Multilevel distillation of magic states for quantum computing

Cody Jones*

Edward L. Ginzton Laboratory, Stanford University, Stanford, California 94305-4088, USA

(Received 17 October 2012; revised manuscript received 6 February 2013; published 4 April 2013)

We develop a procedure for distilling magic states used in universal quantum computing that requires substantially fewer initial resources than prior schemes. Our distillation circuit is based on a family of concatenated quantum codes that possess a transversal Hadamard operation, enabling each of these codes to distill the eigenstate of the Hadamard operator. A crucial result of this design is that low-fidelity magic states can be consumed to purify other high-fidelity magic states to even higher fidelity, which we call multilevel distillation. When distilling in the asymptotic regime of infidelity $\epsilon \to 0$ for each input magic state, the number of input magic states consumed on average to yield an output state with infidelity $O(\epsilon^{2^r})$ approaches $2^r + 1$, which comes close to saturating the conjectured bound in another investigation [Bravyi and Haah, Phys. Rev. A **86**, 052329 (2012)]. We show numerically that there exist multilevel protocols such that the average number of magic states consumed to distill from error rate $\epsilon_{\text{in}} = 0.01$ to ϵ_{out} in the range 10^{-5}–10^{-40} is about $14 \log_{10}(1/\epsilon_{\text{out}}) - 40$; the efficiency of multilevel distillation dominates all other reported protocols when distilling Hadamard magic states from initial infidelity 0.01 to any final infidelity below 10^{-7}. These methods are an important advance for magic-state distillation circuits in high-performance quantum computing and provide insight into the limitations of nearly resource-optimal quantum error correction.

DOI: 10.1103/PhysRevA.87.042305 PACS number(s): 03.67.Ac

I. INTRODUCTION

Quantum computing can potentially solve a handful of otherwise intractable problems such as factoring large integers [1] or simulating quantum physics [2]. Though the number of applications with a known quantum speed-up is small, some are quite valuable, such as the preceding examples. Quantum computations depend on coherent entangled states that are very sensitive to noise, so fault-tolerant quantum computing addresses imperfections in physical hardware with error-correcting codes [3,4], the most studied of which are stabilizer codes [5]. However, while quantum codes protect against noise, no code natively supports a universal set of transversal gates for simulating any quantum circuit [6,7]. To achieve universal quantum computing with error correction, Bravyi and Kitaev proposed a solution [8] that has received considerable attention: Inject faulty magic states into the code, purify them using the error-corrected gates, and then consume them to implement otherwise unavailable quantum circuits. These states are magic because it is possible to distill a subset of high-fidelity states from an ensemble of faulty states and because they enable universal fault-tolerant quantum computation.

Magic-state distillation has been the subject of intense investigation in recent years. Knill introduced a distillation procedure for $|H\rangle$, the $+1$ eigenstate of the Hadamard operation [9], independently of the work by Bravyi and Kitaev [8]. Reichardt showed that these protocols were equivalent and introduced an improvement that increased the threshold error rate [10]. More recently, Meier *et al.* introduced a 10-to-2 distillation procedure based on a code with two encoded qubits [11] and Bravyi and Haah introduced a $(3k + 8)$-to-k procedure using so-called triorthogonal codes with even k encoded qubits [12]. The distillation procedures we develop herein continue this trend of using larger, multiqubit codes. The relationship of magic-state distillation to other aspects of quantum information has also been an area of active study. Fowler and Devitt have proposed methods to reduce the size of distillation circuits when using topological quantum error correction [13]. Magic-state distillation has been demonstrated experimentally in NMR [14]. Additionally, distillation protocols for qudits have been proposed and analyzed [15,16].

For a quantum state, we quantify the probability of it having an error using the infidelity $1 - F$, where $F = \langle\psi|\rho|\psi\rangle$ is the fidelity between some mixed state ρ and the ideal state $|\psi\rangle$. The initial $|H\rangle$ states are prepared in a faulty manner before being injected into a fault-tolerant quantum code and Reichardt proved that the theoretical-limit infidelity for $|H\rangle$ states to be distillable is about 0.146 [10]. Campbell and Browne examined further properties of mixed states that may be distilled [17]. The efficiency of distilling high-infidelity magic states to low infidelity is of great importance to fault-tolerant quantum computing. Although magic states are the widely preferred method for achieving universality, distillation circuits are currently estimated to require the majority of resources in a quantum computer [18,19]. Therefore, advances in distillation protocols are important steps toward making quantum computing possible.

This paper presents two important, related results. First, we specify a family of $[[n,(n-4),2]]$ Calderbank-Shor-Steane (CSS) quantum stabilizer codes [20,21] known as H codes with transversal Hadamard operation, for even $n \geqslant 6$. A transversal quantum operation is one where a gate acting on a logical, encoded qubit is implemented by independent gates on each qubit in that code block (see p. 483 of Ref. [4]). The H codes are dense because the ratio of logical qubits to physical qubits $(n-4)/n \to 1$ as $n \to \infty$. Second, we demonstrate that concatenated versions of H codes allow for distillation of high-fidelity encoded magic states by consuming low-fidelity magic-state ancillas. We call this multilevel distillation because each such protocol takes

[*] ncodyjones@gmail.com

two types of magic-state inputs, which have different levels of infidelity and are applied at different concatenation levels within the distillation circuit. Multilevel protocols lead to the most efficient procedure for distilling magic states reported so far. For suitably small infidelity ϵ in each input magic state with independent errors, there exists a sequence of multilevel protocols that yields output magic states with infidelity $O(\epsilon^{2^r})$ and requires asymptotically $2^r + 1$ input states per distilled output state. This efficiency comes close to the optimality bound conjectured in Ref. [12]. For the purposes of developing quantum devices, this result is useful for exposing limits for optimizing quantum error correction. While this result is interesting theoretically, we also numerically study the distillation efficiency for $\epsilon_{\text{in}} = 0.01$, which is of practical importance to fault-tolerant quantum computing. We find that multilevel distillation is superior to previously reported protocols when the final infidelity is below 10^{-7}.

Throughout this paper, we adopt the following notation for single-qubit Pauli operators, for readability: $\mathrm{X} \equiv \sigma^x$, $\mathrm{Z} \equiv \sigma^z$, and I is the identity operator on a qubit. Additionally, we use "physical qubit" to denote those qubits used to produce a quantum code, whereas "logical qubits" are the protected information inside the code, again for readability. It may be the case that physical qubits for one encoding level are themselves the logical qubits of another code, which is a standard technique of quantum code concatenation [3,4,22].

II. FAMILY OF CODES WITH A TRANSVERSAL HADAMARD OPERATION

We define a family of CSS quantum codes that encode an even number k logical qubits using $k + 4$ physical qubits and possess a transversal Hadamard operation, so we call them collectively H codes and denote H_n as the code using $n = k + 4$ physical qubits. Any H code may be defined as follows. The stabilizer generators are $S_1 = \mathrm{X}_1\mathrm{X}_2\mathrm{X}_3\mathrm{X}_4$, $S_2 = \mathrm{Z}_1\mathrm{Z}_2\mathrm{Z}_3\mathrm{Z}_4$, $S_3 = \mathrm{X}_1\mathrm{X}_2\mathrm{X}_5\mathrm{X}_6 \cdots \mathrm{X}_n$, $S_4 = \mathrm{Z}_1\mathrm{Z}_2\mathrm{Z}_5\mathrm{Z}_6 \cdots \mathrm{Z}_n$, where subscripts index over physical qubits and tensor product between Pauli operators is implicit. The logical Pauli operators (corresponding to logical qubits), denoted with an overbar and indexed by $i = 1, \ldots, k$, are $\overline{\mathrm{X}}_i = \mathrm{X}_1\mathrm{X}_3\mathrm{X}_{i+4}$ and $\overline{\mathrm{Z}}_i = \mathrm{Z}_1\mathrm{Z}_3\mathrm{Z}_{i+4}$. The Hadamard transform exchanges X and Z operators, so application of transversal Hadamard gates at the physical level enacts a transversal Hadamard operation at the logical level, which will be a useful property when we later concatenate these codes. All H codes have distance 2, which means they can detect a single physical Pauli error. The product of two logical Pauli operators of the same type for two distinct logical qubits has weight 2 (number of nonidentity physical, single-qubit Pauli operators); the product of same-type Pauli operators on all logical qubits is also weight 2 at the physical level. The stabilizers come in matched X-Z pairs, so there are no weight-1 logical operators.

The $+1$ eigenstate $|H\rangle = \cos(\pi/8)|0\rangle + \sin(\pi/8)|1\rangle$ of the Hadamard operator $\mathrm{H} = (1/\sqrt{2})(\mathrm{X} + \mathrm{Z})$ is a magic state for universal quantum computing [8–12]. In particular, two of these magic states can be consumed to implement a controlled-H operation [9,11], enabling one to measure in the basis of H [see Fig. 1(a)]. Our distillation procedure is as follows: (a) Encode faulty $|H\rangle$ magic states in an H code, (b) measure in the basis of the transversal Hadamard gate by consuming $|H\rangle$ ancillas, and (c) reject the output states if either the measure-Hadamard or code-stabilizer circuits detect an error. For example, when an H_{k+4} code is used for distillation, k, $|H\rangle$ states are encoded as logical qubits using $k + 4$ physical qubits. Each transversal controlled-Hadamard gate consumes two $|H\rangle$ states [11] and this gate is applied to all physical qubits, which results in the $(3k + 8)$-to-k input-output distillation efficiency of these codes. A diagram of the quantum circuit for distillation using H_6 is shown in Fig. 1(b).

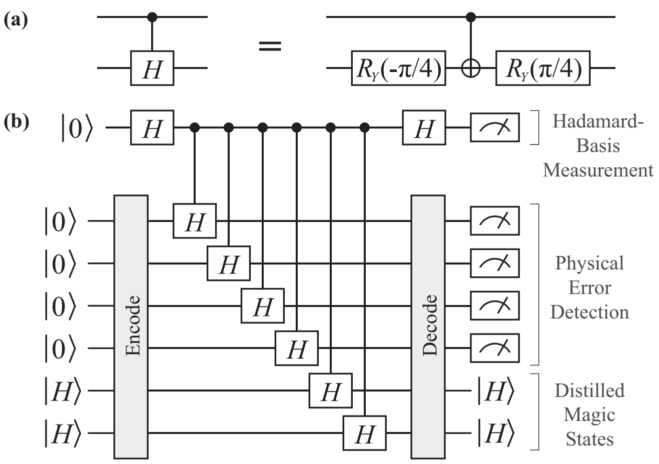

FIG. 1. Distillation of $|H\rangle$ magic states using an H code: (a) controlled-Hadamard gate constructed using $R_Y(-i\pi/4) = \exp(i\pi\sigma^Y/8)$ and its inverse, each requiring one $|H\rangle$ state [11], and (b) initial $|H\rangle$ states (left) encoded with four additional qubits, initialized to $|0\rangle$ here. The boxes "Encode" and "Decode" represent quantum circuits for encoding and decoding, which are not shown here.

III. MULTILEVEL DISTILLATION

Multilevel distillation uses concatenated codes with transversal Hadamard operation for distillation, in such a manner that the protocol takes as input magic states at two different levels of infidelity, and the two types of magic states enter at different concatenation levels in the code. The $|H\rangle$ ancillas consumed for transveral controlled-Hadamard measurement are of lower fidelity than the encoded logical $|H\rangle$ states being distilled. When two quantum codes with transversal Hadamard operation are concatenated, the resulting code also has transversal Hadamard operation. Under appropriate conditions, the distance of the concatenated code is the product of the distances for the individual codes: $d' = d_1 d_2$ [11]. Thus the concatenation of two H codes yields a distance-4 code with transversal Hadamard operation and r-level concatenation has distance 2^r.

The concatenation conditions for H codes are that, through all levels of concatenation, any pair of physical qubits has at most one encoding block (at any level) in common. The reasons for this restriction are that logical errors in the same block are correlated and that the statement above that distance multiplies through concatenation assumes independence of errors, so two qubits from the same encoding block can never be paired again in a different encoding block. The required arrangement of

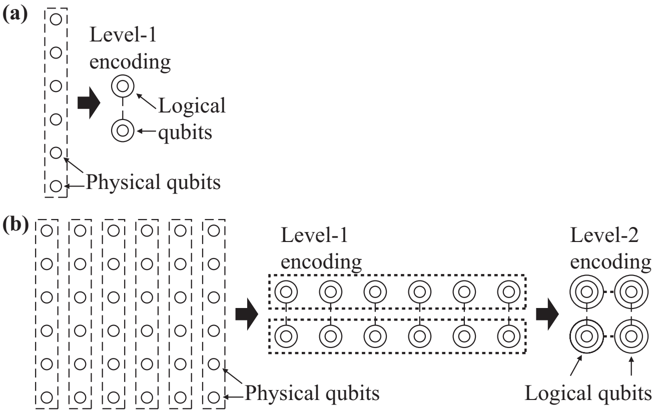

FIG. 2. Concatenation of H codes: (a) six physical qubits coupled into an H_6 code with two logical qubits and (b) a 6×6 array of physical qubits coupled into a concatenated two-level H_6 code.

qubits can be given a geometric interpretation. Arrange all physical qubits at points on a Cartesian grid in the shape of a rectangular solid, with the number of dimensions given by the number of levels of concatenation. A square, cube, or hypercube are possible examples at dimensionality 2, 3, or 4. Each dimension is associated with a level of concatenation and there must be an even $n \geqslant 6$ qubits in each dimension to form an H code. Construct H codes in the first dimension by forming an encoding block with each line of qubits in this direction, as in Fig. 2(a). This will give rise to $k = n - 4$ logical qubits along each line in this direction. Repeat this procedure by grouping these first-level logical qubits in lines along the second dimension to form logical qubits in a two-level concatenated code, as in Fig. 2(b). Continuing in this fashion through all dimensions ensures that any pair of qubits has at most one encoding block in common.

As with the H codes, multilevel codes use a transversal logical Hadamard-basis measurement to detect whether any one encoded qubit has an error (an even number of encoded errors would not be detected). If the logical $|H\rangle$ states have independent error probabilities ϵ_l, then the distilled states will have infidelity $O(\epsilon_l^2)$ with perfect distillation. We must also consider whether the Hadamard-basis measurement has an error. For a two-level code arranged as a square of side length n, the transversal controlled-Hadamard gates at the lowest physical level require $2n^2$ $|H\rangle$ magic states, each of which has infidelity ϵ_p. However, this is a distance-4 code, so for independent input-error rates, the probability of failing to detect errors at the physical level is $O(\epsilon_p^4) + O(\epsilon_l \epsilon_p^2)$ (rigorous analysis is provided later). The code can detect more errors in the magic states at the lower physical level, so these $|H\rangle$ states can be of lower fidelity than the magic states encoded as logical qubits and successfully perform distillation. This is the essential distinction between multilevel distillation and all prior distillation protocols. When multiple rounds of distillation are required [19], low-fidelity magic states are less expensive to produce, so multilevel protocols achieve higher efficiency.

Multilevel distillation protocols are applied in rounds, beginning with a small protocol (such as an H code) and progressing to concatenated multilevel codes. Let us denote the output infidelity from a single round by the function $\epsilon_{\text{out}} = E_t^{n_1 \times \cdots \times n_t}(\epsilon_l, \epsilon_p)$. For each such function, t is the dimensionality (number of levels of concatenation) and $n_1, \ldots, n_t$ are the sizes of each dimension, which need not all be the same. As before, ϵ_l and ϵ_p refer to the independent error probabilities on logical and physical magic states, respectively. A typical progression of rounds using a source of $|H\rangle$ states with infidelity ϵ_0 might be $\epsilon_1 = E_1^n(\epsilon_0, \epsilon_0)$, $\epsilon_2 = E_2^{n \times n}(\epsilon_1, \epsilon_0)$, etc.

Multilevel distillation circuits tend to be much larger in both qubits and gates than other protocols. Because there can be many encoded qubits, the protocol is still very efficient, but the size of the overall circuit may be a concern for some quantum computing architectures. At any number of levels, the distilled output states have correlated errors, so distilled magic-state qubits in our protocol must never meet again in a subsequent distillation circuit (we require that errors are independent within the same encoding block, as in Refs. [11,12]). Let us suppose that one performs two rounds of distillation, where the first round uses one-level distillers with k encoded magic states and the second round uses two-level distillers with k^2 encoded states. Because the inputs to each distiller in the second round must have independent errors, there must be k^2 independent distillation blocks in the first round. Therefore, to distill k^3 output states through two rounds, we require k^3 (logical inputs) $+ 2k^2(k+4)$ (physical inputs) $+ 2k(k+4)^2$ (physical inputs) $= 5k^3 + 24k^2 + 32k$ input states.

Consider a similar sequence through r rounds with each distiller in round q having k^q encoded qubits. The total number of *logical* magic states is $k^r \times k^{r-1} \times \cdots \times k = k^{r(r+1)/2}$ to ensure that errors are independent between logical magic states in every round. In the first round, the number of consumed magic states is $2(k+4)k^{r(r+1)/2-1}$; in any subsequent round $q \geqslant 2$, the number of consumed magic states is $2^{q-1}(k+4)^q k^{r(r+1)/2-q}$ (recall that the Hadamard measurement is implemented 2^{q-2} times, meaning it is repeated for $q \geqslant 3$). The total number of input magic states can thus be expressed as

$$\left[1 + \frac{k+4}{k} + \sum_{q=1}^{r} 2^{q-1}\left(\frac{k+4}{k}\right)^q\right] k^{r(r+1)/2}. \tag{1}$$

For $r = 2$, this reproduces the expression above. What also becomes clear is that the total size of multilevel protocols becomes unwieldy as r and k increase. For example, the case of $r = 3$ and $k = 10$ would require about 1.87×10^7 input magic states and a comparable number of quantum gates to distill 10^6 output magic states. However, since efficient multilevel distillation protocols, measured in the ratio of low-fidelity $|H\rangle$ input states consumed to yield a single high-fidelity $|H\rangle$ output, use $k \gg 1$ and multiple rounds, the greatest benefit from their application is seen in large-scale quantum computing, where a typical algorithm run may require 10^{12} magic states, each with error probability 10^{-12} [19]. Moreover, alternative designs can circumvent these issues. If the first round uses a different protocol without correlated errors across logical magic states, such as Bravyi-Kitaev 15-to-1 distillation, then having multiple distillation blocks is unnecessary in the second round using a two-level concatenated protocol, which would lead to smaller multiround, multilevel protocols. Indeed, Sec. IV shows that optimal protocols found by numerical search happen to take this approach.

The scaling exponent γ of a distillation protocol characterizes its efficiency. Specifically, $O(\log^{\gamma}(\epsilon_{\text{in}}/\epsilon_{\text{out}}))$ input states are required to distill one magic state of infidelity ϵ_{out}. Scaling exponents for previous protocols are $\gamma \approx 2.46$ (15-to-1 [8,9]), $\gamma \approx 2.32$ (10-to-2 [11]), and $\gamma \approx 1.6$ (triorthogonal codes [12]). Moreover, Bravyi and Haah conjecture that no magic-state distillation protocol has $\gamma < 1$ [12]. In this work, if each round of distillation uses one higher level of concatenation in the multilevel protocols, then the number of consumed inputs doubles. In the limits of $k \to \infty$ and $\epsilon \to 0$, multilevel protocols require $2^r + 1$ input states to each output state for r rounds of distillation, where the rth round is a level-r distiller. The final infidelity is $O((\epsilon_{\text{in}})^{2^r})$, so the scaling exponent is $\gamma = \log(2^r + 1)/\log(2^r) \to 1$ as $r \to \infty$, which is the closest any protocol has come to reaching the conjectured bound. We show later through numerical simulation that $\gamma \approx 1$ for error rates relevant to quantum computing.

IV. ANALYSIS

We make the conventional assumption that all quantum circuit operations are perfect, except for the initial $|H\rangle$ magic states we intend to distill. This is a valid approximation if all operations are performed using fault-tolerant quantum error correction where the logical gate error is far below the final infidelity for distilled magic states [3,19]; for a more explicit construction of fault-tolerant distillation circuits, see Ref. [13]. Additionally, following the methodology in Refs. [8,11], we can consider each magic state with infidelity ϵ as the mixed state $\rho = (1-\epsilon)|H\rangle\langle H| + \epsilon|-H\rangle\langle -H|$, where $|-H\rangle$ is the -1 eigenstate of the Hadamard operation.

Determining the infidelity at the output of distillation becomes simply a matter of counting the distinct ways that errors lead to the circuit incorrectly accepting faulty states. This process is aided by the geometric picture from earlier, and details are given in Appendix A. It is essential that error probabilities ϵ_l and ϵ_p for each input magic state are independent. Then a one-level $(3k+8)$-to-k distiller using the H_{k+4} code has output-error rate on each $|H\rangle$ state as

$$E_1^{k+4}(\epsilon_l,\epsilon_p) = (k-1)\epsilon_l^2 + (2k+2)\epsilon_p^2 + \cdots, \qquad (2)$$

where higher-order terms denoted by the ellipsis are omitted. Our numerical results justify the use of lowest-order approximations as higher-order terms are negligible in optimally efficient protocols. The lowest-order error rates are both second order because the Hadamard-basis measurement and H_{k+4} code can together detect a single error in any magic state. The probability of the distiller detecting an error, in which case the output is discarded, is $k\epsilon_l + 2(k+4)\epsilon_p + \cdots$. If $\epsilon_l = \epsilon_p = \epsilon$, then the output-error rate of $(3k+1)\epsilon^2$ conditioned on success is the same as in Ref. [12]. Using the two-level distiller constructed from concatenated H_{k+4} codes, the output infidelity for each distilled $|H\rangle$ state is

$$E_2^{(k+4)\times(k+4)}(\epsilon_l,\epsilon_p) = (k^2-1)\epsilon_l^2 + 8(k^2+4k+3)\epsilon_p^4 + (k+4)^2\epsilon_l\epsilon_p^2 + \cdots. \qquad (3)$$

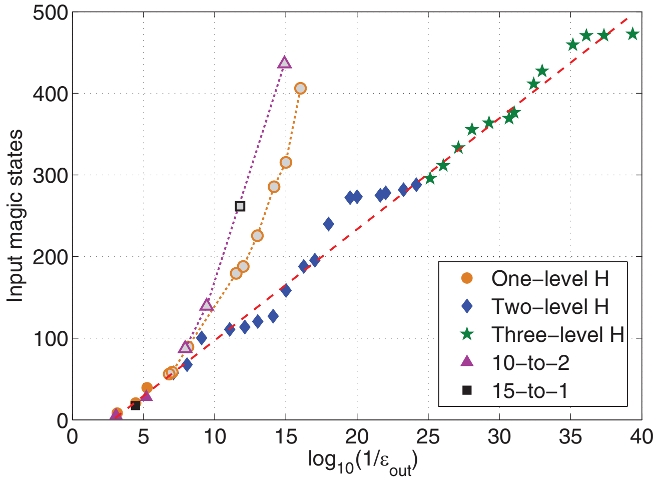

FIG. 3. (Color online) Average number of input $|H\rangle$ states with $\epsilon_{\text{in}} = 0.01$ consumed to produce a single output $|H\rangle$ state with fidelity ϵ_{out}. Multiple-round distillation can use different protocols in each round and the markers indicate just the last round of distillation. The gray-shaded squares, triangles, and circles show, respectively, the best distillation possible with only 15-to-1 [8], 10-to-2 [11], and triorthogonal-code [12] protocols. The dashed line is a linear fit $14\log_{10}(1/\epsilon_{\text{out}}) - 40$.

The probability of the two-level distiller detecting an error is $k^2\epsilon_l + 2(k+4)^2\epsilon_p + 2k^2(k+4)^2\epsilon_l\epsilon_p + \cdots$. Similar error suppression extends to higher multilevel protocols, as examined in Appendix A.

Figure 3 shows the performance of optimal multiround distillation protocols identified by numerical search, indicating the number of input states with $\epsilon_0 = 0.01$ required to reach a desired output infidelity ϵ_{out}. The markers indicate the type of protocol in the last round of distillation, including Bravyi-Kitaev [8], Meier-Eastin-Knill [11], and multilevel H codes (see Appendix B for details). The search attempts to identify the best distillation routines using any combination of known methods. Note that the recent Bravyi-Haah protocols [12] have the same performance as one-level H codes. As expected, there is a trend of using higher-distance multilevel protocols in the last round as the output-error rate ϵ_{out} decreases (earlier rounds may use different protocols). Where present, open markers indicate the best possible performance of previously studied protocols without the advent of multilevel distillation and multilevel distillation is dominant for $\epsilon_{\text{out}} \leqslant 10^{-7}$, which is the regime pertinent to quantum computing. Moreover, in this regime, input-error rates are sufficiently small that only lowest-order terms in the $E(\cdot)$ output-error functions are significant. The linear fit provides empirical evidence that the scaling exponent is $\gamma \approx 1$ in this regime, which demonstrates that multilevel protocols are close to the conjectured optimal performance in practice.

V. CONCLUSION

H codes can distill magic states for $\text{T} = \exp[i\pi(I-\sigma^z)/8]$, which may enable distillation of three-qubit magic states for controlled-controlled-Z, which is locally equivalent to the Toffoli gate [4] (see Appendixes C and D for details). As a first pass at studying distillation protocols, this work considered only average input-to-output efficiency. Future work will more rigorously examine the entire costs in qubits and gates required to fault-tolerantly distill magic states using multilevel codes [23]. Multilevel distillation is an important development for large-scale fault-tolerant quantum computing, where the distillation of magic states is often considered the most costly subroutine [18,19]. Other codes with high-density, high-distance, and transversal Hadamard operations may yet be discovered, though for the present, H codes are useful for their high efficiency and simple construction.

ACKNOWLEDGMENTS

The author would like to thank Sergey Bravyi and Bryan Eastin for assistance in analyzing protocols at levels three and higher. This work was supported by the University of Tokyo Special Coordination Funds for Promoting Science and Technology, NICT, and the Japan Society for the Promotion of Science through its FIRST Program.

APPENDIX A: ERROR ANALYSIS IN MULTILEVEL H-CODE DISTILLATION

The multilevel codes analyzed here use concatenated H codes. When two H codes are concatenated, the logical qubits of the first level of encoding are used as physical qubits for completely distinct codes at the second level. Consider a two-level scheme: If the codes at the first and second levels are $[[n_1,(n_1-4),2]]$ and $[[n_2,(n_2-4),2]]$, respectively, then the concatenated code is $[[n_1 n_2,(n_1-4)(n_2-4),4]]$, as shown in Fig. 2(b) of the main text. This process can be extended to higher levels of concatenation.

Determining the potential errors and their likelihood in multilevel protocols requires careful analysis. Let us enumerate the error configurations that are detected by the protocols; the error probability is given by summing the probability of all error configurations that are not detected and lead to error(s) in the encoded $|H\rangle$ states. As a first step, we may simplify the analysis of multilevel codes by considering each input magic state to our quantum computer as having an independent probability of σ^Y error, as discussed in Refs. [8,11]. This allows us to consider only one type of error stemming from each magic state used in the protocol.

Identifying undetected error events in multilevel distillation, which lead to the output-error rate, is aided by the geometric picture introduced in the main text. Qubits that will form the code are arranged in a rectangular solid and then grouped in lines along each dimension for encoding. There are two error-detecting steps that together implement distillation: the Hadamard-basis measurement and the error detection of the H codes. The Hadamard measurement registers an error for odd parity in the total of encoded state errors and physical-level errors in the first round of $R_Y(-\pi/4)$ gates and there is one of these for each qubit site in the code (see Fig. 1 of the main text).

The second method for H codes to detect errors is by measuring the code stabilizers. The code stabilizers detect any configuration of errors that is not a logical operator in the concatenated code. Because of the redundant structure using overlapping H codes, only a very small fraction of error configurations evade detection. Before moving on, note that at each qubit site, there are two faulty gates applied and two errors on the same qubit will cancel (however, the first error will propagate to the Hadamard-basis measurement). Conversely, a single error in one of the two gates will propagate to the stabilizer-measurement round, but only an error in the first gate will also propagate to the Hadamard measurement. The stabilizer-measurement round will only see the odd-even parity of the number of errors at each qubit site.

One type of error event that occurs at concatenation levels 3 and higher requires special treatment. If there is an error in an encoded magic state and errors on two physical states used for the same controlled-Hadamard gate at the physical level, then this combination of input errors is not detected by the distillation protocol, leading to logical output error. This event leads to the $O(\epsilon_l \epsilon_p^2)$ error probability from the main text, which is not an issue for two-level protocols, but it must be addressed in levels 3 and higher. The solution for t-level distillation, where $t \geqslant 3$, is to repeat the controlled-Hadamard measurement 2^{t-2} times, consuming 2^{t-1} magic states at the physical level. After each transversal controlled-Hadamard operation, the code syndrome checks for detectable error patterns. With this procedure, one encoded-state error would also require at least 2^{t-1} errors in physical-level magic states to go undetected, leading to probability of error that scales as $O(\epsilon_l \epsilon_p^{2^{t-1}})$.

Consider the pattern of errors after the two potentially faulty gates on each qubit in the t-dimensional Cartesian grid arrangement. The many levels of error checking in the H codes can detect a single error in any encoding block at any encoding level. For this analysis, let us separate the $k+4$ qubits in a single H code block into two groups: The first four qubits are preamble qubits, while the remaining k qubits are index qubits. The reason for distinction is that the logical $\overline{\mathrm{Y}}_i$ operators, which would also be undetected error configurations, have common physical-qubit operators in the preamble, with a degeneracy of two: $\overline{\mathrm{Y}}_i = -\mathrm{Y}_1\mathrm{Y}_3\mathrm{Y}_{i+4} = -\mathrm{Y}_2\mathrm{Y}_4\mathrm{Y}_{i+4}$, because of the stabilizer $\mathrm{Y}_1\mathrm{Y}_2\mathrm{Y}_3\mathrm{Y}_4$. Conversely, the logical operators are distinguished by the ith logical Pauli operator having a physical Pauli operator on the ith index qubit (numbered $i+4$ when preamble is included).

With the preamble-index distinction, we can now identify the most likely error patterns. For any size H code, there are two weight-2 errors in the preamble: $\mathrm{Y}_1\mathrm{Y}_2$ and $\mathrm{Y}_3\mathrm{Y}_4$. Logically, these represent the product of $\overline{\mathrm{Y}}$ operators on all encoded qubits. In the index qubits, any pair of errors is logical: $\mathrm{Y}_{i+4}\mathrm{Y}_{j+4} = \overline{\mathrm{Y}}_i\overline{\mathrm{Y}}_j$. However, a pair of errors split with one each in preamble and index is always detectable by the code stabilizers. Thus any single encoded qubit could have a logical error stemming from a pair of errors in two different configurations in the preamble or $k-1$ configurations in the index qubits. There is also one weight-3 error. Each physical-state error configuration is multiplied by a degeneracy factor that is the number of ways an even number of errors occur before the CNOT in Fig. 1, thereby evading the Hadamard measurement. Thus the probability of logical error is $2(k+1)\epsilon_p^2 + 4\epsilon_p^3 + O(\epsilon_p^4)$. The Hadamard measurement fails to detect an even number of errors in the logical input states. There are $k-1$ ways that a pair of encoded input errors could corrupt any given qubit and $(k-1)(k-2)(k-3)/6$ ways four errors could corrupt any given qubit (assuming $k \geqslant 4$). This contributes error terms $(k-1)\epsilon_l^2 + (1/6)(k-1)(k-2)(k-3)\epsilon_l^4 + O(\epsilon^6)$. Finally, it is possible for a single logical error and an odd number of physical errors before the CNOT in Fig. 1 of the main text, potentially in conjunction with other physical errors after CNOT, to occur simultaneously in a way that evades both checks. This contributes a term $(k+4)\epsilon_l\epsilon_p^2 + 8(k-1)\epsilon_l\epsilon_p^3 + O(\epsilon_l\epsilon_p^4)$.

The numerical analysis detailed below shows that efficient use of one-level H codes has similar error rates for ϵ_l and ϵ_p and both are below 0.01, so the relevant terms in the error functions for one-level H codes are $E_1^{k+4}(\epsilon_l,\epsilon_p) = (k-1)\epsilon_l^2 + (2k+2)\epsilon_p^2 + \cdots$, which reproduces Eq. (1) of the main text. As a result, the higher-order terms above can be neglected for this range of parameters so long as k is not too large. Simply put, if the higher-order terms become relevant (i.e., ϵ_l, ϵ_p, or k is sufficiently large in magnitude), then the distillation protocol is being used ineffectively and it may in fact cause more errors than it corrects. These findings are supported by the numerical search for optimal protocols and we proceed using this approximation.

When H codes are concatenated, the analysis of undetected error patterns becomes more complicated. In particular, logical errors from one layer of encoding must be matched with errors from other encoding blocks to go undetected at the next level. Consider the case of the level-2-concatenated square-array distiller and focus on one of the encoded states. As before, a pair of encoded-state input errors evades the Hadamard measurement, which contributes a term $(k^2-1)\epsilon_l^2$. The undetected errors resulting from consumed magic states are more complicated. Within the upper encoding block, there are two ways a logical error could be caused by a pair of errors in the preamble and $k-1$ possibilities for logical error from a pair of index errors. However, each of the inputs to the second level are the logical qubits of distinct H codes at the first level, which has additional error detection. The most likely errors from the first level come in pairs, but these pairs are sent to different codes at the second level. As a result, the error patterns from the first level must come in matched pairs that are also not detected at the second level. For any particular error configuration going into a block at the second level, there are four preamble configurations and $k-1$ index configurations at the first level that could have caused it. There are $k+1$ undetected error configurations at the second level and the degeneracy factor of four physical errors is 8, so the consumed magic states contribute a term $8(k+1)(k+3)\epsilon_p^4$. Finally, the most likely way that physical and encoded errors can occur in conjunction is a logical error on the magic state in question and two physical errors on the same qubit anywhere, which has probability $(k+4)^2\epsilon_l\epsilon_p^2$. Combined, these error terms reproduce the results in Eq. (2): $E_2^{(k+4)\times(k+4)}(\epsilon_l,\epsilon_p) = (k^2-1)\epsilon_l^2 + 8(k^2+4k+3)\epsilon_p^4 + (k+4)^2\epsilon_l\epsilon_p^2 + \cdots$. We drop terms at higher order because they are found to be negligible in optimal protocols. For example, the first optimal two-level protocol has parameters $k=8$, $\epsilon_l = 3.5\times10^{-5}$, and $\epsilon_p = 9\times10^{-4}$, where both input types come from earlier rounds of distillation (Bravyi-Kitaev and Meier-Eastin-Knill, respectively). More details of the numerical search are given below.

Continuing this approach, one can show the significant error terms at level $t \geqslant 3$ are given by

$$E_t^{(k+4)^t}(\epsilon_l,\epsilon_p) = (k^t-1)\epsilon_l^2 + 2^{2^t+t-3}(k+1)(k+3)^{t-1}\epsilon_p^{2^t} + (k+4)^{t(2^t-2)}\epsilon_l\epsilon_p^{2^{t-1}} + \cdots . \qquad \text{(A1)}$$

These terms incorporate degeneracy in error configurations and repeated Hadamard measurements. The coefficients of the second and third terms on the right-hand side of Eq. (1) represent physical error configurations and encoded-physical combinations, respectively, and these grow rapidly as a function of r. Accordingly, the optimal-protocol search does not advocate the use of three-level protocols until the desired output error rate is below 10^{-25}, which is beyond the needs of any quantum algorithm so far conceived. No four-level protocols were found to be optimal for output error rates above 10^{-40}, which under practical considerations means they are not likely to ever be used. The next section considers the size of multilevel distillation circuits, which can also limit their usefulness.

APPENDIX B: OPTIMAL MULTI-ROUND DISTILLATION

The claimed efficiency of multilevel distillation was examined quantitatively with a numerical search for optimal multiround distillation protocols. Each of the protocols is optimal in the sense that, for a given final infidelity ϵ_{out}, no other sequence requires fewer average input states and, for a given average number of input states, no other protocol achieves lower ϵ_{out}. Note that probability of rejection upon detected error is incorporated by considering average cost for distillation when failure-and-repeat steps are included. The protocols plotted in Fig. 3 of the main text are just the last round of a distillation sequence. Earlier rounds can be, and usually are, different protocols. The search space was constrained such that the number of rounds $r \leqslant 5$, number of encoded logical qubits $k \leqslant 20$ for all H codes, and multilevel codes are square $(k+4)\times(k+4)$, etc.

Generally speaking, smaller protocols handle large input-error rates in early rounds better, while larger multilevel protocols are more inefficient at distillation when input-error rates are low enough. For example, the protocols listed in Table I are the same ones plotted in Fig. 3. Note that when ϵ_{out} is smaller than 10^{-7}, multilevel protocols are most efficient. Should one desire $\epsilon_{\text{out}} < 10^{-25}$, level-3 protocols have the highest efficiency. Higher levels of concatenation (up to level 5) were part of the search, but they were not efficient for $\epsilon_{\text{out}} > 10^{-40}$. The error-function notation $E(\cdot)$ from the main text is used to show how the inputs to later rounds of distillation may be the outputs of an earlier round. This notation neglects the total size of the distillers, which must be determined by using parallel distillation blocks whenever correlated errors on logical magic states are present.

For reference, the best achievable results with prior protocols are also shown in Table I, in reverse chronological order of their discovery. The methods are cumulative, so the most recent Bravyi-Haah codes could also use Meier-Eastin-Knill or Bravyi-Kitaev distillation, but the oldest Bravyi-Kitaev distillation is alone in its column. The best achievable results with older protocols are also shown in Fig. 3 for comparison.

TABLE I. Optimal distillation protocols identified by numerical search. Protocols are specified using the error functions $E(\cdot)$ to indicate when inputs to one round are the outputs of another distillation circuit. The data for C_{BH} and C_{MEK} are obtained from Ref. [12].

$-\log_{10}$ $(\epsilon_{\mathrm{target}})$	$-\log_{10}$ $(\epsilon_{\mathrm{out}})$	Protocol	C_{ML}	C_{BH}	C_{MEK}	C_{BK}
4	4.46	$E_{\mathrm{BK}}(\epsilon_0)$	17.44	17.44	17.44	17.44
5	5.14	$E_{\mathrm{MEK}}(E_{\mathrm{BK}}(\epsilon_0))$	27.93	27.86	27.86	261.5
6	6.83	$E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0))$	56.07	56.07	83.99	261.5
7	7.11	$E_2^{12\times12}(E_{\mathrm{BK}}(\epsilon_0),E_{\mathrm{MEK}}(\epsilon_0))$	57.38	58.30	83.99	261.5
8	8.06	$E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{MEK}}(\epsilon_0))$	67.52	89.26	139.3	261.5
9	9.08	$E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{BH}}^2(\epsilon_0)),E_{\mathrm{BH}}^2(\epsilon_0))$	100.3	139.3	139.3	261.5
10	11.1	$E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0))$	110.7	179.4	261.7	261.5
11	11.1	$E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0))$	110.7	179.4	261.7	261.5
12	12.1	$E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BK}}(\epsilon_0),E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0))$	113.7	187.9	418.0	3923.0
13	13.0	$E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BH}}^8(E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0))$	120.4	225.6	418.0	3923.0
14	14.1	$E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0))$	126.9	285.6	419.9	3923.0
15	15.0	$E_2^{14\times14}(E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BH}}^{10}(E_{\mathrm{MEK}}(\epsilon_0)))$	158.5	315.5	696.7	3923.0
16	16.3	$E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	187.9	406.2	696.7	3923.0
17	17.0	$E_2^{22\times22}(E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	195.5	529.5	696.7	3923.0
18	18.0	$E_2^{20\times20}(E_2^{24\times24}(E_{\mathrm{BH}}^{38}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{BH}}^{40}(\epsilon_0)))$	239.8	574.1	1260.0	3923.0
19	19.5	$E_2^{24\times24}(E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	272.1	574.1	1260.0	3923.0
20	20.0	$E_2^{24\times24}(E_2^{24\times24}(E_{\mathrm{BH}}^{30}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	273.3	574.1	1260.0	3923.0
21	21.6	$E_2^{24\times24}(E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BK}}(\epsilon_0), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	275.1	575.9	1260.0	3923.0
22	22.0	$E_2^{24\times24}(E_2^{20\times20}(E_2^{10\times10}(E_{\mathrm{BK}}(\epsilon_0), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	278.0	604.3	1308.0	3923.0
23	23.3	$E_2^{24\times24}(E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BH}}^8(E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	281.9	652.3	2090.0	3923.0
24	24.2	$E_2^{24\times24}(E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BH}}^6(E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_2^{12\times12}(E_{\mathrm{BK}}(\epsilon_0),E_{\mathrm{MEK}}(\epsilon_0)))$	287.9	731.5	2090.0	3923.0
25	25.1	$E_3^{16\times16\times16}(E_2^{22\times22}(E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	295.7	853.1	2090.0	3923.0
26	26.1	$E_3^{16\times16\times16}(E_2^{14\times14}(E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	311.5	914.0	2090.0	3923.0
27	27.1	$E_3^{16\times16\times16}(E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{BH}}^2(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BH}}^6(E_{\mathrm{MEK}}(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	333.3	947.5	2100.0	3923.0
28	28.1	$E_3^{16\times16\times16}(E_2^{18\times18}(E_2^{10\times10}(E_{\mathrm{MEK}}(E_{\mathrm{BH}}^2(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	355.6	1015.0	2181.0	3923.0
29	29.3	$E_3^{16\times16\times16}(E_2^{18\times18}(E_2^{10\times10}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	363.7	1125.0	3483.0	3923.0
30	30.7	$E_3^{16\times16\times16}(E_2^{24\times24}(E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	369.3	1301.0	3483.0	3923.0
31	31.0	$E_3^{16\times16\times16}(E_2^{20\times20}(E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	376.5			3923.0
32	32.4	$E_3^{16\times16\times16}(E_2^{24\times24}(E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{BH}}^2(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	411.5			3923.0
33	33.0	$E_3^{14\times14\times14}(E_2^{20\times20}(E_2^{24\times24}(E_{\mathrm{BH}}^{38}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{MEK}}(E_{\mathrm{BH}}^2(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	427.3			3923.0
34	35.2	$E_3^{16\times16\times16}(E_2^{24\times24}(E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	459.4			58838.0
35	35.2	$E_3^{16\times16\times16}(E_2^{24\times24}(E_2^{24\times24}(E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0))),E_{\mathrm{MEK}}(E_{\mathrm{MEK}}(\epsilon_0)))$	459.4			58838.0
36	36.1	$E_3^{24\times24\times24}(E_2^{24\times24}(E_2^{24\times24}(E_{\mathrm{BH}}^{30}(E_{\mathrm{BK}}(\epsilon_0)), E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0))),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	470.8			58838.0
37	37.3	$E_3^{24\times24\times24}(E_2^{24\times24}(E_2^{24\times24}(E_2^{12\times12}(E_{\mathrm{BK}}(\epsilon_0), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0))),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	471.0			58838.0
38	39.4	$E_3^{24\times24\times24}(E_2^{24\times24}(E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BK}}(\epsilon_0), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0))),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	472.6			58838.0
39	39.4	$E_3^{24\times24\times24}(E_2^{24\times24}(E_2^{24\times24}(E_2^{10\times10}(E_{\mathrm{BK}}(\epsilon_0), E_{\mathrm{MEK}}(\epsilon_0)),E_{\mathrm{BK}}(\epsilon_0)),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0))),E_{\mathrm{BH}}^{40}(E_{\mathrm{BK}}(\epsilon_0)))$	472.6			58838.0

Accordingly, the multilevel protocols can use any of the above protocols wherever the numerical search finds doing so to be optimally efficient.

The numerical simulation uses error functions $E(\cdot)$ for the Bravyi-Kitaev (BK) [8] 15-to-1 and Meier-Eastin-Knill (MEK) [11] 10-to-2 protocols. The first three Taylor-series terms of these functions near $\epsilon = 0$ are

$$E_{\mathrm{BK}}(\epsilon) = 35\epsilon^3 + 105\epsilon^4 + 378\epsilon^5 + \cdots, \tag{B1}$$

$$E_{\mathrm{MEK}}(\epsilon) = 9\epsilon^2 - 56\epsilon^3 + 160\epsilon^4 + \cdots. \tag{B2}$$

In the numerical search, $\epsilon \leqslant 0.01$, so the first term for each dominates. The Bravyi-Haah (BH) triorthogonal codes [12] have the same error function as H codes [see Eq. (1) of the main text]:

$$E_{\mathrm{BH}}^{k}(\epsilon) = E_1^{k+4}(\epsilon,\epsilon) = (3k+1)\epsilon^2 + \cdots. \tag{B3}$$

This correspondence, combined with the results of Reichardt [10], suggests a connection between these code families.

APPENDIX C: DISTILLING MAGIC STATES FOR T GATES WITH H CODES

In addition to distilling Hadamard states $|H\rangle$, H codes can also distill the magic state $|A\rangle = (1/\sqrt{2})(|0\rangle + e^{i\pi/4}|1\rangle)$, which is used to make the T $= \exp[i\pi(I - \sigma^z)/8]$ gate. This construction is useful by itself, but it can also be used to make Toffoli magic states as shown in the next section. State $|A\rangle$

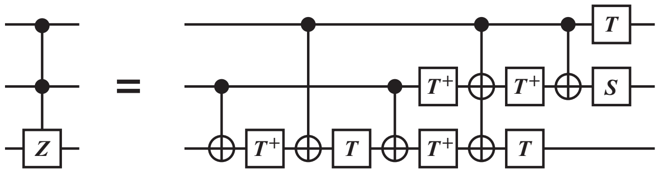

FIG. 4. A controlled-controlled-Z gate, which is locally equivalent to a Toffoli gate, can be decomposed into CNOT and T gates. A CSS + T code has transversal CNOT and T gates, so it could be used to distill three-qubit magic states for the Toffoli gate.

is stabilized by the operator $\mathrm{TXT}^\dagger = (1/\sqrt{2})(\mathrm{X}+\mathrm{Y})$, which is also transversal in H codes. Distillation is performed by encoding $|A\rangle$ states as logical qubits, then measuring the controlled-$\mathrm{TXT}^\dagger$ using $|A\rangle$ states at the physical level, followed by routine error detection.

APPENDIX D: DISTILLING TOFFOLI MAGIC STATES

A CSS quantum code [20,21] necessarily has a transversal CNOT operation. A CSS code with transversal T operation (let us use the shorthand CSS + T) will also have a transversal controlled-controlled-Z (CCZ) operation because the latter quantum gate can be decomposed into CNOT and T (or $\mathrm{T}^\dagger$), as shown in Fig. 4. The CCZ gate is locally equivalent to Toffoli (via Hadamard transforms on the target qubit), so CSS + T codes can distill a magic state for the CCZ gate, which is equivalent to distilling Toffoli magic states (see p. 488 of Ref. [4]), assuming Clifford operations are freely available.

[1] P. W. Shor, SIAM J. Comput. **26**, 1484 (1997).
[2] S. Lloyd, Science **273**, 1073 (1996).
[3] J. Preskill, Proc. R. Soc. London Ser. A **454**, 385 (1998).
[4] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 1st ed. (Cambridge University Press, Cambridge, 2000).
[5] D. Gottesman, Ph.D. thesis, California Institute of Technology, 1997.
[6] B. Zeng, A. Cross, and I. Chuang, IEEE Trans. Inf. Theory **57**, 6272 (2011).
[7] B. Eastin and E. Knill, Phys. Rev. Lett. **102**, 110502 (2009).
[8] S. Bravyi and A. Kitaev, Phys. Rev. A **71**, 022316 (2005).
[9] E. Knill, arXiv:quant-ph/0402171.
[10] B. W. Reichardt, Quant. Inf. Proc. **4**, 251 (2005).
[11] A. M. Meier, B. Eastin, and E. Knill, arXiv:1204.4221v1.
[12] S. Bravyi and J. Haah, Phys. Rev. A **86**, 052329 (2012).
[13] A. G. Fowler and S. J. Devitt, arXiv:1209.0510v3.
[14] A. M. Souza, J. Zhang, C. A. Ryan, and R. Laflamme, Nat. Commun. **2**, 169 (2011).
[15] E. T. Campbell, H. Anwar, and D. E. Browne, Phys. Rev. X **2**, 041021 (2012).
[16] V. Veitch, C. Ferrie, D. Gross, and J. Emerson, New J. Phys. **14**, 113011 (2012).
[17] E. T. Campbell and D. E. Browne, Phys. Rev. Lett. **104**, 030503 (2010).
[18] N. Isailovic, M. Whitney, Y. Patel, and J. Kubiatowicz, in *Proceedings of the 35th International Symposium on Computer Architecture, Beijing, China, 2008* (IEEE, Piscataway, NJ, 2008), pp. 177–188.
[19] N. C. Jones, R. Van Meter, A. G. Fowler, P. L. McMahon, J. Kim, T. D. Ladd, and Y. Yamamoto, Phys. Rev. X **2**, 031007 (2012).
[20] A. R. Calderbank and P. W. Shor, Phys. Rev. A **54**, 1098 (1996).
[21] A. Steane, Proc. R. Soc. London Ser. A **452**, 2551 (1996).
[22] E. Knill and R. Laflamme, arXiv:quant-ph/9608012.
[23] A. G. Fowler, S. J. Devitt, and C. Jones, arXiv:1301.7107v1.