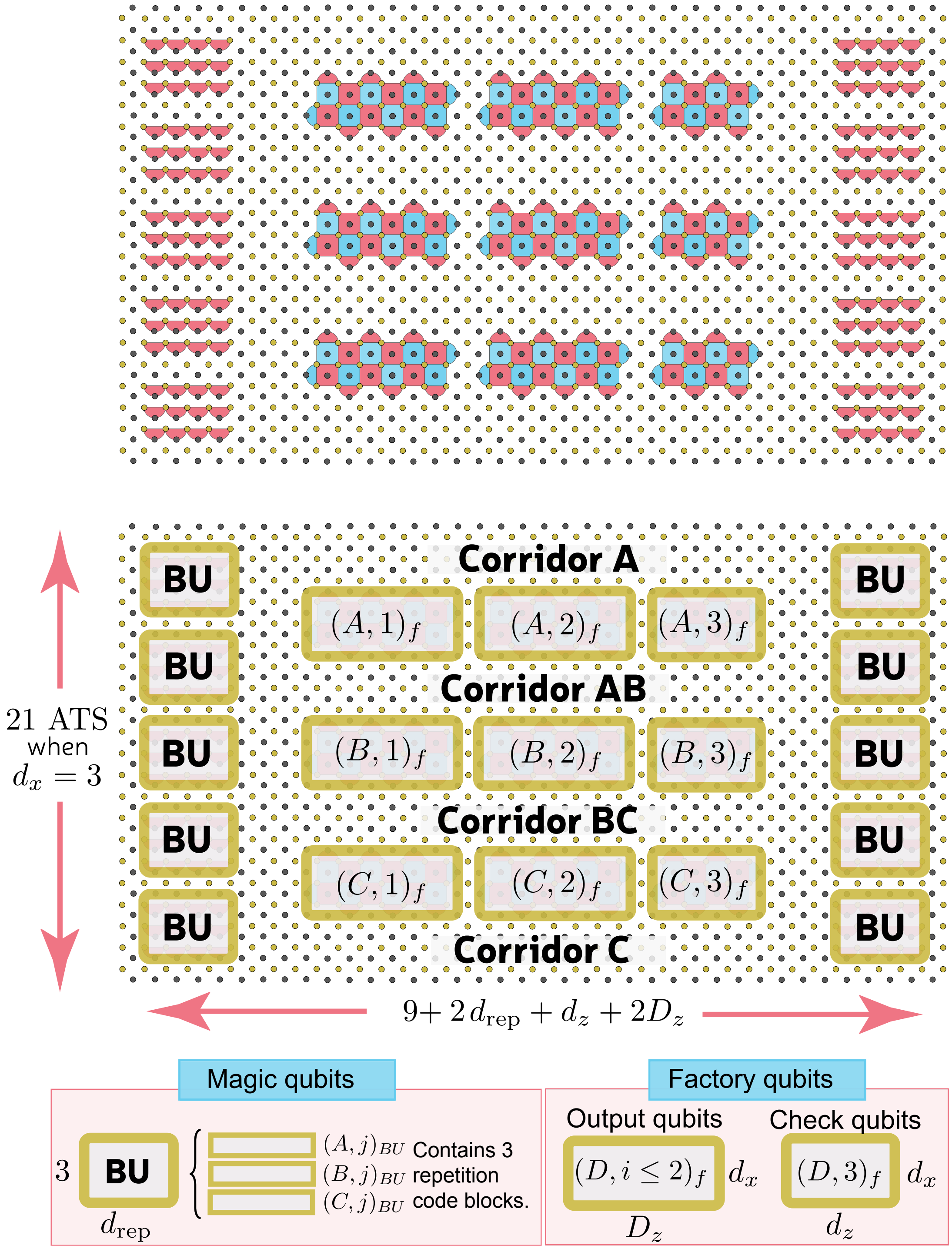


FIG. 60. A 2D layout for realising 8TOF → 2TOF distillation via lattice surgery using a mixture of repetition codes and thin surface code. Example dimensions shown here with: encoding distances $d_x = 3$, $d_{\text{rep}} = 5$, $d_z = 5$ and $D_z = 7$; and $M = 10$ `BUTOF` modules. Additional space between codeblocks is provided for lattice surgery and routing between code blocks (see Appendix L and Fig. 59). We give explicit locations for the 9 factory qubits with labels $(D,i)_f$ following Definition 2. The modules labeled BU consists of 3 repetition codes and provide space to attempt a noisy $|\text{TOF}\rangle$ preparation using the `BUTOF` protocol. Note that `BUTOF` is executed with a distance d_{BU} repetition code (typically we set $d_{BU} = 5, 7$) and then immediately grow to distance $d_z > d_{BU}$.

Factory clock	Corridor A	Corridor AB	Corridor BC	Corridor C
1	$\mathrm{X(C,4)}_{BU}$ $\mathrm{Z(A,3)}_{f}$ then $\mathrm{Z(C,4)}_{BU}$	$\mathrm{X(C,3)}_{BU}$ $\mathrm{Z(B,1)}_{f}$ $\mathrm{Z(B,2)}_{f}$ $\mathrm{Z(B,3)}_{f}$ then $\mathrm{Z(C,3)}_{BU}$	$\mathrm{X(C,2)}_{BU}$ $\mathrm{Z(C,2)}_{f}$ $\mathrm{Z(C,3)}_{f}$ then $\mathrm{Z(C,2)}_{BU}$	$\mathrm{X(C,1)}_{BU}$ $\mathrm{Z(C,1)}_{f}$ $\mathrm{Z(C,2)}_{f}$ $\mathrm{Z(C,3)}_{f}$ then $\mathrm{Z(C,1)}_{BU}$
2	$\mathrm{Z(B,1)}_{BU}$ $\mathrm{Z(A,0)}_{f}$ $\mathrm{Z(A,1)}_{f}$ $\mathrm{Z(A,2)}_{f}$ then $\mathrm{X(B,1)}_{BU}$	$\mathrm{Z(B,2)}_{BU}$ $\mathrm{Z(B,0)}_{f}$ $\mathrm{Z(B,2)}_{f}$ then $\mathrm{X(B,2)}_{BU}$	$\mathrm{Z(B,4)}_{BU}$ $\mathrm{Z(B,0)}_{f}$ $\mathrm{Z(B,2)}_{f}$ then $\mathrm{X(B,4)}_{BU}$	$\mathrm{Z(B,3)}_{BU}$ $\mathrm{Z(C,1)}_{f}$ $\mathrm{Z(C,2)}_{f}$ then $\mathrm{X(B,3)}_{BU}$
3	$\mathrm{Z(A,3)}_{BU}$ $\mathrm{Z(A,0)}_{f}$ $\mathrm{Z(A,2)}_{f}$ then $\mathrm{X(A,3)}_{BU}$	$\mathrm{Z(A,2)}_{BU}$ $\mathrm{Z(A,0)}_{f}$ $\mathrm{Z(A,1)}_{f}$ $\mathrm{Z(A,2)}_{f}$ then $\mathrm{X(A,2)}_{BU}$	$\mathrm{Z(A,1)}_{BU}$ $\mathrm{Z(B,0)}_{f}$ $\mathrm{Z(B,1)}_{f}$ $\mathrm{Z(B,2)}_{f}$ then $\mathrm{X(A,1)}_{BU}$	$\mathrm{Z(A,4)}_{BU}$ $\mathrm{Z(C,1)}_{f}$ $\mathrm{Z(C,2)}_{f}$ then $\mathrm{X(A,4)}_{BU}$
4	$\mathrm{X(C,8)}_{BU}$ $\mathrm{Z(A,3)}_{f}$ then $\mathrm{Z(C,8)}_{BU}$	$\mathrm{X(C,6)}_{BU}$ $\mathrm{Z(B,2)}_{f}$ $\mathrm{Z(B,3)}_{f}$ then $\mathrm{Z(C,6)}_{BU}$	$\mathrm{X(C,7)}_{BU}$ $\mathrm{Z(B,1)}_{f}$ $\mathrm{Z(B,3)}_{f}$ then $\mathrm{Z(C,7)}_{BU}$	$\mathrm{X(C,5)}_{BU}$ $\mathrm{Z(C,1)}_{f}$ $\mathrm{Z(C,3)}_{f}$ then $\mathrm{Z(C,5)}_{BU}$
5	$\mathrm{Z(B,5)}_{BU}$ $\mathrm{Z(A,0)}_{f}$ $\mathrm{Z(A,1)}_{f}$ $\mathrm{Z(A,2)}_{f}$ then $\mathrm{X(B,5)}_{BU}$	$\mathrm{Z(B,7)}_{BU}$ $\mathrm{Z(A,1)}_{f}$ $\mathrm{Z(A,2)}_{f}$ then $\mathrm{X(B,7)}_{BU}$	$\mathrm{Z(B,8)}_{BU}$ $\mathrm{Z(B,2)}_{f}$ then $\mathrm{X(B,8)}_{BU}$	$\mathrm{Z(B,6)}_{BU}$ $\mathrm{Z(C,2)}_{f}$ then $\mathrm{X(B,6)}_{BU}$
6	$\mathrm{Z(A,6)}_{BU}$ $\mathrm{Z(A,0)}_{f}$ $\mathrm{Z(A,2)}_{f}$ then $\mathrm{X(A,6)}_{BU}$	$\mathrm{Z(A,5)}_{BU}$ $\mathrm{Z(B,1)}_{f}$ $\mathrm{Z(B,2)}_{f}$ then $\mathrm{X(A,5)}_{BU}$	$\mathrm{Z(A,7)}_{BU}$ $\mathrm{Z(C,2)}_{f}$ then $\mathrm{X(A,7)}_{BU}$	$\mathrm{Z(A,8)}_{BU}$ $\mathrm{Z(C,2)}_{f}$ then $\mathrm{X(A,8)}_{BU}$
7	Setup exit	$\mathrm{X(A,3)}_{f}$ Clifford corrected	$\mathrm{X(B,3)}_{f}$ Clifford corrected	$\mathrm{X(C,3)}_{f}$ Clifford corrected
8	exit	exit	exit	exit

TABLE XI. The final form of our `TDTOF` protocol for one full cycle of the factory. It executes a variant of Algorithm 4 that has been modified according to the qubit permutations required for noise-tailoring (see Appendix N 6) and embedded within the 2D layout of Fig. 60. Each cell for factory clocks 1-6 has the form A then B. Instruction A specifies a multi-qubit Pauli operator using the qubit notation of Definition 2. For example, $X(C,4)_{BU}, Z(A,3)_f$ means measure the operators $X \otimes Z$ where the X acts on magic input labeled $(C,4)_{BU}$ and the Z acts on factory qubit $(A,3)_f$. Instruction B specifies a single-qubit measurement of a magic input qubit. The B instructions can be realized with physical single-qubit measurements that takes a single surface code cycle. As such, B instructions require negligible time compared to the A instructions, so we present both a and B within a single Factory clock step that has duration $d_m + 1$. In factory clock steps 1 and 4, the role of Z and X are swapped on the magic state qubits to account for the Hadamard difference between $|\mathrm{CCZ}\rangle$ and $|\mathrm{TOF}\rangle$. The column headers "Corridor" indicate which Corridor from Fig. 60 is used to realize the multi-qubit Pauli measurement since lattice surgery requires some workspace to operate. Note that Corridor AB can only be used to access factory qubit with labels of the form $(A,i)_f$ or $(B,i)_f$. The column headers also list which factory blocks $\{A,B,C\}$ the Corridor can be used to access and this constraint it respected in this schedule. Notice that multi-qubit measurements of the form $X(C,4)_{BU}, Z(A,3)_f$ involve different capital letter indices on the factory and magic qubits. In contrast, Item 1a of Algorithm 3 describes multi-qubit Pauli measurements with matching capital letter indices. This is due to the permutation operations required for noise tailoring (see Appendix N 6). In particular, when performing measurements with the $j = 4$ index, the matrix M_4 of Table X instructs us to swap the A and C indices for the magic state qubit. In the cases of M_6 and M_7, these are decomposed into a single Clifford gate W and a permutation. The above table only accounts for the permutation, with the Clifford performed on the input magic state qubits prior to injection into `TDTOF`. Factory clock times 1-3 correspond to batch 1, so that measurements involve only magic state qubits of the form $(D,i)_{BU}$ with $i \in [1,4]$. Factory clock times 4-6 correspond to batch 2, so that measurements involve only magic state qubits of the form $(D,i)_{BU}$ with $i \in [5,8]$. The importance of batching and the related issue of `BUTOF` scheduling is discussed in Appendix N 7. "Exit" refers to factory qubits moving out of the factory.

Pauli measurement in parallel. There are some constraints on which multi-patch Pauli measurements are performed (further discussion in the caption of Fig. 60). The first batch is injected in factory clock steps 1-3. The second batch is injected in factory clock steps 4-6. Factory clock step 7 performs the measurement of the check qubits, and starts the process of exiting some factory qubits out of the factory. Factory clock step 8 completes the process of exiting the factory qubits. Each factory clock step takes a time $(d_m+1)T_{\rm surf}$ where $T_{\rm surf}$ is the duration of 1 surface code cycle and d_m is the number of surface code cycle used per multi-qubit Pauli measurement. The "$+1$" in (d_m+1) provides time to perform high fidelity single qubit measurements and reset between rounds of multi-qubit Pauli measurement. Roughly, a single execution of TDTOF takes time $8(d_m+1)T_{\rm surf}$, though small extra additive timecosts may be incurred to execute BUTOF, which we discuss next.

The BUTOF protocol can have a fairly high failure probability, labeled here by F_{BU}. This failure probability depends on the repetition code distance d_{BU} used in BUTOF. To boost the probability of having ample supply of states from BUTOF, we add redundancy in both time and space. Our illustrations show $M = 20$ modules for BUTOF, but we only need 8 input $|\mathrm{TOF}\rangle$ or $|\mathrm{CCZ}\rangle$ states for the protocol. Not all 8 input $|\mathrm{TOF}\rangle$ or $|\mathrm{CCZ}\rangle$ states need to exist at the same time as they are split into two batches. Rather we aim to prepare 4 $|\mathrm{TOF}\rangle$ at the start of factory clock steps 1 and 4. Therefore, during the factory clock steps 4-8 (a total time of $5(d_m+1)T_{\rm surf}$), we need to prepare 4 $|\mathrm{TOF}\rangle$ for the first batch of the next round of TDTOF. During the factory clock steps 1-3 (a total time of $3(d_m+1)T_{\rm surf}$), we need to prepare 4 $|\mathrm{TOF}\rangle$ for the second batch in the current round of TDTOF. Let us focus our discussion on preparation during steps 1-3 as this is the bottleneck point. Furthermore, our schedule requires that, of these 4 $|\mathrm{TOF}\rangle$ states, 2 are located on the left and 2 are located on the right. Considering just one side, we have $M/2$ BUTOF modules. Of these $M/2$ modules, 2 are busy storing $|\mathrm{TOF}\rangle$ states and performing the required lattice surgery operations. This leaves $(M-1)/2$ modules responsible for preparing 2 $|\mathrm{TOF}\rangle$ states. Each attempt at BUTOF takes a time

$$T_{BU} = 2d_{BU}T_{\rm rep} + \frac{d_{BU}+1}{2}(2+d_{BU}+1)T_{\rm cnot}, \quad \text{(N82)}$$

where $T_{\rm CNOT}$ is the optimal time for a CNOT gate and $T_{\rm rep}$ is the time for a repetition code cycle. Therefore, steps 1-3 provide enough time to fit in $R := \lfloor 3(d_m+1)T_{\rm surf}/T_{BU}\rfloor$ repeated attempts at BUTOF. Given R temporally multiplexed attempts, each BUTOF module has its failure probability reduced from F_{BU} to $\tilde{F}_{BU} := F_{BU}^R$. Each side fails if there are zero or one module successes of the $(M-1)/2$ modules, which occurs with probability

$$\begin{aligned} F_{\rm side} &= \tilde{F}_{BU}^{(M-1)/2} + \frac{M-1}{2}\tilde{F}_{BU}^{(M-1)/2}(1-\tilde{F}_{BU}) \\ &= F_{BU}^{R(M-1)/2} + \frac{M-1}{2}F_{BU}^{R(M-1)/2}(1-F_{BU}^R). \end{aligned} \quad \text{(N83)}$$

For instance, executing BUTOF at distance 5 and using $d_m = 15$ surface code cycles per lattice surgery operation we have $R = 3$ attempts at BUTOF (assuming $\kappa_1/\kappa_2 = 10^{-5}$ and $|\alpha|^2 = 8$). If $F_{BU} = 0.447$ then the temporal redundancy reduces this to $\tilde{F}_{BU} = F_{BU}^3 = 0.089$. Providing $M = 10$ modules in total, there is $(M-1)/2 = 3$ available spatial redundancy on each side, which further suppresses the failure probability to $F_{\rm side} = 0.023$. This is already quite low. We can further reduce the failure probability by either: increasing space cost M; or inserting a small number Q additional rounds of BUTOF between steps 3 and 4. In the latter case, the runtime of TDTOF is extended to

$$T_{TD} = QT_{BU} + 8(d_m+1)T_{\rm surf}, \quad \text{(N84)}$$

where we have assumed $QT_{BU} \leq 2(d_m+1)$. This further reduces $F_{\rm side}$. A coarse bound is obtained by replacing $R \to R+Q$ in Eq. (N83), though actually the suppression is slightly better as there are now $M/2$ modules available for the Q attempts. We do not wish to set Q too high, as the additional delay leads to logical error accumulation due to finite distance choices.

Whenever BUTOF fails to proceed the required $|\mathrm{TOF}\rangle$ states, we count this as a failure of the whole TDTOF protocol. However, we use sufficient redundancy that such occurrences are very rare. Typically, we set $Q = 1$ or $Q = 2$, and we use $M = 10$ when $d_{BU} = 5$ and $M = 20$ when $d_{BU} = 7$.

An additional consideration is that a lattice dislocation is used when performing a multi-qubit Pauli measurement including a Z_L on a repetition encoded logical qubit (see Fig. 59). This dislocation uses a small amount of additional space. However, when using $|\mathrm{TOF}\rangle$ input states (instead of $|\mathrm{CCZ}\rangle$) the third qubit differs by a Hadamard and so the protocol is adjusted to measure X_L and a dislocation is not required. For this reason, we inject the qubits in reverse order: $(C,j)_{BU}$, $(B,j)_{BU}$ then $(A,j)_{BU}$. After $(C,j)_{BU}$ is injected (without needing a dislocation) some space is freed-up for dislocations to be used, enabling $(B,j)_{BU}$ and $(A,j)_{BU}$ to be injected.

8. Clifford noise

Perhaps one of the most importance aspects of magic state factory design is the choice of distance for various code blocks. It is possible to use much smaller code distances within the factory than used inside the main algorithm. Using finite code distances leads to noisy

Fault source and remarks	**Propagated**	**Risk**	**Suppressing parameter**
Z-logical errors on repetition codes during storage	Backwards	not critical	d_{rep}
Z-logical errors on factory qubits during storage	Forwards	critical	D_z
Z-logical errors on check qubits during storage	Forwards	not critical	d_z
X logical on repetition codes during storage	Backwards	not critical	$\lvert\alpha\rvert^2$
X logical on surface codes factory qubits	Stuck	critical	d_x, $\lvert\alpha\rvert^2$
Timelike error during lattice surgery multi-patch measurement. *Remarks:* This flips multi-qubit measurement outcome (denoted ω_j^D in Algorithm 3) but is equivalent to Pauli error on input magic state. See Appendix L for details.	Backwards	not critical	d_m
Measurement failure when reseting after lattice surgery. *Remarks:* This flips some single Pauli measurement outcome m_j^D in Algorithm 3. Equivalent to Pauli error on input magic state.	Backwards	not critical	$\lvert\alpha\rvert^2$

TABLE XII. Fault sources due to imperfect Cliffords. Each error is either propagated forwards or backwards, or it is stuck. We sum the probability of all stuck errors and add to the overall infidelity of TDTOF. Backwards propagated errors modify the noise distribution on the input magic states. Forwards propagated are handled by modifying the formulae (see Eqs. (N85) and (N86)) for the infidelity and acceptance probability. An error is a critical risk it occurs with probability p and contributes to the overall infidelity with probability $O(p)$ rather than $O(p^2)$. Every error source can be exponentially suppressed some parameter, where $\{d_{\text{rep}}, d_z, D_Z, d_x\}$ are code distance illustrated in Fig. 60; d_m is the measurement distance denoting the number of surface code cycles used during lattice surgery (see Appendix L); and $|\alpha|^2$ is the mean photon number in the cat code qubit. For critical risk errors, the associated parameter is typically set higher than the parameters set for non-critical errors. In particular, the parameters $\{d_z, d_m, d_{\text{rep}}\}$ can be safely set at about half the value of D_z though our actual choice is determined by numerical search.

Clifford gates, noisy lattice surgery operations and non-negligible memory noise. This needs to be accounted for in addition to the error estimated by Claim 2 under the assumption of ideal Cliffords. Indeed, typically Clifford noise is the dominate source of errors and the error of Claim 2 should instead be regarded as the minimum achievable error (with 1 round of TDTOF) in the limit of infinite code distances.

Some of the relevant spatial code distance parameters are shown in Fig. 60. An important additional quantity is the "measurement distance" d_m that is increased to suppress the effect of timelike errors during lattice surgery (see Appendix L for further details). A common choice in the literature is to set $d_m = \max[d_z, d_x]$, but this is by no means necessary or optimal.

Rather than a Monte Carlo simulation of Clifford noise, we perform a computer-assisted analytical analysis. It is helpful to distinguish critical and non-critical faults. We say a Clifford fault is a critical risk if (assuming no other errors occur) it leads to an undetected fault on the output magic states. Conversely, a fault is a non-critical risk if it will be detected (assuming no other errors). All sources of Clifford noise can be grouped into one of four classes

1. Backwards propagating and not critical: these are errors that can be commuted towards the start of the circuit, so that they act on a single noisy input $|\text{TOF}\rangle$ state. If ρ is the density matrix with only noise from BUTOF, the backwards propagating noise is applied so $\rho \to \rho'$. Then the effective Z logical error distribution is determined from ρ' using the procedure of Appendix N 5 b.

2. Forwards propagating and not critical: these errors can be commuted to the end of the circuit, so that they act on the check qubits in the factory just before they are measured.

3. Forwards propagating and critical: these errors can be commuted to the end of the circuit, so that they act on the output magic state qubits.

4. Stuck errors and potentially critical: these are errors that are difficult to commute forwards or backwards through the circuit. We sum the probability of these events and add it to the error rate on the output magic states.

Our treatment of stuck errors means that we obtain an upper bound on the performance. One might be concerned that this bound is loose, but in practice the stuck errors are very rare and not, therefore, of major importance. Indeed, if we instead attempted a Monte Carlo simulation, the statistical variance in the error estimate would exceed that of the total stuck error probability.

Therefore, our computer-assisted analytical analysis leads to more accurate results than Monte Carlo methods. We further remark that while a mild amount of truncation of higher order processes is employed, we use the procedure of Appendix N 5 a to monitor this truncation error and verify that it is negligible.

We list all the source of imperfections in Table XII and describe the propagation type and risk level. Let us assume that backwards propagation has been performed and we have accounted for the effect of noise tailoring (recall Appendix N 6) on the error distribution. Following earlier notation of Definition 3 and Appendix N 6, we say j^{th} noisy TOF state suffers fault $Z[\mathbf{e}_j]$ with probability $\mathbb{P}_j(\mathbf{e}_j)$ that we precompute. Then, without any other noise sources, Claim 2 would describe the acceptance probability and output infidelity. However, the factory qubits may be affected by some forwarded propagated error $Z[(\tilde{\mathbf{w}}^A, \tilde{\mathbf{w}}^B, \tilde{\mathbf{w}}^C)]$ where the labels $\{A, B, C\}$ refer to the 3 different blocks of factory qubits. We used similar notation, without the tilde, in Claim 1 to describe how errors due to input magic states impact the protocol. To combine with the forwarded propagated errors we simply replace $\mathbf{w} \rightarrow \mathbf{w} + \tilde{\mathbf{w}}$ to add the effect of the forwarded propagated errors and follow this modification through the analysis of Claim 2. As we did earlier, it will be useful to split $\mathbf{w} = (\mathbf{v}, \mathbf{u})$ to distinguish errors on check qubits and output qubits. If the forwarded propagated error $\tilde{\mathbf{w}}$ on each block occurs with some probability $\mathbb{F}(\tilde{\mathbf{w}})$ then the results of Claim 2 modify to

$$P_{\text{acc}} = \sum_{\substack{\tilde{\mathbf{w}}^D, \mathbf{e}^D \\ [\mathbf{e}^D G_0^D = \tilde{\mathbf{v}}^D] \forall D}} \prod_{\substack{1 \leq j \leq 8 \\ D \in \{A,B,C\}}} \mathbb{P}_j(\mathbf{e}_j) \mathbb{F}(\tilde{\mathbf{w}}^D), \quad \text{(N85)}$$

and

$$F = \frac{1}{P_{\text{acc}}} \sum_{\substack{\tilde{\mathbf{w}}^D, \mathbf{e}^D \\ [\tilde{\mathbf{e}}^D G_1^D = \tilde{\mathbf{u}}^D] \forall D}} \prod_{\substack{1 \leq j \leq 8 \\ D \in \{A,B,C\}}} \mathbb{P}_j(\mathbf{e}_j) \mathbb{F}(\tilde{\mathbf{w}}^D). \quad \text{(N86)}$$

There are three important changes here. First, in both equations we have summed over forwards propagated errors and weighted by the appropriate probability. In the acceptance probability the summation constraint $[\mathbf{e}^D G_0^D = \mathbf{0}] \forall D$ has been replaced by $[\mathbf{e}^D G_0^D = \tilde{\mathbf{v}}^D] \forall D$ since to pass the check measurement any forwards propagated error $\tilde{\mathbf{v}}^D$ must cancel (therefore equal) some other error to go undetected. Similarly, in the fidelity expression we have replaced $[\mathbf{e}^D G_1^D = \mathbf{0}] \forall D$ with $[\mathbf{e}^D G_1^D = \tilde{\mathbf{u}}^D] \forall D$ because to contribute to the fidelity any forwarded propagated error $\tilde{\mathbf{u}}^D$ must cancel (therefore equal) some other error.

Calculating the expressions for $\mathbb{P}_j$, $\mathbb{F}$, performing the summation and adding the stuck error events is too involved to perform by hand. But it is relatively straightforward for a symbolic mathematics package such as Mathematica. Optimizing over various error suppressing parameters, we find the factory designs that achieve a certain target error per Toffoli at the minimum qubit and ATS cost (without making significant sacrifices to acceptance probabilities) and present results in Tables XIII and XIV.

ϵ_{TD}	# ATS	P_{ACC} (%)	Time/Tof (μs)	M_{BU}	d_{BU}	d_{rep}	d_z	D_z	d_x	d_m
$4.2 * 10^{-9}$	2037	100	551	10	5	9	15	27	3	21
$4.3 * 10^{-9}$	1911	100	507	10	5	9	13	25	3	19
$4.5 * 10^{-9}$	1827	100	484	10	5	9	13	23	3	18
$6.3 * 10^{-9}$	1785	100	462	10	5	9	11	23	3	17
$7.6 * 10^{-9}$	1743	100	484	10	5	7	13	23	3	18
$9.6 * 10^{-9}$	1701	100	462	10	5	7	11	23	3	17
$1.2 * 10^{-9}$	1617	100	441	10	5	7	11	21	3	16
$2.3 * 10^{-8}$	1533	100	418	10	5	7	11	19	3	15
$3.8 * 10^{-8}$	1491	100	396	10	5	7	9	19	3	14
$1.1 * 10^{-7}$	1407	99	377	10	5	7	9	17	3	13
$2.6 * 10^{-7}$	1365	99	355	10	5	7	7	17	3	12
$3.7 * 10^{-7}$	1323	99	378	10	5	5	9	17	3	13
$6.2 * 10^{-7}$	1281	99	333	10	5	7	7	15	3	11
$7.4 * 10^{-7}$	1281	98	356	10	5	5	7	17	3	12
$8.3 * 10^{-7}$	1239	98	355	10	5	5	9	15	3	12
$1.1 * 10^{-6}$	1197	98	333	10	5	5	7	15	3	11
$4.2 * 10^{-6}$	1113	98	311	10	5	5	7	13	3	10
$7.5 * 10^{-6}$	1071	93	305	10	5	5	5	13	3	9

TABLE XIII. Assuming $\kappa_1/\kappa_2 = 10^{-5}$, $\kappa_\phi = 0$ and $|\alpha|^2 = 8$. Performance of optimized `TDTOF` factory using `BUTOF` with $d_{BU} = 5$.

ϵ_{TD}	# ATS	P_{ACC} (%)	Time/Tof (μs)	M_{BU}	d_{BU}	d_{rep}	d_z	D_z	d_x	d_m
$4.3 * 10^{-8}$	2289	99	447	5	11	15	31	3	23	10
$4.3 * 10^{-8}$	2205	99	430	5	11	15	29	3	22	10
$4.3 * 10^{-8}$	2121	98	418	5	11	15	27	3	21	10
$4.7 * 10^{-8}$	2121	99	430	5	9	15	29	3	22	10
$4.7 * 10^{-8}$	2037	98	418	5	9	15	27	3	21	10
$6.0 * 10^{-8}$	1995	98	401	5	9	13	27	3	20	10
$6.6 * 10^{-8}$	1911	98	384	5	9	13	25	3	19	10
$8.9 * 10^{-8}$	1827	98	367	5	9	13	23	3	18	10
$1.3 * 10^{-7}$	1785	95	362	5	9	11	23	3	17	10
$1.8 * 10^{-7}$	1743	98	368	5	7	13	23	3	18	10
$2.2 * 10^{-7}$	1701	95	362	5	7	11	23	3	17	10
$4.5 * 10^{-7}$	1617	95	344	5	7	11	21	3	16	10
$6.7 * 10^{-7}$	1575	95	327	5	7	9	21	3	15	10
$8.4 * 10^{-7}$	1533	95	327	5	7	11	19	3	15	10
$1.4 * 10^{-6}$	1491	95	310	5	7	9	19	3	14	10
$2.8 * 10^{-6}$	1449	86	322	5	7	7	19	3	13	10
$3.2 * 10^{-6}$	1407	87	321	5	7	9	17	3	13	10
$8.3 * 10^{-6}$	1365	86	304	5	7	7	17	3	12	10
$9.2 * 10^{-6}$	1323	86	323	5	5	9	17	3	13	10

TABLE XIV. Assuming $\kappa_1/\kappa_2 = 2 * 10^{-5}$, $\kappa_\phi = 0$ and $|\alpha|^2 = 8$. Performance of optimized `TDTOF` factory using `BUTOF` with $d_{BU} = 5$. Zero-dephasing noise.