Resource overheads and attainable rates for trapped-ion lattice surgery

Hudson Leone [1,2] Thinh Le [1] S. Srikara [1] and Simon Devitt [1,3,*]
[1]Centre for Quantum Software and Information, University of Technology Sydney, Sydney 2007, Australia
[2]Centre for quantum computation and Communication Technology (CQC2T)
[3]InstituteQ, Aalto University, 02150 Espoo, Finland

(Received 18 July 2024; accepted 6 December 2024; published 24 April 2025)

We present estimates for the number of ions needed to implement fault-tolerant lattice surgery between spatially separated trapped-ion surface codes. Additionally, we determine attainable lattice surgery rates given a number of dedicated "communication ions" per logical qubit. Because our analysis depends heavily on the rate that syndrome extraction cycles take place, we survey the state-of the art and propose three possible cycle times between 10 and 1000 μs that we could reasonably see realized provided certain technological milestones are met. Consequently, our numerical results indicate that hundreds of resource ions will be needed for lattice surgery in the slowest case, while close to a hundred thousand will be needed in the fastest case. The main factor contributing to these prohibitive estimates is the limited rate that ions can be coupled across traps. Our results indicate an urgent need for optical coupling to improve by one or more orders of magnitude for trapped-ion quantum computers to scale.

DOI: 10.1103/PhysRevResearch.7.023088

I. INTRODUCTION

Trapped ions are among the best studied and most technologically mature type of qubits to date; Their long coherence times and high-fidelity gates alone justify them as a candidate qubit for scalable quantum computing [1]. How a quantum computer is scaled will depend on its underlying architecture. Broadly speaking, an architecture may be monolithic [2] or modular [3]. A monolithic architecture is scaled by increasing the size of the chip, while a modular architecture is scaled by increasing the number of chips. Physical constraints and routing overheads generally cap the number of qubits that a monolithic architecture can reasonably support, which makes modularity something of an informal requirement when scaling quantum computers. Another requirement for scalability is error correction [4]. Industrial applications for quantum computing require programs to run for hours or even days. Since quantum operations introduce small amounts of error into the ion states, the computational qubits must therefore be encoded and periodically corrected for lengthy computations to succeed.

Based on these considerations, we expect that scalable quantum computers will be both modular and error corrected. One obvious disadvantage of the modular architecture is that two-qubit operations are not intrinsically possible between qubits that live in separate modules. Instead, some amount of entanglement has to be shared between modules as a resource for state or gate teleportation [5]. This process is complicated by the fact that entanglement distribution is probabilistic and noisy. To say that distribution is probabilistic means there is a significant chance of failure when attempting to entangle two physical qubits. For intermodular two qubit operations to be reliable, this means that a large number of communication qubits will be required as an overhead to ensure that enough entanglement is collected. On the other hand, to say that distribution is noisy means that errors are inadvertently introduced into the entangled states which must be corrected before any teleportation can take place. This correction is done through entanglement purification [6] which nondeterministically reduces a large ensemble of weakly entangled pairs into a smaller ensemble of strongly entangled pairs.

In this paper, we conduct resource analysis on the number of ions needed to implement reliable two qubit operations between logically encoded qubits in different modules at different speeds. Specifically, we limit our attention to the lattice-surgery operation between surface-code encoded qubits (see Fig. 1 for a simplified schematic). This analysis is significantly influenced by the speed at which lattice surgery can be performed. Faster operations are desirable of course, but myriad factors limit the attainable rate. To ground our analysis, we survey the state-of-the-art in trapped-ion technologies and synthesize this information to propose three possible surgery times together with a rubric of technological milestones that are necessary to achieve each one. Additionally, we determine the attainable rates for lattice surgery operations given a fixed number of ions.

*Contact author: Simon.Devitt@uts.edu.au

II. BACKGROUND

A. Surface code and lattice surgery

An $[[n, k, d]]$ stabilizer code is a quantum error correcting code that uses n so-called data qubits to encode k logical qubits with a code distance d. The code is specified

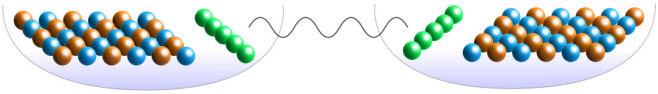

FIG. 1. A collection of trapped ions in two separate elementary logical units (ELUs). Some of the ions in each trap are used to encode a logical surface code qubit while other communication ions collect and refine entanglement to be used for a two-qubit lattice surgery operation.

(nonuniquely) by $n-k$ independent stabilizer generators that define the set of valid code-words. Each stabilizer generator is an observable that is implemented by measuring an ancilla qubit (also called a measure or syndrome qubit) that has been coupled to a number of data qubits. Assuming each stabilizer generator has its own measure qubit, the total number of physical qubits needed for an $[[n,k,d]]$ stabilizer code is $n_{\text{phys}} = n+(n-k) = 2n-k$. A stabilizer code cycle or syndrome extraction cycle is a measurement of all $n-k$ stabilizers (typically implemented in parallel) which results in a collection of measurement outcomes called syndromes ($s \in \{+1,-1\}^{n-k}$). When syndrome extraction cycles are repeated multiple times, changes in syndromes can be used to infer the most likely physical errors that have occurred in the code.

The surface code [7] is a popular choice of stabilizer code because of its high error tolerance (between 0.1% and 1% for unbiased noise [8] and up to 43.7% for biased noise [9]) and because it can be implemented using only nearest-neighbor qubit interactions. Crucially, (and unlike the majority of codes) it is also known how to perform universal quantum computation on surface code encoded qubits [10]. In its original implementation, the surface code is a $[[d^2+(d-1)^2,1,d]]$ stabilizer code. In this paper however, every mention of the surface code will refer to a similar (but slightly more efficient) variant called the rotated surface code which has the parameters $[[d^2,1,d]]$.

One of the necessary conditions for universality is the existence of an entangling gate. Several options exist for implementing two-qubit entangling gates between surface code qubits, though we limit our consideration to just two. A transversal gate is performed by coupling every physical qubit in one code to a counterpart in an adjacent code. These transversal operations are challenging to implement in two-dimensional architectures due to the numerous nonlocal interactions required [11]. An easier alternative is lattice surgery which, unlike transversal gates, can be implemented with only nearest neighbor interactions [12]. In brief, lattice surgery is executed by performing a syndrome extraction over two code patches as if they were one elongated patch (see Fig. 2). The disadvantage of lattice surgery is that, unlike a transversal gate, it requires measurements on a subset of the physical qubits that can introduce new, undetected errors into the ensemble. To correct for this, it is necessary to perform at least d rounds of lattice surgery to build confidence that the operation was done correctly. This makes lattice surgery slow compared to the transversal gate but is still the favored option on account of its feasibility.

An important, though unrelated, fact is that transversal gates and lattice surgery can both be performed on surface codes that are not directly adjacent in space by using shared entanglement. Specifically, maximally entangled qubit pairs can be used to teleport the required two-qubit gates in either case. For surface codes of distance d, transversal gates require a total of d^2 pairs since each qubit in a $d \times d$ lattice must be matched up with its counterpart in the other lattice. Coincidentally, lattice surgery also requires d^2 pairs since d pairs are needed for each of the d rounds that are necessary to account for measurement errors. Although these two operations require the same total amount of entanglement, the rate of required entanglement is much smaller for lattice surgery since only d pairs are required at each time step.

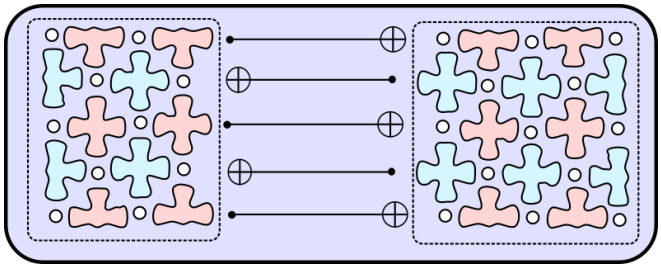

FIG. 2. An illustration of the lattice surgery operation between two surface code encoded qubits. The white circles represent data qubits which work together to encode the logical qubits in either box. The floral tiles represent syndrome extraction circuits which are executed in parallel to obtain information about errors that may have occurred on the data qubits. During a lattice surgery operation, two code patches undergo a syndrome extraction cycle together as if they were one elongated patch. For codes of distance d, lattice surgery requires d entangled pairs along the seam to implement the d CNOTs depicted. The effect of this operation is an XX parity check between the two encoded qubits which, with single qubit gates, is sufficient for universal quantum computing. For an in-depth treatment of lattice surgery see Refs. [12,14].

It was previously thought that the infidelities of the entangled pairs used in lattice surgery must at least match the error threshold of the code, however recent work by Ramette *et al.* indicates that pairs can tolerate an order of magnitude more error than expected [13]. The reason for this robustness stems from the fact that entangled pairs contribute errors along a single spatial dimension of the code during lattice surgery which limits the number of ways in which errors can form paths in the code.

B. Trapped-ion architectures

Monroe *et al.* [15] were among the first to perform a systematic study of a modular trapped-ion architecture, and some of our our terminology follows from their work. The modular trapped-ion computer is a collection of elementary logic units (ELUs) which serve as local processors and or memory banks. In the Monroe framework, each ELU is a linear Coulomb crystal containing an identical quantity of ions. Some of these are communication ions which are coupled to photonic interconnects and can be used to create entangled pairs between ELUs. Still another fraction of the ions are may be used for sympathetic cooling, which is when one ion is brought in proximity to another to absorb some of its vibrational energy.

An attractive advantage of the linear crystal is that it is all-to-all connected: No rearrangement is needed to implement a two-qubit gate between any pair of ions, though there's an effective limit to the crystal's length. Monroe *et al.* believed this limit to be around 100 ions which, ten years later, seems to be a well justified estimate since the longest linear crystal today is around 30 computational qubits [16]. Further corroborating evidence for this limit is discussed in greater detail by Murali *et al.* [17]. Later research from the same Munroe group indicates that their initial 100 ion estimate might have been too optimistic [18]. On the other hand, Ratcliffe *et al.* [19] propose a microtrap based architecture with all-to-all connectivity which they claim can be scaled beyond a linear trap. This is accomplished with a nonadiabatic gate which, at the cost of higher power, circumvents the need to use conventional side-band resolving gates which are slow and more susceptible to vibrational noise. A two dimensional extention to the microtrap architecture has also been studied in some detail [20].

In practice, we do not need to restrict ourselves by limiting the ELU to a single linear crystal. One possible workaround for further scaling is to have multiple crystals "chained together" in the same ELU to form a segmented linear trap [21]. The crystals in this chain can be separated, combined, and rotated to move a qubit in one crystal to any other crystal in the trap. Though the segmented linear trap has a larger qubit capacity, it lacks the all-to-all connectivity of a single crystal—meaning that some amount of physical routing is required for general computation. On average, the number of swap operations needed to move quantum information from one end of the chain to the other scales linearly with the number of modules and the number of qubits per segment. Monroe *et al.* believe more optimistically that a maximum of 1000 ions should be possible in an segmented linear trap [15]. Recent work indicates that stabilizing ions with optical tweezers could significantly improve the scalability of long ion crystals [22,23].

A still more general option for an ELU architecture is a quantum charged coupling device (QCCD) [21,24]. This is a two-dimensional configuration of ion chains which may be routed along a fixed network of corridors and junctions. Theoretically, QCCD architectures are monolithic, well connected, and faster than segmented linear traps, though they are difficult to fabricate. Presently, the largest QCCD is Quantinuum's 32 qubit race-track computer, which is a linear ion trap with periodic boundary conditions [25].

Various theoretic proposals exist for scaling beyond this current best: Malinowski *et al.* report on a QCCD architecture called WISE that addresses a crucial wiring problem and is capable of supporting around 1000 ions [26]. Valentini *et al.* developed the so-called quantum spring array (QSA) where no intertrap routing is required [27]. Sterk *et al.* propose an architecture that specifically addresses the problem of power dissipation when scaling QCCDs [28]. Mehta *et al.* suggest the use of planar-fabricated optics for further improvements [29]. For a survey on the technical challenges of scaling QCCDs, see Murali *et al.* [17]. True two- and three-dimensional ion crystals (also called Wigner crystals) may also be possible to engineer in the future on the scale of hundreds or perhaps even thousands of ions [30,31].

C. Trapped-ion surface code implementations

Specialized architectures for quantum error correction will likely be easier to realize in the near term since error correction is predictable, repetitive, and often nearest neighbor (as is the case for the surface code). Recent experimental efforts to build trapped-ion surface codes are encouraging, though limited in scope. Erhard *et al.* for example used a ten ion quantum computer to perform quantum state teleportation between surface codes of distance two [32] while Egan *et al.* implemented the closely related Bacon-Shor code [33] based in part on theoretical results from [34].

There is also a growing body of theoretical work around this objective of building trapped-ion surface codes. LeBlond *et al.* presented software for compiling surface code operations to trapped-ion hardware [35]. Trout *et al.* conducted extensive simulations of a distance 3 surface code implemented in a trapped-ion linear array [36], and reported a pseudothreshold of 3×10^{-3} with syndrome cycle times ranging from around 3 to 8 milliseconds. Similarly, Li *et al.* studied a surface code modeled on a segmented linear trap [37]. For segment lengths of around 15 qubits, they report an error tolerance of 0.12% but say nothing about syndrome extraction times. Lekitsch *et al.* [2] present a proof-of-concept for a monolithic surface code architecture that closely resembles a QCCD. A crucial difference however is that the Lekitsch architecture relies on global fields for a majority of the operations instead of individual lasers, which they argue is more feasible for monolithic scaling since it circumvents the need to align many optical elements with high precision. Their proposed cycle times are around 300 μs.

Although the focus of this work is on the surface code, we note that there is a strong interest for the so-called color code within the trapped-ion community [27,38,39]; This is a stabilizer code that is closely related to the surface code [40].

III. ESTIMATING SURFACE CODE CYCLE TIMES

The rate we are able to do lattice surgery is upper bounded by the rate at which syndrome extraction cycles can be performed. Some care is therefore required to establish reasonable estimates for the attainable cycle times in trapped-ion systems. Broadly speaking, there are three processes that need to be accounted for. The first is the entangling operations that are required to couple the syndrome qubits with the appropriate data qubits. Ions may or may not need to be routed for this depending on the underlying architecture. The second process is the measurement of the syndrome qubits. As we will soon see, this may require separating the ions a short distance from eachother in order to avoid measurement induced decoherence. The third and final process is ion-cooling, which is necessary to ensure high fidelity two qubit operations. Ion cooling may be a dedicated process during a surface-code cycle (after shuttling or measurement for example) or it may be continuous, as we will later discuss. In brief, the time required for a surface code cycle will depend both on the underlying choice of architecture, but also on various technical factors such as the speeds of single and two-qubit gates, measurements, ion-shuttling and cooling. We consider each of these factors in the following subsections and conclude our

review by establishing three cycle time paradigms we expect are feasible provided that specific technological milestones are met.

A. Trapped-ion gates

The review of Bruzewicz *et al.* presents a thorough comparison of single and two qubit trapped-ion gate times [1]. Typical single qubit gates with fidelities greater than surface code threshold are reported between 2 and 12 μs, though lower fidelity operations have been demonstrated on the order of nanoseconds. Two qubit gates are generally slower and lower fidelity. For the sake of argument, let us assume that we are willing to tolerate operational two-qubit error rates of up to 0.1%, which sits comfortably below the surface code threshold. Typical two-qubit gates with fidelities close to this rate are clocked between 1.6 μs and. It is not unreasonable to assume therefore that the time it takes to implement the gates in a surface code cycle is around 10 μs.

B. Trapped-ion measurements

Single-qubit trapped-ion measurements are commonly implemented via state-dependent fluorescence. In this method, laser light is directed at an ion which exclusively couples the $|1\rangle$ state to a 'cycling transition' that scatters numerous easily detectable photons. Likewise, the absence of photons indicates a measurement of the $|0\rangle$ state [1]. Although fluorescence measurements are relatively fast (on the order of 10 μs [41,42]) and high fidelity, the scattered photons (both from the laser and from the irradiated ion) are likely to decohere nearby qubits that aren't also being measured. This is a significant problem for error correcting circuits which all rely on mid-circuit measurements. Broadly speaking, there are two complementary strategies for mitigating measurement induced decoherence. The first is to incorporate techniques that suppress the decoherence, and the second is to move the ions some distance away to be measured safely. Both of these strategies will be discussed in the following subsections.

1. Techniques for suppressing decoherence

One way to limit measurement induced decoherence is to shorten the amount of time the qubit(s) are illuminated. This comes at the cost of measurement fidelity since there are fewer scattered photons to be detected. Naturally, faster and higher fidelity measurements will be possible with improvements in the photon collection rate and photodetector efficiency. See the introduction of Wolk *et al.* for a brief summary of techniques used to improve sparse detection fidelities [43].

Another approach for protecting against decoherence is to use quantum logic spectroscopy [44]. This is when the information of one qubit is transferred onto an ion of a different species that when measured emits off-resonant photons which are unlikely to disturb the states of neighboring qubits. An accidental benefit of this approach is that preexisting cooling ions may be used for this purpose. The disadvantages of quantum logic spectroscopy are that it is more difficult to maintain coherent control of multiple ion-species simultaneously, and that it requires additional ions (at most double for a one-to-one pairing).

A promising alternative might be to use ions of the same species but have the data and measurement qubits encoded in different energy levels of the ion [45,46]. An alternative strategy would be to suppress the decoherence effects altogether so that additional measurement ions are not required. Gaebler *et al.* [47] demonstrate a technique for reducing measurement cross-talk errors by an order of magnitude using tailored micromotion which may reduce and potentially eliminate the need for logic spectroscopy or shuttling altogether.

2. Ion-shuttling speeds

Perhaps the most intuitive strategy for mitigating measurement induced decoherence is to move the ions a safe distance away before measuring them. Ideally we would like to complete this operation as fast as possible, but faster shuttling introduces more thermal noise which can have a detrimental effect on two-qubit gates in particular. Broadly speaking, the infidelity of a Mølmer-Sørensen gate (Sec. III C) applied between two qubits in an ion-chain is known to depend on both the temperature of the chain and its displacement in phase space. These effects have been fully characterized for two different error metrics [48]. In the ion-shuttling literature, the amount of heat imparted in transport is commonly characterized in terms of how much the expected energy quanta of a particular motional mode increases. In the absence of noise, a linear crystal can withstand several quanta of phonons before there is an appreciable drop in the fidelity of a Mølmer Sørensen gate. As noise is introduced however, this tolerance drops [49]. The fastest and quietest reported shuttling operation at the time of writing is also from Sterk *et al.* who demonstrate a 210 μm one way ion transport in 6 μs with a maximum gain of 0.36 ± 0.08 quanta for an average speed of 35 μm μs^{-1} [50]. Slower, but more conservative routing was also demonstrated in 55 μs with a gain in 0.1 quanta [51].

3. Estimates for shuttling times

How far do ions need to be separated for the effects of measurement induced decoherence to be considered negligible? At the shorter end, Pino *et al.* report a QCCD architecture where a shuttling distance of 110 μm resulted in cross-talk errors between 3.5×10^{-3} and 1.5×10^{-2} [24]. Similarly, Crain *et al.* show that a separation of 370 μm results in cross-talk errors of 2×10^{-5} [41]. If we assume a distance of 300 μm is tolerable, then with the shuttling speed reported by Sterk *et al.*, we can assume that a two-way shuttling time of around 10 μs is sufficient for eliminating decoherence effects.

C. Cooling trapped ions

All gates and operations of trapped ion quantum computers require low temperatures, but this is especially true of the two qubit gates since they depend on vibrational coupling which is highly sensitive to noise. In the early days of trapped ion-quantum computing, two qubit gates required temperatures close to the ground state energy. The breakthrough discovery of Mølmer and Sørrensen [52] shifted this paradigm by introducing a gate that could operate at the Doppler temperature—the temperature regime attainable with

Doppler cooling. In the following sections, we present a brief review of the cooling techniques used to bring collections of ions to Doppler and sub-Doppler temperatures—endeavoring to report approximate cooling times wherever possible. Although cooling single ions is considerably easier than cooling ion crystals (since there are fewer motional modes to be addressed [53]), it is unlikely that single-ion cooling will be a leading technology in the context of quantum computation. This is because virtually all quantum architectures keep their computational ions organized in crystals.

At a high level, Doppler cooling works by shining a laser on an ion with a frequency just below what the ion will absorb. When the ion moves towards the laser, the incoming light is blue-shifted with respect to the ion which causes it to absorb a photon and slow down. Aside from Doppler cooling, other established laser based cooling techniques that operate under similar principles include resolved sideband cooling, Raman sideband cooling, and electromagnetically induced transparency cooling (EIT) with the fastest of these being EIT. Feng *et al.* report cooling a 40 ion chain to a near ground state energy in under 300 μs [54] while Jordan *et al.* reach similar temperatures for a 100 ion Penning trap within 200 μs. Some disadvantages of EIT are that it has a limited range of motional frequencies it can cool, and it is slow at cooling low frequency excitations [55].

Sympathetic cooling, where cold ions are brought into physical contact with computational ions, has been discussed in some detail in the previous sections. This is a well-established cooling method that remains a popular choice today—seeing use, for example, in the race-track architecture by Quantinuum [25]. A disadvantage of sympathetic cooling however is that it is relatively slow compared to other techniques and requires the use of additional ions [56]. An experimental demonstration showed that ion chains up to length 28 could be cooled to the global Doppler cooling limit using only two dedicated cooling ions of the same species [57]. This paper reported relaxation times (defined as the time required for the noise to settle within 5% of noise of the initial state) between 10 and 100 ms. Though these numbers are somewhat discouraging, one promising direction for further study is persistent cooling where a number of sympathetic cooling ions are brought in perpetual contact with computational ions. As the computation proceeds, the cooling ions are continuously chilled with Doppler cooling. This technique could lift the requirement for cooling processes that halt the computation. Lin *et al.* present an analysis of the dynamics of a linear array where a small subset of the ions are continuously cooled [58]. Additionally, a theoretic proposal for sympathetic cooling between one ion and a pre-cooled resource ion can be accomplished on the order of tens of microseconds, which may find some applicability in this context [59].

Rapid exchange cooling is a recently proposed alternative to sympathetic cooling that was suggested by Fallek *et al.* in the context of QCCD [60]. Here, coolant ions in a continuously chilled bank are shuttled to and from the computational ions. The authors of this work perform a proof-of-concept experiment in which two calcium ions are cooled with a round-trip shuttling time of 107.3 μs which, in their words, is "an order of magnitude faster than typical sympathetic cooling durations."

D. Cycle time paradigms

At the beginning of this Sec. III, we mentioned that estimating attainable cycle times is crucial to our resource estimation task. This is because the cycle time has a direct bearing on the number of communication ions we require per ELU; faster cycles mean that more ions are required to collect the necessary entanglement in a shorter period. In this section, we synthesize our findings from the previous review by proposing three cycle time paradigms (1000, 100, and 10 μs) that we could reasonably expect to see achieved for trapped-ion surface codes given various technical assumptions. For a short-hand summary of these paradigms see Table I. The first and slowest cycle time we propose is around 1000 μs. This is several times faster than what was simulated by Trout *et al.* [36] and around three times slower than the architecture proposed by Lekitsch *et al.* [2]. Additionally, Egan *et al.* present a distance three Bacon-Shor code where the X stabilizers are measured in approximately 3 ms, which is within one order of magnitude of 1000 μs [33]. This time scale permits some flexibility in routing and cooling options making it especially suitable for segmented-linear trap and QCCD architectures which require extensive use of both. Here, we expect one or more stages of cooling per cycle and shuttling to avoid measurement induced decoherence. The second time proposed is 100 μs. This is a more optimistic regime, being several times faster than the EIT cooling times reported in Sec. III C. Because of this, we require that fewer than one dedicated round of cooling is made per clock cycle. There is considerably less flexibility permitted for routing or shuttling at this scale. Dedicated zones for multiqubit measurement will likely become significant time and heat savers as will subquanta shuttling. Architectures that are likely to be viable in this paradigm include linear-traps, Wigner crystals, and QCCDs. It is likely as well that at least one strategy for mitigating measurement induced decoherence will be employed (Sec. III B 1). The final, and most optimistic regime is 10 μs. At this timescale, our clock cycle matches the two qubit gate times reported in Sec. III A meaning that no processes are allowed which interrupt syndrome extraction. Persistent cooling is an absolute necessity here, as is an architecture that doesn't require any routing or shuttling. Linear traps, Wigner crystals or microtrap based architectures are the most likely candidates for this regime.

E. Entangling ion pairs

Any modular architecture requires some means of communicating quantum information between the constituent ELUs. A common approach is to establish maximally entangled pairs of qubits between dedicated communication qubits to be used for quantum state teleportation. First, an ion is pulsed to create an entangled pair between an internal state of the ion and an emitted photon. Then, photons emitted from two different ELUs are routed together and fused with a polarization resolving Bell measurement that entangles the separated ions. The maximum theoretical rate at which ion-ion entanglement can be established is fixed by a constant called the photon scattering rate, which is around 100 MHz according to Stephenson *et al.* [61]. The same authors note however that entanglement rates are far lower in practice (up to several kHz though with

TABLE I. A summary of three cycle time paradigms for trapped-ion surface codes and the various technological milestones required for each speed.

Cycle time	System assumptions
1000 μs	This cycle time is comparable to theoretical proposals in literature [2,36], and is one order of magnitude from the syndrome extraction time of an experimentally realized distance 3 Bacon-Shor code [33]. (See Sec. III D for more context.)
100 μs	Less than one dedicated round of cooling per cycle. EIT cooling with subquanta shuttling required. Multiqubit measurement stages recommended. Will likely incorporate at least one suppression technique discussed in Sec. III B 1.
10 μs	Virtually no shuttling allowed. Persistent cooling that doesn't pause syndrome extraction cycle is essential. Multiple suppression techniques from Sec. III B 1 will likely be used together. Purification circuits are low-depth with one round of measurement.

low fidelities [62]) primarily because of low photon collection efficiencies [15]. The best ion coupling at the time of writing comes from the same paper cited previously, and reports 94% fidelity pairs at an average rate of 182 Hz and with a success probability of $p_c = 2.18 \times 10^{-4}$ per attempt.

IV. METHODOLOGY

A. The lattice surgery cycle

Given a surface code cycle time T, there are three steps that need to be completed within this T for lattice surgery to be implemented. The first is entanglement distribution, the second is entanglement purification (see Sec. IV D), and the third is the joint syndrome extraction. All of these processes are illustrated in Fig. 3. A natural strategy is to complete these steps sequentially. This means we divide our time window T into three parts which are then allocated to each process. The disadvantage of this method however is that we cannot make use of the full time window for any of the three steps.

The approach that we opt for instead is to implement these steps in parallel. In other words, the entanglement that we collect in one round is used for purification in the subsequent round which, in turn, is used for the surgery in the following round. Although this gives us the leeway to implement each step within the time frame T, there will necessarily be some dead time when first starting the lattice surgery cycle as the initial entanglement is collected and processed. We do not consider this dead-time in our analysis, but instead suppose that our lattice surgery "engine" is constantly running at a cycle time of T with no starts or stops.

B. A heuristic for fault-tolerant lattice surgery

Our primary objective for this work is to estimate the number of trapped-ions needed to perform fault-tolerant lattice surgery between surface codes at the rates specified in Table I. So far however, we have not addressed the question of what it means for lattice surgery to be fault-tolerant. Here, we formalize a definition by introducing a heuristic consisting of three criteria that must all be met for lattice surgery to be considered fault-tolerant. These conditions correspond to three subroutines of lattice surgery, which are depicted in Fig. 4 (note that these correspond to the steps of Fig. 3). Essentially, the point of this heuristic is to ensure that each of of these necessary subroutines succeeds with high probability.

The first criteria is the promise of a purification protocol that takes n Bell pairs of some initial fidelity F_{in} and, with some probability p, returns one pair at or above the required fidelity threshold F_{ideal}. The second condition is that each purification circuit is multiplexed (run in parallel copies) until the probability of getting at least one pair from the lot exceeds P_{pair}. The final condition is that we are able to collect enough entanglement within the cycle time T to implement d instances of the multiplexed purification protocol (one for each "stitch" of the lattice surgery). We stress that this collection must take place within the time T so that enough entanglement is acquired to be used for the next round of

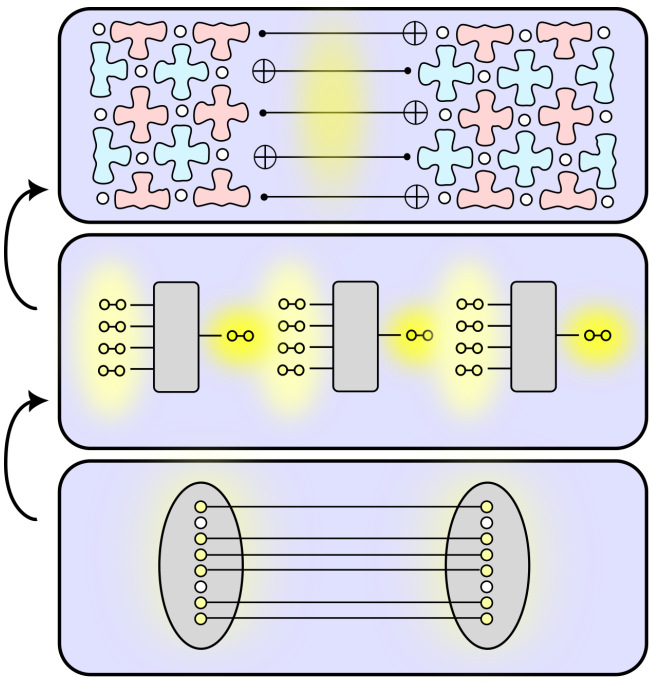

FIG. 3. Three concurrent stages for a single lattice surgery operation that each must succeed within the cycle time T. (Bottom) Entanglement is established between pairs of communication ions in separate ELUs. (Middle) Distributed pairs in storage ions are refined via entanglement purification. (Top) Refined pairs are used to teleport the mediating gates needed for a lattice surgery between two surface code patches.

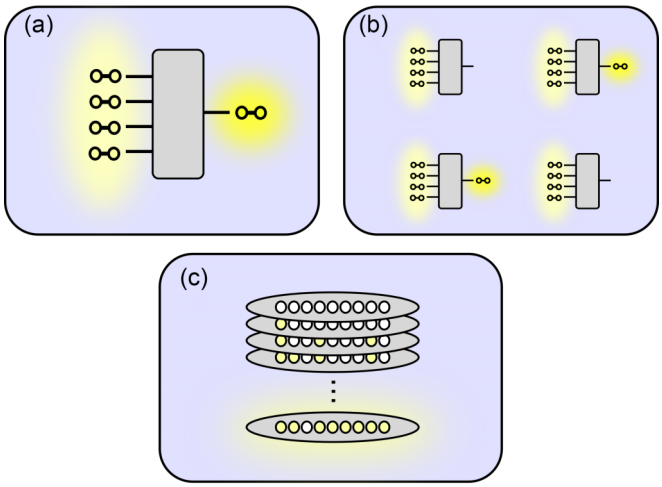

FIG. 4. Illustrations of the three conditions we require for our lattice surgery to be considered fault-tolerant. (a) We are given an $n \to 1$ purification protocol where the output pair meets or exceeds fidelity F_{ideal}. (b) We duplicate the circuit until the probability of getting at least one purified pair exceeds P_{pair}. (c) We have enough communication ions such that after after a collection time T, we are sure up to a confidence of P_{LS} that we have enough entangled implement D instances of the protocol described in (b).

lattice surgery. The probability that this collection succeeds must equal or exceed a threshold P_{LS}.

1. Minimum number of communication ions required given cycle time

Let p be the success probability of a purification circuit. The probability of obtaining at least one success out of n trials is $1-(1-p)^n$. Naturally the minimum number of purification circuits needed to produce at least one pair with a confidence of P_{pair} is then

$$K \equiv \min_n [1-(1-p)^n \geqslant P_{\text{pair}}]. \tag{1}$$

If the purification circuit takes N_p raw pairs as input and returns one pair as output, the total number of raw pairs needed for the lattice surgery according to our heuristic is

$$N_{\text{LS}} = dN_pK. \tag{2}$$

If T is the surface code clock cycle (the time it takes to perform a round of syndrome extraction) and if R is the rate at which entanglement can be attempted between pairs of ions, then we have $A = TR$ attempts to collect N_{LS} pairs. Suppose each ELU has $N_{\text{ions}} > N_{\text{LS}}$ communication ions. During a collection attempt, each of the $v \leqslant N_{\text{ions}}$ vacant (unentangled) ions are pulsed and may become entangled with probability p_e. Entangled ions are not pulsed in subsequent rounds, and we assume the entanglement does not degrade as it waits. Our first objective is to determine the probability that N_{LS} pairs can be collected in A attempts. The probability that a single ion pair is entangled after A attempts is given by

$$P_{\text{onepair}} = 1-(1-p_e)^A. \tag{3}$$

Let $X \sim \mathcal{B}(N_{\text{ions}}, P_{\text{onepair}})$ be the binomial random variable representing the number of ion pairs out of the initial N_{ions} that are entangled after A rounds. The minimum number of communication ions needed to collect at least N_{LS} pairs in a code cycle with confidence P_{LS} is then

$$\min_{N_{\text{ions}}} [P(X \geqslant N_{\text{LS}}) \geqslant P_{\text{LS}}] \tag{4}$$

2. Maximum attainable rate given given number of communication ions

Suppose now that N_{ions} is fixed. Let A_{min} be the minimum number of attempts needed to populate the N_{LS} ions needed for lattice surgery with an overall confidence of P_{LS}.

$$A_{\text{min}} = \min_A [P(X \geqslant dN_pK) \geqslant P_{\text{LS}}]. \tag{5}$$

The maximum attainable rate for our fault-tolerant lattice surgery given N_{ions} is then just A_{min}/R

C. Device parameters and assumptions

From Stephenson *et al.* [61], we assume that we can pulse ions at a rate of 1MHz where each pulse has a 2.18×10^{-4} chance of producing an entangled ion-ion pair of fidelity 0.94. We assume that our surface codes have an operational error rate of 0.1% which, from the results of Ramette *et al.* [13], means we can tolerate Bell pairs with an infidelity of 0.01. We assume that the routing and circuits within the purification stage take a negligible amount of time compared with the entanglement collection. Single and two qubit gate error are approximated as single and two-qubit depolarizing channels that occur with probabilities 1×10^{-5} and 5×10^{-5}, respectively. Measurement errors are taken as bitflip channels that occur with probability 1×10^{-5}. Although ion-trapping lifetimes are extremely good (hours, and even months in extreme cases), they are not indefinite. We do not consider ion loss or replacement in our resource estimation. Neither do we consider leakage errors that are known to accumulate with consecutive surface code cycles [63]. Though a linear crystal architecture has all-to-all connectivity in theory, it may not be possible in practice to perform arbitrary simultaneous two qubit gates as we have assumed. Nevertheless, we note promising results from the current state-of-the-art [64].

D. Optimizing entanglement purification with device level noise

Our need for fault-tolerant lattice surgery highlights the importance of a high yield pair purification protocol. Though all such protocols theoretically asymptote to unit fidelity, the practical reality is that device level noise imposes a cap on the pair fidelities that are attainable with purification. It is essential therefore to find a purification protocol that is able to reach our target fidelity of $F_{\text{ideal}} = 0.99$ despite circuit-level noise. To this end, we decided to search for high-performing purification protocols using recently developed numerical methods. Goodenough *et al.* proposed an exhaustive search over purification protocols by mapping the problem to an enumeration over so called graph codes [66], while Addala *et al.* [65], built on earlier work from Krastanov *et al.* [67] to refine a genetic algorithm for finding purification protocols. We opted to use the genetic optimization over the enumeration because of its ease of use and because the attainable fidelities reported by Addala *et al.* were comparable to those reported

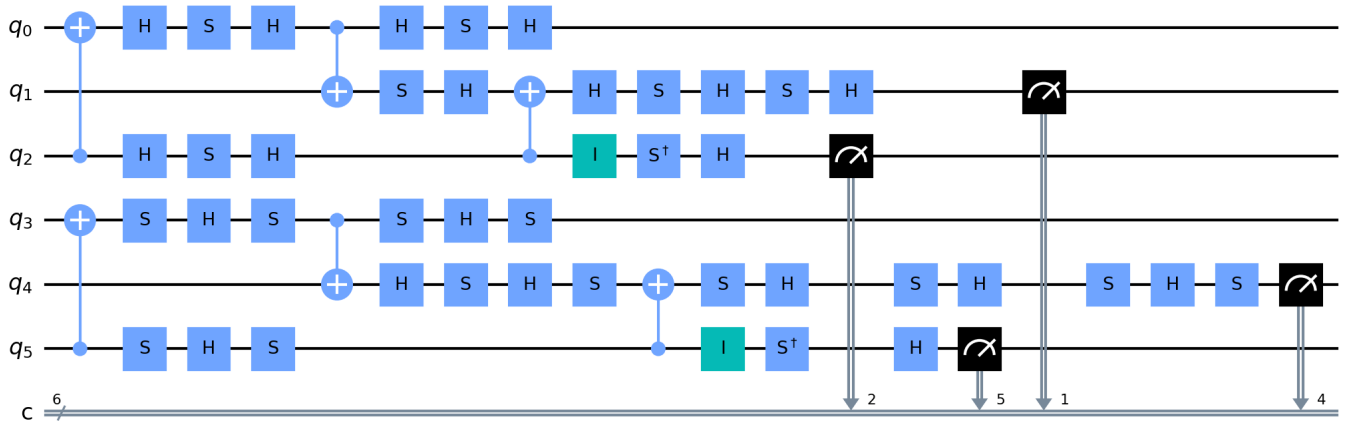

FIG. 5. A high yield purification circuit discovered by the genetic algorithm from Addala *et al.* [65]. This circuit takes three partially entangled pairs as input $\{(q_0, q_3), (q_1, q_4), (q_2, q_5)\}$ and nondeterministically returns a higher fidelity pair at (q_0, q_3). The protocol succeeds if $c_1 = c_4$ and $c_2 \neq c_5$ (in other words, the measurement outcome of qubit q_1 is coincident with the measurement of q_4 and the measurement of q_2 is anticoincident with q_5). When this circuit is run with three copies of the Stephenson pair (see Appendix) as input together with the noise parameters described in Sec. IV C, the purification produces a fidelity $F = 0.9904$ pair with probability $p = 0.819$.

by Goodenough. We used this genetic algorithm to identify several hundred potentially suitable purification circuits. From this initial pool of candidates, we simulated each circuit using the noise parameters detailed in Sec. IV C. For added realism, we modeled our initial $F = 0.94$ entangled pairs after the density matrix of an experimentally realized ion-ion pair reported in the Supplemental Material of Stephenson *et al.* [61] (see Appendix). The highest yield purification circuit we discovered with this genetic algorithm is presented in Fig. 5. This protocol takes three Stephenson pairs as input and produces one output pair with a fidelity of $F = 0.9904$ with an overall success probability of 0.819. A full discussion of our methodology and findings is presented in the Appendix of this paper.

V. RESULTS

Our numerical results are presented as tables in Figs. 6 and 7, respectively. In Fig. 6, we used Eq. (4) to determine the minimum number of communication ions needed to collect sufficiently many entangled pairs for fault-tolerant lattice surgery for a given cycle time T. We considered three cycle times of 10, 100, and 1000 μs according to the technological paradigms we discussed in Sec. III over a small range of code distances. Our calculations indicate that the number of communication ions required is approximately linear with respect to both the code distance and the cycle time within our selected ranges. As the cycle time decreases in orders of magnitude, we find straightforwardly that the number of communication ions increases in orders of magnitude. If we assume that a given ELU may contain around 1000 ions at most, we find that a $d = 9$ code is theoretically supported at a clock cycle 1000 μs, while cycle times considerably faster than this appear out of reach.

In Fig. 7, we used Eq. (5) to calculate the maximum lattice surgery rates that are theoretically possible according to our fault-tolerant heuristic given various numbers of communication ions and various code distances. The whited out squares in the 100 ion column from $d = 7$ onward indicate that fault-tolerant lattice surgery is not possible at these distances, since the number of required pairs exceeds the number of communication ions available. Similar to what we observed in the previous table, we find that there is a tenfold difference between the 1000 and 10 000 ion columns, though we note a slight deviation from this trend at the 100 ion column. This behavior occurs because the number of communication ions is close to the number of required pairs. As the required number of pairs approaches the number of available ions, we expect an exponential increase in the number of attempts needed to collect the entanglement since this collection is done without replacement. Our results indicate that with 100 communication ions, we could expect to support a distance 5 or 6 code at a maximum rate of around 100 Hz. For 1000 and 10 000 communication ions, we find that larger code distances are possible with rates at around 1 and 10 KHz, respectively.

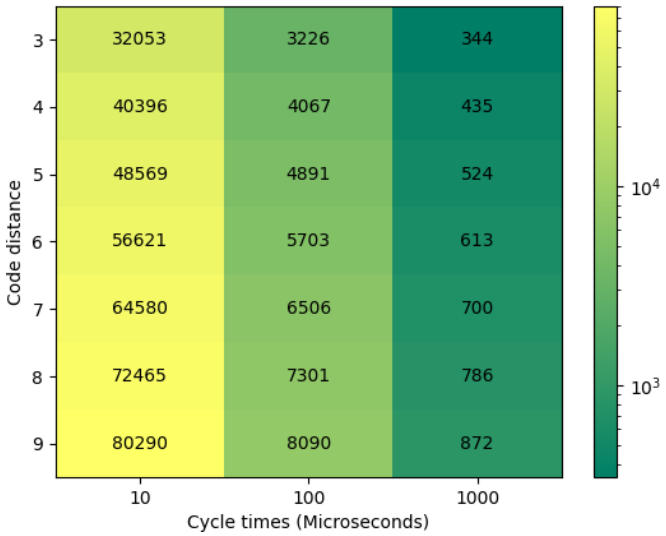

FIG. 6. The minimum number of communication ions that are required to perform fault-tolerant lattice surgery (see Sec. IV B) for a range of code distances and surface code cycle times.

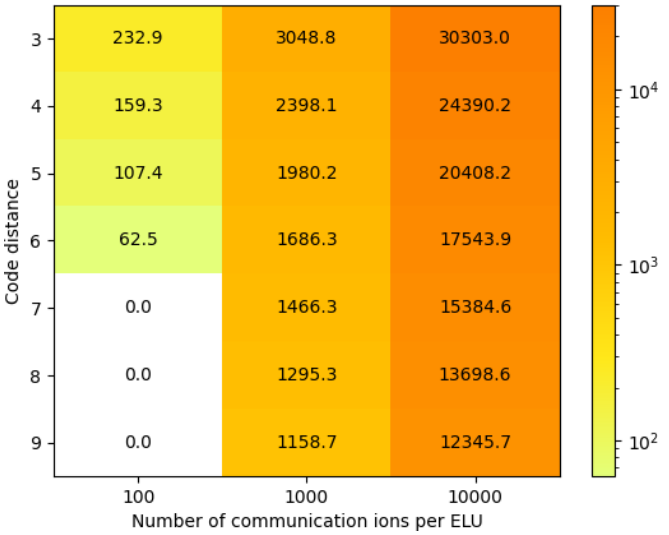

FIG. 7. Average fault-tolerant lattice surgery rates (in hertz) for different code distances and numbers of communication ions. The squares marked with "0.0" signify that fault tolerant lattice surgery is not possible at the specified parameters.

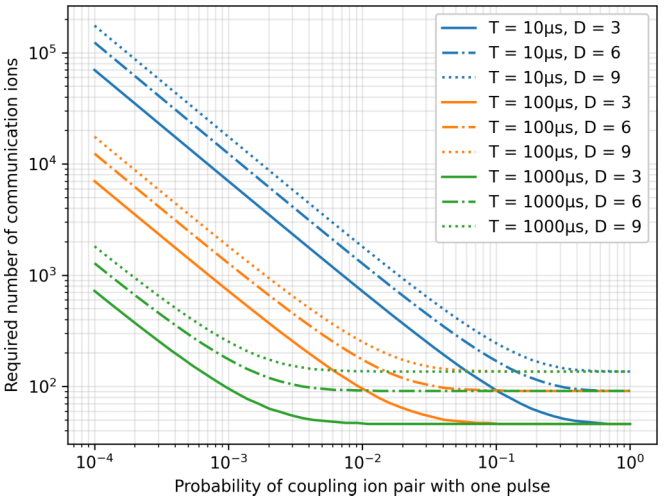

FIG. 8. A selection of line plots that relate the minimum number of communication ions needed per trap with the probability of entangling two intertrap ions with a single "pulse." The three line colors represent the three surface code cycle times proposed in Sec. III D while the three linestyles represent various surface code distances. On the left, the data points begin at probability $p_c = 10^{-4}$, which is comparable to the best known coupling odds of $p_c = 2.18 \times 10^{-4}$ reported by Stephenson *et al.* [61]. Fewer communication ions are needed as the success probability increases since it becomes easier to establish entanglement links between traps. At high probabilities, the lines converge to the same three values for the respective distances; These values represent the minimum number of unrefined pairs that are required for entanglement distillation to yield enough $F = 0.99$ links for lattice surgery. For $d = 3,\ 6$, and 9, these values are 46, 91, and 136, respectively.

Our results indicate long term concerns for scalability due to the large number of physical resources required. If 10 μs is about the fastest clock cycle we can hope for, our findings in Fig. 6 indicate that we would need upwards of 40,000 communication ions per ELU even for modestly sized surface codes. The prohibitive cost of this indicates an urgent need for improved optical coupling; As it stands, the low probability of entangling ions in separate ELUs ($p_c = 2.18 \times 10^{-4}$ [61]) is the leading cause for this inflated resource overhead. A natural question then (and the focus of this section) is how the number of communication ions scales with improvements in the coupling rates. Our results are presented in Fig. 8. Here, we have again used Eq. (4) to calculate the minimum number of ions needed for lattice surgery between a selection of different surface codes while varying the probability p_c of establishing entanglement between a given pair of ions. What we see is that, irrespective of the distances and cycle times we consider, the number of communication ions first decreases as a power law with increasing p_c and eventually tapers off to a fixed value that depends on the code distance and the purification protocol. These plateau values are equal to $N_{\mathrm{LS}} = dN_pK$, which is the number of unpurified pairs that are required between ion traps for lattice surgery to succeed [see Eq. (2)]. We can easily demonstrate this by examining the limiting case behavior when $p_c = 1$. In this scenario, all ions are guaranteed to be entangled after a single round of attempts. Consequently, we only require N_{LS} many ions to ensure that at least N_{LS} many raw pairs are established between traps.

For the sake of transparency, we note that the number of communication ions reported at the plateaus of Fig. 8 are 46, 91, and 136 for distances $d = 3,\ 6$, and 9, respectively. We point this out because for $K = 5$ and $N_p = 3$, these values are equal to $N_{\mathrm{LS}} + 1$. This extra $+1$ comes from a small programming oversight where Eq. (4) was calculated using a strict inequality as opposed to a weak inequality.

The data of Fig. 8 additionally allow us to pin-point the ion coupling probabilities at which we reach these plateaus. For a cycle time of 1000 μs, this occours at around $p_c \approx 10^{-2}$, while for 100 and 10 μs, we find the convergence points at $p_c \approx 0.1$ and ≈ 0.5, respectively.

Let us suppose for the sake of argument that we are allowed 200 communication ions per trap; This is high, but (unlike our data in Fig. 6) is not outside the realm of possibility. From Fig. 8, the ion coupling probabilities that are required to perform lattice surgery for a distance 9 surface code at cycle times of 1000, 100, and 10 μs respectively are $p_c \approx 1.5 \times 10^{-3}$, $p_c \approx 1.5 \times 10^{-2}$, and $p_c \approx 1.5 \times 10^{-1}$. These data suggest that ion-coupling rates need to improve by one or several orders of magnitude depending on the cycle time one wishes to operate at. One way to help meet this demand is to improve the efficiency of photon collection. Carter *et al.* report collection efficiencies of 10% which is roughly an order of magnitude above what was previously possible [68]. Based on this result, it seems that improving entanglement rates by an order of magnitude is feasible target. Whether further improvements are possible however is unclear to us at present. Alternative methods for transporting entanglement via shuttling (therefore bypassing the need for improved coupling) are conceivable, yet speculative. Entanglement distribution using neutral atoms to mediate interactions has been discussed in the context of quantum networking [69].

Entanglement purification is another target for improvement. Our $3 \to 1$ protocol has a 81.9% success rate, which for a confidence threshold of $P_{\text{Pair}} = 0.999$ means that we require $K = 5$ purification circuits [Eq. (1)] per "stitch" in the lattice surgery. This is effectively a $15 \to 1$ purification circuit, which seems somewhat wasteful. Given that this protocol is likely close to optimal for our initial state, the most likely strategy for reducing these overheads is to improve the fidelity at which pairs are distributed. Ideally, purification is eliminated altogether by delivering pairs a fidelity of $F = 0.99$ or higher.

VI. CONCLUSION

In this work, we estimated the number of communication ions needed to perform lattice surgery between two trapped-ion surface code qubits. To this end, we developed three paradigms for syndrome extraction cycle times that are predicated on various technological milestones and presented a heuristic that establishes what it means for a lattice surgery operation to be fault tolerant. With current intertrap coupling rates, we find that hundreds, thousands and tens of thousands of communications ions are required for fault tolerant lattice surgery at cycle times of 1000, 100, and 100 μs, respectively. The primary factor contributing to these prohibitive overheads is poor ion-coupling rates. Our results indicate the need to improve the coupling probability p_c by at least an order of magnitude for lattice surgery to be possible with only a couple hundred resource ions.

ACKNOWLEDGMENTS

We thank Ilia Khait, David Elkouss, Stefan Krastanov, Vaishnavi Addala, and Kenneth Goodenough for helpful discussions. Additional thanks are extended to the Centre for Quantum Computing and Communication Technology (CQC2T). The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. This research was developed with funding from the Defense Advanced Research Projects Agency [under the Quantum Benchmarking (QB) program under Awards No. HR00112230007 and No. HR001121S0026 contracts].

APPENDIX A: NOISY ENTANGLEMENT DISTILLATION

Implementations $\tilde{\mathcal{D}}$ of any entanglement distillation protocol $\mathcal{D}$ are generally subjected to noise in the sense that $0 \neq \|\tilde{\mathcal{D}} - \mathcal{D}\| \leqslant \epsilon$ for small ϵ. Here we benchmark the performance of entanglement distillation protocols obtained with genetic algorithm [65] subjected to noise in ion trap systems. The genetic algorithm takes a Bell-diagonal state

$$F\phi_+ + (1-F)(p_x\psi_+ + p_z\phi_- + p_y\psi_-)$$

as input and searches for optimal $n \to k$ purification protocols for that state by iterating an initial randomly generated population of circuits (describing entanglement distillation protocols) over a number of generations. The fitness of the individuals in the population is evaluated with respect to one of several possible objective functions. We optimized with respect to the "average marginal fidelity," which is the average fidelity of each output pair traced out from the final ensemble. This is determined analytically and may optionally account for single and two qubit depolarizing gate noise along with measurement errors. Each output circuit is returned with its average marginal fidelity and overall success probability.

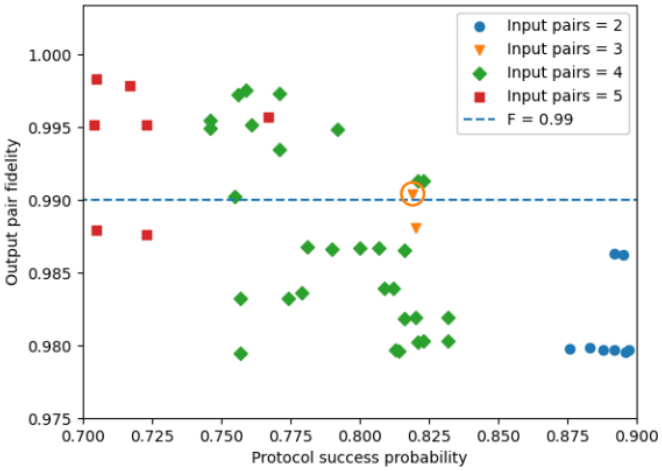

FIG. 9. A scatter plot of the success probabilities and output pair fidelities of a high performing subset of $n \to 1$ purification protocols that were first identified with the genetic algorithm [65] then simulated under circuit level noise with $F = 0.94$ Stephenson pairs [see Eq. (C2)] as inputs. The highest yield purification protocol that exceeds the required fidelity threshold $F_{\text{ideal}} = 0.99$ is circled in orange. This is a $3 \to 1$ purification protocol whose circuit is presented in Fig. 5.

Although the Bell-diagonal states may at first appear to be a somewhat contrived category of entangled pairs, it turns out that all two-qubit mixed states can be deterministically twirled into Bell-diagonal pairs of the same fidelity using local operations and classical communications. This may however cost a small amount of the distillable entanglement depending on the input state, though quantifying the amount of entanglement lost and developing recovery techniques appear to be open research directions. For this reason, and because twirling introduces a small amount of noise, we developed our search methodology around the objective of finding purification protocols that could work without twirling.

Our strategy therefore was to perform an initial search for promising looking protocols using $F = 0.94$ Bell-diagonal pairs with $p_x = p_y = p_z = 1/3$ under the parameter values $n = (3, 4, 5)$ and $k = 1$. Our decision to limit our search to $k = 1$ was motivated by our analysis indicating that the $k > 1$ protocols produced output pairs with some amount of mutual entanglement between them. This is a significant issue for lattice surgery, since it is assumed that each input pair is independent and required us to restrict ourselves to the $k = 1$ case. Simulating beyond $n = 5$ proved to be unnecessary since the average success rate of the protocols can be seen to decrease with increasing n in Fig. 9. For an $(n > 4) \to 1$ protocol to have a higher yield than the

$3 \to 1$ we identified, it would be necessary for the larger purification circuit to have a significantly higher success probability.

Each simulation was performed with a population of 100 circuits evolved for 150 generations. We selected the top performing circuits from each simulation and benchmarked their performance under the same circuit level noise when simulated with $F = 0.94$ Stephenson pairs (see Sec. IV D) as input. Our numerical results are presented in Fig. 9. The broad trend indicates that as the number of pairs increases, the success probability of the protocol decreases while the average success probability increases. None of the protocols we identified were able to exceed the required fidelity threshold of $F_{\text{req}} = 0.999$.

APPENDIX B: THE STEPHENSON PAIR

In Sec. IV D, we alluded to an experimentally realized intertrap ion pair reported by Stephenson *et al.* in the Supplemental Material of their main paper [61]. Strictly speaking, there are four pairs reported which correspond to four possible interferometer detection events. Since these state are all effectively equivalent under local operations and classical communications, we arbitrarily chose to consider the state presented in Eq. (C1). Because our purification protocol works with respect to the target state $|\phi^+\rangle$, it is necessary to rotate our state ρ so its predominant term is $|\phi^+\rangle$. First (for convenience) we apply the following change of basis

$$\left\{|\beta_1\rangle = |\phi^+\rangle,\ \beta_2 = |\phi^-\rangle,\ \beta_3 = \frac{1}{\sqrt{2}}(0,1,i,0)^T,\ \beta_4 = \frac{1}{\sqrt{2}}(0,1,-i,0)^T\right\},$$

After this, we perform the rotation $(I \otimes XZ)\ \rho\ (I \otimes XZ)$ which gives us the density operator ρ' in Eq. (C2). This state is easily verified to have a fidelity of $F = 0.93$ with respect to $|\phi^+\rangle$. This is the Stephenson pair that we refer to in the main body of our paper.

APPENDIX C: TABLES OF CONSTANTS

In this section, we present two tables that detail the most important free parameters and physical constants used throughout the paper. Table II is a summary of the free parameters used and Table III is a summary of the constants.

The "unrotated" Stephenson state, taken from Fig. S4(i) of the Supplemental Material of Ref. [61].

$$\rho = \begin{pmatrix} 0.01 & -0.00487616 + 0.00349614i & 0.0135924 + 0.00634402i & 0.00374015 - 0.00331833i \\ -0.00487616 - 0.00349614i & 0.569 & 0.0542638 + 0.440672i & -0.012985 - 0.0292471i \\ 0.0135924 - 0.00634402i & 0.0542638 - 0.440672i & 0.416 & -0.0225074 - 0.00473484i \\ 0.00374015 + 0.00331833i & -0.012985 + 0.0292471i & -0.0225074 + 0.00473484i & 0.005 \end{pmatrix} \tag{C1}$$

A rotated version of Eq. (C1) obtained with the transformation $\rho' = (I \otimes XZ)\ \rho\ (I \otimes XZ)$.

$$\rho' = \begin{pmatrix} 0.569 + 0.i & -0.00487616 - 0.00349614i & -0.0292471 + 0.012985i & 0.440672 - 0.0542638i \\ -0.00487616 + 0.00349614i & 0.01 + 0.i & -0.00331833 - 0.00374015i & 0.00634402 - 0.0135924i \\ -0.0292471 - 0.012985i & -0.00331833 + 0.00374015i & 0.005 + 0.i & -0.0225074 + 0.00473484i \\ 0.440672 + 0.0542638i & 0.00634402 + 0.0135924i & -0.0225074 - 0.00473484i & 0.416 + 0.i \end{pmatrix} \tag{C2}$$

TABLE II. A summary of the free parameters considered in our analysis.

Parameter	Definition
N_{ions}	The number of communication ions in an ELU
d	Code distance
T	Syndrome extraction cycle time

TABLE III. A catalog of important numerical constants used throughout this paper with justifications.

Parameter	Definition	Value	Justification
p	Purification protocol success probability	0.819	Simulated numerically under circuit level noise
R	Pulse rate	1 MHz	Within the magnitude of what is physically possible [61]
p_c	The probability of entangling two ions with one pulse-attempt	2.18×10^{-4}	State of the art: [61]
F_{ideal}	The fidelity required for pairs used in lattice surgery	0.99	Originally 0.999, but improved thanks to Ref. [13]
N_p	The number of pairs required for the purification circuit	3	Fig. 5
P_{pair}	The required confidence for multiplexed purification circuits to produce at least one pair	0.999	Comfortably below the surface code threshold
K	The required number of purification circuits needed to meet multiplexing confidence	5	Substituting appropriate values into Eq. (1)
P_{LS}	The required confidence for collecting sufficient entanglement within a given clock cycle	0.999	For a surface code of distance $d \leqslant 9$ (the largest distance we consider in our analysis), A P_{LS} of 0.999 is sufficient to guarantee that a logical CNOT between surface codes can be implemented with success probability greater than 99%. This is seen by observing that $(0.999)^9 > 0.99$.

[1] C. D. Bruzewicz, J. Chiaverini, R. McConnell, and J. M. Sage, Trapped-ion quantum computing: Progress and challenges, Appl. Phys. Rev. **6**, 021314 (2019).

[2] B. Lekitsch, S. Weidt, A. G. Fowler, K. Mølmer, S. J. Devitt, C. Wunderlich, and W. K. Hensinger, Blueprint for a microwave trapped ion quantum computer, Sci. Adv. **3**, e1601540 (2017).

[3] K. Nemoto, M. Trupke, S. J. Devitt, A. M. Stephens, B. Scharfenberger, K. Buczak, T. Nöbauer, M. S. Everitt, J. Schmiedmayer, and W. J. Munro, Photonic architecture for scalable quantum information processing in diamond, Phys. Rev. X **4**, 031022 (2014).

[4] J. Roffe, Quantum error correction: an introductory guide, Contemp. Phys. **60**, 226 (2019).

[5] K. S. Chou, J. Z. Blumoff, C. S. Wang, P. C. Reinhold, C. J. Axline, Y. Y. Gao, L. Frunzio, M. H. Devoret, L. Jiang, and R. J. Schoelkopf, Deterministic teleportation of a quantum gate between two logical qubits, Nature (London) **561**, 368 (2018).

[6] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, Purification of noisy entanglement and faithful teleportation via noisy channels, Phys. Rev. Lett. **76**, 722 (1996).

[7] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, Surface codes: Towards practical large-scale quantum computation, Phys. Rev. A **86**, 032324 (2012).

[8] A. M. Stephens, Fault-tolerant thresholds for quantum error correction with the surface code, Phys. Rev. A **89**, 022321 (2014).

[9] D. K. Tuckett, S. D. Bartlett, and S. T. Flammia, Ultrahigh error threshold for surface codes with biased noise, Phys. Rev. Lett. **120**, 050505 (2018).

[10] D. Litinski, A game of surface codes: Large-scale quantum computing with lattice surgery, Quantum **3**, 128 (2019).

[11] M. Vasmer and D. E. Browne, Three-dimensional surface codes: Transversal gates and fault-tolerant architectures, Phys. Rev. A **100**, 012312 (2019).

[12] C. Horsman, A. G. Fowler, S. Devitt, and R. Van Meter, Surface code quantum computing by lattice surgery, New J. Phys. **14**, 123011 (2012).

[13] J. Ramette, J. Sinclair, N. P. Breuckmann, and V. Vuletić, Fault-tolerant connection of error-corrected qubits with noisy links, (2023).

[14] A. Chatterjee, S. Das, and S. Ghosh, Lattice surgery for dummies, arXiv:2404.13202.

[15] C. Monroe, R. Raussendorf, A. Ruthven, K. R. Brown, P. Maunz, L.-M. Duan, and J. Kim, Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects, Phys. Rev. A **89**, 022317 (2014).

[16] J.-S. Chen, E. Nielsen, M. Ebert, V. Inlek, K. Wright, V. Chaplin, A. Maksymov, E. Páez, A. Poudel, P. Maunz, and J. Gamble, Benchmarking a trapped-ion quantum computer with 30 qubits, Quantum **8**, 1516 (2024).

[17] P. Murali, D. M. Debroy, K. R. Brown, and M. Martonosi, Architecting noisy intermediate-scale trapped ion quantum computers, in *Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA '20), Virtual Event* (IEEE Press, 2020), pp. 529–542.

[18] M. Cetina, L. N. Egan, C. Noel, M. L. Goldman, D. Biswas, A. R. Risinger, D. Zhu, and C. Monroe, Control of transverse motion for quantum gates on individually addressed atomic qubits, PRX Quantum **3**, 010334 (2022).

[19] A. K. Ratcliffe, R. L. Taylor, J. J. Hope, and A. R. R. Carvalho, Scaling trapped ion quantum computers using fast gates and microtraps, Phys. Rev. Lett. **120**, 220501 (2018).

[20] Z. Mehdi, A. K. Ratcliffe, and J. J. Hope, Scalable quantum computation with fast gates in two-dimensional mi-

crotrap arrays of trapped ions, Phys. Rev. A **102**, 012618 (2020).

[21] V. Kaushal, B. Lekitsch, A. Stahl, J. Hilder, D. Pijn, C. Schmiegelow, A. Bermudez, M. Müller, F. Schmidt-Kaler, and U. Poschinger, Shuttling-based trapped-ion quantum information processing, AVS Quantum Science **2**, 014101 (2020).

[22] D. Schwerdt, L. Peleg, Y. Shapira, N. Priel, Y. Florshaim, A. Gross, A. Zalic, G. Afek, N. Akerman, A. Stern, A. B. Kish, and R. Ozeri, Scalable architecture for trapped-ion quantum computing using rf traps and dynamic optical potentials, arXiv:2311.01168.

[23] Y.-C. Shen and G.-D. Lin, Scalable quantum computing stabilised by optical tweezers on an ion crystal, New J. Phys. **22**, 053032 (2020).

[24] J. M. Pino, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, M. S. Allman, C. H. Baldwin, M. Foss-Feig, D. Hayes, K. Mayer, C. Ryan-Anderson, and B. Neyenhuis, Demonstration of the trapped-ion quantum ccd computer architecture, Nature (London) **592**, 209 (2021).

[25] S. A. Moses, C. H. Baldwin, M. S. Allman, R. Ancona, L. Ascarrunz, C. Barnes, J. Bartolotta, B. Bjork, P. Blanchard, M. Bohn, J. G. Bohnet, N. C. Brown, N. Q. Burdick, W. C. Burton, S. L. Campbell, J. P. Campora, C. Carron, J. Chambers, J. W. Chan, Y. H. Chen *et al.*, A race-track trapped-ion quantum processor, Phys. Rev. X **13**, 041052 (2023).

[26] M. Malinowski, D. T. C. Allcock, and C. J. Ballance, How to wire a 1000-qubit trapped-ion quantum computer, PRX Quantum **4**, 040313 (2023).

[27] M. Valentini, M. W. van Mourik, F. Butt, J. Wahl, M. Dietl, M. Pfeifer, F. Anmasser, Y. Colombe, C. Rössler, P. Holz, R. Blatt, M. Müller, T. Monz, and P. Schindler, Demonstration of two-dimensional connectivity for a scalable error-corrected ion-trap quantum processor architecture, arXiv:2406.02406.

[28] J. D. Sterk, M. G. Blain, M. Delaney, R. Haltli, E. Heller, A. L. Holterhoff, T. Jennings, N. Jimenez, A. Kozhanov, Z. Meinelt, E. Ou, J. Van D. Wall, C. Noel, and D. Stick, Multi-junction surface ion trap for quantum computing, arXiv:2403.00208.

[29] K. K. Mehta, C. Zhang, M. Malinowski, T.-L. Nguyen, M. Stadler, and J. P. Home, Integrated optical multi-ion quantum logic, Nature (London) **586**, 533 (2020).

[30] Y.-K. Wu, Z.-D. Liu, W.-D. Zhao, and L.-M. Duan, High-fidelity entangling gates in a three-dimensional ion crystal under micromotion, Phys. Rev. A **103**, 022419 (2021).

[31] S.-T. Wang, C. Shen, and L.-M. Duan, Quantum computation under micromotion in a planar ion crystal, Sci. Rep. **5**, 8555 (2015).

[32] A. Erhard, H. P. Nautrup, M. Meth, L. Postler, R. Stricker, M. Stadler, V. Negnevitsky, M. Ringbauer, P. Schindler, H. J. Briegel, R. Blatt, N. Friis, and T. Monz, Entangling logical qubits with lattice surgery, Nature (London) **589**, 220 (2021).

[33] L. Egan, D. M. Debroy, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Newman, M. Li, K. R. Brown, M. Cetina, and C. Monroe, Fault-tolerant control of an error-corrected qubit, Nature (London) **598**, 281 (2021).

[34] D. M. Debroy, M. Li, S. Huang, and K. R. Brown, Logical performance of 9 qubit compass codes in ion traps with crosstalk errors, Quantum Sci. Technol. **5**, 034002 (2020).

[35] T. Leblond, R. S. Bennink, J. G. Lietz, and C. M. Seck, *TISCC: A surface code compiler and resource estimator for trapped-ion processors*, SC-W '23 (Association for Computing Machinery, New York, NY, USA, 2023), pp. 1426–1435.

[36] C. J. Trout, M. Li, M. Gutiérrez, Y. Wu, S.-T. Wang, L. Duan, and K. R. Brown, Simulating the performance of a distance-3 surface code in a linear ion trap, New J. Phys. **20**, 043038 (2018).

[37] Y. Li and S. C. Benjamin, One-dimensional quantum computing with a 'segmented chain' is feasible with today's gate fidelities, npj Quantum Inf. **4**, 25 (2018).

[38] D. Nigg, M. Müller, E. A. Martinez, P. Schindler, M. Hennrich, T. Monz, M. A. Martin-Delgado, and R. Blatt, Quantum computations on a topologically encoded qubit, Science **345**, 302 (2014).

[39] C. Ryan-Anderson, N. C. Brown, M. S. Allman, B. Arkin, G. Asa-Attuah, C. Baldwin, J. Berg, J. G. Bohnet, S. Braxton, N. Burdick, J. P. Campora, A. Chernoguzov, J. Esposito, B. Evans, D. Francois, J. P. Gaebler, T. M. Gatterman, J. Gerber, K. Gilmore, D. Gresh *et al.*, Implementing fault-tolerant entangling gates on the five-qubit code and the color code arXiv:2208.01863.

[40] A. Kubica, B. Yoshida, and F. Pastawski, Unfolding the color code, New J. Phys. **17**, 083026 (2015).

[41] S. Crain, C. Cahall, G. Vrijsen, E. E. Wollman, M. D. Shaw, V. B. Verma, S. W. Nam, and J. Kim, High-speed low-crosstalk detection of a $171Yb^+$ qubit using superconducting nanowire single photon detectors, Commun. Phys. **2**, 97 (2019).

[42] A. H. Myerson, D. J. Szwer, S. C. Webster, D. T. C. Allcock, M. J. Curtis, G. Imreh, J. A. Sherman, D. N. Stacey, A. M. Steane, and D. M. Lucas, High-fidelity readout of trapped-ion qubits, Phys. Rev. Lett. **100**, 200502 (2008).

[43] S. Wölk, Ch Piltz, T. Sriarunothai, and C. Wunderlich, State selective detection of hyperfine qubits, J. Phys. B: At. Mol. Opt. Phys. **48**, 075101 (2015).

[44] P. O. Schmidt, T. Rosenband, C. Langer, W. M. Itano *et al.*, Spectroscopy using quantum logic, Science **309**, 749 (2005).

[45] L. Feng, Y. Y Huang, Y. K. Wu, W. X. Guo, J. Y. Ma, H. X. Yang, L. Zhang, Y. Wang, C. X. Huang, C. Zhang, L. Yao, B. X. Qi, Y. F. Pu, Z. C. Zhou, and L. M. Duan, Realization of a crosstalk-avoided quantum network node with dual-type qubits by the same ion species, Nat. Commun. **15**, 204 (2024).

[46] H.-X. Yang, J.-Y. Ma, Y.-K. Wu, Y. Wang, M.-M. Cao, W.-X. Guo, Y.-Y. Huang, L. Feng, Z.-C. Zhou, and L.-M. Duan, Realizing coherently convertible dual-type qubits with the same ion species, Nat. Phys. **18**, 1058 (2022).

[47] J. P. Gaebler, C. H. Baldwin, S. A. Moses, J. M. Dreiling, C. Figgatt, M. Foss-Feig, D. Hayes, and J. M. Pino, Suppression of midcircuit measurement crosstalk errors with micromotion, Phys. Rev. A **104**, 062440 (2021).

[48] B. P. Ruzic, T. A. Barrick, J. D. Hunker, R. J. Law, B. K. McFarland, H. J. McGuinness, L. P. Parazzoli, J. D. Sterk, J. W. Van Der Wall, and D. Stick, Entangling-gate error from coherently displaced motional modes of trapped ions, Phys. Rev. A **105**, 052409 (2022).

[49] Christopher D. B. Bentley, H. Ball, M. J. Biercuk, Andre R. R. Carvalho, M. R. Hush, and H. J. Slatyer, Numeric optimization for configurable, parallel, error-robust entangling gates in large ion registers, Adv. Quantum Technol. **3**, 2000044 (2020).

[50] J. D. Sterk, H. Coakley, J. Goldberg, V. Hietala, J. Lechtenberg, H. McGuinness, D. McMurtrey, L. P. Parazzoli, J. Van Der Wall, and D. Stick, Closed-loop optimization of fast trapped-ion

shuttling with sub-quanta excitation, npj Quantum Inf. **8**, 68 (2022).

[51] R. Bowler, J. Gaebler, Y. Lin, T. R. Tan, D. Hanneke, J. D. Jost, J. P. Home, D. Leibfried, and D. J. Wineland, Coherent diabatic ion transport and separation in a multizone trap array, Phys. Rev. Lett. **109**, 080502 (2012).

[52] A. Sørensen and K. Mølmer, Quantum computation with ions in thermal motion, Phys. Rev. Lett. **82**, 1971 (1999).

[53] M. Kang, Q. Liang, M. Li, and Y. Nam, Efficient motional-mode characterization for high-fidelity trapped-ion quantum computing, Quantum Sci. Technol. **8**, 024002 (2023).

[54] L. Feng, W. L. Tan, A. De, A. Menon, A. Chu, G. Pagano, and C. Monroe, Efficient ground-state cooling of large trapped-ion chains with an electromagnetically-induced-transparency tripod scheme, Phys. Rev. Lett. **125**, 053001 (2020).

[55] M. K. Joshi, A. Fabre, C. Maier, T. Brydges, D. Kiesenhofer, H. Hainzer, R. Blatt, and C. F. Roos, Polarization-gradient cooling of 1D and 2D ion coulomb crystals, New J. Phys. **22**, 103013 (2020).

[56] Y. Lin, J. P. Gaebler, T. R. Tan, R. Bowler, J. D. Jost, D. Leibfried, and D. J. Wineland, Sympathetic electromagnetically-induced-transparency laser cooling of motional modes in an ion chain, Phys. Rev. Lett. **110**, 153002 (2013).

[57] Z.-C. Mao, Y.-Z. Xu, Q.-X. Mei, W.-D. Zhao, Y. Jiang, Y. Wang, X.-Y. Chang, L. He, L. Yao, Z.-C. Zhou, Y.-K. Wu, and L.-M. Duan, Experimental realization of multi-ion sympathetic cooling on a trapped ion crystal, Phys. Rev. Lett. **127**, 143201 (2021).

[58] G.-D. Lin and L.-M. Duan, Sympathetic cooling in a large ion crystal, Quantum Info. Proc. **15**, 5299 (2016).

[59] T. Sägesser, R. Matt, R. Oswald, and J. P. Home, Robust dynamical exchange cooling with trapped ions, New J. Phys. **22**, 073069 (2020).

[60] S. D. Fallek, V. S. Sandhu, R. A. McGill, J. M. Gray, H. N. Tinkey, C. R. Clark, and K. R. Brown, Rapid exchange cooling with trapped ions, Nat. Commun. **15**, 1089 (2024).

[61] L. J. Stephenson, D. P. Nadlinger, B. C. Nichol, S. An, P. Drmota, T. G. Ballance, K. Thirumalai, J. F. Goodwin, D. M. Lucas, and C. J. Ballance, High-rate, high-fidelity entanglement of qubits across an elementary quantum network, Phys. Rev. Lett. **124**, 110501 (2020).

[62] R. Stockill, M. J. Stanley, L. Huthmacher, E. Clarke, M. Hugues, A. J. Miller, C. Matthiesen, C. L. Gall, and M. Atatüre, Phase-tuned entangled state generation between distant spin qubits, Phys. Rev. Lett. **119**, 010503 (2017).

[63] N. C. Brown and K. R. Brown, Leakage mitigation for quantum error correction using a mixed qubit scheme, Phys. Rev. A **100**, 032325 (2019).

[64] N. Grzesiak, R. Blümel, K. Wright, K. M. Beck, N. C. Pisenti, M. Li, V. Chaplin, J. M. Amini, S. Debnath, J.-S. Chen, and Y. Nam, Efficient arbitrary simultaneously entangling gates on a trapped-ion quantum computer, Nat. Commun. **11**, 2963 (2020).

[65] V. L. Addala, S. Ge, and S. Krastanov, Faster-than-clifford simulations of entanglement purification circuits and their full-stack optimization, arXiv:2307.06354.

[66] K. Goodenough, S. de Bone, V. L. Addala, S. Krastanov, S. Jansen, D. Gijswijt, and D. Elkouss, Near-term n to k distillation protocols using graph codes, IEEE J. Selected Areas Commun. **42**, 1830 (2024).

[67] S. Krastanov, V. V. Albert, and L. Jiang, Optimized Entanglement Purification, Quantum **3**, 123 (2019).

[68] A. L. Carter, J. O'Reilly, G. Toh, S. Saha, M. Shalaev, I. Goetting, and C. Monroe, Ion Trap with In-Vacuum High Numerical Aperture Imaging for a Dual-Species Modular Quantum Computer, arXiv:2310.07058.

[69] J. Hannegan, J. D. Siverns, J. Cassell, and Q. Quraishi, Improving entanglement generation rates in trapped-ion quantum networks using nondestructive photon measurement and storage, Phys. Rev. A **103**, 052433 (2021).