Quantum Random Access Memory

Vittorio Giovannetti,[1] Seth Lloyd,[2] and Lorenzo Maccone[3]

[1]*NEST-CNR-INFM & Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126, Pisa, Italy*
[2]*MIT, RLE and Department of Mechanical Engineering MIT 3-160, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA*
[3]*QUIT, Dipartimento di Fisica "A. Volta," Università di Pavia, via Bassi 6, I-27100 Pavia, Italy*

(Received 10 September 2007; revised manuscript received 12 November 2007; published 21 April 2008)

A random access memory (RAM) uses n bits to randomly address $N = 2^n$ distinct memory cells. A quantum random access memory (QRAM) uses n qubits to address any quantum superposition of N memory cells. We present an architecture that exponentially reduces the requirements for a memory call: $O(\log N)$ switches need be thrown instead of the N used in conventional (classical or quantum) RAM designs. This yields a more robust QRAM algorithm, as it in general requires entanglement among exponentially less gates, and leads to an exponential decrease in the power needed for addressing. A quantum optical implementation is presented.

DOI: 10.1103/PhysRevLett.100.160501 PACS numbers: 03.67.Lx, 03.65.Ud, 89.20.Ff

A fundamental ability of any computing device is the capacity to store information in an array of memory cells [1]. The most flexible architecture for memory arrays is random access memory, or RAM, in which any memory cell can be addressed at will [2]. A RAM is composed of a memory array, an input register ("address register"), and an output register. Each cell of the array is associated with a unique numerical address. When the address register is initialized with the address of a memory cell, the content of the cell is returned at the output register ("decoding"). Just as RAM forms an essential component of classical computers, quantum random access memory, QRAM, will make up an essential component of quantum computers, should large quantum computers eventually be built. It has the same three basic components as the RAM, but the address and output registers are composed of qubits (quantum bits) instead of bits. [The memory array can be either quantum or classical, depending on the QRAM's usage]. The QRAM can then perform memory accesses in coherent quantum superposition [3]: if the quantum computer needs to access a superposition of memory cells, the address register a must contain a superposition of addresses $\sum_j \psi_j |j\rangle_a$, and the QRAM will return a superposition of data in a data register d, correlated with the address register:

$$\sum_j \psi_j |j\rangle_a \overset{\text{QRAM}}{\to} \sum_j \psi_j |j\rangle_a |D_j\rangle_d, \tag{1}$$

where D_j is the content of the jth memory cell. The possibility of efficiently implementing these devices would yield an exponential speedup for pattern recognition algorithms [4–6], period finding, discrete logarithm, and quantum Fourier transform algorithms over classical data. Moreover, QRAMs are required for the implementation of various algorithms, such as quantum searching on a classical database [3], collision finding [7], element distinctness in the classical [8] and quantum [9] settings, and the quantum algorithm for the evaluation of general NAND trees [10]. Finally, QRAMs permit the introduction of new quantum computation primitives, such as quantum cryptographic database searches [11] or the coherent routing of signals through a quantum network of quantum computers [12].

Both classical and quantum RAMs are computationally expensive: If the memory array is disposed in a d-dimensional lattice, conventional architectures involve throwing $O(N^{1/d})$ switches (i.e., two-body interactions) to access one out of the $N = 2^n$ memory slots, where n is the number of bits in the address register [2]. This exponential use of resources translates into a relatively slow speed and high energy usage for classical RAMs during decoding, and to a high decoherence rate for QRAMs. For this reason, up to now little attention has been devoted to developing a QRAM. In this Letter we introduce a new RAM architecture, dubbed "bucket brigade," that reduces the number of switches that must be thrown during a RAM call, quantum or classical, from $O(N^{1/d})$ to $O(\log N)$. If we neglect the travel time of the signals along the wires connecting the device's components, this translates into an exponential reduction in the running-time computational complexity at the information theoretical level, when compared to conventional setups. As will be shown, for QRAMs it entails an exponential reduction in the number of gates that need to be entangled for each memory call, simplifying the QRAM circuit with respect to the conventional architectures [3], and reducing the need for expensive error correction routines. In addition, the reduction in the number of switchings translates into a reduction of the energy employed in the routing, which may yield more efficient RAMs that use less power during decoding than current architectures.

We start by describing the conventional RAM architecture, showing why its direct translation to the quantum realm is inefficient and prone to noise. We then introduce our bucket-brigade architecture and give an account of the

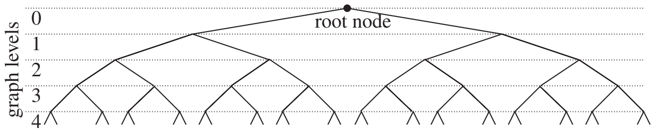

FIG. 1. Bifurcation graph of the RAM addressing.

required resources in the classical and quantum setting. We conclude by introducing an illustrative example.

Quantum RAM.—Even though more elaborate architectures exist [2] (such as ones using d-dimensional memory arrays), the basic RAM addressing scheme is simple: Suppose that the N memory cells are placed at the end of a bifurcation graph, composed by the n levels shown in Fig. 1. The value of the jth bit in the address register can be interpreted as the route to follow when one has reached a node in the jth level of the graph: if the value is 0, the left path must be followed; if it is 1, the right path must be followed (e.g., an address register 010. is interpreted as "left at the 0th level, right at the first level, left at the second," etc.). Each of the N possible values of the address register thus indicates a unique route that crosses the whole graph and reaches one of the memory cells [13]. An electronic implementation requires placing one transistor in each of the two paths following each node in the graph. Each address bit controls all the transistors in one of the graph levels: it activates all the transistors in the left paths if it has value 0, or all the transistors in the right paths if it has value 1 [2]. Thus, an exponential number of transistors must be activated at each memory call to route the signals through the graph (this entails an energy cost exponentially larger than the cost of a single transistor activation).

Direct translations of the above scheme into the quantum realm [3] are quite impractical. The n qubits of the address register coherently control n quantum control lines, each of which acts coherently on an entire level of the bifurcation graph. At each branch of the bifurcation graph, a 0 in the address register for that level shunts signals along the left paths, and a 1 shunts signals along the right paths. Each binary address is correlated with a set of switches that pick out the unique path through the graph associated with that address. A coherent superposition of addresses is coherently correlated, i.e., entangled, with a set of switches that pick out a superposition of paths through the graph. To complete the quantum memory call, a quantum "bus" is injected at the root node and follows the superposition of paths through the graph. Then the internal state of the bus is changed according to the quantum information in the memory slot at the end of the paths (e.g., through a controlled-NOT transformation that correlates the bus and the memory) [14]. Finally, in order to decorrelate the bus position from the address register, the bus returns to the root node by the same path. Like a quantum particle, the bus must be capable of traveling down a coherent superposition of paths. Although not impossible, such a QRAM scheme is highly demanding in practice for any reasonably sized memory. In fact, to query a superposition of memory cells, the address qubits are in general entangled with $O(N)$ switches or quantum gates (or, equivalently, they must control two-body interactions over exponentially large regions of space), i.e., a state of the form $\sum_j \psi_j |j_0 j_1 \cdots j_{n-1}\rangle_a \otimes |j_0\rangle_{s_0} |j_1\rangle_{s_1}^{\otimes 2} \cdots |j_{n-1}\rangle_{s_{n-1}}^{\otimes 2^{n-1}}$, where j_k is the kth bit of the address register, and s_k is the state of the 2^k switches controlled by it. Such a gigantic superposition is highly susceptible to decoherence and requires costly quantum error correction whenever the error rate is bigger than 2^{-n}. In fact, if a single gate out of the $N = 2^n$ gates in the array is decohered, then the fidelity of the state in average is reduced by a factor of 2, and if at least one gate in each of the k lines is decohered, the fidelity in average is reduced by 2^{-k}. (fidelity 保真度)

The bucket brigade is based on sending both the address register and the signal through the bifurcation graph. Like buckets of water passed along a line of improvised fire fighters, they carve a route that crosses the whole graph along which the information can be extracted. With respect to the conventional architecture detailed above, the $O(N)$ active logic gates are replaced by memory elements, most of which are in a passive *wait* state during each memory call. As a result, there is an exponential reduction of active gates and of two-body interactions, from $O(N)$ to $O(\log^2 N)$. This means the bucket-brigade RAM could also be useful in classical computation to reduce the energy needed for the addressing. (Hybrid schemes that combine the two above architectures might be more generally useful.)

The basic idea follows. At each node of the graph of Fig. 1 there is a trit, a three-level memory element. The trit's three levels are labeled *wait*, *left*, and *right*. A trit in the level *wait* will change its value according to the value of any incoming bit: if the incoming bit is 0, it takes the value *left*, while if the incoming bit is 1, it takes the value *right*. A trit in the level *left* or *right* will deviate any incoming signal along the graph according to its value. The protocol starts by initializing all the trits in the state *wait*. Then the first bit of the address register is sent through the graph. It will induce a change in the root node, which will be transferred to *left* or *right* depending on the bit's value. Now the second bit of the address register is sent through the graph. Depending on the value of the first node, it will be deviated left or right and will meet one of the two nodes on the second level of the graph (both of which are in a *wait* state). This node will be transformed according to the bit's value, and so on. After all the $\log N$ bits of the address register have passed through the graph, a single route of $n = \log N$ *left-right* trit states has been carved through the graph (see Fig. 2). All other trits remain in the *wait* state. Now a *bus* signal can easily follow this route (by heeding the indications of the trits it encounters) and find its way to the element in the memory

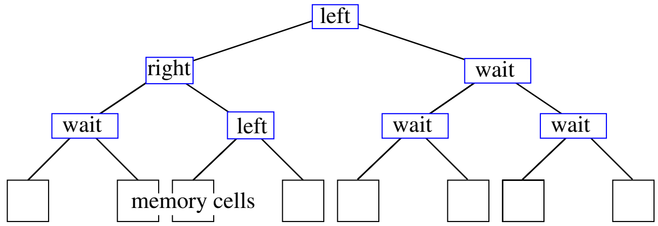

FIG. 2 (color online). Bifurcation graph of the bucket-brigade architecture. Here the third memory cell is addressed (address register 010).

array that the address register was pointing to. Information is then extracted through this route by sending back the bus signal, which must again heed the directions of the trits it encounters while traveling to the graph's root node. In addition, every time the bus signal on its way back encounters a trit, the trit is reset to the *wait* state. Thus, the memory element is addressed by the bus signal, which is then sent back to the root node, and the graph is reset to its initial *wait* state. Only $\log N$ trits have been involved in the memory call.

In the quantum realm the trits must be replaced by qutrits, i.e., three-level quantum systems, described by the vectors $|\text{wait}\rangle$, $|\text{left}\rangle$, and $|\text{right}\rangle$. Now, when the qubits of the address register are sent through the graph, at each node they encounter a unitary encoding transformation U. If the qutrit is initially in the $|\text{wait}\rangle$ state, the unitary swaps the state of the qubit in the two $|\text{left}\rangle$-$|\text{right}\rangle$ levels of the qutrit (i.e., $U|0\rangle|\text{wait}\rangle = |f\rangle|\text{left}\rangle$ and $U|1\rangle|\text{wait}\rangle = |f\rangle|\text{right}\rangle$, where $|f\rangle$ is a fiduciary state of the qubit). If the qutrit is not in the $|\text{wait}\rangle$ state, then it simply routes the incoming qubit according to its state. It is clear that an address register in a quantum superposition will carve a superposition of routes through the graph, so that any incoming qubit will exit the graph in the corresponding superposition of locations. Once all the register qubits are sent through the graph, a bus qubit is injected and it reaches the end of the graph along the requested superposition of paths. It then interacts with the memory cells at such locations changing its state according to their information content. Now the bus qubit is sent back through the graph, exiting at the graph's root node. Finally, starting from the last level of the graph, the qutrits are subject to the inverse of the unitary encoding transformation: a qutrit initially in the states $|\text{left}\rangle$ or $|\text{right}\rangle$ is evolved back to the state $|\text{wait}\rangle$, while sending a qubit (containing the state of the $|\text{left}\rangle$-$|\text{right}\rangle$ levels) back through the graph, i.e., the transformation $U^\dagger|f\rangle|\text{left}\rangle = |0\rangle|\text{wait}\rangle$ or $U^\dagger|f\rangle|\text{right}\rangle = |1\rangle|\text{wait}\rangle$. To activate this transformation at the right moment, various schemes are possible. The simplest one entails activating a classical control over all the qutrits in each level of the tree, sequentially from the last level up to the root node. Alternatively, one can send n control qubits along the superposed path, each of which controls the unitary $U^\dagger$ at one of the tree levels. A further scheme entails introducing counters in each node, which activate the $U^\dagger$ unitary after a level-dependent number of signals have transited. At the end, all qubits of the address register have been ejected from the graph, which is restored to its initial state of all qutrits in the $|\text{wait}\rangle$ state, yielding the transformation of Eq. (1).

Similarly to what happens in quantum computation with atomic ensembles [15], the noise resilience of the bucket brigade stems from the fact that in each branch of the superposition only $\log N$ qutrits are not in the passive $|\text{wait}\rangle$ state. In fact, for a query with a superposition of r memory cells, it is necessary to entangle only $O(r \log N)$ qutrits, as the state of the device is of the type $\sum_j \psi_j |j_0\rangle_{t_0} |j_1\rangle_{t_1(j_0)} \cdots |j_{n-1}\rangle_{t_{n-1}(j_{n-2})} \otimes_{\ell_j} |\text{wait}\rangle_{t_{\ell_j}}$, where t_k represents the state of the one qutrit at the kth level which is aimed to by the non-$|\text{wait}\rangle$ qutrit at the $k-1$ level, and where ℓ_j spans the other qutrits. Even if all of the qutrits are involved in the superposition, the state is still highly resilient to noise: if a fraction ϵ of the gates are decohered (with $\epsilon \log N < 1$) then in average the fidelity of the resulting state is $O(1 - \epsilon \log N)$ (compare this to the $1/2$ fidelity reduction in the conventional QRAM above). The noise resilience is, of course, greater in those algorithms where r is small, such as the quantum private queries (QPQ) [11] or the quantum routing [12]. Moreover, note that the exponentially larger number of $|\text{wait}\rangle$ states could give significant overall errors even if their individual error rates are much lower than those used in the left and right states.

Bucket-brigade implementation.—Like cluster state quantum computation [16], the bucket brigade only assumes the possibility of operating coherently on a small number $O(\log N)$ out of large number $O(N)$ of first-neighbor connected quantum memory elements, and it does not require macroscopic superposition states composed of an exponentially large number of quantum gates. Candidate systems for bucket-brigade QRAMs include optical lattices [17,18], Josephson arrays [19], arrays of coherently coupled quantum dots, or strongly correlated cavity arrays [20]. To be more specific on the nature of the necessary resources, we present a proof-of-principle implementation of the quantum bucket brigade. (It should be considered only as an illustrative example, and not as an experimental proposal. More detailed versions of bucket-brigade implementations will be presented in future work.) The qutrits at the nodes of the graph of Fig. 1 are composed of trapped atoms or ions with the level structure depicted in Fig. 3: a ground state $|\text{wait}\rangle$ and two excited states $|\text{left}\rangle$ and $|\text{right}\rangle$. The register and bus qubits are composed of photons, whose encoding is in the polarization. It is now possible to use a photon in the polarization state $|0\rangle$ to muster a $|\text{wait}\rangle \to |\text{left}\rangle$ atomic transition, and a photon in the polarization state $|1\rangle$ to muster a $|\text{wait}\rangle \to |\text{right}\rangle$ transition. Furthermore, by employing Raman techniques, one uses strong classical pulses that couple $|\text{wait}\rangle$, $|\text{left}\rangle$, and $|\text{right}\rangle$ with extra energy levels (not shown in the picture) to

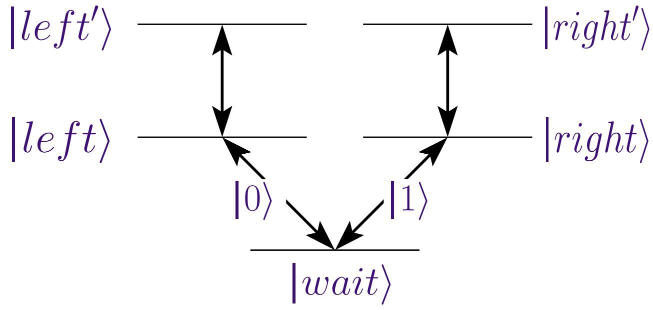

FIG. 3 (color online). Basic level structure of the atoms in a possible bucket-brigade implementation. Some extra energy levels needed to implement Raman transitions are not shown.

externally control the timing of such transitions. Note that, being classical, such pulses do not need to act locally on a single atom but they can interact with all the nodes of each level. Thus, a photon impinging on an atom in the $|\text{wait}\rangle$ state transfers its internal state to the $|\text{left}\rangle$-$|\text{right}\rangle$ atomic levels. A photon impinging on an atom which is in a $|\text{left}\rangle$ state, will excite a cyclic transition (using the level $|\text{left}\rangle$) and is reemitted by the atom. The $|\text{left}\rangle \rightarrow |\text{left}'\rangle$ transition is insensitive to the photon's polarization and is coupled to an outgoing spatial mode departing the trapped atom in the left direction. This means that a photon in any polarization state that impinges onto an atom in the $|\text{left}\rangle$ state is deviated along the graph towards the left. Analogously, a photon in any state impinging on an atom in the $|\text{right}\rangle$ state is deviated towards the right. As in the $|\text{wait}\rangle \rightarrow |\text{left, right}\rangle$ transition, the timing of the whole process can be controlled by coupling the involved states with ancillary levels through strong classical Raman pulses. After all the photons of the address register are sent through the graph, a bus photon (initially in the state $|0\rangle$) is injected. Thanks to the above mechanism, it crosses the graph in a coherent superposition of paths, exiting at the location of the addressed cells and changing its polarization state according to their memory content. It is then reflected back through the graph and is again deflected interacting with the atoms, so that it exits the graph at the root node. To end the protocol, the Raman process is inverted, step by step, starting from the last level in the graph, so that the atomic levels $|\text{left}\rangle$ and $|\text{right}\rangle$ are driven to the $|\text{wait}\rangle$ level, through the emission of a $|0\rangle$ or $|1\rangle$ photon, respectively. Thus the address register photons are emitted one-by-one and coherently driven back through the graph to the root node.

Conclusions.—We have described a RAM architecture where active gates are replaced by three-level memory elements. It could give rise to a significant simplification in the QRAM implementation, to exponentially reduced decoherence rate and energy saving. However, in current RAMs, the primary sources of dissipation are leakage current in the memory cells (for static RAMs) and refreshing memory cells (for dynamic RAMs). Energy costs in the memory access procedure are not currently important enough to warrant accepting the additional delays and memory elements of the bucket brigade. For future, non-CMOS RAMS, however, decoding energy costs may become important, so that the exponential savings of the bucket-brigade architecture may prove significant.

We acknowledge useful feedback from A. Childs and support from Jeffrey Epstein.

[1] R. Feynman, *Feynman Lectures on Computation* (Perseus Books Group, New York, 2000).

[2] R.C. Jaeger and T.N. Blalock, *Microelectronic Circuit Design* (McGraw-Hill, Dubuque, 2003), p. 545.

[3] M.A. Nielsen and I.L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).

[4] G. Schaller and R. Schützhold, Phys. Rev. A **74**, 012303 (2006).

[5] R. Schützhold, Phys. Rev. A **67**, 062311 (2003).

[6] C.A. Trugenberger, Phys. Rev. Lett. **87**, 067901 (2001); **89**, 277903 (2002).

[7] G. Brassard, P. Høyer, and A. Tapp, SIGACT News **28**, 14 (1997).

[8] A. Ambainis, in *Proceedings of the 45th IEEE FOCS'04* (IEEE Computer Society, Rome, 2004), p. 22; arXiv:quant-ph/0311001.

[9] A.M. Childs, A.W. Harrow, and P. Wocjan, in *Proceedings of the 24th Symposium on Theoretical Aspects of Computer Science (STACS 2007)*, Lecture Notes in Computer Science (Springer, New York, 2007), Vol. 4393, p. 598; arXiv:quant-ph/0609110.

[10] A.M. Childs *et al.*, in Proceedings of the 48th IEEE Symposium on Foundations of Computer Science (FOCS'07) (to be published); arXiv:quant-ph/0703015.

[11] V. Giovannetti, S. Lloyd, and L. Maccone, arXiv:quant-ph/0708.2992 (to be published).

[12] V. Giovannetti, S. Lloyd, and L. Maccone (to be published).

[13] The d-dimensional RAM consists of d such graphs, each addressing one side of a $N^{1/d} \times N^{1/d} \times \cdots$ array.

[14] We focus on the case in which we "read" from the memory, the "write" operation being completely analogous.

[15] C.W. Chou *et al.*, Nature (London) **438**, 828 (2005); J.F. Sherson *et al.*, Nature (London) **443**, 557 (2006).

[16] R. Raussendorf and H.J. Briegel, Phys. Rev. Lett. **86**, 5188 (2001).

[17] E. Jané *et al.*, Quantum Inf. Comput. **3**, 15 (2003).

[18] L.-M. Duan, E. Demler, and M.D. Lukin, Phys. Rev. Lett. **91**, 090402 (2003).

[19] A. Romito, R. Fazio, and C. Bruder, Phys. Rev. B **71**, 100501(R) (2005).

[20] M.J. Hartmann, F.G.S.L. Brand, and M.B. Plenio, Nature Phys. **2**, 849 (2006).