Effective Fault-Tolerant Quantum Computation with Slow Measurements

David P. DiVincenzo[1] and Panos Aliferis[2]

[1]IBM Research Division, T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598, USA
[2]Institute for Quantum Information, California Institute of Technology, Pasadena, California 91125, USA

(Received 3 August 2006; published 9 January 2007)

How important is fast measurement for fault-tolerant quantum computation? Using a combination of existing and new ideas, we argue that measurement times as long as even 1000 gate times or more have a very minimal effect on the quantum accuracy threshold. This shows that slow measurement, which appears to be unavoidable in many implementations of quantum computing, poses no essential obstacle to scalability.

DOI: 10.1103/PhysRevLett.98.020501 PACS numbers: 03.67.Pp, 03.67.Lx, 06.20.Dk

Considerable progress has been made towards the physical realization of a working quantum computer in recent years. However, in no existing technology is there an easy pathway to scalability, and the main reason for this is the stringency of the requirements for fault tolerance in quantum computation. Nearly ten years ago, it was established that a quantum circuit whose components are all noisy (including storage, gate operations, state preparation, and quantum measurement) can efficiently simulate any quantum computation to any desired accuracy provided the noise level is lower than an accuracy threshold [1–4]. The initial estimates of this threshold for stochastic noise—around 10^{-5} to 10^{-4} error probability per elementary operation—remain, alas, close to the mark today. However, recent work has investigated what special circumstances would permit the threshold value to be higher: Notably, if qubit transport and storage are assumed to be noiseless, then there is evidence that the noise threshold for depolarizing noise exceeds 10^{-2} [5].

One important parameter whose effect on the accuracy threshold has not been extensively explored concerns the time it takes to complete the measurement of a qubit. Except for Ref. [6], almost all studies have assumed that measurements are fast—that is, that they take no longer than a few gate operation times. This capability definitely increases the effectiveness of error correction, since information about errors is available promptly and any necessary recovery operation can be applied immediately to a logical qubit. Nevertheless, it is known that having this fast-measurement capability is not necessary. In fact, measurements can be avoided altogether and error correction can be implemented fully coherently. The penalty paid in the stringency of the threshold has never been quantified, but it is expected that replacing measurement by coherent operations decreases the noise threshold by a large amount.

In this Letter, we examine a scenario in which accurate quantum measurement is possible but is slow. We will imagine that measurement takes 1000 gate operation times—a reasonable estimate currently for spin qubits [7]—but the arguments developed here will not depend very strongly on the precise value of this number. We find that, by combining several existing strategies for fault-tolerant error correction with a couple of new ''tricks,'' the accuracy threshold value is barely affected by the speed of measurement—that is, the threshold is hardly worse in the slow-measurement setting as compared with the fast-measurement setting. This diminishes one of the principal obstacles to solid-state quantum computing, for which it is difficult to imagine measurement times as short as gate operation times.

Topological quantum computation [8] aside, the best-understood route to fault-tolerant quantum computation (FTQC) uses concatenated quantum codes and gate operations applied directly to the encoded data—our result is developed in this setting. We now first review some of the principal concepts of this approach adopting language due to Ref. [1]. A quantum algorithm, laid out as a quantum circuit, consists of a set of locations, which are elementary components of this circuit: state preparation, one- or two-qubit gate operations (including identity ''wait'' operations when a qubit is stored in memory), or qubit measurements. Next, a ''good'' computation quantum code is chosen. A variety of properties make a code good for computation: It should encode one logical qubit in a not-very-large block of physical qubits while correcting some large number of errors relative to its block size. A large set of logical gate operations should be doable transversally via the application of physical gates to each of the qubits in the code block. Finally, the ancilla quantum states needed for completing universality and for error correction should be relatively easy to prepare in a sufficiently noiseless state. This is typically achieved by verification [9] in which the ancilla, before being coupled to the encoded data, is subject to tests that verify its high fidelity. These tests are done by coupling the ancilla to other verifier ancillae, followed by measurements on the verifier qubits, which confirm the quality of the verified ancilla qubits or reveal the presence of errors. In the latter case, the verified ancilla is typically rejected and the procedure starts anew.

To obtain the encoded quantum circuit, each location in the original circuit executing the desired algorithm is replaced by a rectangle. Rectangles are composite objects consisting of a set of locations: First, locations needed for a fault-tolerant implementation of the ''high-level'' location

(i.e., logical state preparation, gate, or measurement), followed by those locations needed for a full error-correction cycle. If the error rate for elementary operations is below the accuracy threshold, this replacement will result in an encoded circuit whose effective noise rate is lowered with respect to the original unencoded circuit. To lower the noise still further, the replacement procedure can be repeated sufficiently many times for the locations in the encoded circuit itself. Each time, a new circuit is created which is encoded at an increasingly higher level of a concatenated quantum code.

The standard concatenation procedure described above can be varied and optimized in various physical settings. For example, Knill [5] has shown that, in a setting where memory and qubit transport are essentially noiseless, a very inefficient strategy for the generation of ancilla states based on postselection gives a threshold around 3×10^{-2} for depolarizing noise. On the other hand, in the more realistic setting for contemplated solid-state implementations, where memory has a noise level in the same range as gate operations and qubit transport must be accomplished by noisy SWAP gate operations, a different strategy relying less on ancilla post-selection seems to be the best. Such an approach has been analyzed by Aliferis, Gottesman, and Preskill (AGP) [10]—they find noise thresholds in this setting to be somewhat lower than 10^{-4} for stochastic noise. Svore, DiVincenzo, and Terhal (SDT) [11] analyze a variant of this setting with qubits constrained to lie on a fixed two-dimensional square geometry. By modifying and adapting the verification circuits of AGP to this lattice geometry, the penalty on the threshold found by SDT in this setting is only about a factor of 2 compared with the completely unrestricted geometry of AGP.

In all of this work, measurement times and gate operation times have been assumed to be of the same order. In fact, it would seem that the value of the accuracy threshold depends crucially on this assumption: Most importantly, measurement is used in ancilla verification during error correction, and the longer measurement takes, the longer the ancilla qubits need to wait in memory while verification is completed. The problem is illustrated by Fig. 1, which shows a fragment of a circuit that extracts information about errors in the data block according to the scheme introduced by Shor [9,12]. Roughly speaking, if measurement takes 1000 gate operation times, the memory noise level would need to be 1000 times below the gate noise level for the fidelity of the waiting ancilla to remain high enough and the accuracy threshold for gate noise to stay unchanged when slow measurement is taken in consideration. Steane [6] has documented such a decrease of the noise threshold with increasing measurement time, although the effect on the threshold is not as severe as our simple argument implies. There are some physical systems in which the noise for qubit storage (and movement) may indeed be very low, so that measurement-based verification can be used very effectively to obtain high accuracy thresholds [5]. But in other settings (e.g., in solid-state schemes) it is expected that noise levels for gate operations, memory, and moving will be comparable; it would seem that the threshold for FTQC would then be severely compromised by long measurement times.

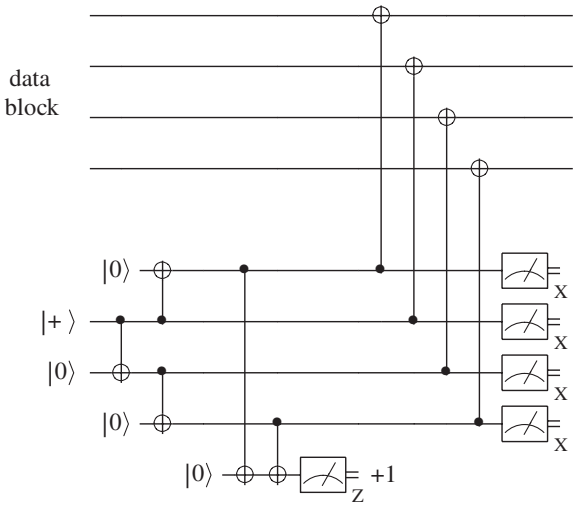

FIG. 1. Fragment of an error-correction circuit in which a "cat state" ancilla [9] is prepared, verified, coupled to the data, and then measured. The first three controlled-NOT (CNOT) gates prepare the four-qubit ancilla state $|0000\rangle + |1111\rangle$. The next two CNOTs and the measurement of $Z \equiv \sigma_z$ comprise the verification of this ancilla. In this protocol, this measurement outcome must be known before the verified ancilla is coupled to the data block: If the measurement outcome is -1, the cat state is to be discarded and ancilla preparation is to be attempted again.

However, in this Letter we show the opposite: Even in these settings, and unlike the conclusion drawn in Ref. [6], the threshold is hardly affected by long measurement times. This is so because, as we discuss below (point 2), one can replace the nondeterministic verification protocol in Fig. 1 with a deterministic protocol that corrects errors in the verified ancilla, and, importantly, this replacement results in an accuracy threshold comparable to that obtained with nondeterministic verification. But the full story involves a combination of existing and new ideas, which we now explain.

1. Use of Pauli frames.—We did not comment above on the use of the measurements of $X \equiv \sigma_x$ in Fig. 1. These measurement bits are combined to yield the code syndrome which indicates errors in the data block and the necessary recovery operation to invert them. For all codes used in FTQC, these recovery operations are tensor products of single-qubit operations in the usual Pauli group. It has been known for some time (e.g., see [5,10]) that it is not necessary to directly apply these recovery operations on the data. Instead, it is sufficient to merely record and keep track of them in a classical memory as a reference frame defined by a Pauli rotation. This is so because the Pauli group is closed under the action of the Clifford group: Pauli operators commute through gates belonging to the Clifford group to give other Pauli operators. Since gates in a fault-tolerant

circuit that determine the accuracy threshold—most importantly, all gates needed for implementing error correction—belong to the Clifford group, the application of the recovery operations specified by the syndrome can usually be delayed a long time.

2. Ancilla decoding instead of verification.—This is a new idea and requires a modification of all existing ancilla verification circuits. But the modification is always simple—Fig. 2 shows the necessary change to the circuit in Fig. 1. The reason that ancilla preverification before interaction with the data has previously been considered necessary is that a single fault, at certain locations in the ancilla preparation circuit, can lead to a multiqubit error in the ancilla state. It has, therefore, always been thought necessary to prevent such ancillae from interacting with the data. But, if the nature of these multiqubit errors can always be determined by postprocessing of the ancilla after its interaction with the data, then a suitable recovery operation can always be devised. The decoding and measurement of the ancilla in Fig. 2 serve to determine such a recovery operation for the data, and this operation is again always a tensor product of single-qubit Pauli operations. Therefore, as in our discussion above, correction of multiqubit errors in the ancilla can always be delayed by incorporating the recovery operation into the Pauli frame. Reference [13] gives further details of this method.

The remaining ideas are needed only to deal with these non-Clifford operations which, together with Clifford-group operations, complete quantum universality. Non-Clifford operations require a different treatment since a Pauli frame cannot be simply propagated through them: Commuting a Pauli operator through such a gate can generally give an operator outside the Pauli group. For this reason, all information determining the current Pauli frame must be known before the application of a non-Clifford gate, so that the restoration operation can be applied immediately before the non-Clifford operation is implemented.

We will now show that, despite this restriction, non-Clifford gates can be executed effectively even when all measurements are slow. First, we recall that logical non-Clifford gates are fault-tolerantly simulated using appropriate ancilla states. Non-Clifford gate operations appear in the subcircuits preparing these ancillae, while the use of the ancillae after preparation and verification involves only Clifford-group operations [14]. This does not immediately lead to a solution to the measurement-time problem, as, e.g., Fig. 3 illustrates. This figure shows how to simulate the $T \equiv \exp(-i\frac{\pi}{8}\sigma_z)$ gate, with the Clifford-group gate $S = T^2$ conditioned on the measurement outcome. Alternatively, the logical Toffoli gate could be simulated, with CNOT gates being conditioned on the measurement outcomes inside the simulation circuit [4,9].

The simulation circuit of Fig. 3 is to be used in an encoded form, and the ancilla block will be prepared in the logical $|a_{\pi/8}\rangle$ state. As the next step, this circuit will also be concatenated in order to decrease the effective noise for the logical T gate to the desired level. When the simulation of the logical T gate occurs at level ℓ, the circuit in Fig. 3 uses a level-ℓ $|a_{\pi/8}\rangle$ ancilla. There is a fault-tolerant rectangle (see [10,11] for the circuit) which prepares the level-ℓ ancilla using level-$(\ell-1)$ T gates. In this standard approach, these alternating replacements are iterated until Fig. 3 is used at level 1, where it contains zeroth-level physical T gates.

However, this circuit is clearly unusable at level 1 if measurements are slow: The data qubits will have to wait in memory too long, since the outcome of the measurement of level-1 logical Z (including all of the preceding Pauli-frame information that determines its meaning) must be known in order to decide if the level-1 logical S gate is to be performed (this decision must be made before this qubit is involved in the next logical CNOT in the circuit, which is usually immediately). Is there a fix to this problem?

Here is the essential idea: In order to get a very low effective error rate for the logical T gate, it is only necessary that the circuit of Fig. 3 appears at sufficiently many high levels of concatenation. But at high levels of concatenation there is no problem with slow measurement. This is easy to see for concatenated error correction: The gate time $t_{\text{gate}}^{(k)}$ at level k of concatenation scales exponentially with k, $t_{\text{gate}}^{(k)} = aC^k$ for some constants a and C. On the other hand, the measurement time t_{meas} is the same at every level of concatenation, since logical measurement is performed by transversal measurements on the physical level. So, even if $t_{\text{meas}} = 1000a$ and for, e.g., $C = 34$ (as in Ref. [11]),

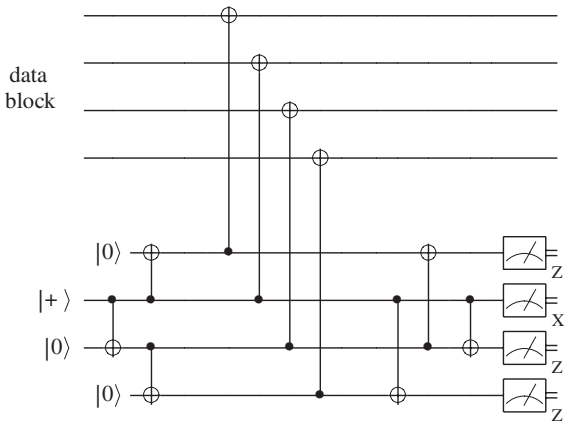

FIG. 2. The modified circuit from Fig. 1: Ancilla verification is removed and is replaced by a decoding and measurement of the ancilla.

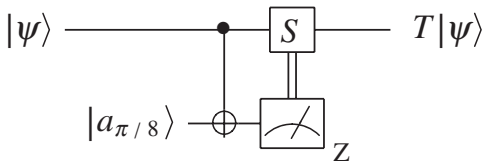

FIG. 3. Simulation of the gate T using the ancilla $|a_{\pi/8}\rangle \equiv T\sigma_x|0\rangle$, Clifford-group operations, and measurement. The gate S is performed only if the measurement outcome is -1.

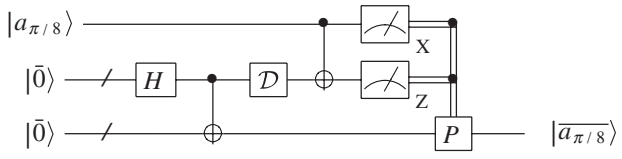

FIG. 4. Creation of the logical ancilla $|\overline{a_{\pi/8}}\rangle$ by teleportation.

measurement is completed in one logical gate time for all $k \geq k_{\min} = 2 \gtrsim \log_C k$. The more general idea is that at some level of coding, because the effective error rate for logical Clifford-group operations decreases quickly with the coding level, the probability for a logical error in memory in the data block in Fig. 3 can be made sufficiently small for the total time it takes to measure the ancilla-block qubits. This can provide a more general criterion for determining $k_{\min}$ in cases where a strategy other than strict concatenation is used or when C is quite small ($C = 5$ in Ref. [10]).

To avoid dealing with the T gate at levels lower than $k_{\min}$, we need an alternative to the iterative replacement described above. One has already been suggested in the literature (see, e.g., [5]); it is referred to as the following.

3. Injection by teleportation.—This is the last ingredient that we need. Figure 4 illustrates the idea: The two logical $|0\rangle$ ($|\bar{0}\rangle$) blocks together with the logical Hadamard and CNOT gates create a logical Bell pair for teleportation. This logical level corresponds to $k_{\min}$. Then one of the code blocks of the Bell pair is decoded to the physical (unencoded) level. Next, a Bell measurement is done at the physical level between a qubit prepared in the $|a_{\pi/8}\rangle$ state and the decoded half of the Bell pair. As a result, up to a Pauli-frame change P, the output block is in the logical $|a_{\pi/8}\rangle$ ($|\overline{a_{\pi/8}}\rangle$) state as desired. The noise level on the $|\overline{a_{\pi/8}}\rangle$ state, which has thus been injected into the level-$k_{\min}$ code block, will not be much greater than that of the original $|a_{\pi/8}\rangle$ state and the physical noise level for Clifford operations (see, e.g., [11]).

The time required for implementing this circuit is independent of t_{meas}, and, since injection occurs at level $k_{\min}$, all measurement outcomes will be available in one logical gate time. With these observations, the threshold analyses of AGP and SDT go through essentially unchanged, so that the accuracy threshold is not effected in the slow-measurement setting. We say "essentially" because there are two changes that slow measurements make for the threshold analysis, neither of which should cause a major change in the threshold value: (i) The circuits will be changed in detail in order to avoid ancilla verification. (ii) In the "local" setting of SDT, where qubits must be moved using SWAP gates on the two-dimensional lattice, extra space must be left for qubits to be measured (since they must remain in place for, say, 1000 time steps before they can be reused). This requires an expansion of the physical patch of the lattice occupied by one logical qubit at the lowest level of concatenation. We estimate that this increases the linear scale of the computer by a small factor of around 2, leading to perhaps a factor of 2 decrease of the threshold. The combined effect of (i) and (ii) leads us to expect that the accuracy threshold value will be only very minimally affected in the slow-measurement setting.

To conclude, we have shown that fault-tolerant quantum computation can be implemented in such a way that, except in a minor way, slow quantum measurements have no effect on the noise threshold at which error correction becomes effective. Our result does not apply to every existing scheme for FTQC; for example, it is not applicable to the (nondeterministic) quantum computation scheme that is possible in a linear-optics setting [15]. Also, it cannot straightforwardly be used in postselected computation [5]; but we are currently studying whether modifications of the approach of Ref. [5] might permit noise thresholds in the vicinity of 10^{-3} to be achievable in the slow-measurement setting.

We thank Daniel Gottesman and Barbara Terhal for discussions. P. A. is supported by the NSF under Grant No. PHY-0456720. D. D. V. is supported by the DTO through ARO Contract No. W911NF-04-C-0098.

[1] D. Aharonov and M. Ben-Or, in *Proceedings of the 29th Annual ACM Symposium on the Theory of Computation* (Association for Computing Machinery, New York, 1998), p. 176.
[2] E. Knill, R. Laflamme, and W. Zurek, Proc. R. Soc. A **454**, 365 (1998).
[3] A. Kitaev, Russ. Math. Surv. **52**, 1191 (1997).
[4] J. Preskill, in *Introduction to Quantum Computation*, edited by H.-K. Lo, S. Popescu, and T. P. Spiller (World Scientific, Singapore, 1998), pp. 213–269.
[5] E. Knill, Nature (London) **434**, 39 (2005).
[6] A. M. Steane, Phys. Rev. A **68**, 042322 (2003).
[7] L. Vandersypen (private communication).
[8] E. Dennis *et al.*, J. Math. Phys. (N.Y.) **43**, 4452 (2002); M. Freedman, M. Larsen, and Z. Wang, Commun. Math. Phys. **227**, 605 (2002); A. Kitaev, Ann. Phys. (N.Y.) **303**, 2 (2003).
[9] P. W. Shor, in *Proceedings of the 37th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, CA, 1996), p. 56.
[10] P. Aliferis, D. Gottesman, and J. Preskill, Quantum Inf. Comput. **6**, 97 (2006).
[11] K. M. Svore, D. P. DiVincenzo, and B. M. Terhal, quant-ph/0604090.
[12] D. P. DiVincenzo and P. W. Shor, Phys. Rev. Lett. **77**, 3260 (1996).
[13] See EPAPS Document No. E-PRLTAO-98-063701 for further details of the method. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[14] D. Gottesman and I. L. Chuang, Nature (London) **402**, 390 (1999).
[15] E. Knill, R. Laflamme, and G. J. Milburn, Nature (London) **409**, 46 (2001).