Dynamic Cooling on Contemporary Quantum Computers

Lindsay Bassman Oftelie ◉,[1,*] Antonella De Pasquale ◉,[2] and Michele Campisi ◉[1]

[1]*NEST, Istituto Nanoscienze-CNR and Scuola Normale Superiore, 56127 Pisa, Italy*
[2]*Graphene Labs, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy*

We study the problem of dynamic cooling whereby a target qubit is cooled at the expense of heating up $N-1$ further identical qubits by means of a global unitary operation. A standard back-of-the-envelope high-temperature estimate establishes that the target qubit temperature can be dynamically cooled by at most a factor of $1/\sqrt{N}$. Here we provide the exact expression for the minimum temperature to which the target qubit can be cooled and reveal that there is a crossover from the high initial temperature regime, where the scaling is $1/\sqrt{N}$, to a low initial temperature regime, where a much faster scaling of $1/N$ occurs. This slow, $1/\sqrt{N}$ scaling, which was relevant for early high-temperature NMR quantum computers, is the reason dynamic cooling was dismissed as ineffectual around 20 years ago; the fact that current low-temperature quantum computers fall in the fast, $1/N$ scaling regime, reinstates the appeal of dynamic cooling today. We further show that the associated work cost of cooling is exponentially more advantageous in the low-temperature regime. We discuss the implementation of dynamic cooling in terms of quantum circuits and examine the effects of hardware noise. We successfully demonstrate dynamic cooling in a three-qubit system on a real quantum processor. Since the circuit size grows quickly with N, scaling dynamic cooling to larger systems on noisy devices poses a challenge. We therefore propose a suboptimal cooling algorithm, whereby relinquishing a small amount of cooling capability results in a drastically reduced circuit complexity, greatly facilitating the implementation of dynamic cooling on near-future quantum computers.

I. INTRODUCTION

Quantum computers offer massive advantages over classical computers in terms of execution time and memory efficiency for a subset of problems, such as optimization and simulation [1,2]. While various physical implementations of quantum computers are still being explored (e.g., superconducting circuits, ion traps, and neutral atoms), all must fulfill a fundamental set of requirements [3]. One of these requirements is the ability to initialize the quantum bits, or qubits, into a pure, fiducial quantum state. Furthermore, pure ancilla qubits will be required for fault-tolerant quantum computers to perform quantum error correction [1,4,5]. The preparation of pure state qubits is therefore a key hurdle in the successful implementation of quantum computers now and in the future.

The problem of initializing a large set of qubits into a pure state was first studied in the context of generating highly polarized qubits in NMR systems to increase signal-to-noise ratios [6–8]. Since large polarization in qubits can be obtained by cooling them down to very low temperatures, scientists began to explore techniques to cool qubits below temperatures that can be achieved with direct, physical cooling methods (e.g., cooling with lasers or large magnetic fields). Schulman and Varizani [9] were the first to propose effective cooling of qubits for quantum computation via the application of certain logic gates on the qubits, which following Ref. [10] we refer to as "dynamic cooling." Their proposal, based on entropy manipulation in a closed system, cools a subset of qubits (e.g., a single target qubit) at the expense of heating the others by performing unitary operations on the entire set of qubits (see Fig. 1).

In the high initial-temperature regime, which was relevant for the NMR-based quantum computers available at the time dynamic cooling was proposed, the Shannon bound establishes that the target qubit can be dynamically cooled by a factor of at most $1/\sqrt{N}$, where N is the total number of qubits [11,12]. This slow scaling led to the dismissal of dynamic cooling as an impractical method for

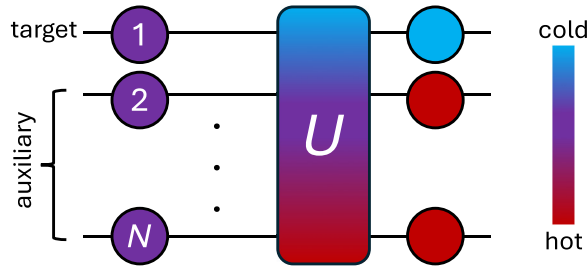*Corresponding author: lindsay.oftelie@nano.cnr.it

FIG. 1. Dynamic cooling of N identical qubits. Here the target qubit is cooled at the expense of heating up the auxiliary qubits via the application of the global unitary operator U.

cooling qubits, and gave thrust to further research aimed at beating Shannon's bound. Since the bound holds for closed systems, subsequent proposals extended the scenario to open systems by allowing a subset of qubits to interact with the environment (i.e., a heat bath), thereby achieving cooling beyond Shannon's bound [11,12]. Such techniques are usually referred to as "heat bath algorithmic cooling" (or sometimes simply as "algorithmic cooling") [13–27].

However, quantum computing technology has undergone a dramatic revolution in the last two decades, with high-temperature NMR quantum computers falling out of favor as newer models that operate at very low temperatures (e.g., superconducting circuits and ion traps) have shown great promise [28,29]. In tandem, as a result of the development of quantum thermodynamics, much interest has grown within the scientific community in regard to the possible advantage, in terms of energy consumption, of quantum technology in general and quantum computing in particular [30].

Here we reexamine dynamic cooling in light of the scientific and technological advances that have been achieved since its inception more than two decades ago. We consider a set of N identical qubits, each initially in thermal equilibrium at some initial temperature T, that undergoes dynamic cooling via a global unitary transformation U, as schematically represented in Fig. 1. After such a transformation, the ground state and excited state populations of the target qubit change, thereby effecting a change in its temperature. We analytically solve the problem of finding the minimum final temperature T' that can be achieved as a function of the initial temperature T, qubit resonant frequency ω, and total number of qubits N. This allows us to unveil a crossover from the expected $1/\sqrt{N}$ scaling at high T to a much faster, unexpected, $1/N$ scaling at low T.

We also provide an analytic expression for the minimal work cost associated with maximal cooling and show that it scales linearly with N (i.e., it is extensive) and displays distinct behaviors at low and high temperature. While it vanishes as $1/T$ in the high-T regime, it vanishes exponentially as $e^{-1/T}$ in the low-T regime. These results evidence that dynamic cooling behaves very differently at high and

low initial temperatures. In particular, at low T, it is much more effective in terms of system-size scaling and energy cost.

Since current quantum computers operate in the low-T regime (unlike early NMR quantum computers), these results reinstate the appeal of dynamic cooling for generating pure state qubits for quantum computation. Given this renewed viability, we discuss the implementation of dynamic cooling in terms of quantum circuits and examine the effect of noise on cooling on near-term quantum computers. We successfully demonstrate dynamic cooling on a real quantum processor on a system of $N = 3$ qubits. While scaling dynamic cooling up to larger systems on noisy quantum computers is a challenge due to the rapid growth of circuit size with N, we demonstrate how this can be overcome by accepting a suboptimal cooling scheme, whereby increased cooling can be achieved at a fixed (low) circuit complexity as the system size is increased. Our reexamination of dynamic cooling suggest that it is a promising technique for preparing pure state qubits on near-future quantum computers.

II. MAXIMAL COOLING

The initial state of the global system reads

$$\rho = \bigotimes_{i=1}^{N} \left(\frac{e^{-\beta H_i}}{Z(\beta)} \right), \quad (1)$$

where $\beta = 1/k_B T$, k_B is Boltzmann's constant, and $Z(\beta) = \text{Tr}\, e^{-\beta H_i} = 2\cosh(\beta\omega/2)$ is the partition function of any of the qubits, whose Hamiltonian $H_i = \hbar\omega\sigma_z^i/2$, is expressed here in terms of the Pauli operator σ_z^i, the reduced Plank constant $\hbar$, and the resonant frequency ω. All qubits are assumed to have the same resonant frequency.

Let $|i_1 i_2 \cdots i_N\rangle$ denote the tensor product $|i_1\rangle_1 \otimes |i_2\rangle_2 \otimes \cdots \otimes |i_N\rangle_N$ of the eigenvectors $|i_k\rangle_k$, of the operators σ_z^k, $k = 1, 2, \ldots, N$, where $i_k = 0$ ($i_k = 1$) denotes the ground (excited) state of qubit k. We assign qubit $k = 1$ to be the target qubit to be cooled and let $P_0 = e^{\beta\hbar\omega/2}/Z(\beta)$ and $P_1 = e^{-\beta\hbar\omega/2}/Z(\beta) = 1 - P_0$ denote the initial ground state and excited state populations of the target qubit, respectively. Similarly, let P'_0 and $P'_1 = 1 - P'_0$ denote the final ground state and excited state populations of the target qubit after application of the cooling unitary.

To maximally cool the target qubit, the goal is to minimize P'_1 over all possible global unitaries U. This problem is equivalent to finding the set of unitaries that minimizes the expectation value of the final energy of the target qubit $u' = \hbar\omega(P'_1 - P'_0) = \hbar\omega(2P'_1 - 1)$. Thus, we must solve the minimization problem

$$u' = \min_{U} \text{Tr} K U \rho U^\dagger, \quad K = H_1 \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1}, \quad (2)$$

where K is the target qubit Hamiltonian expressed in the Hilbert space of the total system. This problem is formally identical to finding the ergotropy of a driven system (ergotropy is the maximum work extractable from a quantum system) [31]. The only difference is that solving for the ergotropy addresses the total system energy, setting K in Eq. (2) to the total system Hamiltonian. Here we address the energy of a subsystem (i.e., the target qubit), using its Hamiltonian for K. Since the specific form of the Hamiltonian is irrelevant to the objective minimization problem, we may borrow techniques used to compute the ergotropy and directly apply them to our problem.

To do so, we note the critical fact that for our system, the initial state ρ commutes with the Hamiltonian K. As discussed in Ref. [31], in such a case the optimization is particularly simple: the minimum is achieved when U is a permutation matrix that maps the eigenstate of ρ with the largest eigenvalue to the eigenstate of K with the smallest eigenvalue, the eigenstate of ρ with the second-largest eigenvalue to the eigenstate of K with the second-smallest eigenvalue, etc. In other words, if e_i are the eigenenergies of K, then if we order the eigenvalues p_i of ρ in nonincreasing fashion, that is,

$$p_i \geq p_k \quad \text{for } i < k, \tag{3}$$

the minimizing unitary U in Eq. (2) is one that performs the permutation σ such that

$$e_{\sigma(i)} \leq e_{\sigma(k)} \quad \text{for } i < k. \tag{4}$$

Note that this holds even in the case of degenerate spectra. Indeed, in our case the spectrum of K is highly degenerate, with only two distinct eigenvalues: $e_i = \pm\hbar\omega/2$. States of the form $|0\, i_2 \cdots i_N\rangle$, which are half of the total 2^N states, have an eigenenergy of K equal to $-\hbar\omega/2$, while states of the form $|1\, i_2 \cdots i_N\rangle$ have an eigenenergy of K equal to $+\hbar\omega/2$. Maximal cooling can thus be implemented by the mapping of the half of the states with the highest occupation probabilities to the half of the states with the lower eigenenergy $e_i = -\hbar\omega/2$. Because of the large degeneracy in the spectrum of K (as well as degeneracy in the spectrum of ρ), there will be many distinct permutations σ, and hence many distinct unitaries U, that achieve maximal cooling. Illustrative examples for $N = 3, 4$ are provided in Appendix A.

The maximum amount by which the target qubit can be cooled is determined by calculating its final excited state population P'_1. Note that P'_1 is simply the sum of the final occupation probabilities of the states $\{|1\, i_2 \cdots i_N\rangle\}$, which are the exact set of states to which the lowest half of the probabilities are mapped. Therefore, to compute P'_1, we simply generate a list of all the occupation probabilities in nondecreasing order and sum the first half of the list.

To do so, note that the occupation probability of a state with k bits set to 0 is given by $(1 - P_1)^k P_1^{N-k}$ (where

P_1 is given by the initial temperature of the qubits), and there will be $\binom{N}{k}$ states with this probability. States with more bits set to 0 (higher k) have higher initial occupation probabilities. Therefore, we can generate a list of the probabilities in nondecreasing order by appending the $\binom{N}{k}$ probabilities of value $(1 - P_1)^k P_1^{N-k}$ to the list as we increase k from 0 to N. Summing the first half of this list will give $P'_1(P_1, N)$.

When N is odd, there are an even number of distinct values of k ranging from 0 to N. The number of probabilities for the first half of the k values ($k = 0, \ldots, \lfloor N/2 \rfloor$) is equal to the number of probabilities for the second half of the k values ($k = \lceil N/2 \rceil, \ldots, N$). Thus,

$$P'_1(P_1, N) = \sum_{0 \leq k < N/2} \binom{N}{k}(1 - P_1)^k P_1^{N-k}, \quad \text{for odd } N. \tag{5}$$

The calculation is slightly more complicated when N is even, since now dividing the list of probabilities in half involves splitting in half the degenerate group of probabilities where $k = N/2$. This means that we must add to Eq. (5) half of the $\binom{N}{N/2}$ degenerate probabilities with value $(1 - P_1)^{N/2} P_1^{N/2}$:

$$P'_1(P_1, N) = \sum_{0 \leq k < N/2} \binom{N}{k}(1 - P_1)^k P_1^{N-k}$$
$$+ \frac{1}{2}\binom{N}{N/2}(1 - P_1)^{N/2} P_1^{N/2}, \quad \text{for even } N. \tag{6}$$

An intriguing observation is that if we start from an odd number of qubits, adding one more qubit will not increase maximal cooling:

$$P'_1(P_1, 2s - 1) = P'_1(P_1, 2s), \quad s \in \mathbb{N} \tag{7}$$

(the proof is provided in Appendix B). This generalizes the fact that a total of at least three (identical) qubits is required to obtain some cooling [32]. To see this, note that with a total of one qubit, no cooling is possible by means of a unitary manipulation, so Eq. (7) implies that cooling with a total of two identical qubits is likewise impossible; a minimum of three qubits is required for dynamic cooling.

Figure 2 shows P'_1 as a function of P_1 for increasing system size N. Note that $P'_1(P_1, N)$ is an increasing function of P_1, meaning that the higher the initial temperature, the higher the final temperature, which agrees with intuition. Note also that in the interval $[0, 1/2[$, $P'_1(P_1, N)$ is a decreasing function of N, namely, the larger N is, the greater is the cooling, in agreement with what one would expect. We have

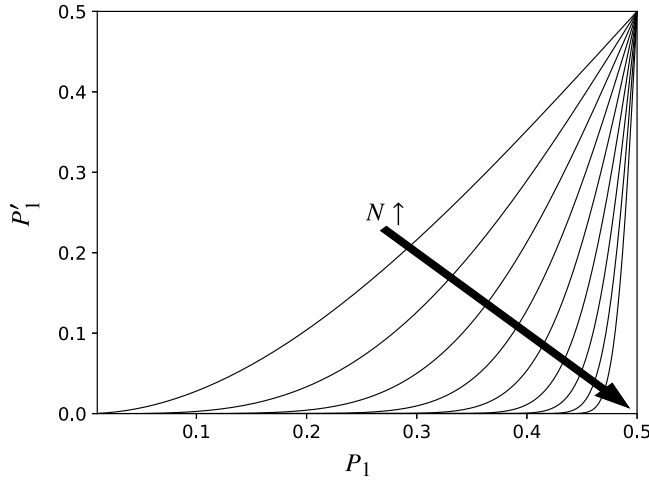$$\lim_{N \to \infty} P'_1(P_1, N) = 0, \quad P_1 \in [0, 1/2[, \tag{8}$$

FIG. 2. Final probability of the excited state of the target qubit after maximal cooling P_1' versus its initial probability P_1 for increasing numbers of total system qubits $N = 2^2, 2^3, \ldots, 2^{10}$.
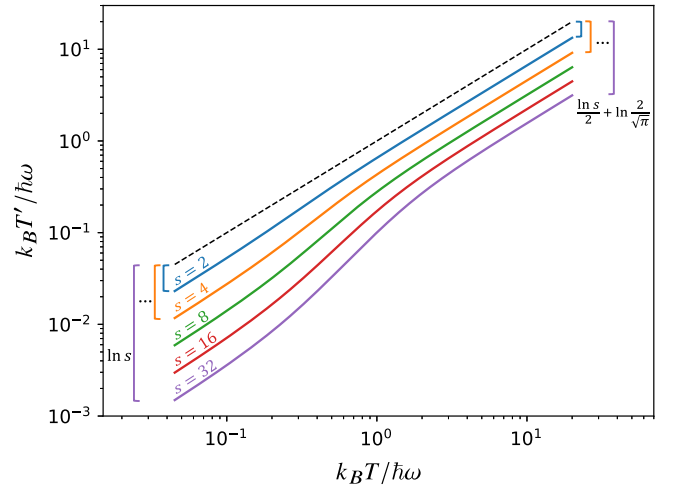

FIG. 3. Log-log plot of the minimal final temperature T' versus the initial temperature T for various system sizes $s = N/2$. The dashed black line plots $T' = T$ (slope equal to 1) to guide the eye in seeing the amount of cooling that occurs. In the low-T regime (bottom left), curves for various system sizes are parallel with a slope of 1, implying a linear relationship between T' and T, and the curve for each system size s has a vertical shift of $\ln s$ from the dashed black line. In the high-T regime (top right), curves for various system sizes are also parallel with a slope of 1, but now the curve for each system size s has a vertical shift of $(\ln s/2) + \ln(2/\sqrt{\pi})$ from the dashed black line.

meaning that as long as the initial temperature is finite and non-negative, by increasing N, one can cool the target qubit arbitrarily close to zero temperature. Note, however, the crucial fact that $P_1'(1/2, N) = 1/2$ for any N. This is because any unitary evolution leaves the completely mixed state unaltered; no cooling is possible if the initial temperature is infinite, regardless of N. This constraint is responsible for the low $1/\sqrt{N}$ scaling at high temperature, which is discussed below.

Using the relation between the initial excited state population P_1 and temperature T

$$\frac{k_B T}{\hbar \omega} = \frac{1}{\ln(1/P_1 - 1)}, \qquad (9)$$

as well as the analogous relation between the final, minimal excited state population P_1' and temperature T', we can write the final minimal temperature as a function of the initial temperature T as:

$$T' = \frac{\hbar \omega}{k_B} \frac{1}{\ln\left(\left[P_1'\left(\frac{1}{e^{\hbar \omega/(k_B T)}+1}, N\right)\right]^{-1} - 1\right)}. \qquad (10)$$

Here the expression "final temperature" is not being used in a strictly thermodynamic sense, i.e., to denote the temperature of the thermal bath surrounding the qubit, but rather it is being used in an "effective" sense, i.e., to denote the temperature that the bath would have if it were in equilibrium with the qubit.

Figure 3 shows a log-log plot of T' versus T for various system sizes $N = 2s$. The dashed black line plots $T' = T$ to guide the eye in seeing the amount of cooling that occurs. In both the low-T regime and the high-T regime, there is a linear relationship between T' and T (the slope of the log-log plots is 1) but the coefficient of proportionality (i.e.,

the vertical shift of the plots) scales differently with s in the two regimes.

In the high-T regime, P_1 is close to $1/2$, and hence we Taylor-expand $P_1'(P_1, 2s - 1) = P_1'(P_1, 2s)$ around $P_1 = 1/2$ to obtain

$$P_1' = 1/2 + c_s(P_1 - 1/2) + O[(P_1 - 1/2)^2], \qquad (11)$$

$$c_s = 2^{2-2s} \sum_{k=0}^{s-1} \binom{2s-1}{k}(2s - 2k + 1). \qquad (12)$$

Expanding the expression $e^{\beta \hbar \omega} = (1 - P_1)/P_1$ to first order around $P_1 = 1/2$, we obtain $\beta \hbar \omega \simeq 4(1/2 - P_1)$. Similarly, for the inverse final temperature, we have $\beta' \hbar \omega \simeq 4(1/2 - P_1') \simeq c_s 4(1/2 - P_1) \simeq c_s \beta \hbar \omega$, or $T' \simeq T/c_s$. This explains the linear relationship between T' and T. It can be proven that $c_s = 2^{2-2s} s\binom{2s-1}{s}$. Then, using Stirling's approximation, $N! \simeq \sqrt{2\pi N}(N/e)^N$, one finds that $c_s \simeq (2/\sqrt{\pi})\sqrt{s}$ in the large-s limit (see Appendix C). Therefore,

$$T' \simeq \frac{\sqrt{\pi}}{2} \frac{T}{\sqrt{s}} = \sqrt{\frac{\pi}{2}} \frac{T}{\sqrt{N}} \quad \text{for } k_B T \gg \hbar \omega. \qquad (13)$$

Note that $T' > T/\sqrt{N}$ because $\sqrt{\pi/2} > 1$, which means that Shannon's bound is obeyed as expected, but not saturated. The finding that the scaling $1/\sqrt{N}$ is realized in

the high-T regime (as opposed to just a theoretical bounding limit) is *per se* a nontrivial result. This slow scaling is clearly visible in the top-right corner of Fig. 3.

In the low-T regime, we have $P_1 \ll 1$. Taylor expansion of $P_1'(P_1, 2s - 1) = P_1'(P_1, 2s)$ around $P_1 = 0$ gives

$$P_1' = a_s P_1^s + O(P_1^{s+1}), \quad a_s = \binom{2s - 1}{s}. \quad (14)$$

For small P_1, we have $e^{-\beta\hbar\omega} = P_1/(1 - P_1) \simeq P_1$, and a small P_1'. Therefore, $e^{-\beta'\hbar\omega} = P_1'/(1 - P_1') \simeq P_1' \simeq a_s P_1^s = e^{-s\beta\hbar\omega + \ln a_s}$. This implies $\hbar\omega\beta' \simeq s\hbar\omega\beta - \ln a_s$. Using Stirling's approximation, we obtain $a_s \simeq s \ln 4 + O(\ln s)$ (see Appendix D); hence, $\hbar\omega\beta' \simeq s(\hbar\omega\beta - \ln 4)$. At low temperature (i.e., $\hbar\omega\beta \gg 1$) the term $\ln 4$ is negligible; therefore $\beta' \simeq s\beta$, or

$$T' \simeq \frac{T}{s} = 2\frac{T}{N} \quad \text{for } k_B T \ll \hbar\omega. \quad (15)$$

This reveals that in the low-T regime, a much faster, $1/N$ scaling holds for dynamic cooling. This superior scaling is clearly visible in the bottom-left corner of Fig. 3.

A characteristic value of $k_B T/\hbar\omega$ for contemporary quantum computers based on superconducting qubits, ion traps, or neutral atoms is approximately 0.2, which places them within the start of the low-T regime. For example, current superconducting qubit quantum computers typically operate at $\omega = 5$ GHz and $P_1 = 0.01$, which equates to an initial temperature T of 8.3 mK. Given these values, we find $T' = 2.1$ mK for $s = 5$ ($N = 9, 10$), which is slightly above the scaling value of $T/s = 8.3/5$ mK $= 1.66$ mK. However, in accordance with our analysis above, the estimate T/s becomes better and better as N increases and/or as T decreases further.

III. MINIMAL WORK

Because of the large degeneracy of the spectrum of K [defined in Eq. (2)], there are many distinct permutations that achieve the desired ordering of eigenvectors for maximal cooling. A natural question is then which among all these permutations have the smallest cost in terms of energy injection into the system, i.e., the work performed on the system, given by

$$W = \text{Tr}\left[H(U\rho U^\dagger - \rho)\right], \quad (16)$$

where $H = \bigotimes_i H_i$ denotes the total system Hamiltonian. We recall that, since the initial state is passive, we have $W \geq 0$. When U realizes a permutation σ, Eq. (16) boils

down to

$$W = \sum_i E_i(p_{\sigma(i)} - p_i), \quad (17)$$

where E_i are the eigenvalues of H. Minimal work cost is thus determined by the following minimization problem:

$$\overline{W} = \min_{\sigma \in \mathcal{C}} \sum_i E_i(p_{\sigma(i)} - p_i), \quad (18)$$

where $\mathcal{C}$ denotes the set of permutations that realize maximal cooling.

Solving this further minimization problem is straightforward. As described above, to achieve maximal cooling it is sufficient to map the half of the states with the highest occupation probabilities to the set of states $\{|0\, i_2 \cdots i_N\rangle\}$. To simultaneously achieve minimal work cost, within this set of states the highest probability should be assigned to the state with the lowest *total system* energy, the second-highest probability should be assigned to the state with the second-lowest total system energy, etc. The probabilities should be mapped in an analogous way for the other half of the states in the set $\{|1\, i_2 \cdots i_N\rangle\}$. This works because states with lower total system energies have higher *initial* occupation probabilities by definition. So assigning the highest *final* probability to the state with the lowest total system energy within each half-list minimizes the differences of the initial and final probabilities $p_{\sigma(i)} - p_i$ in Eq. (17), thereby minimizing work (see Appendix A for more details).

Computing the minimal value of work $\overline{W}$ that must be invested to obtain maximal cooling is conceptually a simple task, but, in practice, it presents some challenges. Note that due to memory limitations, writing the 2^N-dimensional arrays that list the energy eigenvalues E_i, and the populations p_i and $p_{\sigma(i)}$ quickly becomes intractable as N increases (on a desktop computer this happens at around $N \simeq 26$). We overcome this bottleneck by exploiting the sparsity of these arrays, which allows us to encode the relevant information into arrays whose sizes scale linearly, thereby allowing the evaluation of $\overline{W}$ for N into the thousands.

Figure 4 shows the rescaled work $\overline{W}/N$ as a function of P_1 for various values of N. For $P_1 \in [0, 1/2]$, our numerical calculations clearly evidence that as N is increased the solid curves approach the black dashed curve, which plots the following analytic expression:

$$\lim_{N\to\infty} \frac{\overline{W}(P_1, N)}{N} \doteq \bar{w}(P_1) = \frac{\hbar\omega}{2}(P_1 - 2P_1^2). \quad (19)$$

The minimal work is extensive, which evidences a trade-off between cooling power and energetic cost: the further one cools a qubit, the more energy one must expend. At low T, this trade-off is balanced, as the product
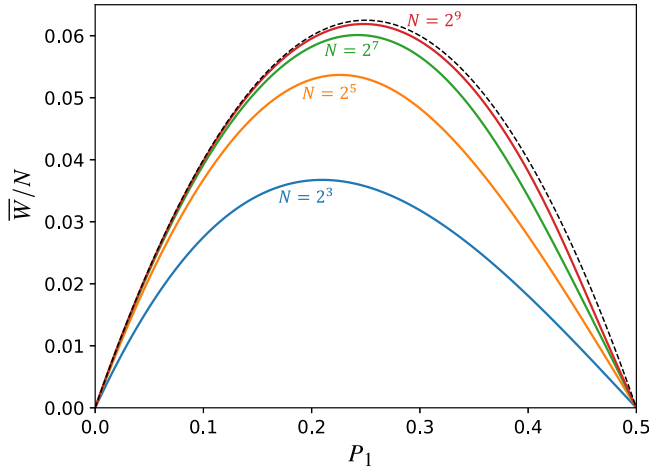
FIG. 4. Rescaled minimum work versus P_1 for various system sizes N. The dashed black line corresponds to Eq. (19).

$\overline{W}T' \sim 2\bar{w}T$ is of order 1. However, in the high-T limit, the trade-off is disadvantageous because the product $\overline{W}T' \sim (\sqrt{\pi}/2)\bar{w}T\sqrt{N}$ scales as $\sqrt{N}$.

Note that $\overline{W}$ goes to zero as it should for $P_1 = 1/2$ (at infinite initial temperature, any U will leave ρ unaltered), and for $P_1 = 0$ (at zero initial temperature, the best you can do is to leave ρ unaltered, i.e., $U = \mathbb{1}$). It is instructive to rewrite the scaling function $\bar{w}$ in terms of initial temperature T:

$$\bar{w}(T) = \frac{\hbar\omega}{2} \frac{\tanh(\frac{\hbar\omega}{2k_BT})}{e^{\hbar\omega/k_BT} + 1}, \qquad (20)$$
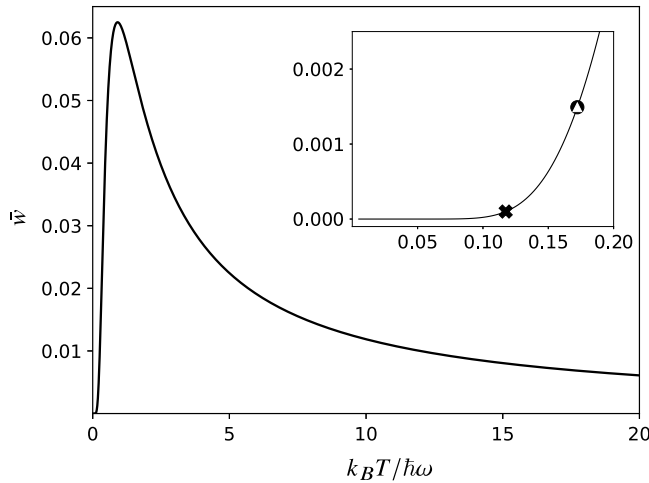

FIG. 5. Rescaled minimum work in the thermodynamic limit versus initial temperature T. The inset shows an enlargement of the exponential vanishing of work at low initial temperatures. Symbols denote state-of-the-art values for qubits on various contemporary quantum computers, including superconducting qubits (black circle), neutral-atom qubits (white triangle), and trapped-ion qubits (black cross).

which is plotted in Fig. 5. In the high-T regime, $\bar{w}$ vanishes as $1/T$, while in the low-T regime, $\bar{w}$ vanishes exponentially, $e^{-\hbar\omega/T}$. The inset in Fig. 5 shows an enlargement of the low-T behavior of $\bar{w}$ and marks state-of-the-art values for qubits on various contemporary quantum computers [33], including superconducting qubits (black circle) [34], neutral-atom qubits (white triangle) [35], and trapped-ion qubits (black cross) [36]. All of them are in the low-T region of the curve, while early NMR qubits are far beyond the full scale of the plot in the high-T regime.

IV. IMPLEMENTATION

To perform dynamic cooling on quantum computers, the cooling unitary U must be translated into a quantum circuit. As stated above, there is a large family of unitaries that can achieve maximal cooling, and different quantum circuits will result from different choices of U. Near-term quantum computers are noisy, with larger circuits accumulating more errors than smaller ones. Therefore, from an implementation perspective, cooling unitaries that can be translated into smaller circuits are more desirable (here and below, the size of a quantum circuit refers to the number of constituent elementary one-qubit and two-qubit gates).

Various protocols exist for generating maximally cooling unitaries. A few protocols of interest include (1) the partner-pairing algorithm (PPA), described in Ref. [13], (2) a minimum work protocol, which generates a unitary with minimal work cost, and (3) a protocol we call the "mirror protocol." The mirror protocol is convenient as it can quickly and automatically generate a unique maximally cooling unitary for arbitrarily large N (other protocols can be significantly more computationally difficult, can be more heuristic, or can have degenerate solutions), the downside being that it generates the largest circuits (see Appendix A for more details on the various protocols). The design of a protocol for generating a maximally cooling unitary with minimal circuit size remains an open question for future research.

It is expected that the size of the cooling circuits will grow exponentially with system size N, as the number of states that must be permuted for maximal cooling likewise grows exponentially in N (subexponential growth, however, has not been disproved). See Appendix E for an expanded discussion of the circuit sizes for dynamic cooling. This scaling has major implications for the practical implementation of dynamic cooling on noisy quantum computers. Namely, while increasing N increases the theoretically optimal cooling capability, increasing N also increases the depth of the associated circuit, and therefore the accumulated error due to noise.

Figure 6 shows the final temperature of the target qubit versus a noise probability parameter p for various system sizes N for two different initial temperatures. The results are derived from quantum circuits simulated with
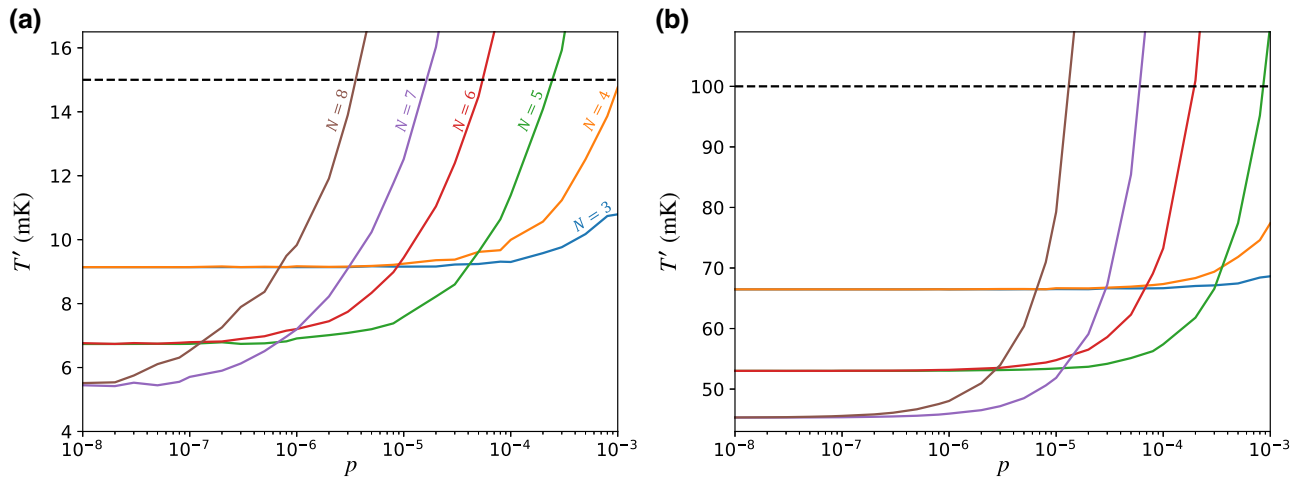
FIG. 6.   Final temperatures T' of the target qubit versus noise probability p of the quantum computer for various system sizes N initialized at (a) $T = 15$ mK or (b) $T = 100$ mK. We assume a value of $\omega = 5$ GHz, which is a typical value for superconducting qubits. The dashed black lines denote initial temperatures of the system. Simulation results from quantum circuits derived from the mirror protocol and executed on a noisy quantum simulator.

a noisy quantum simulator (a classical computer used to simulate the performance of a noisy quantum computer) [37] with the use of a noise model based on a depolarizing channel [38], which can be tuned with a single noise parameter p that effectively sets the probability of error. It is implemented by insertion of a random Pauli operator after each gate in the circuit with probability p. The model is commonly used to emulate the performance of circuits on noisy quantum computers as it approximates well the average noise for large circuits [39–42]. Furthermore, as it is parameterized by only one parameter, the model facilitates study of the scaling of performance versus noise. The quantum circuits were generated with use of the mirror protocol, which was chosen because (1) it generates a unique cooling unitary for each system size N, providing a fair comparison across various system sizes, and (2) it produces larger circuits than other protocols, meaning that if cooling is possible with the mirror protocol, it will certainly be possible with cooling unitaries better optimized for circuit size.

We emphasize that these results should be understood only *qualitatively*, since the noise model does not describe the precise noise present on any particular quantum processor. Moreover, optimizations in terms of selecting a cooling unitary with minimal circuit size and advanced circuit transpilation techniques were not implemented, which would result in shorter, less noisy circuits. As a result, quantitative conclusions from the plots cannot be drawn; rather, Fig. 6 serves to reveal trends in how performance scales with noise.

The initial temperature in each plot is indicated with a horizontal dashed black line. The colored lines indicate the final temperature of the target qubit versus the noise probability p for a range of different system sizes. Given a

system with an odd number of qubits N, both plots show that the addition of one more qubit (which theoretically should exhibit identical cooling capability) impairs cooling capability at higher noise probabilities. While the addition of more than one qubit to the system increases cooling capability at low noise, we see this can actually decrease cooling capability when noise is sufficiently high. Furthermore, for a given system size, the noise probability p at which addition of qubits becomes detrimental as opposed to advantageous is smaller when the system is initialized at a lower temperature. In other words, a system initialized at a lower temperature will be more sensitive to noise.

We conclude that in practice there will be an optimal (finite) number of qubits to use for dynamic cooling, which depends on the level of noise in the quantum hardware as well as the initial temperature of the qubits.

V. DEMONSTRATION

We demonstrate dynamic cooling with $N = 3$ qubits on the IBM quantum computer. Advanced circuit optimization was performed with BQSKit [43] to reduce the cooling circuit down to only nine two-qubit elementary gates. The circuits were executed on the ibmq_brisbane quantum processor, which contains 127 qubits. Dynamic cooling was individually performed within 12 different three-qubit clusters on the chip simultaneously. Figure 7 plots the presumed initial temperature of each target qubit (black curve) and the final temperature of each target qubit (blue curve) after dynamic cooling had been executed within each cluster. The presumed initial temperature was computed by execution of an empty circuit that measured only the target qubit. The measurements allowed us to compute the initial population of the excited state of the target qubit, which
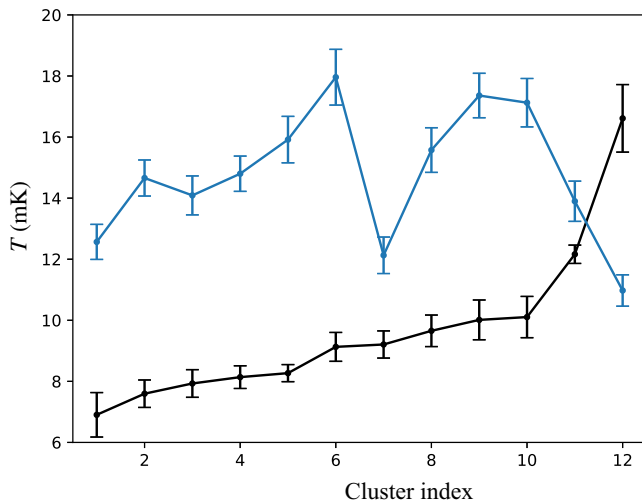
FIG. 7. Initial (black curve) and final (blue curve) effective temperatures of various three-qubit clusters on the IBM quantum computer after dynamic cooling.

was then converted into an initial temperature with use of Eq. (9). This calibration circuit was run five separate times, each with 1024 shots, with the black dots denoting the average value and error bars denoting one standard deviation. The clusters are indexed from 1 to 12, in increasing order of initial temperature. Dynamic cooling was executed in 36 separate runs, each with 1024 shots, on each of the three-qubit clusters, with the blue dots denoting the averaged value and error bars denoting one standard deviation. The final temperature is analogously calculated with use of Eq. (9) with measurements of the final population of the excited state of the target qubit after dynamic cooling. The black and blue curves are drawn to guide the eye in seeing the successful cooling in cluster 12.

The fact that cooling occurs only in the cluster with the highest initial temperature is in line with the trends revealed in Sec. IV; namely, qubits at lower temperature are more sensitive to noise, making them harder to cool at a given level of noise. These results suggest that dynamic cooling might best be used in a scheme that scans the initial temperatures of the qubits (or some estimate thereof) and applies dynamic cooling to only those qubits above some threshold initial temperature.

As noise levels continue to decrease on quantum computers, larger circuits will become more feasible to execute, allowing dynamic cooling to be scaled up to larger system sizes N, thereby cooling the target qubit down to even lower temperatures. While the rapid growth of circuit size with increasing N poses a challenge, the fact that cooling scales much better with system size in the low-T regime may enable sufficient cooling with moderately low N. Another path forward to ameliorate large circuit size is a suboptimal cooling scheme, whereby giving up a

small amount of cooling power results in a large decrease of circuit complexity, as we describe next in Sec. VI.

VI. SUBOPTIMAL DYNAMIC COOLING AT FIXED COMPLEXITY

Currently, the biggest hurdle to the success of dynamic cooling on near-term quantum computers is the size of the cooling circuit. While there is a balanced trade-off between the amount of cooling and energy expenditure at low T, circuit size appears to grow exponentially with N (see Appendix E). While this may seem to be an insurmountable obstacle, here we will see that it can be overcome by relaxing the requirement of optimal (i.e., maximal) cooling and agreeing to achieve a suboptimal final temperature. We find that suboptimal cooling can still (ideally) cool the target qubit down to arbitrarily low temperatures by increasing N, but at a fixed circuit complexity and with a lower work expenditure as compared with optimal cooling. The price that needs to be paid is that the scaling of the final temperature with system size will be slower than $1/N$.

To see this, consider the following suboptimal cooling protocol. Take a system with a total of $N = n^2$ qubits, divided into n clusters, each containing n qubits. Cooling can then be executed in two steps, where first, dynamic cooling is performed within each of the n clusters. Assuming we are operating in the low-T regime, this will bring a total of n qubits to $T' \simeq (2/n)T$ (one qubit from each of the n clusters). In the second step, dynamic cooling is performed among these n cooled qubits, bringing one of them to $T'' \simeq (4/n^2)T = (4/N)T$. While this is less than the maximal cooling $T' \simeq (2/n^2) = (2/N)T$, it requires cooling unitaries acting in a space of dimension only $\sqrt{N}$, drastically reducing the associated circuit complexity. Here and below, the "complexity" of the circuit refers to the maximum Hilbert space dimension on which any of the associated cooling unitaries acts. If we take, for example, $N = 9$, optimal cooling requires a circuit to be generated from a unitary acting on a Hilbert space of dimension 2^9, whereas the suboptimal cooling circuit is generated from unitaries acting on Hilbert spaces of dimension only 2^3. This algorithm can be generalized to $N = n^r$, with use of r steps to obtain a final suboptimal temperature $T^{(r)} = (2^r/N)T$. For a fixed dimension of the clusters n, and hence for a fixed circuit complexity, this amounts to a cooling that scales as

$$T^{(r)} = N^{\frac{\ln 2}{\ln n} - 1}T. \tag{21}$$

For $n > 2$ (which is necessary for cooling to occur in the first place) this implies a negative exponent, ensuring that the qubit can be taken to arbitrarily low final temperature by increasing N, affected by increasing the number of cooling steps r [44].

The key feature here is that cooling of the target qubit can be augmented by increasing N without increasing the complexity of the circuit, which remains fixed. Note that in suboptimal cooling, the total circuit comprises a number of n-qubit subcircuits equal to $\sum_{k=1}^{r} n^{r-k} \leq rn^{r-1} = rN/n = N \ln N/(n \ln n)$. Therefore, while the circuit size still grows with increasing N, it does so only quasilinearly, as opposed to exponentially with N, amounting to a major reduction in circuit size for a given total system size N.

The reduction in circuit complexity of suboptimal cooling significantly increases the feasibility of dynamic cooling on noisy quantum hardware, as evidenced in Fig. 8. For the same total number of qubits $N = 9$, while suboptimal cooling with $n = 3$ and $r = 2$ (dashed-dotted black curve) relinquishes a small amount of cooling capability at very low levels of noise, it has a significant performance advantage for moderate to high levels of noise when compared to optimal cooling (solid black curve). This is due to the substantially reduced circuit complexity. Notice how optimal cooling at the same circuit complexity ($N = 3$, solid blue curve) achieves significantly less cooling than the suboptimal routine. Remarkably, the suboptimal cooling routine achieves more cooling than optimal cooling with $N = 5$ (solid green curve), even though it has a smaller circuit complexity. The advantage of the smaller circuit complexity can also be seen by comparing the noise level at which adverse effects begin to impair cooling: reduction in cooling capability begins at a noise level that is an order of magnitude larger for the suboptimal cooling with $n = 3$
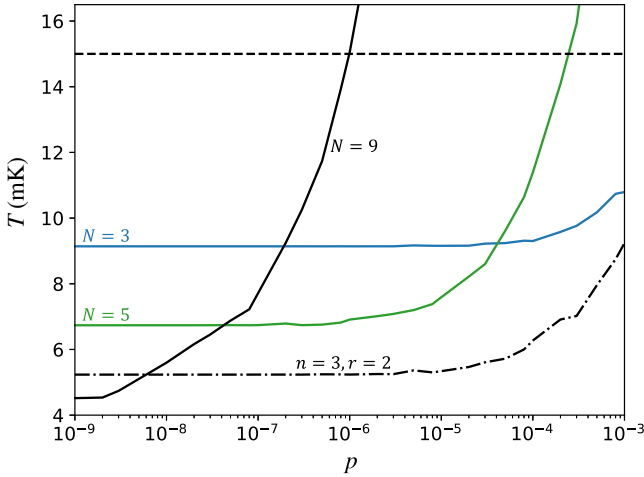
and $r = 2$ versus optimal cooling with $N = 5$, confirming that suboptimal cooling is more resilient to noise.

There is also a reduction in the work cost with suboptimal cooling. The total work cost $W^{(r)}$ is given by the sum of the work costs associated with each step

$$W^{(r)}(P_1, n) = \sum_{k=1}^{r} n^{r-k} \overline{W}(f_n^{k-1}(P_1), n), \qquad (22)$$

where $f_n^k(x) = (f_n \circ f_n \circ \cdots \circ f_n)(x)$ stands for the k-fold application of the function $f_n(x) \doteq P_1'(x, n)$, which we introduce for clarity of notation, and $\overline{W}$ is given in Eq. (18).

The solid curves in Fig. 9 plot the minimal work for suboptimal cooling $W^{(r)}(P_1, n)/n^r$ as a function of P_1 for $n = 3$, with either $r = 3$ (black) or $r = 4$ (red). The corresponding values for work at optimal cooling $\overline{W}(P_1, n^r)/n^r$ obtained with the same total number of qubits $N = n^r$ are also plotted for reference with dotted curves and corresponding colors. Note how, as anticipated, the work associated with the suboptimal r-step cooling is less than the minimal work associated with optimal cooling for the same total system size N. (We remark that a different form of suboptimal cooling was also studied in Ref. [10], where it was also shown to lead to a dramatically reduced work cost.) Note also that the curves $W^{(r)}(P_1, n)/n^r$ collapse onto a single curve for increasing N, meaning that for large N, $W^{(r)}(P_1, n)$ scales linearly with N, namely, $W^{(r)} \simeq N\bar{w}^{(r)}$. This linear scaling can be understood by noting that the sum in Eq. (22) is dominated by the first term $n^{r-1}\overline{W}(P_1, n)$, which is upper-bounded by $n^r\bar{w}(P_1)$, which is linear in $N = n^r$. Subsequent terms are upper-bounded by $n^{r-k}\bar{w}(f_n^{k-1}(P_1))$. While the factor $n^{r-k} = N^{1-k/r}$ is evidently sublinear, the overall scaling is much slower than that, because the factor $\bar{w}$ is evaluated at points $f_n^{k-1}(P_1)$

FIG. 8. Comparison of final temperatures for optimal cooling (solid curves) versus suboptimal cooling (dashed-dotted curve) versus noise probability p. We compare optimal cooling with $N = 3$ qubits (solid blue curve), $N = 5$ qubits (solid green curve), and $N = 9$ qubits (solid black curve) with suboptimal cooling with $r = 2$ steps of cooling with clusters of size $n = 3$ for a total of $N = 9$ qubits (dashed-dotted black curve). An initial temperature of 15 mK is assumed, denoted by the horizontal dashed black line.
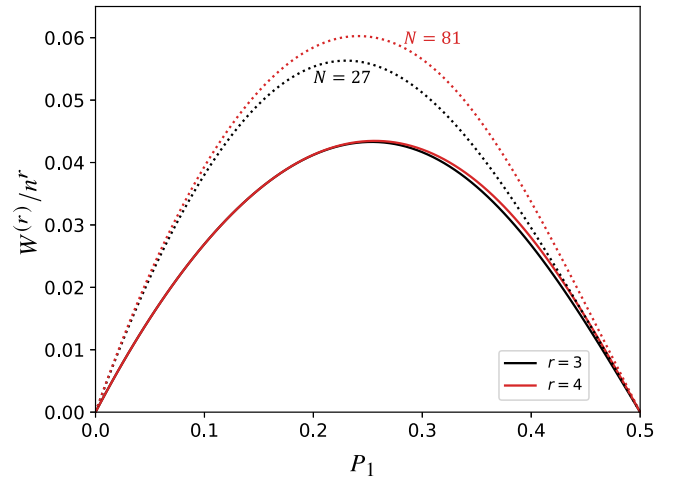

FIG. 9. Rescaled work for r-step suboptimal cooling with cluster size $n = 3$ (solid curves) and the associated optimal cooling with equal total system sizes (dotted curves) for total system sizes $N = 27$ (black) and $N = 81$ (red).

that quickly vanish as k increases [note that $\bar{w}(x) \simeq x/2$ for small x, Eq. (19)].

We have thus shown the nontrivial fact that one can, in principle, cool a target qubit down to arbitrarily low temperature with fixed circuit complexity and at fixed work cost per qubit. The price that must be paid is that of a slower than linear scaling of cooling with system size, Eq. (21)—a price that is, however, counterbalanced by a smaller energy cost and significantly reduced circuit complexity and, therefore, resilience to noise.

VII. CONCLUSION AND OUTLOOK

In light of the major developments in quantum technology, which have moved contemporary quantum computers into the low-temperature regime, we have reexamined dynamic cooling as an effective technique for cooling qubits beyond what is practically achievable with direct, physical cooling methods. We found an analytic expression for the minimum final temperature T' that can be achieved for the target qubit as a function of the initial temperature T. We explored the high-T and low-T regimes and discovered a crossover from a problematic scaling of $1/\sqrt{N}$ at high T to a much more efficient scaling of $1/N$ at low T. We also proposed an analytic expression for the minimal work cost $\overline{W}$ associated with dynamic cooling, which scales linearly with N. In particular, while the work cost vanishes as $1/T$ in the high-T regime, it vanishes exponentially in the low-T regime as $e^{-\hbar\omega/T}$.

We then turned to the implementation of dynamic cooling on noisy quantum computers. We noted that different protocols for generating cooling unitaries will give rise to differing quantum circuit sizes, leaving for future work the problem of finding the cooling unitary with minimal quantum circuit size. We acknowledge that circuit sizes grow rapidly with system size N, and explore the implications with simulations of dynamic cooling on a noisy quantum simulator. The results indicate that there exists an optimal, finite value of N with which to perform dynamic cooling, dependent on the level of noise on the quantum hardware and the initial temperature of the qubits.

Despite high levels of noise of current quantum computers, we were nevertheless able to successfully demonstrate dynamic cooling on the IBM quantum processor. Using a system size of $N = 3$, we performed dynamic cooling on 12 separate three-qubit clusters on the 127-qubit chip, observing cooling in just one of the clusters, which is presumed to have been at a higher initial temperature than all the others. These results suggest that a prudent approach for implementing dynamic cooling on noisy quantum devices may be to scan the initial temperatures (or an estimate thereof) of all the qubits and apply dynamic cooling to only those qubits above a threshold temperature. As noise levels continue to decrease, we expect dynamic cooling will be capable of cooling qubits initialized at lower temperatures and achieve greater cooling with larger system sizes N.

Because of the superior scaling of dynamic cooling with system size in the low-T regime, it may be sufficient to perform dynamic cooling with few-enough qubits to maintain reasonable circuit sizes in near-future quantum devices. However, to overcome the hurdle posed by the rapid growth of circuit size in the near term, we proposed an algorithm for suboptimal dynamic cooling, whereby instead of reaching the optimal final temperature for a given N, we agree to reach a somewhat higher final temperature at the gain of drastically reduced circuit complexity. Surprisingly, cooling a target qubit down to arbitrarily low temperatures is still possible, in principle. While cooling scales more slowly that $1/N$ with this suboptimal routine, the circuit complexity remains fixed while increasing N, yielding the ability to increase cooling capability without increasing the complexity cost.

Recent progress in quantum computing technology has resulted in a slow but steady decrease in noise levels, but a relatively fast increase in the total number of available qubits. Large numbers of moderately-low-noise qubits render the suboptimal cooling described above a very viable scheme for near-future quantum devices. Furthermore, given the demonstration of cooling with a three-qubit cluster on a considerably noisy quantum processor, there is hope that suboptimal cooling with cluster sizes of $n = 3$ could soon be a realizable path for cooling. It should be noted, however, that such schemes may require the connectivity of qubits to be reconsidered in superconducting qubit implementations, which usually provide lattice-shaped connectivity. Instead, a fractal-like network of clustered qubits could greatly facilitate the suboptimal cooling algorithm. Trapped-ion and neutral-atom quantum computers, which provide all-to-all qubit connectivity, are better suited for both optimal and suboptimal dynamic cooling.

All our results support the conclusion that in the low-temperature regime, dynamic cooling is much more effective in terms of scaling and energy cost than at high initial temperatures, and is capable of achieving cooling when noise is reduced to low-enough levels. Given that current quantum computers operate in the low-T regime (unlike early NMR quantum computers), these results reinstate the interest in dynamic cooling for quantum computing applications.

Note added. After submission of our manuscript, it was brought to our attention that Eqs. (5) and (6) had previously been derived using a majorization technique in the Ph.D. thesis of Rodríguez-Briones [45].

ACKNOWLEDGMENTS

APPENDIX A: ILLUSTRATIVE EXAMPLES OF MAXIMAL DYNAMIC COOLING

A simple procedure to implement maximal cooling among N qubits is as follows: Consider listing all 2^N states in increasing lexicographic order, i.e., $|0\cdots000\rangle, |0\cdots001\rangle, |0\cdots010\rangle, |0\cdots011\rangle, \ldots,$ $|1\cdots111\rangle$. The states in the first half of this list all have the target qubit set to 0 ($i_1 = 0$), and thus have the lower eigenenergy of K, $-\hbar\omega/2$. To maximize cooling, we therefore need to construct a unitary U that maps the $2^N/2$ states of ρ with the highest occupation probabilities to the states in the first half of this lexicographically ordered list. We emphasize that, for the case of cooling a single target qubit, the order of probabilities within the first half of the list does not matter; all that matters is that the half of the states with the highest probabilities are mapped to the first half of the lexicographically ordered list.

Among the family of degenerate maximally cooling unitaries, there is a special transformation, which can be derived with the so-called partner-pairing algorithm, which was demonstrated to be optimal in the entropy manipulation step of (heat bath) algorithmic cooling protocols. The PPA generates a permutation such that states in increasing lexicographic order have nonincreasing probabilities. This automatically ensures that the highest probabilities reside in the first half of the lexicographically ordered list. The other special transformations we will examine are the *mirror protocol*, which allows simple preparation of the cooling unitary, and the *minimal work protocol*, which performs maximal cooling with minimal work cost. These will be explained in more detail through the following illustrative examples with $N = 3$ and $N = 4$ qubits.

1. Maximal cooling with $N = 3$ qubits

Consider three identical qubits all initialized at the same temperature T. Our goal is to maximally cool the first qubit below this temperature with the application of a unitary U on all three qubits.

If one lists the $2^N = 8$ states of the total system in lexicographic order, shown in column 2 of Table I, one sees that they are listed in order of increasing eigenenergies of K: the states in the first half of the list, of the form $|0\,i_2\,i_3\rangle$, have eigenenergy $-\hbar\omega/2$, while the states in the second half of the list, of the form $|1\,i_2\,i_3\rangle$, have eigenenergy $+\hbar\omega/2$. Initial occupation probabilities of each state p_i are given in column 3, where $x \equiv P_1$ is the initial occupation

TABLE I. All states of the $N = 3$ qubit system listed in lexicographic order (column 2) with their initial occupation probabilities p_i (column 3). Columns 4 and 6 give various permutations, while columns 5 and 7 give the final occupation probabilities of each state after the respective permutation. For better readability the states that are not being displaced by the permutation are in gray.

i	$|i\rangle$	p_i	$|\sigma(i)\rangle$	$p_{\sigma(i)}$	$|\sigma_B(i)\rangle$	$p_{\sigma_B(i)}$
0	$|000\rangle$	$(1-x)^3$	$|000\rangle$	$(1-x)^3$	$|001\rangle$	$x(1-x)^2$
1	$|001\rangle$	$x(1-x)^2$	$|001\rangle$	$x(1-x)^2$	$|100\rangle$	$x(1-x)^2$
2	$|010\rangle$	$x(1-x)^2$	$|010\rangle$	$x(1-x)^2$	$|010\rangle$	$x(1-x)^2$
3	$|011\rangle$	$x^2(1-x)$	$|100\rangle$	$x(1-x)^2$	$|000\rangle$	$(1-x)^3$
4	$|100\rangle$	$x(1-x)^2$	$|011\rangle$	$x^2(1-x)$	$|101\rangle$	$x^2(1-x)$
5	$|101\rangle$	$x^2(1-x)$	$|101\rangle$	$x^2(1-x)$	$|111\rangle$	x^3
6	$|110\rangle$	$x^2(1-x)$	$|110\rangle$	$x^2(1-x)$	$|110\rangle$	$x^2(1-x)$
7	$|111\rangle$	x^3	$|111\rangle$	x^3	$|011\rangle$	$x^2(1-x)$

probability of the excited state for the target qubit. Notice that the $2^N/2 = 4$ largest occupation probabilities are not in the first half of the list. Specifically, the four-highest probabilities are $(1-x)^3$, $x(1-x)^2$, $x(1-x)^2$, and $x(1-x^2)$. However, one of these values appears in the lower half of the list. Using the PPA, one can achieve maximal cooling by reordering the probabilities in nonincreasing order, which can be accomplished by swapping the two states $|011\rangle$ and $|110\rangle$.

This permutation can be done with a unitary operator U, defined in the computational basis, as

$$U_{011\leftrightarrow100} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad \text{(A1)}$$

To gain a better understanding of how this unitary performs optimal cooling, we examine how the population of the excited state of the target qubit changes before and after this transformation. Recall that P_1 and P'_1 denote the occupation probability of the target qubit's excited state before and after the application of U, respectively. Thus,

$$P_1 = P_{100} + P_{101} + P_{110} + P_{111}, \quad \text{(A2)}$$

$$P'_1 = P'_{100} + P'_{101} + P'_{110} + P'_{111}, \quad \text{(A3)}$$

where $P_{i_1 i_2 i_3}$ is the occupation probability of the state $|i_1 i_2 i_3\rangle$. Now, U implements a transformation that swaps the state $|100\rangle$ with the state $|011\rangle$. Accordingly, it exchanges the populations of the two states, that is, $P'_{100} = P_{011}$ and $P'_{011} = P_{100}$, while leaving all other populations

unaltered. It follows that

$$
\begin{aligned}
P'_1 &= P'_{100} + P'_{101} + P'_{110} + P'_{111} \\
&= P_{011} + P_{101} + P_{110} + P_{111} \\
&< P_{100} + P_{101} + P_{110} + P_{111} \\
&= P_1
\end{aligned}
\tag{A4}
$$

because the probability $P_{011} = e^{-\hbar\omega\beta/2}/Z^3$, featuring two excitations, is lower than the probability $P_{100} = e^{\hbar\omega\beta/2}/Z^3$, featuring only one excitation. From Eq. (A4) it follows that

$$
P'_1(P_1, 3) = 3P_1^2 - 2P_1^3 .
\tag{A5}
$$

Thus, for $0 < P_1 < 1/2$ (i.e., $T > 0$), we have $P'_1 < P_1$—namely, the target qubit is cooled.

Because of the degeneracy in the spectra of ρ and K, there exists a family of degenerate unitaries that can achieve maximal cooling. For example, any unitary of the form in Eq. (A1), but with arbitrary phases replacing the 1's, can also achieve maximal cooling. Furthermore, any unitary that swaps states with equal occupation probabilities (in addition to the swap $|100\rangle \leftrightarrow |011\rangle$) also achieves maximal cooling. There even exist unitaries performing maximal cooling that implement permutations with cycle lengths greater than 2 (note that a swap is a permutation cycle of length 2). An example of such a unitary is given by σ_B in column 6 of Table I, featuring the single cycle of length 6 given by $|000\rangle \rightarrow |001\rangle \rightarrow |100\rangle \rightarrow |101\rangle \rightarrow |111\rangle \rightarrow |011\rangle \rightarrow |000\rangle$. In this case, the probabilities are no longer in nonincreasing order. However, the

four-largest probabilities reside in the first half of the lexicographically ordered list, which implies maximal cooling of the target qubit.

In the case of $N = 3$ qubits, the mirror protocol and the minimal work protocol use the same cooling unitary as the PPA. Therefore, we reserve explanation of these two protocols until the next illustrative example with $N = 4$ qubits, where all three protocols can generate different maximally cooling permutations.

Finally, we remark that cooling cannot be achieved with a total of $N = 2$ identical qubits. If the $2^2 = 4$ states are ordered in increasing lexicographic order, $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$, the states are automatically listed in order of increasing eigenenergy of K (i.e., energy of the target qubit), and we see that the two-highest probabilities already occupy the first half of the list. Thus, the target qubit cannot be cooled.

2. Maximal cooling with $N = 4$ qubits

We now consider the case of $N = 4$ qubits. As before, we list the $2^N = 16$ states of the total system in increasing lexicographic order, as shown in column 2 of Table II. Again, this automatically orders the states by increasing eigenenergies of K: the states in the first half of the list, of the form $|0\, i_2\, i_3\, i_4\rangle$, have eigenenergy $-\hbar\omega/2$, while the states in the second half of the list, of the form $|1\, i_2\, i_3\, i_4\rangle$, have eigenenergy $+\hbar\omega/2$. The occupation probabilities of the states p_i are now denoted by symbols in column 4 of Table II to guide the eye to more quickly recognize patterns, where $\blacksquare \doteq (1-x)^4$, $\blacktriangle \doteq (1-x)^3 x$, $| \doteq (1-x)^2 x^2$, $\bullet \doteq (1-x)x^3$, and $\_ \doteq x^4$. Again, $x \equiv P_1$ is the initial

TABLE II. All states of the $N = 4$ qubit system listed in lexicographic order (column 2) with total state energy (column 3) and their initial occupation probabilities (column 4). Columns 5, 7, and 9 give various permutations, while columns 6, 8, and 10 give the final occupation probabilities of each state after the respective permutation. Occupation probabilities are represented by symbols, where $\blacksquare \doteq (1-x)^4$, $\blacktriangle \doteq (1-x)^3 x$, $| \doteq (1-x)^2 x^2$, $\bullet \doteq (1-x)x^3$, and $\_ \doteq x^4$. For better readability the states that are not being displaced by the permutation are in gray.

i	$|i\rangle$	$E_i[\hbar\omega/2]$	p_i	$|\sigma_{PPA}(i)\rangle$	$p_{\sigma_{PPA}(i)}$	$|\sigma_W(i)\rangle$	$p_{\sigma_W(i)}$	$|\sigma_M(i)\rangle$	$p_{\sigma_M(i)}$
0	$|0000\rangle$	-4	$\blacksquare$	$|0000\rangle$	$\blacksquare$	$|0000\rangle$	$\blacksquare$	$|0000\rangle$	$\blacksquare$
1	$|0001\rangle$	-2	$\blacktriangle$	$|0001\rangle$	$\blacktriangle$	$|0001\rangle$	$\blacktriangle$	$|0001\rangle$	$\blacktriangle$
2	$|0010\rangle$	-2	$\blacktriangle$	$|0010\rangle$	$\blacktriangle$	$|0010\rangle$	$\blacktriangle$	$|0010\rangle$	$\blacktriangle$
3	$|0011\rangle$	0	$|$	$|1000\rangle$	$\blacktriangle$	$|0011\rangle$	$|$	$|0011\rangle$	$|$
4	$|0100\rangle$	-2	$\blacktriangle$	$|0100\rangle$	$\blacktriangle$	$|0100\rangle$	$\blacktriangle$	$|0100\rangle$	$\blacktriangle$
5	$|0101\rangle$	0	$|$	$|0101\rangle$	$|$	$|0101\rangle$	$|$	$|0101\rangle$	$|$
6	$|0110\rangle$	0	$|$	$|0110\rangle$	$|$	$|1000\rangle$	$\blacktriangle$	$|0110\rangle$	$|$
7	$|0111\rangle$	2	$\bullet$	$|1100\rangle$	$|$	$|1010\rangle$	$|$	$|1000\rangle$	$\blacktriangle$
8	$|1000\rangle$	-2	$\blacktriangle$	$|0011\rangle$	$|$	$|0110\rangle$	$|$	$|0111\rangle$	$\bullet$
9	$|1001\rangle$	0	$|$	$|1001\rangle$	$|$	$|1001\rangle$	$|$	$|1001\rangle$	$|$
10	$|1010\rangle$	0	$|$	$|1010\rangle$	$|$	$|0111\rangle$	$\bullet$	$|1010\rangle$	$|$
11	$|1011\rangle$	2	$\bullet$	$|1011\rangle$	$\bullet$	$|1011\rangle$	$\bullet$	$|1011\rangle$	$\bullet$
12	$|1100\rangle$	0	$|$	$|0111\rangle$	$\bullet$	$|1100\rangle$	$|$	$|1100\rangle$	$|$
13	$|1101\rangle$	2	$\bullet$	$|1101\rangle$	$\bullet$	$|1101\rangle$	$\bullet$	$|1101\rangle$	$\bullet$
14	$|1110\rangle$	2	$\bullet$	$|1110\rangle$	$\bullet$	$|1110\rangle$	$\bullet$	$|1110\rangle$	$\bullet$
15	$|1111\rangle$	4	$\_$	$|1111\rangle$	$\_$	$|1111\rangle$	$\_$	$|1111\rangle$	$\_$

occupation probability of the excited state for the target qubit. Roughly, the more vertices the symbol has, the higher the probability it represents. The $2^N/2 = 8$ largest occupation probabilities can thus be represented by a set containing one ■, four ▲'s, and three of the six |'s. The energy of the total system E_i for each state is given in column 3, which is relevant for determining permutations that maximally cool a target qubit with minimal work cost.

There are a number of permutations that will transform the probabilities in the first half of the list into the eight-largest probabilities, three of which are shown in Table II. The first is a permutation generated according to the PPA σ_{PPA}, which permutes all the probabilities into nonincreasing order. Note that there is a degenerate family of permutations that can be generated by the PPA. One such permutation, given in column 5 of Table II, features two cycles of length 2 (i.e., two swaps) given by $|0011\rangle \leftrightarrow |1000\rangle$ and $|0111\rangle \leftrightarrow |1100\rangle$.

The second permutation, σ_W, given in column 7 of Table II, features one of a degenerate family of minimal work protocols, which achieves maximal cooling with minimal work cost. In short, after maximal cooling is achieved by the moving of the highest half of the probabilities to the top half of the lexicographically ordered list, the minimal work protocol sorts the probabilities within each half of the list separately. Within each half-list, the highest probability is assigned to the state with the lowest total system energy, the second-highest probability is assigned to the state with the second-lowest total system energy, etc. It turns out that in the case of $N = 4$ qubits, the PPA also belongs to the family of minimal work protocols, but this is not generally the case.

The third permutation, σ_M, given in column 9 of Table II, enacts what we call the "mirror protocol." In the mirror protocol, states that have the target qubit set to 0 and have a total of $k < N/2$ bits set to 0 are swapped with their mirror image (also called the "negative image"). The idea is that these are the states in the top half of the lexicographically ordered list that have lower probabilities than their mirror-image state in the bottom half of the list. This is because a state with $k < N/2$ bits set to 0 will necessarily have fewer bits set to 0 than its mirror image, and thereby have a lower occupation probability. These mirror-image swaps ensure that all states with the target qubit set to 0 are assigned a higher probability than their mirror-image state, which necessarily has the target qubit set to 1. In turn, this means the highest half of the probabilities will reside in the first half of the lexicographically ordered list. The advantage of the mirror protocol is twofold: (1) the ease with which one can automatically generate the maximally cooling unitary for any system size N and (2) the protocol generates a single, unique cooling unitary for each system size N, as opposed to the PPA and minimum work protocols, which can generate a family of degenerate cooling unitaries.

In the mirror protocol for the case of $N = 4$, we seek states that start with 0 and have $k < N/2 = 2$ total bits set to 0. The only state that adheres to these criteria is the state $|0111\rangle$, which we swap with its mirror image, $|1000\rangle$. Notice that it is not a minimal work protocol as the state in the first half of the list with the highest total energy $|0111\rangle$ is not assigned the lowest probability in the top half of the list. Note, also, that the permutation σ on $N = 3$ qubits, given in Table I, is an instance of the mirror protocol, as well as a minimal work protocol.

To convince ourselves that all three permutations in Table II perform maximal cooling, we can compute the probability of the excited state of the target qubit P_1' after each transformation. In this case, $P_1' = P_{1000}' + P_{1001}' + P_{1010}' + P_{1011}' + P_{1100}' + P_{1101}' + P_{1110}' + P_{1111}'$. By consulting what these constituent probabilities are after each permutation in Table II, one finds that in all cases

$$P_1'(P_1, 4) = 3P_1^2 - 2P_1^3. \tag{A6}$$

This is exactly the same expression found for the $N = 3$ case. Namely, addition of a fourth qubit did not increase the cooling power. This is a special case of a more general result: there is no cooling gain in going from an odd N to $N + 1$. Adding a fourth qubit, however, has the adverse effect of increasing the complexity of the unitary operation needed to implement the cooling (in general, operators acting on larger Hilbert spaces are more complex). For a given system size N, different permutations will carry different complexities in terms of their implementation in quantum circuits. Notice that the permutation σ_M in Table II contains one permutation cycle of length 2 (i.e., a swap), which acts on all the qubits, while σ_W contains two swaps, but each swap acts on only three of the four qubits. Such characteristics of the permutation will alter the complexity of the final quantum circuit, and should therefore be considered from a practical standpoint when one is implementing dynamic cooling on quantum computers.

Another crucial point is that distinct permutations that achieve maximal cooling are generally accompanied by distinct energy costs. For example, the work accompanying the permutation σ_M [see Eq. (16)] is given by

$$\begin{aligned} W_{\sigma_M} &= (\hbar\omega/2)[2(▲ - •) + (-2)(• - ▲)] \\ &= 2\hbar\omega[P_0^3 P_1 - P_0 P_1^3]. \end{aligned} \tag{A7}$$

Similarly, the work accompanying the permutation σ_W is given by

$$\begin{aligned} W_{\sigma_W} &= (\hbar\omega/2)[2(| - •) + (-2)(| - ▲)] \\ &= \hbar\omega[(P_0)^3 P_1 - P_0 P_1^3]. \end{aligned} \tag{A8}$$

Note that the minimal work permutation costs half the work of the mirror protocol. σ_W is more energy efficient than σ_M, while achieving the same cooling power $P_1'(P_1, 4)$. Hence, σ_W may be preferable when it comes to practical applications.

APPENDIX B: PROOF THAT $P_1'(P_1, 2s-1) = P_1'(P_1, 2s)$

Below we give the proof that, in general, $P_1'(P_1, 2s-1) = P_1'(P_1, 2s)$ for $s \in \mathbb{N}$. For ease of notation we set $x \equiv P_1$ and $y \equiv P_1'$. Note that in this notation, $(1-x) \equiv P_0$. For $N = 2s$, we have

$$
\begin{aligned}
y(x, N) &= \sum_{0 \le k \le \frac{N}{2}-1} \binom{N}{k}(1-x)^k x^{N-k} + \frac{1}{2}\binom{N}{\frac{N}{2}}(1-x)^{\frac{N}{2}}x^{\frac{N}{2}} \\
&= \binom{N}{0}x^N + \sum_{1 \le k \le \frac{N}{2}-1}\left[\binom{N-1}{k} + \binom{N-1}{k-1}\right](1-x)^k x^{N-k} + \frac{1}{2}\binom{N}{\frac{N}{2}}(1-x)^{\frac{N}{2}}x^{\frac{N}{2}} \\
&= x\left[\binom{N}{0}x^{N-1} + \sum_{1 \le k \le \frac{N}{2}-1}\binom{N-1}{k}(1-x)^k x^{N-k-1}\right] + \sum_{1 \le k \le \frac{N}{2}-1}\binom{N-1}{k-1}(1-x)^k x^{N-k} + \frac{1}{2}\binom{N}{\frac{N}{2}}(1-x)^{\frac{N}{2}}x^{\frac{N}{2}} \\
&= xy(x, N-1) + (1-x)\left[\sum_{0 \le q \le \frac{N}{2}-1}\binom{N-1}{q}(1-x)^q x^{N-1-q} + \frac{1}{2}\left[\binom{N-1}{\frac{N}{2}-1} + \binom{N-1}{\frac{N}{2}}\right](1-x)^{\frac{N-2}{2}}x^{\frac{N}{2}}\right] \\
&= xy(x, N-1) + (1-x)\left[\sum_{0 \le q \le \frac{N}{2}-1}\binom{N-1}{q}(1-x)^q x^{N-1-q} + \binom{N-1}{\frac{N}{2}-1}(1-x)^{\frac{N}{2}-1}x^{\frac{N}{2}}\right] \\
&= xy(x, N-1) + (1-x)\left[\sum_{0 \le q \le \frac{N}{2}}\binom{N-1}{q}(1-x)^q x^{N-1-q}\right] \\
&= xy(x, N-1) + (1-x)y(x, N-1) \\
&= y(x, N-1).
\end{aligned}
\tag{B1}
$$

In the second line we used the identity $\binom{N}{k} = \binom{N-1}{k} + \binom{N-1}{k-1}$. In the fourth line we used a change of variable $q = k-1$. In the sixth line we used that fact that $\binom{N-1}{(N/2)-1} = \binom{N-1}{N/2}$ for even N.

APPENDIX C: DERIVATION OF $c_s \simeq (2/\sqrt{\pi})\sqrt{s}$

We have

$$
\sum_{k=0}^{s-1}\binom{2s-1}{k}(2s-1-2k)
$$

$$
= (2s-1)\sum_{k=0}^{s-1}\binom{2s-1}{k} - 2\sum_{k=0}^{s-1}k\binom{2s-1}{k}. \tag{C1}
$$

Using the identity

$$
\sum_{k=0}^{n}\binom{n}{k} = 2^n, \tag{C2}
$$

we obtain

$$
\sum_{k=0}^{s-1}\binom{2s-1}{k} = \frac{1}{2}\sum_{k=0}^{2s-1}\binom{2s-1}{k} = \frac{1}{2}2^{2s-1} = 2^{2s-2}, \tag{C3}
$$

where we used the fact that the binomial coefficient is symmetric with respect to reflection about its maximum point.

Using the identity

$$
k\binom{n}{k} = n\binom{n-1}{k-1}, \tag{C4}
$$

we get

$$\sum_{k=0}^{s-1} k \binom{2s-1}{k} = \sum_{k=1}^{s-1} k \binom{2s-1}{k} = (2s-1) \sum_{k=1}^{s-1} \binom{2s-2}{k-1}.$$
(C5)

Furthermore,

$$2^{2s-2} = \sum_{k=0}^{2s-2} \binom{2s-2}{k} = 2 \sum_{k=0}^{s-2} \binom{2s-2}{k} + \binom{2s-2}{s-1},$$
(C6)

where we used the reflection symmetry of the binomial coefficient and kept in mind not to count the mid value $\binom{2s-2}{s-1}$ twice. From the above equation it follows that

$$\sum_{k=1}^{s-1} \binom{2s-2}{k-1} = \sum_{k=0}^{s-2} \binom{2s-2}{k} = \frac{1}{2} \left(2^{2s-2} - \binom{2s-2}{s-1} \right).$$
(C7)

Summing up, we obtain

$$\sum_{k=0}^{s-1} \binom{2s-1}{k} (2s-1-2k)$$

$$= (2s-1)2^{2s-2} - (2s-1) \left(2^{2s-2} - \binom{2s-2}{s-1} \right)$$

$$= (2s-1) \binom{2s-2}{s-1}$$

$$= s \binom{2s-1}{s},$$
(C8)

where we use Eq. (C4) in the last equality. Therefore,

$$c_s = 2^{2-2s} \sum_{k=0}^{s-1} \binom{2s-1}{k} (2s-1-2k)$$

$$= 2^{2-2s} s \binom{2s-1}{s}$$

$$= 2^{2-2s} \, s \, a_s,$$
(C9)

where we used $a_s = \binom{2s-1}{s}$. Using Eq. (D2), we get

$$\ln c_s \simeq 2 \ln 2 - 2s \ln 2 + \ln s - \ln 2 - \ln \sqrt{\pi}$$

$$- \ln \sqrt{s} + s \ln 4$$

$$= \ln 2 - \ln \sqrt{\pi} + \ln \sqrt{s},$$
(C10)

and therefore

$$c_s \simeq \frac{2}{\sqrt{\pi}} \sqrt{s}.$$
(C11)

APPENDIX D: DERIVATION OF $a_s \simeq s \ln 4 + O(\ln s)$

Using Stirling's approximation

$$N! \simeq \sqrt{2\pi N}(N/e)^N,$$
(D1)

one finds

$$\ln a_s = \ln \binom{2s-1}{s}$$

$$\simeq \frac{1}{2} \ln(2\pi) + \frac{1}{2} \ln(2s-1) + (2s-1) \ln(2s-1)$$

$$- (2s-1)$$

$$- \frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln s - s \ln s + s$$

$$- \frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(s-1) - (s-) \ln(s-1) + s - 1$$

$$\simeq - \ln(2\pi) - \frac{1}{2} \ln s + s \ln 4;$$
(D2)

that is, $a_s \simeq s \ln 4 + O(\ln s)$, where $O(\ln s)$ stands for terms that scale at most as $\ln s$.

APPENDIX E: QUANTUM CIRCUIT COMPLEXITY OF DYNAMIC COOLING

We can compare the circuit sizes between various cooling unitaries by defining a systematic way to construct the quantum circuits from the particular unitary. To do this, note that every cooling unitary is defined by a set of cyclic permutations between specified states. The total quantum circuit can be constructed by one building a subcircuit for each permutation cycle and then concatenating all subcircuits.

To describe the construction of a subcircuit for a given permutation cycle, we focus on a swap (a cyclic permutation of length 2), and later explain how to generalize this procedure to cycles of greater lengths. Suppose we wish to construct the quantum circuit that swaps the two states defined by bitstrings b_1 and b_2. First, we define a Gray code between the two bitstrings, which is an ordered list of bitstrings beginning with b_1 and ending with b_2, where each intermediate bitstring differs by only one bit from the previous one [38].

We define the length of the Gray code m as the number of bitstrings in the series. The length of the Gray code will be equal to one more than the number of qubits that differ between the two bitstrings (also known as the "Hamming distance"). Given the Gray code, a quantum circuit that implements the swap can easily be constructed with use of a series of multi-controlled-X (MCX) gates, controlled on $N-1$ qubits, where N is the number of qubits in the system [46]. For a swap with a Gray code of length m, the quantum circuit will contain $2m-3$ MCX

gates (see Ref. [38] and Appendix E for an illustrative example).

Construction of the subcircuit for a cyclic permutation with a length greater than 2 is only slightly more involved. Consider, for example, the cyclic permutation of length 3: $|i_1\rangle \rightarrow |i_2\rangle \rightarrow |i_3\rangle \rightarrow |i_1\rangle$. Let the Gray code length between $|i_1\rangle$ and $|i_2\rangle$ be $m_{1\rightarrow2}$ and the Gray code length between $|i_1\rangle$ and $|i_3\rangle$ be $m_{1\rightarrow3}$. To generate the subcircuit for this permutation, we first apply the $2m_{1\rightarrow2} - 3$ MCX gates to transform the bitstring i_1 to i_2. Next, we apply the $2m_{1\rightarrow3} - 3$ MCX gates to transform the bitstring i_1 to i_3. For a general permutation of length l, the total number of MCX gates required will be $\sum_i[2m_{1\rightarrow i} - 3]$, where $m_{1\rightarrow i}$ is the Gray code length from the first bitstring in the cycle to the i th bitstring in the cycle, where i goes from 2 to l.

To construct the entire circuit, it is only necessary to concatenate all subcircuits for each permutation cycle together in any order. The total circuit will therefore contain a total number of MCX gates equal to $\sum_c \sum_{i_c}[2m_{i_c} - 3]$, where the outer sum runs over all permutation cycles c in the cooling unitary and the inner sum runs over all constituent Gray code lengths m_{i_c} of permutation cycle c.

The mirror protocol (see the definition in Sec. A 2) is, by definition, composed of swaps between states that differ in every qubit. Therefore, the Gray code length of each swap for a system of size N will be the maximum length of $m = N + 1$. In general, the minimal work protocols contain permutations between states that do not differ in every qubit. Therefore, we expect circuits generated with the mirror protocol to be more complex than those generated with minimal work protocols. By tailoring the permutations in the cooling unitary to minimize Gray code lengths, one can minimize circuit sizes. The trade-off is that designing these sets of permutations is currently a heuristic procedure, whereas the mirror protocol can easily and uniquely generate a maximally cooling permutation for each system size N.

We demonstrate the generation of the subcircuit for permutation between the two states $|01111\rangle$ and $|10000\rangle$. To implement the subcircuit for this swap, we let $b_1 = 01111$ and $b_2 = 10000$ and define a Gray code from b_1 to b_2, such as the following:

$$
\begin{aligned}
&01111\\
&11111\\
&10111\\
&10011\\
&10001\\
&10000.
\end{aligned} \tag{E1}
$$

Here the length m of the Gray code is 6. To construct the circuit for the swap, we insert one MCX gate to transform each bitstring to the subsequent one in the Gray code.

After insertion of $m - 1$ MCX gates, the circuit will successfully transform an input state b_1 to b_2. To implement the reverse transformation (since we wish to swap the two states), and uncompute any changes made to other input states not involved in the swap, it is necessary to add the first $m - 2$ MCX gates in reverse order. Thus, a quantum circuit implementing a swap between states with a Gray code of length m will contain $2m - 3$ MCX gates. The quantum circuit implementing the swap between b_1 and b_2 using the Gray code given in Eq. (E1) is depicted in Fig. 10.

Once the quantum circuit has been built with MCX gates, it is necessary to decompose these complex gates into the native gates of the quantum computer (generally these include a two-qubit gate, such as the CNOT gate, and some set of single-qubit gates that render the native gate set universal). Reference [47] describes how such MCX gates can be decomposed into a number of elementary gates (the CNOT gate and arbitrary one-qubit gates) that scales quadratically with system size N. However, if we are not concerned about relative phases between the qubits being conserved, the number of elementary gates scales linearly with N. For cooling, we are concerned not about the relative phases between qubits, rather just about the populations of each state, and thus this linearly scaling transformation can be used.

While the number of elementary gates needed for each MCX gate scales only linearly with N, unfortunately, the total number of MCX gates in the circuit is expected to grow exponentially with N (since the number of permutations in the cooling unitary is expected to grow exponentially with N). This clearly poses a problem for near-term quantum computers with high levels of noise. However,
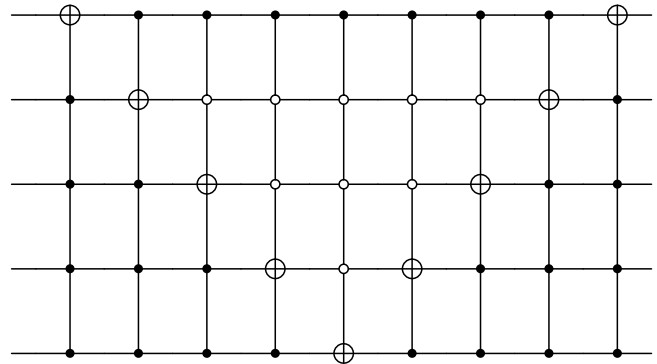

FIG. 10. Quantum circuit implementing a swap between states $|01111\rangle$ and $|10000\rangle$ using the Gray code shown in Eq. (E1). Each wire represents a qubit in the system. The circuit comprises $2m - 3$ MCX gates, where $m = 6$ is the length of the Gray code and $N = 5$ is the number of qubits in the system. Open circles with a cross in the MCX gates are the X (i.e., NOT) gate. closed circles imply the X gate is applied when the corresponding control qubit is in the $|1\rangle$ state, while open circles imply the X gate is applied when the corresponding control qubit is in the $|0\rangle$ state.

since the amount of cooling scales better with system size in the low-T regime, it may be sufficient to cool a target qubit with few-enough auxiliary qubits to maintain reasonable circuit sizes on near-future quantum devices with lower levels of noise. It is also possible to suboptimally cool a target qubit with a fixed complexity, as discussed in Sec. VI.

While it is difficult to derive an analytic expression for the minimal number of elementary gates required for dynamic cooling, Fig. 11 plots a (loose) upper bound on the number of CNOT gates required for dynamic cooling for various system sizes N. The black curve plots the number of CNOT gates calculated with the Gray code method with the mirror protocol, as described above. The red curve plots the number of CNOT gates in circuits derived from circuit synthesis using quantum transpilers (such IBM's Qiskit transpiler and the BQSKit transpiler [43]), which take as input a unitary matrix and as output a circuit. We emphasize that neither of these gate counts describes the minimal gate count for each system size. There is plenty of room for optimization in terms of selecting a cooling unitary with minimal circuit complexity (indeed, we know the mirror protocol is not optimal for complexity), as well as in terms of circuit transpilation techniques. They do, however, reproduce the expected exponential scaling of CNOT gate count with system size. We note that while the circuit transpilation gets extremely computationally expensive as N is increased (we could go up to only $N = 11$ in a reasonable amount of compute time), the CNOT gate count can easily be computed up to any N with use of the Gray code method. Therefore, while computation of the circuit complexity with the Gray code method will not necessarily give the optimal complexity (as evidenced in Fig. 11), it
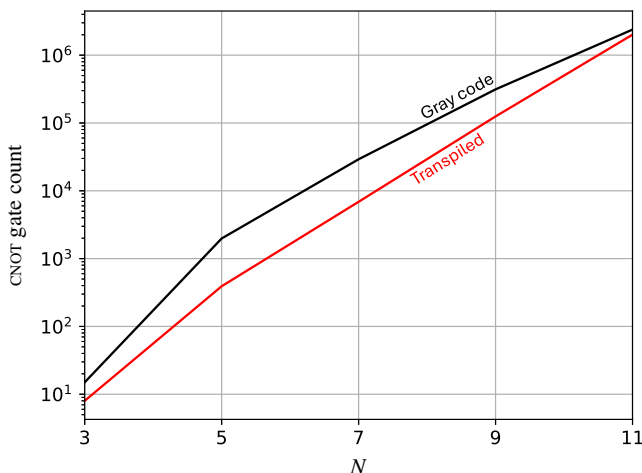
can be useful for quickly comparing complexities between different protocols for large N.

FIG. 11. Number of CNOT gates required for optimal dynamic cooling versus system size N. The black curve plots the gate count derived from the Gray code method, while the red curve plots the gate counts derived from use of a circuit transpiler to synthesize the circuit from the input cooling unitary.

[1] J. Preskill, Reliable quantum computers, Proc. R. Soc. Lond. A **454**, 385 (1998).

[2] L. Bassman, M. Urbanek, M. Metcalf, J. Carter, A. F. Kemper, and W. A. de Jong, Simulating quantum materials with digital quantum computers, Quantum Sci. Technol. **6**, 043002 (2021).

[3] D. P. DiVincenzo, The physical implementation of quantum computation, Prog. Phys. **48**, 771 (2000).

[4] E. Knill, R. Laflamme, and W. H. Zurek, Resilient quantum computation, Science **279**, 342 (1998).

[5] D. Aharonov and M. Ben-Or, in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC '97)* (Association for Computing Machinery, New York, 1997), p. 176.

[6] G. A. Morris and R. Freeman, Enhancement of nuclear magnetic resonance signals by polarization transfer, J. Am. Chem. Soc. **101**, 760 (1979).

[7] O. W. Sørensen, Polarization transfer experiments in high-resolution NMR spectroscopy, Prog. Nucl. Magn. Reson. Spectrosc. **21**, 503 (1989).

[8] L. J. Schulman and U. Vazirani, Scalable NMR quantum computation, ArXiv:quant-ph/9804060.

[9] L. J. Schulman and U. V. Vazirani, in *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing* (ACM, New York, 1999), p. 322.

[10] A. E. Allahverdyan, K. V. Hovhannisyan, D. Janzing, and G. Mahler, Thermodynamic limits of dynamic cooling, Phys. Rev. E **84**, 041109 (2011).

[11] P. O. Boykin, T. Mor, V. Roychowdhury, F. Vatan, and R. Vrijen, Algorithmic cooling and scalable NMR quantum computers, Proc. Natl. Acad. Sci. **99**, 3388 (2002).

[12] J. M. Fernandez, S. Lloyd, T. Mor, and V. Roychowdhury, Algorithmic cooling of spins: A practicable method for increasing polarization, Int. J. Quantum Inf. **2**, 461 (2004).

[13] L. J. Schulman, T. Mor, and Y. Weinstein, Physical limits of heat-bath algorithmic cooling, Phys. Rev. Lett. **94**, 120501 (2005).

[14] J. M. Fernandez, T. Mor, and Y. Weinstein, Paramagnetic materials and practical algorithmic cooling for NMR quantum computing, Int. J. Quantum Inf. **3**, 281 (2005).

[15] J. Baugh, O. Moussa, C. A. Ryan, A. Nayak, and R. Laflamme, Experimental implementation of heat-bath algorithmic cooling using solid-state nuclear magnetic resonance, Nature **438**, 470 (2005).

[16] F. Rempp, M. Michel, and G. Mahler, Cyclic cooling algorithm, Phys. Rev. A **76**, 032325 (2007).

[17] P. Kaye, Cooling algorithms based on the 3-bit majority, Quantum Inf. Process. **6**, 295 (2007).

[18] C. A. Ryan, O. Moussa, J. Baugh, and R. Laflamme, Spin based heat engine: Demonstration of multiple rounds of algorithmic cooling, Phys. Rev. Lett. **100**, 140501 (2008).

[19] Y. Elias, T. Mor, and Y. Weinstein, Semioptimal practicable algorithmic cooling, Phys. Rev. A **83**, 042340 (2011).

[20] F. Ticozzi and L. Viola, Quantum resources for purification and cooling: Fundamental limits and opportunities, Sci. Rep. **4**, 5192 (2014).

[21] G. Brassard, Y. Elias, T. Mor, and Y. Weinstein, Prospects and limitations of algorithmic cooling, Eur. Phys. J. Plus **129**, 1 (2014).

[22] G. Brassard, Y. Elias, J. M. Fernandez, H. Gilboa, J. A. Jones, T. Mor, Y. Weinstein, and L. Xiao, Experimental heat-bath cooling of spins, Eur. Phys. J. Plus **129**, 1 (2014).

[23] S. Raeisi and M. Mosca, Asymptotic bound for heat-bath algorithmic cooling, Phys. Rev. Lett. **114**, 100404 (2015).

[24] N. A. Rodríguez-Briones and R. Laflamme, Achievable polarization for heat-bath algorithmic cooling, Phys. Rev. Lett. **116**, 170501 (2016).

[25] Y. Atia, Y. Elias, T. Mor, and Y. Weinstein, Algorithmic cooling in liquid-state nuclear magnetic resonance, Phys. Rev. A **93**, 012325 (2016).

[26] N. A. Rodríguez-Briones, E. Martín-Martínez, A. Kempf, and R. Laflamme, Correlation-enhanced algorithmic cooling, Phys. Rev. Lett. **119**, 050502 (2017).

[27] S. Raeisi, M. Kieferová, and M. Mosca, Novel technique for robust optimal algorithmic cooling, Phys. Rev. Lett. **122**, 220501 (2019).

[28] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, and D. A. Buell, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

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

[30] A. Auffèves, Quantum technologies need a quantum energy initiative, PRX Quantum **3**, 020101 (2022).

[31] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen, Maximal work extraction from finite quantum systems, Europhys. Lett. **67**, 565 (2004).

[32] A. Solfanelli, A. Santini, and M. Campisi, Quantum thermodynamic methods to purify a qubit on a quantum processing unit, AVS Quantum Sci. **4**, 026802 (2022).

[33] Initial temperatures on commonly available quantum devices can be slightly higher than the state-of-the-art values, but nevertheless still fall within the low-temperature regime. For example, for currently available IBM quantum processors, $k_B T/\hbar\omega \approx 0.2$, which equates to a rescaled minimum work $\bar{w} = 0.005$ in Fig. 5.

[34] A. Opremcak, C. Liu, C. Wilen, K. Okubo, B. Christensen, D. Sank, T. White, A. Vainsencher, M. Giustina, and A. Megrant, *et al.*, High-fidelity measurement of a superconducting qubit using an on-chip microwave photon counter, Phys. Rev. X **11**, 011027 (2021).

[35] S. J. Evered, D. Bluvstein, M. Kalinowski, S. Ebadi, T. Manovitz, H. Zhou, S. H. Li, A. A. Geim, T. T. Wang, and N. Maskara, *et al.*, High-fidelity parallel entangling gates on a neutral-atom quantum computer, Nature **622**, 268 (2023).

[36] T. Harty, D. Allcock, C. J. Ballance, L. Guidoni, H. Janacek, N. Linke, D. Stacey, and D. Lucas, High-fidelity preparation, gates, memory, and readout of a trapped-ion quantum bit, Phys. Rev. Lett. **113**, 220501 (2014).

[37] We assume all-to-all connectivity of the qubits. This means that a two-qubit gate can operate on any pair of qubits in the system. It should be noted that while trapped-ion and neutral-atom qubits feature such connectivity, this is generally not the case for superconducting qubits. Superconducting qubits are usually arranged on a lattice, and two-qubit gates can be applied only to neighboring pairs of qubits. To operate a two-qubit gate on two qubits that are physically separated across the lattice, a series of SWAP gates must be applied, which can dramatically increase circuit depths, further increasing the impact of noise.

[38] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, England, 2002).

[39] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, Phys. Rev. Lett. **119**, 180509 (2017).

[40] M. Urbanek, B. Nachman, V. R. Pascuzzi, A. He, C. W. Bauer, and W. A. de Jong, Mitigating depolarizing noise on quantum computers with noise-estimation circuits, Phys. Rev. Lett. **127**, 270502 (2021).

[41] M. Dupont and J. E. Moore, Quantum criticality using a superconducting quantum processor, Phys. Rev. B **106**, L041109 (2022).

[42] L. B. Oftelie, R. Van Beeumen, D. Camps, W. A. de Jong, and M. Dupont, Simulating dirty bosons on a quantum computer, New J. Phys. **26**, 013030 (2024).

[43] E. Younis, C. C. Iancu, W. Lavrijsen, M. Davis, and E. Smith, Berkeley Quantum Synthesis Toolkit (BQSKit) v1, Tech. Rep., Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA, United States, 2021.

[44] Similarly, in the high-T regime one would get $T^{(r)} = N^{[\ln(\pi/2)/\ln n]-(1/2)}T$, with the exponent being negative for $n > 2$.

[45] N. A. Rodríguez-Briones, Novel heat-bath algorithmic cooling methods, Ph.D. thesis, University of Waterloo, 2020.

[46] A k-controlled-X gate is an X (i.e., NOT) gate controlled on the values of k qubits.

[47] A. Barenco, C. H. Bennett, R. Cleve, D. P. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. A. Smolin, and H. Weinfurter, Elementary gates for quantum computation, Phys. Rev. A **52**, 3457 (1995).