PAPER

Constant depth fault-tolerant Clifford circuits for multi-qubit large block codes

Quantum Science and Technology

PAPER

RECEIVED
4 April 2020

REVISED
28 June 2020

ACCEPTED FOR PUBLICATION
6 July 2020

PUBLISHED
28 July 2020

Constant depth fault-tolerant Clifford circuits for multi-qubit large block codes

Yi-Cong Zheng[1,2,3,9], **Ching-Yi Lai**[4], **Todd A Brun**[5] **and Leong-Chuan Kwek**[2,6,7,8]

[1] Tencent Quantum Lab, Tencent, Shenzhen, Guangdong, 518057, People's Republic of China
[2] Centre for Quantum Technologies, National University of Singapore, 117543, Singapore
[3] Yale-NUS College, 138527, Singapore
[4] Institute of Communications Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan
[5] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, California 90089, United States of America
[6] MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, UMI 3654, Singapore
[7] Institute of Advanced Studies, Nanyang Technological University, 639673, Singapore
[8] National Institute of Education, Nanyang Technological University, 637616, Singapore
[9] Author to whom any correspondence should be addressed.

E-mail: yicongzheng@tencent.com

Keywords: fault-tolerant quantum computation, large block codes, quantum error correction, Clifford circuit

Abstract

Fault-tolerant quantum computation (FTQC) schemes using large block codes that encode $k > 1$ qubits in n physical qubits can potentially reduce the resource overhead to a great extent because of their high encoding rate. However, the fault-tolerant (FT) logical operations for the encoded qubits are difficult to find and implement, which usually takes not only a very large resource overhead but also long *in situ* computation time. In this paper, we focus on Calderbank–Shor–Steane $[[n, k, d]]$ (CSS) codes and their logical FT Clifford circuits. We show that the depth of an arbitrary logical Clifford circuit can be implemented fault-tolerantly in $O(1)$ steps *in situ* via either Knill or Steane syndrome measurement circuit, with the qualified ancilla states efficiently prepared. Particularly, for those codes satisfying $k/n \sim \Theta(1)$, the resource scaling for Clifford circuits implementation on the logical level can be the same as on the physical level up to a constant, which is independent of code distance d. With a suitable pipeline to produce ancilla states, our scheme requires only a modest resource cost in physical qubits, physical gates, and computation time for very large scale FTQC.

1. Introduction

Quantum error-correcting codes (QECCs) [1–5] and the theory of fault-tolerant quantum computation (FTQC) [5–12] have shown that large-scale quantum computation is possible if the noise is not strongly correlated between qubits and its rate is below certain threshold [7, 11–16].

Large QECCs with high encoding rates typically encode many logical qubits with high distance. FTQC architectures based on these codes may potentially outperform smaller codes and topological codes, like surface codes [9, 17] and color codes [18], in terms of the overall resource required and the error correction ability [19–23]. However, for an $[[n, k, d]]$ code with $k, d \gg 1$, it may be extremely difficult (or even impossible) to find all required fault-tolerant (FT) logical gates. For Calderbank–Shor–Steane (CSS) codes [2, 3], one way to resolve this challenge is to implement logical circuits indirectly through Knill or Steane syndrome extraction circuits [11, 24] with additional blocks of encoded ancilla qubits prepared in specific states [21, 24–26]. Unfortunately, the distillation processes for each encoded ancilla state are complicated, and different ancilla states are usually required for each logical gate. As an example, a Clifford circuit on k qubits requires $O(k^2/\log k)$ Clifford gates [27, 28] with circuit depth $O(k)$; if an $[[n, k, d]]$ CSS code is used, it requires $O(k^2/\log k)$ logical Clifford gates [28], and in general, $O(k^2/\log k)$ different ancilla states need to be prepared, and the same number of Knill/Steane syndrome extraction steps are required.

A natural question arises: can one implement logical circuits on those multi-qubit large block codes ($k \gg 1$) in a quicker and more efficient way? In this paper, we show that for Clifford circuits, the answer is positive for CSS codes: one can implement an arbitrary logical Clifford circuit fault-tolerantly using $O(1)$ qualified encoded ancilla states and a constant number of Knill/Steane syndrome measurement steps. Thus the depth of a logical Clifford circuit can be reduced to $O(1)$ *in situ*. Furthermore, we show that with the distillation protocol proposed in [29, 30], these ancilla states can be distilled *off-line* in ancilla factories with yield rate close to $O(1)$ asymptotically, if the physical error rate is sufficiently low. Especially, for those families of large block codes with $k/n \sim \Theta(1)$, the number of physical qubits and physical gates required for an arbitrary logical Clifford circuit can scale as $O(k)$ and $O(k^2/\log k)$ respectively on average. These results suggest that the resource cost of Clifford circuits on the logical level can scale the same as on the physical level, if the distillation circuits and large block quantum codes are carefully chosen. With a proper pipeline structure of ancilla factories to work in parallel, we are also convinced that the scaling of the required resources including the overall number of qubits, physical gates and the computation time, can be very modest for large scale FTQC.

The structure of the paper is as follows. We review preliminaries and set up notation in section 2. In section 3, we propose our scheme to implement FT logical Clifford circuits via a constant number of Knill or Steane syndrome measurement. The resource overhead for the scheme is carefully analyzed. In section 4, we compare our scheme to some other closely-related FTQC schemes according to the resource overhead and real-time computational circuit depth.

2. Preliminaries and notation

2.1. Stabilizer formalism and CSS codes

Let $\mathcal{P}_n = \mathcal{P}_1^{\otimes n}$ denote the n-fold Pauli group, where

$$\mathcal{P}_1 = \{\pm I, \pm iI, \pm X, \pm iX, \pm Y, \pm iY, \pm Z, \pm iZ\},$$

and $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, and $Y = iXZ$ are the Pauli matrices.

Let X_j, Y_j, and Z_j act as single-qubit Pauli matrices on the jth qubit and trivially elsewhere. We also introduce the notation $X^{\mathbf{a}}$, for $\mathbf{a} = a_1 \cdots a_n \in \mathbb{Z}_2^n$, to denote the operator $\otimes_{j=1}^n X^{a_j}$ and let $\mathrm{supp}(\mathbf{a}) = \{j : a_j = 1\}$. For $\mathbf{a}, \mathbf{b} \in \mathbb{Z}_2^n$, define $\mathcal{I}_{\mathbf{ab}} = \mathrm{supp}(\mathbf{a}) \bigcap \mathrm{supp}(\mathbf{b})$ and let $\tau_{\mathbf{ab}} = |\mathcal{I}_{\mathbf{ab}}|$ be the size of $\mathcal{I}_{\mathbf{ab}}$. An n-fold Pauli operator can be expressed as

$$i^l \cdot \otimes_{j=1}^n X^{a_j} Z^{b_j} = i^l X^{\mathbf{a}} Z^{\mathbf{b}}, \quad \mathbf{a}, \mathbf{b} \in \mathbb{Z}_2^n, \ l \in \{0, 1, 2, 3\}. \tag{1}$$

Then $(\mathbf{a}|\mathbf{b})$ is called the *binary representation* of the Pauli operator $i^l X^{\mathbf{a}} Z^{\mathbf{b}}$ up to an overall phase i^l. In particular, $\pm i^{\tau_{\mathbf{ab}}} X^{\mathbf{a}} Z^{\mathbf{b}}$, which is Hermitian, has eigenvalues ± 1. From now on we use the binary representation and neglect the overall phase for simplicity when there is no ambiguity. We define the weight of E, $\mathrm{wt}(E)$, as the number of terms in the tensor product which are not equal to the identity.

Suppose $\mathcal{S}$ is an abelian subgroup of $\mathcal{P}_n$ with a set of $n-k$ independent and commuting generators $\{S_1 = i^{\tau_{\mathbf{a}_1\mathbf{b}_1}} X^{\mathbf{a}_1} Z^{\mathbf{b}_1}, \ldots, S_{n-k} = i^{\tau_{\mathbf{a}_{n-k}\mathbf{b}_{n-k}}} X^{\mathbf{a}_{n-k}} Z^{\mathbf{b}_{n-k}}\}$, and $\mathcal{S}$ does not include $-I^{\otimes n}$. An $[[n, k]]$ quantum stabilizer code $C(\mathcal{S})$ is defined as the 2^k-dimensional subspace of the n-qubit state space ($\mathbb{C}^{2^n}$) fixed by $\mathcal{S}$, which is the joint $+1$ eigenspace of $S_1, \ldots, S_{n-k}$. Then for a codeword $|\psi\rangle \in C(\mathcal{S})$,

$$S|\psi\rangle = |\psi\rangle$$

for all $S \in \mathcal{S}$. We also define $N(\mathcal{S})$ to be the normalizer of the stabilizer group. Thus any non-trivial logical Pauli operator on codewords belongs to $N(\mathcal{S})\backslash\mathcal{S}$ and let $X_{j,L}$, $Y_{j,L}$ and $Z_{j,L}$ be logical Pauli operators acting on the jth logical qubit. The distance d of the code is defined as

$$d = \min_{L \in N(\mathcal{S})\backslash\mathcal{S}} \mathrm{wt}(L).$$

Suppose $\mathcal{S}' \in \mathcal{P}_n$ is another abelian subgroup containing $\mathcal{S}$ with $k = 0$, then $C(\mathcal{S}')$ has only one state $|\psi\rangle$ up to a global phase. This state is called a *stabilizer codeword* of $\mathcal{S}$, whose binary representation is

$$\psi = \left(\begin{array}{c|c} \mathbf{a}_1 & \mathbf{b}_1 \\ \vdots & \vdots \\ \mathbf{a}_n & \mathbf{b}_n \end{array}\right).$$

If a Pauli error E corrupts $|\psi\rangle$, some eigenvalues of $S_1, \ldots, S_{n-k}$ may be flipped, if they are measured on $E|\psi\rangle$. Consequently, we gain information about the error by measuring the stabilizer generators $S_1, \ldots, S_{n-k}$, and the corresponding measurement outcomes (in bits) are called the *error syndrome* of E. A quantum decoder has to choose a good recovery operation based on the measured error syndromes.

CSS codes are an important class of stabilizer codes for FTQC. Their generators are tensor products of the identity and either X or Z operators (but not both) [2, 3]. More formally, consider two classical codes, $\mathcal{C}_Z$ and $\mathcal{C}_X$ with parameters $[n, k_Z, d_Z]$ and $[n, k_X, d_X]$, respectively, such that $\mathcal{C}_X^{\perp} \subset \mathcal{C}_Z$. The corresponding parity-check matrices are H_Z $((n-k_Z)\times n)$ and H_X $((n-k_X)\times n)$ with full rank. One can form an $[[n, k = k_X + k_Z - n, d]]$ CSS code, where $d \geqslant \min\{d_Z, d_X\}$. In general, a logical basis state can be represented as:

$$|u\rangle_L = \sum_{x\in\mathcal{C}_X^{\perp}} |x + uD\rangle,$$

where $u \in \mathbb{Z}_2^k$ and D is a $k \times n$ binary matrix, whose rows are the coset leaders of $\mathcal{C}_Z/\mathcal{C}_X^{\perp}$. The stabilizer generators of a CSS code in binary representation are:

$$\left(\begin{array}{c|c} \mathrm{H}_Z & \mathbf{0} \\ \mathbf{0} & \mathrm{H}_X \end{array}\right),$$

where $\mathrm{H}_X(\mathrm{H}_Z)$ is made of $Z(X)$ type Pauli operators. For the special case that $\mathcal{C}_X = \mathcal{C}_Z$, we call such a code self-dual CSS code.

2.2. Clifford circuits

Clifford circuits are composed solely of Hadamard (H), Phase (P), and controlled-NOT (CNOT) gates, defined as

$$\mathrm{H} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \mathrm{P} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}, \qquad \mathrm{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The n-qubit Clifford circuits form a finite group, which, up to overall phases, is isomorphic to the binary symplectic matrix group defined in [28]:

Definition 1 (Symplectic group). The group of $2n \times 2n$ symplectic matrices over $\mathbb{Z}_2$ is defined in:

$$\mathrm{Sp}(2n, \mathbb{Z}_2) \equiv \{M \in \mathrm{GL}(2n, \mathbb{Z}_2) : MJ_nM^t = J_n\}$$

under matrix multiplication. Here $J_n = \left(\begin{array}{c|c} \mathbf{0} & I_n \\ I_n & \mathbf{0} \end{array}\right)$.

In general, $M \in \mathrm{Sp}(2n, \mathbb{Z}_2)$ has the form

$$M = \left(\begin{array}{c|c} Q & R \\ \hline S & T \end{array}\right),$$

where Q, R, S and T are $n \times n$ square matrices satisfying the following conditions:

$$QR^t = RQ^t, \qquad ST^t = TS^t, \qquad Q^tT + R^tS = I_n.$$

In other words, the rows of $(Q|R)$ are symplectic partners of the rows of $(S|T)$. Thus, an n-qubit Clifford circuit can be represented by a $2n \times 2n$ binary matrix with respect to the basis of the binary representation of Pauli operators in (1). Then $UX^{\mathbf{a}}Z^{\mathbf{b}}U^{\dagger}$ is represented by $(\mathbf{a}, \mathbf{b})M_U$, where M_U is the binary symplectic matrix corresponding to U. For example, the idle circuit (no quantum gates) is represented by I_{2n}, the $2n \times 2n$ identity matrix. The representation of consecutive Clifford circuits $M_1, \ldots, M_j$ is their binary matrix product

$$M = M_1 \cdots M_j.$$

We emphasize here that the symplectic matrix M acts on the binary representation of a Pauli operator from the right. The binary representations of Pauli operators and Clifford unitaries omit the overall phases of full operators. If needed, such overall phases can always be compensated by a single layer of gates consisting solely of Z and X gates [31] on some subsets of qubits [28, 32].

Let $\mathrm{C}(j, l)$ denote a CNOT gate with control qubit j and target qubit l. The actions of appending a Hadamard, Phase, or CNOT gate to a Clifford circuit M can be described as follows:

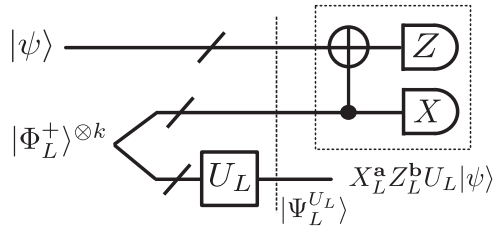

Figure 1. The quantum circuit for Knill syndrome measurement and gate teleportation for an $[[n, k, d]]$ stabilizer code. For a logical Clifford circuit U_L, if $|\Psi_L^{U_L}\rangle$ is prepared before logical Bell measurement, $U_L|\psi\rangle$ can be obtained up to some Pauli correction of $X_L^{\mathrm{a}} Z_L^{\mathrm{b}}$ on the output block, depending on the logical Bell measurement results.

(a) A Hadamard gate on qubit j exchanges columns j and $n + j$ of M.

(b) A Phase gate on qubit j adds column j to column $n + j$ (mod 2) of M.

(c) C(j, l) adds column j to column l (mod 2) of M and adds column $n + l$ to column $n + j$ (mod 2) of M.

3. Constant depth FT Clifford circuit

3.1. FT syndrome measurement

The goal of an error correction protocol in FTQC is to find the most likely errors during computation, based on the extracted syndromes. However, the circuits to perform a syndrome measurement may introduce additional errors to the system or get wrong syndromes with high probability. Therefore, the error correction may fail, if not treated properly.

In this section, we briefly review two major protocols used in this paper—Knill and Steane syndrome measurements [11, 24]. Each scheme has its own advantages in different computation scenarios [33], such as a better threshold or a better ability to handle particular types of noise, and both can be used to construct arbitrary FT logical Clifford circuits.

3.1.1. Knill syndrome measurement

For an arbitrary $[[n, k, d]]$ stabilizer code, one can use the logical teleportation circuit in figure 1 to extract the error syndrome [24], as proposed by Knill [11]. Here, two blocks of ancilla qubits are maximally entangled in a logical Bell state $|\Phi_L^+\rangle^{\otimes k} = \frac{1}{\sqrt{2}}\left(|0_L\rangle \otimes |0_L\rangle + |1_L\rangle \otimes |1_L\rangle\right)^{\otimes k}$. The upper block of ancilla qubits are encoded to the same code protecting the data state, while the lower ones can be protected by an arbitrary stabilizer code encoding k logical qubits. In this paper, we restrict ourselves to the same $[[n, k, d]]$ CSS code for all blocks.

The logical Bell measurement in the dashed box teleports the encoded state to the lower ancilla block up to a logical Pauli correction (depending on the Bell measurement outcomes), and simultaneously obtains the error syndrome of on the input data blocks. Both logical Bell measurement outcomes and syndromes are calculated from the bitwise measurement results. The circuit is intrinsically FT because it consists solely of transversal CNOT gates and bitwise measurements.

One particular virtue of the teleportation syndrome measurement circuit is that it can also provide a straightforward way to produce any logical circuit U_L (on the teleported state) of the Clifford hierarchy C_k (up to a C_{k-1} correction depending on the logical measurement outcomes) via the very same syndrome measurement circuit [25], if one can construct the ancilla state

$$\left|\Psi_L^{U_L}\right\rangle = (I \otimes U_L)\left|\Phi_L^+\right\rangle^{\otimes k}. \tag{2}$$

This construction is very useful when implementing logical circuits for large block codes. In this paper, we focus on $U \in \mathrm{C}_2$, the Clifford circuit. In this case, all the $|\Psi_L^{U_L}\rangle$ are stabilizer states that can be prepared by Clifford circuits.

3.1.2. Steane syndrome measurement

Now we consider a CSS code $[[n, k, d]]$ for convenience in later discussion. For CSS codes, Steane suggested a syndrome measurement circuit as shown in figure 2 [24]. Here, two logical ancilla blocks of the same code are used that protects the data state. Two transversal CNOT gates propagate Z and X errors from the data block to ancilla blocks and corresponding error syndromes are calculated from the bitwise measurement outcomes. If the two ancilla blocks are prepared in a tensor product state $|0_L\rangle^{\otimes k} \otimes |+_L\rangle^{\otimes k}$, the circuit extracts the error syndromes without disturbing the encoded quantum information. Like the Knill syndrome measurement, the circuit is intrinsically FT.

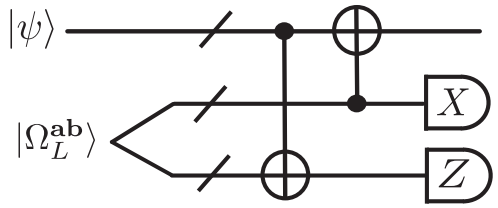

Figure 2. The quantum circuit for Steane syndrome measurement using an $[[n, k, d]]$ CSS code. One can simultaneously measure a Hermitian Pauli operator $i^{\tau_{\mathbf{ab}}} X_L^{\mathbf{a}} Z_L^{\mathbf{b}}$, when the two ancilla blocks are prepared in the state $|\Omega_L^{\mathbf{ab}}\rangle = \frac{1}{\sqrt{2}} \left(I + i^{\tau_{\mathbf{ab}}} X_L^{\mathbf{a}} \otimes Z_L^{\mathbf{b}}\right) |0_L\rangle^{\otimes k} \otimes |+_L\rangle^{\otimes k}$.

Moreover, one can simultaneously measure an arbitrary Hermitian logical Pauli operator of the form $i^{\tau_{\mathbf{ab}}} X_L^{\mathbf{a}} Z_L^{\mathbf{b}}$ while extracting syndromes, if $|\Omega_L^{\mathbf{ab}}\rangle$ is prepared in

$$|\Omega_L^{\mathbf{ab}}\rangle = \frac{1}{\sqrt{2}} \left(I + i^{\tau_{\mathbf{ab}}} X_L^{\mathbf{a}} \otimes Z_L^{\mathbf{b}}\right) |0_L\rangle^{\otimes k} \otimes |+_L\rangle^{\otimes k}. \tag{3}$$

It is easy to prove the functionality of the circuit: start with the joint state $|\psi\rangle|\Omega_L^{\mathbf{ab}}\rangle$, after two transversal CNOTs, the state becomes

$$\frac{1}{\sqrt{2}} \left(|\psi\rangle|0_L\rangle^{\otimes k}|+_L\rangle^{\otimes k} + i^{\tau_{\mathbf{ab}}} X_L^{\mathbf{a}} Z_L^{\mathbf{b}} |\psi\rangle X_L^{\mathbf{a}} |0_L\rangle^{\otimes k} Z_L^{\mathbf{b}} |+_L\rangle^{\otimes k}\right).$$

Let the measurement outcomes of the jth logical qubit in the upper and lower blocks be v_j^x and $v_j^z \in \{0, 1\}$, respectively. Then the joint output state is:

$$\begin{aligned}
&\frac{1}{\sqrt{2}} |\psi\rangle \otimes_{j=1}^k \left(\frac{I + (-1)^{v_j^x} X_L}{2} |0_L\rangle\right) \otimes_{j=1}^k \left(\frac{I + (-1)^{v_j^z} Z_L}{2} |+_L\rangle\right) + \frac{1}{\sqrt{2}} i^{\tau_{\mathbf{ab}}} X_L^{\mathbf{a}} Z_L^{\mathbf{b}} |\psi\rangle \\
&\otimes_{j=1}^k \left(\frac{I + (-1)^{v_j^x} X_L}{2} X_L^{a_j} |0_L\rangle\right) \otimes_{j=1}^k \left(\frac{I + (-1)^{v_j^z} Z_L}{2} Z_L^{b_j} |+_L\rangle\right) \\
&= \frac{1}{\sqrt{2}} \left(I + \prod_{l \in \mathrm{supp}(\mathbf{a})} (-1)^{v_l^x} \prod_{l \in \mathrm{supp}(\mathbf{b})} (-1)^{v_l^z} i^{\tau_{\mathbf{ab}}} X_L^a Z_L^b\right) |\psi\rangle \\
&\otimes_{j=1}^k \left(\frac{I + (-1)^{v_j^x} X_L}{2} |0_L\rangle\right) \otimes_{j=1}^k \left(\frac{I + (-1)^{v_j^z} Z_L}{2} |+_L\rangle\right),
\end{aligned} \tag{4}$$

which is the state after the measurement of $i^{\tau_{\mathbf{ab}}} X_L^{\mathbf{a}} Z_L^{\mathbf{b}}$ on $|\psi\rangle$ with measurement outcome

$$\prod_{l \in \mathrm{supp}(\mathbf{a})} (-1)^{v_l^x} \prod_{l \in \mathrm{supp}(\mathbf{b})} (-1)^{v_l^z}.$$

This circuit also allows measuring several commuting logical Pauli operator simultaneously. Here, we restrict ourselves to a commuting set of $m \leqslant k$ logical Pauli operators and suppose that the set of commuting Pauli operators to be measured is $\{X_L^{\mathbf{e}_1} Z_L^{\mathbf{f}_1}, \ldots, X_L^{\mathbf{e}_m} Z_L^{\mathbf{f}_m}\}$. These operators can be simultaneously measured by replacing $|\Omega_L^{\mathbf{ab}}\rangle$ with:

$$|\Omega_L^{\mathbf{EF}}\rangle = \frac{1}{\sqrt{2^m}} \prod_{j=1}^m \left(I + i^{\tau_{\mathbf{e}_j \mathbf{f}_j}} X_L^{\mathbf{e}_j} \otimes Z_L^{\mathbf{f}_j}\right) |0_L\rangle^{\otimes k} \otimes |+_L\rangle^{\otimes k}. \tag{5}$$

Note that $|\Omega_L^{\mathbf{EF}}\rangle$ is also a stabilizer state. Like logical circuit teleportation, one can also effectively construct any logical Clifford circuit via such Pauli measurements [8, 21].

3.2. Single-shot FT logical circuit teleportation and Pauli measurement

Ideally, if the ancilla qubits are clean and measurements are perfect, one can extract the error syndrome of the data block with logical circuit teleportation or Pauli measurements in a single round of Knill/Steane syndrome measurement.

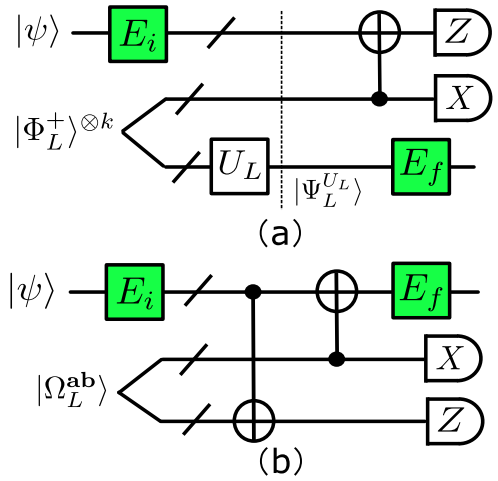

Figure 3. The effective error model of the Knill (part (a)) and Steane (part(b)) syndrome measurement circuits. While E_i and E_f are in general correlated in time, they are spatially uncorrelated, if the ancilla states are qualified.

In practice, ancillas may contain different types of errors after preparation, while the measurement outcomes can also be noisy. One needs to make sure that high weight errors do not propagate from ancilla qubits to data blocks. At the same time, reliable values of syndromes and logical operators must be established from measurement outcomes. For error correction, one can repeat the syndrome measurements several rounds to establish reliable syndromes of the data state via majority vote [6, 22]. However, for the purpose of logical circuit teleportation or Pauli measurements, one needs reliable values of logical operators right after the first round of measurement for subsequent correction. Otherwise, it will cause a logical error on the data state. Thus, a single-shot FT logical circuit teleportation or Pauli measurement protocol is required.

Actually, we will see this is possible if the blocks of ancilla qubits for Knill/Steane syndrome measurements do not contain any *correlated* errors. Here, we define an uncorrelated error as follow [34]:

Definition 2. For an $[[n, k, d]]$ code correcting any Pauli error on $t = \lfloor \frac{d-1}{2} \rfloor$ qubits, we say that an error E on the code block is spatially *uncorrelated* if the probability of E is

$$\Pr(E) \sim O(p^s) : \begin{cases} \text{for some } s \geqslant \mathrm{wt}(E), & \text{if } \mathrm{wt}(E) \leqslant t; \\ \text{for some } s \geqslant t, & \text{if } \mathrm{wt}(E) > t, \end{cases}$$

where the coefficients behind O are not unreasonably large.

Otherwise, E is said to be correlated. For those uncorrelated errors satisfying this definition, they should have a distribution similar to the binomial distribution. Thus, the errors on the code block can be regarded as independent. We say that an ancilla is *qualified* if it is free of correlated errors.

It is obvious that no correlated error will be propagated back to the data blocks if ancilla blocks are qualified. Even more, we have:

Lemma 1 (Effective error model). *During imperfect logical state teleporation via Knill syndrome extraction, or logical Pauli measurements via Steane syndrome measurement, if errors in the same block (data or ancilla) are spatially uncorrelated according to definition 2, then the errors are equivalent to spatially* uncorrelated *effective errors acting only on the data code block before and after the process, as shown in figure 3.*

It has already been shown in reference [21] that this statement is true for Steane syndrome measurement. The basic idea is that failures occurring in any location of the circuit can be commuted forward or backward to the data code block, allowing the ancillas to be treated as clean and the measurements as perfect. Thus we can leave E_f to the next round of syndrome measurements and analyze as if only E_i (and E_f from the previous round) have occurred, followed by perfect syndrome measurements. The same argument is also applicable to Knill syndrome measurements and hence one has:

Theorem 1. *The Knill/Steane syndrome measurement circuit can implement FT logical circuit teleportation/Pauli measurements in a single round.*

After a single-shot Knill/Steane syndrome measurement and correction, the final data state is:

$$|\psi_f\rangle \propto E_f \cdot R \cdot O_L \cdot L_C \cdot E_{\text{comb}}|\psi_i\rangle. \tag{6}$$

Here, E_{comb} includes both E_i in current stage and E_f from previous stage; O_L is the logical operation (either a teleported logical circuit, or logical Pauli measurements); L_C is the Pauli correction based on the outcomes of logical measurements; R is the recovery operation based on the measured syndromes, O_L and L_C.

3.3. Constant depth Clifford circuit via FT circuit teleportation

For a CSS code with k logical data qubits, it requires $O(k^2/\log k)$ logical Clifford gates [27, 28] for a logical Clifford circuit. If we implement these gates one by one in a FT manner, it will require $O(k^2/\log k)$ qualified ancilla states using $O(k^2/\log k)$ times of the Knill/Steane single-shot syndrome measurements circuit. In this and next subsections, we show that $O(1)$ qualified ancilla states and $O(1)$ steps of the Knill/Steane syndrome measurements are sufficient for arbitrary logical Clifford circuits, up to a permutation of qubits.

It is well known that any Clifford circuit has an equivalent circuit comprising 11 stages, each using only one type of gate: -H-C-P-C-P-C-H-P-C-P-C- [28]. That can be further reduced to a nine-stage -C-P-C-P-H-P-C-P-C- [32]. More specifically, one has:

Theorem 2. *(Bruhat decomposition [32]). Any symplectic matrix M of dimension $2k \times 2k$ can be decomposed as*

$$M = M_{\text{C}}^{(1)} M_{\text{P}}^{(1)} M_{\text{C}}^{(2)} M_{\text{P}}^{(2)} M_{\text{H}}^{(1)} \cdot M_{\text{P}}^{(3)} \left(\pi M_{\text{C}}^{(3)} \pi^{-1}\right) M_{\text{P}}^{(4)} \left(\pi M_{\text{C}}^{(4)} \pi^{-1}\right) \pi. \tag{7}$$

Here, $M_{\text{C}}^{(j)}$ are -C- stage matrices containing only CNOT gates $\text{C}(q, r)$ such that $q < r$; $M_{\text{P}}^{(j)}$ and $M_{\text{H}}^{(j)}$ are -P- and -H- stage matrices; π is a permutation matrix.

In a -P- stage, since $\text{P}^4 = I_2$, effectively there are three types of single-qubit gates: P, $\text{P}^2 = Z$ and $\text{P}^3 = \text{P}^\dagger = \text{P}Z$. Note that we will postpone all the Z gates to the final stage, and thus the -P- layer consists of at most k individual Phase gates. The symplectic matrix of a -P- stage on a set of m qubits is in general of the form:

$$M_{\text{P}} = \left(\begin{array}{c|c} I_k & \Lambda_k^m \\ \hline \mathbf{0}_k & I_k \end{array}\right), \tag{8}$$

where Λ_k^m is an $k \times k$ diagonal matrix with m 1s.

Similar to the -P- stage, since $\text{H}^2 = I_2$, an -H- stage contains at most k individual H gates. The symplectic matrix of an -H- stage on an arbitrary set of m qubits can be written as

$$M_{\text{H}} = \left(\begin{array}{c|c} I_k + \Lambda_k^m & \Lambda_k^m \\ \hline \Lambda_k^m & I_k + \Lambda_k^m \end{array}\right). \tag{9}$$

The corresponding symplectic matrix of a -C- stage can be written as:

$$M_{\text{C}} = \left(\begin{array}{c|c} U & \mathbf{0}_k \\ \hline \mathbf{0}_k & \left(U^t\right)^{-1} \end{array}\right), \tag{10}$$

where U is an invertible $k \times k$ upper triangular matrix.

Clearly, if one can implement each stage in $O(1)$ steps fault-tolerantly, an arbitrary logical Clifford circuit can be implemented in $O(1)$ steps. For Knill syndrome measurements, it is straightforward—one can prepare the ancilla for the circuit in each stage directly as:

$$\begin{aligned} \left|\Psi_L^{U_{\text{P}}}\right\rangle &= I \otimes U_{\text{P}}(|0_L\rangle \otimes |0_L\rangle + |1_L\rangle \otimes |1_L\rangle)^{\otimes k}, \\ \left|\Psi_L^{U_{\text{H}}}\right\rangle &= I \otimes U_{\text{H}}(|0_L\rangle \otimes |0_L\rangle + |1_L\rangle \times |1_L\rangle)^{\otimes k}, \\ \left|\Psi_L^{U_{\text{C}}}\right\rangle &= I \otimes U_{\text{C}}(|0_L\rangle \otimes |0_L\rangle + |1_L\rangle \otimes |1_L\rangle)^{\otimes k} \end{aligned} \tag{11}$$

where U_{P}, U_{H} and U_{C} are the corresponding unitaries for the -P-, -H- and -C- stages, respectively. Obviously, these are all CSS states up to local Clifford operations, whose binary representations at the logical level are:

$$\Psi_L^{U_{\text{P}}} = \left(\begin{array}{cc|cc} I_k & I_k & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \mathbf{0} & I_k & I_k \end{array}\right), \tag{12}$$

$$\Psi_L^{U_{\rm H}} = \left(\begin{array}{cc|cc} I_k & I_k+\Lambda_k^m & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \Lambda_k^m & I_k & I_k+\Lambda_k^m \end{array}\right) \tag{13}$$

assuming -P- or -H- is applied to a set of m qubits, and

$$\Psi_L^{U_{\rm C}} = \left(\begin{array}{cc|cc} I_k & U & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_k & \left(U^t\right)^{-1} \end{array}\right). \tag{14}$$

If these states are all qualified for all the stages, by theorems 1 and 2, one can implement an arbitrary logical Clifford circuit in 9 rounds of single-shot Knill syndrome measurements. Later, we will show that all the three types of ancilla states can be prepared fault-tolerantly and efficiently.

3.4. Constant depth Clifford circuit FT Pauli measurement

Unlike Knill syndrome measurement, it is not so obvious how to implement the logical Clifford group using Steane syndrome measurement. In this section, we provide a constructive proof showing that by introducing k extra auxiliary logical qubits (labeled as $\mathrm{A}_1, \ldots, \mathrm{A}_k$), each stage of a logical Clifford circuit on k data logical qubits ($\mathrm{Q}_1, \ldots, \mathrm{Q}_k$) can be implemented via a constant number of Pauli operator measurements, up to a permutation of qubits. We choose an $[[n, 2k, d]]$ CSS code and put the logical qubits in the following order: $\{\mathrm{A}_1, \ldots, \mathrm{A}_k, \mathrm{Q}_1, \ldots, \mathrm{Q}_k\}$.

3.4.1. -P- stage

Consider a pair of qubits $\{\mathrm{A}_j, \mathrm{Q}_j\}$ with A_j in the $|0_L\rangle$ state. Measure operators $X_{\mathrm{A}_j,L}Y_{\mathrm{Q}_j,L}$ and then $Z_{\mathrm{Q}_j,L}$. After swapping A_j and Q_j, the overall effect is a Phase gate on Q_j up to a Pauli correction depending on the measurement outcomes. The swap does not need to be done physically. Instead, one can just keep a record of it in software.

For m Phase gates on a logical qubit set $\mathscr{M}$, since $\{X_{\mathrm{A}_j,L}Y_{\mathrm{Q}_j,L} |\ j \in \mathscr{M}\}$ and $\{Z_{\mathrm{Q}_j,L} |\ j \in \mathscr{M}\}$ are commuting operator sets, it requires only two steps of Pauli measurements by preparing two ancilla states with $4k$ logical qubits:

$$\left|\Omega_L^{\mathrm{P}_1}\right\rangle = \frac{1}{\sqrt{2^m}} \prod_{j\in\mathscr{M}} \left(I + i\left(X_{j,L}X_{j+k,L}\right) \otimes Z_{j+k,L}\right) |0_L\rangle^{\otimes 2k} \otimes |+_L\rangle^{\otimes 2k} \tag{15}$$

and

$$\left|\Omega_L^{\mathrm{P}_2}\right\rangle = \frac{1}{\sqrt{2^m}} |0_L\rangle^{\otimes 2k} \otimes \left(\prod_{j\in\mathscr{M}} \left(I + Z_{j,L}\right) |+_L\rangle^{\otimes 2k}\right), \tag{16}$$

whose binary representations at the logical level are

$$\Omega_L^{\mathrm{P}_1} = \left(\begin{array}{cccc|cccc} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \Lambda_k^m & \mathbf{0} & \mathbf{0} \\ \Lambda_k^m & \Lambda_k^m & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \Lambda_k^m & \mathbf{0} & \mathbf{0}_k \end{array}\right), \tag{17}$$

and

$$\Omega_L^{\mathrm{P}_2} = \left(\begin{array}{ccc|ccc} \mathbf{0}_{2k} & \mathbf{0} & \mathbf{0} & I_{2k} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_k+\Lambda_k^m & \mathbf{0} & \mathbf{0} & \Lambda_k^m \end{array}\right), \tag{18}$$

respectively. Note that $|\Omega_L^{\mathrm{P}_2}\rangle$ is a CSS state. $|\Omega_L^{\mathrm{P}_1}\rangle$ is the joint $+1$ eigenstate of

$$\{Z_{j,L}Z_{j+k,L} \otimes I_{2n}, Z_{j+k,L} \otimes X_{j+k,L}, X_{j,L}Y_{j+k,L} \otimes Z_{j+k,L} \mid j \in \mathscr{M}\},$$

which is also a CSS state up to Phase gates on logical qubits $\{j+k |\ j \in \mathscr{M}\}$ of the upper block, and Hadamard gates on the logical qubits $\{j+k |\ j \in \mathscr{M}\}$ of the lower block.

3.4.2. -H- stage

Like the -P- stage, we consider only a single H on a data qubit. For a pair of qubits $\{\mathrm{A}_j, \mathrm{Q}_j\}$ with A_j in the $|0_L\rangle$ state, measure $X_{\mathrm{A}_j,L}Z_{\mathrm{Q}_j,L}$ and then $X_{\mathrm{Q}_j,L}$. After swapping A_j and Q_j, the overall effect is a Hadamard gate on Q_j with A_j in the $|+_L\rangle$ up to a Pauli correction depending on the measurement outcome.

For m Hadamard gates on a logical qubits set $\mathscr{M}$, since $\{X_{\mathrm{A}_j,L}Z_{\mathrm{Q}_j,L} \mid j \in \mathscr{M}\}$ and $\{X_{\mathrm{Q}_j,L} \mid j \in \mathscr{M}\}$ are both commuting sets, we need just two steps of Pauli measurements and an ancilla state with $4k$ logical qubits for an -H- stage. If Hadamard gates are applied to a set $\mathscr{M}$ of qubits, the required ancilla states are

$$\left|\Omega_L^{\mathrm{H}_1}\right\rangle = \frac{1}{\sqrt{2^m}} \prod_{j\in\mathscr{M}} \left(I + X_{j,L} \otimes Z_{j+k,L}\right) |0_L\rangle^{\otimes 2k} \otimes |+_L\rangle^{\otimes 2k}, \tag{19}$$

and

$$\left|\Omega_L^{\mathrm{H}_2}\right\rangle = \frac{1}{\sqrt{2^m}} \left(\prod_{j\in\mathscr{M}} \left(I + X_{j,L}\right) |0_L\rangle^{\otimes 2k} \right) \otimes |+_L\rangle^{\otimes 2k}, \tag{20}$$

whose binary representations at the logical level are:

$$\Omega_L^{\mathrm{H}_1} = \left(\begin{array}{cccc|cccc} \Lambda_k^m & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k + \Lambda_k^m & \mathbf{0} & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \mathbf{0}_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \Lambda_k^m & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array}\right) \tag{21}$$

and

$$\Omega_L^{\mathrm{H}_2} = \left(\begin{array}{ccc|ccc} \mathbf{0}_k & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda_k^m & \mathbf{0} & \mathbf{0} & I_k + \Lambda_k^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{2k} & \mathbf{0} & \mathbf{0} & \mathbf{0}_{2k} \end{array}\right), \tag{22}$$

respectively. Note that $\left|\Omega_L^{\mathrm{H}_2}\right\rangle$ is a CSS state. $\left|\Omega_L^{\mathrm{H}_1}\right\rangle$ is the joint $+1$ eigenstate of $\{X_{j,L} \otimes Z_{j+k,L}, Z_{j,L} \otimes X_{j+k,L} \mid j \in \mathscr{M}\}$, and thus, it is a CSS state up to Hadamard gates.

3.4.3. -C- stage

We first introduce the generalized stabilizer formalism that is helpful later. Consider a 2^k dimensional subspace $C(\mathcal{G})$ of the N logical qubit Hilbert space, where $\mathcal{G}$ has $N-k$ stabilizer generators. We focus on the effects of Clifford circuits on the k logical qubits stabilized by $\mathcal{G}$. Consider a set of matrices $\mathsf{C}_\mathcal{G}$ of the form:

$$\left(\begin{array}{c|c} Q' & R' \\ \hline S' & T' \\ \hline A & B \end{array}\right). \tag{23}$$

Here, $(A|B)$ corresponds to the stabilizer generators of $\mathcal{G}$; $(Q'|R')$ and $(S'|T')$ are $k \times 2N$ binary matrices orthogonal to $(A|B)$ with respect to the symplectic inner product, and which are symplectic partners of each other. They can be regarded as 'encoded operators' on $C(\mathcal{G})$. We define the following equivalence relation $\mathscr{R}$ in $\mathsf{C}_\mathcal{G}$: two matrices

$$\mathrm{C}_1 = \left(\begin{array}{c|c} Q_1' & R_1' \\ \hline S_1' & T_1' \\ \hline A_1 & B_1 \end{array}\right) \quad \text{and} \quad \mathrm{C}_2 = \left(\begin{array}{c|c} Q_2' & R_2' \\ \hline S_2' & T_2' \\ \hline A_2 & B_2 \end{array}\right),$$

are equivalent if (a) $(A_1|B_1)$ and $(A_2|B_2)$ generate the same stabilizer group $\mathcal{G}$; and (b) $\begin{pmatrix} Q_1' & R_1' \\ S_1' & T_1' \end{pmatrix}$ differs from $\begin{pmatrix} Q_2' & R_2' \\ S_2' & T_2' \end{pmatrix}$ by multiplication of elements in $\mathcal{G}$. Thus, there is a one-to-one correspondence between $\mathsf{C}_\mathcal{G}/\mathscr{R}$ and $\mathrm{Sp}(2k, \mathbb{Z}_2)$. Therefore, $\mathsf{C}_\mathcal{G}/\mathscr{R}$ captures the behavior of stabilizer circuits on $C(\mathcal{G})$. The circuit representation of equation (23) is called the *generalized stabilizer form* (GSF) of $\mathcal{G}$.

The following lemma will also be used in the circuit construction:

Lemma 2. *Let L_1 be an $n \times n$ lower triangular matrix with the diagonal elements being zeros. Suppose*

$$L = (I_n \; L_1).$$

Then there exists a full-rank matrix $L' = (L_2 L_3)$, where L_2 and L_3 are two $n \times n$ lower triangular matrices, such that the rows of L' are linear combinations of rows of L and

$$L' \begin{pmatrix} I_n \\ I_n \end{pmatrix} = L_2 + L_3 = I_n. \tag{24}$$

Proof. See appendix A for details. □

We now construct the sequence of Pauli measurements which can generate any -C- stage on logical qubits $Q_1, \ldots, Q_k$ of an $[[n, N = 2k, d]]$ CSS code. We start with the GSF of an arbitrary -C- circuit with auxiliary logical qubits $A_1, \ldots, A_k$ in $|+_L\rangle^{\otimes k}$:

$$\left(\begin{array}{cc|cc} \mathbf{0}_k & U & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k & (U^t)^{-1} \\ \hline I_k & \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k \end{array}\right), \tag{25}$$

and reduce it to the idle circuit by a series of row operation. This set of operations in reverse will effectively implement the target CNOT circuit.

As mentioned before, U is an invertible upper triangular matrix. The GSF is then equivalent to

$$\left(\begin{array}{cc|cc} U + I_k & U & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k & (U^t)^{-1} \\ \hline I_k & \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k \end{array}\right) \tag{26}$$

since all the nonzero row vectors of $\left(U + I_k \;\; \mathbf{0}_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right)$ can be generated by $\left(I_k \;\; \mathbf{0}_k \mid \mathbf{0}_n \;\; \mathbf{0}_k\right)$ and we add these vectors to the first row.

Since U is of full rank, the diagonal elements of $U + I_k$ must be all zeros. Observe that $\left(\mathbf{0}_k \;\; \mathbf{0}_k \mid I_k \;\; (U^t)^{-1} + I_k\right)$ commutes with the logical operators and is a symplectic partner of the stabilizer generators, since

$$\left(I_k \quad (U^t)^{-1} + I_k\right)(U + I_k \quad U)^t = \mathbf{0}_k,$$

and

$$\left(I_k \; \mathbf{0}_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right)\left(I_k \;\; (U^t)^{-1} + I_k \mid \mathbf{0}_k \; \mathbf{0}_k\right)^t = I_{2k}.$$

One can measure k commuting logical Pauli operators $\left(\mathbf{0}_k \; \mathbf{0}_k \mid I_k \;\; (U^t)^{-1} + I_k\right)$ simultaneously. The GSF will then be transformed into

$$\left(\begin{array}{cc|cc} U + I_k & U & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k & (U^t)^{-1} \\ \hline \mathbf{0}_k & \mathbf{0}_k & I_k & (U^t)^{-1} + I_k \end{array}\right). \tag{27}$$

(Meanwhile, we can perform the Pauli measurements $\left(I_k \;\; \mathbf{0}_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right)$ to reverse the process (from equation (27) to equation (26))).

Now, adding the third row of equation (27) to the second row, one can obtain an equivalent GSF

$$\left(\begin{array}{cc|cc} U + I_k & U & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & I_k & I_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & I_k & (U^t)^{-1} + I_k \end{array}\right). \tag{28}$$

Let $L = (I_k \quad L_1)$, where $L_1 = (U^t)^{-1} + I_n$ is a lower triangular matrix with all the diagonal elements being 0. By lemma 2, the GSF can be equivalently transformed into

$$\left(\begin{array}{cc|cc} U + I_k & U & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & I_k & I_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & L_2 & L_3 \end{array}\right), \tag{29}$$

where $(L_2 \;\; L_3)(I_k \;\; I_k)^t = I_k$. One can measure a set of k Pauli operators $(I_k \;\; I_k \mid \mathbf{0}_k \;\; \mathbf{0}_k)$ simultaneously and transform the GSF into

$$\left(\begin{array}{cc|cc} U + I_k & U & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & I_k & I_k \\ \hline I_k & I_k & \mathbf{0}_k & \mathbf{0}_k \end{array}\right). \tag{30}$$

Meanwhile, measuring $\left(\mathbf{0}_k \;\; \mathbf{0}_k \mid L_2 \;\; L_3\right)$ will transfer the GSF of equation (30) into equation (29). Note that the measurement of $\left(\mathbf{0}_k \;\; \mathbf{0}_k \mid L_2 \;\; L_3\right)$ is equivalent to measuring $\left(\mathbf{0}_k \;\; \mathbf{0}_k \mid I_k \;\; (U^t)^{-1} + I_k\right)$.

Now, since the stabilizer generators in equation (30) are of the form $\left(I_k \;\; I_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right)$, one can add $\left(U + I_k \;\; U + I_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right)$ to the first row of equation (30), which equivalently reduces the GSF to:

$$\left(\begin{array}{cc|cc} \mathbf{0}_k & I_k & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & I_k & I_k \\ \hline I_k & I_k & \mathbf{0}_k & \mathbf{0}_k \end{array}\right). \tag{31}$$

The final step is to eliminate the left-most I_k in the second row of equation (31). This can be done by measuring $\left(\mathbf{0}_k \;\; \mathbf{0}_k \mid I_k \;\; \mathbf{0}_k\right)$ and adding the third row to the second. This will then transform the GSF into the second matrix in:

$$\left(\begin{array}{cc|cc} \mathbf{0}_k & I_k & \mathbf{0}_k & \mathbf{0}_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k & I_k \\ \hline \mathbf{0}_k & \mathbf{0}_k & I_k & \mathbf{0}_k \end{array}\right), \tag{32}$$

Meanwhile, one can measure the set of k logical Pauli operators $\left(I_k \;\; I_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right)$ to transform equation (32) to equation (31).

To reverse the whole procedure above and start from equation (32), we initially set $\mathrm{A}_1, \ldots, \mathrm{A}_k$ to $|0_L\rangle^{\otimes k}$ and perform the following three sets of Pauli measurements:

$$\begin{aligned} &1.\ \left(I_k \;\; I_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right), \\ &2.\ \left(\mathbf{0}_k \;\; \mathbf{0}_k \mid I_k \;\; (U^t)^{-1} + I_k\right), \\ &3.\ \left(I_k \;\; \mathbf{0}_k \mid \mathbf{0}_k \;\; \mathbf{0}_k\right). \end{aligned} \tag{33}$$

The measurements require three $4k$ logical qubits CSS ancilla states, which are

$$\left|\Omega_L^{\mathrm{C}_1}\right\rangle = \frac{1}{\sqrt{2^k}} \left(\prod_{j=1}^{k} \left(I + X_{j,L} X_{j+k,L}\right) |0_L\rangle^{\otimes 2k} \right) \otimes |+_L\rangle^{\otimes 2k}, \tag{34}$$

$$\left|\Omega_L^{\mathrm{C}_2}\right\rangle = \frac{1}{\sqrt{2^k}} |0_L\rangle^{2k} \otimes \left(\prod_{j=1}^{k} \left(I + Z_L^{\mathbf{u}_j}\right) |+_L\rangle^{\otimes 2k}) \right), \tag{35}$$

and

$$\left|\Omega_L^{\mathrm{C}_3}\right\rangle = \left(|+_L\rangle^{\otimes k} |0_L\rangle^{\otimes k}\right) \otimes |+_L\rangle^{\otimes 2k}. \tag{36}$$

Here, $\mathbf{u}_j$ is the jth row of $\left(I_k \quad (U^t)^{-1} + I_k\right)$. The binary representations of these states are:

$$\Omega_L^{\mathrm{C}_1} = \left(\begin{array}{ccc|ccc} I_k & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & I_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{2k} & \mathbf{0} & \mathbf{0} & \mathbf{0}_{2k} \end{array}\right), \tag{37}$$

$$\Omega_L^{\mathrm{C}_2} = \left(\begin{array}{ccc|ccc} \mathbf{0}_{2k} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & (U^t)^{-1} + I_k \\ \mathbf{0} & \left[(U^t)^{-1}\right]^t + I_k & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_{2k} & \mathbf{0} & \mathbf{0} \end{array}\right), \tag{38}$$

and

$$\Omega_L^{\mathrm{C}_3} = \left(\begin{array}{ccc|ccc} I_k & \mathbf{0} & \mathbf{0} & \mathbf{0}_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_k & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{2k} & \mathbf{0} & \mathbf{0} & \mathbf{0}_{2k} \end{array}\right). \tag{39}$$

$\left|\Omega_L^{\mathrm{C}_1}\right\rangle$ is actually a k-fold tensor product of Bell states. $\left|\Omega_L^{\mathrm{C}_2}\right\rangle$ is the key resource state in our procedure to reduce the depth of -C- stage computation. All the ancillas here are CSS states. The net effect is the desired -C- stage acting on $\mathrm{Q}_1, \ldots, \mathrm{Q}_k$, and the auxiliary qubits $\mathrm{A}_1, \ldots, \mathrm{A}_k$ are reset to $|+_L\rangle^{\otimes k}$ (up to logical Z corrections). One can transform $\mathrm{A}_1, \ldots, \mathrm{A}_k$ back into $|0_L\rangle^{\otimes k}$ or just keep them and start with $|+_L\rangle^{\otimes k}$ for the next stage. The procedure with $\mathrm{A}_1, \ldots, \mathrm{A}_k$ initially in the $|+_L\rangle^{\otimes k}$ state for -C- stage is similar. As a conclusion, one has the following theorem:

Theorem 3. *For an $[[n, 2k, d]]$ CSS code, any logical Clifford circuit on k logical qubits can be realized fault-tolerantly by 22 rounds of single-shot Steane syndrome measurement.*

3.5. FT preparation of qualified ancilla states

The ancilla states listed in the last subsection can be prepared fault-tolerantly by using Shor syndrome measurement to measure all stabilizer generators and logical Pauli operators, which will be discussed in section 4.1. In this subsection, we generalize the FT state preparation protocol in [29, 30] to show that all the logical stabilizer ancilla states required in the previous subsection can be prepared fault-tolerantly by distillation with almost constant overhead in terms of the number of qubits.

Since all logical ancilla states we considered are stabilizer states, once the eigenvalues of all their stabilizers are known, one can remove the errors completely. The basic idea of distillation is shown in

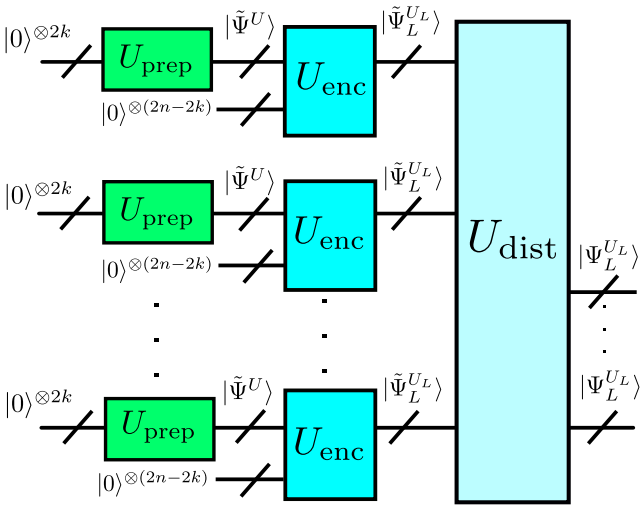

Figure 4. FT ancilla state distillation circuit for $\left|\Psi_L^{U_L}\right\rangle$. Ancilla states are prepared and encoded via noisy quantum circuits U_{prep} and U_{enc}. Then they are fed to the distillation circuit and output the qualified ancilla state. The preparation of $|\Omega_L^{\mathbf{EF}}\rangle$ is similar.

figure 4—many copies of the logical ancilla states are prepared in the physical level via the same U_{prep} and then encoded by U_{enc} to the same large block code. Both U_{prep} and U_{enc} are noisy in practice. They are sent into a distillation circuit (which is also noisy) and certain blocks are measured bitwise. The eigenvalues of all the stabilizers in group $\mathcal{S}$ and the logical Pauli operators of the output blocks can be estimated if the distillation circuit is constructed based on the parity-check matrix of some classical error-correcting code—the flipped eigenvalues caused by errors during state preparation and distillation can be treated as classical noise and thus be decoded. Since correlated errors remaining on a block can cause its estimated eigenvalues not compatible with each other, each output block needs further check for compatibility. Postselection (rejecting the blocks whose estimated eigenvalues are incompatible) is then done to remove the blocks likely containing correlated errors. Error correction is then applied based on the estimated eigenvalues of stabilizers to the remaining blocks.

The distillation circuit can be synthesized according the parity-check matrix of an $[n_c, k_c, d_c]$ classical code, which has the form $\mathsf{H} = (I_{n_c-k_c}|\ \mathsf{A}_c)$. Consider a group of n_c ancilla blocks. Choose the first $r_c = n_c - k_c$ ancillas blocks to hold the classical parity checks, and do transversal CNOTs from the remaining k_c ancillas onto each of the parity-check ancillas according to the pattern of 1s in the rows of A_c: if $[\mathsf{A}_c]_{i,j} = 1$, we apply a transversal CNOT from the $(r_c + j)$th ancilla to the ith ancilla block. Then measure all the qubits on each of the first r_c ancilla blocks in the Z basis, which destroys the states of those blocks and extracts information to estimate the eigenvalues of all Z types stabilizers and logical operators of the remaining k_c blocks. In the low error regime, after filtering out the blocks with incompatible estimated eigenvalues of stabilizers, the output blocks will contain no correlated X errors after subsequent error correction if d_c is larger than the distance of the underlying CSS code [30]. Correlated Z errors can be removed in a similar manner.

One can concatenate two stages of distillation (with the output blocks of the first stage randomly shuffled) to remove both correlated X and Z errors. Figure 5 shows the overall circuit to distill $\left|\Psi_L^{U_C}\right\rangle$ using the [3, 1, 3] code for both stages. For a two-stage distillation protocol based on two classical $[n_{c_1}, k_{c_1}, d_{c_1}]$ and $[n_{c_2}, k_{c_2}, d_{c_2}]$ codes, the number of input and output blocks are $n_{c_1} n_{c_2}$ and $Y(p) \cdot n_{c_1} n_{c_2}$ respectively. Here, $Y(p)$ is the yield rate defined as

$$Y(p) = \frac{k_{c_1} k_{c_2} (1 - R_1(p))(1 - R_2(p))}{n_{c_1} n_{c_2}}, \tag{40}$$

where $R_i(p)$ is the block rejection rate for postselection in the ith stage of distillation for a gate/measurement error rate p. Asymptotically, the rejection rate for each round of distillation is $O(p^2)$, because at least two failures are needed to cause a rejection of the output blocks. Thus, it is likely that $R_1(p)$ and $R_2(p)$ negligible in the small p regime. On the other hand, good capacity-achieving classical codes exist such that $\frac{k_{c_1} k_{c_2}}{n_{c_1} n_{c_2}}$ can be independent of the code distance of the underlying CSS code to ensure $d_c > d$, so that the distillation circuits are still able to output qualified ancilla states. Hence, $Y(p)$ can achieve almost $\Theta(1)$ for sufficiently low p.

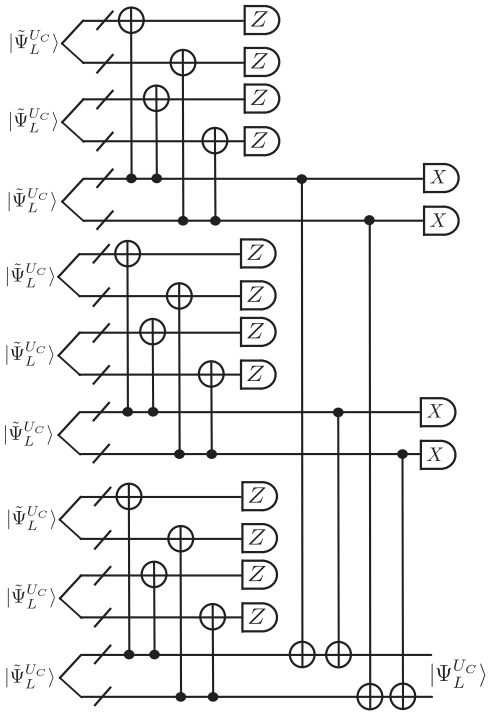

Figure 5. Two-stage FT ancilla state distillation circuit for $|\Psi_L^{U_C}\rangle$. The distillation circuit is based on the parity-check matrix of the [1, 3, 3] code for both stages. Actually this circuit can distill any logical CSS states fault-tolerantly.

As shown in the last subsection, one needs several ancilla states, which can be grouped into three types:

Type I—These are CSS states including $\left|\Psi_L^{U_C}\right\rangle$, $\left|\Omega_L^{P_2}\right\rangle$ $\left|\Omega_L^{H_2}\right\rangle$, $\left|\Omega_L^{C_1}\right\rangle$, $\left|\Omega_L^{C_2}\right\rangle$, and $\left|\Omega_L^{C_3}\right\rangle$. They are stabilized by logical operators which are tensor products of either X or Z. They can be prepared and distilled by the circuit in figure 5: eigenvalues of the Z (X) stabilizer generators and Z (X) logical operators are checked at the first (second) round to remove correlated X (Z) errors.

Type II—These are CSS states up to logical Hadamard gates applied to some logical qubits, including $\left|\Psi_L^{U_H}\right\rangle$ and $\left|\Omega_L^{H_1}\right\rangle$. In this case, we distill the upper block to remove correlated X errors and the lower block to remove correlated Z errors in the first round and reverse the order in the second round, as shown in figure 6.

Type III—This set of states are type I or II states up to logical Phase gates applied to some logical qubits, including $\left|\Psi_L^{U_P}\right\rangle$ and $\left|\Omega_L^{P_2}\right\rangle$. We will confine our attention to doubly even and self-dual CSS codes and set the weight of logical X operators $X_{j,L}$ to odd numbers for all j.

For $\left|\Psi_L^{U_P}\right\rangle$, one could first prepare a qualified CSS state of the form

$$\Psi_L^{U_P'} = \left(\begin{array}{cc|cc} I_k & \Lambda_k^m & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Lambda_k^m & I_k \end{array}\right) \tag{41}$$

via the distillation circuit in figure 5. Then, Phase gates are applied bitwise to the lower ancilla block, which will transform the state to

$$\Psi_L^{U_P''} = \left(\begin{array}{cc|cc} I_k & \Lambda_k^m & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \mathbf{0} & \Lambda_k^m & I_k \end{array}\right). \tag{42}$$

This is because the bitwise Phase gates will preserve the stabilizer group while implementing logical Phase gates on all the logical qubits. Then one can apply logical CNOTs (assisted by some CSS states) from the upper block to the lower block on the remaining $k-m$ logical qubits to obtain a qualified $\left|\Psi_L^{U_P}\right\rangle$.

Similarly, for $\left|\Omega_L^{P_1}\right\rangle$, one can prepare a qualified type II state of the form

$$\Omega_L^{P_1'} = \left(\begin{array}{cccc|cccc} I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda_k^m & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k+\Lambda_k^m & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \Lambda_k^m & \mathbf{0} & \mathbf{0}_k \end{array}\right) \tag{43}$$

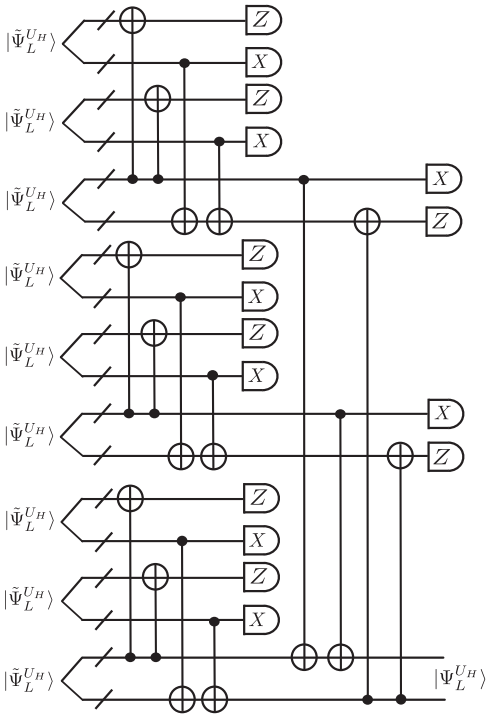

Figure 6. Two-stage FT ancilla state distillation circuit for $\left|\Psi_L^{U_H}\right\rangle$ based on the parity-check matrix of [3, 1, 3] code. This circuit can be used to prepare any logical CSS states up to logical Hadamard gates fault-tolerantly.

by the circuit in figure 6. After that, bitwise Phase gates are applied to the upper block to transform $\left|\Omega_L^{P_1'}\right\rangle$ to

$$\Omega_L^{P_1''} = \left(\begin{array}{cccc|cccc} I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda_k^m & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \Lambda_k^m & \mathbf{0} & \mathbf{0}_k \end{array}\right). \tag{44}$$

We can then measure the operators $(\mathbf{0}\ \ \mathbf{0} \mid I_k\ \Lambda_k^m)$ (assisted by some CSS states) on the upper block and obtain

$$\Omega_L^{P_1} = \left(\begin{array}{cccc|cccc} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \Lambda_k^m & \mathbf{0} & \mathbf{0} \\ \Lambda_k^m & \Lambda_k^m & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \Lambda_k^m \\ \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I_k & \mathbf{0} & \Lambda_k^m & \mathbf{0} & \mathbf{0}_k \end{array}\right) \tag{45}$$

up to logical Pauli corrections.

3.6. Resource overhead

We estimate the average number of qubits and physical gates to implement a logical Clifford circuit in this subsection.

For both Knill and Steane syndrome measurements, one only needs a constant rounds of circuit teleportations or Pauli measurements. Asymptotically, the FT ancilla state distillation protocol dominates the resource cost. We can recycle the ancilla qubits after they are used to further reduce the redundancy. However, it will not change the asymptotic scaling of resource cost.

As discussed in the previous section, the overall number of qubits required for a two-stage state distillation protocol based on the parity-check matrices of two classical codes is

$$N_q = cnn_{c_1}n_{c_2},$$

where c is some constant. The numbers of gates for each U_{prep} and U_{enc} in figure 4 are $O(k^2/\log k)$ and $O(n^2/\log n)$, respectively. Therefore, the total number of gates for noisy logical stabilizer state preparation at the physical level is

$$N_{\text{enc}} = \frac{cn^2 n_{c_1} n_{c_2}}{\log\ n}$$

with depth $O(n)$.

The number of gates for the two-round distillation circuit depends on A_c, which has $O(k_c^2)$ 1s. Hence, the number of gates required for U_{dist} for two stages is

$$N_{\text{dist}} = c_1 k_{c_1}^2 n n_{c_2} + c_2 k_{c_2}^2 n,$$

with depth $O\left(\max\{k_{c_1}, k_{c_2}\}\right)$, where c_1, c_2 are constants. Therefore, the overall depth of ancilla state preparation is

$$O(\max\{n, k_{c_1}, k_{c_2}\}).$$

Note that there are $Y(p)n_{c_1}n_{c_2}$ output ancilla blocks per round of distillation and hence the same number of identical circuits can be implemented. On average, one needs

$$\bar{N}_q = \frac{c n_{c_1} n_{c_2} n}{Y(p) n_{c_1} n_{c_2}} \sim O(n)$$

qubits for a logical Clifford circuit. On the other hand, the physical gates required for raw state preparation is

$$\bar{N}_{\text{enc}} = \frac{c n^2 n_{c_1} n_{c_2}}{Y(p) n_{c_1} n_{c_2} \log\ n} \sim O(n^2/\log\ n).$$

If two good classical capacity-achieving codes are used(e.g., low-density parity-check (LDPC) codes [35]) with $k_{c_i}/n_{c_i} = \Theta(1)$, $i = 1, 2$, and if we restrict ourselves to $k_{c_1} \lesssim n/\log\ n$ and $k_{c_2}/n_{c_1} = \Theta(1)$, the average number of physical gates required for distillation will be

$$\bar{N}_{\text{dist}} = \frac{c_1 k_{c_1}^2 n n_{c_2} + c_2 k_{c_2}^2 n}{Y(p) n_{c_1} n_{c_2}} \sim O(n^2/\log\ n).$$

To sum up, an arbitrary logical Clifford circuit requires $\bar{N}_{\text{gate}} = O(n^2/\log\ n)$ physical gates on average. If one considers the family of large CSS block codes (e.g. quantum LDPC codes) with $k/n \sim \Theta(1)$, only $O(k)$ physical qubits and $O(k^2/\log k)$ physical gates are needed on average to implement any logical Clifford circuit, with off-line circuit depth $O\left(\max\{k, k_{c_1}, k_{c_2}\}\right)$ for ancilla preparation. These results suggest that the numbers of qubits and gates required for logical Clifford circuits have the same scaling as the physical level, when the physical error rate is sufficiently low and good classical LDPC codes are used.

Remark 1. One can achieve very high efficiency of resource utilization with our scheme in the following scenario: a small batch of finite Clifford circuits are repeatedly applied in an algorithm for a certain period of time. This is because our distillation process has high throughput—it can produce a large number of *identical* ancilla states (and thus generate the same number of identical Clifford circuits), and each state preparation is efficient. Several questions are raised here naturally: is there any useful quantum algorithm whose quantum circuits have such structure? In other words, is there an ansatz to adapt our FTQC architecture to design a circuit for some particular algorithm? Is there a good computation architecture to efficiently generate the large amount of identical ancilla states? These questions are all open and need further investigation. One promising candidate here is the Hamiltonian simulation algorithm for quantum simulations [36–41]. In that case, the target Hamiltonian changes slowly during the computation. In a certain period of time, the Trotter decomposition can be regarded as identical. Another candidate is the optimization type algorithms like quantum approximate optimization algorithm (QAOA) [42], which needs to rapidly apply Hadamards. On the other hand, to generate large mount ancilla states, single instruction multiple data (SIMD) style architecture [43, 44] that apply the same quantum gates on multiple qubits in the same region simultaneously, maybe particularly useful.

3.7. Summary

In conclusion, we provided two methods implementing logical Clifford circuits fault-tolerantly via constant number of steps of Knill or Steane syndrome measurement circuits *in situ*. Each method requires certain types of logical stabilizer states as ancillas. We showed that all ancilla states listed can be prepared fault-tolerantly through two-stage distillation. Our method transfers the complexity of Clifford circuits on logical level to the complexity of state preparation U_{prep} on physical level completely, which can be done offline. Surprisingly, if one chooses large block codes with encoding rate $k/n \sim \Theta(1)$, the overall numbers of qubits and physical gates required for a Clifford circuit on $[[n, k, d]]$ CSS circuit are around $O(k)$ and $O(k^2/\log k)$, respectively, which are independent of the distance of the underlying CSS code. This is the same scaling as a perfect Clifford circuit acting on k physical qubits.

Table 1. Resources required for a Clifford circuit on k logical qubits at the physical and logical levels for an $[[n, k, d]]$ CSS code. The physical operations counts all state preparation, gates, measurements including ancilla preparation and error correction. We assume $k/n \sim \Theta(1)$ for the large block codes and w is the maximum weight of the stabilizers. Sufficient parallelization are also considered to minimize the depth.

Method/average resource cost	#. Physical qubits	#. Physical operatation	#. Ancilla states	*in situ* depth	Off-line depth
Standard circuit model	$O(k)$	$O(k^2/\log k)$	N/A	$O(k)$	N/A
Circuit model FTQC (this paper)	$O(k)$	$O(k^2/\log k)$	$O(1)$	$O(1)$	$O(\max\{k, k_{c_1}, k_{c_2}\})$
Circuit model FTQC (as in reference [25, 26])	$O(\max\{k, wd\})$	$O(kwd)$	$O(kd)$	$O(kd)$	$O(kd)$
Circuit model FTQC (as in references [19–21])	$O(k)$	$O\left(\max\{kwd, k^2/\log\ k\}\right)$	$O(k^2/\log k)$	$O(k^2/\log k)$	$O(\max\{k, k_{c_1}, k_{c_2}\})$
one-way QC (as in reference [45])	$O(k^3/\log k)$	$O(k^3/\log k)$	N/A	$O(1)$	N/A
FT one-way QC (as in references [46, 47])	$O\left(k^3d^3/\log\ k\right)$	$O\left(k^3d^3/\log\ k\right)$	N/A	$O(1)$	N/A
Surface code (as in reference [17])	$O(kd^2)$	$O(k^3d^3/\log k)$	N/A	$O(k^2d/\log k)$	N/A

4. Discussion

In this section, we compare the method proposed in this paper with some other related FT protocols in the literature including one-way quantum computation. Then we estimate the numbers of physical qubits and gates required for each scheme. The results are summarized in table 1. Since different FTQC schemes have different working regions and performance, these results only provide a rough insight of resource scaling. We also discuss the potential improvements on ancilla state preparation for further reduce the overhead.

4.1. Related FTQC protocols

In this paper, we implement FT logical circuit teleportation via the single-shot Knill syndrome measurement protocol and FT ancilla distillation. It is worthwhile to compare this with the original teleportation-based FTQC in references [25, 26]. Rather than Knill syndrome measurement, Shor syndrome measurement [6] is used for error correction and ancilla state preparation. Our construction of logical Clifford circuits through a constant number of steps of teleportation is also possible in that scenario, where the ancilla state preparation again dominates the resource cost.

In references [25, 26], $O(1)$ logical ancilla states of size $O(n)$ is required. To prepare qualified logical ancilla states, one applies Shor syndrome measurement to measure n stabilizers, including stabilizer generators and logical operators. Each measurements needs one cat state. Each cat state consists $O(w)$ qubits, which takes $O(w)$ CNOTs to prepare, where w is the maximum weight of the stabilizers. A verification is also required after the raw preparation of each cat state, which also takes $O(w)$ CNOTs and rejects the states with probability around $O(p)$. Transversal CNOTs from the verified cat state to the code block are then applied to extract the eigenvalues of the stabilizers, which takes $O(w)$ CNOTs. To establish reliable eigenvalues for the stabilizers of an $[[n, k, d]]$ code, $O(d)$ rounds of Shor syndrome measurements and a majority vote are required for each stabilizer. Thus the overall number of physical gates required for state preparation is $O(\max\{nwd, n^2/\log n\})$ with depth $O(nd)$.

For large block codes with $k/n \sim \Theta(1)$, the number of physical gates required for ancilla state preparation is $O(\max\{kwd, k^2/\log k\})$ with depth $O(kd)$. Meanwhile, $O(1)$ logical ancilla states are required, which needs $O(k)$ physical ancilla qubits altogether. It also takes $O(kd)$ rounds of serial Shor syndrome measurements (since the stabilizers are in general highly overlapped) to do error correction on the data block, each round consumes a verified cat states. Hence, the depth for a logical Clifford circuit is $O(kd)$ and the same number of verified cat states are needed. Assuming the qubits supporting cat states are recycled after they are measured, one needs $O(wd)$ ancilla qubits for cat states throughout the process after parallelization. The number of all ancilla qubits is thus $O(\max\{k, wd\})$. This way of implementing logical Clifford circuits needs more physical gates when w is large and takes a much longer computation time for large k and d.

Our scheme also greatly simplifies the block-code based FTQC using Steane syndrome measurement in references [19–21]. There, logical Clifford gates are implemented one by one, and thus $O(k^2/\log k)$ different ancilla states of size $2n$ qubits are required and $O(k^2/\log k)$ rounds of Steane syndrome measurements are needed. With the same ancilla distillation protocol and ancilla recycling, one needs $k^4/(\log k)^2$ gates and $O(k)$ qubits for every single circuit on average for finite rate codes with $k/n \sim \Theta(1)$.

4.2. One-way quantum computing

For one-way QC, one initially prepares a cluster state consisting of a large number of qubits. Quantum information is then loaded onto the cluster and processed through single-qubit measurements on the cluster state substrate. It can be shown that all quantum circuits can be mapped to the form of one-way QC.

In general, one-qubit measurements are performed in a certain temporal order and in a spatial pattern of adaptive measurement bases based on previous measurement outcomes. Interestingly, for those qubits supporting Clifford circuits, no measurement bases have to be adjusted (i.e., those of which the operator X, Y or Z is measured). Thus, for any given quantum circuit, all of its Clifford gates can be realized *simultaneously* in the first round of single-qubit measurements, regardless of their space-time locations in the circuit [45], if a sufficiently large cluster state is provided to support the whole computation circuit. Specifically, for a cluster state in 2D, it requires $O(k^3/\log k)$ supporting cluster qubits and single-qubit measurements for an instantaneous Clifford circuit without error correction.

However, it is difficult to control errors if a cluster state large enough to support the entire quantum computation is used, since the computation might reach certain qubits only after a long time, so that these qubits would already suffer significant errors. This scheme is not FT. By contrast, if the computation is split, then the size of sub-circuits may be adjusted so that each of them can be performed within some constant time. The measured qubits are then recycled to entangle with the unmeasured qubits to form a new cluster for the next computation step. In this way, each cluster qubit is exposed to constant decoherence time before being measured and the error rate is bounded. In this case, FT one-way QC is possible [45]. Note that it is still possible to perform a Clifford circuit during the computation in one time step, if $O(k^3/\log k)$ qubits are provided at the same time, but it is no longer possible to finish all the Clifford gates in the computation in a single time step.

To complete the discussion, here we consider FT one-way QC in 3D lattice as in references [46, 47]. The Clifford circuits are performed through single-qubit measurements in the Z basis to create topologically entangled defects in the 3D lattice. The remaining qubits are measured in the X basis to provide syndrome information for 3D topological error correction [48]. For k encoded qubits in [47] with distance d boundary surface codes, the number of cluster qubits needed in a single 2D slice is $O(kd^2)$ and it requires $O(k^2/\log k)$ slices for an arbitrary Clifford circuit in the worst case. Thus, it takes the volume of a cluster state comprising $O(k^3d^3/\log k)$ qubits and the same number of single-qubit measurements. As a comparison, the variants of FT one-way QC in 2D based on the surface code [17] need kd^2 physical qubits for encoding and each logical CNOT gate takes $O(d)$ time steps.

In conclusion, even though one-way QC can in principle implement the Clifford gates of a circuit in a single time step, it requires many more physical qubits, whether it is implemented in a FT manner or not. It is worth noting that the FT one-way QC and its 2D variants require only local operation, which is a great practical advantage, since the codes considered in our scheme are highly non-local in general. However, our results suggest the potential for huge resource reduction for FTQC if non-local operations are allowed.

4.3. More efficient ancilla state preparation

The distillation protocol mentioned in this paper is basically the same as the one in reference [30]. The main difference is that the ancilla states distilled here can generate a whole circuit rather than a single gate on the data code block. These will cause an extra complexity on U_{prep} stage in figure 4 up to $O(k^2/\log k)$ gates with an extra depth $O(k)$. Meanwhile, the overall number of gates and depth for the whole distillation protocol (U_{prep}, U_{enc} and U_{dist} combined) also scale as $O(k^2/\log k)$ and $O(k)$. Note that the extra depth of ancilla preparation is negligible if they are produced in a pipeline manner. Thus, in the worst case, it will cause a constant decrease of distillation quality. But in practice, U_{prep} may only take a small portion of the whole protocol. On the other hand, since we generate a circuit rather than a single gate at one time, it will reduce the quality requirement of the output ancilla states to support FTQC, and hence the error rate requirement for each operation as well. The overall effect of extra preparation complexity on distillation needs further exploration.

The two-stage distillation protocol gives $O(1)$ yield rate on average. However, the number of input ancilla blocks required simultaneously is $O(n_{c_1}n_{c_2})$, which can be huge in practice. Consequently, the distillation circuit is still relatively complicated and error can occur in many positions. As a result, the typical acceptable error rate (or threshold for distillation) is less than 10^{-4} [30], which is challenging with current technologies like superconducting qubits [49]. Meanwhile, in many cases, one does not need as many as $O(k_{c_1}k_{c_2})$ identical ancilla states to generate that large number of the same Clifford circuits.

There are two ways to further simplify the distillation process and reduce overall number of qubits: in stead of two-stage distillation, one can filter out one type of correlated errors at the beginning by single-round stabilizer measurements with Steane Latin rectangle method [34], which takes advantage of the fact that only a small set of correlated errors needs to be removed according to the degeneracy of quantum code. Then we remove the other type of correlated errors through distillation. The depth of such preparation is also $O(\max\{k, k_c\})$. Here, only $O(n_c)$ input code blocks are required simultaneously and it

generates $O(k_c)$ identical output states. It not only reduces the complexity of preparation circuit but also gives more flexibility. The second method is to take advantage of the symmetry of the underlying CSS codes: the qubits of different code blocks are permuted in different ways after raw preparation (though finding such permutation may be challenging), so that the correlation of errors between code blocks can be suppressed [50]. Consequently, it may require less input code blocks for distillation. In principle, these two methods can also be combined together and their effect needs further investigation.

Acknowledgments

The funding support from the National Research Foundation and Ministry of Education, Singapore, is acknowledged. This work is also supported by the National Research Foundation of Singapore and Yale-NUS College (through Grant Number IG14-LR001 and a startup Grant). CYL was supported by the Ministry of Science and Technology, Taiwan under Grant MOST108-2636-E-009-004. TAB was supported by NSF Grants No. CCF-1421078 and No. MPS-1719778, and by an IBM Einstein Fellowship at the Institute for Advanced Study.

Appendix A. Proof of lemma 2

Proof. Let l'_j denote the jth row vector of L' and c_p be the pth column vector of $(I_n\ I_n)^t$. Equation (24) is equivalent to

$$l'_j c_p = \delta_{jp}, \qquad 1 \leqslant j, p \leqslant n, \tag{A1}$$

where δ is the Kroneker delta function.

Let l_j denote the jth row vector of L. Obviously, $l_1 = (1, 0, \ldots, 0)$, satisfying $l_1 c_p = \delta_{1p}$. Let $l'_1 = l_1$.

It is easy to see that $l_j c_p = 0$ for $p > j$, since L_1 is a lower triangular matrix. With all the diagonal elements of L_1 being 0, one has

$$l_j c_j = 1. \tag{A2}$$

Define the set $\mathscr{I}_j = \{p \mid l_j c_p = 1, p < j\}$. For $j = 2, \ldots, n$, let

$$l'_j = l_j + \sum_{p \in \mathscr{I}_j} l'_p. \tag{A3}$$

We also define a matrix $L'^{(j)}$ that contains the rows $l'_1, \ldots, l'_j$:

$$L'^{(j)} = \begin{pmatrix} l'_1 \\ \vdots \\ l'_j \end{pmatrix}.$$

Since L_1 is lower triangular, and the summation of l_p in equation (A3) only counts the terms with $p < j$, $L'^{(j)}$ can be written as

$$L'^{(j)} = \left(L_2^{(j)}\ L_3^{(j)}\right),$$

where $L_2^{(j)}$ and $L_3^{(j)}$ are also lower triangular matrices. Eventually, we have $L_2 = L_2^{(n)}$ and $L_3 = L_3^{(n)}$.

It remains to prove equation (A1). We prove this by induction. For $j = 2$, if $l_2 c_1 = 1$, one has $l'_2 = l_2 + l_1$. Thus $l'_2 c_1 = 0$ and $l'_2 c_2 = 1$, since $l_1 c_1 = 1$ and $l_1 c_2 = 0$. Also, $l'_2 c_p = 0$ for $p > 2$ since $L_2'^{(2)}$ and $L_3'^{(2)}$ are lower triangular matrices. So $l'_2 c_p = \delta_{2p}$ holds for $1 \leqslant p \leqslant n$.

Now assume $l'_1 c_p = \delta_{1p}, \ldots, l'_j c_p = \delta_{jp}$ holds. Then

$$l'_{j+1} c_q = l_{j+1} c_q + \sum_{p \in \mathscr{I}_{j+1}} l'_p c_q.$$

Consider $q < j + 1$ first. If $l_{j+1} c_q = 1$, then $q \in \mathscr{I}_{j+1}$ and

$$\sum_{p \in \mathscr{I}_{j+1}} l'_p c_q = \sum_{p \in \mathscr{I}_{j+1}} \delta_{pq} = 1.$$

Then $l'_{j+1} c_q = 0$. If $l_{j+1} c_q = 0$, then $q \notin \mathscr{I}_{j+1}$ and $\sum_{p \in \mathscr{I}_{j+1}} l'_p c_q = 0$. Again, $l'_{j+1} c_q = 0$. When $q = j + 1$,

$l'_{j+1}c_{j+1} = l_{j+1}c_{j+1} = 1$ by equation (A2). For $q > j+1$, since $L_2^{(j+1)}$ and $L_3^{(j+1)}$ are both lower triangular, $l'_{j+1}c_q = 0$. Thus, $l'_j c_p = \delta_{jp}$ holds for $1 \leqslant j, p \leqslant n$.

□

ORCID iDs

Yi-Cong Zheng https://orcid.org/0000-0003-2680-8000
Ching-Yi Lai https://orcid.org/0000-0003-1970-8167
Todd A Brun https://orcid.org/0000-0002-8807-3495
Leong-Chuan Kwek https://orcid.org/0000-0002-0879-0591

References

[1] Shor P W 1995 *Phys. Rev.* A **52** R2493
[2] Steane A M 1996 *Phys. Rev. Lett.* **77** 793
[3] Calderbank A R and Shor P W 1996 *Phys. Rev.* A **54** 1098
[4] Gaitan F 2008 *Quantum Error Correction and Fault Tolerant Quantum Computing* (Boca Raton, FL: CRC Press)
[5] Lidar D and Brun T 2013 *Quantum Error Correction* (Cambridge: Cambridge University Press)
[6] Shor P 1996 *Proc. 37th Annual Symp. on Foundations of Computer Science* (Los Alamitos, CA: IEEE Computer Society Press) p 56
[7] Aharonov D and Ben-Or M 1997 *Proc. 29th Annual ACM Symp. on the Theory of Computation* (New York: ACM) p 176
[8] Gottesman D 1997 Stabilizer codes and quantum error correction *PhD Thesis* California Institute of Technology arXiv:quant-ph/9705052
[9] Kitaev A 2003 *Ann. Phys.* **303** 2
[10] DiVincenzo D P and Shor P W 1996 *Phys. Rev. Lett.* **77** 3260
[11] Knill E 2005 *Nature* **434** 39
[12] Aharonov D, Kitaev A and Preskill J 2006 *Phys. Rev. Lett.* **96** 050504
[13] Terhal B M and Burkard G 2005 *Phys. Rev.* A **71** 012336
[14] Aliferis P, Gottesman D and Preskill J 2006 *Quantum Inf. Comput.* **6** 97
[15] Cross A W, Divincenzo D P and Terhal B M 2009 *Quantum Inf. Comput.* **9** 0541
[16] Aliferis P, Gottesman D and Preskill J 2008 *Quantum Inf. Comput.* **8** 181
[17] Fowler A G, Mariantoni M, Martinis J M and Cleland A N 2012 *Phys. Rev.* A **86** 032324
[18] Bombin H and Martin-Delgado M 2006 *Phys. Rev. Lett.* **97** 180501
[19] Steane A M 1999 *Nature* **399** 124
[20] Steane A M and Ibinson B 2005 *Phys. Rev.* A **72** 052335
[21] Brun T A, Zheng Y-C, Hsu K-C, Job J and Lai C-Y 2015 arXiv:1504.03913
[22] Steane A M 2003 *Phys. Rev.* A **68** 042322
[23] Gottesman D 2014 *Quantum Inf. Comput.* **14** 1338
[24] Steane A M 1997 *Phys. Rev. Lett.* **78** 2252
[25] Gottesman D and Chuang I 1999 *Nature* **402** 390
[26] Zhou X, Leung D W and Chuang I L 2000 *Phys. Rev.* A **62** 052316
[27] Patel K N, Markov I L and Hayes J P 2008 *Quantum Inf. Comput.* **8** 0282
[28] Aaronson S and Gottesman D 2004 *Phys. Rev.* A **70** 052328
[29] Lai C-Y, Zheng Y-C and Brun T A 2017 *Phys. Rev.* A **95** 032339
[30] Zheng Y-C, Lai C-Y and Brun T A 2018 *Phys. Rev.* A **97** 032331
[31] Such extra layer has depth $O(1)$. Throughout the paper, Pauli gates are assumed to be free and can be directly applied to qubits. This is also true in FTQC using stabilizer codes, where logical Pauli operators are easy to realize.
[32] Maslov D and Roetteler M 2018 *IEEE Trans. Inf. Theory* **64** 4729
[33] Chamberland C and Ronagh P 2018 *Quantum Sci. Technol.* **3** 044002
[34] Steane A M 2002 arXiv:quant-ph/0202036
[35] MacKay D J C 2003 *Information Theory, Inference and Learning Algorithms* (Cambridge: Cambridge University Press)
[36] Lloyd S 1996 *Science* **273** 1073
[37] Aspuru-Guzik A, Dutoi A D, Love P J and Head-Gordon M 2005 *Science* **309** 1704
[38] Wecker D, Bauer B, Clark B K, Hastings M B and Troyer M 2014 *Phys. Rev.* A **90** 022305
[39] Hastings M B, Wecker D, Bauer B and Troyer M 2015 *Quantum Inf. Comput.* **15** 1
[40] Poulin D, Hastings M B, Wecker D, Wiebe N, Doberty A C and Troyer M 2015 *Quantum Inf. Comput.* **15** 361
[41] Bauer B, Bravyi S, Motta M and Chan G K 2020 arXiv:2001.03685
[42] Farhi E, Goldstone J and Gutmann S 2014 arXiv:1411.4028
[43] Heckey J, Patil S, JavadiAbhari A, Holmes A, Kudrow D, Brown K R, Franklin D, Chong F T and Martonosi M 2015 *Proc. 29th Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS)* (New York: ACM) pp 445–56
[44] Risque R and Jog A 2016 *IEEE Int. Symp. on Workload Characterization (IISWC)* (Los Alamitos, CA: IEEE Computer Society Press) pp 1–9
[45] Raussendorf R, Browne D E and Briegel H J 2003 *Phys. Rev.* A **68** 022312
[46] Raussendorf R, Harrington J and Goyal K 2006 *Ann. Phys.* **321** 2242
[47] Raussendorf R, Harrington J and Goyal K 2007 *New J. Phys.* **9** 199
[48] Raussendorf R, Bravyi S and Harrington J 2005 *Phys. Rev.* A **71** 062313
[49] Arute F *et al* 2019 *Nature* **574** 505
[50] Paetznick A and Reichardt B W 2012 *Quantum Inf. Comput.* **12** 1034