

Accelerating Fault-Tolerant Quantum Computation with Good Quantum Low-Density Parity-Check Codes

Guo Zhang, Yuanye Zhu, and Ying Li^{*}

Graduate School of China Academy of Engineering Physics, Beijing 100193, China



(Received 17 November 2025; revised 5 April 2026; accepted 3 June 2026; published 8 July 2026)

We propose a fault-tolerant quantum computation scheme that is broadly applicable to quantum low-density parity-check (qLDPC) codes. The scheme achieves constant qubit overhead and a time overhead of $O(d^{a+o(1)})$ for any $[[n, k, d]]$ qLDPC code with constant encoding rate and distance $d = \Omega(n^{1/a})$. For good qLDPC codes, the time overhead is minimized and reaches $O(d^{1+o(1)})$. In contrast, code surgery based on gauging measurement and brute-force branching requires a time overhead of $O(dw^{1+o(1)})$, where $d \leq w \leq n$. Thus, our scheme is asymptotically faster for all codes with $a < 2$. This speedup is achieved by developing techniques that enable parallelized code surgery under constant qubit overhead and leverage classical locally testable codes for efficient resource state preparation. These results establish a new paradigm for accelerating fault-tolerant quantum computation on qLDPC codes, while maintaining low overhead and broad applicability.

DOI: 10.1103/3fj5-gn1h

I. INTRODUCTION

Quantum error correction is essential to realizing large-scale quantum computation [1–4]. According to the fault-tolerance theorem, when the physical error rate is below a constant threshold, arbitrarily high computational accuracy can be achieved by choosing a quantum error correction code with sufficiently large code distance d [5,6]. However, fault-tolerant quantum computation (FTQC) inevitably introduces additional resource overheads: each logical qubit requires multiple physical qubits for encoding and for supporting its computational operations, referred to as *qubit overhead*; each round of logical-qubit operations must be implemented through multiple rounds of physical-qubit operations, referred to as *time overhead*. For instance, when computation is performed using the surface code via lattice surgery, the qubit overhead scales as $O(d^2)$, while the time overhead scales as $O(d)$ [7,8].

Substantial effort has been devoted to reducing the resource overhead of FTQC. A major advance in this direction is the development of low-overhead quantum low-density parity-check (qLDPC) codes, such as hypergraph product (HGP) codes, which achieve a constant encoding rate [9]. This property ensures that the logical

qubit number k scales linearly with the physical qubit number n , i.e. $k = \Theta(n)$. Recent breakthroughs have led to the discovery of good qLDPC codes, which further improve the code distance from $d = \Theta(n^{1/2})$ to $d = \Theta(n)$ while maintaining a constant encoding rate [10–14].

To enable logical computational operations on qLDPC codes, two main strategies have been developed: the concatenated codes combined with gate teleportation (CC + GT) approach [15–18] and the code surgery approach (a generalization of lattice surgery) [8,19–29]. For general qLDPC codes, the most efficient known method is the gauging measurement combined with brute-force branching (GM + BFB)—a form of code surgery that achieves constant qubit overhead and a time overhead scaling as $O(d^{2+o(1)})$ [assuming the logical operator weight $w = O(d)$] [22–24,28]. A lower time overhead of $O(d^{1+o(1)})$ can be attained using a refined variant of the CC + GT scheme [18]; however, its applicability is restricted to (almost) good quantum locally testable codes (qLTCs)—also known as c^3 -qLTCs—which, beyond having constant encoding rate and constant relative distance, must also possess constant soundness [18,30].

In this work, we propose a new scheme applicable to general qLDPC codes, achieving lower resource overhead than the GM + BFB protocol and extending the $O(d^{1+o(1)})$ time overhead to broader code families. For qLDPC codes with a constant encoding rate and code distance $d = \Omega(n^{1/a})$, our scheme attains a time overhead of $O(d^{a+o(1)})$ while maintaining constant qubit overhead. Compared with GM + BFB, our method achieves a smaller time overhead for all constant-rate qLDPC codes

^{*}Contact author: yli@giscaep.ac.cn

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](#). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

whose relative distance exceeds that of HGP codes (i.e., $a < 2$). In particular, for good qLDPC codes, the time overhead is further reduced to $O(d^{1+o(1)})$. Although this minimal overhead matches the best performance of the CC + GT approach, our scheme achieves it under weaker conditions—requiring only a constant encoding rate and a constant relative distance.

Our scheme integrates code surgery with gate teleportation by performing the parity-check measurements required in code surgery through gate teleportation. To minimize resource overhead, we introduce two key techniques: parallelized code surgery (PCS) and locally testable state preparation (LTSP). In PCS, multiple code blocks share a common ancilla system (a set of mutually coupled qubits) to implement computational operations, thereby reducing the qubit overhead of code surgery to a constant. In LTSP, the resource states required for gate teleportation are prepared with constant qubit and time overhead, employing a classical locally testable code (LTC) that requires only two constants—constant encoding rate and constant soundness [13,31,32]. Leveraging local testability, we can effectively control measurement errors in parity-check measurements: when performing code surgery with resource states prepared via LTSP, only a single round of parity-check measurements is needed, reducing the time overhead of code surgery from $O(d)$ to $O(1)$. The introduction of PCS and LTSP inevitably leads to serialization of logical operations, resulting in a final time overhead of $O(d^{a+o(1)})$. As general overhead-reduction techniques, PCS and LTSP not only improve the asymptotic scaling of FTQC but also offer practical means to lower overhead in near-term, finite-size implementations.

II. RELATED WORKS

Two general approaches have been developed for fault-tolerant quantum computation with qLDPC codes: CC + GT and code surgery.

Gate teleportation implements logical gates by consuming pre-prepared resource states [15,17,18]. Gottesman proposed using gate teleportation to operate on logical qubits encoded in qLDPC codes, with resource states generated via concatenated codes [15]. This approach is made possible by the existence of well-established fault-tolerant quantum computation protocols in certain concatenated codes, such as the concatenated Steane codes [16]. In Gottesman's approach, the resource state is prepared by simulating its preparation circuit using fault-tolerant operations of the concatenated code, producing an encoded resource state. This is then decoded using the decoding circuit of the concatenated code to yield the unencoded resource state needed for gate teleportation. Since the simulation is fault-tolerant and the decoding circuit introduces only local stochastic errors at an asymptotically constant rate, the final resource state can

be made sufficiently accurate to enable reliable gate teleportation.

Several variants of the CC + GT approach have been developed, all achieving constant qubit overhead [17,18]. Their primary difference lies in the time overhead. Consider a quantum computer consisting of M memory blocks, each encoded using an $[[n, k, d]]$ qLDPC code. To implement a quantum circuit of size $|C|$ on the $K = Mk$ logical qubits with a target failure rate ϵ , the logical error rate per operation must be reduced to $p_L = O(\epsilon/|C|)$. Achieving this requires the concatenated code to have sufficiently large distance, ensuring that the encoded resource states are prepared with error rate $O(p_L)$. This requirement imposes a significant qubit cost on resource state production. To maintain a constant total qubit overhead, resource states are provided to only a subset of memory blocks at any given time. As a result, each layer of logical operations must be executed over multiple time steps. The overall time overhead is determined by both the cost of preparing resource states and the complexity due to serialization. In Gottesman's original protocol, and its subsequent improvement [15,16], this overhead scales as $O(\text{poly}(K))$. A more recent advancement [17] reduces the time overhead to

$$O\left(\log^{2a+\gamma_1+\gamma_2}\left(\frac{|C|}{\epsilon}\right)\right) = O(d^{2a+\gamma_1+\gamma_2}),$$

where γ_1 and γ_2 are parameters characterizing the concatenated code; note that $|C|/\epsilon = O(\text{poly}(K))$ in typical scenarios. Here, we have re-expressed the time overhead by assuming an exponential suppression of the logical error rate with code distance, i.e., $p_L = e^{-\Omega(d)}$, allowing for a direct comparison with code-surgery protocols.

A further breakthrough was reported in Ref. [18], which introduced state distillation protocols capable of achieving an almost constant spacetime overhead. Before distillation, a concatenated code with modest distance is used to produce resource states with a sufficiently small constant error rate. These are subsequently distilled to achieve the target error rate of $O(p_L)$. Due to the efficiency of the distillation process, the time cost of preparing resource states becomes negligible, making serialization cost the dominant contribution to the overall time overhead. Consequently, the total time overhead is reduced to

$$O\left(\log^{1+o(1)}\left(\frac{|C|}{\epsilon}\right)\right) = O(d^{1+o(1)}).$$

However, since the efficient distillation relies on local testability, achieving this $O(d^{1+o(1)})$ overhead requires the underlying code to be an (almost) good qLTC, characterized by triple constants: constant encoding rate, constant relative distance, and constant soundness.

TABLE I. Qubit and time overheads in fault-tolerant schemes using quantum low-density parity-check (qLDPC) codes. The exponent b in the overhead expression $O(d^{b+o(1)})$ is referred to as the overhead exponent. Techniques developed for code surgery include gauging measurement (GM) [23,24], brute-force branching (BFB) [22], and devised sticking (DS) [22]. The overhead estimates for the GM + BFB protocol are predicated on the assumption that $w = O(d)$; any departure from this regime only increases the overhead exponents. In contrast, our protocol and other methods in the table do not rely on this assumption. Typical protocols based on gate teleportation (GT) utilize concatenated codes (CC) [15,17,18]. The overhead results for the polylog CC + GT [17] protocol assume the use of quantum codes possessing single-shot properties. Lattice surgery (LS) [33] is a method for quantum computation on the surface code (SC). This work introduces two new techniques for reducing time overhead: parallelized code surgery (PCS) and locally testable state preparation (LTSP).

Code	Protocol	Qubit overhead exponent	Time overhead exponent	Refs.
qLDPC codes with $k = \Theta(n)$ and $d = \Omega(n^{1/a})$	GM + BFB	0	≥ 2	[25,28]
	DS	1	1	[22]
	polylog CC + GT	0	$\geq 2a$	[17]
	PCS + LTSP + GT	0	a	This work
Good qLTCs	log CC + GT	0	1	[18]
	PCS + LTSP + GT	0	1	This work
Surface codes	LS	2	1	[33]

Code surgery is a generalization of lattice surgery, originally developed for implementing logical operations on surface codes [8,19,20,22–26,28,29]. In code surgery, the original code is temporarily deformed by coupling it to an ancilla system, enabling the measurement of one or more selected logical Pauli operators. These logical measurements can then be used to implement Clifford gates. Code surgery for qLDPC codes was first proposed in Ref. [19], known as the CKBB protocol, named after its authors. In this protocol, a logical operator is measured using an ancilla system of $O(dw)$ physical qubits, where w is the weight of the logical operator. Since $w \geq d$, the required ancilla size is comparable to that of a surface code. Subsequent works introduced gauging measurements, which reduced the ancilla size to $O(w^{1+o(1)})$ physical qubits [20,23,24]. Gauging measurements achieve this by transforming the relevant checks into those of a repetition code and constructing the ancilla system using an expander graph, thereby achieving robustness against certain errors that the CKBB protocol mitigates using additional qubits.

Code surgery faces an inherent limitation in time efficiency: the parity-check measurements of the deformed code must be repeated for $\Theta(d)$ rounds to reliably correct measurement errors. As a result, lattice surgery on surface codes incurs a time overhead of $O(d)$. In the CKBB and gauging measurement protocols, when logical operators are measured sequentially—one in each block at a time—the total time overhead for measuring $\Theta(k)$ logical operators (a layer of operations on k logical qubits) is $O(kd) = O(d^{1+a})$.

To improve time efficiency, general techniques have been developed to construct deformed codes that support the simultaneous measurement of multiple logical operators. Two main approaches are devised sticking and brute-force branching [22]. In devised sticking, an arbitrary

number of logical operators can be measured in parallel using a single ancilla system of $O(nd)$ physical qubits [34]. This approach has a time overhead of $O(d)$ and a qubit overhead of $O(d)$, assuming the underlying code has constant encoding rate. In brute-force branching, an intermediate deformed code is constructed to distribute the logical operators across disjoint supports, enabling their measurements to be performed in parallel. A protocol that combines gauging measurements with brute-force branching (GM + BFB) [28] achieves a time overhead of $O(dw^{1+o(1)})$ while reducing the qubit overhead to constant.

Note that code surgery operations alone can only implement Clifford logical gates. A parallelized magic state injection protocol enables the preparation of magic states with (almost) the same resource scaling as code surgery [25].

In addition to typical CC + GT and code surgery protocols, Ref. [17] proposes preparing gate-teleportation resource states using surface codes (or alternatives such as color codes, rather than concatenated codes), and then implementing gate teleportation on high-rate qLDPC code blocks via code surgery.

III. SCHEME

We now sketch our scheme for FTQC. In our scheme, we adopt the code-surgery approach and propose to implement code surgery via gate teleportation. We introduce two key techniques to minimize qubit and time overheads: PCS and LTSP; see Fig. 1 and Sec. IV.

In the PCS framework, we simultaneously operate on multiple logical qubits while maintaining a strictly constant qubit overhead. This is achieved by coupling an ancilla system to the memory code (which encodes the data logical qubits) to form a deformed qLDPC code. The ancilla system is constructed as a hypergraph product of

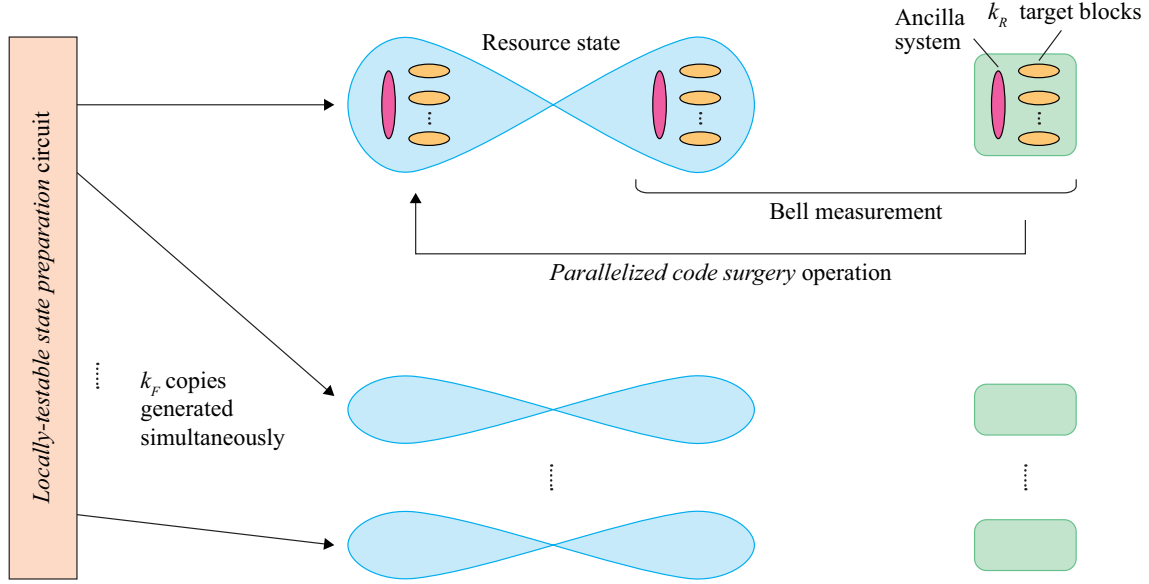


FIG. 1. Schematic illustration of our scheme. Each execution of the locally testable state preparation circuit produces k_F copies of the resource state, which are then used in gate teleportation to implement parallelized code surgery (PCS). PCS applies logical operations simultaneously on k_R target code blocks, using a single ancilla system shared across the k_R blocks. The locally testable state preparation circuit has almost constant qubit and time overhead, while gate-teleportation-based PCS also maintains constant qubit and time overhead.

two linear codes. The first linear code determines the specific logical operations to be applied, while the second linear code—referred to as the R code—is a classical LDPC code with a constant encoding rate. A central finding of this work is that the codewords of the R code establish parallel channels for logical operations: if the R code has a logical dimension k_R , a single ancilla system can be coupled to k_R distinct memory-code blocks, enabling parallel logical operations across all blocks. Since both the memory code and the R code possess constant encoding rates, the ancilla system has a total size of $O(kk_R)$, where k is the number of logical qubits per block. This approach permits the simultaneous execution of logical operations on $O(kk_R)$ logical qubits distributed across the blocks, thereby achieving high parallelism with constant qubit overhead.

To ensure high time efficiency, we propose the LTSP scheme to circumvent the temporal bottlenecks of conventional code surgery. Standard implementations typically require $O(d)$ rounds of repeated stabilizer measurements to achieve fault tolerance against measurement errors. We overcome this requirement by introducing *single-shot* properties. Rather than requiring the memory code or the deformed code itself to be single-shot [35,36], we demonstrate that single-shot measurements can be effectively realized by utilizing an independent single-shot code within some state preparation circuits. Specifically, we implement stabilizer measurements in a single round via gate teleportation. By utilizing a carefully prepared gate-teleportation resource state with bounded error weight, a

single measurement round is sufficient to reliably extract the stabilizer information. This approach shifts the single-shot requirement from the code surgery circuit to the circuit preparing resource states. To control errors on the resource state, we encode each qubit in the preparation circuit as a logical qubit of an F code. By selecting the F code to be a classical LTC with a constant encoding rate and constant soundness, we can prepare resource states with (almost) constant qubit overhead in constant depth.

By combining PCS and LTSP, logical operations can be implemented in parallel with constant qubit overhead and circuit depth. However, the architecture naturally imposes a constraint, requiring identical operations to be performed across multiple blocks simultaneously. For general heterogeneous circuits, this constraint necessitates the *serialization* of logical operations, introducing a time overhead factor of $O(k) = O(d^a)$. Additional polylogarithmic factors $O(d^{o(1)})$ arise from the requirements for robustness against stochastic noise and magic state distillation. Ultimately, through a time-space trade-off, we achieve a strictly constant qubit overhead while maintaining a total time complexity within the $O(d^{a+o(1)})$ class. These results are summarized in the following theorem.

Theorem 1. Consider a classical-input/classical-output quantum circuit C composed of Clifford and T gates, with size $|C| \leq WD$, where W denotes the width and $D = \text{poly}(W)$ denotes the depth. There exists a constant $\epsilon_* \in (0, 1)$ such that for any $\epsilon_L \in (0, 1)$, one can efficiently construct a fault-tolerant version C_{FT} with width W_{FT} and depth D_{FT} satisfying

$$\frac{W_{\text{FT}}}{W} = O_{W \rightarrow \infty}(1), \quad (1)$$

$$\frac{D_{\text{FT}}}{D} = O_{W \rightarrow \infty}\left(\log^{a+o(1)}\left(\frac{|C|}{\epsilon_L}\right)\right), \quad (2)$$

where a is a constant associated with the memory code block. If C_{FT} operates under a local stochastic noise model and physical error rate $p \leq \epsilon_*$ at every location, then C_{FT} simulates C up to output error ϵ_L , in the sense that their output distributions differ by at most ϵ_L .

A detailed proof is provided in Appendix I.

Remark. The theorem assumes a physical qubit network with all-to-all connectivity, where two-qubit gates can be applied between any pair of physical qubits. In this regime, the logical qubit network inherits this all-to-all connectivity, allowing the fault-tolerant circuit C_{FT} to implement any logical circuit C with the stated overheads. Notably, our protocol can be adapted to much more stringent architectural constraints. Specifically, on a physical network with almost constant vertex degree $D_P = O(d^{o(1)})$, we still achieve an almost constant qubit overhead and a time overhead of $O(d^{a+o(1)})$. In this restricted setting, however, the logical qubit network is characterized by a constant vertex degree $D_L = \Theta(1)$. For further details, see Appendix E.

A. Codes

Our scheme employs a set of error correction codes, as listed in Table II. A qLDPC code, referred to as the memory code, is used to encode logical qubits. We perform operations on these logical qubits via code surgery, which requires the construction of corresponding

deformed codes. A classical LDPC code, called the R code, is used in constructing the deformed codes, replacing the role of the repetition code in the CKBB protocol. Another classical LDPC code, referred to as the F code, serves as the core component of the resource-state factory, which produces the resource states required for gate teleportation. The requirements for these codes, along with representative examples, are presented in Table II. Additionally, our scheme also employs a surface code and a color code for various auxiliary tasks.

B. Quantum computer

As depicted in Fig. 2, the quantum computer comprises three main components: the memory, the resource-state factory, and the magic-state factory. The memory consists of multiple memory-code blocks. M of the blocks store the logical information, giving a total of $K = Mk$ data logical qubits. Here, k is the logical dimension of the memory code. The remaining memory-code blocks serve as ancillas. The resource-state factory produces the resource states required for gate teleportation, while the magic-state factory prepares encoded magic states.

C. Logical measurements—Parallelized code surgery

We achieve universal quantum computation using measurements of logical Pauli operators, supplemented with encoded magic states. Table III summarizes all logical measurements used in our scheme. These measurements are realized through PCS; see Fig. 3 for the circuit. Each logical measurement corresponds to a specifically tailored deformed code. The construction of the deformed codes is detailed in Sec. IV A. A key feature of PCS is that the deformed codes require only a constant qubit overhead.

TABLE II. Quantum (Q) and classical (C) codes utilized in our protocol. The quantum computer is structured into multiple blocks, each encoded in a qLDPC code, called memory code. The quantum computation is performed on logical qubits within these memory-code blocks via code surgery, where each operation corresponds to a specific deformed code. The deformed code is generated utilizing a classical linear code, called R code. To implement the parity-check measurements of a deformed code (or the memory code), we prepare the corresponding gate-teleportation resource states using another classical linear code, referred to as the F code. This code has to be a locally testable code, and s denotes its soundness. In examples, ϵ denotes an arbitrary positive number. Additionally, we employ a low-distance surface code to suppress errors during resource state preparation and to inject T -gate magic states, and a high-distance color code to prepare S -gate magic states. We use d_S and d_C to denote distances of the surface and color codes, respectively.

Codes	Parameters	Examples
Memory code (Q)	$k = \Theta(n)$ and $d = \Omega(n^{1/a})$	Hypergraph product, $a = 2$ [9] Quantum Tanner, $a = 1$ [37] Expander LP, $a = 1 + \epsilon$ [13] DLV, $a = 1 + \epsilon$ [30] LH, $a = 1$ [38] DHLV, $a = 1$ [14]
R code (C)	$k_R = \Theta(n_R)$ and $d_R = \Omega(n_R^{1/a_R})$	Expander, $a_R = 1$ [39] Tanner, $a_R = 1$ [40]
F code (C)	$k_F = \Theta(n_F)$, $d_F = \Omega(n_F^{1/a_F})$, and $s = \Theta(1)$	PK, $a_F = 1$ [13] Lossless-expander LTCs, $a_F = 1$ [32]

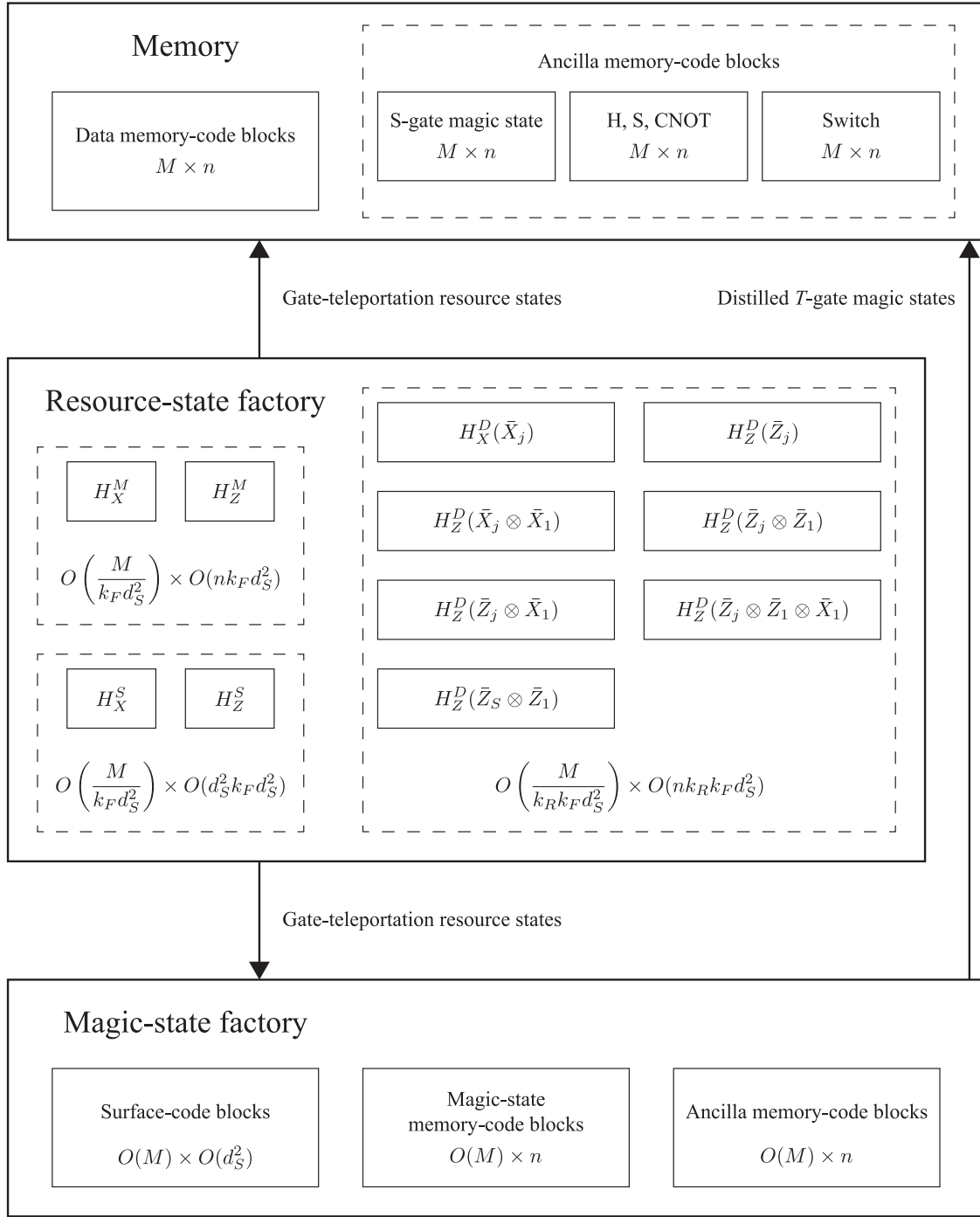


FIG. 2. Quantum computer and allocation of physical qubits. In the memory, the memory-code blocks are divided into four groups according to their functions. Each group contains M blocks, and each block occupies n physical qubits. The resource-state factory produces resource states specified by the corresponding check matrices, which are organized into three families (indicated by the dashed boxes). In each dashed box, the notation $O(A) \times O(B)$ denotes the total number of physical qubits used for generating resource states within that family: each type of resource state is prepared by a tailored locally testable state preparation circuit requiring $O(B)$ physical qubits, and $O(A)$ such circuits are run in parallel. In the magic-state factory, $O(M)$ surface-code blocks are used to inject T -gate magic states into $O(M)$ memory-code blocks, while an additional $O(M)$ ancillary memory-code blocks are employed for magic-state distillation.

TABLE III. List of logical measurements. $\bar{X}_j$ ($\bar{Z}_j$) denotes the X (Z) logical operator of the j th logical qubit in a memory-code block; $\bar{Z}_S$ is the Z logical operator of a distance- d_S surface-code block; and $\bar{Z}_C$ is the Z logical operator of a distance- d_C color-code block. A resource state is specified by the corresponding check matrix H : the state H is used to measure Z stabilizer operators represented by H . The notation $H \times m$ denotes m copies of the resource state. Measurements of X stabilizer operators are obtained by applying transversal Hadamard gates to rotate the basis. We use H_P^c with $c = M, D, S, C$ to represent the $P \in \{X, Z\}$ check matrix of the memory, deformed, surface, and color codes, respectively; and $H_Z^D(\sigma)$ denotes the Z check matrix of the deformed code tailored to the measurement of logical operator(s) σ . Each set of resource states enables the simultaneous application of the same logical measurement on k_R identical systems (target blocks).

Logical operator (set) σ	Resource states
$\bar{X}_j$	$H_Z^D(\bar{X}_j), H_Z^M \times k_R$
$\bar{Z}_j$	$H_Z^D(\bar{Z}_j), H_X^M \times k_R$
$\bar{X}_j \otimes \bar{X}_1$	$H_Z^D(\bar{X}_j \otimes \bar{X}_1), H_Z^M \times 2k_R$
$\bar{Z}_j \otimes \bar{Z}_1$	$H_Z^D(\bar{Z}_j \otimes \bar{Z}_1), H_X^M \times 2k_R$
$\bar{Z}_j \otimes \bar{X}_1$	$H_Z^D(\bar{Z}_j \otimes \bar{X}_1), H_X^M \times k_R, H_Z^M \times k_R$
$\bar{Z}_j \otimes \bar{Z}_1 \otimes \bar{X}_1$	$H_Z^D(\bar{Z}_j \otimes \bar{Z}_1 \otimes \bar{X}_1), H_X^M \times 2k_R, H_Z^M \times 2k_R$
$\bar{Z}_S \otimes \bar{Z}_1$	$H_Z^D(\bar{Z}_S \otimes \bar{Z}_1), H_X^S \times k_R, H_X^M \times k_R$
$\bar{Z}_C \otimes \bar{Z}_1$	$H_Z^D(\bar{Z}_C \otimes \bar{Z}_1), H_X^C \times k_R, H_X^M \times k_R$

We refer to the collection of code blocks involved in a logical measurement as a target block. For example, the measurement $\bar{Z}_j \otimes \bar{Z}_1 \otimes \bar{X}_1$ acts on a target block composed of three memory-code blocks. In conventional code surgery, a logical measurement is performed on a single target block at a time. In contrast, PCS performs identical

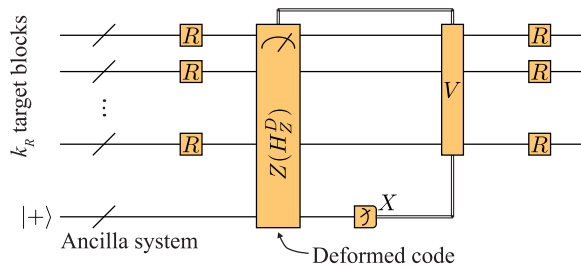


FIG. 3. Circuit for parallelized code surgery. The ancilla system is transversally initialized in the state $|+\rangle$ and measured in the X basis. The R gate rotates the logical measurement basis to Z : for each memory-code block on which the measured Pauli operator acts as X , a transversal Hadamard gate is applied. The central step of code surgery is a single round of parity-check measurements of the deformed code, specifically measuring Z stabilizer operators $Z(H_Z^D)$. Implemented via gate teleportation with a resource state from locally testable state preparation, this single round suffices to guarantee fault tolerance. The feedback gate V is a Pauli gate.

logical measurements simultaneously across k_R identical target blocks, where k_R is the logical dimension of the R code (see Fig. 1). In this context, “parallelized” refers specifically to the batching of identical operations applied to multiple blocks.

D. Gate teleportation—Locally-testable state preparation

The fundamental operations in error correction and code surgery are parity-check measurements. In our scheme, we perform parity-check measurements on various quantum codes: the memory code, surface code, color code, and deformed codes. In the standard CSS code formalism, each code corresponds to two sets of stabilizer operators— X -type and Z -type—represented by the corresponding check matrices H_X and H_Z . We measure the X and Z generators separately. For each of H_X and H_Z , the parity-check measurements are implemented via gate teleportation, each consuming a dedicated resource state; see Fig. 4 for the circuit.

We prepare the gate-teleportation resource states using LTSP. Each type of resource state is prepared by a tailored state-preparation circuit. Each execution of such a circuit generates k_F identical copies of the corresponding resource state, where k_F denotes the logical dimension of the F code; see Figs. 1 and 5. We note that each qubit in the circuit is encoded via the concatenation of the F code and a surface code to protect against both types of Pauli errors. The detailed construction of these circuits is provided in Sec. IV B.

A key feature of LTSP is that, by encoding with a constant-soundness LTC, the prepared resource states possess a crucial property: the parity-check measurements performed using these resource states are effectively free of measurement errors. Consequently, only a single round

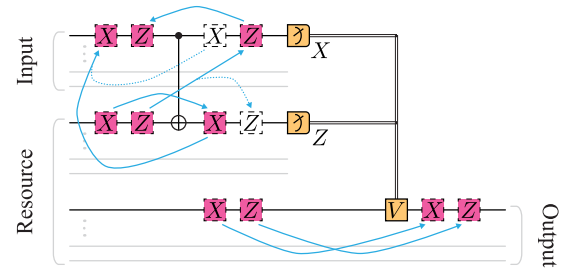


FIG. 4. Gate teleportation. The circuit involves three sets of qubits: the first set carries the input state, while the second and third sets encode the resource state. The diagram illustrates operations on a representative qubit from each set (black lines), with the same operations applied in parallel to the remaining qubits (gray lines). Depending on measurement outcomes, a Pauli gate V is applied on the output qubit. Dashed boxes indicate Pauli errors, and white boxes denote trivial errors. Arrows trace the paths of error propagation through the circuit.

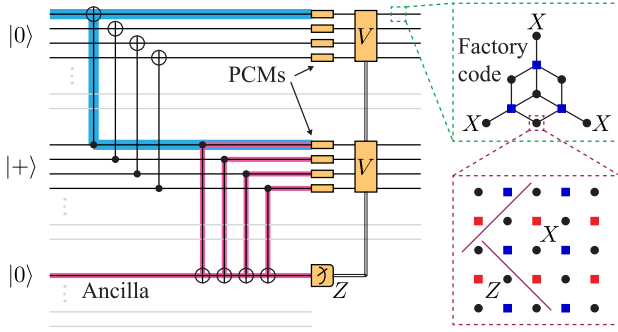


FIG. 5. Circuit for generating resource states used to measure Z stabilizers. The diagram illustrates operations on selected representative qubits (black lines), with similar operations applied in parallel to the remaining qubits (gray lines). The procedure begins by applying transversal controlled-NOT gates between two sets of qubits to prepare Bell states. An ancilla qubit is then used to measure each Z -stabilizer operator of the code, with the illustrated example assuming a weight-four stabilizer for the representative ancilla qubit. Measurement outcomes of the ancilla qubits determine Pauli gates V , which are subsequently applied. Each qubit in this circuit is encoded using an F code, which has only Z -type stabilizers. An example Tanner graph is shown, where blue (red) squares represent Z (X) stabilizers. Note that the depicted graph is from the $[7, 4, 3]$ Hamming code; in practice, the F code should be an LTC with large enough distance. Furthermore, each qubit of the F code is itself encoded in a low-distance surface code, forming a two-level encoding scheme. All operations are performed via transversal gates within both the factory and surface codes. After preparing the encoded resource states, decoding is performed to obtain unencoded resource states. This is achieved by measuring selected qubits in the X and Z bases as illustrated in the figure and applying Pauli gates to the remaining qubits based on the outcomes. The measurement pattern is determined by the standard form of the generator matrices. The entire decoding process can be completed in $O(1)$ time, producing k_F copies of unencoded resource states per circuit run. To correct errors, parity-check measurements (PCMs) of the F code are performed before decoding. Since the F code is an LTC, measurement errors in PCMs are equivalent to low-weight Pauli errors, and a single round of PCMs suffices to reliably detect and correct X errors (including measurement errors on ancilla qubits) by exploiting the propagation of F -code Z stabilizers (highlighted in bold cyan and magenta lines). As a result, each resource state copy has only low-weight residual errors.

of measurements is required in PCS (and other code-surgery protocols).

E. Universal quantum computation

We adopt the following universal set of logical gates: Hadamard gate, S gate, T gate, and controlled-NOT gate. Using logical measurements, we can realize controlled-NOT and Hadamard gates directly, while S and T gates are implemented via logical measurements assisted by their corresponding magic states. In addition to the logical gates, initialization and measurement of memory-code

logical qubits are also required. Table IV summarizes all logical operations, with relevant circuits given in Appendix B.

For the S -gate magic state, we prepare it using a color code with a sufficiently large code distance d_C . The state is first created on a color-code block using the transversal S gate and then transferred to a memory-code block via logical measurements. Because the operations are fault-tolerant, the resulting magic state attains fault-tolerant-level fidelity, eliminating the need for further magic-state distillation.

For the T -gate magic state, we prepare it using a surface code with a small code distance d_S . A noisy magic state is first generated on a surface-code block following the protocol in Ref. [41] and then transferred to a memory-code block. Since the prepared state is noisy, further distillation is required, which is performed using a constant-spacetime-overhead protocol [18]. We note that the protocol naturally distills CCZ-gate magic states. To align with our description of the universal gate set in the Clifford+ T form, the distillation protocol must incorporate a conversion between T -gate and CCZ-gate magic states, which incurs only a constant overhead [42]. Alternatively, our protocol can be straightforwardly adapted to the Clifford + CCZ gate set.

Each logical S gate requires one S -gate magic state as input. Since the magic state is preserved after use, it can be reused. To exploit this, we prepare M copies of the S -gate magic state prior to computation, with each copy stored in an ancilla memory-code block. These ancilla blocks serve S gates for data memory-code blocks throughout the entire computation.

In contrast, each T gate consumes one copy of the T -gate magic state. Therefore, these magic states must be generated on demand during computation.

F. Cost analysis

We now present a scheme for qubit allocation and logical circuit compilation, followed by an analysis of the associated resource overhead. While the scheme leaves significant room for further optimization, it is sufficient to establish our main conclusions: a constant qubit overhead and a time overhead of $O(kd^{o(1)}) = O(d^{a+o(1)})$.

The allocation of qubits in the quantum computer is illustrated in Fig. 2. In addition to data memory-code blocks, the memory includes ancilla memory-code blocks organized into three sectors: one stores S -gate magic states, another supports the implementation of H , S , and controlled-NOT gates, and the third manages slot switching in PCS. For the latter, when logical operations are performed on $q < k_R$ target blocks, ancilla blocks are used to fill the remaining $k_R - q$ slots. Each sector contains M ancilla blocks. Since the memory code has constant encoding rate, the total number of physical qubits in

TABLE IV. Universal set of logical operations employed in our scheme. Initialization prepares all logical qubits of a memory-code block transversally in the state $|0\rangle$. Mea_j , H_j , S_j , and T_j denote, respectively, the Z-basis measurement, Hadamard gate, S gate, and T gate acting on the j th logical qubit of a memory-code block. Note that Mea_j can be used to reinitialize a logical qubit. $\text{CNOT}_{a,b}$ is the controlled-NOT gate with control the a th logical qubit of block A and target the b th logical qubit of block B , where A and B may be the same or different blocks. Each logical operation is implemented through a set of logical measurements and therefore consumes corresponding resource states; the table also lists any additional resource states consumed in addition to those used in the logical measurements. A T gate consumes one copy of the T -gate magic state, whereas an S gate requires one copy of the S -gate magic state as input but does not consume it. In addition to the universal operation set, we also list the magic-state preparation operations.

Logical operations	Logical measurements	Extra resource states
Initialization		H_X^M
Mea_j	$\bar{Z}_j$	
H_j	$\bar{Z}_j \otimes \bar{Z}_1, \bar{X}_j, \bar{Z}_j \otimes \bar{X}_1, \bar{Z}_1$	H_Z^M
S_j	$\bar{Z}_1 \otimes \bar{Z}_1, \bar{Z}_j \otimes \bar{Z}_1 \otimes \bar{X}_1, \bar{X}_1$	H_Z^M
T_j	$\bar{Z}_j \otimes \bar{Z}_1, \bar{X}_1 \otimes 2, \bar{Z}_1 \otimes \bar{Z}_1, \bar{Z}_j \otimes \bar{Z}_1 \otimes \bar{X}_1$	T -gate magic state, H_Z^M
$\text{CNOT}_{a,b}$	$\bar{Z}_a \otimes \bar{Z}_1, \bar{X}_b \otimes \bar{X}_1, \bar{Z}_1$	H_Z^M
Noisy T -gate magic state preparation	$\bar{Z}_S \otimes \bar{Z}_1$	H_X^S, H_Z^S, H_Z^M
Fault-tolerant S -state magic state preparation	$\bar{Z}_C \otimes \bar{Z}_1$	H_X^C, H_Z^C, H_Z^M

memory is $O(Mk)$, where Mk is the total number of logical qubits in the data memory-code blocks.

The resource-state and magic-state factories also require physical qubits. Assume the F code has a constant encoding rate. To serve all M data blocks simultaneously, the factories would naively require $O(Mkd_S^2)$ physical qubits, where the factor d_S^2 arises from the encoding with the surface code (see Fig. 5). To recover a *strictly constant* qubit overhead, we implement a time-space trade-off: by using fewer physical qubits in the factories and serving data blocks in batches of $O(M/d_S^2)$, the qubit requirement for the factories is reduced to $O(Mk)$ (see Fig. 2). This trade-off incurs an additional time overhead factor of $O(d_S^2) = O(\text{polylog}(d))$, which is absorbed into the $O(d^{o(1)})$ term of our final complexity results. Altogether, the quantum computer utilizes $O(Mk)$ physical qubits, achieving the target of constant qubit overhead.

We now address the time overhead of the protocol. As we will show in Sec. IV, the LTSP scheme generates resource states with almost constant qubit overhead and circuit depth. Utilizing these resource states, the PCS gadget can then be implemented with strictly constant qubit overhead and circuit depth. Therefore, the integration of PCS and LTSP enables the parallel processing of $O(Mk)$ logical qubits using almost $O(Mk)$ physical qubits (without the time-space trade-off) in almost constant depth, potentially achieving a nearly constant spacetime cost.

However, PCS and LTSP impose specific compilation constraints: each execution of PCS implements identical operations across k_R code blocks, and each run of the LTSP circuit generates k_F identical resource state copies.

These characteristics introduce a single-instruction-multiple-data constraint on the logical circuit compilation. For structured circuits with repetitive gate motifs—such as the Trotterized simulation of Hamiltonians with translational symmetry—it may be possible to approach nearly constant spacetime overhead. Conversely, for general circuits with heterogeneous patterns, these constraints necessitate the *serialization* of logical operations. This serialization process is the primary source of the $O(d^{a+o(1)})$ time overhead reported in our final results.

Logical operations are serialized as follows. Consider compiling a single layer of heterogeneous logical operations acting on disjoint data logical qubits. These operations are serialized into multiple sub-layers so that, in each sub-layer, every data memory-code block performs at most one logical operation. Consequently, the number of operations per sub-layer is at most M . This serialization reduces to a graph coloring problem, and a full layer of operations can be scheduled within $O(k)$ sub-layers; see Ref. [18] and Appendix C. If the time overhead of each sub-layer is T , then the total time overhead is $O(kT)$. This introduces a time overhead factor of $O(k) = O(d^a)$, which represents the primary contributor to the overall time complexity. Next, we analyze the time overhead of each sub-layer.

The physical circuit depth of all logical operations is $O(1)$, assuming access to sufficient gate-teleportation resource states and distilled magic states. The circuits of logical operations are detailed in Appendix B, where logical measurements are implemented using the PCS circuit shown in Fig. 3, and the parity-check measurements in PCS are realized via the gate-teleportation circuit in

Fig. 4. Consequently, the time overhead T is entirely determined by the efficiency of the two state factories.

Gate-teleportation resource states used in logical operations fall into two categories: those for parity checks of the memory code and those for parity checks of deformed codes. In PCS, a deformed code can simultaneously act on k_R code blocks. Accordingly, to implement up to M operations in a sub-layer, the number of required memory-code resource states is $O(M)$, while that of deformed-code resource states is $O(M/k_R)$. In LTSP, each execution of the preparation circuit produces k_F identical copies of a resource state. For memory-code resource states, the preparation circuit requires $O(kk_F d_S^2)$ physical qubits; for deformed-code resource states, it requires $O(kk_R k_F d_S^2)$ qubits. Here, $d_S = \text{polylog}(d)$ is the surface code distance, which will be justified in Sec. IV B. In both cases, the circuit depth is $O(d_S)$. Therefore, with $O(Mk)$ physical qubits allocated to the resource-state factory (consider the time-space trade-off), one can generate all gate-teleportation resource states needed for a sub-layer within time $O(d_S^3)$.

For each sub-layer of logical operations, at most M copies of distilled T -gate magic states are required. The preparation of these states proceeds in two stages. In the first stage, we prepare $O(M)$ copies of noisy magic states on distance- d_S surface-code blocks, which are subsequently transferred to memory-code blocks—each copy is transferred from a surface-code block to the first logical qubit of a memory-code block. The noisy-state preparation circuit requires $O(Mk)$ physical qubits and has a depth of $O(1)$, while the necessary resource states are generated in the resource-state factory with a time overhead of $O(d_S^3)$. In the second stage, we distill the noisy magic states. Using a constant-overhead distillation protocol [18], we can obtain M copies of distilled magic states from the noisy inputs. The distillation stage incurs a time overhead of $O(d^{o(1)} d_S^3)$, where $d^{o(1)}$ originates from the distillation circuit, and the d_S^3 factor arises from the resource-state generation.

Based on the above analysis, the total time overhead for each sub-layer is $O(d^{o(1)} d_S^3)$ (the $d^{o(1)}$ factor originates from the magic state distillation protocol [18]; see Theorem 2), resulting in a time overhead of $O(kd^{o(1)} d_S^3) = O(d^{a+o(1)})$ for an entire layer. Note that $d_S = \text{polylog}(d)$ (an explicit analytical expression for d_S is provided in Appendix J 1). In addition to executing logical operations, the computation also requires the preparation of S -gate magic states and the initialization of data memory-code blocks prior to computation. These tasks can be completed in time $O(kd_S^3)$ —the same order as that of an operation layer—given $O(Mk)$ physical qubits. A detailed analysis is provided in Appendix D.

IV. LOW-OVERHEAD CODE SURGERY

In the resource-efficient code surgery protocol, specifically devised sticking, the overall spacetime overhead scales as $O(d^2)$. This overhead arises from two sources. One factor of d comes from the qubit cost of constructing the ancilla system in code surgery. In devised sticking, the ancilla size scales as $O(kd)$ to measure $\Theta(k)$ logical operators simultaneously. The other factor of d reflects the inherent time-efficiency limitation of conventional code surgery, which requires $\Theta(d)$ rounds of parity-check measurements to suppress measurement errors. In this work, we develop an approach that eliminates both d -scaling factors, enabling a constant-spacetime-overhead implementation of code surgery. However, it imposes constraints on the operation set, resulting in an overall time overhead of $O(d^{a+o(1)})$ when compiling a general logical circuit.

We address the qubit cost with PCS. In this approach, code surgery is applied simultaneously across multiple memory blocks using a single ancilla system. Consequently, with an ancilla of size $O(k\text{poly}(d))$, we can operate on $\Theta(k\text{poly}(d))$ —the same polynomial—logical qubits in parallel, resulting in a qubit overhead that remains constant.

We overcome the inherent time-efficiency limitation of code surgery through the combination of gate teleportation and LTSP. The central challenge is to reduce the time overhead without compromising the constant qubit overhead [15,28]. To this end, we utilize gate teleportation to perform parity-check measurements and propose generating the required resource states using a classical LTC. Leveraging the local testability property, we demonstrate that the overhead from state preparation contributes only an almost constant factor to the qubit cost, while reducing the time overhead of code surgery to almost constant.

Although the d^2 factor is eliminated, the use of PCS and LTSP imposes constraints on the compilation of logical circuits. In particular, only an operation set of size $O(1)$ can be applied to each memory block, similar to the CC + GT protocol using state distillation [18]. This limited operation set restricts intra-block parallelization and necessitates serialization of logical operations. As a result, the overall time overhead scales as $O(kd^{o(1)}) = O(d^{a+o(1)})$.

A. Parallelized code surgery

Before presenting our scheme, we first review the qubit overhead in several existing code surgery protocols; see Fig. 6. In the CKBB protocol, the ancilla system is constructed as a hypergraph product code, which is a product of two linear codes, referred to as the glue code and the R code throughout this paper. To measure a logical operator via code surgery, a layer of qubits along the

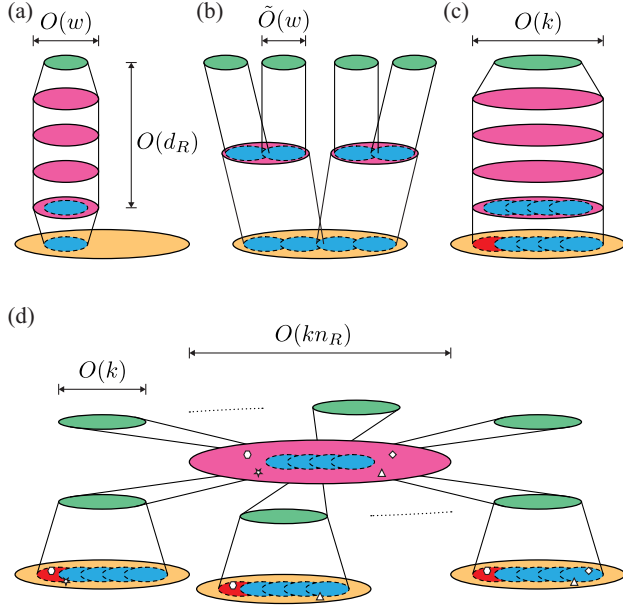


FIG. 6. Code surgery schemes. Each orange circle represents a memory block. Magenta and green circles denote two types of subsystems that constitute the ancilla system. Dashed cyan circles indicate logical operators to be measured. (a) CKBB protocol. The ancilla system comprises multiple subsystems and is coupled to the support of the target logical operator. (b) GM + BFB protocol. The ancilla system consists of two components: branch-sticker subsystems (magenta), which directly couple to memory blocks to isolate logical operators, and gauging-measurement subsystems (green), which couple to the branch stickers to facilitate readout. Although only one layer of branch stickers is shown, measuring q logical operators in parallel requires $O(\log d)$ layers in practice. (c) Devised sticking. The ancilla system is coupled to a subset of the memory block that may overlap with logical operators beyond the targeted ones. By appropriately designing the glue code, any subset of logical operators can be selectively measured, while unmeasured logical operators (dashed red circles) are preserved. (d) Parallelized code surgery. Logical operators on multiple memory blocks are measured simultaneously. The ancilla system comprises two parts: the large magenta circle corresponds to the second columns of H_X^D and H_Z^D in Eqs. (3) and (4); the green circles correspond to the third columns. There are n_R green-circle subsystems in total, with k_R of them coupled to memory blocks. Errors on the magenta subsystem (indicated by white-face markers) are equivalent to errors on the memory blocks.

glue-code direction [horizontal direction in Fig. 6(a)] is coupled to the support of the operator to extract its eigenvalue. Consequently, the glue code has size $\Theta(w)$, where w is the weight of the logical operator. The R code is chosen as a repetition code of length $\Theta(d)$ to suppress errors in the ancilla, thereby setting the size of the ancilla in the other direction [vertical direction in Fig. 6(a)]. Together, these yield an overall ancilla size of $O(wd)$ per measured logical operator. In the gauging measurement protocol, the R-code dimension is minimized by using an

expander-based construction that controls errors through graph-theoretic properties, reducing the ancilla size to $O(w^{1+o(1)})$ per measured operator. In devised sticking, the ancilla is also constructed as a hypergraph product code. However, the glue-code layer is coupled to multiple logical operators simultaneously. By carefully designing the glue code, the ancilla can be programmed to measure an arbitrary subset of logical operators. When measuring up to k logical operators in parallel, the glue code has size $O(k)$, resulting in an overall ancilla size of $O(kd)$. Thus, the qubit overhead per logical operator is reduced to $O(d)$.

In our approach, we further reduce the qubit overhead by coupling a single ancilla system to multiple memory blocks; see Fig. 6(d). Specifically, we select the R code to be a constant-rate LDPC code (see Table II). We require the R code to have distance $d_R = \Omega(n_R^{1/a_R})$ for some finite constant a_R , ensuring that its block length satisfies $n_R = \text{poly}(d)$ when $d_R = \Theta(d)$. Each independent codeword of the R code functions as a dedicated channel for reading out the eigenvalues from a corresponding memory block. Consequently, an R code of length n_R enables simultaneous measurement of $\Theta(n_R)$ memory blocks and $\Theta(kn_R)$ logical operators. This construction reduces the eventual qubit overhead per logical operator to $O(1)$.

The idea of coupling an ancilla system to the memory multiple times may have broader applications. It has been discussed as a potential strategy for measuring high-weight logical operators [43]. This technique may also offer a pathway to reducing qubit overhead when measuring multiple operators within a single code block. However, such applications remain underdeveloped due to the absence of a rigorous justification of fault tolerance—specifically, there is no guarantee that the resulting deformed code maintains a well-lower-bounded distance. In contrast, the approach presented in this work provides a systematic framework for constructing the ancilla system and designing its coupling to memory blocks, ensuring a provable code distance while faithfully implementing the desired logical operations.

We now present the scheme of PCS. Following the standard formalism for CSS codes, we represent the target code by the tuple (H_X, H_Z, J_X, J_Z) . This corresponds to the code of the target block on which the logical measurement is applied. The target block could be a single memory-code block or a composite of multiple blocks. To simplify notation, we use $[[n, k, d]]$ to denote the parameters of the target code, although this notation has also been used for the memory code. Here, H_X and H_Z (J_X and J_Z) are the check (generator) matrices, representing X and Z stabilizer (logical) operators of the code, respectively. Formal definitions of these matrices and notation can be found in Appendix A 1. These matrices satisfy the standard

conditions: $H_X H_Z^T = H_X J_Z^T = H_Z J_X^T = 0$ and $J_X J_Z^T = E_k$, where E_k is the $k \times k$ identity matrix.

In PCS, we perform simultaneous logical measurements on multiple target blocks. On each block, we measure q independent Z logical operators, which can be compactly represented by the matrix αJ_Z . Here, $\alpha \in \mathbb{F}_2^{q \times k}$ is a full-rank matrix, with each row specifying a Z logical operator. Given a R code with logical dimension k_R , the same single-block measurement specified by αJ_Z can be applied in parallel to k_R target blocks. While we focus on single-block measurements of Z logical operators, the generalization to multi-block measurements and X logical operators is straightforward.

To realize the parallelized measurement, we employ a deformed code defined by the following check matrices:

$$H_X^D = \begin{pmatrix} E_{k_R} \otimes H_X & 0 & G_R^r \otimes T \\ 0 & E_{r_R} \otimes H_G & H_R \otimes E_{r_G} \end{pmatrix} \quad (3)$$

and

$$H_Z^D = \begin{pmatrix} E_{k_R} \otimes H_Z & 0 & 0 \\ G_R^r \otimes S & H_R^T \otimes E_{n_G} & E_{n_R} \otimes H_G^T \end{pmatrix}. \quad (4)$$

Here, r_X is the number of rows in the matrix H_X , and the superscript “r” denotes the right inverse of a matrix. The matrices H_R , G_R , H_G , S , and T will be explained later. In each check matrix, the first column corresponds to the k_R target blocks, while the second and third columns represent the ancilla system. In the code surgery procedure, the ancilla-system qubits are initialized in the $|+\rangle$ state. Parity-check measurements associated with the deformed code are then performed, followed by X -basis measurements on the ancilla-system qubits. This protocol implements the desired Z logical measurements across the target blocks.

The matrices H_R and G_R define the R code, where H_R is the check matrix and G_R is the generator matrix. We choose G_R in its standard form, so that its right inverse takes the form $G_R^r = (E_{k_R} \ 0)$, which is sparse. The ancilla system used in the protocol is a hypergraph product code generated from the R code and the glue code, the latter specified by the check matrix H_G . The X and Z check matrices of the hypergraph product code are given by the last two columns in the second row of H_X^D and H_Z^D , respectively.

The matrices H_G , S , and T are constructed following the framework of devised sticking [22], such that they satisfy the following conditions: (i) $H_X S^T = T H_G$; (ii) there exists a matrix R such that $\alpha J_Z R S = \alpha J_Z$ and $H_G (\alpha J_Z R)^T = 0$; and (iii) there exists a matrix β such that $\alpha_\perp J_X S^T = \beta H_G$. Here, $\alpha_\perp \in \mathbb{F}_2^{(k-q) \times k}$ is a full-rank matrix satisfying $\alpha_\perp \alpha^T = 0$. These conditions ensure that the ancilla system

is correctly programed and coupled to the target blocks: the intended logical operators are measured, while the unmeasured operators are preserved throughout the code surgery procedure. Furthermore, the matrices H_G , S , and T are constructed to be sparse: their row and column weights are bounded above by a constant, ensuring that the resulting deformed code remains a qLDPC code. In particular, S is designed to have both row and column weights exactly equal to one, a property that contributes to maintaining a favorable distance in the deformed code. Finally, when the generator matrices J_X and J_Z are expressed in standard form, and assuming that the logical operators selected for measurement act on mutually disjoint sets of logical qubits (i.e. the logical thickness is one), the number of rows and columns in H_G scales as $O(k)$. This ensures that the qubit overhead of code surgery is constant.

We can measure the target logical operators because they belong to the stabilizer group of the deformed code. This is demonstrated by the relation

$$(E_{k_R} \otimes (\alpha J_Z) \ 0 \ 0) = (0 \ G_R \otimes (\alpha J_Z R)) H_Z^D. \quad (5)$$

This equation also specifies how to extract the eigenvalues of the target logical operators from the measurement outcomes of the stabilizers.

Next, we sketch the proof for a lower bound on the distance of the deformed code. The key idea is to use the technique that establishes an equivalence between errors on the ancilla system and errors on the target blocks; see Fig. 6(d). On each target block, the weight of the equivalent error is upper bounded by the weight of the original error on the ancilla system. Consequently, the distance of the deformed code is lower bounded by the distance of the target code. We note, however, that an error on the ancilla system may be equivalent to multiple distinct errors on different target blocks, resulting in error amplification. In the present protocol, such amplification is tolerable because error correction is performed independently on each target block. Finally, we emphasize that the validity of the error equivalence argument assumes that the error weight is below the distance of the R code. Therefore, the R code must have sufficiently large distance to ensure fault tolerance.

Lemma 1. Suppose the target code is a qLDPC code with constant encoding rate, and the R code is an LDPC code also with constant encoding rate. Then, the resulting deformed code supports PCS with constant qubit overhead. In particular, the deformed code defined by Eqs. (3) and (4) is a qLDPC code whose length is upper bounded by $(k_R + 2n_R)n = O(k_R k)$. This code enables the intended operation: the parallel measurement of an identical set of up to k logical operators, specified by αJ_Z , across k_R target

blocks. Moreover, if H_R is full-rank, the distance of the deformed code satisfies $d_D \geq \min\{d, d_R\}$.

See Appendix F for the proof.

B. Locally-testable state preparation

The inherent time-efficiency limitation of code surgery arises from the need to correct measurement errors. In conventional code surgery, a subset of logical operators is measured by coupling the target block to an ancilla system, resulting in a deformed code whose stabilizer group includes the logical operators to be measured. Eigenvalues of these operators are then extracted by performing parity-check measurements of the deformed code, i.e. measuring its stabilizer operators. However, due to the presence of measurement errors, these measurements must be repeated for $\Theta(d)$ rounds to ensure accurate error correction and reliable eigenvalue extraction. Consequently, if parity-check measurements could be performed without measurement errors, this would eliminate the $\Theta(d)$ repetition factor and significantly reduce the overall time overhead.

Parity-check measurements free of measurement errors can be effectively realized through gate teleportation. The key insight is that, in gate teleportation, all errors can be interpreted as acting either on the input or output state, as illustrated in Fig. 4. This leads to the following effective picture: the input state may carry some error, followed by a single round of error-free parity-check measurements, after which additional errors may occur on the output state. Although the measurements are effectively error-free in this picture, it remains crucial to ensure that the effective errors on both the input and output states are low-weight and therefore correctable. In particular, this requires that the resource state used in gate teleportation contains only low-weight errors.

Errors in the resource state can be managed either by using a high-distance concatenated code or by applying state distillation [15–18]. While high-distance concatenated codes incur substantial spacetime overhead, constant-overhead state distillation requires the quantum memory to be encoded in a good qLTC. Since our objective is a general method of reducing spacetime overhead for qLDPC codes, an alternative approach is necessary. Furthermore, even if the memory was encoded in a good qLTC, constructing a deformed code that also satisfies local testability is a highly nontrivial task, further limiting the practical applicability of constant-overhead state distillation in our context.

Our approach, LTSP, offers a generic method to control measurement errors in qLDPC-code quantum error correction, while requiring only almost constant overhead; see Fig. 5. Central to this approach is generating resource states using a classical LTC, which we refer to as the F code because it serves as the core component of the resource state factory (see Table II). In our scheme, each resource state is dedicated to measuring only Z stabilizer

operators— X stabilizer operators can be measured by rotating the basis. When measuring Z stabilizers, we simulate the corresponding measurement circuit by encoding logical qubits in the F code. Specifically, we treat the classical code as a quantum code that protects only Z operators. Then, the measurement circuit is simulated with transversal operations. This simulation produces an encoded resource state, which can be efficiently transformed into the eventual unencoded resource state. Our method reliably produces resource states meeting the requirement that errors remain low-weight, as summarized in Lemma 2. Importantly, this approach is broadly applicable to any qLDPC code.

Lemma 2. Consider the generation of gate-teleportation resource states for measuring Z stabilizer operators of a deformed code. Suppose the F code is an LDPC code with constant soundness. Let $|\bar{e}_X^{sp}|$ and $|\bar{e}_Z^{sp}|$ denote the weights of X and Z errors occurring during the state preparation circuit, respectively. Let $|e_X^{rs}|$ and $|e_Z^{rs}|$ denote the weights of X and Z errors present on one copy of the prepared resource state. If these errors are undetectable and satisfy $|\bar{e}_X^{sp}| < d_F/C_1$, then the following bounds hold: $|e_X^{rs}| \leq C_2|\bar{e}_X^{sp}|$ and $|e_Z^{rs}| \leq C_3|\bar{e}_Z^{sp}|$, where C_1 , C_2 , and C_3 are constants determined by the weight and soundness parameters of the involved codes. The state preparation circuit uses $O(n_D n_F d_S^2)$ physical qubits, has depth $O(d_S)$, and produces k_F copies of the resource state. While d_S can be arbitrary positive integer, in practice it is chosen to scale as $\text{polylog}(n_D n_F)$. These conclusions naturally generalize from deformed codes to arbitrary CSS-type qLDPC codes.

See Appendix G for a detailed explanation of the protocol and the corresponding proof. Explicit expressions for constants C_1 , C_2 , and C_3 are provided in Lemma 15.

Now, we explain why errors on the resource state remain low-weight. Recall that our goal is to measure stabilizer operators of a qLDPC code (the deformed code). Consequently, Pauli errors generated directly by circuit operations remain low-weight. The challenge arises in the feedback operations: to prepare the resource state within the logical subspace of the deformed code, we apply Pauli gates conditioned on measurement outcomes identifying the subspace sectors. However, these measurements can be faulty. Since these feedback Pauli gates may have large weight, even a low-weight measurement error can propagate into a high-weight Pauli error on the output resource state.

We address this vulnerability by protecting the stabilizer-measurement circuit using the F code. We treat the F code as a quantum CSS code that only protects against X errors (assuming the circuit measures Z stabilizers). Then, each qubit in the circuit is encoded as a logical qubit of the F code. Operations within the circuit are then implemented as transversal logical operations on the F code. This protection ensures that measurement errors—and the subsequent propagation into high-weight Pauli errors—are suppressed. As established in Lemma 2, as long as the

weight of the physical errors in the circuit is bounded by $\Theta(d_F)$ (specifically, $|\bar{e}_X^{sp}| < d_F/C_1$), measurement errors do not occur. This prevents the triggering of dense feedback and ensures that the errors on the output resource state remain bounded and proportional to the weight of the physical errors.

However, introducing an additional level of encoding could potentially increase the circuit depth. Standard implementations of error correction for the F code would typically require $O(d_F)$ rounds of stabilizer measurements, thereby amplifying the circuit depth by a factor proportional to the distance. We circumvent this bottleneck by selecting a constant-soundness LTC as the F code. Such codes are characterized by their *single-shot* property and have linear soundness functions [35]. Due to this single-shot nature, a single round of stabilizer measurement is sufficient to achieve the required error suppression, ensuring that the circuit depth remains constant.

Next, we analyze the qubit costs. To reduce the qubit overhead for generating resource states, we propose using a double-constant classical LTC with constant encoding rate and constant soundness [13,31,32]. Similar to the R code, we require the F code to have distance $d_F = \Omega(n_F^{1/a_F})$ for some finite constant a_F , ensuring that its block length satisfies $n_F = \text{poly}(d)$ when $d_F = \Theta(d)$. Because the encoding rate is constant, we can produce $\Theta(n_F)$ copies of the resource state simultaneously, ensuring constant qubit overhead.

Finally, we address a critical issue neglected in the above discussion. Because the F code is classical, it can only protect against one type of error, allowing the other type to accumulate. In fact, it cannot correct either error type to the fault-tolerant level up to our circuit design. This error accumulation typically results in a vanishing fault-tolerant threshold. To overcome this, we further encode each physical qubit of the F code into a logical qubit of a low-distance surface code (or any other low-distance code supporting a certain logical operation set). The measurement circuit has size $O(n_D)$, where n_D is the length of the deformed code. When encoding this circuit with the F code, the circuit size becomes $O(n_D n_F)$. To control error accumulation, it suffices to choose a surface code distance $d_S = \text{polylog}(n_D n_F)$. Incorporating the surface code increases the circuit size to $O(n_D n_F \text{poly}(d_S))$, thus contributing only polylogarithmic overhead factors to the qubit overhead and circuit depth.

V. DISCUSSIONS

We have presented a general scheme for FTQC on qLDPC codes that combines code surgery with gate teleportation and introduces optimizations in deformed-code construction and resource-state preparation. Our approach achieves constant qubit overhead while significantly reducing time overhead. For all constant-rate qLDPC codes whose code distance grows faster than that

of HGP codes, the time overhead is lower than that of the GM + BFB protocol; for good qLDPC codes, it reaches the $O(d^{1+o(1)})$ scaling. These results improve the asymptotic resource efficiency of FTQC.

While the primary objective of this work is to optimize scaling behavior in the limit of large code distance, the techniques developed for this purpose—specifically PCS and LTSP—are applicable for lowering the time overhead in near-term, intermediate-distance implementations, offering promising avenues for future investigation (see Appendix J). PCS is uniquely suited for implementing identical logical operations across multiple code blocks, which arise frequently in practical applications such as the simulation of many-body quantum systems with translational invariance. In these scenarios, PCS enables the simultaneous manipulation of all logical qubits with constant qubit overhead, even for code families like HGP codes, where the asymptotic advantage of our protocol over the GM + BFB method for general heterogeneous circuits vanishes. The resource efficiency of LTSP is governed by the parameters of the chosen LTC, underscoring the importance of identifying code instances with optimized parameters at moderate distances to fully realize the performance gains. Notably, LTSP is a versatile framework; for example, utilizing a repetition code with an expander-graph-based check matrix as the F code provides a mechanism for trading qubit overhead for enhanced temporal efficiency. Overall, compared with existing approaches, our protocol offers distinct advantages in scenarios involving large code distances, highly symmetric gate patterns, or time-critical applications.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grants No. 12225507, No. 12088101) and NSAF (Grant No. U1930403).

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX A: PRELIMINARIES

1. Notations

We use $|\bullet|$ and $\|\bullet\|$ to denote the Hamming weight and the matrix norm induced by Hamming weight, respectively, i.e. the norm of a matrix $A \in \mathbb{F}_2^{m \times n}$ is

$$\|A\| = \max \left\{ \frac{|uA|}{|u|} : u \in \mathbb{F}_2^m - \{0\} \right\}. \quad (\text{A1})$$

We denote by $A_{j,\bullet}$ and $A_{\bullet,j}$ the j -th row and j -th column of the matrix A , respectively, and we denote by A^t the right inverse.

Let $b = (b_1, b_2, \dots, b_n) \in \mathbb{F}_2^n$. We use

$$\sigma(b) = \sigma^{b_1} \otimes \sigma^{b_2} \otimes \dots \otimes \sigma^{b_n} \quad (\text{A2})$$

to represent n -qubit $\sigma = X, Z$ Pauli operators. When $B \in \mathbb{F}_2^{m \times n}$,

$$\sigma(B) = \begin{pmatrix} \sigma(B_{1,\bullet}) \\ \sigma(B_{2,\bullet}) \\ \vdots \\ \sigma(B_{m,\bullet}) \end{pmatrix} \quad (\text{A3})$$

is a list of Pauli operators.

An $[[n, k, d]]$ CSS code is represented by four matrices $H_X \in \mathbb{F}_2^{r_X \times n}$, $H_Z \in \mathbb{F}_2^{r_Z \times n}$ and $J_X, J_Z \in \mathbb{F}_2^{k \times n}$ that satisfy conditions $H_X H_Z^T = H_X J_Z^T = J_X H_Z^T = 0$, $J_X J_Z^T = E_k$. The stabilizer of the code is generated by operators $X(H_X)$ and $Z(H_Z)$. The X and Z logical operators of the j -th logical qubit are $X((J_X)_{j,\bullet})$ and $Z((J_Z)_{j,\bullet})$, respectively.

We use $(H_X^c, H_Z^c, J_X^c, J_Z^c)$ with $c = M, D, S, C$ to represent the memory, deformed, surface, and color codes, respectively. Regarding parameters of a quantum code, we use notations n_c, k_c, d_c, r_X^c , and r_Z^c . Check and generator matrices of the R and F codes are $H_c \in \mathbb{F}_2^{r_c \times n_c}$ and $G_c \in \mathbb{F}_2^{k_c \times n_c}$ with $c = R, F$, respectively, which satisfy $H_c G_c^T = 0$; and we also use d_c to denote the distance of a classical linear code.

2. Primitive operations, spacetime locations, and error models

The primitive operations are illustrated in Fig. 7. Given a quantum circuit, there are two sets of spacetime locations. Quantum locations are the qubits during time

intervals between primitive operations; see Figs. 13, 15, and 16 for examples, in which red boxes represent the quantum locations. Classical locations are outcomes of projective measurements: In the measurement of $\sigma(A)$, the outcome is a vector $\nu \in \mathbb{F}_2^m$, and $(-1)^{\nu_j}$ is the eigenvalue of $\sigma(A_{j,\bullet})$, i.e. each element corresponds to an operator; such a measurement introduces m locations, and each location corresponds to one element in the outcome.

We assume that errors can occur independently at any location. We now justify the error model. For initializations and gates with errors, each of them can be modeled as an ideal operation with errors occurring at the locations immediately after the operation. For qubit measurements with errors, they can be modeled as ideal measurements with errors occurring at locations immediately before the measurements: these errors effectively flip measurement outcomes. For a projective measurement with errors, it can be modeled as a measurement that projects the state onto the correct subspace with errors occurring in the measurement outcome and at locations immediately after the measurement: the errors at locations after the measurement may induce an effective incorrect projection. Compared with taking spacetime locations in circuits that projective measurements and generalized transversal controlled-NOT gates are decomposed into standard single-qubit and two-qubit operations (see Lemma 3), our choice of spacetime locations could overestimate effective circuit distances and the fault-tolerance threshold by a constant factor, which does not change our conclusions.

Lemma 3. Projective measurements and generalized transversal controlled-NOT gates specified by a matrix A can be implemented using a circuit composed of single-qubit and two-qubit operations. If the row and column weights of A are bounded above by a positive constant $\omega_{\max}$, then the circuit has depth at most $\omega_{\max} + 2$.

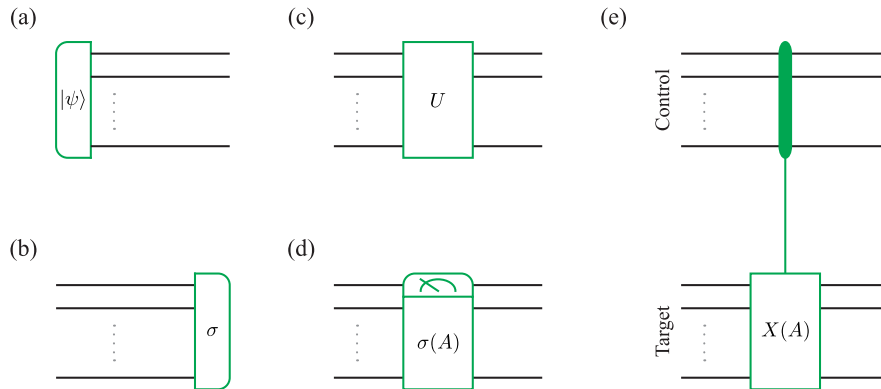


FIG. 7. (a) Transversal initialization that prepares each qubit in the state $|\psi\rangle$. (b) Transversal measurement that measures each qubit in the basis $\sigma = X, Z$. (c) Gate U . Here, U is a layer of single-qubit gates, i.e. $U = U_1 \otimes U_2 \otimes \dots$, where $U_1, U_2, \dots$ are single-qubit gates on qubits 1, 2, ..., respectively. (d) Projective measurement on n qubits that measures Pauli operators $\sigma(A)$, where $A \in \mathbb{F}_2^{m \times n}$. (e) Generalized transversal controlled-NOT gate. The control (target) is a set of m (n) qubits. The transformation $X(A_{j,\bullet})$ is (not) applied on the target when the j -th control qubit is in the state $|1\rangle$ ($|0\rangle$). We assume that the matrix A has rows and columns of weight at most $\omega_{\max}$, where $\omega_{\max}$ is a positive constant.

Moreover, an error occurring at a single location in this circuit may propagate to at most $\omega_{\max}$ locations.

Proof. We only need to consider generalized transversal controlled-NOT gates because projective measurements can be implemented using them. Specifically, to measure the operators $X(A)$, we initialize the control qubits in the $|+\rangle$ state, apply the generalized transversal controlled-NOT gate defined by the matrix A , and then measure the control qubits in the X basis. Similarly, to measure $Z(A^T)$, we initialize the target qubits in the $|0\rangle$ state, apply the same gate, and then measure the target qubits in the Z basis.

A generalized transversal controlled-NOT gate can be implemented by a circuit of depth at most $\omega_{\max}$, since scheduling the controlled-NOT gates corresponds to an edge-coloring problem on a bipartite graph [44,45].

Each qubit in the circuit is involved in at most $\omega_{\max}$ controlled-NOT gates. On a control (target) qubit, an X (Z) error can propagate to the corresponding target (control) qubits, while Z (X) errors do not propagate. Therefore, a single-qubit error may only spread to qubits within the corresponding support set. As a result, the weight of the propagated error is upper bounded by $\omega_{\max}$. ■

3. Locally testable codes

Definition 1. (ω, s) -**locally testable codes** [31]. A linear code $C \subset \mathbb{F}_2^n$ is called (ω, s) -locally testable if it has a check matrix $H \in \mathbb{F}_2^{r \times n}$ with rows of weight at most ω such that for any vector $u \in \mathbb{F}_2^n$ we have

$$\frac{1}{r} |Hu^T| \geq \frac{s}{n} \min_{c \in C} |u - c|. \quad (\text{A4})$$

$$\text{supp}(B, F) = \begin{cases} \{(u, j)\}, & \text{if } B = u, \\ \{(u, a), (v, b)\}, & \text{if } B = (u, v), \end{cases}$$

Similarly, we define the *block support* of (B, F) as

$$\text{supp}(B) = \begin{cases} \{u\}, & \text{if } B = u, \\ \{u, v\}, & \text{if } B = (u, v). \end{cases} \quad (\text{C2})$$

Consider an operation set $OS = \{(B_l, F_l) | l = 1, 2, \dots, L\}$. We say that OS is *qubit-disjoint* if and only if the qubit supports are disjoint, i.e.,

$$\text{supp}(B_{l_1}, F_{l_1}) \cap \text{supp}(B_{l_2}, F_{l_2}) = \emptyset \quad \forall l_1 \neq l_2. \quad (\text{C3})$$

Similarly, OS is *block-disjoint* if and only if the block supports are disjoint, i.e.,

$$\text{supp}(B_{l_1}) \cap \text{supp}(B_{l_2}) = \emptyset \quad \forall l_1 \neq l_2. \quad (\text{C4})$$

The parameters ω and s are positive real numbers called the locality and soundness, respectively.

Lemma 4. If $H \in \mathbb{F}_2^{r \times n}$ is the check matrix of an (ω, s) -locally testable code, for every vector $v^T \in \text{colsp} H$, there exists a vector $u_\star \in \mathbb{F}_2^n$ such that $v^T = Hu_\star^T$ and

$$|u_\star| \leq \frac{n}{rs} |v|. \quad (\text{A5})$$

Proof. Because $v^T \in \text{colsp} H$, there exists $w \in \mathbb{F}_2^n$ such that $v^T = Hw^T$. Then, there exists $c_\star \in \ker H$ such that

$$\frac{1}{r} |v| \geq \frac{s}{n} |w - c_\star|. \quad (\text{A6})$$

The lemma is proved by taking $u_\star = w - c_\star$. ■

APPENDIX B: LOGICAL OPERATIONS

The circuits implementing logical operations are shown in Fig. 8, while the circuits for magic-state preparation are shown in Fig. 9.

APPENDIX C: SERIALIZATION

We use a pair (B, F) to denote a logical operation: the operation $F \in \{\text{Mea}_j, H_j, S_j, T_j, \text{CNOT}_{a,b}\}$ applied on the block(s) B . Specifically, if $F \in \{\text{Mea}_j, H_j, S_j, T_j\}$ is applied on the u -th block, then $B = u$; if $F = \text{CNOT}_{a,b}$ is applied on the a -th logical qubit in the u -th block and the b -th logical qubit in the v -th block, then $B = (u, v)$.

We define the *qubit support* of the operation (B, F) as

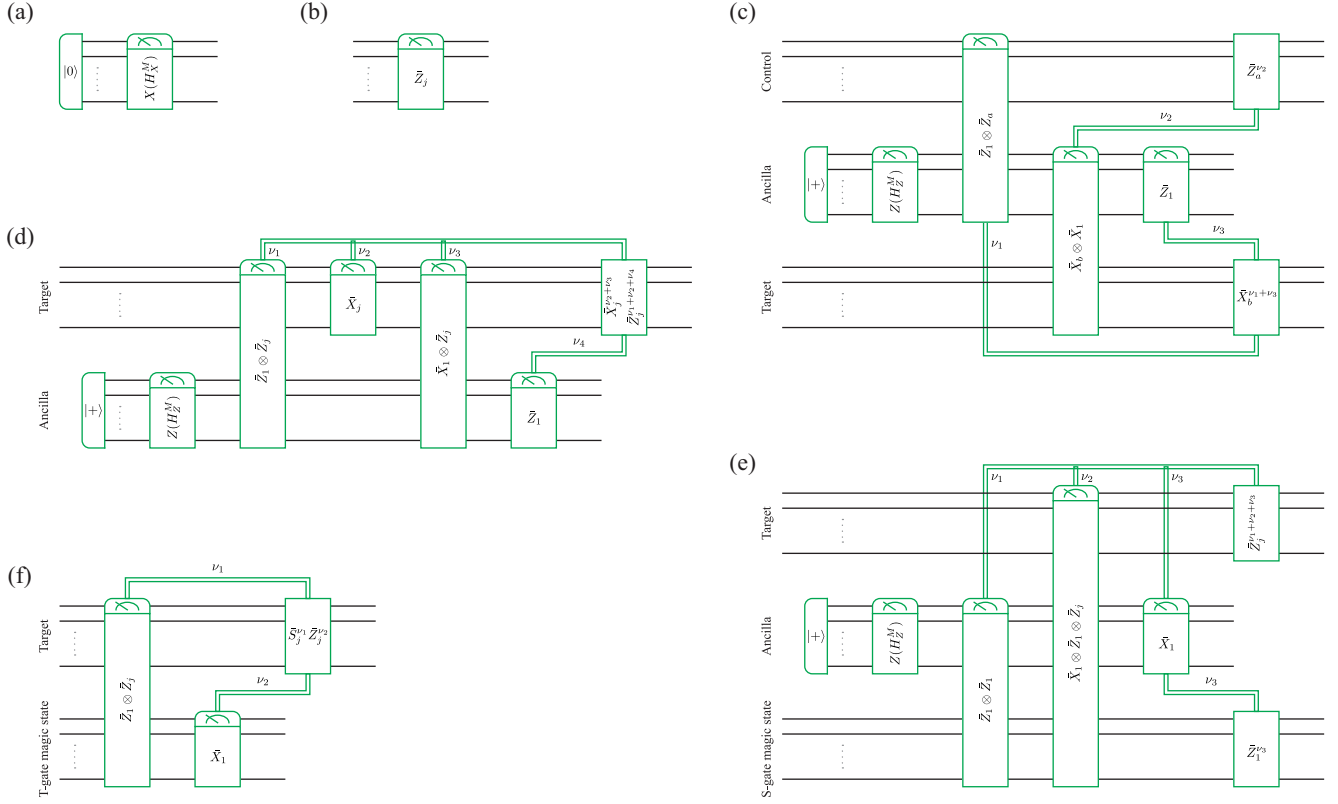
$$\begin{aligned} \text{supp}(B, F) &= \begin{cases} \{(u, j)\}, & \text{if } B = u, F \in \{\text{Mea}_j, H_j, S_j, T_j\}, \\ \{(u, a), (v, b)\}, & \text{if } B = (u, v), F = \text{CNOT}_{a,b}. \end{cases} \end{aligned} \quad (\text{C1})$$

Lemma 5. For any qubit-disjoint operation set $OS = \{(B_l, F_l) : l = 1, 2, \dots, L\}$ (an operation layer) acting on code blocks, each encoding k logical qubits, there exists a partition P of OS such that $|P| = O(k)$, and every element of P (a sub-layer) is block-disjoint.

Proof. We construct the partition as follows (see also Refs. [17,18]).

First, we map the operation set OS to a multigraph with loops $G = (V, E, r)$. Suppose the operation set acts on M code blocks; then the vertex set is $V = \{1, 2, \dots, M\}$. The edge set $E = \{1, 2, \dots, L\}$ has the same cardinality as the operation set, and $r(l) = \text{supp}(B_l)$ specifies the endpoint vertices of the l -th edge.

Next, we find a proper edge coloring of G , i.e., a mapping $f: E \rightarrow C$, where C is the set of colors, such that



any two edges incident to the same vertex receive different colors. Since OS is qubit-disjoint, the maximum degree of G is k . Then, the minimum number of colors required for the coloring, the chromatic index of G , is $O(k)$ [18,46].

Finally, we construct the desired partition as $P = \{U_c : c \in C\}$, where $U_c = \{(B_l, F_l) : f(l) = c\}$. Each U_c is block-disjoint by construction, and $|P| = O(k)$, completing the proof. ■

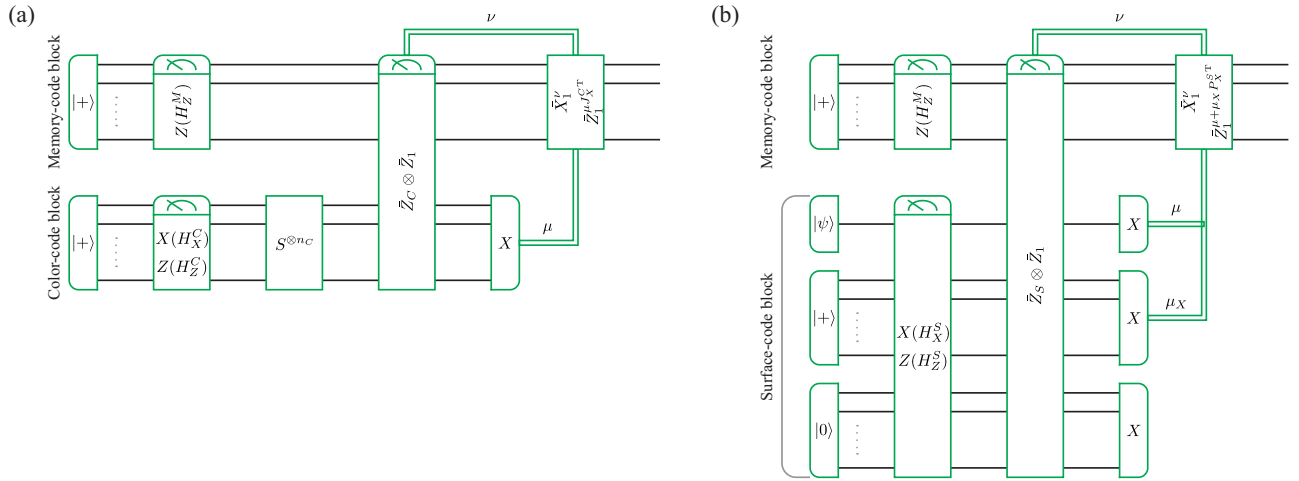


FIG. 9. (a) Preparation of the S -gate magic state: a fault-tolerant copy is prepared on the first logical qubit of the memory-code block. (b) Preparation of the T -gate magic state: a noisy copy is prepared on the first logical qubit of the memory-code block.

APPENDIX D: DETAILED COST ANALYSIS

In this section, we analyze the time overhead under the constraint of a constant qubit overhead. We assume that the number of logical qubits k is small compared to M . Specifically, we assume that the memory code distance scales as $d = O(\text{polylog}(Mk))$, which determines both the block length n and the corresponding logical qubit number k . This distance is sufficient to achieve a logical error rate per operation p_L satisfying $1/p_L = O(\text{poly}(Mk))$. If the logical circuit size is polynomial, i.e. $|C| = O(\text{poly}(Mk))$, then this choice of code distance is also sufficient to achieve any target total logical error rate ϵ satisfying $1/\epsilon = O(\text{poly}(Mk))$.

Lemma 6. Suppose $d, d_R, d_F = O(\text{polylog}(Mk))$. Assuming a sufficient supply of distilled magic states, any block-disjoint operation set (sub-layer) OS can be implemented in time $O(d_S^3)$, requiring $O(Mk)$ physical qubits in the resource-state factory.

Proof. Let $\text{num}(F)$ denote the number of occurrences of operation $F \in \text{Mea}_j, H_j, S_j, T_j, \text{CNOT}_{a,b}$ in OS . Because OS is block-disjoint, $\sum_F \text{num}(F) = |OS| \leq M$. We define

$$\text{batch}(F) = \left\lceil \frac{1}{d_S^2} \left\lceil \frac{1}{k_F} \left\lceil \frac{1}{k_R} \text{num}(F) \right\rceil \right\rceil \right\rceil. \quad (\text{D1})$$

Since

$$\text{batch}(F) \leq \frac{1}{d_S^2} \left\lceil \frac{1}{k_F} \left(\frac{1}{k_R} \text{num}(F) + 1 \right) + 1 \right\rceil + 1, \quad (\text{D2})$$

we obtain

$$\sum_F \text{batch}(F) \leq \frac{M + \Theta(k^2)(k_R + k_R k_F + k_R k_F d_S^2)}{k_R k_F d_S^2}, \quad (\text{D3})$$

where we have used that $|\text{Mea}_j, H_j, S_j, T_j, \text{CNOT}_{a,b}| = \Theta(k^2)$.

The operation set OS is implemented in d_S^2 steps, where each step contains at most $\text{batch}(F)k_R k_F$ copies of operation F acting on different code blocks.

To illustrate, consider $F = \text{Mea}_j$. This operation requires two types of resource states: $H_Z^D(\bar{Z}_j)$ and H_X^M . For each step, at most $\text{batch}(F)k_F$ copies of $H_Z^D(\bar{Z}_j)$ are needed. These can be generated by running $\text{batch}(F)$ copies of the corresponding LTSP circuit in parallel, requiring at most $\text{batch}(F) \times O(nk_R k_F d_S^2)$ physical qubits. Similarly, up to $\text{batch}(F)k_R k_F$ copies of H_X^M are needed, which can be produced by running $\text{batch}(F)k_R$ copies of the corresponding LTSP circuit in parallel, requiring at most $\text{batch}(F)k_R \times O(nk_F d_S^2)$ physical qubits. Therefore,

the total number of physical qubits required for operation F is $\text{batch}(F)O(nk_R k_F d_S^2)$.

The same scaling applies to all operations, yielding

$$N_{RSF,1} \leq \sum_F \text{batch}(F)O(nk_R k_F d_S^2) \quad (\text{D4})$$

$$\leq O(Mk) + O(k^3 k_R k_F d_S^2). \quad (\text{D5})$$

Assuming $d, d_R, d_F = O(\text{polylog}(Mk))$, we have $k^3 k_R k_F d_S^2 = O(\text{polylog}(Mk))$, under assumptions on code rate and distance. Consequently, $N_{RSF,1} = O(Mk)$, i.e., the total number of physical qubits required in the resource-state factory to implement a block-disjoint operation set scales linearly with the logical qubit count.

Each step involves running LTSP circuits in parallel, which requires a time overhead of $O(d_S)$. Since there are d_S^2 steps, the total time overhead for implementing a block-disjoint operation set is $O(d_S^3)$. ■

Lemma 7. Suppose $d, d_R, d_F = O(\text{polylog}(Mk))$ and $d_S = O(\text{polylog}(d))$. Then M copies of distilled T -gate magic states can be generated in time $O(d^{o(1)} d_S^3)$, requiring $O(Mk)$ physical qubits in the resource-state and magic-state factories.

Proof. The generation of distilled magic states consists of two stages. First, we prepare noisy magic states, and then we distill them.

The circuit for noisy magic-state preparation is shown in Fig. 9(b). We begin by preparing a noisy magic state on a surface-code block, which is subsequently transferred to the first logical qubit of a memory-code block. Using the constant spacetime-overhead distillation protocol, to obtain M copies of distilled magic states, we require $O(M)$ copies of noisy magic states [41]. Consequently, $O(M)$ surface-code blocks and $O(M)$ memory-code blocks are needed to store these states, corresponding to $O(M d_S^2)$ and $O(Mk)$ physical qubits in the magic-state factory, respectively.

The $M' = O(M)$ copies of noisy magic states are prepared over d_S^2 steps, where each step produces at most $m' k_R k_F$ copies, with

$$m' = \left\lceil \frac{1}{d_S^2} \left\lceil \frac{1}{k_F} \left\lceil \frac{1}{k_R} M' \right\rceil \right\rceil \right\rceil. \quad (\text{D6})$$

We require four types of gate-teleportation resource states: H_Z^M, H_X^S, H_Z^S , and $H_Z^D(\bar{Z}_S \otimes \bar{Z}_1)$. In each step, up to $m' k_R k_F$ ($m' k_R k_F, m' k_R k_F$, and $m' k_F$) copies of H_Z^M (H_X^S, H_Z^S , and $H_Z^D(\bar{Z}_S \otimes \bar{Z}_1)$) are required. These resource states can be generated by running $m' k_R$ ($m' k_R, m' k_R$, and m') copies of the corresponding LTSP circuits in parallel, which demand at most $m' k_R \times O(nk_F d_S^2)$ ($m' k_R \times O(d_S^2 k_F d_S^2)$),

$m'k_R \times O(d_S^2 k_F d_S^2)$, and $m' \times O(nk_R k_F d_S^2)$ physical qubits. Since $d_S = O(\text{polylog}(n))$, the total number of physical qubits is $O(m'nk_R k_F d_S^2) = O(M'n) = O(Mk)$. The time cost of each step is $O(d_S)$, therefore, the overall time overhead of preparing noisy magic states is $O(d_S^3)$.

To prepare M copies, the distillation circuit requires $O(M)$ logical qubits, has a depth of $O(\text{polylog}(d))$, and is repeated $O(d^{o(1)})$ times [18]. We propose to utilize only the first logical qubit of each memory-code block for distillation. Consequently, $O(M)$ memory-code blocks are required in the magic-state factory, corresponding to $O(Mk)$ physical qubits.

Each layer of the distillation circuit constitutes a block-disjoint operation set, as operations are restricted to the first logical qubit of each block. Therefore, the entire circuit can be executed in time $O(d^{o(1)} d_S^3)$, requiring $O(Mk)$ physical qubits in the resource-state factory (Lemma 6).

Note that the implementation of the distillation circuits also requires ancillary memory-code blocks; however, this contributes only a constant factor to the overall qubit overhead. ■

Lemma 8. Suppose $d, d_R, d_F = O(\text{polylog}(Mk))$, $d_C = O(d)$, and $d_S = O(\text{polylog}(d))$. The quantum computer can be initialized in time $O(kd_S^3)$, requiring $O(Mk)$ physical qubits.

Proof. There are two tasks required to initialize the quantum computer: preparing M copies of high-fidelity S -gate magic states and storing them in M memory-code blocks—each stored in the first logical qubit of a block; and initializing the Mk logical qubits in the data memory-code blocks in the state $|0\rangle$.

As shown in Fig. 9(a), a high-fidelity S -gate magic state is first prepared on a color code block, which is subsequently transferred to the first logical qubit of a memory-code block. We prepare the M copies in k steps with at most $M'' = \lceil M/k \rceil$ copies prepared in each step; then, we need M'' color-code blocks, i.e. at most $M'' \times O(d_C^2) = O(Mk)$ physical qubits. Note that $d_C = O(k)$.

In each step, we prepare the M'' copies over d_S^2 sub-steps, where each sub-step produces at most $m''k_R k_F$ copies, with

$$m'' = \left\lceil \frac{1}{d_S^2} \left\lceil \frac{1}{k_F} \left\lceil \frac{1}{k_R} m'' \right\rceil \right\rceil \right\rceil. \quad (\text{D7})$$

Similar to the preparation of noisy T -gate magic states, we require four types of gate-teleportation resource states: H_Z^M , H_X^C , H_Z^C , and $H_Z^D(\tilde{Z}_C \otimes \tilde{Z}_1)$. In each step, up to $m''k_R k_F$ ($m''k_R k_F$, $m''k_R k_F$, and $m''k_F$) copies of H_Z^M (H_X^C , H_Z^C , and $H_Z^D(\tilde{Z}_C \otimes \tilde{Z}_1)$) are required. These resource states can be generated by running $m''k_R$ ($m''k_R$, $m''k_R$, and m'') copies of the corresponding LTSP circuits in parallel,

requiring at most $m''k_R \times O(nk_F d_S^2)$ ($m''k_R \times O(d_C^2 k_F d_S^2)$, $m''k_R \times O(d_C^2 k_F d_S^2)$, and $m'' \times O((n + d_C^2)k_R k_F d_S^2)$) physical qubits. Since $d_S = O(\text{polylog}(n))$, the total number of physical qubits is $O(m''n^2 k_R k_F d_S^2) = O(M''n^2) = O(Mk)$, noting that $d_C = O(n)$.

According to the above analysis, the overall time overhead for preparing high-fidelity S -gate magic states is $O(kd_S^3)$.

The initialization of the data memory-code blocks is similar to a block-disjoint operation set and can therefore be accomplished in time $O(d_S^3)$ with $O(Mk)$ physical qubits. ■

APPENDIX E: IMPLEMENTATION ON ALMOST-CONSTANT-DEGREE QUBIT NETWORKS

In this section, we adapt our protocol to a physical qubit network characterized by an almost constant vertex degree:

- (i) **Physical qubit network:** The physical network is modeled as a graph where each vertex represents a physical qubit. Two-qubit physical gates can be implemented between any pair of qubits connected by an edge. The maximum vertex degree of this graph is $D_P = O(d^{o(1)})$, meaning that each physical qubit is coupled to at most an almost constant number of neighbors.
- (ii) **Logical qubit network:** On this physical network, we realize a fault-tolerant quantum computer defined by a *logical coupling graph*. On this graph, a memory-code block is assigned to each vertex. Inter-block logical operations are supported between any two blocks connected by an edge. The maximum vertex degree of the logical graph is strictly bounded by a constant, $D_L = \Theta(1)$.
- (iii) **Qubit overhead:** Under these connectivity constraints, the protocol maintains an almost constant qubit overhead of $O(d^{o(1)})$.
- (iv) **Time overhead and logical operations:** The time overhead remains $O(d^{a+o(1)})$ for the implementation of logical circuits consisting of the following operations:

- (1) **Single-qubit operations:** These include state initialization and measurement in the X or Z basis, single-qubit Clifford gates, and the preparation of noisy magic states for each logical qubit.
- (2) **Intra-block operations:** Two-qubit gates can be implemented between any arbitrary pair of logical qubits residing within the same memory-code block.
- (3) **Inter-block operations:** Two-qubit gates can be implemented between logical qubits belonging to distinct blocks, provided those blocks are adjacent on the logical graph.

1. Operation diversity

Since our protocol exclusively utilizes (q)LDPC codes, one might intuitively expect it to require only a constant-degree physical qubit network, i.e., $D_P = O(1)$. In practice, however, a single physical qubit often participates in multiple distinct logical operations across successive circuit layers. This necessitates interactions with various resource states at different stages of the protocol. Consequently, the total set of required physical couplings for a given qubit can accumulate beyond $O(1)$, leading to a total vertex degree that exceeds a strict constant.

As detailed in Table III, the number of distinct logical measurement types, denoted by σ , scales as $O(k)$. A naïve implementation would allocate dedicated ancilla qubits and resource-state preparation circuits for each measurement type, referred to as a resource-state factory in what follows. As illustrated in Fig. 10, each code block would require multiple dedicated factories to support single-block measurement types. Similarly, inter-block measurements would require additional specialized factories. This leads to two major challenges:

- (i) **Qubit overhead:** Supporting $O(k)$ distinct measurement types with dedicated factories would incur a total qubit overhead of $O(k) \times O(Mkd_S^2)$, assuming no reuse of factory qubits. This scaling fundamentally violates the target of constant qubit overhead.

- (ii) **Degree explosion:** Since each code block must interface with every specialized factory (as indicated by the arrows in Fig. 10), the physical vertex degree D_P scales as $O(k)$, preventing the realization on a constant-degree qubit network.

Next, we show how to address these issues.

2. Rotation of logical qubit array

We can significantly reduce the number of operation types by introducing a *logical qubit array rotation* scheme within each memory-code block, which is proposed in Ref. [18]. Initially, $O(k)$ distinct operation types are required to access every logical qubit in a block. However, the same functionality can be achieved using only $O(\text{polylog}(k))$ operation types.

To implement this, we model the logical qubits within a block as a one-dimensional ring with indices from 1 to k . In this configuration, logical measurements are restricted to the logical qubit at the first position. To operate on any other logical qubit, we rotate the ring until the target qubit reaches the first position. By utilizing only $O(\text{polylog}(k))$ elementary rotation types (e.g., rotations by powers of two), we can realize any arbitrary rotation of the ring in at most $O(\text{polylog}(k))$ steps. This introduces an additional $O(d^{o(1)})$ factor to the time overhead but reduces the variety of required physical resource-state factories. As illustrated

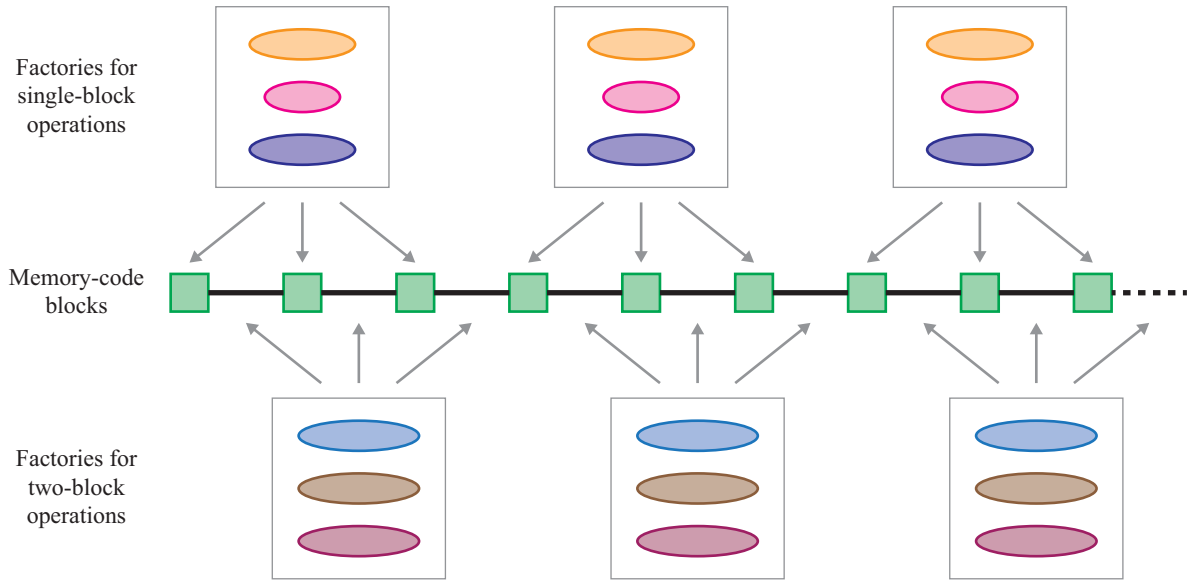


FIG. 10. Network of memory-code blocks and resource-state factories. The architecture is mapped onto an almost-constant-degree physical qubit network where two-block measurements are restricted to pairs connected by black edges. Although illustrated as a one-dimensional chain, the network can be generalized to any graph with a constant vertex degree. Each distinct single-block and two-block measurement type is assigned a dedicated resource-state factory. Consequently, each physical qubit within a data block must potentially interface with every relevant specialized factory, leading to a physical vertex degree that scales linearly with the number of logical measurement types.

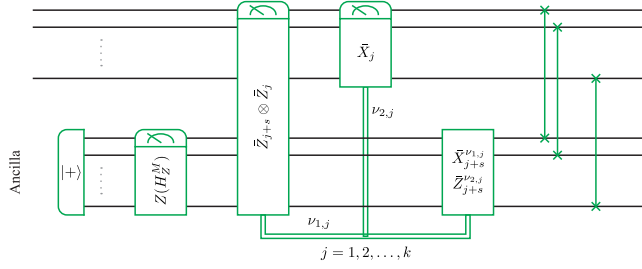


FIG. 11. Logical rotation operation. The state of each j -th logical qubit is transferred to the $(j + s \pmod k)$ th logical qubit.

in Fig. 11, these logical rotations can be implemented efficiently via code surgery.

By adopting this rotation-based protocol, we reduce the total number of operation types to $O(\text{polylog}(k)) = O(d^{o(1)})$, which is almost constant with respect to the code distance.

3. Overhead and connectivity scaling

We define a logical measurement as the pair (σ, B) , where σ denotes the measurement type and B specifies the participating code blocks. Let $LM_\sigma = \{(\sigma, B)\}$ be the set of logical measurements of type σ . Since the logical coupling graph is characterized by a constant degree, it follows that $|LM_\sigma| = O(M)$. Then, we allocate physical qubits as follows:

- (i) **Data and auxiliary blocks:** We utilize M data blocks alongside $O(M)$ auxiliary blocks. Because the underlying memory code has a constant encoding rate, the total number of physical qubits required for these blocks is $O(Mk)$.
- (ii) **Ancilla qubits for deformed codes:** For each measurement type σ , we partition the set LM_σ into $O(|LM_\sigma|/k_R)$ groups, where each group contains up to k_R logical measurements. Each group is assigned a dedicated set of physical qubits to construct the corresponding ancilla system for the deformed code. The number of physical qubits per group is $O(k_R k)$. Given that the rotation scheme reduces the number of distinct measurement types to $O(\text{polylog}(k))$, the total number of physical qubits for all ancilla systems is $O(Mk \text{polylog}(k))$.
- (iii) **Resource-state factories:** For each measurement type σ , we allocate a specialized factory to every set of k_F groups. This requires $O(|LM_\sigma|/(k_F k_R))$ factories to serve all measurements of a given type. Including the qubit cost of surface-code encoding (d_S^2), each factory requires $O(k_F k_R k d_S^2)$ physical qubits. Summing over all types, the total qubit requirement for the factories is $O(Mk d_S^2 \text{polylog}(k))$.

As a result, the total qubit overhead for the protocol is $O(d_S^2 \text{polylog}(k)) = O(d^{o(1)})$, which is almost constant with respect to the code distance.

The time overhead incurs an additional $O(\text{polylog}(k))$ factor from the rotation operations, maintaining an overall complexity of $O(d^{a+o(1)})$.

The connectivity is summarized below:

- (i) **Data and auxiliary blocks:** These physical qubits are coupled to their respective factory qubits. Since the number of distinct measurement types is $O(\text{polylog}(k))$, the vertex degree for these qubits is bounded by $O(\text{polylog}(k))$.
- (ii) **Ancilla qubits for deformed codes:** Each ancilla qubit is coupled to exactly one resource-state factory; thus, its vertex degree is $O(1)$.
- (iii) **Resource-state factories:** Each factory is dedicated to its corresponding code blocks, resulting in a constant vertex degree of $O(1)$.

Consequently, the maximum vertex degree of the physical qubit network is $D_P = O(\text{polylog}(k)) = O(d^{o(1)})$, satisfying the requirement for an almost constant-degree architecture.

APPENDIX F: PARALLELIZED CODE SURGERY—PROOF OF LEMMA 1

Lemma 9. [25] Consider a CSS code (H_X, H_Z, J_X, J_Z) . Let $Z(\alpha J_Z)$ be the set of logical operators to be measured, where $\alpha \in \mathbb{F}_2^{q \times k}$ is a full-rank matrix, and $q \leq k$. We can realize the measurement using a deformed code, which is a CSS code with check matrices

$$H_X^D = \begin{pmatrix} H_X & T \\ 0 & H_M \end{pmatrix} \quad (\text{F1})$$

and

$$H_Z^D = \begin{pmatrix} H_Z & 0 \\ S & H_G^T \end{pmatrix}, \quad (\text{F2})$$

where $S \in \mathbb{F}_2^{n_G \times n}$, $T \in \mathbb{F}_2^{r_X \times r_G}$, $H_G \in \mathbb{F}_2^{r_G \times n_G}$ and $H_G \in \mathbb{F}_2^{r_M \times r_G}$ satisfy the following conditions:

- (i) $H_X S^T = T H_G$;
- (ii) There exists a matrix $R \in \mathbb{F}_2^{n \times n_G}$ such that $\alpha J_Z R S = \alpha J_Z$ and $H_G (\alpha J_Z R)^T = 0$;
- (iii) There exists a matrix $\beta \in \mathbb{F}_2^{(k-q) \times r_G}$ such that $\alpha_\perp J_X S^T = \beta H_G$, where $\alpha_\perp \in \mathbb{F}_2^{(k-q) \times k}$ is a full-rank matrix satisfying $\alpha_\perp \alpha^T = 0$; and
- (iv) $H_M H_G = 0$.

The unmeasured X and Z logical operators are $X(\alpha_\perp J_X)$ and $Z(\alpha_\perp^T J_Z)$, respectively.

Lemma 10. [25] Given an arbitrary set of Z logical operators $Z(\alpha J_Z)$ to be measured, we can generate matrices S , T , and H_G such that code surgery conditions i), ii), and iii) are satisfied, following the devised sticking approach. Furthermore, row and column weights of the matrices H_G , S , and T are bounded above by a constant,

and $\|S\| = 1$. When the generator matrices J_X and J_Z are expressed in standard form, and assuming that the logical operators selected for measurement act on mutually disjoint sets of logical qubits (i.e. the logical thickness is one), the number of rows and columns in H_G scales as $O(k)$.

According to Lemma 10, the check matrices H_X^D and H_Z^D given by Eqs. (3) and (4) have the following properties:

- (i) Row and column weights are bounded above by a constant, i.e. the deformed code is a qLDPC code;
- (ii) The number of rows and columns scales as $O(k_R k)$. According to Eqs. (3) and (4), the row numbers of H_X^D and H_Z^D are $k_R r_X + r_R r_G$ and $k_R r_Z + n_R n_G$, respectively, and their column numbers are both $k_R n + r_R n_G + n_R r_G$. Since $n_G, r_G \leq n$, $r_R \leq n_R$, and $r_X, r_Z \leq n$ (assuming the check matrices are full-rank), all these numbers are upper bounded by $(k_R + 2n_R)n$. Hence, using $n = O(k)$ and $n_R = O(k_R)$, we conclude that the numbers of rows and columns are $O(k_R k)$.

1. Logical measurements associated with the deformed codes

Now, we verify that the deformed code given by Eqs. (3) and (4) is valid for realizing the desired measurement. Let's define matrices

$$\tilde{H}_X = E_{k_R} \otimes H_X, \quad (\text{F3})$$

$$\tilde{H}_Z = E_{k_R} \otimes H_Z, \quad (\text{F4})$$

$$\tilde{J}_X = E_{k_R} \otimes J_X, \quad (\text{F5})$$

$$\tilde{J}_Z = E_{k_R} \otimes J_Z, \quad (\text{F6})$$

$$\tilde{\alpha} = E_{k_R} \otimes \alpha, \quad (\text{F7})$$

$$\tilde{\alpha}_\perp = E_{k_R} \otimes \alpha_\perp, \quad (\text{F8})$$

and

$$\tilde{S} = G_R^T \otimes S, \quad (\text{F9})$$

$$\tilde{T} = (0 \quad G_R^{T^T} \otimes T), \quad (\text{F10})$$

$$\tilde{H}_G = \begin{pmatrix} H_R \otimes E_{n_G} \\ E_{n_R} \otimes H_G \end{pmatrix}, \quad (\text{F11})$$

$$\tilde{H}_M = (E_{r_R} \otimes H_G \quad H_R \otimes E_{r_X}), \quad (\text{F12})$$

$$\tilde{R} = G_R \otimes R, \quad (\text{F13})$$

$$\tilde{\beta} = (0 \quad G_R^T \otimes \beta). \quad (\text{F14})$$

Then, we can find that check matrices in Eqs. (3) and (4) are consistent with Eqs. (F1) and (F2), and the matrices satisfy the code surgery conditions (i), (ii), (iii), and (iv), i.e.

$$\tilde{H}_X \tilde{S}^T = \tilde{T} \tilde{H}_G, \quad (\text{F15})$$

$$\tilde{\alpha} \tilde{J}_Z \tilde{R} \tilde{S} = \tilde{\alpha} \tilde{J}_Z, \quad (\text{F16})$$

$$\tilde{H}_G (\tilde{\alpha} \tilde{J}_Z \tilde{R})^T = 0, \quad (\text{F17})$$

$$\tilde{\alpha}_\perp \tilde{J}_X \tilde{S}^T = \tilde{\beta} \tilde{H}_G, \quad (\text{F18})$$

$$\tilde{H}_M \tilde{H}_G = 0. \quad (\text{F19})$$

Therefore, according to Lemma 9, with this deformed code, we can simultaneously measure logical operators $Z(\tilde{\alpha} \tilde{J}_Z)$.

2. Distances of deformed codes

Without loss of generality, we consider the R-code generator matrix in the form

$$G_R = (E_{k_R} \quad P_R), \quad (\text{F20})$$

and then

$$G_R^T = (E_{k_R} \quad 0). \quad (\text{F21})$$

Lemma 11. Suppose H_R is full-rank. For every vector $v \in \mathbb{F}_2^{r_R n_G + n_R r_X}$ satisfying $\tilde{H}_M v^T = 0$ and $|v| < d_R$, there exists a vector $u_\star \in \mathbb{F}_2^{k_R n_G}$ such that $v^T = \tilde{H}_G u_\star^T$ and

$$|u_\star[(G_R^T)_{\bullet, l} \otimes S]| \leq \|S\| \times |v| \quad (\text{F22})$$

for all $l = 1, 2, \dots, k_R$.

Proof. Let $v = (ab)$. Then, the condition $\tilde{H}_M v^T = 0$ becomes

$$(E_{r_R} \otimes H_G) a^T = (H_R \otimes E_{r_X}) b^T. \quad (\text{F23})$$

Let's introduce matrices A , B , and U such that

$$\text{vec}(A) = a^T, \quad (\text{F24})$$

$$\text{vec}(B) = b^T, \quad (\text{F25})$$

$$\text{vec}(U) = u_\star^T. \quad (\text{F26})$$

We can rewrite the condition as

$$H_G A = B H_R^T. \quad (\text{F27})$$

Similarly, we can rewrite $v^T = \tilde{H}_G u_\star^T$ as

$$A = U H_R^T, \quad (\text{F28})$$

$$B = H_G U. \quad (\text{F29})$$

Since H_R is full-rank, there exists $H_R^r \in \mathbb{F}_2^{n_R \times r_R}$ such that $H_R H_R^r = E_{r_R}$. We remark that $n_R > r_R$ when H_R is full-rank and k_R is not zero. Then, we can take

$$U = A H_R^{r_T} \quad (\text{F30})$$

to satisfy Eq. (F28). Such a matrix U also satisfies the inequality (F22). Note that

$$\begin{aligned} [(G_R^r)_{\bullet, l} \otimes S]^T u_\star^T &= \text{vec}(S^T U (G_R^r)_{\bullet, l}) = (S^T U)_{\bullet, l} \\ &= S^T U_{\bullet, l} = S^T A (H_R^r)^T_{\bullet, l}. \end{aligned} \quad (\text{F31})$$

Therefore,

$$\begin{aligned} |[(G_R^r)_{\bullet, l} \otimes S]^T u_\star^T| &\leq \|S\| \times |A (H_R^r)^T_{\bullet, l}| \leq \|S\| \times |a| \\ &\leq \|S\| \times |v|. \end{aligned} \quad (\text{F32})$$

Now, we prove that the matrix U in Eq. (F30) also satisfies Eq. (F29) when $|v| < d_R$. Using the expression of U , we have

$$H_G U = H_G A H_R^{r_T} = B H_R^T H_R^{r_T}. \quad (\text{F33})$$

Here, we have used Eq. (F27). If

$$B = B H_R^T H_R^{r_T}, \quad (\text{F34})$$

Equation (F30) holds. Note that the right inverse matrix of H_R is not unique. Next, we will prove that there always

exists a right inverse matrix of H_R such that Eq. (F34) is true.

Because $|v| < d_R$, the matrix B has at most $d_R - 1$ non-zero columns. Suppose the indexes of the $d_R - 1$ columns are $i_1, i_2, \dots, i_{d_R-1}$ and all other columns are zero. We introduce the matrices

$$B^{(0)} = (B_{\bullet, i_1} \quad B_{\bullet, i_2} \quad \cdots \quad B_{\bullet, i_{d_R-1}}), \quad (\text{F35})$$

$$H_R^{(0)} = ((H_R)_{\bullet, i_1} \quad (H_R)_{\bullet, i_2} \quad \cdots \quad (H_R)_{\bullet, i_{d_R-1}}). \quad (\text{F36})$$

Later, we will prove that the $d_R - 1$ columns in $H_R^{(0)}$ are linearly independent, and then $r_R \geq d_R - 1$. Because H_R is full-rank and has r_R linearly-independent columns, we can find additional $r_R - (d_R - 1)$ columns to constitute a complete set together with columns in $H_R^{(0)}$. Let $i_{d_R}, i_{d_R+1}, \dots, i_{r_R}$ be indexes of the additional $r_R - (d_R - 1)$ columns, we define

$$H_R^{(1)} = ((H_R)_{\bullet, i_{d_R}} \quad (H_R)_{\bullet, i_{d_R+1}} \quad \cdots \quad (H_R)_{\bullet, i_{r_R}}). \quad (\text{F37})$$

Then, there exists a permutation matrix $\pi \in \mathbb{F}_2^{n_R \times n_R}$ such that matrices B and H_R are in the forms

$$B = (B^{(0)} \quad 0 \quad 0)\pi, \quad (\text{F38})$$

$$H_R = (H_R^{(0)} \quad H_R^{(1)} \quad H_R^{(2)})\pi. \quad (\text{F39})$$

Because $H_R^{(0,1)} = (H_R^{(0)} H_R^{(1)})$ is invertible, a right inverse matrix of H_R is

$$H_R^r = \pi^{-1} \begin{pmatrix} H_R^{(0,1)-1} & 0 \\ 0 & 0 \end{pmatrix} = \pi^{-1} \begin{pmatrix} (H_R^{(0,1)-1})_{1: d_R-1, \bullet} \\ (H_R^{(0,1)-1})_{d_R: r_R, \bullet} \\ 0 \end{pmatrix}. \quad (\text{F40})$$

Note that $\pi^{-1} = \pi^T$. Taking this right inverse matrix, we have

$$B H_R^T H_R^{r_T} = (B^{(0)} \quad 0 \quad 0) \begin{pmatrix} E_{d_R-1} & 0 & 0 \\ 0 & E_{r_R-(d_R-1)} & 0 \\ H_R^{(2)T} [(H_R^{(0,1)-1})_{1: d_R-1, \bullet}]^T & H_R^{(2)T} [(H_R^{(0,1)-1})_{d_R: r_R, \bullet}]^T & 0 \end{pmatrix} \pi = B. \quad (\text{F41})$$

Equation (F34) has been proved.

Lastly, we prove that columns in $H_R^{(0)}$ are linearly independent. Suppose columns in $H_R^{(0)}$ are not linearly independent, there exists a vector $x \in \mathbb{F}_2^{d_R-1} - \{0\}$ such that $H_R^{(0)} x^T = 0$. Then, we can construct a vector $y \in \mathbb{F}_2^{n_R}$,

$$y = (x \ 0 \ 0)\pi. \quad (\text{F42})$$

We have $H_R y^T = 0$, $y \neq 0$, and $|y| < d_R$. This is in contradiction with the code distance of the R code. The lemma has been proved. ■

Lemma 12. For the deformed code with check matrices in Eqs. (3) and (4), the generator matrices of X and Z logical operators are

$$J_X^D = (\tilde{\alpha}_\perp \tilde{J}_X \ \tilde{\beta}) \quad (\text{F43})$$

and

$$J_Z^D = (\tilde{\alpha}_\perp^T \tilde{J}_Z \ 0), \quad (\text{F44})$$

respectively. Let d be the distance of the code (H_X, H_Z, J_X, J_Z) . If H_R is full-rank, the distance of the deformed code satisfies

$$d_D \geq \min\{d/\|S\|, d_R\}. \quad (\text{F45})$$

Proof. The generator matrices satisfy conditions $H_X^D J_Z^{D^T} = H_Z^D J_Z^{D^T} = 0$ and $J_X^D J_Z^{D^T} = E_{k_R(k-q)}$, i.e. they are valid generator matrices of the deformed code.

Let $e = (uv)$ be an X logical error, i.e. $H_Z^D e^T = 0$ and $J_Z^D e^T \neq 0$. Then, we have

$$\tilde{H}_Z u^T = 0 \quad (\text{F46})$$

and

$$\tilde{\alpha}_\perp^T \tilde{J}_Z u^T \neq 0. \quad (\text{F47})$$

Let's introduce the matrix U such that

$$\text{vec}(U) = u^T. \quad (\text{F48})$$

Then,

$$H_Z U = 0 \quad (\text{F49})$$

and

$$\alpha_\perp^T J_Z U \neq 0. \quad (\text{F50})$$

Suppose the m -th column in $\alpha_\perp^T J_Z U$ is not zero, we have

$$H_Z U_{\bullet, m} = 0 \quad (\text{F51})$$

and

$$\alpha_\perp^T J_Z U_{\bullet, m} \neq 0. \quad (\text{F52})$$

According to the code distance d ,

$$|e| \geq |U_{\bullet, m}| \geq d. \quad (\text{F53})$$

Let $e = (uv)$ be a Z logical error, i.e. $H_X^D e^T = 0$ and $J_X^D e^T \neq 0$. Then, we have

$$\tilde{H}_X u^T = \tilde{T} v^T, \quad (\text{F54})$$

$$\tilde{H}_M v^T = 0, \quad (\text{F55})$$

and

$$\tilde{\alpha}_\perp \tilde{J}_X u^T + \tilde{\beta} v^T \neq 0. \quad (\text{F56})$$

If $|v| < d_R$, there exists a vector $u_\star \in \mathbb{F}_2^{k_R n_G}$ such that $v^T = \tilde{H}_G u_\star^T$ and the inequality (F22) holds. Then,

$$\tilde{H}_X u^T = \tilde{T} \tilde{H}_G u_\star^T \quad (\text{F57})$$

and

$$\tilde{\alpha}_\perp \tilde{J}_X u^T + \tilde{\beta} \tilde{H}_G u_\star^T \neq 0. \quad (\text{F58})$$

Furthermore, we have

$$\tilde{H}_X (u + u_\star \tilde{S})^T = 0 \quad (\text{F59})$$

and

$$\tilde{\alpha}_\perp \tilde{J}_X (u + u_\star \tilde{S})^T \neq 0. \quad (\text{F60})$$

Here, we have used Eqs. (F15) and (F18). Let's introduce the matrix U such that

$$\text{vec}(U) = (u + u_\star \tilde{S})^T. \quad (\text{F61})$$

Then,

$$H_X U = 0 \quad (\text{F62})$$

and

$$\alpha_\perp J_X U \neq 0. \quad (\text{F63})$$

Suppose the m -th column in $\alpha_\perp J_X U$ is not zero, we have

$$H_X U_{\bullet, m} = 0 \quad (\text{F64})$$

and

$$\alpha_\perp J_X U_{\bullet, m} \neq 0. \quad (\text{F65})$$

According to the code distance d ,

$$|U_{\bullet,m}| \geq d. \quad (\text{F66})$$

Note that

$$\begin{aligned} U_{\bullet,m} &= U(E_{k_R})_{\bullet,m} = [(E_{k_R})_{m,\bullet} \otimes E_n](u + u_\star \tilde{S})^T \\ &= [(E_{k_R})_{m,\bullet} \otimes E_n]u^T + [(G_R^r)_{\bullet,m} \otimes S]^T u_\star^T, \end{aligned} \quad (\text{F67})$$

therefore,

$$\begin{aligned} \|S\| \times |e| &\geq |u| + \|S\| \times |v| \\ &\geq |[(E_{k_R})_{m,\bullet} \otimes E_n]u^T| + |[(G_R^r)_{\bullet,m} \otimes S]^T u_\star^T| \geq d. \end{aligned} \quad (\text{F68})$$

■

APPENDIX G: LOCALLY-TESTABLE STATE PREPARATION—PROOF OF LEMMA 2

In LTSP, we generate resource states used to measure one type of stabilizer operator—either X or Z —of a CSS code. Without loss of generality, we focus on the measurement of Z stabilizer operators, specifically the operators $Z(H_Z^D)$ in the deformed code. The conclusions are straightforwardly applicable to X stabilizer operators or to other CSS codes, such as the memory code.

1. Stabilizer of the resource state

To implement the measurement of $Z(H_Z^D)$ using gate teleportation, we need to prepare a resource state encoded in two blocks of the deformed code, denoted by B and C . Each block contains n_D qubits. Ideally, the resource state can be prepared as follows; see Fig. 12. First, we generate n_D copies of the Bell state between the two blocks, which constitute a $2n_D$ -qubit stabilizer state, and the stabilizer is generated by operators $X(H_X^{BS})$ and $Z(H_Z^{BS})$, where

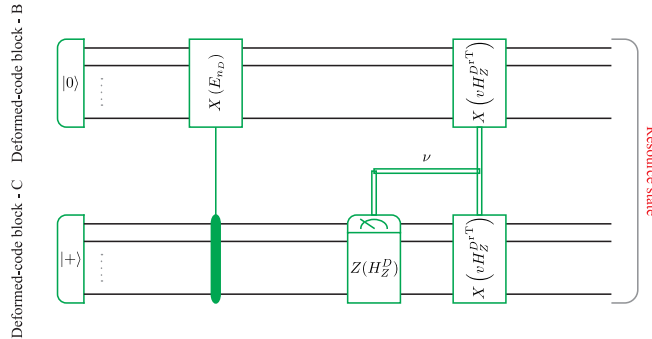


FIG. 12. Circuit that defines the resource state for the $Z(H_Z^D)$ measurement. Outcome of the $Z(H_Z^D)$ measurement in the circuit is $\nu \in \mathbb{F}_2^{r_Z^D}$.

$$H_X^{BS} = H_Z^{BS} = (E_{n_D} \quad E_{n_D}). \quad (\text{G1})$$

We can rewrite the matrices as

$$\tilde{H}_X^{BS} = \begin{pmatrix} H_X^D & H_X^D \\ J_X^D & J_X^D \\ H_Z^{DrT} & H_Z^{DrT} \end{pmatrix} \quad (\text{G2})$$

and

$$\tilde{H}_Z^{BS} = \begin{pmatrix} H_X^{DrT} & H_X^{DrT} \\ J_Z^D & J_Z^D \\ H_Z^D & H_Z^D \end{pmatrix}, \quad (\text{G3})$$

such that $\text{rowsp} H_{X/Z}^{BS} = \text{rowsp} \tilde{H}_{X/Z}^{BS}$. Without loss generality, we can choose H_X^{Dr} and H_Z^{Dr} such that $J_X^D H_X^{Dr} = J_Z^D H_Z^{Dr} = H_X^{DrT} H_Z^{Dr} = 0$. Then, we apply the measurement of operators $Z(H_Z^D)$ on block- C , followed by a feedback Pauli gate. Note that $X(\nu H_Z^{DrT}) Z((H_Z^D)_{j,\bullet}) X(\nu H_Z^{DrT}) = (-1)^{\nu_j} Z((H_Z^D)_{j,\bullet})$. These operations transform Bell states into another stabilizer state with the stabilizer generated by operators $X(H_X^{RS})$ and $Z(H_Z^{RS})$, where

$$H_X^{RS} = \begin{pmatrix} H_X^D & H_X^D \\ J_X^D & J_X^D \end{pmatrix} \quad (\text{G4})$$

and

$$H_Z^{RS} = \begin{pmatrix} E_{n_D} & E_{n_D} \\ 0 & H_Z^D \end{pmatrix}. \quad (\text{G5})$$

The resource state to be prepared is the stabilizer state defined by the generators $X(H_X^{RS})$ and $Z(H_Z^{RS})$.

2. State preparation using a classical locally testable code

We prepare the resource state with the circuit shown in Fig. 13. The circuit constitutes of three sets of F-code blocks, denoted by B , C , and D , containing n_D , n_D , and r_Z^D blocks, respectively. In each F-code block, we encode k_F logical qubits according to the following CSS code: Stabilizer operators are $X(0)$ and $Z(H_F)$, and logical operators are $X(G_F)$ and $Z(G_F^r)$. Without loss of generality, we consider the F-code generator matrix in the form

$$G_F = (E_{k_F} \quad P_F), \quad (\text{G6})$$

and then

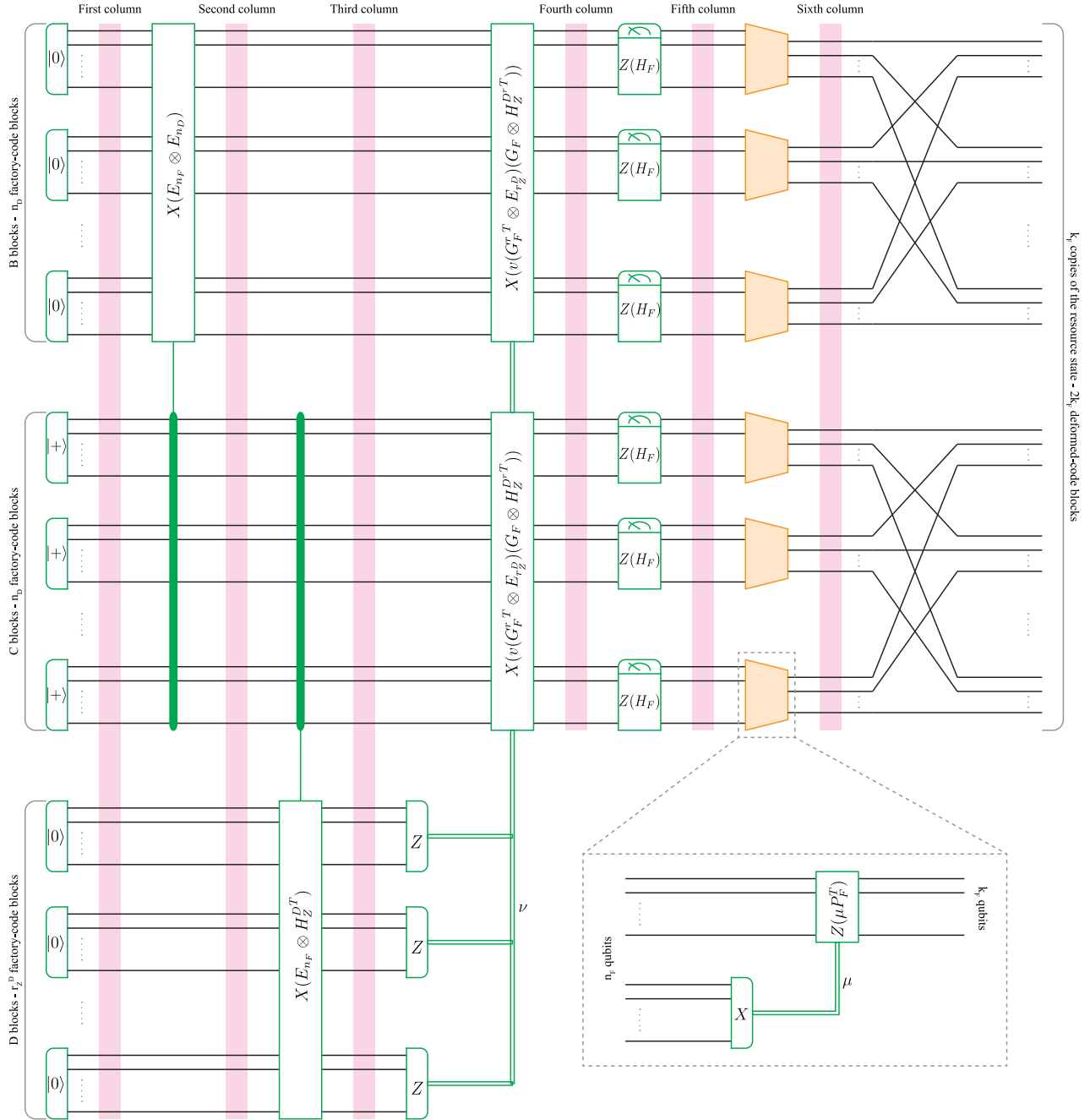


FIG. 13. Locally-testable state preparation. The outcome of the transversal measurement on D blocks is $\nu = \text{vec}(V)^T$, where $V \in \mathbb{F}_2^{r_F^D \times n_F}$, and $V_{i,j}$ is the outcome of the measurement on the j -th qubit in the i -th block. The outcome of the partial transversal measurement on one of B and C blocks in decoding is $\mu \in \mathbb{F}_2^{n_F - k_F}$, and μ_j is the outcome of the measurement on the $(k_F + j)$ -th qubit in the block.

$$G_F^{r_F^T} = (E_{k_F} \quad 0). \quad (\text{G7})$$

This circuit yields n_F copies of the resource state and is supported by the following lemmas.

Lemma 13. Let $|e_Z^{sp}|$ denote the total weight of Z errors occurring at the spacetime locations in the state preparation circuit shown in Fig. 13. Then, each copy of the

prepared resource state contains Z errors with total weight at most $|e_Z^{RS}| \leq |e_Z^{sp}|$.

Lemma 14. Let $|e_X^{sp}|$ denote the total weight of X errors (including incorrect outcomes in projective measurements of Z operators) occurring at the spacetime locations in the state preparation circuit shown in Fig. 13. Suppose the F code is (ω, s) -locally testable, and that H_Z^D has rows and columns of weights at most ω_Z^D . If the spacetime X errors

are undetectable by the two sets of checks illustrated in Fig. 5, and

$$|e_X^{sp}| < \frac{d_F}{\omega_Z^D \max\{1, \frac{n_F}{r_{FS}}\}}, \quad (\text{G8})$$

then each copy of the prepared resource state contains X errors with total weight at most

$$|e_X^{RS}| \leq \max\left\{1, \frac{n_F}{r_{FS}}\right\} \times |e_X^{sp}|. \quad (\text{G9})$$

a. Proof of Lemma 13— X operators and Z errors

Focusing on the j -th copy of the prepared resource state, i.e. j -th logical qubits in B and C F-code blocks, the propagation of X operators in spacetime is described by the generator matrix

$$J_{X,j}^{sp} = (J_{X,B,j}^{sp} \quad J_{X,C,j}^{sp} \quad J_{X,D,j}^{sp}), \quad (\text{G10})$$

where

$$J_{X,B,j}^{sp} = \begin{pmatrix} 0 & (G_F)_{j,\bullet} \otimes H_X^D & (G_F)_{j,\bullet} \otimes H_X^D & (G_F)_{j,\bullet} \otimes H_X^D & (G_F)_{j,\bullet} \otimes H_X^D & (E_{k_F})_{j,\bullet} \otimes H_X^D \\ 0 & (G_F)_{j,\bullet} \otimes J_X^D & (G_F)_{j,\bullet} \otimes J_X^D & (G_F)_{j,\bullet} \otimes J_X^D & (G_F)_{j,\bullet} \otimes J_X^D & (E_{k_F})_{j,\bullet} \otimes J_X^D \end{pmatrix}, \quad (\text{G11})$$

$$J_{X,C,j}^{sp} = \begin{pmatrix} (G_F)_{j,\bullet} \otimes H_X^D & (G_F)_{j,\bullet} \otimes H_X^D & (G_F)_{j,\bullet} \otimes H_X^D & (G_F)_{j,\bullet} \otimes H_X^D & (G_F)_{j,\bullet} \otimes H_X^D & (E_{k_F})_{j,\bullet} \otimes H_X^D \\ (G_F)_{j,\bullet} \otimes J_X^D & (G_F)_{j,\bullet} \otimes J_X^D & (G_F)_{j,\bullet} \otimes J_X^D & (G_F)_{j,\bullet} \otimes J_X^D & (G_F)_{j,\bullet} \otimes J_X^D & (E_{k_F})_{j,\bullet} \otimes J_X^D \end{pmatrix}, \quad (\text{G12})$$

and

$$J_{X,D,j}^{sp} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (\text{G13})$$

correspond to spacetime locations on B , C , and D blocks, respectively: The six columns correspond to spacetime locations as illustrated in Fig. 13. We remark that operators on C blocks do not propagate to D blocks by the second transversal controlled-NOT gate because of $[(G_F)_{j,\bullet} \otimes H_X^D][E_{n_F} \otimes H_Z^{DT}] = [(G_F)_{j,\bullet} \otimes J_X^D][E_{n_F} \otimes H_Z^{DT}] = 0$. This generator matrix describes that operators $X((G_F)_{j,\bullet} \otimes H_X^D)$ and $X((G_F)_{j,\bullet} \otimes J_X^D)$ on C blocks at the beginning are transformed into operators $X((G_F)_{j,\bullet} \otimes H_X^D) \otimes X((G_F)_{j,\bullet} \otimes H_X^D)$ and $X((G_F)_{j,\bullet} \otimes J_X^D) \otimes X((G_F)_{j,\bullet} \otimes J_X^D)$ on B and C blocks at the end, respectively. Because qubits in C blocks are initialized in the state $|+\rangle$, the final state is an eigenstate of operators $X((G_F)_{j,\bullet} \otimes H_X^D) \otimes X((G_F)_{j,\bullet} \otimes H_X^D)$ and $X((G_F)_{j,\bullet} \otimes J_X^D) \otimes X((G_F)_{j,\bullet} \otimes J_X^D)$ on B and C blocks with the eigenvalue $+1$.

We represent Z errors in spacetime with a vector

$$e_Z^{sp} = (e_{Z,B}^{sp} \quad e_{Z,C}^{sp} \quad e_{Z,D}^{sp}), \quad (\text{G14})$$

where

$$e_{Z,B/C}^{sp} = (u_{B/C,1} \quad u_{B/C,2} \quad u_{B/C,3} \quad u_{B/C,4} \quad u_{B/C,5} \quad u_{B/C,6}) \quad (\text{G15})$$

and

$$e_{Z,D}^{sp} = (u_{D,1} \quad u_{D,2} \quad u_{D,3}). \quad (\text{G16})$$

The errors flip operators during the propagation according to

$$\begin{aligned} J_{X,j}^{sp} e_Z^{spT} &= \begin{pmatrix} \text{vec}(H_X^D U[(G_F)_{j,\bullet}]^T) \\ \text{vec}(J_X^D U[(G_F)_{j,\bullet}]^T) \end{pmatrix} \\ &\quad + \begin{pmatrix} \text{vec}(H_X^D U[(E_{k_F})_{j,\bullet}]^T) \\ \text{vec}(J_X^D U[(E_{k_F})_{j,\bullet}]^T) \end{pmatrix} \\ &= \begin{pmatrix} H_X^D u_{\text{eff}}^T \\ J_X^D u_{\text{eff}}^T \end{pmatrix} = H_X^{RS} \begin{pmatrix} 0 \\ u_{\text{eff}}^T \end{pmatrix}, \end{aligned} \quad (\text{G17})$$

where

$$\text{vec}(U) = (u_{B,2} + u_{B,3} + u_{B,4} + u_{B,5} + u_{C,1} + u_{C,2} + u_{C,3} + u_{C,4} + u_{C,5})^T, \quad (\text{G18})$$

$$\text{vec}(U') = (u_{B,6} + u_{C,6})^T, \quad (\text{G19})$$

$$u_{\text{eff}}^T = U[(G_F)_{j,\bullet}]^T + U'[(E_{k_F})_{j,\bullet}]^T. \quad (\text{G20})$$

We can find that the vector

$$e_Z^{RS} = (0 \quad u_{\text{eff}}) \in \mathbb{F}_2^{2n_D} \quad (\text{G21})$$

satisfies $H_X^{RS} e_Z^{RST} = J_X^{sp} e_Z^{spT}$ as illustrated in Eq. (G17). Therefore, e_Z^{RS} represents Z errors on the prepared resource state. According to Eq. (G20), it satisfies the inequality $|e_Z^{RS}| \leq |e_Z^{sp}|$.

b. Proof of Lemma 14—Z operators and X errors

Focusing on the j -th copy of the prepared resource state, i.e. j -th logical qubits in B and C F-code blocks, the propagation of Z operators in spacetime is described by the generator matrix

$$J_{Z,j}^{sp} = \begin{pmatrix} J_{Z,B,j}^{sp} & J_{Z,C,j}^{sp} & J_{Z,D,j}^{sp} & J_{Z,\text{mea},j}^{sp} \end{pmatrix}, \quad (\text{G22})$$

where

$$J_{Z,B,j}^{sp} = \begin{pmatrix} (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (E_{k_F})_{j,\bullet} \otimes E_{n_D} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\text{G23})$$

$$J_{Z,C,j}^{sp} = \begin{pmatrix} 0 & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (G_F^r)^T_{j,\bullet} \otimes E_{n_D} & (E_{k_F})_{j,\bullet} \otimes E_{n_D} \\ 0 & 0 & (G_F^r)^T_{j,\bullet} \otimes H_Z^D & (G_F^r)^T_{j,\bullet} \otimes H_Z^D & (G_F^r)^T_{j,\bullet} \otimes H_Z^D & (E_{k_F})_{j,\bullet} \otimes H_Z^D \end{pmatrix}, \quad (\text{G24})$$

and

$$J_{Z,D,j}^{sp} = \begin{pmatrix} 0 & 0 & 0 \\ (G_F^r)^T_{j,\bullet} \otimes E_{r_Z^D} & (G_F^r)^T_{j,\bullet} \otimes E_{r_Z^D} & (G_F^r)^T_{j,\bullet} \otimes E_{r_Z^D} \end{pmatrix}, \quad (\text{G25})$$

correspond to spacetime locations on B , C , and D blocks, respectively, and

$$J_{Z,\text{mea},j}^{sp} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad (\text{G26})$$

corresponds to measurements of $Z(H_F)$: The two columns correspond to measurements on B and C blocks, respectively. Note that $[(G_F^r)^T_{j,\bullet} \otimes H_Z^D][(G_F^r \otimes E_{r_Z^D})(G_F \otimes H_Z^{DrT})]^T = (G_F^r)^T_{j,\bullet} \otimes E_{r_Z^D}$, which justifies the propagation through the feedback gate depending on the outcome ν . This generator matrix describes that operators $Z((G_F^r)^T_{j,\bullet} \otimes E_{n_D})$ on B blocks and $Z((G_F^r)^T_{j,\bullet} \otimes E_{r_Z^D})$ on D blocks at the beginning are transformed into operators $Z((E_{k_F})_{j,\bullet} \otimes E_{n_D}) \otimes Z((E_{k_F})_{j,\bullet} \otimes E_{n_D})$ on B and C blocks and $Z((E_{k_F})_{j,\bullet} \otimes H_Z^D)$ on C blocks at the end, respectively. Because qubits in C and D blocks are initialized in the state $|0\rangle$, the final state is an eigenstate of

operators $Z((E_{k_F})_{j,\bullet} \otimes E_{n_D}) \otimes Z((E_{k_F})_{j,\bullet} \otimes E_{n_D})$ on B and C blocks and $Z((E_{k_F})_{j,\bullet} \otimes H_Z^D)$ on C blocks with the eigenvalue $+1$.

We detect errors with two sets of checks. First, let $\eta \in \mathbb{F}_2^{r_F}$ be the outcome of a $Z(H_F)$ measurement, it satisfies $\eta \in \text{colsp} H_F$ if the circuit is error-free. Therefore, we can detect errors that flip the measurement outcome through $H_M^F \eta$, where H_M^F is the check matrix of the code $\text{colsp} H_F$. Second, because qubits in C and D blocks are initialized in the state $|0\rangle$, outcomes of $Z(H_F)$ measurements on B and C blocks as well as transversal Z measurements on D blocks satisfy certain equations if the circuit is error-free. Accordingly, the check matrix is

$$H_Z^{sp} = \begin{pmatrix} H_{Z,B}^{sp} & H_{Z,C}^{sp} & H_{Z,D}^{sp} & H_{Z,\text{mea}}^{sp} \end{pmatrix}, \quad (\text{G27})$$

where

$$H_{Z,B}^{sp} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ H_F \otimes E_{n_D} & H_F \otimes E_{n_D} & H_F \otimes E_{n_D} & H_F \otimes E_{n_D} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\text{G28})$$

$$H_{Z,C}^{sp} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & H_F \otimes E_{n_D} & H_F \otimes E_{n_D} & H_F \otimes E_{n_D} & 0 & 0 \\ 0 & 0 & H_F \otimes H_Z^D & H_F \otimes H_Z^D & 0 & 0 \end{pmatrix}, \quad (\text{G29})$$

$$H_{Z,D}^{sp} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ H_F \otimes E_{r_Z^D} & H_F \otimes E_{r_Z^D} & H_F \otimes E_{r_Z^D} \end{pmatrix}, \quad (\text{G30})$$

and

$$H_{Z,\text{mea}}^{sp} = \begin{pmatrix} H_M^F \otimes E_{n_D} & 0 \\ 0 & H_M^F \otimes E_{n_D} \\ E_{r_F} \otimes E_{n_D} & E_{r_F} \otimes E_{n_D} \\ 0 & E_{r_F} \otimes H_Z^D \end{pmatrix}. \quad (\text{G31})$$

We represent X errors in spacetime with a vector

$$e_X^{sp} = (e_{X,B}^{sp} \quad e_{X,C}^{sp} \quad e_{X,D}^{sp} \quad e_{X,\text{mea}}^{sp}), \quad (\text{G32})$$

where

$$e_{X,B/C}^{sp} = (u_{B/C,1} \quad u_{B/C,2} \quad u_{B/C,3} \quad u_{B/C,4} \quad u_{B/C,5} \quad u_{B/C,6}), \quad (\text{G33})$$

$$e_{X,D}^{sp} = (u_{D,1} \quad u_{D,2} \quad u_{D,3}). \quad (\text{G34})$$

and

$$e_{X,\text{mea}}^{sp} = (v_B \quad v_C). \quad (\text{G35})$$

The errors flip operators during the propagation according to

$$J_{Z,j}^{sp} e_X^{spT} = \begin{pmatrix} \text{vec}((U_B + U'_B + U_C + U'_C)[(G_F^T)_{j,\bullet}]^T) \\ \text{vec}((H_Z^D U_C + H_Z^D U'_C + U_D)[(G_F^T)_{j,\bullet}]^T) \\ \text{vec}((U''_B + U''_C)[(E_{k_F})_{j,\bullet}]^T) \\ \text{vec}(H_Z^D U'_C[(E_{k_F})_{j,\bullet}]^T) \end{pmatrix}, \quad (\text{G36})$$

where

$$\text{vec}(U_B) = (u_{B,1} + u_{B,2} + u_{B,3} + u_{B,4} + u_{C,2})^T, \quad (\text{G37})$$

$$\text{vec}(U'_B) = u_{B,5}^T, \quad (\text{G38})$$

$$\text{vec}(U''_B) = u_{B,6}^T, \quad (\text{G39})$$

$$\text{vec}(U_C) = (u_{C,3} + u_{C,4})^T, \quad (\text{G40})$$

$$\text{vec}(U'_C) = u_{C,5}^T, \quad (\text{G41})$$

$$\text{vec}(U''_C) = u_{C,6}^T, \quad (\text{G42})$$

$$\text{vec}(U_D) = (u_{D,1} + u_{D,2} + u_{D,3})^T. \quad (\text{G43})$$

If the spacetime error e_X^{sp} is undetectable by H_Z^{sp} , it satisfies the condition $H_Z^{sp} e_X^{spT} = 0$. Let

$$\text{vec}(V_{B/C}) = v_{B/C}^T, \quad (\text{G44})$$

the condition can be rewritten as

$$V_{B/C} H_M^{FT} = 0, \quad (\text{G45})$$

$$(U_B + U_C) H_F^T = V_B + V_C \quad (\text{G46})$$

$$H_Z^D U_C H_F^T + U_D H_F^T = H_Z^D V_C. \quad (\text{G47})$$

When Eq. (G45) holds, there exists $W_{B/C} \in \mathbb{F}_2^{n_D \times n_F}$ such that $V_{B/C} = W_{B/C} H_F^T$, and

$$|(W_{B/C})_{a,\bullet}| \leq \frac{n_F}{r_{FS}} |(V_{B/C})_{a,\bullet}| \quad (\text{G48})$$

for all $a = 1, 2, \dots, n_D$; see Lemma 4. Then, we have

$$(U_B + U_C) H_F^T = (W_B + W_C) H_F^T, \quad (\text{G49})$$

$$H_Z^D U_C H_F^T + U_D H_F^T = H_Z^D W_C H_F^T. \quad (\text{G50})$$

If the inequality (G8) is true,

$$\begin{aligned} & |(U_B + U_C)_{a,\bullet} + (W_B + W_C)_{a,\bullet}| \\ & \leq \max \left\{ 1, \frac{n_F}{r_{FS}} \right\} \times |e_X^{sp}| \leq d_F \end{aligned} \quad (\text{G51})$$

and

$$\begin{aligned} & |(H_Z^D U_C + U_D)_{b,\bullet} + (H_Z^D W_C)_{a,\bullet}| \\ & \leq \omega_Z^D \max \left\{ 1, \frac{n_F}{r_{FS}} \right\} \times |e_X^{sp}| \leq d_F \end{aligned} \quad (\text{G52})$$

hold for all $a = 1, 2, \dots, n_D$ and $b = 1, 2, \dots, r_Z^D$. Then, according to the F-code distance, we have

$$U_B + U_C = W_B + W_C, \quad (\text{G53})$$

$$H_Z^D U_C + U_D = H_Z^D W_C. \quad (\text{G54})$$

of physical operations. The logical operation has errors only if physical operations have errors: Let w_L and w_P be the weights of logical errors and physical errors, respectively; then $w_L = 0$ if $w_P = 0$. For single-qubit logical operations, i.e. initializations, measurements, and Pauli gates, we have $w_L \leq w_P$. Note that w_L is either zero or one. For a logical controlled-NOT gate, because it acts on two surface-code logical qubits, $w_L \leq 2$, and we have $w_L \leq 2w_P$. Multiple logical controlled-NOT gates constitute the generalized transversal controlled-NOT gates in Fig. 13. The error propagation in the second generalized transversal controlled-NOT gate may transform an error on one surface-code logical qubit into an error on multiple surface-code logical qubits, i.e. increase the weight of logical errors. Because H_Z^D has rows and columns of weight at most ω_Z^D , the logical-error weight is amplified by a factor of at most ω_Z^D . Therefore, for all these logical operations, we have $w_L \leq 2\omega_Z^D w_P$. Decoding operations may also cause errors, which are the same as single-qubit logical operations; they cause extra errors, and we can take into account decoding errors in the sixth column in Fig. 13.

Lemma 15. (Formal version of Lemma 2). Suppose the F code is (ω, s) -locally testable, and H_Z^D has rows and columns of weight at most ω_Z^D . Let $\bar{e}_X^{sp}$ and $\bar{e}_Z^{sp}$ be the X and Z spacetime (physical) errors in the state preparation circuit shown in Fig. 13, when simulated using a surface code of arbitrary code distance. If the spacetime errors are undetectable and satisfy

$$|\bar{e}_X^{sp}| < \frac{d_F}{2\omega_Z^{D^2} \max\left\{1, \frac{n_F}{r_F s}\right\}}, \quad (\text{G59})$$

the following inequalities hold:

$$|e_X^{RS}| \leq 2\omega_Z^D \max\left\{1, \frac{n_F}{r_F s}\right\} \times |\bar{e}_X^{sp}|, \quad (\text{G60})$$

$$|e_Z^{RS}| \leq 2\omega_Z^D |\bar{e}_Z^{sp}|. \quad (\text{G61})$$

Here, $\frac{n_F}{r_F s} \leq (w/s)$, i.e., $w r_F \geq n_F$; otherwise the distance reduces to $d_F = 1$.

4. Overhead in the state preparation

The state preparation circuit shown in Fig. 13 involves $2n_D + r_Z^D$ F-code blocks, so the number of data qubits is $(2n_D + r_Z^D)n_F$. Additionally, we require r_F ancilla qubits to implement parity-check measurements on each F-code block, and such measurements are applied to $2n_D$ blocks. Thus, the total number of qubits is $(2n_D + r_Z^D)n_F + 2n_D r_F = O(n_D n_F)$. In this circuit, all operation layers have depth one, except for two layers of parity-check measurements: the layer between the second and third columns, and the layer between the fourth and fifth columns. According to Lemma 3, each of these layers has depth $O(1)$ because the

deformed code is a qLDPC code and the F code is an LDPC code. Therefore, the overall circuit depth is $O(1)$.

When encoding the state preparation circuit using a surface code, each qubit in the circuit is encoded in a surface code of distance d_S , requiring $(2d_S - 1)^2 = O(d_S^2)$ physical qubits. Since the circuit involves $O(n_D n_F)$ qubits, the total number of physical qubits is $O(n_D n_F d_S^2)$. Each operation on surface-code logical qubits is implemented using a physical circuit of depth $O(d_T)$. Therefore, the overall depth of the physical circuit for state preparation is $O(d_T) = O(d_S)$.

APPENDIX H: PARALLELIZED CODE SURGERY REALIZED THROUGH LOCALLY-TESTABLE STATE PREPARATION

1. Parity-check measurements

With the resource state prepared on blocks B and C , we can transfer the state of block- A to block- C and effectively realize the $Z(H_Z^D)$ measurement; see Fig. 15.

a. X operators and Z errors

The propagation of X operators in spacetime is described by the generator matrix

$$J_X^{pc} = (J_{X,A}^{pc} \quad J_{X,B}^{pc} \quad J_{X,C}^{pc}), \quad (\text{H1})$$

where

$$J_{X,A}^{pc} = \begin{pmatrix} H_X^D & H_X^D \\ J_X^D & J_X^D \end{pmatrix}, \quad (\text{H2})$$

$$J_{X,B}^{pc} = \begin{pmatrix} H_X^D & 0 \\ J_X^D & 0 \end{pmatrix}, \quad (\text{H3})$$

and

$$J_{X,C}^{pc} = \begin{pmatrix} H_X^D & H_X^D & H_X^D \\ J_X^D & J_X^D & J_X^D \end{pmatrix}, \quad (\text{H4})$$

correspond to spacetime locations on A , B , and C blocks, respectively: The three columns correspond to spacetime locations as illustrated in Fig. 15. This generator matrix describes that operators $X(H_X^D) \otimes X(H_X^D) \otimes X(H_X^D)$ and $X(J_X^D) \otimes X(J_X^D) \otimes X(J_X^D)$ on three blocks at the beginning are transformed into operators $X(H_X^D)$ and $X(J_X^D)$ on block- C at the end, respectively. Note that blocks B and C are prepared in the resource state, which is an eigenstate of operators $X(H_X^D) \otimes X(H_X^D)$ and $X(J_X^D) \otimes X(J_X^D)$ with the eigenvalue $+1$. Therefore, such a process effectively transforms operators $X(H_X^D)$ and $X(J_X^D)$ on block- A into operators $X(H_X^D)$ and $X(J_X^D)$ on block- C , respectively.

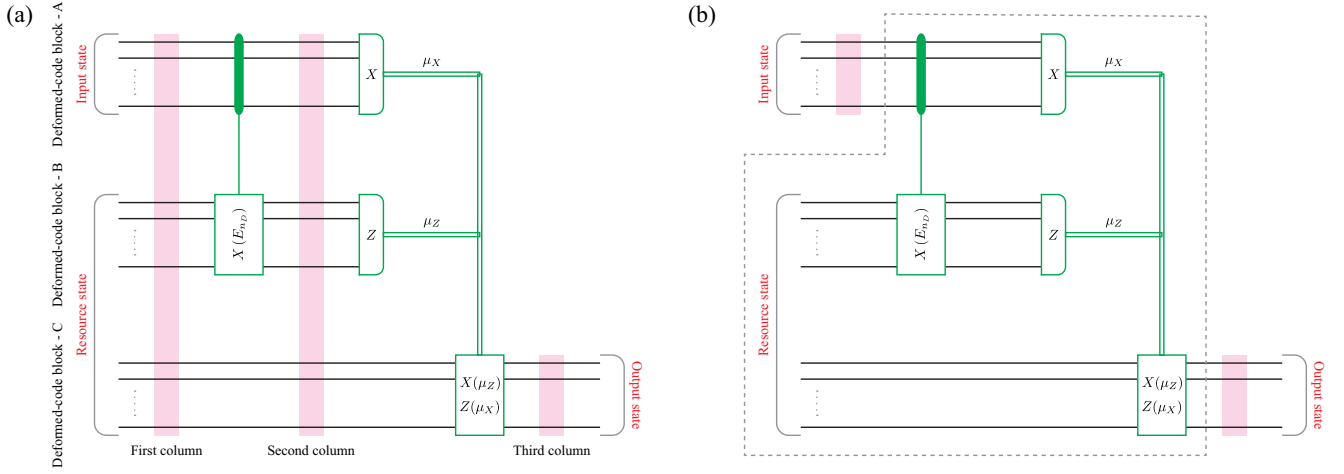


FIG. 15. (a) Measurement of $Z(H_Z^D)$ based on the resource state. Outcomes of transversal measurements on blocks A and B are $\mu_X, \mu_Z \in \mathbb{F}_2^{n_D}$, respectively. Outcome of the $Z(H_Z^D)$ measurement is $H_Z^D \mu_Z$. (b) Effective error in the measurement of $Z(H_Z^D)$ based on the resource state. Operations in the dashed box are effectively error-free.

We represent Z errors in spacetime with a vector

$$e_Z^{pc} = (e_{Z,A}^{pc} \quad e_{Z,B}^{pc} \quad e_{Z,C}^{pc}) \quad (\text{H5})$$

where

$$e_{Z,A/B}^{pc} = (u_{A/B,1} \quad u_{A/B,2}) \quad (\text{H6})$$

and

$$e_{Z,C}^{pc} = (u_{C,1} \quad u_{C,2} \quad u_{C,3}). \quad (\text{H7})$$

The errors flip operators during the propagation according to

$$J_X^{pc} e_Z^{pcT} = \begin{pmatrix} H_X^D \\ J_Z^D \end{pmatrix} u_{\text{eff}}^T, \quad (\text{H8})$$

where

$$u_{\text{eff}} = u_{A,1} + u_{A,2} + u_{B,1} + u_{C,1} + u_{C,2} + u_{C,3}. \quad (\text{H9})$$

Lemma 16. Let e_Z^{pc} be the Z spacetime error in the $Z(H_Z^D)$ measurement based on the resource state. There exists an effective spacetime error in the form

$$e_{Z,\text{eff}}^{pc} = (0 \quad 0 \quad e_{Z,C,\text{eff}}^{pc}) \quad (\text{H10})$$

where

$$e_{Z,C,\text{eff}}^{pc} = (0 \quad 0 \quad u_{\text{eff}}), \quad (\text{H11})$$

such that the effective error $e_{Z,\text{eff}}^{pc}$ is equivalent to the error e_Z^{pc} , i.e. $J_X^{pc} e_{Z,\text{eff}}^{pcT} = J_X^{pc} e_Z^{pcT}$, and $|e_{Z,\text{eff}}^{pc}| \leq |e_Z^{pc}|$. In the effective spacetime error, errors only occur on block-C at the end.

Proof. In the effective spacetime error, we take u_{eff} according to Eq. (H9). Then, $J_X^{pc} e_{Z,\text{eff}}^{pcT} = J_X^{pc} e_Z^{pcT}$ is true, and $|e_{Z,\text{eff}}^{pc}| \leq |e_Z^{pc}|$. ■

b. Z operators and X errors

The propagation of unmeasured Z operators in spacetime is described by the generator matrix

$$J_Z^{pc} = (J_{Z,A}^{pc} \quad J_{Z,B}^{pc} \quad J_{Z,C}^{pc}), \quad (\text{H12})$$

where

$$J_{Z,A}^{pc} = \begin{pmatrix} (H_X^D)^{rT} & 0 \\ J_Z^D & 0 \end{pmatrix}, \quad (\text{H13})$$

$$J_{Z,B}^{pc} = \begin{pmatrix} (H_X^D)^{rT} & (H_X^D)^{rT} \\ J_Z^D & J_Z^D \end{pmatrix}, \quad (\text{H14})$$

and

$$J_{Z,C}^{pc} = \begin{pmatrix} (H_X^D)^{rT} & (H_X^D)^{rT} & (H_X^D)^{rT} \\ J_Z^D & J_Z^D & J_Z^D \end{pmatrix}. \quad (\text{H15})$$

Similar to J_X^{pc} , the generator matrix J_Z^{pc} describes the transformation of operators $Z((H_X^D)^{rT})$ and $Z(J_Z^D)$ on block-A into operators $Z((H_X^D)^{rT})$ and $Z(J_Z^D)$ on block-C, respectively.

The propagation of measured Z operators in spacetime is described by the generator matrix

$$J_{mz}^{pc} = \begin{pmatrix} J_{mz,A}^{pc} & J_{mz,B}^{pc} & J_{mz,C}^{pc} \end{pmatrix} \quad (\text{H16})$$

where

$$J_{mz,A}^{pc} = \begin{pmatrix} H_Z^D & 0 \end{pmatrix}, \quad (\text{H17})$$

$$J_{mz,B}^{pc} = \begin{pmatrix} H_Z^D & H_Z^D \end{pmatrix}, \quad (\text{H18})$$

and

$$J_{mz,C}^{pc} = \begin{pmatrix} H_Z^D & H_Z^D & H_Z^D \end{pmatrix}. \quad (\text{H19})$$

This generator matrix describes the transformation of operators $Z(H_Z^D)$ on block-A into operators $Z(H_Z^D)$ on block-C.

Outcome of the $Z(H_Z^D)$ measurement is justified by the following operator propagation,

$$J_{oc}^{pc} = \begin{pmatrix} J_{oc}^{pc} & J_{oc,B}^{pc} & J_{oc,C}^{pc} \end{pmatrix} \quad (\text{H20})$$

where

$$J_{oc,A}^{pc} = \begin{pmatrix} H_Z^D & 0 \end{pmatrix}, \quad (\text{H21})$$

$$J_{oc,B}^{pc} = \begin{pmatrix} H_Z^D & H_Z^D \end{pmatrix}, \quad (\text{H22})$$

and

$$J_{oc,C}^{pc} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}. \quad (\text{H23})$$

Note that the resource state is an eigenstate of operators $Z(H_Z^D)$ on block-B. This generator matrix describes the transformation of operators $Z(H_Z^D)$ on block-A into operators $Z(H_Z^D)$ on block-B, which is subsequently measured in the transversal measurement on block-B.

We represent X errors in spacetime with a vector

$$e_X^{pc} = \begin{pmatrix} e_{X,A}^{pc} & e_{X,B}^{pc} & e_{X,C}^{pc} \end{pmatrix} \quad (\text{H24})$$

where

$$e_{X,A/B}^{pc} = \begin{pmatrix} u_{A/B,1} & u_{A/B,2} \end{pmatrix} \quad (\text{H25})$$

and

$$e_{X,C}^{pc} = \begin{pmatrix} u_{C,1} & u_{C,2} & u_{C,3} \end{pmatrix}. \quad (\text{H26})$$

The errors flip operators and measurement outcomes during the propagation according to

$$J_Z^{pc} e_X^{pcT} = \begin{pmatrix} (H_X^D)^{rT} \\ J_Z^D \end{pmatrix} (u_{A,\text{eff}} + u_{C,\text{eff}})^T, \quad (\text{H27})$$

$$J_{mz}^{pc} e_X^{pcT} = H_Z^D (u_{A,\text{eff}} + u_{C,\text{eff}})^T, \quad (\text{H28})$$

$$J_{oc}^{pc} e_X^{pcT} = H_Z^D u_{A,\text{eff}}^T, \quad (\text{H29})$$

where

$$u_{A,\text{eff}} = u_{A,1} + u_{B,1} + u_{B,2}, \quad (\text{H30})$$

$$u_{C,\text{eff}} = u_{C,1} + u_{C,2} + u_{C,3}. \quad (\text{H31})$$

Lemma 17. Let e_X^{pc} be the X spacetime error in the $Z(H_Z^D)$ measurement based on the resource state. There exists an effective spacetime error in the form

$$e_{X,\text{eff}}^{pc} = \begin{pmatrix} e_{X,A,\text{eff}}^{pc} & 0 & e_{X,C,\text{eff}}^{pc} \end{pmatrix} \quad (\text{H32})$$

where

$$e_{X,A,\text{eff}}^{pc} = \begin{pmatrix} u_{A,\text{eff}} & 0 \end{pmatrix}, \quad (\text{H33})$$

$$e_{X,C,\text{eff}}^{pc} = \begin{pmatrix} 0 & 0 & u_{C,\text{eff}} \end{pmatrix}, \quad (\text{H34})$$

such that the effective error $e_{X,\text{eff}}^{pc}$ is equivalent to the error e_X^{pc} , i.e.

$$\begin{pmatrix} J_Z^{pc} \\ J_{mz}^{pc} \\ J_{oc}^{pc} \end{pmatrix} e_{X,\text{eff}}^{pcT} = \begin{pmatrix} J_Z^{pc} \\ J_{mz}^{pc} \\ J_{oc}^{pc} \end{pmatrix} e_X^{pcT}, \quad (\text{H35})$$

and $|e_{X,\text{eff}}^{pc}| \leq |e_X^{pc}|$. In the effective spacetime error, errors only occur on block-A at the beginning and block-C at the end.

Proof. In the effective spacetime error, we take $u_{A,\text{eff}}$ and $u_{C,\text{eff}}$ according to Eqs. (H30) and (H31). Then, Eq. (H35) is true, and $|e_{X,\text{eff}}^{pc}| \leq |e_X^{pc}|$. ■

2. Code surgery

The circuit of code surgery is illustrated in Fig. 16, which is applied on k_F blocks of the code (H_X, H_Z, J_X, J_Z) , called memory system. Check and generator matrices of the memory system are $(\tilde{H}_X, \tilde{H}_Z, \tilde{J}_X, \tilde{J}_Z)$. This code surgery realizes the measurement of $Z(\tilde{\alpha}\tilde{J}_Z)$.

The input state of the memory system may carry errors. We model them as errors occurring at locations on the memory system in the first column; see Fig. 16. Accordingly, the effective input state (the state of the memory system before the first column) is in the codeword subspace defined by stabilizer operators $X(\tilde{H}_X)$ and $Z(\tilde{H}_Z)$.

We realize the measurements of $Z(H_Z^D)$ and $X(\tilde{H}_X)$ with corresponding resource states. According to Lemmas 16

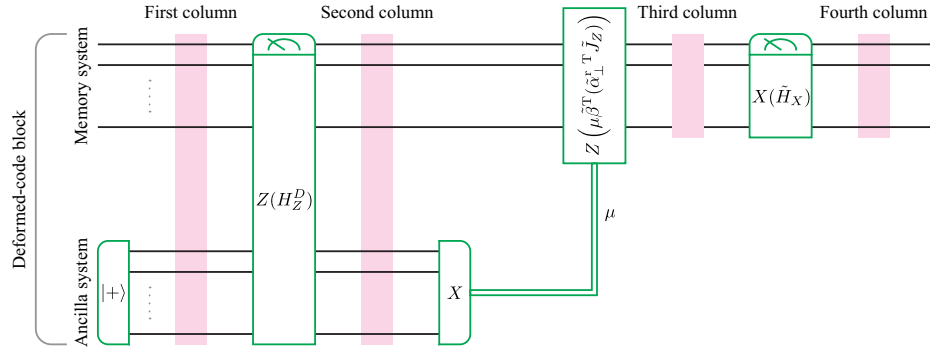


FIG. 16. Code surgery.

and 17, errors effectively occur on the input and output states of these measurements, i.e. outcomes of these measurements are effectively error-free; see Fig. 15(b).

a. X operators and Z errors

The propagation of X operators in spacetime is described by the generator matrix

$$J_X^{cs} = (J_{X,M}^{cs} \quad J_{X,A}^{cs} \quad J_{X,\text{mea}}^{cs}), \quad (\text{H36})$$

where

$$J_{X,M}^{cs} = (\tilde{\alpha}_\perp \tilde{J}_X \quad \tilde{\alpha}_\perp \tilde{J}_X \quad \tilde{\alpha}_\perp \tilde{J}_X \quad \tilde{\alpha}_\perp \tilde{J}_X) \quad (\text{H37})$$

and

$$J_{X,A}^{cs} = (\tilde{\beta} \quad \tilde{\beta}) \quad (\text{H38})$$

correspond to spacetime locations on the memory and ancilla systems, respectively, and

$$J_{X,\text{mea}}^{cs} = 0 \quad (\text{H39})$$

corresponds to the measurement of $X(\tilde{H}_X)$. The columns in $J_{X,M}^{cs}$ and $J_{X,A}^{cs}$ correspond to spacetime locations as illustrated in Fig. 16. This generator matrix describes that operators $X(\tilde{\alpha}_\perp \tilde{J}_X) \otimes X(\tilde{\beta})$ on two systems at the beginning are transformed into operators $X(\tilde{\alpha}_\perp \tilde{J}_X)$ on the memory system at the end. Note that operators $X(\tilde{\alpha}_\perp \tilde{J}_X) \otimes X(\tilde{\beta}) = X(J_X^D)$ commute with the measurement of $Z(H_Z^D)$. Because qubits in the ancilla system are prepared in the state $|+\rangle$, such a process effectively implements a trivial transformation on operators $X(\tilde{\alpha}_\perp \tilde{J}_X)$.

We detect errors in the following way. Let $\tilde{\nu} \in \mathbb{F}_2^{k_{FrX}}$ be the measurement outcome of $X(\tilde{H}_X)$. We introduce the vector

$$\nu = (\tilde{\nu} + \mu \tilde{T}^T \quad \mu \tilde{H}_M^T). \quad (\text{H40})$$

Then, $(-1)^{\nu_j}$ is the eigenvalue of the operator $X((H_X^D)_{j,\bullet})$. Note that operators $X(\tilde{H}_X)$ $[X(H_X^D)]$ commute with the feedback gate [measurement of $Z(H_X^D)$]. Therefore, we have $\nu = 0$ if the circuit is error-free. Accordingly, the check matrix is

$$H_X^{cs} = (H_{X,M}^{cs} \quad H_{X,A}^{cs} \quad H_{X,\text{mea}}^{cs}), \quad (\text{H41})$$

where

$$H_{X,M}^{cs} = \begin{pmatrix} \tilde{H}_X & \tilde{H}_X & \tilde{H}_X & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\text{H42})$$

$$H_{X,A}^{cs} = \begin{pmatrix} \tilde{T} & \tilde{T} \\ \tilde{H}_M & \tilde{H}_M \end{pmatrix}, \quad (\text{H43})$$

and

$$H_{X,\text{mea}}^{cs} = \begin{pmatrix} E_{k_{FrX}} \\ 0 \end{pmatrix}. \quad (\text{H44})$$

We represent Z errors in spacetime with the vector

$$e_Z^{cs} = (e_{Z,M}^{cs} \quad e_{Z,A}^{cs} \quad 0) \quad (\text{H45})$$

represents errors before the $X(\tilde{H}_X)$ measurement, where

$$e_{Z,M}^{cs} = (u_{M,1} \quad u_{M,2} \quad u_{M,3} \quad u_{M,4}) \quad (\text{H46})$$

and

$$e_{Z,A}^{cs} = (u_{A,1} \quad u_{A,2}). \quad (\text{H47})$$

We decompose this spacetime error into two components:

$$e_Z^{cs-b} = (u_{M,1} \quad u_{M,2} \quad u_{M,3} \quad 0 \quad u_{A,1} \quad u_{A,2} \quad 0) \quad (\text{H48})$$

represents errors occurring before the $X(\tilde{H}_X)$ measurement, and

$$e_Z^{cs-a} = (0 \quad 0 \quad 0 \quad u_{M,4} \quad 0 \quad 0 \quad 0) \quad (\text{H49})$$

represents errors occurring after the $X(\tilde{H}_X)$ measurement. Note that $e_Z^{cs} = e_Z^{cs-b} + e_Z^{cs-a}$. The errors flip operators during the propagation according to

$$J_X^{cs} e_Z^{csT} = J_X^D u_{\text{eff}}^T + (\tilde{\alpha}_\perp \tilde{J}_X) u_{\text{res}}^T, \quad (\text{H50})$$

where

$$u_{\text{eff}} = (u_{M,1} + u_{M,2} + u_{M,3} \quad u_{A,1} + u_{A,2}), \quad (\text{H51})$$

$$u_{\text{res}} = u_{M,4}. \quad (\text{H52})$$

If the spacetime error e_Z^{cs} is undetectable through H_X^{cs} , it satisfies the condition $H_X^{cs} e_Z^{csT} = 0$. This condition can be rewritten as $H_X^D u_{\text{eff}}^T = 0$.

Lemma 18. Let e_Z^{cs-b} and e_Z^{cs-a} be the spacetime error vectors representing Z errors that occur before and after the $X(\tilde{H}_X)$ measurement, respectively, during code surgery. If errors are undetectable and satisfy $|e_Z^{cs-b}| < d_D$, then the weight of the residual Z errors on the output state of the memory system is upper bounded by $|e_Z^{cs-p}|$.

Proof. Note that d_D is the distance of the deformed code. Because $|u_{\text{eff}}| \leq |e_Z^{cs-b}| < d_D$ and $H_X^D u_{\text{eff}}^T = 0$, we have $J_X^D u_{\text{eff}}^T = 0$. Then,

$$J_X^{cs} e_Z^{csT} = (\tilde{\alpha}_\perp \tilde{J}_X) u_{\text{res}}^T, \quad (\text{H53})$$

i.e. the vector u_{res} represents the residual errors, and $|u_{\text{res}}| = |e_Z^{cs-a}|$. ■

b. Z operators and X errors

The propagation of unmeasured Z operators in spacetime is described by the generator matrix

$$J_Z^{cs} = (J_{Z,M}^{cs} \quad J_{Z,A}^{cs} \quad J_{Z,\text{mea}}^{cs}), \quad (\text{H54})$$

where

$$J_{Z,M}^{cs} = (\tilde{\alpha}_\perp^T \tilde{J}_Z \quad \tilde{\alpha}_\perp^T \tilde{J}_Z \quad \tilde{\alpha}_\perp^T \tilde{J}_Z \quad \tilde{\alpha}_\perp^T \tilde{J}_Z) \quad (\text{H55})$$

and

$$J_{Z,A}^{cs} = (0 \quad 0) \quad (\text{H56})$$

correspond to spacetime locations on the memory and ancilla systems, respectively, and

$$J_{Z,\text{mea}}^{cs} = 0 \quad (\text{H57})$$

corresponds to the measurement of $Z(H_Z^D)$. Note that operators $Z(\tilde{\alpha}_\perp^T \tilde{J}_Z)$ commute with the measurement of $X(\tilde{H}_X)$.

Similarly, the propagation of measured Z operators in spacetime is described by the generator matrix

$$J_{mz}^{cs} = (J_{mz,M}^{cs} \quad J_{mz,A}^{cs} \quad J_{mz,\text{mea}}^{cs}), \quad (\text{H58})$$

where

$$J_{mz,M}^{cs} = (\tilde{\alpha} \tilde{J}_Z \quad \tilde{\alpha} \tilde{J}_Z \quad \tilde{\alpha} \tilde{J}_Z \quad \tilde{\alpha} \tilde{J}_Z), \quad (\text{H59})$$

$$J_{mz,A}^{cs} = (0 \quad 0), \quad (\text{H60})$$

and

$$J_{mz,\text{mea}}^{cs} = 0. \quad (\text{H61})$$

Outcome of the $Z(\tilde{\alpha} \tilde{J}_Z)$ measurement is described by the following generator matrix,

$$J_{oc}^{cs} = (J_{oc,M}^{cs} \quad J_{oc,A}^{cs} \quad J_{oc,\text{mea}}^{cs}), \quad (\text{H62})$$

where

$$J_{oc,M}^{cs} = (\tilde{\alpha} \tilde{J}_Z \quad 0 \quad 0 \quad 0), \quad (\text{H63})$$

$$J_{oc,A}^{cs} = (0 \quad 0), \quad (\text{H64})$$

and

$$J_{oc,\text{mea}}^{cs} = \tilde{\alpha} \tilde{J}_Z \tilde{R} \gamma_2. \quad (\text{H65})$$

Here,

$$\gamma_1 = (E_{k_R r_Z} \quad 0) \quad (\text{H66})$$

and

$$\gamma_2 = (0 \quad E_{r_Z^D - k_R r_Z}). \quad (\text{H67})$$

Note that the measurement of $Z(H_Z^D)$ measures operators $Z(\tilde{\alpha} \tilde{J}_Z)$ because

$$\tilde{\alpha} \tilde{J}_Z \tilde{R} \gamma_2 H_Z^D = (\tilde{\alpha} \tilde{J}_Z \quad 0). \quad (\text{H68})$$

Let $\nu \in \mathbb{F}_2^{r_Z^D}$ be the outcome of the $Z(H_Z^D)$ measurement. Then, $\nu \gamma_{1/2}^T$ is the outcome of $Z(\gamma_{1/2} H_Z^D)$ operators, and $\nu(\tilde{\alpha} \tilde{J}_Z \tilde{R} \gamma_2)^T$ is the outcome of $Z(\tilde{\alpha} \tilde{J}_Z)$ operators.

If the circuit is error-free, we have $\nu \gamma_1^T = 1$. Accordingly, the check matrix is

$$H_Z^{cs} = (H_{Z,M}^{cs} \quad H_{Z,A}^{cs} \quad H_{Z,\text{mea}}^{cs}), \quad (\text{H69})$$

where

$$H_{Z,M}^{cs} = (\tilde{H}_Z \ 0 \ 0 \ 0), \quad (H70)$$

$$H_{Z,A}^{cs} = (0 \ 0), \quad (H71)$$

and

$$H_{Z,\text{mea}}^{cs} = \gamma_1. \quad (H72)$$

We represent X errors in spacetime with the vector

$$e_X^{cs} = (e_{X,M}^{cs} \ e_{X,A}^{cs} \ 0) \quad (H73)$$

represents errors before the $X(\tilde{H}_X)$ measurement, where

$$e_{X,M}^{cs} = (u_{M,1} \ u_{M,2} \ u_{M,3} \ u_{M,4}) \quad (H74)$$

and

$$e_{X,A}^{cs} = (u_{A,1} \ u_{A,2}). \quad (H75)$$

Similar to Z errors, we decompose this spacetime error into two components:

$$e_X^{cs-b} = (u_{M,1} \ 0 \ 0 \ 0 \ u_{A,1} \ 0 \ 0) \quad (H76)$$

represents errors occurring before the $Z(H_Z^D)$ measurement, and

$$e_X^{cs-a} = (0 \ u_{M,2} \ u_{M,3} \ u_{M,4} \ 0 \ u_{A,2} \ 0) \quad (H77)$$

represents errors occurring after the $Z(H_Z^D)$ measurement. The errors flip operators and measurement outcomes during the propagation according to

$$J_Z^{cs} e_X^{csT} = (\tilde{\alpha}_\perp^T \tilde{J}_Z) u_{\text{eff}}^T + (\tilde{\alpha}_\perp^T \tilde{J}_Z) u_{\text{res}}^T, \quad (H78)$$

$$J_{mz}^{cs} e_X^{csT} = (\tilde{\alpha} \tilde{J}_Z) u_{\text{eff}}^T + (\tilde{\alpha} \tilde{J}_Z) u_{\text{res}}^T, \quad (H79)$$

$$J_{oc}^{cs} e_X^{csT} = (\tilde{\alpha} \tilde{J}_Z) u_{\text{eff}}^T, \quad (H80)$$

where

$$u_{\text{eff}} = u_{M,1}, \quad (H81)$$

$$u_{\text{res}} = u_{M,2} + u_{M,3} + u_{M,4}. \quad (H82)$$

If the spacetime error e_X^{cs} is undetectable through H_Z^{cs} , it satisfies the condition $H_Z^{cs} e_X^{csT} = 0$. This condition can be rewritten as $\tilde{H}_Z u_{\text{eff}}^T = 0$.

Lemma 19. Let e_X^{cs-b} and e_X^{cs-a} be the spacetime error vectors representing X errors that occur before and after

the $Z(H_Z^D)$ measurement, respectively, during code surgery. If errors are undetectable and satisfy $|e_X^{cs-b}| < d$, then the measurement outcome of $Z(\tilde{\alpha} \tilde{J}_Z)$ is correct, and the weight of the residual X errors on the output state of the memory system is upper bounded by $|e_X^{cs-p}|$.

Proof. Note that d is the distance of the code (H_X, H_Z, J_X, J_Z) , which is also the distance of the code $(\tilde{H}_X, \tilde{H}_Z, \tilde{J}_X, \tilde{J}_Z)$. Because $|u_{\text{eff}}| \leq |e_X^{cs-b}| < d$ and $\tilde{H}_Z u_{\text{eff}}^T = 0$, we have $\tilde{J}_Z u_{\text{eff}}^T = 0$. Then,

$$J_Z^{cs} e_X^{csT} = (\tilde{\alpha}_\perp^T \tilde{J}_Z) u_{\text{res}}^T, \quad (H83)$$

$$J_{mz}^{cs} e_X^{csT} = (\tilde{\alpha} \tilde{J}_Z) u_{\text{res}}^T, \quad (H84)$$

$$J_{oc}^{cs} e_X^{csT} = 0, \quad (H85)$$

i.e. the measurement outcome is correct, and the vector u_{res} represents the residual errors. Regarding the weight, $|u_{\text{res}}| \leq |e_X^{cs-a}|$. ■

APPENDIX I: THRESHOLD THEOREM

In this section, we provide the proof of the threshold theorem (Theorem 1) for our protocol. We begin by defining the underlying stochastic noise model. Subsequently, we analyze the primitive circuits by leveraging the technical lemmas established in Appendices G and H. Building on these results, we finally establish and prove the existence of a fault-tolerant threshold.

1. Noise model

Definition 2. Standard local stochastic noise model. Let C be the set of operations constituting a quantum circuit. Each element in C is also referred to as a **location**. A fault configuration is described by a random subset $F \subseteq C$, representing locations where operations are replaced by faulty versions. Such a set F is referred to as a **fault path**. The local stochastic noise model specifies that for any subset of locations $S \subseteq C$, the probability that S is contained within the random fault path F satisfies:

$$\Pr[S \subseteq F] \leq \prod_{j \in S} p_j, \quad (I1)$$

where each location $j \in C$ has an associated error probability p_j .

We establish the fault-tolerance theorem under the local stochastic noise model. To enhance clarity, we first present the derivation using an intuitive, simplified noise model

before demonstrating that the theorem remains valid under the standard local stochastic model.

a. Generalized local stochastic noise model

As a proof technique, we introduce a *generalized local stochastic noise model* to capture correlated errors and the propagation of faults. In this framework, we introduce *virtual locations* where faults occur and subsequently propagate to *real locations*.

Definition 3. ν -confined generalized local stochastic noise model. The noise model is specified by the tuple $(A, B, f, p, \delta, \text{con})$, where A denotes the set of real locations, B is the set of virtual locations, and $f: A \rightarrow \mathcal{P}(B)$ is a mapping to the power set of B that describes the propagation of faults. We define the following notations to characterize this mapping:

- (i) For any subset $S \subseteq A$, let $f_{\cup}(S) \equiv \bigcup_{a \in S} f(a)$ denote the set of all virtual locations that can influence the real locations in S .
- (ii) Let $\|f\|_{\rightarrow} \equiv \max_{a \in A} |f(a)|$ be the maximum number of virtual locations associated with any single real location.
- (iii) Let $\|f\|_{\leftarrow} \equiv \max_{b \in B} |\{a \in A | b \in f(a)\}|$ be the maximum number of real locations influenced by any single virtual location.
- (iv) A mapping f is said to be *disjoint* if for any $a, a' \in A$ with $a \neq a'$, the sets $f(a)$ and $f(a')$ are disjoint, i.e., $f(a) \cap f(a') = \emptyset$.

The model is said to be ν -confined if the mapping f satisfies $\|f\|_{\leftrightarrow} \leq \nu$, where $\|f\|_{\leftrightarrow} \equiv \max\{\|f\|_{\rightarrow}, \|f\|_{\leftarrow}\}$.

Faults at virtual locations are distributed according to a standard local stochastic noise model, with error probabilities specified by the function $p: B \rightarrow [0, 1]$. Faults at real locations arise from two sources:

- (1) **Intrinsic faults:** Conditioned on the event con , at most δ faults may occur directly at real locations.
- (2) **Propagated faults:** A virtual fault at location $b \in B$ may induce a fault at any real location $a \in A$ for which $b \in f(a)$.

Formally, let $F_A \subseteq A$ and $F_B \subseteq B$ denote the real and virtual fault paths, respectively. For any subset of virtual locations $S \subseteq B$, the probability that S is contained within the virtual fault path F_B satisfies:

$$\Pr[S \subseteq F_B] \leq \prod_{b \in S} p_b \leq p^{|S|}. \quad (I2)$$

The real fault path F_A is always contained within the union of two sub-paths, $F_A \subseteq F_{A,I} \cup F_{A,P}$, where $F_{A,I}$ and $F_{A,P}$ represent the intrinsic and propagated fault paths, respectively. Conditioned on the event con , the intrinsic fault path satisfies

$$|F_{A,I}| \leq \delta. \quad (I3)$$

The propagated fault path satisfies

$$F_{A,P} \subseteq \{a \in A : f(a) \cap F_B \neq \emptyset\}. \quad (I4)$$

For brevity, we often denote the generalized noise model as $(A, B, f, p, \delta, \text{con})$, replacing the mapping p with its upper bound p .

The standard local stochastic noise model is a 1-confined generalized local stochastic noise model given by $(A = C, B = C, f = \text{identity}, p, \delta = 0, \text{con} = \text{trivial})$.

Lemma 20. Probability bound under the generalized noise model. Let $(A, B, f, p, \delta, \text{con})$ be a ν -confined generalized local stochastic noise model. Let $F_A \subseteq A$ be the real fault path. Conditioned on the event con , for any subset of real locations $S \subseteq A$ with cardinality $s = |S| \geq \delta$, the probability that S contains at least $t \geq \delta$ faults satisfies:

$$\Pr[|S \cap F_A| \geq t | \text{con}] \leq \left(\frac{\nu s e}{l} p \right)^l, \quad (I5)$$

where $l = \lceil (t - \delta)/\nu \rceil$. Furthermore, if the mapping f is disjoint, the same upper bound holds with $l = \lceil t - \delta \rceil$.

Proof. By definition, the real fault path is partitioned such that $F_A \subseteq F_{A,I} \cup F_{A,P}$, where $F_{A,I}$ and $F_{A,P}$ represent the intrinsic and propagated fault paths, respectively. Conditioned on con , the intrinsic component contributes at most δ faults to the set S , i.e., $|S \cap F_{A,I}| \leq \delta$.

Consequently, to satisfy the condition $|S \cap F_A| \geq t$, the propagated component must account for the remaining faults:

$$|S \cap F_{A,P}| \geq t - |S \cap F_{A,I}| \geq t - \delta. \quad (I6)$$

Since each virtual fault in B can influence at most ν real locations ($\|f\|_{\leftarrow} \leq \nu$), achieving at least $t - \delta$ propagated faults in S requires at least $l = \lceil (t - \delta)/\nu \rceil$ distinct faults in the virtual set $f_{\cup}(S)$.

The probability of having at least l faults in the virtual set is upper bounded by

$$\Pr[|f_{\cup}(S) \cap F_B| \geq l] \leq \binom{|f_{\cup}(S)|}{l} p^l \leq \left(\frac{|f_{\cup}(S)| e}{l} p \right)^l. \quad (I7)$$

Using the confinement property $\|f\|_{\rightarrow} \leq \nu$, we have $|f_{\cup}(S)| \leq \nu s$. Substituting this into the inequality yields the inequality in the lemma.

Now, consider the case where the mapping f is disjoint. Because the images $\{f(a) : a \in S\}$ are mutually disjoint, each virtual fault can influence at most one real location in S . Achieving at least $t - \delta$ propagated faults in S necessitates at least $l' = \lceil t - \delta \rceil$ virtual faults, which must be

distributed across l' distinct subsets $f(a)$ with $a \in S$. This probability is upper-bounded by:

$$\begin{aligned} \Pr[|\{f(a) \cap F_B \neq \emptyset : a \in S\}| \geq l'] &\leq \binom{s}{l'} \nu^{l'} p^{l'} \\ &\leq \left(\frac{s e}{l'} \nu p\right)^{l'}. \end{aligned} \quad (18)$$

■

Lemma 21. Merging of generalized noise models. Consider two generalized noise models $(A_j, B_j, f_j, p_{j,\bullet}, \delta_j, \text{con}_j)$ for $j \in \{1, 2\}$, where the j th model is ν_j -confined. Suppose two virtual location sets are disjoint, i.e., $B_1 \cap B_2 = \emptyset$. Let A be a set of locations, and assume there exist merging mappings $g_j: A \rightarrow \mathcal{P}(A_j)$ such that an arbitrary location $a \in A$ is faulty only if $g_1(a) \cup g_2(a)$ contains at least one fault.

Then, the effective noise on A is described by a ν -confined model $(A, B, f, p_\bullet, \delta, \text{con})$, where:

- (i) $B = B_1 \cup B_2$ and $f(a) = f_{1,\cup} \circ g_1(a) \cup f_{2,\cup} \circ g_2(a)$;
- (ii) $p_\bullet|_{B_j} = p_{j,\bullet}$ for $j \in \{1, 2\}$, and $p = \max\{p_1, p_2\}$;
- (iii) $\delta = \delta_1 \|g_1\|_{\leftarrow} + \delta_2 \|g_2\|_{\leftarrow}$;
- (iv) $\text{con} = \text{con}_1 \cap \text{con}_2$;
- (v) $\nu = \nu_1 \|g_1\|_{\leftrightarrow} + \nu_2 \|g_2\|_{\leftrightarrow}$.

If f_1, f_2, g_1 , and g_2 are disjoint, f is disjoint.

Proof. Let $F_{j,A}, F_{j,B}, F_{j,I}$, and $F_{j,P}$ denote the real, virtual, intrinsic, and propagated fault paths of model $j \in \{1, 2\}$, respectively. We use the same notation without the subscript j to represent the fault paths of the resulting effective model on A .

Based on the merging mappings, a location $a \in A$ is faulty only if at least one of its associated locations in A_1 or A_2 is faulty. Formally:

$$F_A \subseteq \{a \in A : g_1(a) \cap F_{1,A} \neq \emptyset \text{ or } g_2(a) \cap F_{2,A} \neq \emptyset\}. \quad (19)$$

Since $F_{j,A} \subseteq F_{j,I} \cup F_{j,P}$, we can decompose the effective fault path into intrinsic and propagated components. We define F_I as the set of real locations in A affected by intrinsic faults in A_1 and A_2 :

$$F_I \subseteq \{a \in A : g_1(a) \cap F_{1,I} \neq \emptyset \text{ or } g_2(a) \cap F_{2,I} \neq \emptyset\}. \quad (110)$$

Conditioned on $\text{con} = \text{con}_1 \cap \text{con}_2$, the cardinality of the intrinsic fault set F_I is bounded by the total number of errors propagated from the intrinsic sets $F_{1,I}$ and $F_{2,I}$, i.e., $|F_I| \leq \delta_1 \|g_1\|_{\leftarrow} + \delta_2 \|g_2\|_{\leftarrow} = \delta$.

Next, we characterize the propagated fault path F_P . Using the definition of propagated faults in the constituent models, $F_{j,P} \subseteq \{a \in A_j : f_j(a_j) \cap F_{j,B} \neq \emptyset\}$, we have:

$$\begin{aligned} F_P \subseteq \{a \in A : f_{1,\cup} \circ g_1(a) \cap F_{1,B} \neq \emptyset \text{ or } \\ f_{2,\cup} \circ g_2(a) \cap F_{2,B} \neq \emptyset\}. \end{aligned} \quad (111)$$

This expression justifies the definitions of the virtual location set, mapping function, and error probability for the effective model. Specifically, the virtual fault path is given by $F_B = F_{1,B} \cup F_{2,B}$. Finally, the confinement parameter ν is obtained by calculating the norms $\|f\|_{\rightarrow}$ and $\|f\|_{\leftarrow}$ based on the individual norms of f_j and g_j , yielding the result stated in the lemma. ■

Remark. This result can be generalized to the merging of N models with disjoint virtual location sets. For instance, if $\|g_j\|_{\leftrightarrow} = 1$ for all $j \in \{1, \dots, N\}$, the parameters of the resulting effective model are:

$$\begin{aligned} p &= \max\{p_j : j = 1, 2, \dots, N\}, & \delta &= \sum_{j=1}^N \delta_j, \\ \text{con} &= \bigcap_{j=1}^N \text{con}_j, & \nu &= \sum_{j=1}^N \nu_j. \end{aligned}$$

Remark. In all generalized noise models presented in Appendixes I2, I3, and I4, the propagation and merging mapping, i.e., f and g , are all disjoint. Note that a sufficient condition for g to be disjoint is that $\|g\|_{\leftarrow} = 1$, which is satisfied by all merging mappings utilized in our analysis. For a 1-confined model, f is always disjoint.

Lemma 22. Splitting of a generalized noise model. Consider a ν -confined generalized noise model $(A, B, f, p_\bullet, \delta, \text{con})$. If the mapping f is disjoint, then for any arbitrary partition of the real locations $A = A_1 \cup A_2$, the noise in each subset A_j ($j \in \{1, 2\}$) is described by a ν_j -confined model $(A_j, B_j, f_j, p_{j,\bullet}, \delta_j, \text{con}_j)$, characterized by:

- (i) $B_j = f_{\cup}(A_j) \subseteq B$ is the set of associated virtual locations;
- (ii) $f_j = f|_{A_j}$ is the restriction of the mapping f to the domain A_j ;
- (iii) $p_{j,\bullet} = p_\bullet|_{B_j}$ is the restriction of the probability mapping to B_j , with $p_j = p$;
- (iv) $\delta_j = \delta$, $\text{con}_j = \text{con}$, and $\nu_j = \nu$.

Furthermore, the resulting propagation mappings f_1 and f_2 are disjoint, and the virtual location sets B_1 and B_2 are disjoint.

Proof. Given an arbitrary fault path $F_A \subseteq A$ and its decomposition into intrinsic and propagated sub-paths $F_A \subseteq F_I \cup F_P$, this induces a corresponding decomposition of the fault path restricted to each A_j . Specifically, we define $F_j = F_A \cap A_j = (F_I \cap A_j) \cup (F_P \cap A_j)$, where $F_I \cap A_j$ and $F_P \cap A_j$ represent the intrinsic and propagated components of the j -th sub-model, respectively. Under the condition con , the cardinality of the intrinsic

component satisfies $|F_I \cap A_j| \leq |F_I| \leq \delta$. Furthermore, the distribution of the propagated component $F_P \cap A_j$ satisfies the model of A_j described in the lemma. ■

b. Simplified local stochastic noise model

In our protocol, universal quantum computation is implemented through the following primitive circuits:

- (i) **Resource state preparation:** The preparation of gate-teleportation resource states, as depicted in Fig. 13;
- (ii) **Parity-check measurement:** Measurement of stabilizer operators using the prepared resource states, as depicted in Fig. 15(a);
- (iii) **Code surgery:** Measurement of logical operators performed via code surgery, as depicted in Fig. 16;
- (iv) **Additional transversal operations:** Transversal initialization, Pauli and S gates, and measurement of code blocks, as shown in Figs. 8 and 9; the parity-check measurement and code surgery circuits apply specifically to Z -type operators, and X -type operators can be measured by applying transversal Hadamard gates to the code block both before and after a Z -type measurement.

Definition 4. Simplified local stochastic noise model. In this model, we assume that error events occur between transversal operations. To simplify the analysis, we effectively subsume the noise from all additional constant-depth transversal operations into the state preparation and parity-check measurement circuits. Errors in the code surgery circuits are likewise accounted for within these two primary circuits. Specifically, each location is either a qubit-time tuple (q, t) or a measurement operation, defined as follows:

- (i) **State preparation (Fig. 13):** The locations include the qubit-time positions in the six indicated columns and all parity-check measurement operations shown in the circuit. We partition the locations into two disjoint subsets: the *preparationlocations*, consisting of qubit-time locations in the first five columns and all measurement locations, and the *decodinglocations*, consisting of qubit-time locations in the sixth column.
- (ii) **Parity-check measurement [Fig. 15(a)]:** The locations include the qubit-time positions in block A of the first column and blocks A and B of the second column. Errors occurring at other qubit-time positions in the first and second columns are effectively subsumed into the preceding state preparation circuit, while errors in the third column are attributed to the subsequent parity-check measurement circuit.

A fault at a qubit-time location indicates that an error is applied to the qubit at that time; a fault at a measurement location indicates that the measurement outcome is

flipped. In this model, faults are distributed according to a standard local stochastic noise model characterized by a physical error parameter p as follows:

(i) In the state preparation circuits:

- (1) For preparation locations, the error probability is $p_{L,S}(p)$. This represents the logical error probability of a surface-code logical qubit. We assume a threshold $p_{th,S}$ exists such that for $p < p_{th,S}$, this probability is $p_{L,S}(p) \leq e^{-\Theta(d_S)}$, where d_S is the surface-code distance.
- (2) For decoding locations, the error probability is $p_{D,S} \leq \lambda_S p$. This represents the error probability of the decoding operation that transfers the state of a surface-code logical qubit to a physical qubit, where λ_S is a positive constant. We define $\lambda = \max\{1, \lambda_S\}$.

(ii) In the stabilizer measurement circuits:

The error probability for every location is simply p .

This simplified noise model is consistent with the error weight analysis presented in Appendices G and H. However, one specific simplification warrants clarification: we have neglected error propagation during the generalized transversal controlled-NOT gates (including those used for parity-check measurements). We will demonstrate that omitting explicit error propagation does not impact the validity of our final threshold conclusions (see Sec. 15). This is justified by our exclusive use of (q)LDPC codes, where the sparse connectivity of the operations ensures that any error spread remains strictly bounded.

Regarding the surface-code assumptions, we acknowledge that this simplified model neglects potential temporal correlations of surface-code logical errors. However, as the state preparation circuit has constant depth, these correlations remain strictly bounded and do not affect the existence of a finite fault-tolerant threshold (see Sec. 15).

2. Fault-tolerant operations

Lemma 23. Resource state preparation for fault-tolerant operations. Let d be the distance of the memory code. Assume the factory code is an LDPC-type LTC with parameters as specified in Table II. Assume that the circuit noise follows the simplified local stochastic noise model. There exists a decoding algorithm such that the state preparation circuit in Fig. 13, using a factory-code distance $d_F = \Theta(d)$ and surface-code distance $d_S = O(\text{polylog}(n_D n_F))$, satisfies the following:

- (i) The circuit outputs k_F copies of the resource state H_Z^D .
- (ii) Under the simplified local stochastic noise model with physical error rate $p < p_{S,th}$, the circuit fails with probability $\Pr(\text{fail}_{SP}) \leq \exp(-\Omega(d))$.
- (iii) If the circuit succeeds, each copy of the output state is subject to noise characterized as follows:

- (1) Let A_{RS} be the set of physical qubits for a given copy of the state.
- (2) The noise on the output state is effectively characterized by a 1-confined model $(A_{RS}, B_{RS}, f_{RS}, p_{D,S}, t_{RS} = d/16, \text{success}_{SP})$ for some B_{RS} and disjoint f_{RS} .

The virtual location sets associated with the k_F resource state copies are mutually disjoint; furthermore, the virtual location sets associated with resource states generated by different state preparation circuits are also disjoint.

Proof. Let C denote the total set of locations, partitioned into the subsets of preparation locations, C_P , and decoding locations, C_D . Consequently, any fault path $F \subseteq C$ can be partitioned into $F = F_P \cup F_D$, where $F_P = F \cap C_P$ and $F_D = F \cap C_D$ represent preparation and decoding faults, respectively.

The check matrix of the state preparation circuit, H_Z^{sp} , only detects X -type errors at preparation locations. Following Lemma 14, the X -type error vector is $e_X^{sp} \in \mathbb{F}_2^{C_X}$, where $C_X \subseteq C$ is the subset of locations relevant to X -type errors. The support of this vector, restricted to preparation locations, satisfies $\text{supp}_{C_X \cap C_P}(e_X^{sp}) \subseteq C_X \cap F_P$.

Given the observed syndrome $v^T = H_Z^{sp} e_X^{spT}$, the decoding algorithm identifies the minimum weight vector $c \in \mathbb{F}_2^{C_X}$ such that $v^T = H_Z^{sp} c^T$. Since H_Z^{sp} only monitors locations in $C_X \cap C_P$, the weight-minimizing solution satisfies $\text{supp}(c) \subseteq C_X \cap C_P$ and $|c| \leq |\text{supp}_{C_X \cap C_P}(e_X^{sp})| \leq |C_X \cap F_P| \leq |F_P|$.

Error correction is performed by effectively applying X gates on the support of c . Let $F' = F'_P \cup F_U$ be the post-correction fault path, where F'_P denotes post-correction preparation faults. Since $F'_P \subseteq \text{supp}(c) \cup F_P$, $|F'_P| \leq |c| + |F_P| \leq 2|F_P|$.

The circuit is defined to succeed if the total number of preparation faults is smaller than t ($|F_P| < t$), where

$$t = \left\lceil \min \left\{ \frac{t_{RS} + 1}{4 \max \left\{ 1, \frac{n_F}{r_{FS}} \right\}}, \frac{d_F}{2\omega_Z^D \max \left\{ 1, \frac{n_F}{r_{FS}} \right\}} \right\} \right\rceil. \quad (I12)$$

Under this success condition, $|F'_P| < 2t$. By Lemmas 13 and 14, these preparation faults induce X (or Z) errors on at most $t_{RS}/2$ physical qubits in each copy of the output state: the post-correction errors are undetectable (the corresponding error vector is in the kernel of H_Z^{sp}), and the potential occurrence of decoding faults does not interfere with the propagation of errors originating from the preparation locations. Let $F_{RS,P} \subseteq A_{RS}$ be the set of physical qubits having errors from preparation faults. It follows that $|F_{RS,P}| \leq t_{RS}$.

The total number of preparation locations is $O(n_D n_F)$. Using a tail bound for the binomial distribution, the

probability that the circuit fails (i.e., $|F_P| \geq t$) is:

$$\begin{aligned} \Pr(\text{fail}_{SP}) &\leq \binom{O(n_D n_F)}{t} p_{L,S}(p)^{[t]} \\ &\leq \left(\frac{O(n_D n_F) e}{[t]} p_{L,S}(p) \right)^{[t]}. \end{aligned} \quad (I13)$$

By choosing a surface-code distance $d_S = O(\text{polylog}(n_D n_F))$ such that $p_{L,S}(p)$ is sufficiently small, we ensure the term inside the parentheses is less than 1. Given $t_{RS} = d/16$ and $d_F = \Theta(d)$, it follows that $t = \Theta(d)$, yielding a failure probability suppressed exponentially in the memory code distance: $\Pr(\text{fail}_{SP}) \leq e^{-\Omega(d)}$.

Each fault at a decoding location maps directly to a physical qubit in the output state. This correspondence is described by a bijective mapping $f_{RS}: A_{RS} \rightarrow C'_D$, where $C'_D \subseteq C_D$ denotes the subset of decoding locations associated with the physical qubits of a given copy of the output state. Accordingly, we take $B_{RS} = C'_D$ as the set of virtual locations for the resulting output state.

Let $F_{RS,D} \subseteq A_{RS}$ denote the set of output faults originating from the decoding locations. The total output fault path satisfies $F_{RS} \subseteq F_{RS,P} \cup F_{RS,D}$. Within this framework, the preparation faults $F_{RS,P}$ constitute the intrinsic faults, satisfying $|F_{RS,P}| \leq t_{RS}$ conditioned on the event of circuit success. The decoding faults $F_{RS,D}$ represent the propagated faults induced by the virtual fault path $F_D \cap C'_D$. Consequently, the resulting noise on the output state precisely satisfies the generalized stochastic noise model.

Since the virtual locations of each output state copy are the decoding locations on the relevant qubit set, the virtual location sets associated with different copies are mutually disjoint. These virtual locations correspond to the physical components of each preparation circuit; consequently, the virtual location sets for different circuits are disjoint. ■

Lemma 24. Fault-tolerant parity-check measurement. Consider the circuit in Fig. 15(a). Let A_{in} and A_{out} denote the sets of qubits in the input and output states, respectively. Assume that the circuit noise follows the simplified local stochastic noise model. Suppose the resource state is generated by a fault-tolerant resource state preparation circuit depicted by Lemma 23.

Under these conditions, all errors introduced by the noise sources are equivalent to effective errors on the input and output states, characterized as follows:

- (i) The circuit is treated as effectively error-free.
- (ii) The effective noise on the input state is described by a 4-confined model $(A_{\text{in}}, B_{\text{in}}, f_{\text{in}}, \lambda p, t_{RS}, \text{success}_{SP})$ for some B_{in} and disjoint f_{in} .
- (iii) The effective noise on the output state is described by a 2-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, t_{RS}, \text{success}_{SP})$ for some B_{out} and disjoint f_{out} .

The virtual location sets B_{in} and B_{out} are disjoint; furthermore, the virtual location sets associated with different parity-check measurements are also disjoint.

Proof. Let $C_{\alpha,j}$ denote the set of qubit-time locations for block α in the j -th column of the circuit, and define $C_{\text{in}} = \bigcup_{\phi \in \Phi} C_{\phi}$, where the index set is $\Phi = \{(A, 1), (B, 1), (B, 2)\}$. By Lemmas 16 and 17, every single-qubit fault at a location in C_{in} is equivalent to a single-qubit error on the input state, and every single-qubit fault at a location in $C_{A,2}$ is equivalent to a single-qubit error on the output state. Hence every single-qubit fault within the circuit is equivalent to a single-qubit error on either the input or output state.

Partition the resource-state as $A_{RS} = A_{RS,1} \cup A_{RS,2}$, where errors on $A_{RS,1}$ are equivalent to errors at locations $C_{B,1}$ and hence to input errors, while errors on $A_{RS,2}$ are equivalent to errors at locations $C_{C,1}$ and hence to output errors. By Lemma 22, this induces corresponding disjoint virtual location sets $B_{RS,1}$ and $B_{RS,2}$.

According to the above analysis, the error equivalences are described by the following merging mappings: $g_{\phi \in \Phi}: A_{\text{in}} \rightarrow C_{\phi}$, $g_{A,2}: A_{\text{out}} \rightarrow C_{A,2}$, $g_{RS,1}: A_{\text{in}} \rightarrow A_{RS,1}$ and $g_{RS,2}: A_{\text{out}} \rightarrow A_{RS,2}$. All these mappings satisfy $\|g\|_{\leftrightarrow} = 1$. The input and output virtual location sets are $B_{\text{in}} = C_{\text{in}} \cup B_{RS,1}$ and $B_{\text{out}} = C_{A,2} \cup B_{RS,2}$, which are disjoint. By applying Lemma 21 to these parameters, the proof is complete. Note that all the mappings are disjoint.

The sets B_{in} and B_{out} are comprised of locations within the current circuit or virtual locations of the resource state. Since the virtual locations for distinct resource states are mutually disjoint, they are disjoint from the virtual locations for all other parity-check measurements. ■

Remark. While we only explicitly analyze the deformed code parity-check matrix H_2^D in Lemmas 23 and 24, our results and conclusions generalize naturally to all other check matrices utilized in the protocol.

Lemma 25. Lemma 1 in [15]. Consider a set S of t nodes in a graph where each node has degree at most z . Let $M_z(\mathbf{s}, \mathbf{S})$ be the number of sets containing $\mathbf{S}$ and in total $\mathbf{s}$ nodes ($\mathbf{s} - t$ nodes beyond those in $\mathbf{S}$), which form a union of connected clusters, each containing at least one node from S . Then $M_z(\mathbf{s}, \mathbf{S}) \leq e^{t-1}(ze)^{\mathbf{s}-t}$, where e is the base of the natural logarithm.

Lemma 26. Fault-tolerant code surgery. Let d be the distance of the memory code. Consider the code surgery circuit illustrated in Fig. 16. We assume that the deformed code is a qLDPC code constructed such that $n_D = O(\text{poly}(d))$ and $d_D = d$. Let A_{in} and A_{out} denote the sets of qubits in the input and output states, respectively. Assume that the circuit noise follows the simplified local stochastic noise model. Furthermore, assume the noise on the input state is characterized by a ν -confined model $(A_{\text{in}}, B_{\text{in}}, f_{\text{in}}, \lambda p, \eta t_{RS}, \text{success}_{SP})$ for some B_{in} and disjoint f_{in} , where $\nu = O(1)$, B_{in} is disjoint from the virtual

location sets of all other noise sources in the circuit, $\eta < 8$, and success_{SP} denotes the event that all preceding fault-tolerant resource state preparation circuits have been successful. Suppose the parity-check measurements are fault-tolerant as depicted by Lemma 24, and that the condition success_{SP} is satisfied.

Under these conditions, there exists a decoding algorithm such that:

- (i) There exists a finite threshold $\epsilon > 0$ such that when $p < \epsilon$, the error correction fails with probability $\Pr(\text{fail}_{EC}) \leq \exp(-\Omega(d))$. If the error correction succeeds, the errors undergoing correction, combined with the application of recovery gates, result in trivial operations on the encoded state, ensuring that the logical measurement outcomes are correct.
- (ii) Errors that bypass the correction process result in effective noise on the output state characterized by an 8-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, 4t_{RS}, \text{success}_{SP})$ for some B_{out} and disjoint f_{out} .

Furthermore, B_{out} is disjoint from the virtual location sets of all other code surgery circuits and parity-check measurements, with the exception of the two specifically employed in the current circuit.

Proof. Let C_j denote the set of qubit-time locations in the j th column of the circuit. Following the proofs of Lemmas 18 and 19, X-type errors at C_1 undergo correction through the Z-type parity-check measurement, while X errors at C_2, C_3 , and C_4 bypass the correction procedure. Similarly, Z-type errors at C_1, C_2 , and C_3 undergo correction through the X-type parity-check measurement, whereas Z errors at C_4 bypass the correction procedure.

We now analyze the effective noise model for each subset C_j . Errors on the input state and the effective input errors of the Z-type parity-check measurement are equivalent to errors at C_1 . Taking these two sources of errors into account, the effective noise at C_1 is characterized by a $(\nu + 4)$ -confined model with $(C_1, B_1, f_1, \lambda p, \eta t_{RS}, \text{success}_{SP})$ for some B_1 and disjoint f_1 . The errors at C_2 and C_4 are the effective output errors of corresponding parity-check measurements, therefore, characterized by a 2-confined model with $(C_j, B_j, f_j, \lambda p, t_{RS}, \text{success}_{SP})$ for some B_j and disjoint f_j , where $j = 2, 4$; similarly, the effective errors at C_3 are characterized by a 4-confined model with $(C_3, B_3, f_3, \lambda p, t_{RS}, \text{success}_{SP})$ for some B_3 and disjoint f_3 . Note that virtual location sets in these models are mutually disjoint.

Error correction. Our derivation follows the threshold proof presented in Ref. [15], adapted for our ν -confined generalized noise model.

Since the parity-check measurements are effectively error-free, the logical measurement outcomes are always correct provided that no logical X errors occur prior to the Z-type parity-check measurement. Consequently, the threshold proof reduces to analyzing the probability of logical errors.

First, we consider X errors acting on the memory system in the first column, which are corrected by the check matrix $\tilde{H}_Z$ (see Lemma 19): Although the circuit measures stabilizers for the complete H_Z^D check matrix, only the $\tilde{H}_Z$ component can detect errors. This is because only the stabilizers of $\tilde{H}_Z$ have well-defined values in the input state, enabling a direct syndrome comparison to identify faults. Note that measurement outcomes for these stabilizers are treated as effectively error-free.

The proof is established on the adjacency graph of $\tilde{H}_Z$. Let $C'_1 \subseteq C_1$ be the subset of qubits in the memory system. Each column of $\tilde{H}_Z$ corresponds to a physical qubit at a location in C'_1 . In this graph, qubits are represented as vertices, and an edge exists between two vertices if and only if they share a common parity check. For qLDPC codes, the maximum degree of this graph is bounded by a constant $z = O(1)$. Under the minimum-weight decoding algorithm, error correction fails only if there exists a connected cluster $S \subseteq C'_1$ of $s \geq d$ qubits in the adjacency graph such that S contains at least $\lceil s/2 \rceil$ errors.

Given $S \subseteq C_1$, we apply Lemma 20. The probability that S contains at least $\lceil s/2 \rceil$ errors is bounded by:

$$P_s = \left[\frac{(\nu+4)s e}{l} \lambda p \right]^l, \quad \text{where } l = \lceil \lceil s/2 \rceil - \eta t_{RS} \rceil. \quad (\text{I14})$$

Utilizing $t_{RS} = d/16$, we observe that for $s \geq d$, $l \geq \kappa s$, where $\kappa = 1/2 - \eta/16$. For a sufficiently small physical error rate $p < \epsilon$, where

$$\epsilon = \frac{\kappa}{\lambda(\nu+4)e(ze)^{1/\kappa}}, \quad (\text{I15})$$

the term inside the parentheses is bounded by $(ze)^{1/\kappa} > 1$. Then, we obtain:

$$P_s \leq \left[\frac{(\nu+4)e}{\kappa} \lambda p \right]^{\kappa s}. \quad (\text{I16})$$

The number of connected clusters of size s containing a fixed vertex in the adjacency graph is at most $(ze)^{s-1}$ (Lemma 25). Summing over all n_D possible starting vertices and all cluster sizes $s \geq d$, the total probability of failure for X -type error correction is bounded by:

$$\begin{aligned} \Pr(\text{fail}_{EC,X}) &\leq \sum_{s \geq d} n_D (ze)^{s-1} P_s \leq \frac{n_D}{ze} \sum_{s \geq d} \left(\frac{p}{\epsilon} \right)^{\kappa s} \\ &= \exp(-\Omega(d)). \end{aligned} \quad (\text{I17})$$

This result establishes the existence of a finite fault-tolerant threshold in the correction of X errors.

Next, we consider Z errors acting on the first three columns, which are corrected by the check matrix H_X^D (see Lemma 18). Although the circuit directly measures only the stabilizers in $\tilde{H}_X$, the eigenvalues of the full H_X^D stabilizer set can be reconstructed from the outcomes of the transversal X -basis measurements on the ancilla system.

The Z -type errors occurring in the first three columns are equivalent to effective Z errors localized at C_1 . By merging the noise models for these three columns, we obtain a $(\nu+10)$ -confined generalized noise model $(C_1, B'_1, f'_1, \lambda p, (\eta+3)t_{RS}, p, \text{success}_{SP})$ for some B'_1 and f'_1 . Note that the distance of the code used for decoding is $d_D = d$. Following an analysis analogous to that for X -type errors, we can establish the existence of a finite threshold $\epsilon_Z > 0$. Specifically, for any physical error rate $p < \epsilon_Z$, the total failure probability for Z -type error correction is exponentially suppressed: $\Pr(\text{fail}_{EC,Z}) = \exp(-\Omega(d))$.

Faults on the output state. X -type errors at C_2 and C_3 may cause faults on the output state: each single-qubit X error on the memory system can result in fault on a physical qubit in the output state; and X errors on the ancilla system are trivial because of the transversal X measurement. Errors at C_4 directly act on physical qubits in the output state. Consequently, the corresponding merging mappings g each satisfy $\|g\|_{\leftrightarrow} = 1$, which implies they are disjoint. By applying Lemma 21 to these parameters, we have the effective noise model of the output state.

The set B_{out} is comprised of locations within the current circuit and the virtual locations of its two constituent parity-check measurements. Since the virtual sets for distinct parity-check measurements are mutually disjoint, they are disjoint from the virtual locations of all other parity-check measurement and code surgery circuits. ■

Remark. In Appendix I 4 a, we establish that the input noise model parameters are $\nu = 8$ and $\eta = 4$, resulting in a threshold $\epsilon = \frac{\kappa}{12\lambda e(ze)^{1/\kappa}}$, where $\kappa = 1/2 - 4/16 = 1/4$. Numerical studies in Ref. [41] demonstrate that magic states can be injected into surface codes with effective error rates potentially lower than those of individual physical gates, suggesting a parameter value of $\lambda \approx 1$. The factor of $1/16$ in the definition of κ originates from the specific definition of t_{RS} . By logarithmically increasing the surface-code distance d_S , this factor can be reduced such that κ approaches $1/2$. In the limit $\kappa = 1/2$, this threshold coincides with the bound reported in Theorem 3 of Ref. [15], subject to an additional factor of $(6e)^{-1}$. This discrepancy is comprised of a factor of $2/e$ arising from the specific combinatorial bounding techniques employed, and a factor of $1/12$ which accounts for the accumulation of errors throughout the circuit.

The final fault-tolerant operations to be analyzed are the initializations of a code block in the logical $|0\rangle$ or $|+\rangle$ state.

We focus our discussion on the $|+\rangle$ state, as the results are symmetrically applicable to the $|0\rangle$ state.

Lemma 27. Fault-tolerant block initialization. Apply the parity-check measurement of H_X^M to a memory-code block where each physical qubit is initialized in the $|+\rangle$ state, the circuit yields a block with all logical qubits prepared in the logical state $|+\rangle$. Let A_{out} denote the set of physical qubits in the resulting output state. Assume that the circuit noise follows the simplified local stochastic noise model. Suppose the parity-check measurement is fault-tolerant as depicted by Lemma 24.

Under these conditions, the output state is subject to effective noise described by a 6-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, t_{RS}, \text{success}_{SP})$ for some B_{out} and disjoint f_{out} . Furthermore, B_{out} is disjoint from the virtual location sets of all other initialization circuits, code surgery circuits and parity-check measurements, with the exception of the one specifically employed in the current circuit.

Proof. By considering the logical X operators, one can verify that the logical qubits are initialized in the logical $|+\rangle$ state. Specifically, if the entire circuit is error-free, the output state is an eigenstate of both the logical X operators and the X -type stabilizers with eigenvalue $+1$. Furthermore, the state is projected into an eigenstate of the Z -type stabilizers, with eigenvalues determined by the outcomes of the parity-check measurement.

The noise in the parity-check measurement is effectively described by the input and output noise models established in Lemma 24. Input X errors are trivial in this context, as they act on physical qubits already in the $|+\rangle$ state. Conversely, input Z errors are equivalent to Z errors on the final output state, as they commute with the parity-check measurement. Therefore, the effective noise model for the fault-tolerant initialization operation is obtained by merging the two effective noise models of the parity-check measurement according to Lemma 21. Note that merging mappings g are disjoint.

The set B_{out} consists of the virtual locations from its constituent parity-check measurement; consequently, it is disjoint with virtual location sets for all other fault-tolerant operations. ■

3. Noisy preparation of encoded magic states

Remark. Throughout the following discussion, referring to certain operations as “noisy” does not necessarily imply that their physical fidelity is low. Rather, it signifies that we have adopted a modified success condition: while the probability of failure may be higher, the weight of the residual errors is guaranteed to be low, conditioned on success.

To distinguish between the events of fault-tolerant operations and noisy operations, we let $\text{success}'$ and fail' denote events regarding noisy operations, whereas

success and fail denote events relevant to fault-tolerant operations.

In our protocol, surface codes are employed in two distinct roles. In the resource state preparation circuits, they protect qubits against both X - and Z -type errors. In the context of magic state preparation, they are utilized for the initial encoding of the magic states. For clarity, we denote the distances of these two types of surface codes as d_S and d'_S , respectively.

In this section, we adopt the notation and conditions established in Sec. I 2; thus, they will not be redundantly restated. For simplicity, we omit explicit statements regarding the disjointness of propagation mappings and virtual location sets, although these properties remain valid for all noisy operations.

Lemma 28. Resource state preparation for noisy operations. Under the same assumptions as Lemma 23, but with a modified success condition, the following statements hold:

- (i) If $p < p_{S,th}$, the circuit fails with a probability $\Pr(\text{fail}'_{SP}) \leq p/16$.
- (ii) Conditioned on the circuit succeeding, each copy of the output state is subject to noise characterized by a 1-confined model $(A_{RS}, B_{RS}, f_{RS}, p_{D,S}, 0, \text{success}'_{SP})$.

Proof. The proof follows the same logic as that of Lemma 23, by setting $t_{RS} = 0$. Given that $\lceil t \rceil \geq 1$, the failure probability bound $\Pr(\text{fail}'_{SP}) \leq p/16$ can be satisfied by selecting a surface-code distance $d_S = O(\text{polylog}(n_D n_F))$, provided the physical error rate is below the surface-code threshold. ■

Lemma 29. Noisy parity-check measurement. Under the same assumptions as Lemma 24, but suppose the resource state is generated by a noisy resource state preparation circuit depicted by Lemma 28. The circuit is effectively error-free, with effective input and output noise described by a 4-confined model $(A_{\text{in}}, B_{\text{in}}, f_{\text{in}}, \lambda p, 0, \text{success}'_{SP})$ and a 2-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, 0, \text{success}'_{SP})$, respectively.

Proof. The proof is the same as that of Lemma 24. ■

Lemma 30. Noisy code surgery. Consider the code surgery circuit illustrated in Fig. 16. We assume a surface-code distance $d'_S = O(\text{polylog}(d))$ and a deformed code distance $d_D = d'_S$. Assume the noise on the input state is characterized by a ν -confined model $(A_{\text{in}}, B_{\text{in}}, f_{\text{in}}, \lambda p, 0, \text{success}'_{SP})$ with $\nu = O(1)$. Suppose the parity-check measurements are noisy as depicted by Lemma 29.

Under these conditions, there exists a decoding algorithm such that:

- (i) There exist a finite threshold $\epsilon' > 0$ and a constant $\lambda'_S > 0$ such that for $p < \epsilon'$, the error correction fails with probability $\Pr(\text{fail}'_{EC}) \leq \lambda'_S p$.
- (ii) Errors that bypass the correction process result in effective noise on the output state characterized

by an 8-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, 0, \mathfrak{Success}'_{SP})$.

Proof. The difference between fault-tolerant and noisy parity-check measurements lies in the effective output noise. Specifically, in the fault-tolerant (noisy) measurement, the intrinsic fault parameter is $\delta = t_{RS}$ ($\delta = 0$). Consequently, the proof for Lemma 26 applies here, yielding a finite threshold ϵ' , and if $p < \epsilon'$, the failure probability is $\Pr(\mathfrak{fail}'_{EC}) \leq n_D \exp(-\Omega(d'_S))$. By choosing the surface-code distance $d'_S = O(\text{polylog}(d))$, we can ensure that this probability is suppressed below $\lambda'_S p$ for any fixed physical error rate $p < \epsilon'$ and positive constant λ'_S . ■

Lemma 31. Noisy block initialization. Under the same conditions as in Lemma 27 with the exception that the parity-check measurement is noisy as depicted by Lemma 29. Then, the output state is subject to effective noise described by a 6-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, 0, \mathfrak{Success}'_{SP})$.

Proof. The proof is the same as that of Lemma 27. ■

Lemma 32. Noisy surface-code state injection. Consider a surface-code block of distance d'_S . Suppose each physical qubit is initialized according to the circuit in Fig. 9(b), following the layout pattern given in Fig. 14(e). Upon applying parity-check measurements to extract X - and Z -type stabilizers, the circuit prepares the surface-code block in the logical state $|\psi\rangle$. Assume that the circuit noise follows the simplified local stochastic noise model. Suppose the parity-check measurements are noisy as depicted by Lemma 29.

Under these conditions, there exists a decoding algorithm such that:

- (i) There exist a finite threshold $\epsilon' > 0$ and a constant $\lambda'_S > 0$ such that for $p < \epsilon'$, the error correction fails with probability $\Pr(\mathfrak{fail}'_{EC}) \leq \lambda'_S p$.
- (ii) Errors that bypass the correction process result in effective noise on the output state characterized by an 8-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, 0, \mathfrak{Success}'_{SP})$.

Proof. Every parity-check measurement is subject to both input and output effective noise. The error correction procedure specifically addresses X errors preceding the Z -stabilizer measurement and Z errors preceding the X -stabilizer measurement. Consequently, the effective noise on the output state is described by a model formed by merging one input noise model and two output noise models. Notably, in the merging of models, the merging mappings g are disjoint.

The noise addressed during error correction is characterized by a model formed by merging two input noise models and one output noise model, which is a 10-confined model $(A_{\text{in}}, B_{\text{in}}, f_{\text{in}}, \lambda p, 0, \mathfrak{Success}'_{SP})$. Error correction is performed by comparing the stabilizer measurement outcomes with expected eigenvalues derived from the initialization pattern shown in Fig. 14(e). The initialization pattern is partitioned into two triangular regions, where qubits are

initialized in either the $|0\rangle$ or $|+\rangle$ state. In the $|0\rangle$ -state region, Z -type stabilizer eigenvalues can be deduced and utilized for error correction, while in the $|+\rangle$ -state region, X -type stabilizer eigenvalues are similarly available. Consequently, this initialization scheme provides only partial error-correction capability compared to the standard surface code.

Henceforth, we focus on the $|+\rangle$ -state region, defined by the top horizontal boundary and the diagonal line of the surface-code lattice [see Fig. 14(e)]. We observe that a logical Z error occurs if a connected cluster of Z errors spans the distance from the top boundary to the diagonal line. Let x denote the horizontal coordinate of a qubit on the top boundary, indexed from left to right. There exists a positive constant L such that any such cluster containing qubit- x must have a minimum size of x/L ; otherwise, the cluster cannot span the geometric distance between the top boundary and the diagonal line.

The number of connected clusters of size s containing a fixed qubit on the top boundary is at most $(ze)^{s-1}$ (see Lemma 25). Since a cluster of size s can only reach the diagonal line if its starting point x on the top boundary satisfies $x \leq Ls$, there are at most Ls possible starting positions for such a cluster. Therefore, the total number of connected size- s clusters spanning from the top boundary to the diagonal line is upper-bounded by $Ls(ze)^{s-1}$.

Following an analysis analogous to the proof of Lemma 26, the probability of having sufficient errors within a cluster of size s to cause an error correction failure is upper bounded: there exist positive constants ϵ' and κ' such that for $p < \epsilon'$,

$$P_s \leq \left(\frac{10e}{\kappa'} \lambda p \right)^{\max\{\kappa' s, 1\}}. \quad (\text{I18})$$

Note that we have used $l \geq 1$. Consequently, the probability of an error correction failure can be bounded by splitting the sum over cluster sizes at a constant $s_0 = \lfloor 1/\kappa' \rfloor$:

$$\begin{aligned} \Pr(\mathfrak{fail}'_{EC}) &\leq \frac{L}{ze} \sum_{s=1}^{s_0} s \left(\frac{p}{\epsilon'} \right)^1 + \frac{L}{ze} \sum_{s>s_0} s \left(\frac{p}{\epsilon'} \right)^{\kappa' s} \\ &\leq \frac{L}{ze} \left[\frac{s_0(s_0+1)}{2\epsilon'} p + \frac{(s_0+1)(p/\epsilon')^{\kappa'(s_0+1)}}{[1 - (p/\epsilon')^{\kappa'}]^2} \right]. \end{aligned} \quad (\text{I19})$$

When $(p/\epsilon')^{\kappa'} \leq 1/2$, which defines the threshold, there exists a positive constant λ'_S such that $\Pr(\mathfrak{fail}'_{EC}) \leq \lambda'_S p$.

It is similar for the correction of X errors. ■

4. Proof of the theorem under the simplified noise model

a. Fault-tolerant operations

The lemmas in Sec. I 2 establish the fault tolerance of the protocol's operations, specifically including all operations illustrated in Figs. 8 and 9(a) (taking $d_C = \Theta(d)$). These operations, which consist of parity-check measurements and code surgery, each fail with a probability upper-bounded by $e^{-\Omega(d)}$, conditioned on the success of all preceding fault-tolerant operations. Consequently, the probability that any failure occurs within a set of L_{FT} fault-tolerant operations is at most $L_{FT}e^{-\Omega(d)}$. This total failure probability can be efficiently suppressed by increasing the code distance d .

The final remaining condition to examine is the input noise model for code surgery operations; see Lemma 26. In a general code surgery, the input memory system may be composed of multiple code blocks, each originating from a preceding operation—such as a parity-check measurement, code surgery, or block initialization. A single preceding operation may produce multiple output code blocks, with noise collectively described by a unified noise model. Since the propagation mapping f in all our output noise models is inherently disjoint, we can apply Lemma 22 to decompose the collective model into independent sub-models for each block. Consequently, each input memory-code block participating in a code surgery operation possesses noise characterized by an 8-confined model $(A, B, f, \lambda p, 4t_{RS}, \mathfrak{Succes}\mathfrak{S}_{SP})$.

However, a critical issue arises: a single code surgery operation may involve as many as $\Theta(k_R)$ code blocks. Since each constituent input block contributes an intrinsic fault path of weight $4t_{RS}$, the cumulative weight of the intrinsic faults across the entire input memory system scales as $\Theta(k_R)t_{RS}$. Consequently, the joint noise model for the memory system is described by an 8-confined model $(A', B', f', \lambda p, \Theta(k_R)t_{RS}, \mathfrak{Succes}\mathfrak{S}_{SP})$, which is an invalid input when k_R is large.

We can effectively control the accumulation of intrinsic faults through joint error correction. Whenever multiple code blocks are provided as input to a code surgery operation, a preceding round of joint error correction must be performed. This is implemented by executing a code surgery operation with the check matrix configured as $H_Z^D = \tilde{H}_Z$. While the parity-check measurements are executed jointly across multiple blocks, the subsequent error correction is performed individually for each code block. Although the total intrinsic error weight may reach $\Theta(k_R)t_{RS}$ across the full input memory system, the intrinsic weight localized to any single block remains bounded by $4t_{RS}$. Consequently, the joint error correction round maintains a failure probability upper-bounded by $e^{-\Omega(d)}$. Following this procedure, the noise on the resulting

memory system is characterized by an 8-confined model $(A', B', f', \lambda p, 4t_{RS}, \mathfrak{Succes}\mathfrak{S}_{SP})$. This provides a valid input for the subsequent code surgery operation.

b. Magic states

The noisy encoded magic states are prepared using the circuit illustrated in Fig. 9(b). Most constituent operations within this circuit have been previously analyzed in Sec. I 3. Specifically, each operation yields an effective output noise model that serves as a valid input for the subsequent step, with logical failure probabilities bounded by $O(p)$.

The remaining operation, the transversal X -basis measurement on the surface-code block, corrects Z -type errors to ensure that the feedback gate applied to the memory-code block is logically correct. Since the surface code possesses a finite fault-tolerant threshold, the probability of a logical failure at this stage is upper-bounded by $e^{-\Omega(d'_s)}$. By choosing a surface-code distance $d'_s = O(\text{polylog}(d))$, this probability can be efficiently suppressed to $O(p)$.

In summary, the circuit produces an encoded magic state with a logical error probability of $O(p)$ and residual physical errors characterized by an 8-confined model $(A_{\text{out}}, B_{\text{out}}, f_{\text{out}}, \lambda p, 0, \mathfrak{Succes}\mathfrak{S}'_{SP})$. Hence, the prepared state is a valid noisy input for the subsequent fault-tolerant stages of the protocol.

Theorem 2. (Theorem 1.4 in [18]). Let $|\text{CCZ}\rangle = \text{CCZ}|+\rangle^{\otimes 3}$. There exists a distillation protocol (using noiseless Clifford operations and measurements) $\text{MSD}(p_L)$ and a constant noise threshold p_{noisy} such that the following holds. Upon receiving N (independent) noisy states whose infidelity with $|\text{CCZ}\rangle$ is $p \leq p_{\text{noisy}}$, where $N = O(\log^{1+o(1)}(1/p_L))$ is a sufficiently large number, $\text{MSD}(p_L)$ produces $\Theta(\frac{N}{\log^{o(1)}(1/p_L)})$ states each of which has infidelity at most p_L with $|\text{CCZ}\rangle$. Furthermore, the quantum depth and classical depth of $\text{MSD}(p_L)$ are both $\text{polyloglog}(1/p_L)$. Here, the $o(1)$ exponent scales as $O(\frac{1}{\log \log \log(1/p_L)})$.

Although our protocol is formulated in terms of T -gate magic states, it can be straightforwardly adapted to CCZ -gate magic states. Alternatively, one could utilize the CCZ state distillation protocol (described in Theorem 2) to distill T states, as these states can be converted into each other with constant overhead [42]. Given that the error rate of the injected magic state is $O(p)$, there exists a finite threshold for p below which the conditions of Theorem 2 are satisfied. Consequently, magic states can be distilled to the target error rate with almost constant overhead.

c. Summary

The overhead analysis is provided in Sec. III F and Appendix D. Our analysis concludes that the protocol achieves strictly constant qubit overhead and a time

overhead of $O(d^{a+o(1)})$, provided that the code distance $d, d_R, d_F = O(\text{polylog}(Mk))$.

We consider the usual setting where classical information is provided as input, processed by the quantum computation, and finally returned as classical output. We use C_{FT} to simulate the quantum circuit C , and measure the simulation error by the discrepancy between their resulting output distributions, with target accuracy ϵ_L .

To ensure that the output distributions of C and C_{FT} differ by at most ϵ_L , it suffices to require that the logical failure probability per operation satisfies $p_L \leq \epsilon_L/|C|$. In our construction, $p_L = e^{-\Omega(d)}$, so we choose $d = d_R = d_F = \Theta(\log(|C|/\epsilon_L))$. Given that the circuit size is polynomial in the width, $|C| = O(\text{poly}(W))$, and the total number of logical qubits satisfies $Mk = \Theta(W)$, the polylogarithmic condition assumed in the overhead analysis is fully justified. Consequently, the total time overhead of the protocol scales as $O(\log^{a+o(1)}(|C|/\epsilon_L))$.

5. Justification of the simplified noise model

In this section, we demonstrate that the threshold theorem remains valid under the standard local stochastic noise model, assuming each operation in the circuit can be faulty. In this framework, every circuit operation is treated as a potential fault location, and the probability of a specific fault path F is exponentially suppressed by the number of faulty operations, i.e., $\Pr[F] \leq p^{|F|}$.

Error propagation. In our simplified model, we assume that noise is concentrated at specific locations that subsume all errors generated by the preceding or subsequent constant-depth transversal operations. These locations are treated under a 1-confined model with $\delta = 0$, where the probability of s faults is exponentially suppressed as p^s .

However, in a more realistic model where every operation acts as a potential error source, a single faulty operation may propagate to multiple locations (of the simplified model), and conversely, a single location may be influenced by multiple operations. This error propagation is naturally captured by our generalized local stochastic noise model, where each physical operation corresponds to a virtual location and the locations in the simplified model represent real locations.

To account for this propagation, we transition from a 1-confined model to a ν -confined model (with $\delta = 0$). The confinement parameter ν remains a constant due to the (q)LDPC nature of the circuits: the sparsity of the constant degree of the gate network ensures that each faulty operation affects at most a constant number of qubits, and each qubit is affected by at most a constant number of operations. While increasing the confinement parameter ν necessarily reduces the numerical value of the fault-tolerant threshold, it does not invalidate its existence; the threshold remains finite as long as ν is bounded.

Surface-code logical errors. We now examine our model for surface-code logical errors. The existence of a finite fault-tolerant threshold for the surface code is well-established (see Ref. [15]). If the physical error rate is below this threshold, the logical error probability is exponentially suppressed with the code distance d_S . However, because our protocol employs a small d_S , the logical error rate is non-negligible.

In the following, we demonstrate that although these logical errors are potentially *temporally non-local*, their impact can be effectively mitigated. Specifically, since the surface code is utilized within a constant-depth state preparation circuit, any temporal non-locality is strictly bounded. This limited temporal range ensures that such correlations can be successfully managed during the error correction process without compromising the threshold theorem.

In the fault-tolerance proof presented in Ref. [15], the essential task is to bound both the probability of error-correction failure and the weight of the residual errors following correction. If error correction succeeds, the residual errors are effectively described by a (standard) local stochastic noise model, which ensures that the logical error probability for subsequent operations remains suppressed by $e^{-\Omega(d_S)}$. Conversely, if an error-correction failure occurs, this exponential suppression can no longer be guaranteed for subsequent operations on the surface-code block.

This necessitates a more careful treatment of logical errors: an error-correction failure at a qubit-time position (q, t) may induce faults throughout its *causal cone*, which consists of all subsequent positions reachable through the circuit's operations. We can naturally incorporate these effects into our generalized noise model by treating error-correction failures as *virtual faults*. Each such virtual fault is then mapped to a set of *real faults* representing the logical errors at the affected qubit-time positions.

Critically, since the surface code is utilized within a constant-depth state preparation circuit, each causal cone contains at most a constant number of locations. Consequently, while this propagation increases the confinement parameter ν , it remains an $O(1)$ constant, thereby preserving the existence of a finite fault-tolerant threshold.

Lastly, we address the errors associated with decoding operations. Analogous to the analysis in Lemma 28, if the noise on a surface-code block follows a (standard) local stochastic model, the decoding error rate is upper-bounded by $\lambda_S p$. However, due to potential temporal correlations of logical errors, this $\lambda_S p$ bound is strictly applicable only to qubits lying outside the causal cones of any prior error-correction failures.

If t such error-correction failures occur within the circuit, they may propagate to $O(t)$ qubits at the decoding locations. To incorporate this effect, we can model the decoding-operation errors as a 1-confined model with an

intrinsic fault parameter $\delta = O(t)$. Since the number of logical failures t is exponentially suppressed by increasing the surface-code distance d_S , the parameter δ can be rigorously controlled in the same manner as t_{RS} . Including a finite intrinsic error parameter δ in the decoding-operation noise model merely results in an enlarged value for the parameter t_{RS} in the generalized noise model of the output resource state in Lemma 23.

APPENDIX J: PERFORMANCE AT FINITE DISTANCE

While our protocol is optimized for asymptotic performance, its feasibility for moderate code distances is of practical interest. This section evaluates performance of the proposed methods in the finite-distance regime.

1. The surface-code overhead

Although the surface-code distance d_S is sub-polynomial in the large-distance limit, it may significantly influence the physical qubit requirements and the overall performance in the finite-distance regime. In this section, we provide a more explicit analytical estimation of d_S .

The resource-state preparation circuit, prior to surface-code encoding, comprises $(n_D + r_Z^D)(n_F + r_F)$ qubits and possesses a depth of $\omega_Z^D + \omega_F + 3$. Here, n_D and n_F denote the lengths of the deformed and factory codes, respectively; r_Z^D and r_F are the number of rows in the Z -type check matrix of the deformed code and the check matrix of the factory code, respectively. The parameters ω_Z^D and ω_F represent the maximum row/column weights of these respective matrices. To ensure that the logical error contribution in Eq. (I13) is suppressed below $p/16$, it is sufficient to satisfy the condition:

$$p_{L,S}(p) < \frac{[t]p}{16(\omega_Z^D + \omega_F + 3)(n_D + r_Z^D)(n_F + r_F)e}, \quad (\text{J1})$$

where $[t] \geq 1$. Using the standard phenomenological formula for the surface-code logical error rate, $p_{L,S}(p) \approx p_0(p/p_{th,S})^{\lfloor (d_S+1)/2 \rfloor}$ [47], where p_0 is a positive constant and $p_{th,S}$ is the surface-code threshold, we can derive an estimate for the required surface-code distance:

$$d_S \approx 2 \frac{\ln(\frac{p/p_0}{16(\omega_Z^D + \omega_F + 3)(n_D + r_Z^D)(n_F + r_F)e})}{\ln(p/p_{th,S})}. \quad (\text{J2})$$

Since $n_D, r_Z^D = O(d^{a+a_R})$ and $n_F, r_F = O(d^{a_F})$, while the weights ω_Z^D and ω_F remain constant, it follows that the required distance scales logarithmically with the memory-code distance, $d_S = O(\log d)$.

2. Application of PCS to finite-length codes

In this section, we analyze the application of PCS to quantum codes with finite length. We demonstrate its performance advantage for implementing certain logical operations, even when the underlying code is not drawn from a good qLDPC family.

Consider the following logical circuit layer. We have M memory-code blocks, each encoding k logical qubits. Within each block, we label the logical qubits as a one-dimensional array indexed from 1 to k . The layer consists of $Z \otimes Z$ logical measurements applied to nearest-neighboring logical qubit pairs. Assuming k is even, the measurement is performed on each pair of logical qubits with indices $(2m-1, 2m)$, where m ranges from 1 to $k/2$.

Before illustrating the benefits of PCS for implementing such a layer, we first demonstrate its utility within practical quantum algorithms using a concrete example. Consider the simulation of time evolution under a one-dimensional Heisenberg model, where each spin is represented by a logical qubit. According to the Trotter decomposition, this simulation requires the implementation of two-qubit unitary gates of the form $e^{-iZ \otimes Z \theta}$ on nearest-neighboring spins. When decomposed into standard universal gates, the $e^{-iZ \otimes Z \theta}$ interaction can be realized using two controlled-NOT gates and an $e^{-iZ \theta}$ gate on the target qubit. Consequently, to implement this interaction across all spin pairs, we must perform controlled-NOT gates on each pair of nearest neighbors simultaneously. In the code surgery framework, each controlled-NOT gate requires an ancilla logical qubit and $Z \otimes Z$ measurements on the control and ancilla qubits. Therefore, to implement the $Z \otimes Z$ measurement layer on a one-dimensional logical qubit array, we allocate the logical qubits as follows: logical qubits with odd indices are used to represent the spins, and logical qubits with even indices are used as ancilla qubits. In this arrangement, the $e^{-iZ \otimes Z \theta}$ interaction between each spin pair involves the simultaneous execution of the specified operations across all blocks. The use of ancilla qubits leads to a scaling in the number of memory-code blocks $M = \lceil 2N/k \rceil$, where N is the number of spins.

To estimate the qubit and time overhead for this $Z \otimes Z$ measurement layer under PCS, we recall that the deformed code has a length bounded by $n_D \leq (k_R + 2n_R)n$. Given that PCS simultaneously implements the operation layer on k_R blocks, the qubit overhead is given by:

$$\frac{n_D}{k_R k} \leq \frac{\rho_R + 2}{\rho_R \rho}, \quad (\text{J3})$$

where $\rho = k/n$ and $\rho_R = k_R/n_R$ are encoding rates for the memory code and R code, respectively. Here, we have neglected ancilla qubits for implementing stabilizer measurements. Absent the use of LTSP, the time overhead for PCS is simply $O(d)$. When the codes have constant

TABLE V. Examples of classical codes, the corresponding symmetric HGP memory codes, and the resulting PCS qubit-overhead bounds. The (3, 4)-regular Gallager code examples and the (3, 5)-regular quasi-cyclic code examples are taken from Refs. [48,49].

Code family	Classical code	HGP code	(ρ_R, ρ)	PCS qubit overhead
Hamming code	[7, 4, 3]	[[58, 16, 3]]	$(\frac{4}{7}, \frac{8}{29})$	≤ 17
(3, 4)-regular Gallager codes	[12, 3, 4]	[[225, 9, 4]]	$(\frac{1}{4}, \frac{1}{25})$	≤ 225
	[20, 5, 6]	[[625, 25, 6]]		
	[28, 7, 8]	[[1225, 49, 8]]		
	[40, 10, 12]	[[2500, 100, 12]]		
	[60, 15, 16]	[[5625, 225, 16]]		
	[80, 20, 18]	[[10000, 400, 18]]		
(3, 5)-regular quasi-cyclic codes	[80, 34, 12]	[[8704, 1160, 12]]	$(\geq \frac{2}{5}, \geq \frac{2}{17})$	≤ 51
	[105, 44, 16]	[[14994, 1940, 16]]		
	[150, 62, 20]	[[30600, 3844, 20]]		
	[210, 86, 24]	[[59976, 7400, 24]]		

encoding rates, the overall spacetime overhead is $O(1)$, outperforms code surgery methods listed in Table I.

To provide concrete numbers, we take HGP codes as an example. Assume the memory code is an HGP code generated by a linear code with rate ρ_R (the same code used for the R code). Assuming the check matrix of the linear code is full-rank, the HGP rate is $\rho = \rho_R^2 / (1 + (1 - \rho_R)^2)$. Substituting this into the PCS overhead formula yields:

$$\frac{n_D}{k_R k} \leq \frac{(\rho_R + 2)[1 + (1 - \rho_R)^2]}{\rho_R^3}. \quad (\text{J4})$$

Notably, the qubit overhead remains constant, resulting in a total spacetime overhead of $O(d)$. Table V provides explicit qubit overhead values for several example linear codes.

3. LTSP with more general factory codes

To achieve constant qubit and time overheads for resource state preparation, we have primarily focused on the asymptotic regime where the F code is selected from a family of LTCs with constant encoding rate and constant soundness. Although theoretical constructions for such codes exist [13,32], identifying concrete code instances optimized for the finite-distance regime remains a subject that requires further investigation.

Critically, neither the constant encoding rate nor the constant soundness requirement is strictly necessary for finite-distance applications. Any LDPC code with a non-zero soundness parameter can be utilized as an F code. For example, one could employ a repetition code with a check matrix defined by a bounded-degree expander graph, where vertices represent bits and edges represent checks enforcing equality between adjacent bits. In this case, the soundness of the code is controlled by the expansion properties of the graph. Specifically, the soundness satisfies $s = \frac{n_E}{r_F} h$, where h denotes the edge expansion of the graph.

In particular, for a constant-degree expander family with $h = \Omega(1)$, this yields $s = \Omega(1)$.

Using such a construction for the F code allows for a significant reduction in time overhead at the expense of additional qubits. For an F code based on a length- d repetition code, the circuit depth for state preparation remains almost constant at $O(d_S)$, while the qubit overhead increases to $O(dd_S^2)$, where the factor d originates from the vanishing encoding rate of the repetition code. This allows for a trade-off between qubit overhead and temporal efficiency, illustrating the flexibility of the LTSP approach, where even codes that do not satisfy strictly constant asymptotic requirements remain viable for applications in fault-tolerant quantum computation.

-
- [1] P. W. Shor, *Scheme for reducing decoherence in quantum computer memory*, *Phys. Rev. A* **52**, R2493 (1995).
 - [2] A. Y. Kitaev, *Quantum computations: Algorithms and error correction*, *Russ. Math. Surv.* **52**, 1191 (1997).
 - [3] D. Gottesman, *Theory of fault-tolerant quantum computation*, *Phys. Rev. A* **57**, 127 (1998).
 - [4] F. Gaitan, *Quantum Error Correction and Fault Tolerant Quantum Computing* (CRC Press, Boca Raton, 2018).
 - [5] E. Knill, R. Laflamme, and W. H. Zurek, *Resilient quantum computation*, *Science* **279**, 342 (1998).
 - [6] D. Aharonov and M. Ben-Or, *Fault-tolerant quantum computation with constant error rate*, *SIAM J. Comput.* **38**, 1207 (2008).
 - [7] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation*, *Phys. Rev. A* **86**, 032324 (2012).
 - [8] D. Horsman, A. G. Fowler, S. Devitt, and R. V. Meter, *Surface code quantum computing by lattice surgery*, *New J. Phys.* **14**, 123011 (2012).
 - [9] J.-P. Tillich and G. Zemor, *Quantum LDPC codes with positive rate and minimum distance proportional to the square root of the blocklength*, *IEEE Trans. Inf. Theory* **60**, 1193 (2014).

- [10] P. Panteleev and G. Kalachev, *Degenerate quantum LDPC codes with good finite length performance*, [Quantum](#) **5**, 585 (2021).
- [11] N. P. Breuckmann and J. N. Eberhardt, *Balanced product quantum codes*, [IEEE Trans. Inf. Theory](#) **67**, 6653 (2021).
- [12] N. P. Breuckmann and J. N. Eberhardt, *Quantum low-density parity-check codes*, [PRX Quantum](#) **2**, 040101 (2021).
- [13] P. Panteleev and G. Kalachev, *Asymptotically good quantum and locally testable classical LDPC codes*, in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, STOC '22* (ACM, 2022).
- [14] I. Dinur, M.-H. Hsieh, T.-C. Lin, and T. Vidick, *Good quantum LDPC codes with linear time decoders*, [arXiv:2206.07750](#).
- [15] D. Gottesman, *Fault-tolerant quantum computation with constant overhead*, [Quantum Inf. Comput.](#) **14**, 1339 (2014).
- [16] H. Yamasaki and M. Koashi, *Time-efficient constant-space-overhead fault-tolerant quantum computation*, [Nat. Phys.](#) **20**, 247 (2024).
- [17] S. Tamiya, M. Koashi, and H. Yamasaki, *Polylog-time- and constant-space-overhead fault-tolerant quantum computation with quantum low-density parity-check codes*, [arXiv:2411.03683](#).
- [18] Q. T. Nguyen and C. A. Pattison, *Quantum fault tolerance with constant-space and logarithmic-time overheads*, [arXiv:2411.03632](#).
- [19] L. Z. Cohen, I. H. Kim, S. D. Bartlett, and B. J. Brown, *Low-overhead fault-tolerant quantum computing using long-range connectivity*, [Sci. Adv.](#) **8**, eabn1717 (2022).
- [20] A. Cross, Z. He, P. Rall, and T. Yoder, *Improved QLDPC surgery: Logical measurements and bridging codes*, [arXiv:2407.18393](#).
- [21] A. Cowtan and S. Burton, *CSS code surgery as a universal construction*, [Quantum](#) **8**, 1344 (2024).
- [22] G. Zhang and Y. Li, *Time-efficient logical operations on quantum low-density parity check codes*, [Phys. Rev. Lett.](#) **134**, 070602 (2025).
- [23] B. Ide, M. G. Gowda, P. J. Nadkarni, and G. Dauphinais, *Fault-tolerant logical measurements via homological measurement*, [Phys. Rev. X](#) **15**, 021088 (2025).
- [24] D. J. Williamson and T. J. Yoder, *Low-overhead fault-tolerant quantum computation by gauging logical operators*, [Nat. Phys.](#) **22**, 598 (2026).
- [25] G. Zhang, Y. Zhu, X. Yuan, and Y. Li, *Constant-overhead magic state injection into QLDPC codes with error independence guarantees*, [arXiv:2505.06981](#).
- [26] Z. He, A. Cowtan, D. J. Williamson, and T. J. Yoder, *Extractors: QLDPC architectures for efficient pauli-based computation*, [arXiv:2503.10390](#).
- [27] E. Swaroop, T. Jochym-O'Connor, and T. J. Yoder, *Universal adapters between quantum low-density parity check codes*, [PRX Quantum](#) **7**, 010324 (2026).
- [28] A. Cowtan, Z. He, D. J. Williamson, and T. J. Yoder, *Parallel logical measurements via quantum code surgery*, [PRX Quantum](#) **7**, 020325 (2026).
- [29] N. Baspın, L. Berent, and L. Z. Cohen, *Fast surgery for quantum LDPC codes*, [arXiv:2510.04521](#).
- [30] I. Dinur, T.-C. Lin, and T. Vidick, *Expansion of high-dimensional cubical complexes: With application to quantum locally testable codes*, in *Proceedings of the 2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2024), pp. 379–385.
- [31] A. Leverrier, V. Londe, and G. Zémor, *Towards local testability for quantum coding*, [Quantum](#) **6**, 661 (2022).
- [32] T.-C. Lin and M.-H. Hsieh, *c^3 -locally testable codes from lossless expanders*, [arXiv:2201.11369](#).
- [33] D. Litinski, *A game of surface codes: Large-scale quantum computing with lattice surgery*, [Quantum](#) **3**, 128 (2019).
- [34] Throughout this paper, when discussing devised sticking, we refer specifically to the case where the logical operators to be measured are (i) expressed in standard form, and (ii) have logical thickness one, meaning they act on mutually disjoint sets of logical qubits [22].
- [35] E. T. Campbell, *A theory of single-shot error correction for adversarial noise*, [Quantum Sci. Technol.](#) **4**, 025006 (2019).
- [36] A. Cowtan, Z. He, D. J. Williamson, and T. J. Yoder, *Fast and fault-tolerant logical measurements: Auxiliary hypergraphs and transversal surgery*, [arXiv:2510.14895](#).
- [37] A. Leverrier and G. Zémor, *Quantum tanner codes*, in *Proceedings of the 2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2022), pp. 872–883.
- [38] T.-C. Lin and M.-H. Hsieh, *Good quantum LDPC codes with linear time decoder from lossless expanders*, [arXiv:2203.03581](#).
- [39] M. Sipser, and D. A. Spielman, *Expander codes*, [IEEE Trans. Inf. Theory](#) **42**, 1710 (1996).
- [40] R. Tanner, *A recursive approach to low complexity codes*, [IEEE Trans. Inf. Theory](#) **27**, 533 (1981).
- [41] Y. Li, *A magic state's fidelity can be superior to the operations that created it*, [New J. Phys.](#) **17**, 023037 (2014).
- [42] C. Gidney and A. G. Fowler, *Efficient magic state factories with a catalyzed $|CCZ\rangle$ to $2|T\rangle$ transformation*, [Quantum](#) **3**, 135 (2019).
- [43] In a discussion with Armands Strikis (unpublished data, 2024).
- [44] D. B. West, *Introduction to Graph Theory*, 2nd ed. (Prentice Hall, Upper Saddle River, NJ, 2001), pp. 533–568, includes bibliographical references and indexes.
- [45] M. A. Tremblay, N. Delfosse, and M. E. Beverland, *Constant-overhead quantum error correction with thin planar connectivity*, [Phys. Rev. Lett.](#) **129**, 050504 (2022).
- [46] E. Arjomandi, *An efficient algorithm for colouring the edges of a graph with $\delta + 1$ colours*, [INFOR: Inf. Syst. Operat. Res.](#) **20**, 82 (1982).
- [47] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation*, [Phys. Rev. A](#) **86**, 032324 (2012).
- [48] Q. Xu, J. P. Bonilla Ataides, C. A. Pattison, N. Raveendran, D. Bluvstein, J. Wurtz, B. Vasić, M. D. Lukin, L. Jiang, and H. Zhou, *Constant-overhead fault-tolerant quantum computation with reconfigurable atom arrays*, [Nat. Phys.](#) **20**, 1084 (2024).
- [49] M. P. C. Fossorier, *Quasi-cyclic low-density parity-check codes from circulant permutation matrices*, [IEEE Trans. Inf. Theory](#) **50**, 1788 (2004).