Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects

C. Monroe,[1] R. Raussendorf,[2] A. Ruthven,[2] K. R. Brown,[3] P. Maunz,[4,*] L.-M. Duan,[5] and J. Kim[4]

[1]*Joint Quantum Institute, University of Maryland Department of Physics and National Institute of Standards and Technology, College Park, Maryland 20742, USA*

[2]*Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia V6T1Z1, Canada*

[3]*Schools of Chemistry and Biochemistry; Computational Science and Engineering; and Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*

[4]*Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina 27708, USA*

[5]*Department of Physics and MCTP, University of Michigan, Ann Arbor, Michigan 48109, USA and Center for Quantum Information, Tsinghua University, Beijing 100084, China*

(Received 22 June 2013; published 13 February 2014)

The practical construction of scalable quantum-computer hardware capable of executing nontrivial quantum algorithms will require the juxtaposition of different types of quantum systems. We analyze a modular ion trap quantum-computer architecture with a hierarchy of interactions that can scale to very large numbers of qubits. Local entangling quantum gates between qubit memories within a single register are accomplished using natural interactions between the qubits, and entanglement between separate registers is completed via a probabilistic photonic interface between qubits in different registers, even over large distances. We show that this architecture can be made fault tolerant, and demonstrate its viability for fault-tolerant execution of modest size quantum circuits.

DOI: 10.1103/PhysRevA.89.022317 PACS number(s): 03.67.Lx, 03.67.Pp, 32.80.Qk, 42.50.Ex

I. INTRODUCTION

A quantum computer is composed of at least two quantum systems that serve critical functions: a reliable quantum memory for hosting and manipulating coherent quantum superpositions, and a quantum bus for the conveyance of quantum information between memories. Quantum memories are typically formed out of matter such as individual atoms, spins localized at quantum dots or impurities in solids, or superconducting junctions [1]. On the other hand, the quantum bus typically involves propagating quantum degrees of freedom such as electromagnetic fields (photons) or lattice vibrations (phonons). A suitable and controllable interaction between the memory and the bus is necessary to efficiently execute a prescribed quantum algorithm. The current challenge in any quantum-computer architecture is to scale the system to very large sizes, where errors are typically caused by speed limitations and decoherence of the quantum bus or its interaction with the memory. The most advanced quantum bit (qubit) networks have thus been established only in very small systems, such as individual atomic ions bussed by the local Coulomb interaction [2] or superconducting Josephson junctions coupled capacitively or through microwave striplines [3,4]. In this paper, we propose and analyze a hierarchy of quantum information processing units in a modular quantum-computer architecture that may allow the scaling of high performance quantum memories to useful sizes [5]. This architecture compares to the "multicore" classical information processor, and is suitable for the implementation of complex quantum circuits utilizing the flexible connectivity provided by a reconfigurable photonic interconnect network. Unlike previous related proposals [6–13], all of the rudiments of a modular universal scalable ion trap quantum-computer (MUSIQC) architecture have been experimentally demonstrated in small-scale trapped ion systems. Furthermore, we show this reconfigurable architecture can be made fault tolerant over a wide range of system parameters, using a variety of fault-tolerant schemes.

We specialize to the use of atomic ion qubit memories, due to the outstanding qubit properties demonstrated to date. Qubits stored in ions enjoy a level of coherence that is unmatched in any other physical system, underlying the reason such states are also used as high performance atomic clocks. Moreover, atomic ions can be initialized and detected with nearly perfect accuracy using conventional optical pumping and state-dependent fluorescence techniques [14]. There have been many successful demonstrations of controlled entanglement of several-ion quantum registers in the past decade involving the use of qubit state-dependent forces supplied by laser beams [2,15]. These experiments exploit the collective motion of a small number of trapped ion qubits, but as the size of the ion chain grows, such operations are more susceptible to external noise, decoherence, or speed limitations.

One promising approach to scaling trapped ion qubits is the quantum charge-coupled device (QCCD), where physical shuttling of ions between trapping zones in a multiplexed trap is used to transfer qubits between (short) chains of ions [14,16]. This approach involves advanced ion trap structures, perhaps with many times more discrete electrodes as trapped ion qubits, and therefore motivates the use of micrometer-scale surface traps [17–19] and novel fabrication techniques [20–22]. The shuttling approach requires careful control of the time-varying

*Present address: Sandia National Laboratories, Albuquerque, NM 87123, USA.

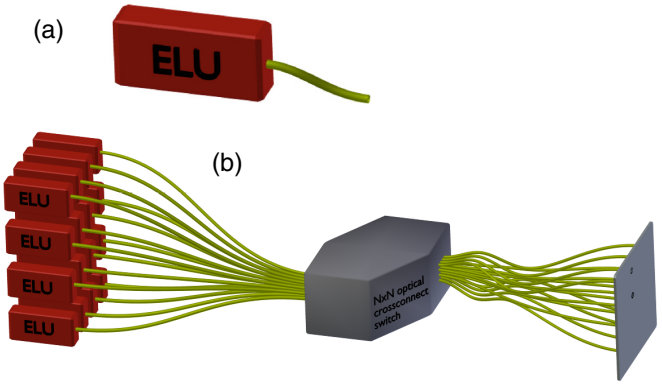

FIG. 1. (Color online) Hierarchical modular quantum-computer architecture hosting $N = N_{\rm ELU} N_q$ qubits. (a) The elementary logic units (ELU) consist of a register of N_q trapped atomic ion qubits, whereby entangling quantum logic gates are mediated through the local Coulomb interaction between qubits. (b) One or more atomic qubits within each of the $N_{\rm ELU}$ registers are coupled to photonic quantum channels, and through a reconfigurable optical crossconnect switch (OXC, center), fiber beamsplitters, and position sensitive imager (right), qubits between different registers can be entangled.

trapping potential to manipulate the position of the atomic ion, and cannot easily be extended over large distances for quantum communications applications. The QCCD approach is expected to enable a quantum information processing platform where basic quantum error correction and quantum algorithms can be realized. Further scaling in the near future is likely limited by the complexity of the trap design, diffraction of optical beams [23], and the hardware controllers to operate the system.

Here we describe and analyze a MUSIQC architecture that may enable construction of quantum processors with up to 10^6 qubits utilizing component technologies that have already been demonstrated. This architecture features two elements described in Sec. II: stable trapped ion multiqubit registers that can further be connected with ion shuttling, and scalable photonic interconnects that can link these registers in a flexible configuration over large distances, as shown in Fig. 1. We highlight two unique features enabled by this hardware platform. In Sec. III, we articulate architectural advantages of this approach that allow significant speedup and resource reduction in quantum circuit execution over other hardware architectures, enabled by the ability to operate quantum gates between qubits throughout the entire processor regardless of their relative location. In Sec. IV, we prove how a quantum network such as MUSIQC can support fault-tolerant error correction even in the face of probabilistic and slow interconnects. Section V discusses the experimental challenges and technological developments necessary for its realization. While we focus our discussions on quantum registers composed of trapped atomic ions, the networking aspect of this architecture is applicable to other qubit platforms that feature strong optical transitions, such as quantum dots, neutral atoms, or nitrogen-vacancy (NV) color centers in diamond [1].

II. QUANTUM COMPUTING IN A MODULAR ARCHITECTURE

A. The modular elementary logic unit (ELU)

The base unit of MUSIQC is a collection of N_q qubit memories with local interactions, called the elementary logic unit (ELU). Quantum logic operations within the ELU are ideally fast and deterministic, with error rates sufficiently small that fault-tolerant error correction within an ELU is possible [24]. We represent the ELU with a crystal of $N_q \gg 1$ trapped atomic ions as shown in Fig. 2(a), with each qubit comprising internal energy levels of each ion, labeled as $|\uparrow\rangle$ and $|\downarrow\rangle$, separated by frequency ω_0. We assume the qubit levels are coupled through an atomic dipole operator $\hat{\mu} = \mu(|\uparrow\rangle\langle\downarrow| + |\downarrow\rangle\langle\uparrow|)$. The ions interact through their external collective modes of quantum harmonic motion. Such phonons can be used to mediate entangling gates through application of qubit-state-dependent optical or microwave dipole forces [25–27]. There are many known protocols for phonon-based gates between ions, and here we summarize the main points relevant to the size of the ELU and the larger architecture.

An externally applied near-resonant running wave field with amplitude $E(\hat{x}) = E_0 e^{ik\hat{x}}$ and wave number k couples to the atomic dipole through the interaction Hamiltonian $\hat{H} = -\hat{\mu} E(\hat{x})$, and by suitably tuning the field near sidebands induced by the harmonic motion of the ions [14] a qubit-state-dependent force results. In this way, qubits can be mapped onto phonon states [14,25] and then onto other qubits for entangling operations with characteristic speed $R_{\rm gate} = \eta\Omega$, where $\eta = \sqrt{\hbar k^2/(2m_0 N_q \omega)}$ is the Lamb Dicke parameter, m_0 is the mass of each ion, ω the frequency of harmonic oscillation of the collective phonon mode, and $\Omega = \mu E_0/2\hbar$ is the Rabi frequency of the atomic dipole independent of motion. For optical Raman transitions between qubit states (e.g., atomic hyperfine ground states) [14], two fields are each detuned by

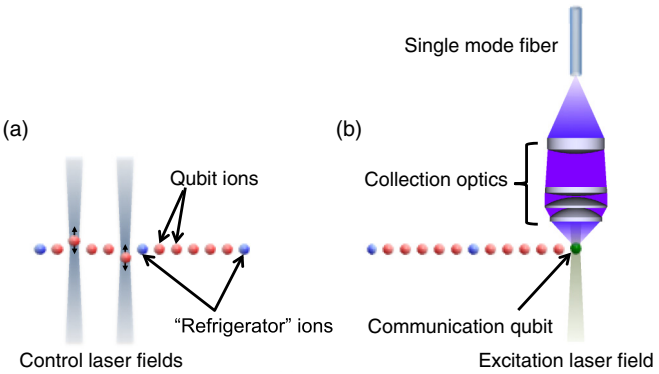

FIG. 2. (Color online) Elementary logic unit (ELU) composed of a single crystal of N_q trapped atomic ion qubits coupled through their collective motion. (a) Classical laser fields impart qubit state-dependent forces on one or more ions, affecting entangling quantum gates between the memory qubits. Second ion species is introduced as refrigeration ions. (b) One or more of the ions (rightmost in the figure) are coupled to a photonic interface, where a classical laser pulse maps the state of these communication qubits onto the states of single photons (e.g., polarization or frequency), which then propagate along an optical fiber to be interfaced with other ELUs.

Δ from an excited state of linewidth $\gamma \ll \Delta$, and when their difference frequency is near resonant with the qubit frequency splitting ω_0, we use instead $\Omega = (\mu E_0)^2/(2\hbar^2\Delta)$.

The typical gate speed within an ELU therefore slows down with the number of qubits N_q as $R_{\rm gate} \sim N_q^{-1/2}$. As the size of the ELU grows, so will the coupling between the modes of collective motion that could lead to crosstalk. However, through the use of pulse-shaping techniques [28], the crosstalk errors need not be debilitating, although the effective speed of a gate will slow down with size N_q. Changes of the ions' motional states during the gate, arising from sources like heating of the motional modes [29–31] or fluctuating fields, will degrade the quality of the gates, leading to practical limits on the size of the ELU on which high performance gates can be realized. It is likely that long chains will require periodic "refrigerator" ions to remove motional excitations between gates. Since cooling is a dissipative process, these cooling ions should be chosen to be different isotope or species of ions and quench motional heating through sympathetic cooling [32]. We estimate that ELUs ranging from $N_q = 10$–100 should be possible [2,15]. More than one ELU chain can be integrated into a single chip by employing ion shuttling through more complex ion trap structures [16]. Such extended ELUs (EELUs) consisting of N_E ELU chains can contain a total of $N_q N_E = 20$–1000 physical qubits. For simplicity, we focus the remainder of the article on systems with one ELU per chip ($N_E = 1$).

B. Probabilistic linking of ELUs

Two qubits from a pair of ELUs (or EELUs) can be entangled by each emitting photon that interferes with each other. Entanglement generated between these "communication qubits" can be utilized as a resource to perform a two-qubit gate between any pair of qubits, one from each ELU, using local qubit gates, measurements, and classical communication between the ELUs. In this scheme, the communication qubit is driven to an excited state with fast laser pulses whose duration $\tau_e \ll 1/\gamma$, so that no more than one photon is emitted from each qubit per excitation cycle following the atomic radiative selection rules [Fig. 2(b)]. The photon can be postselected so that one of its degrees of freedom (polarization, frequency, etc.) is entangled with the state of the communication qubit [33–36]. When the photons from two communication qubits are mode matched and interfere on a 50:50 beamsplitter, detectors on the output modes of the beamsplitter can herald the creation of entanglement between the memory qubits [37–41].

We consider two types of photonic connections, characterized by the number of total photons used in the entanglement protocol between two communication qubits [42]. For type I connections [shown in Fig. 3(a)], each communication qubit with an index i (or j) is weakly excited with probability $p_e \ll 1$ and the state of the ion+photon qubit pair is approximately written (ignoring the higher-order excitation probabilities) as $\sim \sqrt{1-p_e}|\downarrow\rangle_i|0\rangle_i + e^{ikx_i}\sqrt{p_e}|\uparrow\rangle_i|1\rangle_i$ where $|n\rangle_i$ denotes the state of n photons radiating from the communication qubit into an optical mode i, x_i is the path length from the emitter i to a beamsplitter, and k the optical wave number [37]. When two communication qubits i and j are excited in this way and the photons interfere at the beamsplitter,

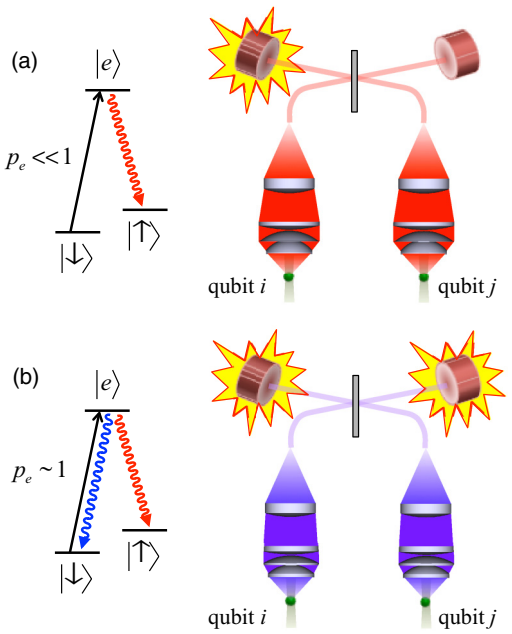

FIG. 3. (Color online) (a) Type I interference from photons emitted from two communication qubits. Each qubit is weakly excited so that single-photon emission has a very small probability yet is correlated with the final qubit state. The output photonic channels are mode matched with a 50:50 beamsplitter and subsequent detection of a photon from either output port heralds the entanglement of the communication qubits. The probability of two photons present in the system is much smaller than that of detecting a single photon. (b) Type II interference involves the emission of one photon from each communication qubit, where the internal state of the photon (e.g., its color) is correlated with the qubit state. After two-photon interference at the beamsplitter, coincidence detection of photons at the two detectors heralds the entanglement of the communication qubits.

the detection of a single photon in either detector placed at the two output ports of the beamsplitter heralds the creation of the state $[e^{ikx_j}|\downarrow\rangle_i|\uparrow\rangle_j \pm e^{ikx_i}|\uparrow\rangle_i|\downarrow\rangle_j]/\sqrt{2}$ with success probability $p = p_e F \eta_D$, where F is the fractional solid angle of emission collected, η_D is the detection efficiency including any losses between the emitter and the detector, and the sign in this state is determined by which one of the two detectors fires. Following the heralding of a single photon, the (small) probability of errors from double excitation and detector dark counts are given, respectively, by p_e^2 and $R_{\rm dark}/\gamma$ where $R_{\rm dark}$ is the rate of detector dark counts. For type I connections to be useful, the relative optical path length $x_i - x_j$ must be stable to much better than the optical wavelength $\sim 2\pi/k$.

For type II connections [shown in Fig. 3(b)], each communication qubit is excited with near unit probability $p_e \sim 1$ and the single photon carries its qubit through two distinguishable internal photonic states (e.g., polarization or optical frequency). For example, the state of the system containing both communication and photonic qubits is written as $[e^{ik_\downarrow x_i}|\downarrow\rangle_i|\nu_\downarrow\rangle_i + e^{ik_\uparrow x_i}|\uparrow\rangle_i|\nu_\uparrow\rangle_i]/\sqrt{2}$, where $|\nu_\downarrow\rangle_i$ and $|\nu_\uparrow\rangle_i$ denote the frequency qubit states of a single-photon emitted by the ith communication qubit with wave numbers $k_\downarrow$ and $k_\uparrow$ associated with optical frequencies $\nu_\uparrow$ and $\nu_\downarrow$, respectively.

Here, $|\nu_\uparrow - \nu_\downarrow| = \omega_0 \gg \gamma$ so that these two frequency qubit states are distinguishable. The coincidence detection of photons from two such communication qubits i and j after interfering at a 50:50 beamsplitter herald the successful entanglement of the communication qubits, creating the state $[e^{i(k_\downarrow x_i + k_\uparrow x_j)}|\downarrow\rangle_i|\uparrow\rangle_j - e^{i(k_\uparrow x_i + k_\downarrow x_j)}|\uparrow\rangle_j|\downarrow\rangle_i]/\sqrt{2}$ with success probability $p = (p_e F \eta_D)^2/2$ [38,39]. Other schemes have also been proposed that scale similar to the type II connections [43,44].

The success probability of the two-photon type II connection may be lower than that of the type I connection when the light collection efficiency is low, but type II connections are much less sensitive to optical path length fluctuations. The stability requirement of the relative path length $x_i - x_j$ is only at the level of the wavelength associated with the difference frequency $2\pi c/\omega_0$ of the photonic frequency qubit, which is typically at the centimeter scale for hyperfine-encoded communication qubits.

In both cases, the mean connection time is given by $\tau_E = 1/(Rp)$ where R is the repetition rate of the initialization/excitation process and p is the success probability of generating the entanglement. For atomic transitions, $R \sim 0.1(\gamma/2\pi)$, and for typical free-space light collection ($F \sim 10^{-2}$) and taking $\eta_D \sim 0.2$, we find for a type I connection $\tau_E \sim 5$ ms ($p_e = 0.05$) and for a type II connection $\tau_E \sim 250$ ms where we have assumed $\gamma/2\pi = 20$ MHz. Type II connections eventually outperform that of type I with more efficient light collection, which can be accomplished by integrating optical elements with the ion trap structure without any fundamental loss in fidelity. Eventually, $\tau_E \sim 1$ ms should be possible in both types of connections [45].

The process to generate ion-ion entanglement using photon interference requires resonant excitation of the communication qubits, and steps must be taken to isolate the communication qubit from other memory qubits so that scattered light from the excitation laser and the emitted photons do not disturb the spectator memory qubits. It may be necessary to physically separate or shuttle the communication qubit away from the others, invoking some of the techniques from the QCCD approach. This crosstalk can also be eliminated by utilizing a different atomic species for the communication qubit [46], so that the excitation and emitted light is sufficiently far from the memory qubit optical resonance to avoid causing decoherence. The communication qubits do not require excellent quantum memory characteristics, because once the entanglement is established between the communication qubits in different ELUs, they can immediately be swapped with neighboring memory qubits in each ELU.

C. Reconfigurable connection network in MUSIQC

The MUSIQC architecture allows a large number N_{ELU} of ELUs (or EELUs) to be connected with each other using the photonic channels, as shown in Fig. 1. The connection is made through an optical crossconnect (OXC) switch [47] with N_{ELU} input and output ports. The photon emitted from the communication qubit in each ELU is collected into a single-mode fiber and directed to a corresponding input port of the OXC switch. Up to $N_{\mathrm{ELU}}/2$ Bell state detectors, each comprising two fibers interfering on a beamsplitter and two detectors, are connected to the output ports of the OXC switch. The OXC switch is capable of providing an optical path between any input fiber to any output fiber that is not already connected to another input fiber. An ideal OXC switch achieves full nonblocking connectivity with uniform optical path lengths. This optical network provides fully reconfigurable interconnect network for the photonic qubits, allowing entanglement generation between any pair of ELUs in the processor with up to $N_{\mathrm{ELU}}/2$ such operations running in parallel. OXC switches that support 200–1100 ports utilizing microelectromechanical systems (MEMS) technology have been constructed and are readily available [47,48]. In practice, the photon detection can be accomplished in parallel with a conventional charge-coupled-device (CCD) imager or an array of photon-counting detectors, with pairs of regions on the CCD or the array elements associated with particular pairs of output ports from the fiber beamsplitters, as shown in Fig. 1.

D. Current status of ion qubit experiments

Trapped ion experiments feature high quality qubits, and have demonstrated high quality quantum logic operations in the past two decades. Hyperfine qubits utilizing two ground states of an atom are shown to routinely exhibit long coherence times of a few seconds [49,50], and more than an order of magnitude longer when operated in the "field-independent" regime where the energy splitting of the two qubit states is independent of the magnetic field fluctuations to first order [51,52]. Optical qubits between a ground state and a metastable excited state are also compelling candidates for qubits, when stable laser systems can be constructed to control the transition [14,15]. Recent experiments showed substantial progress in improving the fidelity of individual operations necessary for the quantum computation processes.

High fidelity qubit preparation with near-unity fidelity is routinely achieved by optical pumping, although the experimental characterization is typically limited by the qubit state detection process. This is commonly referred to as state preparation and measurement (SPAM) errors. High fidelity qubit state detection with errors in the 10^{-4} range are available in the optical qubit [53] with an average detection time of 150 μs, while a direct detection of hyperfine qubits can be performed with 10^{-3} errors [54] with an average detection time of 50 μs. Single qubit gates on hyperfine qubits driven by microwave sources show the lowest level of error, in the 10^{-5}–10^{-6} range [55,56]. The best performance of two-qubit gate demonstrated to date features errors in the 7×10^{-3} range [57], while recent progress indeed is approaching closer to the 10^{-3} range [56].

The prospect for further reduction of SPAM errors and higher fidelity remains highly positive in the future, as the researchers continue to search for protocols that are robust against experimental errors, and improve experimental imperfections that lead to residual errors.

III. PERFORMANCE ADVANTAGE OF MUSIQC ARCHITECTURE

In this section we examine the performance of the MUSIQC architecture under the assumption of large ELUs and low

TABLE I. Assumptions on the time scales of quantum operation primitives used in the model.

Quantum primitive	Single-qubit gate	Two-qubit gate	Toffoli gate	Qubit measurement	Remote entanglement generation
Operation time (μs)	1	10	10	30	3000

errors. This allows us to directly compare our results to previous studies on ion traps using the Steane [7, 1,3] code and the quantum logic array (QLA) architecture [58,59]. In Sec. IV, we will examine the limits of small ELUs and large errors.

A. Computation model in MUSIQC

In the circuit model of quantum computation, execution of two-qubit gates creates the entanglement necessary to exploit the power of quantum physics in computation [24]. In the alternate model of measurement-based cluster-state quantum computation, all of the entanglement is generated at the beginning of the computation, followed by conditional measurements of the qubits [60]. The MUSIQC architecture presented here follows the circuit model of computation within each ELU, but the probabilistic connection between ELUs is carried out by generation of entangled Bell pairs similar to the cluster-state computation model. In this sense, MUSIQC realizes a hybrid model of quantum computation, driven by the generation rate and burn rate of entanglement between the ELUs. In the event the generation rate of entangled Bell pairs between ELUs is lower than the burn rate, each ELU would require the capacity to store enough initial entanglement so that the end of the computation can be reached at the given generation and burn rates of entanglement. The hybrid nature of MUSIQC provides a unique hardware platform with three distinct advantages: fully reconfigurable connectivity to dynamically adjust the connectivity graph, constant time scale to perform operations between distant qubits, and moderate ELU size adequate for practical implementation. One can further reduce the entanglement generation time by time-division multiplexing (TDM) the communication ports at the expense of added qubits. Moreover, the temporal mismatch between the remote entanglement generation and local gates is reduced as the requirement of error correction increases the logical gate time.

For a complex quantum algorithm associated with a problem size of n bits, logical operations between spatially distant qubit pairs are necessary. In a hardware architecture where only local gate operations are allowed (e.g., nearest neighbor gates), gate operations between two (logical) qubits separated by long distances can be implemented with resource overhead (number of qubits, parallel operations, and/or communication time) polynomial in the distance between qubits, $O(n^k)$. When a large number of parallel operations is available, one can employ entanglement swapping protocols to efficiently distribute entanglement with communication times scaling either polylogarithmically [61], or even independent of the communication distance [62]. This procedure requires extra qubits that are used to construct quantum buses for long-distance entanglement distribution, and the architecture adopting such buses was referred to as quantum logic array (QLA) [58]. We construct a simple model that provides a direct comparison between the QLA and MUSIQC architectures in terms of the resources required to execute useful quantum algorithms. Despite the slow entanglement generation times, we find that the performance of the MUSIQC architecture is comparable to QLA (and its variations [63]), with substantial advantage in required resources and feasibility for implementation.

In our simplified model, we consider hardware (1) capable of implementing a Steane [7, 1,3] quantum error correction code to multiple levels of concatenation, and (2) where all gate operations are performed following fault-tolerant procedures. This simplified model is designed to estimate the execution time of the circuits in select architectures, and is not intended to provide the complete fault-tolerant analysis of the quantum circuit. For this model, we therefore require that the physical error levels are sufficiently low ($\sim 10^{-7}$) to produce the correct answer with order-unity probability using only up to three levels of concatenation of Steane code. We also assume that the quality of entangled pairs that are generated in MUSIQC architecture is high enough that error correction schemes can improve its fidelity sufficient to achieve fault tolerance [64]. The hardware is based on trapped ion quantum computing with the assumptions for the time scales for quantum operation primitives summarized in Table I. The details of fault-tolerant implementation of universal gate set utilized in this analysis is summarized in Appendix A.

B. Construction of efficient arithmetic circuits

The example quantum circuit we analyze is an adder circuit that computes the sum of two n-bit numbers. Simple adder circuits form the basis of more complex arithmetic circuits, such as the modular exponentiation circuit that dominates the execution time of Shor's factoring algorithm [65]. Quantum adder circuits can be constructed using X, CNOT, and Toffoli gates. When only local interactions are available without dedicated buses for entanglement distribution, a quantum ripple-carry adder (QRCA) is the adequate adder of choice [66], for which the execution time goes as $O(n)$. For QLA and MUSIQC architectures, one can implement quantum carry-lookahead adder (QCLA) that is capable of reducing the runtime to $O(\log n)$ [67,68], at the expense of extra qubits and parallel operations. QCLA dramatically outperforms the QRCA for n above $\sim$100 in terms of execution time. Practical implementation of large-scale QCLAs are hindered by the requirement of executing Toffoli gates among qubits that are separated by long distances within the quantum computer. MUSIQC architecture flattens the communication cost between qubits in different ELUs, providing a suitable platform for implementing QCLAs. Alternatively, QLA architecture can also efficiently execute QCLAs using dedicated

communication bus that reduces the connection time between two qubits (defined as the time it takes to generate entangled qubit pairs that can be used to teleport one of the qubits or the gate itself) to increase only as a logarithmic function of the separation between them [58].

C. MUSIQC implementation

In order to implement the QCLA circuit in MUSIQC architecture, each ELU should be large enough to accommodate the generation of the $|\phi_+\rangle_L$ state shown in Fig. 9(a). This requires a minimum of three logical qubits and a seven-qubit cat state, and sufficient ancilla qubits to support the state preparation. We balance the qubit resource requirements with computation time by requiring four ancilla qubits per logical qubit, so that the four-qubit cat states necessary for the stabilizer measurement can be created in parallel. Implementation of each Toffoli gate is realized by allocating a fresh ELU and preparing the $|\phi_+\rangle_L$ state, then teleporting the three qubits from other ELUs into this state. Once the gate is performed, the original logical qubits from the other ELUs are freed up and become available for another Toffoli gate. We find that $6n$ logical qubits placed on $6n/4 = 1.5n$ ELUs is sufficient to compute the sum of two n-bit integers using the QCLA circuit at the first concatenation level of Steane code encoding.

Teleportation of qubits into the ELU containing the prepared $|\phi_+\rangle_L$ state requires generation of entangled states via photon interference. In order to minimize the entanglement generation time, one should provide at least three optical ports to connect to these ELUs in parallel. In order to successfully teleport the gate, we need to create seven entangled pairs to each ELU holding the input qubits. The entanglement generation time can be reduced by running multiple optical ports to other ELUs in parallel (we call this the *port multiplexity* m_p). In a typical entanglement generation procedure, the ion is prepared in an initial state, and then excited using a short pulse laser (∼5 ps). The ion emits a photon over a spontaneous emission lifetime (∼10 ns), and the photon detection process will determine whether the entanglement generation from a pair of such ions is successful. If the entanglement generation is successful, the pair is ready for use in the computation. If not, the ions will be re-initialized (∼1 μs) and the process is repeated. Since the initialization time of the ion is ∼100 times longer than the time a photon is propagating in the optical port, one can utilize multiple ions per optical port and "pipeline" the photon emission process. In this time-division-multiplex (TDM) scheme, another ion is brought into the optical port to make another entanglement generation attempt while the initialization process is proceeding for the unsuccessful ion. This process can be repeated m_T times using as many extra ions, before the first ion can be brought back (we call m_T the *TDM multiplexity*). Using the port and TDM multiplexity, we can reduce the entanglement generation time by a factor of $m_p m_T$.

In our example, we assume multiplexities $m_p = 2$ and $m_T = 10$ that require 100 qubits ($= 3 \times 7 + 3 \times 4 + 3 \times 2 \times 10$) and 12 parallel operations per ELU as shown in Fig. 4(a). This choice adequately speeds up the communication time between ELUs to balance out other operation times in the hardware. Multiple ELUs are connected by an optical switch

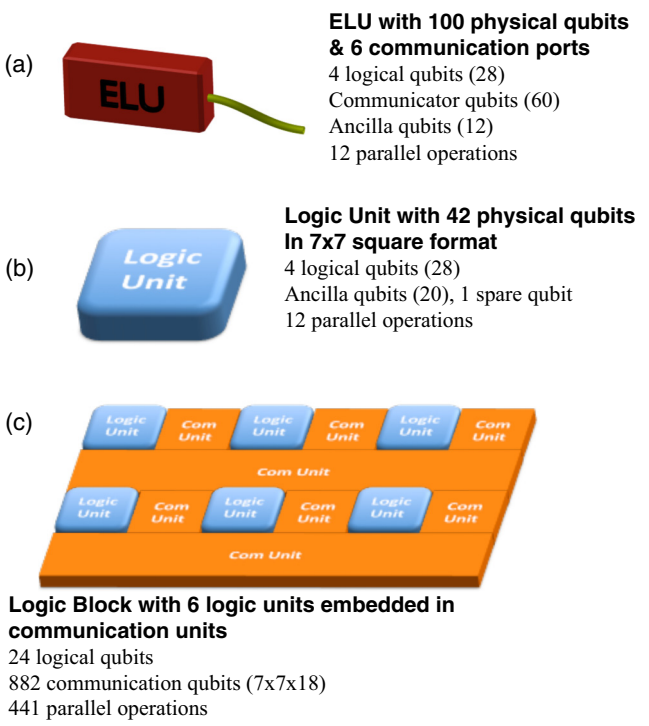

FIG. 4. (Color online) Example of the MUSIQC and QLA hardware considered. (a) Each ELU in MUSIQC is made up of 100 physical qubits and six communication ports (only one shown in the figure), where 60 qubits are used to increase the bandwidth of the remote entanglement generation. These ELUs are connected through an OXC switch as shown in Fig. 1. (b) For QLA, each logic unit is made up of 49 physical qubits hosting four logical qubits and necessary ancilla qubits. (c) A logic block is six such logic units embedded in communication units. Communication units are square arrangements of 7×7 qubits, and eight such units fully surround the logic unit.

to complete the MUSIQC hardware [Fig. 1(b)]. With these resources, an efficient implementation of the QCLA circuit can be realized by executing all necessary logic gates in parallel. Under these circumstances, the depth of the n-bit in-place adder circuit is given by [67]

$$\lfloor \log_2 n \rfloor + \lfloor \log_2 (n-1) \rfloor + \left\lfloor \log_2 \frac{n}{3} \right\rfloor + \left\lfloor \log_2 \frac{n-1}{3} \right\rfloor + 14, \tag{1}$$

for sufficiently large n ($n > 6$) where $\lfloor x \rfloor$ denotes the largest integer not greater than x. Out of these, two time steps contain X gates, four contain CNOT gates, and the rest contain Toffoli gates which dominate the execution time of the circuit. We assume an error correction step is performed on all qubits after each time step, by measuring all stabilizers of the Steane code and making necessary corrections based on the measurement outcome.

Once the basic operational primitives outlined in the previous section are modeled at the first level of code concatenation, we can construct all of these primitives at the second level of concatenation using the primitives at the first level. We can recursively construct the primitives at higher levels of code concatenation. Since the cost of remote CNOT gates

between ELUs are independent of the distance between them, recursive estimation of circuit execution at higher levels of code concatenation is straightforward on MUSIQC hardware.

D. QLA implementation

We consider a concrete layout of a QLA device optimized for n-bit adder with one level of Steane [7, 1,3] encoding, which can be used to construct circuits at higher levels of code concatenation. In order to implement the fault-tolerant Toffoli gate described in Fig. 9, one should assemble four logical qubits into a single tight unit, as we did for the ELUs in the MUSIQC architecture. In the QLA implementation, a "Logic Unit (LU)" consists of a square of 49 ($= 7 \times 7$) qubits, where a block of 12 ($= 3 \times 4$) qubits form a logical qubit with seven physical qubits and five ancilla qubits [Fig. 4(b)]. Just like in the MUSIQC example, $6n$ logical qubits placed on $1.5n$ LUs are necessary for adding two n-bit numbers. Therefore, we organize six LUs into a logical block (LB), capable of adding two 4-bit numbers. Each LU in the LB is surrounded by eight blocks of 7×7 communication units dedicated for distributing entanglement using the quantum repeater protocol [Fig. 4(c)]. We assume that the communication of the qubits within each LU is "free," and do not consider the time it takes for such communication. This simplified assumption is justified as the communication time between LUs utilizing the qubits in the communication units dominates the computation time, and therefore does not change the qualitative conclusion of this estimate.

Similar to the MUSIQC hardware example, a Toffoli gate execution involves the preparation of the $|\phi_+\rangle_L$ state in an "empty" LU, then teleporting three qubits onto this LU to complete the gate operation. The execution time of the Toffoli gate therefore comprises the time it takes to prepare the $|\phi_+\rangle_L$ state, the time it takes to distribute entanglement between adequate pairs of LUs, and then utilizing the distributed entanglement to teleport the gate operation. Among these, the distribution time for the entanglement is a function of the distance between the two LUs involved, while the other two are independent of the distance.

The QCLA circuit involves various stages of Toffoli gates characterized by the "distance" between qubits that goes as 2^t, where $1 \leqslant t \leqslant \lfloor \log_2 n \rfloor$ [67]. In a two-dimensional (2D) layout as considered in Fig. 4(c), the linear distance between these two qubits goes as $2^{t/2}$, in units of the number of communication units that the entanglement must be generated over. A slightly more careful analysis shows that the linear distance is approximately given by $d(t) \approx 3 \times 2^{t/2} + 1$ when t is even, and $d(t) \approx 2^{(t+1)/2} + 1$ when t is odd. Since each communication unit has seven qubits along a length, the actual teleportation distance is $L(t) = 7d(t)$ in units of the length of ion chain. The nested entanglement swapping protocol can create entanglement between the two end ions in $\lfloor \log_2 L(t) \rfloor$ time steps, where each time step consists of one CNOT gate, two single-qubit gates, and one qubit measurement process. Using the expression for $d(t)$, we approximate $\log_2 L(t) \approx t/2 + 4$ for both even and odd t, without loss of much accuracy. Unlike in the case of MUSIQC, the entanglement generation time is now dependent on the distance between the qubits (although only in a logarithmic way), and the resulting time steps needed for entanglement distribution within the QCLA is (approximately) given by

$$\lfloor \log_2 n \rfloor (\lfloor \log_2 n \rfloor + 17)/4 + \lfloor \log_2 (n-1) \rfloor (\lfloor \log_2 (n-1) \rfloor + 17)/4 + \left\lfloor \log_2 \frac{n}{3} \right\rfloor \left(\left\lfloor \log_2 \frac{n}{3} \right\rfloor + 17 \right) \Big/ 4 + \left\lfloor \log_2 \frac{n-1}{3} \right\rfloor \left(\left\lfloor \log_2 \frac{n-1}{3} \right\rfloor + 17 \right) \Big/ 4.$$

It should be noted that in order to achieve this logarithmic time, one has to have the ability to perform two qubit gates between every pair of qubits in the entire communication unit in parallel. The addition of two n-qubit numbers requires $n/4$ LBs. Since each LB has 18 communication units, there are a total of $7 \times 7 \times 18 = 882$ communication qubits in an LB. The number of parallel operations necessary is therefore 441 simultaneous CNOT operations per LB, or $441n/4 \approx 110n$ parallel operations for n-bit QCLA. The number of X, CNOT, and Toffoli gates that have to be performed remains identical to the MUSIQC case since we are executing an identical circuit. We assume that error correction is performed after every logic gate, but the entanglement distribution process has high enough fidelity so that no further distillation process is necessary.

Similar to the MUSIQC case, one can generate basic operational primitives at higher levels of code concatenation in the QLA model. Unlike the first encoding level, one may not have to explicitly provide communication channels for the second level of code concatenation if the quality of the distributed entanglement is sufficiently high so that neither entanglement purification [69] nor error correction of the entangled pairs [64] is needed. This type of "interlevel optimization" can be justified because the remote interaction between two logical qubits at the second level of code concatenation occurs very rarely, and the communication units at the first level can be used to accommodate this communication at higher level without significant time overhead. If dedicated communication qubits were provided in addition, these qubits might sit idle most of the time leading to inefficient use of the qubit resources. The number of physical qubits therefore scales much more favorably at higher levels of code concatenation than in the first level of the QLA architecture. The distance-dependent gate operation at higher levels of code concatenation is somewhat difficult to predict accurately, but the logarithmic scaling of communication time allows effective estimation of the gate operation time with only small errors.

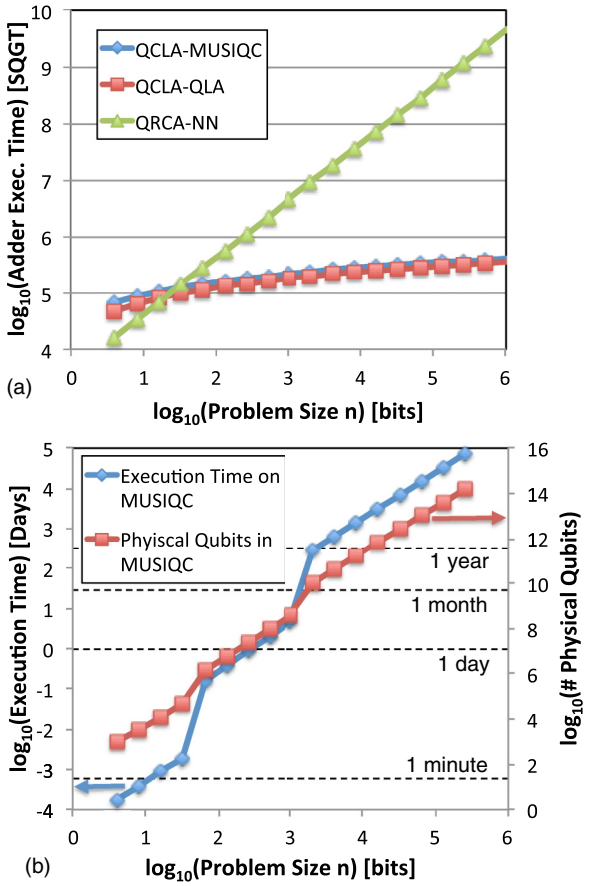

FIG. 5. (Color online) (a) Execution time comparison of quantum ripple-carry adder (QRCA) on a nearest-neighbor architecture (green triangles), and quantum carry-lookahead adder (QCLA) on QLA (red squares) and MUSIQC (blue diamonds) architectures, as a function of the problem size n. All three circuits considered are implemented fault tolerantly, using one level of Steane [7,1,3] code. The execution time is measured in units of single-qubit gate time (SQGT), assumed to be 1 μs in our model. (b) Execution time (blue diamonds, left axis) and number of required physical qubits (red squares, right axis) of running fault-tolerant modular exponentiation circuit, representative of executing the Shor algorithm.

E. Results and comparison

Figure 5(a) and Table II summarize the resource requirements and performance of the QCLA circuit on MUSIQC and QLA architecture, as well as the QRCA circuit on a nearest neighbor (NN) quantum hardware, where multiqubit gates can only operate on qubits sitting right next to one another. Although the QLA architecture considered in this example is also an NN hardware, presence of the dedicated communication units (quantum bus) allows remote gate operation with an execution time that depends only logarithmically on the distance between qubits, enabling fast execution of the QCLA. The cost in resources, however, is significant: Realization of efficient communication channels requires $\sim$3 times as many physical qubits as used for storing and manipulating the qubits in the first level of encoding, and requires a large number of parallel operations as well as the necessary control hardware to run them. The execution time can be fast compared to the MUSIQC architecture, which is hampered by the probabilistic nature of the photonic network in establishing the entanglement. We have dedicated substantial resources in MUSIQC to speed up the entanglement generation time as described in the previous section. Although MUSIQC architecture will take $\sim$15%–30% more time to execute the adder circuit, the resources it requires to operate the same task is only about 13% of that required in the QLA architecture. In both cases, we note the importance of moving qubits between different parts of a large quantum computer. The speed advantage in adder circuits translate directly to faster execution of the Shor algorithm, so we adopted QCLA for further analysis.

Once the execution time and resource requirements are identified for the adder circuit, one can adopt the analyses provided in Ref. [68] to estimate the performance metrics of running the Shor algorithm. The execution time and total number of physical qubits necessary to run the Shor algorithm depends strongly on the level of code concatenation required to successfully obtain the correct answer. We first estimate the number of logical qubits (Q) and the total number of logic gate operations (K) required to complete the Shor algorithm of a given size, to obtain the product KQ. In order to obtain correct results with a probability of order unity, the individual error rate corresponding to one logic gate operation must be on the order of $1/KQ$ [58]. From this consideration, we determine the level of code concatenation to be used. Table III summarizes the comparison on the number of physical qubits and the execution time of running the Shor algorithm on MUSIQC and QLA architectures for factoring 32, 512, and 4096 bit numbers [59].

TABLE II. Summary of the resource estimation and execution times of various adders in MUSIQC and QLA architecture.

Performance metrics	QCLA on MUSIQC	QCLA on QLA	QRCA on NN
Physical qubits	$150n$	$1176n$	$20(n+1)$
No. of parallel operations	$18n$	$110n$	$8n+43$
Logical Toffoli (μs)	3250	2327[a]	2159
128-bit addition	0.16 s	0.13 s	0.56 s
1024-bit addition	0.22 s	0.18 s	4.5 s
16 384-bit addition	0.29 s	0.25 s	72 s

[a]Does not include entanglement distribution time.

TABLE III. Estimated execution time and physical qubits necessary to complete Shor algorithm of a given size. The numbers on top (bottom) correspond to MUSIQC (QLA) architecture.

Performance metrics		$n=32$	$n=512$	$n=4096$
Code level		1	2	3
No. of physical qubits	MUSIQC	4.7×10^4	9.2×10^7	4.1×10^{10}
	QLA	3.7×10^5	7.2×10^8	3.2×10^{11}
Execution time	MUSIQC	2.5 min	2.1 days	650 days
	QLA	2.2 min	1.5 days	520 days

Figure 5(b) shows the execution time (in days) and the total number of necessary physical qubits for completing the modular exponentiation circuit on a MUSIQC hardware, which is a good representation of running the Shor algorithm. The discrete jumps in the resource estimate correspond to addition of another level of code concatenation, necessary for maintaining the error rates low enough to obtain a correct result as the problem size increases. Using 2 levels of concatenated Steane code, we expect to be able to factor a 128-bit integer in less than 10 h, with less than 6×10^6 physical qubits in the MUSIQC system. The execution time on QLA architecture is comparable to that on MUSIQC architecture (within 20%), but the number of required physical qubits is higher by about a factor of 10. Furthermore, the total size of the single ELU necessary to implement the QLA architecture grows very quickly (over 4.5×10^7 physical qubits for a machine that can factor a 128-bit number), while the ELU size in MUSIQC architecture is fixed at moderate numbers ($\approx$58 000 ELUs with 100 qubits per ELU). Therefore, although still daunting, the MUSIQC architecture substantially lowers the practical technological barrier in integration levels necessary for a large-scale quantum computer.

IV. FAULT TOLERANCE OF PROBABILISTIC PHOTONIC GATES

In the previous section, we examined how the MUSIQC architecture could be used to perform algorithms in the limit of large ELUs and low errors using Steand code. In this section, we turn our attention to the following fundamental question: Given a finite coherence time, how slow can the creation of entanglement be to still allow for fault tolerance? In this context, it is adequate to consider a MUSIQC system where many small ELUs are connected through the photonic network. Here, we focus on the demonstration of fault-tolerant circuit construction on MUSIQC architecture, rather than quantitative analysis of the resource overhead for these schemes.

Naïvely, it would appear that the average entanglement creation time τ_E must be much smaller than the decoherence time scale τ_D to achieve fault tolerance, but we find that scalable fault-tolerant quantum computation is possible for *any* ratio τ_E/τ_D, even in the presence of additional gate errors. While large values of τ_E/τ_D would lead to impractical levels of overhead in qubits and time (similar to the case of conventional quantum fault tolerance near threshold error levels [70]), this result is still remarkable and indicates that fault tolerance is always possible in the MUSIQC architecture. In this section, we provide a complete description of the strategies used to secure fault tolerance in MUSIQC architecture in this limit.

A. Analysis of fault tolerance for fast entangling gates

First, we consider the case where $\tau_E/\tau_D \ll 1$, where fault-tolerant coding is more practical. When each ELU is large enough to accommodate several logical qubits encoded with a conventional error correcting code, one can implement full fault-tolerant procedure within an ELU as in the example presented in the previous section. When the ELUs are too small to fit a logical qubit, fault tolerance can be achieved by mapping to three-dimensional (3D) cluster states, a known approach

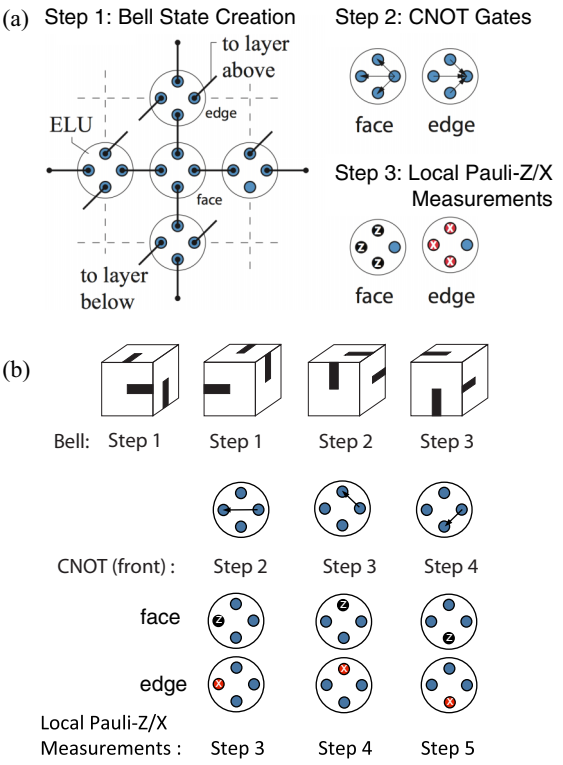

FIG. 6. (Color online) (a) Three steps of creating a 3D cluster state in the MUSIQC architecture, for fast entangling gates. (Step 1) Creation of Bell pairs between different ELUs, all in parallel. (Step 2) CNOT gates (head of arrow, target qubit; tail of arrow, control qubit). (Step 3) Measuring of three out of four qubits per ELU. If the ELU represents a face (edge) qubit in the underlying lattice, the measurements are in the Z-(X) basis. The resulting state is a 3D cluster state, up to Hadamard gates on the edge qubits. (b) Schedule for the creation of a 3D cluster state in the MUSIQC architecture. (Upper line) Schedule for Bell pair production between ELUs representing face and edge qubits. (Lower line) Schedule for the CNOT gates within the ELUs corresponding to the front faces of the lattice cell. Schedules for the ELUs on other faces and on edges are similar.

for supporting fault-tolerant universal quantum computation [71]. This type of encoding is well matched to the MUSIQC architecture, because the small degree of their interaction graph leads to small ELUs. A similar approach has recently been explored in Refs. [12,13].

Scheduling. For $\tau_E \ll \tau_D$, the 3D cluster state with qubits on the faces and edges of a three-dimensional lattice can be created using the procedure displayed in Fig. 6(a). The procedure consists of three basic steps: (1) creation of Bell states between different ELUs via the photonic link, (2) CNOT gates within each ELU, and (3) local measurement of three out of four qubits in each ELU. As can be easily shown using standard stabilizer arguments, the resulting state is a 3D cluster state, up to local Hadamard gates on the edge qubits.

The operations can be scheduled such that (a) qubits are never idle, and (b) no qubit is acted upon by multiple gates (even commuting ones) at the same time. The latter is required in some proposals for realizing quantum gates with ion qubits. To this end, the schedule [71] for 3D cluster state generation

is adapted to the MUSIQC architecture, and the three-step sequence shown in Fig. 6(a) is expanded into the five-step sequence shown in Fig. 6(b). In steps 1–3 the Bell pairs across the ELUs are created. In steps 2–4 the CNOTs within each ELU are performed, and in steps 3–5 three qubits in each ELU are measured. The sequence of operations is such that each of the three ancilla qubits in every ELU lives for only three time steps: initialization (to half of a Bell pair), CNOT, measurement. No qubit is ever idle in this protocol.

What remains to complete the computation is the local measurement of the 3D cluster state [71]. All remaining measurements are performed in step 5 of the above procedure. This works trivially for cluster qubits intended for topological error correction or the implementation of topologically protected encoded Clifford gates [72], since these measurements require no adjustment of the measurement basis. To avoid delay in the measurement of qubits for the implementation of non-Clifford gates, it is necessary to break the 3D cluster states into overlapping slabs of bounded thickness [71].

Fault-tolerance threshold. We assume the following error model. (1) Every gate operation, i.e., preparation and measurement of individual qubits, gates within an ELU, and Bell pair creation between different ELUs, can all be achieved within a clock cycle of duration T. An erroneous one-qubit (two-qubit) gate is modeled by the perfect gate followed by a partially depolarizing one-qubit (two-qubit) channel. In the one-qubit channel, X, Y, and Z errors each occur with probability $\epsilon/3$. In the two-qubit channel, each of the 15 possible errors $X_1, X_2, X_1X_2, \ldots, Z_1Z_2$ occurs with a probability of $\epsilon/15$. All gates have the same error ϵ. (2) In addition, the effect of decoherence per time step T is described by local probabilistic Pauli errors X, Y, Z, each happening with a probability $T/3\tau_D$.

A criterion for the error threshold of measurement-based quantum computation with cluster states that has been established numerically for a variety of error models is

$$\langle K_{\partial q}\rangle(\{\text{error parameters}\}) = 0.70. \tag{2}$$

Therein, $K_{\partial q}$ is a cluster state stabilizer operator associated with the boundary of a single volume q, consisting of six faces. Let f be a face of the three-dimensional cluster, and $K_f = \sigma_x^{(f)} \bigotimes_{e\in\partial f} \sigma_z^{(e)}$ as shown in Fig. 7(a). Then, $K_{\partial q} = \prod_{f\in\partial q} K_f = \bigotimes_{f\in\partial q} \sigma_x^{(f)}$. Furthermore, for the above criterion to apply, all errors—for preparation of local states, local and entangling unitaries, and measurement—are propagated forward or backward in time, to solely affect the 3D cluster state.

The above criterion applies for a phenomenological error model with local memory error and measurement error (the threshold error probability per memory step and measurement is 2.9% [73]), for a gate-based error model (the threshold error probability per gate is 0.67% [71]), and further error models with only low-order correlated error. Specifically, the criterion (2) has numerically been tested for cluster state creation procedures with varying relative strength of local vs 2-local gate error [71], with excellent agreement. In all cases, the error correction was performed using Edmonds' perfect matching algorithm.

The detailed procedure for calculating the error probability of the stabilizer measurement process for the 3D cluster state

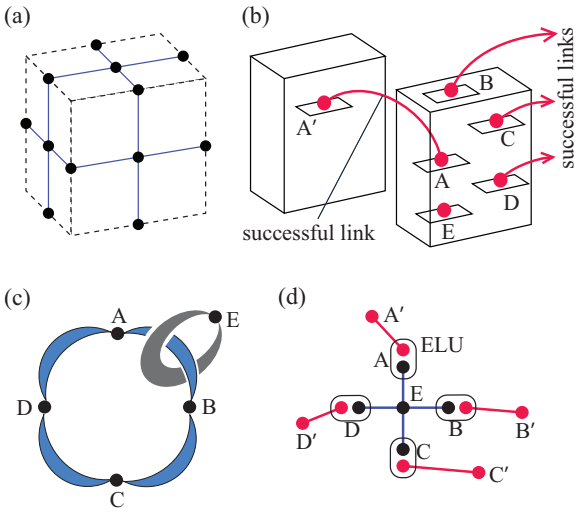

FIG. 7. (Color online) Hypercell construction II. (a) Lattice cell of a three-dimensional four-valent cluster state. The dashed lines represent the edges of the elementary cell and the solid lines represent the edges of the connectivity graph. The three-dimensional cluster state is obtained by repeating this elementary cell in all three spatial directions. (b) Creating probabilistic links between several 3D cluster states. (c) Reduction of a 3D cluster state to a five-qubit graph state, via Pauli measurements. The shaded regions represent measurements of Z; the blank regions represent measurements of X. The qubits represented as black dots remain unmeasured. For details, see [71]. (d) Linking graph states by Bell measurements in the remaining ELUs. Four-valent, 3D cluster states of arbitrary size can be created.

is provided in Appendix B. In combination with criterion (2), we obtain the threshold condition:

$$\epsilon + \frac{55}{32}\frac{T}{\tau_D} < 2.9\times 10^{-3}. \tag{3}$$

Overhead. The operational cost of creating a 3D cluster state and then locally measuring it for the purpose of computation is 24 gates per elementary cell in the standard setting, and 54 gates per elementary cell in MUSIQC. Here the elementary cell of a 3D four-valent cluster state is shown in Fig. 7(b). The overhead of the MUSIQC architecture over fault-tolerant cluster state computation is thus constant. The operational overhead for fault tolerance in the latter is poly-logarithmic [71], as described in detail in Ref. [72].

B. Analysis of fault tolerance for slow entangling gates

The above construction fails for $\tau_E/\tau_D \geqslant 1$, where decoherence occurs while waiting for Bell-pair entanglement. However, scalable fault-tolerant computing can still be achieved in the MUSIQC architecture for *any* ratio τ_E/τ_D, even for ELUs of only three qubits. Compared to the case of $\tau_E \ll \tau_D$, the operational cost of fault tolerance is increased by a factor that depends strongly on τ_E/τ_D but is *independent* of the size of the computation. Thus, while quantum computation becomes more costly when $\tau_E \geqslant \tau_D$, it remains scalable. This surprising result shows that there is no hard threshold for the ratio τ_E/τ_D, and opens up the possibility for efficient fault-tolerant constructions with slow entangling gates. Here we show that

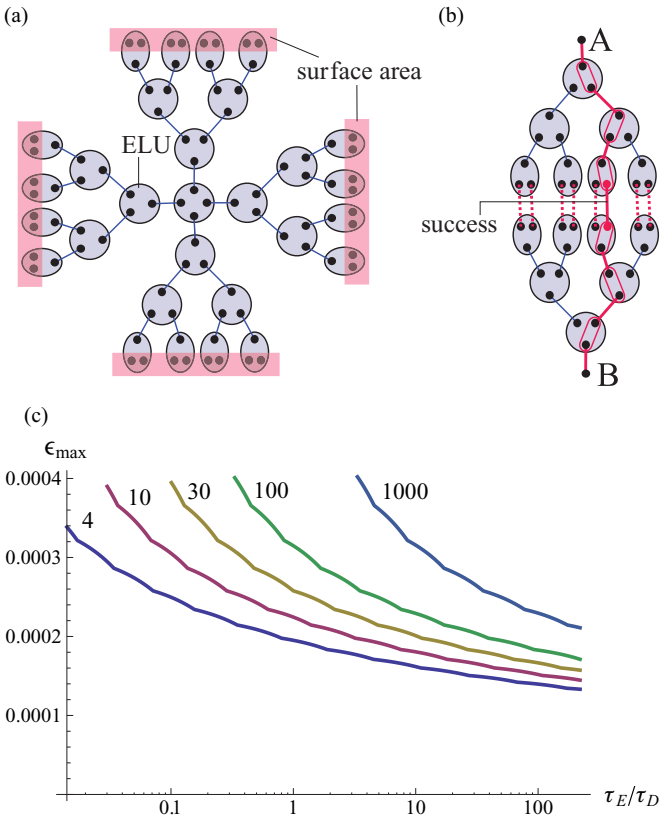

FIG. 8. (Color online) Hypercell construction I. (a) Snowflake design of Refs. [74,75]. (b) Connecting two hypercells. If the surface area is large, with high probability one or more Bell pairs are created between the surface areas via the photonic link. By Bell measurements within individual ELUs (indicated by ovals) one such Bell pair is teleported to the roots A and B. (c) Boundary of the fault-tolerance region for gate error ϵ and ratio τ_E/τ_D, for various ELU sizes. The threshold for the gate error ϵ depends only weakly on τ_E/τ_D.

scalable quantum computation can be achieved for arbitrarily slow entangling gates.

The main idea is to construct a "hypercell" out of several ELUs. A hypercell has the same storage capacity for quantum information as a single ELU, but with the ability to become (close to) deterministically entangled with four other hypercells. Fault-tolerant universal quantum computation can then be achieved by mapping to a four-valent, three-dimensional cluster state [71]. First, we show that arbitrarily large ratios τ_E/τ_D can be tolerated in the limiting case where the gate error rate $\epsilon = 0$ (construction I). Then, we show how to tolerate arbitrarily large ratios τ_E/τ_D with finite gate errors $\epsilon > 0$ (construction II).

Hypercell construction I is based on the snowflake design [74,75], as shown in Fig. 8(a). The difference is that in the present case, each node in the connectivity tree represents an entire ELU, not a single qubit as in Refs. [74,75]. At the root of the tree is an ELU that contains the qubit used in the computation, while multiple layers of bifurcating branches lead to a large "surface area" with many ports from which entanglement generation between two trees can be attempted. Once a Bell pair is created, it can be converted to a Bell pair between the root qubits A and B via teleportation as shown in Fig. 8(b).

The links (each representing a Bell pair) within a snowflake structure are created probabilistically, each with a probability p of heralded success. The success probability of each hypercell is small, but if the surface area between two neighboring hypercells is large enough, the probability of creating a Bell pair between them via a probabilistic photonic link approaches unity. Thus, the cost of entangling an entire grid of hypercells is linear in the size of the computation, as opposed to the exponential dependence that would be expected if the hypercells could not be entangled deterministically. Correspondingly, the operational cost of creating a hypercell is large, but the cost of linking this qubit into the grid is independent of the size of the computation. The hypercell offers a qubit which can be near-deterministically entangled with a constant number of other qubits *on demand*. A quantum computer made up of such hypercells can create a four-valent, 3D cluster state with few missing qubits, and is thus fault tolerant [71,76].[1] Hypercells can readily be implemented in the modular ion trap quantum computer since the probability of entanglement generation does not depend on the physical distance between the ELUs.

We call the part of the hypercell needed to connect to a neighboring hypercell a "tree." For ELUs of coordination number 3, the number m of ports that are available to connect two hypercells is twice the number of ELUs in the top layer of the tree. The probability for all m attempts to generate entanglement between two trees to fail is $P_{\text{fail}} = (1-p)^m \approx \exp(-mp)$. (In practice, we will allow a constant probability of failure which is tolerable in 3D cluster states [76]). In addition, the number of ELUs in the top layer is $2^{\#\text{layers}}$, and the path length l (number of Bell pairs between the roots) is $l = 2\log_2 m + 1$. Combining the above, we find that $l = 2\log_2 \frac{c}{p} + 1$, for $c = -\ln P_{\text{fail}}$. For simplification we assume that the time t for attempting entanglement generation is the same when creating the trees and when connecting the trees. Then, $p = t/\tau_E$ in both cases. From the beginning of the creation of the trees to completion of entangling two trees, a time $2t$ has passed. The Bell pairs within the trees have been around, on average, for a time $3t/2$, and the Bell pairs between the two trees for an average time of $t/2$. If overall error probabilities remain small, the total probability of error for creating a Bell pair is proportional to l. The memory error alone is

$$\epsilon_{\text{mem}} = \frac{t}{\tau_D}\left[3\log_2\left(c\frac{\tau_E}{t}\right) + \frac{1}{2}\right]. \tag{4}$$

This function is monotonically increasing with t, and $\epsilon_{\text{mem}}(t = 0) = 0$. The task now is to suppress the memory error rate ϵ_{mem} below the error threshold ϵ_{crit} that applies to fault-tolerant quantum computation with 3D cluster states. From Eq. (2) we know that $\epsilon_{\text{crit}} > 0$.

From Eq. (4) we find that, for any ratio τ_E/τ_D, we can make t small enough such that $\epsilon_{\text{mem}} < \epsilon_{\text{crit}}$. The operational

[1] If ELUs of size $N_q = 3$ are used, resulting in hypercells of valency 3, then two such hypercells can be combined into one of valency 4.

cost for creating a hypercell with sufficiently many ports is $O(\text{hypercell}) \sim (\frac{1}{p})^{\frac{9/2c}{p}}$. This cost is high for small $p = t/\tau_E$, but independent of the size of the computation. Thus, whenever decoherence on waiting qubits is the only source of error, scalable fault-tolerant QC is possible for arbitrarily slow entangling gates.

We now discuss how the above hypercell construction I fares in the presence of additional gate error ϵ. We model every noisy one- (two-)qubit operation by the perfect operation followed by a SU(2)- [SU(4)-] invariant partial depolarizing channel with strength ϵ, the same as that used in Sec. IV A. If $\epsilon > 0$ then every entanglement swap adds error to the computation. We must swap entanglement in every ELU on the path between the roots A and B, and because there are $2\log_2 m$ of them ($m \geqslant 2$), for $\epsilon \ll 1$ the total error is

$$\epsilon_{\text{total}} = \frac{t}{\tau_D}\left[3\log_2\left(c\frac{\tau_E}{t}\right) + \frac{1}{2}\right] + 2\epsilon\log_2\left(c\frac{\tau_E}{t}\right). \quad (5)$$

Now it is no longer true that for any choice of τ_E/τ_D we can realize $\epsilon_{\text{crit}} > \epsilon_{\text{total}}$. A nonvanishing gate error sets an upper limit to the tree depth, because the accumulated gate error is proportional to the tree depth [Fig. 8(b)]. This implies an upper bound on the size of the top layer of the tree, which further implies a lower bound on the time t needed to attempt entangling the two trees [see Eq. (6) below] and thus a lower bound on the memory error caused by decoherence during the time interval t. The accumulated memory error alone may be above or below the error threshold, depending on the ratio τ_E/τ_D.

In more detail, suppose that $\epsilon_{\text{crit}} > \epsilon_{\text{total}}$ holds. Considering only gate errors, $\epsilon_{\text{crit}} > 2\epsilon\log_2(c\frac{\tau_E}{t})$, and hence,

$$t > c\tau_E 2^{-\frac{\epsilon_{\text{crit}}}{2\epsilon}}. \quad (6)$$

Now, recalling that $c\frac{\tau_E}{t} = m \geqslant 2$, with Eq. (5) we find that $\epsilon_{\text{crit}} > 3t/\tau_D + 2\epsilon$, or

$$t < \tfrac{1}{3}(\epsilon_{\text{crit}} - 2\epsilon)\tau_D. \quad (7)$$

The two conditions Eqs. (6) and (7) can be simultaneously obeyed only if

$$\frac{\tau_E}{\tau_D} < \frac{\epsilon_{\text{crit}} - 2\epsilon}{3c} 2^{\frac{\epsilon_{\text{crit}}}{2\epsilon}}. \quad (8)$$

We see that there is now an upper bound to the ratio τ_E/τ_D. Equation (8) is a necessary but not sufficient condition for fault-tolerant quantum computation using the hypercells of Fig. 8(b).

We have numerically simulated the process of constructing these hypercells for various values of the decoherence parameters ϵ and τ_E/τ_D. The boundary of the fault-tolerance region in the τ_E/τ_D,ϵ plane is shown in Fig. 8(c). In the above, for simplicity, we have considered hypercells in which all constituent ELUs are entangled in a single time step t. However, there are various possible refinements. (1) The computational overhead can be significantly decreased by creating the hypercell in stages, starting with the leaves of the trees and iteratively combining them to create the next layers [74]. (2) Using numerical simulations it was found that if each of the four trees making up a hypercell has coordination number 4 or 5 rather then 3 (i.e., a ternary tree instead of a binary tree), the overhead can be further reduced. These optimizations were used to produce Fig. 8(c).

Hypercell construction II allows fault tolerance for finite gate errors $\epsilon > 0$. In construction I, the accumulated error for creating a Bell pair between the roots A and B is linear in the path length l between A and B. This limits the path length l, and thereby the surface area of the hypercell. This limitation can be overcome by invoking 3D cluster states already at the level of creating the hypercell. Three-dimensional cluster states have an intrinsic capability for fault tolerance [71] related to quantum error correction with surface codes [77,78]. For hypercell construction II, we employ a 3D cluster state nested within another 3D cluster state. Therein, the "outer" cluster state is created near-deterministically from the hypercells. Its purpose is to ensure fault tolerance of the construction. The "inner" 3D cluster state is created probabilistically. Its purpose is to provide a means to connect distant qubits in such a way that the error of the operation does not grow with distance. Specifically, if the local error level is below the threshold for error correction with 3D cluster states, the error of (quasi-) deterministically creating a Bell pair between two root qubits A and B in distinct 3D cluster states is independent of the path length between A and B.

The construction is as follows. We start from a three-dimensional grid with ELUs on the edges and on the faces. Each ELU contains four qubits and can be linked to four neighboring ELUs. Such a grid of ELUs (of suitable size) is used to probabilistically create a 4-valent cluster state by probabilistic generation of Bell pairs between the ELUs, postselection and local operations within the ELUs.

After such cluster states have been successfully created, in each ELU three qubits are freed up, and can now be used for near-deterministic links between different 3D cluster states, as shown in Fig. 7(b). After four probabilistic links to other clusters have succeeded (the size of the cluster states is chosen such that this is a likely event), the cluster state is transformed into a star-shaped graph state via X and Z measurements [Fig. 7(c)]. This graph state contains five qubits, shared between the four ELUs at which the successful links start, and an additional ELU. Due to the topological error-correction capability of 3D cluster states, the conversion from the 3D cluster state to the star-shaped graph state is fault tolerant [71]. By further measurement in the ELUs, the graph states created in different hypercells can now be linked, e.g., to form again a 4-valent 3D cluster state which is a resource for fault-tolerant quantum computation [71], as shown in Fig. 7(d). This final linking step is prone to error. However, the error level is independent of the size of the hypercell, which was not the case for hypercell construction I.

The only error sources remaining after error correction in the 3D cluster stem from (i) the (two) ports per link, and (ii) the two root qubits A and B, which are not protected topologically. The total error ϵ_{total} of a Bell pair created between A and B in this case is given by $\epsilon_{\text{total}} = c_1 t/\tau_D + c_2\epsilon$, where t is the time spent attempting Bell pair generation, and c_1 and c_2 are algebraic constants which do not depend on the time scales τ_E and τ_D, and not on the distance between the root qubits A and B. Then, if the threshold error rate ϵ_{crit} for fault tolerance of the outer 3D cluster state is larger than $c_2\epsilon$, we can reach an overall error ϵ_{total} below the threshold value ϵ_{crit} by making t

sufficiently small. Smaller t requires larger inner 3D cluster states, but does not limit the success probability for linking construction II hypercells. Thus, fault tolerance is possible for all ratios τ_E/τ_D, even in the presence of small gate errors.

V. OUTLOOK

The success of silicon-based information processors in the past five decades hinged upon the scalability of integrated circuits (IC) technology characterized by Moore's law [79]. IC technology integrated all the components necessary to construct a functional circuit, using the same conceptual approach over many orders of magnitude in integration levels. The hierarchical modular ion trap quantum-computer architecture discussed here promises scalability, not only in the number of physical systems (trapped ions) that represent the qubits, but also in the entire control structure to manipulate each qubit at such integration levels.

The technology necessary to realize each and every component of the MUSIQC architecture described in Sec. II is already available, although the performance is still far from being able to realize the features discussed in Secs. III and IV. The recognition that ion traps can be mapped onto a two-dimensional surface that can be fabricated using standard silicon microfabrication technologies [17,20] has led to a rapid development in complex surface trap technology [21,22]. Present-day trap development exploits extensive electromagnetic simulation codes to design optimized trap structures and control voltages, allowing sufficient control and stability of ion positioning. Integration of optical components into such microfabricated traps will enable stronger interaction between the ions and photons for better photon collection and qubit detection [23] through the use of high numerical aperture optics or integration of an optical cavity with the ion trap [45]. Moreover, electro-optic and MEMS-based beam steering systems allows the addressing of individual atoms in a chain with tightly focused laser beams [80,81] and an optical interconnect network can be constructed using large-scale all-optical cross-connect switches [47]. While technical challenges such as the operation of narrow-band (typically ultraviolet) lasers or the presence of residual heating of ion motion [14] still remain, they do not appear to be fundamental roadblocks to scalability. Within the MUSIQC architecture we have access to a full suite of technologies to realize the ELU in a scalable manner, where the detailed parameters of the architecture such as the number of ions per ELU, the number of ELUs, or the number of photonic interfaces per ELU can be adapted to optimize performance of the quantum computer.

ACKNOWLEDGMENTS

We thank D. Bacon, M. Biercuk, B. Blinov, S. Flammia, D. L. Moehring, and R. E. Slusher for helpful discussions. This work was supported by the Intelligence Advanced Research Projects Activity, the Army Research Office MURI Program on Hybrid Quantum Optical Circuits, and the NSF Physics Frontier Center at JQI. L.M.D. acknowledges support from NBRPC (973 Program) Grant No. 2011CBA00300.

APPENDIX A: UNIVERSAL FAULT-TOLERANT QC USING STEANE CODE

We utilize the basic operational primitives of universal quantum computation using Steane [7,1,3] code [82] fully outlined in Ref. [24], summarized below.

(1) The preparation of logical qubit $|0\rangle_L$ is performed by measuring the six stabilizers of the code using four-qubit cat state $|\text{cat}\rangle_4 \equiv (|0000\rangle + |1111\rangle)/\sqrt{2}$, following the procedure that minimizes the use of ancilla qubits as outlined in Ref. [83]. The stabilizer measurement is performed up to three times to ensure that the error arising from the measurement process itself can be corrected. We perform a sequential measurement of the six stabilizers re-using the four ancilla qubits for each logical qubit, which reduces the number of physical qubits and parallel operations necessary for the state preparation at the expense of the execution time. Once all the stabilizers are measured, a three-qubit cat state is used to measure the logical Z_L operator to finalize qubit initialization process. This procedure requires 11 physical qubits to complete preparation of logical qubit $|0\rangle_L$.

(2) In Steane [7,1,3] code considered here, all operations in the Pauli group $\{X_L, Y_L, Z_L\}$ and the Clifford group $\{H_L, S_L, \text{CNOT}_L\}$ can be performed transversally (i.e., in a bitwise fashion). We assume seven parallel operations are available, so that these logical operations can be executed in one time step corresponding to the single- or two-qubit operation. The transversal CNOT_L considered here is between two qubits that are close by, so the operation can be performed locally without further need for qubit communication.

(3) In order to construct effective arithmetic circuits, we need Toffoli gate (a.k.a. CCNOT_L) which is not in the Clifford group. Since a transversal implementation of this gate is not possible in Steane code, fault-tolerant implementation requires preparation of a special three (logical)-qubit state,

$$|\phi_+\rangle_L = \tfrac{1}{2}(|000\rangle_L + |010\rangle_L + |100\rangle_L + |111\rangle_L), \qquad \text{(A1)}$$

and "teleport" the gate into this state [84]. This state can be prepared by measuring its stabilizer operator using a seven-qubit cat state on three logical qubits $|0\rangle_L$, as shown in Fig. 9(a). Successful preparation of this state requires a bitwise Toffoli gate (at the physical level), which we assume can only be performed locally among qubits that are close to one another. Once this state is prepared, the three qubits $|x\rangle_L$, $|y\rangle_L$, and $|z\rangle_L$ participating in the Toffoli gate can be teleported to execute the gate, as shown in Fig. 9(b). Therefore, a successful Toffoli gate operation requires three logical qubits (which in turn require extra ancilla qubits to initialize) and seven physical qubits as ancillary qubits, in addition to the three logical qubits on which the gate operates.

(4) When a CNOT gate is necessary between two qubits that are separated by large distances, we take the approach where the two qubits of a maximally entangled state is each distributed to the vicinity of the two qubits, and then the gate is teleported using the protocol proposed in Ref. [85]. Efficient distribution of the entangled states makes this approach much more effective than where the qubits themselves are transported directly.

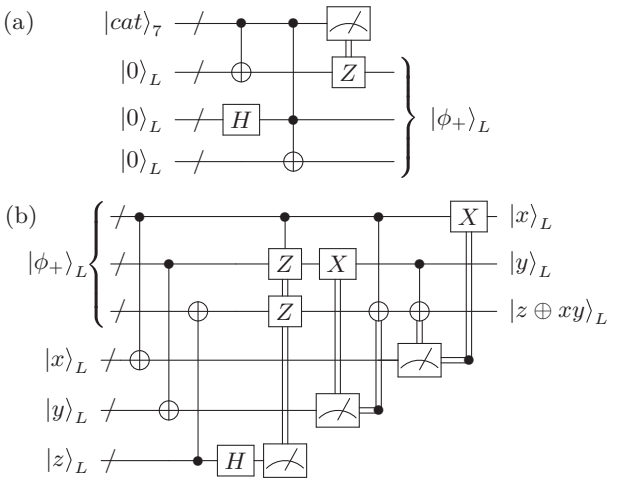

FIG. 9. Circuit diagram for realizing fault-tolerant Toffoli gate using Steane code. (a) The initial state $|\phi_+\rangle_L$ is prepared by measuring the X_1 and CNOT_{12} of three-qubit state $|0\rangle_1(|0\rangle_2 + |1\rangle_2)|0\rangle_3/\sqrt{2}$. Note that the Toffoli gate shown here is a bitwise Toffoli between the seven-qubit cat state and the two logical qubit states. (b) Using the state prepared in (a), the Toffoli gate can be realized using only measurement, Clifford group gates, and classical communication, all of which can be implemented fault tolerantly in the Steane code.

APPENDIX B: ERROR PROBABILITY FOR 3D CLUSTER STATES WITH FAST ENTANGLING GATES

Here we calculate the total error probability of the stabilizer measurement process for the model considered in Sec. IV A, assuming independent strengths for the local errors and 2-local gate errors. We have local errors with strength T/τ_D, and 2-local gate errors with strength ϵ. The expectation value of the stabilizer operator $K_{\partial q}$ in Eq. (2) is

$$\langle K_{\partial q}\rangle = \prod_{E\in\text{error sources}} 1 - 2p_E. \tag{B1}$$

Therein, p_E is the total probability of those Pauli errors in the error source E which, after (forward) propagation to the endpoint of the cluster state creation procedure, anticommute with the stabilizer operator $K_{\partial q}$. The right-hand side of Eq. (B1) is simply a product due to the statistical independence of the individual error sources. Since the cluster state creation procedure is of bounded temporal depth and built of local and nearest-neighbor gates only, errors can only propagate a finite distance. Therefore, only a finite number of error sources contribute in Eq. (B1).

To simplify the bookkeeping, we make the following observations. (a) A Bell state preparation, two CNOT gates (one on either side), and two local measurements on the qubits of the former Bell pair (one in the Z and one in the X basis) amount to a CNOT gate between remaining participating qubits. Therein, the qubit on the edge of the underlying lattice is the target; the qubit on the face is the control. We call this a teleported CNOT link. (b) Errors can only propagate once from face qubit to an edge qubit or vice versa, but never farther than that. To see this, consider, e.g., a face qubit. There, an X or Y error can get propagated (face = control of CNOTs). In either case it causes an X error on a neighboring edge qubit. But X errors are not propagated from edge qubits (edge = target of all CNOTs). (c) The stabilizer $K_{\partial q}$ has only support on face qubits, and is not affected by X errors.

Based on these observations, we subdivide the error sources affecting $\langle K_{\partial q}\rangle$ into three categories, namely, type 1, first Bell pair created on each face (according to the five-step schedule); type 2, the CNOT links, consuming the remaining Bell pairs; and type 3, the final measurements of the cluster qubits (1 per ELU).

Type-2 contributions. For every CNOT link we only need to count Z errors (and $Y \cong Z$) on both the control (= face) and target (= edge), because on the face qubit the Z errors are the ones that matter [with (c)], and on the edge qubit, such errors may still propagate to a neighboring face qubit [with (b)] and matter there. With these simplifications, the effective error of each CNOT link between two neighboring ELUs is described by the probabilities p_{ZI} for a Z error on the face qubit, p_{IZ} for a Z error on the edge qubit, and p_{ZZ} for the combined error; and

$$p_{ZI} = 2\epsilon + \frac{10}{3}\frac{T}{\tau_D},\ p_{IZ} = p_{ZZ} = \frac{4}{15}\epsilon + \frac{2}{3}\frac{T}{\tau_D}. \tag{B2}$$

Herein, we have only kept contributions up to linear order in $\epsilon, T/\tau_D$. The contributions to the error come from (1) the Bell pair, (2) a first round of memory error on all qubits, (3) the CNOT gates, (4) a second round of memory error on all qubits, and (5) the two local measurements per link.

Now we need to discuss the effect of each of the above gates on $\langle K_{\partial q}\rangle$, taking into account propagation effects. For example, consider the link established between the face qubit of a front face f with its left neighboring edge qubit. (The Bell pair for this link is created in step 1, the required CNOTs are performed in step 2, and the local measurements in step 3). The Z error on f does not propagate further. The Z error on e is propagated in later steps to a neighboring face [see Fig. 6(b)]. Thus, the errors Z_f and Z_e of this gate affect $\langle K_{\partial q}\rangle$, and Z_eZ_f doesn't. With Eq. (B1), the gate in question reduces $\langle K_{\partial q}\rangle$ by a factor of $1 - 68/15\,\epsilon - 8T/\tau_D$.

The following links contribute: three for every face in ∂q from within the cell, and three more per face of ∂q from the neighboring cells (links ending in an edge belonging to the cell q can affect $\langle K_{\partial q}\rangle$ by propagation). (i) Contributions from within the cell. If a Z_e error of the link propagates to an even (odd) number of neighboring faces in q, the total error probability affecting $\langle K_{\partial q}\rangle$ is $p_{ZZ} + p_{ZI}$ ($p_{IZ} + p_{ZI}$). But since $p_{IZ} = p_{ZZ}$, all 18 contributions from within the cell q are the same, irrespective of propagation. (ii) Contributions from neighboring cells. Each of the 18 links in question contributes an effective error probability $p_{IZ} + p_{ZZ}$ if an error on the edge qubit of the link propagates to an odd number of face qubits in ∂q. By inspection of Fig. 6(b), this happens for six links. With Eq. (B2), all the type-2 errors reduce $\langle K_{\partial q}\rangle$ by a factor of

$$1 - 160\frac{T}{\tau_D} - 88\epsilon. \tag{B3}$$

Type-1 contributions. Each of the initial Bell pair creations carries a two-qubit gate error of strength ϵ, and memory error of strength T/τ_D on either qubit. Similar to the above case, we can group the 15 possible Pauli errors into the equivalence classes I, Z_f ($Z_eZ_f \equiv I$ and $Z_e \equiv Z_f$ for Bell states). The

single remaining error probability, for Z_f, is

$$p_{ZI} = \frac{8}{15}\epsilon + \frac{4}{3}\frac{T}{\tau_D}. \tag{B4}$$

For each face of ∂q, there is one Bell pair within the face that reduces $\langle K_{\partial q}\rangle$ by a factor of $1 - 2p_{ZI}$. Bell pairs from neighboring cells do not contribute an error here. Thus, all the type-1 errors reduce $\langle K_{\partial q}\rangle$ by a factor of

$$1 - 8\frac{T}{\tau_D} - \frac{16}{5}\epsilon. \tag{B5}$$

Again, only the contributions to linear order in ϵ, T/τ_D were kept.

Type-3 contributions. The only remaining error source is in the measurement of the one qubit per ELU which is part of the 3D cluster state. The strength of the effective error on each face qubit is $p_Z = 2/3\,\epsilon$. Each of the six faces in ∂q is affected by this error. Thus, all the type-3 errors reduce $\langle K_{\partial q}\rangle$ by a factor of

$$1 - 8\epsilon. \tag{B6}$$

Combining the contributions Eq. (B3), (B5), and (B6) of error types 1–3 yields

$$\langle K_{\partial q}\rangle = 1 - \frac{512}{5}\epsilon - 176\frac{T}{\tau_D} \tag{B7}$$

for the expectation value $\langle K_{\partial q}\rangle$.

[1] T. D. Ladd *et al.*, Nature (London) **464**, 45 (2010).
[2] D. Wineland and R. Blatt, Nature (London) **453**, 1008 (2008).
[3] M. Neeley *et al.*, Nature (London) **467**, 570 (2010).
[4] L. DiCarlo *et al.*, Nature (London) **467**, 574 (2010).
[5] C. Monroe and J. Kim, Science **339**, 1164 (2013).
[6] D. D. Thaker, T. S. Metodi, and F. T. Chong, in *Proceedings of the High Performance Computing—HiPC 2006, 13th International Conference, Bangalore, India, December 18–21, 2006* (Springer, Heidelberg/New York, 2006), pp. 111–122.
[7] L. Jiang, J. M. Taylor, A. S. Sørensen, and M. D. Lukin, Phys. Rev. A **76**, 062323 (2007).
[8] D. L. Moehring *et al.*, J. Opt. Soc. Am. B **24**, 300 (2007).
[9] F. Helmer *et al.*, Europhys. Lett. **85**, 50007 (2009).
[10] N. Y. Yao *et al.*, Nat. Commun. **3**, 800 (2012).
[11] K. Fujii, T. Yamamoto, M. Koashi, and N. Imoto, arXiv:1202.6588.
[12] Y. Li and S. C. Benjamin, New J. Phys. **14**, 093008 (2012).
[13] N. H. Nickerson, Y. Li, and S. C. Benjamin, Nat. Commun. **4**, 1756 (2013).
[14] D. J. Wineland *et al.*, J. Res. Nat. Inst. Stand. Tech. **103**, 259 (1998).
[15] H. Häffner, C. Roos, and R. Blatt, Phys. Rep. **469**, 155 (2008).
[16] D. Kielpinski, C. Monroe, and D. Wineland, Nature (London) **417**, 709 (2002).
[17] J. Chiaverini *et al.*, Quantum Inf. Comput. **5**, 419 (2005).
[18] S. Seidelin *et al.*, Phys. Rev. Lett. **96**, 253003 (2006).
[19] S. X. Wang, J. Labaziewicz, Y. Ge, R. Shewmon, and I. L. Chuang, Phys. Rev. A **81**, 062332 (2010).
[20] J. Kim *et al.*, Quantum Inf. Comput. **5**, 515 (2005).
[21] D. L. Moehring *et al.*, New J. Phys. **13**, 075018 (2011).
[22] J. T. Merrill *et al.*, New J. Phys. **13**, 103005 (2011).
[23] J. Kim and C. Kim, Quantum Inf. Comput. **9**, 181 (2009).
[24] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).
[25] J. I. Cirac and P. Zoller, Phys. Rev. Lett. **74**, 4091 (1995).
[26] A. Sørensen and K. Mølmer, Phys. Rev. Lett. **82**, 1971 (1999).
[27] C. Ospelkaus *et al.*, Phys. Rev. Lett. **101**, 090502 (2008).
[28] S.-L. Zhu, C. Monroe, and L.-M. Duan, Europhys. Lett. **73**, 485 (2006).
[29] Q. A. Turchette *et al.*, Phys. Rev. A **61**, 063418 (2000).
[30] L. Deslauriers *et al.*, Phys. Rev. Lett. **97**, 103007 (2006).
[31] J. Labaziewicz *et al.*, Phys. Rev. Lett. **101**, 180602 (2008).
[32] G.-D. Lin and L.-M. Duan, New J. Phys. **13**, 075015 (2011).
[33] B. Blinov, D. L. Moehring, L.-M. Duan, and C. Monroe, Nature (London) **428**, 153 (2004).
[34] E. Togan *et al.*, Nature (London) **466**, 730 (2010).
[35] K. De Greve *et al.*, Nature (London) **491**, 421 (2012).
[36] W. B. Gao *et al.*, Nature (London) **491**, 426 (2012).
[37] C. Cabrillo, J. I. Cirac, P. García-Fernández, and P. Zoller, Phys. Rev. A **59**, 1025 (1999).
[38] L.-M. Duan and H. J. Kimble, Phys. Rev. Lett. **90**, 253601 (2003).
[39] C. Simon and W. T. M. Irvine, Phys. Rev. Lett. **91**, 110405 (2003).
[40] D. L. Moehring *et al.*, Nature (London) **449**, 68 (2007).
[41] L.-M. Duan and C. Monroe, Rev. Mod. Phys. **82**, 1209 (2010).
[42] L.-M. Duan, B. B. Blinov, D. L. Moehring, and C. Monroe, Quantum Inf. Comput. **4**, 165 (2004).
[43] S. C. Benjamin, D. E. Browne, J. Fitzsimons, and J. J. L. Morton, New J. Phys. **8**, 141 (2006).
[44] E. T. Campbell and S. C. Campbell, Phys. Rev. Lett. **101**, 130502 (2008).
[45] T. Kim, P. Maunz, and J. Kim, Phys. Rev. A **84**, 063423 (2011).
[46] P. O. Schmidt *et al.*, Science **309**, 749 (2005).
[47] J. Kim *et al.*, IEEE Photon. Technol. Lett. **15**, 1537 (2003).
[48] D. Neilson *et al.*, J. Lightwave Technol. **22**, 1499 (2004).
[49] S. Olmschenk *et al.*, Phys. Rev. A **76**, 052314 (2007).
[50] E. Mount *et al.*, New J. Phys. **15**, 093018 (2013).
[51] C. Langer *et al.*, Phys. Rev. Lett. **95**, 060502 (2005).
[52] D. M. Lucas *et al.*, arXiv:0710.4421.
[53] A. H. Myerson *et al.*, Phys. Rev. Lett. **100**, 200502 (2008).
[54] R. Noek *et al.*, Opt. Lett. **38**, 4735 (2013).
[55] K. R. Brown *et al.*, Phys. Rev. A **84**, 030303 (2011).
[56] D. Lucas (private communication, 2013).
[57] J. Benhelm, G. Kirchmair, C. F. Roos, and R. Blatt, Nature Physics **4**, 463 (2008).
[58] T. S. Metodi *et al.*, in *Proceedings of the 38th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'05)* (IEEE, Washington, DC, 2005), pp. 305–318.
[59] C. R. Clark, T. S. Metodi, S. D. Gasster, and K. R. Brown, Phys. Rev. A **79**, 062314 (2009).

[60] R. Raussendorf and H. J. Briegel, Phys. Rev. Lett. **86**, 5188 (2001).
[61] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, Phys. Rev. Lett. **81**, 5932 (1998).
[62] R. Beals *et al*., Proc. R. Soc. A **469**, 20120686 (2013).
[63] T. S. Metodi *et al*., ACM J. Emerg. Tech. Com. **4**, 1 (2008).
[64] L. Jiang *et al*., Phys. Rev. A **79**, 032325 (2009).
[65] V. Vedral, A. Barenco, and A. Ekert, Phys. Rev. A **54**, 147 (1996).
[66] S. A. Cuccaro, T. G. Draper, S. A. Kutin, and D. P. Moulton arXiv:cond-mat/0410184.
[67] T. G. Draper, S. A. Kutin, E. M. Rains, and K. M. Svore, Quantum Inf. Comput. **6**, 351 (2006).
[68] R. V. Meter and K. M. Itoh, Phys. Rev. A **71**, 052320 (2005).
[69] W. Dür, H.-J. Briegel, J. I. Cirac, and P. Zoller, Phys. Rev. A **59**, 169 (1999).
[70] E. Knill, Nature (London) **434**, 39 (2005).
[71] R. Raussendorf, J. Harrington, and K. Goyal, Ann. Phys. **321**, 2242 (2006).
[72] R. Raussendorf, J. Harrington, and K. Goyal, New J. Phys. **9**, 199 (2007).
[73] C. Wang, J. Harrington, and J. Preskill, Ann. Phys. **303**, 31 (2003).
[74] Y. Li, S. D. Barrett, T. M. Stace, and S. C. Benjamin, Phys. Rev. Lett. **105**, 250502 (2010).
[75] K. Fujii and Y. Tokunaga, Phys. Rev. Lett. **105**, 250503 (2010).
[76] S. D. Barrett and T. M. Stace, Phys. Rev. Lett. **105**, 200502 (2010).
[77] A. Kitaev, Ann. Phys. **303**, 2 (2003).
[78] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill, J. Math. Phys. **43**, 4452 (2002).
[79] G. E. Moore, Electronics **38**, 114 (1965).
[80] F. Schmidt-Kaler *et al*., Nature (London) **422**, 408 (2003).
[81] C. Knoernschild *et al*., Appl. Phys. Lett. **97**, 134101 (2010).
[82] A. M. Steane, Phys. Rev. Lett. **77**, 793 (1996).
[83] D. P. DiVincenzo and P. Aliferis, Phys. Rev. Lett. **98**, 020501 (2007).
[84] X. Zhou, D. W. Leung, and I. L. Chuang, Phys. Rev. A **62**, 052316 (2000).
[85] D. Gottesman and I. L. Chuang, Nature (London) **402**, 390 (1999).