Unifying paradigms of quantum refrigeration: Fundamental limits of cooling and associated work costs

Clivaz, Fabien; Silva, Ralph; Haack, Geraldine; Brask, Jonatan Bohr; Brunner, Nicolas; Huber, Marcus

Published in: Physical Review E

Link to article, DOI: 10.1103/physreve.100.042130

Publication date: 2019

Document Version Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Unifying paradigms of quantum refrigeration: Fundamental limits of cooling and associated work costs

Fabien Clivaz,1,3 Ralph Silva,1 Géraldine Haack,1 Jonatan Bohr Brask,1,2 Nicolas Brunner,1 and Marcus Huber3
1Department of Applied Physics, University of Geneva, 1211 Geneva 4, Switzerland
2Department of Physics, Technical University of Denmark, Fysikvej, Kongens Lyngby 2800, Denmark
3Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria

(Received 9 November 2017; revised manuscript received 16 July 2018; published 22 October 2019)

In classical thermodynamics the work cost of control can typically be neglected. On the contrary, in quantum thermodynamics the cost of control constitutes a fundamental contribution to the total work cost. Here, focusing on quantum refrigeration, we investigate how the level of control determines the fundamental limits to cooling and how much work is expended in the corresponding process. We compare two extremal levels of control: first, coherent operations, where the entropy of the resource is left unchanged, and, second, incoherent operations, where only energy at maximum entropy (i.e., heat) is extracted from the resource. For minimal machines, we find that the lowest achievable temperature and associated work cost depend strongly on the type of control, in both single-cycle and asymptotic regimes. We also extend our analysis to general machines. Our work provides a unified picture of the different approaches to quantum refrigeration developed in the literature, including algorithmic cooling, autonomous quantum refrigerators, and the resource theory of quantum thermodynamics.

DOI: 10.1103/PhysRevE.100.042130

I. INTRODUCTION

Characterizing the ultimate performance limits of thermal machines is directly connected to the problem of understanding the fundamental laws of thermodynamics. The development of classical thermodynamics was instrumental for the realization of efficient thermal machines. Similarly, understanding the thermodynamics of quantum systems is closely related to the fundamental limits of quantum thermal machines. An intense research effort has been devoted to these questions [1–4], resulting in the formulation of the basic laws of quantum thermodynamics, a resource theory perspective, and a large body of work on quantum thermal machines, including first experimental demonstrations.

When trying to establish fundamental limits on quantum thermodynamics tasks, one is always faced with the problem of identifying the relevant resources. For instance, one may consider different classes of allowed operations on a quantum system, or equivalently different levels of control. This challenge is particular to the quantum regime, where monitoring and manipulating systems generally affects the dynamics. Conceptually different approaches have been pursued in parallel to explore this question.

One approach is via the development of a general theory of quantum thermodynamics that aims at placing upper bounds on the performance limits of quantum thermal machines. By establishing fundamental laws, this abstract perspective provides limits that hold for any possible quantum process (hence to all transformations achievable by quantum thermal machines). Typically, such upper bounds are obtained by characterizing possible state transitions, focusing on the single-cycle regime. The intuition being that a machine cannot perform better than a perfect cycle. Here one can distinguish two paradigms. In the first, free operations are given by “thermal operations” [5–10], i.e., energy-conserving unitaries applied to the system and a thermal bath. The implicit assumptions are access to (1) a perfect timing device, (2) arbitrary spectra in the bath, and (3) interaction Hamiltonians of arbitrary complexity. This perspective led to derivations of the second law [11–13]—i.e., the removal of system entropy in a thermally equilibrated environment comes at an inevitable work cost—and general formulations of the third law [14–16]—cooling to temperatures approaching absolute zero requires a diverging amount of resources. In the second paradigm, one considers an increased amount of classical control over a single quantum system, but no access to bath degrees of freedom; i.e., the implicit assumptions are (1) a perfect timing device and (2) the ability to implement any cyclic change in the Hamiltonian of a quantum system. This led to the concepts of passive states [17–21] and algorithmic cooling [22–26] and more generally to fundamental limits on single-cycle performance of coherently driven quantum machines [27].

Another approach is via explicit models of quantum thermal machines that provide lower bounds on their performance. A wide range of such models have been discussed. In general terms, a quantum thermal machine makes use of external resources (e.g., thermal baths) to accomplish a specific task, such as work extraction or cooling. More formally, these machines are modeled as open quantum systems, where the machine consists of few interacting quantum systems coupled to external baths. Performance is usually evaluated in the asymptotic regime of nonequilibrium steady states. Machines with very different levels of control must be distinguished.

Autonomous quantum thermal machines feature the lowest level of control [28–36]. Here the machine subsystems are coupled to thermal baths at different temperatures, and interact via time-independent Hamiltonians, thus requiring no external source of work or control. In the opposite regime,
TABLE I. Summary of the important properties of both paradigms. Complexity means the number of components the machine is allowed to have. Each component is in principle allowed to be a qudit of arbitrary dimension. In the limit of infinitely many ancillas the single-cycle incoherent paradigm becomes the thermal operations (TOs) used in the resource theory of thermodynamics (RTT) and in the single-cycle coherent paradigm one is allowed to apply any CPTP map to the target.

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Control</th>
<th>Incoherent</th>
<th>Coherent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancillas</td>
<td>At (T_R)</td>
<td>(m)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>At (T_R)</td>
<td>(n-m)</td>
<td>(n)</td>
</tr>
<tr>
<td>Limit (n \to \infty)</td>
<td>TOs</td>
<td>CPTP</td>
<td></td>
</tr>
</tbody>
</table>

machines requiring a high level of control have been considered, such as quantum Otto engines [37–40]. Here one assumes the ability to implement complex unitary cycles, which generally require time-dependent Hamiltonians or well-timed access to a coherent battery [41–43]. Nonetheless similar statements of the second and third law are also possible in this regime [44,45].

Each of the above approaches represents a perfectly reasonable paradigm for discussing the ultimate limitations of quantum thermodynamics, each featuring its own merits and drawbacks. Comparing these approaches is thus a natural and important question. It is, however, also a challenging one, due to the fact that each approach works within its own respective framework and set of assumptions. Recently, several works established preliminary connections between some of these approaches. References [46,47] studied autonomous machines in the transient regime and showed that a single cycle can achieve more cooling than the steady-state regime. Quantum machines powered by finite-size baths have been studied [48] to understand the impact of finite resources, and the control cost of achieving a shortcut to adiabaticity was studied in Ref. [49]. In Ref. [50] the authors explored the implications of finite-size systems, i.e., thermal operations not at the thermodynamic limit. In the single-cycle regime, Refs. [51–53] discussed thermodynamic performance under restricted sets of thermal operations, with limited complexity. Finally, even the assumption of perfect timing control, inherent to all paradigms except autonomous machines, should arguably carry a thermodynamic cost [54].

The above paradigms can instructively be split into two types of assumed control over the quantum system. For a single cycle of a thermodynamic process, we can either assume to be capable of engineering time-dependent Hamiltonians, dubbed coherent control or just turning on time-independent interactions, which we call incoherent control. We explicitly model each bath constituent that we have access to and refer to it as machine size. Thus for an infinite machine size, the incoherent control paradigm exactly captures the resource theory of thermodynamics (see Table I for an overview of machine complexity in our two paradigms). On the other hand, the explicit modeling of size adds another layer to the analysis of thermodynamic processes in terms of size and complexity.

In the accompanying article [55], we used this framework to derive a universal bound for quantum refrigeration and proved that it could be obtained by all types of control, provided that complex enough machines and corresponding interactions are available. In the present work we dig deeper and reveal the intricate connection between machine complexity, control and add the amount of resources consumed in the process to the picture. The latter, in turn, is connected to the entropy change associated with the energy drawn from the resource. Consider our two extremal levels of control—first, the coherent scenario, where the entropy of the resource is left unchanged and, second, the incoherent scenario, where only energy at maximum entropy (i.e., heat) is extracted from the resource. Within each level of control, we investigate the lowest attainable temperature, and the work cost for attaining a certain temperature. These quantities allow us to give a direct and insightful comparison between the different approaches for quantum refrigeration.

To tackle these questions, it is natural to consider machines of a given size (i.e., the number of systems to which one has access), since the size in itself also represents a form of control. We analyze this aspect of control starting from the smallest possible machines. It turns out that the two-qubit machine is the smallest one where the coherent and incoherent scenarios can be compared in a meaningful way. We also discuss the case of general machines, and in particular the limit of asymptotically large machines.

Our results clearly demonstrate the expected crucial role of control for quantum cooling performance, but surprisingly unifies the different operational approaches through machine complexity.

II. SETTING AND SUMMARY OF RESULTS

Cooling a quantum system could have several meanings. For a system initially in a thermal state, one can drive it to a thermal state of lower temperature. Alternatively, one could consider increasing the ground-state population, or decreasing the entropy or the energy. These notions are in general inequivalent for target systems of arbitrary dimension. Determining the fundamental limits to cooling is therefore a complex problem in general. It turns out, however, that for the case of a qubit target, all the above notions of cooling coincide. Because of the clarity that this offers but also because the bounds set on target qubits imply bounds for target qudits (see our accompanying article [55]) we focus on qubit targets only in this article. Specifically we consider cooling a single qubit which is initially in a thermal state set by the environment temperature \(T_R\) and then isolated from any environment [56]. The goal is to increase the ground state of the qubit (without changing its energy gap). In order to cool the target qubit, we couple it to a quantum thermal machine. We consider two scenarios for the operation of this machine, that represent the two extremal levels of control (coherent and incoherent) introduced above. For each of these scenarios we are interested in the limits to cooling performance (see our accompanying article [55] for a complementary treatment of this) as well as in the associated work cost.

Work cost

We characterize the work cost by the free energy change of the resource with respect to the environment temperature
W := ΔFR = ΔE − TRΔS, where ΔE is the energy change of the resource (hot bath or battery), ΔS the corresponding entropy change of the resource and TR the temperature of the environment. Since all of our free energy differences will be taken with respect to the same temperature, namely, TR, we will in the rest of the paper denote ΔFR by ΔF for ease of notation. ΔF is a well-established monotone across thermodynamic paradigms (see, e.g., Refs. [57,58]). In a resource theoretic sense it quantifies the maximum extractable work from a resource in the presence of an environment at equilibrium. This measure thus quantifies how much work the resource could have performed with the energy invested in the machine, if there were no further restrictions. This quantifier is uniquely suited to capture the dimensional restrictions coming from finite-size machines. If one could engineer machines operating at the saturation of the second law of thermodynamics, there would be no difference between both scenarios studied here (indeed, one could use the energy in the hot bath to run a thermal machine at Carnot efficiency to produce “pure” work corresponding to the free energy change of the hot bath). As we will see in the following, for finite-size machines, there can be a drastic difference in cooling performance for the same amount of invested work potential of the resource.

More precisely, the two scenarios are defined as follows:

Scenario 1: Incoherent operations. The source of free energy is a hot bath at a temperature TH > TR. The machine (or any of its subsystems) can be coupled to the hot bath or rethermalized with the environment at any stage. The machine interacts with the target qubit via an energy-conserving unitary operation. The work cost of the operation corresponds to the decrease in free energy of the hot bath. The source of free energy is coherent in the sense of allowing for energy-nonconserving unitary operations between the machine and the target qubit. This effectively assumes a coherent battery or classical control field as the source of free energy. There is no additional thermal bath, and the machine may only be coupled to the environment (at temperature TR). As the entropy is unaffected, the work cost, i.e., the change in free energy, is simply the change in energy.

In order to compare these two scenarios and understand the fundamental limits to cooling performance, we investigate

1. The lowest attainable temperature T∗
2. The work cost for attaining any given temperature, in particular T∗.

In contrast to our accompanying article [55] where we focus on the unbounded number of cycles regime, we are here interested in the single cycle, repeated and asymptotic regimes. In the single-cycle regime, an initial thermalization step is followed by a single unitary operation on the machine and the target qubit (energy-conserving or arbitrary, for scenarios 1 and 2, respectively). In the repeated operations regime, thermalization and unitary operations are alternated a finite number of times. In the asymptotic regime, this cycle of steps is repeated indefinitely.

Turning our attention to the machine more closely, we consider that distinct subsystems of the machine can connect to baths at different temperatures, but we do not allow individual transitions in the machine to be separately thermalized at different temperatures. With that in mind, while bounds on the performance of general machines can be set for both paradigms (see our accompanying article [55]), the incoherent paradigm is trivial unless the machine has a tensor product structure. Since we are here focusing in comparing both paradigms, in particular with respect to their associated work cost, we will consider machines with such a structure only. Furthermore, besides the more practical aspect of small machines, which are arguably easier to realize, especially in the incoherent scenario where increasing the machine size usually comes at the price of decreased interaction strengths [33,35], they also already suffice to saturate the cooling bounds of each scenario, see our accompanying article [55]. This as such motivates our interest to focus most our analysis on the minimal settings. The two smallest possible machines consist of either a single qubit or two qubits. Of these, only the latter allows for a nontrivial comparison between the incoherent and coherent scenarios as a single-qubit machine allows for cooling only in scenario 2.

Figure 2 summarizes the results of our comparison, and demonstrates the crucial role of control for the fundamental limits of quantum refrigeration. It shows the minimal achievable temperature of the target qubit versus the associated work cost in each scenario and for the single-cycle, finite-repetition, and asymptotic regimes.

Surprisingly, in the single-cycle regime, we find that neither scenario is universally superior. While scenario 2 always achieves the lowest temperature when no restriction is placed on the work cost, there is a threshold work cost below which scenario 1 outperforms scenario 2.

1By the second law of thermodynamics we mean that every state transformation on the system ρ → ρ′ such that F(ρ′) − F(ρ), where F is the Helmholtz free energy, is smaller than the free energy difference of the resource, denoted by ΔF here, is allowed. This law holds true for asymptotically large dimensional resources; see [58].
In Sec. VII we discuss repeated operations and the asymptotic regime of the two-qubit machine, focusing on the single-cycle regime. We then investigate the cooling performance and associated work cost of the two-qubit machine augmented with the ability to individually rethermalize the machine qubits to the environment temperature . We then discuss the saturation of the second law by more general machines in Sec. VIII before concluding in Sec. IX.

In the asymptotic regime scenario 1 corresponds to the minimal autonomous quantum thermal refrigerator, as shown in Ref. [59] and discussed in our accompanying article [55]. Scenario 2 leads to heat bath algorithmic cooling, when augmented with the ability to individually rethermalize the machine qubits to the environment temperature . Moreover, like in the single-cycle regime, scenario 2 always achieves a lower temperature, although generally at a higher work cost.

While minimal machines saturate the cooling bounds, they do so in a very ineffective way from a work cost perspective. Extending our analysis to the case of -qubit machines, by considering cooling to a fixed target temperature, we finally work out the two scenarios for the two-qubit machine case for all our relevant operations, the target has a well defined temperature and this is a valid way to extract the ground-state population and inverting the relation (3). When the target qubit is diagonal, which will turn out to be the case for all our relevant operations, the target has a well defined temperature and this is a valid way to extract the ground-state population and inverting the relation (3). When the target qubit is diagonal, which will turn out to be the case for all our relevant operations, the target has a well defined temperature and this is a valid way to extract the ground-state population and inverting the relation (3). When the target qubit is diagonal, which will turn out to be the case for all our relevant operations, the target has a well defined temperature and this is a valid way to extract the ground-state population and inverting the relation (3).

For finite repetitions, additional cooling starts from the end points of maximal single-cycle cooling in each scenario. For scenario 1, one can think of this as repeated thermal operations with a locality restriction, i.e., access to a single qubit from each of the two baths in every round, and for scenario 2 it corresponds to multiple cycles of coherently driven quantum machines (such as, e.g., quantum Otto cycles).

In the asymptotic regime scenario 1 corresponds to the minimal autonomous quantum thermal refrigerator, as shown in Ref. [59] and discussed in our accompanying article [55]. Scenario 2 leads to heat bath algorithmic cooling, when augmented with the ability to individually rethermalize the machine qubits to the environment temperature . Moreover, like in the single-cycle regime, scenario 2 always achieves a lower temperature, although generally at a higher work cost.

While minimal machines saturate the cooling bounds, they do so in a very ineffective way from a work cost perspective. Extending our analysis to the case of -qubit machines, by considering cooling to a fixed target temperature, we finally show that both coherent and incoherent machines can achieve minimal work cost, i.e., saturate the second law, in the limit of large size.

The rest of the paper is organized as follows. In Sec. III we introduce notation and definitions. Section IV deals with the case of the one-qubit machine. In Secs. V and VI we investigate the cooling performance and associated work cost of the two-qubit machine, focusing on the single-cycle regime. In Sec. VII we discuss repeated operations and the asymptotic regime of the two-qubit machine. We then discuss the saturation of the second law by more general machines in Sec. VIII before concluding in Sec. IX.

III. NOTATION AND DEFINITIONS

As argued in Sec. II, we consider machines consisting of a given number of qubits. We take the energies of all qubit ground states to be zero, denote the excited state energy of qubit by , and the energy eigenstates by , , . Thus, the local Hamiltonian for each qubit is , , and the total Hamiltonian of target and machine is

\[ H = \sum_i E_i |1\rangle\langle 1|, \]  

(1)

The initial state, prior to cooling, is the same for the incoherent and coherent scenarios. Every qubit is in a thermal state of its local Hamiltonian at the environment temperature . In general, a thermal state of a qubit with energy gap and temperature is given by

\[ \tau(\epsilon, T) = r(\epsilon, T) |0\rangle\langle 0| + [1 - r(\epsilon, T)] |1\rangle\langle 1|, \]  

(2)

where the populations are determined by the Boltzmann distribution (throughout the paper we work in natural units, )

\[ r(\epsilon, T) = \frac{1}{1 + e^{-\epsilon/T}} = \frac{1}{Z(\epsilon, T)}, \]  

(3)

where \( Z(\epsilon, T) \) is the partition function corresponding to the qubit Hamiltonian and temperature.

We denote the ground-state populations at the environmental temperature by , , , and the corresponding thermal states by . We will refer to the target to be cooled as qubit , but for convenience, we will generally drop the subscript for the target qubit, such that

\[ E := E_A, \quad r := r_A, \quad \tau := \tau_A. \]  

(4)

Note that we can choose a unit of energy such that \( E = 1 \) without loss of generality, which we do for all our numerical analysis.

In scenario 1, one (or more) of the machine qubits is first heated to a higher temperature . This is followed by an energy-conserving unitary acting jointly on the target and the machine, i.e., any unitary \( U \) for which \( [U, H] = 0 \). In scenario 2, an energy-nonconserving unitary is applied directly to the initial state of target and machine.

We extract the temperature of the target qubit by reading its ground-state population and inverting the relation (3). When the target qubit is diagonal, which will turn out to be the case for all our relevant operations, the target has a well defined temperature and this is a valid way to extract it. When the target state is not diagonal it strictly speaking has no temperature. One way to nevertheless extend the notion of temperature to these states also is as presented above. The work cost is accounted for from the perspective of the work reservoir, i.e., the free energy change of the resource. This is not necessarily equal to the free energy change of the system itself, but is nonetheless the appropriate way to quantify consumed resources. For completeness, we have also worked out the two scenarios for the two-qubit machine case from a system perspective in Appendix F.
IV. ONE-QUBIT MACHINE

Denoting the machine qubit by \( B \), the Hamiltonian is \( H = H_A + H_B \), and the initial state is simply
\[
\rho^{in} = \tau \otimes \tau_B. \tag{5}
\]

A. Scenario 1: Incoherent operations

In this scenario, the machine qubit is first heated to a higher temperature \( T_B \), resulting in the state
\[
\rho = \tau \otimes \tau_B^H, \tag{6}
\]
where \( \tau_B^H = \tau(E_B, T_B) \) is the thermal state of qubit B at the temperature of the hot bath. This is followed by an energy-conserving unitary. However, there is no such unitary that can cool the target, as we demonstrate now.

For the action of the unitary to be nontrivial (and hence, for any cooling of the target to happen), the spectrum of the joint Hamiltonian \( H \) must have some degeneracy, allowing one to shift population between distinct energy eigenstates of the same energy. The only possibilities are that (1) one of the energies vanish \( E = 0 \) or \( E_B = 0 \) or (2) the gaps are equal \( E = E_B \). In case (1) the thermal state \( \rho^H \) will be proportional to the identity in the degenerate subspace, and hence \( U \rho^H U^\dagger = \rho^H \) for any energy-conserving unitary \( U \). In case (2) because the matrix elements (in the product basis of \( H_A, H_B \)) fulfill
\[
\rho_{00,00}^H = r(1 - r_B^H) > r_B^H(1 - r) = \rho_{10,10}^H, \tag{12}
\]
unitalies acting on the degenerate subspace can only heat up the target.

Thus, for the single-qubit machine, cooling is impossible in the incoherent scenario.

B. Scenario 2: Coherent operations

Using coherent operations, it is possible to cool, and we now derive the minimal attainable temperature of the target, and the work cost of cooling. This will also provide some intuition for how to tackle the two-qubit machine, where coherent and incoherent cooling can be compared.

Cooling corresponds to increasing the ground-state population of the target using an arbitrary joint unitary \( U \) on target and machine. This population \( r_{coh} \) is given by the sum of the two first diagonal entries of the final state \( U \rho^{in} U^\dagger \), when expressed in the product basis of \( H_A \otimes H_B \). From the Schur-Horn theorem one learns that this sum can at most be the sum of the two greatest eigenvalues of \( U \rho^{in} U^\dagger \), which, since \( U \) cannot change the eigenvalues of the state and since \( \rho^{in} \) is diagonal, are the sum of the two largest diagonal entries of \( \rho^{in} \). Maximal cooling is thus achieved when \( r_{coh} \) equals the sum of the two largest diagonal entries of \( \rho^{in} \). One readily sees that \( \rho^{in}_{00,00} = r_B \) is the largest element and \( \rho^{in}_{11,11} = (1 - r)(1 - r_B) \) the smallest, while
\[
\frac{\rho^{in}_{00,00}}{\rho^{in}_{10,10}} = \frac{r(1 - r_B)}{(1 - r)r_B} = e^{-s_B}. \tag{7}
\]

Cooling is only possible if the initial ground-state population \( r = \rho^{in}_{00,00} + \rho^{in}_{01,01} \) is not already maximal, i.e., if \( E < E_B \). In this case, the maximal final population is \( r_{coh}^* = \rho^{in}_{00,00} + \rho^{in}_{10,10} = r_B \) corresponding to [from (3)]
\[
T^*_{coh} = \frac{E}{\ln \left( \frac{r_B}{1 - r_B} \right)} = \frac{E}{E_B}T_B. \tag{8}
\]

This temperature can be achieved by a unitary which swaps the states \( |01\rangle \) and \( |10\rangle \), and in fact this also minimizes the associated work cost. More generally, we can identify an optimal unitary which minimizes the work cost of cooling to any temperature in the attainable range, i.e., any ground-state population \( r_{coh} \) between \( r \) and \( r_{coh}^* \). The optimal work cost is given by
\[
\Delta F_{coh} = (r_{coh} - r)(E_B - E), \tag{9}
\]
and it is achieved by a unitary of the form \( U = e^{-u L} \).

\[
L = i(01\rangle\langle 10) - i(10\rangle\langle 01) \tag{11}
\]
is a Hamiltonian which generates swapping of excitations between the target and machine qubits, and \( t = \arcsin(\sqrt{\mu}) \) with
\[
\mu = \frac{r_{coh} - r}{r_B - r} \tag{12}
\]
The optimality of (9) can be proven using the Shur-Horn theorem and majorization \([60,61]\). The idea of the proof is as follows. By scanning through all the unitarily attainable \( \rho^{coh} = U \rho^{in} U^\dagger \) we are looking at all the Hermitian matrices \( \rho^{coh} \) with spectrum \( \bar{\rho}^{in} \) (given an \( n \times n \) matrix \( \mu = (\mu_{ij}) \) we generically denote its vectorized diagonal \( (\mu_1, \ldots, \mu_{nn}) \) by \( \vec{\mu} \)). According to the Schur-Horn theorem there exists a Hermitian matrix \( \rho^{coh} \) with spectrum \( \bar{\rho}^{coh} \) and if only if the majorization condition \( \bar{\rho}^{coh} \prec \bar{\rho}^{in} \) holds. Hence, a state \( \rho^{coh} \) is reachable by a unitary starting from \( \rho^{in} \) if and only if the diagonals fulfill \( \bar{\rho}^{coh} \prec \bar{\rho}^{in} \). In the coherent scenario, the free energy difference and hence the work cost are simply given by
\[
\Delta F = \Tr[(\rho^{coh} - \rho^{in})H] = (\bar{\rho}^{coh} - \bar{\rho}^{in}) \cdot \vec{H}, \tag{13}
\]
where \( \vec{H} \) is the diagonal of \( H \). The last term is constant, and so minimizing the work cost for a given final \( r_{coh} \) is equivalent to
\[
\min_{\bar{\rho} < \bar{\rho}^{in}} \bar{\rho} \cdot \vec{H} \quad \text{s.t.} \quad \rho_1 + \rho_2 = r_{coh}. \tag{14}
\]
As shown in Appendix C, this minimization can be solved analytically, leading to (9) and (10).

V. TWO-QUBIT MACHINE: MODEL

When considering the two-qubit machine, the total Hamiltonian of target and machine is \( H = H_A + H_B + H_C \), with qubits B and C forming the machine. The setup, as well as the two scenarios, are illustrated in Fig. 1. The starting point for both scenarios 1 and 2 is the initial state
\[
\rho^{in} = \tau \otimes \tau_B \otimes \tau_C. \tag{15}
\]
In scenario 2, an energy nonconserving unitary is applied directly to the initial state \( \rho^\text{in} \), while in scenario 1, qubit C is first heated to a higher temperature \( T_H \), resulting in the state

\[
\rho^H = \tau \otimes \tau_B \otimes \tau_C^H,
\]

where \( \tau^H_C = \tau(E_C, T_H) \) is the thermal state of qubit C at the temperature of the hot bath. This is followed by an energy-conserving unitary acting on the three qubits. To allow for nontrivial energy-conserving unitaries, there must be a degeneracy in the spectrum of \( H \) with an associated degenerate subspace. In Appendix A we show that the only degeneracy which enables cooling of the target is obtained by setting

\[
E = E_B - E_C.
\]

Hence, we work with this convention throughout the following.

VI. TWO-QUBIT MACHINE: SINGLE-CYCLE REGIME

In this section, we discuss the single-cycle regime of the two-qubit machine. We show that scenario 2 (coherent operations) always reaches lower temperatures when the work cost is unrestricted. However, for sufficiently low work cost, it turns out that scenario 1 (incoherent operations) outperforms scenario 2.

A. Scenario 1: Incoherent operations

We first identify the energy-conserving unitary that is optimal for cooling the target qubit. From the relation (17) it follows that there is only one subspace that is degenerate in energy (relevant for cooling), which is spanned by the states \(|010\rangle\) and \(|101\rangle\). Optimal cooling is simply achieved by swapping these two states, i.e., the unitary is given by (see Appendix A for more details)

\[
U = |010\rangle\langle101| + |101\rangle\langle010| + I_{\text{non-deg}},
\]

where \( I_{\text{non-deg}} \) is the identity operation on the complement space. We can thus directly compute the final temperature of the target qubit. We first compute the final ground-state population \( r_{\text{inc}} \)

\[
r_{\text{inc}}(T_H) = r r_B + [(1-r) r_B + r (1-r_B)](1-r_C^H),
\]

where \( r_C^H = r(E_C, T_H) \) denotes the ground-state population of qubit C after heating and \( r \) and \( r_B \) denote the ground-state populations of the target qubit and qubit B at room temperature \( T_B \). The final temperature is found by inverting Eq. (3):

\[
T_{\text{inc}}(T_H) = \frac{E}{\ln \left( \frac{r_{\text{inc}}}{1-r_{\text{inc}}} \right)}. \tag{20}
\]

Not limiting the work cost, optimal cooling is obtained in the limit \( T_H \rightarrow \infty \). In this case \( r_C^H = \frac{1}{2} \), and thus

\[
r^*_{\text{inc}} = \lim_{T_H \rightarrow \infty} r_{\text{inc}}(T_H) = \frac{1}{2} (r + r_B). \tag{21}
\]

We thus obtain the lowest achievable temperature for scenario 1:

\[
T^*_{\text{inc}} = \lim_{T_H \rightarrow \infty} T_{\text{inc}}(T_H) = \frac{E}{\ln \left( \frac{r+ r_B}{2- (r+r_B)} \right)}. \tag{22}
\]

We are now interested in the work cost of cooling. For scenario 1, the hot bath is the only resource, implying that the free energy decrease in the hot bath represents the cooling cost. The free energy difference is \( \Delta F = \Delta U - T_H \Delta S \), where \( \Delta U \) is the internal energy change. For a thermal bath \( \Delta U \) is defined as the heat drawn from the bath, \( Q \), which from the first law equals the change in energy of qubit C. We follow the convention of counting as positive what is taken from the bath. The change in entropy \( \Delta S \) also takes a simple form for a thermal bath, \( \Delta S = Q / T_H \). This gives

\[
\Delta F_{\text{inc}}(T_H) = Q \left( 1 - \frac{T_B}{T_H} \right) = E_C \left( r_C - \frac{r_C^H}{2} \right). \tag{23}
\]

The above equation shows that the work cost is determined directly by the hot bath temperature \( T_H \). The work cost associated to maximal cooling is given by

\[
\Delta F^*_{\text{inc}} = \lim_{T_H \rightarrow \infty} \Delta F_{\text{inc}}(T_H) = E_C \left( r_C - \frac{1}{2} \right). \tag{24}
\]

Note that despite appearances, the above expression is not independent of \( E_B \), as the machine qubits are mutually constrained by the degeneracy condition (17).

More generally, as the ground-state population \( r_{\text{inc}} \) is monotonic in \( r_C^H \) [see Eq. (19)] and thus in \( T_H \), one can cool to any temperature between \( T_B \) and \( T^*_{\text{inc}} \) by varying \( T_H \) continuously between \( T_B \) and infinity. The associated work cost is given by Eq. (23); see Fig. 3.
Note that the minimum achievable temperature in this scenario is lower bounded away from absolute zero. Taking the limits $T_H \to \infty$ and then $E_B \to \infty$, $r_{\text{inc}}$ tends to $(1 + r)/2$. The work cost diverges in this limit. This is in contrast to scenario 2 presented in the following section, where for an unbounded work cost, one can cool arbitrarily close to absolute zero.

### B. Scenario 2: Coherent operations

We now turn to the second scenario, where any joint unitary operation can be applied to the target and machine qubits. The freedom in unitary operation means that the resonance condition $E_B = E + E_C$ is in principle not required to allow cooling, in contrast to scenario 1. However, as the cooling in either scenario depends on the choice of machine qubits, the freedom to choose them represents an extra level of control. In order to make a meaningful comparison between coherent and incoherent operations, we will therefore enforce the resonance condition for scenario 2 as well.

We first investigate the lowest achievable temperature. By definition this is obtained by maximizing the ground-state population of the target qubit. If we express the state of all three qubits as a density matrix population of the target qubit, if we denote the populations whose ordering depends on $\{|000\}, |001\>, |010\>, |101\>\}$. Making use of the Schur-Horn theorem as argued in Sec. IV B one reaches optimal cooling by unitarily rearranging the populations such that the four largest populations of the initial state are mapped to the four levels contributing to the ground-state population of the target. Labeling the population of the state $|i j k\>$ in the corresponding initial density operator $\rho_{\text{in}}$ by $p_{ijk}$, and arranging them in decreasing order of magnitude, we find

$$p_{000} > |p_{001}, p_{100}\> > p_{100} = p_{011},$$

where $\{\}$ denotes populations whose ordering depends on whether $E_C > E$ or $E_C < E$. Thus the only change necessary to optimize cooling is to swap the populations of $|100\>$ and $|011\>$, and this leads to a final ground-state population of $r_{\text{coh}}^* = r_B$, corresponding to the remarkably simple final temperature

$$T_{\text{coh}}^* = T_B \frac{E}{E_B}.$$  

This is the lowest achievable temperature in scenario 2, when the work cost is unrestricted.

We now turn to the question of optimizing the work cost. Indeed, on inspection of the end point of the above procedure, one finds that within the ground and excited subspace of the target qubit, one can perform unitaries that rearrange populations without affecting cooling, but that extract energy back from the system, hence decreasing the work cost of the cooling procedure.

We illustrate this subtlety with the end point of the simple swap above. The only modified populations after the swap are those of the states $|100\>$ and $|011\>$. Denoting the new population of energy level $|i j k\>$ by $p'_{ijk}$, we have that $p'_{011} = p_{100}$ and $p'_{100} = p_{011}$, with the rest unchanged. Thus the new ordering is

$$p_{000} > |p_{001}, p_{100}\> > p_{100} = p_{011},$$

Although the ground-state population is maximized by this swap, one sees that its energy is not minimal, since, e.g., $p_{011} > p_{010}$. As a consequence, one could now extract energy without changing the ground-state population by simply swapping the levels $|011\>$ and $|010\>$. Formally, this implies that within each subspace of the target qubit (ground and excited), the state is not passive [17], i.e., the populations are not ordered in decreasing order in energy within each subspace. This showcases the general fact that the state the optimal unitary drives the system to, necessarily has to be passive within each of these subspaces. In the maximal cooling case, as shown in Appendix C, this passivity condition remarkably turns out to also be sufficient. If one thus follows performing the unitary that reorders each subspace to be passive, and subtracts the energy extracted from the work cost, we arrive at the optimal work cost corresponding to maximal cooling, $\Delta F_{\text{coh}}^*$.

We find that there are two cases. If $E_C \leq E$, then

$$\Delta F_{\text{coh}}^* = E_C(r_B - r).$$

Note that this end point can be achieved by simply performing the unitary that swaps the states of qubits $A$ and $B$. On the other hand, if $E_C > E$, then

$$\Delta F_{\text{coh}}^* = (E_C - E)(r_C - r) + E_C(r_B - r_C).$$

The unitary that achieves this result is the sequence of two swaps: first, the swap between the target and qubit $C$, followed by the swap between the target and qubit $B$.

Remarkably, these two expressions can be intuitively understood in the following manner. In order to achieve cooling on the target qubit, one would swap its state with a qubit of the machine (or qubit subspace from the machine, also called a “virtual qubit” [31] (see Appendix G) that has a larger energy gap between its ground and excited states than the target qubit. However, doing so requires moving population against the energy difference between the target and the specific machine qubit. Minimizing the work cost of the cooling procedure therefore amounts to swapping the state of the target qubit with the state of the machine qubit with the minimal energy gap as long as this one is bigger than the energy $E$ of the target qubit.

If $E_C \leq E$, then the smallest qubit subspace of the machine that has a higher energy gap than the target is qubit $B$, and the optimal procedure is to swap the states of these two qubits. This has a work cost $E_B - E = E_C$ per population. In contrast, when $E_C > E$, then qubit $C$ is the machine qubit with the smallest energy gap bigger than $E$ ($E_C < E_B$ by definition). We thus begin by swapping the target qubit with qubit $C$, at a work cost per population of $E_C - E < E_C$, and only after proceed to cool further by swapping the target qubit with qubit $B$, at higher work cost. This two cases, respectively, lead to Eqs. (28) and (29) when the work cost is unrestricted.

We now move to the case where the work cost is restricted. Equivalently, we consider the problem of cooling to a certain temperature (above $T_{\text{coh}}^*$), and derive the minimal associated work cost. Intuitively, as the lowest temperature given by
Eq. (26) can be reached by a full swap (or a sequence of two full swaps if $E_C > E$), we might expect that an optimal strategy for reaching an intermediate temperature will be a partial swap.

This is indeed the case. In analogy with (30), the minimal work cost for a given target temperature $T_{coh}$ and corresponding ground-state population $r_{coh}$ is given by

$$\min_{\hat{\rho} \to \rho^{ib}} \hat{\rho} \cdot \hat{H} \quad \text{s.t.} \quad \sum_{i=1}^{4} \rho_i = r_{coh},$$

(30)

where $\hat{\rho}$, $\rho^{ib}$, and $\hat{H}$ represent the diagonals of $\rho$, $\rho^{ib}$, and $H$. This minimization can be solved analytically, as shown in Appendix C. The optimal unitary and associated work cost depends on whether $E_C \leq E$ or $E_C > E$.

For the case $E_C \leq E$, we can parametrize a partial swap of the target with machine qubit B as in (11) by

$$U_{\mu}(\mu) = e^{-iL_{AB}},$$

(31)

where

$$L_{AB} = i|01\rangle\langle 01| - i|10\rangle\langle 10|.$$  

(32)

It is useful to define $t = \arcsin(\sqrt{\mu})$, where $\mu \in [0, 1]$ is a swapping parameter. The ground-state population of the target qubit and the free energy cost are given by

$$r_{coh, <}(\mu) = r + \mu (r_B - r),$$

(33)

$$\Delta F_{coh, <}(\mu) = \mu E_C (r_B - r),$$

(34)

with $\mu = 0$ corresponding to no swap and $\mu = 1$ to a full swap, which is the limit of maximal cooling, as previously discussed; see Eqs. (26) and (28).

Similarly, for the case $E_C > E$ we employ the unitary that first swaps qubits A with $C$ until the required temperature is reached, and if this is not the case after the full swap, continue by swapping the new state of qubit A with qubit B. This unitary can be parametrized as

$$U_{\mu}(\mu) = e^{-i\mu L_{AB}} e^{-i(1-\mu)L_{AC}},$$

(35)

where $f(\mu) = \arcsin(\sqrt{\mu})$, $g(\mu) = \arcsin(\sqrt{\max(2\mu - 1, 0)})$, and $L_{AC}$ is defined analogously to Eq. (32). Again $\mu \in [0, 1]$ such that for $\mu \leq \frac{1}{2}$, a partial swap between $A$ and $C$ is performed and for $\frac{1}{2} < \mu \leq 1$, an additional partial swap between $A$ and $B$ is performed.

The ground-state population for the strategy defined by $U_{\mu}$ is

$$r_{coh, >}(\mu) = \begin{cases} r + 2\mu (r_C - r), & \mu \in [0, \frac{1}{2}] \\ r_C + (2\mu - 1)(r_B - r_C), & \mu \in (\frac{1}{2}, 1] \end{cases},$$

(36)

and the work cost for the same strategy is given by

$$\Delta F_{coh, >}(\mu) = \begin{cases} 2\mu (E_C - E)(r_C - r), & \mu \in [0, \frac{1}{2}] \\ (E_C - E)(r_C - r), & \mu \in (\frac{1}{2}, 1] \end{cases}.$$  

(37)

The final temperature can again be computed by inverting Eq. (3) using the ground-state population $r_{coh}$ as given by Eq. (33) or Eq. (36) according to the relative size of $E$ and $E_C$. Since both $\Delta F_{coh}$ and $T_{coh}$ are given as functions of $\mu$, by varying $\mu$ from 0 to 1, we can parametrically map out the amount of cooling and the associated work cost, as shown in Fig. 3 and discussed in Sec. V1C.

**C. Comparison of scenarios 1 and 2**

Our main results in the single-cycle regime are summarized in Fig. 3. There we map out the amount of cooling versus the associated work cost for both scenarios 1 and 2. In the first case, the curve is generated from Eqs. (19) and (23) [inverting Eq. (3) to extract $T_{inc}$] and is parametric in the hot bath temperature $T_H$. In the second case, the curve is generated from Eqs. (34) and (33) [inverting Eq. (3) to extract $T_{coh}$] and is parametrized by the swapping parameter $\mu$. We selected $E_C \leq E$ for Fig. 3 but note that the behavior of the curve for $E_C > E$ is similar, changing only by the fact that the coherent curve has a discontinuity in the first derivative at $\mu = \frac{1}{2}$.

The plot illustrates several interesting observations. First, comparing the endpoints of the curves, we see that coherent operations achieve a lower minimal temperature (i.e., stronger cooling) and that the associated work cost is lower than the one for achieving the minimal temperature with incoherent operations. This is true generally. As can be seen by comparing Eqs. (22) and (26), $T_{coh}^* < T_{inc}^*$ since

$$\ln \left[ \frac{r + r_B}{2 - (r + r_B)} \right] < \frac{E_B}{T_R},$$

(38)

where we use that $E_B > E$. Similarly, comparing Eqs. (24) and (28), we see that $\Delta F_{coh}^* < \Delta F_{inc}^*$; see Appendix D. Thus, for maximal cooling, coherent operations always perform better than incoherent ones in the single-cycle regime.

Second, perhaps surprisingly, for nonmaximal cooling with low work cost, incoherent operations may outperform coherent ones. In fact, for sufficiently low work cost, this is always the case. This can be seen by looking at the derivatives of the two curves in Fig. 3 with respect to $\Delta F$, close to $\Delta F = 0$. For the incoherent scenario, using the parametrization w.r.t. $T_H$, we have

$$\lim_{\Delta F_{inc} \to 0} \frac{dT_{inc}}{d\Delta F_{inc}} = \lim_{\Delta F_{inc} \to 0} \frac{dT_{inc}}{dT_H} \left( \frac{d\Delta F_{inc}}{dT_H} \right)^{-1} = -\infty.$$  

(39)

On the other hand, for the coherent scenario, using the parametrization in terms of $\mu$, we find that

$$\lim_{\Delta F_{coh} \to 0} \frac{dT_{coh}}{d\Delta F_{coh}} = -\frac{1}{E_C^* r (1 - r) \ln^2 \left( \frac{\mu^*}{\mu} \right)},$$

(40)

where $E_C^* = E_C$ for $E_C \leq E$ and $E_C^* = E_C - E$ if $E > E_C$. This expression is negative but finite. Hence, since both curves begin at the same point, the incoherent curve must lie below the coherent one for sufficiently small $\Delta F$. From the previous observations, it follows that the curves must cross at least once. Numerically we find that there is always exactly one such crossing. Hence, there exists a critical work cost $\Delta F_{crit}^*$ below which incoherent operations perform better than coherent ones, while the reverse is true above some $\Delta F_{crit}^* \geq \Delta F_{crit}$. With $\Delta F_{crit} = \Delta F_{crit}^*$ numerically strongly supported to be true. We denote the temperature of the target qubit at the crossing point by $T_{crit}$. In Appendix E we study the behavior of $T_{crit}$ and $\Delta F_{crit}$ as functions of $T_R$ and $E_C$.  

042130-8
VII. TWO-QUBIT MACHINE: REPEATED OPERATIONS AND ASYMPTOTIC REGIME

In this section we go beyond the single-cycle regime discussed above. In the repeated and asymptotic regimes, the cooling unitaries of either scenario can be repeated a finite number of times or indefinitely, interspaced by steps in which the machine qubits (B and C) are rethermalized to the temperatures of their baths, i.e., respectively, $T_B$ and $T_C$ in scenario 1 and $T_R$ for both machine qubits in scenario 2. The target qubit is assumed not to rethermalize during the cooling process. In this way, the bounds we obtain on achievable temperature and work cost are general. Moreover, these bounds can be attained in the limit where the thermal coupling of the target qubit is much smaller than other couplings in the system.

Before going into detail, we first summarize the main results of this section.

(1) Repeated operations do enhance the cooling, as the lowest achievable temperatures in both scenarios are strictly lower than in the single-cycle case.

(2) For incoherent operations (scenario 1), the asymptotic regime (the limit of infinite repetitions) corresponds to autonomous refrigeration. Specifically, we recover the cooling and work cost obtained in the steady state of a three-qubit autonomous refrigerator [29,62].

(3) For coherent operations (scenario 2), the asymptotic regime corresponds to algorithmic cooling. In particular, the cooling bounds correspond to known results [24,25].

(4) In the asymptotic regime, incoherent operations (scenario 1, autonomous cooling) achieve the same maximal cooling (for $T_H \to \infty$) as that of a single-cycle coherent operation (scenario 2). See our accompanying article [55] for more details on this relation.

(5) In both scenarios, the approach to the asymptotic state of the target qubit (w.r.t. its ground-state population) is exponential in the number of repetitions.

In the following, we will start by discussing repeated operations in scenario 1 and then move to scenario 2.

A. Scenario 1: Repeated incoherent operations

As mentioned above, the scenario of repeated incoherent operations involves a rethermalization of the machine qubits to their respective baths in every step. This is followed by an energy-conserving unitary operation between the machine and their respective baths, i.e., respectively, $TR$ and $TH$ in every step. This is followed by an rethermalization of the machine qubits (B and C) in every step. This is followed by an rethermalization of the machine qubits (B and C) in every step.

Before going into detail, we first summarize the main results of this section.

(1) Repeated operations do enhance the cooling, as the lowest achievable temperatures in both scenarios are strictly lower than in the single-cycle case.

(2) For incoherent operations (scenario 1), the asymptotic regime (the limit of infinite repetitions) corresponds to autonomous refrigeration. Specifically, we recover the cooling and work cost obtained in the steady state of a three-qubit autonomous refrigerator [29,62].

(3) For coherent operations (scenario 2), the asymptotic regime corresponds to algorithmic cooling. In particular, the cooling bounds correspond to known results [24,25].

(4) In the asymptotic regime, incoherent operations (scenario 1, autonomous cooling) achieve the same maximal cooling (for $T_H \to \infty$) as that of a single-cycle coherent operation (scenario 2). See our accompanying article [55] for more details on this relation.

(5) In both scenarios, the approach to the asymptotic state of the target qubit (w.r.t. its ground-state population) is exponential in the number of repetitions.

In the following, we will start by discussing repeated operations in scenario 1 and then move to scenario 2.

A. Scenario 1: Repeated incoherent operations

As mentioned above, the scenario of repeated incoherent operations involves a rethermalization of the machine qubits to their respective baths in every step. This is followed by an energy-conserving unitary operation between the machine and the target. Thus, the cooling cycle consists in the following steps (see Fig. 4), which can be repeated any number of times.

1. Environment reset and resource input. Qubit C is heated to $T_H$ after the machine has been brought back to the environment temperature $T_R$.

2. Cooling step. The energy-preserving unitary operation given by Eq. (18) (swapping the degenerate states $|10\rangle \leftrightarrow |01\rangle$) is applied.

Prior to the first step, all three qubits are at temperature $T_R$. Then qubit C is heated to $T_H$. After this, every cooling step lowers the temperature of the target qubit A, but also cools down qubit C while heating qubit B, which necessitates the reset of B to $T_R$ and the heating of C to $T_H$ before the swap can be repeated. This process can be conveniently characterized using the notion of a virtual qubit [31]. The virtual qubit corresponds to the subspace of the machine which is involved in the cooling swap with the target qubit. See Appendix G and Appendix H for a detailed explanation. It is thus the properties of the virtual qubit that determine the cooling in each step. For the unitary operation here, the virtual qubit is spanned by the states $|01\rangle_{BC}, |10\rangle_{BC}$. In each repetition, the rethermalization of qubits B and C (Step 1) resets the virtual qubit.

In the asymptotic limit of infinite repetitions, we find that the ground-state population of the target qubit goes to (see Appendix H)

$$r_{inc,\infty} = \frac{1}{1 + e^{-E_{inc,\infty}}},$$

where $T_{inc,\infty}$ is equal to the temperature of the virtual qubit,

$$T_{inc,\infty} = T_{V,inc} = \frac{E}{E_B - E_C}.$$  \hspace{1cm} (42)

For a finite number $n$ of repetitions, the ground-state population of the target qubit approaches the asymptotic value as

$$r_{inc,n} = r_{inc,\infty} - (r_{inc,\infty} - r)(1 - N_{V,inc})^n,$$ \hspace{1cm} (43)

where $N_{V,inc} = T_{inc}-T_{R}+(1-T_{inc})$ is the norm of the virtual qubit (i.e., the total population in the subspace $|01\rangle_{BC}, |10\rangle_{BC}$). Note that all of the quantities in the above expressions are functions of $T_H$.

As argued also in Appendix I, the asymptotic temperature given by Eq. (42) is exactly equal to the temperature obtained in the steady state of an autonomous refrigerator [29], and thus the asymptotic state of the target qubit under repeated incoherent operations is the same as the steady state of the autonomous fridge. More precisely,

$$r_{inc,\infty} = r_{auto}, \quad \text{i.e.,} \quad T_{inc,\infty} = T_{auto}.$$ \hspace{1cm} (44)

This highlights an interesting connection between discrete and continuous cooling procedures; see also Ref. [59].

Furthermore, showcasing one of the result of our accompanying article [55], the maximal cooling in either case, obtained in the limit $T_H \to \infty$, is the same as for a single-cycle coherent operation [cf. Eq. (26)]

$$T_{auto}^* = \lim_{T_H \to \infty} T_{auto} = \frac{E_B}{E_B - E_C}.$$ \hspace{1cm} (45)
Note that in this limit we have that $N_{V_{inc}} = \frac{1}{2}$. Hence in each repetition the difference between the current and asymptotic ground-state population is halved.

Finally, we discuss the work cost of cooling. Detailed calculations are given in Appendix H. Intuitively, the free energy drawn from the hot bath can be divided into two parts: (1) the energy required in the first instance of step 1, to initially heat up qubit C to temperature $T_H$, and (2) the energy required in all subsequent repetitions of step 1, to bring qubit C back to $T_H$. This is straightforwardly calculated from the change in population of qubit C, which is equal to the change in population of qubit A, due to the form of the energy-preserving unitary in step 2. The total heat drawn from the hot bath for $n$ repetitions is

$$Q^H_n = E_C(r_C - r_C^H) + E_C(r_{inc,n-1} - r). \quad (46)$$

In the asymptotic case, we find that the total heat drawn from the hot bath is exactly the same as if we had run the autonomous refrigerator beginning from the initial state, i.e., $Q^H_{\infty} = Q^H_{auto}$. See Appendix I for a detailed proof.

In order to cool to a given temperature, it is possible to vary the number of repetitions as well as the temperature of the hot bath $T_H$. One may therefore ask which is the most cost-efficient strategy. Generically, we observe (see Fig. 5) that for a given final temperature, implementing many cooling swaps has a lower work cost than using fewer swaps (at higher temperature $T_H$). As implementing a higher number of swaps would take longer time, this observation is reminiscent of the power versus efficiency trade-off in continuously operated machines [63].

**B. Scenario 2: Repeated coherent operations**

When discussing single-cycle cooling via coherent operations in Sec. VI B, we found that according to the relative size of $E_C$ and $E_A$, there were two different sets of unitaries which lead to the lowest achievable temperature $T_{coh}^*$ of the target qubit. The first procedure involved only qubits A and B, and maximal cooling could be achieved with a single-qubit machine (i.e., without qubit C). This procedure was found to be optimal when $E_C \leq E_A$. However, although this procedure was also valid when $E_C > E_A$ we showed that in this case a different procedure, involving all three qubits, could reach the same temperature, but at a lower work cost.

In the present section we discuss cooling via repeated coherent operations. We find that after the first cycle a procedure similar to the second procedure in the single-cycle case must be applied in order to cool further. In fact, one can immediately see that for a single-qubit machine, repetitions do not lower the temperature further beyond the single-cycle case. Since the single-qubit machine simply swaps qubits A and B, there is no unitary operation that can cool further, even after B is rethermalized to the ambient temperature $T_R$.

On the contrary, using a two-qubit machine one can enhance the cooling beyond the single-cycle case. This is achieved by repeating the following steps (see Fig. 6 and Appendix J for more details):

1. **Environment reset.** Qubits B and C are brought back to the environment temperature $T_R$.

2. **Cooling step.** The unitary swapping the populations of the states $|100\rangle, |011\rangle$ is applied.

As qubit A is cooled by swapping with the subspace $|00\rangle_{BC}, |11\rangle_{BC}$ of the machine, we identify this subspace as the relevant virtual qubit of the machine, and denote its norm as $N_{V_{coh}}$. Following calculations given in Appendix J, one finds that in the asymptotic limit (infinite repetitions), the ground-state population of the target goes to

$$r_{coh,\infty} = \frac{1}{1 + e^{-E/E_{coh,\infty}}}, \quad (47)$$

where the asymptotic temperature takes the simple form

$$T_{coh,\infty} = \frac{T_R}{1 + \frac{E}{E_B + E_C}}. \quad (48)$$

This recovers the result of our accompanying article [55] and the results of heat bath algorithmic cooling with no compression qubit. Note that in the coherent case, the temperature of the virtual qubit is just $T_R$, since both the machine qubits are at $T_R$ after rethermalization. However, due to the swap, the final temperature of the target qubit is not simply the virtual temperature, but rather is modulated by the ratio of energies of
the target and virtual qubits; see Appendix G for more details. This is why maximal cooling in the asymptotic case is attained by picking the virtual qubit of the largest energy gap, which for the two qubit machine is $|00\rangle_{BC}, |11\rangle_{BC}$.

For a finite number $n$ of repetitions, the ground-state population of the target approaches its asymptotic value as

$$r_{coh,n}^* = r_{coh,\infty}^* - (r_{coh,\infty}^* - r)(1 - N_{V,coh})^n. \quad (49)$$

Thus we see that cooling is enhanced compared to the single-cycle case, i.e., $T_{coh,n}^* < T_{coh}^*$. (Note that we use the asterisk here to denote the lowest achievable temperature for a fixed number of repetitions.)

We proceed to discuss the work cost of this process. Note that the optimal work cost of the first coherent operation has already been discussed in Sec. VII B and is denoted by $\Delta F_{coh}^*$. For further repetitions of the steps presented above, free energy is needed to implement the unitary in step 2, as populations of states with different energies are swapped. (Step 1 is free as it involves thermalization of the machine qubits to the environment temperature $T_R$.) The work cost of $n$ full repetitions of the cycle is given by (details in Appendix J)

$$\Delta F_{coh,n} = \Delta F_{coh}^* + 2E_C(r_{coh,n}^* - rB), \quad (50)$$

where $\Delta F_{coh}^*$ is the work cost in the single-cycle regime given by Eq. (28). In the asymptotic regime, the work cost becomes

$$\Delta F_{coh,n} = \Delta F_{coh}^* + 2E_C(r_{coh,n}^* - rB), \quad (51)$$

where $r_{coh,\infty}^*$ is the final ground-state population for the target qubit corresponding to $T_{coh,\infty}^*$. Following the argument expanded in full detail in Appendix J, the steps presented above are the only way to cool the target after the first (optimal) coherent operation, and thus $\Delta F_{coh,n}^*$ represents the minimum work cost given the lowest achievable temperature after $n$ repetitions.

**C. Scenario 2: Algorithmic cooling**

It turns out that even stronger cooling can be obtained, by increasing the level of control compared to the above model of repeated coherent operations, specifically, by allowing for individual rethermalization of each machine qubit separately. This model is equivalent to heat bath algorithmic cooling, this time with a compression qubit, as we will demonstrate shortly. The procedure consists in repeating the following steps, shown schematically in Fig. 7:

1. **Environment reset.** Qubit $B$ is brought back to the environment temperature $T_R$.

2. **Precooling.** The states of qubits $B$ and $C$ are swapped.

3. **Environment reset.** Qubit $B$ is brought back to the environment temperature $T_R$.

4. **Cooling step.** The unitary swapping the populations of the states $|00\rangle \leftrightarrow |01\rangle$ is applied.

As before, the target qubit is swapped with the qubit subspace of the machine that has the highest energy gap, spanned by $|00\rangle_{BC}$ and $|11\rangle_{BC}$. However, because of the precooling step, the virtual temperature of this coldest qubit subspace is decreased, from $T_R$ to

$$T_{V,algo} = T_R - \frac{E_B + E_C}{2E_B}. \quad (52)$$

The final temperature is again determined by the virtual temperature. Following calculations given in Appendix K, in the asymptotic limit of infinite repetitions, the ground-state population of the target qubit tends to

$$r_{algo,\infty}^* = \frac{1}{1 + e^{-E/E_{F,algo}}}, \quad (53)$$

where the asymptotic temperature is given by

$$T_{algo,\infty}^* = T_R \frac{E}{2E_B} = T_{coh}^* \frac{2}{2}. \quad (54)$$

The final temperature is thus half the temperature achieved via single-cycle coherent operations. Note that it is also half of the minimal achievable temperature $T_{auto}$ in the asymptotic incoherent regime. Moreover, since $E_B > E_C$, we see that the lowest achievable temperature of algorithmic cooling is strictly colder than that of repeated coherent operations. It is worth noting that the expression for the minimal temperature of Eq. (54) perfectly matches known results in algorithmic cooling: specifically Eq. (7) of Ref. [24] (for the case of two reset qubits), as well as Eq. (10) of Ref. [25].

For a finite number of repetitions of the above cycle of steps, one finds that the ground-state population of the target approaches $r_{algo,\infty}^*$ as

$$r_{algo,n} = r_{algo,\infty}^* - (r_{algo,\infty}^* - r_0)(1 - N_{V,algo})^n, \quad (55)$$

where $r_0$ is the population of the ground state before the first application of the procedure, and $N_{V,algo}$ is the norm of the virtual qubit $|00\rangle_{BC}, |11\rangle_{BC}$ right before step 4 (i.e., after qubit $C$ has been precooled and qubit $B$ rethermalized).

Finally, we discuss the work cost of this process. Free energy is needed to implement the unitaries in steps 2 and 4, as populations of states with different energies are swapped. Steps 1 and 3 have zero cost, since they involve only the environment bath. As detailed in Appendix K, the work cost after $n$ full repetitions is given by

$$\Delta F_{algo,n} = E(r_B - r_C) + 2EC(r_{algo,n} - r_0) + E(r_{algo,n-1} - r_0). \quad (56)$$
an autonomous refrigerator can cool to any $T \geq T^*_\text{auto}$ more efficiently.

VIII. SATURATING THE SECOND LAW

Upon comparing the cooling performance of the minimal machines presented in this article with the ultimate performance bound set by each paradigm in our accompanying article [55], it is quite striking to notice that as simple as they are, the minimal machines already suffice to saturate the bound. The next natural question to ask is if these machines are also optimal in terms of the associated work cost. We in this section answer this question by the negative.

Clearly, fundamental limitations on the work cost arise from the second law. Specifically, the free energy change of the target qubit is a lower bound on the work cost. Here we present a family of $N$-qubit coherent machines which asymptotically saturate this bound. These machines have been introduced in Ref. [12] for demonstrating optimal work extraction from quantum states. Moreover, for any machine in the family, we construct an incoherent machine of $2N$ qubits achieving the same temperature. In the limit where the hot bath becomes infinite, the associated work cost is the same up to a constant offset that can be made arbitrarily small.

As we have learned from Sec. IV, a given temperature $T$ can be achieved via a single-qubit machine with energy gap $E_N = E\frac{2}{N}$. In order to minimize the work cost, we now introduce $N - 1$ additional qubits with energy gaps (evenly) spaced between $E$ and $E_N$. The single swap is now replaced by a sequence of swaps between the target qubit and machine qubits in order of increasing energy gaps. This can be understood intuitively by noticing that the energy difference $\Delta E$ when swapping two qubits represents the work cost per unit of population transferred $\Delta r$ (see Appendix G)

$$\Delta F = \Delta r \Delta E.$$  

(58)

Hence, for a given final population transfer, replacing one single swap at large $\Delta E$ by $N$ swaps at smaller $\Delta E$ reduces the work cost. As shown in Ref. [12], the total work cost of this procedure is given by

$$\Delta F = \Delta F_{\text{target}} + O\left(\frac{1}{N}\right),$$  

(59)

where $\Delta F_{\text{target}}$ is the increase in the free energy of the target qubit. In the limit $N \to \infty$, the work cost is exactly the free energy transferred to the system, which is the lower bound provided by the second law. See Appendix M for details.

The next question is whether we can find an incoherent machine which also saturates the second law. A first possibility is to transform the above coherent machine into an incoherent one, using the same idea as discussed in our accompanying article [55]. Specifically, each swap can be made energy preserving by adding an extra qubit to the machine. Therefore, the $N$-qubit coherent machine discussed above, can be made incoherent by adding $N$ extra qubits. The temperature achieved by the incoherent machine will match that of the coherent if either $T_H \to \infty$, or if the energies of the machine qubits are increased so as to match the desired temperature on the target qubit. In the former case, the work cost will diverge when $N \to \infty$, as each additional qubit must now be heated.
from $T_R$ to infinite temperature. Nevertheless, in the second case, this problem can be circumvented by noting that these $N$ additional qubits do not need to correspond to physical qubits, but can be taken as virtual qubits. For instance, one can consider a single evenly spaced $(N+1)$-level system providing all these virtual qubits. By embedding this $(N+1)$-level system into a larger system, the initial work cost can be made arbitrarily small, and we can thus approach the work cost of the corresponding coherent machine arbitrarily closely. Consequently, we have constructed an incoherent machine which also saturates the second law in the limit of $N \to \infty$. See Appendix M2 for an explicit construction and proof.

**IX. CONCLUSION AND OUTLOOK**

We have presented a unified view of quantum refrigeration, allowing us to compare various paradigms. In particular, our framework incorporates autonomous quantum thermal machines, algorithmic cooling, single-cycle refrigeration and the resource theory of thermodynamics.

We characterize fundamental limits of cooling, in terms of achievable temperature and work cost, for both coherent and incoherent operations, in single-cycle, finite repetitions, and asymptotic regimes. The main formulas are summarized in Fig. 9.

We find that, contrary to classical thermodynamics, the fundamental limits crucially depend on the level of control available. In particular, this implies that the free energy does not uniquely determine the minimal achievable temperature. Moreover, the size of the machine represents an additional form of control, which also influences thermodynamic performance. On the one hand, for minimal machines, the difference between coherent and incoherent control is strongly pronounced. On the other, in the asymptotic limit, the two scenarios become mostly equivalent.

While we focused here on the task of cooling a single qubit, it is natural to ask what the fundamental limits to cooling larger systems are. Understanding the qubit case already provides significant insight into the general case. For the task of increasing the ground-state population, we showed that qubit bounds apply in general. Repeating every scenario, for every possible notion of cooling would, while possible, not add much insight without a more physical motivation of the respective target Hamiltonian and setting. It would
furthermore be interesting to characterize the work cost of cooling general systems, although this will also depend on the exact Hamiltonian structure of the target, and so one cannot expect to obtain a single-letter formula (as in the qubit case).

Finally, it would indeed be interesting to discuss different tasks than cooling, e.g., work extraction, and determine if similar conclusions can be drawn.

ACKNOWLEDGMENTS

We are grateful to Flavien Hirsch, Patrick P. Hofer, Marc-Olivier Renou, and Tamás Krivachy for fruitful discussions. We would also like to acknowledge the referees of our initial submission for productive and challenging comments. Several authors acknowledge the support from the Swiss NFS and the NCCR QSIT: R.S. through the Grants No. 200021-169002 and No. 200020_165843, N.B. through Grant No. 200021_169002, F.C. through the AMBIZIONE Grant No. PZ00P2_161351, and G.H. through the starting grant PRIMA PR00P2_179748 and Marie-Heim Vöglin Grant No. 164466. M.H. acknowledges support from the Austrian Science Fund (FWF) through the START Project Y879-N27. J.B.B. acknowledges support from the Independent Research Fund Denmark.

APPENDIX A: DEGENERACY CONDITION FOR COOLING

We want here to investigate the conditions for cooling to be possible in the incoherent scenario. First, we will see that given an arbitrary machine, the system Hamiltonian of target and machine together must have some degeneracy in that scenario for cooling to be possible at all with that machine. This is the content of Lemma 1. We will then move on to the specific case of the two-qubit machine and prove that given such a machine, cooling is only possible when $E_B = E_A + E_C$. This is what Lemma 2 proves.

Lemma 1. As $[U, H] = 0$, $U$ can only cool the target by acting on the degenerate eigenspaces of $H$.

Proof. Let $\text{Eig}_H(E)$ be the eigenspace of $H$ with eigenvalue $E$. Let $|v\rangle \in \text{Eig}_H(E)$. Per definition $H|v\rangle = E|v\rangle$. Furthermore as $[U, H] = 0$ we have

$$UH|v\rangle = HU|v\rangle, \quad \Leftrightarrow E(U|v\rangle) = H(U|v\rangle), \quad (A1)$$

which shows that $U|v\rangle \in \text{Eig}_H(E)$. This means that every energy eigenspace is invariant under $U$ and so as the whole vector space can be decomposed as a direct sum of $\text{Eig}_H(E), U = \bigoplus E U_E$. We now have left to prove that if $\dim(\text{Eig}_H(E)) = 1$, $U_E$ does not affect the temperature of the target. For this, let $E$ be an eigenvalue of $H$ with $\dim(\text{Eig}_H(E)) = 1$. Let $|v\rangle \in \text{Eig}_H(E)$. From Eq. (A1) and $\dim(\text{Eig}_H(E)) = 1$, $U_E|v\rangle = U|v\rangle = \lambda|v\rangle$, meaning that $|v\rangle$ is an eigenvector of $U_E$ with eigenvalue $\lambda$. Since $U_E$ is unitary, $\lambda = e^{i\theta}$ and so $U_E|v\rangle = |v\rangle U_E^\dagger = |v\rangle|v\rangle$, which proves our assertion as only the diagonal elements of the density matrix $U_E \rho U_E^\dagger$ contribute to the temperature of the target and that we can write any $\rho$ as

$$\rho = \sum_{ij} a_{ij}|v_i\rangle\langle v_j|, \quad (A2)$$

with $(|v_i\rangle)_i$ being an ONB of eigenvectors of $H$ and $|v_1\rangle = |v\rangle$. So

$$U_E \rho U_E^\dagger = U_E \left( \sum_{ij} a_{ij}|v_i\rangle\langle v_j| \right) U_E^\dagger$$

$$= U_E \left( \sum_{i,j\neq 1} a_{ij}|v_i\rangle\langle v_j| + \sum_{j\neq 1} a_{1j}|v_i\rangle\langle v_j| + \sum_{i\neq 1} a_{1i}|v_i\rangle\langle v_j| + \sum_{j\neq 1} a_{i1}|v_i\rangle\langle v_j| \right) U_E^\dagger$$

$$= \sum_{i,j\neq 1} a_{ij}|v_i\rangle\langle v_j| + \sum_{j\neq 1} \lambda a_{1j}|v_i\rangle\langle v_j| + \sum_{i\neq 1} a_{1i}\lambda |v_i\rangle\langle v_j| + a_{i1}|v_i\rangle\langle v_j|, \quad (A3)$$

implying that

$$(U_E \rho U_E^\dagger)_{kk} = a_{kk}|v_k\rangle\langle v_k| = \rho_{kk}, \quad (A4)$$

i.e., the diagonal elements of $U_E \rho U_E^\dagger$ are the original ones.■

Next we want to argue the following lemma.

Lemma 2. Among all the possible degeneracies of $H$, only $E_B = E_A + E_C$ enables cooling of qubit A.

Proof. Going through all the possible eigenvalue degeneracies of $H = H_A + H_B + H_C$, $H_i = E_i|1\rangle\langle 1| \otimes I_2, i \in \{A, B, C\}$, we see that we can have three different types of degeneracies:

1) $E_i = E_j, \quad i, j \in \{A, B, C\}$
2) $E_i = 0, \quad i \in \{A, B, C\}$
3) $E_i = E_j + E_k, \quad i, j, k \in \{A, B, C\}$

We first look at type 2. Imposing $E_A = 0$ we get four degenerate subspaces: $A_{mn} := \text{span}(|0\rangle_A|mn\rangle_B, |1\rangle_A|mn\rangle_C)$, where $m, n \in \{0, 1\}$. Our unitary can act within each $A_{mn}$ subspace on $\rho^H = \tau \otimes \tau_B \otimes \tau_C^H$. However, note that as $r_A = \frac{1}{1+e^{\frac{\theta}{T_A}}} = \frac{1}{2} = 1 - r_A$, we have

$$\rho^H_{mn,0m} = r_A [m + (1)^m r_B [n + (1)^n r_C^H]] = (1 - r_A) [m + (1)^{m} r_B [n + (1)^n r_C^H]]$$

$$= \rho^H_{mn,1m} \quad (A5)$$

such that in each of the $A_{mn}$ $\rho$ is proportional to the identity and hence $U \rho U^\dagger = \rho$ for all unitaries $U$ acting only within those subspaces. Note that this argument actually holds for any permutation of $A, B,$ and $C$ thus also treating the $E_B = 0$ and $E_C = 0$ cases and showing that Type 2 degeneracies do not enable cooling qubit A.

Turning to type 1, if $E_B = E_C$, we have two 2-dim. degenerate subspaces $\text{span}(|00\rangle, |01\rangle)$ and $\text{span}(|10\rangle, |11\rangle)$. In order to cool qubit A, one should maximize $\langle T_{BC}(U \rho^H U^\dagger)\rangle_{1,1} = \langle U \rho^H U^\dagger \rangle_{00,00}$ +
(U \rho^H U^\dagger)_{001,001} + (U \rho^H U^\dagger)_{010,010} + (U \rho^H U^\dagger)_{011,011}. As units are trace preserving, acting with U within the first subspace leaves \rho^H_{001,001} + \rho^H_{010,010} unchanged. Acting with U within the second one does not alter any term in \{Tr_{BC}(\rho^H)\}_{1,1}, meaning that this degeneracy does not allow us to cool qubit A. For \rho_{=} = \rho_B, the degenerate subspaces are span(001), [100]) and span([011], [110]). Doing the same analysis as before shows that in general the unitary does not leave \{Tr_{BC}(\rho)\}_{1,1} invariant, unfortunately it does for our \rho^H since with this condition
\rho^H_{010,010} = r_A(1 - r_B)\rho_C^H = r_B(1 - r_A)\rho_C^H = \rho_{100,100}^H (A6)
and similarly \rho_{011,011} = \rho_{011,101}. Imposing \rho_A = \rho_C we have as degenerate subspaces span(001), [100]) and span([011], [110]). As above the unitary does not in general leave \{Tr_{BC}(\rho)\}_{1,1} invariant. For our \rho^H it is also the case but since \mathcal{T}_H \geq \mathcal{T}_K, we have \frac{-E_A}{T_A} \geq \frac{-E_B}{T_B} meaning \rho_C^H \leq \rho_A, such that
\rho_{011,001}^H = r_Ar_B(1 - r_C^H) \geq r_C^H r_B(1 - r_A) = \rho_{100,100}^H (A7)
and
\rho_{011,101}^H = r_A(1 - r_B)(1 - r_C^H) \geq r_C^H(1 - r_B)(1 - r_A) \geq \rho_{110,110}^H (A8)
The unitary that maximizes \{Tr_{BC}(U \rho^H U^\dagger)\}_{1,1} is therefore the trivial one. Indeed any two-dimensional unitary can be written as
U = \begin{pmatrix} a & b e^{i\theta} \\ -\bar{b} e^{-i\theta} & a \end{pmatrix}, (A9)
with |a|^2 + |b|^2 = 1 and \theta \in [0, 2\pi). And so
\begin{align*}
E_A + E_C. All the ways of selecting two degeneracies can be listed as
(a) \ E_i = E_j = 0, \quad i, j \in \{A, B, C\}
(b) \ E_i = E_j, E_k = 0, \quad \{i, j, k\} = \{A, B, C\}
(c) \ E_A = E_B = E_C = 0
(d) \ E_A = E_B = E_C
(e) \ E_i = E_j, E_k = 2E_i, \quad \{i, j, k\} = \{A, B, C\}.
In case (a), \rho \propto 1 within the degenerate subspaces and so no cooling can occur. In case (b) the degenerate subspaces are span(001)[100|00]}_{10}, [101][1111]}_{11}) and span([011][01]}_{10}, [111][111]}_{11}). In the first two subspaces \rho \propto 1. In the third if (i, j, k) = (A, B, C) then \rho \propto 1, if (i, j, k) = (B, C, A) then cooling is possible as \rho_{011,001}^H = \rho_{100,100}^H \geq \rho_{110,110}^H, but this is no contradiction to our claim since in this case \mathcal{E}_B = \mathcal{E}_A + \mathcal{E}_C holds, and if (i, j, k) = (C, A, B), \rho_{100,100}^H = \rho_{110,110}^H \geq \rho_{011,001}^H, meaning that no cooling is possible. In case (c) \rho \propto 1 and so no cooling is possible. In case (d) the degenerate subspaces are span(001), [100] and span([011], [110] and as \rho_{011,011}^H \geq \rho_{000,001}^H \geq \rho_{010,100}^H = \rho_{011,011}^H, \rho_{011,101}^H \geq \rho_{011,101}^H, \rho_{011,011}^H = \rho_{011,011}^H \geq \rho_{011,101}^H, no cooling is possible. Finally in case (e) the degenerate subspaces are span([011][01]}_{10}, [101][111]}_{11}) and span([001][00]}_{10}, [101][10]}_{11}) and as \rho_{011,011}^H \geq \rho_{000,001}^H \geq \rho_{010,100}^H \geq \rho_{011,011}^H, \rho_{011,101}^H \geq \rho_{011,101}^H \geq \rho_{011,011}^H \geq \rho_{011,101}^H, no cooling is possible either. If (i, j, k) = (C, A, B), then \rho_{011,001}^H \geq \rho_{010,100}^H \geq \rho_{011,011}^H \geq \rho_{011,101}^H \geq \rho_{011,011}^H \geq \rho_{011,101}^H, no cooling is possible. In such a case \mathcal{E}_B = \mathcal{E}_A + \mathcal{E}_C also happens to hold; this again is no contradiction to our claim.
If one selects more than two different degeneracies from the list 1, 2, and 3, either three linearly independent degeneracies are selected, which results in \mathcal{E}_A = \mathcal{E}_B = \mathcal{E}_C = 0 and leads to no cooling as shown above, or less than three of
the selected degeneracies are linearly independent and the situation reduces to one of the above treated case. This ends the proof.

**APPENDIX B: OPTIMAL INCOHERENT THERMALIZATION**

Here we want to argue that for the case of the two-qubit machine in order to cool the target qubit maximally, the best way to make use of both thermal baths at $T_R$ and $T_H$ respectively, is to thermalize qubit B at $T_R$ and qubit C at $T_H$.

To begin, note that the only allowed unitaries are those within the degenerate subspace, as these are the only ones that preserve energy; see Sec. A. Any unitary within this qubit subspace can be viewed as a partial swap between the populations of the two levels (up to a change in relative phase, which does not affect cooling). Thus the maximum cooling is achieved by either swapping the populations fully, or not at all, since these are the two extremes of the achievable populations.

Thus given two thermal baths, at temperatures $T_R$ and $T_H$, the optimal manner of cooling would be to thermalize qubits B and C in such a way as to maximize the difference in the populations of the two degenerate levels $|101⟩$ and $|010⟩$ before applying a full swap; i.e., maximize $p_{101} - p_{010}$, where $p_{ijk}$ denotes the population of level $|ijk⟩$.

Consider that we thermalize B and C to temperatures between those of the environment and of the hot bath (these are the extremes of temperatures available to us, and thus any temperature in between is also attainable). It is straightforward to check that

$$
\frac{d}{dT_B}(p_{101} - p_{010}) = -\frac{E_{BR}(1 - r_B)}{T_B}(rr_C + (1 - r)(1 - r_C)) < 0 \ \forall T_B,
$$

and

$$
\frac{d}{dT_C}(p_{101} - p_{010}) = +\frac{E_{CR}(1 - r_C)}{T_C}(r(1 - r_B) + (1 - r)r_B) > 0 \ \forall T_C.
$$

(B1)

Therefore, it is optimal to have qubit B as be as cold as possible (the environment temperature $T_R$), and qubit C as be as hot as possible (the temperature of the hot bath $T_H$). Thus, although the whole machine has access to a hotter thermal bath at temperature $T_H$, it is best to only put qubit C in contact with it, leaving B at the room temperature $T_R$.

Note that the above argument also holds if the population on qubit A is set to be some other value than $r$, meaning that in the repeated incoherent operations one should also rethermalize qubit B to $T_R$ and qubit C to $T_H$ before applying the swap operation in order to maximally cool the target qubit.

**APPENDIX C: SINGLE-CYCLE COHERENT MACHINES**

We want here to discuss the solution of the single-cycle coherent machines presented in the main text. More precisely, we are interested in finding the unitary $U_{opt}$ (or equivalently the state $\rho_{opt} = U_{opt}\rho U_{opt}^\dagger$) that enables us to cool the target to a given temperature, i.e., ground state, $\rho_{coh} \in \{l^m, r^m\}$ at a minimal work cost. From the discussion of the main text we know that using the Schur-Horn theorem, finding $\rho_{opt}$ for a system comprises a target qubit and a machine of $n/2$ energy gaps amounts to solving

$$
\min_{\rho < \rho^m} \tilde{\rho} \cdot \tilde{H}, \quad \text{s.t.} \sum_{i=1}^{n/2} \rho_i = r_{coh}.
$$

(C1)

Indeed, the solution of Eq. (C1) gives us $\rho_{coh}$ from which we can easily reconstruct $\rho_{coh}$ and $U_{opt}$.

We in the following solve the problem for the one qubit machine ($n = 4$) and the two qubit machine ($n = 8$). We then show that, given a general machine, it is sufficient to solve two marginal problems in order to find the optimal unitary cooling the target to the lowest temperature $r^m_{coh}$. We also provide $\rho^\sharp_{coh}$.

1. **Coherent one-qubit machine**

We want to solve

$$
\min_{\rho < \rho^m} \tilde{\rho} \cdot \tilde{H}, \quad \text{s.t.} \rho_1 + \rho_2 = r_{coh},
$$

(C2)

where the majorization conditions are simply given by

$$
\sum_{i=1}^{l} \rho_i^{\downarrow} \leq \sum_{i=1}^{l} \rho_i^{\downarrow}, \ \forall l = 1, \ldots, 4,
$$

(C3)

with equality for $l = 4$. First, note that $\rho_1 + \rho_2 = r$ with the trace condition implies that $\rho_3 + \rho_4 = 1 - r$ such that

$$
\tilde{\rho} \cdot \tilde{H} = \rho_2 E_B + \rho_3 E_A + \rho_4 (E_A + E_B) = (r - \rho_1 + \rho_4)E_B + (1 - r)E_A
$$

(C4)

such that in order to minimize $\tilde{\rho} \cdot \tilde{H}$, one should minimize $\rho_4 - \rho_1$. This means that $\rho_1$ should be the greatest component of $\tilde{\rho}$, namely, $\rho_1 = \rho_1^\dagger$ and $\rho_4$ the smallest, namely, $\rho_4 = \rho_4^\dagger$. From Eq. (C3) with $l = 1$ we have $\rho_1 = \rho_1^{\downarrow} \leq \rho_1^{\uparrow}$ and combining Eq. (C3) with $l = 3$ with the trace condition we get

$$
\rho_1^{\uparrow} + \rho_2^{\uparrow} + \rho_3^{\uparrow} + \rho_4^{\uparrow} = \rho_1^\dagger + \rho_2^\dagger + \rho_3^\dagger + \rho_4^\dagger,
$$

(C5)

such that $\rho_4 = \rho_4^\dagger \geq \rho_4^{\downarrow}$. In order to minimize $\rho_4 - \rho_1$ we therefore have to choose $\tilde{\rho}$ such that

$$
\rho_1 = \rho_1^{\uparrow}, \quad \rho_4 = \rho_4^{\downarrow}.
$$

(C6)

Plugging these values in the majorization conditions [Eq. (C3)], we are left with

$$
\rho_1^\dagger \leq \rho_3^\dagger, \quad \rho_2^\dagger + \rho_4^\dagger = \rho_2^m + \rho_4^m.
$$

(C7)

As $\rho_2^\dagger = \max(\rho_2, \rho_3)$ and $\rho_2^\dagger + \rho_4^\dagger = \rho_2 + \rho_3$, these are exactly the conditions for

$$
(\rho_2, \rho_3) \prec (\rho_2^m, \rho_3^m),
$$

(C8)

which means that one can get the majorized vector by applying a T-transform to the initial vector; that is, for some
This simply follows from the fact that in general $r < s$ iff there exists some doubly stochastic matrix $D$ such that $r = Ds$, and that the most general $2 \times 2$ doubly stochastic matrices are $T$-transforms. Now we just have to choose $t$ such that $\rho_1 + \rho_2 = r$, that is,

$$t = \frac{\rho_{1m}^i + \rho_{2m}^j - r}{\rho_{3m}^k - \rho_{2m}^j}.$$  \hfill (C9)

or equivalently

$$1 - t = \frac{r - \rho_{1m}^i}{\rho_{2m}^j - \rho_{1m}^i}.$$  \hfill (C10)

The associated work cost is

$$\Delta F = (\vec{\rho} - \vec{\rho}_m) \cdot \vec{H} = (1 - t)(\rho_{1m}^i - \rho_{1m}^i)(\rho_{2m}^j - \rho_{A})$$

$$= (1 - t)(\rho_{1m}^i - \rho_{2m}^j)(\rho_{A} - \rho_{E}).$$  \hfill (C11)

A unitary $U$ such that $\vec{\rho} = \text{Diag}(U\rho_{m}^i U^\dagger)$ is for example given by

$$U = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & \sqrt{1 - \mu} & \sqrt{\mu} & 0 \\
0 & -\sqrt{\mu} & \sqrt{1 - \mu} & 0 \\
0 & 0 & 0 & 1
\end{pmatrix},$$  \hfill (C12)

where $\mu = 1 - t$ and can be written more compactly as

$$U = e^{-i\text{arcsin}(\sqrt{\mu}) L_{AB}},$$  \hfill (C13)

with $L_{AB} = |01\rangle \langle 10| - |10\rangle \langle 01|$.  

2. Coherent two-qubit machine

We want to solve

$$\min_{\vec{\rho} \in \rho_{m}^i} \Delta F, \quad \text{s.t.} \quad \sum_{i=1}^{4} \rho_i = r_{\text{coh}},$$  \hfill (C14)

where the majorization conditions are simply given by

$$\sum_{i=1}^{k} \rho_{i}^{\downarrow} \leq \sum_{i=1}^{k} \rho_{i}^{\downarrow}, \quad \forall k = 1, \ldots, 7,$$

$$\sum_{i=1}^{8} \rho_{i}^{\downarrow} = \sum_{i=1}^{8} \rho_{i}^{\downarrow}.$$  \hfill (C15)

The ordering of the original entries is crucial to the solving of the problem. There are hence two regimes that one needs to investigate, namely, $E_C \leq E_A$ and $E_C > E_A$. We begin with the $E_C \leq E_A$ regime.

\textit{a. The $E_C \leq E_A$ regime.} In this regime the ordering of the original diagonal entries is given by

$$\rho_{1m}^i > \rho_{2m}^j > \rho_{3m}^k > \rho_{6m}^m > \rho_{4m}^m > \rho_{5m}^m > \rho_{8m}^m.$$  \hfill (C16)

We furthermore have

$$\vec{p} \cdot \vec{H} = \rho_1 \rho_{E,1} + \rho_2 \rho_{E,2} + \rho_3 \rho_{E,3} + \rho_4 \rho_{E,4} + \rho_5 \rho_{E,5} \rho_6 \rho_{E,6}$$

$$+ \rho_7 \rho_{E,7} + \rho_8 \rho_{E,8}.$$  \hfill (C17)

We will next rewrite Eq. (C18), keeping in mind the ordering of the original state of Eq. (C18), in a way that the majorization conditions of Eq. (C17) can easily be applied. First, we use that $\rho_5 + \rho_6 + \rho_7 + \rho_8 = 1 - r$ and get

$$\vec{p} \cdot \vec{H} = (\rho_2 + \rho_3) \rho_{E,1} + \rho_4 \rho_{E,2} + \rho_5 \rho_{E,3} + \rho_6 \rho_{E,4} + \rho_7 \rho_{E,5}$$

$$+ (1 - r) \rho_8 \rho_{E,8}.$$  \hfill (C18)

where in the second step we used that $E_B = E_A + E_C$. Note that the sum of the the minima of each summand of $\vec{p} \cdot \vec{H}$ is for sure a lower bound to the minimum of $\vec{p} \cdot \vec{H}$, such that if one can pick a $\rho$ achieving this lower bound, we will have reached the minimum of $\vec{p} \cdot \vec{H}$. Using the last reformulation of $\vec{p} \cdot \vec{H}$, this is luckily possible, indeed:

$$\rho_{1m}^i \leq \rho_{1m}^i \leq \rho_{1m}^i = \rho_{1m}^i;$$

$$\rho_{1m}^i + \rho_{2m}^j + \rho_{3m}^k \leq \sum_{i=1}^{4} \rho_{i}^{\downarrow} \leq \sum_{i=1}^{4} \rho_{i}^{\downarrow} = \rho_{1m}^i + \rho_{1m}^i + \rho_{1m}^i + \rho_{1m}^i = \rho_{1m}^i + \rho_{1m}^i;$$

$$\rho_{8m}^m \geq \rho_{8m}^m \geq \rho_{8m}^m = \rho_{8m}^m;$$

$$\rho_{8m}^m \geq \rho_{8m}^m + \rho_{8m}^m \geq \rho_{8m}^m = \rho_{8m}^m.$$  \hfill (C19)

To minimize the first summand we hence have to choose $\rho_1 = \rho_{1m}^i$. To minimize the third summand, since $\rho_1 = \rho_{1m}^i$, we have to pick $\rho_2 = \rho_{1m}^i$. To minimize the fourth summand we have to choose $\rho_5 = \rho_{1m}^i$ which forces us to choose $\rho_2 = \rho_{1m}^i$ in order to minimize the last summand. We are hence left with the minimization of the second summand that is achieved if

$$\rho_{5m}^m + \rho_{3m}^m = \rho_{5m}^m + \rho_{3m}^m.$$  \hfill (C20)

is satisfied.

Now note that we have

$$\rho_{1m}^i + \rho_{2m}^j + \max\{\rho_3, \rho_5\} \leq \rho_{1m}^i + \rho_{1m}^i + \rho_{1m}^i$$

$$\leq \rho_{1m}^i + \rho_{1m}^i + \rho_{1m}^i$$

$$= \rho_{1m}^i + \rho_{1m}^i + \rho_{1m}^i$$

$$= \rho_{1m}^i + \rho_{1m}^i + \rho_{1m}^i.$$  \hfill (C21)
such that
\[
\max\{\rho_3, \rho_5\} \leq \rho_5^{in}.
\] (C24)

Equations (C24) and (C22) together mean that \((\rho_3, \rho_5) < (\rho_5^{in}, \rho_3^{in})\), which we know from Sec. C.1 to be equivalent to
\[
\begin{pmatrix} \rho_3 \\ \rho_5 \end{pmatrix} = T_1 \begin{pmatrix} \rho_5^{in} \\ \rho_3^{in} \end{pmatrix}, \quad T_1 \begin{pmatrix} t_1 \\ 1 - t_1 \end{pmatrix} = \begin{pmatrix} 1 - t_1 \\ t_1 \end{pmatrix}.
\] (C25)

for some \(t_1 \in [0, 1]\). Similarly, as \(\sum_{i=1}^{8} \rho_i = \sum_{i=1}^{8} \rho_i^{in}\) we find that
\[
\rho_4 + \rho_6 = \rho_4^{in} + \rho_6^{in}
\] (C26)

and using that the second line of Eq. (C21) is satisfied with equality we find that
\[
\sum_{i=1}^{4} \rho_i^{in} + \max\{\rho_4, \rho_6\} \leq \sum_{i=1}^{5} \rho_i^{in} - \sum_{i=1}^{4} \rho_i^{in} + \rho_6^{in}
\] (C27)

such that
\[
\max\{\rho_4, \rho_6\} \leq \rho_6^{in}.
\] (C28)

Now Eqs. (C26) and (C28) together mean that \((\rho_4, \rho_6) < (\rho_4^{in}, \rho_6^{in})\), which as before is equivalent to
\[
\begin{pmatrix} \rho_4 \\ \rho_6 \end{pmatrix} = T_2 \begin{pmatrix} \rho_6^{in} \\ \rho_4^{in} \end{pmatrix}, \quad T_2 \begin{pmatrix} t_2 \\ 1 - t_2 \end{pmatrix} = \begin{pmatrix} 1 - t_2 \\ t_2 \end{pmatrix}.
\] (C29)

for some \(t_2 \in [0, 1]\). This means that for any \(t_1\) and \(t_2\), we have found a vector \(\rho\) that achieves the minimum of each summands in (C20) and that therefore is the solution of our problem. Of course, for a given \(r\), only some \(t_1\) and \(t_2\) will solve our problem, namely the ones satisfying
\[
\begin{align*}
&\quad r = \rho_1 + \rho_2 + \rho_3 + \rho_4 \\
&= \rho_1^{in} + \rho_2^{in} + t_1 \rho_3^{in} + (1 - t_1) \rho_5^{in} + t_2 \rho_4^{in} + (1 - t_2) \rho_6^{in} \\
&= \sum_{i=1}^{4} \rho_i^{in} + (t_1 - 1) \rho_5^{in} + (1 - t_1) \rho_2^{in} + (t_2 - 1) \rho_4^{in} + (1 - t_2) \rho_6^{in} \\
&= r^{in} + (1 - t_1) (\rho_5^{in} - \rho_2^{in}) + (1 - t_2) (\rho_4^{in} - \rho_6^{in}).
\end{align*}
\] (C30)

Next note that
\[
\begin{align*}
\rho_3^{in} - \rho_5^{in} &= (1 - r_A) r_B r_C - r_A (1 - r_B) r_C \\
&= r_B r_C - r_A r_B r_C - r_A r_B + r_A r_B r_C \\
&= (r_B - r_A) r_C, \\
\rho_4^{in} - \rho_6^{in} &= (1 - r_A) r_B (1 - r_C) - r_A (1 - r_B) (1 - r_C) \\
&= (1 - r_C) (r_B - r_A r_B - r_A + r_A r_B) \\
&= (1 - r_C) (r_B - r_A)
\end{align*}
\] (C31)

such that
\[
\begin{align*}
r &= r^{in} + [(1 - t_1) r_C + (1 - t_2) (1 - r_C)] (r_B - r_A). \\
&= (1 - t_1) (r_B - r_A) r_C.
\end{align*}
\] (C32)

If we were to choose \(t_1 = t_2 = t\), then we would have
\[
r = r^{in} + (1 - t) (r_B - r_A).
\] (C33)

Now the work cost of carrying this process is
\[
\Delta F = (\bar{\rho} - \rho^{in}) \cdot \tilde{H} = [t_1 \rho_3^{in} + (1 - t_1) \rho_5^{in} - \rho_3^{in}] E_B + [t_2 \rho_4^{in} + (1 - t_2) \rho_6^{in} - \rho_4^{in}] (E_B + E_C) + [(1 - t_1) \rho_3^{in} + t_1 \rho_5^{in} - \rho_3^{in}] E_A + [(1 - t_2) \rho_4^{in} + t_2 \rho_6^{in} - \rho_4^{in}] (E_A + E_C)
\]
\[
= (1 - t_1) (\rho_3^{in} - \rho_5^{in}) E_B + (1 - t_2) (\rho_4^{in} - \rho_6^{in}) (E_B + E_C) + (1 - t_1) (\rho_3^{in} - \rho_5^{in}) E_A + (1 - t_2) (\rho_4^{in} - \rho_6^{in}) (E_A + E_C)
\]
\[
= [(1 - t_1) (\rho_3^{in} - \rho_5^{in}) + (1 - t_2) (\rho_4^{in} - \rho_6^{in})] E_C
\]
\[
= [(1 - t_1) r_C + (1 - t_2) (1 - r_C)] (r_B - r_A) E_C.
\] (C34)

If we choose \(t_1 = t_2 = t\), then
\[
\Delta F = (1 - t) (r_B - r_A) E_C.
\] (C35)

A unitary \(U\) such that \(\bar{\rho} = \text{Diag}(U \rho^{in} U^+)\) is, for example, given by
\[
U = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \sqrt{1 - \mu_1} & 0 & 0 & \sqrt{\mu_1} & 0 \\
0 & 0 & 0 & \sqrt{1 - \mu_2} & 0 & 0 & \sqrt{\mu_2} \\
0 & 0 & -\sqrt{\mu_1} & 0 & 0 & 0 & \sqrt{1 - \mu_1} \\
0 & 0 & 0 & -\sqrt{\mu_2} & 0 & 0 & \sqrt{1 - \mu_2} \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix},
\] (C36)
where \( \mu_1 = 1 - t_1 \) and \( \mu_2 = 1 - t_2 \). If we choose \( t_1 = t_2 = t \), then \( \mu_1 = \mu_2 = \mu \) and

\[
U = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & \sqrt{1 - \mu} & 0 & 0 & 0 \\
0 & 0 & 0 & \sqrt{1 - \mu} & 0 & 0 \\
0 & 0 & 0 & 0 & \sqrt{1 - \mu} & 0 \\
1 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

(C37)

can be compactly written as

\[
U = e^{-i\arcsin(\sqrt{\mu})\hat{A}}, \quad L_{AB} = i|01\rangle\langle 10|_{AB} - i|01\rangle\langle 01|_{AB}.
\]

(C38)

b. The \( E_C > E_A \) regime. In this regime the ordering of the original diagonal entries is given by

\[
\rho_1^{in} > \rho_5^{in} > \rho_2^{in} > \rho_3^{in} > \rho_4^{in} > \rho_8^{in}.
\]

As before, we would like to reshuffle the terms of

\[
\tilde{\rho} \cdot \hat{H} = \rho_2 E_C + \rho_3 E_B + \rho_4 (E_B + E_C) + \rho_5 E_A + \rho_6 (E_A + E_C) + \rho_7 (E_A + E_B) + \rho_8 (E_A + E_B + E_C)
\]

such that each summand can be minimized. So we get

\[
\tilde{\rho} \cdot \hat{H} = (\rho_2 + \rho_4)E_C + (r - \rho_1 - \rho_2) \frac{E_B}{E_C} + \rho_3 E_A + \rho_6 (E_A + E_C) + (1 - r - \rho_5 - \rho_6) (E_A + E_B) + \rho_8 E_C
\]

\[
= -\rho_1 E_C + (-\rho_1 - \rho_2) E_A + (-\rho_3) (E_A + E_C) + \rho_4 E_C + (-\rho_6) E_A + \rho_8 E_C + (1 - r) (E_A + E_B) + \rho_8 E_B
\]

\[
= -\rho_1 E_C + (-\rho_1 - \rho_5 - \rho_6) E_A + (r - 1 + \rho_6 + \rho_7 + \rho_8) E_C + (\rho_4 + \rho_8) E_C + (-\rho_6) E_A + c
\]

\[
= -\rho_1 E_C + (-\rho_1 - \rho_5 - \rho_6) E_A + \rho_7 (E_A - E_C) + (\rho_7 + \rho_4 + \rho_8) E_C + \rho_8 E_C + c + (r - 1) E_C
\]

\[
= -\rho_1 E_C + (-\rho_1 - \rho_5 - \rho_6) E_A + d + \rho_8 E_C + (\rho_6 + \rho_7 + \rho_4 + \rho_8) (E_C - E_A) + (\rho_7 + \rho_4 + \rho_8) E_A.
\]

(C40)

Now looking at each summand and applying the majorization conditions with the order of the original vector that we know we get

\[
\rho_1 \leq \rho_1^{in} \\
\rho_1 + \rho_2 + \rho_3 \leq \rho_1^{in} + \rho_2^{in} + \rho_3^{in} \Rightarrow \rho_1 = \rho_1^{in} \\
\rho_5 \geq \rho_5^{in} \Rightarrow \rho_5 = \rho_5^{in} \\
\rho_7 + \rho_4 + \rho_8 \geq \rho_7^{in} + \rho_4^{in} + \rho_8^{in} \Rightarrow \rho_7 + \rho_4 = \rho_7^{in} + \rho_4^{in} \\
\rho_6 + \rho_7 + \rho_4 + \rho_8 \geq \rho_6^{in} + \rho_7^{in} + \rho_4^{in} + \rho_8^{in} \Rightarrow \rho_6 = \rho_6^{in}.
\]

(C42)

Furthermore, note that out of \( \sum_{i=1}^{8} \rho_i = \sum_{i=1}^{8} \rho_i^{in} \) and the above fixed values we get

\[
\rho_3 = \rho_3^{in}.
\]

(C43)

Also using the majorization conditions, we have

\[
\rho_1 + \max\{\rho_5, \rho_6\} \leq \rho_1^{in} + \rho_5^{in} \Rightarrow \max\{\rho_5, \rho_6\} \leq \rho_5^{in}
\]

\[
\min\{\rho_5, \rho_6\} + \rho_8 \geq \rho_4^{in} + \rho_8^{in} \Rightarrow \min\{\rho_5, \rho_6\} \geq \rho_4^{in},
\]

(C44)

which together with (C42) means that \( (\rho_5, \rho_6) < (\rho_5^{in}, \rho_6^{in}) \) and \( (\rho_4, \rho_7) < (\rho_4^{in}, \rho_7^{in}) \), which is equivalent to

\[
\begin{pmatrix}
\rho_5 \\
\rho_2
\end{pmatrix} = T_1 \begin{pmatrix}
\rho_5^{in} \\
\rho_2^{in}
\end{pmatrix}, \quad T_1 \begin{pmatrix}
1 & t_1 \\
1 - t_1 & t_1
\end{pmatrix}
\]

(C45)

and

\[
\begin{pmatrix}
\rho_5 \\
\rho_7
\end{pmatrix} = T_2 \begin{pmatrix}
\rho_5^{in} \\
\rho_7^{in}
\end{pmatrix}, \quad T_2 \begin{pmatrix}
t_2 & 1 - t_2 \\
1 - t_2 & t_2
\end{pmatrix}.
\]

(C46)

for some \( t_1 \in [0, 1] \) and \( t_2 \in [0, 1] \). This means that for any \( t_1 \) and \( t_2 \), the vector \( \rho \) is the solution of our problem. Of course, for a given \( r \), only some \( t_1 \) and \( t_2 \) will solve our problem, namely, the ones satisfying

\[
r = \rho_1 + \rho_2 + \rho_3 + \rho_4
\]

\[
= r^{in} + (1 - t_1)(\rho_5^{in} - \rho_2^{in}) + (1 - t_1)(\rho_7^{in} - \rho_4^{in})
\]

\[
= r^{in} + [(1 - t_1)r_B + (1 - t_1)(1 - r_B)](\rho_C - \rho_A).
\]

(C47)

If we were to choose \( t_1 = t_2 = t \), then we would have

\[
r = r^{in} + (1 - t)(\rho_C - \rho_A).
\]

(C48)

Now the work cost of carrying this process is

\[
\Delta F = \langle \tilde{\rho} - \rho^{in} \rangle \cdot \hat{H}
\]

\[
= [(1 - t_1)r_B + (1 - t_2)(1 - r_B)](\rho_C - \rho_A).
\]

(C49)
If we choose $t_1 = t_2 = t$, then
\[ \Delta F = (1 - t)(r_C - r_A)(E_C - E_A). \]  

A unitary $U$ such that $\tilde{p} = \text{Diag}(U \rho^{in}U^\dagger)$ is for example given by
\[ U = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \sqrt{1 - \mu_1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{1 - \mu_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{1 - \mu_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \]  

where $\mu_1 = 1 - t_1$ and $\mu_2 = 1 - t_2$. If we choose $t_1 = t_2 = t$, then $\mu_1 = \mu_2 = \mu$ and then $U$ can be compactly written as
\[ U = e^{-i \arcsin(\sqrt{\mu})} \mathcal{A}, \quad L_{AC} = -i|01\rangle\langle 10|_{AC} - i|10\rangle\langle 01|_{AC}. \]  

Note, however, that upon applying this procedure one only finds the solution of our problem for
\[ r_A \leq r \leq \rho_1^{in} + \rho_2^{in} + \rho_1^{in} + \rho_1^{in}. \]  

To find the solution for
\[ \rho_1^{in} + \rho_2^{in} + \rho_3^{in} + \rho_1^{in} \leq r \leq \rho_1^{in} + \rho_2^{in} + \rho_3^{in} + \rho_3^{in}, \]  

we use another rearrangement of terms of $\tilde{p} \cdot \mathcal{H}$, namely,
\[ \tilde{p} \cdot \mathcal{H} = \rho_2 E_C + (r - \rho_1 - \rho_2) E_B + \rho_4 E_C + (1 - r - \rho_1 - \rho_3) E_A + \rho_5 E_C + (\rho_1 + \rho_3)(E_A + E_B) + \rho_8 E_C \]
\[ = -\rho_1 E_C + (-\rho_1 - \rho_2)E_A + r E_B + (1 - r)E_A(\rho_4 + \rho_6 + \rho_7 + \rho_8)E_C + (\rho_1 + \rho_3)E_A + \rho_8 E_C. \]

By looking at each summand individually we find that
\[ \begin{align*}
\rho_1 &\leq \rho_3^{in} \\
\rho_1 + \rho_2 &\leq \rho_1^{in} + \rho_3^{in} \\
\rho_6 &\geq \rho_8^{in} \\
\rho_6 + \rho_8 &\geq \rho_6^{in} + \rho_8^{in} \\
\rho_1 + \rho_1 + \rho_4 + \rho_8 &\geq \rho_1^{in} + \rho_1^{in} + \rho_4^{in} + \rho_8^{in} \\
\Rightarrow \rho_1 &= \rho_3^{in} \\
\Rightarrow \rho_2 &= \rho_5^{in} \\
\Rightarrow \rho_8 &= \rho_8^{in} \\
\Rightarrow \rho_7 &= \rho_8^{in} \\
\Rightarrow \rho_4 + \rho_6 &= \rho_6^{in} + \rho_7^{in}. \end{align*} \quad (C56) \]

Using the trace condition we find that
\[ \rho_3 + \rho_5 = \rho_1^{in} + \rho_3^{in}. \]  

As before this leads to $(\rho_5, \rho_3) < (\rho_3^{in}, \rho_3^{in})$ and $(\rho_6, \rho_4) < (\rho_6^{in}, \rho_7^{in})$, which is equivalent to
\[ \begin{pmatrix} \rho_3 \\ \rho_5 \end{pmatrix} = T_1 \begin{pmatrix} \rho_3^{in} \\ \rho_3^{in} \end{pmatrix}, \quad T_1 \begin{pmatrix} t_1 \\ t_1 - t_1 \\ t_1 \end{pmatrix} \]  

and
\[ \begin{pmatrix} \rho_6 \\ \rho_4 \end{pmatrix} = T_2 \begin{pmatrix} \rho_6^{in} \\ \rho_7^{in} \end{pmatrix}, \quad T_2 \begin{pmatrix} t_2 \\ t_2 - t_2 \\ t_2 \end{pmatrix}, \]

for some $t_1 \in [0, 1]$ and $t_2 \in [0, 1]$. This means that for any $t_1$ and $t_2$, the vector $\rho$ is the solution of our problem. Of course, for a given $r$, only some $t_1$ and $t_2$ will solve our problem, namely, the ones satisfying
\[ r = \rho_1 + \rho_2 + \rho_3 + \rho_4 \]
\[ = r_C + [(1 - t_1)r_A + (1 - t_1)(1 - r_A)](r_B - r_C). \]  

If we were to choose $t_1 = t_2 = t$, then we would have
\[ r = r_C + (1 - t)(r_B - r_C). \]  

The work cost of carrying this process is
\[ \Delta F = (\tilde{p} - \tilde{\rho}^{in}) \cdot \mathcal{H} \]
\[ = (r_C - r_A)(E_C - E_A) \]
\[ + [(1 - t_1)r_A + (1 - t_2)(1 - r_A)](r_B - r_C)E_C. \]  

If we choose $t_1 = t_2 = t$, then
\[ \Delta F = (r_C - r_A)(E_C - E_A) + (1 - t)(r_B - r_C)E_C. \]  

A unitary $U$ such that $\tilde{p} = \text{Diag}(U \rho^{in}U^\dagger)$ is given by
\[ U = U_{25}(\mu_1)U_{46}(\mu_2)U_{25}(1)U_{27}(1), \]  

where $\mu_1 = 1 - t_1$ and $\mu_2 = 1 - t_2$ and $\mu \in [0, 1]$
\[ U_{ij}(\mu) := \begin{pmatrix} \sqrt{1 - \mu} & \sqrt{1 - \mu} \\ -\sqrt{1 - \mu} & \sqrt{1 - \mu} \end{pmatrix} \oplus \mathcal{L}_{ij}. \]  

If we choose $t_1 = t_2 = t$ then $\mu_1 = \mu_2 = \mu$, then $U$ can be written as
\[ U = e^{-i \arcsin(\sqrt{\mu})\mathcal{A} + e^{-in}/2L_{AC}}, \]

with $L_{AC} = -i|01\rangle\langle 10|_{AC} - i|10\rangle\langle 01|_{AC}$ and $L_{AB} = -i|01\rangle\langle 10|_{AB} - i|10\rangle\langle 01|_{AB}$. We can also summarize both
parts of the solution in one unitary. Then $U$ looks like

$$U = U_{35}(t_2)U_{46}(t_2)U_{57}(t_2)U_{57}(t_1),$$

(C67)

with $\mu_1 = \min(2\mu, 1)$, $\mu_2 = \max(2\mu - 1, 0)$, and $\mu \in [0, 1]$,
or

$$U = e^{-i\arcsin(\sqrt{\pi/2})} e^{-i\arcsin(\sqrt{\pi/2})}.$$

(C68)

3. Endpoint of arbitrary single-cycle machines

We here want to find the solution of the problem of Eq. (C1) when $r$ is chosen to be the maximally allowed $r$. We set $k = n/2$. We know that $r$ is at most $r_{\text{coh}}^* = \sum_{i=1}^k \rho_i^{\text{in}, i}$ since

$$r = \sum_{i=1}^k \rho_i \leq \sum_{i=1}^k \rho_i^{\text{in}, i} \leq \sum_{i=1}^k \rho_i^{\text{in}, i}$$

(C69)

and choosing $\rho_i = \rho_i^{\text{in}, i}$, $i = 1, \ldots, n$ achieves this upper bound, i.e., then $\tilde{\rho} < \rho^\text{in}$ and $r = r_{\text{coh}}^*$. We next want to show the following lemma.

**Lemma 3.** For any state $\rho$ such that $r_{\rho} = r_{\text{coh}}^*$, the first $k$ entries of the state are its biggest entries.

**Proof.** Suppose not, i.e., there exists a $p$ as in the statement for which there exist $i \leq k$, and $j > k$ such that $\rho_i < \rho_j$. Then $\tilde{\rho} := \rho_j \rho_j \prec \rho^\text{in}$ and $r' = \sum_{i=1}^k \rho_i = \sum_{i=1, i \neq j}^k \rho_i + \rho_j > \sum_{i=1, i \neq j}^k \rho_i + \rho_j = \sum_{i=1}^k \rho_i = r$. As $\tilde{\rho} \prec \rho^\text{in}$, $r' < r_{\text{coh}}^*$ so $r < r_{\text{coh}}^*$ contradicting the assumption.

Writing $\tilde{\rho}$ as

$$\tilde{\rho} = (v_{p}, w),$$

(C70)

with

$$v_p = ((v_{p_1}), \ldots, (v_{p_l})) := (\rho_1, \ldots, \rho_k)$$

$$w = ((w_{p_1}), \ldots, (w_{p_{n-k}})) := (\rho_{k+1}, \ldots, \rho_n),$$

(C71)

the above lemma can be reformulated as

$$v_{p}^i = (\rho_{k+1}^i, \ldots, \rho_k^i).$$

(C72)

What makes the above equation nontrivial is that on the left-hand side only the first $k$ entries of $\rho$ are reordered whereas on the right-hand side all the entries of $\rho$ are reordered. Using that $\tilde{\rho} < \rho^\text{in}$ we have for all $l = 1, \ldots, k$ that

$$\sum_{i=1}^l (v_{p}^i) = \sum_{i=1}^l \rho_i^\text{in} \leq \sum_{i=1}^l \rho_i^\text{in},$$

(C73)

with equality for $l = k$. This is equivalent to $v_{p} < v_{p^\text{in}, i}$. Also note that the lemma implies that the last $n - k$ entries of $\tilde{\rho}$ are the smallest $n - k$ ones, that is,

$$w_{p}^i = (\rho_{k+1}^i, \ldots, \rho_k^i).$$

(C74)

Again using that $\tilde{\rho} < \rho^\text{in}$ we find that for all $l = 1, \ldots, n - k$,

$$\sum_{i=1}^l (w_{p}^i) = \sum_{i=1}^l \rho_i^{\text{in}, i} \leq \sum_{i=1}^l \rho_i^{\text{in}, i}$$

(C75)

with equality for $l = n - k$. This is equivalent to $w_{p} < w_{p^{\text{in}, i}}$. So we have proven the following lemma.

**Lemma 4.** If $\tilde{\rho}$ satisfies $r_{\rho} = r_{\text{coh}}^*$, then

$$\tilde{\rho} < \rho^\text{in} \iff v_{p} < v_{p^\text{in}, i} \text{ and } w_{p} < w_{p^{\text{in}, i}},$$

(C76)

where $\tilde{\rho} = (\rho_1, \ldots, \rho_k, \rho_{k+1}, \ldots, \rho_n)$.

Indeed, the reverse implication is trivially satisfied. This means that $\rho$ is the solution of

$$\min_{\rho < \rho^\text{in}} \mathcal{H}, \text{ s.t. } \sum_{i=1}^k \rho_i = r_{\text{coh}}^*$$

(C77)

iff $v_{p}$ and $w_{p}$ are solutions of

$$\min_{v_{p} < v_{p^\text{in}, i}} v_{H} \cdot v_{p}, \min_{w_{p} < w_{p^{\text{in}, i}}} w_{H} \cdot w_{p},$$

(C78)

where we split $H$ in the same way as $\rho$ in $H = (v_{H}, w_{H})$. That is, we have reformulated the original constraint problem into two marginal unconstraint problems. The minimums of $v_{H} \cdot v_{p}$ and $w_{H} \cdot w_{p}$ are attained by $\sum_{i=1}^k (v_{H}^i)v_{p}^i$ and $\sum_{i=1}^{n-k} (w_{H}^i)w_{p}^i$, that is when the entries of $v_{p}$ and $w_{p}$ are inversely ordered with respect to the ones of $v_{H}$, respectively, $w_{H}$. This uniquely defines the $\tilde{\rho}$ that minimizes (C77) and solves the endpoint problem.

APPENDIX D: SINGLE-CYCLE ENDPOINT FREE ENERGY

We want here to argue that in the two-qubit machine one always needs less free energy to reach the endpoint in the single-cycle coherent scenario than in the single-cycle incoherent scenario. This is formulated in the following.

**Claim 1.** $\Delta F_{\text{coh}}^* \leq \Delta F_{\text{inc}}^*$ with equality iff $T_R \to +\infty$, $E_0 \to 0$, $E_C \to 0$, or $E_C \to +\infty$.

**Proof.** If $T_R \to +\infty$, one sees directly that for both cases of $E_C \leq E$ and $E > E_C$, $\Delta F^* = 0 = \Delta F_{\text{inc}}^*$. If $E \to 0$, then also in both cases $\Delta F^* = \Delta F_{\text{inc}}^*$. If $E_C \to 0$ we also trivially have $\Delta F_{\text{coh}}^* = 0 = \Delta F_{\text{inc}}^*$. If $E_C \to +\infty$, both terms go to infinity as $O(E)$ and are in that sense equal. Similarly one sees that if $T_R \to 0$ or $E \to +\infty$, $E_C(r_{B} - r) < \frac{1}{2} = \Delta F_{\text{inc}}^*$. Else, assuming $E_C, E, T_R \notin \{0, +\infty\}$, note that as for $E_C > E$ we have

$$\Delta F_{\text{coh}}^* = E_C(r_{B} - r) - E(r_C - r),$$

(D1)

the work cost in the coherent scenario is always bounded by $E_C(r_{B} - r)$. In order to prove our point we hence need only to prove that $E_C(r_{B} - r) < \Delta F_{\text{inc}}^*$. To do so we look at

$$f(T_R) = r_C + r - r_{B} - \frac{1}{2}$$

(D2)

for fixed $E_C, E \in [0, +\infty]$. As

$$f(0) = \frac{1}{2}, \quad f(\infty) = 0,$$

(D3)
if \( f'(T_R) < 0 \) our point is proven. We hence calculate

\[
\begin{align*}
  f'(T_R) &= -\frac{1}{T_R^2} \left[ E_C r_C (1 - r_C) + E r (1 - r) - E_B r_B (1 - r_B) \right] \\
  &= -\frac{1}{T_R^2} \left\{ E_C \left[ r_C (1 - r_C) - r_B (1 - r_B) \right]_{>0} \right\} < 0, \quad (D4)
\end{align*}
\]

where in the second step we used that \( g(r) = r (1 - r) \) is strictly decreasing on \([\frac{1}{2}, 1]\) as well as \( \frac{1}{2} < r_C < r_B < 1 \) and \( \frac{1}{2} < r < r_B < 1 \). This ends the proof. ■

**APPENDIX E: CROSSING POINT**

In Sec. VI C we demonstrate the existence of a critical point (\( \Delta F_{\text{crit}}, T_{\text{crit}} \)) beyond which the coherent scenario outperforms the incoherent one in the single-cycle regime. Note that as both curves start at the same point, this critical point is not the only crossing point of both curves. Our numerical results though strongly suggest that those are the only two. We want here to study the behavior of the more interesting crossing point, (\( \Delta F_{\text{crit}}, T_{\text{crit}} \)), when one varies the environment temperature \( T_R \) and the energy gap \( E_C \). In Fig. 10 we analyze the behavior of \( \Delta F_{\text{crit}} \) as a function of \( T_R \) for fixed \( E_C \). Apart from the fact that the curves seem smooth, it is interesting to note that they all exhibit a maximum for some environmental temperature. This point corresponds to the environmental temperature for which the crossing between coherent and incoherent occurs at the lowest temperature of the target qubit (i.e., at maximum cooling).

**FIG. 10.** \( \Delta F_{\text{crit}} \) is plotted as a function of \( T_R \) for various fixed \( E_C \).

**APPENDIX F: TREATING THE RESOURCE INTERNALLY**

Instead of treating the resource as an external supply, one can instead consider part of the machine to be the resource itself. We showcase here what such a standpoint would lead to for the two-qubit machine when considering qubit C to be the resource. One can then ask the same question, namely how do the fully entropic (incoherent) versus the nonentropic (coherent) supply of free energy scenarios compare in terms of

Reachable temperatures and

Reachable temperatures for a given work cost.

The incoherent scenario translates to exchanging qubit C with a qubit at a hotter temperature \( T_H \) and then performing the energy-conserving unitary in the subspace span(\(|010\rangle, |101\rangle\)). The free energy difference is now calculated in terms of the system state since the state itself is the resource. We hence have for the final free energy

\[
F^{\text{fin}} = (H)_{\rho^{\text{fin}}} - T_R S_{\rho^{\text{fin}}}
\]

\[
= \text{Tr}(\rho^{\text{fin}} H) + T_R \text{Tr}(\rho^{\text{fin}} \ln(\rho^{\text{fin}}))
\]

\[
= T_R \ln(r_B r_C^H) + E_C \left( 1 - \frac{T_R}{T_H} \right) (1 - r_C^H).
\]

To calculate the initial free energy note that the initial state \( \rho^{\text{in}} = \tau \otimes \tau_B \otimes \tau_C \) is the same as \( \rho^H = \tau \otimes \tau_B \otimes \tau_C^H \) with \( T_H = T_R \). Hence by setting \( T_H = T_R \) in the above result,

\[
F^{\text{in}} = T_R \ln(r_B r_C^H).
\]

Therefore

\[
\Delta F_{\text{inc,int}} = F^{\text{fin}} - F^{\text{in}} = E_C \frac{T_H - T_R}{T_H} (1 - r_C^H) + T_R \ln \left( \frac{r_C^H}{r_C} \right).
\]

The temperature achieved on the target qubit is the same as in the single-cycle incoherent scenario of Sec. VI A and reads

\[
r_{\text{inc,int}} = r_B + (1 - r_C^H)(1 - r_B) + r (1 - r_B),
\]

\[
T_{\text{inc,int}} = \frac{E}{\ln \frac{r_{\text{inc,int}}}{1 - r_{\text{inc,int}}}.
\]

The coherent scenario allows one to implement any unitary on qubit C and then performing the energy-conserving unitary on the three-qubit system in the subspace span (\(|010\rangle, |101\rangle\)). After applying the unitary to qubit C the state looks like

\[
\rho^U = \tau \otimes \tau_B \otimes U \tau_C U^\dagger,
\]

where

\[
U = \begin{pmatrix} a & b \\ -b^* e^{i\theta} & a^* e^{i\theta} \end{pmatrix}.
\]
with $\theta \in [0, 2\pi]$ and $|a|^2 + |b|^2 = 1$, $a, b \in \mathbb{C}$. Hence,
\[
U_{T_\mu} U^\dagger = \left(\begin{array}{cc}
1 - |b|^2 & |b|^2 (1 - r_C) \\
a^* b^* e^{i\theta} & (1 - 2r_C)
\end{array}\right),
\]
where in the second step we made the choice of $a$ and $b$ being real, $\theta = 0$, and $b^2 = \mu$. Note that making this choice does not influence the performance of $U$ since for this only the value of $r_C^U$, which is not altered by the choice, matters. In any case, to maximally cool the target qubit for a given state of qubit $C$, one notices that the energy-conserving unitary $U_{\text{cons}}$ need be chosen as $U_{\text{cons}} = \left(\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right)$ in the span($\{|0\rangle, |1\rangle\}$) subspace and as identity elsewhere, such that for the final state $\rho_{\text{fin}} := U_{\text{cons}} \rho U_{\text{cons}} \hat{\rho}$ need not be, either one of the machine qubits. The effect of this swap operation can be very simply understood in terms of the “virtual qubit” subspace of the machine qubits. This section presents the cooling effect of the swap in terms of the virtual qubit, as was done in Ref. [45]. All of the results in the case of repeated operations (and some of those in the single-cycle regime) follow from this argument. For a proof of the statement, see Ref. [45], Appendix A.

**APPENDIX G: THE SWAP OPERATION AND THE VIRTUAL QUBIT AS A BASIS FOR COOLING OPERATIONS**

In all of the paradigms discussed in this work, the operation that causes the target qubit to be cooled down is a swap operation between the target qubit and a qubit subspace in the joint system of the machine qubits. The latter can, but need not be, either one of the machine qubits. The effect of this swap operation can be very simply understood in terms of the “virtual qubit” subspace of the machine qubits. This section presents the cooling effect of the swap in terms of the virtual qubit, as was done in Ref. [45]. All of the results in the case of repeated operations (and some of those in the single-cycle regime) follow from this argument. For a proof of the statement, see Ref. [45], Appendix A.

Let $A$ be a real (target) qubit system that begins in a state that is diagonal w.r.t. the energy eigenbasis (denoted by $\{|0\rangle, |1\rangle\}$, with the population of its ground state (i.e., the corresponding diagonal element in the density matrix)

![Plot](image-url)

**FIG. 11.** The internal resource versions of the coherent (solid blue) and the incoherent (dashed red) scenarios are compared. The energy gaps are fixed to $E = 1$ and $E_C = \frac{1}{4}$ and the environment temperature to $T_R = 1$. One sees that the incoherent scenario always outperforms the coherent one for temperatures that are reachable to both scenarios.
denoted by \( r \). Denote the energy difference between the two levels by \( E \). In addition, consider another system \( M \) (representing the machine), that has in particular a two-dimensional subspace spanned by the energy eigenstates \((|E_0\rangle, |E_1\rangle)\), this subspace is referred to as the “virtual qubit.” We denote by \( E_V = E_0 - E_1 \), the energy gap of the virtual qubit. The initial state of the machine, expressed as a density matrix in the energy eigenbasis, is assumed to have no coherence w.r.t. the eigenstates of the virtual qubit, i.e., the coefficients of \(|E_0\rangle\langle E_i|\) are zero for all \( i \) (except the diagonal element \( i = g \)), and similarly for \(|E_1\rangle\langle E_i|\).

Let the population in the \(|E_0\rangle\) state (the coefficient of \(|E_0\rangle\langle E_1|\) in the density matrix) be denoted as \( p_g \), and that in the \(|E_1\rangle\) state be denoted by \( p_e \). We label by \( N_V \) (the “norm” of the virtual qubit), the total population in the virtual qubit, \( N_V = p_g + p_e \).

\( r_V \) the normalized ground-state population of the virtual qubit, \( r_V = p_g/N_V \), i.e., the population if the virtual qubit was normalized to have \( N_V = 1 \).

\( E_g \) the bias of the virtual qubit, also normalized, \( E_g = (p_g - p_e)/N_V \).

\( T_V \) the virtual temperature of the virtual qubit, calculated by inverting its Gibb’s ratio,

\[
\frac{p_g}{p_e} = e^{-E_g/T_V}.
\] (G1)

Alternatively, the virtual temperature can also be expressed in terms of the bias, via the relation

\[
Z_V = \tanh \left( \frac{E_V}{2T_V} \right).
\] (G2)

Let a swap operation be performed between the real and virtual qubits, described by the unitary

\[
U = 1_{AM} - |0, E_g\rangle\langle 0, E_e| - |1, E_g\rangle\langle 1, E_e| + |1, E_g\rangle\langle 0, E_e| + |0, E_g\rangle\langle 1, E_e|.
\] (G3)

Then the final reduced state of the target qubit will have a new ground-state population given by

\[
r' = r + (1 - r)p_g - rp_e = N_V r_V + (1 - N_V) r,
\] (G4)

i.e., with probability \( N_V \), the new populations of the target qubit are those of the normalized virtual qubit, and with probability \( 1 - N_V \), there is no change. We assume \( N_V \neq 0 \), as this corresponds to the virtual qubit being empty.

One can also express the above in the form

\[
\frac{r_V - r'}{r_V - r} = 1 - N_V.
\] (G5)

Thus, if after a single swap, the machine is restored to its state before the unitary, and then the swap is repeated, the recursive relation between \( r \) and \( r' \) will hold for the new population \( r'' \) in terms of \( r' \). In general, if the reset of the machine and the swap are repeated in turn \( n \) times, then the ground-state population of the target qubit after the \( n \)th step will be

\[
\frac{r_V - r^{(n)}}{r_V - r} = (1 - N_V)^n,
\] (G6a)
equivalently,

\[
r^{(n)} = r_V - (r_V - r)(1 - N_V)^n.
\] (G6b)

In the asymptotic limit of infinite swaps, \( r \rightarrow r_V \). This is equivalent to the Gibbs ratio of the target qubit approaching that of the virtual qubit, and the bias of the target qubit approaching \( Z_V \).

In terms of temperature, if the target qubit and virtual qubit have the same energy gap (\( E = E_V \)), then the temperature of the target qubit approaches the virtual temperature with each swap, and in the asymptotic limit, \( T \rightarrow T_V \). However, if the energies are unequal, then

\[
T \rightarrow T_V \frac{E}{E_V}.
\] (G7)

since it is the Gibbs ratio that equilibrates to that of the virtual qubit.

Finally, one can calculate the work cost of the swap operation. Since it is unitary, the energy difference and free energy difference are the same, and given by

\[
\Delta F = \text{Tr}(\rho' H) - \text{Tr}(\rho H),
\] (G8)

where \( \{\rho, \rho'\} \) are the initial and final states of the system and machine, and \( H \) is the Hamiltonian of the system and machine.

For the degenerate case, i.e., \( E = E_V \), one finds the work cost to be zero. For the nondegenerate case, the work cost of a single swap is given by

\[
\Delta F = (r' - r)(E_V - E).
\] (G9)

To end this section, we list the relevant virtual qubits for each of the paradigms used in this work (see further Appendixes for details):

1. For single-shot and repeated incoherent operations, the virtual qubit is spanned by the two levels \(|01\rangle_{BC}, |10\rangle_{BC}\) of the machine qubits, with the energy gap of the virtual qubit equal to that of the target qubit \( E_B - E_C \).

2. For repeated coherent operations and algorithmic cooling, the virtual qubit is spanned by the levels \(|00\rangle_{BC}, |11\rangle_{BC}\), with the energy gap being \( E_B + E_C \).

3. For single-shot coherent operations, one requires the swap between the target qubit \( A \) and the machine qubit \( B \), which also falls under the above analysis, here the virtual qubit is simply the machine qubit \( B \) (thus \( N_V = 1 \)). The energy gap is thus \( E_B \). If \( E_C > E \), one also requires the swap between qubits \( A \) and \( C \), where \( C \) can be treated as a virtual qubit.

**APPENDIX II: REPEATED INCOHERENT OPERATIONS**

1. The rate of cooling with repeated incoherent operations

In the case of incoherent operations, the relevant virtual qubit (see Appendix G) is the subspace \(|01\rangle_{BC}, |10\rangle_{BC}\) of the machine qubits. When qubit \( B \) is at the environment temperature \( T_R \) and qubit \( C \) at the hot temperature \( T_H \), one can
calculate the populations and variables of the virtual qubit:

\[ p_{01} = r_B (1 - r_C^H) \], \hspace{1cm} (H1) \\
\[ p_{10} = (1 - r_B)r_C^H \], \hspace{1cm} (H2) \\
\[ N_{V, \text{inc}} = p_{01} + p_{10} = r_B (1 - r_C^H) + (1 - r_B)r_C^H \], \hspace{1cm} (H3) \\
\[ r_{V, \text{inc}} (\equiv r_{\text{inc, } \infty}) = \frac{r_B (1 - r_C^H)}{r_B (1 - r_C^H) + (1 - r_B)r_C^H} \], \hspace{1cm} (H4)

where the labeling of \( r_{V, \text{inc}} \) as \( r_{\text{inc, } \infty} \) will become clear shortly. Equivalently, \( r_{V, \text{inc}} \) can be expressed in terms of the virtual temperature of the virtual qubit,

\[ r_{V, \text{inc}} = \frac{1}{1 + e^{-E/T_r}} \], \hspace{1cm} \text{where} \hspace{1cm} T_{V, \text{inc}}(\equiv T_{\text{inc, } \infty}) = \frac{E}{r_E - r_C^H}. \hspace{1cm} (H5) \]

Thus following the argument in Appendix G, the ground-state population after \( n \) repetitions of the incoherent cycle will be given by

\[ r_{\text{inc, } n} = r_{V, \text{inc}} - (r_{V, \text{inc}} - r) (1 - N_{V, \text{inc}})^n \]. \hspace{1cm} (H6)

Thus in the asymptotic limit of infinite repetitions, as \( 0 < N_V \leq 1 \), we recover \( r_{\text{inc, } \infty} = r_{V, \text{inc}} \), and the temperature of the target qubit in this limit is the virtual temperature, i.e., \( T_{\text{inc, } \infty} \approx T_{V, \text{inc}} \).

In particular, in the limit that the hot bath is at infinite temperature, \( T_H \rightarrow \infty \),

\[ N_{V, \text{inc}}^* = \frac{1}{\pi} \], \hspace{1cm} (H7) \\
\[ r_{\text{inc, } \infty}^* = r_B \], \hspace{1cm} (H8) \\
\[ T_{\text{inc, } \infty}^* = \frac{T_R}{E_B} \], \hspace{1cm} (H9) \\
\[ r_{\text{inc, } n}^* = r_B - \frac{(r_B - r)}{2^n} \]. \hspace{1cm} (H10)

2. The free energy cost of repeated incoherent operations

Here we calculate the free energy cost of repeating the incoherent operations a finite number of times. Since the resource is the hot bath, we will account for \( Q_B \), the heat drawn from it. Among all of the steps involved, only the thermalization of qubit \( C \) involves the hot bath, and so it is sufficient to keep track of the populations of the reduced state of qubit \( C \) in order to calculate \( Q_H^* \).

We can divide the total heat current into two parts, first off, the amount required to heat up qubit \( C \) from the environment temperature \( T_R \) to the temperature of the hot bath \( T_H \). Following that, there is the repeated heat current required to bring back qubit \( C \) to \( T_H \) after a cooling swap has been performed.

The first heat current is trivial to calculate, from the difference in the ground-state population of \( C \) due to heating:

\[ Q_H^* = E_C (r_C - r_C^H). \hspace{1cm} (H11) \]

For the second part, we have to determine the population change, specifically the reduction in the excited state population of qubit \( C \), every time that the cooling swap is performed. However, since the swap is between the levels \( |010 \rangle \) and \( |101 \rangle \), we see that whatever the change in the reduced state populations of \( C \), the change in the corresponding reduced state populations of \( A \) is exactly the same. More precisely, the heat required to reset qubit \( C \) before the \( n \)th swap [i.e., after the \((n - 1)\)th cooling swap] is

\[ Q_H^* = E_C (r_{\text{inc, } n-1} - r_{\text{inc, } n-2}). \hspace{1cm} (H12) \]

which holds for \( n \geq 2 \). From the above two expressions, we thus have the cumulative heat current required for \( n \) cooling steps,

\[ Q_H^* = E_C (r_C - r_C^H) + E_C (r_{\text{inc, } n-1} - r). \hspace{1cm} (H13) \]

In the asymptotic limit of infinite repetitions, \( r_{\text{inc, } n-1} \) goes to \( r_{\text{inc, } \infty} \), and the resultant expression demonstrates that the total heat current is asymptotically finite.

The work cost is given by the decrease in the free energy of the hot bath w.r.t. the temperature of the environment, which is defined as

\[ \Delta F_{\text{inc, } n} = Q_H^* - T_R \Delta S_{H,n} \], \hspace{1cm} (H14) \]

where \( \Delta S_{H,n} \) is the decrease in the entropy of the bath after \( n \) repetitions of the swap. For a thermal bath that stays at equilibrium, as we assume throughout, \( \Delta S_{H,n} = Q_H^*/T_H \), leading to

\[ \Delta F_{\text{inc, } n} = Q_H^* \left( 1 - \frac{T_R}{T_H} \right). \hspace{1cm} (H15) \]

In particular, for the case that \( T_H \rightarrow \infty \), in the asymptotic limit of infinite repetitions of the swap,

\[ \Delta F_{\text{inc, } \infty} = E_C (r_C - \frac{1}{2} + r_B - r). \hspace{1cm} (H16) \]

APPENDIX I: ASYMPTOTIC EQUIVALENCE OF INCOHERENT OPERATIONS AND AUTONOMOUS REFRIGERATOR

In this section, we show that in the two-qubit machine the final state, and hence the final temperature of the target, as well as the total work cost, are the same as if we had run an autonomous refrigerator between the three qubits and waited for the steady state. In other words, since the autonomous refrigerator runs continuously, repeated incoherent operations can be understood as a discretized version of the continuous process. For a discussion on the connection between continuous and discretized versions of quantum thermal machines, see Ref. [59]. Here we simply review the autonomous three-qubit fridge introduced in Ref. [29] and the equivalence of its steady-state parameters with the asymptotic end point of repeated incoherent operations.

In the case of the autonomous fridge, rather than having repeated unitary operations, there is a time-independent interaction Hamiltonian between the three qubits given by

\[ H_{\text{int}} = g(|010\rangle_{ABC} \langle 101| + H.c), \hspace{1cm} (I1) \]

which acts on the degenerate subspace. Note that this Hamiltonian is a generator of the unitary that swaps the population of the degenerate levels, specifically, \( U = \exp(-i \frac{\pi}{2} H_{\text{int}}) \).

At the same time, each qubit is coupled to a thermal bath, qubit \( B \) to the environment, qubit \( C \) to the hot bath. For completeness one could also consider qubit \( A \) to be coupled to its own environment, but for simplicity we ignore this effect here. This is to be consistent with the repeated incoherent
operations picture, where we did not take into account any coupling between qubit A and an environment in between the cooling operations.

As proven in Ref. [29], the three qubits approach a steady state, which is particularly simple in the case that qubit A has no coupling to a bath,

$$\tau_{\text{auto}} \otimes \tau_B \otimes \tau_C^H.$$  \hspace{1cm} (12)

That is, the steady state is a tensor product state, with qubits B and C thermal at the temperatures of the baths they are, respectively, coupled to, and qubit A in Gibbs state with temperature

$$T_{\text{auto}} = \frac{E_{r_B} - E_{r_C}}{E_{r_B} - E_{r_C}}.$$  \hspace{1cm} (13)

This is the same as $T_{\text{inc,\infty}}$ [see Eq. (H5)], which is the asymptotic limit of repeated incoherent operations. Furthermore, it is clear that in the repeated operations, when the number of operations approaches infinite, the cooling swaps stop having an effect, and thus the final states of qubits B and C are Gibbs states at $T_R$ and $T_{\text{eff}}$, respectively, as these are the temperatures they are reset to after each cooling cycle. Thus the final state of all three qubits is the same in both the autonomous and repeated operations scenario.

**Free energy equivalence**

Here we calculate the free energy consumed by the autonomous fridge to go from the initial state to the final state. As the resource is the hot bath, we will calculate the free energy from $Q_{\text{auto}}^H$, the heat drawn from the hot bath. The initial state is that of all three qubits being at the environment temperature $T_R$, while the final state is the tensor product of Gibbs states derived above; see Eq. (12).

Consider the entire system to comprise three parts. Each part consists one of the qubits and the bath that it is attached to (in the case of qubits B and C). The only way that energy is exchanged between the different parts is via the energy-preserving interaction Hamiltonian given by Eq. (H1). This swaps the populations of the two energy eigenstates $|010\rangle$ and $|101\rangle$, and thus the change in population of qubit A due to the interaction is exactly the same as that in qubit C. Since the energy change is given by the population times the energy gap this implies that the energy change of the three parts (at all times during the operation of the machine) must be in proportion to $E : -E_B : E_C$, from the form of $H_{\text{int}}$.

Since part A consists only of the target qubit, the total energy change is simply the difference in energy from the initial to the final state, $E(r - r_{\text{auto}})$. For part C, the total energy change is the sum of that of qubit C, and that of the hot bath, $E_C(r_C - r_{\text{auto}}^C) - Q_{\text{auto}}^H$. Via the preceding argument,

$$E(r - r_{\text{auto}}) = E_C(r_C - r_{\text{auto}}^C) - Q_{\text{auto}}^H.$$  \hspace{1cm} (14)

Solving for $Q_{\text{auto}}^H$, we find that

$$Q_{\text{auto}}^H = E_C(r_C - r_{\text{auto}}^C - r_{\text{auto}} + r).$$  \hspace{1cm} (15)

As $r_{\text{auto}} = r_{\text{inc,\infty}}$, this is the same heat current as in the asymptotic limit of infinite repetitions of the incoherent operation; see Eq. (H13).

**APPENDIX J: REPEATED COHERENT OPERATIONS**

1. **Choosing the best virtual qubit from the machine**

In this section we investigate the effect and optimal strategy for repeated coherent operations. Here we are allowed to repeatedly perform arbitrary unitary operations on the joint system of the target and machine qubits, with the machine qubits being reset to the temperature of the environment in between (see Fig. 12). To begin with, we demonstrate that in terms of asymptotic cooling, the best virtual qubit of the machine to choose is that spanned by $|00\rangle_{BC}, |11\rangle_{BC}$.

First off, w.r.t. the virtual qubit picture, here we can choose any qubit subspace of the machine qubits to swap with the target qubit, unlike in the incoherent case, where we were forced to choose the subspace $|01\rangle_{BC}, |10\rangle_{BC}$, so as to be degenerate ($E_V = E$) with the target system.

However, in the coherent case, there is only a single temperature available ($T_B$), thus the state of the machine after it is rethermalized to the environment will simply be the thermal state of qubits B and C at $T_B$. Given that the entire state of B and C is thermal, every qubit subspace of the machine has the same virtual temperature, $T_V = T_B$.

From Appendix G, Eq. (G7), we conclude that if we pick a virtual qubit from the machine with energy gap $E_V$, then the temperature of the target system after many repetitions of the swap between the virtual qubit and the target (with the reset of the machine in between) will tend to

$$T \rightarrow T_B \frac{E}{E_V}.$$  \hspace{1cm} (J1)

Thus to cooling the maximum amount in the asymptotic case of infinite repetitions, we should pick the largest energy gap, i.e., the qubit subspace $|00\rangle_{BC}, |11\rangle_{BC}$. In what follows, we show that in fact, after the first coherent operation (which was dealt with in Sec. VI B), this is the only virtual qubit that allows for cooling the target.

2. **The target qubit after n repetitions of coherent operations**

Consider the state of the three qubits at the end of the single coherent operation. The initial state before the operation was the thermal state of all three qubits at the environment temperature $T$. If the energies satisfy $E \geq E_C$, then the optimal coherent operation is simply to swap the states of A and B,
leaving the three qubits in the state
\[ \rho' = \tau_B \otimes \tau_A \otimes \tau_C, \]  
whereas if \( E < E_C \), then the optimal coherent operation is to first swap the states of \( A \) and \( C \), and then proceed by swapping \( A \) with \( B \), leading to the final state of
\[ \rho' = \tau_B \otimes \tau_C \otimes \tau_A. \]

In either case, there is no further cooling on qubit \( A \) possible by any unitary operation. Thus the only option to continue is to bring the machine back to the environment temperature. At this point, the state is now given by
\[ \tilde{\rho} = \tau_B \otimes \tau_B \otimes \tau_C. \]

Recall that the first four populations (i.e., eigenvalues) are those in the ground state of qubit \( A \). Labeling all of the populations from \( p_{000} \) to \( p_{111} \), one can verify (using \( E_B > E_C \)) that
\[ p_{000} > p_{001} = p_{100} > p_{011} = p_{101} > p_{110} > p_{111}. \]

Thus from the perspective of maximizing the ground-state population of \( A \), the only two populations that are not in the optimal location are \( p_{011} \) and \( p_{100} \), which should be swapped, corresponding to unitarily swapping the two energy levels \( |011 \rangle \) and \( |100 \rangle \). This is unlike the initial state before the first coherent operation, where there were a number of possible level swaps that achieved cooling. There one had to optimize over all possible swap operations to minimize the work cost, whereas here there is only one possible cooling swap. Thus the second coherent operation continues with the \( |100 \rangle \leftrightarrow |011 \rangle \) swap, cooling down qubit \( A \) further, followed by bringing back the machine qubits \( B \) and \( C \) to the environment temperature. One can verify that after resetting the machine qubits, the populations once again satisfy \( p_{011} < p_{100} \), allowing cooling to continue by repetition of this cycle of steps. In the same manner as for repeated incoherent operations, from the arguments of Appendix G, one can identify the properties of the relevant virtual qubit in this case, the states \( |00 \rangle_{BC} \) and \( |11 \rangle_{BC} \) of the machine,
\[ p_{00} = r_B r_C, \]
\[ p_{11} = (1 - r_B)(1 - r_C), \]
\[ p_{01} + p_{10} = r_B r_C + (1 - r_B)(1 - r_C), \]
where \( N_{V,coh} = p_{00} + p_{11} = r_B r_C + (1 - r_B)(1 - r_C) \),
\[ r_{V,coh} = \frac{r_B r_C}{r_B r_C + (1 - r_B)(1 - r_C)}, \]
where the labeling of \( r_{V,coh} \) as \( r_{coh, \infty} \) will become clear shortly. Equivalently, \( r_{V,coh} \) can be expressed in terms of the virtual temperature of the virtual qubit,
\[ r_{V,coh} = \frac{1}{1 + E^{-E_C/T_{V,coh}}}, \]
where \( E_V = E_B + E_C \) and
\[ T_{V,coh} = T_R. \]

Thus following the argument in Appendix G, the ground-state population after \( n \) repetitions of the incoherent cycle will be given by
\[ r_{coh,n} = r_{V,coh} - (r_{V,coh} - r)(1 - N_{V,coh})^n. \]

Thus in the asymptotic limit of infinite repetitions, \( r_{coh,\infty} = r_{V,inc} \), and the temperature of the target qubit in this limit is
\[ T_{coh,\infty} = T_{V,coh} \frac{E}{E_{V,coh}} = T_R \frac{E}{E_B + E_C}. \]

3. The free energy cost of cooling with repeated coherent operations

In the case of repeated coherent operations, the work cost is only calculated from the unitary swap operations, as the other step is the thermalization of the machine to the environment temperature, which comes at no cost. To calculate the work cost of the unitary operations, we follow the argument in Appendix G. From the argument therein [Eq. (G9)], the free energy in each repeated coherent operation is given by
\[ \Delta F_{coh,n}^* - \Delta F_{coh,n-1}^* = \frac{r_{coh,n}^*}{(E_B + E_C - E)} \frac{r_{coh,n}^*}{(1 - r_{coh,n}^*)}. \]

This only applies for \( n \geq 2 \) since the first coherent operation is different, and the optimal work cost of the latter (\( \Delta F_{coh}^* \)) has been calculated in Sec. VI.B. Recalling that the ground-state population of the target qubit after a single coherent operation is \( r_B \), we can calculate the work cost of \( n \) repetitions of coherent operations,
\[ \Delta F_{coh,n}^* = \Delta F_{coh}^* + 2E_C(r_{coh,n}^* - r_B), \]
where
\[ \Delta F_{coh}^* = \begin{cases} E_C(r_B - r) & \text{if } E_C \leq E, \\ (E_C - E)(r_C - r) + E_C(r_B - r_C) & \text{if } E_C \geq E. \end{cases} \]
environment temperature, and then its state is swapped with that of qubit C, thus cooling the state of C. Qubit B is then rethermalized to $T_R$, and then the cooling swap involving all three qubits is repeated.

The state of the machine qubits prior to the swap is now a tensor product of two copies of the thermal state of qubit $B$ w.r.t. $T_R$, and so the virtual qubit $\{00\}_BC,\{11\}_BC$ has the following properties:

$$p_{00} = r_B^2,$$  \quad (K1)

$$p_{11} = (1 - r_B)^2,$$  \quad (K2)

$$N_{V,algo} = p_{00} + p_{11} = r_B^2 + (1 - r_B)^2,$$  \quad (K3)

$$r_{V,algo} (= r_{algo,\infty}^*),$$

$$r_{V,algo} = \frac{1}{1 + e^{-E_{V,algo}/r_{V,algo}}},$$  \quad (K4)

where the labeling of $r_{V,algo}$ as $r_{algo,\infty}^*$ will become clear shortly. Equivalently, $r_{V,algo}$ can be expressed in terms of the virtual temperature of the virtual qubit,

$$r_{V,algo} = \left\{\begin{array}{ll}
\frac{1}{1 + e^{-E_{V,algo}/r_{V,algo}}}, & \text{where } E_{V,algo} = E_B + E_C \\
\frac{1}{1 + e^{-E_{V,algo}/r_{V,algo}}}, & \text{and } T_{V,algo} = \frac{E_B + E_C}{2E_B}.
\end{array}\right. \quad (K5)$$

Thus following the argument in Appendix G, the ground-state population after $n$ repetitions of algorithmic cooling will be given by

$$r_{algo,n} = r_{V,algo} - (r_{V,algo} - r_0)(1 - N_{V,algo})^n,$$  \quad (K6)

where $r_0$ is the ground-state population of the target before starting the algorithmic cooling procedure. $r_0$ can be $r$, in the case that we begin with algorithmic cooling from the initial state, but can also be anything else, in particular some population greater than $r$, corresponding to the endpoint of a different type of cooling operation. Finally note that in the asymptotic limit of infinite repetitions, $r_{algo,\infty}^* = r_{V,algo}$, and the temperature of the target qubit in this limit is given by

$$T_{algo,\infty}^* = T_{V,algo} \frac{E}{E_{V,algo}} = T_R \frac{E}{2E_B},$$  \quad (K7)

which is independent of $r_0$, the initial ground-state population of the target.

**The free energy cost of algorithmic cooling**

Analogous to the case of repeated coherent operations, here the work cost is invested during the unitary operations. However, in addition to the cooling swap involving all three qubits, whose cost is calculated in exactly the same way as in the repeated coherent case[see Appendix J3], there is also the precooling of qubit $C$, which is a nonenergy-preserving unitary operation. Since this is effected by a swap between qubits $B$ and $C$, the work cost per population swapped (in the direction of cooling $C$) is $E_B - E_C = E$.

The work cost of precooling $C$ can be split into two contributions: first, the initial cost of cooling $C$ from the environment temperature $T_R$ to the state that has the same populations as that of the initial state of $B$, that costs $E(r_B - r_C)$, and then the work cost of returning it to the precooled state after every successive three-qubit swap. Since the three qubit swap is between the states $|011\rangle$ and $|100\rangle$, we see that whatever the change in the population of the ground state of the target qubit, there is exactly the same decrease in the ground-state population of qubit $C$.

Adding up all of these contributions, one finds that the free energy cost of algorithmic cooling is given by

$$\Delta F_{algo,n} = 2E_C(r_{algo,n} - r_0) + E(r_B - r_C) + E(r_{algo,n-1} - r_0).$$  \quad (K8)

where the first term is the total work cost of the cooling swap on all three qubits, the second term is the cost of precooling qubit $C$ from its initial state thermal at $T_R$, and the third represents the cost of returning qubit $C$ to the precooled state prior to the $n$th cooling swap. As before, $r_0$ is the ground-state population of the target before starting the algorithmic cooling procedure.

**APPENDIX L: OPTIMIZING THE REPEATED OF COHERENT OPERATIONS W.R.T. THE WORK COST**

In the case of coherent operations, we now have a number of different procedures for cooling. Recall that in the single-cycle case, we found that we could cool by simply swapping the target qubit $A$ with $B$. Furthermore, if it is the case that $E < E_C$, then a lower work-cost can be achieved by swapping the target qubit with $C$ to begin with. For repeated coherent operations, we have to swap the target qubit with the virtual qubit $\{00\}_BC,\{11\}_BC$. And finally, to cool the maximum we should precool qubit $C$ (which is a swap between qubits $B$ and $C$) prior to the same cooling swap.

Each of these processes has a different work cost, and it is illuminating to construct the optimal manner of combining them to have the minimum work cost. Following the argument in Appendix G, Eq. (G9), we understand that to optimize the work cost, we should always seek to swap the target qubit with a virtual (or real) qubit of as small an energy gap as we can find, given it has a greater normalized ground-state population $r_V$ than the ground-state population of the target. This way we minimize the energy gradient over which we move population, and thus minimize the work cost.

At the beginning, when the target and machine qubits are at the environment temperature, if $E_C > E$, then one can verify from the machine state that among all the virtual qubits of the machine with greater normalized ground-state population $r_V$ than $r$, qubit $C$ (here it is a real qubit, rather than virtual) is the one that has the smallest energy difference with $E$, $E_V - E$. Thus the minimal cost of cooling is to swap these states, taking $r \rightarrow r_C$ at a gradient of $E_C - E$.

Once this procedure is exhausted and the ground-state population of qubit $A$ has become $r_C$, we find that among the above virtual qubits of the machine, qubit $B$ has the second smallest energy difference with $E$, and so one proceeds by swapping the target qubit with qubit $B$, taking $r \rightarrow r_B$, at a gradient of $E_B - E$. One then rethermalizes the machine qubits to $T_R$. Note that qubit $C$ could have equivalently been rethermalized at any point between the end of the first swap and now without affecting the cooling and the work cost of the procedure.
At this point, after resetting the machine qubits, we find that the only virtual or real qubit in the machine that allows for cooling is the virtual qubit $|00\rangle_{BC}, |11\rangle_{BC}$, and one proceeds by repeatedly swapping the target qubit with this virtual qubit, until $r \rightarrow r_{V,\text{coh}}$. This is performed at a gradient of $E_B + E_C - E$.

Finally, one proceeds via algorithmic cooling, where one precools qubit $C$, at a gradient of $E_B - E_C$, before applying the same cooling swap as in the case of repeated coherent operations. The reason one exhausts the repeated coherent operations procedure before proceeding with algorithmic cooling is that precooling qubit $C$ has a work cost that arguably enables one to cool more but still at the same energy rate, $2E_C$. Thus, as long as cooling without this extra work-cost is possible, it is more efficient to do so.

The work cost at an intermediate stage in this process can be simply calculated from the above, we present here the total work cost of the entire procedure:

$$
\Delta F_{\text{algo, } \infty}^* = (r_C - r)(E_C - E) + (r_B - r)(E_B - E) \\
+ (r_{\text{coh, } \infty} - r_B)2E_C \\
+ (r_B - r_C)(E_B - E_C) \\
+ (r_{\text{coh, } \infty} - r_{\text{coh}})(E_B - E_C) + 2E_C,
$$

where the first, second and third lines correspond to the work cost of the single-cycle, repeated and algorithmic sections of the protocol respectively. In the case of $E_C \leq E$, the single-shot case simplifies to directly swapping the target qubit with qubit $B$, and thus the first line of the work cost becomes $(r_B - r)(E_B - E).

It is interesting to observe that subdividing the entire procedure in this manner, the temperature of the target qubit evolves due to each subsequence as

$$(r_{\text{coh}}^* - r_B)2E_C \\
+ (r_B - r_C)(E_B - E_C) \\
+ (r_{\text{coh}, \infty}^* - r_{\text{coh}}^*)(E_B - E_C) + 2E_C.
$$

An optimal cooling sequence in the regime of algorithmic cooling

In the analysis above, we noted that algorithmic cooling is more expensive as it requires the precooling of qubit $C$. Furthermore, if one precools $C$ via a full swap with $B$, as presented above, this represents an initial work cost which does not cool down the target at all, representing a discontinuity in the curve of cooling versus work cost. This is especially relevant if the desired final temperature is not that corresponding to algorithmic cooling, but is rather somewhere in between algorithmic cooling and the endpoint of repeated coherent operations.

In this case, one can optimize the work cost by using the same cycle of steps as in Fig. 13, but only partially precooling qubit $C$ in Step 2, to exactly the temperature required to achieve the desired final temperature on the target.

More precisely, consider that during Step 2, one performs a partial swap between qubits $B$ and $C$, such that the final ground-state population of qubit $C$ is given by

$$r_C(v) = r_C + v(r_B - r_C),$$

where $v \in [0, 1]$. On inspection of the virtual qubit $|00\rangle_{BC}, |11\rangle_{BC}$, we can calculate the normalized ground-state population $r_C(v)$,

$$
r_{V,\text{algo}} = \frac{r_B \cdot r_C(v)}{r_B \cdot r_C(v) + (1 - r_B)(1 - r_C(v))}.
$$

Note that $r_{\text{coh}, \infty}^* < r_{V,\text{algo}} < r_{\text{algo}, \infty}^*$, with $r_{V,\text{algo}} = r_{\text{coh}, \infty}^*$ and $r_{V,\text{algo}} = r_{\text{algo}, \infty}^*$, and thus $v$ parameterizes the entire regime of cooling between the endpoint repetitive coherent operations, and full algorithmic cooling.

In the limit of infinite repetitions of the cycle of steps, the ground-state population of the target becomes $r_{V,\text{algo}}$, such that, given the desired final temperature of the target, $T_{\text{v,algo, } \infty}$, the swapping parameter $v$ need be chosen such that

$$r_{V,\text{algo}} = \frac{1}{1 + e^{-\frac{1}{T_{\text{v,algo, } \infty}}}}.
$$

The work cost of cooling the target to $r_{V,\text{algo}}$, given that we began with the target ground-state population of $r_0$, is found by adding up the cost of precooling qubit $C$, the cost of returning it the precooled state, and the cost of the repeated cooling swaps on the target,

$$
\Delta F_{\text{algo, } \infty} = E(r_C(v) - r_C) + E(r_{V,\text{algo}} - r_0) \\
+ 2E_C(r_{V,\text{algo}} - r_0).
$$

Thus given the endpoint of repeated coherent operations, (where $r_0 = r_{\text{coh}, \infty}^*$), the above expression represents the optimal extra work cost for cooling the target to a ground-state population [Eq. (L4)] that is between the end points of repeated coherent operations and algorithmic cooling. The total work cost of the optimal sequence is in this case therefore given by

$$
\Delta F_{\text{algo, } \infty} = (r_C - r)(E_C - E) + (r_B - r_C)(E_B - E) \\
+ (r_{\text{coh, } \infty}^* - r_B)2E_C
$$

FIG. 13. The cycle of steps corresponding to algorithmic cooling. Steps 1 and 3 thermalize qubit $B$ to the environment. Step 2 is the precooling of qubit $C$ by a swap with $B$. Step 4 is the cooling of the target qubit via the usual coherent operation. In the case of optimizing algorithmic cooling w.r.t. the work cost (see Appendix L.1), Step 2 is replaced by a partial rather than full swap.
+ (r_C(v) - r_C)(E_R - E_C) \\
+ (r_{v,\text{algo}} - r_{\text{coh},\infty})(E_R - E_C) + 2E_C). \quad \text{(L7)}

Note that for $v = 1$, we recover the previously discussed total work cost of the optimal sequence of coherent operations of Eq. \text{(L1)}.

**APPENDIX M: OPTIMIZING THE WORK COST**

1. The $N$-qubit coherent machine

In this section we review a result demonstrated in Ref. \[12\] (within a different context), that given a final cold temperature, there exists a family of coherent machines, each member of increasing size, that can attain the final temperature, and that saturate the second law of thermodynamics in the limit of infinite size. We do this for coherent machines first, and prove the same for incoherent machines in the next section.

Consider the system to be a qubit of energy $E$ (the result may be generalized by cooling individual qubit subspaces of a more complicated system), that begins at the environment temperature $T_R$. The final temperature that we would like to attain is labeled $T_C$, where $T_C < T_R$.

The simplest machine to do so would be a single qubit of energy

$$E_{\text{coh, max}} = E \frac{T_R}{T_C}, \quad \text{(M1)}$$

as in Sec. \text{IV B}, and perform a swap in the energy eigenbasis. Note that the machine is assumed, as always, to begin at $T_R$.

As discussed in the main text and in Appendix G on the virtual qubit, the work cost of this protocol involves pushing population against the energy gradient between the machine and system, $E_{\text{coh, max}} - E$,

$$W = (r_{\text{max}} - r)(E_{\text{coh, max}} - E) = (r_{\text{max}} - r)E \left( \frac{T_R}{T_C} - 1 \right), \quad \text{(M2)}$$

where $r$ and $r_{\text{max}}$ are the initial and final ground-state populations of the target.

One can reduce this work cost by splitting the protocol into a number of steps. Consider that the machine is constructed out of a sequence of $N$ qubits, with linearly increasing energy,

$$E_{\text{coh}, i} = E + \frac{i}{N}(E_{\text{coh, max}} - E) = E \left[ 1 + \frac{i}{N} \left( \frac{T_R}{T_C} - 1 \right) \right], \quad \text{i} \in \{1, 2, \ldots, N\}. \quad \text{(M3)}$$

The protocol now consists in performing swap operations between the target qubit and each of the machine qubits in sequence. The final temperature is the same as before, since the final qubit has energy $E_{\text{coh, max}}$. At each intermediate step, the temperature attained by the target is given by

$$E \frac{T_i}{T_R} = \frac{E_{\text{coh}, i}}{T_R} = E + \frac{i}{N}(E_{\text{coh, max}} - E) \quad \text{(M4)}$$

\[ \vdots \]

$$\Rightarrow \quad \frac{1}{T_i} = \frac{1}{T_R} + \frac{i}{N} \left( \frac{1}{T_C} - \frac{1}{T_R} \right). \quad \text{(M5)}$$

Correspondingly, the ground-state population of the target after the $i$th step is given by

$$r_i = \frac{1}{1 + e^{-E/T_i}}. \quad \text{(M6)}$$

The work cost of the $i$th step is now

$$W_{\text{coh}, i} = (r_i - r_{i-1})(E_{\text{coh}, i} - E), \quad \text{(M7)}$$

from which the total cost follows as

$$W_{\text{coh}} = \sum_i W_{\text{coh}, i} = \sum_{i=1}^N (r_i - r_{i-1})(E_{\text{coh}, i} - E) \quad \text{(M8)}$$

$$= E \sum_{i=1}^N (r_i - r_{i-1}) \frac{i}{N} \left( \frac{T_R}{T_C} - 1 \right). \quad \text{(M9)}$$

In Ref. \[12\] this protocol was studied, and it was shown that the total work cost was equal to

$$W_{\text{coh}} = \Delta F + O\left( \frac{1}{N} \right). \quad \text{(M10)}$$

where $\Delta F$ is the increase in free energy of the system from its initial to final temperature, and where the free energy is defined w.r.t. the environment temperature,

$$F = \langle E \rangle - T_R S, \quad \text{(M11)}$$

where $\langle E \rangle$ and $S$ being the average energy and entropy of the system. Thus one can get arbitrarily close to saturating the second law of thermodynamics by increasing the number of steps involved in the protocol.

Note that the qubits in the coherent machine need not be real qubits, but qubit subspaces (virtual qubits) embedded in a larger space. In this case, rather than a single swap for each of the machine qubits, one would require repeated swaps (interspersed with rethermalization of the machine) to approach the asymptotic temperature corresponding to that qubit. This does not, however, change the work cost of the procedure, since the cost is always given by the amount of population changed multiplied by the energy gradient, so repeating the swap with the same virtual qubit a number of times to achieve the same population difference as with a real qubit of the same energy gap results in the same work cost.

2. The $2N$-qubit incoherent machine

Consider as before that we wish to cool a target qubit of energy $E$ from the environment temperature $T_R$ to $T_C$, but only using incoherent operations, that include energy-preserving unitaries, and heating up parts of our machine to a given hot temperature $T_H$.

The simplest manner of achieving this temperature is via the simplest possible incoherent machine, comprises two qubits (as in Sec. \text{VII A}), of energies $E_R = E_{\text{inc, max}}$ and $E_C = E_{\text{inc, max}} - E$. The machine may be run in the repeated operations regime, where the energy-preserving swap operation between the states $|01\rangle_{ABC}$ and $|10\rangle_{ABC}$ is repeatedly applied, interspersed by rethermalizing qubits $B$ and $C$ to the environment $T_R$ and hot bath $T_H$ respectively, or in the autonomous mode, where the Hamiltonian that generates the
The final temperature achieved by such a machine is given by

$$\frac{E}{T^f} = \frac{E_{\text{inc,max}}}{T_R} - \frac{E}{T_H},$$

(M12)

and so we choose $E_{\text{inc,max}}$ such that $T^f = T_C$, the final desired cold temperature, resulting in

$$E_{\text{inc,max}} = E \left(\frac{\frac{1}{T_C} + \frac{1}{T_R}}{\frac{1}{T_C} + \frac{1}{T_H}}\right) \Leftrightarrow E_{\text{inc,max}} - E = E \left(\frac{\frac{1}{T_C} - \frac{1}{T_R}}{\frac{1}{T_C} + \frac{1}{T_R}}\right).$$

(M13)

The work cost of this protocol was discussed in Sec. I, and may be calculated from the heat drawn from the hot bath during the protocol. The heat comprises two parts, the preheating of qubit bath during the protocol. The heat drawn in each step again comprises the two contributions of preheating and maintenance of the $i$th qubit $C$,

$$Q_i = Q_{\text{init},i} + (r_i - r_{i-1}) E_{\text{coh}},$$

(M19)

where we label the initial heat drawn for preheating as $Q_{\text{init},i}$. Simplifying the rest of the expression using (M13), and summing up for the total heat,

$$Q = \sum_{i=1}^{N} Q_{\text{init},i} + E \sum_{i=1}^{N} (r_i - r_{i-1}) \frac{i}{N} \left(\frac{\frac{1}{r_C} - \frac{1}{r_R}}{\frac{1}{r_C} + \frac{1}{r_R}}\right).$$

(M21)

using (M9) for the total work cost in the coherent case. We have denoted by $Q_{\text{init}}$ the total cost of preheating each of the qubits $\{C, i\}$.

The above is the heat drawn from the hot bath. To compare with the coherent case we take the work cost instead, which is the decrease in free energy of the hot bath,

$$W = \Delta F_H = Q - T_R \Delta S = Q \left(1 - \frac{r_R}{r_H}\right).$$

(M23)

Thus the work cost in the incoherent case is

$$W_{\text{inc}} = Q_{\text{init}} \left(1 - \frac{r_R}{r_H}\right) + W_{\text{coh}}.$$

(M24)

Thus the incoherent cost is very similar to the coherent cost, with the sole addition of bringing the additional qubits from $T_R$ to $T_H$. At first glance, this may appear to be a finite disadvantage; however, it is possible to make this additional cost as small as possible, as we now demonstrate.

The $N$ qubits $\{C, i\}$ need not be real qubits, but virtual ones. Consider for instance, a system with Hamiltonian

$$H_C = -E_g \langle E_g \rangle + \sum_{i=0}^{N} \frac{i}{N} (E_{\text{inc,max}} - E),$$

(M25)

which is an evenly spaced ladder of $N+1$ levels plus a single ground state that lies at an energy $E_g$ below the ladder. Labeling the levels by $\{g, 0, 1, 2, \ldots, N\}$, we observe that $E_i - E_0 = E_{\text{coh},i}$, and thus the pair of levels 0 and $i$ may be employed as the $i$th virtual qubit $C$ in the incoherent machine.

However, for any fixed $N$, $T_H$, and $E_{\text{inc,max}}$, the cost of preheating this system can be made as small as we like, by pushing the ground-state energy further downward. For high enough values of $E_g$, the population in the ladder will be small enough for both $T_H$ and $T_R$ such that the difference in average energy is vanishingly small.

Note that this implies that the machine will run slower (in the autonomous case) or require many more repeated operations in the discrete case. However, in principle the final temperature attained is still the same and thus the incoherent machine can achieve as close a work cost as one likes to the coherent case. Together with the fact that the coherent
machine can get as close to saturating the second law, we thus have the statement that for arbitrary sized machines, both coherent and incoherent machines can approach the limit of the second law of thermodynamics.

[34] P. P. Hofer, J.-R. Souquet, and A. A. Clerk, Quantum heat engine based on photon-assisted Cooper pair tunneling, Phys. Rev. B 93, 041418(R) (2016).


[56] We consider the system to be isolated from the environment during the operations. Taking the system-environment interaction into account can only weaken cooling and requires a comparison of the timescales of different operations, which we do not go into in this work.


[62] In recent literature, the term self-contained is sometimes used to mean the same as autonomous, i.e., a machine with no external source of work or control.