Constructions for measuring error syndromes in Calderbank-Shor-Steane codes between Shor and Steane methods

Shilin Huang [1,*] and Kenneth R. Brown [1,2,3,†]

[1]*Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina 27708, USA*
[2]*Department of Physics, Duke University, Durham, North Carolina 27708, USA*
[3]*Department of Chemistry, Duke University, Durham, North Carolina 27708, USA*

(Received 13 February 2021; accepted 29 July 2021; published 25 August 2021)

In another work [S. Huang and K. R. Brown, Phys. Rev. Lett. **127**, 090505 (2021)] we introduced syndrome extraction methods for Calderbank-Shor-Steane quantum error-correction codes that interpolate between the well-known Shor and Steane syndrome extraction methods. Here we provide detailed proofs of the main theorems and show that up to gate ordering there is a one-to-one correspondence between extraction gadgets and partitions of the parity check matrix. Operationally, all the circuits in our framework can be obtained by merging ancilla qubits of Shor error correction with certain rules, which enables us to design fault-tolerant syndrome extraction circuits for given specific codes. We then apply our construction to the toric code and provide a detailed analysis of the time overhead of fault tolerance. In particular, we construct a syndrome extraction family whose time overhead smoothly varies from Shor to Steane error correction. To understand the potential advantage, we consider two simplified error models: no errors on the ancilla block and uncorrelated errors on the ancilla block. We study the threshold behavior of the family for both error models and show its potential advantage for quantum architectures with long coherence times.

DOI: 10.1103/PhysRevA.104.022429

I. INTRODUCTION

The theory of quantum error correction [1–10] opens the path towards large-scale quantum computation. Stabilizer codes are a conventional choice of quantum error-correcting codes [7,8,11], where error correction is performed conditional on the measurement outcomes of a set of code stabilizers, also known as the error syndrome bit string. The syndrome extraction circuits need to be fault tolerant [9,12–18] as the act of measuring syndromes also introduces extra errors in the quantum system.

The first fault-tolerant syndrome extraction scheme was proposed by Shor [9,10]. In Shor's scheme, each syndrome bit is extracted from the data qubits to a verified ancilla cat state by transversal two-qubit gates. As transversal operations limit the error propagation, no high-weight correlated errors can occur on the data qubits if the cat states are verified by postselection. The value of the syndrome bit is the parity of the transversal measurement outcome of the cat state. As any measurement error will flip the syndrome bit, for a stabilizer code of distance d, one needs to repeat the syndrome extraction for $O(d^2)$ rounds to guarantee fault tolerance. Utilizing the information of the code structure, the time overhead can be significantly reduced on particular codes [19–24].

Optimizing Shor's scheme is an active area of research building off substantial progress since its invention. For example, for low-weight stabilizers, ancilla postselection could be avoided by decoding the ancilla cat states [16,25,26]. On specific stabilizer codes, the space overhead can be reduced by allowing nontransversal data-ancilla interactions that preserve the code distance [19,26–32]. The time overhead can also be reduced by careful choice of sequential extractions [23,24]. Notably, as an alternative to cat-state measurements, flag error-correction gadgets [16–18,30] circumvent the need of ancilla postselection for arbitrary stabilizer codes, while having a low-qubit overhead for low-distance codes or high-distance codes with low-weight stabilizer measurements.

The extraction gadgets for Shor's scheme are arguably the smallest. As a tradeoff, a large number of two-qubit gates are applied between data and ancilla qubits. For many quantum devices, two-qubit gates are usually the most challenging operations to implement with high fidelity [33,34]. To minimize the data-ancilla interaction, Steane [13] and Knill [15] suggested the use of transversal two-qubit gates to transfer the errors from the data block to an ancilla code block and then measurement of the ancilla block to gain error information. Steane's scheme is specialized for Calderbank-Shor-Steane (CSS) codes [5,6] and needs two separate rounds of transversal two-qubit gates for extracting the X- and Z-type stabilizers, respectively. Knill's scheme works for arbitrary stabilizer codes and only needs one round of the transversal gate to extract all the stabilizers. Using a constant number of Steane or Knill syndrome extractions, an arbitrary logical Clifford circuit can be implemented fault tolerantly in $O(1)$ steps [35]. Both Steane's and Knill's schemes are single shot, i.e., no repetition of measurements is required. Indeed, each data qubit is touched by $O(1)$ two-qubit gates. However, comparing to a cat state, the ancilla blocks for both Steane's and

*shilin.huang@duke.edu
†ken.brown@duke.edu

Knill's schemes are much harder to prepare, as they are as large as the data block [14,36–39].

A natural question to raise is whether it is possible to balance the complexities of ancilla preparation and data-ancilla interaction. In Ref. [40] we gave a positive answer for CSS codes and in this companion paper we elaborate on the theory behind the method and present additional numerical details. We have generated a family of extraction circuits including Shor's and Steane's constructions as its two extremes. This family increases the complexity of ancilla construction in exchange for reducing the number of two-qubit gates between data and ancilla qubits required to fault-tolerantly measure the error. Applying our construction on the toric code, we are able to use a single ancilla block to measure the plaquette operators (Z-stabilizer elements) inside any connected sublattice. In particular, one can partition the $L \times L$ toric lattice into patches, each of which contains $m \times m$ plaquettes. Shor's and Steane's schemes correspond to the special cases $m = 1$ and $m = L$, respectively. Moreover, by offsetting the partition periodically, one can achieve fault tolerance within $O(L/m)$ measurement rounds. Indeed, the parameter m simultaneously controls the ancilla-block size and the time overhead of fault tolerance. Our result is compatible with the fact that Shor's and Steane's schemes require $O(L)$ and $O(1)$ measurement rounds on the toric code, respectively.

Our paper is organized as follows. In Sec. II we review the essential background and present our notation. In Sec. III we discuss the construction of our family of extraction circuits. In particular, a subfamily called transversal gadgets is proposed. In Sec. IV we discuss the application of transversal gadgets on the toric code and analyze the time overhead of fault-tolerant syndrome measurements. We conclude with a summary and discussion in Sec. V.

II. BACKGROUND

A. Linear codes

We start by developing our notation for connecting sets and binary vector spaces. Given a set Ω, we can define a vector space V_Ω where each vector corresponds to a finite subset $A \subseteq \Omega$, and vector addition is defined as the symmetric difference of two finite subsets of Ω, i.e., for any two finite subsets $A, B \subseteq \Omega$ we define $A + B := (A \cup B) \setminus (A \cap B)$. The symmetric difference yields $A + A = \emptyset$ and as a result V_Ω is a binary vector space. For convenience, each element $w \in \Omega$ is associated with two other meanings: a subset $\{w\} \subseteq \Omega$ and indeed a vector $\{w\} \in V_\Omega$. We can then simplify the equation $A = \sum_{w \in A} \{w\}$ as $A = \sum_{w \in A} w$ for every $A \in V_\Omega$. As a result, we have

$$V_\Omega = \left\{ \sum_{v \in A} v \middle| A \subseteq \Omega, |A| < \infty \right\} = \mathbb{F}_2[\Omega], \tag{1}$$

i.e., V_Ω is the binary vector space spanned by the set Ω, denoted by $\mathbb{F}_2[\Omega]$. The cardinality of a finite subset $A \subseteq \Omega$, $|A|$, is also referred to as the weight of the vector $A \in \mathbb{F}_2[\Omega]$.

The (standard) inner product of $\mathbb{F}_2[\Omega]$ is an $\mathbb{F}_2$-bilinear form

$$\langle \cdot, \cdot \rangle_\Omega : \mathbb{F}_2[\Omega] \times \mathbb{F}_2[\Omega] \to \mathbb{F}_2$$

such that for any $a, b \in \Omega$, $\langle a, b \rangle_\Omega = 1$ if and only if $a = b$. Two vectors $\phi, \psi \in \mathbb{F}_2[\Omega]$ are said to be orthogonal if and only if $\langle \phi, \psi \rangle_\Omega = 0$. More generally, two vector subspaces $V, W \subseteq \mathbb{F}_2[\Omega]$ are said to be orthogonal (denoted by $V \perp W$) if we have $\langle v, w \rangle_\Omega = 0$ for every $v \in V$ and $w \in W$.

Let Ω and Θ be two sets and $M : \mathbb{F}_2[\Omega] \to \mathbb{F}_2[\Theta]$ be a linear map. The kernel and image of M are denoted by $\ker M$ and $\operatorname{im} M$, respectively. When Ω and Θ are finite, the transpose of M is defined to be a linear map $M^\mathsf{T} : \mathbb{F}_2[\Theta] \to \mathbb{F}_2[\Omega]$ such that

$$\langle a, M^\mathsf{T} b \rangle_\Omega = \langle Ma, b \rangle_\Theta$$

for every $a \in \Omega$ and $b \in \Theta$. In other words, under the standard bases Ω and Θ, the matrix M^T is the transpose of the matrix M. We also define the transpose of a vector $\phi \in \mathbb{F}_2[\Omega]$, denoted by ϕ^T, as the map

$$\psi \mapsto \phi^\mathsf{T} \psi := \langle \phi, \psi \rangle_\Omega. \tag{2}$$

We now briefly discuss linear codes. Let Ω be a set of n classical bits. A linear code on Ω is a subspace $\mathcal{C} \subseteq \mathbb{F}_2[\Omega]$. The vectors of $\mathcal{C}$ are called codewords. A parity-check matrix H of $\mathcal{C}$ is a linear map from $\mathbb{F}_2[\Omega]$ to some other $\mathbb{F}_2$-vector space, $\mathbb{F}_2^m$, such that $\mathcal{C} \subseteq \ker H$. In this work, we are not required to have $\mathcal{C} = \ker H$. The dual code of $\mathcal{C}$ is defined by

$$\mathcal{C}^\perp := \{ v \in \mathbb{F}_2[\Omega] : c^\mathsf{T} v = 0 \, \forall \, c \in \mathcal{C} \}. \tag{3}$$

Note that $\mathcal{C}^\perp \supseteq \operatorname{im} H^\mathsf{T}$. This is because for each $x \in \mathbb{F}_2^m$ we have

$$c^\mathsf{T} H^\mathsf{T} x = (Hc)^\mathsf{T} x = 0 \tag{4}$$

for every $c \in \mathcal{C}$.

B. CSS stabilizer codes

Let Ω be a set of n qubits. The state space of the quantum system Ω is denoted by $\mathcal{H}_\Omega = \mathbb{C}[\mathbb{F}_2[\Omega]] \cong \mathbb{C}^{2^n}$. The Pauli group on Ω is denoted by $\mathcal{P}_\Omega$. For each qubit $q \in \Omega$, the Pauli X and Z operators on $\mathcal{H}_q$ are denoted by X_q and Z_q, respectively. For each subset $\psi \subseteq \Omega$, the X-type and Z-type operators supporting on ψ are defined by

$$X[\psi] := \bigotimes_{q \in \Omega} X_q^{\psi^\mathsf{T} q} \in \mathcal{P}_\Omega \tag{5}$$

and

$$Z[\psi] := \bigotimes_{q \in \Omega} Z_q^{\psi^\mathsf{T} q} \in \mathcal{P}_\Omega, \tag{6}$$

respectively. Note that $X_q^0 = Z_q^0 = \mathbb{1}_q$.

A quantum code on Ω is a subspace $\mathcal{C}_Q \subseteq \mathcal{H}_\Omega$. A logical operator of $\mathcal{C}_Q$ is an operator L on $\mathcal{H}_\Omega$ such that $L(\mathcal{C}_Q) \subseteq \mathcal{C}_Q$. A stabilizer code on Ω is defined by an Abelian subgroup $\mathcal{S} \subseteq \mathcal{P}_\Omega$ such that every operator $P \in \mathcal{S}$ has eigenvalues ± 1 and $-\mathbb{1}_{\mathcal{H}_\Omega} \notin \mathcal{S}$. The corresponding code space $\mathcal{C}_Q$ is the common $+1$ eigenspace of all the operators in $\mathcal{S}$. In particular, if $\dim \mathcal{C}_Q = 1$, the unique state in $\mathcal{C}_Q$ (up to a constant) is said to be a stabilizer state. If $\mathcal{S} = \mathcal{S}_X \oplus \mathcal{S}_Z$, where elements in $\mathcal{S}_X$ ($\mathcal{S}_Z$) are all X type (Z type), we say that $\mathcal{S} = \mathcal{S}_X \oplus \mathcal{S}_Z$ is a Calderbank-Shor-Steane code [5,6] and $\mathcal{S}_X$ and $\mathcal{S}_Z$ are the X and Z stabilizers, respectively. We can represent $\mathcal{S}_X$ and $\mathcal{S}_Z$ by

two binary vector subspaces

$$\mathcal{C}_X := \{\psi \in \mathbb{F}_2[\Omega] : X[\psi] \in \mathcal{S}_X\} \cong \mathcal{S}_X \tag{7}$$

and

$$\mathcal{C}_Z := \{\psi \in \mathbb{F}_2[\Omega] : Z[\psi] \in \mathcal{S}_Z\} \cong \mathcal{S}_Z, \tag{8}$$

respectively. Note that for any $\psi, \phi \subseteq \Omega$, $X[\psi]$ and $Z[\phi]$ commute with each other if and only if ψ and ϕ are orthogonal. As $\mathcal{S} = \mathcal{S}_X \oplus \mathcal{S}_Z$ is Abelian, we must have $\mathcal{C}_X \perp \mathcal{C}_Z$. In this paper, we prefer the binary vector space representation of CSS codes and use the notation $\mathcal{C}_X \perp \mathcal{C}_Z$ to denote a CSS code.

For a CSS code $\mathcal{C}_X \perp \mathcal{C}_Z$ with a codespace $\mathcal{C}_Q \subseteq \mathcal{H}_\Omega$, the dimension of $\mathcal{C}_Q$ is 2^k, where

$$k = n - \dim \mathcal{C}_X - \dim \mathcal{C}_Z. \tag{9}$$

Thus we say that $\mathcal{C}_X \perp \mathcal{C}_Z$ encodes k logical qubits. For any vector $\psi \in \mathcal{C}_Z^\perp$, $X[\psi]$ commutes with any stabilizer and hence must be a logical operator. Similarly, for any $\phi \in \mathcal{C}_X^\perp$, $Z[\phi]$ is a logical operator. In fact, we can always find k vectors $\psi_1, \ldots, \psi_k \in \mathcal{C}_Z^\perp$ and another k vectors $\phi_1, \ldots, \phi_k \in \mathcal{C}_X^\perp$ such that $\psi_i^\mathsf{T} \phi_j = \delta_{ij}$. Defining $\bar{X}_i := X[\psi_i]$ and $\bar{Z}_j := Z[\phi_j]$, we have

$$\bar{X}_i \bar{Z}_j = (-1)^{\delta_{ij}} \bar{Z}_j \bar{X}_i. \tag{10}$$

Indeed, $\bar{X}_i$ and $\bar{Z}_i$ can be regarded as the logical Pauli X and Z operators of the ith logical qubit, respectively.

III. SYNDROME EXTRACTIONS FOR CSS CODES

On a stabilizer code, the errors are detected by measuring the stabilizer elements. For CSS codes, it is natural to measure the Z- and X-stabilizer elements so that the X errors (bit flips) and the Z errors (phase flips) can be handled separately. Here we are regarding a Y error as the combination of an X error and a Z error. In this paper, we focus on Z-stabilizer measurements, since an analysis of X-stabilizer measurements would be the same up to a Hadamard transform.

Let $\mathcal{C}_X \perp \mathcal{C}_Z$ be a CSS code on a set of data qubits D with a codespace $\mathcal{C}_Q \subseteq \mathcal{H}_D$. Suppose we have an X error $X[\psi]$ ($\psi \subseteq D$) and we are measuring the Z-stabilizer elements $\{Z[\phi_b]\}_{b\in B}$, where $\phi_b \in \mathcal{C}_Z$ and B is a finite set of syndrome bits. For each syndrome bit $b \in B$, the outcome of the measurement $Z[\phi_b]$ is determined by the value $\phi_b^\mathsf{T} \psi \in \mathbb{F}_2$. The binary vector

$$H\psi := \sum_{b\in B} b \phi_b^\mathsf{T} \psi \in \mathbb{F}_2[B] \tag{11}$$

is called the syndrome. The map $H = \sum_{b\in B} b\phi_b^\mathsf{T}$ is called a Z-check matrix of the CSS code $\mathcal{C}_X \perp \mathcal{C}_Z$. Note that we do not require the condition $\operatorname{im} H^\mathsf{T} = \mathcal{C}_Z$. However, we do require that $\operatorname{im} H^\mathsf{T} \subseteq \mathcal{C}_Z$, as $H^\mathsf{T} = \sum_{b\in B} \phi_b b^\mathsf{T}$ maps every $b \in B$ to $\phi_b \in \mathcal{C}_Z$.

If we are at the end of the computation, the syndrome of a Z-check matrix H can be obtained by measuring all qubits of D in the Z basis: If D is measured to be in the state $|\psi\rangle \in \mathcal{H}_D$, where $\psi \in \mathbb{F}_2[D]$, then the syndrome is simply $H\psi$. In the intermediate steps, however, we are not allowed to measure the data qubits directly, as the single-qubit Z measurements anticommute with the X-stabilizer elements. A general idea shared by both Shor's and Steane's syndrome extraction protocols is to transfer the X errors to a set of ancilla qubits A by CNOT gates: As the CNOT gate propagates the X errors on the control qubit to the target qubit, we can perform CNOT gates with controls in D and targets in A and then apply Z measurements on all ancilla qubits to detect these errors. Of course, this is far from being a valid construction. In the following, we explore the necessary and sufficient conditions for a valid extraction circuit. To simplify our problem, we do not initially consider the challenge of fault tolerance.

As a first step, we encode the information of the CNOT gates by a matrix $\Gamma : \mathbb{F}_2[D] \to \mathbb{F}_2[A]$, where for each data qubit $d \in D$ and ancilla qubit $a \in A$, $a^\mathsf{T} \Gamma d = 1$ if and only if a CNOT gate with control d and target a is applied. As these CNOT gates commute with each other and we do not consider the fault-tolerance properties, the order of these CNOT gates does not matter. The product of all CNOT gates is denoted by U_Γ. One can easily verify the identities

$$U_\Gamma X[\psi] X[\psi'] = X[\psi] X[\psi' + \Gamma\psi] U_\Gamma, \tag{12}$$

$$U_\Gamma Z[\phi] Z[\phi'] = Z[\phi + \Gamma^\mathsf{T}\phi'] Z[\phi'] U_\Gamma, \tag{13}$$

where $\psi, \phi \in \mathbb{F}_2[D]$ and $\psi', \phi' \in \mathbb{F}_2[A]$.

Suppose the ancilla block A is prepared in a state $|\text{anc}\rangle$. When there are no errors on the data block D, the action of U_Γ is required to be trivial, i.e., for any code state $|\omega\rangle \in \mathcal{C}_Q$, we should have

$$U_\Gamma |\omega\rangle |\text{anc}\rangle = |\omega\rangle |\text{anc}\rangle. \tag{14}$$

Note that the code space $\mathcal{C}_Q$ is spanned by

$$\{Z[\phi] |\overline{+^k}\rangle : \phi \in \mathcal{C}_X^\perp\}, \tag{15}$$

where $|\overline{+^k}\rangle$, the logical $|+^k\rangle$ state, is the CSS stabilizer state of $\mathcal{C}_Z^\perp \perp \mathcal{C}_Z$. This is true because we have enumerated all the logical Z-type operators. From (13), $Z[\phi]$ commutes with U_Γ for any $\phi \subseteq D$. Thus (14) can be simplified as

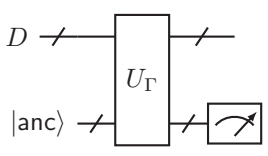
$$U_\Gamma |\overline{+^k}\rangle |\text{anc}\rangle = |\overline{+^k}\rangle |\text{anc}\rangle. \tag{16}$$

Let $X[\psi]$ ($\psi \subseteq D$) be an X error on a codeword $|\omega\rangle \in \mathcal{C}_Q$. From (12) we have

$$U_\Gamma (X[\psi]|\omega\rangle) |\text{anc}\rangle = (X[\psi]|\omega\rangle)(X[\Gamma\psi]|\text{anc}\rangle). \tag{17}$$

If $|\text{anc}\rangle$ has a nontrivial Z stabilizer represented by $\tilde{\mathcal{C}}_Z \subseteq \mathbb{F}_2[A]$, fixing a Z-check matrix $\tilde{H} : \mathbb{F}_2[A] \to \mathbb{F}_2[B]$ with $\operatorname{im} \tilde{H}^\mathsf{T} \subseteq \tilde{\mathcal{C}}_Z$, the syndrome of $\tilde{H}$ will be $\tilde{H}\Gamma\psi \in \mathbb{F}_2[B]$. If $H = \tilde{H}\Gamma$, we can obtain $H\psi = \tilde{H}\Gamma\psi$, the syndrome of H, by measuring all qubits of A in the Z basis. One could naturally ask the following question.

Question 1. Given two matrices $\tilde{H} : \mathbb{F}_2[A] \to \mathbb{F}_2[B]$ and $\Gamma : \mathbb{F}_2[D] \to \mathbb{F}_2[A]$ such that $H = \tilde{H}\Gamma$ is a Z-check matrix of $\mathcal{C}_X \perp \mathcal{C}_Z$, can we find a state $|\text{anc}\rangle \in \mathcal{H}_A$ such that $U_\Gamma |\overline{+^k}\rangle |\text{anc}\rangle = |\overline{+^k}\rangle |\text{anc}\rangle$ and $Z[\phi]|\text{anc}\rangle = |\text{anc}\rangle$ for every $\phi \in \operatorname{im} \tilde{H}^\mathsf{T}$? In other words, can the circuit



extract the syndrome of H?

Suppose we already have a satisfying ancilla state $|\text{anc}\rangle$ for a given $\tilde{H}$ and Γ. In (17), if we take $\psi \in \mathcal{C}_Z^\perp$ and combine (16), we will have

$$\overline{|+^k\rangle} \otimes (X[\Gamma\psi]|\text{anc}\rangle) = \overline{|+^k\rangle} \otimes |\text{anc}\rangle,$$

i.e., $X[\Gamma\psi]|\text{anc}\rangle = |\text{anc}\rangle$. Using $\tilde{\mathcal{C}}_X \subseteq \mathbb{F}_2[A]$ to represent the X stabilizer of $|\text{anc}\rangle$, $X[\Gamma\psi]|\text{anc}\rangle = |\text{anc}\rangle$ is equivalent as $\Gamma\psi \in \tilde{\mathcal{C}}_X$. Therefore, we must have

$$\Gamma(\mathcal{C}_Z^\perp) \subseteq \tilde{\mathcal{C}}_X. \tag{18}$$

On the other hand, let $\tilde{\mathcal{C}}_Z \subseteq \mathbb{F}_2[A]$ represents the Z-stabilizer group of $|\text{anc}\rangle$. For every $\phi \in \tilde{\mathcal{C}}_Z$ we have

$$\overline{|+^k\rangle}|\text{anc}\rangle = U_\Gamma \overline{|+^k\rangle}(Z[\phi]|\text{anc}\rangle) = (Z[\Gamma^\mathsf{T}\phi]\overline{|+^k\rangle})|\text{anc}\rangle.$$

Therefore, $\Gamma^\mathsf{T}\phi \in \mathcal{C}_Z$ and hence

$$\Gamma^\mathsf{T}(\tilde{\mathcal{C}}_Z) \subseteq \mathcal{C}_Z. \tag{19}$$

The conditions (18) and (19) imply that U_Γ preserves the CSS stabilizer group

$$(\mathcal{C}_Z^\perp \oplus \tilde{\mathcal{C}}_X) \perp (\mathcal{C}_Z \oplus \tilde{\mathcal{C}}_Z). \tag{20}$$

If $\tilde{\mathcal{C}}_X = \tilde{\mathcal{C}}_Z^\perp$, i.e., $|\text{anc}\rangle$ is a CSS stabilizer state and hence $\overline{|+^k\rangle}|\text{anc}\rangle$ is the stabilizer state of (20), the condition (16) must hold.

Our problem now becomes finding a CSS stabilizer state $|\text{anc}\rangle$ with X and Z stabilizers $\tilde{\mathcal{C}}_X$ and $\tilde{\mathcal{C}}_Z$ ($\tilde{\mathcal{C}}_X = \tilde{\mathcal{C}}_Z^\perp$), respectively, satisfying (18), (19), and the condition $\text{im}\,\tilde{H}^\mathsf{T} \subseteq \tilde{\mathcal{C}}_Z$. A natural strategy is to minimize $\tilde{\mathcal{C}}_Z$ and maximize $\tilde{\mathcal{C}}_X = \tilde{\mathcal{C}}_Z^\perp$. The minimal choice of $\tilde{\mathcal{C}}_Z$ is just $\text{im}\,\tilde{H}^\mathsf{T}$ as required. The following lemma shows that $\tilde{\mathcal{C}}_Z = \text{im}\,\tilde{H}^\mathsf{T}$ and $\tilde{\mathcal{C}}_X = \ker\tilde{H}$ always satisfy (18) and (19).

Lemma 1. $\Gamma^\mathsf{T}(\text{im}\,\tilde{H}^\mathsf{T}) \subseteq \mathcal{C}_Z$ and $\Gamma(\mathcal{C}_Z^\perp) \subseteq \ker\tilde{H}$.

Proof. The first part follows from a direct calculation

$$\Gamma^\mathsf{T}(\text{im}\,\tilde{H}^\mathsf{T}) = \Gamma^\mathsf{T}\tilde{H}^\mathsf{T}(\mathbb{F}_2[B]) = \text{im}\,H^\mathsf{T} \subseteq \mathcal{C}_Z.$$

The latter part is also straightforward: For any $\psi \in \mathcal{C}_Z^\perp$, we have $\tilde{H}\Gamma\psi = H\psi = 0$. Thus $\Gamma\psi \in \ker\tilde{H}$.

From the discussion above, we can conclude that given a Z-check matrix H of $\mathcal{C}_X \perp \mathcal{C}_Z$, any decomposition $H = \tilde{H}\Gamma$ corresponds a valid gadget extracting the syndrome of H. All the information of the gadget is determined by $\tilde{H}$ and Γ, which allows us to define a gadget in an abstract way.

Definition 1 (Z-extraction gadgets). A Z-extraction gadget of a CSS code $\mathcal{C}_X \perp \mathcal{C}_Z$ on the block D is a tuple $(A, B, \tilde{H}, \Gamma)$, where A is the set of ancilla qubits, B is the set of syndrome bits, and $\tilde{H} : \mathbb{F}_2[A] \to \mathbb{F}_2[B]$ and $\Gamma : \mathbb{F}_2[D] \to \mathbb{F}_2[A]$ are two matrices such that $H = \tilde{H}\Gamma$ is a Z-check matrix of $\mathcal{C}_X \perp \mathcal{C}_Z$. Here H, $\tilde{H}$, and Γ are referred to as the data check matrix, ancilla check matrix, and gate matrix, respectively. The ancilla block A is prepared in the CSS state of $\ker\tilde{H} \perp \text{im}\,\tilde{H}^\mathsf{T}$, while the gate applied between D and A is U_Γ.

If we apply two gadgets $\mathcal{G}_1 = (A_1, B_1, \tilde{H}_1, \Gamma_1)$ and $\mathcal{G}_2 = (A_2, B_2, \tilde{H}_2, \Gamma_2)$, we can view them as a united gadget $\mathcal{G} = (A, B, \tilde{H}, \Gamma)$ such that $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$. The gate matrix Γ and the ancilla check matrix $\tilde{H}$ are defined by

$$\Gamma = \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \end{pmatrix} \tag{21}$$

and

$$\tilde{H} = \begin{pmatrix} \tilde{H}_1 & 0 \\ 0 & \tilde{H}_2 \end{pmatrix}, \tag{22}$$

respectively. The total parity check matrix

$$H = \tilde{H}\Gamma = \begin{pmatrix} \tilde{H}_1\Gamma_1 \\ \tilde{H}_2\Gamma_2 \end{pmatrix}. \tag{23}$$

We say that $\mathcal{G}$ is a sum of $\mathcal{G}_1$ and $\mathcal{G}_2$, denoted by $\mathcal{G} = \mathcal{G}_1 \oplus \mathcal{G}_2$.

We now review Shor's and Steane's constructions in our notation.

Example 1 (Shor's scheme). In Shor's scheme [9,10], each syndrome bit $b \in B$ corresponding to the Z-stabilizer element $Z[\phi_b]$ ($\phi_b \subseteq D$) is extracted by a separate gadget $\mathcal{G}_b = (A_b, \{b\}, \tilde{H}_b, \Gamma_b)$ such that $\Gamma_b^\mathsf{T}\tilde{H}_b^\mathsf{T} b = \phi_b$. The whole gadget is therefore $\mathcal{G} = \bigoplus_{b\in B} \mathcal{G}_b$ and we say that $\mathcal{G}$ is a Shor-style gadget.

The simplest choice of $\mathcal{G}_b$ is to set $A_b = \{a\}$ to have one ancilla qubit a, and $\tilde{H}_b^\mathsf{T} b = a$ and $\Gamma_b^\mathsf{T} a = \phi_b$. Since $\text{im}\,\tilde{H}_b^\mathsf{T} \cong \mathbb{F}_2$ and $\ker\tilde{H}_b = 0$, the ancilla qubit a is stabilized by the single-qubit Pauli Z operator. This is known as the bare-ancilla gadget. If all the $\mathcal{G}_b$ are bare ancilla, the gate matrix and the data check matrix of $\mathcal{G}$ are identical, while the ancilla check matrix is the identity matrix of dimension $|B|$.

Another choice of $\mathcal{G}_b$ is to set $|A_b| = |\phi_b|$. We fix a set bijection $\gamma : A_b \leftrightarrow \phi_b$ and define the gate matrix Γ_b by $\Gamma_b^\mathsf{T} a = \gamma(a)$ for every $a \in A_b$. The ancilla check matrix is defined by $\tilde{H}_b^\mathsf{T} b = A_b \in \mathbb{F}_2[A_b]$. One can verify that the ancilla state is the cat state $|+^{|\phi_b|}\rangle + |-^{|\phi_b|}\rangle$. This is known as the cat-state gadget.

Example 2 (Steane's scheme). In Steane's scheme [13], all the Z-stabilizer elements are extracted by one ancilla block. The ancilla block A is a set identical to the data block D. The gate matrix Γ is an identity matrix under the standard bases, and the ancilla check matrix $\tilde{H}$ is identical to the data check matrix H. The ancilla state is the CSS stabilizer state of $\mathcal{C}_Z^\perp \perp \mathcal{C}_Z$, which is the logical $|+^k\rangle$ state, where k is the number of logical qubits.

Shor's scheme benefits from having small transversal cat-state gadgets with ancilla blocks whose sizes are defined by the weight of the measured stabilizers, $|A_b| = |\phi_b|$. If the stabilizers are measured in a serial manner, one only needs a number of ancilla qubits equal to the highest-weight stabilizer $\max_b(|\phi_b|)$. Parallel syndrome measurement is desirable and one will need F ancilla qubits where $F = \sum_{b\in B} |\phi_b|$. If each one of the n data qubits interacts on average with f measured stabilizers, then we can also write $F = fn$. Measurements need to be repeated T times until errors due to measurement can be distinguished from errors on the data; T can scale as $O(d^2)$ [9,23,24] in the worst case, but scales as $O(d)$ for topological codes [19].

Steane's scheme requires large ancilla-block size equal to the data block. However, only one block is needed for Z extraction. The total number of ancilla qubits is just n, which is less than the fn needed for the parallel Shor scheme. In addition, Steane's scheme is a single-shot method, so T is constant. The cost of Steane's scheme is the preparation and verification of the ancilla block, which becomes challenging as the code distance increases. However, if we can prepare the ancilla, it greatly reduces the number of interactions

between the data block and the ancilla block before correction is applied.

Our goal is to construct gadgets other than Shor- and Steane-style ones that result in simplified ancilla blocks compared to Steane's and fewer interactions between data and ancilla than Shor's. Conceptually, we can start with Shor ancilla blocks and merge ancilla blocks in a way that removes ancilla qubits. We now describe this procedure in detail.

We start from Shor's cat-state syndrome measurement circuit, denoted by $\mathcal{G} = (A, B, \tilde{H}, \Gamma)$. For every ancilla qubit $a \in A$, it interacts with exactly one data qubit and transfers the error to exactly one syndrome bit, i.e., $|\Gamma^{\mathsf{T}} a| = |\tilde{H} a| = 1$. In matrix language, each row of Γ and each column of $\tilde{H}$ contains exactly one nontrivial entry. We construct gadgets by applying a series of steps, each of which is described by one of the following operations.

(i) Pick two ancillas $a_1, a_2 \in A$ with $\Gamma^{\mathsf{T}} a_1 = \Gamma^{\mathsf{T}} a_2$. Merge a_1 and a_2 as one ancilla qubit a. Update Γ by adding a row $a^{\mathsf{T}}\Gamma = a_1^{\mathsf{T}}\Gamma$ and deleting the rows $a_1^{\mathsf{T}}\Gamma$ and $a_2^{\mathsf{T}}\Gamma$. Update $\tilde{H}$ by adding a column $\tilde{H} a = \tilde{H}(a_1 + a_2)$ and then deleting the columns $\tilde{H} a_1$ and $\tilde{H} a_2$.

(ii) Pick two ancillas $a_1, a_2 \in A$ with $\tilde{H} a_1 = \tilde{H} a_2$. Merge a_1 and a_2 as one ancilla qubit a. Update Γ by adding a row $a^{\mathsf{T}}\Gamma = a_1^{\mathsf{T}}\Gamma + a_2^{\mathsf{T}}\Gamma$ and then deleting the rows $a_1^{\mathsf{T}}\Gamma$ and $a_2^{\mathsf{T}}\Gamma$. Update $\tilde{H}$ by adding a column $\tilde{H} a = \tilde{H} a_1$ and then deleting $\tilde{H} a_1$ and $\tilde{H} a_2$.

It is easy to check that the updated $\tilde{H}$ and Γ satisfy the condition $H = \tilde{H}\Gamma$. If we only apply (i), each row of Γ will always have exactly 1 nontrivial entry. Indeed, the gadget will be transversal on the ancilla. Assuming the ancilla block can be fault-tolerantly prepared by postselection so that no correlated errors can occur, a transversal gadget will not introduce correlated errors to the data block. If we keep doing (i), we will eventually obtain Steane's scheme. On the other hand, if we only apply (ii), we make the cat states smaller and introduce nontransversal interactions. The extreme case where no ancilla can be further merged is the bare-ancilla scheme.

By allowing both types of ancilla merging, we can generate a large number of gadgets. In fact, arguably, all the gadgets can be generated in this way. To see this, for any gadget $\mathcal{G} = (A, B, \tilde{H}, \Gamma)$ we can split each ancilla qubit $a \in A$ into $|\Gamma^{\mathsf{T}} a| \times |\tilde{H} a|$ ancilla qubits: With each data qubit $q \in \Gamma^{\mathsf{T}} a$ and syndrome bit $b \in \tilde{H} a \subseteq B$, we associate an ancilla qubit to pass the X error on q to b. The obtained gadget is Shor's cat-state syndrome measurement scheme, and the reverse of the splitting provides us the merging steps from Shor's scheme to $\mathcal{G}$.

However, such an argument has a few exceptions. First, if $|\Gamma^{\mathsf{T}} a| = 0$ or $|\tilde{H} a| = 0$, a will be split into 0 qubits so that the process is irreversible. Second, for each data qubit q and each syndrome bit b, the number of ancilla qubits connecting them after splitting, whose parity is $b^{\mathsf{T}} H q$, can be more than 1. In contrast, for Shor's scheme, there must be exactly $b^{\mathsf{T}} H q \in \{0, 1\}$ ancilla qubits. To handle these exceptions, we can introduce pairs of redundant ancilla qubits to Shor's scheme before merging: Each pair of ancilla qubits passes the error on the same data qubit to the same syndrome bit. After merging, we can add ancilla qubits that only connect syndrome bits. Of course, we can add ancilla qubits that only connect data qubits as well. These ancilla qubits, however, are not helpful for gaining error information.

The tradeoffs of two merging operations are different: Operation (i) reduces the number of CNOT gates between data and ancillas while making the ancilla block more entangled; (ii) simplifies the ancilla-block preparation while taking the risk of introducing correlated errors and breaking fault tolerance. If we wish to have postselection-free ancillas, the stabilizer generators of the ancilla block should have weight no more than 3 so that the residual error of each single error in the preparation circuit can be reduced to a weight-1 error. Bare ancillas, Bell states, and three-qubit Greenberger-Horne-Zeilinger states are example states that satisfy this requirement. More generally, any CSS state equivalent to a one-dimensional cluster state (up to Hadamard gates) can be directly prepared. To preserve the code distance, the correlated errors introduced by these simple gadgets need to be handled by either a well-designed decoder or modifications to the decoding circuits, such as flag qubits [16,18,30] and DiVincenzo-Aliferis ancilla decoding [25]. However, as the matrix decomposition $H = \tilde{H}\Gamma$ inherits the code structure from H, the detailed fault-tolerance design will be code specific.

In Fig. 1 we illustrate how to obtain syndrome extraction methods for Steane's seven-qubit code by applying merging operations on Shor's scheme. Instead of applying only one type of operation, we can apply both. As an example, we apply (i) on Shor's scheme twice to obtain a transversal gadget (referred to as scheme A in Fig. 1) and then apply (ii) five times to obtain a non-fault-tolerant circuit (referred to as scheme B) whose ancilla state can be prepared without verification. As Steane's code has distance 3, one can apply a DiVincenzo-Aliferis decoding circuit [25] to make scheme B fault tolerant. The details of the protocol are given in Appendix A.

When measurement errors are being considered, the use of different extraction gadgets will lead to different decoding problem details. In the next section, as an example, we will study the behavior of transversal gadgets on Kitaev's toric code [4] thoroughly. In particular, we will show that by switching between different gadgets, the time overhead of fault tolerance can be reduced without increasing the ancilla complexity.

IV. BLOCK EXTRACTIONS OF THE TORIC CODE

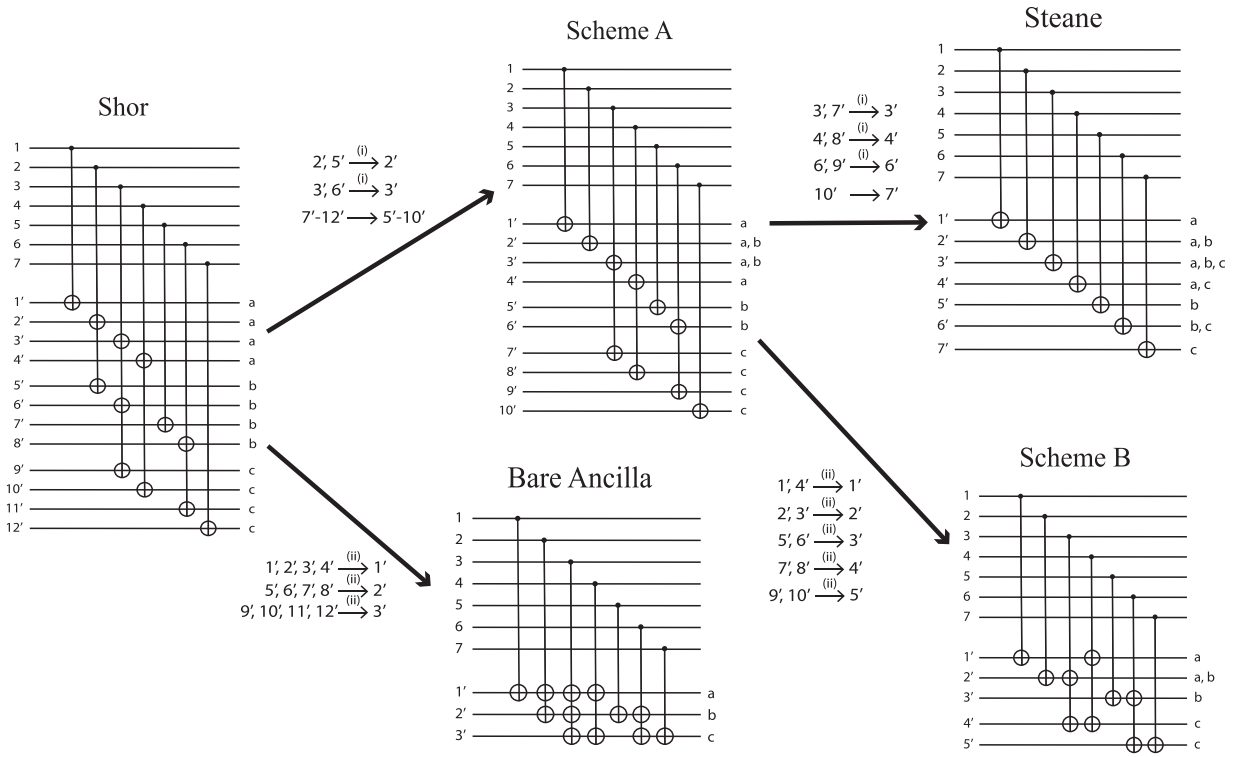
We briefly review the construction of the toric code [4]. A toric code is a CSS code defined on an $L \times L$ periodic lattice on the torus. The lattice has a set V of L^2 vertices, a set E of $2L^2$ edges, and a set F of L^2 faces. Define the boundary map

$$\begin{aligned} \partial : \mathbb{F}_2[F] &\mapsto \mathbb{F}_2[E], \\ F \ni f &\mapsto \{e \in E | e \text{ borders } f\} \end{aligned} \tag{24}$$

and the coboundary map

$$\begin{aligned} \delta : \mathbb{F}_2[V] &\mapsto \mathbb{F}_2[E], \\ V \ni v &\mapsto \{e \in E | e \text{ is incident to } v\}. \end{aligned} \tag{25}$$

For each $f \in F$ and $v \in V$, since $|\partial f \cap \delta v| = 0$ or 2, we must have $\langle \partial f, \delta v \rangle_E = 0$. Indeed, $\text{im}\,\delta \perp \text{im}\,\partial$. Taking $\mathcal{C}_X = \text{im}\,\delta$ and $\mathcal{C}_Z = \text{im}\,\partial$, $\mathcal{C}_X \perp \mathcal{C}_Z$ is a well-defined CSS code on the



	Bare Ancilla	Cat State	Steane	Scheme A	Scheme B
H			$\begin{bmatrix}1&1&1&1&0&0&0\\0&1&1&0&1&1&0\\0&0&1&1&0&1&1\end{bmatrix}$		
$\tilde{H}$	$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	$\begin{bmatrix}1&1&1&1&&&&&&&&\\&&&&1&1&1&1&&&&\\&&&&&&&&1&1&1&1\end{bmatrix}$	H	$\begin{bmatrix}1&1&1&1&0&0&\\0&1&1&0&1&1&\\&&&&&&1&1&1&1\end{bmatrix}$	$\begin{bmatrix}1&1&0&&&\\0&1&1&&&\\&&&1&1&\end{bmatrix}$
Γ	H	$\begin{bmatrix}1&0&0&0&0&0&0\\0&1&0&0&0&0&0\\0&0&1&0&0&0&0\\0&0&0&1&0&0&0\\0&1&0&0&0&0&0\\0&0&1&0&0&0&0\\0&0&0&0&1&0&0\\0&0&0&0&0&1&0\\0&0&1&0&0&0&0\\0&0&0&1&0&0&0\\0&0&0&0&0&1&0\\0&0&0&0&0&0&1\end{bmatrix}$	$\begin{bmatrix}1&0&0&0&0&0&0\\0&1&0&0&0&0&0\\0&0&1&0&0&0&0\\0&0&0&1&0&0&0\\0&0&0&0&1&0&0\\0&0&0&0&0&1&0\\0&0&0&1&0&0&0\\0&0&0&0&0&1&0\\0&0&0&0&0&0&1\end{bmatrix}$	$\begin{bmatrix}1&0&0&0&0&0&0\\0&1&1&0&0&0&0\\0&0&0&0&1&1&0\\0&0&1&1&0&0&0\\0&0&0&0&0&1&1\end{bmatrix}$
Z-stabilizer generators	$Z_{1'}, Z_{2'}, Z_{3'}$	$Z_{1'}Z_{2'}Z_{3'}Z_{4'}, Z_{5'}Z_{6'}Z_{7'}Z_{8'}, Z_{9'}Z_{10'}Z_{11'}Z_{12'}$	$Z_{1'}Z_{2'}Z_{3'}Z_{4'}, Z_{2'}Z_{3'}Z_{5'}Z_{6'}, Z_{3'}Z_{4'}Z_{6'}Z_{7'}$	$Z_{1'}Z_{2'}Z_{3'}Z_{4'}, Z_{2'}Z_{3'}Z_{5'}Z_{6'}, Z_{7'}Z_{8'}Z_{9'}Z_{10'}$	$Z_{1'}Z_{2'}, Z_{2'}Z_{3'}, Z_{4'}Z_{5'}$
X-stabilizer generators	None	$X_{1'}X_{2'}, X_{2'}X_{3'}, X_{3'}X_{4'}, X_{5'}X_{6'}, X_{6'}X_{7'}, X_{7'}X_{8'}, X_{9'}X_{10'}, X_{10'}X_{11'}, X_{11'}X_{12'}$	$X_{1'}X_{2'}X_{3'}X_{4'}, X_{2'}X_{3'}X_{5'}X_{6'}, X_{3'}X_{4'}X_{6'}X_{7'}, X_{1'}X_{2'}X_{5'}$	$X_{1'}X_{2'}X_{3'}X_{4'}, X_{2'}X_{3'}X_{5'}X_{6'}, X_{3'}X_{4'}X_{6'}, X_{1'}X_{2'}X_{5'}, X_{7'}X_{8'}, X_{8'}X_{9'}, X_{9'}X_{10'}$	$X_{1'}X_{2'}X_{3'}, X_{4'}X_{5'}$

FIG. 1. For the Steane [[7,1,3]] code, we show how to construct syndrome extraction circuits by merging ancilla qubits in Shor's scheme. In the circuit diagram, the labeled sets of the data qubits, ancilla qubits, and syndrome bits are $\{1, \ldots, 7\}$, $\{1', \ldots, 12'\}$, and $\{a, b, c\}$, respectively. The syndrome bits can be determined by a collective measurement of the labeled qubit outputs. Two ancilla qubits can be merged if they (i) talk to the same set of data qubits or (ii) contribute to the same set of syndrome bits. If we repeat application of rule (i), we will eventually obtain Steane's scheme. Bare-ancilla extraction is the limit for applying rule (ii) only. For example, we demonstrate scheme A, a transversal gadget obtained by applying (i) to Shor's scheme twice, and scheme B, obtained by applying (ii) on scheme A five times. We illustrate the matrix decomposition description $H = \tilde{H}\Gamma$ of these gadgets. Empty elements in $\tilde{H}$ are zeros, which are not shown to emphasize the block structure. The Z- and X-stabilizer generators are generated from im $\tilde{H}^{\mathsf{T}}$ and ker $\tilde{H}$, respectively. We note that the ancilla state for scheme B is $(|000\rangle + |111\rangle) \otimes (|00\rangle + |11\rangle)$, which can be directly prepared without verification. However, as the circuit is not transversal on the ancilla block, we will require a DiVincenzo-Aliferis decoding circuit [25] to achieve fault tolerance. The details are given in Appendix A.

edge set E. A nontrivial logical Z-type (X-type) operator is represented by a noncontractible loop (cut) on the torus, which has minimum length L. In other words, the toric code has distance L. The Z- and X-check matrices of the toric code are ∂^{T} and δ^{T}, respectively. A Z-extraction gadget is therefore represented by a decomposition $\partial^{\mathsf{T}} = \tilde{\partial}^{\mathsf{T}}\gamma^{\mathsf{T}}$, or $\partial = \gamma\tilde{\partial}$. The matrices $\tilde{\partial}^{\mathsf{T}}$ and γ^{T} are the ancilla check and gate matrices, respectively.

The Shor-style extraction gadget, denoted by $(\tilde{E}_0, F, \tilde{\partial}_0^\mathsf{T}, \gamma_0^\mathsf{T})$, can be understood as cutting the torus into L^2 disjoint square faces. The set $\tilde{E}_0$ contains the $4L^2$ edges of these squares. For two edges $\tilde{e}_1, \tilde{e}_2 \in \tilde{E}_0$, if they were the same edge $\gamma_0\tilde{e}_1 = \gamma_0\tilde{e}_2 \in E$ before the torus was cut, we can glue the two edges back, which corresponds to a type-(i) merging of ancilla qubits $\tilde{e}_1$ and $\tilde{e}_2$. Therefore, any transversal gadgets can be constructed by cutting the torus along some chosen edges. If we do not make any cut, the gadget will be Steane style; if we choose to cut along all edges, we will obtain the Shor-style gadget. In general, we will obtain a new topological space, with an edge set $\tilde{E}$ and a boundary map $\tilde{\partial} : \mathbb{F}_2[F] \to \mathbb{F}_2[\tilde{E}]$. The map $\gamma : \mathbb{F}_2[\tilde{E}] \to \mathbb{F}_2[E]$ maps an edge $\tilde{e} \in \tilde{E}$ back to its corresponding edge $e \in E$ on the torus. The boundary of the space is $\tilde{\partial}F \subseteq \tilde{E}$. Here, as a reminder, F is viewed as an $\mathbb{F}_2$-vector. If we cut along an edge $e \in E$ so that $|\gamma^\mathsf{T} e| = 2$, we must have $\gamma^\mathsf{T} e \subseteq \tilde{\partial}F$. If we cut the torus into several connected components so that the face set F is decomposed as $F = \bigcup_i F_i$, where each F_i is the face set of a component, then the gadget $(\tilde{E}, F, \tilde{\partial}^\mathsf{T}, \gamma^\mathsf{T})$ can be decomposed as

$$(\tilde{E}, F, \tilde{\partial}^\mathsf{T}, \gamma^\mathsf{T}) = \bigoplus_i (\tilde{E}_i, F_i, (\tilde{\partial}|_{F_i})^\mathsf{T}, (\gamma|_{\tilde{E}_i})^\mathsf{T}), \tag{26}$$

where each $\tilde{E}_i := \bigcup_{f\in F_i} \tilde{\partial} f$ is the set of edges of the component F_i, $\tilde{\partial}|_{F_i}$ is the restriction of $\tilde{\partial}$ on F_i, and $\gamma|_{\tilde{E}_i}$ is the restriction of γ on $\tilde{E}_i$. That is, the ancilla block $\tilde{E}$ is decomposed as subblocks $\tilde{E}_i$, and two different blocks are not entangled.

A. Errors in space-time

We now describe our model of fault-tolerant error correction following the presentation in Ref. [40]. As X and Z errors can be corrected separately, we can ignore the Z errors on the data qubits and the X-stabilizer measurements. Suppose our computation starts from time 0 and never ends. We extract the Z-check matrix ∂ at every positive integer time. For every $t \in \mathbb{Z}_+$, each data qubit could have an X error at time $t - \frac{1}{2}$, and the measurement outcome of each ancilla qubit at time t could have a classical bit-flip error. For now, we ignore the data errors between two CNOT gates in the same extraction round.

Suppose the transversal gadget applied at time $t \in \mathbb{N}$ is $\mathcal{G}_t = (\tilde{E}_t, F, \tilde{\partial}_t^\mathsf{T}, \gamma_t^\mathsf{T})$, where $\gamma_t\tilde{\partial}_t = \partial$. An X error on the data qubit $e \in E$ at time $t - \frac{1}{2}$ is denoted by the pair (e, t). The measurement error on the ancilla qubit $\tilde{e} \in \tilde{E}_t$ at time $t \in T$ is denoted by the pair $(\tilde{e}, t)$. The set of all single data-qubit errors, denoted by D, is

$$D = E \times \mathbb{N}, \tag{27}$$

while the set of measurement errors, denoted by M, is

$$M = \bigsqcup_{t\in\mathbb{N}} \tilde{E}_t := \bigcup_{t\in\mathbb{N}} \tilde{E}_t \times \{t\}. \tag{28}$$

An error history is defined to be a finite subset $\psi \subseteq D \cup M$, or equivalently a vector $\psi \in \mathbb{F}_2[D \cup M]$. The evaluation of data errors at time $t - \frac{1}{2}$ ($t \in \mathbb{N}$) is a map

$$\mathcal{D}_t : \mathbb{F}_2[D \cup M] \mapsto \mathbb{F}_2[E], \quad \psi \mapsto \{e : (e, t) \in \psi \cap D\}. \tag{29}$$

For later convenience, the time coordinates of the errors are discarded. Similarly, the evaluation of measurement errors at time t is a map

$$\mathcal{M}_t : \mathbb{F}_2[D \cup M] \mapsto \mathbb{F}_2[\tilde{E}_t], \quad \psi \mapsto \{\tilde{e} : (\tilde{e}, t) \in \psi \cap M\}. \tag{30}$$

The data error propagating to the ancilla block $\tilde{E}_t$ is the accumulation of all data errors before time t, which can be evaluated by the map

$$\bar{\mathcal{D}}_t := \sum_{t'\leqslant t} \mathcal{D}_{t'}. \tag{31}$$

In particular, we define $\bar{\mathcal{D}} := \sum_{t\in\mathbb{N}} \mathcal{D}_t$. The syndrome at time t can therefore be evaluated by the map

$$\sigma_t = \tilde{\partial}_t^\mathsf{T}(\gamma_t^\mathsf{T}\bar{\mathcal{D}}_t + \mathcal{M}_t) = \partial^\mathsf{T}\bar{\mathcal{D}}_t + \tilde{\partial}_t^\mathsf{T}\mathcal{M}_t. \tag{32}$$

We can take the difference of the syndrome sequence $\{\sigma_t\}_{t\in\mathbb{N}}$,

$$\begin{aligned}\Delta_t &:= \sigma_t - \sigma_{t-1}\\ &= \partial^\mathsf{T}\bar{\mathcal{D}}_t + \tilde{\partial}_t^\mathsf{T}\mathcal{M}_t - \partial^\mathsf{T}\bar{\mathcal{D}}_{t-1} - \tilde{\partial}_{t-1}^\mathsf{T}\mathcal{M}_{t-1}\\ &= \partial^\mathsf{T}(\bar{\mathcal{D}}_t - \bar{\mathcal{D}}_{t-1}) + \tilde{\partial}_t^\mathsf{T}\mathcal{M}_t - \tilde{\partial}_{t-1}^\mathsf{T}\mathcal{M}_{t-1}\\ &= \partial^\mathsf{T}\mathcal{D}_t + \tilde{\partial}_t^\mathsf{T}\mathcal{M}_t - \tilde{\partial}_{t-1}^\mathsf{T}\mathcal{M}_{t-1}\\ &= \partial^\mathsf{T}\mathcal{D}_t + \tilde{\partial}_t^\mathsf{T}\mathcal{M}_t + \tilde{\partial}_{t-1}^\mathsf{T}\mathcal{M}_{t-1}.\end{aligned} \tag{33}$$

In the above calculation, we set $\sigma_0 = 0$, $\bar{\mathcal{D}}_0 = 0$, and $\mathcal{M}_0 = 0$ for convenience. We can see that the data error $\mathcal{D}_t$ only contributes to Δ_t, while the measurement error $\mathcal{M}_t$ contributes to both Δ_t and Δ_{t+1}. The syndrome history is defined as a map

$$\Sigma : \mathbb{F}_2[D \cup M] \mapsto \mathbb{F}_2[F \times \mathbb{N}], \quad \psi \mapsto \bigsqcup_{t\in\mathbb{N}} \Delta_t\psi. \tag{34}$$

We now analyze the behavior of each single error. For each data error (e, t), one can verify that

$$\Sigma(e, t) = (\partial^\mathsf{T} e) \times \{t\} = \{f_1, f_2\} \times \{t\}, \tag{35}$$

where $f_1, f_2 \in F$ are the two faces sharing e as their borders. For each measurement error $(\tilde{e}, t) \in M$, where $\tilde{e} \in \tilde{E}_t$, one can verify that

$$\Sigma(\tilde{e}, t) = (\tilde{\partial}_t^\mathsf{T}\tilde{e}) \times \{t, t + 1\}. \tag{36}$$

If $\tilde{e} \in \tilde{\partial}_t F$, i.e., $\tilde{e}$ is an split edge, $\tilde{\partial}_t^\mathsf{T}\tilde{e}$ has only one face and we say that $(\tilde{e}, t)$ is a type-I error. Otherwise, $|\tilde{\partial}_t^\mathsf{T}\tilde{e}| = 2$ and $(\tilde{e}, t)$ is said to be a type-II error. The set of type-I errors, denoted by M_1, is

$$M_1 = \bigsqcup_{t\in\mathbb{N}} \tilde{\partial}_t F, \tag{37}$$

while the set of type-II errors is denoted by $M_2 = M - M_1$.

The syndrome bit flips can be viewed as defects in the $(2 + 1)$-dimensional lattice: Each data error creates two defects in the same time slice: Each type-I error creates two defects on the same location, but in two consecutive time slices, and each type-II error creates four defects. We give an example of the effect of different error types in Fig. 2. If $\mathcal{G}_t$ is a Shor-style gadget, all measurement errors at time t will be of type I, and the data and measurement errors are referred to as spacelike

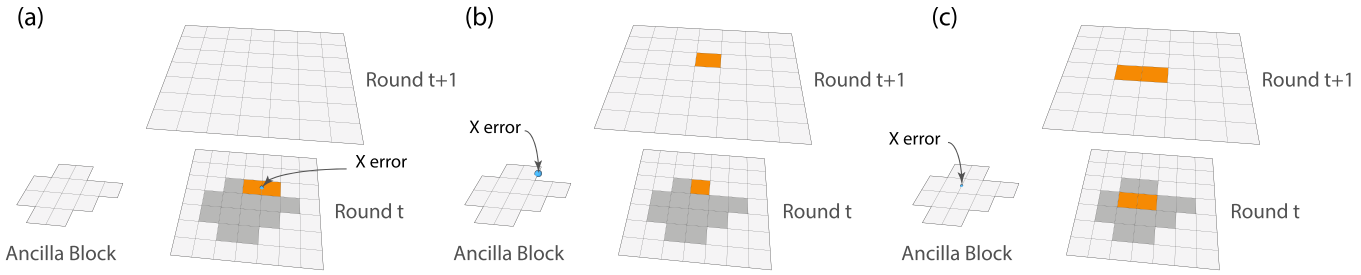


FIG. 2. Example of toric code syndrome extraction via our transversal gadget construction. At time t, a verified ancilla block consisting of 13 faces is used to extract the syndrome bits (face operators) in the gray region via transversal CNOT gates. (a) A bit-flip data error creates a pair of defects at time t. (b) A measurement error on the boundary of the ancilla block (type-I error) creates a defect at time t and another one at $t+1$. (c) A measurement error in the bulk of the ancilla block (type-II error) creates a defect pair at time t and another pair at time $t+1$.

and timelike errors, respectively [19]. If $\mathcal{G}_t$ is a Steane-style gadget, however, all measurement errors at time t will be of type II.

B. Minimum-weight perfect matching decoder

Given an error history $\psi \in \mathbb{F}_2[D \cup M]$ with observed syndrome history $\Sigma\psi \in \mathbb{F}_2[F \times \mathbb{N}]$, a decoder

$$\mathrm{Dec} : \mathbb{F}_2[F \times \mathbb{N}] \to \mathbb{F}_2[D \cup M] \tag{38}$$

will take $\Sigma\psi$ as the input and output an estimation of error history $\psi' = \mathrm{Dec}(\Sigma\psi)$ such that $\Sigma\psi' = \Sigma\psi$. The optimal choice of the decoder is the minimum-weight-error (MWE) decoder $\mathrm{Dec}_{\mathrm{MWE}}$ defined by

$$Dec_{\mathrm{MWE}}(\Sigma\psi) = \mathrm{argmin}_{\Sigma\psi'=\Sigma\psi}|\psi'|. \tag{39}$$

If the gadgets are all Shor style so that $M = M_1$, the decoding problem can be visualized by a decoder graph G with vertex set $F \times \mathbb{N}$ and edge set $D \cup M_1$: Each $(f,t) \in F \times \mathbb{N}$ is a vertex and each error $\psi \in D \cup M_1$ is an edge connecting the two vertices (defects) in $\Sigma\psi$. The error histories are also called error chains, as they can be visualized as sums of chains in G. Note that the map Σ evaluates the boundary of a given error chain. Decoding a syndrome $\Sigma\psi$ is essentially finding an error chain ψ' whose boundary coincides with $\Sigma\psi$. As an error chain ψ always matches the defects in $\Sigma\psi$ into pairs, the MWE decoder returns a minimum-weight perfect matching (MWPM) of the defects [19,27,41].

The existence of type-II errors complicates our problem, as they create four defects instead of two. We have to use hyperedges to represent these errors in the decoder (hyper)graph. Although we can still use a MWE decoder, the geometric meaning will be less clear. Notice that for any type-II error $(\tilde{e},t) \in M_2$ is equivalent to two data errors $(\gamma_t\tilde{e},t) + (\gamma_t\tilde{e},t+1) \in \mathbb{F}_2[D]$. As an approximation, we can pretend that the type-II errors do not exist and use the MWPM decoder

$$\mathrm{Dec}_{\mathrm{MWPM}} : \mathbb{F}_2[F \times \mathbb{N}] \to \mathbb{F}_2[D \cup M_1]$$

on the decoder graph with an edge set $D \cup M_1$. For convenience, we define a map

$$\begin{aligned} \Pi : \mathbb{F}_2[D \cup M] &\to \mathbb{F}_2[D \cup M_1], \\ D \cup M_1 \ni \psi &\mapsto \psi, \\ M_2 \ni (\tilde{e},t) &\mapsto (\gamma_t\tilde{e},t) + (\gamma_t\tilde{e},t+1) \end{aligned} \tag{40}$$

that projects all the errors onto the decoder graph. For each error history $\psi \in \mathbb{F}_2[D \cup M]$, the MWPM decoder will regard it as an error chain $\Pi\psi$ on the decoder graph with total length $|\Pi\psi| = |\psi| + |\psi \cap M_2|$, or more explicitly

$$|\psi \cap D| + |\psi \cap M_1| + 2|\psi \cap M_2|.$$

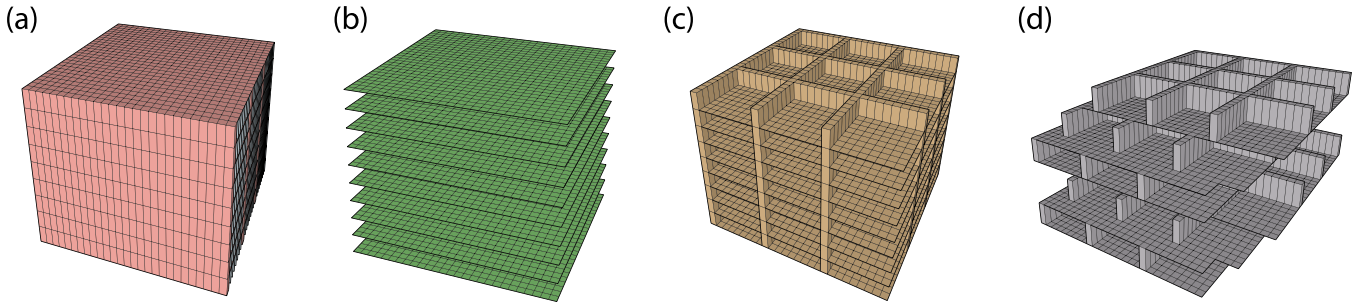


FIG. 3. Decoder graphs of the toric code. The syndrome bits are vertices, the data errors are horizontal edges, and the type-I measurement errors are vertical edges. (a) The decoder graph of the Shor error correction is a homogeneous three-dimensional lattice. (b) The decoder graph of the Steane error correction is a stack of two-dimensional lattices. Different layers are not connected. (c) Decoder graph of block extraction. The ancilla blocks when aligned lead to timelike correlations between direct repeated measurements. (d) By offsetting the ancilla blocks, the timelike correlations require spacelike errors in order to correlate defects from top to bottom. The images in (c) and (d) have been reproduced from Ref. [40].

Indeed, the MWPM decoder can only guarantee that

$$|\mathrm{Dec}_{\mathrm{MWPM}}(\Sigma\psi)| \leqslant |\Pi\psi| = |\psi| + |\psi \cap M_2|. \tag{41}$$

However, as shown below, $\mathrm{Dec}_{\mathrm{MWPM}}$ can preserve the code distance L.

Theorem 1. If $|\psi| < L/2$, $\mathrm{Dec}_{\mathrm{MWPM}}$ will correct ψ without introducing a logical error.

Proof. Let $\psi' = \mathrm{Dec}_{\mathrm{MWPM}}(\Sigma\psi)$. By applying the correction ψ', the data error will become an undetectable error $\bar{\mathcal{D}}(\psi + \psi')$. Since the toric code has distance L, it suffices to show that $|\bar{\mathcal{D}}(\psi + \psi')| < L$.

From (41) and the fact that $|\psi' \cap D| \leqslant |\psi'|$, we obtain $|\psi' \cap D| \leqslant |\psi| + |\psi \cap M_2|$. Moreover,

$$\begin{aligned} &|\psi' \cap D| + |\psi \cap D| \\ &\leqslant |\psi| + |\psi \cap D| + |\psi \cap M_2| \\ &\leqslant 2|\psi| < L. \end{aligned} \tag{42}$$

The theorem is proved by combining (42) and the fact that

$$|\bar{\mathcal{D}}(\psi + \psi')| \leqslant |\psi' \cap D| + |\psi \cap D|. \tag{43}$$

C. Fault-tolerant error correction in finite time

In practice, the circuit will always end in some finite time T. All the errors and defects can only occur before T. Indeed, the MWPM decoder will have a finite decoder graph of size $O(TL^2)$. As the MWPM algorithm runs in polynomial time to the input size, for the purpose of reliable quantum memory, we can delay the decoding until the end of the circuit execution. However, this will be impractical for quantum computation tasks with non-Clifford gates [42]. Instead, we need to process the syndrome bits and correct the errors as soon as possible. As the information provided by the latest syndrome bits is always unreliable, they will be processed only when further syndrome bits are gathered. For example, we can divide the time axis by some chosen time

$$1 = t_1 < t_2 < \cdots < t_i < \cdots .$$

In the ith round of error correction, we decode the syndrome bits in the time interval $[t_i, t_{i+2})$ but only correct the errors in $[t_i - \frac{1}{2}, t_{i+1} - \frac{1}{2})$; then we discard the syndrome bits in $[t_i, t_{i+1})$ while keeping the syndrome bits in $[t_{i+1}, t_{i+2})$ for the next round of correction. Note that the syndrome bits at time t_{i+1} need to be updated. This is known as the overlapping recovery method [19,23]. The details are formally described in Procedure 1, for which we define

$$\psi[t, t'] = \{(e, t'') \in \psi | t \leqslant t'' < t'\} \tag{44}$$

and

$$\Sigma[t, t']\psi = \bigsqcup_{t \leqslant t'' < t'} \Delta_{t''}\psi \tag{45}$$

for convenience.

Procedure 1 Overlapping recovery.

1: $i \leftarrow 1$.

2: Use MWPM to find an error history ψ' such that

$$\Sigma[t_i, t_{i+2}](\Pi\psi + \psi') = 0.$$

3: $\psi \leftarrow \psi + \psi'[t_i, t_{i+1}]$. // Correct the error history. In practice, this is equivalent to applying error correction on the data qubits and updating the syndrome at time t_{i+1}.

4: $i \leftarrow i + 1$, goto 2.

In the ith iteration of Procedure 1, the decoder graph for MWPM only contains the vertices (syndrome bits) and edges (errors) in $[t_i - \frac{1}{2}, t_{i+2} - \frac{1}{2})$. In particular, the type-I errors at time $t_{i+2} - 1$ will become open edges, i.e., edges connecting to the time boundary. The defects not only can be paired with each other, but also can be fused with the time boundary so that its lifetime is extended to the next round. Intuitively, if the distance from time slice t_i to t_{i+1} on the decoder graph, denoted by $d(t_i, t_{i+1})$, is too small, it would be too easy for the decoder to fuse a defect at time t_i to the time boundary. As a consequence, errors can hardly be corrected. If we use the Shor-style gadget all the time, we will have $d(t_i, t_{i+1}) = |t_{i+1} - t_i|$, and it has been shown that $|t_{i+1} - t_i| = O(L)$ suffices for fault tolerance [19]. Naturally, the result should be generalized as $d(t_i, t_{i+1}) = O(L)$ for arbitrary choices of gadgets. To show this, we prove the following.

Theorem 2. In the ith round of error correction, if the correction ψ' creates a propagating error, i.e., $\Pi\psi + \psi'$ contains a path from the time slice t_i to the time slice t_{i+1}, then $|\psi[t_i, t_{i+1}]| \geqslant \frac{1}{2}[d(t_i, t_{i+1}) - L]$.

Proof. Suppose $\Pi\psi + \psi'$ contains a path P_1 from a vertex (f_1, t_i) to some vertex (f_2, t_{i+1}), where $f_1, f_2 \in F$. Then $\Pi\psi + \psi'$ must contain another path P_2 from a vertex (f_1', t_i) to a vertex (f_2', t_{i+1}), where $f_1', f_2' \in F$ and $P_1 \cap P_2 = \emptyset$. We must have

$$|P_i| = |\Pi\psi \cap P_i| + |\psi' \cap P_i|. \tag{46}$$

Let P_3 be the shortest path from (f_1, t_i) to (f_1', t_i) and P_4 be the shortest path from (f_2, t_{i+1}) to (f_2', t_{i+1}). We must have $|P_3|, |P_4| \leqslant L$. Consider the cycle $C = P_1 + P_2 + P_3 + P_4$. By the definition of MWPM decoder, $|\psi'| \leqslant |\psi' + C|$, which is equivalent to

$$|\psi' \cap C| \leqslant |C| - |\psi' \cap C|. \tag{47}$$

Combining $|\psi' \cap C| \geqslant |\psi' \cap P_1| + |\psi' \cap P_2|$, (46), and (47), we have

$$\begin{aligned} &|\psi' \cap P_1| + |\psi' \cap P_2| \\ &\leqslant |\Pi\psi \cap P_1| + |\Pi\psi \cap P_2| + |P_3| + |P_4| \\ &\leqslant |\Pi\psi \cap P_1| + |\Pi\psi \cap P_2| + 2L. \end{aligned} \tag{48}$$

Adding $|\Pi\psi \cap P_1| + |\Pi\psi \cap P_2|$ to both side of (48) and combining (46), we obtain

$$2(|\Pi\psi \cap P_1| + |\Pi\psi \cap P_2| + L) \geqslant |P_1| + |P_2|. \tag{49}$$

Therefore,

$$\begin{aligned} |\Pi\psi[t_i, t_{i+1}]| &\geqslant |\Pi\psi \cap P_1| + |\Pi\psi \cap P_2| \\ &\geqslant \tfrac{1}{2}(|P_1| + |P_2| - 2L) \\ &\geqslant d(t_i, t_{i+1}) - L. \end{aligned} \tag{50}$$

Finally, as $|\Pi\psi[t_i, t_{i+1}]| \leqslant 2|\psi[t_i, t_{i+1}]|$, we have

$$|\psi[t_i, t_{i+1}]| \geqslant \tfrac{1}{2}[d(t_i, t_{i+1}) - L]. \tag{51}$$

Given the gadgets $\{\mathcal{G}_t\}_{t\in\mathbb{N}}$, we can choose the time slices $\{t_i\}_{i\in\mathbb{N}}$ such that $d(t_i, t_{i+1}) \geqslant \alpha L = O(L)$ for some constant $\alpha \gg 1$. By Theorem 2, to make a propagating error happen, the system should have at least $\lceil\frac{(\alpha-1)L}{2}\rceil$ errors. As $\lceil\frac{L}{2}\rceil \ll \lceil\frac{(\alpha-1)L}{2}\rceil$ errors can already lead to a logical error, assuming the errors are independent, the probability of a propagating error is negligible compared to that of a logical error.

Now we can move from these theorems to comparing different transversal gadgets by visualizing their decoder graph in time and space. The Shor gadget leads to a uniform three-dimensional lattice decoder graph where the vertices are syndrome bits as shown in Fig. 3(a). As proved above, we need the time dimension to be comparable to the space dimension. Steane-style syndrome extraction is capable of single-shot error correction [14]. In our method, this is clear from the lack of timelike edges in the decoder graph in Fig. 3(b). What intermediate schemes with block decoding generate are decoder graphs that only have timelike edges on the edges of the block.

We create the blocks by partitioning the $L \times L$ toric code into $m \times m$ blocks, where m divides L. We then use a small $m \times m$ surface-code ancilla block with $2m(m + 1)$ ancilla qubits. At the edges of the blocks, there are timelike boundaries as shown in Fig. 3(c).

In this picture, it is clear that we can reduce timelike edges by shifting the pattern of blocks. To simplify the shifting, we choose $m = 3k$ for an integer k. At every time step, we shift the position of the blocks up and to the right by k as shown in Fig. 3(d). By construction, the error gadgets repeat every three steps, $\mathcal{G}_t = \mathcal{G}_{t+3}$, and we can verify that $d(t, t + 3) = \Omega(m)$. This enables us to achieve fault tolerance with $|t_{i+1} - t_i| = O(L/m)$.

D. Numerical results

We test our methods using numerical simulations as described in Ref. [40]. In Appendix B we include the raw data of the simulation used to determine the thresholds in [40]. We reproduce the details of the numerical method below for convenience. We also expand the simulations to consider the case with no ancilla-block error.

We study the circuit-level performance of our fault-tolerant error-correction schemes by Monte Carlo simulations. The X- and Z-syndrome extractions are applied alternatively. Our error model is parametrized by a single error parameter p and consists of three parts.

(i) Gate errors. With probability p, each two-qubit CNOT gate is followed by a Pauli error drawn uniformly at random from the set $\{I, X, Y, Z\}^{\otimes 2} \setminus \{I \otimes I\}$.

(ii) Measurement errors. With probability $2p/3$, a measurement outcome in either the Z or X basis is flipped.

TABLE I. Comparison of thresholds when $p_1 = p$. These results are also presented in Ref. [40]. For Shor and Steane ancilla, there is no offset strategy.

	Method						
	Shor		Block extraction				
Strategy	Cat	Bare	$m = 3$	$m = 6$	$m = 9$	$m = 12$	Steane
offset			0.68	0.89	1.04	1.13	
aligned	0.57	0.83	0.74	0.89	0.97	1.04	2.05

(iii) Preparation errors. Ancilla preparation can lead to correlated errors that need to be removed through verification or syndrome measurement decoding. Here we assume a simple error model where the complicated ancilla blocks are generated perfectly and then each qubit undergoes an independent depolarizing channel with probability p_1, which we set to either p or 0.

We further simplify by ignoring idling errors, which enables us to avoid complications due to scheduling. Comparison of these syndrome extraction methods for practical application would require a detailed multiparameter error model, a procedure for ancilla generation and verification, and the connectivity constraints of the quantum processor. To accelerate our simulation, instead of the standard MWPM decoding algorithm, we use a weighted variant of the union-find decoder [43,44]. For the surface code with bare ancilla, weighted union-find relative to MWPM decreases the threshold from 0.72% to 0.62% for a standard depolarizing error model with idling errors [44]. Tables I and II compare our block extraction schemes and Shor's and Steane's schemes for the two cases $p_1 = p$ and $p_1 = 0$, respectively. In our noise model, the thresholds of transversal-gadget extraction is lower (upper) bounded by that of Shor's cat-state extraction (Steane's method). For block extraction, we fix the ancilla-block size m when $L \to \infty$. In this case, we will still need $O(L)$ rounds of extractions even if we offset the blocks. However, we observe that offsetting the blocks yields different threshold values than aligning them. This makes sense as the two strategies provide different decoder graph symmetries. When m gets larger, the offset version starts to yield higher threshold values. We also calculate the threshold of the conventional bare-ancilla extraction scheme [41] for a comparison.

V. CONCLUSION

An ideal fault-tolerant syndrome extraction circuit would have minimal interaction with the data, easy to prepare ancilla blocks, and require a small number of measurement rounds

TABLE II. Comparison of thresholds when $p_1 = 0$. For Shor and Steane ancilla, there is no offset strategy.

	Method						
	Shor		Block extraction				
Strategy	Cat	Bare	$m = 3$	$m = 6$	$m = 9$	$m = 12$	Steane
offset			1.15	1.48	1.73	1.89	
aligned	0.91	0.86	1.2	1.46	1.6	1.71	3.12

to make a decision. In this work, we have shown a family of extraction circuits that produces methods between Shor's [9] and Steane's [13] schemes. These circuits allow us to trade off the complexity of ancilla-block preparation for reduced interactions with the data. Furthermore, by shifting the choice of ancilla blocks in time, we can reduce the number of measurement rounds to achieve fault tolerance while maintaining constant ancilla-block complexity. Specifically for the toric code of distance L, we can use offset blocks of size $O(m^2)$ to achieve fault tolerance in $O(L/m)$ rounds, as presented in Ref. [40].

When ancilla postselections are allowed, we found that our error-correction schemes could yield higher thresholds for certain error models assuming negligible idling errors and independent ancilla errors. For a more realistic threshold estimation, we need to choose a specific ancilla preparation protocol. For the toric code example, the ancilla blocks inherit the toric code structure and one can use the bare-ancilla extraction circuit with postselections to prepare them [41]. There also exist protocols for preparing a general CSS stabilizer state by state distillations without leaving any correlated errors [39]. The detailed simulation of these preparation protocols is beyond the scope of this work and the utility strongly depends on the physical error model.

We have presented the toric code as an example because it is well studied and enables us to separate the advantages and disadvantages of our methods from advantages and disadvantages of new codes. The toric code allows for fault-tolerant syndrome extraction with bare ancilla on a nearest-neighbor two-dimensional lattice. Our methods are not a natural choice for the toric code, because the methods require the architecture to break out of two dimensions and also creates new challenges for generating ancilla blocks. On the other hand, we know that finite-rate quantum error-correction codes are not compatible with Euclidean two-dimensional architectures. We hope that our framework will enable the development of high-threshold fault-tolerant extraction circuits for these more qubit efficient codes.

There are a number of directions for further study. Codes that typically use Shor-style extraction, such as two-dimensional color codes [30,45], can be decoded with ancilla blocks to improve the threshold. Concatenated codes that have high thresholds with postselected Knill or Steane schemes [15,46] also have high ancilla rejection rates and block methods can examine trading a reduced threshold for less ancilla verification. The non-fault-tolerant schemes developed here can be made fault tolerant using flag methods [17,18,30]. The time optimization and the choice of ancilla blocks can be analyzed using the framework recently applied to Shor-style extraction [23,24]. Finally, these methods need to be tested in the face of more realistic errors as experimental systems approach the complexity capable of generating and utilizing large ancilla blocks.

ACKNOWLEDGMENTS

We thank Michael Newman and Rui Chao for helpful discussion. This work was sponsored by the NSF EPiQC Expeditions in Computing (Grant No. 1832377), the Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity, through the Army Research Office (Contract No. W911NF-16-1-0082), and the Army Research Office (Contract No. W911NF-21-1-0005).

APPENDIX A: FAULT-TOLERANT ERROR CORRECTION VIA SCHEME B

In this Appendix we provide a fault-tolerant error-correction protocol based on scheme B in Fig. 1. Our protocol is an adaptive flagged circuit [16,47]. The extraction circuit for Z stabilizers is described in Fig. 4. The extraction circuit of X stabilizers is identical, up to a Hadamard transformation. A DiVincenzo-Aliferis decoding circuit is applied before ancilla measurement. As a result, in addition to the syndrome bits, we also obtain two flags $f_{1,Z}$ ($f_{1,X}$) and $f_{2,Z}$ ($f_{2,X}$) in the Z (X) stabilizer measurement circuit. The flag $f_{i,Z}$ ($f_{i,X}$) can

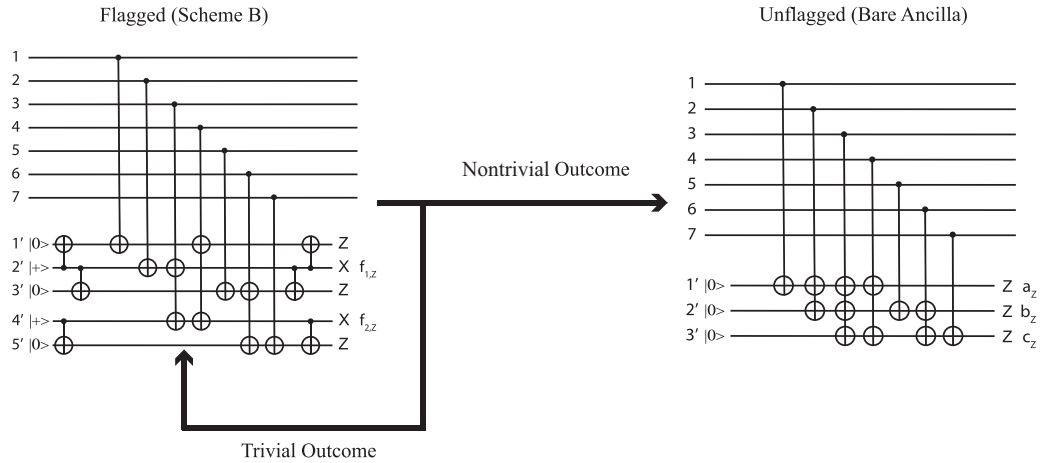


FIG. 4. Adaptive fault-tolerant error-correction circuit, using scheme B in Fig. 1. Here, only the extraction of the Z stabilizer is demonstrated. The circuit for X-stabilizer extraction is identical up to a Hadamard transformation. We keep running the flagged circuit (scheme B) until a nontrivial syndrome is observed or some flags are raised. We then measure the syndrome again via an unflagged (bare-ancilla) circuit and decode the errors using the new syndrome and the flags in the previous flagged round.

TABLE III. Lookup table decoder when no flags are on. If the flag $f_{1,X}$ is raised, we change the correction for $a_Z b_Z c_Z = 001$ to X_2X_3. If the flag $f_{2,X}$ is raised, we change the correction for $a_Z b_Z c_Z = 010$ to X_3X_4.

$a_Z b_Z c_Z$	Correction	$a_Z b_Z c_Z$	Correction
000	none	110	X_2
100	X_1	101	X_4
010	X_5	011	X_6
001	X_7	111	X_3

be used to detect Z (X) errors on the ancilla block, which could propagate to the data block as weight-2 Z (X) errors. We repeatedly execute the scheme B circuit until we observe a nontrivial outcome. As we have detected at least one error and we cannot correct any two errors for a distance-3 code, we can use the bare-ancilla circuit to extract the syndrome bits a_Z, b_Z, and c_Z (a_X, b_X, and c_X), as shown in Fig. 4. We use Table III to decode the X errors based on the syndromes measured in the unflagged circuit when no flags were raised. However, if $f_{1,X}$ ($f_{1,Z}$) is flagged and the syndrome $a_Z b_Z c_Z$ ($a_X b_X c_X$) is 001, the most probable error should be X_2X_3 (Z_2Z_3) instead of X_7 (Z_7). Similarly, if $f_{2,X}$ ($f_{2,Z}$) is flagged, we need to change the correction for the syndrome 010 to X_3X_4 (Z_3Z_4). The flags tell us to expect correlations and decode appropriately.

APPENDIX B: SIMULATION DATA

In this Appendix we present the simulated data for a series of syndrome extraction methods in Figs. 5–10. The decoder used is the weighted union-find decoder [44] and for the bare ancilla, cat decoders, and $m \times m$ block decoders we repeat the extraction for L rounds. For Steane's scheme, we compare the extraction method for $T = 2, 5,$ and 8 rounds regardless of L.

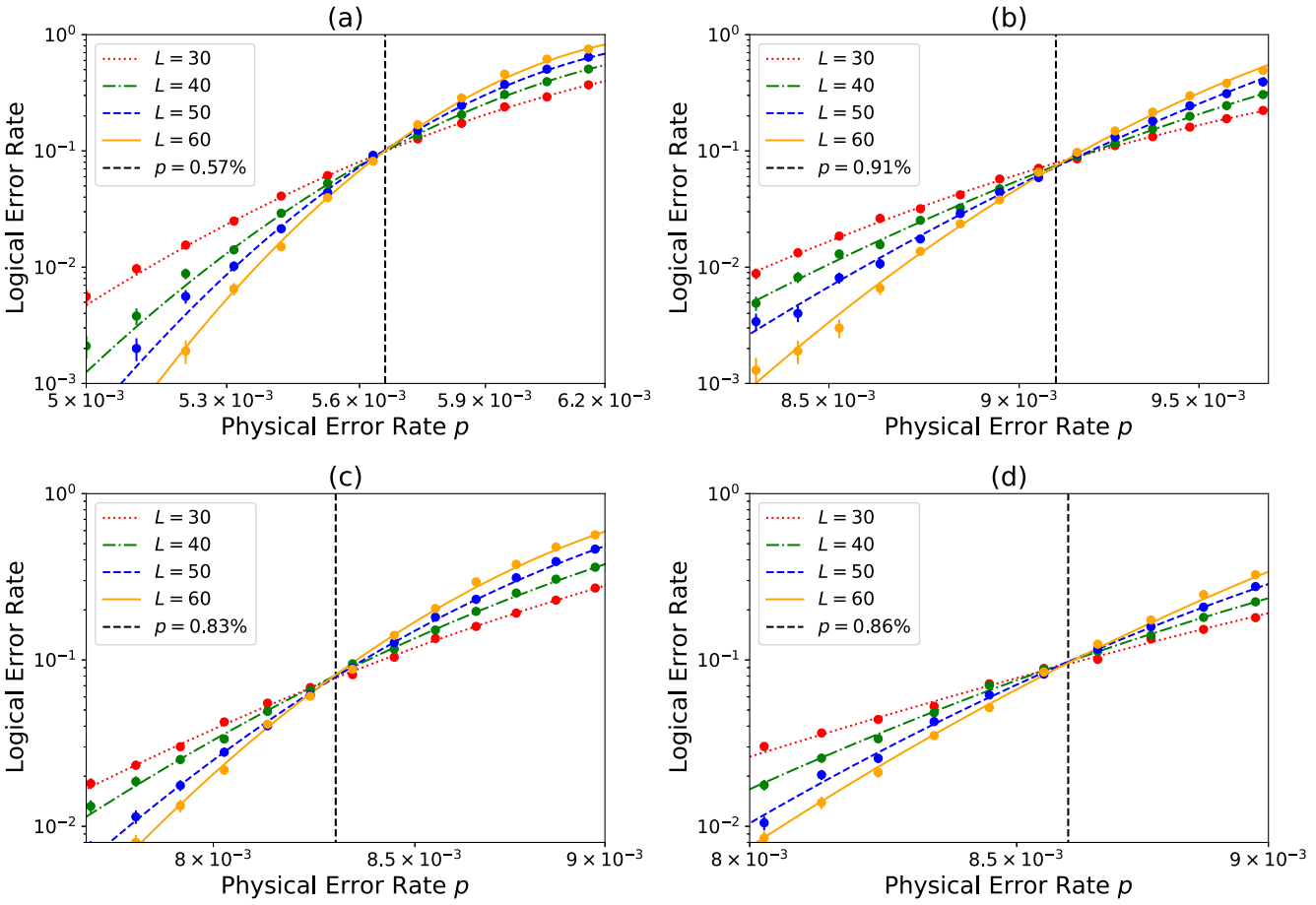


FIG. 5. Threshold behavior of Shor error correction with (a) $p_1 = p$, a cat state; (b) $p_1 = 0$, a cat state; (c) $p_1 = p$, a bare ancilla; and (d) $p_1 = 0$, a bare ancilla, where p_1 is the preparation error rate. The syndrome extraction is repeated for L noisy rounds and one ideal round. Each data point is obtained from 10^4 trials.

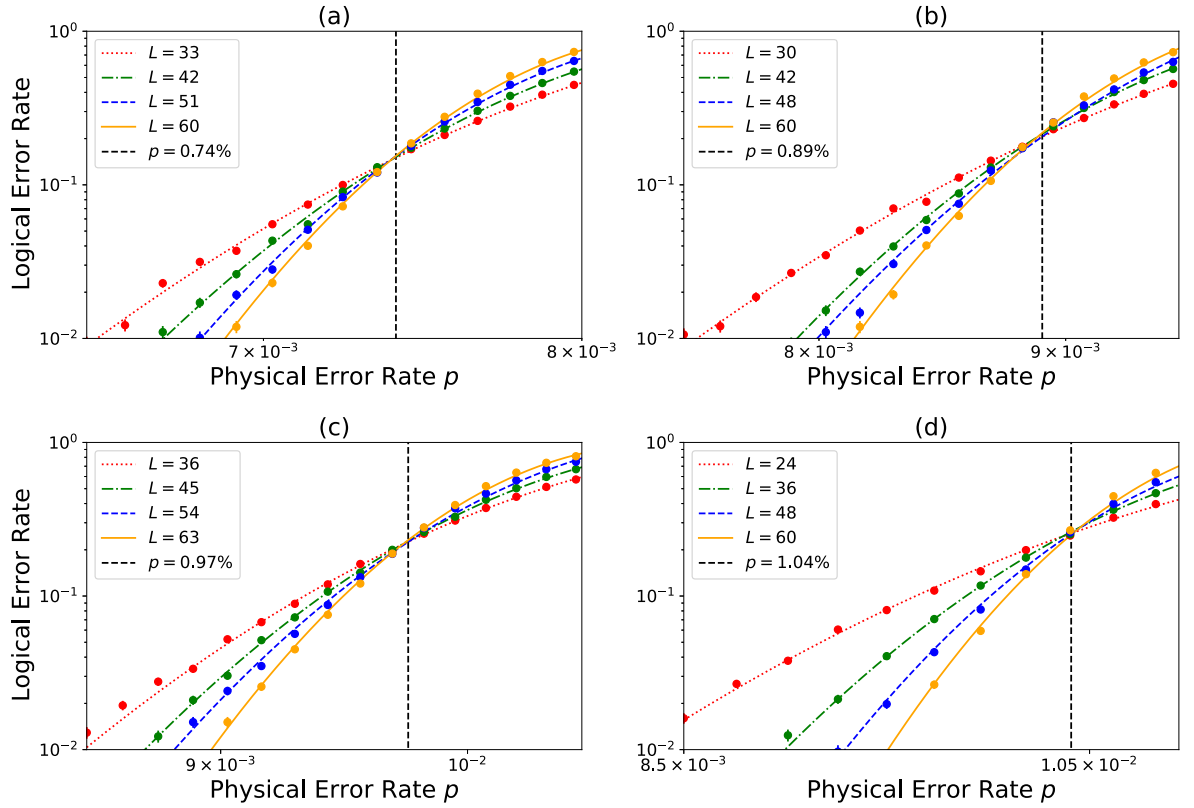


FIG. 6. Threshold behavior of aligned block syndrome extraction with $p_1 = p$ and block sizes (a) $m = 3$, (b) $m = 6$, (c) $m = 9$, and (d) $m = 12$. We repeat the syndrome extraction for L rounds. Each data point is obtained from 10^4 trials.

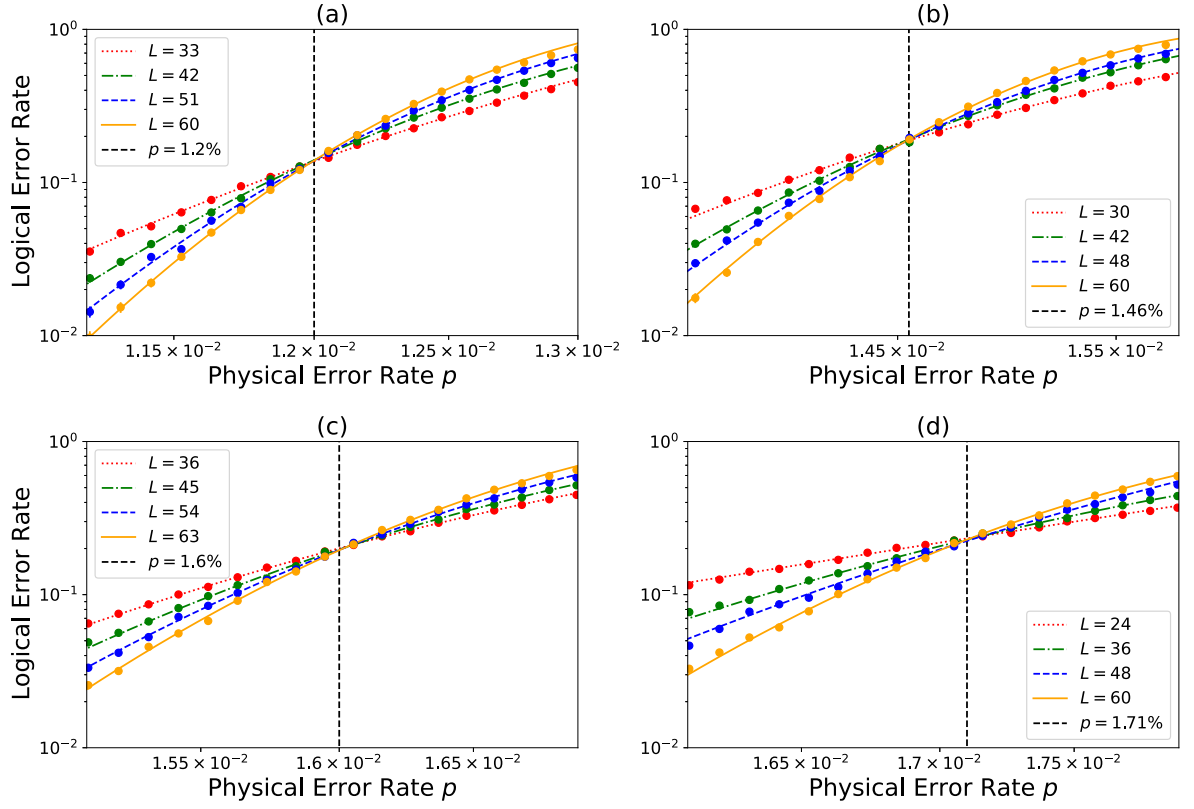


FIG. 7. Threshold behavior of aligned block syndrome extraction with $p_1 = 0$ and block sizes (a) $m = 3$, (b) $m = 6$, (c) $m = 9$, and (d) $m = 12$. We repeat the syndrome extraction for L rounds. Each data point is obtained from 10^4 trials.

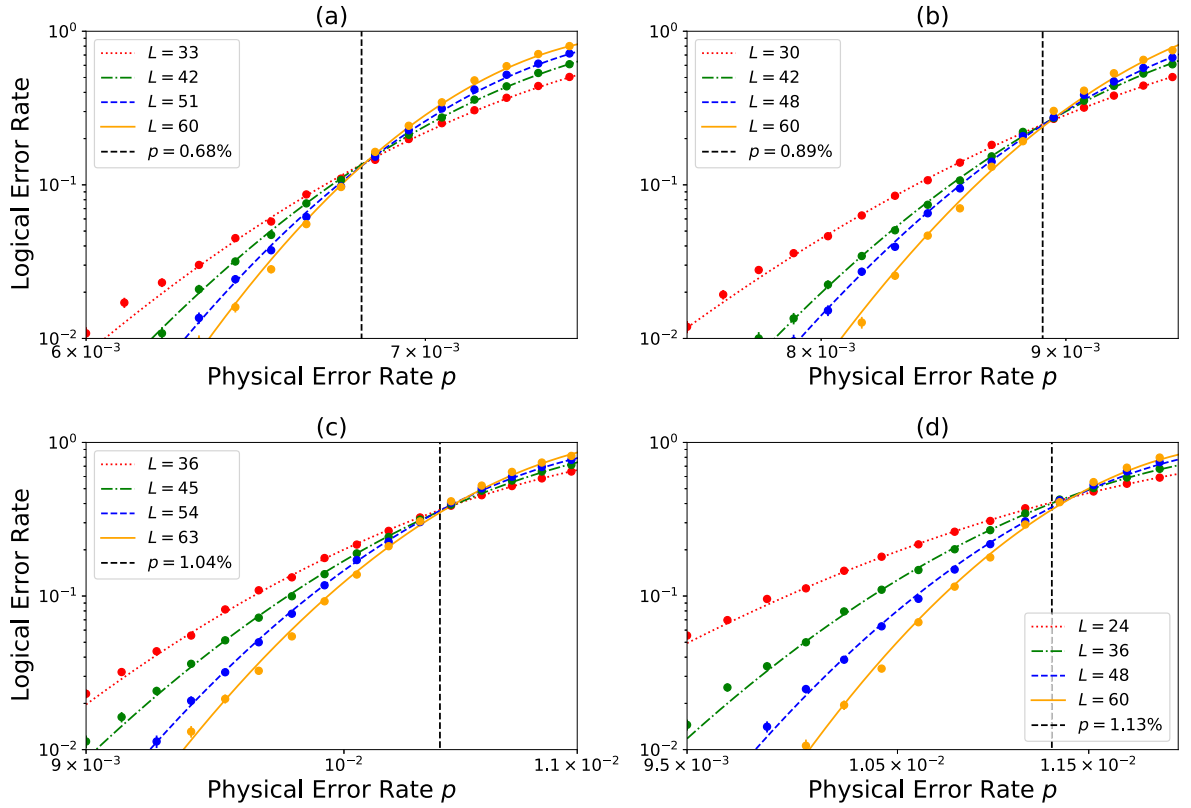


FIG. 8. Threshold behavior of offset block syndrome extraction with $p_1 = p$ and block sizes (a) $m = 3$, (b) $m = 6$, (c) $m = 9$, and (d) $m = 12$. We repeat the syndrome extraction for L rounds. Each data point is obtained from 10^4 trials.

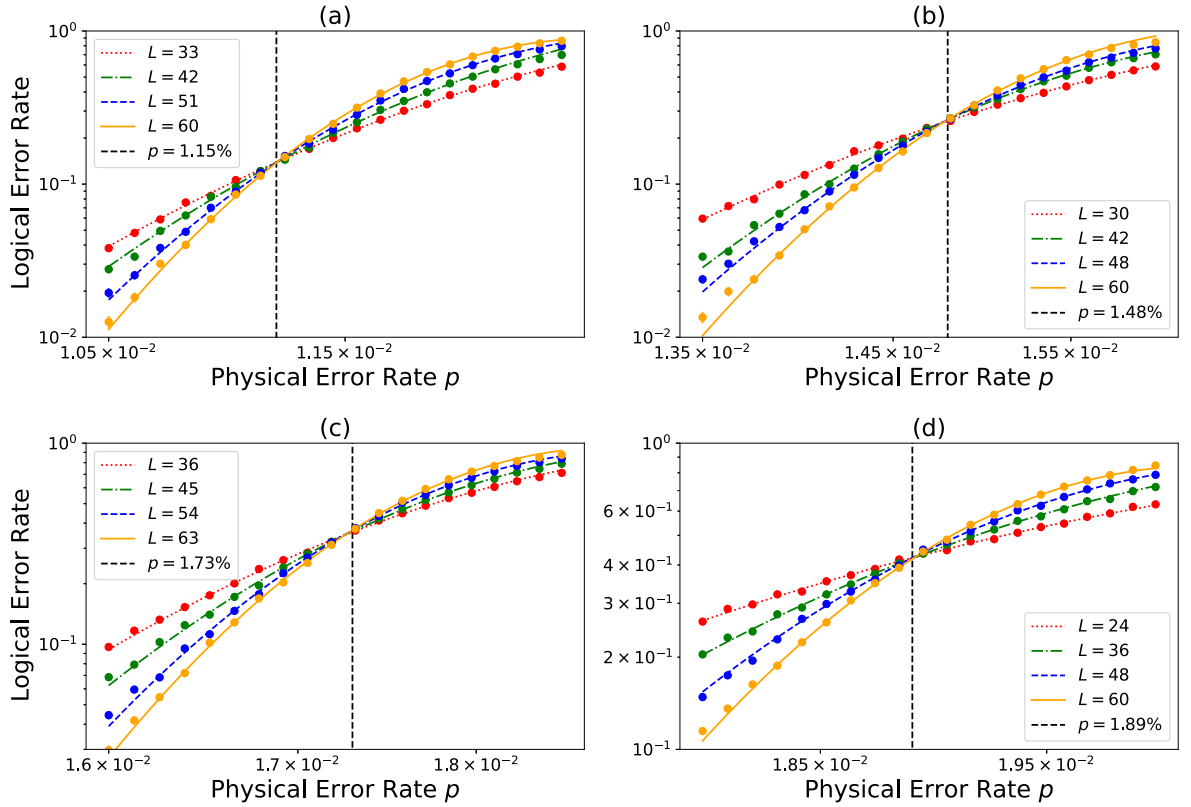


FIG. 9. Threshold behavior of offset block syndrome extraction with $p_1 = 0$ and block sizes (a) $m = 3$, (b) $m = 6$, (c) $m = 9$, and (d) $m = 12$. We repeat the syndrome extraction for L rounds. Each data point is obtained from 10^4 trials.

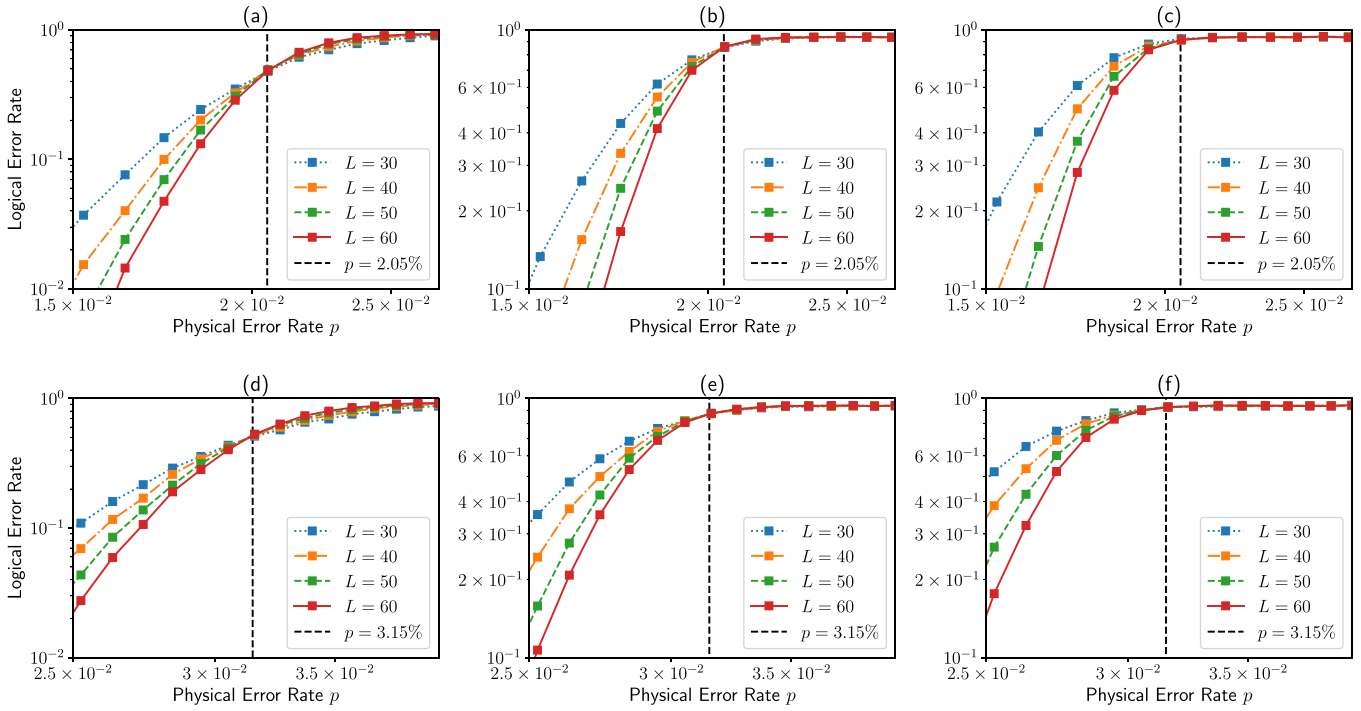


FIG. 10. Threshold behavior of Steane error correction for (a)–(c) $p_1 = p$ and (d)–(f) $p_1 = 0$. As Steane EC is single shot, the number of syndrome measurement rounds T is set to be finite. Each data point is obtained from 10^4 trials. We estimate the threshold for (a) and (d) $T = 2$, (b) and (e) $T = 5$, and (c) and (f) $T = 8$ and do not find a decrease of the threshold values.

[1] P. W. Shor, Phys. Rev. A **52**, R2493 (1995).
[2] D. Aharonov and M. Ben-Or, in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing, El Paso, 1997* (ACM, New York, 1997), p. 176.
[3] E. Knill and R. Laflamme, Phys. Rev. A **55**, 900 (1997).
[4] A. Y. Kitaev, Ann. Phys. (NY) **303**, 2 (2003).
[5] A. R. Calderbank and P. W. Shor, Phys. Rev. A **54**, 1098 (1996).
[6] A. M. Steane, Phys. Rev. Lett. **77**, 793 (1996).
[7] A. R. Calderbank, E. M. Rains, P. W. Shor, and N. J. A. Sloane, Phys. Rev. Lett. **78**, 405 (1997).
[8] A. R. Calderbank, E. M. Rains, P. M. Shor, and N. J. A. Sloane, IEEE Trans. Inf. Theory **44**, 1369 (1998).
[9] P. W. Shor, in *Proceedings of the 37th Symposium on Foundations of Computer Science* (IEEE, Piscataway, 1996), p. 56.
[10] D. P. DiVincenzo and P. W. Shor, Phys. Rev. Lett. **77**, 3260 (1996).
[11] D. E. Gottesman, Ph.D. thesis, California Institute of Technology, 1997.
[12] P. Aliferis, D. Gottesman, and J. Preskill, Quantum Inf. Comput. **6**, 97 (2005).
[13] A. M. Steane, Phys. Rev. Lett. **78**, 2252 (1997).
[14] A. M. Steane, Phys. Rev. A **68**, 042322 (2003).
[15] E. Knill, Nature (London) **434**, 39 (2005).
[16] R. Chao and B. W. Reichardt, Phys. Rev. Lett. **121**, 050502 (2018).
[17] R. Chao and B. W. Reichardt, npj Quantum Inf. **4**, 42 (2018).
[18] R. Chao and B. W. Reichardt, PRX Quantum **1**, 010302 (2020).
[19] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, J. Math. Phys. **43**, 4452 (2002).
[20] H. Bombín, Phys. Rev. X **5**, 031043 (2015).
[21] O. Fawzi, A. Grospellier, and A. Leverrier, in *Proceedings of the 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS), Paris, 2018* (IEEE, Piscataway, 2018), p. 743.
[22] E. T. Campbell, Quantum Sci. Technol. **4**, 025006 (2019).
[23] N. Delfosse, B. W. Reichardt, and K. M. Svore, arXiv:2002.05180.
[24] N. Delfosse and B. W. Reichardt, arXiv:2008.05051.
[25] D. P. DiVincenzo and P. Aliferis, Phys. Rev. Lett. **98**, 020501 (2007).
[26] T. J. Yoder and I. H. Kim, Quantum **1**, 2 (2017).
[27] A. G. Fowler, A. M. Stephens, and P. Groszkowski, Phys. Rev. A **80**, 052312 (2009).
[28] M. Li, M. Gutiérrez, S. E. David, A. Hernandez, and K. R. Brown, Phys. Rev. A **96**, 032341 (2017).
[29] M. Li, D. Miller, and K. R. Brown, Phys. Rev. A **98**, 050301(R) (2018).
[30] C. Chamberland and M. E. Beverland, Quantum **2**, 53 (2018).
[31] C. Chamberland, G. Zhu, T. J. Yoder, J. B. Hertzberg, and A. W. Cross, Phys. Rev. X **10**, 011022 (2020).
[32] S. Huang and K. R. Brown, Phys. Rev. A **101**, 042312 (2020).
[33] P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, Appl. Phys. Rev. **6**, 021318 (2019).
[34] C. D. Bruzewicz, J. Chiaverini, R. McConnell, and J. M. Sage, Appl. Phys. Rev. **6**, 021314 (2019).

[35] Y.-C. Zheng, C.-Y. Lai, T. A. Brun, and L.-C. Kwek, Quantum Sci. Technol. **5**, 045007 (2020).
[36] A. Paetznick and B. W. Reichardt, Quantum Inf. Comput. **12**, 1034 (2012).
[37] H. Goto, Sci. Rep. **6**, 19578 (2016).
[38] C.-Y. Lai, Y.-C. Zheng, and T. A. Brun, Phys. Rev. A **95**, 032339 (2017).
[39] Y.-C. Zheng, C.-Y. Lai, and T. A. Brun, Phys. Rev. A **97**, 032331 (2018).
[40] S. Huang and K. R. Brown, Phys. Rev. Lett. **127**, 090505 (2021).
[41] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Phys. Rev. A **86**, 032324 (2012).
[42] B. M. Terhal, Rev. Mod. Phys. **87**, 307 (2015).
[43] N. Delfosse and N. H. Nickerson, arXiv:1709.06218.
[44] S. Huang, M. Newman, and K. R. Brown, Phys. Rev. A **102**, 012419 (2020).
[45] H. Bombin and M. A. Martin-Delgado, Phys. Rev. Lett. **97**, 180501 (2006).
[46] B. W. Reichardt, arXiv:quant-ph/0406025.
[47] B. W. Reichardt, Quantum Sci. Technol. **6**, 015007 (2020).