Physical Limits of Heat-Bath Algorithmic Cooling

Leonard J. Schulman,[1] Tal Mor,[2] and Yossi Weinstein[2]
[1]*California Institute of Technology, MC 256-80, Pasadena, CA 91125, USA*
[2]*Technion – Israel Institute of Technology, Haifa, Israel*

(Received 30 March 2004; revised manuscript received 20 October 2004; published 1 April 2005)

Simultaneous near-certain preparation of qubits (quantum bits) in their ground states is a key hurdle in quantum computing proposals as varied as liquid-state NMR and ion traps. "Closed-system" cooling mechanisms are of limited applicability due to the need for a continual supply of ancillas for fault tolerance, and to the high initial temperatures of some systems. "Open-system" mechanisms are therefore required. We describe a new, efficient initialization procedure for such open systems. With this procedure, an n-qubit device that is originally maximally mixed, but is in contact with a heat bath of bias $\varepsilon \gg 2^{-n}$, can be almost perfectly initialized. This performance is optimal due to a newly discovered threshold effect: for bias $\varepsilon \ll 2^{-n}$ no cooling procedure can, even in principle (running indefinitely without any decoherence), significantly initialize even a single qubit.

DOI: 10.1103/PhysRevLett.94.120501 PACS numbers: 03.67.Lx, 03.65.Yz, 76.60.–k, 82.56.–b

Quantum computation poses a difficult experimental challenge. Simultaneous near-certain preparation of qubits (quantum bits) in their ground states is a key hurdle in proposals as varied as NMR and ion traps [1–6]. Such "cooling" (also known as "biasing" or "polarizing") is required both for initiation of the computation [7] and in order to supply ancillas for fault tolerance as the computation proceeds.

Cooling of quantum systems has long been essential in a variety of experimental contexts unrelated to quantum computation, and is performed by processes that directly cool the system such as laser cooling in ion traps or application of strong magnetic fields in NMR. Spin exchange has also been employed in order to transfer highly-cooled states into the desired system from another that is more readily directly cooled [8–10]. In all these methods, the temperature is limited by the original cooling process.

Algorithmic cooling.—It is in principle possible, however, to reach even lower temperatures by application of certain logic gates among the qubits [11]. (Even prior to quantum computation, the need for signal amplification in NMR imaging led to the implementation of a basic 3-qubit logic gate [12].) In several quantum computation proposals this kind of improvement in cooling is necessary due to the requirement that a large number of qubits all be, with high probability, simultaneously in their ground states.

We distinguish between closed- and open-system algorithmic cooling methods. In the former [3,4,11] an initial phase of physical cooling is performed, which reduces the entropy of the system. In the second phase an entropy preserving (unitary) algorithmic process is performed on the qubits. By contrast, in an open process [13] some of the qubits of the system can be cooled by external interaction even during (or at interruptions in) the quantum computation. Open-system cooling places an additional experimental difficulty: computation qubits must not decohere during the process of cooling other qubits with which, at another stage, they must interact. Nonetheless, closed-system cooling appears to be insufficient for two reasons. The first applies specifically to liquid-state NMR quantum computing, where the initial entropy-reducing preparation is quite weak: the probability of the ground state of each qubit exceeds the probability of the excited state by the small factor of $e^{2\varepsilon} \approx 1 + 10^{-5}$. In the subsequent (unitary) phase an ε^2 fraction of the qubits can be prepared in highly-cooled states [11] (and see [14] for experimental demonstration); for information-theoretic reasons, this fraction is the best possible, but at the current value of ε it is too small for effective implementation of a quantum computer. The second reason applies more broadly. Any quantum computing implementation must cope with noise. Fault-tolerance mechanisms have been designed that can do so [15] if the noise level is below a specified threshold (estimated to be between 10^{-4} and 10^{-2} per qubit per operation [16]) and if a continual supply of "ancillas" (qubits initialized in a known state) is available. Ancilla initialization need not be perfect but the error cannot exceed the same fault-tolerance threshold. In ion traps, for example, direct cooling can place qubits in their ground states with probability ≈ 0.95, a level that necessitates further cooling to exceed the threshold [17,18]. Since fresh ancillas are needed in each time step, either a large supply must be chilled in advance and maintained without substantial decoherence, or—more likely—an open-system approach must be adopted in which registers are cooled on a regular basis.

It is necessary, therefore, to study effective means for open-system algorithmic cooling. A suggested framework (the "heat-bath" approach) was made in [13] (see also [19]) and is related to bias amplification methods in current liquid-state NMR experiments (e.g., in ^{13}C-labeled trichloroethylene), as well as proposals for solid-state NMR experiments in malonic acid [20]. A heat-bath device comprises two types of qubits—some that are hard to cool (but relax slowly), and others that are readily cooled (but relax rapidly). The former are computation qubits and

© 2005 The American Physical Society

the latter are "refrigerants." At chosen times, the computation and refrigerant qubits can undergo joint unitary interaction (such as spin exchange). A similar framework is contemplated for ion trap quantum computers [18]—the computation ions are not cooled directly due to the decoherence that this causes; instead they are cooled by interaction with separate refrigerant ions that have been directly laser-cooled.

Results.—In this Letter we establish the theoretical limits for cooling on heat-bath devices. We introduce a cooling mechanism achieving much higher bias amplification than given previously. We bound the number of cooling steps required in our process, a crucial matter since any cooling process must be carried out within the relaxation times of the computation qubits. Finally, we show that our method is optimal in terms of entropy extraction per cooling step. In the course of doing so we discover a threshold phenomenon: significant initialization cannot be achieved at all unless ε, the bias that can be imparted to the rapidly relaxing qubits, is asymptotically above 2^{-n}. The proof uses majorization inequalities to convert the problem to analysis of a certain combinatorial "chip game."

For specificity we assume that the quantum computer has $n - 1$ computation qubits and an nth refrigerant qubit that is in contact with the heat bath. The cooling step, ι, changes the traced density matrix on the nth qubit to

$$\rho_\varepsilon = \frac{1}{e^\varepsilon + e^{-\varepsilon}} \begin{pmatrix} e^\varepsilon & 0 \\ 0 & e^{-\varepsilon} \end{pmatrix} \tag{1}$$

(no matter what the previous state was). In between cooling steps, reversible (unitary) quantum logic gates can be applied to the register of n qubits. Let $\mathcal{I}_n$ be the density matrix of the maximally mixed state over the 2^n-dimensional Hilbert space. The question is: starting from $\mathcal{I}_n$, and using these operations, how different from $\mathcal{I}_n$ can we make the density matrix of the device?

Theorem 1 (Physical limit).—No heat-bath method can increase the probability (i.e., $|\text{amplitude}|^2$) of any basis state from its initial value, 2^{-n}, to any more than $\min\{2^{-n}e^{\varepsilon 2^{n-1}}, 1\}$. This conclusion holds even under the idealization that an unbounded number of cooling and logic steps can be applied without error or decoherence.

This shows that if $\varepsilon \ll 2^{-n}$ then the variation distance between the uniform distribution, and any distribution reachable by cooling, is $\ll 1$.

We establish a converse to this statement using a specific cooling procedure, the partner pairing algorithm (PPA). For convenience let $\tilde{\varepsilon} = \tanh\varepsilon$. (For small ε, $\tilde{\varepsilon} \approx \varepsilon$.)

Theorem 2 (Cold qubit extraction).—Within $4n\tilde{\varepsilon}^{-2}[1 + \log(1/\tilde{\varepsilon})]$ cooling steps, the PPA creates a probability distribution in which with probability $1 - O(\frac{1}{1+\log 1/\tilde{\varepsilon}})$, all of the first $n - [1 + o(1)]\log_2 1/\tilde{\varepsilon}$ bits are $|0\rangle$'s (where $o(1)$ denotes a term tending to 0 as $\tilde{\varepsilon}$ tends to 0).

This extraction procedure is useful for quantum computing (it extracts qubits of bias almost 1, i.e., that are almost certainly in their ground state) as long as $\varepsilon \gg 2^{-n}$. (For comparison, the previous heat-bath procedure [19] amplifies bias of a qubit by only $(3/2)^n$. At comparable levels of amplification it also requires more cooling steps.)

The notion that the computation qubits are entirely insulated from the environment is, of course, merely a simplification good for moderate time spans. To be useful, algorithms must converge within the relaxation time of the computation qubits. Next we show that the PPA is near-optimal in terms of the number of cooling steps:

Theorem 3 (Cooling steps required).—Any algorithm that creates a bit of constant bias requires a number of cooling steps proportional to ε^{-2}.

Other applications of algorithmic cooling.—A central point of this Letter is the firm limit that Theorem 1 sets on the cooling parameter ε in order that the heat-bath method be useful for quantum computation. However, it is important to note that heat-bath cooling algorithms (the PPA or others) may be viable for other applications even at smaller ε. Specifically, algorithmic cooling is likely to find significant application in the scientific and medical imaging applications for which NMR technology is already in wide use. The signal-to-noise ratio in NMR imaging is proportional to the polarization of the nuclear spins and to the square root of the duration of the scan; since the duration is often limited in medicine by the need to immobilize the patient, improved sensitivity demands increased polarization. In other applications the benefit of increased polarization is in decreased scan times. Algorithmic cooling of a few nuclear spins may therefore be highly beneficial even in the range $\varepsilon \ll 2^{-n}$ that is not adequate for quantum computation. For example, perfect implementation of the PPA on a 5-qubit molecule (4 computation qubits and one refrigerant) would yield a qubit of bias 16ε, implying a 256-fold decrease in scan duration compared to cooling without algorithmic amplification.

Method of proof of Theorem 1.—The eigenvalues of a density matrix are the probabilities with which the spectral basis states are measured; the spectral basis gives measurement probabilities that are furthest from uniform in the sense of majorization [21]. A probability vector $(p_1, \ldots)$ is said to majorize another $(p'_1, \ldots)$ if there exists a doubly stochastic matrix D such that $(p_1, \ldots)D = (p'_1, \ldots)$. This is a partial (pre-)order on probability distributions in which the singular distribution $(1, 0, 0, \ldots)$ dominates all others, while the uniform distribution is dominated by all. A density matrix h is said to majorize another h' if the eigenvalues of h majorize those of h'.

Domination in majorization implies domination in any of the other measures we are interested in, such as variation distance from uniform or the sum of the largest K probabilities (for a fixed K). So our concern is: if $\bar{u}_1, \ldots, \bar{u}_r$ represent the reversible actions of an algorithm between its cooling steps (each acting on the density matrix as conjugation by a unitary operator), how different can the eigenvalues of $\iota\bar{u}_r\iota \cdots \bar{u}_1\iota\mathcal{I}_n$ be from those of $\mathcal{I}_n$ (in which all equal 2^{-n})?

A *classical* cooling algorithm is one that uses only reversible (deterministic) classical logic gates between cooling steps. In this case each operator $\overline{u}_i$ acts on the density matrix as conjugation by a permutation matrix. The first step of the analysis shows:

Proposition.—Given any quantum logic steps $\overline{u}_1, \ldots, \overline{u}_r$, there are classical steps $\overline{\pi}_1, \ldots, \overline{\pi}_r$ such that $\iota\overline{\pi}_r\iota \cdots \overline{\pi}_1\iota\mathcal{I}_n$ majorizes $\iota\overline{u}_r\iota \cdots \overline{u}_1\iota\mathcal{I}_n$.

We may therefore restrict attention to classical cooling algorithms. Observe that every intermediate density matrix created by a classical algorithm is diagonal. Hence the classical cooling steps are equivalent to the following discrete process on probability distributions on the set $\{0, 1\}^n$: begin with the uniform distribution on $\{0, 1\}^n$. The only tool for modifying the probability distribution is "discrete cooling steps," which have the effect of transforming the current distribution (denoted p) to a new distribution (denoted p'), related to p by:

$$\left.\begin{array}{l} p'_{w0} = (p_{w0} + p_{w1})\frac{e^{\varepsilon}}{e^{\varepsilon}+e^{-\varepsilon}} \\ p'_{w1} = (p_{w0} + p_{w1})\frac{e^{-\varepsilon}}{e^{\varepsilon}+e^{-\varepsilon}} \end{array}\right\} \begin{array}{c}\text{for each binary string} \\ w \text{ of length } n-1. \end{array} \tag{2}$$

There is no way of *directly* cooling the first $n - 1$ bits, but in between cooling steps we can perform arbitrary permutations of the binary strings. Because of the proposition, Theorem 1 is equivalent to showing that the above discrete process cannot increase any probability from its initial value, 2^{-n}, to any more than $2^{-n}e^{\varepsilon 2^{n-1}}$. In the discrete process, the role of a permutation of the basis is to pair off the current probabilities before the next cooling step.

If the basis states of the computer are relabeled so that their probabilities are $p_0 \geq \ldots \geq p_{2^n-1}$ (ties broken in arbitrary but fixed fashion), then for each even i we will refer to the states i and $i + 1$ as each other's "partners." The PPA is simply: in each cooling step, pair partners together.

The second step in demonstrating Theorem 1 is establishing a relation between the output of an arbitrary classical cooling algorithm B and that of the PPA. *Lemma:* given any initial probability distribution $p = \{p_0, \ldots, p_{2^n-1}\}$, and any cooling algorithm B, the distribution which results from applying the PPA for t cooling steps majorizes the distribution which results from applying B for t cooling steps.

As a consequence, in pursuit of Theorem 1's upper bound on the achievable probability of any one string, we can focus solely on the PPA. The remainder of the proof is a detailed analysis of the PPA under the dynamics of Eq. (2). These dynamics are difficult to analyze directly, but can be linearized in the following chip game: 2^n chips are placed initially at the origin of the real line. In each step you choose a pairing of the chips, and then the positions of each pair of chips (say x and y) are moved to $(x + y)/2 \pm \varepsilon$. Your goal is to move any one chip as far to the right as possible.

In this linearization, a probability p is mapped to a chip at $\log(2^n p)$, and the above dynamics replace the true physical dynamics which carry the chips to the pair (x', y') satisfying $y' - x' = 2\varepsilon$ and $e^{x'} + e^{y'} = e^x + e^y$.

The theorem rests on showing: (a) the maximum probability $p_{\max}$ achieved by the PPA and the maximum chip position $x_{\max}$ achieved by the linearized PPA are related by $\log(2^n p_{\max}) \leq x_{\max}$. (b) The linearized PPA cannot carry any chip beyond $\varepsilon 2^{n-1}$.

Part (a) follows from: *Lemma*: suppose $x = (x_0 \geq \ldots \geq x_{2^n-1})$ and $y = (y_0 \geq \ldots \geq y_{2^n-1})$ are two possible sets of chips, such that $x_i \leq y_i$ for all i. Apply the PPA to x, resulting in x'; apply the linearized PPA to y, resulting in y'. Then $x'_i \leq y'_i$ for all i. Part (b) follows by showing that certain combinatorial structures of a set of chips are preserved by the linearized PPA. Let $c(S)$ denote the mean of a subset S of the chips. Such a subset is called an *assembly* if either: (1) it is a pair of partners or (2) it is the union (or "merger") of two assemblies S_1 and S_2 such that the closed intervals $[c(S_1) - |S_1|\varepsilon/2, c(S_1) + |S_1|\varepsilon/2]$ and $[c(S_2) - |S_2|\varepsilon/2, c(S_2) + |S_2|\varepsilon/2]$ intersect. A maximal assembly is one which cannot be merged with any other assembly. *Lemma:* maximal assemblies are preserved by the linearized PPA. (This is the most technically complex part of the argument.)

Method of proof of Theorem 2.—The runtime analysis relies on tracking the distribution entropy. For $0 \leq \delta \leq 1$ let $H(\delta) = \frac{1-\delta}{2}\log\frac{2}{1-\delta} + \frac{1+\delta}{2}\log\frac{2}{1+\delta}$. Let $(1 \pm \delta)p/2$ be two probabilities paired in a cooling step of the PPA. The change in their contribution to the distribution entropy due to the cooling step is $[H(\tilde{\varepsilon}) - H(\delta)]p$. We show that in the PPA, any pair of partners satisfy $\delta \leq \tilde{\varepsilon}$, so this contribution is nonpositive, and hence the distribution entropy is nonincreasing in each cooling step. For pairs separated by $\delta \leq \tilde{\varepsilon}/2$, the decrease in entropy is strictly positive, and on this basis we show that within $\frac{n\log 2}{[H(\tilde{\varepsilon}/2)-H(\tilde{\varepsilon})]\gamma}$ cooling steps, at least $1 - \gamma$ of the probability resides in partners $\{p, p'\}$ for which $|\log p - \log p'| \geq \tilde{\varepsilon}$. We also show that once this condition is satisfied, for any positive even y, at least $(1 - \gamma)(1 - e^{-(y+2)\tilde{\varepsilon}/2})$ of the probability resides in just y of the states. Finally, Theorem 2 follows by setting $\gamma = \frac{\log 2}{1+\log 1/\tilde{\varepsilon}}$ and $y = \frac{2\log 1/\gamma}{\tilde{\varepsilon}}$. The total probability of these y most likely states is $1 - O(\frac{1}{1+\log 1/\tilde{\varepsilon}})$, and once indexed lexicographically in decreasing likelihood from 0 to $2^n - 1$, they all share $|0\rangle$'s in their first $n - \log_2 y \geq n - [1 + o(1)]\log_2 1/\tilde{\varepsilon}$ bits.

Method of proof of Theorem 3.—The entropy of the initial distribution is $n\log 2$; a distribution in which some bit has bias bounded away from 0 has entropy $(n - c)\log 2$ for a constant $c > 0$. The entropy can decrease by at most $\log 2 - H(\tilde{\varepsilon}) \leq \tilde{\varepsilon}^2$ in each cooling step.

Numerical estimates.—We depict a specific way of using the PPA. Consider an ion trap quantum computer in which four qubits are reserved for preparation of ancillas, all others being devoted to the main quantum algorithm (including the fault-tolerance mechanism). Of the reserved qubits, three are "computation qubits" and one is the "refrigerant." Ion trap technology is capable of placing the refrigerant in its ground state with probability 0.95 (i.e., $\bar{\varepsilon} = \text{arctanh}\, 0.9 \approx 1.47$). Calculation shows that application of the PPA on the quadruple for just nine cooling steps suffices to prepare one of the qubits in the ground state with probability $1 - 10^{-4}$. This is at the conservative end of the estimates for the fault-tolerance threshold for quantum computation. Hence after every nine cooling steps the PPA can prepare an ancilla, ready to be moved by spin exchange into the main bank of qubits (in place of a "warm" qubit generated by the fault-tolerance mechanism).

Implementation objectives.—It is necessary to study the sensitivity of the model to imperfections in the cooling steps, as well as in the logic gates between cooling steps, in specific experimental implementations.

Experimental algorithmic cooling also has the opportunity to produce a physically meaningful result well before producing a quantum computer. A series of papers [22–24] show that if k qubits have bias less than 2^{-2k} then their joint state is separable. Conversely, in the ball of radius $2^{-k/2}$ about the maximally mixed state there exist nonseparable states. Liquid-state NMR experiments have not, to date, produced a demonstrably nonseparable state. Achieving this goal will require some combination of an increase in the number of coherently-manipulated qubits and an increase in the individual polarization of these qubits. The latter demands implementation of new cooling techniques.

In the simple model adopted in this Letter we have assumed that there is only a single refrigerant qubit. One may ask how the model is affected if the number of such qubits is proportional to the number of computation qubits. (In liquid-state NMR, for example, we can expect that nuclei of various types will be present in fixed proportions.) The answer is that while some gain is likely, the fundamental limits of the model are unchanged because with a slowdown in the cooling process by a factor of $O(n)$, the same effect can be achieved by spin exchange with a single refrigerant qubit.

The necessity of cooling many qubits for quantum computation.—In view of the difficulty of cooling certain kinds of quantum computers, the question was posed of whether this was truly necessary [25]. Quantum-over-classical computational speedups may indeed be possible on devices that are initialized in a highly (though not completely) mixed state; see [25,26]. However, general-purpose quantum computers cannot be directly simulated on such devices [7], so the need for effective cooling is unlikely to be circumvented. The necessity of using ancillas to compensate for noise buttresses this conclusion.

Summary.—We have studied the fundamental limits of open-system "heat-bath" cooling, with a view to the significance of such methods for quantum computation as well as for imaging tasks limited by imperfect state preparation. We have provided a cooling (bias amplification) method and have shown that: (a) the bias it achieves is substantially higher than in previous methods, and the ground-state probability after any number of cooling steps is highest possible. (b) The number of cooling steps it requires is asymptotically close to best possible. (c) There is a sharp threshold for the heat-bath temperature, above which substantial cooling is impossible in any method, and below which it is achieved by ours.

Thanks to R. Laflamme and J. Fernandez for helpful discussions. L. S. was supported in part by the NSF, the Mathematical Sciences Research Institute, and the Okawa Foundation. T. M. and Y. W. were supported in part by the Israel Ministry of Defense.

[1] J. I. Cirac and P. Zoller, Phys. Rev. Lett. **74**, 4091 (1995).
[2] C. Monroe *et al.*, Phys. Rev. Lett. **75**, 4714 (1995).
[3] D. G. Cory, A. F. Fahmy, and T. F. Havel, Proc. Natl. Acad. Sci. U.S.A. **94**, 1634 (1997).
[4] N. A. Gershenfeld and I. L. Chuang, Science **275**, 350 (1997).
[5] D. P. DiVincenzo, in *Mesoscopic Electron Transport* (Kluwer, Dordrecht, 1997).
[6] D. P. DiVincenzo, Fortschr. Phys. **48**, 771 (2000).
[7] A. Ambainis, L. J. Schulman, and U. Vazirani, in *Proceedings of the 32nd STOC*, edited by F. Yao (ACM, New York, 2000), pp. 705–714.
[8] C. M. Bowden, J. P. Dowling, and S. P. Hotaling, in *SPIE Proceedings 3076: Photonic Quantum Computing*, edited by S. P. Hotaling and A. R. Pirich (SPIE, Bellingham, Washington, 1997), pp. 173–182.
[9] M. Iinuma *et al.*, Phys. Rev. Lett. **84**, 171 (2000).
[10] A. S. Verhulst *et al.*, Appl. Phys. Lett. **79**, 2480 (2001).
[11] L. J. Schulman and U. Vazirani, in *Proceedings of the 31st STOC*, edited by T. Leighton (ACM, New York, 1999), pp. 322–329.
[12] O. W. Sørensen, Progress in NMR Spectroscopy **21**, 503 (1989).
[13] P. O. Boykin *et al.*, Proc. Natl. Acad. Sci. U.S.A. **99**, 3388 (2002).
[14] D. E. Chang, L. M. K. Vandersypen, and M. Steffen, Chem. Phys. Lett. **338**, 337 (2001).
[15] D. Aharonov and M. Ben-Or, in *Proceedings of the 29th STOC*, edited by P. Shor (ACM, New York, 1997), pp. 176–188.
[16] E. Knill, quant-ph/0404104.
[17] B. E. King *et al.*, Phys. Rev. Lett. **81**, 1525 (1998).
[18] M. D. Barrett *et al.*, Phys. Rev. A **68**, 042302 (2003).
[19] J. M. Fernandez *et al.*, Int. J. Quantum. Inform. **2**, 461 (2004).
[20] R. Laflamme (personal communication).
[21] A. W. Marshall and I. Olkin, *Inequalities* (Academic, New York, 1979).
[22] K. Zyczkowski *et al.*, Phys. Rev. A **58**, 883 (1998).
[23] G. Vidal and R. Tarrach, Phys. Rev. A **59**, 141 (1999).
[24] S. L. Braunstein *et al.*, Phys. Rev. Lett. **83**, 1054 (1999).
[25] E. Knill and R. Laflamme, Phys. Rev. Lett. **81**, 5672 (1998).
[26] D. Poulin *et al.*, Phys. Rev. Lett. **92**, 177906 (2004).