Distributed quantum computation based on small quantum registers

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:41467456">http://nrs.harvard.edu/urn-3:HUL.InstRepos:41467456</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Distributed Quantum Computation Based-on Small Quantum Registers

Liang Jiang\textsuperscript{1}, Jacob M. Taylor\textsuperscript{1,2}, Anders S. Sørensen\textsuperscript{3}, Mikhail D. Lukin\textsuperscript{1}

\textsuperscript{1} Department of Physics, Harvard University, Cambridge, Massachusetts 02138
\textsuperscript{2} Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and
\textsuperscript{3} Quantop and The Niels Bohr Institute, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark

(Dated: October 26, 2018)

We describe and analyze an efficient register-based hybrid quantum computation scheme. Our scheme is based on probabilistic, heralded optical connection among local five-qubit quantum registers. We assume high fidelity local unitary operations within each register, but the error probability for initialization, measurement, and entanglement generation can be very high (∼5%). We demonstrate that with a reasonable time overhead our scheme can achieve deterministic non-local coupling gates between arbitrary two registers with very high fidelity, limited only by the imperfections from the local unitary operation. We estimate the clock cycle and the effective error probability for implementation of quantum registers with ion-traps or nitrogen-vacancy (NV) centers. Our new scheme capitalizes on a new efficient two-level pumping scheme that in principle can create Bell pairs with arbitrarily high fidelity. We introduce a Markov chain model to study the stochastic process of entanglement pumping and map it to a deterministic process. Finally we discuss requirements for achieving fault-tolerant operation with our register-based hybrid scheme, and also present an alternative approach to fault-tolerant preparation of GHZ states.

I. INTRODUCTION

The key challenge in experimental quantum information science is to identify isolated quantum mechanical systems with good coherence properties that can be manipulated and coupled together in a scalable fashion. Recently, considerable advances have been made towards interfacing of individual qubits in the optical and microwave regimes. These include advances in cavity QED as well as in probabilistic techniques for entangling remote qubits. At the same time, substantial progress has been made towards the physical implementation of few-qubit quantum registers using systems of coupled trapped ions, neutral atoms, or solid-state qubits based on either electronic and nuclear spins in semiconductors or superconducting islands.

While the precise manipulation of large, multi-qubit systems still remains an outstanding challenge, various approaches for connecting such few qubit sub-systems into large scale circuits have been investigated. These studies suggest that hybrid schemes, which benefits from short range interactions for local coupling and (optical) long range interactions for non-local coupling, might be an effective way toward large scale quantum computation: small local few-qubit quantum systems may be controlled with very high precision using optimal control techniques, and in practice it may be more feasible to operate several such small sub-systems compared to the daunting task of high-precession control of a single large quantum system with thousands of qubits. Optical techniques for quantum communication can then be used to connect any two sub-systems. For example, we may directly transfer a quantum state from one sub-system to another via an optical channel, which immediately provides an efficient way to scale-up the total number of the physical qubits we can manipulate coherently. In particular the use of optical means for connecting different subsystems has the advantage that it allows for fast non-local operations over large distances. This is advantageous for quantum error correction since the existence of such non-local coupling operations alleviates the threshold requirement for fault-tolerant quantum computation.

In practice, however, it is very difficult to have a perfect optical connection. In particular, there is excitation loss associated with the optical channel, due to scattering or absorption. For a lossy channel, it is therefore more desirable to use it to generate entanglement between different sub-systems (see Fig. 1), rather than for direct state transfer. The entanglement generation is then heralded by the click patterns from the photon detectors. Such detection-based scheme is intrinsically robust against excitation loss in the channel, since it only reduces the success probability but does not affect the entanglement fidelity. This entanglement can then be used as a resource to teleport quantum state from one sub-system to another. More generally, entanglement provides a physical resource to implement non-local unitary coupling gates (such as CNOT gate). If only one physical qubit for each sub-system, so-called cluster states can be created based on probabilistic, heralded entanglement generation (see Ref. and references therein). Such cluster states can be used for universal quantum computation. If there are two physical qubits available for each sub-system, cluster states can be created deterministically; meanwhile one can also use these two-qubit sub-systems to implement any quantum circuit directly.

Furthermore, realistic optical channel connecting sub-systems has other imperfections beside excitation loss, such as the distortion of the polarization or shape of the wave-packet. These imperfections will reduce the fidelity of the heralded entanglement generated between the sub-
three logical operations. In Ref. [15] it was found that purified Bell pair shared by the sub-systems can be very useful for quantum computation. In principle, the infidelity of the local physical resources, and Dur and Briegel [15] first provided a very efficient purification scheme in terms of quantum communication over long distances [31, 35, 36]. In particular, the entanglement which may create some high-fidelity Bell pairs from many entanglement purification schemes have been proposed, systems. To overcome these imperfections in the channel, entanglement purification schemes have been proposed, which may create some high-fidelity Bell pairs from many low-fidelity ones [33, 34]. In particular, the entanglement pumping scheme originally presented in the context of quantum communication over long distances [31, 35, 36] provides a very efficient purification scheme in terms of local physical resources, and Dur and Briegel [15] first proposed to use such entanglement pumping scheme for quantum computation. In principle, the infidelity of the purified Bell pair shared by the sub-systems can be very low, and is only limited by the error probability from local operations. In Ref. [15] it was found that three auxiliary qubits (requiring a total of five qubits including the storage and communication qubits) for each sub-system provide enough physical resources to obtain high fidelity entangled pairs via entanglement pumping.

In order to implement the idea of distributed quantum computation using realistic optical channel and imperfect operations, it is necessary to consider the following questions: What are the minimal local physical resources needed for robust entanglement generation? What is the time overhead associated with entanglement generation? Can we extend the robustness to other imperfections, such as errors from initialization and measurement?

Motivated by these considerations, we study the practical implementation of robust quantum registers for scalable applications. In Ref. [37] we have proposed an entanglement purification scheme that only requires two auxiliary qubits for robust entanglement generation. We have found that the time overhead associated with entanglement generation ranges from a factor of 10 to a few 100, depending on the initial and targeting infidelities. We have also suggested to use one more auxiliary qubit to suppress errors from initialization and measurement. Thus, our hybrid scheme also requires only 5 (or fewer) qubits with local deterministic coupling, while providing additional improvements over the protocol of Ref. [15]: reduced measurement errors, higher fidelity, and more efficient entanglement purification. In this paper, we will provide a detailed discussion on the register-based, hybrid quantum computation scheme presented in Ref. [37].

The paper is organized as follows. In Sec. II we will introduce the concept of quantum register and discuss two experimental implementations. In Sec. III we will review the idea of universal quantum computation based on two-qubit quantum registers. In Sec. IV we will specify the error models for imperfections and provide the basic ideas underlying our robust operations. In Sec. V we will describe the robust measurement/initialization scheme. In Sec. VI we will present our bit-phase two-level entanglement pumping scheme. In Sec. VII we will introduce the Markov chain model to quantitatively analyze the time overhead and residual infidelity associated with the stochastic process of entanglement pumping, and discuss further improvement upon our two-level entanglement pumping scheme. In Sec. VIII we will map our stochastic, hybrid, and distributed quantum computation scheme to a deterministic computation model that is characterized by two quantities (the clock cycle and effective error probability), and estimate the practical values for these quantities. We will also consider the constraint set by the finite memory lifetime and determine the achievable performance of hybrid distributed quantum computation. Finally, in Sec. IX we will discuss using our hybrid scheme for fault-tolerant quantum computation with quantum error correcting codes, and provide a resource-efficient approach for fault-tolerant preparation of the GHZ states.

II. QUANTUM REGISTER AND EXPERIMENTAL IMPLEMENTATIONS

We define a quantum register as a few-qubit device (see Fig. 2(a)) that contains one communication qubit (c), with a photonic interface; one storage qubit (s), with very good coherence times; and several auxiliary qubits (a1, a2, ……), used for purification and error correction. A critical requirement for a quantum register is high-fidelity unitary operations between the qubits within a register.

The quantum registers considered here can be implemented in several physical systems, but in this paper we shall focus on two specific systems where these considerations are particularly relevant. First, ion traps have been used to demonstrate all essential elements of quantum registers. (1) The ion qubits may play the role of communication qubits: they can be initialized and measured efficiently using optical pumping and cycling transitions respectively, and they can also be prepared in highly entangled states with the polarization of single photons [39, 40]. Very recently, entanglement generation between ion qubits from two remote traps has been demonstrated.
which is at the order of \(\mu s\) has been demonstrated in ion traps \([43, 44]\), which is qubits as well. Coherence time of approximately 10 seconds has been demonstrated, allowing gates to be implemented among the qubits in the register. (4) High fidelity operations between the ion qubits within the ion trap has also been demonstrated \([45, 46]\). A second promising candidate for implementing quantum registers is the nitrogen vacancy (NV) centers in diamond. Each NV center can be regarded as an ion trap confined by the diamond crystal, which can be treated as a single register. The qubits for each NV register consists of one electronic spin associated with the defect and several nuclear spins associated with the proximal C-13 nuclei. The electronic spin is optically active, so that it can be measured and initialized optically. With optical cavities or diamond-based photonic crystal micro-cavities \([47]\), one could enhance photon collection efficiency towards unity. Furthermore, the electron spin can be coherently manipulated by microwave pulses \([10]\). The electronic spin is thus suitable as the communication qubit. The nuclear spins are coupled to the electronic spin via hyperfine interactions. One can either use these hyperfine interaction to directly rotate the nuclear spins \([48, 49]\), or apply radio-frequency pulses to address individual nuclear spins spectroscopically \([50]\). These nuclear spins have very long coherence times approaching seconds \([12]\), and can be good storage and auxiliary qubits. Furthermore, the optical manipulation of the electronic spin can be well decoupled from the nuclear spins \([51]\). It can be inferred from the recent experiment \([12]\) that the fidelity of local operations between electronic and nuclear spins is higher than 90%. While the fidelity is still low for the procedures considered here, we believe that it can be significantly improved (to higher than 0.999) by optimal control techniques \([20, 50]\), such as composite pulses \([52]\) and numerically optimized GRAPE pulses \([53]\).

### III. Universal Quantum Computation with Two-Qubit Registers: Fundamentals

We now consider universal quantum computation via the simplest possible two-qubit registers \([11, 30]\). Each register has one qubit for communication and the other qubit for storage. We can use probabilistic approaches from quantum communication \([31]\) and references therein) to generate entanglement between communication qubits from two arbitrary non-local registers. The probabilistic entanglement generation creates a Bell pair conditioned on certain measurement outcomes, which are distinct from outcomes of unsuccessful entanglement generation. If the entanglement generation fails, it can be re-attempted until success, with an exponen-

![Image]

**FIG. 2:** (Color online) Distributed quantum computer. (a) Illustration of distributed quantum computer based on many quantum registers. Each register has five physical qubits, including one communication qubit \((c)\), one storage qubit \((s)\), and three auxiliary qubits \((a_1, a_2, a_3)\). Local operations for qubits within the same register have high fidelity. Entanglement between non-local registers can be generated probabilistically \([3, 31, 32]\). Devices of optical micro-electro-mechanical systems (MEMS) \([35]\) can efficiently route photons and couple arbitrary pair of registers. Detector array can simultaneously generate entanglement for many pairs of registers. (b) An ion trap coupled to a cavity also provides a promising candidate for distributed quantum computation. A single ion is resonantly coupled to the cavity and serves as the communication qubit; while the others can be storage or auxiliary qubits. (c) Nitrogen-vacancy (NV) defect center in photonic crystal micro-cavity. The inset shows the atomic structure of the NV center \([10]\), which forms a quantum register. The electronic spin localized at the cavity is optically accessible (measurement/initialization) and can play the role of the communication qubit. The nuclear spins from proximal \(^{13}\)C atoms constitute the storage and auxiliary qubits, which are coherently controlled via hyperfine interaction and rf pulses.
measurement result for qubit $i$ from quantum registers $[4]$. This task is performed by probabilistic entanglement scheme and two-qubit transfer directly $[21]$, but we are able to achieve the same possibilities of entanglement generation. It is experimentation not quantum computation $[29]$. We emphasize that deterministic entanglement generation between different quantum registers is in principle sufficient for universal quantum computation $[4]$. Similar approaches are also known for deterministic generation of graph states $[30]$ — an essential resource for one-way quantum computation $[29]$.

By consuming one Bell pair, one can implement any non-local controlled-U gate between two storage qubits $[27]$, as shown in Fig. 4. Since operations on a single qubit can be performed within a register, the CNOT operation between the $i$th qubit as control and the $j$th qubit as target, and $m_i = 0, 1$ is the measurement result for qubit $i$ from the circuit in Fig. 3.

IV. ERRORS AND IMPERFECTIONS

In practice, the qubit measurement, initialization, and entanglement generation can be noisy with error probabilities as high as ~ 5%, due to practical limitations such as imperfect cycling transitions, finite collection efficiency, and poor interferometric stability. As a result, there will be a large error probability in non-local gate circuits. In contrast, local unitary operations may fail infrequently ($p_L \lesssim 10^{-4}$) when quantum control techniques for small quantum system are utilized $[19, 20]$. We now show that the most important sources of imperfections, such as imperfect initialization and measurement errors for individual qubits in each quantum register, and entanglement generation errors between registers, can be corrected with a modest increase in register size. We determine that with just three additional auxiliary qubits and high-fidelity local unitary operations, all these errors can be efficiently suppressed by repeated quantum non-demolition (QND) measurement $[54]$ and entanglement purification $[35, 36]$. This provides an extension of Ref. 13 that mostly focused on suppressing errors from entanglement generation.

We will use the following error model for the entire paper: (1) The imperfect local two-qubit operation $U_{ij}$ is

$$U_{ij} = (1 - p_L) U_{ij} + \frac{p_L}{4} \text{Tr}_{ij} [\rho] \otimes I_{ij},$$

where $\text{Tr}_{ij} [\rho]$ is the partial trace over the qubits $i$ and $j$, and $I_{ij}$ is the identity operator for qubits $i$ and $j$. This error model describes that with a probability $1 - p_L$ the gates perform the correct operation and with a probability $p_L$ the gates produce a complete random output for the two involved qubits $[71]$. (2) The imperfect initialization of state $|0\rangle$ will prepare a mixed state

$$\rho_0 = (1 - p_I) |0\rangle \langle 0| + p_I |1\rangle \langle 1|,$$

which has error probability $p_I$, i.e., it prepares the wrong state with a probability $p_I$. (3) The imperfect measurement of state $|0\rangle$ will correspond to the projection oper-
This operator describes that a qubit prepared in state $|0\rangle$ or $|1\rangle$ will give rise to the opposite measurement output with the measurement error probability $p_M$. Finally, the entanglement fidelity for a non-ideal preparation of state $|\Phi^+\rangle$ is defined as

$$F = \langle \Phi^+ | \rho | \Phi^+ \rangle,$$

and the infidelity is just $1 - F$. The fidelity does, however, not completely characterize the produced entangled state. Depending on the exact method used to generate the entangled state, one can in some situations argue that the error will predominantly be, e.g., only a phase error $\delta \Phi$ [3, 32, 55], whereas in other situations it will be a combination of phase and bit flip errors (see [31] and references therein). Below we shall therefore both consider the situation where we only have a dephasing error as well the situation, where we have a more complicated depolarizing error (exact definition given later). As we shall see, the knowledge that the error is of a particular type (e.g. only dephasing error) provides a significant advantage for purification.

We will also assume a separation of error probabilities: any internal, unitary operation within the register fails with extremely low probability, $p_L$, while all operations connecting the communication qubit to the outside world (initialization, measurement, and entanglement generation) fail with error probabilities that can be several orders of magnitude higher.

$$p_L \ll p_I, p_M, 1 - F.$$  \hspace{1cm} (6)

In terms of these quantities the error probability in the non-local CNOT gate in Fig. 5 is

$$p_{CNOT} \sim (1 - F) + 2p_L + 2p_M,$$

because we use one entangled state, two local operations, and two measurements. In the next two sections, we will show how to use robust operations to dramatically improve the fidelity for these non-local coupling gates.

Robust measurement can be implemented by repeated QND measurement, i.e., a majority vote among the measurement outcomes (Fig. 5), following a sequence of CNOT operations between the auxiliary/storage qubit and the communication qubit. This also allows robust initialization by measurement. High-fidelity, robust entanglement generation is achieved via entanglement pumping [13, 33, 36] (Figs. 6 and 7), in which lower fidelity entanglement between the communication qubits is used to purify entanglement between the auxiliary qubits, which can then be used for non-local CNOT operations. To make the most efficient use of physical qubits, we introduce a new entanglement pumping scheme. In our bit-phase two-level entanglement pumping scheme, we first use unpurified Bell pairs to repeatedly pump (purify) against bit-errors (Fig. 7a), and then use the bit-purified Bell pairs to repeatedly pump against phase-errors (Fig. 7b).

Entanglement pumping, like entanglement generation, is probabilistic; however, failures are detected. In computation, where each logical gate should be completed within the allocated time (clock cycle), failed entanglement pumping can lead to gate failure. To demonstrate the feasibility of our approach for quantum computation, we will analyze the time required for robust initialization, measurement and entanglement generation, and show that the failure probability for these procedures can be made sufficiently small with reasonable time overhead.

V. ROBUST MEASUREMENT & INITIALIZATION

In this section, we will analyze the robust measurement scheme based on repeated QND measurement, discuss the recent experimental demonstration of robust measurement in the ion-trap system, and present two approaches to robust initialization.

The measurement circuit shown in Fig. 5 yields the correct result based on a majority vote from $2m + 1$ consecutive readouts (bit-verification). Since the evolution of the system (CNOT gate) commutes with the measured observable ($Z$ operator) of the auxiliary/storage qubit, it is a quantum non-demolition (QND) measurement [54], which can be repeated many times. The error probability for such a majority vote measurement scheme is:

$$\varepsilon_M \approx \sum_{j=m+1}^{2m+1} \left( \frac{2m + 1}{2} \right) (p_I + p_M)^j + \frac{2m + 1}{2}p_L$$  \hspace{1cm} (8)

where the last term conservatively estimates the probability for bit-flip error of the auxiliary/storage qubit during the repeated QND measurement. For simplicity, we will use Eq. (3) for our conservative estimate of error probability for repeated QND measurement. Suppose $p_I = p_M = 5\%$, we can achieve $\varepsilon_M \approx 8 \times 10^{-4}$ by choosing $m^* = 6$ for $p_L = 10^{-4}$, or even $\varepsilon_M \approx 12 \times 10^{-6}$ by choosing $m^* = 10$ for $p_L = 10^{-6}$. For convenience of discussion, we shall add $\varepsilon_M$ to the set of imperfection parameters: $(1 - F, p_I, p_M, p_L, \varepsilon_M)$. The time for robust
measurement is
\[ t_M = (2m + 1) (t_I + t_L + t_M). \] (9)
where \( t_I, t_L, \) and \( t_M \) are times for initialization, local unitary gate, and measurement, respectively.

Measurements with very high fidelity \( (\varepsilon_M \text{ as low as } 6 \times 10^{-4}) \) have recently been demonstrated in the ion-trap system \[56\], using similar ideas as above. There are several possibilities to further improve the performance of the repeated QND measurements. (1) We may use maximum likelihood estimate (MLE) to replace the majority vote for repeated measurements with multi-value outcomes (e.g., fluorescent intensity) \[56\]. (2) We may keep updating the error probability using MLE after each measurement. Once the estimated error probability is below some fixed error rate, we stop the repetition of the QND measurement to avoid errors from redundant operations \[56\]. (3) We may use the implementation of CNOT gate that has small/vanishing bit-flip errors to the control qubit, which will reduce/eliminate the last term in Eq. (9).

The robust measurement scheme also allows to achieve robust initialization by measurement, i.e., by measuring the state of a qubit with the robust measurement scheme, we initialize the qubit into the result of the measurement outcome with an effective initialization error
\[ \varepsilon_I \approx \varepsilon_M. \] (10)
Besides the above measurement-based scheme, we may achieve robust initialization using verification-based scheme \[72\]. For clarity, we shall assume the measurement-based initialization [Eq. (10)] for the rest of the paper.

VI. ROBUST NON-LOCAL TWO-QUBIT GATE

With high-fidelity local unitary gate and repeated QND measurement, the error probability for non-local coupling gates (e.g., Fig. 3 and 4) is
\[ p_{\text{CNOT}} \sim (1 - F) + 2p_L + 2\varepsilon_M, \] (11)
which is dominated by the infidelity of the Bell pair \( 1 - F \), since we assume \( p_L \sim \varepsilon_M \ll 1 - F \). In this section, we will show how to create high-fidelity Bell pairs between two registers with a reasonable time overhead. We will first briefly review entanglement pumping \[31, 35, 36\] with high-fidelity local unitary gate and robust measurement. During the entanglement pumping process, we first store one unpurified Bell pair [Fig. 6 (a,b)], and then generate another unpurified Bell pair to pump (purify) the previously stored pair. (d) If the purification is successful, we obtain a purified Bell pair (level-1 pair) with higher fidelity; otherwise we discard the stored Bell pair and start the entire pumping process from the beginning. (e-f) The second level of entanglement pumping uses previously purified pairs to purify a stored Bell pair, to obtain a Bell pair with higher fidelity (level-2 pair).

A. Entanglement pumping

We now consider entanglement pumping \[31, 35, 36\] with high-fidelity local unitary gate and robust measurement. During the entanglement pumping process, we first store one unpurified Bell pair [Fig. 6 (a,b)], and then generate another unpurified Bell pair to pump (purify) the previously stored pair [Fig. 6 (c,d)]. If the purification is successful, we will obtain a purified Bell pair with higher fidelity, which can be further purified by repeating the process in Fig. 6 (c,d) with new unpurified Bell pairs; otherwise we discard the stored Bell pair, and start the entire pumping process from the beginning. Sometimes, we may want to introduce a second level of entanglement pumping; that is to use previously purified pairs to purify a stored Bell pair [Fig. 6 (e-h)].

B. Fidelity of entanglement pumping

We now analyze the performance of entanglement pumping for different errors of the unpurified Bell pairs. If the unpurified Bell pair is dominated by one type of error (e.g., dephasing error with density matrix \( \rho_{\text{dephasing}} = \text{diag} \{ F, 1 - F, 0, 0 \} \) in the Bell basis \( \{ |\Phi^\pm \rangle, |\Psi^- \rangle, |\Psi^+ \rangle, |\Psi^- \rangle \} \), defined as \( |\Phi^\pm \rangle = (|00\rangle \pm |11\rangle)/\sqrt{2} \), \( |\Psi^\pm \rangle = (|01\rangle \pm |10\rangle)/\sqrt{2} \), we can skip the first level pumping. The unpurified pair then immediately becomes a level-1 pair and is purified with the circuit in Fig. 7b. In Fig. 5, we plot the fidelity curve (purified fidelity v.s. number of successful pumping steps) for the one-level pumping process (i.e. \( n_b = 0 \)), where a very high fidelity pair can be created after \( n_p = 3 \) suc-
successful pumping steps. Note that we consider the full density matrix for all numerical calculation of entanglement fidelities \[\epsilon\text{[57]}\], with the error models given in Eqs. \[\epsilon\text{[23]}\].

If the unpurified Bell pair contains errors from both bit-flip and dephasing processes (e.g., depolarizing error with density matrix \(\rho_{\text{depolarizing}} = \text{diag} [F, 1- F, 1- F, 1- F/4]\) in the Bell basis), two-level entanglement pumping is needed. We introduce the following bit-phase two-level pumping scheme — the first level has \(n_b\) steps of bit-error pumping using raw Bell pairs (Fig. 7a) to produce a bit-error-purified entangled pair, and the second level uses these bit-error-purified pairs for \(n_p\) steps of phase-error pumping (Fig. 7b). In Fig. 8, we plot the fidelity curves for the first level (thin blue curve) and the second level (thick red curve) of entanglement pumping. One-level pumping is insufficient to achieve high fidelity, but two-level pumping can achieve very high fidelity. With the parameters specified for Fig. 8b, the maximum fidelity is achieved via the optimal choice of control parameters \((n_b^*, n_p^*) = (1, 3)\) for successful pumping steps of the first and second levels, respectively. For successful purification, the infidelity of the purified pair, \(\epsilon(\rho, n_b, n_p)\), depends on both the control parameters \((n_b, n_p)\) and the imperfection parameters \((F, p_L, \epsilon_M)\).

For depolarizing error, we find

\[
\epsilon_{\text{max}} = 99.98\%
\]

Therefore, for very small \(p_L\) and \(\epsilon_M\), the new pumping scheme is crucial to minimize the number of qubits per register.

In Fig. 9, we show the contours of the infidelity \(\epsilon_{\text{max}} = 99.97\%\) as a function of \(n_b\) and \(n_p\), where the contours are labeled by values of \(\log_{10} \epsilon_{\text{max}}\). The parameters for the contour plots are \(F = 0.90\) (left) and \(F = 0.95\) (right); \(\epsilon_M = p_L = 10^{-4}\) (up) and \(\epsilon_M = p_L = 10^{-6}\) (down). For optimal choice of \((n_b^*, n_p^*)\), the minimal infidelity is limited by \(\epsilon_M\) and \(p_L\).

For dephasing error, one level pumping is sufficient (i.e. no bit-error purification, \(n_b = 0\)). The infidelity is approximately

\[
\epsilon(\rho, n_p^*, 1) = 99.97\%
\]

by expanding to the leading order in \(p_L\) and \(\epsilon_M\).
FIG. 9: (Color online) The contours of the infidelity \( \epsilon^{(n_b,n_p)} \) as a function of \( n_b \) and \( n_p \) for depolarizing error. We use \( \log_{10} \epsilon^{(n_b,n_p)} \) to label the contours. The other parameters are \( F = 0.90 \) (left) and \( F = 0.95 \) (right); \( \bar{\epsilon}_M = p_L = 10^{-4} \) (up) and \( \bar{\epsilon}_M = p_L = 10^{-6} \) (down). With optimal choice of \((n_b, n_p)\), the minimal infidelity is comparable to the corresponding value of \( p_L \).

VII. MARKOV CHAIN MODEL

The overall success probability can be defined as the joint probability that all successive steps succeed. We use the model of finite-state Markov chain [55] to directly calculate the failure probability of \((n_b,n_p)\)-two-level entanglement pumping using \( N_{\text{tot}} \) raw Bell pairs, denoted as \( \epsilon^{(n_b,n_p)}_{\text{fail}}(N_{\text{tot}}) \).

A. Markov chain model for entanglement pumping

We first use the Markov chain model to study the \( n \)-step one-level entanglement pumping. As shown in Fig. 10, we use "0" to denote the initial state with no Bell pairs, "1" for the state with one stored unpurified pair, \((j + 1)\) for the state with one purified pair surviving \( j \) steps of pumping, and "\( \star \)" for the final state with the purified pair surviving \( n \) steps of pumping. The (success) transition probability from state \( j \) to state \( j + 1 \) is \( q_j \), while the (failure) transition probability from state \( j \) to state 0 is \( 1 - q_j \), for \( j = 0, 1, \ldots, n \). Here \( q_0 \equiv 1 \) (corresponding to deterministic state transfer as shown in Fig. 9 (a,b)) and \( q_{j \geq 1} \) can be calculated with the density matrix of the purified Bell pair surviving \( j \) steps of pumping [55]. The final state is self-trapped, and goes back to itself with unit probability. Each transition attempt consumes one unpurified Bell pair.

FIG. 10: (Color online) Markov chain model for one-level entanglement pumping. We use "0" to denote the initial state with no Bell pairs, "1" for the state with one stored unpurified pair, \((j + 1)\) for the state with one purified pair surviving \( j \) steps of pumping, and "\( \star \)" for the final state with the purified pair surviving \( n \) steps of pumping [55]. The final state is self-trapped, and goes back to itself with unit probability. Each transition attempt consumes one unpurified Bell pair.

The overall success probability can be defined as the probability vector evolves from \( \vec{P}_t \) to \( \vec{P}_{t+1} \) according to the rule

\[ \vec{P}_{t+1} = M \vec{P}_t, \]

with the transition matrix

\[ M = \begin{pmatrix}
0 & 1 - q_1 & 1 - q_2 & \cdots & 1 - q_n & 0 \\
1 & 0 & 0 & \cdots & 0 & q_1 \\
0 & q_1 & 0 & \cdots & 0 & q_2 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & q_n & 1 \\
q_n & 0 & \cdots & 0 & 1
\end{pmatrix}. \]

Since the initial probability vector is \( \vec{P}_0 = (1, 0, \cdots, 0)^T \), we can calculate the probability vector after \( N_{\text{tot}} \) attempts

\[ \vec{P}_{N_{\text{tot}}} = M^{N_{\text{tot}}} \vec{P}_0. \]
The probability vector $\vec{P}$ describes the entire probability distribution over all states of the Markov chain. The last element of $\vec{P}_{N_{tot}}$ is the success probability of reaching the final state "0" after $N_{tot}$ attempts; the failure probability after $N_{tot}$ attempts is then

$$\varepsilon^{(n_b,n_p)}_{E,\text{fail}}(N_{tot}) = 1 - (P_{N_{tot}})_{n+2}. \quad (17)$$

For two-level entanglement pumping, the state transition diagram is shown in Fig. 11. $n_b$ and $n_p$ are the number of pumping steps used to purify bit-errors and phase-errors, respectively. As detailed in Appendix B, we may use a (column) vector $\vec{P}$ with $(n_b+1)(n_p+1) + 1$ elements to characterize the probability distribution among all $(n_b+1)(n_p+1) + 1$ states. From the $t$-th attempt to the $(t+1)$-th attempt, the probability vector evolves from $\vec{P}(t)$ to $\vec{P}(t+1)$ according to the same rule as above [Eq. (14)], but with the transition matrix $M$ given in Eq. (15).

Similar to one-level pumping, we can calculate the probability vector after $N_{tot}$ attempts using Eq. (16). The probability vector $\vec{P}_{N_{tot}}$ describes the entire probability distribution over all states of the Markov chain.

FIG. 11: (Color online) Markov chain model for two-level entanglement pumping. The required pumping steps are $n_b$ and $n_p$ for the two levels, respectively. We use ”0, 0” to denote the initial state with no Bell pairs, ”0, j + 1” for the state with one purified pair surviving $k$ steps of pumping at the first level, ”k + 1, j + 1” for the state with one purified pair surviving $k$ steps of pumping at the second level and one purified pair surviving $j$ steps of pumping at the first level, ”k + 1, *” for the state with one purified pair surviving $k$ steps of pumping at the second level and one purified pair surviving $n_b$ steps of pumping at the first level, and ”*, 0” for the final state with one purified pair surviving $n_p$ steps of pumping at the second level. For the first level pumping, the (success) transition probability from state ”k, j” to state ”k, j + 1” is $q_j$, while the (failure) transition probability from state ”k, j” to state ”k, 0” is $1 - q_j$, for $j = 0, 1, \ldots, n_b$. For the second level pumping, the (success) transition probability from state ”k, *” to state ”k + 1, *” is $Q_k$, while the (failure) transition probability from state ”k, *” to state ”0, 0” is $1 - Q_k$, for $k = 0, 1, \ldots, n_p$. The final state is self-trapped, and goes back to itself with unit probability.

The last element of $\vec{P}_{N_{tot}}$ is the success probability of reaching the final state "0" after $N_{tot}$ attempts; the failure probability after $N_{tot}$ attempts is then

$$\varepsilon^{(n_b,n_p)}_{E,\text{fail}}(N_{tot}) = 1 - (P_{N_{tot}})_{n+2}. \quad (18)$$

In Fig. 12 we plot the failure probability $\varepsilon^{(n_b,n_p)}_{E,\text{fail}}(N_{tot})$ v.s. $N_{tot}$ for control parameters $(n_b, n_p) = (2, 3)$ and $(3, 4)$. For $N_{tot}$ sufficiently large, the failure probability decreases exponentially to zero. For any given parameters, we can efficiently suppress the failure probability with some reasonably large $N_{tot}$.

B. Total error probability & average infidelity

We now introduce the total error probability (TEP) approximated by the sum of the failure probability and the infidelity of the purified Bell pair

$$\varepsilon^{(n_b,n_p)}_{E}(N_{tot}) \approx \varepsilon^{(n_b,n_p)}_{E,\text{fail}}(N_{tot}) + \varepsilon^{(n_b,n_p)}_{E,\text{infid}}. \quad (19)$$

This is a very conservative estimate, since sometimes we do create some partially purified Bell pair though not the targeted purified Bell pair. And here we just say that the state has fidelity zero in these cases.

To consider the possibility of using a partially purified Bell pair for output, we may introduce another useful quantity—the average infidelity (AIF)—for output Bell pair from the robust entanglement generation, where we take into account these partially purified pairs. The average infidelity of the output pair is the weighted average of the infidelity of the Markov chain

$$\delta^{(n_b,n_p)}_{E}(N_{tot}) \equiv 1 - \langle \delta^{(n_b,n_p)}_{E}(t) \rangle$$

$$= \sum_{n_b=0}^{n_b} \sum_{n_p=0}^{n_p} \varepsilon^{(n_b,n_p)}_{E,\text{infid}} P(N_{tot})(n_b+1)(n_p+1)+2 + \frac{1}{2} P(N_{tot})_1. \quad (20)$$

FIG. 12: (Color online) Failure probability $\varepsilon_{E,\text{fail}}$ as a function of $N_{tot}$. We assume a depolarizing error with $F = 0.95$, $p_L = \varepsilon_M = 10^{-4}$. We choose $(n_b, n_p) = (2, 3)$ for the lower curve, and $(3, 4)$ for the upper curve. For large $N_{tot}$, the failure probability $\varepsilon_{E,\text{fail}}$ decreases exponentially with $N_{tot}$. We observe a rapid decrease of the failure probability for $N_{tot}$.
The optimized total error probability $\varepsilon_E$ [Eq. (21)] (upper plots) and the optimized average infidelity $\delta_E$ [Eq. (22)] (lower plots) as a function $N_{\text{tot}}$. The error probability for the local coupling gates is $p_L = 10^{-4}$ (left plots) and $p_L = 10^{-6}$ (right plots). One-level pumping is used for dephasing error (thin blue curves); two-level pumping is used for depolarizing error (thick red curves). The other parameters $F = 0.95, p_l = p_M = 5\%$ are the same for all plots. Both $\varepsilon_E$ and $\delta_E$ saturate for large $N_{\text{tot}}$. In each plot, we also show the initial infidelity $1-F$ (upper blue dashed lines) and the local error probability $p_L$ (lower violet dashed lines).

Here the first term sums over all states of the Markov chain (except for the initial one), each of which has at least one partially purified pair with infidelity $\varepsilon_{E,\text{infl}}$ and probability $P(N_{\text{tot}}) \binom{n_b+1}{n_b}n_b^* n_p^* \binom{n_b}{n_b} n_b + 2$; the last term comes from the situation that none of the partially purified Bell pairs remain after the last attempt of the entanglement purification and we just use a classically correlated pair with infidelity $1/2$. Generally, the average infidelity is smaller than the total error probability.

We may also optimize the choice of the control parameters $(n_b, n_p)$

\[
\varepsilon_E(N_{\text{tot}}) \equiv \min_{(n_b, n_p)} \varepsilon_E(n_b, n_p) (N_{\text{tot}}) \quad (21)
\]

and

\[
\delta_E(N_{\text{tot}}) \equiv \min_{(n_b, n_p)} \delta_E(n_b, n_p) (N_{\text{tot}}). \quad (22)
\]

In Fig. 13 we plot both the optimized total error probability $\varepsilon_E$ and the optimized average infidelity $\delta_E$ as a function of $N_{\text{tot}}$. Both quantities asymptotically approach the same minimum value

\[
\lim_{N_{\text{tot}} \to \infty} \varepsilon_E(N_{\text{tot}}) = \lim_{N_{\text{tot}} \to \infty} \delta_E(N_{\text{tot}}) = \Delta_{\text{min}}. \quad (23)
\]

Here the minimum value is simply the minimal infidelity of the entanglement purification

\[
\Delta_{\text{min}} \equiv \min_{(n_b, n_p)} \varepsilon_E(n_b, n_p), \quad (24)
\]

which is achieved by the control parameters $(n_b, n_p) \equiv \binom{n_b^*, n_p^*}{n_b, n_p}$, for the imperfection parameters $(p_L, 1-F, \varepsilon_M)$.

We remark that a faster and less resource intensive approach may be used if the unpurified Bell pair is dominated by dephasing error. Then, one-level pumping is sufficient (i.e. no bit-error purification, $n_b = 0$). The optimized total error probability and average infidelity (thin blue curves) for this situation are plotted as a function of $N_{\text{tot}}$ in Fig. 13.

\section*{C. Total time for robust entanglement generation}

The total time for robust entanglement generation $\tilde{t}_E$ is proportional to the average number of raw Bell pairs generated $\langle N_{\text{tot}} \rangle$

\[
\tilde{t}_E \approx \langle N_{\text{tot}} \rangle \times (t_E + t_L + \tilde{t}_M), \quad (25)
\]

where $t_E$ is the average generation time of the unpurified Bell pair. Note that the entanglement generation itself is a stochastic process. In principle, we may also include the stochastic nature of the entanglement generation by introducing a sub-level of Markov chain to characterize the stochastic entanglement generation. Since each entanglement generation either succeeds or fails, the sub-level Markov chain only involves two states, which can be easily incorporated into the Markov chain models discussed above. After incorporating the sub-level into the Markov chain, each transition corresponds to one attempt of entanglement generation, instead of one attempt of entanglement purification that consumes one unpurified Bell pair previously.

Nevertheless, the number of Bell pairs generated in a given period of time (i.e. $N_{\text{tot}}$) has a distribution. Since the relative deviation of this distribution ($\sim N_{\text{tot}}^{-1/2}$) is fairly small for large $N_{\text{tot}}$ (> 20), this only has a minor influence. Thus we replace $\langle N_{\text{tot}} \rangle$ by $N_{\text{tot}}$.

\section*{D. Trade-off between gate quality and time overhead}

We now consider the balance between the "quality" of the robustly generated entangled pairs and the time overhead $N_{\text{tot}}$ associated with the robust generation process. We may use either the optimized total error probability $\varepsilon_E(N_{\text{tot}})$ or the optimized average infidelity $\delta_E(N_{\text{tot}})$ to characterize the quality. Since both quantities approach the same asymptotic minimum $\Delta_{\text{min}}$, according to Eq. (23), there is only little improvement in the quality of the robust entanglement generation once $\varepsilon_E(N_{\text{tot}})$ or $\delta_E(N_{\text{tot}})$ is comparable to $\Delta_{\text{min}}$ (say 2$\Delta_{\text{min}}$). Thus, we find the value for $N_{\text{tot}}$ by imposing the relations

\[
\varepsilon_E(N_{\text{tot}}) = 2\Delta_{\text{min}} \quad (26)
\]
or
\[
\delta_E(N_{\text{tot}}) = 2\Delta_{\text{min}}. \tag{27}
\]

First, we consider the total error probability \(\varepsilon_E(N_{\text{tot}})\). The relation in Eq. (26) can be simplified, if we assume fixed control parameters \((n_b, n_p)\) for the left hand side (rather than minimizing over all possible choices of \((n_b, n_p)\)). Combined with Eq. (19), the failure probability should be comparable to the minimal infidelity \(\varepsilon_{E_{\text{fail}}}(N_{\text{tot}}) \approx \Delta_{\text{min}}. \tag{28}\)

Since both the variable \(\Delta_{\text{min}}\) and the parameters \((n_b^*, n_p^*)\) depend on \(\{p_L, p_I, p_M, 1 - F\}\), the above relation implicitly determines \(N_{\text{tot}}\) as a function a function of \(\{p_L, p_I, p_M, 1 - F\}\).

In Fig. 14 we plot the contours of \(\varepsilon_E\) [Eq. (26)] and \(N_{\text{tot}}\) [Eq. (28)] with respect to the imperfection parameters \(p_L\) and \(1 - F\), while assuming \(p_I = p_M = 5\%\). Actually the choice of \(p_I\) and \(p_M\) (< 10\%) has negligible effect on the contours, since they only modify \(\varepsilon_M\) marginally. For initial fidelity \(F_0 > 0.95\), the contours of \(\varepsilon_E\) are very close to vertical lines; that is \(\varepsilon_E\) is mostly limited by \(p_L\) with an overhead factor (about 10) very insensitive to \(F_0\). The contours of \(N_{\text{tot}}\) indicate that the entanglement pumping needs about tens or hundreds of raw Bell pairs to ensure a very high success probability.

Similarly, we may also numerically obtain the value \(N_{\text{tot}}\) from Eq. (27). The contour plot of \(N_{\text{tot}}\) with respect to the imperfection parameters \(p_L\) and \(1 - F\) is also shown in Fig. 14 (c,f). We compare \(N_{\text{tot}}\)s obtained from two estimates (total error probability [Eq. (26)] and average infidelity [Eq. (27)]). As we expected, the \(N_{\text{tot}}\) obtained from total error probability is approximately 1.2 ∼ 2 time larger than the \(N_{\text{tot}}\) obtained from average infidelity, since the former is a more conservative estimate and requires more unpurified Bell pairs. Nevertheless, the difference is small and can be easily accounted by a prefactor of order unity. For clarity, in the rest of the paper we will use the \(N_{\text{tot}}\) estimated by using total error probability, and sometimes quote the values estimated by using average infidelity.

E. Entanglement pumping with non-post-selective (NPS) scheme

We now consider another entanglement pumping protocol, proposed by Campbell [59]. The entanglement pumping scheme we have considered so far is post-selective (PS); that is we discard the Bell pair if one step of entanglement pumping is not successful. However, the Bell pair may still be highly entangled even if the entanglement pumping failed at some intermediate step. The non-post-selective (NPS) entanglement pumping scheme

![FIG. 14: (Color online) Contours of the total error probability \(\varepsilon_E\) (or average infidelity \(\delta_E\)) after purification (left), the total number of unpurified Bell pairs \(N_{\text{tot}}\) associated with \(\varepsilon_E\) [Eq. (26)] (middle), and \(N_{\text{tot}}\) associated with \(\delta_E\) [Eq. (27)] (right). The contours are drawn with respect to the imperfection parameters \(p_L\) (horizontal axis) and \(F\) (vertical axis). Two-level pumping (up) is used for depolarizing error, and one-level pumping (down) for dephasing error. \(p_I = p_M = 5\%\) is assumed.](image-url)
FIG. 15: (Color online) Markov chain model for one-level entanglement pumping with non-post-selective (NPS) scheme. The key difference from the previous Markov chain model with post-selective (PS) pumping scheme (see Fig. 10) is that here the transition for unsuccessful pumping reduces the chain label (score) by 1, while in the previous model the transition for unsuccessful pumping goes back to state ”0” (restart of the entire pumping scheme). [59] keeps track of the evolution of the density matrix of the Bell pair after each step of pumping. The NPS scheme avoids the inefficient restart (i.e., discarding intermediately purified Bell pairs), and it may reduce the time overhead, especially when the unpurified Bell pairs have relatively low fidelity ($F < 0.9$). In Ref. [59], the NPS pumping is discussed in the context of generating a graph state.

We now describe how to use the NPS pumping scheme to generate purified Bell pairs. To simplify the discussion, we first assume that the errors from local measurements and operations are negligible. This assumption enables us to establish a connection between the Markov chain model and the NPS pumping scheme.

Suppose the unpurified Bell pairs have only phase errors, then one level of entanglement pumping is sufficient. For this error model, one can show that a failed attempt produces an EPR pair with a density matrix identical to the one in the previous step [59]. One may introduce an accumulated score associated with entanglement pumping. The score increases by one unit for each attempt of successful pumping, and decrease by one unit for an attempt of unsuccessful pumping. The score for no Bell pair is 0, and for one unpurified Bell pair it is 1. The score exactly corresponds to the state label of the Markov chain (see Fig. 15). After each attempt of pumping, the score changes by ±1. If the score drops to 0 (i.e., no Bell pair left), it gets back to 1 in the next attempt (i.e., creating a new unpurified Bell pair). The pumping procedure continues, until the score reaches $n + 1$ (i.e., the final state ”$n$” in the Markov chain). The key different from the previous Markov chain for post-selective pumping scheme (see Fig. 10) is that here the score decrease by 1 for unsuccessful pumping rather than restart from 0. This modification increases the success probability of the robust entanglement generation.

FIG. 16: (Color online) Markov chain model for two-level entanglement pumping with NPS scheme. We still use PS entanglement pumping scheme at the first level to have minimal accumulation of phase errors. Only at the second level, does the NPS scheme work more efficiently than the original PS scheme (see Fig. 11).

When the unpurified Bell pairs have both bit-flip and phase errors (e.g., depolarizing error), we may use the bit-phase two-level pumping scheme (see Sec. VI B), which purifies the bit error at the first level and then the phase error at the second level. Since the phase error is not purified at the first level, it accumulates after each attempt of pumping. Therefore, it is better to use PS entanglement pumping scheme at the first level to have minimal accumulation of phase errors. At the second level, the NPS scheme works more efficiently than the PS scheme. The Markov chain circuit for such mixed PS-NPS pumping schemes is shown in Fig. 16.

In practice, the error probability for the local operations is always finite. Then our simple Markov chain model only provides an approximate description for the real process. The approximation comes from the fact that the score is now insufficient to specify the density matrix for intermediate Bell pairs, in the presence of local operational errors. In order to obtain the density matrix for the intermediate state, we need to have the entire list of all previous pumping outcomes. Nevertheless, when the local operational errors are small compared to the infidelity of the intermediate Bell pairs, the Markov chain model still provides an (optimistic) estimate for the total error probability and the average fidelity.

We now compare the $N_{\text{tot}}$ associated with the PS and NPS schemes. The contours of the ratio between the two $N_{\text{tot}}$ is plotted as a function of $p_L$ and $F$ in Fig. 17. As pointed out in Ref. [59], there is a significant improvement by using the NPS scheme (more than a factor of 3), for $F < 0.9$ and $p_L < 10^{-4}$. 


with a photon collection/detection efficiency $\eta$, vacuum radiative lifetime $\tau$, and the cooperativity (Purcell) factor $C$ for cavity-enhanced radiative decay [60, 61]. Eq. (31) is obtained from the estimate for the measurement error probability $p_M \approx (1 - \eta)N_{\text{photon}}$ with $N_{\text{photon}} \approx t_M/ (\tau/C)$. We assume that the entanglement is generated based on detection of two photons [3, 32], which takes time

$$t_E \approx (t_I + \tau/C)/\eta^2.$$  \hfill (32)

Generally entanglement fidelity is higher for the two-photon schemes than one-photon schemes [31]. In addition, some two-photon schemes have intrinsic purification against bit-flip errors [55]. The time for robust measurement is given in Eq. (9), and the total time for robust entanglement generation is given in Eq. (25).

Combining Eqs. (29), (31), (32), (9) and (25), we obtain the clock cycle time (in units of the local operation time) as a function of other parameters

$$\frac{t_C}{t_L} = f\left[\frac{\tau}{t_L C}, p_m, \eta, m, N_{\text{tot}}\right].$$  \hfill (33)

Meanwhile, we may obtain the relation $m = m [p_L, p_I, p_M]$ by minimizing $\varepsilon_M$ with Eq. (8), and find the relation $N_{\text{tot}} = N_{\text{tot}} [p_L, F, \varepsilon_M] = N_{\text{tot}} [p_L, F, 2\Delta_{\text{min}} [p_L, p_I, p_M]]$ using Eqs. (24) and (26). Therefore, we have

$$\frac{t_C}{t_L} = f\left[\frac{\tau}{t_L C}, p_M, \eta, p_L, F\right].$$  \hfill (34)

The dimensionless parameter is the ratio between the times of emitting a single photon and performing a local unitary operation. For systems such as ion-traps and NV centers, this ratio is usually much less than unity ($< 0.01$).

Similarly, we can obtain the effective error probability in terms of imperfection parameters

$$\gamma = g [p_L, p_M, F],$$  \hfill (35)

by combining Eqs. (8), (9) and (26).

In Fig. 18, we plot the clock cycle time $t_C$ and effective error probability $\gamma$, for two-level pumping against depolarizing error. Assuming $\eta = 0.2$, we consider the two choices of parameters $1 - F = p_I = p_M = 5\%$ (left) and $1\%$ (right). For each case, we plot the contours of the normalized clock cycle time $t_C/t_L$ as a function of $p_L$ and $1/t_C$, and the effective error probability $\gamma$ as a function of $p_L$. The clock cycle time can be reduced by having a fast radiative decay rate $\tau/C$, which can be facilitated by having a large cooperativity factor $C$. The reduction of the clock cycle time stops once this ratio is below certain value, approximately 0.003 (left) and 0.001 (right), where local gate operation becomes the dominant time consuming step. Similarly, we plot the clock cycle time $t_C$ and effective error probability $\gamma$, for one-level pumping against dephasing error in Fig. 19.
The error probability ($p$) and/or the error probability ($p$) for two-level pumping against depolarizing errors. Upper plots: contours of the normalized clock cycle time ($t_C/t_L$) as a function of $p_L$ and $t_C$ (the normalized effective radiative lifetime). Lower plots: the effective error probability $\gamma$ as a function of $p_L$. We assume $1 - F = p_I = p_M = 5\%$ (left) and 1% (right), and $\eta = 0.2$.

In the limit of negligible radiative decay time, we obtain the lower bound for the normalized clock cycle time

$$
\lim_{\tau \to \infty} \frac{t_C}{t_L} = \begin{cases} 
N_{\text{tot}} & \text{for } m = 0 \\
(2m + 2)N_{\text{tot}} & \text{for } m \geq 1
\end{cases},
$$

where for $m \geq 1$ there is a time overhead $2m + 2$ associated with local operation and robust measurement; while there is no such overhead for $m = 0$.

### B. Estimated numbers for experimental setups

Suppose the parameters are $(t_L, \tau, \eta, C) = (0.1\ \mu s, 10\ ns, 0.2, 10)$ and $(1 - F, p_I, p_M, p_L, \epsilon_M) = (5\%, 5\%, 5\%, 10^{-4}, 8 \times 10^{-4})$ for our quantum registers (based on ion-traps or NV centers). For depolarizing errors, two-level pumping can achieve $(t_C, \gamma) = (200\ \mu s, 2.7 \times 10^{-3})$. For some entanglement generation schemes in principle only dephasing error exists, because they have intrinsic purification against bit-flip errors. If all bit-flip errors are suppressed, then one-level pumping is sufficient and $(t_C, \gamma) = (42\ \mu s, 2.2 \times 10^{-3})$. In Table 1 we have listed $(t_C, \gamma)$ for parameters $1 - F = p_M = p_I = 5\%$ or 1%, and $p_L = 10^{-3}, 10^{-4}, 10^{-5}$, or $10^{-6}$. As expected, we find that $t_C$ gets longer, if the fidelity $F$ is lower and/or the error probability ($p_M$ or $p_I$) is higher: $t_C$ significantly reduces if the error for the unpurified Bell pairs changes from depolarizing error to dephasing error.

We remark that $t_C$ should be much shorter than the memory time of the storage qubit, $t_{\text{mem}}$. Because the memory error probability for each clock cycle is approximately $t_C/t_{\text{mem}}$, which should be small (say $10^{-4}$) in order to achieve fault-tolerant quantum computation. This is indeed the case for both trapped ions (where $t_{\text{mem}} \sim 10$ s has been demonstrated [43, 44]), as well as for proximal nuclear spins of NV centers (where $t_{\text{mem}}$ approaching a second can be inferred [12]). So far, we have justified the feasibility of the hybrid distributed quantum computation scheme. In the next subsection, we will provide a criterion for hybrid distributed quantum computation.

### C. Constraints from finite memory life time

Above we have mostly ignored the effect of finite memory time, and with the various sequences of purification of imperfections the final fidelity of the operations have then been limited only be the local operation. All of these purifications, however, increase the time of the operations and eventually the system may become limited by the finite life time of the memory. In this subsection we shall evaluate this constraint set by the finite memory lifetime.

To simplify the discussion we assume that we have a very short radiative lifetime $\tau$ or that we are able to achieve a very large Purcell factor so that $\gamma/C$ becomes negligible. All the time scales are then proportional to the local gate time $t_L$. With a finite memory time, i.e., some fixed $t_{\text{mem}}/t_L$, there is a limit to have many operations we can do before we are limited by the memory error. To get an estimate for this limit we assume that the ideal number of operations is roughly given by
point, where the memory error probability is the same as the effective error probability for the non-local coupling gate:

\[ t_C / t_{\text{mem}} = \gamma. \]  

(37)

Then according to Eq. (36), we have

\[ \frac{t_{\text{mem}}}{t_L} = \frac{t_{\text{mem}}}{t_C} \frac{t_C}{t_L} = \gamma^{-1} (2m + 2 - \delta_{m,0}) N_{\text{tot}}, \]  

(38)

where the variables \( \{\gamma, m, N_{\text{tot}}\} \) are all determined by the imperfection parameters \( \{1 - F, p_M, p_I, p_L\} \). We further reduce the imperfection parameters by assuming \( 1 - F = p_M = p_I, \) and get the contour plot of \( t_{\text{mem}} / t_L \) in terms of the imperfection parameter \( p_L \) and \( 1 - F \) in Fig. 20. In the plot, we consider both the situation of depolarizing or dephasing error during entanglement generation.

For given \( t_{\text{mem}} / t_L \), we may use Fig. 20 to find the valid region in the parameter space of \( p_L \) and \( 1 - F, \) and then identify the achievable effective error probability \( \gamma \). For example, with ion-trap systems it may be possible to achieve \( t_{\text{mem}} / t_L \sim 10^6 \) \cite{43, 62}, and the region left of the shaded contour line (\( \log_{10} t_{\text{mem}} / t_L = 8 \)) can then be accessed, which enables us to obtain a wide range effective error probability \( \gamma \) depending on the practical values of \( p_L \) and \( 1 - F. \) For NV centers, it should be feasible to achieve \( t_{\text{mem}} / t_L \sim 10^7 \) by having \( t_{\text{mem}} \sim 10^6 \) s and \( t_L \sim 10^{-6} \) s \cite{12}; the region on the left side of the shaded contour line (\( \log_{10} t_{\text{mem}} / t_L = 7 \)) still covers a large portion of the parameter space. For a given experiment situation with a finite memory time as well as other imperfections, we can thus use Fig. 20 to determine the achievable performance of hybrid distributed quantum computation.

### IX. APPROACHES TO FAULT TOLERANCE

The entanglement based approach discussed in this paper provides a method to make gates between any quantum registers and this can be used to implement arbitrary quantum circuits, once the errors in the gates are sufficiently small. The errors can be further suppressed by using quantum error correction. For example, as shown in Table 1, \( (p_L, F) = (10^{-4}, 0.95) \) can achieve \( \gamma \approx 2.7 \times 10^{-3}, \) well below the 1% threshold for fault tolerant computation based on approaches such as the \( C_4 / C_6 \) code \cite{65} or 2D toric codes \cite{66}; \( (p_L, F) = (10^{-6}, 0.99) \) can achieve \( \gamma \approx 1.7 \times 10^{-5}, \) which allows efficient codes such as the \( \text{BCH (}[127, 43, 13]\) code to be used without concatenation.

Following Ref. \cite{67} we estimate 20 registers per logical qubit to be necessary for a calculation involving \( K = 10^4 \) logical qubits and \( Q = 10^6 \) logical operations, assuming the memory failure rate and effective error probability are \( t_C / t_{\text{mem}} \approx \gamma \approx 1.7 \times 10^{-5} \) (e.g., achieved by \( t_{\text{mem}} \approx 10^6 \) s, \( t_C \approx 16 \) ms). This estimate is based on Fig. 10b of Ref. \cite{67}. Assuming that error correction is applied after each logical operation, and that logical operation and following recovery take approximately 2 – 16 clock cycles depending on the type of operation and the coding scheme (see section II.A of Ref. \cite{67}), the total running time of this computation would then be approximately 400 – 3000 s.

We remark that one important property of distributed quantum computation is that the measurement time is relatively fast compared with the non-local coupling gate, because the measurement does not rely on the time-consuming processes of entanglement generation and purification while the non-local coupling gate does. This property is different from the conventional model of quantum computation, where the measurement is usually a slow process that induces extra overhead in both time and physical resources \cite{67}. Thus, instead of reconciling slow measurements \cite{65}, it might also be interesting to study possible improvements using fast measurements for fault-tolerant quantum computation.

The above estimates have been performed assuming that our hybrid register based approach is mapped directly to the standard circuit model. In some situations this may, however, not be the most advantageous way to proceed, since the register architecture may allow for more efficient performance of certain tasks. As a particular example, we now briefly discuss an alternative new approach to fault-tolerant preparation of GHZ states (e.g., \([00 \cdots 0]+[11 \cdots 1]\)), which are a critical component both for syndrome extraction and construction of universal gates in quantum error correcting codes \cite{65, 69}. This

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
 & \multicolumn{2}{c|}{Depolarizing} & \multicolumn{2}{c|}{Dephasing} \\
 & \( F = 0.95 \) & \( F = 0.99 \) & \( F = 0.95 \) & \( F = 0.99 \) \\
\hline
\( p_L = 10^{-3} \) & 65 & 1.9 \times 10^{-2} & 19 & 9.1 \times 10^{-3} & 20 & 1.7 \times 10^{-2} & 8 & 8.7 \times 10^{-3} \\
\hline
\( p_L = 10^{-4} \) & 200 & 2.7 \times 10^{-3} & 49 & 1.2 \times 10^{-3} & 42 & 2.2 \times 10^{-3} & 17 & 9.9 \times 10^{-4} \\
\hline
\( p_L = 10^{-5} \) & 387 & 3.5 \times 10^{-4} & 65 & 1.3 \times 10^{-4} & 80 & 2.8 \times 10^{-4} & 22 & 1.2 \times 10^{-4} \\
\hline
\( p_L = 10^{-6} \) & 997 & 4.5 \times 10^{-5} & 162 & 1.7 \times 10^{-5} & 140 & 3.4 \times 10^{-5} & 39 & 1.4 \times 10^{-5} \\
\hline
\end{tabular}
\caption{We list the values of \( t_C \) and \( \gamma \) as a function of \( p_L \) (rows) and \( F \) (columns) for depolarizing and dephasing errors of the unpurified Bell pairs. We also assume \( p_M = p_I = 1 - F, \) and \( (t_L, \tau, \eta, C) = (0.1 \\mu s, 10 \text{ ns}, 0.2, 10). \) Note that \( t_C \) estimated by using average infidelity is approximately 1.3 – 1.6 times less than the numbers listed here.}
\end{table}
new approach relies upon the observation that the EPR pairs from entanglement generation can be used for deterministic partial Bell measurement (PBM), which is achieved by applying local coupling gates and projective measurements as shown in Fig. 21. After the PBM, the two storage qubits are projected to the subspace spanned by the Bell states $|\Phi^\pm\rangle = (|00\rangle \pm |11\rangle)/\sqrt{2}$ if the measurement outcomes are the same, or they are projected to the subspace spanned by the Bell states $|\Psi^\pm\rangle = (|01\rangle \pm |10\rangle)/\sqrt{2}$ if the measurement outcomes are different. For the latter case, we may further flip one of the storage qubits, so that they are projected to the subspace spanned by $|\Phi^\pm\rangle$. By using PBMs, we can perform fault-tolerant preparation of GHZ state efficiently (up to single qubit rotations) as detailed below.

Fault-tolerant state preparation requires that the probability to have errors in more than one qubit in the prepared state is $O(p^2)$, with the error probability for each input qubit or quantum gate being $O(p)$; that is multiple errors only occur at the second or higher orders. The regular circuit to prepare a four-qubit GHZ state fault-tolerantly [65] is shown in Fig. 21a. If this circuit should be implemented with quantum registers, the CNOT gates in Fig. 21a should be created by using the circuit detailed in Fig. 3, and eight quantum registers would be required. By using PBMs, however, only four quantum registers are needed in order to generate GHZ states fault-tolerantly as shown in Fig. 21b. The fault-tolerance comes from the last (redundant) PBM between the second and fourth register (Fig. 21b), which detects bit-errors from earlier PBMs. The advantage of the PBMs is that it propagates neither bit- nor phase-errors. The circuit of Fig. 21c indicates that the only way to propagate error from one input to another (say, $S_1$ to $S_2$) is via some initial error in the entangled pair between $C_1$ and $C_2$. However, for Bell states $|\Phi^\pm\rangle$, we have the following identities

$$X_{C_2} |\Phi^\pm\rangle_{C_1,C_2} = \pm X_{C_1} |\Phi^\pm\rangle_{C_1,C_2} \quad (39)$$

$$Z_{C_2} |\Phi^\pm\rangle_{C_1,C_2} = Z_{C_1} |\Phi^\pm\rangle_{C_1,C_2}. \quad (40)$$

Similar identities also exist for Bells states $|\Psi^\pm\rangle$. Suppose $S_1$ has an error, because of the above identities, we can always treat the imperfection of the Bell pair $|\Phi^\pm\rangle_{C_1,C_2}$ as an error in $C_1$ (the qubit from the same register as $S_1$). Therefore, only the first register has errors and they
FIG. 22: Circuit for fault-tolerant preparation of the 8-qubit GHZ state with only 8 quantum registers.

never propagate to the second one. (PBM may induce errors to unconnected but entangled qubits.)

The present scheme may be expanded to larger numbers of qubits and generally we may fault-tolerantly prepare $2^n$-qubit GHZ state with only $2^n$ quantum registers, by recursively using Fig. 21b with the two dashed boxes replaced by two $2^{n-1}$-qubit GHZ states. The circuit for fault-tolerant preparation of the 8-qubit GHZ state is shown in Fig. 22. Note that we can perform PBMs acting on different registers in parallel. Suppose each PBM takes one clock cycle, the preparation time is 2 (clock cycles) for a 4-qubit GHZ state shown in Fig. 21b. The two PBMs in the orange boxes are performed in the first clock cycle, and the rest for the second clock cycle. Generally, for a $2^n$-qubit GHZ state with $n \geq 3$ (see discussion in Appendix C), the preparation time is only 3 (clock cycles), and the error probability for each register is only approximately $3p/2$. Therefore, the PBM-based scheme for fault-tolerant preparation of the GHZ state is efficient in both time and physical-resources.

X. CONCLUSION

In conclusion, we have proposed an efficient register-based, hybrid quantum computation scheme. Our scheme requires only five qubits (or less) per register, and it is robust against various kinds of imperfections, including imperfect initialization/measurement and low fidelity entanglement generation. We presented a Markov chain model to analyze the time overhead associated with the robust operations of measurement and entanglement generation. We found reasonable time overhead and considered practical implementation of quantum registers with ion traps or NV centers. We also provided an example using partial Bell measurement to prepare GHZ states for fault-tolerant quantum computation. It might be possible to further facilitate fault-tolerant quantum computation with systematic optimization using dynamic programming [70].

Acknowledgements

The authors wish to thank Gurudev Dutt, Lily Childress, Paola Cappellaro, Earl Campbell, Wolfgang Dür, Phillip Hemmer, and Charles Marcus. This work is supported by NSF, DTO, ARO-MURI, the Packard Foundations, Pappalardo Fellowship, and the Danish Natural Science Research Council.

APPENDIX A: BIT-PHASE TWO-LEVEL PUMPING SCHEME

In this Appendix we show that our bit-phase two-level entanglement pumping scheme can create pairs with fidelity arbitrarily close to unity, if we have perfect local operations. Numerical indication of this is shown in Fig. 23, and in the following we will provide a rigorous proof to this claim.

We assume that the initial state is a mixed state that has only diagonal terms in the Bell basis

$$\rho = a |\Phi^+\rangle \langle \Phi^+| + b |\Phi^-\rangle \langle \Phi^-| + c |\Psi^+\rangle \langle \Psi^+| + d |\Psi^-\rangle \langle \Psi^-|,$$

(A1)
where \( |\Phi^\pm\rangle = ((00) \pm |11\rangle)/\sqrt{2}, \quad |\Psi^\pm\rangle = ((01) \pm |10\rangle)/\sqrt{2} \) and the coefficients are non-negative and sum to unity. (This assumption is only made to simplify the presentation. For a general density matrix, only the diagonal elements given in Eq. (A1) are important [31].) After the purification the fidelity density retains this form but with new coefficients. Therefore, we only need four coefficients for each state using Bell basis, denoted as the fidelity vector \( \vec{F} = (a, b, c, d) \). We use the lower index to keep track of the pumping steps, so that the fidelity vector for the unpurified state is \( \vec{F}_0 = (a_0, b_0, c_0, d_0) \), and the vector for the purified state after \( n \) steps of entanglement pumping is \( \vec{F}_n = (a_n, b_n, c_n, d_n) \).

Suppose we use the \( \vec{F}_0 \) state to pump the state \( \vec{F}_n \) against bit-errors, the success probability is

\[
p_{n+1} = (a_0 + b_0) (a_n + b_n) + (c_0 + d_0) (c_n + d_n), \quad (A2)
\]

and the fidelity vector is

\[
\vec{F}_{n+1} = \frac{1}{p_{n+1}} (a_0 a_n + b_0 b_n, a_0 b_n + b_0 a_n, c_0 c_n + d_0 d_n, c_0 d_n + d_0 c_n), \quad (A3)
\]

for perfect local operations (measurement and CNOT gate).

Similarly, for pumping against phase-errors, the success probability is

\[
p_{n+1}' = (a_0 + c_0) (a_n + c_n) + (b_0 + d_0) (b_n + d_n), \quad (A5)
\]

and the fidelity vector is

\[
\vec{F}_{n+1}' = \frac{1}{p_{n+1}'} (a_0 a_n + c_0 c_n, b_0 b_n + d_0 d_n, a_0 c_n + c_0 a_n, b_0 d_n + d_0 b_n). \quad (A6)
\]

In general, any state can be turned into a so-called Werner state with the same fidelity, and as a worst case scenario we shall assume the unpurified Bell state to be a Werner state:

\[
\vec{F}_0 = \left( F_0, \frac{1 - F_0}{3}, \frac{1 - F_0}{3}, \frac{1 - F_0}{3} \right). \quad (A8)
\]

with \( F_0 > 1/2 \) to ensure that it contains distillable entanglement. For convenience of later discussion, we rewrite \( \vec{F}_0 \) as

\[
\vec{F}_0 = (1 - 3\alpha/2, \alpha/2, \alpha/2, \alpha/2) \quad (A9)
\]

where \( \alpha = \frac{2}{3} (1 - F_0) < \frac{1}{3} \).

### 1. First level pumping

For the first level of pumping against bit-errors, we may parameterize the fidelity vector

\[
\vec{F}_n = (a_n, b_n, c_n, d_n) = \left( \frac{1}{2} + \delta_n, \frac{1}{2} - \delta_n - 2\eta_n, \eta_n, \eta_n \right) \quad (A10)
\]

in terms of two variables \( \delta_n \) and \( \eta_n \), and obtain some bounds for these variables. Since we are pumping against bit-errors, \( \eta_n \) decreases with \( n \)

\[
\eta_{n+1} < \eta_n < \cdots < \eta_0 = \alpha/2 < 1/6. \quad (A11)
\]

The success probability for the \((n+1)\)th step of pumping is

\[
p_{n+1} = (1 - \alpha)(1 - 2\eta_n) + 2\alpha\eta_n \quad (A12)
\]

\[
< 1 - \alpha - 2\alpha\eta_n < 1 - \alpha, \quad (A13)
\]

and on the other hand we have

\[
p_{n+1} > (1 - \alpha)(1 - 2\eta_n) > \frac{2}{3}(1 - \alpha) \quad (A14)
\]

where the second inequality follows from \( \eta_n < 1/6 \) [73]. The third element of the fidelity vector is

\[
\eta_{n+1} = c_n = \frac{\alpha \eta_n + 2\alpha \eta_n}{p_{n+1}} = \frac{\alpha}{p_{n+1}} \eta_n. \quad (A15)
\]

Since \( p_{n+1} > \frac{2}{3}(1 - \alpha) \), we have

\[
\eta_{n+1} < \left( \frac{\alpha}{3(1 - \alpha)} \right) \eta_n < \cdots < \left( \frac{3\alpha}{2(1 - \alpha)} \right)^{n+1} c_0 \quad (A16)
\]

which indicates that \( \eta_n \) approaches zeros exponentially with respect to \( n \). When \( n \) is large enough, \( \eta_n \) is negligible and \( p_n \approx 1 - \alpha - O(\eta_n) \).

Similarly, we obtain the recursive relation for \( \delta_n \).

\[
\frac{1}{2} + \delta_{n+1} = a_{n+1} = \frac{1}{p_{n+1}} (a_0 a_n + b_0 b_n) \quad (A17)
\]

\[
= \frac{1}{2} \left( 1 - \alpha - 2\alpha \eta_n \right) + \frac{1 - 2\alpha}{p_{n+1}} \delta_n \quad (A18)
\]

\[
> \frac{1}{2} + \frac{1 - 2\alpha}{1 - \alpha} \delta_n. \quad (A19)
\]

Thus

\[
\delta_{n+1} > \frac{1 - 2\alpha}{1 - \alpha} \delta_n > \cdots > \left( \frac{1 - 2\alpha}{1 - \alpha} \right)^{n+1} \delta_0 \quad (A20)
\]

and \( \delta_0 = \frac{1 - 3\alpha}{2} \). We have therefore

\[
\frac{\eta_n}{\delta_n} < \left( \frac{2 - 4\alpha}{3\alpha} \right)^n \frac{\alpha}{1 - 3\alpha} \quad (A21)
\]
2. Second level pumping

After \( n_b \) steps of first level pumping, we have the fidelity vector

\[
\vec{F}_0 = \left( \frac{1}{2} + \delta'_0, \frac{1}{2} - \delta'_0 - 2 \eta'_0, \eta'_0, \eta'_0 \right)
\]

where \( \delta'_0 = \delta_{na} > \left( \frac{3\alpha}{1-\alpha} \right)^{n+1} \delta_0 \) and \( \eta'_0 = \eta_{na} < \left( \frac{3\alpha}{1-\alpha} \right)^{n+1} \eta_0 \).

The fidelity vector after \( n_p \) steps of pumping against phase errors is

\[
\vec{F}_{n} = \left( \frac{1}{2} + \delta'_n, \frac{1}{2} - \delta'_n - d_n, \eta'_n, d_n \right),
\]

where \( \delta'_n < 1/2 \) and \( \eta'_n < 1/2 \).

The success probability for the \((n+1)\)th step is

\[
\begin{align*}
p'_{n+1} & = \left( \frac{1}{2} + \delta'_0 + \eta'_0 \right) \left( \frac{1}{2} + \delta'_n + \eta'_n \right) \\
& \quad + \left( \frac{1}{2} - \delta'_0 - \eta'_0 \right) \left( \frac{1}{2} - \delta'_n - \eta'_n \right) \\
& = \frac{1}{2} + 2 (\delta'_0 + \eta'_0) (\delta'_n + \eta'_n)
\end{align*}
\]

If \( \frac{1}{2} > \delta'_n > \delta'_0 > \eta'_n > \eta'_0 \), we have

\[
\delta'_0 \eta'_n + \delta'_n \eta'_0 + \eta'_0 \eta'_n < \delta'_0 \eta'_n + \eta'_0 < \delta'_0 \zeta,
\]

where we in the second inequality introduced a number \( \zeta \) such that \( \delta'_0 \eta'_n + \eta'_0 < \delta'_0 \zeta \). In the next subsection, we will show that we may choose \( \zeta \) small (i.e., \( \zeta = 2\varepsilon \)) such that

\[
\begin{align*}
\delta'_{n+1} & > \frac{\delta'_n + \delta'_0 (1 - 2\zeta)}{1 + 4\delta'_0 \lambda'_{na} (1 + \zeta)} \\
\lambda'_{n+1} & > \frac{\lambda'_n + \lambda'_0 (1 - 2\zeta)}{1 + 4\lambda'_0 \lambda'_{na} (1 - 2\zeta)}
\end{align*}
\]

We now consider the elements of \( \vec{F}_{n} \). On one hand, the erroneous admixture of \( |\Psi^-\rangle \) described by \( \{d_n\} \) keep decreasing with \( n \), since \( |\Psi^-\rangle \) errors are also purified during the second level pumping. On the other hand, the erroneous admixture of \( |\Psi^+\rangle \) described by \( \{\eta'_n\} \) may increase with \( n \), but it is upper-bounded by the following relation:

\[
\begin{align*}
\eta'_{n+1} & = \frac{a_0 c_n + c_0 a_n}{p'_{n+1}} \\
& < 2 \left( \frac{1}{2} + \delta'_0 \right) \eta'_n + \eta'_0 \left( \frac{1}{2} + \delta'_n \right) \\
& < (1 + 2\delta'_0) \eta'_n + 2\eta'_0 \\
& < (1 + 2\delta'_0)^{n+1} \eta'_0,
\end{align*}
\]

and one can also show the lower bound for \( \eta'_{n+1} \) by induction:

\[
\eta'_{n+1} > a_0 c_n + c_0 a_n > \frac{1}{2} (\eta'_n + \eta'_0) > \eta'_0.
\]

However, \( \delta'_n \) approaches \( 1/2 \) much faster:

\[
\begin{align*}
\frac{1}{2} + \delta'_{n+1} & = a_{n+1} = \frac{a_0 a_n + c_0 c_n}{p'_{n+1}} \\
& = \left( \frac{1}{2} + \delta'_0 \right) \left( \frac{1}{2} + \delta'_n \right) \eta'_0 + 2 \eta'_0 \eta'_n \\
& = \frac{1}{2} + 2 (\delta'_0 + \eta'_0) (\delta'_n + \eta'_n)
\end{align*}
\]

3. \( \varepsilon - N \) argument

For \( \forall \varepsilon > 0 \) and \( \alpha < 2/7 \), we may choose \( n_b \)

\[
n_b \geq \max \left\{ \frac{\ln \varepsilon - \ln n_b}{\ln \varepsilon - \ln x}, \frac{3\ln \varepsilon - \ln \frac{2}{1 - \alpha}}{\ln \varepsilon - \ln \frac{2}{1 - \alpha}} \right\},
\]

such that

\[
\eta'_0 < \left( \frac{3\alpha}{2(1 - \alpha)} \right)^{n_b} \frac{\alpha}{2} < \varepsilon^3
\]

\[
\frac{\eta'_0}{\delta'_0} \frac{\delta'_0}{\delta_{na}} < \left( \frac{3\alpha}{2 - 4\alpha} \right)^{n_b} \frac{\alpha}{1 - 3\alpha} < \varepsilon
\]

and

\[
\delta'_0 > \left( \frac{1 - 2\alpha}{1 - \alpha} \right)^{n_b} \frac{1 - 3\alpha}{2}
\]

Then we choose for \( n_p \)

\[
\frac{2 \ln \varepsilon}{\ln (1 - 2\delta'_0)} > n_p > \frac{\ln \varepsilon}{\ln (1 - 2\lambda_0(1 - 2\zeta))}.
\]
Such \( n_p \) always exists, for \( \frac{2 \ln \varepsilon}{\ln(1-2\delta'_0)} \approx 2 \ln \varepsilon - 2 \delta'_0 > -2\delta'_0 \sqrt{(1+\zeta)(1-2\zeta)} \approx \ln(1-2\delta'_0(1-2\zeta)). \) Thus, we have
\[
\eta'_{n_p} < (1 + 2\delta'_0)^{n_p} \eta'_0 < (1 - 2\delta'_0)^{-n_p} \eta'_0 < \varepsilon^{-2} \eta'_0 < \delta'_0' \tag{A37}
\]
and
\[
1/2 - \lambda'_{n_p} < (1 - 2\delta'_0 (1 - 2\zeta))^{n_p} (1/2 - \lambda'_{0}) < \varepsilon/2 \tag{A38}
\]
We now verify the two required relations which are underlined in the previous discussion. The relation \( \frac{1}{2} > \delta'_n > \delta'_0 > \eta'_n > \eta'_0 \) is satisfied for all \( 0 < n \leq n_p \).
And the relation \( \eta'_0\delta'_n + \xi'_0 < \delta'_0\xi'_0 \) is also implied, since \( \eta'_n + \eta'_0/\delta'_0 < \varepsilon^{-2}\eta'_0 + \varepsilon < 2\varepsilon \equiv \xi'_0 \).
Finally, the achievable fidelity for the above choice of \( n_b \) and \( n_p \) is
\[
1 - F_{n_p,n_b} = 1/2 - \delta'_n = 1/2 - \left( \frac{1 - 2\zeta}{1 + \zeta} \right)^{1/2} \lambda'_{n_p} < 1/2 - \left( \frac{1 - 2\zeta}{1 + \zeta} \right)^{1/2} (1/2 - \varepsilon) < 4\varepsilon. \tag{A39}
\]
The bit-phase two-level pumping thus allows us to approach \( F = 1 \) with arbitrary good precision.

**APPENDIX B: MARKOV CHAIN MODEL FOR TWO-LEVEL PUMPING**

Here we present the Markov chain model for two-level entanglement pumping.

The state transition diagram for two-level entanglement pumping is shown in Fig. We assume that the required pumping steps are \( n_b \) and \( n_p \) for the two levels, respectively. Since two entangled pairs are stored, we need two labels to track the intermediate state for two-level entanglement pumping.

Here we use "0,0" to denote the initial state with no Bell pairs, "0,j+1" for the state with one purified pair surviving \( j \) steps of pumping at the first level, "k+1,j+1" for the state with one purified pair surviving \( k \) steps of pumping at the second level and one purified pair surviving \( j \) steps of pumping at the first level, "k+1,*" for the state with one purified pair surviving \( k \) steps of pumping at the second level and one purified pair surviving \( n_b \) steps of pumping at the first level, and "*,0" for the final state with one purified pair surviving \( n_p \) steps of pumping at the second level.

For the first level pumping, the (success) transition probability from state "k,j" to state "k,j+1" is \( q_{j} \), while the (failure) transition probability from state "k,j" to state "k,0" is \( 1 - q_{j} \), for \( j = 0,1,\cdots,n_b \). For the second level pumping, the (success) transition probability from state "k,*,*" to state "k+1,0" is \( Q_{k} \), while the (failure) transition probability from state "k,*,*" to state "0,0" is \( 1 - Q_{k} \), for \( k = 0,1,\cdots,n_p \). Here the transition probabilities \( \{q_{j}\} \) and \( \{Q_{k}\} \) can be calculated according to the density matrices of the intermediate purified Bell pairs \( \text{[36]} \). The final state is self-trapped, and goes back to itself with unit probability. Altogether there are \( (n_b + 1)(n_p + 1) + 1 \) states.

In order to fulfill the requirement that each transition attempt consumes one unpurified Bell pair, we need to contract the states of "k,*,*" and "k+1,0" into one state, since this transition does not consume any unpurified Bell pair. After the contraction, there are \( (n_b + 1)(n_p + 1) + 1 \) states remaining.

Therefore, we may use a (column) vector \( \vec{P} \) with \( (n_b + 1)(n_p + 1) + 1 \) elements to characterize the probability distribution among all \( (n_b + 1)(n_p + 1) + 1 \) states. From the \( t \)-th attempt to the \( (t+1) \)-th attempt, the probability vector evolves from \( \vec{P}(t) \) to \( \vec{P}(t+1) \) according to the following rule
\[
\vec{P}(t+1) = M \vec{P}(t), \tag{B1}
\]
and the transition matrix is
\[
M = \begin{pmatrix}
M_1 + q_{n_b} (1 - Q_{1}) N_1 & q_{n_b} (1 - Q_{1}) N_1 & \cdots & q_{n_b} (1 - Q_{2}) N_1 & 0 \\
q_{n_b} N_1 & M_1 & \cdots & M_1 & \cdots & M_1 \\
0 & \cdots & M_1 & q_{n_b} Q_{n_p} N_1 & 1
\end{pmatrix},
\tag{B2}
\]
with sub-matrices:
\[
M_1 = \begin{pmatrix}
0 & 1 - q_{1} & \cdots & 1 - q_{n_b} & 0 \\
1 & 0 & \cdots & 0 & 1 \\
q_{1} & \cdots & 0 & 1 & \cdots & 1 \\
\cdots & 0 & 1 & \cdots & \cdots & \cdots \\
q_{n_b} & 0 & 1 & \cdots & \cdots & \cdots
\end{pmatrix}_{(n_b+1)\times(n_b+1)}.
\tag{B3}
\]
and

\[
N_1 = \begin{pmatrix}
0 & \cdots & 0 & 1 \\
0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots \\
0 & \cdots & 0
\end{pmatrix}_{(n_k+1) \times (n_k+1)}.
\] (B4)

**APPENDIX C: FAULT-TOLERANT PREPARATION OF 2ⁿ-QUBIT GHZ STATE**

We consider fault-tolerant preparation of a 2ⁿ-qubit GHZ state with \( n \geq 3 \). We label 2ⁿ registers by 0, 1, \ldots, 2ⁿ − 1. The GHZ state can be prepared in just 3 (clock cycles): in the first clock cycle, we perform PBM for register pairs \( (2m, 2m + 1) \) with integer \( m \); in the second clock cycle, we perform PBM for pairs \( (2m, 2m + 2) \) and \( (2m + 1, 2m + 3) \); in the last clock cycle, we perform PBM for pairs \( (8m + 3, 8m + 5), (8m + 4, 8m + 6), (16m' + 7, 16m' + 9), (16m' + 8, 16m' + 10), \ldots, (2^{n-1} + 2^{n-2} - 1, 2^{n-1} + 2^{n-2} + 1), \) and \( (2^{n-1} + 2^{n-2}, 2^{n-1} + 2^{n-2} + 2) \) in parallel.

In order to prepare the specific GHZ state \(|00\cdots0\rangle + |11\cdots1\rangle\), we still need to perform bit-flip operations for individual registers, which are determined by the measurement outcomes for all PBMs. Suppose the error probability for each PBM is \( p \). Since the redundancy checks of the PBMs impose consistency requirements for measurement outcomes (error detection), the probability for undetected errors in measurement outcomes has been suppressed to \( O(p^3) \) for each PBM. To the leading order of \( p \), we only need to consider the phase errors from PBM that are not detected by the redundancy check. Thus, the total error probability scales as \( 2^{n-3}p \), and the error probability for each register is only approximately \( 3p/2 \). Therefore, we have demonstrated a PBM-based scheme to prepare the GHZ state fault-tolerantly, which is efficient in both time and physical-resources.

[71] The error model introduced in Eq. 1 can be regarded as the worst case error, since it in principle includes all possible errors that can happen to the system.
[72] The verification-based scheme requires $k \geq 0$ verifications for a good initialization and suppresses the initialization error to $\varepsilon_I \approx (p_M + p_{AM})^{k+1} + \frac{2k+1}{k+1}p_L$, a smaller error probability than that of the measurement-based scheme. However, the verification-based scheme may fail at some intermediate step and require reiteration of the verification from scratch, which is very similar to the one-level entanglement pumping, described by the Markov chain model in Sec. [VII].
[73] For depolarizing or dephasing errors, the estimated $t_C$ is approximately 1.3 times less by using the average fidelity.
[74] Note that the lower bound for $p_{n+1}$ can be arbitrarily close to $1 - \alpha$ for sufficiently large $n$. Here, for clarity we use the loose bound $\frac{2}{3} (1 - \alpha) < p_{n+1}$, which later imposes a constraint $\alpha < 2/7$ (i.e. $F > 4/7 \approx 0.57$). In principle, by choosing tighter lower bounds for $p_{n+1}$, the following proof works for all $\alpha < 1/3$. 
[75] For depolarizing or dephasing errors, the estimated $t_C$ is approximately 1.3 times less by using the average fidelity.